SCALAR LIMITS AND GEOMETRIC RECOVERY IN SOFTMAX GRADIENT BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Softmax gradient bandits can converge to the best arm, but convergence alone does not determine the cumulative reward lost during learning. This leaves open whether adaptive learning rates and reward baselines can achieve square-root worst-case regret, or whether the update structure must change. The analysis addresses this question for bounded, independent and identically distributed rewards with a unique best arm. The analysis identifies a coupling in the mean update: a common scalar rescales corrective and competing drift terms together, while the baseline cancels from the conditional mean. Beyond this local mechanism, matching lower and upper polynomial exponents characterize the entire bounded scalar class, including history- and horizon-dependent learning rates and common baselines with nonnegative steps. For each fixed number of arms $K \geq 3$, its minimax regret is of order $T^{1-1/K}$ up to logarithmic factors, uniformly over an explicit neighborhood of biased initial logits. To change this coupling, an explicit diagonal update reweights coordinates while preserving softmax sampling and single-reward feedback. It achieves square-root worst-case regret up to logarithmic factors and polylogarithmic fixed-gap regret from every fixed finite initialization, without gap or horizon inputs. These guarantees establish a polynomial separation on the common initialization domain: changing coordinate weights achieves a regret exponent unavailable through bounded scalar tuning alone.

1 INTRODUCTION

Convergence to the best arm does not determine the reward lost while learning it. Softmax gradient bandits update action preferences, or logits, using only the chosen arm's reward. Existing convergence and regret guarantees analyze important rules (Mei et al., 2023; 2024; Robertson et al., 2025; Baudry et al., 2025), but leave a broader question: can adapting the learning rate and reward baseline repair poor worst-case regret, or does the update structure itself impose a limit? Regret measures cumulative expected reward lost relative to always choosing the best arm.

A lower bound for one schedule may indicate poor tuning; a bound for the whole class identifies what tuning cannot repair. The original update scales the selected action's log-probability gradient, or score, without reweighting its coordinates. The question is what coordinate weighting gains under the same softmax sampling and single-reward feedback.

An inferior third arm can reinforce a nearly optimal rival's lead in conditional mean. A nonnegative scalar scales this effect and the opposing gap correction together; the baseline cancels in mean. The class bound must also cover waiting and history adaptation.

Scalar limit. For fixed $K \geq 3$, the entire bounded scalar class has minimax regret of order $T^{1-1/K}$ up to logarithmic factors, uniformly on the biased-logit neighborhood $\mathcal{N}$ (Theorem 1). For $K = 2$, every fixed finite start has rate $\Theta(\sqrt{T})$.

Geometric recovery. An explicit anytime diagonal gradient bandit (DGB) achieves square-root-polylog worst-case regret and polylogarithmic fixed-gap regret from every fixed finite start, without gap or horizon inputs (Theorem 3). On $\mathcal{N}$, it improves the exponent to $1/2$ while retaining the sampler, feedback, and scalar-step and baseline bounds (Corollary 4).

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.

Theorems 1 and 3 establish the core comparison. Two supporting results describe its scope and cost. For $K \geq 3$, every fixed *anytime* scalar rule, from each fixed finite start including uniform logits, has a worst-case regret lower bound with polynomial exponent above $1/2$; the exponent and constants may depend on the rule and initialization (Theorem 2). Section 5 studies the exact fixed-instance expected regret coefficient and the cost of finite geometric activity.

Finite-horizon experiments identify conditions under which a fixed DGB design improves on thirteen scalar implementations, then test persistence and boundary failures. A separate allocation decomposition explains how best-arm gains can be offset by outside-arm sampling costs (Section 6).

Related work and the missing class comparison

From a specified rule to an adaptive class. The closest regret comparisons differ first in which controllers they cover. Baudry et al. (2025) establish polynomial regret for some constant-step rules (their Theorem 3), alongside near-square-root two-arm guarantees and an equal-suboptimal-gap upper bound (their Lemma 15). Lattimore (2026b) analyzes a fixed step from zero logits with gap- and horizon-dependent restrictions. The present lower bound instead covers vanishing and history-dependent steps, unknown gaps and adaptive baselines.

A separate distinction concerns dimension and feedback. The wrong-rival mechanism is related to the diffusion analysis of Lattimore (2026a), whose lower bound requires $K \gtrsim \log(T/\Delta)$, and the exact-advantage, full-action updates of Mei & Osband (2026). Here K is fixed and feedback is sampled.

Efficiency through a different update interface. Fast bandit regret is already possible outside the scalar class. Efficient probability-space alternatives include SAMBA, Tsallis-INF and TS-Prod (Walton & Denisov, 2023; Zimmert & Seldin, 2021; Zimmert & Marinov, 2024). Alternative parameterizations (Labbi et al., 2026) and arm-specific Boltzmann scales (Cesa-Bianchi et al., 2017) also change how actions are updated or selected. The contribution here is a finite-step potential and running-cap analysis for a specified softmax logit update: one process has both gap-uniform and fixed-gap guarantees, enabling the class comparison above. Appendix N details these differences in updates, feedback and guarantees.

2 Preliminaries and problem formulation

Bandit interaction and regret. There are $K \geq 2$ arms. Arm i supplies an independent pull-indexed IID stream $Y_n^{(i)} \in [0,1]$ with unknown mean μ_i. At round t, the agent samples an arm I_t from its probability vector p_t and observes only that arm's next unused reward Y_t. The best arm j is unique; write $\Delta_i = \mu_j - \mu_i$, so $\Delta_j = 0$ and all other gaps are positive. Cumulative regret measures the reward lost by sampling other arms: realized pseudo-regret and its unconditional expectation are

$$\widehat{R}_T = \sum_{t=1}^{T} \Delta_{I_t}, \qquad R_T = \mathbb{E}\widehat{R}_T = \sum_{t=1}^{T} \sum_{i \neq j} \Delta_i \mathbb{E} p_{i,t}.$$

Softmax sampling and score updates. The policy is parameterized by logits $\theta_t \in \mathbb{R}^K$, initialized at a finite deterministic $\theta_1 = \vartheta$. Softmax gives positive sampling probabilities, and its log-probability gradient, or *score*, is

$$p_i(\theta) = \frac{e^{\theta_i}}{\sum_k e^{\theta_k}}, \qquad \nabla_\theta \log p_i(\theta) = e_i - p(\theta),$$

Here e_i is the ith standard basis vector. A scalar gradient-bandit update multiplies the score by the centered reward $Y_t - b_t$ and a learning rate η_t shared by all coordinates; b_t is a common baseline.

The scalar controller class. Let $\mathcal{F}_t$ contain the history before the current action, including controller randomization. Fix $u > 0$ and $B_{\max} \geq 1$, independently of the unknown instance and T. The original-score interface is

$$I_t \mid \mathcal{F}_t \sim p_t, \qquad \theta_{t+1} = \theta_t + \eta_t (Y_t - b_t)(e_{I_t} - p_t), \quad 0 \leq \eta_t \leq u, \quad |b_t| \leq B_{\max}. \tag{1}$$

Both controls are $\mathcal{F}_t$-measurable. They may use the entire past and internal randomization, but not the current action or reward. Fresh action and unused reward randomness retain their conditional product law. Controls may know the evaluation horizon, but not the reward law, gaps or best-arm label; the same history-to-control map applies across unknown environments. Zero and vanishing steps are allowed. The agent cannot reset logits, force an arm, change the sampler or add a direction to the score. An *anytime* rule is one infinite history-to-control process without the evaluation horizon T as an input; it may depend on the current round t and the observed history.

A diagonal alternative. Fix a round t and suppress its subscript. The expected reward is $J(\theta) = \sum_i \mu_i p_i(\theta)$, with gradient $g_i = p_i(\mu_i - J)$. Since the baseline is chosen before the action and $\mathbb{E}[e_I - p \mid \mathcal{F}_t] = 0$,

$$\mathbb{E}[(Y - b)(e_I - p) \mid \mathcal{F}_t] = g.$$

The scalar mean increment is ηg. Replacing the score by $D_\gamma(p)(e_I - p)$, with $D_\gamma(p) = \mathrm{diag}(p_i^{-\gamma})$, instead gives $\eta D_\gamma(p) g$. Section 4 fixes $\gamma = 1/2$; the supplement treats $0 < \gamma \le 1/2$. For $\eta > 0$ and $g \neq 0$, this is an ascent direction: $g^\top \eta D_\gamma(p) g = \eta \sum_i p_i^{-\gamma} g_i^2 > 0$. The weights amplify rarely sampled arms while retaining the sampler; Section 4 controls finite random steps to obtain regret guarantees. Shared scalar-step and baseline bounds do not equalize update norms, covariance or computational cost.

The uniform comparison. Let $\mathfrak{E}_K$ be the IID instance class above and $\mathcal{C}_0^T(u, B_{\max}; \vartheta)$ denote the scalar rules, where the superscript permits horizon knowledge. The minimax value is

$$\mathcal{M}_0(T; \vartheta) = \inf_{\mathcal{A} \in \mathcal{C}_0^T(u, B_{\max}; \vartheta)} \sup_{\mathcal{E} \in \mathfrak{E}_K} R_T^{\mathcal{A}}(\mathcal{E}). \tag{2}$$

The infimum already selects the best permitted controller, rather than a particular schedule. The supremum may select a different fixed IID instance at each T. We fix $K, u, B_{\max}, \vartheta$ before taking $T \to \infty$, write $L_t = 1 + \log t$, and suppress fixed parameters when unambiguous.

Table 1: Initialization domains and controller quantifiers. Resources and K are fixed. Exponents concern worst-case expected regret; logarithms are suppressed here.

Result	Initialization	Controller quantifier	Exponent
Scalar, $K \ge 3$	All starts in biased neighborhood $\mathcal{N}$	Entire horizon-aware class	$1 - 1/K$
DGB	Any fixed finite	One explicit anytime rule	$1/2$
Scalar, $K = 2$	Any fixed finite	Entire horizon-aware class	$1/2$
Scalar extension, $K \ge 3$	Any fixed finite	Fix one anytime rule	Lower bound: rule-dependent $> 1/2$

The scalar $K \ge 3$ and DGB results in Table 1 give the core separation on $\mathcal{N}$: one explicit anytime diagonal rule improves on the entire horizon-aware scalar class's regret exponent. The anytime scalar extension is not a premise of this comparison.

3 Why scalar adaptivity has a dimension-dependent limit

A third arm can favor the wrong competitor. Consider Bernoulli means $\mu_1 = 1/2 + \delta$, $\mu_2 = 1/2$ and $\mu_k = 1/4 - (k-3)/(16K)$ for $k = 3, \dots, K$, with $0 < \delta \le 1/32$. Arm 1 is best, arm 2 is its close competitor, and the others are outside arms. At a fixed round, suppress the subscript t on policy coordinates. Put $X = \theta_2 - \theta_1$ and $r_o = \sum_{k \ge 3} (1/2 - \mu_k) p_k$. The exact conditional drift is

$$\mathbb{E}[X_{t+1} - X_t \mid \mathcal{F}_t] = \eta_t \{ \underbrace{(p_2 - p_1) r_o}_{\text{outside feedback}} - \underbrace{\delta p_1 (1 + p_2 - p_1)}_{\text{true-gap correction}} \}.$$

If the competitor is favored, the first term can reinforce its lead even though the second term points toward the best arm. The common scalar multiplies both. A baseline changes the random increments but cancels from this mean. Indeed the mean increment of coordinate i is $\eta_t p_i(\mu_i - J)$: inferior outside arms lower the policy's expected reward J, so the more probable rival can gain more than

the best arm. At $K = 3$, $\theta = (0, 1, 0)$ gives $p = (1, e, 1)/(e + 2)$ and drift $\eta_t\{(e - 1)/4 - \delta(2e + 1)\}/(e + 2)^2$, whose bracket is positive for $0 < \delta \leq 1/32$. Thus every positive scalar gives positive conditional drift at this state; setting the scalar to zero stops, but does not reverse, that drift (Figure 1). This is a statement about the conditional mean, not every sample.

The class-level result. The next result turns this local mechanism into a uniform class obstruction, covering controllers that wait or use noise to escape.

Theorem 1 (Scalar minimax bounds; Appendix Theorem A.1)**.** *Under Section 2, fix $u > 0$, $B_{\max} \geq 1$ and K. For $K = 2$, $\mathcal{M}_0(T) = \Theta(\sqrt{T})$ at every fixed finite initialization. For $K \geq 3$, let $\mathcal{N} = \{\vartheta : \|\vartheta - (0, 1, 0, \ldots, 0)\|_\infty \leq 1/8\}$. There are $c_- > 0$, $C_+, T_0 < \infty$, depending only on $K, u, B_{\max}$, such that simultaneously for every $\vartheta \in \mathcal{N}$ and $T \geq T_0$,*

$$c_- T^{1-1/K} / L_T^{3-2/K} \leq \mathcal{M}_0(T; \vartheta) \leq C_+ T^{1-1/K}.$$

Both bounds have the same polynomial exponent, uniformly on $\mathcal{N}$. The closed neighborhood has nonempty full-dimensional interior and permits independent logit perturbations. Its center favors arm 2 with probability $e/(e + K - 1)$ (0.576 at $K = 3$); it excludes uniform logits. Fixed-instance logarithmic regret with tuning or additional moment conditions is compatible with this result: the worst-case gap can shrink with T, whereas pointwise constants need not be uniform.

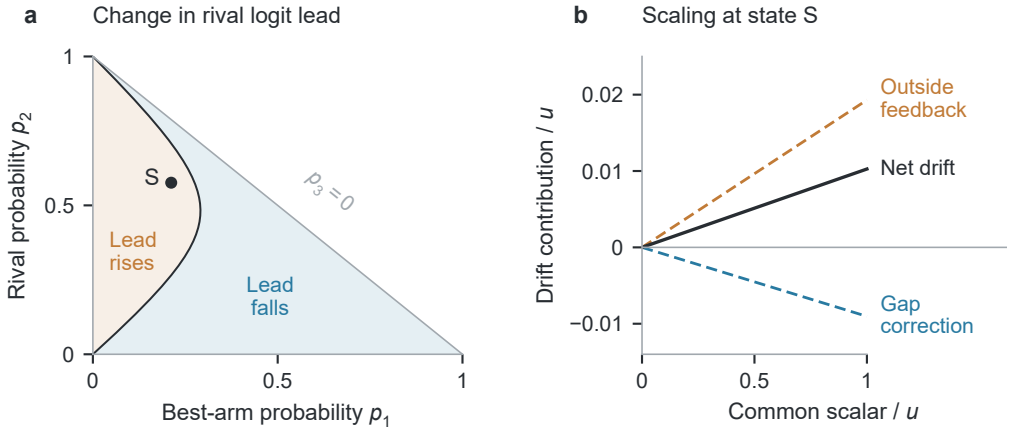


Figure 1: Competing drift terms under a common scalar. (a) Sign of the conditional rival-minus-best logit drift for means $(17/32, 1/2, 1/4)$. The curved boundary is zero drift; the diagonal edge has zero third-arm probability. The regions show whether the rival's logit lead increases or decreases, not reward gains. (b) At $S = (1, e, 1)/(e + 2)$, varying η/u scales the positive outside feedback and negative gap correction together. The net remains positive for every $\eta > 0$; zero stops both terms. These are exact local mean relationships, not learning trajectories or a class-level proof.

Proof idea: costly occupation and limited information. The task is to lower-bound costly time despite waiting or noise-driven escape. Fix the gap δ_T, to be chosen below. Under the real IID law $\mathbb{P}$, the best arm has probability below $1/3$ until the history-defined exit time τ constructed in Appendix A, giving $R_T \geq (2\delta_T/3)\mathbb{E}_{\mathbb{P}}[T \wedge \tau]$.

Step 1. Establish persistence under an auxiliary law. For this lower bound, introduce $\mathbb{Q}$: it agrees with $\mathbb{P}$ until entry into a low-best-arm-probability region (cold entry), then changes only the best arm's rewards. This proof device is not the theorem's IID instance. Stopped phase potentials and a uniform tail bound yield $\mathbb{Q}(\tau > T) \gtrsim L_T^{-2}$ for sufficiently large T. Thus $\mathbb{E}_{\mathbb{Q}}[T \wedge \tau] \gtrsim T L_T^{-2}$. This covers waiting and multiple crossings in one update (Appendix A); next, transfer occupation to the real law.

Step 2. Transfer occupation, not only a rare event. To retain a comparable duration under $\mathbb{P}$, write $q_{\mathrm{occ}} = \mathbb{E}_{\mathbb{Q}}[T \wedge \tau]/T$ and $p_{\mathrm{occ}} = \mathbb{E}_{\mathbb{P}}[T \wedge \tau]/T$. These are normalized expected occupation times, not arm probabilities. The laws share the initial state and stopping rule. If their per-round conditional KL in direction $\mathbb{Q}\|\mathbb{P}$, before exit, is at most ϵ_{KL}, Lemma A.6 gives

$$\mathrm{kl}(q_{\mathrm{occ}}, p_{\mathrm{occ}}) \le \epsilon_{\mathrm{KL}} T q_{\mathrm{occ}}, \qquad \epsilon_{\mathrm{KL}} T \le 1/8 \implies p_{\mathrm{occ}} \ge q_{\mathrm{occ}}/2.$$

Scaling information with occupation avoids a fixed error overwhelming rare survival. The next step makes this budget small enough.

Step 3. Balance regret cost against distinguishability. The scalar update preserves the sum of logits. Together with the controlled logit contrasts before exit, this links the best arm's probability to the $K-2$ outside coordinates. After cold entry, it yields $p_1 \lesssim L_T^{K-2}\delta^{\kappa}$, with $\kappa = K - 2 - O(L_T^{-1})$ as specified in Appendix A. Only the best arm's reward law changes, and its conditional reward KL is $O(\delta^2)$. Weighting by its sampling probability gives $\epsilon_{\mathrm{KL}} T \lesssim T L_T^{K-2}\delta^{\kappa+2}$. Thus $\delta_T = d_* T^{-1/K} L_T^{-(K-2)/K}$ keeps this budget bounded: the $O(L_T^{-1})$ exponent correction contributes only a bounded factor. One sufficiently small fixed d_* makes the budget at most $1/8$ (Appendix A). Combining per-round cost, occupation transfer and survival gives

$$R_T \gtrsim \delta_T T L_T^{-2} = d_* T^{1-1/K} / L_T^{3-2/K}.$$

The same constant and onset cover all permitted controllers and $\vartheta \in \mathcal{N}$, without continuity of their control maps. Each horizon selects one fixed IID instance; it does not change during a run.

Why the exponent is attainable inside the scalar class. For $K \ge 3$, the upper bound balances two costs. For analysis only, order the means from largest to smallest. If every adjacent gap is below ε, all arms are within $(K-1)\varepsilon$ of the best, so $O(T\varepsilon)$ already suffices. Otherwise, form the top group above the first adjacent gap of size at least ε. Its within-group loss is $O(T\varepsilon)$. To control sampling outside the group, we adapt the subset-exponential potential of Jia & Ouyang (2026), equation (65), replacing the best-arm logit by the group's mean logit. Its drift controls outside-group pulls weighted by their individual gaps, rather than dividing by the smallest gap. The potential starts at $O(\varepsilon^{-(K-2)})$ uniformly on $\mathcal{N}$; with step $\eta = \min\{u, 1/(8eK)\}\varepsilon$, its regret contribution is $O(\varepsilon^{-(K-1)})$. Balancing gives

$$T\varepsilon \asymp \varepsilon^{-(K-1)}, \qquad \varepsilon = T^{-1/K}, \qquad R_T = O(T^{1-1/K}).$$

The remaining additive $O(T^{1/K}\log T)$ term is absorbed for fixed $K \ge 3$ (Appendix A.3). The rule uses $b_t = 0$ and the stated step, without receiving the group or means. For $K = 2$, empirical-gap testing commits to a scalar, not an arm: natural sampling continues even when the scalar is zero. Testing and safe-update bounds give $O(\sqrt{T})$. Both upper rules may know T.

Scope extension: arbitrary finite initialization. The next result covers every finite start, with a different quantifier order.

Theorem 2 (Every finite initialization; Appendix Theorem D.1)**.** *Fix $K \ge 3$, resources $u, B_{\max}$, finite ϑ, and one anytime rule $\mathcal{A}$ in the original-score class. There is a fixed arm relabeling of the Bernoulli family above (now $0 < \delta \le 1/8$), and $c_{\mathcal{A}} > 0$, $T_{\mathcal{A}} < \infty$, $\zeta_{\mathcal{A}} > 1/2$, such that*

$$\sup_{0<\delta\le 1/8} R_T^{\mathcal{A}}(\delta) \ge c_{\mathcal{A}} T^{\zeta_{\mathcal{A}}} \quad (T \ge T_{\mathcal{A}}).$$

The constants and exponent may depend on the rule and initialization, but are independent of the subsequently varied δ, T.

This rules out square-root-polylog worst-case regret from uniform logits for each fixed anytime rule, without a common margin above $1/2$ across rules. The relabeling stays fixed as δ, T vary. At a uniform policy, the contrast drift is $-\eta_t\delta/K \le 0$, so the biased-state mechanism cannot prove this extension. Appendix D gives the separate argument, including rules that never move. Theorem 1, not this extension, supplies the sharp horizon-aware exponent used below.

4 RECOVERING THE SQUARE-ROOT EXPONENT BY DIAGONAL SCALING

For DGB at $\gamma = 1/2$, inverse square-root weights change the mean increment per unit scalar from $p_i(\mu_i - J)$ to $\sqrt{p_i}(\mu_i - J)$, partly offsetting suppression of rarely sampled actions.

To control the resulting large weights at small probabilities, the rule remembers the largest amplification measure seen so far:

$$C_{1/2}(p) = \Big(\sum_i \sqrt{p_i}\Big)\Big(\sum_i p_i^{-1/2}\Big), \qquad \overline{C}_t = \max_{s \le t} C_{1/2}(p_s). \tag{3}$$

Algorithm 1 combines this cap with a fixed step scale $0 < h \le u$ and logarithmic decay exponent $1/2 < \omega \le 1$. The running cap limits amplification; the time factor controls accumulated noise. The baseline $b = 1$ makes $Y - 1 \le 0$, a sign used to control cap increments. This baseline is allowed in both classes. Since $C_{1/2}(p) \ge p_{\min}^{-1/2}$, $\eta_t p_{\min,t}^{-1/2} \le h L_t^{-\omega} \le h$. Thus the rule respects $0 < \eta_t \le u$, has bounded coordinate steps, and preserves finite logits and full support at every finite time. Cauchy–Schwarz also gives $C_{1/2}(p) \ge (\sum_i p_i^{1/4} p_i^{-1/4})^2 = K^2$, with equality only at the uniform policy. Thus $\eta_1 \le h/(1 + K^2)$: the cap already constrains the initial step.

Algorithm 1: DGB at $\gamma = 1/2$.

Input: finite θ_1, $1/2 < \omega \le 1$, $0 < h \le u$; set $\overline{C}_0 = 0$.
For $t = 1, 2, \ldots$:
1. Compute $p_t = \mathrm{softmax}(\theta_t)$ and $\overline{C}_t = \max\{\overline{C}_{t-1}, C_{1/2}(p_t)\}$.
2. Set $\eta_t = h/[L_t^{\omega}(\sqrt{t} + \overline{C}_t)]$.
3. Sample $I_t \sim p_t$ and observe its next unused reward Y_t.
4. Update $\theta_{t+1} = \theta_t + \eta_t(Y_t - 1)D_{1/2}(p_t)(e_{I_t} - p_t)$.

Theorem 3 (Geometric recovery; Theorem B.1 at $\gamma = 1/2$)**.** *Fix $K \ge 2$, $1/2 < \omega \le 1$, $u > 0$, $0 < h \le u$ and finite deterministic ϑ. Under Section 2, Algorithm 1 satisfies, on every permitted instance and for every $T \ge 1$,*

$$R_T \le C_{\mathrm{init}} + 2C_{\mathrm{rate}}\sqrt{(K-1)T}\,L_T^{\omega}, \tag{4}$$

$$R_T \le 2C_{\mathrm{init}} + C_{\mathrm{rate}}^2 L_T^{2\omega+1} \sum_{i \ne j} \Delta_i^{-1}. \tag{5}$$

The constants in Appendix B.2, evaluated at $\gamma = 1/2$, depend only on K, ω, h, ϑ, not on gaps, variances, T or an unproved recovery time.

C_{init} pays for initialization and C_{rate} for drift and cap control. One process achieves both bounds for every fixed $0 < h \le u$.

Corollary 4 (Polynomial separation; Appendix Corollary C.1)**.** *For fixed $K \ge 3$ and the neighborhood $\mathcal{N}$ of Theorem 1, the same DGB parameters satisfy, for all sufficiently large T,*

$$\inf_{\vartheta \in \mathcal{N}} \frac{\mathcal{M}_0(T;\vartheta)}{\sup_{\mathcal{E}} R_T^{\mathrm{DGB},\vartheta}(\mathcal{E})} \ge c\frac{T^{1/2-1/K}}{L_T^{3-2/K+\omega}} \longrightarrow \infty.$$

Indeed $p_{\min,1} := \min_i p_{i,1} \ge e^{-5/4}/K$ on $\mathcal{N}$, giving common finite bounds for the initial potentials. Combining Theorems 1 and 3 proves the claim. Together with the statistical bandit lower bound, this restores the optimal fixed-K polynomial exponent. For $K = 2$, the diagonal process is policy-equivalent to a bounded original-score rule (Appendix C.4); this separation starts at three arms.

Proof idea: compare policies while paying for the running cap. *Step 1. Turn policy progress into a drift inequality.* For analysis, let j be best and use the nonnegative potential $W_j(p) = \sum_{i \ne j} \sqrt{p_i/p_j}$. It measures suboptimal mass relative to the best arm. Write $Z = \sum_i e^{\theta_i}$ and $\Phi = \sum_i e^{\theta_i/2}$, so $1 + W_j = e^{-\theta_j/2}\Phi$. For the actual increment v,

$$\nabla\Phi(\theta)^{\top} v = \tfrac{1}{2}\eta(Y-1)\sqrt{Z}\sum_i (\mathbf{1}\{I = i\} - p_i) = 0.$$

The identity connects this cancellation to the best-arm coordinate in W_j; Φ itself need not remain constant after a finite update. Controlling the finite-step remainder and reward-weighted score square gives

$$\mathbb{E}_t W_{j,t+1} \le (1 + d_1\eta_t^2)W_{j,t} - c_{1/2}\eta_t r_t, \qquad r_t = \sum_{i\neq j} \Delta_i p_{i,t}.$$

Here $\mathbb{E}_t$ conditions on $\mathcal{F}_t$; positive constants are specified in Appendix B.2. This controls step-weighted regret, not yet regret per observed round.

Step 2. Pay for the random inverse step. Since $\eta_t^{-1} = L_t^{\omega}(\sqrt{t} + \overline{C}_t)/h$, dividing the conditional drift bound by the predictable step introduces $w_t = 1/\eta_t$. The summation-by-parts identity is

$$\sum_{t=1}^{T} w_t(W_{j,t} - W_{j,t+1}) = w_1 W_{j,1} + \sum_{t=2}^{T}(w_t - w_{t-1})W_{j,t} - w_T W_{j,T+1}.$$

The middle term couples inverse-step changes and the potential through the same history. Square summability of η_t and the baseline-one sign yield fourth-moment bounds, controlling the same-time cap–potential mixed term without independence. The correlated cap-increment term is handled separately: the conditional positive-increment bound and the pathwise estimate $W_{j,t+1} \le e^h W_{j,t}$ control its product with the next potential (Appendix B.6). Together with the deterministic time-factor changes, these estimates give the actual-time master bound

$$R_T \le C_{\text{init}} + C_{\text{rate}} \sum_{t=1}^{T} \frac{L_t^{\omega}}{\sqrt{t}} \sum_{i\neq j} \sqrt{\mathbb{E}p_{i,t}}. \tag{6}$$

Step 3. Read both rates from the same inequality. Cauchy–Schwarz applied to (6) gives (4). Taking twice (6) minus the regret identity and applying $2a\sqrt{x} - \Delta x \le a^2/\Delta$ for each arm and time gives (5). The full finite-step and cap-moment proof is in Appendix B.

5 Exact fixed-instance cost and the endpoint choice

The diagonal rule changes the worst-case exponent. Its fixed-instance analysis also identifies the price of controlling amplified updates. For every fixed positive-gap instance and fixed design of Algorithm 1, Theorem F.1 gives

$$\lim_{T\to\infty} \frac{R_T}{L_T^{2\omega+1}} = \frac{1}{(2\omega+1)h^2} \sum_{i\neq j} \Delta_i^{-1}. \tag{7}$$

This exact *expected* coefficient resolves the dependence on gaps, the step scale h, and the logarithmic decay ω. The proof first identifies pathwise allocations, then controls all fixed integer moments of the potential. A fourth-moment bound yields uniform integrability, which transfers the pathwise limit to expectation (Appendix F). Corollary F.4 collects the endpoint clock and allocation formulas; Theorem F.1 supplies the exact expectation limit.

Why the half endpoint? Within the full $0 < \gamma \le 1/2$ family of Appendix B, $\gamma = 1/2$ minimizes the fixed-instance coefficient at the same h and ω: strictly for $K \ge 3$, with equality across γ for $K = 2$ (Corollary F.3). This supplies a design rationale for the endpoint in Algorithm 1. At this endpoint the family's potentials reduce to W_j, and the step needs no tail-mass correction (Appendix B.1). The comparison concerns fixed parameters as $T \to \infty$, not a finite-horizon optimizer.

The logarithmic cost of finite activity. Define $\mathcal{E}_\gamma = \sum_t \eta_t^2 a_\gamma(p_t)$, where $a_\gamma(p) = \sum_i p_i^{1-2\gamma}(1 - p_i)$ is a predictable score-square budget, not an exact covariance. For Algorithm 1, $\mathcal{E}_{1/2} = (K-1)\sum_t \eta_t^2 < \infty$. Theorem E.1 shows that any one horizon-free rule in this interface, with fixed resources and finite initialization, has $\liminf_T R_T/L_T^2 = \infty$ on each permitted instance with $\mathbb{P}(\mathcal{E}_\gamma < \infty) > 0$. Thus the extra logarithmic cost is structural for the finite-activity subclass; (7) quantifies it for this process.

6 EXPERIMENTS: CONDITIONAL GAINS AND THEIR COSTS

The experiments ask where a fixed DGB design improves finite-horizon regret, whether that improvement persists, and which costs can reverse it. The fixed-design study uses $u = 16$; a separate legacy $u = 1$ study accounts for allocation gains and costs. Their cohorts are kept separate.

A fixed design and its comparison set. The new study fixes $(h, \omega, \gamma) = (8, 0.51, 1/2)$ and compares it with 13 locked original-score scalar implementations under $u = 16$, $B_{\max} = 1$. These span step schedules, adaptive baselines, norm matching, and history-dependent controls; their definitions are in Appendix G. They test specific alternatives, not an empirical optimizer over the scalar class. Figure 2 uses S2 (constant step $\eta = 0.3$, $b = 0.5$) and S10 (confidence-scaled step) as scalar references. UCB1 and Thompson sampling provide task-level references. The representative Bernoulli instance has $K = 5$, $\mu = (0.625, 0.5, 0.05, 0.05, 0.05)$ and $\vartheta = (0, 3, 0, 0, 0)$. It tests DGB away from the scalar lower-bound neighborhood $\mathcal{N}$; it does not enlarge that theorem's domain.

Where does the fixed design help? At $T = 65{,}536$, 4,096 paired runs give DGB mean regret 808.74, against 2401.27 for the lowest of the 13 scalar sample means, a 66.32% reduction. All 13 mean savings have positive lower bounds in the specified simultaneous 95% confidence family; the smallest is 1232.04. Across a separate 192-run-per-environment scan, DGB has lower sample means than all 13 controls in all 47 target environments, but in only three of six boundary environments. The scan maps the empirical range of the effect: simultaneous lower bounds against all 13 are positive in only two of the 47 target environments. Figure 2 shows absolute trajectories at the anchor and a boundary failure; the full scan, intervals and numerical protocol are reported in Appendix G.

A separate interval-validated anchor cohort checks each sampled action against the real-arithmetic rule using interval arithmetic. It gives at least 43.2% lower expected regret than both S2 and S10 at $T = 65{,}536$, on one 97.5% confidence event. Appendix G.7 explains the validated sampling decisions and reports all four methods in this 1,024-run-per-method cohort.

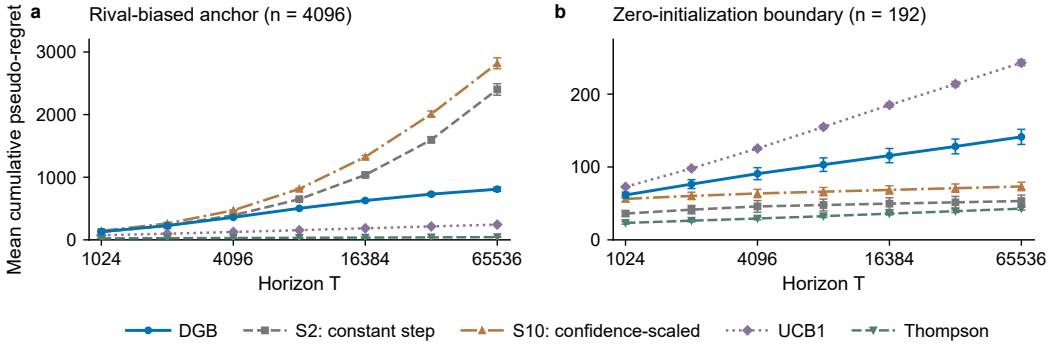


Figure 2: Conditional finite-horizon gains of one fixed DGB design. Means with pointwise approximate 95% t error bars; lines join recorded checkpoints. (a) Biased anchor, $n = 4096$. (b) Same means, zero initial logits, $n = 192$: selected post hoc as the largest terminal relative mean loss against the best scalar sample mean among six boundary configurations. The panels have different vertical scales. Methods are paired within each cohort; S2 and S10 are the scalar controls defined above. Appendix G gives all configurations and the separate simultaneous mean comparisons.

Do the gains persist at longer horizons? At $T = 1{,}048{,}576$, a matched 4,096-run cohort gives mean regret 1004.73 for DGB versus 10471.68 and 5585.76 for the original constant-step scalar S2 and confidence-scaled S10: reductions of 90.41% and 82.01%. The prescribed joint confidence calculation and monotone endpoint bounds give at least 45.38% improvement in expected regret of the stated numerical process against both on every integer $T \in [65{,}536, 1{,}048{,}576]$ on one 95% event. In a separate 512-run reference cohort at the final horizon, DGB, UCB1 and Thompson have means 854.57, 345.98 and 58.14 respectively. Thus the gains over these scalar implementations persist, while standard bandit methods remain stronger. Cohorts are not pooled.

Which costs offset allocation gains? The legacy $u = 1$ study examines this accounting with Scalar-NM, which matches the action-averaged squared projected score norm at its own state before reward weighting (Appendix I.1). For paired pull-count differences $dN_i = N_i^{\mathrm{DGB}} - N_i^{\text{Scalar-NM}}$ and $\delta = \mu_1 - \mu_2$, the exact accounting is

$$d\widehat{R}_T = -\delta dN_1 + \sum_{i \geq 3} (\mu_2 - \mu_i) dN_i.$$

The first term captures best-arm allocation; the second captures the cost of sampling outside arms. At $h = 0.1$, biased $K = 5$ and gap $1/16$, best-arm savings 1278.84 exceed outside costs 928.64, leaving net savings 350.21 (paired SE 22.08). Yet both biased gap-$1/4$ cells reverse early gains into terminal losses. DGB wins in mean in only three of eight legacy cells; the complete eight-panel decomposition is retained in Figure 3.

Three further diagnostics test step size, small gaps and reward noise. Larger steps lower both methods' absolute regrets but do not eliminate the trade-off: in a disjoint 32-pair cohort, DGB's lower-mean count falls from three to two to one as h increases from 0.1 to 0.4 to 1. The separate small-gap diagnostic is favorable for $K = 3$ but not $K = 5$: DGB exceeds the scalar upper rule by 45.7–72.8% in every tested $K = 5$ cell. Low-noise and deterministic checks of the new design also retain failures (Appendix G). All numerical studies diagnose finite-scale behavior; the class-level exponent comparison is supplied by the theory, not these fixed-gap runs.

7 Discussion and conclusion

For fixed $K \geq 3$ and starts in $\mathcal{N}$, the optimal scalar exponent $1 - 1/K$ contrasts with DGB's $1/2$ under the same sampling, feedback and scalar-step and baseline bounds. The fixed-instance analysis identifies DGB's exact expected logarithmic coefficient; the experiments identify conditional gains over fixed scalar rules and explain which allocation gains survive offsetting costs.

Scope and next questions. *Theory.* The fixed-anytime extension leaves open the sharp uniform-start exponent for horizon-aware scalar rules. Optimal logarithmic factors and dimension-dependent constants also remain open. The activity cost concerns only the finite-activity subclass. *Finite scales.* Quantitative crossover bounds remain open: one evaluated upper bound exceeds $100T$ (Appendix M.1). A sufficient onset is not a necessary recovery time. Fixed-gap trajectories and finite clocks do not certify asymptotic entry. Reaching a probability threshold once does not show that the policy remains above it. *Implementation.* Exact-real full support at finite times does not ensure binary64 reliability indefinitely (samplers and guards: Appendices H and G).

AI USE STATEMENT

Generative AI assisted mathematical exploration, proof drafting and criticism, experimental design and interpretation, code and tests, literature work, figures and writing. Simulations generated the empirical results, with failures retained. The theoretical proofs and experiments have undergone human review. The authors are responsible for the content, including all AI-assisted material.

REPRODUCIBILITY STATEMENT

Proofs and constants appear in Appendices A–F; cohort definitions are in Appendices I.2 and G. The current source package rebuilds the paper from the included figures and data extracts. Its README separates this rebuild from figure regeneration, raw reanalysis and new simulation, and identifies the unchanged predecessor archives required for each. New and legacy numerical cohorts remain separate.

REFERENCES

Shipra Agrawal and Navin Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pp. 39.1–39.26, 2012. URL https://proceedings.mlr.press/v23/agrawal12.html.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002. doi: 10.1023/A:1013689704352.

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, 2025. doi: 10.52202/085713-3787.

Nicolò Cesa-Bianchi, Claudio Gentile, Gábor Lugosi, and Gergely Neu. Boltzmann exploration done right. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/2017/hash/b299ad862b6f12cb57679f0538eca514-Abstract.html.

Leonardo Cesani, Matteo Papini, and Marcello Restelli. How log-barrier helps exploration in policy optimization. *arXiv preprint arXiv:2603.15001*, 2026. URL https://arxiv.org/abs/2603.15001v2.

Wesley Chung, Valentin Thomas, Marlos C. Machado, and Nicolas Le Roux. Beyond variance reduction: Understanding the true impact of baselines on policy optimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 1999–2009. PMLR, 2021. URL https://proceedings.mlr.press/v139/chung21a.html.

Shivam Garg, Samuele Tosatto, Yangchen Pan, Martha White, and Rupam Mahmood. An alternate policy gradient estimator for softmax policies. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pp. 6630–6689. PMLR, 2022. URL https://proceedings.mlr.press/v151/garg22b.html.

Yanwei Jia and Du Ouyang. Convergence and regret of the policy gradient for multi-armed bandits in diffusion environment. *arXiv preprint arXiv:2607.29593*, 2026. URL https://arxiv.org/abs/2607.29593v1.

Safwan Labbi, Daniil Tiapkin, Paul Mangold, and Eric Moulines. Beyond softmax and entropy: Convergence rates of policy gradients with f-SoftArgmax parameterization & coupled regularization. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/17a5dab745d625415c119f127b9fd121-Abstract-Conference.html.

Tor Lattimore. A diffusion analysis of policy gradient for stochastic bandits. *arXiv preprint arXiv:2603.10219*, 2026a. URL https://arxiv.org/abs/2603.10219v1.

Tor Lattimore. A lyapunov analysis of softmax policy gradient for stochastic bandits. *arXiv preprint arXiv:2603.26547*, 2026b. URL https://arxiv.org/abs/2603.26547v1.

Michael Lu, Matin Aghaei, Anant Raj, and Sharan Vaswani. Towards principled, practical policy gradient for bandits and tabular MDPs. *Reinforcement Learning Journal*, 1:216–282, 2024. URL https://rlj.cs.umass.edu/2024/papers/Paper35.html.

Jincheng Mei and Ian Osband. Delightful gradients accelerate corner escape. *arXiv preprint arXiv:2605.11908*, 2026. URL https://arxiv.org/abs/2605.11908v1.

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, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/a12f69495f41bb3b637ba1b6238884d6-Abstract.html.

Jincheng Mei, Wesley Chung, Valentin Thomas, Bo Dai, Csaba Szepesvári, and Dale Schuurmans. The role of baselines in policy gradient optimization. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/718d02a76d69686a36eccc8cde3e6a41-Abstract-Conference.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.

Emerson Melo and David Müller. Beyond softmax: A new perspective on gradient bandits. *arXiv preprint arXiv:2510.03979*, 2025. URL https://arxiv.org/abs/2510.03979v1.

Samuel Robertson, Thang Chu, Bo Dai, Dale Schuurmans, Csaba Szepesvári, and Jincheng Mei. REINFORCE converges to optimal policies with any learning rate. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0541.

Neil Walton. A short note on soft-max and policy gradients in bandits problems. *arXiv preprint arXiv:2007.10297*, 2020. URL https://arxiv.org/abs/2007.10297.

Neil Walton and Denis Denisov. Regret analysis of a Markov policy gradient algorithm for multiarm bandits. *Mathematics of Operations Research*, 48(3):1553–1588, 2023. doi: 10.1287/moor.2022.1311. Preprint: arXiv:2007.10229, version 3.

Julian Zimmert and Teodor V. Marinov. PRODuctive bandits: Importance weighting no more. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/9b14a88c6e50068a97256696902521bf-Abstract-Conference.html.

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

SUPPLEMENTARY MATERIAL: SCALAR LIMITS AND GEOMETRIC RECOVERY IN SOFTMAX GRADIENT BANDITS

Anonymous authors
Paper under double-blind review

SUPPLEMENTARY MATERIAL: READING GUIDE

The main text states the results and their proof ideas; the complete proofs are collected below. Main-text Theorems 1, 2 and 3 correspond respectively to Theorem A.1, the result in Appendix D, and Theorem B.1 at $\gamma = 1/2$. Main-text Corollary 4 specializes Corollary C.1 to the same parameter. The appendix retains the full $0 < \gamma \leq 1/2$ family. DGB and Scalar-NM correspond to legacy PG and NM identifiers, retained unchanged in raw records.

Question	Proof or evidence	Start page
Scalar obstruction	Two-arm case, subset upper bound, stopped-phase lower bound	13
Geometric recovery	Potential and cap moments; two regret bounds	24
Connecting the results	Common-domain separation and update interfaces	29
Other initializations	Separate arbitrary-initialization theorem	32
Cost of finite activity	Clock dichotomy, allocation and Hardy bounds	39
Exact expected coefficient	General family, moments, uniform integrability; half endpoint	45, 53
Experimental evidence	Fixed-design study and interval validation; legacy cost diagnostics	54, 62
Limits and context	Endpoint feasibility; failures; prior-work interfaces	80, 81, 82

Core proof paths. Scalar: phase induction (43) $\rightarrow$ survival $\rightarrow$ occupation transfer (Lemma A.6) $\rightarrow$ Theorem A.7; the upper bound is separate. Recovery: comparison drift (Appendix B.3) $\rightarrow$ cap and mixed moments (Appendices B.4–B.5) $\rightarrow$ master inequality (68) $\rightarrow$ Theorem B.1. These two routes meet only in Corollary C.1. Finite activity and the endpoint law explain cost; neither is a premise of that separation.

Key quantities and modes. j denotes the best arm; the lower-bound family relabels it as arm 1. $X = \theta_2 - \theta_1$ is the rival lead, τ the common history-defined exit time, and σ the cold-entry time. The phase parameter ρ sets $\kappa = (K-2)(1-\rho)/(1+(K-1)\rho)$ in (39). The normalized expected occupations $q_{\mathrm{occ}}, p_{\mathrm{occ}}$ belong to the auxiliary and real laws, respectively (Lemma A.6); neither is a policy coordinate. $B_{\max}$ bounds the baseline, whereas $B_i(p)$ in Appendix B is a comparison-potential ingredient. The two-arm constant C_2 in (16) includes the fixed initial policy. For recovery, $\overline{C}_t$ is a running maximum, not a count; A_t is a cumulative scalar clock. Appendix F proves pathwise limits for the fixed family, then uses moment control and uniform integrability for the exact expectation limit. The half-endpoint specialization is collected in Appendix F.6; its almost-sure clock and count limits are distinct from convergence in expectation.

The arbitrary-initialization theorem fixes one anytime rule before its constants and exponent; it is not a premise of the neighborhood minimax result. Bounds concern unconditional expected pseudo-regret unless another mode is stated. Table 10 and Appendix G separate the experimental cohorts; they are not pooled. Complete current proofs and evidence follow.

A THE FIXED-RESOURCE SCALAR BARRIER

This appendix proves the scalar minimax bounds, including the auxiliary inequalities used below. All rewards are independent, pull-indexed IID streams in $[0,1]$. The best arm is unique. The initial state is the finite deterministic vector $\theta_1 = \vartheta$, and $\mathcal{F}_t$ is the history before action t. The current action seed and the next unused rewards have their original conditional product law given $\mathcal{F}_t$. An admissible rule uses the exact update

$$\begin{aligned} p_t &= \mathrm{softmax}(\theta_t), & I_t \mid \mathcal{F}_t &\sim p_t, \\ \theta_{t+1} &= \theta_t + \eta_t(Y_t - b_t)(e_{I_t} - p_t), & 0 \le \eta_t &\le u, \quad |b_t| \le B_{\max}. \end{aligned} \tag{8}$$

where $u > 0$ and $B_{\max} \ge 1$ are fixed. Both controls are common to all coordinates and are chosen before the action. Independent control randomization is allowed. The rule may know the horizon, but its map from observed histories to controls is the same in every unknown environment. There are no forced actions, resets, projections, changes to the sampler or score, or additional reward observations.

Let $\mu_i = \mathbb{E}Y_1^{(i)}$. If i^* is the best arm, write $\Delta_i = \mu_{i^*} - \mu_i$ and

$$N_i(T) = \sum_{t=1}^{T} \mathbf{1}\{I_t = i\}, \qquad \widehat{R}_T = \sum_i \Delta_i N_i(T), \qquad R_T = \mathbb{E}\widehat{R}_T = \mathbb{E}\sum_{t=1}^{T}\sum_i \Delta_i p_{i,t}.$$

Thus R_T is unconditional expected pseudo-regret, not realized reward loss. We write $\mathbb{E}_t$ for conditional expectation given $\mathcal{F}_t$. In one-step calculations a prime denotes the state after the observation and $\Delta\theta = \theta' - \theta$ denotes its increment. Put $L_T = 1 + \log T$ and define

$$\mathcal{M}_0(T;\vartheta) = \inf_{\mathcal{A}} \sup_{\mathcal{E}} R_T^{\mathcal{A}}(\mathcal{E}),$$

where the infimum is over the rules just specified and the supremum is over the stated IID environments.

Theorem A.1 (Fixed-resource scalar minimax bounds)**.** *For $K = 2$ and every fixed finite ϑ, $\mathcal{M}_0(T;\vartheta) = \Theta(\sqrt{T})$ as $T \to \infty$. For every fixed $K \ge 3$, put $\vartheta^* = (0,1,0,\dots,0)$ and $\mathcal{N} = \{\vartheta : \|\vartheta - \vartheta^*\|_\infty \le 1/8\}$. There are $c_- > 0$, $C_+ < \infty$ and $T_0 < \infty$, depending only on $K, u, B_{\max}$, such that simultaneously for all $\vartheta \in \mathcal{N}$,*

$$c_- \frac{T^{1-1/K}}{L_T^{3-2/K}} \le \mathcal{M}_0(T;\vartheta) \le C_+ T^{1-1/K} \qquad (T \ge T_0). \tag{9}$$

Consequently its polynomial exponent is $1 - 1/K$. The assertion for $K \ge 3$ does not identify an optimal logarithmic factor.

The proof has four components: a statistical lower bound valid for all bandit rules; a horizon-aware two-arm upper bound; an analysis-only grouping argument for the multi-arm upper bound; and a multi-stage stopped geometric construction followed by occupation-weighted KL transfer for the lower bound. No assertion about a common exponent at uniform initialization is needed.

A.1 ELEMENTARY STATISTICAL TOOLS AND THE UNIVERSAL LOWER BOUND

For $x, y \in (0,1)$ let $\mathrm{kl}(x,y) = x\log(x/y) + (1-x)\log((1-x)/(1-y))$, with the usual continuous or infinite endpoint values. We use relative entropy in the displayed order of its arguments. For two laws on the same finite history, identical conditional control and action kernels cancel from their likelihood ratio. Multiplication of the remaining one-step reward densities and conditional expectation give

$$D(\mathbb{Q}^{\tau\wedge n} \| \mathbb{P}^{\tau\wedge n}) = \mathbb{E}_{\mathbb{Q}} \sum_{s < \tau\wedge n} D\big(\mathbb{Q}(\cdot \mid \mathcal{H}_s) \| \mathbb{P}(\cdot \mid \mathcal{H}_s)\big), \tag{10}$$

where $\mathcal{H}_s$ is the history after s observations and τ is a common history-defined stopping time. This identity involves only finitely many summands; no unbounded stopping theorem is used. Randomized controls are included in the history with identical conditional laws.

For completeness, KL data processing for an event follows by applying Jensen's inequality separately on the event and its complement to the density ratio. In particular, $D(\mathbb{Q}\|\mathbb{P})$ is at least the

binary relative entropy of their event probabilities. For fixed y, the second derivative of $\mathrm{kl}(x, y)$ in x is $1/[x(1-x)] \geq 4$ and its value and first derivative vanish at $x = y$. Hence $\mathrm{kl}(x, y) \geq 2(x-y)^2$. Taking the supremum over events gives Pinsker's inequality

$$\|\mathbb{Q} - \mathbb{P}\|_{\mathrm{TV}} \leq \sqrt{D(\mathbb{Q}\|\mathbb{P})/2}.$$

For any $Z \in [0, T]$, its layer-cake representation then gives $|\mathbb{E}_{\mathbb{Q}} Z - \mathbb{E}_{\mathbb{P}} Z| \leq T\|\mathbb{Q} - \mathbb{P}\|_{\mathrm{TV}}$.

We will also use a fixed-sample bound without applying it directly to a random sample size. If a random variable is supported on an interval of length w, its centered log-MGF is at most $\lambda^2 w^2/8$. Indeed, the second derivative of the log-MGF is the variance under an exponential tilt, at most $w^2/4$; integrate this bound twice from zero. For the difference of two independent means of n observations in $[0, 1]$, Chernoff optimization therefore yields

$$\mathbb{P}\{|\widehat{\mu}_{1,n} - \widehat{\mu}_{2,n} - (\mu_1 - \mu_2)| > z\} \leq 2e^{-nz^2}. \tag{11}$$

Lemma A.2 (Universal finite-horizon lower bound)**.** *For $K \geq 2$ and $T \geq 1$, every unknown-environment bandit rule has a unique-best Bernoulli environment satisfying*

$$R_T \geq \frac{1}{32} \min\{T, \sqrt{KT}\}.$$

Proof. Let $\mathbb{P}^{(0)}$ have all means $1/2$. For $i \in [K]$, let $\mathbb{P}^{(i)}$ change only arm i to mean $1/2 + \epsilon$, where $\epsilon = \frac{1}{8}\min\{1, \sqrt{K/T}\}$. The true alternatives have a unique best arm; the tie law is used only for comparison. Set $n_i = \mathbb{E}_0 N_i(T)$, so $\sum_i n_i = T$. The history chain rule and $\mathrm{kl}(1/2, 1/2 + \epsilon) = -\frac{1}{2}\log(1 - 4\epsilon^2) \leq 4\epsilon^2$ give $D(\mathbb{P}^{(0)}\|\mathbb{P}^{(i)}) \leq 4\epsilon^2 n_i$. The preceding total-variation bound implies

$$\mathbb{E}_i N_i(T) \leq n_i + T\epsilon\sqrt{2n_i}.$$

Average over i and use Cauchy–Schwarz:

$$\frac{1}{K}\sum_i \frac{\mathbb{E}_i N_i(T)}{T} \leq \frac{1}{K} + \epsilon\sqrt{2T/K} \leq \frac{1}{2} + \frac{\sqrt{2}}{8} < \frac{3}{4}.$$

Thus some alternative has regret at least $\epsilon T/4$, the claimed quantity. The argument permits horizon knowledge and arbitrary sampling, so it also applies to the scalar interface and to geometric extensions. □

A.2 A HORIZON-AWARE TWO-ARM SCALAR RULE

For this subsection only, label the true best arm 1 in the analysis, write $\Delta = \mu_1 - \mu_2 > 0$, and let $\pi_0 = p_{1,1} \in (0, 1)$. Define the logit contrast $\mathsf{D}_t = \theta_{1,t} - \theta_{2,t}$ and odds $\Omega_t = e^{-\mathsf{D}_t}$, $\Lambda_t = e^{\mathsf{D}_t}$. Here and below the subscript zero on $\pi_0, \Omega_0, \Lambda_0$ denotes their initial values, so $\Omega_0 = (1 - \pi_0)/\pi_0$. Put $M = 1 + B_{\max}$ and

$$\Psi(x) = 2\sum_{k=0}^{\infty} \frac{x^k}{(k+2)!}.$$

Lemma A.3 (Two-arm safe-interval bound)**.** *Suppose that on a fresh continuation from a finite observed history, $0 < \eta_- \leq \eta_t \leq \eta_+$ and $\chi_+ = \eta_+ M^2 \Psi(2\eta_+ M) < \Delta$. If the initial inverse odds of that continuation are Ω_0, then*

$$R_T \leq \frac{1}{2\eta_-}\log\{1 + 2\eta_+(\Delta + \chi_+)\Omega_0 T\} + \frac{\Delta\Omega_0}{2\eta_-(\Delta - \chi_+)}. \tag{12}$$

The bound is conditional on that observed history, with the unchanged future product reward kernel.

Proof. For $|Z| \leq M$ and $|v| \leq 2\eta_+$, the absolutely convergent exponential series gives $\mathbb{E}_t e^{vZ} \leq 1 + v\mathbb{E}_t Z + (v^2/2)M^2\Psi(2\eta_+ M)$. The contrast increment is $2\eta_t(Y^{(1)} - b_t)(1 - p_{1,t})$ on arm1 and $-2\eta_t(Y^{(2)} - b_t)p_{1,t}$ on arm2. Averaging over the action, the baseline cancels from the first-order term. Applying the exponential bound with both signs gives

$$\mathbb{E}_t\Lambda_{t+1} \leq \Lambda_t + 2\eta_+(\Delta + \chi_+), \qquad \mathbb{E}_t\Omega_{t+1} \leq \Omega_t - 2\eta_-(\Delta - \chi_+)(1 - p_{1,t})^2.$$

Finite summation, nonnegativity and Jensen's inequality imply

$$\mathbb{E}(\mathsf{D}_{T+1} - \mathsf{D}_1) \le \log\{1 + 2\eta_+(\Delta + \chi_+)\Omega_0 T\}, \qquad \mathbb{E}\sum_{t\le T}(1 - p_{1,t})^2 \le \frac{\Omega_0}{2\eta_-(\Delta - \chi_+)}.$$

The exact conditional mean contrast is $2\Delta\eta_t p_{1,t}(1 - p_{1,t})$. Use its sum and the identity $1 - p_{1,t} = p_{1,t}(1 - p_{1,t}) + (1 - p_{1,t})^2$ to obtain (12). All sums are finite. The same calculation works after a finite stopping history because its unused rewards retain their original laws. □

Theorem A.4 (Two-arm square-root upper bound)**.** *For every fixed finite initial state and $u > 0$, there is a horizon-aware unknown-gap and unknown-best rule satisfying* (8) *and $R_T \le C_2\sqrt{T}$ on every two-arm IID environment in $[0, 1]$. The finite constant C_2 is independent of the gap and horizon.*

Proof. Set $\mathsf{H} = \max\{32M^2, 1/u\}$. Use baseline zero throughout. Before activation set $\eta_t = 0$, while continuing natural softmax sampling. For each $n_j = 2^j \le T/2$, test once when both arms have at least n_j observations, using their first n_j rewards. Define

$$\widehat{\Delta}_j = \widehat{\mu}_{1,n_j} - \widehat{\mu}_{2,n_j}, \qquad \mathfrak{r}_j = \sqrt{\frac{2}{n_j}\log\frac{eT}{n_j}}.$$

At the first test with $|\widehat{\Delta}_j| > 2\mathfrak{r}_j$, commit to the common scalar $\eta = |\widehat{\Delta}_j|/\mathsf{H}$ for the remaining actions. If no test triggers, keep $\eta = 0$. This commits to a scalar, not an action. Since $|\widehat{\Delta}_j| \le 1$, the scalar is at most $1/\mathsf{H} \le u$. Neither its value nor the test needs the best-arm identity or the true gap.

Let $x = T\Delta^2$ and $c_\pi = 1/\pi_0 + 1/(1 - \pi_0)$. If $x < 4096$, the trivial bound $R_T \le \Delta T$ is less than $64\sqrt{T}$. Suppose $x \ge 4096$, and put $L_x = \log(ex)$. Let n_* be the smallest power of two no smaller than $32L_x/\Delta^2$. Then

$$\frac{32L_x}{\Delta^2} \le n_* < \frac{64L_x}{\Delta^2} \le T/2, \qquad \mathfrak{r}_* \le \Delta/4. \tag{13}$$

The third inequality follows from $128\log(ex) \le x$ for $x \ge 4096$; the radius bound also uses $n_* \ge 1/\Delta^2$, so $\log(eT/n_*) \le L_x$.

Consider a virtual continuation which remains frozen forever. The expected time until each arm has n_* observations is at most $c_\pi n_*$: the two individual waiting-time expectations are n_*/π_0 and $n_*/(1 - \pi_0)$, and their maximum is at most their sum. The actual and virtual histories coincide up to activation. If the n_* test is reached without a trigger, then $|\widehat{\Delta}_*| \le 2\mathfrak{r}_* \le \Delta/2$, and consequently $|\widehat{\Delta}_* - \Delta| \ge \Delta/2$. By (11), this has probability at most $2e^{-n_*\Delta^2/4} \le 2(ex)^{-8}$. If σ_{fr} is the number of frozen actions, truncated at T, then

$$\mathbb{E}\sigma_{\mathrm{fr}} \le c_\pi n_* + 2T(ex)^{-8}.$$

Because the indicator that an action remains frozen is pre-action measurable, its expected regret is exactly $\Delta(1 - \pi_0)\mathbb{E}\sigma_{\mathrm{fr}}$. In particular,

$$R_{\mathrm{explore}} \le 64c_\pi L_x/\Delta + 2\Delta T(ex)^{-8} \le (128c_\pi + 2)\sqrt{T}, \tag{14}$$

using $L_x \le 2\sqrt{x}$ for $x \ge 1$.

Call a triggering test bad if $|\widehat{\Delta}_j - \Delta| > \mathfrak{r}_j$. For a trigger before or at the n_* test, the union bound and (11) give probabilities at most $2(n_j/(eT))^2$. A later trigger requires the n_* test not to have triggered. The sum of squares of the dyadic sizes through n_* is at most $4n_*^2/3$. Thus

$$\mathbb{P}\{\text{bad trigger}\} \le \frac{8}{3e^2}(n_*/T)^2 + 2(ex)^{-8}.$$

Charging at most ΔT on that branch yields

$$R_{\mathrm{bad}} \le \sqrt{T}\left\{\frac{8\cdot 4096}{3e^2}\frac{L_x^2}{x^{3/2}} + 2e^{-8}x^{-15/2}\right\} \le 100\sqrt{T}. \tag{15}$$

For the last bound use $L_x \le 2\sqrt{x}$, $\sqrt{x} \ge 64$ and $e^2 > 7$.

On a good triggering history, put $d = |\widehat{\Delta}_j|$. The relations $d > 2\mathfrak{r}_j$ and $|\widehat{\Delta}_j - \Delta| \le \mathfrak{r}_j$ imply $2\Delta/3 \le d \le 2\Delta$. The committed scalar therefore belongs to $[\eta_-, \eta_+] = [2\Delta/(3\mathsf{H}), 2\Delta/\mathsf{H}]$. Using $\Psi(z) \le e^z$, $M \ge 1$, $\Delta \le 1$ and $\mathsf{H} \ge 32M^2$,

$$\chi_+ \le \frac{2\Delta M^2}{\mathsf{H}} e^{4\Delta M/\mathsf{H}} \le \frac{e^{1/8}}{16}\Delta < \Delta/4.$$

The logits have not moved before commitment, so the initial inverse odds remain Ω_0. The good-trigger event is measurable at that stopping history. Lemma A.3, conditionally on that history and using at most T further actions, gives

$$R_{\mathrm{good,post}} \le \frac{3\mathsf{H}}{4\Delta}\log(1 + 5\Omega_0\Delta^2 T/\mathsf{H}) + \frac{\mathsf{H}\Omega_0}{\Delta} \le \left(\frac{3}{2}\sqrt{5\mathsf{H}\Omega_0} + \mathsf{H}\Omega_0\right)\sqrt{T}.$$

Here $\log(1+z) \le 2\sqrt{z}$ and $\Delta^{-1} \le \sqrt{T}$ on this branch of the argument. Conditioning is only on observed histories, not on a future event of perpetual correct certification. Combining the three costs and exchanging the two best-arm labels if necessary proves the result, for example with

$$\begin{aligned} C_2 &= 128\left(\frac{1}{\pi_0} + \frac{1}{1-\pi_0}\right) + 128 + \frac{3}{2}\sqrt{5\mathsf{H}\Omega_{\max}} + \mathsf{H}\Omega_{\max}, \\ \Omega_{\max} &= \max\{\pi_0/(1-\pi_0), (1-\pi_0)/\pi_0\}. \end{aligned} \tag{16}$$

This constant also covers $T = 1$ and $x < 4096$.

For the baseline-zero rule just constructed, the sharper residual envelope $M_{\mathrm{actual}} = 1$ is valid since $|Y_t - b_t| = Y_t \le 1$. Every use of M in the safe-interval proof bounds this residual, so the same argument permits $\mathsf{H} = \max\{32, 1/u\}$. Explicitly, on a good triggering history its only safety check becomes

$$\chi_+ \le \frac{2\Delta}{\mathsf{H}} e^{4\Delta/\mathsf{H}} \le \frac{e^{1/8}}{16}\Delta < \Delta/4.$$

The testing, exploration and bad-trigger bounds are unchanged, and (16) uses this specialized H. □

A.3 The multi-arm upper bound without oracle information

Theorem A.5 (Fixed-horizon multi-arm upper bound)**.** *For $K \ge 3$ and every $\vartheta \in \mathcal{N}$, the rule*

$$b_t = 0, \qquad \eta_t = \upsilon T^{-1/K}, \qquad \upsilon = \min\{u, 1/(8eK)\}$$

has $R_T \le C_+ T^{1-1/K}$ for all $T \ge 1$, uniformly over the reward environments and initializations in $\mathcal{N}$, with a common C_+. It does not use the gaps or best-arm identity.

Proof. Put $\varepsilon = T^{-1/K}$ and $\eta = \upsilon\varepsilon$. In this proof alone write θ for the actual logits minus the constant $(\sum_i \vartheta_i - 1)/K$ in every coordinate. This changes neither probabilities nor updates and does not feed normalized logits to the controller. The coordinate sum is one, the initial maximum is at most $5/4$, and the initial span is at most $5/4$, uniformly on $\mathcal{N}$. Sort the means in descending order, only for analysis: $\mu_{(1)} \ge \cdots \ge \mu_{(K)}$, where (r) denotes the arm of rank r (break ties by arm index). If every adjacent difference is less than ε, their total range is less than $(K-1)\varepsilon$, and $R_T \le (K-1)\varepsilon T$. Otherwise take $g = \min\{r \in \{1, \dots, K-1\} : \mu_{(r)} - \mu_{(r+1)} \ge \varepsilon\}$ and $G = \{(1), \dots, (g)\}$: this is the top cluster above the highest qualifying gap, not the cluster above an arbitrary gap. Put $J = G^{\mathrm{c}}$ and define

$$\begin{aligned} P_t &= \sum_{i \in G} p_{i,t}, & Q_t &= \sum_{j \in J} p_{j,t}, \\ \delta_j &= \min_{i \in G} \mu_i - \mu_j \ge \varepsilon, & r_t^- &= \sum_{j \in J} \delta_j p_{j,t}, \\ \Delta_G^{\mathrm{in}} &= \mu_{i^*} - \min_{i \in G} \mu_i \le (K-1)\varepsilon. \end{aligned}$$

The true one-step pseudo-regret is at most $\Delta_G^{\mathrm{in}} + r_t^-$. Write $\bar{\theta}_{G,t} = g^{-1}\sum_{i \in G}\theta_{i,t}$ and $z_{G,t} = e^{\bar{\theta}_{G,t}}/\sum_i e^{\theta_{i,t}} \le P_t/g$. Let $C = 6/\varepsilon$ and set

$$\mathfrak{h}_A(\theta) = \exp\left(\sum_{j \in A}\theta_j - |A|\bar{\theta}_G\right), \qquad \mathcal{V}(\theta) = \sum_{\varnothing \ne A \subseteq J} C^{|A|-1}\mathfrak{h}_A(\theta).$$

This adapts the subset-exponential potential of Jia & Ouyang (2026, Section 2.3, equation (65)): the distinguished best-arm logit is replaced by the mean logit of the near-optimal group, and subsets are taken over its complement. The highest-gap choice controls within-group regret while the following layer calculation controls outside regret. The group and potential are proof quantities, not algorithm inputs.

For a fixed exponent vector $\mathbf{a}$ with oscillation at most K/g, let $\mathfrak{m}_i = \mathbb{E}[(Y^{(i)})^2] \le 1$. The first conditional moment of $\mathbf{a}^\top \Delta\theta$ is $\eta \sum_{i<j} p_i p_j (\mathbf{a}_i - \mathbf{a}_j)(\mu_i - \mu_j)$. Jensen gives $(\mathbf{a}_i - p^\top \mathbf{a})^2 \le \sum_j p_j (\mathbf{a}_i - \mathbf{a}_j)^2$. After multiplication by $p_i \mathfrak{m}_i$ and summation this bounds its conditional second moment by $\eta^2 \sum_{i<j} p_i p_j (\mathfrak{m}_i + \mathfrak{m}_j)(\mathbf{a}_i - \mathbf{a}_j)^2$. The exponential Taylor remainder is therefore bounded as follows:

$$\mathbb{E}_t e^{\mathbf{a}^\top \Delta\theta} \le 1 + \eta \sum_{i<j} p_i p_j (\mathbf{a}_i - \mathbf{a}_j)(\mu_i - \mu_j) + \frac{\eta^2 e^{K\eta/g}}{2} \sum_{i<j} p_i p_j (\mathfrak{m}_i + \mathfrak{m}_j)(\mathbf{a}_i - \mathbf{a}_j)^2. \quad (17)$$

Take $\mathbf{a}_i = -s/g$ on G, $\mathbf{a}_i = 1$ on A and $\mathbf{a}_i = 0$ on $J \setminus A$, where $s = |A| \le K - g$. Its oscillation is $1 + s/g \le K/g$, which verifies the hypothesis of the preceding bound. Pairs within G disappear. For a pair from G to J, the second-order term consumes at most half the negative first-order term: its ratio is at most $\eta e^{K\eta/g}(K/g)/\varepsilon \le 1/2$. For pairs crossing A and $J \setminus A$, the positive coefficient is at most $2 + \eta e^{K\eta/g} \le 3$. Consequently

$$\mathbb{E}_t \Delta \mathfrak{h}_A \le \eta \mathfrak{h}_A \left[-\frac{P_t}{2} \left\{ (1 + s/g) \sum_{i \in A} \delta_i p_{i,t} + (s/g) \sum_{j \in J \setminus A} \delta_j p_{j,t} \right\} + 3 \Big(\sum_{i \in A} p_{i,t} \Big) \Big(\sum_{j \in J \setminus A} p_{j,t} \Big) \right]. \quad (18)$$

For $i \in A$, $j \in J \setminus A$, $\mathfrak{h}_A p_i p_j = \mathfrak{h}_{A \cup \{j\}} z_G p_i \le \mathfrak{h}_{A \cup \{j\}} (P_t/g) p_i$. For a target subset A of size $r \ge 2$, each fixed $i \in A$ has $r - 1$ positive predecessors. After factoring out $\eta \mathfrak{h}_A p_{i,t}$, the inherited positive and inside negative coefficients combine as

$$\begin{aligned} &\frac{3P_t}{g}(r-1)C^{r-2} - \frac{P_t}{2}(1 + r/g)\delta_i C^{r-1} \\ &\qquad \le \frac{3P_t}{g} C^{r-2}\{(r-1) - (g+r)\} \le 0, \end{aligned}$$

because $\delta_i \ge \varepsilon$ and $C\varepsilon/2 = 3$. Retain all singleton negative terms. Since $P_t/z_G \ge g$, they sum to

$$\mathbb{E}_t \Delta \mathcal{V} \le -\frac{\eta}{2} \left\{ g \sum_{j \in J} \delta_j p_{j,t}^2 + Q_t r_t^- \right\}, \qquad \mathbb{E} \sum_{t \le T} \eta Q_t r_t^- \le 2\mathcal{V}(\vartheta). \quad (19)$$

In detail, singleton inside terms yield $(g+1)\sum_j \delta_j p_j^2$; their other negative terms yield $\sum_{i \ne j} p_i \delta_j p_j = Q_t r_t^- - \sum_j \delta_j p_j^2$. All finite-time expectations exist because increments are bounded.

The coordinate sum remains one. Taylor applied to $e^{-\theta_i}$, using $\mathbb{E}_t \Delta\theta_i = \eta p_i (\mu_i - \sum_j p_j \mu_j)$ and $\mathbb{E}_t (\Delta\theta_i)^2 \le \eta^2 p_i (1 - p_i)$, gives

$$\mathbb{E}_t e^{-\theta_{i,t+1}} \le e^{-\theta_{i,t}} + d_0 \eta, \qquad d_0 = \frac{e^{-1/K}}{K} \left(2 + \frac{\eta_0 e^{\eta_0}}{2} \right), \quad \eta_0 = 1/(8eK).$$

Here $e^{-\theta_i} p_i = (\sum_j e^{\theta_j})^{-1} \le e^{-1/K}/K$. Iteration and Jensen imply

$$\mathbb{E}(\theta_{i,1} - \theta_{i,T+1}) \le \log(1 + d_0 \eta e^{\theta_{i,1}} T).$$

The group coordinate sum has conditional mean increment $\eta \sum_{i \in G, j \in J} p_i p_j (\mu_i - \mu_j) \ge \eta P_t r_t^-$. Its total expected increase equals the expected decrease outside G. Combine this fact with (19), $P_t + Q_t = 1$ and $e^{\theta_{i,1}} \le e^{5/4}$. With $D_\theta = d_0 e^{5/4}$,

$$R_T \le \Delta_G^{\text{in}} T + \frac{K-1}{\eta} \log(1 + D_\theta \eta T) + \frac{2\mathcal{V}(\vartheta)}{\eta}. \quad (20)$$

The initial logit span is at most $5/4$. Thus $\mathfrak{h}_A(\vartheta) \le e^{5|A|/4}$ and

$$\mathcal{V}(\vartheta) \le V_\theta \varepsilon^{-(K-2)}, \qquad V_\theta = 6^{K-2}(1+e^{5/4})^{K-1}.$$

Use $T^{1/K}\log T \le K\{e(K-2)\}^{-1}T^{1-1/K}$ in (20). One permissible constant is

$$C_+ = (K-1) + \frac{1}{\upsilon}\left\{(K-1)\log(1+D_\theta \upsilon) + \frac{(K-1)^2}{e(K-2)} + 2V_\theta\right\}. \tag{21}$$

It is independent of the reward laws, their gaps, T, and the initialization within $\mathcal{N}$. □

A.4 Occupation-weighted relative-entropy transfer

For the lower-bound proof it is convenient to index states by completed observations. Define, without changing the process,

$$\theta^{[n]} = \theta_{n+1}, \qquad p^{[n]} = p_{n+1}, \qquad \eta^{[n]} = \eta_{n+1}, \qquad b^{[n]} = b_{n+1}, \qquad \mathcal{H}_n = \mathcal{F}_{n+1} \quad (n \ge 0).$$

Thus $n = 0$ means the specified initial state and the action indexed n below is the original action $n+1$. For a law used in the lower-bound construction, $\mathbb{E}_n$ denotes conditional expectation given $\mathcal{H}_n$ under that law.

Lemma A.6 (Stopped-occupation transfer)**.** *Let $\mathbb{P}, \mathbb{Q}$ be laws on the same history through T observations, with the same initial law. Let τ be a common $\mathcal{H}_n$-stopping time. If before τ their one-step conditional relative entropy, in the direction $\mathbb{Q}\|\mathbb{P}$, is at most a deterministic ϵ_{KL}, then*

$$\mathbb{E}_{\mathbb{P}}(T \wedge \tau) \ge \varrho_{\epsilon_{\mathrm{KL}} T}\, \mathbb{E}_{\mathbb{Q}}(T \wedge \tau),$$

where $\varrho_c \in (0,1]$ is the unique solution of $\varrho_c - 1 - \log \varrho_c = c$. In particular $\epsilon_{\mathrm{KL}} T \le 1/8$ gives a factor $1/2$.

Proof. Add an independent uniform index $U \in \{0, \dots, T-1\}$, not revealed to the controller. Record U and the stopped history through $U \wedge \tau$, including the stopping decision at that state. The event $\{\tau > U\}$ is measurable from this record. Set $q_{\mathrm{occ}} = \mathbb{E}_{\mathbb{Q}}(T \wedge \tau)/T$ and $p_{\mathrm{occ}} = \mathbb{E}_{\mathbb{P}}(T \wedge \tau)/T$. These are its probabilities under the two laws. The finite stopped chain rule (10) gives

$$\mathrm{kl}(q_{\mathrm{occ}}, p_{\mathrm{occ}}) \le \epsilon_{\mathrm{KL}} \mathbb{E}_{\mathbb{Q}}(U \wedge \tau) \le \epsilon_{\mathrm{KL}} T q_{\mathrm{occ}}.$$

The conclusion is immediate if $q_{\mathrm{occ}} = 0$ or $p_{\mathrm{occ}} \ge q_{\mathrm{occ}}$. Otherwise finite KL implies $p_{\mathrm{occ}} > 0$. The inequality $\log(1+x) \le x$ gives

$$\mathrm{kl}(q_{\mathrm{occ}}, p_{\mathrm{occ}}) \ge q_{\mathrm{occ}} \log(q_{\mathrm{occ}}/p_{\mathrm{occ}}) - q_{\mathrm{occ}} + p_{\mathrm{occ}}.$$

Divide by q_{occ} and use that $r - 1 - \log r$ decreases from infinity to zero on $(0,1]$. This proves the bound, including $q_{\mathrm{occ}} = 1$ by continuity. Finally $\log 2 - 1/2 > 1/8$: the two equal-length midpoint rectangles under $1/x$ on $[1,2]$ give $\log 2 \ge 2/5 + 2/7 > 5/8$. □

If a nonnegative predictable loss is at least g_{loss} before τ, then its unconditional expected total loss is at least

$$g_{\mathrm{loss}} \mathbb{E}_{\mathbb{P}}(T \wedge \tau) \ge \frac{g_{\mathrm{loss}}}{2} T \mathbb{Q}(\tau > T) \quad \text{when } \epsilon_{\mathrm{KL}} T \le 1/8. \tag{22}$$

This uses stopped occupation, not conditional sampling on a future survival event.

A.5 A uniform initialization-neighborhood lower bound

For the lower bound fix $K \ge 3$ and put $B = B_{\max}$. The neighborhood $\mathcal{N}$ is defined in Theorem A.1. All controls and the model remain those of (8).

Theorem A.7 (Uniform neighborhood lower bound)**.** *There are $d, c > 0$ and $T_0 < \infty$, depending only on K, u, B, such that for every integer $T \ge T_0$, every $\vartheta \in \mathcal{N}$, and every admissible controller $\mathcal{A}$, the* same *IID Bernoulli environment*

$$\mu_1 = \tfrac{1}{2} + \delta_T, \qquad \mu_2 = \tfrac{1}{2}, \qquad \mu_j = \tfrac{1}{4} - \frac{j-3}{16K} \quad (3 \le j \le K), \qquad \delta_T = d\, T^{-1/K} L_T^{-(K-2)/K} \tag{23}$$

satisfies

$$R_T^{\mathcal{A},\vartheta}(\mathcal{E}_{\delta_T}) \geq c\,\frac{T^{1-1/K}}{L_T^{3-2/K}}. \tag{24}$$

Consequently this lower bound holds for $\inf_{\vartheta\in\mathcal{N}} \mathcal{M}_0(T;\vartheta)$ *as well. The constants and hard gap do not depend on either the initialization inside* $\mathcal{N}$ *or the controller, including a horizon-dependent one.*

The set $\mathcal{N}$ has nonempty K-dimensional interior; it is not just a family of common logit shifts. For example, $\vartheta^* + s(e_1 - e_3) \in \mathcal{N}$ for $|s| \leq 1/8$, and these initializations change the softmax policy. No assertion at uniform initialization or about an optimal logarithmic factor is made.

A.5.1 Normalization and a uniform initial potential

The construction below is made separately for each fixed T, controller, and initialization. Its auxiliary law $\mathbb{Q}_\delta$, entry and exit times σ, τ, and potential W_n may depend on these choices. By contrast, the final d, c, T_0 depend only on K, u, B: the bounds used to select them are uniform over the controller and $\vartheta \in \mathcal{N}$. Write $\vartheta = \vartheta^* + e$, $|e_i| \leq 1/8$, and $\bar{e} = K^{-1}\sum_i e_i$. For analysis only let

$$\phi_n = \theta_{n+1} - \bar{e}\mathbf{1}, \quad p_n = p_{n+1}, \quad \eta_n = \eta_{n+1}, \quad b_n = b_{n+1}, \quad \mathcal{H}_n = \mathcal{F}_{n+1} \quad (n \geq 0).$$

Thus n counts completed observations, $\sum_i \phi_{i,n} = 1$, and $p_n = \mathrm{softmax}(\phi_n)$. The actual controls are unchanged. This is not an algorithmic reset and does not require translation-equivariance of the controller: it merely represents the same realized increments in shifted coordinates. In particular

$$\phi_{i,0} \geq -\tfrac{1}{4}, \qquad X_0 := \phi_{2,0} - \phi_{1,0} \in [\tfrac{3}{4}, \tfrac{5}{4}], \qquad V_0^{3/4} := \phi_{1,0} - \tfrac{3}{4}\phi_{2,0} \leq -\tfrac{1}{2}. \tag{25}$$

The last bound follows directly from $V_0^{3/4} = -3/4 + e_1 - (3/4)e_2 - (1/4)\bar{e} \leq -3/4 + 2/8$. Let

$$h_0 = \log 2 < \tfrac{3}{4}, \qquad f_* := \tfrac{1}{2}(\tfrac{3}{4} - \log 2) > 0. \tag{26}$$

Then $(X_0 - h_0)(-V_0^{3/4}) \geq f_*$ for every initialization in $\mathcal{N}$. The strict inequality $\log 2 < 3/4$ follows, for example, by retaining the first four terms of $e^{3/4}$, whose sum exceeds 2.

Put $m = K - 2$ and $a = u(1 + B)$. We prove the lower bound in the larger class $\eta_n \geq 0$, $\eta_n(1 + |b_n|) \leq a$. Thus $|\eta_n(Y - b_n)| \leq a$ for every reward used below. For $0 < \delta \leq 1/32$, let $\mathbb{P}_\delta$ be the environment in (23) with gap δ. Define

$$z_n = \sum_{j\geq 3} p_{j,n}, \qquad r_{o,n} = \sum_{j\geq 3}(\tfrac{1}{2} - \mu_j)p_{j,n} \geq z_n/4, \qquad X_n = \phi_{2,n} - \phi_{1,n}, \quad V_n^\rho = \phi_{1,n} - \rho\phi_{2,n}.$$

A.5.2 One-step identities and the killed product

Conditional on the pre-action history, direct averaging gives

$$\mathbb{E}_{\mathbb{P}_\delta}[\Delta X_n \mid \mathcal{H}_n] = \eta_n\{(p_2 - p_1)r_o - \delta p_1(1 + p_2 - p_1)\}, \tag{27}$$

$$\mathbb{E}_{\mathbb{P}_\delta}[\Delta V_n^\rho \mid \mathcal{H}_n] = \eta_n\{(p_1 - \rho p_2)r_o + \delta p_1(1 - p_1 + \rho p_2)\}. \tag{28}$$

Here and below probabilities on a one-step display refer to the current state. The baseline cancels only from these conditional means, not from sample-path increments.

If $p_1 < \rho p_2$ and $0 < \rho < 1$, then on every action and reward outcome $\Delta V^\rho = -k_I \Delta X$, with

$$k_1 = \frac{1 - p_1 + \rho p_2}{1 + p_2 - p_1}, \quad k_2 = \frac{p_1 + \rho(1 - p_2)}{1 - p_2 + p_1}, \quad k_j = \frac{\rho p_2 - p_1}{p_2 - p_1}\ (j \geq 3), \qquad 0 \leq k_I \leq 1.$$

For $x, y > 0$, $k \geq 0$, and real d,

$$(x + d)_+(y + kd)_+ \geq xy + yd + xkd. \tag{29}$$

If both new factors are positive, their difference from the right side is kd^2. If $x + d \leq 0$, the right side is at most $-kx^2$; if $y + kd \leq 0$ and $k > 0$, it is at most $-y^2/k$. The case $k = 0$ is immediate. Consequently, if the conditional drifts of X and V^ρ have signs $\mathbb{E}_n\Delta X \geq 0$ and $\mathbb{E}_n\Delta V^\rho \leq 0$, the positive product $(X - h)_+(v - V^\rho)_+$, permanently set to zero at the first nonpositive factor, has nonnegative conditional mean increments. Indeed the linear terms in (29) have mean $y\mathbb{E}_n\Delta X - x\mathbb{E}_n\Delta V^\rho \geq 0$. No independence of k_I and ΔX is assumed.

A.5.3 A TAIL BOUND UNIFORM IN INITIALIZATION AND THE REWARD LAW

The following estimate holds also under the history-dependent auxiliary law constructed later. It uses only actual softmax sampling, the amplitude bound, and conservation of $\sum_i \phi_i = 1$. We have

$$|\Delta\phi_i| \le a, \qquad \mathbb{E}_n|\Delta\phi_i| \le 2ap_i(1-p_i) \le 2ap_i.$$

For $x_{i,n} = e^{-\phi_{i,n}}$, use $|e^v - 1| \le e^a|v|$ for $|v| \le a$, and AM–GM, to get

$$\mathbb{E}_n x_{i,n+1} \le x_{i,n} + c_a^{\mathrm{tail}}, \qquad c_a^{\mathrm{tail}} = \frac{2ae^a}{Ke^{1/K}}, \qquad e^{-\phi_i}p_i = \frac{1}{\sum_j e^{\phi_j}} \le \frac{e^{-1/K}}{K}.$$

For a fixed finite T, $x_{i,n} + c_a^{\mathrm{tail}}(T-n)$ is a nonnegative supermartingale for $0 \le n \le T$. Stop at the first crossing of e^y and use its bounded stopping-time inequality. Summing over i and using $\sum_i e^{-\phi_{i,0}} \le Ke^{1/4}$ yields

$$Z_T := \max_{n\le T,i}(-\phi_{i,n})_+, \qquad \Pr(Z_T > y) \le \min\{1, e^{\Xi_T - y}\}, \qquad \Xi_T = \log(Ke^{1/4} + 2ae^aT). \tag{30}$$

The notation Pr here denotes either law used in the proof. Also, for all $n \le T$,

$$|\phi_{i,n}| \le 1 + (K-1)Z_T, \qquad |X_n| \le 1 + KZ_T. \tag{31}$$

All variables at any fixed horizon are integrable; this also follows directly from bounded increments and finite deterministic initialization.

A.5.4 UNIFORM WARM PHASES AND PAID SWITCHES

Fix T and write $L = L_T$. Set

$$\begin{aligned}
J &= \max\{0, \lceil \log_2(3L/4)\rceil\}, \qquad \rho_j = \tfrac{3}{4}2^{-j},\\
h_j &= \log\frac{3}{2\rho_j} = \log 2 + j\log 2 \quad (0 \le j \le J),\\
H_j &= h_{j+1} + (j+1)^2, \qquad A_j = \frac{1-\rho_{j+1}}{1-\rho_j} \quad (0 \le j < J),\\
v_0 &= 0, \qquad v_{j+1} = A_j v_j + (A_j - 1)(H_j + 2a),\\
w_0 &= 1, \qquad w_{j+1} = \frac{w_j}{A_j}\left(1 + \frac{\log 2}{(j+1)^2}\right).
\end{aligned}$$

In particular $(2L)^{-1} < \rho_J \le L^{-1}$. Phase j is alive while $X > h_j$ and $V^{\rho_j} < v_j$. Failure causes permanent absorption of the analysis process. If $j < J$, an alive state with $X \ge H_j$ switches to $j+1$. Process all thresholds crossed by one actual update in order, with absorption taking precedence. There are no artificial observations or state changes. The initial state is alive by (25), and $X_0 \le 5/4 < H_0$ when $J > 0$.

The intercept recursion divided by $1 - \rho_{j+1}$ telescopes, and

$$\frac{\rho_j - \rho_{j+1}}{(1-\rho_j)(1-\rho_{j+1})} \le 6\,2^{-j}.$$

The sums of 2^{-j}, $2^{-j}(j+1)$, and $2^{-j}(j+1)^2$ are $2, 4, 12$, respectively. Hence

$$0 \le v_j \le v_* := 12h_0 + 24\log 2 + 72 + 24a, \qquad \tfrac{1}{4} \le w_j \le 4. \tag{32}$$

For the weights, use the telescoping A_j product and

$$\prod_{n\ge 1}(1 + \log 2/n^2) \le \exp\left(\log 2 \sum_{n\ge 1} n^{-2}\right) \le 4.$$

Here the elementary bound $\sum_{n\ge 1} n^{-2} \le 2$ suffices.

Let $F_j = (X - h_j)_+(v_j - V^{\rho_j})_+$. Every processed crossing has $X \le H_j + 2a$, since $|\Delta X| \le 2a$ and the later thresholds are larger. The identity

$$V^{\rho_{j+1}} = A_j V^{\rho_j} + (A_j - 1)X$$

and the intercept recursion give the exact switch slack

$$(v_{j+1} - V^{\rho_{j+1}}) - A_j(v_j - V^{\rho_j}) = (A_j - 1)(H_j + 2a - X) \geq 0.$$

Here $A_j > 1$ and $X \leq H_j + 2a$ at every processed crossing, including later thresholds crossed by the same update. At the same state, $(X - h_j)/(X - h_{j+1}) \leq 1 + \log 2/(j+1)^2$. Thus

$$w_{j+1}F_{j+1} \geq w_j F_j. \tag{33}$$

This controls overshoots for arbitrary admissible controllers.

Write $H_{\max} = 0$ if $J = 0$, and $H_{\max} = H_{J-1}$ otherwise. Define

$$z_{\mathrm{pre}} = \left[1 + \frac{2}{m}\exp\left\{\frac{4K(v_* + H_{\max}) - 1}{m}\right\}\right]^{-1}. \tag{34}$$

At a pre-action state in a nonfinal phase, $\phi_2 = (V^{\rho_j} + X)/(1 - \rho_j) \leq 4(v_* + H_{\max})$ and $\phi_1 < \phi_2$. The external exponential sum is at least $m\exp((1 - \phi_1 - \phi_2)/m)$ by AM–GM. Comparing it to $e^{\phi_1} + e^{\phi_2}$ gives $z \geq z_{\mathrm{pre}}$. Suppose

$$\delta \leq z_{\mathrm{pre}}/16. \tag{35}$$

Then $r_o \geq 4\delta$. Moreover, throughout an alive phase, $q := p_1/p_2 = e^{-X} < 2\rho_j/3 \leq 1/2$. Equations (27)–(28) imply

$$\mathbb{E}_n \Delta X \geq \eta_n(p_2 r_o/2 - 2\delta p_1) \geq 0, \tag{36}$$

$$\mathbb{E}_n \Delta V^{\rho_j} \leq \eta_n(-\rho_j p_2 r_o/3 + 2\delta p_1) \leq 0. \tag{37}$$

Thus the wider opening still gives the required drift signs.

In the final phase put $\rho = \rho_J$, $v = v_J$. It has no upper height. Its first alive state satisfying

$$r_o \leq \frac{6\delta}{\rho}e^{-X} \tag{38}$$

is the cold entry σ. If the process is absorbed before entry, set $\sigma = \infty$. Before entry, the strict reverse of (38) implies the two drift signs in (36)–(37): for V substitute $r_o > 6\delta q/\rho$ into its negative term; for X this gives $p_2 r_o/2 - 2\delta p_1 > \delta p_1(3/\rho - 2) > 0$. All decisions are pre-action measurable.

A.5.5 Cold-entry geometry and the auxiliary law

Let

$$\alpha = \frac{m(1-\rho)}{K - 1 + \rho}, \qquad \kappa = \frac{\alpha}{1-\alpha} = \frac{m(1-\rho)}{1 + (K-1)\rho}, \qquad c_y = \frac{K}{1 + (K-1)\rho} \geq 1.$$

On finite entry let $s_\sigma = v - V^\rho_\sigma > 0$ and define

$$\Gamma_\rho = \max\left\{1, m^{-\alpha}\exp\left(\frac{Kv - (1-\rho)}{K - 1 + \rho}\right)\right\}, \qquad D_\rho = (3\Gamma_\rho)^{1/(1-\alpha)}(24/\rho)^\kappa.$$

Conservation of the normalized coordinate sum gives the exact identity

$$(K - 1 + \rho)\log p_1 = K(v - s_\sigma) - (1-\rho) + [1 + (K-1)\rho]\log p_2 + (1-\rho)\sum_{j \geq 3}\log p_j.$$

Drop the nonpositive $\log p_2$ term and use AM–GM on the external probabilities. This yields $p_1 \leq \Gamma_\rho e^{-Ks_\sigma/(K-1+\rho)} z^\alpha$. By (38),

$$z \leq 24\delta q/\rho < 16\delta \leq \tfrac{1}{2}, \qquad q = e^{-X_\sigma} < 2\rho/3 \leq \tfrac{1}{2}, \qquad p_2 = \frac{1-z}{1+q} \geq \tfrac{1}{3}.$$

Consequently $q \leq 3\Gamma_\rho e^{-Ks_\sigma/(K-1+\rho)}(24\delta q/\rho)^\alpha$. Divide by $q^\alpha > 0$ and raise to $1/(1-\alpha)$ to obtain

$$e^{-X_\sigma} \leq D_\rho \delta^\kappa e^{-c_y s_\sigma}. \tag{39}$$

Define $\mathbb{Q}_\delta$ on the same histories and with the same initialization: keep all control and action kernels unchanged; keep rewards unchanged before σ; from σ onward change only arm 1 to $\mathrm{Ber}(1/2)$. Once

set, this flag stays set even after analysis absorption. This is an auxiliary history-dependent law, *not* the IID environment used in the theorem. Set

$$\ell_\sigma = 1 + (c_y/2)s_\sigma, \qquad h_{\mathrm{c}} = X_\sigma - \ell_\sigma.$$

Suppose also

$$\kappa \log(1/\delta) - \log D_\rho - 1 \geq h_0. \tag{40}$$

Equation (39) then implies $h_{\mathrm{c}} \geq h_0$. In the cold phase survival means $X > h_{\mathrm{c}}$, and the potential is

$$W_n = \frac{w_J(X_\sigma - h_J)s_\sigma}{\ell_\sigma}(X_n - h_{\mathrm{c}})_+, \tag{41}$$

permanently set to zero at cold absorption. At entry this equals the preceding weighted warm potential. Under $\mathbb{Q}_\delta$, the contrast drift while cold and alive is $\eta_n(p_2 - p_1)r_o \geq 0$. Convexity of the positive part gives the conditional submartingale inequality, with the entry coefficient kept inside the history. While cold and alive, (39) also gives

$$p_1 \leq e^{-X_n} \leq eD_\rho \delta^\kappa e^{-c_y s_\sigma/2} \leq eD_\rho \delta^\kappa. \tag{42}$$

These geometric bounds hold at any alive cold history under either law.

A.5.6 Uniform survival and its quadratic envelope

Let τ be the first absorbed state. Combine the warm potentials $w_j F_j$ with (41), and put $W_n = 0$ after absorption. More precisely, define W_n at the state after processing all thresholds and the cold-entry test, before drawing action $n+1$. On an absorbed history $W_{n+1} = W_n = 0$. On a warm alive history in phase j, first take one actual update and evaluate the old-phase positive product. If either old-phase factor is nonpositive, this product is zero and absorption is permanent. Otherwise every crossed threshold satisfies $X_{n+1} \geq H_j > h_{j+1}$ and $v_{j+1} - V_{n+1}^{\rho_{j+1}} \geq A_j(v_j - V_{n+1}^{\rho_j}) > 0$. Thus a legal switch leaves the next phase alive, and (33) increases its weighted potential. Repeating this argument handles all thresholds crossed by the same update; the number of repetitions is at most $J - j$. A subsequent cold-entry test either leaves the warm potential unchanged or replaces it by the equal value in (41). Hence the processed W_{n+1} is at least the old-phase post-update positive product times w_j. Its conditional mean is at least W_n by (29) and the warm drift signs. Under $\mathbb{Q}_\delta$ the reward law at this action is still the original law whenever the processed state is warm.

On a cold alive history, its entry coefficient and h_{c} are $\mathcal{H}_n$-measurable and nonnegative (the coefficient is positive). Convexity and the nonnegative cold contrast drift give $\mathbb{E}_{\mathbb{Q}_\delta}[W_{n+1} \mid \mathcal{H}_n] \geq W_n$; hitting $X_{n+1} \leq h_{\mathrm{c}}$ produces zero, with no later revival. The random entry coefficient remains inside this conditional expectation. These cases partition all histories, including zero increments and equality at an absorption boundary. The finite-horizon bounded-increment bound makes every term integrable. Iterating the conditional inequality from $n = 0$ to $T - 1$ proves

$$\mathbb{E}_{\mathbb{Q}_\delta} W_T \geq W_0 \geq f_*. \tag{43}$$

This remains valid if there is no cold entry, or several phases are crossed by a single update. No entrance event is conditioned on and no probability of a prescribed action sequence is needed.

For the warm potential, (32) and (31) bound its two factors by $K(1 + Z_T)$ and $(v_* + 2K)(1 + Z_T)$, respectively. For the cold potential, $s_\sigma/\ell_\sigma \leq 2/c_y \leq 2$, so its entry coefficient is at most $8|X_\sigma|$, and its last factor is at most $|X_T|$ since $h_{\mathrm{c}} > 0$. Thus, with

$$C_W = 128K^3(1 + v_*), \qquad 0 \leq W_T \leq C_W(1 + Z_T)^2 \mathbf{1}_{\{\tau > T\}}. \tag{44}$$

There is no separate integrability assumption about a random cold-entry coefficient.

For an event E of probability $p > 0$, split the layer-cake integral at $y_* = \Xi_T + \log(1/p)$. Equation (30) gives

$$\begin{aligned}\mathbb{E}[(1 + Z_T)^2 \mathbf{1}_E] &\leq p(1 + y_*)^2 + \int_{y_*}^{\infty} 2(1 + y)e^{\Xi_T - y}\, dy \\ &= p\{(1 + y_*)^2 + 2(y_* + 2)\} \leq 5p[A_T + \log(1/p)]^2, \qquad A_T = 2 + \Xi_T.\end{aligned}$$

Choose the least integer $N_* \geq 1$ with

$$5C_W e^{-N_*}(N_* + 3)^2 \leq f_*/2. \tag{45}$$

It exists. For $A_T \geq 2$, the function $p \mapsto p[A_T + \log(1/p)]^2$ is nondecreasing on $(0, 1]$. If $\mathbb{Q}_\delta(\tau > T) \leq e^{-N_*}/A_T^2$, the last bound and (44) imply

$$\mathbb{E}_{\mathbb{Q}_\delta} W_T \leq 5C_W e^{-N_*} \left(1 + \frac{N_* + 2\log A_T}{A_T}\right)^2 \leq 5C_W e^{-N_*}(N_* + 3)^2 \leq f_*/2,$$

contradicting (43). The case of zero survival probability contradicts (43) directly. Hence

$$\mathbb{Q}_\delta(\tau > T) \geq \frac{e^{-N_*}}{A_T^2} \geq \frac{e^{-N_*}}{C_A^2 L^2}, \qquad C_A = 5 + \log K + 3a. \tag{46}$$

For the last inequality, note $Ke^{1/4} + 2ae^a T \leq Ke^{1/4+3a}T$ for $T \geq 1$: $2ae^a \leq e^{3a} - 1$ and $Ke^{1/4} \geq 1$. The displayed value of C_A is deliberately generous. All constants in (46) are uniform over $\mathcal{N}$ and the full controller class, provided (35) and (40) hold.

A.5.7 One common hard gap and an explicit uniform onset

The phase bounds and $(2L)^{-1} < \rho \leq L^{-1}$ give

$$D_\rho \leq D_a L^m, \quad D_a = 3^{K-1} e^{Kv_*} 48^m, \qquad 0 \leq m - \kappa = \frac{mK\rho}{1 + (K-1)\rho} \leq \frac{mK}{L}. \tag{47}$$

Indeed $1/(1-\alpha) \leq K - 1$, and the exponential part of $\Gamma_\rho^{1/(1-\alpha)}$ has denominator $1 + (K-1)\rho \geq 1$; the remaining factor is at most $(48L)^m$. Define

$$E_K = e^{(m+1)^2}, \qquad d = \min\left\{\frac{1}{32}, \sqrt{\frac{3}{128 e D_a E_K}}\right\}, \qquad \delta_T = dT^{-1/K} L^{-m/K}. \tag{48}$$

If $\xi = m - \kappa$, then

$$D_\rho T \delta_T^{\kappa+2} \leq D_a d^{\kappa+2} T^{\xi/K} L^{m\xi/K} \leq D_a E_K d^2. \tag{49}$$

To see the last inequality, use $d \leq 1$, $\xi \leq mK/L$, $\log T \leq L$, and $\log L/L \leq 1/e$. The extra exponential factor is at most $e^{m+m^2/e} \leq E_K$.

Here is an explicit onset verifying both domain conditions rather than assuming them. The threshold definition implies $H_{\max} \leq h_0 + J\log 2 + J^2 \leq 40\sqrt{L}$ for $L \geq 1$. For example use $J \leq 1 + \log L/\log 2$, $\log L/\sqrt{L} \leq 2/e$, and $(\log L)^2/\sqrt{L} \leq 16/e^2$. Consequently

$$z_{\mathrm{pre}}/16 \geq \frac{m}{16K} e^{-4Kv_*/m - 160K\sqrt{L}/m}, \qquad \delta_T \leq e\, e^{-L/K}.$$

Therefore (35) holds whenever

$$\sqrt{L} \geq 320K^2/m, \qquad L \geq 2K\{4Kv_*/m + \log(16Ke/m)\}.$$

When $L \geq 2K$ and $\sqrt{L} \geq 4K$, we have $\kappa \geq m/2$ and $\log L \leq L/(4K)$. Equations (47)–(48) give

$$\kappa \log(1/\delta_T) - \log D_\rho - 1 \geq \frac{mL}{4K} - \frac{m}{2K} - \log D_a - 1.$$

Thus (40) holds if $L \geq (4K/m)\{h_0 + m/(2K) + \log D_a + 1\}$. One permissible choice is

$$\begin{aligned} L_0 = \max\Big\{ & 1,\ 2K,\ 16K^2,\ (320K^2/m)^2, \\ & 2K\{4Kv_*/m + \log(16Ke/m)\}, \\ & (4K/m)\{h_0 + m/(2K) + \log D_a + 1\} \Big\}, \qquad T_0 = \lceil e^{L_0 - 1} \rceil. \end{aligned} \tag{50}$$

These are finite, controller-independent, initialization-independent constants. They are not proposed as sharp or practically predictive onset bounds.

A.5.8 STOPPED-OCCUPATION TRANSFER AND CONCLUSION

Use Lemma A.6; its sufficient condition $\epsilon T \le 1/8$ includes the $1/16$ budget below.

Apply the lemma for each fixed actual initialization ϑ and controller, with $\mathbb{Q} = \mathbb{Q}_{\delta_T}$ and $\mathbb{P} = \mathbb{P}_{\delta_T}$. Before cold entry the laws are identical. During the cold suffix their conditional KL, before absorption, is

$$p_{1,n}\,\mathrm{kl}(1/2, 1/2+\delta_T) \le \frac{8e}{3} D_\rho \delta_T^{\kappa+2} =: \epsilon.$$

Here $\mathrm{kl}(1/2, 1/2+\delta) = -\frac{1}{2}\log(1-4\delta^2) \le (8/3)\delta^2$ for $\delta \le 1/32$; use $-\log(1-x) \le x/(1-x)$. Control and action likelihoods cancel at the *same observed history*; this does not compare trajectories or controllers at two nearby initializations. For clarity, let $I^{[n]}, Y^{[n]}$ denote the action and reward between $\mathcal{H}_n$ and $\mathcal{H}_{n+1}$, and let $f_\delta(y) = (1/2+\delta)^y(1/2-\delta)^{1-y}$ for $y \in \{0,1\}$. For any deterministic $r \le T$, the finite-history density is

$$\frac{d\mathbb{Q}_\delta|_{\mathcal{H}_r}}{d\mathbb{P}_\delta|_{\mathcal{H}_r}} = \prod_{0\le n<r} \left\{ \frac{1/2}{f_\delta(Y^{[n]})} \right\}^{\mathbf{1}_{\{\sigma\le n,\ I^{[n]}=1\}}}.$$

The common initial state, controller randomization kernels and action kernels contribute no density factor. For the stopped record in Lemma A.6, the product ends at $r = U \wedge \tau$; the independent uniform U also has the same law. The full-history product can continue after absorption, since the reward switch remains active, but those later factors are absent from this stopped record. Empty products equal one. Taking the expected log of the stopped product gives precisely the occupation-weighted KL used in that lemma. Equation (49) and the choice of d ensure $\epsilon T \le 1/16$.

Every alive warm or cold state has $X > h_0 = \log 2$, so $p_1/p_2 < 1/2$, whence $p_1 < 1/3$. Under the actual IID law the predictable pseudo-regret is at least $\delta_T(1-p_1) \ge 2\delta_T/3$. It follows that

$$\begin{aligned} R_T^{\mathcal{A},\vartheta}(\mathcal{E}_{\delta_T}) &\ge \frac{2\delta_T}{3}\mathbb{E}_{\mathbb{P}_{\delta_T}}(T\wedge\tau) \\ &\ge \frac{\delta_T}{3}\mathbb{E}_{\mathbb{Q}_{\delta_T}}(T\wedge\tau) \ge \frac{\delta_T T}{3}\mathbb{Q}_{\delta_T}(\tau > T) \\ &\ge \frac{de^{-N_*}}{3C_A^2}\frac{T^{1-1/K}}{L_T^{2+m/K}}. \end{aligned}$$

Since $2 + m/K = 3 - 2/K$, Theorem A.7 holds with

$$c = \frac{de^{-N_*}}{3C_A^2} > 0. \tag{51}$$

The loss is summed against the predictable event $\{n < \tau\}$; we have not assumed that the sampler remains unchanged after conditioning on a future survival event. The actual final environment is exactly the IID law (23). The same d, c, T_0 work for every member of $\mathcal{N}$ and every controller, so taking first the environment supremum and then the controller infimum preserves the result. □

Proof of Theorem A.1. For $K = 2$, combine Lemma A.2 and Theorem A.4 at the same fixed initialization. For $K \ge 3$, Theorem A.5 belongs to the stated fixed-resource class. That class is a subset of the amplitude class with $a = u(1 + B_{\max})$, while the Bernoulli hard family belongs to $[0,1]$. Theorem A.7 therefore supplies the lower side of (9). Taking logarithms and dividing by $\log T$ gives the exponent $1 - 1/K$, since $\log L_T / \log T \to 0$. This operation does not remove the unresolved logarithmic gap. □

B EXPLICIT GEOMETRIC RECOVERY: COMPLETE FINITE-TIME PROOF

This section supplies the complete proof behind main-text Theorem 3: comparison drift and joint cap control yield one master regret inequality, from which both rates follow for one gap- and horizon-independent process. The proof covers the full parameter family; the main text uses its inverse-square-root endpoint. Fix $K \ge 2$, $0 < \gamma \le 1/2$, $1/2 < \omega \le 1$, a scalar ceiling $u > 0$, a fixed design scalar $0 < h \le u$, and a deterministic finite initialization $\theta_1 = \vartheta$. Rewards lie in $[0,1]$ and have

independent pull-indexed IID streams, with unique best arm j. All conditional expectations below are with respect to the pre-action history $\mathcal{F}_t$; write $\mathbb{E}_t$ for this operation. In particular,

$$r_t = \sum_{i \neq j} \Delta_i p_{i,t}, \qquad R_T = \mathbb{E} \sum_{t=1}^{T} \Delta_{I_t} = \sum_{t=1}^{T} \mathbb{E} r_t.$$

No conditional action law is changed by conditioning on a future event.

B.1 Rule and well-definedness

Endpoint dictionary for the main text. To check the main-text process, set $\gamma = 1/2$ throughout this appendix: $\beta = 1/2$, $a = \nu = 1$, $d = 0$, and $c_\gamma = 1/2$. Then $\eta_t = \kappa_t$; the tail-mass correction is one, $C_\gamma(p) = (\sum_i \sqrt{p_i})(\sum_i p_i^{-1/2})$, and $B_i = U_i = W_i = \sum_{k \neq i} \sqrt{p_k/p_i}$. Thus the main-text potential W_j and running cap are exactly the endpoint of the definitions below, not additional constructions.

Use the abbreviations

$$\beta = 1 - \gamma, \qquad a = \beta/\gamma, \qquad \nu = (2\gamma)^{-1}, \qquad d = \tfrac{1}{2} - \gamma,$$
$$c_\gamma = \nu\beta = a/2, \qquad L_t = 1 + \log t.$$

Here $a, \nu \geq 1$, $d \geq 0$ and $h > 0$ are fixed design quantities, not unknown environment parameters. For a strictly positive probability vector set

$$S_\gamma(p) = \sum_i p_i^\gamma, \qquad q(p) = 1 - \max_i p_i,$$
$$H_\gamma(p) = S_\gamma(p)^a \sum_i p_i^{-\beta}, \qquad C_\gamma(p) = H_\gamma(p)^{1/(2\beta)}. \tag{52}$$

At round t, compute $p_t = \mathrm{softmax}(\theta_t)$, update the running maximum $\overline{C}_t = \max_{s \leq t} C_\gamma(p_s)$, and choose

$$\kappa_t = \frac{h}{L_t^\omega(\sqrt{t} + \overline{C}_t)}, \qquad \eta_t = \kappa_t q(p_t)^{-d}, \qquad b_t = 1. \tag{53}$$

Then draw $I_t \sim p_t$, observe $Y_t^{(I_t)}$, put $\ell_t = 1 - Y_t^{(I_t)}$, and execute the explicit update

$$\theta_{i,t+1} = \theta_{i,t} - \eta_t \ell_t p_{i,t}^{-\gamma}(\mathbf{1}\{I_t = i\} - p_{i,t}). \tag{54}$$

Thus the only matrix is the specified diagonal D_γ; there is no mirror root, logit reset, forced action, or extra exploration distribution. The rule never inputs j, a gap, a variance, or a final horizon.

Since $S_\gamma(p) \geq 1$, $C_\gamma(p) \geq p_{\min}^{-1/2}$. Moreover $p_{\min} \leq q(p)$, whence

$$\eta_t p_{\min,t}^{-\gamma} \leq \frac{h}{L_t^\omega}(p_{\min,t}/q(p_t))^d \leq h. \tag{55}$$

Every coordinate increment is at most h in absolute value and $0 < \eta_t \leq h \leq u$. Induction gives finite logits and strictly positive probabilities at every finite time. All controls are pre-action measurable. The finite-time deterministic logit bounds also justify every expectation of a state function used below. They are not used as regret constants. The sequence κ_t is nonincreasing, and integral comparison gives

$$\sum_{t \geq 1} \kappa_t^2 \leq h^2 \sum_{t \geq 1} \frac{1}{t(1 + \log t)^{2\omega}} \leq \mathcal{S}_\omega := h^2 \left(1 + \frac{1}{2\omega - 1}\right). \tag{56}$$

B.2 Constants and theorem

For each comparison arm define the positive finite functions

$$B_i(p) = S_\gamma(p) p_i^{-\gamma} - 1, \qquad U_i(p) = (1 + B_i(p))^a - 1, \qquad W_i(p) = U_i(p)^\nu.$$

The subscripted function B_i is not the common-baseline resource bound $B_{\max}$. Let $W_*^0 = \max_i W_i(p_1)$ and $C_*^0 = C_\gamma(p_1)$. The following deterministic constants specify the finite-time bound:

$$
\begin{aligned}
d_r &= 2\gamma^2\big[(r\nu a)^2 + r\nu a\big]\exp\big[(2\gamma r\nu a + 4\gamma)h\big], \quad r \in \{1,2,4\},\\
M_r &= \big[2^{r-1}\{1 + (W_*^0)^r e^{d_r \mathcal{S}_\omega}\}\big]^{1/r}, \quad r \in \{2,4\},\\
A_C &= \beta e^{2\beta h}(K-1)K^{\beta^2/\gamma}, \qquad Q_C = 2(1+K^\gamma),\\
B_C &= \tfrac{1}{2}a\gamma^2 e^{\beta h} Q_C, \qquad D_C = (A_C + B_C h)/(2\beta),\\
k_2 &= (1+e^h)D_C h, \qquad k_4 = 4e^{3h} D_C h, \qquad C_R = (C_*^0)^2 + k_4/2,\\
C_W &= a^\nu (K-1)^{\beta\nu} K^{\beta(a-1)\nu}, \qquad C_I = 2^{\nu-1},\\
A_0 &= C_W C_I M_2, \qquad A_1 = C_W C_I \sqrt{2M_2^2 + 2C_R M_4^2},\\
C_{\mathrm{init}} &= \frac{(1+C_*^0)W_*^0}{h c_\gamma}, \qquad C_{\mathrm{rate}} = \frac{1}{c_\gamma}\left[\frac{2A_1 + A_0}{h} + e^{ah} D_C A_0 + d_1 h A_0\right].
\end{aligned}
$$

They are finite and depend only on $K, \gamma, \omega, h, \vartheta$; u restricts the permitted choice $h \le u$. They do not depend on gaps, variances, a random entry time, or T. Their poor dependence as $\gamma \downarrow 0$ or $\omega \downarrow 1/2$ is not suppressed. No smallness condition $h \le \gamma/64$ is needed: all along-segment exponential factors are retained, rather than absorbed into a small-step approximation. The constants can grow rapidly with h; the theorem is not uniform over an unbounded range of h, nor does it authorize a history-dependent choice h_t.

Theorem B.1 (Finite-time explicit geometric recovery)**.** *The same process* (53)–(54) *satisfies, for every permitted IID instance and every integer* $T \ge 1$*,*

$$
R_T \le \min\left\{ T,\ C_{\mathrm{init}} + 2C_{\mathrm{rate}}\sqrt{(K-1)T}\,L_T^\omega,\ 2C_{\mathrm{init}} + C_{\mathrm{rate}}^2 L_T^{2\omega+1} \sum_{i\neq j} \Delta_i^{-1} \right\}. \tag{57}
$$

To prove the theorem, we first derive a regret-linked drift for W_j, then bound the running cap and its mixed moment with W_j. Summation by parts yields a single master regret inequality. Two different deterministic estimates of that inequality give the two rates.

B.3 Comparison-potential drift and moments

Freeze a state and write $v = \theta' - \theta$ for its actual increment. For $\Phi(\theta) = \sum_i e^{\gamma\theta_i}$, each action/reward outcome obeys

$$
\nabla\Phi(\theta)^\top v = -\gamma\eta\ell \sum_i e^{\gamma\theta_i} p_i^{-\gamma}(\mathbf{1}\{I = i\} - p_i) = 0.
$$

This cancels a directional derivative, not the full finite-step change: convexity instead implies $\Phi(\theta') \ge \Phi(\theta)$. Put $F_j = 1 + U_j = \Phi^a e^{-\beta\theta_j}$. Since $a\gamma = \beta$, $\nabla F_j^\top v = -\beta F_j v_j$. The softmax normalizer cancels exactly:

$$
F_j p_j^\beta = \Phi^a e^{-\beta\theta_j}\frac{e^{\beta\theta_j}}{(\sum_i e^{\theta_i})^\beta} = S_\gamma(p)^a.
$$

For $k = r\nu$, the exact action kernel and conditional reward means give

$$
\begin{aligned}
\mathbb{E}_t v_j &= \eta_t p_{j,t}^\beta r_t,\\
\mathbb{E}_t[\nabla {W_j^r}^\top v] &= -k\beta\eta_t S_\gamma(p_t)^a U_j^{k-1} r_t \le 0.
\end{aligned} \tag{58}
$$

For any j, $U_j \ge B_j \ge q^\gamma$, with $q = 1 - p_{\max}$: if j maximizes p, concavity gives $\sum_{i\neq j} p_i^\gamma \ge q^\gamma$ and $p_j^{-\gamma} \ge 1$; otherwise B_j contains a ratio at least one. Consequently $U_j^{\nu-1} \ge q^d$. Taking $r = 1$ in (58) yields the linear drift at most $-c_\gamma\kappa_t r_t$.

For $i \neq j$, the actual conditional square increment satisfies

$$
\mathbb{E}_t(v_i - v_j)^2 \le \eta_t^2\left[p_i^{1-2\gamma} + p_j^{1-2\gamma} - (p_i^{1-\gamma} - p_j^{1-\gamma})^2\right] \le 4\kappa_t^2. \tag{59}
$$

The first bound sums the nonnegative score squares with their actual action probabilities and uses $\ell^2 \le 1$, without equating different arms' reward second moments. For the second, put $\xi = 1 - 2\gamma \in [0, 1)$. If $p_{\max} \le 1/2$, the bracket is at most $2 \le 4q^\xi$. If neither coordinate maximizes p, both are at most q and the bracket is at most $2q^\xi$. In the remaining case write the two probabilities as $z = 1 - q$ and $x \le q$; the bracket is

$$z^\xi(1-z) + x^\xi(1-x) + 2z^{1-\gamma}x^{1-\gamma} \le q + q^\xi + 2q^{1-\gamma} \le 4q^\xi.$$

Now $\eta^2 q^\xi = \kappa^2$. This includes $\gamma = 1/2$, since $q > 0$.

Here is the remainder bound in full. For $a, k \ge 1$ and $x > 0$ let $f(x) = ((1+x)^a - 1)^k$. Differentiating and using $x(1+x)^{a-1} \le (1+x)^a - 1 \le ax(1+x)^{a-1}$ gives

$$0 \le xf'(x)/f(x) \le ka, \qquad 0 \le x^2 f''(x)/f(x) \le (ka)^2.$$

For the second inequality, separate $k(k-1)(xU'/U)^2 + kx^2U''/U$ and use $x^2U''/U \le a(a-1)$. For $B_j = \sum_{i \ne j} e^{\gamma(\theta_i - \theta_j)}$ set $\alpha_i^{(j)} = e^{\gamma(\theta_i - \theta_j)}/B_j$. They sum to one, and the chain rule and Cauchy give

$$D^2[f(B_j)](\theta)[v, v] \le \gamma^2((ka)^2 + ka) f(B_j) \sum_{i \ne j} \alpha_i^{(j)} (v_i - v_j)^2.$$

Along $\theta + sv$, $0 \le s \le 1$, the ratio of B_j is between $e^{-2\gamma h}$ and $e^{2\gamma h}$, the ratio of $f(B_j)$ is at most $e^{2\gamma kah}$, and each weight ratio is at most $e^{4\gamma h}$. Taylor's integral remainder, whose integration weight is $1/2$, combined with (59) therefore proves

$$\mathbb{E}_t W_{j,t+1}^r \le W_{j,t}^r + \mathbb{E}_t[\nabla W_j^{r\top} v] + d_r \kappa_t^2 W_{j,t}^r. \tag{60}$$

In particular

$$\mathbb{E}_t W_{j,t+1} \le W_{j,t} - c_\gamma \kappa_t r_t + d_1 \kappa_t^2 W_{j,t}. \tag{61}$$

For $r \in \{2, 4\}$, divide (60) by $\prod_{s \le t}(1 + d_r \kappa_s^2)$ after discarding the nonpositive linear term. This is a nonnegative supermartingale normalization; the denominator at step t is pre-action measurable. It lies in $[1, e^{d_r \mathcal{S}_\omega}]$, so

$$\sup_t \mathbb{E} W_{j,t}^r \le (W_*^0)^r e^{d_r \mathcal{S}_\omega}, \qquad \mathbb{E}(1 + W_{j,t})^r \le M_r^r. \tag{62}$$

For clarity, the pathwise estimate used below follows without dropping the subtraction in U_j. Put $f(x) = ((1+x)^a - 1)^\nu$. The coordinate increment bound gives $B_j(p_{t+1}) \le e^{2\gamma h} B_j(p_t)$. For $c \ge 1$ the logarithmic derivative just established yields

$$\log \frac{f(cx)}{f(x)} = \int_x^{cx} \frac{f'(s)}{f(s)}\, ds \le a\nu \log c, \qquad W_{j,t+1} \le e^{2\gamma h a \nu} W_{j,t} = e^{ah} W_{j,t}.$$

Here monotonicity of f handles the first inequality and $2\gamma\nu = 1$ the last identity. This is a pathwise bound for a potential multiplied by a correlated cap increment in Appendix B.6; it is distinct from the fourth-moment estimate used to control the mixed cap–potential moment.

B.4 Positive cap increments and cap moments

Write $Z_-(\theta) = \sum_i e^{-\beta\theta_i}$. The shift-invariant quantity $H_\gamma = \Phi^a Z_-$ satisfies $H_\gamma' \le e^{2\beta h} H_\gamma$ and hence $C_\gamma' \le e^h C_\gamma$. As $\Phi' \ge \Phi$,

$$(H_\gamma' - H_\gamma)_+ \le (\Phi')^a (Z_-' - Z_-)_+ + Z_-((\Phi')^a - \Phi^a). \tag{63}$$

Baseline one makes only the selected logit nonincreasing. Therefore

$$(Z_-' - Z_-)_+ \le \beta\eta\ell e^{\beta h} e^{-\beta\theta_I} p_I^{-\gamma}(1 - p_I).$$

Multiply by $(\Phi')^a \le e^{\beta h}\Phi^a$ and sum with the true action probabilities. The exponents cancel because $\beta + \gamma = 1$, giving

$$\mathbb{E}_t[(\Phi')^a (Z_-' - Z_-)_+] \le \beta e^{2\beta h} \eta S_\gamma(p)^a \sum_i (1 - p_i) \le A_C \eta.$$

For the other term, cancellation of the first derivative of Φ gives

$$\mathbb{E}_t(\Phi'/\Phi - 1) \le \tfrac{1}{2}\gamma^2 e^{\gamma h}\eta^2 \frac{\sum_i p_i^\beta(1-p_i)}{S_\gamma(p)}.$$

The fraction is at most $Q_C q^\beta$. To see this when $p_{\max} \le 1/2$, bound the numerator by K^γ and note $q^\beta \ge 1/2$; otherwise separate the largest coordinate and use the bound $[1 + (K-1)^\gamma]q^\beta$ on the numerator. Since $1 \le \Phi'/\Phi \le e^{\gamma h}$, the inequality $x^a - 1 \le a e^{(a-1)\gamma h}(x-1)$ on this interval implies

$$\mathbb{E}_t[Z_-((\Phi')^a - \Phi^a)] \le B_C \eta^2 q^\beta H_\gamma.$$

The exponent $1/(2\beta) \le 1$, so concavity applied to the positive increment in (63) gives

$$\mathbb{E}_t(C'_\gamma - C_\gamma)_+ \le \frac{1}{2\beta}\{A_C \eta C_\gamma^{-2d} + B_C \eta^2 q^\beta C_\gamma\}.$$

Here $\eta C_\gamma^{-2d} = \kappa(qC_\gamma^2)^{-d} \le \kappa$, while $\eta^2 q^\beta C_\gamma = \kappa^2 q^\gamma C_\gamma \le h\kappa$. Consequently

$$\mathbb{E}_t(\overline{C}_{t+1} - \overline{C}_t) \le D_C \kappa_t, \qquad \overline{C}_{t+1} \le e^h \overline{C}_t. \tag{64}$$

The running-maximum increment is bounded by $(C_{t+1} - C_t)_+$. Thus this argument controls a positive increment, not merely signed drift.

For $x \le x' \le e^h x$, use $x'^2 - x^2 \le (1+e^h)x(x'-x)$ and $x'^4 - x^4 \le 4e^{3h}x^3(x'-x)$. Since $\kappa_t \overline{C}_t \le h$, finite-time induction yields

$$\mathbb{E}\overline{C}_t^2 \le (C_*^0)^2 + k_2(t-1),$$
$$\mathbb{E}\overline{C}_t^4 \le (C_*^0)^4 + k_4\left[(t-1)(C_*^0)^2 + \frac{k_2(t-1)(t-2)}{2}\right] \le C_R^2 t^2.$$

The last inequality follows from $k_4 \ge 2k_2$ by comparison with $[(C_*^0)^2 + (k_4/2)t]^2$. No unbounded stopping-time moment is assumed.

B.5 Probability-weighted mixed moments

The comparison-arm tail $q_j = 1 - p_j$ need not equal the control tail q. Power-sum inequalities give

$$U_j \le aB_j(1+B_j)^{a-1}, \quad B_j \le (K-1)^\beta q_j^\gamma p_j^{-\gamma}, \quad 1 + B_j \le K^\beta p_j^{-\gamma}.$$

After taking the νth power,

$$W_j \le C_W \sqrt{q_j}\, p_j^{-a/2}, \qquad p_j^{-a/2} \le (1+U_j)^\nu \le C_I(1+W_j). \tag{65}$$

The second bound uses $1 + U_j \ge p_j^{-\beta}$ and convexity of the νth power. Write $s_t = \sum_{i\ne j}\sqrt{\mathbb{E}p_{i,t}}$. By $\sqrt{q_j} \le \sum_{i\ne j}\sqrt{p_i}$, Cauchy and (62),

$$\mathbb{E}W_{j,t} \le A_0 s_t. \tag{66}$$

The fourth moments above control the correlated cap/potential product:

$$\mathbb{E}[\overline{C}_t^2(1+W_{j,t})^2] \le \sqrt{\mathbb{E}\overline{C}_t^4\,\mathbb{E}(1+W_{j,t})^4} \le C_R t M_4^2.$$

Thus Cauchy gives

$$\mathbb{E}[(\sqrt{t} + \overline{C}_t)^2 p_j^{-a}] \le C_I^2\{2t\mathbb{E}(1+W_j)^2 + 2\mathbb{E}[\overline{C}_t^2(1+W_j)^2]\}$$
$$\le C_I^2 t(2M_2^2 + 2C_R M_4^2).$$

Apply Cauchy once more, separately to each $\sqrt{p_i}$ in (65), to obtain

$$\mathbb{E}[(\sqrt{t} + \overline{C}_t)W_{j,t}] \le A_1\sqrt{t}\, s_t. \tag{67}$$

These are unconditional estimates; cap and comparison potential are not assumed independent.

B.6 Calendar-time summation and both rates

Let $w_t = 1/\kappa_t = L_t^\omega(\sqrt{t} + \overline{C}_t)/h$. Multiplying (61) by this pre-action measurable weight and summing finite expectations gives

$$c_\gamma R_T \le \mathbb{E}\sum_{t=1}^{T} w_t(W_{j,t} - W_{j,t+1}) + d_1 \sum_{t=1}^{T} \mathbb{E}[\kappa_t W_{j,t}].$$

The exact pathwise summation-by-parts identity is

$$\sum_{t=1}^{T} w_t(W_t - W_{t+1}) = w_1 W_1 + \sum_{t=2}^{T}(w_t - w_{t-1})W_t - w_T W_{T+1}.$$

Only the final nonpositive term may be discarded. Split the middle increment as

$$\begin{aligned} w_t - w_{t-1} = \frac{L_t^\omega - L_{t-1}^\omega}{h}(\sqrt{t} + \overline{C}_t) \\ + \frac{L_{t-1}^\omega}{h}(\sqrt{t} - \sqrt{t-1}) + \frac{L_{t-1}^\omega}{h}(\overline{C}_t - \overline{C}_{t-1}). \end{aligned}$$

For $t \ge 2$, $L_t^\omega - L_{t-1}^\omega \le 2/t$ and $\sqrt{t} - \sqrt{t-1} \le t^{-1/2}$. Thus (66)–(67) bound the first two terms by $[(2A_1 + A_0)/h]L_t^\omega s_t/\sqrt{t}$ in expectation. The third term is correlated with W_t. Use its pathwise growth bound and then condition on $\mathcal{F}_{t-1}$:

$$\mathbb{E}[(\overline{C}_t - \overline{C}_{t-1})W_t] \le e^{ah} D_C \mathbb{E}[\kappa_{t-1} W_{t-1}].$$

Since $L_{t-1}^\omega \kappa_{t-1}/h \le (t-1)^{-1/2}$, its summed contribution is at most $e^{ah} D_C A_0 \sum_{t=1}^{T-1} s_t/\sqrt{t}$. The remaining drift-error contribution is at most $d_1 h A_0 \sum_{t=1}^{T} s_t/(L_t^\omega \sqrt{t})$. Using $W_j(p_1) \le W_*^0$, the stated constants therefore give

$$R_T \le C_{\mathrm{init}} + C_{\mathrm{rate}} \sum_{t=1}^{T} \frac{L_t^\omega}{\sqrt{t}} \sum_{i \ne j} \sqrt{\mathbb{E}p_{i,t}}. \tag{68}$$

This closes the cap and low-scalar costs without leaving a random entry or recovery assumption.

For the first rate, $s_t \le \sqrt{K-1}$ and $\sum_{t \le T} t^{-1/2} \le 2\sqrt{T}$ in (68) suffice. For the second let $x_{i,t} = \mathbb{E}p_{i,t}$. Twice the master bound minus the exact IID identity $R_T = \sum_{t,i \ne j} \Delta_i x_{i,t}$ gives

$$\begin{aligned} R_T &\le 2C_{\mathrm{init}} + \sum_{t, i \ne j} \left(2C_{\mathrm{rate}} L_t^\omega \sqrt{x_{i,t}/t} - \Delta_i x_{i,t} \right) \\ &\le 2C_{\mathrm{init}} + C_{\mathrm{rate}}^2 \sum_{i \ne j} \Delta_i^{-1} \sum_{t=1}^{T} L_t^{2\omega}/t \\ &\le 2C_{\mathrm{init}} + C_{\mathrm{rate}}^2 L_T^{2\omega+1} \sum_{i \ne j} \Delta_i^{-1}. \end{aligned}$$

For clarity, with $v = C_{\mathrm{rate}} L_t^\omega/\sqrt{t}$, $2v\sqrt{x} - \Delta_i x = v^2/\Delta_i - \Delta_i(\sqrt{x} - v/\Delta_i)^2 \le v^2/\Delta_i$. Thus the middle step introduces no additional factor of two in C_{rate}^2; the last uses $\sum_{t \le T} 1/t \le L_T$. Together with $R_T \le T$, this proves Theorem B.1. Both rates refer to the same rule and the same unconditional expectation. No adversarial guarantee, exact expected asymptotic coefficient, or uniform-in-γ constant is asserted.

C Matching interfaces and interpreting the separation

C.1 Classes and polynomial separation

Write $\mathfrak{E}_K$ for the reward class and $\mathcal{C}_0^T(u, B_{\max}; \vartheta)$ for the original-score rules specified in Appendix A; use its minimax value $\mathcal{M}_0(T)$. In particular, $u > 0$ and $B_{\max} \ge 1$ are fixed, the rule

may know T, and its pre-action history map is shared across unknown IID environments. Sampling, independent random sources, finite deterministic initialization and the no-reset restriction are unchanged. Replace only the score by $D_\gamma(p)(e_I - p)$, with $D_\gamma(p) = \mathrm{diag}(p_i^{-\gamma})$, to define $\mathcal{C}_\gamma^T$ and $\mathcal{M}_\gamma(T)$, leaving all other interface conditions unchanged. This is an explicit change of model class, not a scalar tuning inside $\mathcal{C}_0^T$.

For $K = 2$, Theorem A.1 gives $\mathcal{M}_0(T) = \Theta(\sqrt{T})$ at every fixed finite initialization ϑ; the constants may depend on that initialization. For $K \geq 3$, the same theorem gives

$$c\,T^{1-1/K}/L_T^{3-2/K} \leq \mathcal{M}_0(T) \leq CT^{1-1/K} \quad (K \geq 3)$$

uniformly over $\vartheta \in \mathcal{N}$ for sufficiently large T, with $K, u, B_{\max}$ fixed first. The bounds use the Bernoulli lower family and the gap-free upper policy proved in Appendix A. Taking logarithms shows $\log \mathcal{M}_0(T)/\log T \longrightarrow 1 - 1/K$. Only the two-arm case has no remaining logarithmic gap.

For fixed $\gamma \in (0, 1/2]$ and $\omega \in (1/2, 1]$, the process in Theorem B.1 belongs to $\mathcal{C}_\gamma^T$ for every T: its finite-state and scalar bounds were proved in (55), its baseline is one, and it uses only the pre-action history. Its C_{init} and C_{rate} are gap-independent for every fixed finite ϑ and, for $K \geq 3$, admit common bounds on $\mathcal{N}$, as shown below. The universal statistical bandit lower bound used in Theorem A.1 also applies to this subclass of all single-observation bandit algorithms. Hence

$$c\sqrt{T} \leq \mathcal{M}_\gamma(T) \leq C\sqrt{T}L_T^\omega \quad \text{for sufficiently large } T,$$

and the geometric minimax polynomial exponent is $1/2$. In particular, for $K \geq 3$ the concrete horizon-free DGB process satisfies

$$\frac{\mathcal{M}_0(T)}{\sup_{\mathcal{E}\in\mathfrak{E}_K} R_T^{\mathrm{DGB}}(\mathcal{E})} \geq c' \frac{T^{1/2-1/K}}{L_T^{3-2/K+\omega}} \longrightarrow \infty. \tag{69}$$

The denominator is positive by the statistical lower bound. All comparisons use the same initialization, reward class and resource ceilings; constants are uniform on $\mathcal{N}$. No supremum is interchanged with expectation or a random limit. The constant is not uniform as $\gamma \downarrow 0$, and no such separation is claimed for $K = 2$.

C.2 Uniform initial constants for the constructive comparison

Corollary C.1 (Uniform neighborhood separation). *Fix $K \geq 3$, $u > 0$, $B_{\max} \geq 1$, $0 < \gamma \leq 1/2$, $1/2 < \omega \leq 1$ and $0 < h \leq u$. The constants in* (57) *can be bounded by common finite $\overline{C}_{\mathrm{init}}, \overline{C}_{\mathrm{rate}}$ over $\vartheta \in \mathcal{N}$, independently of the IID instance and T. In particular, for all $T \geq T_0$,*

$$\inf_{\vartheta\in\mathcal{N}} \frac{\mathcal{M}_0(T;\vartheta)}{\sup_{\mathcal{E}\in\mathfrak{E}_K} R_T^{\mathrm{DGB},\vartheta}(\mathcal{E})} \geq \frac{c_-}{G_*} \frac{T^{1/2-1/K}}{L_T^{3-2/K+\omega}}, \qquad G_* = \overline{C}_{\mathrm{init}} + 2\overline{C}_{\mathrm{rate}}\sqrt{K-1}.$$

Proof. Every $\vartheta \in \mathcal{N}$ has span at most $5/4$, so its initial policy satisfies $p_{i,1} \geq p_* := e^{-5/4}/K$. Concavity gives $S_\gamma(p_1) \leq K^{1-\gamma} = K^\beta$. Using the notation in Appendix B.2, define

$$\overline{W} = \left[(K^\beta p_*^{-\gamma})^a - 1\right]^\nu, \qquad \overline{C} = \left[K^{\beta a} K p_*^{-\beta}\right]^{1/(2\beta)}.$$

Both are finite and positive, and the displayed definitions give $W_*^0 \leq \overline{W}$, $C_*^0 \leq \overline{C}$ uniformly in the initial policy and the identity of the best arm. Obtain $\overline{C}_{\mathrm{init}}, \overline{C}_{\mathrm{rate}}$ by replacing W_*^0, C_*^0 with $\overline{W}, \overline{C}$ in the full constant list of Appendix B.2. These replacements dominate the original constants: M_2, M_4 increase in W_*^0, C_R increases in C_*^0, and $A_0, A_1, C_{\mathrm{init}}, C_{\mathrm{rate}}$ are compositions of sums, products and square roots of nonnegative increasing terms. All remaining factors depend only on the fixed design parameters. This checks the actual formulas rather than inferring uniformity from pointwise finite constants.

Theorem B.1 therefore yields both stated rates uniformly on $\mathcal{N}$. Since $\sqrt{T}L_T^\omega \geq 1$, its gap-free bound is at most $G_*\sqrt{T}L_T^\omega$. Theorem A.1 supplies a common lower constant c_- and T_0 on the same set. Dividing the two inequalities for each ϑ and then taking the infimum proves the corollary. The denominator is positive by the initialization-independent statistical lower bound in Lemma A.2. No limit is interchanged with a supremum, infimum or expectation. The displayed deterministic lower bound diverges for fixed $K \geq 3$. □

The common constants still depend on h, γ, ω; this is not a uniform limit as $h \downarrow 0$ or $\gamma \downarrow 0$. The corollary adds no claim at uniform initialization or outside $\mathcal{N}$.

C.3 A NECESSARY STARTUP COST UNDER AN EXPLICIT DECOMPOSITION

Consider the difficult Bernoulli family $\mu_1 = 1/2 + \delta$, $\mu_2 = 1/2$, and $\mu_i = 1/4 - (i-3)/(16K)$ for $i \geq 3$, at $\vartheta^* = (0, 1, 0, \ldots, 0)$. The lower proof supplies constants $d_*, c_* > 0$, independent of the admissible controller, such that

$$\delta_T = d_* T^{-1/K} L_T^{-(K-2)/K}, \qquad R_T(\delta_T) \geq c_* T^{(K-1)/K} L_T^{-(3-2/K)} \tag{70}$$

for all sufficiently large T. In particular $0 < \delta_T \leq 1/16$ eventually.

Proposition C.2 (Startup cost for a low-gap-coefficient decomposition)**.** *Let $K \geq 3$. Suppose an admissible unknown-environment rule has a bound*

$$R_T(\delta) \leq D(\delta) + C\delta^{-s} L_T^q \quad (T \geq 1,\ 0 < \delta \leq 1/16),$$

where $0 \leq s < K - 1$, $q \geq 0$, $C < \infty$ are fixed, and D is finite and nonnegative. Then, along the sequence in (70)*,*

$$D(\delta_T) \geq c'' \, \frac{\delta_T^{-(K-1)}}{[\log(1/\delta_T)]^K} \quad \textit{for all sufficiently large } T.$$

Proof. The second term of the assumed upper bound at δ_T is $C d_*^{-s} T^{s/K} L_T^{q+s(K-2)/K}$. Dividing this term by the lower bound in (70) gives a negative power of T times a fixed power of L_T, which tends to zero since $s < K - 1$. It is therefore at most half that lower bound eventually. Thus

$$\begin{aligned} D(\delta_T) &\geq (c_*/2) T^{(K-1)/K} L_T^{-(3-2/K)} \\ &= \tfrac{1}{2} c_* d_*^{K-1} \delta_T^{-(K-1)} L_T^{-K}. \end{aligned}$$

The equality uses $3 - 2/K + (K-1)(K-2)/K = K$. Moreover

$$\log(1/\delta_T) = K^{-1} \log T + (K-2) K^{-1} \log L_T - \log d_*,$$

so $L_T / \log(1/\delta_T) \to K$. An eventual upper bound $2K$ for this ratio yields $c'' = c_* d_*^{K-1} / [2(2K)^K]$. □

The proposition does not assert that every pointwise logarithmic algorithm admits this decomposition. A larger gap dependence in its leading coefficient may pay the cost instead. Nor is a bound along δ_T a lower bound for the asymptotic logarithmic coefficient at one fixed positive gap.

C.4 TWO-DIMENSIONAL SCALAR EQUIVALENCE HAS A PRECISE BOUNDARY

Proposition C.3 (Projected diagonal score)**.** *Let $P = I - \mathbf{1}\mathbf{1}^\top / K$, let $D = \operatorname{diag}(d_1, \ldots, d_K)$, and fix an interior probability vector p. If the same scalar λ is to satisfy $PD(e_i - p) = \lambda(e_i - p)$ for every action i, then for $K = 2$ the identity always holds with $\lambda = (d_1 + d_2)/2$. For $K \geq 3$ it holds if and only if every d_i is equal.*

Proof. For $K = 2$ each score has form $(z, -z)$; projection of $D(z, -z)$ equals $((d_1 + d_2)z/2, -(d_1 + d_2)z/2)$. For $K \geq 3$, subtract the identities for i and j. At a coordinate $k \notin \{i, j\}$, the left side of $PD(e_i - e_j) = \lambda(e_i - e_j)$ is $-(d_i - d_j)/K$, while the right side is zero. Thus $d_i = d_j$ for every pair. Conversely a scalar matrix commutes with this projection and gives the required identity because $e_i - p$ sums to zero. □

For D_γ with $\gamma > 0$ and $K \geq 3$, absorption is possible at the uniform state only, not at every state. This obstructs universal actionwise matching at a nonzero update: baseline one and reward zero give a legal nonzero residual for every action. When the scalar is zero both updates vanish, and no obstruction is claimed. In two dimensions the algebra alone does not prove bounded-resource equivalence of whole algorithms, since the induced scalar factor can be unbounded near the boundary.

For the particular DGB controller of Theorem B.1, the required resource check is available. In two dimensions set $\lambda_t = (p_{1,t}^{-\gamma} + p_{2,t}^{-\gamma})/2$ and $\widetilde{\eta}_t = \eta_t \lambda_t$. Then

$$0 < \widetilde{\eta}_t \le \eta_t p_{\min,t}^{-\gamma} \le h \le u.$$

Run the original score with baseline one and this predictable scalar, using the DGB cap functional of its own probability history. Couple action seeds and arm-indexed rewards between the two rules. Starting from the same policy, Proposition C.3 says their next logits differ only by a common translation; thus their next softmax probabilities are identical. Induction gives the same probability, action and reward histories, so the cap functional and the next scalar are also matched. This is an exact-real policy equivalence for this specified two-arm controller, not an equivalence for all diagonal rules. In particular, a two-arm ablation using η_t instead of $\widetilde{\eta}_t$ also changes the effective scalar magnitude. Its empirical difference is not evidence of a multidimensional direction change. The polynomial separation in (69) remains restricted to $K \ge 3$.

C.5 Small positive exponents are not uniformly small perturbations

For fixed K and $0 < \alpha \le 1/K$, diagonal-matrix norms give

$$\sup_{p_i \ge \alpha} \|D_\gamma(p) - I\|_{\mathrm{op}} \le \alpha^{-\gamma} - 1 \longrightarrow 0 \quad (\gamma \downarrow 0).$$

On the whole open simplex, however, this supremum is infinite for every fixed $\gamma > 0$: set $p_1 = \varepsilon$ and $p_i = (1-\varepsilon)/(K-1)$ for $i > 1$, then send $\varepsilon \downarrow 0$. Consequently the positive theorem for each fixed arbitrarily small $\gamma > 0$ does not extend by continuity to $\gamma = 0$. This distinction is relevant both to the theorem constants and to finite-horizon experiments.

D Arbitrary initialization for one fixed anytime rule

This appendix proves the arbitrary-start result: after fixing any finite initialization and one horizon-free rule, a polynomial obstruction above the square-root scale remains, with rule-dependent constants and exponent. Joint tilting handles an unfavorable prefix. Nonuniform starts already supply one; uniform starts split into a no-update branch and a positive-probability finite prefix. The output is rule-dependent, not the common sharp exponent on the neighborhood in Theorem A.1. The auxiliary sampler is tilted here, unlike the unchanged sampler in that theorem's occupation-KL proof.

D.1 Statement, model and observation indexing

Fix $K \ge 3$, $u > 0$, $B_{\max} \ge 1$, and finite deterministic $\theta_1 = \vartheta \in \mathbb{R}^K$. A fixed horizon-free rule $\mathcal{A}$ chooses common controls $0 \le \eta_t \le u$, $|b_t| \le B_{\max}$ before action t and uses

$$p_t = \mathrm{softmax}(\theta_t), \qquad I_t \mid \mathcal{F}_t \sim p_t, \qquad \theta_{t+1} = \theta_t + \eta_t (Y_t - b_t)(e_{I_t} - p_t).$$

The rule has the same observed-history map in every unknown reward environment. Environment-independent internal randomization is allowed. For analysis it can be represented by a random tape included in the pre-action filtration; it reveals no action seed or unused reward. Current actions and unused rewards retain their conditional product law. The rule does not reset or project logits, force an action, change the sampler or score, or use an arm-specific scalar or post-action baseline. Both resource bounds are fixed as the gap and horizon vary.

To keep the global convention $\theta_1 = \vartheta$, define local states and probabilities after n observations by

$$\psi_n = \theta_{n+1}, \qquad \pi_n = p_{n+1} = \mathrm{softmax}(\psi_n), \qquad \mathcal{H}_n = \mathcal{F}_{n+1} \quad (n \ge 0).$$

Thus $\psi_0 = \vartheta$. The nth transition below uses the global action I_{n+1} and controls η_{n+1}, b_{n+1}, and

$$\psi_{n+1} - \psi_n = \eta_{n+1}(Y_{n+1} - b_{n+1})(e_{I_{n+1}} - \pi_n). \tag{71}$$

Put $M = 1 + B_{\max}$ and $S_0 = \sum_i \vartheta_i$. Every coordinate increment has absolute value at most uM, and $\sum_i \psi_{i,n} = S_0$ at every observation index.

Theorem D.1 (Arbitrary initialization, each fixed anytime rule)**.** *For every initialization and rule just specified, there is a fixed relabeling of the arms under which the independent Bernoulli family*

$$\mu_1(\delta) = \tfrac{1}{2} + \delta, \qquad \mu_2 = \tfrac{1}{2}, \qquad \mu_j = \tfrac{1}{4} - \frac{j-3}{16K} \quad (3 \le j \le K), \qquad 0 < \delta \le \tfrac{1}{8} \tag{72}$$

satisfies, for some $c_{\mathcal{A}} > 0$, $T_{\mathcal{A}} < \infty$ and $\zeta_{\mathcal{A}} > 1/2$,

$$\sup_{0<\delta\le 1/8} R_T^{\mathcal{A}}(\delta) \ge c_{\mathcal{A}} T^{\zeta_{\mathcal{A}}} \qquad (T \ge T_{\mathcal{A}}). \tag{73}$$

The constants and exponent may depend on the fixed rule, initialization, $K, u, B_{\max}$. They are independent of the subsequently varied δ, T. The family and relabeling are also fixed before varying T. The regret is unconditional expected pseudo-regret.

Each positive-gap family member has a unique best arm, overlapping Bernoulli support and strictly positive variances. Relabeling assigns the family to the actual arm identities; it does not modify the controller's history map. The theorem does not place an infimum over different horizon-indexed rules outside the displayed supremum. It does not replace $\zeta_{\mathcal{A}}$ by $1 - 1/K$.

The proof first establishes a lower bound after a fixed, positive-probability unfavorable prefix. Its ingredients are a single normalized joint tilt, two stopped exponential supermartingales, moments under the genuinely tilted action kernel, a paid prefix likelihood, an actual-count cutoff and a self-consistency inequality for worst-case regret. A final case split constructs such a prefix from each finite initialization.

D.2 The real family and finite-history continuity

For the moment fix one labeling in (72). Let $\mathbb{P}_\delta$ be the history law of the same rule in that family, and let $\mathbb{P}_0$ be the reference law with arm 1 also $\mathrm{Ber}(1/2)$. Only this reference has two tied best arms. For $j \ge 3$ put $d_j = 1/2 - \mu_j \ge 1/4$. Define

$$N_i(T) = \sum_{n=0}^{T-1} \mathbf{1}\{I_{n+1} = i\}, \quad Q_i(T) = \sum_{n=0}^{T-1} \pi_{i,n},$$
$$z_n = \sum_{j=3}^{K} \pi_{j,n}, \qquad Q_o(T) = \sum_{n=0}^{T-1} z_n.$$

These $N_i(T)$ are the global actual counts. On every path,

$$\widehat{R}_T(\delta) = \delta\{T - N_1(T)\} + \sum_{j=3}^{K} d_j N_j(T), \qquad R_T(\delta) = \mathbb{E}_\delta \widehat{R}_T(\delta).$$

Set

$$\mathcal{R}_T = \sup_{0<\delta\le 1/8} R_T(\delta), \qquad \mathcal{O}_T = \mathbb{E}_0 Q_o(T).$$

Conditional sampling gives $\mathbb{E}_\delta N_i(T) = \mathbb{E}_\delta Q_i(T)$; we will not apply this identity after conditioning on a future event.

Lemma D.2 (Continuity at the tie reference)**.** *For each fixed finite T,*

$$R_T(\delta) \longrightarrow R_T(0) \quad (\delta \downarrow 0), \qquad \mathcal{R}_T \ge R_T(0) \ge \tfrac{1}{4}\mathcal{O}_T.$$

Proof. At the same history and internal random tape, both environments use the same controls and action probabilities. Thus the finite-history density is

$$\frac{d\mathbb{P}_\delta^T}{d\mathbb{P}_0^T} = \prod_{\substack{0\le n<T \\ I_{n+1}=1}} (1+2\delta)^{Y_{n+1}} (1-2\delta)^{1-Y_{n+1}}.$$

It lies between $(1-2\delta)^T$ and $(1+2\delta)^T$, so converges uniformly to one at fixed T. The pathwise costs at δ and zero differ by at most δT, and the costs are bounded at this fixed horizon. This proves continuity without assuming continuity of the controller as a function of its observations. At zero gap the remaining cost is $\sum_{j\ge 3} d_j N_j(T)$; unconditional sampling compensation and $d_j \ge 1/4$ give the last inequality. □

D.3 A NORMALIZED JOINT TILT AND TWO SUPERMARTINGALES

Fix $0 < \rho < 1$, let $h = \log(1/\rho) > 0$, and define two coordinates of the actual state:

$$X_n = \psi_{2,n} - \psi_{1,n}, \qquad V_n = \psi_{1,n} - \rho\psi_{2,n}.$$

The region $X_n \geq h$ is exactly $\pi_{1,n} \leq \rho\pi_{2,n}$. We write $\mathbb{E}_{0,n}$ for expectation under $\mathbb{P}_0$ conditional on $\mathcal{H}_n$. For a one-step calculation, suppress the index n, write $\pi = \pi_n$, $I = I_{n+1}$ and $Y = Y_{n+1}$, write $\Delta X = X_{n+1} - X_n$, $\Delta V = V_{n+1} - V_n$, and set $r_o(\pi) = \sum_{j\geq 3} d_j\pi_j$. Averaging (71) over its original action/reward kernel gives $\mathbb{E}_{0,n}\Delta\psi_i = \eta_{n+1}\pi_i(\mu_i - \sum_j \pi_j\mu_j)$. Hence

$$\begin{aligned} \mathbb{E}_{0,n}\Delta X &= \eta_{n+1}(\pi_2 - \pi_1)r_o(\pi) \geq 0, \\ \mathbb{E}_{0,n}\Delta V &= \eta_{n+1}(\pi_1 - \rho\pi_2)r_o(\pi) \leq 0 \quad \text{when } X_n \geq h. \end{aligned} \tag{74}$$

The baseline cancels only from these conditional first moments.

The actual score coefficients of the two coordinates are

$$\mathsf{s}_i^X = \mathbf{1}\{i = 2\} - \mathbf{1}\{i = 1\} - (\pi_2 - \pi_1), \qquad \mathsf{s}_i^V = \mathbf{1}\{i = 1\} - \rho\mathbf{1}\{i = 2\} - (\pi_1 - \rho\pi_2).$$

Thus $\Delta X = \eta_{n+1}(Y - b_{n+1})\mathsf{s}_I^X$ and $\Delta V = \eta_{n+1}(Y - b_{n+1})\mathsf{s}_I^V$. Within the region, for every possible action,

$$\begin{aligned} &\mathsf{s}_i^V = -k_i\mathsf{s}_i^X, \qquad 0 \leq k_i \leq 1, \\ &k_1 = \frac{1 - \pi_1 + \rho\pi_2}{1 + \pi_2 - \pi_1}, \qquad k_2 = \frac{\pi_1 + \rho(1 - \pi_2)}{1 - \pi_2 + \pi_1}, \\ &k_j = \frac{\rho\pi_2 - \pi_1}{\pi_2 - \pi_1} \quad (j \geq 3). \end{aligned} \tag{75}$$

All denominators are positive because $\pi_1 < \pi_2$. Nonnegativity of the first two numerators is direct; their upper bounds follow by replacing ρ by one. The last ratio lies in $[0, 1]$ from $\pi_1 \leq \rho\pi_2 \leq \pi_2$. In particular $\Delta V = -k_I\Delta X$ holds for each actual reward, regardless of its residual's sign.

Fix $\lambda > 0$. Let P_i^0 denote the reference Bernoulli reward law of arm i. The following kernel changes the joint action and reward law:

$$\begin{aligned} \mathsf{Z}_n^\lambda &= \mathbb{E}_{0,n}e^{\lambda\Delta X_n}, \\ \mathbb{Q}_n(i, dy \mid \mathcal{H}_n) &= \frac{\pi_{i,n}P_i^0(dy)}{\mathsf{Z}_n^\lambda}\exp\{\lambda\eta_{n+1}(y - b_{n+1})\mathsf{s}_i^X\}. \end{aligned} \tag{76}$$

Because $|\Delta X_n| \leq 2uM$, the normalizer is positive and finite at every history, including histories outside the region. The control map and state transition are unchanged. The actual action marginal under this auxiliary kernel generally differs from π_n; no ordinary-SGB convergence theorem is applied to it.

Lemma D.3 (The two tilted supermartingales)**.** *Whenever $X_n \geq h$,*

$$\mathsf{Z}_n^\lambda \geq 1, \qquad \mathbb{E}_{\mathbb{Q},n}e^{-\lambda\Delta X_n} \leq 1, \qquad \mathbb{E}_{\mathbb{Q},n}e^{\lambda\Delta V_n} \leq 1.$$

Consequently, for $\tau = \inf\{n \geq s : X_n < h\}$ and an initial fixed finite $\mathcal{H}_s$-history with $X_s \geq h$, $e^{-\lambda X_{n\wedge\tau}}$ and $e^{\lambda V_{n\wedge\tau}}$, $n \geq s$, are nonnegative $\mathbb{Q}$-supermartingales.

Proof. Jensen and (74) give $\mathsf{Z}_n^\lambda \geq 1$. The first tilted exponential expectation is exactly $1/\mathsf{Z}_n^\lambda$. For the other expectation use, separately for every outcome,

$$e^{\lambda(\Delta X + \Delta V)} = e^{(1-k_I)\lambda\Delta X} \leq (1 - k_I)e^{\lambda\Delta X} + k_I.$$

Since $k_I \geq 0$ and $e^x - 1 \geq x$,

$$k_I(e^{\lambda\Delta X} - 1) \geq \lambda k_I\Delta X = -\lambda\Delta V.$$

After expectation, the previous display therefore is at most $\mathsf{Z}_n^\lambda + \lambda\mathbb{E}_{0,n}\Delta V \leq \mathsf{Z}_n^\lambda$. Divide by the normalizer to obtain the second tilted inequality. The outcome-dependent k_I has not been removed from its expectation. Stopping keeps the increment precisely when the pre-action state has $X_n \geq h$. Finite-time increments are bounded, so the stopped exponentials are integrable on every finite interval, including its exit overshoot. This proves the supermartingale statements without assuming an integrable exit time. □

D.4 MOMENTS UNDER THE TILTED ACTION KERNEL

Write $q^{\mathbb{Q}}_{i,n}$ for the action marginal of (76). Its density bounds imply

$$e^{-4\lambda uM}\pi_{i,n} \le q^{\mathbb{Q}}_{i,n} \le e^{4\lambda uM}\pi_{i,n}. \tag{77}$$

Put $C_{\mathbb{Q}} = e^{4\lambda uM}$ and $x_{i,n} = e^{-\psi_{i,n}}$. Within a one-step calculation let $\Delta\psi_i = \psi_{i,n+1} - \psi_{i,n}$. For each integer $k \ge 1$,

$$|e^{-k\Delta\psi_i} - 1| \le ke^{kuM}|\Delta\psi_i|,$$

whereas the actual marginal under the tilted law gives

$$\begin{aligned}\mathbb{E}_{\mathbb{Q},n}|\Delta\psi_i| &\le uM\{q^{\mathbb{Q}}_{i,n}(1-\pi_{i,n}) + (1-q^{\mathbb{Q}}_{i,n})\pi_{i,n}\}\\ &\le uM(C_{\mathbb{Q}}+1)\pi_{i,n}.\end{aligned}$$

The score still uses π_n, not $q^{\mathbb{Q}}_n$. Using $x^k_{i,n}\pi_{i,n} = x^{k-1}_{i,n}/\sum_j e^{\psi_{j,n}}$ and $\sum_j e^{\psi_{j,n}} \ge Ke^{S_0/K}$, we obtain

$$\mathbb{E}_{\mathbb{Q},n}x^k_{i,n+1} \le x^k_{i,n} + A_k x^{k-1}_{i,n}, \qquad A_k = \frac{ke^{kuM}uM(C_{\mathbb{Q}}+1)}{Ke^{S_0/K}}. \tag{78}$$

Suppose the starting distribution at observation index s is supported on $\psi_{i,s} \ge -L$ for a deterministic finite L. Define $D_0 = 1$ and $D_k = e^{kL} + A_k D_{k-1}$ for $1 \le k \le K$. Induction in k and finite summation of (78) show that

$$\mathbb{E}_{\mathbb{Q}}x^k_{i,n} \le D_k(n-s+1)^k \quad (n \ge s).$$

Indeed, the accumulated bound is $e^{kL} + A_k D_{k-1}\sum_{j=1}^{n-s} j^{k-1}$, which is at most the displayed expression. Coordinate-sum conservation gives

$$e^{X_n} = e^{S_0}x^2_{1,n}\prod_{j=3}^{K} x_{j,n}.$$

The monomial has total degree K. Weighted AM–GM bounds it by $e^{S_0}\{(2/K)x^K_{1,n} + (1/K)\sum_{j=3}^{K} x^K_{j,n}\}$, and hence

$$\mathbb{E}_{\mathbb{Q}}e^{X_n} \le C_X(n-s+1)^K, \qquad C_X = e^{S_0}D_K. \tag{79}$$

This derivation uses relative marginal domination, not an assumption that the auxiliary sampler is the original sampler. No favorable-gap, small-step relative to the gap, or certification premise was required.

D.5 PAYING FOR AN UNFAVORABLE PREFIX

We now assume a deterministic observation index $s \ge 0$, an event $H \in \mathcal{H}_s$, and fixed constants satisfying

$$\begin{aligned}&\mathbb{P}_0(H) \ge \pi_H > 0, \qquad X_s \ge \underline{X} > h, \qquad V_s \le v_0,\\ &\qquad \psi_{i,s} \ge -L \quad (1 \le i \le K) \qquad \text{on } H.\end{aligned} \tag{80}$$

Here $\underline{X}$ is a contrast lower bound, not the inverse coordinate $x_{i,n}$. The prefix quantities are fixed independently of T and δ. We will establish this premise for each initialization in the final subsection.

Start $\mathbb{Q}$ at index s with history law $\mathbb{P}_0(\cdot \mid H)$, and use the joint kernel (76) subsequently. Put

$$\begin{aligned}\varepsilon_0 &= 1 - e^{-\lambda(\underline{X}-h)} > 0,\\ v_* &= v_0 + \lambda^{-1}\log(4/\varepsilon_0),\\ \ell_T &= \log(4C_X/\varepsilon_0) + K\log(T+1), \qquad \varpi = K\lambda.\end{aligned} \tag{81}$$

Take $\tau = \inf\{n \ge s : X_n < h\}$. For $T \ge s$, finite stopping of the first supermartingale gives

$$\mathbb{Q}(\tau \le T) \le e^{-\lambda(\underline{X}-h)} = 1 - \varepsilon_0,$$

because $X_\tau < h$ at exit. Finite maximal stopping of the second gives

$$\mathbb{Q}\{\max_{s\le n\le T} V_{n\wedge\tau} > v_*\} \le e^{\lambda(v_0-v_*)} = \varepsilon_0/4.$$

The moment bound (79) and Markov's inequality give $\mathbb{Q}(X_T > \ell_T) \le \varepsilon_0/4$. Therefore

$$E_T = H\cap\{\tau > T\}\cap\{V_n \le v_* \text{ for } s\le n\le T\}\cap\{X_T\le \ell_T\} \quad \text{satisfies} \quad \mathbb{Q}(E_T)\ge \varepsilon_0/2. \tag{82}$$

The prefix probability cannot be omitted when returning to the reference law. Define the nonnormalized restriction $(\mathbb{P}_0^T\restriction_H)(A) = \mathbb{P}_0^T(A\cap H)$. On the support H of $\mathbb{Q}$, the exact density is

$$\frac{d(\mathbb{P}_0^T\restriction_H)}{d\mathbb{Q}^T} = \mathbb{P}_0(H)e^{-\lambda(X_T-X_s)}\prod_{n=s}^{T-1}\mathsf{Z}_n^\lambda. \tag{83}$$

The exponential increments telescope; the prefactor comes from the conditional initial law. This is not an asserted density for the whole reference law outside H. On E_T, all states in the product have $X_n \ge h$, so every $\mathsf{Z}_n^\lambda \ge 1$. Thus (83), (80) and (82) give

$$\mathbb{P}_0(E_T) \ge c_e(T+1)^{-\varpi}, \qquad c_e = \min\left\{\frac{1}{2},\ \frac{\pi_H\varepsilon_0}{2}e^{\lambda\underline{X}}\left(\frac{\varepsilon_0}{4C_X}\right)^\lambda\right\} > 0. \tag{84}$$

The normalization product has been retained and bounded in its valid direction, not set to one.

D.6 Geometry converts exterior occupation into an information budget

Put

$$m = K-2, \qquad \alpha = \frac{m(1-\rho)}{K-1+\rho}\in(0,1).$$

For a state with coordinate sum S_0 and $V\le v_*$, write $\mathcal{Z} = \sum_i e^{\psi_i}$ and $\pi_i = e^{\psi_i}/\mathcal{Z}$. The identity $V = (1-\rho)\log\mathcal{Z} + \log\pi_1 - \rho\log\pi_2$ implies $\mathcal{Z}^{1-\rho} \le e^{v_*}\pi_2^\rho/\pi_1$. Since $\prod_i \pi_i = e^{S_0}\mathcal{Z}^{-K}$, it follows that

$$\prod_{j=3}^{K}\pi_j \ge e^{S_0 - Kv_*/(1-\rho)}\pi_1^{K/(1-\rho)-1}\pi_2^{-K\rho/(1-\rho)-1}.$$

Taking the geometric mean of these m exterior probabilities gives

$$z \ge me^{S_0/m - Kv_*/((1-\rho)m)}\pi_1^{1/\alpha}\pi_2^{-(1/\alpha-1)}.$$

The last factor is $\pi_2^{-(1/\alpha-1)} \ge 1$, since $0<\pi_2<1$ and $0<\alpha<1$. Consequently

$$\pi_1 \le C_{\mathrm{geo}} z^\alpha, \qquad C_{\mathrm{geo}} = \max\left\{1, m^{-\alpha}\exp\left(\frac{Kv_* - (1-\rho)S_0}{K-1+\rho}\right)\right\}. \tag{85}$$

On E_T this holds at every index from s to T. Concavity of z^α therefore implies

$$Q_1(T) \le s + C_{\mathrm{geo}}T^{1-\alpha}Q_o(T)^\alpha \quad \text{on } E_T. \tag{86}$$

The prefix is charged by s; none of its observations are removed.

Let $z_0 = \sum_{j=3}^K \pi_{j,0} > 0$. For $T\ge 1$ we have $\mathcal{O}_T \ge z_0$. Introduce deterministic analysis cutoffs

$$\begin{aligned} \mathsf{O}_T &= \frac{4\mathcal{O}_T(T+1)^\varpi}{c_e},\\ n_T^{\mathrm{pred}} &= s + C_{\mathrm{geo}}T^{1-\alpha}\mathsf{O}_T^\alpha,\\ n_T^{\mathrm{cut}} &= \left\lceil 2n_T^{\mathrm{pred}}\right\rceil. \end{aligned} \tag{87}$$

They depend on the reference occupation of the given controller, and are not algorithm inputs. Markov's inequality gives

$$\mathbb{P}_0\{Q_o(T) > \mathsf{O}_T\} \le (c_e/4)(T+1)^{-\varpi}.$$

Also $\exp\{(\log 2)N_1(T) - Q_1(T)\}$ is a nonnegative reference-law supermartingale with initial value one: conditional expectation of $2^{\mathbf{1}\{I_{n+1}=1\}}$ is $1 + \pi_{1,n} \le e^{\pi_{1,n}}$. It follows that

$$\mathbb{P}_0\{N_1(T) > n_T^{\mathrm{cut}}, Q_1(T) \le n_T^{\mathrm{pred}}\} \le e^{-(2\log 2-1)n_T^{\mathrm{pred}}}. \tag{88}$$

We specify a finite onset for this error to be smaller than the reference event probability. Let

$$\xi = 1 - \alpha, \quad c_m = C_{\mathrm{geo}}(4z_0/c_e)^\alpha, \quad c_{\exp} = (2\log 2 - 1)c_m > 0, \quad C_{\log} = \log(4/c_e).$$

Then $n_T^{\mathrm{pred}} \ge c_m T^\xi$. Take T_{pre} no smaller than the upward integer rounding of

$$\max\left\{2,\, 2s,\, \left(\frac{\max\{s,1\}}{c_m}\right)^{1/\xi},\, \left[\max\left\{1, \sqrt{\frac{2C_{\log}}{c_{\exp}}}, \frac{8\varpi}{\xi c_{\exp}}\right\}\right]^{2/\xi}\right\}. \tag{89}$$

For $T \ge T_{\mathrm{pre}}$, the inequality $\log(T+1) \le (4/\xi)T^{\xi/2}$ and the last term of this maximum imply

$$c_{\exp} T^\xi \ge C_{\log} + \varpi \log(T+1).$$

The logarithm bound follows from $T + 1 \le 2T$ and $\log T \le [2/(e\xi)]T^{\xi/2}$, the latter obtained by maximizing $(\log T)/T^{\xi/2}$; since $0 < \xi < 1$, their coefficient is at most $4/\xi$. Indeed, writing $x = T^{\xi/2}$, the two summands on the right are at most $c_{\exp}x^2/2$ each. This proves

$$e^{-(2\log 2-1)n_T^{\mathrm{pred}}} \le (c_e/4)(T+1)^{-\varpi}.$$

The other terms in the maximum ensure $T \ge 2s$, $n_T^{\mathrm{pred}} \ge 1$ and $s \le C_{\mathrm{geo}}T^{1-\alpha}\mathsf{O}_T^\alpha$. Thus

$$n_T^{\mathrm{cut}} \le 6C_{\mathrm{geo}}T^{1-\alpha}\mathsf{O}_T^\alpha. \tag{90}$$

Combining (84), (86) and the two error bounds, the event

$$E_T^{\mathrm{cnt}} = E_T \cap \{Q_o(T) \le \mathsf{O}_T\} \cap \{N_1(T) \le n_T^{\mathrm{cut}}\} \quad \text{satisfies} \quad \mathbb{P}_0(E_T^{\mathrm{cnt}}) \ge (c_e/2)(T+1)^{-\varpi}. \tag{91}$$

No step assumes $n_T^{\mathrm{cut}} < T$. In particular, all actual best-arm observations, including those in the prefix, are included in this event.

D.7 Actual-count likelihood and the self-bounding closure

Lemma D.4 (Actual-count-normalized likelihood)**.** *For fixed T, n and a history event $A \subseteq \{N_1(T) \le n\}$,*

$$\mathbb{P}_\delta(A) \ge (1 - 4\delta^2)^n \{\mathbb{P}_0(A)\}^2.$$

Proof. Let $\mathfrak{L}_t = d\mathbb{P}_0^t/d\mathbb{P}_\delta^t$ and put $\chi_\delta = (1-4\delta^2)^{-1}$. Given the same history, the controller and action densities cancel. Let f_0, f_δ be the Bernoulli mass functions with parameters $1/2$ and $1/2+\delta$. For one observed reward of arm1,

$$\mathbb{E}_\delta\left(\frac{f_0(Y)}{f_\delta(Y)}\right)^2 = \frac{1/4}{1/2+\delta} + \frac{1/4}{1/2-\delta} = \chi_\delta.$$

The corresponding factor for another arm is one. Consequently $\mathfrak{L}_t^2 \chi_\delta^{-N_1(t)}$ is a finite-time martingale of mean one; conditioning on the actual action proves the identity one step at a time. On A, its normalization yields

$$\mathbb{E}_\delta[\mathfrak{L}_T^2 \mathbf{1}_A] \le \chi_\delta^n \mathbb{E}_\delta[\mathfrak{L}_T^2 \chi_\delta^{-N_1(T)}] = \chi_\delta^n.$$

Cauchy–Schwarz applied to $\mathbb{P}_0(A) = \mathbb{E}_\delta[\mathfrak{L}_T \mathbf{1}_A]$ gives the claim. The random actual count has not been treated as a fixed independent sample size. □

For $T \ge T_{\mathrm{pre}}$, choose

$$\delta_T = \frac{1}{8\sqrt{n_T^{\mathrm{cut}}}} \in (0, 1/8]. \tag{92}$$

This may depend on the controller through $\mathcal{O}_T$. The inequality $(1-1/(16n))^n \geq 15/16$ for integers $n \geq 1$ and Lemma D.4 give

$$\mathbb{P}_{\delta_T}(E_T^{\mathrm{cnt}}) \geq \frac{15c_e^2}{64}(T+1)^{-2\varpi}. \tag{93}$$

Each selected parameter specifies one fixed IID environment for that trajectory; it does not change the actual environment during a run.

On E_T^{cnt}, $X_t \geq h$ for every $s \leq t < T$, and thus $\pi_{1,t} \leq \rho\pi_{2,t} \leq \rho(1-\pi_{1,t})$. The predictable conditional pseudo-regret in the true environment is

$$r_t^\delta = \delta(1-\pi_{1,t}) + \sum_{j=3}^{K} d_j \pi_{j,t},$$

so it is at least $\delta/(1+\rho)$ there. First use the unconditional identity $R_T(\delta) = \mathbb{E}_\delta \sum_{t<T} r_t^\delta$, and then restrict the nonnegative random sum to E_T^{cnt}. Since $T \geq 2s$, (93) gives

$$R_T(\delta_T) \geq \frac{\delta_T(T-s)}{1+\rho}\mathbb{P}_{\delta_T}(E_T^{\mathrm{cnt}}) \geq \frac{15c_e^2\delta_T T}{128(1+\rho)}(T+1)^{-2\varpi}. \tag{94}$$

This does not use a conditional equality between actual and predictable counts on E_T^{cnt}.

From (90) and (92),

$$\delta_T \geq \frac{T^{-(1-\alpha)/2}\mathsf{O}_T^{-\alpha/2}}{8\sqrt{6C_{\mathrm{geo}}}}.$$

Substitute (87) into (94) and use $T+1 \leq 2T$. This gives

$$R_T(\delta_T) \geq C_* T^{(1+\alpha)/2-(2+\alpha/2)\varpi}\mathcal{O}_T^{-\alpha/2}, \tag{95}$$

where the following explicit positive constant suffices:

$$C_* = \frac{15c_e^2}{1024(1+\rho)\sqrt{6C_{\mathrm{geo}}}}\left(\frac{c_e}{4}\right)^{\alpha/2} 2^{-(2+\alpha/2)\varpi}. \tag{96}$$

Lemma D.2 implies $\mathcal{O}_T \leq 4\mathcal{R}_T$, while $\mathcal{R}_T \geq R_T(\delta_T)$. Both are positive, since $\mathcal{O}_T \geq z_0 > 0$, and finite. Therefore

$$\mathcal{R}_T^{1+\alpha/2} \geq C_* 4^{-\alpha/2} T^{(1+\alpha)/2-(2+\alpha/2)\varpi}.$$

We have proved the fixed-prefix conclusion

$$\begin{aligned} \mathcal{R}_T &\geq c_* T^\zeta, \qquad T \geq T_{\mathrm{pre}}, \\ c_* &= (C_* 4^{-\alpha/2})^{1/(1+\alpha/2)}, \\ \zeta &= \frac{1+\alpha-(4+\alpha)\varpi}{2+\alpha}. \end{aligned} \tag{97}$$

Choose

$$0 < \varpi < \frac{\alpha}{2(4+\alpha)}, \qquad \lambda = \varpi/K. \tag{98}$$

Then $\zeta > 1/2$. These choices are compatible with the preceding construction: first fix the prefix and ρ, hence α; next choose ϖ, λ; then form the constants and the finite onset above. Their dependence is only on the fixed prefix and $K, u, B_{\max}, \rho, \lambda$. In particular no unknown-gap or expected recovery-time bound has been inserted.

D.8 Obtaining the prefix from every finite initialization

Proof of Theorem D.1. Suppose first that the initial logits are not all equal. Choose an initially lower-logit arm as arm1 and a higher-logit arm as arm2, and assign the other family means to the remaining arm identities. Let

$$\underline{X} = \vartheta_2 - \vartheta_1 > 0, \qquad h = \underline{X}/2, \qquad \rho = e^{-h}.$$

Take $s = 0$, H the whole history space, $\pi_H = 1$, $v_0 = \vartheta_1 - \rho\vartheta_2$ and $L = \max_i |\vartheta_i|$. These satisfy (80). Choose (98) and apply (97). This proves the required statement in this case without assigning the controller an unknown best-arm label.

Now suppose $\vartheta = c\mathbf{1}$. Fix a labeling of the family and consider the same infinite rule under $\mathbb{P}_0$. If $\mathbb{P}_0\{\exists t \geq 1 : \eta_t > 0\} = 0$, its reference policy stays uniform, and

$$R_T(0) = \frac{T}{K}\sum_{j=3}^{K} d_j \geq \frac{K-2}{4K}T.$$

Lemma D.2 transfers this to a linear lower bound on $\mathcal{R}_T$. Thus the theorem holds with exponent one in this branch.

Otherwise countability gives a deterministic observation index $t \geq 0$ and a constant $\eta_{\mathrm{pos}} > 0$ for which the pre-action event

$$F = \{\eta_1 = \cdots = \eta_t = 0, \quad \eta_{t+1} \geq \eta_{\mathrm{pos}}\} \in \mathcal{H}_t$$

has positive reference probability. Indeed the first positive step has some finite index, and its positive value exceeds $1/m$ for some integer m. This selection is made for the fixed infinite rule, independently of the final horizon and gap. On F, $\psi_t = c\mathbf{1}$ and $\pi_{i,t} = 1/K$. Define an event after the next actual observation by

$$\begin{aligned} H = F \cap \big(&\{b_{t+1} \leq 1/2,\ I_{t+1} = 2,\ Y_{t+1} = 1\} \\ &\cup \{b_{t+1} > 1/2,\ I_{t+1} = 1,\ Y_{t+1} = 0\}\big). \end{aligned}$$

Which branch is required is known before the action. Under the tie reference its conditional probability on F is $1/(2K)$, because each of the first two arms has reward mean $1/2$ and is selected with probability $1/K$. Consequently

$$H \in \mathcal{H}_{t+1}, \qquad \mathbb{P}_0(H) = \mathbb{P}_0(F)/(2K) > 0.$$

No action has been forced; H is a naturally occurring analysis event.

At uniform logits the score contrast has no probability-difference term. In the first branch its increment is $\eta_{t+1}(1 - b_{t+1}) \geq \eta_{t+1}/2$. In the second it is $\eta_{t+1}b_{t+1} > \eta_{t+1}/2$. Thus

$$X_{t+1} \geq \eta_{\mathrm{pos}}/2, \qquad |\psi_{i,t+1} - c| \leq uM.$$

Set

$$\begin{aligned} s &= t+1, \qquad \underline{X} = \eta_{\mathrm{pos}}/2, \qquad h = \eta_{\mathrm{pos}}/4, \\ \rho &= e^{-\eta_{\mathrm{pos}}/4}, \quad L = |c| + uM, \quad v_0 = (1-\rho)c + (1+\rho)uM, \\ \pi_H &= \mathbb{P}_0(H). \end{aligned}$$

These parameters verify all of (80). For example $V_s = \psi_{1,s} - \rho\psi_{2,s} \leq (1-\rho)c + (1+\rho)uM$, while $\underline{X} > h > 0$. Apply (98) and (97). The prefix length and probability are paid in the displayed density, the occupation relation retains the first s observations, and the regret lower bound retains $T - s$. They have not been treated as a free restart or as an integrable recovery time. This proves (73). □

The uniform-initialization prefix may have arbitrarily small probability or step size as the fixed rule varies. Its resulting $c_{\mathcal{A}}, T_{\mathcal{A}}, \zeta_{\mathcal{A}}$ therefore cannot be moved inside an infimum over different horizon-aware rules. The lower bound selects a small fixed IID instance separately for each horizon; it does not imply polynomial asymptotic regret at one fixed gap. It also does not separately refute every inverse-gap logarithmic bound. The joint tilt, its moments and the actual-count normalization in this appendix are required for this supporting statement; replacing them by the different occupation-KL construction would not be a faithful consumption of the present result.

E THE FINITE-ACTIVITY BOUNDARY

This appendix concerns a subclass of the explicit geometric rules, not the entire class. In particular, its obstruction does not rule out a controller whose geometric activity diverges. Throughout, a rule is one horizon-free history mapping, used without access to the true gaps or the best-arm label on every reward instance. It is not a separate choice of rule for each evaluation horizon.

E.1 Model and statement

Fix $K \geq 2$, $u > 0$, $B_{\max} \geq 1$, a deterministic finite vector $\vartheta \in \mathbb{R}^K$, and $\gamma \in (0, 1/2]$. Each arm has an independent pull-indexed IID reward stream taking values in $[0, 1]$. Write μ_i for its mean, let j be the unique best arm, and set $\Delta_i = \mu_j - \mu_i$, including $\Delta_j = 0$. Thus $\Delta_i > 0$ for $i \neq j$. The filtration $\mathcal{F}_t$ is the information available before action I_t. It may include independent controller randomization, but the current action seed and unused rewards retain their original conditional product law. The controls η_t, b_t are $\mathcal{F}_t$-measurable and obey $0 \leq \eta_t \leq u$ and $|b_t| \leq B_{\max}$.

With $\theta_1 = \vartheta$, the rule is exactly

$$p_{i,t} = \frac{e^{\theta_{i,t}}}{\sum_{k=1}^K e^{\theta_{k,t}}}, \qquad I_t \mid \mathcal{F}_t \sim p_t, \qquad \theta_{i,t+1} = \theta_{i,t} + \eta_t (Y_t - b_t) p_{i,t}^{-\gamma} \big(\mathbf{1}\{I_t = i\} - p_{i,t}\big), \tag{99}$$

where Y_t is the next unused reward of the selected arm. There is no additional exploration, projection, mirror normalization, logit reset, or action-dependent scalar or baseline. Define

$$a_\gamma(p) = \sum_{i=1}^K p_i^{1-2\gamma}(1 - p_i), \qquad \mathcal{E}_\gamma = \sum_{t=1}^\infty \eta_t^2 a_\gamma(p_t), \qquad E = \{\mathcal{E}_\gamma < \infty\}, \tag{100}$$

$$A_t = \sum_{s<t} \eta_s, \qquad Q_i(T) = \sum_{t=1}^T p_{i,t}, \qquad N_i(T) = \sum_{t=1}^T \mathbf{1}\{I_t = i\}. \tag{101}$$

Here $A_\infty = \sum_{t \geq 1} \eta_t$ and $N_i(0) = Q_i(0) = 0$. Geometric activity is a predictable squared-update budget; it is not an assertion of an exact covariance identity. In particular, bounded baseline randomness remains present in the second-moment estimate below.

The instantaneous predictable pseudo-regret, realized pseudo-regret, and unconditional expected pseudo-regret are, respectively,

$$r_t = \sum_{i \neq j} \Delta_i p_{i,t}, \qquad \widehat{R}_T = \sum_{i \neq j} \Delta_i N_i(T), \qquad R_T = \mathbb{E}\widehat{R}_T = \mathbb{E}\sum_{t=1}^T r_t. \tag{102}$$

Theorem E.1 (Positive-probability finite activity excludes quadratic logarithmic regret)**.** *Under the preceding assumptions, the following statements hold almost surely on E.*

1. *On $\{A_\infty < \infty\}$, every logit converges to a finite real number and every action probability has a strictly positive limit. Both $\widehat{R}_T/T$ and $T^{-1}\sum_{t \leq T} r_t$ converge to the same strictly positive, possibly random, number.*
2. *On $\{A_\infty = \infty\}$, $p_{j,t} \to 1$, but nevertheless*

$$\frac{\widehat{R}_T}{\log^2(eT)} \longrightarrow \infty, \qquad \frac{\sum_{t=1}^T r_t}{\log^2(eT)} \longrightarrow \infty. \tag{103}$$

Consequently, (103) *holds almost surely on all of E. If $\mathbb{P}(E) > 0$, then*

$$\liminf_{T \to \infty} \frac{R_T}{\log^2(eT)} = \infty. \tag{104}$$

In particular, an $O(\log^2(eT))$ expected-regret bound on an instance requires $\mathcal{E}_\gamma = \infty$ almost surely on that instance. This is a necessary condition, not a sufficient one.

All conditional calculations in the proof are made under the original filtration and sampling law in (99). The event E depends on the future; we never condition the current action distribution on E and then pretend that it remains p_t.

E.2 Two localization facts

Lemma E.2 (Conditional summability)**.** *Let $W_{t+1} \geq 0$ be $\mathcal{F}_{t+1}$-measurable and let $\nu_t = \mathbb{E}[W_{t+1} \mid \mathcal{F}_t] < \infty$. Then $\sum_t W_{t+1} < \infty$ almost surely on $\{\sum_t \nu_t < \infty\}$.*

Proof. For each integer $\ell \geq 1$, the indicator $\chi_t^{(\ell)} = \mathbf{1}\{\sum_{s \leq t} \nu_s \leq \ell\}$ is $\mathcal{F}_t$-measurable. The retained conditional means sum to at most ℓ, so Tonelli's theorem and conditional expectation give

$$\mathbb{E} \sum_t W_{t+1} \chi_t^{(\ell)} = \mathbb{E} \sum_t \nu_t \chi_t^{(\ell)} \leq \ell.$$

Thus the retained nonnegative sum is finite almost surely. On $\{\sum_t \nu_t \leq \ell\}$ no term is removed. The countable union over ℓ proves the claim. Because an overshooting term is excluded using its already predictable conditional mean, no bound on the size of that overshoot is required. □

Lemma E.3 (Martingale convergence on a finite variance budget)**.** *Suppose ξ_{t+1} are square-integrable martingale differences with $\mathbb{E}[\xi_{t+1} \mid \mathcal{F}_t] = 0$. Then $\sum_t \xi_{t+1}$ converges almost surely on $\{\sum_t \mathbb{E}[\xi_{t+1}^2 \mid \mathcal{F}_t] < \infty\}$.*

Proof. Use the preceding predictable indicators with $\nu_t = \mathbb{E}[\xi_{t+1}^2 \mid \mathcal{F}_t]$. The retained differences still have conditional mean zero. Orthogonality of martingale differences gives

$$\sup_T \mathbb{E}\left[\left(\sum_{t \leq T} \xi_{t+1} \chi_t^{(\ell)}\right)^2\right] = \sup_T \mathbb{E} \sum_{t \leq T} \nu_t \chi_t^{(\ell)} \leq \ell.$$

The retained martingale is therefore bounded in L^2 and converges almost surely by the L^2 martingale convergence theorem. On a total variance budget at most ℓ it coincides with the original martingale. Taking the countable union over ℓ proves the assertion. □

E.3 Finite activity bounds the actual jumps and exponential scale

Write $v_{i,t} = \theta_{i,t+1} - \theta_{i,t}$ and $L_{\max} = 1 + B_{\max}$. Since $|Y_t - b_t| \leq L_{\max}$, the original conditional action/reward kernel implies

$$\begin{aligned}\mathbb{E}[v_{i,t}^2 \mid \mathcal{F}_t] &\leq L_{\max}^2 \eta_t^2 p_{i,t}^{-2\gamma} \mathbb{E}[(\mathbf{1}\{I_t = i\} - p_{i,t})^2 \mid \mathcal{F}_t] \\ &= L_{\max}^2 \eta_t^2 p_{i,t}^{1-2\gamma}(1 - p_{i,t}), \qquad \mathbb{E}[\|v_t\|_2^2 \mid \mathcal{F}_t] \leq L_{\max}^2 \eta_t^2 a_\gamma(p_t). \end{aligned} \tag{105}$$

The finite-time integrability needed for localization holds without any uniform bound on preconditioned jumps. Indeed, if $\max_i |\theta_{i,t}| \leq H_t$ for a deterministic finite H_t, then $p_{i,t} \geq e^{-2H_t}/K$ and $|v_{i,t}| \leq u L_{\max} K^\gamma e^{2\gamma H_t}$. Starting from $H_1 = \max_i |\vartheta_i|$, the recursion $H_{t+1} = H_t + u L_{\max} K^\gamma e^{2\gamma H_t}$ gives such a finite bound at every deterministic finite time. This bound is used only for integrability, not as a regret constant.

Lemma E.2 gives, almost surely on E,

$$\sum_t \|v_t\|_2^2 < \infty, \qquad v_t \longrightarrow 0. \tag{106}$$

For each arm define the coordinate martingale

$$M_i(t) = \sum_{s<t} \big(v_{i,s} - \mathbb{E}[v_{i,s} \mid \mathcal{F}_s]\big). \tag{107}$$

Its conditional variance is bounded by the corresponding second moment in (105); hence Lemma E.3 shows that every $M_i(t)$ converges to a finite $M_i(\infty)$ on E.

Define the common exponential scale and its normalized weights by

$$\Phi_t = \sum_i e^{\gamma \theta_{i,t}}, \qquad w_{i,t} = e^{\gamma \theta_{i,t}}/\Phi_t.$$

For every realized action and reward,

$$\sum_i w_{i,t} v_{i,t} = 0. \tag{108}$$

To see this, set $Z_t = \sum_i e^{\theta_{i,t}}$ and observe that $e^{\gamma \theta_{i,t}} p_{i,t}^{-\gamma} = Z_t^\gamma$, independently of i; the remaining score coordinates sum to zero. Jensen's inequality now gives $\Phi_{t+1}/\Phi_t = \sum_i w_{i,t} e^{\gamma v_{i,t}} \geq 1$. On E,

(106) ensures $\|v_t\|_\infty \le 1$ eventually. The pathwise Taylor bound $e^x \le 1 + x + e^\gamma x^2/2$ for $|x| \le \gamma$, followed by $\log(1+x) \le x$, yields

$$0 \le \log \frac{\Phi_{t+1}}{\Phi_t} \le \frac{\gamma^2 e^\gamma}{2} \sum_i w_{i,t} v_{i,t}^2 \le \frac{\gamma^2 e^\gamma}{2} \|v_t\|_2^2. \tag{109}$$

The right-hand side is summable. The finite prefix consists of finite positive values, so

$$0 < \Phi_1 \le \Phi_t \longrightarrow \Phi_\infty < \infty \quad \text{almost surely on } E. \tag{110}$$

In particular, every coordinate is bounded above on each such path. Since $0 < \gamma \le 1$, subadditivity of $x \mapsto x^\gamma$ on nonnegative arguments gives $Z_t^\gamma \le \Phi_t$, so Z_t is bounded above as well. This step used Taylor's inequality only for realized small jumps; it did not infer a conditional Taylor-remainder estimate from eventual pathwise smallness.

E.4 The scalar-clock dichotomy and allocation law

Put $\beta = 1 - \gamma \in [1/2, 1)$. The conditional mean of the update is

$$\mathbb{E}[v_{i,t} \mid \mathcal{F}_t] = \eta_t p_{i,t}^\beta \left(\mu_i - \sum_k p_{k,t} \mu_k \right) = \eta_t p_{i,t}^\beta (r_t - \Delta_i). \tag{111}$$

The common predictable baseline cancels in this first moment: the two terms containing b_t are $-p_{i,t} b_t$ and $+p_{i,t} b_t$. It does not cancel from $L_{\max}$ in the second-moment bound.

For the best coordinate,

$$\theta_{j,t} = \vartheta_j + \sum_{s<t} \eta_s p_{j,s}^\beta r_s + M_j(t).$$

The middle term is nonnegative and nondecreasing. Since $M_j(t)$ converges on E and $\theta_{j,t}$ is bounded above there, the middle sum has a finite limit, and $\theta_{j,t}$ converges to a finite real number. Boundedness of Z_t then gives a path-dependent $c > 0$ such that $p_{j,t} \ge c$ eventually. Consequently,

$$\sum_t \eta_t r_t < \infty \quad \text{almost surely on } E. \tag{112}$$

For any arm, its compensated coordinate can be written as

$$\theta_{i,t} - M_i(t) = \vartheta_i + \sum_{s<t} \eta_s p_{i,s}^\beta r_s - \Delta_i \sum_{s<t} \eta_s p_{i,s}^\beta.$$

The positive sum converges by (112), whereas the last sum is nondecreasing. Thus every $\theta_{i,t}$ has either a finite limit or limit $-\infty$. At least the best coordinate has a finite limit, so

$$Z_t \longrightarrow Z_\infty \in (0, \infty), \quad \text{and every } p_{i,t} \text{ has a limit, on } E. \tag{113}$$

If $A_\infty < \infty$, then $\sum_t |\mathbb{E}[v_{i,t} \mid \mathcal{F}_t]| \le \sum_t \eta_t < \infty$ because $p_{i,t}^\beta \le 1$ and $|r_t - \Delta_i| \le 1$. All logit limits are therefore finite and all probability limits are strictly positive. If $A_\infty = \infty$, a positive limit of r_t would contradict (112). Hence $r_t \to 0$; the unique-best-arm assumption implies $p_{j,t} \to 1$ and $p_{i,t} \to 0$ for $i \ne j$. These limits were derived directly, without assuming a low-regret conclusion or importing a policy-convergence theorem.

We next work on $E \cap \{A_\infty = \infty\}$. For each $i \ne j$, define $y_{i,t} = \theta_{i,t} - M_i(t)$ and $X_{i,t} = e^{-\beta y_{i,t}}$. By (111),

$$y_{i,t+1} - y_{i,t} = -\eta_t p_{i,t}^\beta (\Delta_i - r_t).$$

This difference tends to zero because $\eta_t \le u$, $p_{i,t} \to 0$, and $\beta > 0$. Expanding the exponential with a relative remainder tending to zero, on rounds with $\eta_t > 0$ we obtain

$$\begin{aligned} \frac{X_{i,t+1} - X_{i,t}}{\eta_t} &= \beta(\Delta_i - r_t) X_{i,t} p_{i,t}^\beta (1 + o(1)) \\ &= \beta(\Delta_i - r_t) e^{\beta M_i(t)} Z_t^{-\beta} (1 + o(1)) \longrightarrow \beta \Delta_i e^{\beta M_i(\infty)} Z_\infty^{-\beta}. \end{aligned} \tag{114}$$

On zero-scalar rounds the increment of $X_{i,t}$ is zero. Summing with nonnegative weights η_t and dividing by $A_t \to \infty$ gives the same limit for $X_{i,t}/A_t$: any fixed prefix vanishes, and all sufficiently late weighted slopes lie arbitrarily close to their displayed limit. Since $p_{i,t}^{-\beta} = X_{i,t} e^{-\beta M_i(t)} Z_t^{\beta}$, the random scale cancels, yielding

$$\frac{p_{i,t}^{-\beta}}{A_t} \longrightarrow \beta\Delta_i \qquad (i \neq j). \tag{115}$$

Define the transformed scalar clock $z_t = (1 + A_t)^{1/(2\beta)}$ and constants $c_i = (\beta\Delta_i)^{-1/\beta}$ for $i \neq j$. Then

$$p_{i,t} \sim c_i z_t^{-2}. \tag{116}$$

Let $\alpha = 1 - 2\gamma \in [0, 1)$ and $V_\Delta = \sum_{i \neq j} (\beta\Delta_i)^{-\alpha/\beta} > 0$. Equation (115) gives

$$a_\gamma(p_t) \sim V_\Delta A_t^{-\alpha/\beta}. \tag{117}$$

Indeed, the suboptimal-coordinate terms have precisely this asymptotic sum; the best-coordinate term is $p_{j,t}^{\alpha}(1 - p_{j,t}) = O(A_t^{-1/\beta})$, which is of smaller order because $\alpha < 1$. At $\gamma = 1/2$, this also follows from the exact identity $a_{1/2}(p) = K - 1$.

Boundedness of η_t implies $\eta_t/A_t \to 0$. On positive-scalar rounds, expansion of the power defining z_t and $1/\beta - 2 = -\alpha/\beta$ now show

$$\frac{(z_{t+1} - z_t)^2}{\eta_t^2 a_\gamma(p_t)} \longrightarrow \frac{1}{4\beta^2 V_\Delta} \in (0, \infty). \tag{118}$$

No $0/0$ ratio is used on zero-scalar rounds; their clock increments are zero. The finite prefix is finite, and the tail ratio is bounded. Therefore

$$\sum_t (z_{t+1} - z_t)^2 < \infty. \tag{119}$$

The gap-dependent constants above are fixed-instance asymptotic comparison constants, not a gap-uniform numerical lower-bound coefficient.

E.5 A DISCRETE HARDY ARGUMENT

Lemma E.4 (Finite squared increments force superquadratic logarithmic reciprocal mass)**.** *If $z_n > 0$ and $\sum_{n \geq 1} (z_{n+1} - z_n)^2 < \infty$, then*

$$\sum_{n \geq 1} \frac{z_n^2}{n^2} < \infty, \qquad \frac{\sum_{n=1}^{T} z_n^{-2}}{\log^2(eT)} \longrightarrow \infty.$$

The increments need not be nonnegative.

Proof. Set $d_k = z_{k+1} - z_k$ and $S_n = \sum_{k=1}^{n} |d_k|$. Weighted Cauchy–Schwarz gives

$$S_n^2 \leq \left(\sum_{k \leq n} \sqrt{k}\, d_k^2 \right) \left(\sum_{k \leq n} k^{-1/2} \right) \leq 2\sqrt{n} \sum_{k \leq n} \sqrt{k}\, d_k^2.$$

By Tonelli and the bound $\sum_{n \geq k} n^{-3/2} \leq k^{-3/2} + 2k^{-1/2}$,

$$\sum_{n \geq 1} \frac{S_n^2}{n^2} \leq 2 \sum_{k \geq 1} \sqrt{k}\, d_k^2 \sum_{n \geq k} n^{-3/2} \leq 6 \sum_{k \geq 1} d_k^2 < \infty.$$

Since $z_n^2 \leq 2z_1^2 + 2S_n^2$ and $\sum_n n^{-2} < \infty$, this proves the first assertion. For each fixed integer $N \geq 1$ and $T \geq N$, Cauchy–Schwarz also gives

$$\left(\sum_{n=N}^{T} \frac{1}{n} \right)^2 \leq \left(\sum_{n=N}^{T} z_n^{-2} \right) \left(\sum_{n=N}^{T} \frac{z_n^2}{n^2} \right).$$

Divide by $\log^2(eT)$ and let $T \to \infty$ with N fixed. The normalized harmonic sum tends to one, so

$$\liminf_{T\to\infty} \frac{\sum_{n=1}^{T} z_n^{-2}}{\log^2(eT)} \geq \left(\sum_{n=N}^{\infty} \frac{z_n^2}{n^2} \right)^{-1}.$$

The positive tail on the right tends to zero as $N \to \infty$. Its reciprocal therefore tends to infinity, proving the claimed limit. □

E.6 Actual counts and unconditional expectation

We first justify, under the original probability law, the counting relation

$$\frac{N_i(T) - Q_i(T)}{Q_i(T)} \longrightarrow 0 \quad \text{almost surely on } \{Q_i(T) \to \infty\}. \tag{120}$$

The process

$$\sum_{t=1}^{T} \frac{\mathbf{1}\{I_t = i\} - p_{i,t}}{1 + Q_i(t)}$$

is a martingale: its denominator includes the current $p_{i,t}$ but is still $\mathcal{F}_t$-measurable. Its total conditional variance is at most

$$\sum_{t\geq 1} \frac{p_{i,t}}{(1+Q_i(t))^2} \leq \int_0^{\infty} \frac{dx}{(1+x)^2} = 1.$$

The inequality follows by integrating the decreasing integrand over each interval $[Q_i(t-1), Q_i(t)]$. Thus the martingale is bounded in L^2 and converges almost surely. Kronecker's summation lemma applies pathwise to the nondecreasing denominators $1 + Q_i(t)$ whenever they diverge, and gives (120). For completeness, the lemma here follows directly from summation by parts: if the partial sums of a convergent series after some fixed index have absolute value at most ε, their weighted sum with nondecreasing positive weights, divided by the last weight, has absolute value at most 2ε; the finite prefix vanishes as the last weight diverges. Letting $\varepsilon \downarrow 0$ gives the required normalized sum of centered counts.

Completion of Theorem E.1. On $E \cap \{A_\infty < \infty\}$, let $p_{i,\infty} > 0$ denote the probability limits already established. Cesaro averaging gives $Q_i(T)/T \to p_{i,\infty}$, and (120) gives $N_i(T)/T \to p_{i,\infty}$. Therefore both normalized regrets converge to $\sum_{i\neq j} \Delta_i p_{i,\infty} > 0$.

On $E \cap \{A_\infty = \infty\}$, apply Lemma E.4 to (119). By (116), for every $i \neq j$ the probability $p_{i,t}$ is eventually at least $(c_i/2)z_t^{-2}$. Hence $Q_i(T)/\log^2(eT) \to \infty$. In particular $Q_i(T) \to \infty$, so (120) also gives $N_i(T)/\log^2(eT) \to \infty$. Multiplication by the positive gaps and summation prove both limits in (103).

Finally, apply Fatou's lemma to the nonnegative random variables $\widehat{R}_T/\log^2(eT)$, without conditioning on E. If $\mathbb{P}(E) > 0$, then

$$\liminf_{T\to\infty} \frac{R_T}{\log^2(eT)} \geq \mathbb{E}\left[\liminf_{T\to\infty} \frac{\widehat{R}_T}{\log^2(eT)} \right] = \infty,$$

because the inside limit is infinite on an event of positive probability. This proves (104). Its contrapositive gives the stated necessary divergence condition. □

E.7 The feasible logarithmic powers in the finite-activity subclass

Fix the same $K, u, B_{\max}, \vartheta, \gamma$. Consider only rules satisfying (99) that, on every $[0,1]$ IID instance with a unique best arm, have $\mathcal{E}_\gamma < \infty$ almost surely. Each rule is a single horizon-free, unknown-gap history mapping as specified above. Call a real exponent ρ feasible if there exist one such rule, a finite predetermined $m \geq 0$, and finite nonnegative constants C_0, C_1, C_2 such that, simultaneously on every instance and for every $T \geq 1$,

$$R_T \leq C_0 + C_1 \log^\rho(eT) \sum_{i\neq j} \Delta_i^{-1}, \qquad R_T \leq C_0 + C_2\sqrt{T}\log^m(eT). \tag{121}$$

The constants may depend on the fixed design parameters, K, and ϑ, but not on the reward instance or the horizon.

Corollary E.5 (Unattained infimum of feasible logarithmic powers)**.** *The feasible powers defined by* (121) *have infimum* 2*, and* 2 *is not feasible. Every* $2+\epsilon$ *with* $0<\epsilon\le 1$ *is feasible.*

Proof. Fix any instance. Membership in the subclass gives $\mathbb{P}(E)=1$. If $\rho\le 2$, the first bound in (121) would make $R_T/\log^2(eT)$ bounded, contradicting Theorem E.1. Thus no $\rho\le 2$ is feasible.

For $0<\epsilon\le 1$, choose the controller in Theorem B.1 with $\omega=(1+\epsilon)/2\in(1/2,1]$. It is one horizon-free, unknown-gap rule with baseline 1, legal because $B_{\max}\ge 1$. Its two finite-time bounds give (121) with $\rho=2\omega+1=2+\epsilon$ and $m=\omega$, after enlarging the common additive constant if needed. It remains to verify its membership in the finite-activity subclass; the finite-time rate theorem alone is not used as a substitute for that verification.

For this controller write $q_t=1-\max_i p_{i,t}$, any fixed $h\in(0,u]$, and

$$\eta_t=\kappa_t q_t^{-(1/2-\gamma)},\qquad \kappa_t=\frac{h}{\log^{\omega}(et)(\sqrt{t}+\overline{C}_t)}\le\frac{h}{\sqrt{t}\log^{\omega}(et)},$$

where $\overline{C}_t$ is the positive running cap of that controller. Set $\alpha=1-2\gamma\in[0,1)$. Choose a coordinate attaining the maximum probability. Each other coordinate satisfies $p_i\le q_t$, and hence $p_i^{\alpha}(1-p_i)\le q_t^{\alpha}$. The maximizing coordinate contributes at most $q_t\le q_t^{\alpha}$. Therefore, including the endpoint $\gamma=1/2$,

$$a_\gamma(p_t)\le Kq_t^{1-2\gamma},\qquad \eta_t^2 a_\gamma(p_t)\le K\kappa_t^2.$$

The integral comparison for the decreasing function $x\mapsto[x\log^{2\omega}(ex)]^{-1}$ gives the deterministic bound

$$\mathcal{E}_\gamma\le Kh^2\sum_{t\ge1}\frac{1}{t\log^{2\omega}(et)}\le Kh^2\left(1+\int_1^\infty\frac{dx}{x(1+\log x)^{2\omega}}\right)=Kh^2\left(1+\frac{1}{2\omega-1}\right)<\infty. \tag{122}$$

Thus the controller belongs to the required subclass on every instance. Powers arbitrarily close to 2 from above are feasible, while 2 and all smaller powers are excluded, which proves the corollary. □

This corollary is about finite geometric activity along a single infinite process. It is neither a statement about choosing a different horizon-dependent rule at each T nor a single-logarithmic impossibility theorem for the full explicit geometric class. No assertion at $\gamma=1$ or a limit through $\beta=0$ is used in this appendix.

F Exact expected coefficients for the fixed geometric family

This appendix proves the exact pathwise and expected coefficients for the full fixed family in Appendix B. The controller, filtration, and reward model are unchanged. The path calculation comes first; an all-order potential bound then supplies the uniform integrability needed for expectation. Appendix F.6 collects the half-endpoint specialization after the general proof.

F.1 Fixed-instance statement and notation

Fix $K\ge 2$, a finite deterministic initialization $\vartheta\in\mathbb{R}^K$, independent pull-indexed IID reward streams in $[0,1]$, and a unique best arm j. Set $\Delta_j=0$ and $\Delta_i=\mu_j-\mu_i>0$ for $i\ne j$. Fix $u>0$, $0<h\le u$, $1/2<\omega\le 1$, and $0<\gamma\le 1/2$ before the process starts. Use exactly (52)–(54), with baseline one and the original pre-action action/reward law. In particular, with

$$\beta=1-\gamma,\quad a=\beta/\gamma,\quad \nu=(2\gamma)^{-1},\quad d=\tfrac{1}{2}-\gamma,\quad L_t=1+\log t,$$

the general-family scalar is

$$\kappa_t=\frac{h}{L_t^{\omega}(\sqrt{t}+\overline{C}_t)},\qquad \eta_t=\kappa_t q_t^{-d},\qquad q_t=1-\max_i p_{i,t}.$$

The tail correction is retained when $\gamma < 1/2$; this is not the endpoint scalar used with a different diagonal. All parameters remain fixed, and h is a design scale subject to the resource ceiling u, not an online choice h_t.

Write

$$A_t = \sum_{s<t} \eta_s, \quad W_\beta = \sum_{i \neq j} \Delta_i^{-1/\beta}, \quad S_\Delta = \sum_{i \neq j} \Delta_i,$$

$$a_i^{\mathrm{asy}} = \frac{W_\beta^{2\beta-1} \Delta_i^{-1/\beta}}{h^2} \quad (i \neq j), \tag{123}$$

$$c_\gamma^{\mathrm{asy}}(\Delta, h, \omega) = \frac{W_\beta^{2\beta-1} \sum_{i \neq j} \Delta_i^{1-1/\beta}}{(2\omega+1)h^2}. \tag{124}$$

The coefficient c_γ^{asy} is distinct from the drift constant $c_\gamma = \nu\beta$ in Appendix B. Retain $Q_i(T), N_i(T), r_t, \widehat{R}_T, R_T$ from (101) and (102), and write $\mathcal{R}_T = \sum_{t \le T} r_t$ for cumulative predictable pseudo-regret.

Theorem F.1 (Exact fixed-instance expected coefficient)**.** *Under the fixed assumptions above, simultaneously almost surely,*

$$\kappa_t \sim \frac{h}{\sqrt{t} L_t^\omega}, \qquad A_t \sim \frac{h^{2\beta}}{\beta W_\beta^{\beta(2\beta-1)}} \frac{t^\beta}{L_t^{2\beta\omega}}, \tag{125}$$

$$\eta_t \sim \frac{h^{2\beta}}{W_\beta^{\beta(2\beta-1)}} \frac{t^{-\gamma}}{L_t^{2\beta\omega}}, \qquad \overline{C}_t \sim h S_\Delta^{1/(2\beta)} W_\beta^{-(2\beta-1)/2} \frac{\sqrt{t}}{L_t^\omega}, \tag{126}$$

$$\frac{t p_{i,t}}{L_t^{2\omega}} \longrightarrow a_i^{\mathrm{asy}}, \qquad \frac{Q_i(T)}{L_T^{2\omega+1}}, \quad \frac{N_i(T)}{L_T^{2\omega+1}} \longrightarrow \frac{a_i^{\mathrm{asy}}}{2\omega+1} \quad (i \neq j). \tag{127}$$

Both $\widehat{R}_T / L_T^{2\omega+1}$ and $\mathcal{R}_T / L_T^{2\omega+1}$ converge to c_γ^{asy} almost surely and in L^1. In particular,

$$\lim_{T \to \infty} \frac{R_T}{L_T^{2\omega+1}} = c_\gamma^{\mathrm{asy}}(\Delta, h, \omega). \tag{128}$$

For each $i \neq j$, also

$$\frac{t\, \mathbb{E} p_{i,t}}{L_t^{2\omega}} \longrightarrow a_i^{\mathrm{asy}}, \qquad \frac{\mathbb{E} N_i(T)}{L_T^{2\omega+1}} \longrightarrow \frac{a_i^{\mathrm{asy}}}{2\omega+1}. \tag{129}$$

All limits concern one fixed positive-gap instance and fixed design. Neither the convergence onset nor the moment constants below are asserted uniform as a gap vanishes, $\gamma \downarrow 0$, $\omega \downarrow 1/2$, $h \downarrow 0$, or the initialization varies without a bound.

We first prove the path limits, then the all-order estimate that permits the transfer to expectation. The supporting finite-time inequalities are proved for the same process in Appendix B; their extension to every fixed integer power is supplied below.

F.2 The general-parameter path calculation

Proof of the almost-sure assertions in Theorem F.1. **Step 1. Finite activity and the compensated allocation law.** Every nonmaximal probability is at most q_t. With $\alpha = 1 - 2\gamma \in [0, 1)$, the maximizing contribution to $a_\gamma(p_t)$ is at most $q_t \le q_t^\alpha$, and every other contribution is at most q_t^α. Thus

$$a_\gamma(p_t) \le K q_t^{1-2\gamma}, \qquad \sum_t \eta_t^2 a_\gamma(p_t) \le K \sum_t \kappa_t^2 \le K \mathcal{S}_\omega < \infty, \tag{130}$$

where $\mathcal{S}_\omega = h^2(1 + (2\omega-1)^{-1})$ is the deterministic budget in (56). The endpoint $\alpha = 0$ is included, because the finite-time policy is interior.

For completeness, the relevant compensated argument can be specialized directly to this process. Put $v_{i,t} = \theta_{i,t+1} - \theta_{i,t}$. Its true action kernel, $0 \le \ell_t \le 1$, and (55) give

$$|v_{i,t}| \le h, \qquad \mathbb{E}_t \|v_t\|_2^2 \le \eta_t^2 a_\gamma(p_t).$$

Consequently $\sum_t \|v_t\|_2^2 < \infty$ almost surely by Tonelli, and each martingale $M_i(t) = \sum_{s<t}(v_{i,s} - \mathbb{E}_s v_{i,s})$ is bounded in L^2 and converges almost surely. There are finitely many coordinates, so these conclusions hold simultaneously on one probability-one event.

Let $\Phi_t = \sum_i e^{\gamma\theta_{i,t}}$, $w_{i,t} = e^{\gamma\theta_{i,t}}/\Phi_t$, and $Z_t = \sum_i e^{\theta_{i,t}}$. The update gives $\sum_i w_{i,t} v_{i,t} = 0$ exactly. Jensen's inequality and the bound $e^x \le 1 + x + e^{\gamma h}x^2/2$ for $|x| \le \gamma h$ imply

$$0 \le \log \frac{\Phi_{t+1}}{\Phi_t} \le \frac{\gamma^2 e^{\gamma h}}{2}\|v_t\|_2^2.$$

Thus Φ_t has a finite positive limit. Each logit is bounded above on the path, and $Z_t^\gamma \le \Phi_t$ bounds Z_t above. The exact conditional drift is

$$\mathbb{E}_t v_{i,t} = \eta_t p_{i,t}^\beta (r_t - \Delta_i), \qquad \theta_{i,t} = \vartheta_i + \sum_{s<t} \eta_s p_{i,s}^\beta (r_s - \Delta_i) + M_i(t). \tag{131}$$

For the best coordinate the drift sum is nondecreasing. Its logit is bounded above and its martingale converges, so this sum is finite and $\theta_{j,t}$ has a finite real limit. The upper bound on Z_t then gives an eventual positive path-dependent lower bound on $p_{j,t}$. It follows that $\sum_t \eta_t r_t < \infty$. For every coordinate, the positive sum $\sum_{s<t} \eta_s p_{i,s}^\beta r_s$ converges and the subtracted sum $\Delta_i \sum_{s<t} \eta_s p_{i,s}^\beta$ is nondecreasing. Therefore every logit converges to a finite value or to $-\infty$. The best logit has a finite limit, so $Z_t \to Z_\infty \in (0, \infty)$; all probabilities and r_t have limits.

If $A_\infty < \infty$, the conditional drifts in (131) are absolutely summable. All logits then have finite limits and every limiting probability is positive. Continuity of C_γ on the interior implies that its running maximum is bounded, and q_t^{-d} has a finite positive limit. The scalar would satisfy $\eta_t \sim c/(\sqrt{t} L_t^\omega)$ for a positive path-dependent c. That series diverges, a contradiction. Hence $A_t \to \infty$. Since r_t has a limit and $\sum_t \eta_t r_t < \infty$, its limit is zero. The best arm is unique, so $p_{j,t} \to 1$ and $p_{i,t} \to 0$ for $i \ne j$.

For $i \ne j$ put $y_{i,t} = \theta_{i,t} - M_i(t)$ and $X_{i,t} = e^{-\beta y_{i,t}}$. The compensated drift gives $y_{i,t+1} - y_{i,t} = -\eta_t p_{i,t}^\beta (\Delta_i - r_t)$. The quantity $x_{i,t} = \beta\eta_t p_{i,t}^\beta (\Delta_i - r_t)$ tends to zero because $\eta_t \le h$ and $p_{i,t} \to 0$. With $g(x) = (e^x - 1)/x$ for $x \ne 0$ and $g(0) = 1$, this yields

$$\begin{aligned}\frac{X_{i,t+1} - X_{i,t}}{\eta_t} &= \beta(\Delta_i - r_t) X_{i,t} p_{i,t}^\beta g(x_{i,t}) \\ &= \beta(\Delta_i - r_t) e^{\beta M_i(t)} Z_t^{-\beta} g(x_{i,t}) \longrightarrow \beta\Delta_i e^{\beta M_i(\infty)} Z_\infty^{-\beta}.\end{aligned}$$

All η_t are positive. Summing with weights η_t and dividing by $A_t \to \infty$ transfers this limit to $X_{i,t}/A_t$: a fixed prefix vanishes and all sufficiently late slopes are within any fixed error of the displayed limit. As $p_{i,t}^{-\beta} = X_{i,t} e^{-\beta M_i(t)} Z_t^\beta$, the random scale cancels and we recover the manuscript's allocation law (115):

$$\frac{p_{i,t}^{-\beta}}{A_t} \longrightarrow \beta\Delta_i \quad (i \ne j). \tag{132}$$

No deterministic-flow approximation or future-conditioned sampler has entered this calculation.

Step 2. Current cap and running maximum. For $t \ge 2$ define

$$Q_\Delta = \sum_{i \ne j} (\beta\Delta_i)^{-1/\beta} = \beta^{-1/\beta} W_\beta, \qquad z_t = A_t^{1/(2\beta)}.$$

Here $A_2 = \eta_1 > 0$. Eventually the best arm maximizes the policy, so (132) gives $q_t \sim Q_\Delta z_t^{-2}$. Also $S_\gamma(p_t) \to 1$, $p_{j,t}^{-\beta}/A_t \to 0$, and the suboptimal terms converge to $\beta\Delta_i$. The cap definition therefore gives

$$\frac{C_\gamma(p_t)}{z_t} \longrightarrow c_C := (\beta S_\Delta)^{1/(2\beta)} > 0.$$

For $0 < \varepsilon < c_C$, choose a finite path-dependent n_0 after which $(c_C - \varepsilon) z_s \le C_\gamma(p_s) \le (c_C + \varepsilon) z_s$. Let $H_0 = \max_{s<n_0} C_\gamma(p_s)$, with value zero if the prefix is empty. Monotonicity of z_t gives, for $t \ge n_0$,

$$(c_C - \varepsilon) z_t \le \overline{C}_t \le \max\{H_0, (c_C + \varepsilon) z_t\}.$$

Divide by $z_t \to \infty$, then let $\varepsilon \downarrow 0$. Thus $\overline{C}_t/z_t \to c_C$; the running maximum has not been replaced by the current cap without proof.

Step 3. The transformed clock and calendar time. Put $\rho = 1/(2\beta)$. Since $\eta_t \le h$ and $A_t \to \infty$, $\eta_t/A_t \to 0$. The relative power expansion gives

$$z_{t+1} - z_t = \rho A_t^{\rho-1}\eta_t(1+o(1)), \qquad \frac{z_{t+1}-z_t}{\kappa_t} = \rho A_t^{\rho-1} q_t^{-d}(1+o(1)) \longrightarrow \frac{Q_\Delta^{-d}}{2\beta}.$$

The last identity uses $\rho - 1 + d/\beta = 0$. This finite positive slope and $z_t \to \infty$ imply $\sum_t \kappa_t = \infty$; otherwise the increments of z_t would be summable after a finite prefix. Weighted summation then gives $z_t \sim [Q_\Delta^{-d}/(2\beta)] \sum_{s<t} \kappa_s$.

First use only the eventual upper slope bound and $\kappa_s \le h/(\sqrt{s} L_s^\omega)$. The deterministic comparison is

$$\sum_{s<t} \frac{1}{\sqrt{s} L_s^\omega} \sim \frac{2\sqrt{t}}{L_t^\omega}. \tag{133}$$

Indeed $f(x) = x^{-1/2}(1+\log x)^{-\omega}$ is positive and decreasing, whereas $F(x) = 2\sqrt{x}(1+\log x)^{-\omega}$ diverges and satisfies $F'(x) = f(x)(1-2\omega/(1+\log x))$. Integrating eventual upper and lower ratio bounds on F'/f proves the integral equivalence, and the monotone sum–integral error is bounded. Hence $z_t = O(\sqrt{t}/L_t^\omega)$ on the path. Together with $\overline{C}_t/z_t \to c_C$, this proves $\overline{C}_t/\sqrt{t} \to 0$ before using the precise scalar asymptotic. Now

$$\kappa_t \sim \frac{h}{\sqrt{t} L_t^\omega}, \qquad \sum_{s<t} \kappa_s \sim \frac{2h\sqrt{t}}{L_t^\omega}, \qquad z_t \sim \frac{h Q_\Delta^{-d}}{\beta} \frac{\sqrt{t}}{L_t^\omega}.$$

Raising the last expression to 2β, and substituting $Q_\Delta = \beta^{-1/\beta} W_\beta$ and $2d = 2\beta - 1$, gives the A_t formula in (125). Multiplication by c_C gives the cap formula in (126). Substitution of $q_t \sim Q_\Delta z_t^{-2}$ into $\eta_t = \kappa_t q_t^{-d}$ gives the scalar formula there.

Step 4. Counts and pathwise regret. From (132), $p_{i,t} \sim (\beta\Delta_i)^{-1/\beta} z_t^{-2}$. The coefficient simplifies using

$$\beta^2 Q_\Delta^{2d} (\beta\Delta_i)^{-1/\beta} = W_\beta^{2\beta-1} \Delta_i^{-1/\beta}.$$

This proves the probability limit in (127). The function $(1+\log x)^{2\omega}/x$ is decreasing after a finite index and has antiderivative $(1+\log x)^{2\omega+1}/(2\omega+1)$. Consequently

$$\sum_{t \le T} \frac{L_t^{2\omega}}{t} \sim \frac{L_T^{2\omega+1}}{2\omega+1}. \tag{134}$$

Eventual ratio bounds transfer the probability limit to $Q_i(T)$; in particular $Q_i(T) \to \infty$.

Under the unchanged action law, the process

$$\sum_{t=1}^{T} \frac{\mathbf{1}\{I_t = i\} - p_{i,t}}{1 + Q_i(t)}$$

is a martingale. Its total conditional variance is bounded by

$$\sum_t \frac{p_{i,t}}{(1+Q_i(t))^2} \le \int_0^\infty (1+x)^{-2}\, dx = 1,$$

where each summand is bounded by the integral over $[Q_i(t-1), Q_i(t)]$. It converges almost surely by the L^2 martingale convergence theorem. Summation by parts, or the pathwise Kronecker lemma with the nondecreasing divergent denominators $1+Q_i(t)$, then gives $(N_i(T)-Q_i(T))/Q_i(T) \to 0$. To see the summation step directly, choose a tail of the convergent series whose partial sums are all at most ε in absolute value. Its weighted sum with nondecreasing denominators, divided by the last denominator, is at most 2ε in absolute value; the fixed prefix vanishes. Let $\varepsilon \downarrow 0$. This proves the count limit. Multiplying the finitely many limits by Δ_i and summing proves the two almost-sure regret limits with coefficient (124). □

F.3 ALL-ORDER POTENTIAL DECAY WITH AN EXPLICIT FINITE-TIME BOUND

Use the comparison potential already defined in Appendix B, with the local alias

$$U_{j,t} = (S_\gamma(p_t)p_{j,t}^{-\gamma})^a - 1, \qquad V_t = W_j(p_t) = U_{j,t}^\nu.$$

Set $V_0 = W_j(p_1) > 0$ and $C_0 = C_\gamma(p_1) > 0$; the subscript zero here denotes a deterministic initial constant, not an extra round. Let $\delta = \min_{i\neq j}\Delta_i > 0$, $\Delta_{\max} = \max_{i\neq j}\Delta_i$, and $c_{\mathrm{d}} = \nu\beta$. For every positive integer r, extend the finite-step constant to

$$d_r = 2\gamma^2\big[(r\nu a)^2 + r\nu a\big]e^{(2\gamma r\nu a+4\gamma)h}. \tag{135}$$

For $r = 1, 2, 4$ this is the identical constant of Appendix B. Retain its constants C_W, C_I, D_C and put $G = C_W C_I$. Define deterministic cap-moment bounds by

$$\mathcal{B}_0 = 1, \qquad e_k = ke^{(k-1)h}D_C h \quad (k = 2, 4, \ldots), \qquad \mathcal{B}_{2m} = \max\{C_0^{2m}, e_{2m}\mathcal{B}_{2m-2}/m\}. \tag{136}$$

For each positive integer r, set

$$\begin{aligned} \mathcal{M}_{4r} &= 2^{4r-1}(1 + V_0^{4r}e^{d_{4r}\mathcal{S}_\omega}), \\ \mathcal{H}_r &= h^{-r}\{2^{2r-1}(1+\mathcal{B}_{2r})\mathcal{M}_{4r}\}^{1/2}, & \mathcal{D}_r &= \mathcal{H}_r^{1/r}, \\ k_r &= \frac{c_{\mathrm{d}}\delta}{G^2\mathcal{D}_r(1+d_rh^2)}, & P_t^{(r)} &= \prod_{s<t}\left(1 + \frac{d_rh^2}{sL_s^{2\omega}}\right), \end{aligned} \tag{137}$$

$$\Psi_r(t) = P_t^{(r)}\left[V_0^{-1} + k_r\sum_{s<t}\frac{1}{\sqrt{s}L_s^\omega}\right]^{-r}. \tag{138}$$

An empty product is one and an empty sum is zero. These constants are finite for each fixed instance and design, but may be very large.

Lemma F.2 (All-order decay)**.** *For every positive integer r there is a deterministic finite A_r such that, for every $t \geq 1$,*

$$\mathbb{E}V_t^r \leq \Psi_r(t) \leq A_r L_t^{r\omega}t^{-r/2}. \tag{139}$$

Consequently the same fixed instance satisfies the finite-time bounds

$$\mathbb{E}r_t \leq \Delta_{\max}\min\{1, \Psi_2(t)\}, \qquad R_T \leq \Delta_{\max}\sum_{t\leq T}\min\{1, \Psi_2(t)\}. \tag{140}$$

Proof. **Step 1. Every integer power retains the regret drift.** The finite-step computation of Appendix B.3 extends to every integer $r \geq 1$ because $k = r\nu \geq 1$. Here are the needed derivative and remainder bounds. For $U(x) = (1+x)^a - 1$ and $f(x) = U(x)^k$, $a, k \geq 1$ and $x > 0$,

$$x(1+x)^{a-1} \leq U(x) \leq ax(1+x)^{a-1}, \qquad 0 \leq \frac{xf'(x)}{f(x)} \leq ka, \qquad 0 \leq \frac{x^2f''(x)}{f(x)} \leq (ka)^2.$$

For the last inequality, differentiation separates $k(k-1)(xU'/U)^2 + kx^2U''/U$; the bounds $xU'/U \leq a$ and $x^2U''/U \leq a(a-1)$ give at most $k(k-1)a^2 + ka(a-1) \leq (ka)^2$.

Write $B_j = \sum_{i\neq j}e^{\gamma(\theta_i-\theta_j)}$ and $\alpha_i^{(j)} = e^{\gamma(\theta_i-\theta_j)}/B_j$. The weights are positive and sum to one. The chain rule and Cauchy's inequality give, for the actual increment v,

$$D^2[f(B_j)](\theta)[v, v] \leq \gamma^2((ka)^2 + ka)f(B_j)\sum_{i\neq j}\alpha_i^{(j)}(v_i - v_j)^2.$$

Along $\theta + sv$ for $0 \leq s \leq 1$, $|v_i| \leq h$ bounds the B_j ratio by $e^{2\gamma h}$, the $f(B_j)$ ratio by $e^{2\gamma kah}$, and each weight ratio by $e^{4\gamma h}$. The second bound follows by integrating $f'/f \leq ka/x$; it does not drop the subtraction in U. The original-kernel contrast estimate (59) is $\mathbb{E}_t(v_i - v_j)^2 \leq 4\kappa_t^2$. Taylor's integral remainder has total integration weight $1/2$, so the conditional remainder is at most $d_r\kappa_t^2V_t^r$ with (135).

The exact first derivative in (58), using $\nabla\Phi^\top v = 0$ and the original reward means, is $-r\nu\beta\eta_t S_\gamma(p_t)^a U_{j,t}^{r\nu-1} r_t$. Thus

$$\mathbb{E}_t V_{t+1}^r \le V_t^r - r\nu\beta\eta_t S_\gamma(p_t)^a U_{j,t}^{r\nu-1} r_t + d_r\kappa_t^2 V_t^r.$$

Here $S_\gamma \ge 1$ and $U_j \ge B_j \ge q_t^\gamma$. If j maximizes the policy, the latter inequality follows from $\sum_{i\neq j} p_i^\gamma \ge q_t^\gamma$ and $p_j^{-\gamma} \ge 1$; otherwise B_j contains a ratio at least one. Since $\gamma(\nu-1) = d$, we have $\eta_t U_{j,t}^{\nu-1} \ge \kappa_t$. Therefore

$$\mathbb{E}_t V_{t+1}^r \le (1 + d_r\kappa_t^2) V_t^r - r c_{\mathrm{d}} \kappa_t V_t^{r-1} r_t. \tag{141}$$

Discarding only the nonpositive term, and using the deterministic bound $\kappa_t^2 \le h^2/(tL_t^{2\omega})$, gives by finite induction

$$\sup_t \mathbb{E} V_t^s \le V_0^s e^{d_s \mathcal{S}_\omega} \qquad (s = 1, 2, \ldots). \tag{142}$$

No independence of the cap and potential is used.

Step 2. Global comparison of potential and policy error. Write $q_{j,t} = 1 - p_{j,t}$, which need not equal the control tail q_t. Subadditivity of x^γ gives $B_j \ge (q_j/p_j)^\gamma$, and $U_j \ge B_j$ since $a \ge 1$. Because $\gamma\nu = 1/2$, $V \ge (q_j/p_j)^{1/2} \ge \sqrt{q_j}$. In the other direction, (65) gives $V \le C_W\sqrt{q_j}\, p_j^{-a/2}$ and $p_j^{-a/2} \le C_I(1+V)$. These inequalities follow from $U_j \le aB_j(1+B_j)^{a-1}$, $B_j \le (K-1)^\beta q_j^\gamma p_j^{-\gamma}$, $1 + B_j \le K^\beta p_j^{-\gamma}$, and $1 + U_j \ge p_j^{-\beta}$ followed by $(1+U_j)^\nu \le 2^{\nu-1}(1+U_j^\nu)$. Thus, globally on the interior simplex,

$$\frac{V_t^2}{G^2(1+V_t)^2} \le q_{j,t} \le V_t^2, \qquad \frac{\delta V_t^2}{G^2(1+V_t)^2} \le r_t \le \Delta_{\max}\min\{1, V_t^2\}. \tag{143}$$

The best arm is not assumed already to dominate.

Step 3. Every even moment of the running cap. The positive-increment estimate (64) states

$$\mathbb{E}_t(\overline{C}_{t+1} - \overline{C}_t) \le D_C\kappa_t, \qquad \overline{C}_t \le \overline{C}_{t+1} \le e^h \overline{C}_t, \qquad \kappa_t \overline{C}_t \le h.$$

For even $k \ge 2$ and $x \le x' \le e^h x$, the mean-value theorem gives $x'^k - x^k \le k e^{(k-1)h} x^{k-1}(x' - x)$. Conditioning before the action consequently yields

$$\mathbb{E}_t \overline{C}_{t+1}^{\,k} \le \overline{C}_t^{\,k} + e_k \overline{C}_t^{\,k-2}.$$

Induction on m and then on time proves

$$\mathbb{E}\overline{C}_t^{\,2m} \le \mathcal{B}_{2m} t^m \qquad (m = 0, 1, \ldots). \tag{144}$$

At $t = 1$ this is the initial bound in (136). For the induction step, the added term is at most $e_{2m}\mathcal{B}_{2m-2}t^{m-1}$, which is at most $m\mathcal{B}_{2m}t^{m-1} \le \mathcal{B}_{2m}((t+1)^m - t^m)$. This proves the recurrence without assuming any moment of a random entry time.

Step 4. Keep cap–potential dependence in the moment estimate. Let $y_t = \mathbb{E}V_t^r$. Cauchy–Schwarz, followed by (142) and (144), gives

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{(1+V_t)^2}{\kappa_t}\right)^r\right] &= h^{-r} L_t^{r\omega} \mathbb{E}[(\sqrt{t} + \overline{C}_t)^r (1+V_t)^{2r}] \\
&\le h^{-r} L_t^{r\omega} \{\mathbb{E}(\sqrt{t} + \overline{C}_t)^{2r} \mathbb{E}(1+V_t)^{4r}\}^{1/2} \\
&\le \mathcal{H}_r t^{r/2} L_t^{r\omega}.
\end{aligned} \tag{145}$$

Indeed the two factors inside braces are at most $2^{2r-1}(1+\mathcal{B}_{2r})t^r$ and $\mathcal{M}_{4r}$, respectively. There is no factorization of a joint expectation by independence. Apply Hölder's inequality with conjugate exponents $(r+1)/r$ and $r+1$ to the identity

$$V_t^r = \left(\frac{\kappa_t V_t^{r+1}}{(1+V_t)^2}\right)^{r/(r+1)} \left(\frac{(1+V_t)^2}{\kappa_t}\right)^{r/(r+1)}.$$

Taking expectations, raising to $(r+1)/r$, and using (145) proves

$$\mathbb{E}\frac{\kappa_t V_t^{r+1}}{(1+V_t)^2} \geq \frac{y_t^{1+1/r}}{\mathcal{D}_r\sqrt{t}L_t^{\omega}}. \tag{146}$$

Substitute this and (143) into (141). With $\alpha_t = d_r h^2/(tL_t^{2\omega})$ and $b_r = rc_{\mathrm{d}}\delta/(G^2\mathcal{D}_r)$, we obtain the deterministic recurrence

$$y_{t+1} \leq (1+\alpha_t)y_t - \frac{b_r}{\sqrt{t}L_t^{\omega}}y_t^{1+1/r}. \tag{147}$$

Step 5. Solve the recurrence discretely. Normalize by $P_t^{(r)}$ from (137) and put $\widetilde{y}_t = y_t/P_t^{(r)}$. Since $1 \leq P_t^{(r)} \leq e^{d_r\mathcal{S}_\omega}$ and $(P_t^{(r)})^{1/r}/(1+\alpha_t) \geq 1/(1+d_r h^2)$, (147) implies

$$\widetilde{y}_{t+1} \leq \widetilde{y}_t - \frac{b_r}{(1+d_r h^2)\sqrt{t}L_t^{\omega}}\widetilde{y}_t^{1+1/r}.$$

The policy is interior and $K \geq 2$, so $V_t > 0$ at each finite time. Thus $y_t, \widetilde{y}_t > 0$, and the factor subtracted from one in this inequality is strictly below one. For $0 \leq x < 1$, convexity gives $(1-x)^{-1/r} \geq 1 + x/r$. Taking inverse powers consequently gives

$$\widetilde{y}_{t+1}^{-1/r} \geq \widetilde{y}_t^{-1/r} + \frac{k_r}{\sqrt{t}L_t^{\omega}}.$$

Sum this inequality, using $y_1 = V_0^r$ and $P_1^{(r)} = 1$. The resulting bound is exactly $\mathbb{E}V_t^r \leq \Psi_r(t)$. Furthermore

$$\sum_{s<t}\frac{1}{\sqrt{s}L_s^{\omega}} \geq \frac{2(\sqrt{t}-1)}{L_t^{\omega}} \geq \frac{\sqrt{t}}{L_t^{\omega}} \quad (t \geq 4).$$

The first inequality follows by $L_s \leq L_t$ and the integral lower bound on $\sum_{s<t} s^{-1/2}$. Thus, for example,

$$A_r = e^{d_r\mathcal{S}_\omega}\max\{k_r^{-r}, 3^{r/2}V_0^r\}$$

works in (139), including $t = 1, 2, 3$. Finally (143) gives $\mathbb{E}r_t \leq \Delta_{\max}\min\{1, \mathbb{E}V_t^2\}$. The $r = 2$ bound and finite summation prove (140). □

F.4 Uniform integrability and transfer to expectation

Completion of Theorem F.1. By (143) and the $r = 4$ case of Lemma F.2,

$$\|r_t\|_2 \leq \Delta_{\max}(\mathbb{E}V_t^4)^{1/2} \leq \Delta_{\max}\sqrt{A_4}\frac{L_t^{2\omega}}{t}.$$

Minkowski's inequality and $\sum_{t\leq T} L_t^{2\omega}/t \leq L_T^{2\omega+1}$ imply

$$\|\mathcal{R}_T\|_2 \leq \Delta_{\max}\sqrt{A_4}L_T^{2\omega+1}. \tag{148}$$

The centered action-regret sum $M_T^R = \widehat{R}_T - \mathcal{R}_T$ is a martingale under the original filtration, with increments $\Delta_{I_t} - r_t$. Orthogonality of its increments and $0 \leq \Delta_{I_t} \leq \Delta_{\max}$ give

$$\mathbb{E}(M_T^R)^2 = \sum_{t\leq T}\mathbb{E}\,\mathrm{Var}(\Delta_{I_t} \mid \mathcal{F}_t) \leq \Delta_{\max}\sum_{t\leq T}\mathbb{E}r_t = \Delta_{\max}R_T.$$

The $r = 2$ case of the same lemma yields $R_T \leq \Delta_{\max}A_2L_T^{2\omega+1}$. Since $L_T \geq 1$, this and (148) show

$$\sup_{T\geq 1}\mathbb{E}\left(\frac{\widehat{R}_T}{L_T^{2\omega+1}}\right)^2 < \infty, \qquad \sup_{T\geq 1}\mathbb{E}\left(\frac{\mathcal{R}_T}{L_T^{2\omega+1}}\right)^2 < \infty. \tag{149}$$

Bounded second moments imply uniform integrability: for a variable X in either normalized family, $\mathbb{E}[|X|\mathbf{1}\{|X| > M\}] \leq \mathbb{E}X^2/M$, uniformly. The already proved almost-sure limits are the deterministic finite number c_γ^{asy}. Truncation at M, bounded convergence on the truncated variables, and the uniform tail estimate therefore give convergence in L^1. Taking expectations proves (128).

For the allocation limit, $p_{i,t} \le q_{j,t} \le V_t^2$ and the same $r = 4$ bound give

$$\sup_{t\ge 1} \mathbb{E}\left(\frac{tp_{i,t}}{L_t^{2\omega}}\right)^2 \le A_4 < \infty.$$

Uniform integrability thus transfers the pathwise probability limit in (127) to its expectation. Finite-time sampling compensation gives $\mathbb{E}N_i(T) = \sum_{t\le T} \mathbb{E}p_{i,t}$. Deterministic weighted summation using (134) then proves the count expectation in (129). □

At $\gamma = 1/2$, the exact expected coefficient is $\sum_{i\ne j} \Delta_i^{-1}/[(2\omega+1)h^2]$. Corollary F.4 specializes this theorem's clock, cap and allocation formulas; it is a consequence, not a premise, of this proof. The exact expectation limit uses (149). For fixed means and design, this leading coefficient does not depend on the bounded IID reward distributions beyond their means. This does not assert that finite-time performance is independent of reward noise or that convergence onset is uniform across distributions.

F.5 The endpoint minimizes the fixed-order coefficient

Corollary F.3 (Fixed-order coefficient comparison)**.** *Fix the positive gaps, $h > 0$, and $\omega \in (1/2, 1]$. For $K \ge 3$, the coefficient c_γ^{asy} is strictly decreasing in $\gamma \in (0, 1/2]$; its unique minimum is at $\gamma = 1/2$. For $K = 2$, it is independent of γ on this interval. At fixed ω and common resource restriction $0 < h \le u$,*

$$\min_{0<\gamma\le 1/2,\ 0<h\le u} c_\gamma^{\mathrm{asy}}(\Delta, h, \omega) = \frac{\sum_{i\ne j}\Delta_i^{-1}}{(2\omega+1)u^2}. \tag{150}$$

For $K \ge 3$ the minimizer is uniquely $(\gamma, h) = (1/2, u)$; for $K = 2$ every permitted γ minimizes at $h = u$. This is both pathwise- and expected-asymptotic-coefficient optimality within this fixed family and fixed logarithmic order, not a finite-time optimizer or a comparison with other controllers.

Proof. Set $a_i = \Delta_i^{-1}$, $a_{\max} = \max_{i\ne j} a_i$, $x_i = a_i/a_{\max} \in (0, 1]$, and $\rho = 1/(1-\gamma) \in (1, 2]$. Up to the fixed positive factor $a_{\max}/[(2\omega+1)h^2]$, (124) becomes

$$f(\rho) = A(\rho)^{2/\rho-1}B(\rho), \qquad A(\rho) = \sum_{i\ne j} x_i^\rho, \qquad B(\rho) = \sum_{i\ne j} x_i^{\rho-1}.$$

These are positive differentiable finite sums. Direct differentiation gives

$$\frac{d}{d\rho}\log f(\rho) = -\frac{2\log A(\rho)}{\rho^2} + \left(\frac{2}{\rho} - 1\right)\frac{\sum_{i\ne j} x_i^\rho \log x_i}{A(\rho)} + \frac{\sum_{i\ne j} x_i^{\rho-1}\log x_i}{B(\rho)}.$$

Since one x_i is one, $A(\rho) \ge 1$. Also $\log x_i \le 0$ and $2/\rho - 1 \ge 0$, so every term is nonpositive. If $K \ge 3$, there are at least two positive terms in A, one equal to one; hence $A(\rho) > 1$ and the first term is strictly negative. Because ρ strictly increases with γ, this proves strict decrease. If $K = 2$, the sole x_i equals one, so the derivative vanishes and the coefficient equals $[(2\omega+1)h^2\Delta_i]^{-1}$ for every γ. At $\gamma = 1/2$, $2\beta - 1 = 0$ and $\sum_{i\ne j}\Delta_i^{1-1/\beta} = \sum_{i\ne j}\Delta_i^{-1}$. Finally the strictly decreasing factor h^{-2} for $h > 0$ gives (150) and the stated equality cases. □

For equal gaps Δ and $m = K - 1$ suboptimal arms, substitution into (124) gives

$$c_\gamma^{\mathrm{asy}} = \frac{m^{2(1-\gamma)}}{(2\omega+1)h^2\Delta}, \qquad \frac{c_\gamma^{\mathrm{asy}}}{c_{1/2}^{\mathrm{asy}}} = m^{1-2\gamma}.$$

This simplified ratio is not asserted for heterogeneous gaps. There is no attained eventual expected-regret minimizer over the permitted interval of fixed ω, whose lower endpoint is open. For any fixed $\omega > 1/2$ choose $\omega' \in (1/2, \omega)$ and keep γ, h and the instance fixed. Let $R_T^{(\omega)}$ denote expected

regret with decay parameter ω and all other parameters held fixed. Theorem F.1 gives two positive coefficients and therefore

$$\frac{R_T^{(\omega')}}{R_T^{(\omega)}} \sim \frac{c_\gamma^{\mathrm{asy}}(\Delta, h, \omega')}{c_\gamma^{\mathrm{asy}}(\Delta, h, \omega)} L_T^{2(\omega'-\omega)} \longrightarrow 0.$$

This comparison has no onset uniformity near $\omega = 1/2$. The finite-time envelope (140) may have very conservative constants; it is not a practical crossover guarantee. No part of this appendix identifies the exact-real process with a floating-point implementation, expands the initialization domain of the scalar separation theorem, or proves universal finite-time dominance.

F.6 Exact pathwise asymptotics at $\gamma = 1/2$

This section specializes the same infinite exact-real process of Appendix B; it does not introduce another controller. Fix a finite $K \geq 2$, a deterministic finite initialization ϑ, independent pull-indexed IID reward streams in $[0,1]$, and a unique best arm j. Write $\Delta_i = \mu_j - \mu_i > 0$ for $i \neq j$. Fix $u > 0$, $0 < h \leq u$, $\gamma = 1/2$, and $1/2 < \omega \leq 1$, with baseline $b_t = 1$. In this specialization, $\beta = 1/2$, $a = \nu = 1$, and $d = 0$ in Appendix B. Thus, with $L_t = 1 + \log t$,

$$C(p) = \left(\sum_k \sqrt{p_k}\right)\left(\sum_k p_k^{-1/2}\right), \qquad \overline{C}_t = \max_{s \leq t} C(p_s),$$
$$\eta_t = \kappa_t = \frac{h}{L_t^\omega(\sqrt{t} + \overline{C}_t)}, \qquad A_t = \sum_{s<t} \eta_s. \tag{151}$$

The actual update remains $\theta_{i,t+1} = \theta_{i,t} - \eta_t \ell_t p_{i,t}^{-1/2}(\mathbf{1}\{I_t = i\} - p_{i,t})$, where $\ell_t = 1 - Y_t \in [0,1]$ and $I_t \mid \mathcal{F}_t \sim p_t = \mathrm{softmax}(\theta_t)$. Define $Q_i(T) = \sum_{t \leq T} p_{i,t}$, $N_i(T) = \sum_{t \leq T} \mathbf{1}\{I_t = i\}$, $r_t = \sum_{i \neq j} \Delta_i p_{i,t}$, and $\widehat{R}_T = \sum_{i \neq j} \Delta_i N_i(T)$, so $R_T = \mathbb{E}\widehat{R}_T = \mathbb{E}\sum_{t \leq T} r_t$.

Corollary F.4 (Exact pathwise coefficient and fixed-instance expected order)**.** *For every fixed instance and design just specified, the following limits hold simultaneously almost surely:*

$$A_t \sim \frac{2h\sqrt{t}}{L_t^\omega}, \qquad \frac{L_t^\omega \overline{C}_t}{\sqrt{t}} \longrightarrow h \sum_{i \neq j} \Delta_i, \tag{152}$$

$$\frac{h^2 \Delta_i^2 t\, p_{i,t}}{L_t^{2\omega}} \longrightarrow 1 \qquad (i \neq j), \tag{153}$$

$$\frac{Q_i(T)}{L_T^{2\omega+1}}, \quad \frac{N_i(T)}{L_T^{2\omega+1}} \longrightarrow \frac{1}{(2\omega+1)h^2\Delta_i^2} \qquad (i \neq j). \tag{154}$$

Consequently, with the deterministic positive coefficient $c_\Delta = [(2\omega+1)h^2]^{-1} \sum_{i \neq j} \Delta_i^{-1}$,

$$\frac{\widehat{R}_T}{L_T^{2\omega+1}} \longrightarrow c_\Delta, \qquad \frac{\sum_{t \leq T} r_t}{L_T^{2\omega+1}} \longrightarrow c_\Delta \quad \textit{almost surely}. \tag{155}$$

For the unconditional expectation, the following bound also holds

$$0 < c_\Delta \leq \liminf_{T \to \infty} \frac{R_T}{L_T^{2\omega+1}} \leq \limsup_{T \to \infty} \frac{R_T}{L_T^{2\omega+1}} \leq C_{\mathrm{rate}}^2 \sum_{i \neq j} \Delta_i^{-1} < \infty, \tag{156}$$

where C_{rate} *is the existing constant of Theorem B.1 evaluated at* $\gamma = 1/2$*. In particular,* $R_T = \Theta(L_T^{2\omega+1})$ *on this fixed instance. No uniform finite-time onset is asserted.*

Proof. Apply Theorem F.1 at $\gamma = 1/2$. Its fixed K, finite deterministic initialization, bounded independent pull-indexed IID streams, unique best arm, u, h, ω, baseline and pre-action filtration are exactly those specified above. Here $\beta = 1/2$, $a = \nu = 1$, $d = 0$, so $q_t^{-d} = 1$ and the general controller and running cap are precisely (151). In particular,

$$W_\beta = \sum_{i \neq j} \Delta_i^{-2}, \qquad W_\beta^{2\beta-1} = 1, \qquad a_i^{\mathrm{asy}} = \frac{1}{h^2\Delta_i^2}, \qquad c_{1/2}^{\mathrm{asy}} = c_\Delta.$$

Substitution in (125) and (126) gives $A_t \sim 2h\sqrt{t}/L_t^{\omega}$ and $\overline{C}_t \sim h(\sum_{i\neq j}\Delta_i)\sqrt{t}/L_t^{\omega}$. Equation (127) gives (153)–(154). The two almost-sure regret limits of that theorem give (155). Thus the running-maximum, calendar-time and actual-count steps are supplied by the single general-parameter proof in Appendix F.2.

For (156), the same theorem gives the stronger equality $R_T/L_T^{2\omega+1} \to c_\Delta$. This transfer uses the moment estimates and uniform integrability in Appendices F.3–F.4, not an exchange of an almost-sure limit with expectation alone. The finite-time bound of Theorem B.1, at the same endpoint, is

$$R_T \le 2C_{\mathrm{init}} + C_{\mathrm{rate}}^2 L_T^{2\omega+1} \sum_{i\neq j} \Delta_i^{-1}.$$

Its constants are deterministic and finite for the fixed design and initialization. Division by $L_T^{2\omega+1} \to \infty$ proves the stated upper comparison with C_{rate}. □

All limits retain the fixed-instance and fixed-design scope of Theorem F.1. The coefficient h^{-2} does not rank finite-horizon performance as h varies; gaps enter the analysis, not the online controller.

G FIXED-SCHEME FINITE-SCALE EXPERIMENTS

This section specifies the additional fixed-scheme experiments in Section 6. They use a different resource setting, initialization and numerical sampler from the earlier scale experiment; the earlier experiments and their unfavorable outcomes remain separate evidence. All costs below are realized pseudo-regret reconstructed from integer action counts, $Z_T = \sum_i(\mu_* - \mu_i)N_i(T)$, not the sum of realized reward losses. Their population target is $R(T) = \mathbb{E}Z_T$ for the stated numerical process under the independent replicate-stream model. The true means are used in reward generation and offline evaluation, never by a learner.

G.1 FIXED PRIMARY RULE AND NUMERICAL PROTOCOL

The primary DGB rule is fixed throughout at

$$h = 8, \qquad \omega = 0.51, \qquad \gamma = 1/2, \qquad u = 16, \qquad B_{\max} = 1.$$

The anchor has five Bernoulli arms with

$$\mu = (0.625, 0.5, 0.05, 0.05, 0.05), \qquad \theta_1 = (0, 3, 0, 0, 0).$$

The initial logit bias 3 is not the baseline bound. In particular, this initialization is not presented as belonging to the small neighborhood used by the scalar lower-bound theorem. These are fixed-scheme finite-time comparisons, not an empirical optimization over the scalar controller class. The primary rule and the comparator list were inherited from the locked campaign following its earlier screening stages; no environment-specific retuning is made here.

Let p_t denote the vector used by the implementation at action round t, after exactly $t-1$ observations. Write

$$s_t = \sum_i \sqrt{p_{i,t}}, \qquad C_t = \max_{r\le t}\left\{\left(\sum_i \sqrt{p_{i,r}}\right)\left(\sum_i p_{i,r}^{-1/2}\right)\right\}.$$

The primary control is $\eta_t = 8/[(1+\log t)^{0.51}(\sqrt{t}+C_t)]$, baseline one, and diagonal factor $p_{i,t}^{-1/2}$ in coordinate i. Every scalar comparator instead updates $\theta_{t+1} = \theta_t + \eta_t(Y_t - b_t)(e_{I_t} - p_t)$. All its controls are selected from observed history before the next action. The code checks $0 \le \eta_t \le 16(1+10^{-12})$, $|b_t| \le 1$, and finite states; DGB additionally checks $\eta_t p_{i,t}^{-1/2} \le h(1+10^{-8})$. These are numerical check tolerances, not a change to the mathematical resource domain. A violation stops and records the trajectory; it does not clamp an invalid result back into the experiment. The explicit minima in the rules below are part of those rules.

A distinct finite numerical process. Subtract-maximum binary64 softmax is evaluated first. Each nonmaximal positive evaluated probability is represented exactly as an integer dyadic mass. The maximal coordinate receives the positive normalization residual. The denominator starts at 2^{120} and is enlarged when needed, using additional random bits rather than a probability floor. The residual is converted back to binary64 for the stored control vector; the recorded relative distortion guard is 10^{-10}. Thus exact sampling of the constructed dyadic masses does not assert exact softmax sampling or exact real-arithmetic score centering. Nonpositive evaluated probabilities, nonfinite values and failed guards stop a trajectory. There is no clipping, forced exploration, redraw, or logit reset. This is not the older 2^{52}-mass kernel in Appendix H.1.

Rewards use pull-indexed arm streams: a stream advances only when that arm is pulled. Bernoulli outcomes compare a top-53-bit uniform to the stored mean, so a non-grid threshold has a discretization error smaller than 2^{-53}. Action, arm-reward and Thompson-posterior keys are separate. Within an environment, methods share replicate keys and arm-reward streams; logit-based methods share action keys. This coupling supports paired descriptive standard errors. It is not independence across methods. Different environments and the high-repeat, long-joint and 512-repeat cohorts use different environment IDs or namespaces. Statistical coverage assumes independent replicate streams with a common within-cell law; a deterministic pseudorandom generator is not itself a proof of that model.

Checkpoints are recorded after updates $2^{10}, 2^{11}, \ldots, T$, hence seven checkpoints for $T = 2^{16}$ and eleven for $T = 2^{20}$. Counts and logits are post-update. Checkpoints from one trajectory are repeated measurements, not extra independent samples. The primary inference times are $2^{14}, 2^{15}, 2^{16}$ for the broad and high-repeat cohorts, and $2^{16}, 2^{18}, 2^{20}$ for the long cohorts.

G.2 The thirteen scalar comparators

All statistics in this subsection use only observations preceding round t. Let n_i and s_i^Y be arm counts and reward sums, and $\widehat{\mu}_i = (s_i^Y + 1)/(n_i + 2)$. Let j be the first maximizer of $\widehat{\mu}$, let $\widehat{\mu}_{(2)}$ be the second-largest value, and write $g_t = \widehat{\mu}_j - \widehat{\mu}_{(2)}$. Two recurrent baselines are

$$b_t^{\max} = \widehat{\mu}_j, \qquad b_t^{\mathrm{var}} = \frac{\sum_i v_i \widehat{\mu}_i}{\sum_i v_i}, \qquad v_i = p_i \left[\left(\sum_{k \neq i} p_k \right)^2 + \sum_{k \neq i} p_k^2 \right].$$

Both lie in $[0, 1]$. In archived method names, `b-5` and `b-8` are selector codes for these two baselines, not baselines equal to -5 or -8.

Table 2: Locked scalar rules. S1–S13 are stable display aliases for the listed archive names. All warm-up rules continue natural sampling.

ID	Archive name	Predictable control
S1	`S_NM_h16`	Baseline one. Norm-matched scalar with the formula below.
S2	`S_old_b0.5_eta0.3`	$\eta_t = 0.3,\ b_t = 0.5$.
S3	`S_const_b-5_eta0.2`	$\eta_t = 0.2,\ b_t = b_t^{\mathrm{var}}$.
S4	`S_fine_b-5_eta0.26`	$\eta_t = 0.26,\ b_t = b_t^{\mathrm{var}}$.
S5	`S_fine_b0.6_eta0.4`	$\eta_t = 0.4,\ b_t = 0.6$.
S6	`S_nwait8_b-8_eta0.3`	$\eta_t = 0$ until every arm has at least 8 observations, then 0.3; $b_t = b_t^{\max}$.
S7	`S_nwait32_b-5_eta0.3`	The same count trigger at 32, then $\eta_t = 0.3$; $b_t = b_t^{\mathrm{var}}$.
S8	`S_nwait32_b-8_eta0.3`	The same count trigger at 32, then $\eta_t = 0.3$; $b_t = b_t^{\max}$.
S9	`S_wait2048_b-8_eta0.3`	$\eta_t = 0$ for $t \leq 2048$, then 0.3; $b_t = b_t^{\max}$.
S10	`S_conf_eta1_a1`	$b_t = b_t^{\max}$, confidence-scaled control defined below; this name does not certify the learner.

ID	Archive name	Predictable control
S11	`S_switch_w512_eta0.3_promote0_freeze0.5`	Zero updates through round 512, then the empirical-target switching rule below.
S12	`S_gap_c1_b-5`	$\eta_t = \min\{16, g_t\},\ b_t = b_t^{\mathrm{var}}$.
S13	`S_decay_c4_p0.25_b-5`	$\eta_t = 4t^{-1/4},\ b_t = b_t^{\mathrm{var}}$.

The exact S1 scalar is

$$\eta_t = \frac{16}{(1+\log t)^{0.6}(\sqrt{t}+C_t)}\sqrt{\frac{K-2+s_t^2/K}{2\sum_{i<j} p_{i,t}p_{j,t}}}.$$

It is checked against the resource bound, not silently clipped. S10 uses

$$r_t = \sqrt{\log(t+1)\left(\frac{1}{n_j+2}+\frac{1}{\min_{i\neq j}(n_i+2)}\right)},$$
$$\eta_t = \min\{16, \min\{1, \max\{0, g_t/(r_t+10^{-15})\}\}\}.$$

For S11 after its warm-up, let k be the first maximizer of current policy mass. Set $b_t = 0$ if $k = j$, and $b_t = 1$ otherwise; set $\eta_t = 0.3$, except that $\eta_t = 0$ whenever $p_{j,t} \geq 1 - \min\{0.1, t^{-1/2}\}$. This last test is reevaluated each round; it is not an irreversible freeze.

The two long-joint scalar comparators are exactly S2 and S10, without horizon-specific changes. UCB1 initializes by pulling each unobserved arm in index order and then maximizes $s_i^Y/n_i + \sqrt{2\log t/n_i}$, with first-index ties. Bernoulli Thompson sampling uses independent $\mathrm{Beta}(s_i^Y + 1, n_i - s_i^Y + 1)$ posterior draws generated from two gamma draws. These standard bandit references are not asserted to lie in the restricted sampled-score scalar class.

G.3 Cohorts, inferential quantities and simultaneous bounds

Table 3: Separate numerical cohorts. A repeat is a complete trajectory for one method and environment. M counts mean targets, including methods not selected for a display. Family levels are separate, not one global 95% event.

Cohort	Environments	Methods	Repeats/cell	Terminal T	M
Broad panel	53	18	192	65,536	2,862
High-repeat anchor	1	36	4,096	65,536	108
Long joint	1	3	4,096	1,048,576	9
Long reference	1	14	512	1,048,576	42

The broad panel includes the primary DGB, its $(h, \omega, \gamma) = (8, 0.6, 0.5)$ and $(10, 0.51, 0.45)$ variants, all thirteen scalars, UCB1 and Thompson sampling. The high-repeat family contains 21 DGB candidates: the 18 combinations $h \in \{6, 8, 10, 12, 14, 16\}$, $\omega \in \{0.51, 0.55, 0.6\}$, $\gamma = 0.5$, and $(10, 0.51, 0.45), (16, 0.51, 0.4), (16, 0.6, 0.3)$, along with the same thirteen scalars and two bandit references. The primary is not replaced by the sample-minimizing candidate. The 512-repeat family contains the three broad-panel DGB rules, S1–S4, S6–S8, S10–S11 and the two bandit references. The long joint is assembled from the two-method anchor and one-method challenger files with one shared DGB cohort, not duplicated DGB observations.

For each cell the displayed SE is the sample standard deviation, with denominator $n-1$, divided by $\sqrt{n}$. It is descriptive and is not a 95% confidence interval. The simultaneous bounds instead use the deterministic support $[0, B_T]$, where $B_T = T(\max_i \mu_i - \min_i \mu_i)$, never a sample range. They are recomputed from the count-derived costs using the original bounded-mean procedure. Each family uses $\alpha = 0.05$, split equally between fixed-mixture betting and empirical Bernstein bounds, with $\delta = \alpha/(4M)$ for each directional bound in each procedure.

For completeness, for normalized costs $z_r = Z_r/B_T$, the betting mixture at a proposed mean $m \in (0, 1]$ is

$$E_n(m) = \frac{1}{16} \sum_{c \in \mathcal{C}} \prod_{r=1}^{n} (1 - c + c z_r/m),$$

$$\mathcal{C} = \{.0001, .0003, .001, .003, .01, .03, .1, .2, .4, .6, .8, .9, .97, .99, .999, 1\}.$$

If $\mathbb{E}z_r \le m$, independence across replicates implies $\mathbb{E}E_n(m) \le 1$. Markov's inequality bounds rejection at $E_n(m) \ge 1/\delta$ by δ. Monotonicity in m therefore permits inversion for a lower mean limit; applying the same construction to $1 - z_r$ gives an upper limit. The empirical Bernstein radius implemented by the protocol is

$$e_n = \sqrt{\frac{2s^2 \log(2/\delta)}{n}} + \frac{7B_T \log(2/\delta)}{3(n-1)},$$

where s^2 is the sample variance of unnormalized costs. Intersect the two intervals with each other and with $[0, B_T]$, giving $[L_a(t), U_a(t)]$. The union bound uses all $4M$ directional targets and needs no independence across methods or checkpoints. The implementation uses log-sum-exp and 55 bisection steps, retaining the lower bisection endpoint. These reported numerical limits are not formally outward-rounded interval-arithmetic certificates.

For a scalar S versus the primary D, the reported mean saving is $\overline{Z}_S - \overline{Z}_D$, its simultaneous lower bound is $L_S - U_D$, and the relative mean saving is $1 - \overline{Z}_D/\overline{Z}_S$. If $L_S > 0$, its relative lower bound is $1 - U_D/L_S$. These bounds are differences or ratios of simultaneous marginal mean limits, not Gaussian intervals inferred from plotted SEs or intervals for individual trajectories. Paired SEs reported for the long joint are computed separately from within-replicate cost differences. No uncertainty budget is shared across the broad, high-repeat, long-joint and long-reference families.

Every reconstructed trajectory completed with status zero. The analysis would abort on any failure rather than select successful trajectories. The archived failure-aware rule is to retain prefix cost as a lower variable and add $(T - T_{\text{completed}})\Delta_{\max}$ for an upper variable before an explicitly justified continuation of inference.

G.4 High-repeat anchor and long-horizon continuation

At $T = 65,536$, the 4,096-repeat primary mean is 808.7446 (SE 16.7706), with simultaneous mean interval $[730.8157, 919.1278]$ in the original $M = 108$ family. Table 4 retains every scalar comparator. The minimum sample-mean relative saving is 66.3201%, and every scalar has a positive simultaneous saving lower bound, the minimum being 1232.0442. This conclusion concerns the thirteen stated rules at this anchor and horizon.

Table 4: All thirteen scalar anchor comparisons, $n = 4096$, $T = 65,536$. Parentheses are SEs, not confidence limits. The last column is $L_S - U_D$ from the $M = 108$ family.

Rule	Mean (SE)	Mean saving (%)	Saving lower bound
S1	3187.58 (54.91)	74.63	1982.91
S2	2401.27 (47.08)	66.32	1232.04
S3	2426.98 (44.98)	66.68	1274.54
S4	2418.49 (46.80)	66.56	1251.92
S5	2499.28 (49.93)	67.64	1311.73
S6	2436.69 (46.57)	66.81	1272.44
S7	2731.74 (49.14)	70.39	1557.86
S8	2584.77 (46.83)	68.71	1423.89
S9	2720.15 (45.82)	70.27	1571.07
S10	2818.88 (44.74)	71.31	1677.59
S11	3240.15 (53.05)	75.04	2051.30
S12	3195.85 (47.06)	74.69	2045.47
S13	4004.08 (59.25)	79.80	2786.23

At $T = 1,048,576$, the joint-cohort savings over S2 and S10 are respectively 9466.9539 and 4581.0369, with paired SEs 428.1250 and 215.7458. Their relative mean savings are 90.4053%

Table 5: Long-horizon evidence at the same anchor. Upper panel: the same 4,096-repeat joint cohort for DGB, S2 and S10. Lower panel: the distinct 512-repeat cohort; its DGB entry must not be replaced with the upper cohort's DGB mean. Values are mean (SE).

4,096-repeat rule	$T = 65,536$	$T = 262,144$	$T = 1,048,576$
DGB	811.83 (16.22)	920.59 (27.83)	1004.73 (49.07)
S2	2483.85 (47.88)	5334.32 (152.56)	10471.68 (437.82)
S10	2827.45 (44.63)	4453.65 (113.62)	5585.76 (218.65)
512-repeat rule	$T = 1,048,576$, separate reference cohort		
DGB	854.57 (42.31)		
UCB1	345.98 (1.80)		
Thompson	58.14 (0.66)		

and 82.0127%. In the $M = 9$ family, the corresponding absolute lower bounds are 6535.0227 and 2589.9494, and relative lower bounds are 75.0886% and 54.4335%.

The same event also covers every integer horizon between 65,536 and 1,048,576. For $a \le t \le b$, nonnegative count costs imply

$$R_S(t) - R_D(t) \ge L_S(a) - U_D(b), \qquad 1 - \frac{R_D(t)}{R_S(t)} \ge 1 - \frac{U_D(b)}{L_S(a)} \quad \text{if } L_S(a) > 0.$$

Taking the minimum over S2 and S10 and the two adjacent segments $[65,536, 262,144]$, $[262,144, 1,048,576]$ gives an absolute lower bound 1035.0665 and a relative lower bound 45.3811% throughout the integer interval. This uses monotonicity of each expected cumulative cost, not monotonicity of their difference, and does not retune a rule for each intermediate horizon. It is a consequence of the existing nine-mean event, not another independent confidence family or an exact-real regret theorem. The independent-namespace 512-repeat references show lower sample-mean regret for UCB1 and Thompson sampling than for DGB. They prevent interpreting the scalar gains as standard-bandit superiority.

G.5 All 53 broad-panel environments and outcomes

Every environment is Bernoulli, with 192 repetitions per method and $T = 65,536$. Use $\mathbf{0}_m$ and $c\mathbf{1}_m$ for repeated coordinates. Environment IDs determine the following complete panel; the machine-readable protocol retains the full stored binary64 thresholds and logits.

- IDs 0–26 are the Cartesian product $B \in \{2.75, 3, 3.25\}$, $\delta \in \{0.1, 0.125, 0.15\}$, $d \in \{0.4, 0.425, 0.45\}$, in that lexicographic order with d varying fastest. Set $\mu = (0.5 + \delta, 0.5, (0.5 - d)\mathbf{1}_3)$ and $\theta_1 = (0, B, \mathbf{0}_3)$.
- IDs 27–34 have $K = 4, 6, 8, 10$, in pairs for each K, and $\mu = (0.625, 0.5, 0.05\mathbf{1}_{K-2})$. The first member uses $\theta_1 = (0, 3, \mathbf{0}_{K-2})$; the second uses $\theta_1 = (0, 3, \log(3/(K-2))\mathbf{1}_{K-2})$, preserving the anchor's total initial mass on the low-mean arms.
- IDs 35–40 have $K = 5$, $\theta_1 = (0, 3, \mathbf{0}_3)$ and $\mu = (m + 0.125, m, (m - d)\mathbf{1}_3)$, for the ordered pairs $(m, d) = (.45, .35), (.55, .35), (.65, .35), (.75, .35), (.6, .45), (.7, .45)$.
- IDs 41–46 are the six fixed heterogeneous environments in Table 6; the table rounds their definitions to six decimals, while the source protocol retains every digit.
- IDs 47–52 are the six boundary configurations. Relative to the anchor, they respectively use initial bias $B = 0, 1, 6$; best-to-rival gap $\delta = .01, .25$; and three low-arm means .45, with all other anchor coordinates unchanged.

IDs 0–46 are the 47 target environments; IDs 47–52 are the boundary panel. In all 47 targets, the primary mean is below each of the thirteen scalar means, with the minimum within-environment relative saving ranging from 54.1012% to 75.6424%. However, only IDs 31 and 33 have $L_S - U_D > 0$ simultaneously for every scalar in the full $M = 2862$ family. Thus 47/47 is a sample-mean statement, not 47 equally strong confirmations. All six boundary cases are retained: only three have lower DGB mean than every scalar. The failures at IDs 47, 50 and 52 are visible in Table 7.

Table 6: The six fixed heterogeneous broad-panel configurations. The listed vectors are rounded descriptions, not replacement simulator inputs.

ID	Mean vector / initial logit vector
41	$(.629550, .516160, .085899, .113436, .056960)$ $(.119590, 2.902054, -.181902, .137273, .192286)$
42	$(.624820, .493628, .093445, .112896, .113598)$ $(.140478, 3.289573, .059688, .143392, .067044)$
43	$(.619875, .492563, .103943, .095477, .108169)$ $(.146686, 3.121779, .060143, .093856, -.166987)$
44	$(.623692, .497879, .064420, .036106, .068405)$ $(.032060, 2.904919, -.073318, -.030385, -.065065)$
45	$(.653360, .507452, .068368, .061608, .115230)$ $(-.112065, 2.788307, .005145, -.227030, -.127518)$
46	$(.645302, .500816, .118040, .044477, .039688)$ $(-.058278, 2.960947, .081547, -.199745, -.169177)$

Table 7: All broad-panel terminal outcomes, $n = 192$ per cell. $S_{\min}$ is the scalar with lowest *sample* mean in that row, identified using Table 2; it is not an oracle estimate of the best expected scalar cost. Parentheses are SEs. Saving is $100(1 - \overline{Z}_D / \min_S \overline{Z}_S)$. The final column is $\min_S(L_S - U_D)$ over all thirteen scalars, not necessarily the scalar selected in column three.

ID	DGB mean (SE)	$S_{\min}$: mean (SE)	Saving (%)	Min. lower bound
0	786.6 (77.1)	S6: 2124.3 (185.8)	62.97	-1780.1
1	887.0 (70.4)	S4: 2547.8 (196.2)	65.19	-1645.4
2	868.5 (77.9)	S5: 2621.5 (205.0)	66.87	-1732.3
3	614.1 (50.1)	S6: 1586.9 (176.0)	61.30	-2112.2
4	630.4 (64.3)	S3: 1774.8 (189.2)	64.48	-2146.5
5	605.3 (53.0)	S8: 1926.6 (185.6)	68.58	-2061.5
6	531.7 (52.6)	S2: 1337.9 (182.2)	60.26	-2347.9
7	532.4 (44.6)	S3: 1496.2 (179.5)	64.41	-2351.1
8	546.1 (39.6)	S6: 1516.0 (179.9)	63.97	-2433.4
9	894.5 (67.1)	S9: 2546.0 (182.1)	64.87	-1445.9
10	910.0 (70.8)	S4: 2760.7 (192.4)	67.04	-1493.6
11	1062.8 (96.3)	S8: 2974.3 (194.8)	64.27	-1531.7
12	562.9 (32.3)	S2: 1716.3 (184.0)	67.20	-1984.1
13	787.0 (72.9)	S6: 2498.0 (232.7)	68.49	-1831.3
14	839.1 (81.1)	S8: 2425.8 (210.4)	65.41	-1947.7
15	608.6 (44.2)	S2: 1702.6 (204.2)	64.25	-2205.7
16	672.3 (62.0)	S4: 1790.4 (202.2)	62.45	-2290.4
17	611.2 (46.2)	S8: 1977.8 (210.4)	69.10	-2190.5
18	967.7 (70.6)	S2: 2535.3 (189.8)	61.83	-1589.4
19	1194.6 (90.3)	S8: 2968.0 (194.7)	59.75	-1525.1
20	1291.3 (95.4)	S7: 2982.4 (196.5)	56.70	-1733.2
21	807.1 (68.2)	S8: 2451.1 (215.7)	67.07	-1676.5
22	956.6 (82.4)	S8: 2563.9 (216.8)	62.69	-1836.6
23	846.2 (48.8)	S6: 2770.1 (224.6)	69.45	-1708.8
24	817.6 (68.7)	S8: 2254.4 (209.0)	63.73	-2006.7
25	711.3 (55.6)	S6: 2181.3 (222.0)	67.39	-2079.5
26	751.4 (56.9)	S2: 2341.3 (235.3)	67.91	-2128.6
27	577.8 (68.0)	S3: 1277.8 (150.3)	54.78	-2519.5
28	710.2 (79.5)	S6: 1547.4 (171.3)	54.10	-2448.1
29	897.8 (56.3)	S7: 3685.9 (239.7)	75.64	-972.8
30	975.9 (66.3)	S6: 2976.7 (230.2)	67.22	-1675.3
31	1625.3 (74.5)	S5: 5875.7 (239.8)	72.34	324.0

ID	DGB mean (SE)	$S_{\min}$: mean (SE)	Saving (%)	Min. lower bound
32	1333.5 (56.9)	S8: 3707.8 (240.0)	64.04	-1393.3
33	2476.7 (91.8)	S13: 7120.9 (194.4)	65.22	858.1
34	1728.1 (47.5)	S6: 3863.3 (254.7)	55.27	-1654.6
35	699.1 (64.6)	S3: 1771.9 (181.6)	60.55	-1864.6
36	571.4 (40.5)	S3: 1633.0 (173.0)	65.01	-1795.3
37	478.9 (21.3)	S3: 1333.5 (151.7)	64.09	-1910.0
38	489.2 (22.5)	S6: 1137.1 (131.0)	56.98	-1972.7
39	622.2 (49.7)	S4: 1946.6 (184.1)	68.04	-2068.6
40	604.3 (38.2)	S6: 1773.6 (178.5)	65.93	-2163.8
41	745.8 (76.2)	S3: 2031.7 (189.1)	63.29	-2166.0
42	843.5 (79.3)	S4: 2126.7 (210.7)	60.34	-2023.4
43	813.8 (71.4)	S3: 2023.1 (197.8)	59.78	-2005.0
44	683.0 (41.3)	S6: 2297.3 (220.0)	70.27	-2016.2
45	556.8 (36.0)	S8: 1726.0 (185.4)	67.74	-2267.6
46	656.9 (53.6)	S6: 1655.4 (187.9)	60.31	-2475.9
47	141.3 (5.3)	S2: 53.3 (4.1)	-165.25	-2699.9
48	229.1 (15.9)	S10: 336.2 (65.0)	31.86	-2661.4
49	5860.2 (155.1)	S13: 6387.8 (221.4)	8.26	-3107.0
50	622.3 (9.4)	S13: 531.3 (18.4)	-17.13	-2250.7
51	371.1 (18.5)	S10: 507.2 (35.7)	26.83	-3216.1
52	419.9 (27.4)	S3: 293.4 (46.2)	-43.12	-1015.5

G.6 Numerical provenance and retained adverse evidence

The broad, high-repeat and long results above are reconstructed trajectories from the archived source, original seeds, horizons and cohort definitions, run on Windows with MinGW GCC 13.1 and `-O3`, without fast-math. The original run used Linux GCC 14.2. The reconstructed observations have new identities: they are neither recovered original raw bytes nor additional independent confirmation samples. Original damaged archives, retained prefixes, fully recovered members and difference reports are preserved. The long scalar challenger is replayed on the same Windows platform as both other long-joint methods, rather than mixing an original Linux scalar with reconstructed Windows DGB.

Across the five missing scientific batches and the two same-platform supporting batches, reconstruction contains 358,272 trajectories, 42,605,740,032 updates and 2,585,728 checkpoint rows, all passing recorded status, uniqueness, count-total and count-regret checks. The separate 172,032-row diagnostic trace contains 42 trajectories and is not counted as additional scientific repetitions. These are reproducibility checks, not semantic acceptance of a theorem.

The largest broad-panel original-versus-reconstructed summary-mean difference is about 10.2212. Of 806,218 complete checkpoint records that can be compared with surviving original prefixes, 1,620 differ in a count or state field. The high-repeat and long-anchor maximum summary-mean differences are about 0.05699 and 0.004132. Aggregate agreement therefore does not imply pathwise or bitwise agreement. The first divergence of every path has not been localized, so the platform change is a known difference, not a fully established causal explanation.

In the same-Windows precision diagnostic, the primary at $T = 65,536$ has mean 766.0697 in binary64 and 766.1162 in the extended implementation on 512 paired streams. The mean extended-minus-binary64 difference is 0.04648 (paired SE 0.04766), but only 40.04% of terminal action-count vectors agree exactly. A $\gamma = 0.45$ candidate has a difference about -0.1625 (paired SE 0.06405). The local long-double type has 64 significand bits, not arbitrary precision; parameter parsing also differs between implementations. These diagnostics establish neither numerical-process equivalence nor an exact-real-to-floating-point bridge.

The fully recovered noise and finite-grid records retain their original raw-data identities and are not pooled with reconstructed records. The noise experiment keeps the anchor means and initial logits, but returns either the average of 4 or 16 independent Bernoulli draws as one reward per action, or the deterministic reward $Y = \mu_i$. Deterministic rewards do not make actions deterministic. Each

cell has $n = 8192$ and $T = 65,536$. The comparator is S5, $b = .6, \eta = .4$, not the S2 long-joint comparator. The peak DGB candidate has $(h, \omega, \gamma) = (12, .5001, .5)$.

Table 8: All three recovered noise laws. Entries are terminal mean (SE). The inferential statements in the text use the recovered $M = 27$ simultaneous mean family, not paired-difference intervals.

Reward law	Fixed DGB	Peak DGB	S5
Average of 4	625.46 (5.45)	520.84 (7.36)	1040.15 (17.97)
Average of 16	592.66 (4.38)	473.28 (5.58)	535.60 (5.28)
Deterministic	583.76 (3.88)	463.37 (5.12)	430.26 (2.53)

The original family includes all three laws, three methods and three checkpoints $\{16,384, 32,768, 65,536\}$, hence $M = 27$; it uses the same $\alpha = .05$ split and interval construction specified above. With 16-draw average rewards, the peak saving lower bound is -7.1099, despite its lower sample mean. With deterministic rewards even the peak sample mean is worse than S5, while the fixed primary's saving interval is $[-207.0513, -95.0494]$. Changing a reward distribution does not by itself isolate a pure causal effect of variance.

The separate recovered finite-grid analysis uses 4,096 repeats at the Bernoulli anchor. Its 16 DGB candidates are the fourteen combinations $h \in \{9, 10, 11, 12, 13, 14, 16\}$, $\omega \in \{.5001, .51\}$, plus $(h, \omega) = (8, .6), (8, .51)$, all at $\gamma = .5$. The seventeenth simulated method, a scalar, is excluded from the grid certificate. It selects the sample-best $(12, .5001, .5)$, with mean 744.8412. Its original post-outcome reanalysis uses the 120 unordered-pair comparisons and 16 individual lower-mean bounds, allocating $\alpha = .05$ to 256 one-sided targets and giving an additive suboptimality upper bound 91.6583 and a ratio upper bound 1.14018 relative to the best expected cost *within that finite grid*. This family uses the fixed-mixture bound alone at $\delta = .05/256$, applying it to $(Z_a - Z_b + B_T)/(2B_T)$ for paired differences and to Z_a/B_T for individual lower limits. If U_{ab} bounds $R_a - R_b$, then any selected candidate $\widehat{a}$ has additive bound $\max\{0, \max_b U_{\widehat{a}b}\}$; dividing that bound by $\min_b L_b = 653.8700$ and adding one gives the reported ratio. Simultaneous coverage permits selecting the reported candidate after observing the fixed table, but it gives no continuous-parameter optimality claim. The historical five-group, $M = 270$ aggregation is retained only as an exploratory reanalysis: selection-valid coverage for the post-outcome grouping has not been established, and group averages cannot substitute for environment-wise guarantees. No likelihood-transport or continuous-region claim is added by the present discrete-panel integration.

The accompanying source-data directory preserves exact copies of the compact confidence tables, protocol dictionaries and cohort manifest. All displayed means and SEs come from the count-derived tables, all displayed inferential limits from their original-family recalculation, and every scalar alias maps to the unchanged method name in Table 2. No new simulation, tuning, dropped boundary configuration or replacement of an old unfavorable result is used here.

G.7 Interval-validated finite-horizon evidence

This separate cohort checks whether the anchor advantage persists when each sampled action is validated against the real-arithmetic rule, rather than sampled from rounded softmax masses. It uses the same Bernoulli means $\mu = (5/8, 1/2, 1/20, 1/20, 1/20)$, initial logits $(0, 3, 0, 0, 0)$ and resource bounds $u = 16$, $B_{\max} = 1$. The four locked methods are the primary DGB $(h, \omega, \gamma) = (8, .51, 1/2)$, its $(12, .5001, 1/2)$ variant, and the scalar rules S2 and S10 from Table 2. Each has 1,024 replicates, observed at $T =$ 16,384, 32,768 and 65,536. These replicates are not pooled with the earlier binary64 cohorts, and the primary DGB is not replaced by the lower-mean variant.

Validating the original updates and sampling decisions. The implementation represents $x_i = \exp(\theta_i/2) > 0$ and $Z = (\sum_i x_i^2)^{1/2}$, so that $p_i = x_i^2/Z^2$ and the instantaneous cap is $(\sum_i x_i)(\sum_i 1/x_i)$. For DGB at $\gamma = 1/2$, with loss $\ell = 1 - Y$, the unchanged logit update is equivalently

$$x_i' = x_i \exp\left(\frac{\eta \ell x_i}{2Z}\right) \quad (i \neq I), \qquad x_I' = x_I \exp\left(-\frac{\eta \ell \sum_{i \neq I} x_i^2}{2 x_I Z}\right).$$

For the scalar rules it is $x_i' = x_i \exp[\eta(Y - b)(\mathbf{1}\{i = I\} - p_i)/2]$. Directed MPFR rounding encloses all rational inputs, elementary functions, controls and running maxima. Induction over these operations encloses the exact state before each action. A finite random-bit prefix encloses the action uniform in an interval; an action is accepted only when that entire interval lies within one arm's certified cumulative-probability cell. Thus every completion of the prefix selects the same arm as exact softmax sampling. Ambiguity triggers more bits or higher-precision replay of the same random tape, not a replacement replicate. Bernoulli rewards use integer rejection sampling at the exact rational means, with separate pull-indexed arm streams. Induction over rounds then identifies the completed action counts with those of the real-arithmetic process on that tape. All 4,096 planned trajectories completed; earlier incomplete arithmetic attempts remain in the archive. This validation relies on the interval implementation, MPFR and runtime correctness; statistical coverage assumes independent replicate tapes, as distinct from the counter-based pseudorandom implementation.

Simultaneous bounds and retained comparisons. The count-derived cost is exactly $Z_T = (5N_2(T) + 23\sum_{i=3}^{5} N_i(T))/40$, with support $[0, B_T]$, $B_T = 23T/40$. The locked confidence family has 36 one-sided endpoints: a lower and upper mean regret limit and an upper mean outside-arm cost limit for each of four methods at three times. Here outside-arm cost means $(23/40)\sum_{i=3}^{5} N_i(T)$. The full family retains $\alpha = .025$, including endpoints not displayed in the terminal table. For normalized replicate costs $z_r \in [0, 1]$, it uses the fixed mixture

$$E_n(m) = \frac{1}{25}\sum_{j=0}^{24}\prod_{r=1}^{n}\big(1 - 2^{-j} + 2^{-j} z_r/m\big), \qquad 0 < m \le 1.$$

Under $\mathbb{E}z_r \le m$, independence across replicates gives $\mathbb{E}E_n(m) \le 1$. Markov's inequality at threshold $36/.025 = 1440$, inversion in m, and the same construction for $1 - z_r$ give lower and upper mean limits. A union bound yields simultaneous 97.5% coverage without independence across methods or times. All 36 rejection inequalities were checked with outward-rounded arithmetic and independently with exact integer products; descriptive SEs are not used for these limits.

Table 9: Separate interval-validated anchor cohort at $T = 65{,}536$, $n = 1024$ per method. Limits belong to the full 36-endpoint simultaneous 97.5% family. Means and SEs are descriptive; all four locked methods are retained.

Method	Mean regret	SE	Mean regret limits
DGB $(8, .51, 1/2)$	773.55	28.72	$[663.616, 1161.778]$
DGB $(12, .5001, 1/2)$	738.57	38.05	$[594.033, 1136.786]$
Scalar S2	2452.32	93.76	$[2046.812, 2946.222]$
Scalar S10	2909.49	91.72	$[2508.496, 3392.840]$

For primary DGB versus S2, subtracting the DGB upper limit from the scalar lower limit gives an expected saving of at least 885.034; the relative saving is at least $1 - 1161.778/2046.812 > 43.2\%$. The corresponding S10 bounds are 1346.718 and 53.6%. Both comparisons hold on the same event. This is finite-horizon evidence for the specified rules at the anchor, not validation of the earlier binary64 trajectories. The accompanying `source_data/validated_anchor/` retains the protocol, all three checkpoints, interrupted attempts, completed counts and the 36 endpoint checks.

H LEGACY $u = 1$ EXPERIMENTS: PROTOCOLS AND STATISTICAL UNITS

This section specifies the legacy $u = 1$ experiments. The separate $u = 16$ study uses the protocol in Appendix G, including a different numerical sampler. The mathematical rules use exact softmax probabilities; the experiments use the finite implementation below. No numerical diagnostic upgrades an exact-real theorem to an infinite-time floating-point guarantee.

H.1 REWARD STREAMS, PROBABILITIES AND CLOCKS

Each environment has independent pull-indexed Bernoulli streams with the stated means and a unique best arm. The learner sees only its chosen reward; true means enter reward generation and offline diagnostics, not its controls. Within a cohort, paired methods share environment/replicate/arm

reward keys and action uniforms. Biased and uniform initializations have different environment keys and are not paired with each other. A stream advances only when its arm is pulled. Action, posterior and reward generators are separate. Rewards use $U = (\text{top 53 random bits})/2^{53}$ and $Y = \mathbf{1}\{U < \mu_i\}$. The induced success probability differs from the stored threshold by less than 2^{-53} unless it is already on the grid.

For a logit-based method, subtract-maximum binary64 softmax gives $p_{i,t}^{\mathrm{eval}}$. The categorical sampler then uses

$$w_{i,t} = \lfloor 2^{52} p_{i,t}^{\mathrm{eval}} \rfloor \ (i < K), \qquad w_{K,t} = 2^{52} - \sum_{i<K} w_{i,t}, \qquad \pi_{i,t} = w_{i,t}/2^{52}. \tag{157}$$

It requires positive integer masses and $\max_i |\pi_{i,t}/p_{i,t}^{\mathrm{eval}} - 1| \le 10^{-10}$. Support loss, nonfinite states or controls, or a failed distortion guard stop that trajectory as a numerical failure: no clipping, probability floor, redraw or logit reset is applied. Sampling and score centering use the same π_t. At action round t, exactly $t-1$ pulls have been observed. The original-score transition is

$$\theta_{t+1} = \theta_t + \eta_t(Y_t - b_t)(e_{I_t} - \pi_t), \tag{158}$$

with controls determined before that action. A zero scalar stops the update, not natural sampling. Checkpoint T is recorded after the Tth update: counts, regret and θ_{T+1} are post-update; fields ending in `_pre` refer to the state used for update T.

H.2 Implemented rules

DGB evaluates equations (53)–(54) on π_t. It accumulates q_t from all coordinates except the first maximizer, avoiding cancellation in $1 - \max_i \pi_{i,t}$. The focused and confirmation runners use ordinary-domain cap and update products with explicit nonfinite/underflow checks. The older reference kernel stores $\log \overline{C}_t$ and evaluates magnitudes in the log domain; these are not claimed to be bitwise-identical evaluation paths. The fixed-h confirmation, scale extension and small-gap diagnostic use $\gamma = 1/2$, $\omega = 0.6$, $u = 1$, with their separately specified h grids rather than imposing $h = \gamma/64$.

The unscaled scalar ablation uses the DGB scalar formula at its own policy and cap with baseline one but omits the diagonal score weighting. The norm-matched ablation is defined below. The horizon-aware scalar upper rule uses a distinct trajectory for each terminal horizon and sets

$$\eta_t = \min\{1, 1/(8eK)\} T^{-1/K}, \qquad b_t = 0. \tag{159}$$

It does not reuse a maximum-horizon trajectory as shorter tuned runs. The legacy UCB1/Thompson batch and its independent sampling protocol are specified in Appendix I.2.

H.3 Cohorts and statistical units

Means and sample SE use independent replicates. Paired differences use matched replicate IDs within a cohort and preserve cross-method covariance; SE is not a confidence interval. Shared uniforms and pull-indexed rewards couple comparisons, not realized actions. No independence is inferred between successive checkpoints or between methods on the same replicate. Failures retain prefixes and unobserved suffix bounds (Appendix M); missing outcomes are never replaced by zero. Local design locks are not public preregistrations. Cohorts are not pooled into a larger effective sample.

H.4 Scale-matched controls and focused design

The additional controls match projected score scale in an end-to-end comparison; they do not isolate direction from reward-weighted noise. All controls are computed before the action from the corresponding algorithm's own state and probability-history cap. Sharing the cap formula does not mean sharing its realized value, scalar sequence or trajectory. These diagnostics do not modify the theorem algorithms or prove worst-case minimax rates.

Table 10: Distinct legacy $u = 1$ data cohorts, separate from the current $u = 16$ study in Appendix G. A stream/replicate, not a checkpoint, is the statistical unit. The pilot is not pooled with confirmation. Counts refer to the declared design; failed original paths retain their observed prefixes.

Cohort	Repeats and cells	DGB design scale	Evidence role
Original campaign	256 streams/config.; 606 configs	$h = \min\{\gamma/64, u\}$	Task and allocation diagnostics; five failures
Focused comparison	64 streams/config.; 24 configs	$h = 1/128$	Eight environments, three methods; paired controls
Two-arm check	16 pairs/setting; eight settings	$h = \min\{\gamma/64, u\}$	Coupled projection check; separate sample
Fixed-h confirmation	32 pairs/config.; 64 method cells	$1/128, 0.03, 0.06, 0.1$	Eight environments, DGB and Scalar-NM; new namespace
Scale extension	32 pairs/config.; 48 method cells	$0.1, 0.4, 1$	Eight environments; direct clocks; separate namespace
External pilot	Four streams	Motivated the h grid	Exploratory only; no pooled estimate
Neighborhood diagnostic	32 streams/cell; 48 cells	$h = 0.1$	16 small-gap environments; DGB, Scalar-NM, Upper
Task references	32 streams/cell; eight cells	Not applicable	Independent UCB1/Thompson; shared across starts

A legal scale-matched scalar. For an interior probability vector p, set $g_i = e_i - p$, $P = I - \mathbf{1}\mathbf{1}^\top/K$, and $D = D_\gamma(p)$, and define

$$s(p)^2 = \frac{\sum_i p_i \|PDg_i\|_2^2}{\sum_i p_i \|g_i\|_2^2}. \tag{160}$$

The denominator is $1 - \sum_i p_i^2 > 0$: every coordinate is positive and $K \geq 2$. The numerator is positive because each Dg_i has a positive coordinate at i and negative coordinates elsewhere and thus cannot be a constant vector annihilated by P. Since P is an orthogonal projection, $\|PDg_i\| \leq \|D\|_{\mathrm{op}} \|g_i\| = p_{\min}^{-\gamma} \|g_i\|$. Consequently $0 < s(p) \leq p_{\min}^{-\gamma}$. Choose $\widetilde{\eta}_t = \eta_t^{\mathrm{DGB}} s(p_t)$, where the DGB scalar formula is evaluated using this control's own probabilities and cap. The pointwise small-step estimate (55) implies $\widetilde{\eta}_t \leq h \leq u$, independently of reward means. The control is therefore a predictable bounded common scalar for the *original* score, with baseline one.

For reproducible numerical evaluation, writing $d_i = p_i^{-\gamma}$ gives

$$\sum_i p_i \|PDg_i\|^2 = \sum_i p_i d_i^2 (1 - p_i) - \frac{\sum_i p_i d_i^2 - (\sum_i p_i d_i)^2}{K}.$$

This follows by expanding the squared norm and subtracting $K^{-1}\mathbb{E}(\mathbf{1}^\top Dg_I)^2$; $\mathbb{E}g_I = 0$. The implementation uses sums of other probabilities for $1 - p_i$. Matching is over action scores at a given state, not over reward-weighted update variances. Subsequent states differ, so this is an end-to-end control, not a causal experiment holding the entire trajectory or scalar sequence fixed.

Two-arm positive control. When $K = 2$, Proposition C.3 gives $PDg_i = \lambda(p) g_i$, with $\lambda(p) = (p_1^{-\gamma} + p_2^{-\gamma})/2 > 0$. Thus $s(p) = \lambda(p)$, and using the projected scalar yields the same policy as DGB under coupled actions and rewards in exact arithmetic. The common logit shift can differ. This identity is checked numerically over every round, not inferred from a coincident terminal regret. It does not imply equivalence of the multi-arm algorithms.

Shared eight-environment family. The legacy focused, fixed-h confirmation and scale-extension cohorts cross $K \in \{3, 5\}$, $\delta \in \{1/4, 1/16\}$, and uniform zero or biased $(0, 1, 0, \ldots, 0)$ logits, with Bernoulli means $\mu_1 = 1/2 + \delta$, $\mu_2 = 1/2$, and $\mu_i = 1/4 - (i - 3)/(16K)$ for $i \geq 3$.

Locked 64-stream design. The methods are DGB, the original unrescaled scalar-only control, and (160), with $(\gamma, \omega, u) = (1/2, 0.6, 1)$ and $h = 1/128$. Each of the 24 configurations has 64 replicate streams and a fixed horizon 2^{18}, with dyadic checkpoints from 2^{10}. The separate two-arm check uses $\gamma \in \{1/4, 1/2\}$, both gaps and initializations, and 16 paired paths per setting through 8192 steps. These observations are not pooled with the original 256-stream experiment. Configuration, seeds, metrics, and horizons were fixed before the new outcomes were observed; this was a local design lock, not a public preregistration.

The shared law and failure policy are those of Appendix H.1; controls never use unobserved rewards or true means. Means enter only offline regret and mean-direction diagnostics.

Mechanism measurements and limits. At checkpoints we retain per-arm probabilities and counts, optimal-versus- competitor logit contrasts, outside-group occupation, scalar magnitude, cap divided by $\sqrt{t}$, and projected score norms. First passages of $p_{\mathrm{best}} \geq 1/2$ or 0.9 include time zero and retain right censoring. For vectors $v = PDg_i$ and $g = g_i$, direction residual is $\|v - \alpha g\|/\|v\|$, $\alpha = \langle v, g\rangle/\|g\|^2$; the analogous mean-direction diagnostic replaces g by $(\operatorname{diag} p - pp^\top)\mu$. A zero mean vector is recorded with zero norm and zero residual. These are local geometry diagnostics, not new regret guarantees or proof that geometry alone causes an observed difference. All configurations are retained, using the statistical conventions of Appendix H.3.

The original 256-stream campaign also contains fixed-scalar, mirror and data-dependent controls not used to establish the current main comparisons. Its retained plot records are in `data/LEGACY_POINTS.json` and the `figures/integrated/Legacy_⋆` assets of the source package. Appendix M retains all five recorded failures and all four censored endpoint rows; these historical records are not pooled with the later cohorts.

I LEGACY $u = 1$ CONFIRMATION, ALLOCATION COSTS AND CONTROLS

Design and numerical semantics. This legacy confirmation fixes all eight environments of Appendix H.4, DGB and the own-state norm-matched scalar, four fixed values $h \in \{1/128, 0.03, 0.06, 0.1\}$, 32 paired replicates, $T = 262144$, and nine dyadic checkpoints starting at 1024. The 2048 paths contain 536870912 completed updates and 18432 checkpoints. An external four-stream exploratory pilot motivated the grid; this confirmation uses a disjoint random namespace. The design was locked before confirmation outcomes, not publicly preregistered. All h values and environments are retained. Within an environment/replicate, h and method variants share action uniforms and pull-indexed reward streams, not necessarily realized actions.

The shared simulator semantics are specified in Appendix H.1; explicit h changes only the fixed numerator, not the sampling law, score or feedback interface. No path records a numerical failure. Aggregate runtime was approximately 137 seconds with four workers on the recorded host; the artifact retains runtime/platform details. Deterministic tests include the original kernel at its original h, linear h-scaling at a fixed state, resource bounds, enumerated conditional drift, two-arm coupled equivalence and injected-failure accounting. These tests are not mathematical or asymptotic acceptance.

Numerical accounting. Post-update states, pre-action drift and accumulated residuals use the distinct measurements defined in Appendix K.1. The regret decomposition uses all arm counts. Repeated binary64 addition differs from count-based evaluation by at most 1.17×10^{-7} in these records; the analyzer checks the nonnegative-summation roundoff bound at every checkpoint. No numerical residual is presented as a theoretical failure or silently rounded into an exact machine identity.

Uncertainty and interpretation. Means and sample SE use 32 independent replicate pairs. The relative-regret panel divides both the paired mean difference and its SE by observed mean NM regret; the latter is a descriptive scaled SE, not a confidence interval for a ratio of expectations. Total-difference SE is computed from pairwise totals, retaining component covariance. Counts of first hits retain right censoring and include time zero. Neither hit fractions nor mean curves establish an asymptotic polylogarithmic regime. The larger h values improve finite-scale motion in these settings but do not imply monotone performance in h in general or superiority over the full scalar class.

I.1 THE LEGACY $h = 0.1$ ENDPOINT: PERFORMANCE AND ALLOCATION COSTS

This subsection concerns only the $u = 1$, $h = 0.1$, 32-pair confirmation endpoint, not the current $u = 16$ study. Its eight environments and own-state Scalar-NM control are defined in Appendix H.4; its horizon is $T = 2^{18}$ and $(\gamma, \omega) = (1/2, 0.6)$. Scalar-NM matches projected score scale, not reward-weighted variance or future states. UCB1 and Thompson sampling provide independent task-level references (Table 15); none of these comparisons empirically optimizes over the entire scalar class.

Absolute and relative performance. DGB has lower mean regret than Scalar-NM in three of eight environments; relative differences range from -9.88% to $+46.98\%$. Absolute means are 12,297–37,767 for DGB and 12,131–31,562 for Scalar-NM (rounded; means and SEs: Table 15). DGB regret is 33–161 times UCB1's (Auer et al., 2002) and 206–932 times Thompson sampling's (Agrawal & Goyal, 2012): sample-mean ratios from independent reference streams, not paired effects.

One allocation identity for all legacy comparisons. For each paired run, write $dN_i = N_i^{\mathrm{DGB}} - N_i^{\text{Scalar-NM}}$. Since both methods make T pulls, $\sum_i dN_i = 0$ gives

$$d\widehat{R}_T = \underbrace{-\delta dN_1}_{B_T\text{: best-arm term}} + \underbrace{\sum_{i\geq 3}(1/2 - \mu_i)dN_i}_{O_T\text{: outside-arm term}}. \tag{161}$$

Negative terms save DGB regret relative to Scalar-NM; positive terms add cost. The identity is exact in count arithmetic; the sequential binary64 regret check is specified above. Total SE uses paired totals, retaining component covariance. This is allocation accounting, not a causal intervention that fixes a trajectory or isolates direction from cap effects. Figure 3 shows the contributions and their sum over time; Figure 4 places the endpoint in the complete four-scale grid.

For biased $K = 5$, gap $1/16$ (panel g), best-arm allocation saves 1278.84 ± 10.07 regret units against 928.64 ± 29.70 extra outside cost: a net saving of 350.21 ± 22.08 (sample SE). Components are rounded separately. For uniform $K = 3$, gap $1/16$ (panel d), lower outside cost instead offsets a worse best-arm term.

Both biased gap-$1/4$ cells reverse early DGB gains into terminal losses, despite terminal mean best-arm probability 0.796 for $K = 5$. Uniform $K = 5$, gap $1/4$ has the largest loss (Appendix I.2.1). The alternative net-only, common-axis view remains in the source package as `figures/relations/Figure_09.pdf`; it uses the same 72 paired checkpoint summaries as Figure 3, not a separate cohort. At frozen terminal DGB states, projected update covariance traces are 1.127–1.279 times the hypothetical norm-matched scalar's (means of per-stream ratios; Table 24). These finite-scale comparisons do not estimate the scalar minimax exponent.

Intervals for the paired differences. Table 14 augments the existing sample SE with approximate per-cell intervals at the prespecified grid endpoint $h = 0.1$. For each environment we first subtract NM from DGB within each of the 32 stream pairs, then use $\bar{d} \pm 2.0395134464\, s_d/\sqrt{32}$. Student-$t$ coverage is exact for independent normal differences; without that distributional assumption it is a normal-approximation diagnostic, not a distribution-free finite-sample guarantee. These intervals do not simultaneously cover the 32 environment/scale comparisons. The artifact retains all 32 intervals and paired values, without selecting by sign. The smallest absolute mean/SE ratio in that full grid is 7.106; this describes the observed signal relative to its SE, not a tail guarantee. We do not construct a percentage interval by treating the observed NM denominator as fixed. No checkpoint is counted as an extra replicate.

The dimension pattern depends on the cohort. At $h = 1/128$, both the earlier 64-stream comparison and the separate 32-pair confirmation have three favorable $K = 3$ means and one favorable $K = 5$ mean out of four each. At $h = 0.1$, these counts are two and one; over all four fixed scales they are ten of sixteen and three of sixteen. The separate neighborhood diagnostic instead has eight of eight favorable $K = 3$ means and none at $K = 5$. Thus its all-adverse five-arm pattern must not be extrapolated to every tested five-arm environment. The best-arm and weighted outside-count terms provide exact accounting of each reversal; their scaling with dimension and the causal effect of changing a trajectory are not identified by these comparisons.

Table 11: Fresh-seed fixed-h confirmation for $K = 3$. All entries are means $\pm$ sample SE over 32 pairs; hits count paths reaching $p_{\text{best}} \geq 0.5/0.9$ by $T = 2^{18}$. B/U denotes biased/uniform initialization.

Gap/init	h	DGB regret	NM regret	DGB−NM	Hits DGB / NM
1/4/B	0.007812	61837.55 ± 18.27	62401.59 ± 17.95	-564.05 ± 2.71	0/0 : 0/0
1/4/B	0.03	50767.70 ± 43.89	52161.76 ± 46.87	-1394.05 ± 11.02	0/0 : 0/0
1/4/B	0.06	35624.77 ± 85.28	35424.11 ± 109.11	200.66 ± 28.24	32/0 : 32/0
1/4/B	0.1	21649.03 ± 94.03	19859.26 ± 111.63	1789.77 ± 21.17	32/0 : 32/32
1/4/U	0.007812	58254.52 ± 24.54	58221.02 ± 24.64	33.49 ± 1.06	0/0 : 0/0
1/4/U	0.03	40158.88 ± 48.58	39144.38 ± 53.13	1014.51 ± 7.14	32/0 : 32/0
1/4/U	0.06	24592.76 ± 59.53	22272.26 ± 64.74	2320.50 ± 7.18	32/0 : 32/0
1/4/U	0.1	14525.99 ± 52.96	12131.12 ± 54.62	2394.88 ± 6.68	32/32 : 32/32
1/16/B	0.007812	25102.59 ± 9.54	25285.16 ± 9.42	-182.57 ± 1.33	0/0 : 0/0
1/16/B	0.03	21413.65 ± 17.59	22117.77 ± 15.26	-704.12 ± 3.73	0/0 : 0/0
1/16/B	0.06	18105.35 ± 26.38	19361.38 ± 21.88	-1256.03 ± 5.70	0/0 : 0/0
1/16/B	0.1	15166.38 ± 40.42	16829.46 ± 39.26	-1663.07 ± 10.10	0/0 : 0/0
1/16/U	0.007812	29730.14 ± 15.14	29770.23 ± 14.89	-40.09 ± 0.70	0/0 : 0/0
1/16/U	0.03	22882.62 ± 36.63	23222.79 ± 34.46	-340.17 ± 4.74	0/0 : 0/0
1/16/U	0.06	16921.34 ± 51.28	17453.86 ± 52.79	-532.52 ± 11.53	32/0 : 32/0
1/16/U	0.1	12297.22 ± 58.39	12654.14 ± 72.74	-356.92 ± 22.32	32/0 : 32/0

Table 12: Fresh-seed fixed-h confirmation for $K = 5$. All entries are means $\pm$ sample SE over 32 pairs; hits count paths reaching $p_{\text{best}} \geq 0.5/0.9$ by $T = 2^{18}$. B/U denotes biased/uniform initialization.

Gap/init	h	DGB regret	NM regret	DGB−NM	Hits DGB / NM
1/4/B	0.007812	82801.81 ± 17.69	82998.80 ± 17.55	-196.99 ± 1.87	0/0 : 0/0
1/4/B	0.03	71900.74 ± 30.61	72793.85 ± 30.33	-893.12 ± 3.64	0/0 : 0/0
1/4/B	0.06	56198.42 ± 67.97	54939.68 ± 103.37	1258.74 ± 44.17	32/0 : 32/0
1/4/B	0.1	37767.22 ± 95.48	31562.02 ± 132.62	6205.20 ± 47.21	32/0 : 32/32
1/4/U	0.007812	88316.08 ± 18.69	88174.62 ± 18.62	141.46 ± 1.24	0/0 : 0/0
1/4/U	0.03	70285.45 ± 45.02	67060.89 ± 60.86	3224.56 ± 18.97	0/0 : 32/0
1/4/U	0.06	47345.55 ± 77.49	38227.37 ± 100.45	9118.18 ± 26.86	32/0 : 32/0
1/4/U	0.1	29042.89 ± 79.28	19759.40 ± 79.86	9283.49 ± 13.51	32/0 : 32/32
1/16/B	0.007812	42371.95 ± 10.18	42183.28 ± 10.40	188.67 ± 1.78	0/0 : 0/0
1/16/B	0.03	36511.21 ± 26.04	35976.28 ± 29.09	534.94 ± 7.13	0/0 : 0/0
1/16/B	0.06	30412.74 ± 40.82	30011.15 ± 41.83	401.60 ± 16.09	0/0 : 0/0
1/16/B	0.1	24882.83 ± 48.80	25233.04 ± 41.37	-350.21 ± 22.08	0/0 : 0/0
1/16/U	0.007812	51803.60 ± 14.56	51780.70 ± 14.60	22.90 ± 0.41	0/0 : 0/0
1/16/U	0.03	44133.13 ± 30.47	43805.79 ± 31.88	327.34 ± 3.33	0/0 : 0/0
1/16/U	0.06	34605.29 ± 51.38	33462.04 ± 62.08	1143.25 ± 17.89	0/0 : 32/0
1/16/U	0.1	25257.18 ± 67.51	22873.72 ± 96.10	2383.46 ± 41.85	32/0 : 32/0

Table 13: Terminal states and drift/cost decomposition at the prespecified endpoint $h = 0.1$. Contrast is post-update; drift is cumulative pre-action conditional mean. All differences are DGB minus NM. Each entry is mean $\pm$ sample SE over 32 pairs.

K/gap/init	dX	$d\sum_t \mathbb{E}_t \Delta X_t$	Best-arm term	Outside cost
3/1/4/B	0.56 ± 0.0029	0.56 ± 0.0015	2336.20 ± 15.61	-546.42 ± 7.41
3/1/4/U	0.62 ± 0.0028	0.62 ± 0.0012	2246.66 ± 5.40	148.22 ± 5.63
3/1/16/B	-0.14 ± 0.0061	-0.14 ± 0.0042	-581.01 ± 12.49	-1082.06 ± 7.89
3/1/16/U	0.31 ± 0.0049	0.31 ± 0.0038	427.51 ± 9.88	-784.43 ± 13.34
5/1/4/B	1.13 ± 0.0047	1.14 ± 0.0030	5383.83 ± 36.35	821.37 ± 11.69
5/1/4/U	1.36 ± 0.0052	1.36 ± 0.0018	6290.87 ± 9.83	2992.62 ± 10.04
5/1/16/B	-0.72 ± 0.0109	-0.73 ± 0.0094	-1278.84 ± 10.07	928.64 ± 29.70
5/1/16/U	0.74 ± 0.0087	0.73 ± 0.0076	1212.45 ± 15.16	1171.01 ± 27.20

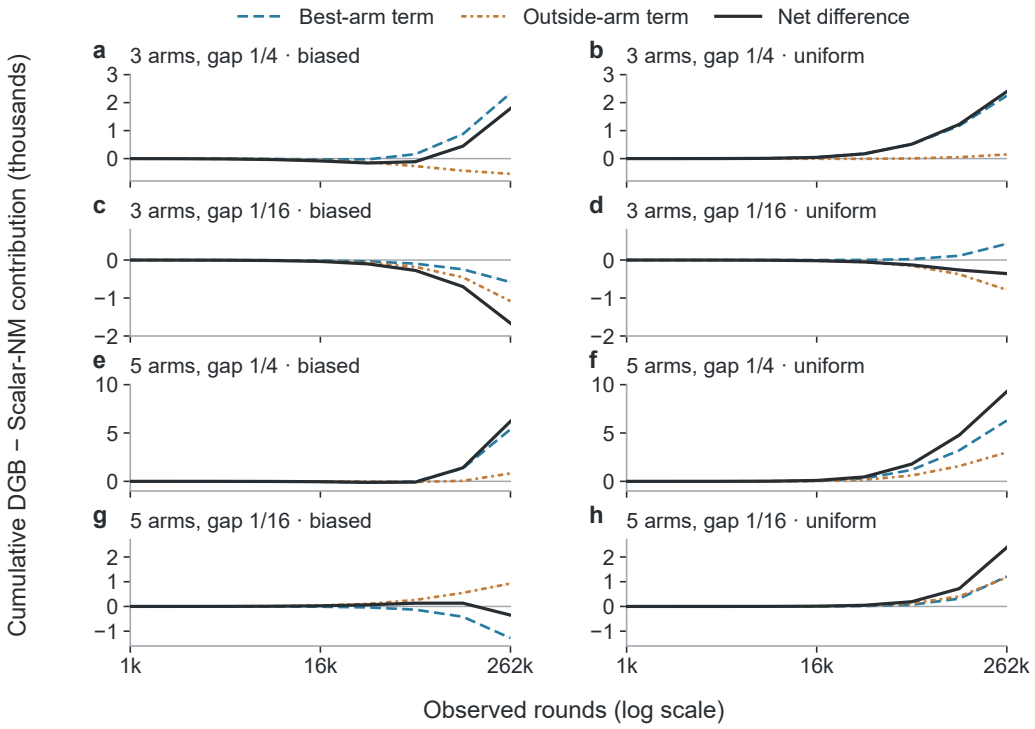


Figure 3: Allocation contributions in the legacy $u = 1$, $h = 0.1$ confirmation. Blue and orange sum to black; negative values save DGB regret relative to Scalar-NM, and positive values add cost. All eight configurations are retained. Lines join nine recorded checkpoints of 32 paired streams; bands are pointwise approximate 95% t_{31} intervals, not simultaneous bands. Net SEs use paired totals. Vertical scales are shared within rows, not across rows; some bands are narrower than lines. Segments do not measure crossing times, and a sign describes a sample mean, not significance. Table 15 reports absolute performance and standard baselines; Table 14 gives numerical endpoint intervals.

Table 14: Absolute paired regret differences at $h = 0.1$: DGB minus NM, $T = 2^{18}$, 32 independent stream pairs per cell. Approximate per-cell 95% Student-t intervals use $\bar{d} \pm t_{31,0.975}\mathrm{SE}(d)$; they are not simultaneous intervals or distribution-free finite-sample guarantees. No interval is asserted for the percentage heatmap.

K	δ	Initialization	Mean $\pm$ sample SE	Approximate 95% interval
3	1/4	Biased	1789.77 ± 21.17	$[1746.60,\ 1832.94]$
3	1/4	Uniform	2394.88 ± 6.68	$[2381.25,\ 2408.50]$
3	1/16	Biased	-1663.07 ± 10.10	$[-1683.66,\ -1642.48]$
3	1/16	Uniform	-356.92 ± 22.32	$[-402.44,\ -311.39]$
5	1/4	Biased	6205.20 ± 47.21	$[6108.90,\ 6301.49]$
5	1/4	Uniform	9283.49 ± 13.51	$[9255.92,\ 9311.05]$
5	1/16	Biased	-350.21 ± 22.08	$[-395.25,\ -305.17]$
5	1/16	Uniform	2383.46 ± 41.85	$[2298.11,\ 2468.81]$

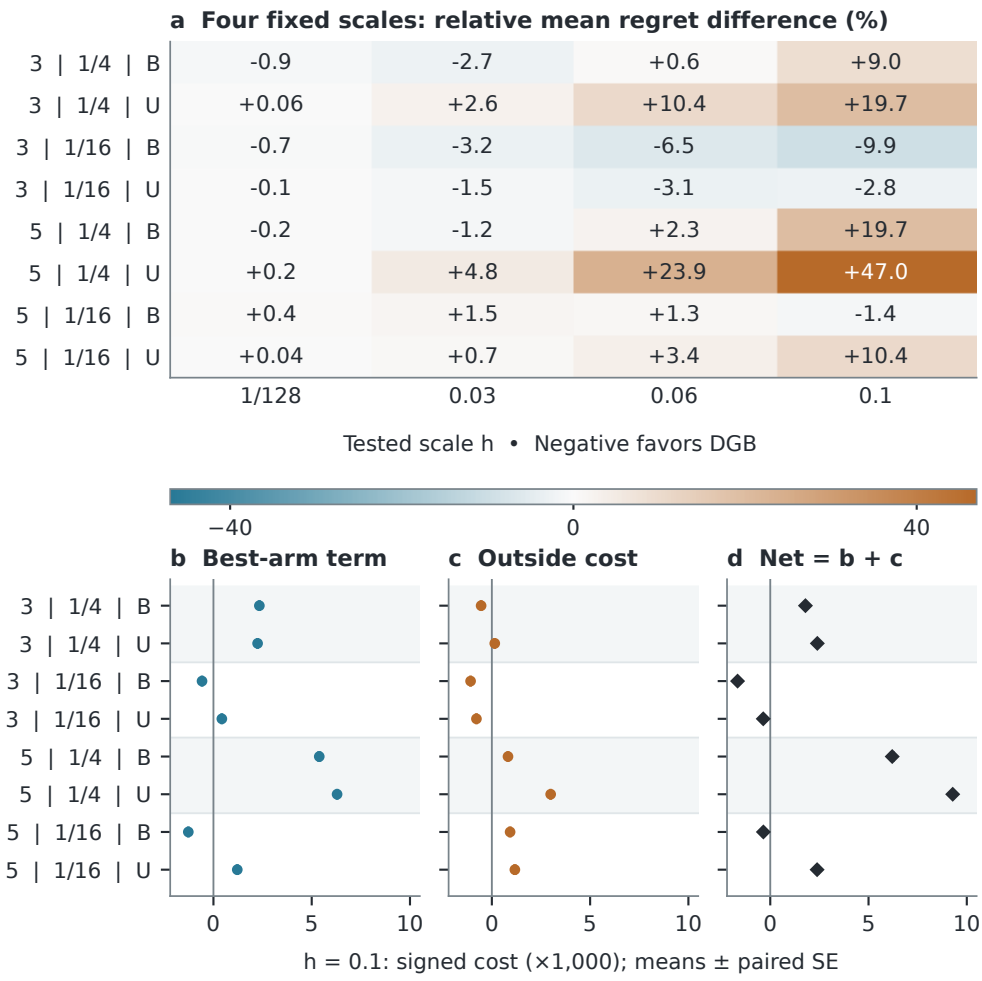


Figure 4: Scale and allocation accounting in the legacy $u = 1$, 32-pair confirmation cohort. (a) All four tested scales: 100 times paired DGB-minus-Scalar-NM mean regret divided by mean Scalar-NM regret. The color scale is symmetric and linear; the small positive entries retain their sign. These descriptive ratios are not confidence intervals. (b–d) At $h = 0.1$, signed best-arm and outside-cost terms sum to net regret difference. The common axis is in thousands of regret units; bars are one paired sample SE. Net SE uses the recorded paired total, retaining covariance. B/U denote biased/uniform starts, and the environments match Figure 3, not the current $u = 16$ experiment.

Table 15: Matched task-level comparison at $T = 2^{18}$: mean cumulative pseudo-regret ± sample SE over 32 streams. DGB/NM retain the legacy $u = 1$, $h = 0.1$ records of Figure 3; B/U denote biased/uniform logits. UCB1/Thompson use independent streams; each B/U pair shares the single reference estimate printed in its B row. Lower is better. Appendix I.2 specifies the locked protocol.

K, gap, init.	DGB	NM	UCB1	Thompson
3, 1/4, B	21649.0 ± 94.0	19859.3 ± 111.6	137.5 ± 3.4	23.5 ± 1.3
3, 1/4, U	14526.0 ± 53.0	12131.1 ± 54.6	shared with B above	
3, 1/16, B	15166.4 ± 40.4	16829.5 ± 39.3	369.4 ± 9.1	59.8 ± 4.5
3, 1/16, U	12297.2 ± 58.4	12654.1 ± 72.7	shared with B above	
5, 1/4, B	37767.2 ± 95.5	31562.0 ± 132.6	234.0 ± 2.5	40.5 ± 1.4
5, 1/4, U	29042.9 ± 79.3	19759.4 ± 79.9	shared with B above	
5, 1/16, B	24882.8 ± 48.8	25233.0 ± 41.4	528.2 ± 10.1	77.9 ± 4.8
5, 1/16, U	25257.2 ± 67.5	22873.7 ± 96.1	shared with B above	

I.2 Legacy learning states and matched standard references

Figures 5 and 6 use the legacy fixed-h cohort at $u = 1$, $h = 0.1$: eight environments, two methods and 32 streams. They report means and sample SEs at nine same-path checkpoints from 2^{10} to 2^{18}, not additional independent replicates. The full four-scale grid and cost decomposition appear above.

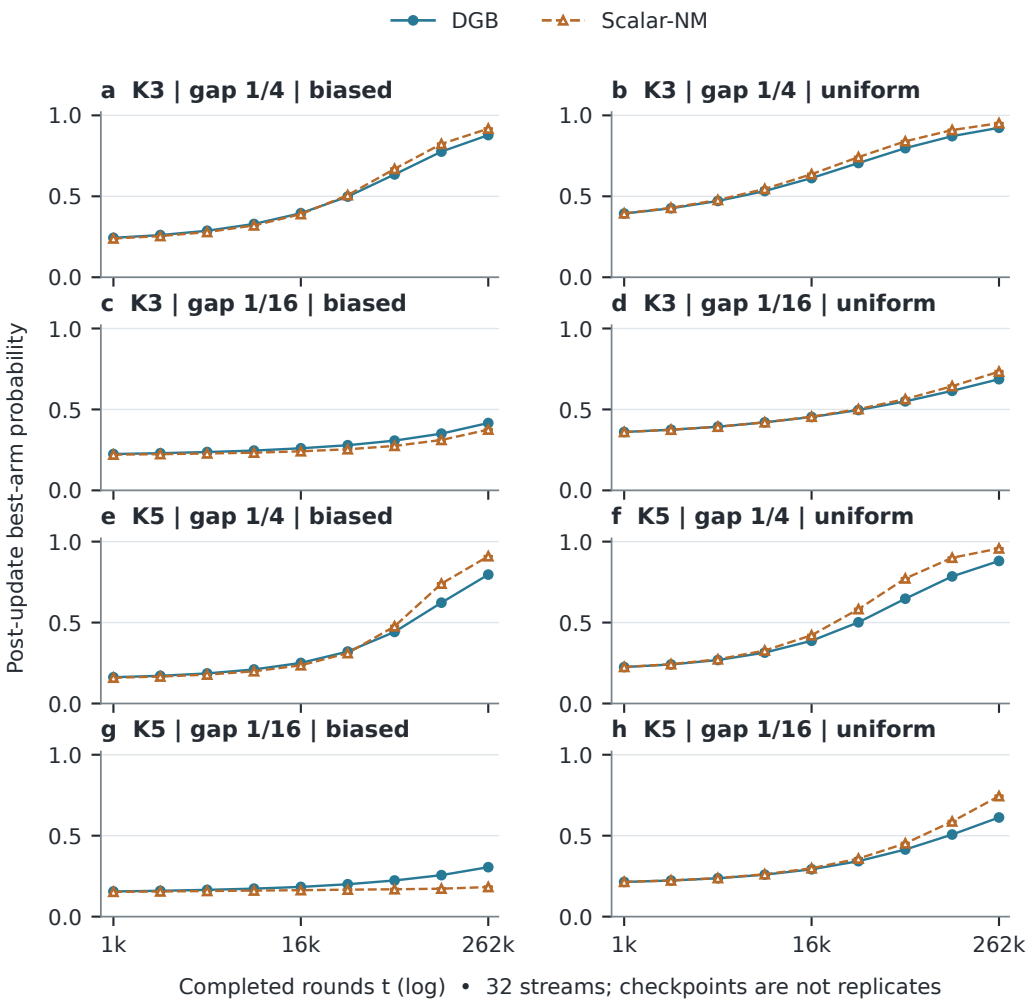


Figure 5: Post-update best-arm probability in all eight legacy confirmation environments. DGB/Scalar-NM use circles/triangles and solid/dashed lines; $h = 0.1$, $\gamma = 1/2$, $\omega = 0.6$, $u = 1$. Each of nine checkpoints reports a mean and one sample SE across 32 same-path streams. Time is logarithmic and probability is linear, with common axes. A mean crossing is not a path's first hitting time; checkpoints do not increase the sample count.

I.2.1 Learning states in the legacy 32-stream confirmation

Table 16 separates terminal mean post-update $p_{j,T+1}$, paths reaching probability $1/2$ by $T = 2^{18}$ (including time zero), and the last-window slope $s = \log_2(\bar{R}_{2^{18}}/\bar{R}_{2^{17}})$ of this cohort's mean regret. The slopes span 0.272–0.857 across 16 method/environment cells; they are neither asymptotic exponents nor confidence intervals. A threshold hit need not persist: the six environments where all paths hit still have substantial regret. No path hits in the two biased, gap-$1/16$ environments. These 32-stream results are separate from the 64-stream conservative-step censored hitting times in the reliability section.

Table 16: Legacy confirmation learning states: $u = 1$, $h = 0.1$, $T = 2^{18}$, 32 paths per method and environment. B/U mean biased/uniform logits. Probabilities are terminal means, hits are cumulative path counts, and slopes describe the last doubling window. Settings match Figure 3.

K, gap	Init.	$\bar{p}_j^{\text{DGB}}$	$\bar{p}_j^{\text{Scalar-NM}}$	Hits DGB	Hits NM	s_{DGB}	$s_{\text{Scalar-NM}}$
3, 1/4	B	0.878	0.918	32/32	32/32	0.484	0.402
3, 1/4	U	0.923	0.952	32/32	32/32	0.469	0.388
3, 1/16	B	0.416	0.375	0/32	0/32	0.820	0.857
3, 1/16	U	0.687	0.733	32/32	32/32	0.674	0.667
5, 1/4	B	0.796	0.910	32/32	32/32	0.548	0.370
5, 1/4	U	0.880	0.958	32/32	32/32	0.458	0.272
5, 1/16	B	0.305	0.183	0/32	0/32	0.779	0.813
5, 1/16	U	0.612	0.744	32/32	32/32	0.649	0.572

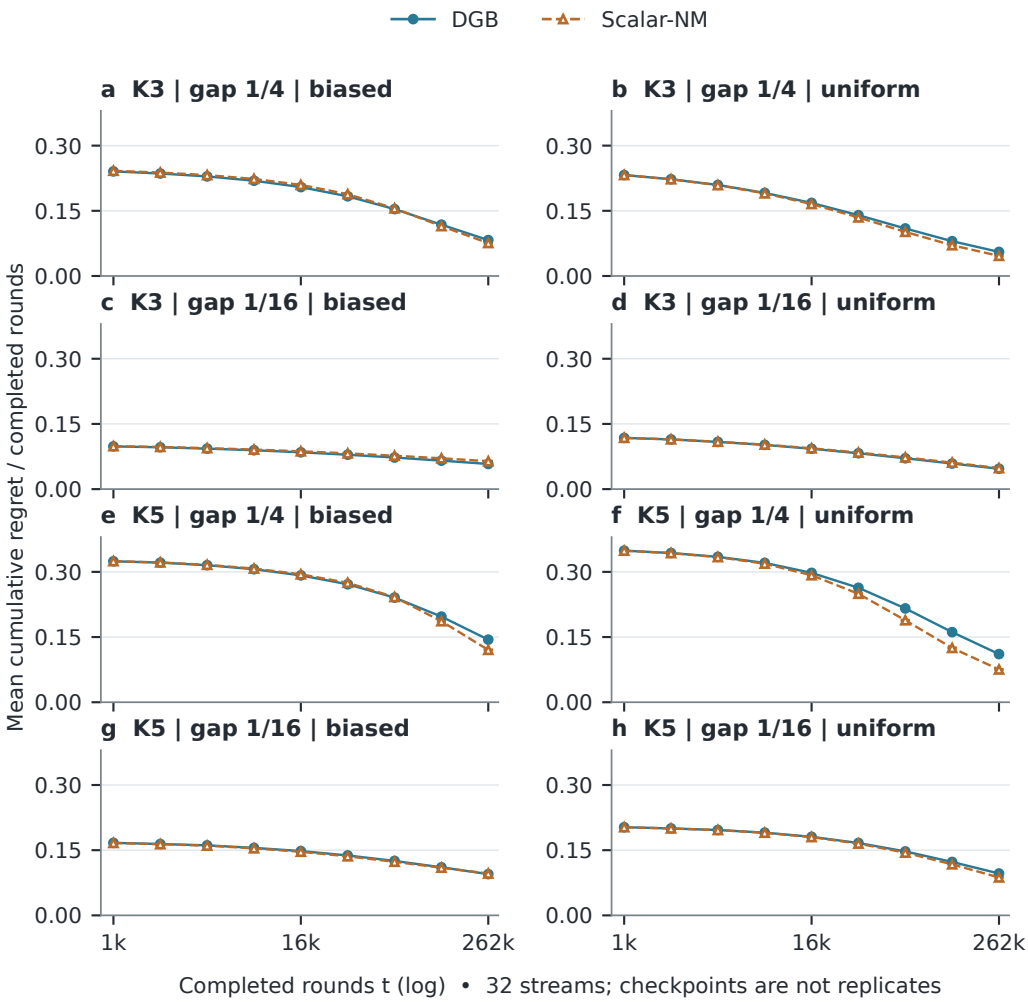


Figure 6: Average cumulative pseudo-regret per completed round in all eight legacy confirmation environments at $u = 1$, $h = 0.1$. Each point is the existing mean $\bar{R}_t/t$, with sample SE divided by the same deterministic t; 32 streams and nine same-path checkpoints are unchanged. Time is logarithmic, the common vertical scale is linear, and symbols match Figure 5. This is not instantaneous regret or a fitted asymptotic rate. The original cumulative log–log view is retained as `Companion_cumulative.pdf` in the source package's figure gallery.

I.2.2 MATCHED STANDARD-BASELINE PROTOCOL

The legacy reference batch uses the Bernoulli family of Appendix H.4, $T = 2^{18}$ and 32 replicates per method and mean vector. UCB1 pulls each arm once in index order, then maximizes $\widehat{\mu}_i + \sqrt{2\log(t)/N_i}$ with index tie-breaking. Thompson sampling has independent $\mathrm{Beta}(1,1)$ priors and updates only the selected arm's posterior. Neither receives logits, gaps or an optimal-arm label. Streams are independent of DGB/NM, so cross-family comparisons are unpaired. Biased/uniform rows share a reference estimate because neither reference method uses initialization.

The runner verifies the protocol/source/runner hashes fixed before execution. All 256 paths, 67,108,864 action/reward steps and 2,304 checkpoints are retained. Every checkpoint checks total pulls, $0 \leq$ successes $\leq$ counts, and regret against gap-weighted counts. SE is sample standard deviation divided by $\sqrt{32}$. Historical DGB/NM data, pairing and parameters were unchanged, with no reruns. This is a same-setting finite-scale reference, not a causal explanation or minimax experiment.

I.3 EARLIER CONTROLS: A SEPARATE 64-STREAM COMPARISON

At the conservative scale $h = 1/128$ and $u = 1$, this independent cohort compares DGB with the unscaled and norm-matched scalar controls. It is not pooled with the 32-pair confirmation above. The additional campaign executed 404750336 updates in 1792 trajectories: 1536 main paths and 256 two-arm positive-control paths. All transactions completed, with zero recorded numerical failures or missing trajectories. The two-arm pairs had identical action sequences and counts. These finite checks are not new independent evidence for multi-arm regret rates.

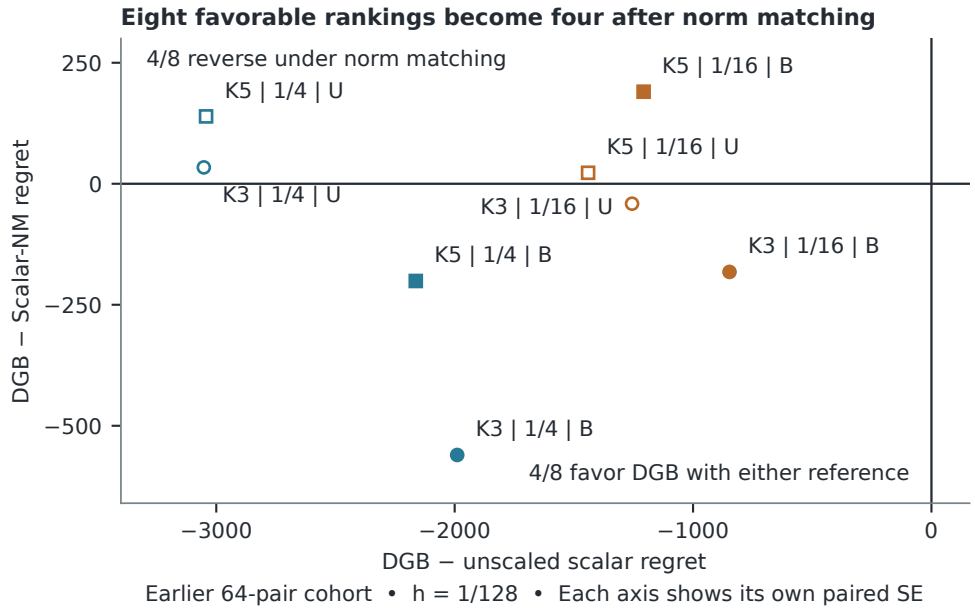


Figure 7: Norm matching changes eight favorable sample-mean rankings into four. Each point is one environment of the earlier 64-pair cohort at $h = 1/128$, $T = 2^{18}$. Its horizontal and vertical coordinates are the paired DGB-minus-unscaled and DGB-minus-Scalar-NM mean regret differences. All eight horizontal differences are negative; four vertical differences are positive. B/U denote biased/uniform logits. Each axis bar is its own paired sample SE, not a joint confidence region. The controls follow their own histories; this is not a common-state causal comparison.

Table 17: Terminal actual-count pseudo-regret at $T = 262144$: mean $\pm$ one sample SE, not a confidence interval. Each environment uses 64 paired replicate runs. Negative DGB-minus-Scalar-NM differences favor DGB; the last column uses the sample SE of paired differences, not combined marginal SEs.

K	δ	Init.	DGB	Scalar only	Scalar-NM	DGB − Scalar-NM
3	0.25	Bias	$61,841.56$ ± 10.97	$63,830.96$ ± 11.04	$62,401.91$ ± 10.59	-560.35 ± 1.67
3	0.25	Uniform	$58,269.66$ ± 15.26	$61,321.89$ ± 12.89	$58,235.83$ ± 15.28	33.82 ± 0.96
3	0.0625	Bias	$25,109.12$ ± 6.95	$25,956.20$ ± 6.08	$25,291.24$ ± 6.82	-182.12 ± 0.97
3	0.0625	Uniform	$29,739.83$ ± 9.92	$30,995.72$ ± 8.15	$29,780.98$ ± 9.86	-41.15 ± 0.41
5	0.25	Bias	$82,831.15$ ± 11.29	$84,994.26$ ± 11.60	$83,032.01$ ± 11.26	-200.86 ± 1.37
5	0.25	Uniform	$88,305.92$ ± 12.10	$91,348.74$ ± 11.99	$88,166.75$ ± 12.27	139.17 ± 0.96
5	0.0625	Bias	$42,366.93$ ± 8.55	$43,574.31$ ± 8.79	$42,176.73$ ± 8.87	190.20 ± 1.54
5	0.0625	Uniform	$51,808.41$ ± 8.99	$53,248.70$ ± 8.03	$51,785.76$ ± 8.99	22.65 ± 0.24

J LEGACY SCALE EXTENSION AND SMALL-GAP DIAGNOSTICS

J.1 A SEPARATE $u = 1$ SCALE-EXTENSION COHORT

In the fixed-h confirmation, raising h from $1/128$ to 0.1 lowers both methods' mean regret in every environment. The question was fixed before the separate scale-extension batch: does increasing the legal scale beyond 0.1 improve finite-scale regret, and who bears the allocation cost? We retained all eight environments, $\gamma = 1/2$, $\omega = 0.6$, $u = 1$, $T = 2^{18}$ and 32 paired streams, and fixed $h \in \{0.1, 0.4, 1\}$. The new random namespace is independent of earlier cohorts; its $h = 0.1$ rows are new controls, not pooled historical records. This is a locally locked design, not a public preregistration. No outcomes were used to choose additional scales, seeds or stopping times.

The sampler and update kernel are inherited from Appendix H.1. Across methods and scales, each environment/replicate shares action uniforms and pull-indexed per-arm reward streams. All 1536 trajectories completed 402653184 updates with zero numeric failures in about 153 seconds on the recorded four-worker host. Failed prefixes would have been retained without replacement.

Timing and uncertainty. We accumulate $A_{t+1} = \sum_{s \le t} \eta_s$ and $Q_{t+1} = \sum_{s \le t} \eta_s^2$ at every round, not by interpolating checkpoint step sizes. Pre-action checkpoints retain $A_t, Q_t, p_t, \overline{C}_t, \eta_t$; post-round checkpoints retain A_{t+1}, Q_{t+1}, counts, regret and p_{t+1}. For DGB at $\gamma = 1/2$, geometric activity is $(K-1)Q_{t+1}$. For Scalar-NM, this number is only a quadratic-clock diagnostic and is not assigned the geometric theorem's meaning. The data file contains all nine checkpoints and all methods/scales. Sample SE uses 32 independent replicate streams. Difference intervals use within-replicate differences and $t_{31} = 2.0395134464$; they are approximate individual 95% intervals, not joint coverage or distribution-free bounds.

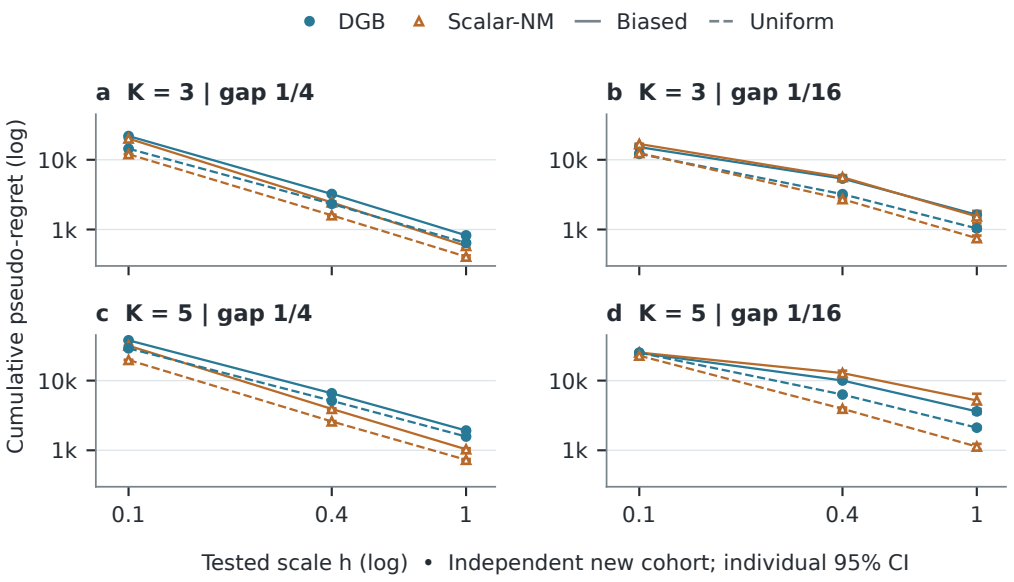


Figure 8: Larger legal scales reduce absolute regret without ensuring DGB superiority. All 48 terminal means in the independent scale-extension cohort: 32 paired streams, $T = 2^{18}$, $\gamma = 1/2$, $\omega = 0.6$, $u = 1$. Color/marker identify method; solid/dashed identify biased/uniform start. Log–log segments connect tested scales, not times. Bars are individual approximate 95% t_{31} intervals; Appendix J gives all 24 paired differences and costs. This cohort is not pooled with the legacy 32-pair fixed-h confirmation cohort of Appendix I.

Table 18: New fixed-scale cohort: terminal regret (mean ± SE), paired difference and measured clocks. B/U denote biased/uniform logits; all rows have 32 pairs.

K, gap, start; h	DGB	Scalar-NM	DGB − NM	A_{T+1}^{DGB}	$\bar{C}_T^{\mathrm{DGB}}/\sqrt{T}$
3, 1/4, B; 0.1	21905 ± 112.9	20163 ± 132.9	1742 ± 24.5	21.8	0.031
3, 1/4, B; 0.4	3229 ± 46.8	2445 ± 46.8	784 ± 9.4	83.6	0.083
3, 1/4, B; 1	824 ± 18.1	580 ± 17.4	244 ± 5.1	193.2	0.170
3, 1/4, U; 0.1	14430 ± 56.3	12037 ± 57.8	2393 ± 7.2	21.8	0.033
3, 1/4, U; 0.4	2329 ± 20.0	1598 ± 18.4	731 ± 4.8	83.5	0.085
3, 1/4, U; 1	642 ± 10.4	407 ± 8.4	235 ± 3.7	192.9	0.175
3, 1/16, B; 0.1	15200 ± 49.8	16848 ± 48.4	-1648 ± 7.6	21.9	0.025
3, 1/16, B; 0.4	5373 ± 123.3	5628 ± 197.8	-255 ± 83.5	85.6	0.053
3, 1/16, B; 1	1627 ± 89.3	1538 ± 152.8	90 ± 76.2	204.2	0.106
3, 1/16, U; 0.1	12168 ± 51.8	12505 ± 67.9	-337 ± 21.3	21.9	0.024
3, 1/16, U; 0.4	3212 ± 59.6	2723 ± 66.4	490 ± 12.6	85.8	0.053
3, 1/16, U; 1	1035 ± 35.2	751 ± 33.3	284 ± 8.8	204.6	0.105
5, 1/4, B; 0.1	37857 ± 101.7	31669 ± 146.3	6188 ± 54.7	19.5	0.082
5, 1/4, B; 0.4	6575 ± 42.6	3924 ± 42.3	2650 ± 13.8	71.8	0.199
5, 1/4, B; 1	1921 ± 20.3	1033 ± 19.4	888 ± 7.2	157.7	0.372
5, 1/4, U; 0.1	29050 ± 51.9	19755 ± 51.3	9295 ± 14.1	19.6	0.083
5, 1/4, U; 0.4	5132 ± 19.7	2595 ± 18.2	2538 ± 7.5	71.9	0.199
5, 1/4, U; 1	1575 ± 13.6	730 ± 10.0	845 ± 6.9	157.7	0.373
5, 1/16, B; 0.1	24986 ± 41.1	25258 ± 38.3	-272 ± 26.7	19.6	0.069
5, 1/16, B; 0.4	10029 ± 159.3	12822 ± 513.9	-2792 ± 366.1	74.5	0.145
5, 1/16, B; 1	3614 ± 138.4	5187 ± 629.3	-1573 ± 505.3	169.3	0.272
5, 1/16, U; 0.1	25175 ± 58.7	22795 ± 92.3	2381 ± 43.1	19.8	0.066
5, 1/16, U; 0.4	6307 ± 67.2	3940 ± 80.7	2366 ± 24.1	74.9	0.146
5, 1/16, U; 1	2112 ± 52.1	1129 ± 52.2	983 ± 15.3	170.0	0.268

What improves, and what does not. Both $h = 0.4$ and $h = 1$ reduce each method's mean regret relative to its new $h = 0.1$ control in all eight environments. Relative rankings do not follow that

improvement: DGB wins 3/8, 2/8 and 1/8 mean comparisons. At $h = 1$, absolute terminal means are 642–3,614 for DGB and 407–5,187 for Scalar-NM (rounded; Table 18). These sample-mean rankings are not significance tests. In the biased three-arm, gap-$1/16$, $h = 1$ cell, its paired excess is 89.64 with approximate interval $[-65.68, 244.95]$. For biased five-arm, gap $1/16$, $h = 0.4$ improves the paired DGB advantage to 2792.42 regret units, but larger $h = 1$ reduces that mean advantage to 1573.31. Larger absolute improvements need not improve a relative margin. Table 19 retains the decomposition for all 24 settings. Larger h increases both directly accumulated update effort and cap attenuation; the retained clocks distinguish these quantities.

Table 19: New scale cohort: paired allocation costs and quadratic clocks. Costs are means $\pm$ sample SE. Activity is $(K-1)\sum \eta_t^2$ for DGB; the Scalar-NM column is a diagnostic only.

Setting; h	Best-arm term	Outside cost	Net	DGB activity	Scalar quadratic
3, 1/4, B; 0.1	2306 ± 19.5	-564 ± 6.5	1742 ± 24.5	0.0090	0.0335
3, 1/4, B; 0.4	803 ± 10.0	-19 ± 4.2	784 ± 9.4	0.1379	2.3703
3, 1/4, B; 1	238 ± 5.3	6 ± 1.5	244 ± 5.1	0.7781	62.0184
3, 1/4, U; 0.1	2241 ± 6.3	151 ± 6.8	2393 ± 7.2	0.0093	0.0391
3, 1/4, U; 0.4	664 ± 5.6	67 ± 2.7	731 ± 4.8	0.1408	3.0265
3, 1/4, U; 1	211 ± 4.4	23 ± 1.5	235 ± 3.7	0.7906	71.2084
3, 1/16, B; 0.1	-559 ± 14.1	-1088 ± 11.1	-1648 ± 7.6	0.0090	0.0314
3, 1/16, B; 0.4	376 ± 70.8	-631 ± 15.1	-255 ± 83.5	0.1408	0.5633
3, 1/16, B; 1	248 ± 69.4	-158 ± 10.2	90 ± 76.2	0.8284	8.6844
3, 1/16, U; 0.1	447 ± 9.6	-784 ± 12.2	-337 ± 21.3	0.0093	0.0294
3, 1/16, U; 0.4	733 ± 10.4	-243 ± 13.3	490 ± 12.6	0.1451	0.7637
3, 1/16, U; 1	348 ± 10.9	-64 ± 6.7	284 ± 8.8	0.8518	13.3994
5, 1/4, B; 0.1	5372 ± 42.4	816 ± 13.0	6188 ± 54.7	0.0104	0.0663
5, 1/4, B; 0.4	2106 ± 12.4	544 ± 4.3	2650 ± 13.8	0.1501	7.5265
5, 1/4, B; 1	688 ± 6.9	200 ± 2.5	888 ± 7.2	0.7787	153.8846
5, 1/4, U; 0.1	6282 ± 10.7	3013 ± 12.1	9295 ± 14.1	0.0107	0.0892
5, 1/4, U; 0.4	1745 ± 8.4	792 ± 3.4	2538 ± 7.5	0.1529	10.1385
5, 1/4, U; 1	587 ± 6.9	259 ± 2.2	845 ± 6.9	0.7904	184.8612
5, 1/16, B; 0.1	-1296 ± 11.4	1024 ± 34.9	-272 ± 26.7	0.0104	0.0634
5, 1/16, B; 0.4	-2469 ± 384.4	-323 ± 33.6	-2792 ± 366.1	0.1561	1.1775
5, 1/16, B; 1	-1490 ± 499.1	-83 ± 23.3	-1573 ± 505.3	0.8519	17.8581
5, 1/16, U; 0.1	1216 ± 17.3	1164 ± 26.5	2381 ± 43.1	0.0108	0.0581
5, 1/16, U; 0.4	1554 ± 21.9	812 ± 27.9	2366 ± 24.1	0.1611	2.5865
5, 1/16, U; 1	693 ± 18.3	290 ± 14.0	983 ± 15.3	0.8789	53.5709

The necessary capacity test in Appendix L must not be read as a sufficient maturity certificate. Raising h relaxes that budget condition, but neither passing it nor observing a larger clock proves entry into the asymptotic allocation law.

Table 20: Necessary joint count capacity in the scale-extension design at $T = 2^{18}$. Values do not depend on initialization. Passing the budget test is not evidence of convergence.

K	Gap	$h = 0.1$	$h = 0.4$	$h = 1$
3	$1/4$	0.944	15.100	94.374
3	$1/16$	0.071	1.134	7.089
5	$1/4$	0.688	11.008	68.797
5	$1/16$	0.066	1.062	6.635

For a concrete tolerance, all normalized suboptimal-arm counts can lie within 10% of one only if $U_{\text{joint}} \geq 0.9$. At gap $1/16$, $h = 0.4$ already clears this necessary test for both K; $h = 0.1$ does not. This changes feasibility, not observed maturity. The artifact's scale-capacity table pairs these deterministic capacities with the directly recorded clock and cap summaries for all 24 settings. It does not reconstruct per-path normalized counts or probabilities from aggregate means, nor apply this count constraint to every asymptotic quantity.

J.2 Small gaps and four neighborhood initializations

The earlier grids include small-gap and shrinking-gap experiments, but not the following fixed combination of gaps, initialization perturbations and controls. This is a separate diagnostic, not a replacement cohort or an exponent-estimation experiment. Before execution we fixed $K \in \{3, 5\}$, $\delta \in \{1/32, 1/64\}$, and four logits: center $\vartheta_* = (0, 1, 0, \ldots)$ (C), $\vartheta_* + (1/8, -1/8, 0, \ldots)$ (+), $\vartheta_* + (-1/8, 1/8, 0, \ldots)$ (−), and $\vartheta_* + (0, 0, 1/8, 0, \ldots)$ (3rd). All belong to $\mathcal{N}$, but four points do not test its uniform quantifier. Rewards use the Bernoulli hard family in Appendix A.

DGB and NM retain the specified updates with $\gamma = 0.5$, $\omega = 0.6$, $h = 0.1$, $u = 1$, $b = 1$. The third comparator, Upper, is the fixed-horizon scalar construction in Theorem A.1: $b = 0$, $\eta = \min\{u, 1/(8eK)\}T^{-1/K}$. It receives T, not the gap or best-arm identity. Each of 48 cells uses $T = 2^{18}$ and 32 new streams. Methods are paired within an environment; initialization is part of the random key, so there is no cross-initialization pairing. No parameter or seed was replaced based on the results. Checkpoints, dyadic sampling, guard conditions and prefix-failure accounting follow Appendix H. All 1536 paths and 402653184 updates completed without numerical failure (78.05 wall-clock seconds with six workers, excluding compilation and analysis). Changed-kernel checks enumerated actions and binary rewards, including Upper's baseline-zero update and its conditional contrast drift; these are finite code checks, not proof acceptance.

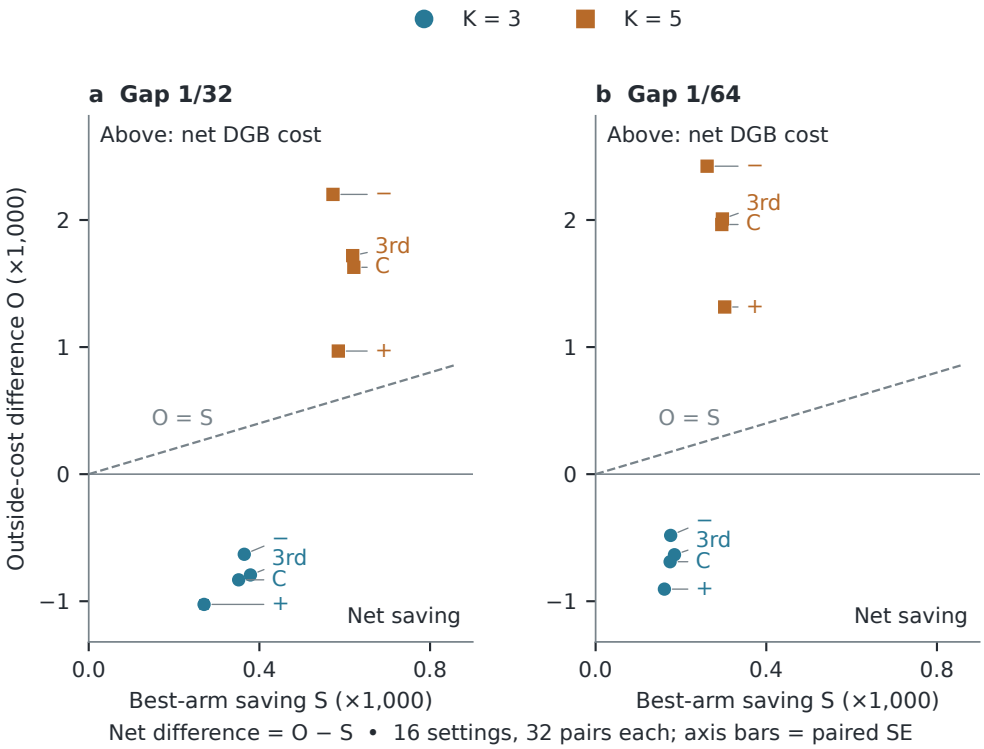


Figure 9: Best-arm savings versus outside-arm costs in all 16 small-gap settings. For the DGB-minus-Scalar-NM accounting, S is the negative best-arm term and O is the signed outside-cost term, so the net regret difference is $O - S$. Above $O = S$ the outside cost exceeds the saving; below it DGB saves net regret. Negative O means outside allocations also save regret. Each coordinate has its original paired sample SE from 32 streams, not a joint confidence region. C/+/−/3rd denote the four starts defined in this subsection; 3rd perturbs the third logit and is distinct from the outside-cost variable O. The two panels share linear axes. All comparisons with Scalar-NM and the upper rule remain in the tables. These finite settings do not test the full uniform-neighborhood quantifier.

Relative to NM, DGB improves mean regret by 7.91%–12.53% across the eight $K = 3$ settings, but worsens it by 2.04%–15.95% across the eight $K = 5$ settings. Relative to Upper, the respective

changes are -15.18% to -11.50%, and $+45.68\%$ to $+72.77\%$. Tables 21–22 give absolute regrets and paired differences, without selecting a favorable comparator.

Against Upper, the same accounting explains the five-arm losses: the best-arm term saves 386.78–873.76 regret units across the eight settings, while the outside term costs 6759.02–7366.35. For example, the central $K = 5$, gap-$1/32$ setting saves 873.76 and pays 6937.78, giving 6064.02 extra regret (rounded paired means). In all eight three-arm settings both terms are negative. These values use this cohort's existing paired records, not cap measurements from another cohort; they do not identify why the update produced those allocations.

The count identity (161), with the chosen comparator in place of Scalar-NM when needed, is particularly informative here. Against NM, DGB's best-arm term is negative in every setting. For $K = 3$ its outside term is also negative; for $K = 5$ the outside penalty exceeds the best-arm saving in every setting. Its terminal contrast and accumulated nominal drift are lower in both dimensions (Table 23). Thus better leader discrimination does not alone imply lower cumulative regret. The identity is exact count accounting, not a causal intervention on a single update component. The code checks every stored count sum and cost identity at all nine checkpoints, with rounding tolerance for the sequential regret sum. Checkpoints and the paired quantities share paths. Raw counts, probabilities, hitting-time censoring flags and full-precision statistics are retained. These outcomes do not contradict the asymptotic class separation, predict its onset, or establish a rate for any fixed finite-scale instance. At fixed $h = 0.1$, $T = 2^{18}$, the pull budget excludes joint per-arm asymptotic normalization (Appendix L); more seeds cannot remove that constraint.

Table 21: Neighborhood diagnostic, $K = 3$: mean realized pseudo-regret and sample SE, 32 streams per method; differences use the paired SE. Upper is the horizon-aware scalar upper rule. No cross-initialization pairing.

$1/\delta$, init	DGB	NM	Upper	DGB−NM	DGB−Upper
32, C	10395.65 ± 31.38	11578.67 ± 24.70	11827.70 ± 16.01	-1183.03 ± 9.98	-1432.06 ± 36.73
32, +	10002.33 ± 38.97	11297.63 ± 32.16	11427.36 ± 21.93	-1295.30 ± 8.75	-1425.02 ± 43.98
32, −	10685.96 ± 31.78	11681.03 ± 27.69	12074.83 ± 15.94	-995.07 ± 11.29	-1388.88 ± 35.73
32, 3rd	10747.85 ± 30.00	11920.06 ± 24.23	12235.32 ± 12.63	-1172.21 ± 11.41	-1487.48 ± 28.04
64, C	7606.95 ± 23.88	8470.65 ± 21.78	8784.40 ± 15.27	-863.70 ± 13.42	-1177.45 ± 28.87
64, +	7440.21 ± 35.09	8506.09 ± 30.87	8771.84 ± 14.84	-1065.88 ± 9.33	-1331.63 ± 41.92
64, −	7656.58 ± 25.15	8314.49 ± 24.48	8756.38 ± 13.80	-657.91 ± 13.67	-1099.79 ± 30.95
64, 3rd	7918.13 ± 27.06	8737.08 ± 23.33	9157.50 ± 13.95	-818.95 ± 14.39	-1239.37 ± 31.20

Table 22: Neighborhood diagnostic, $K = 5$: mean realized pseudo-regret and sample SE, 32 streams per method; differences use the paired SE. Upper is the horizon-aware scalar upper rule. No cross-initialization pairing.

$1/\delta$, init	DGB	NM	Upper	DGB−NM	DGB−Upper
32, C	19215.00 ± 35.32	18208.57 ± 35.24	13150.98 ± 15.88	1006.43 ± 35.63	6064.02 ± 36.93
32, +	19235.85 ± 33.04	18852.14 ± 32.25	12857.66 ± 49.99	383.71 ± 27.42	6378.19 ± 66.59
32, −	19006.19 ± 41.13	17377.17 ± 54.37	13046.43 ± 13.60	1629.02 ± 27.06	5959.75 ± 45.11
32, 3rd	19576.66 ± 34.57	18475.72 ± 39.58	13289.67 ± 16.57	1100.94 ± 27.71	6286.99 ± 39.00
64, C	16110.20 ± 41.08	14441.28 ± 42.83	9443.07 ± 18.54	1668.92 ± 25.24	6667.13 ± 41.88
64, +	16301.11 ± 35.65	15288.14 ± 36.62	9650.42 ± 23.61	1012.97 ± 21.65	6650.69 ± 51.11
64, −	15714.76 ± 31.72	13552.66 ± 36.45	9153.95 ± 17.67	2162.10 ± 13.81	6560.81 ± 37.13
64, 3rd	16401.24 ± 38.18	14690.41 ± 52.45	9493.02 ± 21.03	1710.83 ± 27.25	6908.21 ± 41.82

Table 23: DGB minus NM: cost components, terminal state contrast and accumulated nominal drift, with paired sample SE. These are dependent views of the same 32 paths, not independent evidence.

$K, 1/\delta$, init	Best-arm term	Outside term	dX	Drift difference
3,32,C	-351.40 ± 4.66	-831.62 ± 11.09	-0.2787 ± 0.0055	-0.2807 ± 0.0033
3,32,+	-270.32 ± 8.37	-1024.98 ± 13.65	-0.1613 ± 0.0083	-0.1659 ± 0.0062
3,32,-	-364.54 ± 3.73	-630.53 ± 11.61	-0.3596 ± 0.0060	-0.3653 ± 0.0027
3,32,3rd	-379.02 ± 3.16	-793.20 ± 12.71	-0.3099 ± 0.0049	-0.3070 ± 0.0037
3,64,C	-174.59 ± 1.52	-689.11 ± 13.83	-0.3198 ± 0.0055	-0.3274 ± 0.0033
3,64,+	-161.21 ± 2.47	-904.67 ± 10.27	-0.2416 ± 0.0052	-0.2391 ± 0.0034
3,64,-	-175.88 ± 1.53	-482.03 ± 13.36	-0.3997 ± 0.0044	-0.3983 ± 0.0024
3,64,3rd	-184.81 ± 1.49	-634.14 ± 14.86	-0.3521 ± 0.0046	-0.3570 ± 0.0032
5,32,C	-621.32 ± 3.33	1627.76 ± 35.04	-0.9094 ± 0.0091	-0.9077 ± 0.0064
5,32,+	-584.99 ± 6.17	968.70 ± 32.65	-0.6833 ± 0.0124	-0.6858 ± 0.0089
5,32,-	-572.44 ± 3.60	2201.46 ± 25.51	-1.0583 ± 0.0073	-1.0518 ± 0.0027
5,32,3rd	-618.65 ± 2.90	1719.60 ± 27.65	-0.9228 ± 0.0083	-0.9253 ± 0.0051
5,64,C	-295.71 ± 1.55	1964.63 ± 24.83	-0.9614 ± 0.0054	-0.9640 ± 0.0039
5,64,+	-302.53 ± 1.60	1315.50 ± 22.25	-0.7945 ± 0.0061	-0.7863 ± 0.0050
5,64,-	-261.29 ± 1.50	2423.39 ± 13.49	-1.0663 ± 0.0052	-1.0682 ± 0.0014
5,64,3rd	-297.15 ± 1.73	2007.98 ± 27.44	-0.9676 ± 0.0081	-0.9712 ± 0.0038

K LEGACY STATE, DRIFT AND SAME-STATE MOMENT DIAGNOSTICS

The allocation identity and its sign convention are given once in (161). Figure 3 and Table 13 report the legacy $u = 1$, $h = 0.1$, 32-pair confirmation; Figure 4 retains all four scales. The following state and moment diagnostics complement that count accounting.

K.1 STATE AND DRIFT ARE DIFFERENT MEASUREMENTS

Checkpoints retain post-update contrast X_{t+1}, nominal pre-action drift m_t evaluated at the simulator's actual action probabilities, and $\sum_{s \le t} m_s$ accumulated at every update. In the exact model, $m_t = \mathbb{E}_t[X_{t+1} - X_t]$, so $X_{t+1} - X_1 - \sum_{s \le t} m_s$ is a martingale residual. The logged difference also includes binary64 rounding and accumulation errors: it is not an exact machine-process martingale certificate, and a path need not equal its drift sum. No unlogged phase entry is inferred. For the scalar score, the drift of $X = \theta_2 - \theta_1$ is $\eta\{(p_2 - p_1)\sum_{i \ge 3}(1/2 - \mu_i)p_i - \delta p_1(1 + p_2 - p_1)\}$. For DGB it is instead $\eta\{p_2^{1-\gamma}(\mu_2 - p^\top \mu) - p_1^{1-\gamma}(\mu_1 - p^\top \mu)\}$. Conditional drift is enumerated from the actual DGB or scalar action kernel; the original-score $O - G$ formula is not applied to DGB. Diagnostics evaluate each nonlinear expression before averaging, using true means only offline and retaining the pre-action/post-update distinction.

Conditional moments at the same frozen state. Let $g_i = e_i - p$, $P = I - \mathbf{1}\mathbf{1}^\top/K$, $D = D_\gamma(p)$, and $C = \operatorname{diag} p - pp^\top$. For Bernoulli rewards and baseline one, $\mathbb{E}[(Y-1)^2 \mid I = i] = 1 - \mu_i$. With the same policy, cap and base scalar η, the two projected updates have second moments and means

$$M_{\mathrm{DGB}} = \eta^2 \sum_i p_i(1 - \mu_i)\|PDg_i\|^2, \qquad m_{\mathrm{DGB}} = \eta PDC\mu,$$

$$M_{\text{Scalar-NM}} = \eta^2 s(p)^2 \sum_i p_i(1 - \mu_i)\|g_i\|^2, \qquad m_{\text{Scalar-NM}} = \eta s(p) C\mu,$$

where $s(p)$ is defined in (160). Their covariance traces are $V_A = M_A - \|m_A\|^2$. Table 24 reports $M_{\mathrm{DGB}}/M_{\text{Scalar-NM}}$ and $V_{\mathrm{DGB}}/V_{\text{Scalar-NM}}$ at the terminal DGB states, averaged over streams. True means enter only this offline calculation, not either online controller. Score-norm matching does not impose equality of either reward-weighted quantity.

K.2 THE SAME-STATE COMPARISON AT THE FIXED-SCALE ENDPOINT

We repeat the offline calculation at all retained DGB states in the eight $h = 0.1$ confirmation environments, using their existing 32 paired streams. The 512 DGB/NM journals contain 4608 checkpoints;

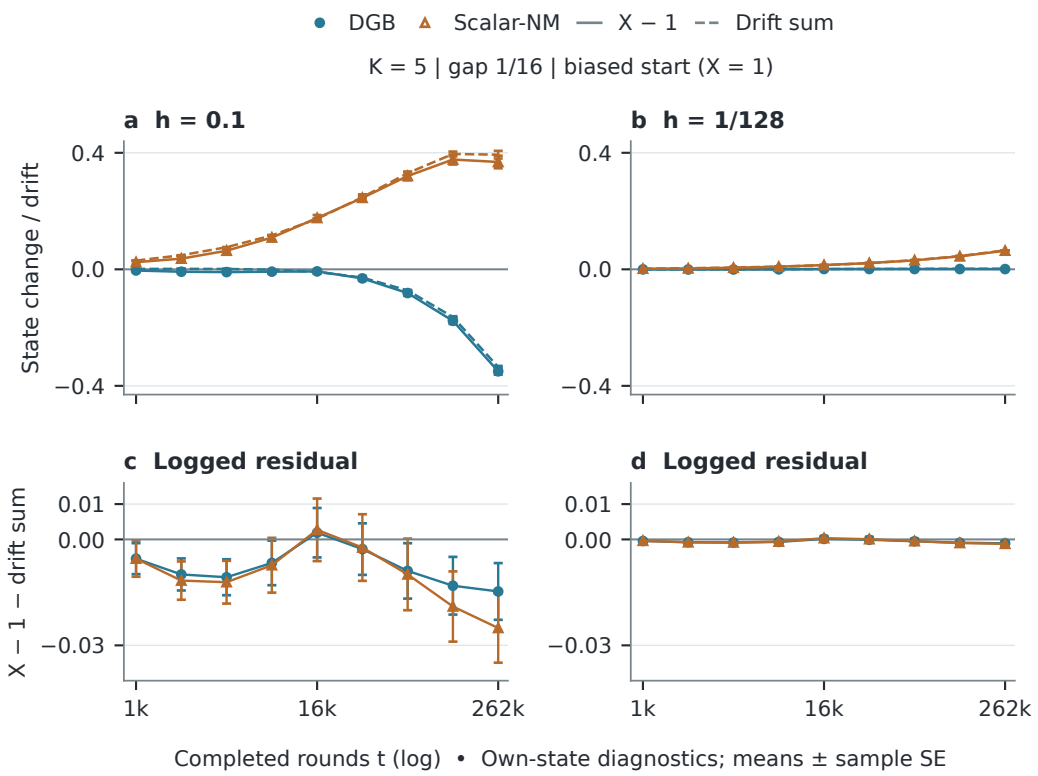


Figure 10: State change and accumulated nominal drift for $K = 5$, gap $1/16$, biased start, $\gamma = 1/2$, $\omega = 0.6$, $u = 1$. Columns use $h = 0.1$ and $h = 1/128$. (a–b) Post-update rival-minus-best contrast centered at its initial value, $X_{t+1} - 1$, and accumulated pre-action drift. Zero denotes no change from initialization, not best-arm overtaking. (c–d) The directly logged residual, with its own sample SE; it is not obtained by adding marginal variances. All bars are one sample SE from 32 same-path streams at nine checkpoints. At $h = 0.1$, terminal mean X remains positive for both methods (DGB 0.650, Scalar-NM 1.368). These rounded own-state diagnostics are neither a causal intervention nor a new machine-precision martingale certificate.

the hypothetical same-state comparisons use the 2304 DGB states, not 2304 independent replicates. All selected paths completed. The parameters are $\gamma = 1/2$, $\omega = 0.6$, $u = 1$ and $T = 2^{18}$, with means and initialization as in Appendix I. We do not pool these records with the earlier cohort.

Table 24: Same-state ratios at $h = 0.1$: projected DGB versus a hypothetical norm-matched scalar at the same DGB policy, cap and base step. Values are means ± sample SE of 32 per-stream ratios, not ratios of averaged moments. Covariance trace subtracts the squared conditional mean. True reward means enter only this offline diagnostic; neither column measures rollout regret.

K	δ	Initialization	Second-moment ratio	Covariance-trace ratio
3	1/4	Biased	1.1754 ± 0.0007	1.1837 ± 0.0007
3	1/4	Uniform	1.1213 ± 0.0006	1.1266 ± 0.0006
3	1/16	Biased	1.1929 ± 0.0004	1.1917 ± 0.0004
3	1/16	Uniform	1.1602 ± 0.0007	1.1597 ± 0.0007
5	1/4	Biased	1.2594 ± 0.0007	1.2790 ± 0.0008
5	1/4	Uniform	1.1427 ± 0.0008	1.1559 ± 0.0009
5	1/16	Biased	1.2460 ± 0.0007	1.2458 ± 0.0007
5	1/16	Uniform	1.1988 ± 0.0008	1.2012 ± 0.0008

Both kinds of ratio differ from one in Table 24; score-norm matching is therefore not reward-weighted second-moment or covariance matching. This frozen-state diagnostic is distinct from the

own-trajectory drifts displayed above. It does not identify which discrepancy causes a regret difference. The artifact retains every per-state value and all 256 paired terminal regret differences, including medians, q90 and observed extrema. Within each environment all 32 observed differences have the same sign (three environments favor DGB and five NM), but this does not bound an unobserved population tail. No confidence interval for a ratio of population means is inferred from these per-stream summaries.

K.3 Earlier conservative-scale controls

The separate 64-stream comparison uses $h = 1/128$, not the confirmation's $h = 0.1$ endpoint. At biased $K = 5$, $\delta = 1/16$, DGB's best-arm gain is 61.4531 ± 0.3009 but its extra outside cost is 251.6578 ± 1.5852, leaving 190.2047 ± 1.5391 more regret. Both terms improve in two biased $K = 3$ settings; both worsen in the two uniform $K = 5$ settings. The source package retains this older cohort's paired records in `data/RESULTS.json`, distribution and checkpoint views in `figures/integrated/S1_distributions.pdf` and `figures/integrated/S2_dynamics.pdf`, and its offline summaries in `figures/integrated/offline_table.tex`; they are not the 32-pair confirmation records. The maximum checked stored-versus-count regret discrepancy is below 1.33×10^{-7}, consistent with the retained binary64 accounting check. Neither a favorable local drift nor small cap attenuation alone identifies the regret of a counterfactual process with a different full history.

L Can the retained horizon reveal the pathwise coefficient?

Corollary F.4 is a fixed-instance theorem, not a forecast that its leading term is visible at the experimental horizon. We first check necessary finite-count constraints, then analyze retained records; we add no simulations or fitted slopes. Fix $\gamma = 1/2$ and write

$$Z_i(T) = \frac{(2\omega+1)h^2\Delta_i^2 N_i(T)}{L_T^{2\omega+1}}, \quad P_i(t) = \frac{h^2\Delta_i^2 t p_{i,t}}{L_t^{2\omega}},$$
$$V_T = \frac{\widehat{R}_T}{c_\Delta L_T^{2\omega+1}}, \qquad c_\Delta = \frac{\sum_{i\neq j}\Delta_i^{-1}}{(2\omega+1)h^2}.$$

All three normalized quantities converge to one in the indicated exact-real, fixed-instance limits. They need not be close at the same finite time.

Capacity before convergence. The deterministic inequality $N_i(T) \le T$ gives

$$Z_i(T) \le U_i(T) := \frac{(2\omega+1)h^2\Delta_i^2 T}{L_T^{2\omega+1}}.$$

The multi-arm constraint is stronger: since $\sum_{i\neq j} N_i(T) \le T$, the weighted average of the $Z_i(T)$ with weights Δ_i^{-2} is at most

$$\min_{i\neq j} Z_i(T) \le \frac{\sum_{i\neq j}\Delta_i^{-2} Z_i(T)}{\sum_{i\neq j}\Delta_i^{-2}} \le U_{\text{joint}}(T) := \frac{(2\omega+1)h^2 T}{L_T^{2\omega+1}\sum_{i\neq j}\Delta_i^{-2}}.$$

In particular, all normalized counts can be at least $1-\varepsilon$ only if $U_{\text{joint}}(T) \ge 1-\varepsilon$. This is a necessary budget test, not a sufficient onset theorem or a promised convergence tolerance.

Table 25: Single nearest-rival count capacity at $h = 0.1$, $\omega = 0.6$, $T = 262144$. The bound uses only $N_2(T) \le T$; it does not use a fitted trajectory or depend on initialization.

Nearest gap $\delta = \Delta_2$	1/4	1/16	1/32	1/64
$U_2(T)$	1.179675	0.073730	0.018432	0.004608

For the three smaller gaps, a normalized rival count near one is impossible at this horizon, regardless of how many seeds are used. Even when the single-arm capacity exceeds one at $\delta = 1/4$, joint capacity is only 0.943740 for $K = 3$ and 0.687973 for $K = 5$ under the fixed-h reward vectors. Initializations do not change these capacities. Passing a necessary capacity test would still not establish maturity; the running cap and transient allocations can delay it.

What the logs actually permit. We read all 2048 journals in the fixed-h cohort, checking their cell, replicate, checkpoint and completion bindings against retained hashes. These endpoint diagnostics use the eight $h = 0.1$ DGB environments: 32 paths each, with nine checkpoints on each path (2304 states, not 2304 independent replicates). The logs retain pre-action probabilities, running cap and scalar at those checkpoints, as well as cumulative arm counts and regret. They do *not* retain $A_t = \sum_{s<t} \eta_s$, the total scalar effort before the action. Sparse values of η_t do not reconstruct this sum, so the normalized scalar-clock curve is unavailable and is not imputed. The artifact retains all arms and checkpoints for P_i, Z_i, V_T, and $L_t^{\omega}\overline{C}_t/(h\sqrt{t}\sum_{i\neq j}\Delta_i)$.

Table 26: Endpoint diagnostics from the eight fixed-h environments at $h = 0.1$, $\omega = 0.6$, $T = 2^{18}$. Each entry is the median [q10, q90] over 32 DGB paths. Arm 2 is the nearest rival. These empirical quantiles are not confidence intervals.

K	δ	Init.	$Z_2(T)$	$P_2(T)$	V_T
3	$1/4$	Biased	0.279 [0.266, 0.289]	0.743 [0.706, 0.776]	0.261 [0.252, 0.269]
3	$1/4$	Uniform	0.141 [0.134, 0.144]	0.418 [0.402, 0.432]	0.175 [0.169, 0.179]
3	$1/16$	Biased	0.044 [0.042, 0.044]	0.247 [0.236, 0.254]	0.057 [0.056, 0.058]
3	$1/16$	Uniform	0.023 [0.022, 0.025]	0.121 [0.114, 0.131]	0.046 [0.045, 0.047]
5	$1/4$	Biased	0.329 [0.322, 0.338]	1.020 [0.995, 1.063]	0.276 [0.272, 0.281]
5	$1/4$	Uniform	0.140 [0.135, 0.145]	0.442 [0.426, 0.466]	0.212 [0.208, 0.217]
5	$1/16$	Biased	0.042 [0.041, 0.043]	0.265 [0.255, 0.271]	0.071 [0.070, 0.072]
5	$1/16$	Uniform	0.020 [0.019, 0.021]	0.117 [0.109, 0.124]	0.072 [0.071, 0.073]

Table 26 illustrates why a local allocation value near one does not establish global maturity. In the biased $K = 5$, $\delta = 1/4$ cell, median $P_2(T)$ is 1.020, while median $Z_2(T)$ is 0.329 and median V_T is 0.276. Over all eight environments, median V_T ranges from 0.046 to 0.276. These quantities describe different stages of accumulated behavior; a near-unit local value is not evidence that the asymptotic regret coefficient has become operative. The recorded probabilities are pre-action dyadic simulator probabilities; counts and regret are post-round quantities. Small numerical discrepancies from the exact-real model retain the semantics of Appendix H.1. Neither these quantiles nor the capacity bounds prove uniform integrability, an exact expected coefficient, or a uniform onset over shrinking gaps.

M Legacy numerical reliability, censoring and finite-scale limits

Failure-aware summaries. All five failures are pre-action relative-distortion guard stops in the original fixed-scalar, baseline-one comparator. They do not establish support loss, quantify the rejected distortion, or refute an exact-real theorem. Four $T = 2^{20}$ summary rows are censored. For each failed path we retain the regret accumulated by its last completed update, and bound its unobserved suffix by the remaining number of rounds times the largest gap. Averaging these per-path bounds with the observed complete paths gives the planned finite-sample bounds in Table 27. These bounds are not confidence intervals for the population expectation. Successful-path moments are explicitly conditional; the unconditional sample mean field is unavailable, not replaced by zero or by the successful-path mean.

Compute and finite numerical checks. The artifact's environment snapshot records an Intel Core i7-13620H CPU, Windows, PowerShell 7.6.5 and .NET 10.0.11; the launch configuration requests eight worker cells. The core campaign completed 84959564952 updates in 25226.65 seconds of process wall time (7.01 hours) and 146009.38 seconds of process CPU time (40.56 CPU-hours), measured from the operating-system process. These are not a sum of overlapping trajectory times. The native C# simulation has no GPU backend. A separate, previously specified $K = 10$ auxiliary ran 12 unchanged method variants through 512 warm-up and 65536 measured updates each (792576 total), with no exception or numerical failure. Its timing overlapped the core campaign and is not an isolated speed comparison or additional regret sample. Finite high-precision checks made 17656 comparisons at 120 decimal digits with no threshold violation; the exact-softmax evaluator and actual dyadic probability/update checks were distinguished. These finite checks do not

Table 27: All censored endpoint rows. Fixed scalar, baseline one throughout; "shrinking" denotes the predeclared shrinking-gap diagnostic. Conditional moments use complete endpoint observations only. Outward-rounded brackets bound the mean over the planned 256 paths, not a population expectation.

K	Gap / initialization	Observed	Failures	Conditional mean	Planned bounds
5	$1/4$ / biased	254	2	574.08	[573.92,745.02]
5	$1/16$ / biased	255	1	1214.52	[1215.40,1335.57]
5	$1/4$ / uniform	255	1	484.88	[484.94,504.61]
3	shrinking / biased	255	1	477.79	[478.14,549.30]

certify infinite-time numerical robustness, outward-rounded confidence intervals, or portability of floating-point trajectory bits to untested platforms.

M.1 Representative constants and a conservative onset condition

Table 28: Direct evaluation of the constants in Appendix B.2 at $K = 3$, $u = 1$, $\omega = 0.6$, $h = \gamma/64$, and $\vartheta = (0, 1, 0)$. The last column is the gap-free bound expression at $T = 2^{20}$ before taking the minimum with T. These are analytical bounds, not experimental measurements.

γ	C_{init}	C_{rate}	Gap-free expression
0.5	7.13×10^3	4.53×10^4	6.63×10^8
0.25	4.19×10^6	2.66×10^{10}	3.88×10^{14}
0.1	9.33×10^{26}	1.44×10^{52}	2.10×10^{56}

For the first row, the three summands in the printed rate constant are

$$C_{\mathrm{rate}} = \underbrace{\frac{2A_1 + A_0}{hc_\gamma}}_{45291.36} + \underbrace{\frac{e^{ah} D_C A_0}{c_\gamma}}_{20.00} + \underbrace{\frac{d_1 h A_0}{c_\gamma}}_{0.09}.$$

The first term accounts for approximately 99.956% of this bound. It combines comparison-potential and cap-moment estimates with the reciprocal design scale $1/h$; it is not a measurement that the running cap causes 99.956% of observed regret. The terminal cap attenuation in Appendix K measures a different, statewise quantity. This decomposition locates conservatism in the proof expression. For the expanded fixed-h family with $\gamma = 1/2$, the same formula gives $C_{\mathrm{rate}} = 11995.59, 6380.01, 4647.17$ at $h = 0.03, 0.06, 0.1$. At $h = 0.1$, $C_{\mathrm{init}} = 556.79$ and the gap-free expression at $T = 2^{18}$ remains $3.20 \times 10^7 > T$. Thus these smaller constants still do not provide a useful finite-scale regret estimate or a tuning oracle.

One conservative sufficient onset condition in the scalar lower proof is $\sqrt{1 + \log T} \geq 320K^2/(K - 2)$. For $K = 3$ this alone requires $\log_{10} T \geq (8294400 - 1)/\log 10 \approx 3.60 \times 10^6$. This is a sufficient condition of the proof, not a necessary time for actual separation or an empirical estimate. Neither bound justifies extrapolating finite curves to an observed asymptotic regime.

Censoring in the 64-stream comparison. All 24 main configurations have 0 observed hits, 64 right-censored endpoints and 0 missing endpoints at both best-arm thresholds 0.5 and 0.9. The complete machine-readable tables retain every row; plotting a set of all-censored curves would not identify differences in hitting times.

N Extended interface comparisons

The comparison concerns update and information interfaces, not a claim that non-scalar updates or stochastic regret guarantees are new individually. Table 29 separates the nearest prior interfaces; Table 30 separates quantifiers within this paper.

Table 29: Why existing efficiency results do not settle the present comparison. The distinctions are at the algorithm interface and order of quantifiers, not a claim that earlier methods lack efficient bandit guarantees.

Line of work	Update or control	Relation to this question
Original-score upper bounds (Baudry et al., 2025; Lattimore, 2026b; Jia & Ouyang, 2026)	Specific scalars, with the initialization, tuning and moment conditions discussed here	These upper bounds do not establish the fixed-resource multi-arm class separation proved here.
SAMBA; Tsallis-INF (Walton & Denisov, 2023; Zimmert & Seldin, 2021)	Direct probability-coordinate or implicit mirror updates	Established efficiency outside the original score; not the specified explicit diagonal logit process.
LB-SGB (Cesani et al., 2026)	Softmax with an added log-barrier gradient	A closely related structural exploration mechanism, not a common scalar or pure diagonal-score update.
Present comparison	Same sampler and feedback; original score versus explicit D_γ	A bounded adaptive-class lower bound, constructive recovery, and a separate finite-activity boundary.

Table 30: Controller and quantifier interfaces. No rule receives the unknown gaps. Resources and K are fixed; guarantees concern expected regret.

Result	Initialization	Horizon interface	Conclusion
Scalar, $K = 2$	Any fixed finite	May know T	$\Theta(\sqrt{T})$
Scalar, $K \geq 3$	Fixed radius-$1/8$ biased neighborhood	May know T	Exponent $1 - 1/K$; log gap
Explicit DGB	Any fixed finite	One anytime rule	Two stochastic bounds
Finite activity	Any fixed finite	One anytime rule	No $O(\log^2 T)$ in subclass
Scalar, Theorem D.1	Any fixed finite	Fix one anytime rule	Rule-dependent exponent $> 1/2$

Original-score convergence and uniform regret. Exploration and noise arguments underpin stochastic-gradient-bandit convergence results (Mei et al., 2023; 2024; Robertson et al., 2025); almost-sure convergence alone does not give uniform finite-time regret. Baudry et al. (2025), Proposition 4 and Remark 4, give unknown-gap, horizon-free two-arm bounds from uniform logits; the remark improves the bound to $O(\sqrt{T}\log T)$ using $\eta_t = 1/\sqrt{t}$. Our two-arm $O(\sqrt{T})$ construction may know T, so removes logarithmic slack under a different horizon interface and covers every fixed finite initialization. This is a supporting refinement, not our main novelty. The old schedule's initial step is one; membership in a prescribed $\eta_t \leq u$ class is not automatic when $u < 1$. Lattimore (2026b) analyzes discrete fixed-scalar, gap/horizon-tuned updates from zero logits. In a continuous-time Gaussian diffusion, Lattimore (2026a), Section 4, identifies how eliminating inferior arms can select the wrong near-best arm and create an unfavorable effective initialization. Its Theorem 10 uses a fixed scalar and an arm count growing at least logarithmically with horizon/gap. Our distinction is not discovering that mechanism: it is a discrete bounded-IID obstruction for every fixed $K \geq 3$, uniform over the bounded predictable scalar/baseline class and the specified logit neighborhood. Jia & Ouyang (2026), Section 2.3, treat discrete updates with predictable baselines: Theorem 3 and Remark 1 in the version-1 PDF give fixed-small-scalar logarithmic regret under a residual-moment condition, including finite deterministic initialization. Theorem 3 additionally requires the fixed step ℓ to satisfy $\ell K \Gamma_K \leq \Delta_{\min}$, with Γ_K controlling residual moments. Bounded rewards do not remove this gap-dependent restriction. It therefore does not certify a single positive step uniformly along our shrinking-gap family. Nor does it automatically satisfy the shared-rule and additive-constant decomposition premises of Proposition C.2; those must be checked separately. We explicitly adapt their subset potential, equation (65), to an analysis-only near-optimal group in Appendix A.3. Lu et al. (2024) use a horizon-dependent decreasing stochastic schedule for optimization error. These upper bounds for particular learning-rate choices under their stated conditions do not settle the uniform cost over unknown instances of the full bounded adaptive class.

Exploration, baselines and alternative estimators. Baselines, stochasticity and commitment affect policy geometry (Chung et al., 2021; Mei et al., 2022; 2021); Garg et al. (2022) address saturation with a different estimator. These mechanisms do not supply the bounded-class minimax comparison.

Earlier geometric routes to efficiency. Walton (2020) gives ODE arguments for softmax and a probability-coordinate alternative, motivating state-dependent step powers. These precedents are distinct from a discrete adaptive-class minimax theorem. For Bernoulli bandits, SAMBA achieves gap-unknown $O(\log T \log\log T)$ regret through probability-coordinate steps and leader-relative importance weighting (Walton & Denisov, 2023); its preprint Theorem 4 (v3, Appendix C) further allows slowly varying factors. Its interface is neither original-score scalar nor our explicit D_γ rule, so its rate does not contradict our activity boundary. Tsallis-INF uses regularized OMD (Zimmert & Seldin, 2021); our mirror/IW reference uses that geometry with its stated bias and scale. LB-SGB adds a log-barrier gradient to softmax sampling (Cesani et al., 2026), with gap/horizon-tuned parameters in v2, Corollary C.10. TS-Prod already combines explicit arithmetic probability updates with stochastic/adversarial guarantees (Zimmert & Marinov, 2024). Neither is the diagonal logit process studied here.

Melo & Müller (2025) use established GEV/GNL choice models, changing sampling and preference dynamics; their adversarial GBPA theorem is separate from the stochastic construction. Labbi et al. (2026) change parameterization, regularization and a projection-like monotone-improvement operator for discounted MDPs; Mei & Osband (2026) use action-specific gates and exact advantages in a full-action update. Our novelty is the uniform original-score obstruction paired with this geometric recovery, not geometric exploration, explicit arithmetic or polylogarithmic bandit regret in general.

Why the main text uses $\gamma = 1/2$. At this endpoint the tail-mass factor $q_t^{-(1/2-\gamma)}$ equals one, making the controller simplest to state. Every proved $0 < \gamma \le 1/2$ has the same fixed-K polynomial exponent; this is not a finite-scale optimality claim for $\gamma = 1/2$. The algebraic score $D_1(p)(e_I - p)$ lies outside that proof range. Similarities to natural-gradient or importance-weighted scores do not identify every stochastic implementation with this logit process: the baseline, state coordinates, cap and sampling rule also specify the interface.