SHARP SAFETY AND CRITICAL EXPECTATION DEFECTS IN SOFTMAX GRADIENT BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Softmax gradient bandits adapt their exploration through updates driven by sampled rewards. Choosing the learning rate therefore requires understanding how these updates shape cumulative regret. Almost-sure convergence alone leaves this question unresolved, because rare trajectories can retain substantial expected loss even as typical trajectories learn the optimal arm. For K IID arms with a unique optimum, a common predictable baseline, reward-residual bound M, and gap floor $0 < \Delta \leq M$, we identify the exact learning-rate ceiling for class-uniform logarithmic expected regret, including the endpoint. Below this ceiling, every fixed positive rate η yields cumulative pulls $\log T/(K\eta\Delta_i)$ to leading order for each suboptimal arm with gap Δ_i, almost surely and in L^1. Thus typical and expected allocations agree to leading order, with equal leading regret contributions from all suboptimal arms. At the ceiling, an explicit two-arm family with signed rewards retains logarithmic expected regret, but rare, prolonged excursions toward the suboptimal arm raise its leading coefficient above the typical-path value. A positive harmonic function quantifies this excess and its dependence on initialization. A uniform near-critical expansion for the same family connects the critical correction to polynomial expected regret above the ceiling.

1 INTRODUCTION

Softmax gradient bandits update action logits from a single sampled reward, using the softmax score vector and, optionally, a common reward baseline. This update belongs to the classical likelihood-ratio approach to reinforcement learning (Williams, 1992). Unlike confidence-bound or posterior-sampling bandit methods (Auer et al., 2002; Agrawal & Goyal, 2012), its exploration is generated by its own stochastic logit dynamics. Understanding when those dynamics yield low regret remains a distinct theoretical question.

Recent results show that convergence and regret can part ways. Stochastic gradient can succeed for bandits (Mei et al., 2023), and almost-sure convergence can hold at arbitrary fixed positive learning rates (Mei et al., 2024; Robertson et al., 2025). Yet constant-step gradient bandits can exhibit either logarithmic or polynomial expected regret (Baudry et al., 2025). A policy may therefore converge on almost every path while increasingly rare paths retain enough loss to change its expected performance. Convergence alone does not specify a safe learning rate, an expected regret coefficient, or the reliability of a typical learning curve.

We study the unchanged sampled softmax score update with a common learning rate and a common predictable reward baseline. A rate is *safe* here when it guarantees logarithmic expected regret uniformly over a specified reward-residual and gap class. Safety answers an order-of-growth question. It leaves two finer questions: how the algorithm allocates its logarithmic exploration, and whether the coefficient seen on typical paths also describes the mean. Our results connect these three questions.

Contributions.

- **The exact safe range, including its endpoint.** For K arms, residual bound M and gap floor $0 < \Delta \leq M$, we identify the exact ceiling $\eta_c = s_*(\Delta/M)/(KM)$ for class-uniform logarithmic expected regret. The guarantee holds for every predictable rate sequence in a fixed interval with

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.

positive lower endpoint if and only if its upper endpoint is at most η_c. Above the ceiling, a bounded IID instance has polynomial expected regret. This sharpens sufficient small-rate conditions to a necessary and sufficient boundary (Section 4).

- **How exploration is shared below the ceiling.** At every fixed $0 < \eta < \eta_c$ in that class, $N_i(T)/\log T \to 1/(K\eta\Delta_i)$ almost surely and in L^1. Smaller-gap arms receive more pulls, but all suboptimal arms contribute the same leading regret. Beyond the inverse-gap profile discussed by Baudry et al. (2025), we establish cumulative limits and justify their passage to expectations (Section 5).
- **Why the critical mean exceeds the typical coefficient.** In a specified signed two-arm family, the almost-sure logarithmic coefficient persists at all fixed positive rates. At η_c, the expected coefficient acquires a positive, initialization-dependent correction. A normalized harmonic function determines this correction, and the transition variable $(1 - \eta_c/\eta)\log T$ connects critical logarithmic behavior to supercritical polynomial growth (Section 6).

The mechanism connecting the results. The distinction comes from rare trajectories that spend a long time assigning substantial probability to suboptimal arms. If q_t is their total probability, the square-occupancy sum $\sum_{t<T} q_t^2$ records this behavior and enters the drift analysis. Strictly below the ceiling its infinite sum has finite expectation; together with logit moment bounds, this allows the cumulative allocation limit to pass from paths to expectations. At the ceiling, the sum can be finite almost surely while its finite-horizon mean grows logarithmically. An exact two-arm log-odds decomposition (Baudry et al., 2025) locates the additional regret in this term; our analysis determines its positive coefficient.

The main results use a residual budget; we also retain separate raw-reward and baseline conditions where these give a different admissible class. The experiments examine the visible consequences: mean–median separation, tail concentration, and finite-horizon departures from the allocation law. They make concrete why the safety of an expected regret order and the representativeness of a typical learning curve are distinct properties.

2 Model and the safety question

At each round the learner selects one of $K \geq 2$ arms, observes its reward, and updates the logits that determine its next sampling probabilities. Arm i has mean reward μ_i; arm 1 is uniquely optimal, a label used only in the analysis, and $\Delta_i = \mu_1 - \mu_i > 0$ for $i > 1$. The softmax probabilities are $p_{i,t} = e^{\theta_{i,t}}/\sum_j e^{\theta_{j,t}}$. Writing Y_t for the observed reward, the sampled-score update is

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

Here $e_{I_t} - p_t = \nabla_\theta \log p_{I_t}(\theta_t)$ is the score vector. The common learning rate η_t scales the update, and the common baseline b_t centers its reward. Both are chosen before observing the current action and reward. At a positive learning rate, a positive residual raises the selected arm's logit relative to the others; a negative residual reverses the update.

Cumulative loss. Let $N_i(T) = \sum_{t<T} \mathbf{1}_{\{I_t=i\}}$. We measure the opportunity cost of selecting suboptimal arms through $\widehat{R}_T = \sum_{i>1} \Delta_i N_i(T)$ and $R_T = \mathbb{E}\widehat{R}_T$. The first is random realized pseudo-regret; the second is its expectation. Neither includes accumulated reward noise. Put $q_t = 1 - p_{1,t}$ and $r_t = \sum_{i>1} \Delta_i p_{i,t}$, so $R_T = \mathbb{E}\sum_{t<T} r_t$. Asymptotic statements fix model parameters as $T \to \infty$ unless a joint limit is explicitly stated.

Rewards and admissible control. Each arm has an IID reward stream, independent across arms, and its n-th pull returns its n-th draw. The history $\mathcal{F}_t$ includes all information available before the action, determining finite logits, a nonnegative η_t and a bounded b_t. Conditional on this history, the action-drawing randomness and next unused rewards are independent and retain their original laws. Initialization is finite and deterministic. Control may depend on history and independent controller randomization; the update has no resets, projection, forced pulls, arm-specific rates or action-dependent baseline. Appendix E describes the numerical implementation.

The residual-budget class. For $M > 0$ and $0 < \Delta \leq M$, $\mathcal{C}_K(\Delta, M)$ consists of the instances and admissible baselines satisfying

$$|Y^{(i)} - b_t| \leq M \text{ conditionally a.s. for every next unused reward,} \qquad \Delta_i \geq \Delta \quad (i > 1).$$

A *class-uniform* logarithmic guarantee fixes these budgets, the initialization and the rate interval, then uses one constant over all admissible instances and control rules. It is stronger than an instance-dependent guarantee and keeps a positive gap floor, unlike a gap-uniform minimax question. Section 4 identifies its exact ceiling η_c; fixed rates $0 < \eta < \eta_c$ define the *strict residual-safe region* used for allocation.

Separate bounds $|Y| \leq R$ and $|b_t| \leq B$ embed in this class by taking $M = R + B$ and $0 < \Delta \leq \min\{\Delta_{\min}, M\}$, where $\Delta_{\min} = \min_{i>1} \Delta_i$. Here R denotes a reward bound. These bounds specify a smaller class; for $K = 2$, its exact boundary is treated separately in Section 4.

Table 1: Scope of the results. All rows use IID rewards, a unique best arm and finite deterministic initialization. The first three rows follow the main argument; the last two give separate-budget refinements. The signed family is defined in Section 6, and the raw-budget margin c in (3).

Result	Reward/baseline scope	Learning rate	Conclusion
Safety (Thm. 1)	$K \geq 2$; residual bound M; gap floor $\Delta \leq M$	Predictable $\eta_t \in [\eta_-, \eta_+]$, $\eta_- > 0$	Exact ceiling for class-uniform $R_T = O(\log T)$
Allocation (Cor. 3)	$\mathcal{C}_K(\Delta, M)$	Fixed $0 < \eta < \eta_c$	Cumulative coefficients a.s. and in L^1
Critical defect (Thm. 4)	Specified two-arm signed family; $b_t = 0$	Fixed $\eta > 0$ in each process	Harmonic correction and near-critical crossover
Allocation (Thm. 2)	Raw bounds $\|Y\| \leq R$, $\|b_t\| \leq B$	Fixed $\eta > 0$ with $c > 0$	Same cumulative coefficients
Separate budgets (App. C)	$K = 2$; $\|Y\| \leq R$, $\|b_t\| \leq B$; $\Delta \leq 2R$	Predictable interval with positive floor	Exact four-corner budget criterion

3 Positioning relative to prior work

Convergence and regret are different guarantees. The stochastic-gradient bandit analyses of Mei et al. (2023), Mei et al. (2024) and Robertson et al. (2025) establish increasingly broad convergence guarantees. Almost-sure convergence at arbitrary fixed rates remains compatible with a large contribution from rare trajectories. Baudry et al. (2025) make this distinction explicit through logarithmic and polynomial regret regimes, critical-scale analysis and time-varying two-arm guarantees. Their Appendix G.3 discusses inverse-gap exploration and equal per-arm logarithmic regret contributions. We establish cumulative allocation throughout the strict residual-safe region. Our result makes the cumulative almost-sure and L^1 limits precise; a pointwise probability limit is not required.

What is sharpened and what is imported. Jia & Ouyang (2026) also prove a discrete-time theorem (arXiv v1 PDF, Section 2.3, Theorem 3 and Remark 1). Their subset products, layer weights and cross-layer absorption inform our multi-arm proof (Section 4). We sharpen the bounded-residual question to an exact ceiling, necessity and a closed endpoint for predictable rates in a fixed interval. That work's conditional exponential-moment model is broader in a different direction. Lattimore (2026b) studies discrete-time regret with a horizon-dependent restriction: its Theorem 1 assumes $\eta \leq \Delta_{\min}^2/[120\Delta_{\max}\log(TK)]$ for rewards in $[0, 1]$. The continuous-time analysis of Lattimore (2026a) supplies another perspective on the dynamics, rather than the discrete endpoint identity studied here. Table 2 summarizes the closest distinctions.

Table 2: Position of the present results. In the last row, $u = K\eta M$ and $d = \Delta/M$; the comparison substitutes the same bounded residual budget into the prior moment condition. Here ψ_d is the exponential drift function defined in (2); the prior full moment class is broader.

Prior work	Relevant guarantee or mechanism	Question addressed here
Robertson et al. (2025)	Almost-sure convergence with arbitrary fixed positive rates	Expected regret and its coefficient; convergence is an input to our converse proof
Baudry et al. (2025)	Logarithmic/polynomial regimes; inverse-gap profile discussion	Exact class ceiling with equality; cumulative L^1 limits; critical harmonic correction
Jia & Ouyang (2026)	Subset potentials; sufficient budget condition $ue^u \le d$ for a fixed rate	Necessary and sufficient: $\psi_d(u) \le 0$, including equality; predictable rate intervals

The update and baseline matter. Changing the gradient estimator (Garg et al., 2022) or coupling a different parameterization with regularization (Labbi et al., 2026) changes the learning dynamics. Here both softmax and its sampled score estimator are fixed. A common baseline cancels from the conditional mean gradient, but still changes stochastic increments. This accords with the broader observation that baseline effects are not exhausted by variance reduction (Chung et al., 2021). Our separate-budget criterion asks the specific class-uniform safety question under bounds on both rewards and baselines. Finally, the inverse-gap allocation describes a fixed gradient algorithm and learning rate; it is distinct from the information-dependent optimal allocation of classical bandit theory (Lai & Robbins, 1985; Lattimore & Szepesvári, 2020).

4 The sharp safety boundary

Let $s_*(d) > 0$ be the unique positive solution of

$$\psi_d(s) := \cosh s + (1-d)\sinh s - 1 - s = 0, \qquad d = \Delta/M \in (0, 1]. \tag{2}$$

Strict convexity, $\psi_d'(0) = -d$ and divergence at infinity ensure uniqueness. Equivalently, $\psi_d(s) = \mathbb{E}e^{sZ} - 1 - s$ for a ± 1-valued Z with mean $1-d$: Appendix A.1 shows that it is the supremum of the normalized binary exponential drift factor over the class, achieved in the limit by two-point residuals. The sign of this exponential remainder, rather than a quadratic approximation alone, determines the exact boundary. Define $\eta_c = s_*(d)/(KM)$.

Theorem 1 (Closed class-uniform safety)**.** *Fix K, M, Δ, θ_0 and $0 < \eta_- \le \eta_+ < \infty$. There is a finite constant depending only on these quantities such that $R_T \le C\log(e+T)$ uniformly over $\mathcal{C}_K(\Delta, M)$ and every predictable $\eta_t \in [\eta_-, \eta_+]$ if and only if $\eta_+ \le \eta_c$. Strictly below the boundary, $\mathbb{E}\sum_{t\ge 0} q_t^2 < \infty$; at equality, $\mathbb{E}\sum_{t<T} q_t^2 = O(\log(e+T))$. In both cases square occupancy is finite a.s. and $p_{1,t} \to 1$ a.s. Above the boundary, for every fixed $\eta > \eta_c$ and $0 < \gamma < 1 - \eta_c/\eta$, there is a bounded IID zero-baseline instance with $R_T \ge c_\gamma T^\gamma$ eventually.*

The theorem identifies a class boundary, not a pointwise classification of all instances. Its small-gap scale is $\eta_c = 2\Delta/(KM^2)(1 + O(\Delta/M))$. The critical endpoint is included in the expected-logarithmic set but not automatically in the set with integrable infinite square occupancy.

Proof idea. The expected regret bound needs two controls: logarithmic logit growth and expected square occupancy. Inverse-logit moments give the first. For the second, the proof resolves three successive obstacles. *Two arms: accumulate negative drift.* Below the ceiling, the suboptimal odds have negative drift proportional to the squared suboptimal probability. Summing this drift bounds square occupancy; ψ_d captures the worst-case exponential factor. *More arms: absorb their interactions.* Cross terms between distinct suboptimal arms can be positive, so a separate two-arm bound for each arm does not suffice. For each nonempty subset A of suboptimal arms use $h_{A,z} = e^{z\theta_1}\prod_{i\in A} e^{\theta_i - \theta_1}$, grouping equal-size subsets into layers. The identity $h_{A,z}p_ip_j = h_{A\cup\{j\},z}p_1p_i$ for $i \in A$, $j \notin A$, $j > 1$ moves a positive cross term to the next layer. Suitable layer weights absorb it into that layer's negative drift. *At the ceiling: recover the lost margin.* The top layer loses its margin. The proof factor $e^{z\theta_1}$ restores a margin of order z without changing the algorithm. The

resulting bound costs $1/z$; choosing $z \asymp 1/\log T$ therefore yields logarithmic square occupancy. A higher power of $1/z$ would not give the endpoint bound.

For necessity, a reward tilt and zero-reward embedding produce an auxiliary stochastic gradient bandit to which Robertson et al. (2025) applies. Its likelihood cost supplies a supercritical bad instance. Together these arguments establish both sides of the boundary; Appendix A gives the full proof.

Separate reward and baseline budgets. Raw bounds $|Y| \le R$ and $|b_t| \le B$ imply residual bound $R + B$, but the smaller class has its own necessity question. For $K = 2$, $0 < \Delta \le 2R$, Theorem 10 gives the exact criterion $\Phi_{R,B,\Delta}(\eta_+) \le 0$, including equality. The four expressions in (24) retain reward and baseline budgets separately; Appendix C supplies their extremal laws and both directions.

5 Cumulative allocation

Below the safety ceiling, the theory gives more than a logarithmic regret order. For every fixed $0 < \eta < \eta_c$ in $\mathcal{C}_K(\Delta, M)$, the cumulative pulls of each suboptimal arm satisfy $N_i(T)/\log T \to 1/(K\eta\Delta_i)$ almost surely and in L^1 (Corollary 3). Thus an arm with half the gap receives twice as many pulls to leading order, while each suboptimal arm contributes the same $1/(K\eta)$ to the logarithmic regret coefficient. This identifies how the unchanged gradient update distributes exploration.

The precise statements below compare actual pulls $N_i(T)$ with their accumulated conditional probabilities $Q_i(T) = \sum_{t<T} p_{i,t}$ and the interaction exposure $\mathcal{E}_i(T) = \eta \sum_{t<T} p_{1,t} p_{i,t}$. The same limits also hold under a separate sufficient condition on raw rewards and baselines. For $|Y^{(i)}| \le R$, $|b_t| \le B$, put $M = R + B$ and $\Delta_{\min} = \min_{i>1} \Delta_i$, and define

$$\Psi(x) = 2 \sum_{n \ge 0} \frac{x^n}{(n+2)!}, \qquad c = \Delta_{\min} - \frac{K\eta M^2}{2} \Psi(K\eta M) > 0. \tag{3}$$

Here $\Psi(x) = 2(e^x - 1 - x)/x^2$ for $x > 0$, and $\Psi(0) = 1$.

Theorem 2 (Cumulative allocation)**.** *Under* (3)*, for every suboptimal arm $i > 1$,*

$$\frac{N_i(T)}{\log T}, \ \frac{Q_i(T)}{\log T} \longrightarrow \frac{1}{\eta K \Delta_i}, \qquad \frac{\mathcal{E}_i(T)}{\log T} \longrightarrow \frac{1}{K\Delta_i}, \tag{4}$$

almost surely and in L^1. In particular $\widehat{R}_T/\log T \to (K-1)/(\eta K)$ in both modes and $R_T/\log T \to (K-1)/(\eta K)$.

Corollary 3 (Strict residual-safe allocation)**.** *For an instance in $\mathcal{C}_K(\Delta, M)$ and a constant $0 < \eta < \eta_c$, the allocation and regret limits of Theorem 2 hold with an admissible common predictable baseline and finite deterministic initialization.*

The two statements give alternative sufficient domains for the same allocation law. Corollary 3 combines the strict square-occupancy bound with the expectation argument below (Appendix B.5). On the overlap $\Delta_{\min} \le M$, the raw-budget condition is more restrictive: for $u = K\eta M$, it requires $\Delta_{\min}/M > (e^u - 1 - u)/u$, whereas the strict residual condition requires $\Delta_{\min}/M > (e^u - 1 - u)/\sinh u$, with $\sinh u \ge u$. The raw-budget result also covers $\Delta_{\min} > M$.

From balanced clocks to logarithmic allocation. Define the gap-weighted clocks $a_i(T) = \eta\Delta_i Q_i(T)$ and $a(T) = \sum_{i>1} a_i(T)$. Integrable square occupancy bounds the accumulated second-order drift corrections. The centered logit martingales have conditional variance controlled by increments of a, so their contribution is $o(a)$ almost surely. To leading order the best logit grows like a and tail logit i decreases like a_i. An under-allocated tail arm then has a larger relative softmax weight. The resulting clock-balancing argument gives $a_i/a \to 1/(K-1)$. Inverting the relation between these clocks and calendar time yields

$$a(T) \sim \frac{K-1}{K} \log T, \qquad a_i(T) \sim \frac{1}{K} \log T \quad \text{a.s.}$$

Thus the inverse-gap allocation follows from balancing gap-weighted exploration, rather than from an assumed smooth probability trajectory.

Why the expectation statement needs another argument. An almost-sure clock limit does not determine its mean. The proof separately bounds inverse-logit moments, beginning with $\mathbb{E}e^{-\theta_{i,T}} \leq C_i(1+T)$. Tail integration controls the squared negative logit at scale $\log^2 T$. Together with the clock martingale bound and integrable square occupancy, this establishes uniform integrability of the normalized clocks. Finally $\mathbb{E}|N_i(T) - Q_i(T)| = O(\sqrt{\log(e+T)})$ transfers the limit to actual counts. The uniform-integrability step is what fails in the critical family below; the normalized count-compensator error still vanishes in L^1 there.

The cumulative formulation is essential. Even for three deterministic, positive-reward arms satisfying (3), the quantity $tp_{i,t}$ need not converge: the infinitely many pulls of arm i cause nonvanishing multiplicative jumps. Appendix B supplies an explicit example. The theorem resolves cumulative allocation, without asserting a pointwise probability equivalent.

The strict inequality in the corollary has a concrete role. At the critical endpoint, square occupancy can retain a logarithmic expectation even though its infinite sum is finite almost surely. The next section computes the resulting coefficient difference.

6 When typical and expected logarithmic coefficients differ

The safety theorem includes the critical ceiling. Yet in the family below, the expected logarithmic coefficient is strictly larger than the typical one and depends on initialization. The *signed family* has $K = 2$, baseline zero, $Y^{(2)} = -M$, and $Y^{(1)} \in \{-M, M\}$ with probability $d/2$ of $+M$. Its gap is $\Delta = dM$. This two-point law realizes the limiting extremal construction in Appendix A.1; its threshold equals the class ceiling. Fix $\eta > 0$ and put $s = 2\eta M$, so $s/s_* = \eta/\eta_c$.

Where the extra coefficient comes from. Let $X_t = e^{\theta_{2,t} - \theta_{1,t}}$ be the suboptimal odds, $X_0 = x > 0$, $L_t = -\log X_t$, and $H_T = \sum_{t<T} q_t^2$. The two-arm conditional score drift gives the exact decomposition

$$R_T = \frac{M}{s}\big(\mathbb{E}L_T - L_0\big) + \Delta \mathbb{E}H_T. \tag{5}$$

Log-odds growth describes learning; square occupancy records time spent giving substantial probability to the suboptimal arm. Squaring suppresses small probabilities relative to large ones. Indeed, $\mathbb{E}_t(L_{t+1} - L_t) = 2\eta\Delta p_{1,t} q_t$ and $q_t = p_{1,t} q_t + q_t^2$. Baseline cancellation gives the same identity for any two-arm instance at fixed η with a common predictable baseline, using $M/s = 1/(2\eta)$. Following this decomposition (Baudry et al., 2025), we determine the signed family's critical occupancy term exactly.

Typical paths drive X_t to zero, while rare large-odds excursions can retain expectation mass. With $\kappa = s_*/s$, define

$$h_s(x) = \lim_{n\to\infty} \mathbb{E}_{s,x} X_n^{\kappa}.$$

This positive harmonic function measures the surviving mass from initial odds x. Harmonicity means that its value equals its one-step conditional average. Appendix D proves existence and the normalization that fixes its scale.

Theorem 4 (Critical expectation defect)**.** *For every fixed $s > 0$, $\widehat{R}_T / \log T \to M/s$ a.s. For $s < s_*$ the convergence is also in L^1, with $R_T = (M/s)\log T + O(1)$. At $s = s_*$, let $v_* = s_* \psi_d'(s_*)$ and $a_*(x) = h_{s_*}(x)/v_* > 0$. Then*

$$R_T = \left[\frac{M}{s_*} + \Delta a_*(x)\right]\log T + O(1), \qquad \mathbb{E}H_T = a_*(x)\log T + O(1). \tag{6}$$

Here $H_\infty < \infty$ a.s. but $\mathbb{E}H_\infty = \infty$; normalized counts are not uniformly integrable. For fixed $s > s_*$, $R_T = \Theta(T^{1 - s_*/s})$ although the a.s. logarithmic coefficient persists.

At criticality, $\mathbb{E}L_T = \log T + O(1)$, so (5) places the extra coefficient entirely in $\mathbb{E}H_T$. Each typical path has finite total square occupancy, but rare paths keep its expectation growing logarithmically. This failure of uniform integrability is why taking an almost-sure limit does not recover the mean. Table 3 compares the three fixed-rate regimes.

Table 3: Exact asymptotic statements for the specified two-arm endpoint family, at each fixed M, d, s, x. The first column is not an all-instance classification. At criticality $a_*(x) > 0$.

Rate	Typical $\widehat{R}_T / \log T$	Expected regret R_T
$s < s_*$ ($\eta < \eta_c$)	M/s	$(M/s)\log T + O(1)$
$s = s_*$ ($\eta = \eta_c$)	M/s_*	$[M/s_* + \Delta a_*(x)]\log T + O(1)$
$s > s_*$ ($\eta > \eta_c$)	M/s	$\Theta(T^{1-s_*/s})$

Finite-horizon crossover. A finite horizon can conceal the change of asymptotic regime: rates close to the ceiling can look critical on either side. The relevant scale compares distance from criticality with logarithmic observation time. Put $\epsilon = 1 - s_*/s$ and $F_\epsilon(T) = (T^\epsilon - 1)/\epsilon$, continuously defined as $\log T$ at zero. Define $a_s(x) = \epsilon h_s(x)/\psi_d(s)$ for $s \neq s_*$, with continuous value $a_{s_*} = a_*$. Uniformly for $7/8 \leq s_*/s \leq 9/8$ with fixed M, d, x,

$$\left| \frac{R_T - (M/s)\log T}{\Delta F_\epsilon(T)} - a_s(x) \right| \leq C \left(|\epsilon| + \frac{1}{\log T} + T^{-1/3} \right). \tag{7}$$

The crossover scale can be read directly from $F_\epsilon(T) = \log T\,(e^{\epsilon \log T} - 1)/(\epsilon \log T)$. When $|\epsilon| \log T$ is small, finite horizons can resemble the critical logarithmic law even on either side of the boundary. Each member of this parameterized family uses a fixed learning rate throughout its trajectory. If $s_T \to s_*$ and $\epsilon_T \log T \to z$, the expected logarithmic coefficient tends to $M/s_* + \Delta a_*(x)(e^z - 1)/z$. The proof subtracts exact finite-time power-moment identities before dividing by the vanishing margin; fractional moments control the remainder uniformly (Appendix D).

7 Experiments: paths, means and finite horizons

The experiments ask what the asymptotic distinctions look like in finite samples. We first compare typical regret with the mean near criticality, then examine cumulative allocation below the ceiling. A separate safety sweep shows how an individual instance can behave beyond a class guarantee. All displays use the existing simulations and protocols in Appendix E.

Settings and measurement. The critical and crossover experiments use the unperturbed signed two-arm family of Section 6. The allocation experiment and signed safety sweep use the distinct multi-arm construction with a fixed perturbation in Appendix E.1. For criticality and crossover, we report original-law means, medians and path quantiles. Because rare trajectories can dominate the mean, we also use separate proposal paths that favor large suboptimal odds, then reweight their contributions by prefix likelihood ratios. The proposal tilt uses $X^\kappa_{t+1}/X^\kappa_t$; Appendix E.3 specifies the estimator. Original and proposal samples are never pooled. An independent trajectory is the statistical unit.

Tail visibility at nominal criticality. At $d = 0.75, x = 10, T = 262,144$, the original-law mean of normalized pseudo-regret is 6.63 (SE 2.90), whereas its median is 1.80. The largest 21 of 2,048 paths contribute 58.0% of observed regret. The separate prefix-weighted mean is 6.09 (SE 0.40; 512 proposal paths). The estimates are compatible at this observed precision. The mean–median separation and tail contribution show how a small fraction of paths can dominate finite-horizon mean regret.

Reading the crossover. Figure 1(b) subtracts the typical logarithmic term from mean regret and divides the excess by $\Delta F_\epsilon(T)$. The theory predicts that, at each fixed $z = \epsilon \log T$, this rescaled excess tends to the same $a_*(x)$ as the horizon grows. Colors distinguish horizons; each point uses a rate chosen before its run. Across z at one finite horizon, the reference $a_s(x)$ still varies with the rate. Over the observed horizons, the normalized curves retain visible dependence on z, with larger uncertainty on the negative-z side. The complete grid, including noisy estimates, appears in Appendix F.

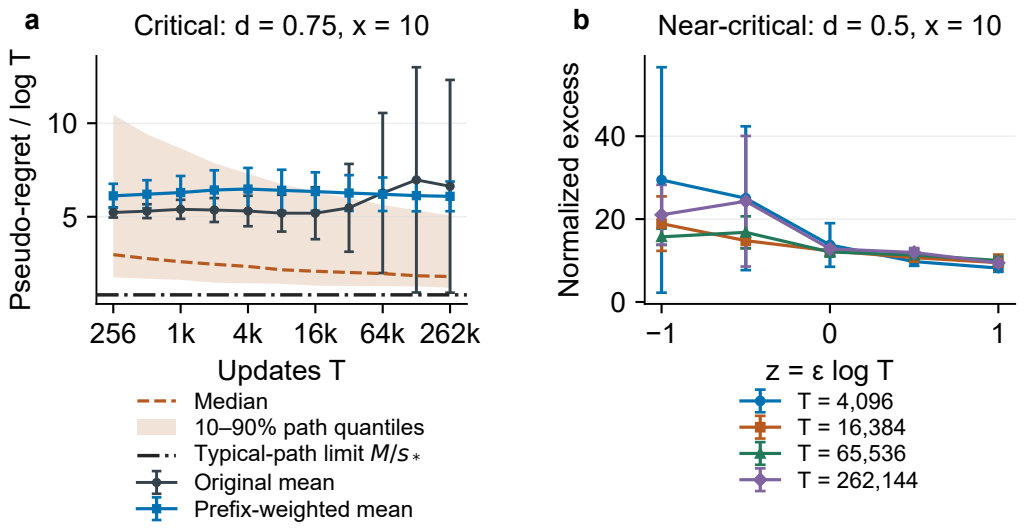


Figure 1: Criticality and crossover diagnostics. (a) $d = 0.75, x = 10$, nominal critical rate: 2,048 original-law and 512 proposal paths. Mean error bars are descriptive 1.96-SE intervals; the shaded band spans 10th–90th path quantiles, not a confidence interval for the median. The horizontal line is the typical-path limit. (b) Each point estimates $[R_T - (M/s)\log T]/[\Delta F_\epsilon(T)]$ from 128 proposal paths for one (T, z) condition on the crossover grid ($d = 0.5$). The rescaled rate $s = s_*/(1 - z/\log T)$ is fixed within each run; curves connect distinct conditions at a common horizon. By (7), at fixed z the quantity estimated at each point tends to $a_*(x)$ as T grows.

What the tail diagnostics show. At the displayed critical cell the mean terminal likelihood-ratio weight is 4.38×10^{-5}, against exact-arithmetic expectation 1, and terminal ESS is 18.33 of 512 paths. These concern terminal weights, not the prefix-sum estimator. They show why a small reported SE cannot exclude unseen tail mass. The simulator uses binary64 states and finite-bit action and reward laws; root brackets and executed-law discrepancies are recorded in Appendix E. The plots illustrate finite conditions, without identifying the limiting harmonic coefficient.

How allocation approaches its reference. Figure 2 follows all suboptimal arms in four safety cells satisfying (3): $K = 2, 3, 5, 10$, $d = 0.5$, logits $(0, 1, 0, \ldots, 0)$ and $\eta/\eta_c = 0.5$. The normalization $\eta K \Delta_i N_i(T)/\log T$ converts each arm's predicted coefficient to one. At $T = 65,536$, its sample means range from 1.035 to 4.000. Thus the available horizons still show substantial departures from the asymptotic allocation, especially for arm 2 at $K = 10$. The four cells were selected by the declared parameter rule before extracting allocation outcomes; all 512 paths and all 16 suboptimal arms are retained. Appendix G gives the compensators, descriptive SEs and complete selection details.

The class ceiling and individual instances. At $K = 10, d = 0.5, T = 65,536$, increasing η/η_c from 1 to 1.5 lowers mean regret$/\log T$ for the signed multi-arm construction from 22.52 to 16.70. A class ceiling need not produce a visible transition in every instance at finite time. Figure 3 retains the sweep, including the cell where two of 128 paths contribute 68.6% of regret.

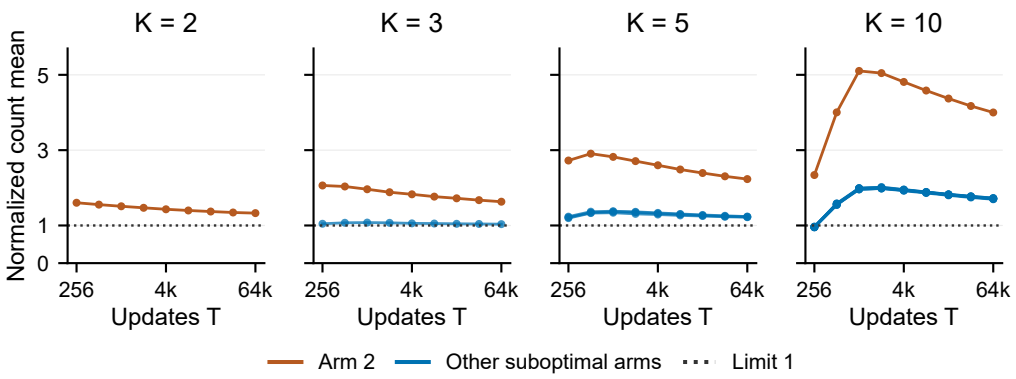


Figure 2: Cumulative allocation below the ceiling. Each curve is the sample mean of $\eta K \Delta_i N_i(T)/\log T$ across 128 independent paths at nine checkpoints. Orange denotes arm 2; blue denotes every other suboptimal arm, shown separately even where curves overlap. The dotted line is the exact-model limit one. These finite-horizon curves retain all selected arms; they are not fitted asymptotes. Pointwise descriptive SEs are supplied in the artifact, with terminal SEs in Table 6.

8 Discussion and Limitations

The three results separate the safety of a regret order from agreement of its pathwise and expected coefficients. Integrable square occupancy and logit moment control connect them below the ceiling; the signed two-arm family shows where this agreement breaks at the endpoint. Extending its exact critical formulas to broader reward families or within-run adaptive rates remains open.

Using the ceiling to select a learning rate with a logarithmic regret guarantee requires valid budget bounds and a positive rate floor. Data-driven rate selection requires further work, as does making the harmonic coefficient efficiently computable: its certified finite-tree evaluation has exponential cost (Appendix D.7).

For empirical evaluation, median learning curves can describe typical success while missing a leading contribution to mean regret. Mean uncertainty, path quantiles and tail concentration therefore carry different information. The critical endpoint changes the interpretation of the logarithmic coefficient, even while the expected regret remains logarithmic.

Reproducibility Statement

Complete proofs and protocols are in the supplement. The anonymous `g4h3-artifact.zip` provides logs, plotting sources, a figure/table map and verification/replay commands in its `README.md`.

AI Use Statement

Generative AI was used extensively for ideation, theory and hypotheses, proof drafting and critique, counterexample and literature searches, translation, experimental design, simulation/reporting code, data analysis, figures and manuscript preparation. Measurements came from the archived simulator. The theoretical proofs and experimental results have undergone human review. The authors retain responsibility for accuracy, originality, citations and conclusions.

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.

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 of *Proceedings of Machine Learning Research*, pp. 1999–2009, 2021. URL https://proceedings.mlr.press/v139/chung21a.html.

Shivam Garg, Samuele Tosatto, Yangchen Pan, Martha White, and A. Rupam Mahmood. An alternate policy gradient estimator for softmax policies. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 6630–6689, 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.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

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

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

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

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.

Samuel Robertson, Thang D. 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. URL https://papers.neurips.cc/paper_files/paper/2025/hash/17a84369cf6e2803101294b5d74dd92c-Abstract-Conference.html.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992. doi: 10.1007/BF00992696.

A THE EXACT RESIDUAL-CLASS SAFETY BOUNDARY

We use the exact update with a bounded predictable common baseline and scalar, independent pull-indexed IID reward streams, finite deterministic initial logits, and unique best arm 1. Let $\Delta_i = \mu_1 - \mu_i$, $q_t = 1 - p_{1,t}$, $r_t = \sum_{i>1} \Delta_i p_{i,t}$, and $S_0 = \sum_i \theta_{i,0}$, which is conserved. The conditional residual support condition is

$$|Y^{(i)} - b_t| \le M \quad \text{a.s. given } \mathcal{F}_t \quad \text{for every unused arm reward,} \qquad \Delta_i \ge \Delta, \quad 0 < \Delta \le M. \tag{8}$$

This is not merely an empirical bound on observed rewards. For fixed $0 < \eta_- \le \eta_+ < \infty$, let ULog denote existence of one finite $C = C(K, M, \Delta, \eta_-, \eta_+, \theta_0)$ such that $R_T \le C \log(e + T)$ for every instance/baseline satisfying (8) and every predictable scalar in $[\eta_-, \eta_+]$.

Theorem 5 (Sharp class-uniform safety)**.** *Let $d = \Delta/M$ and $\psi_d(s) = \cosh s + (1-d)\sinh s - 1 - s$. There is a unique positive root $s_*(d)$, and*

$$\mathsf{ULog} \quad \Longleftrightarrow \quad \eta_+ \le \eta_c := \frac{s_*(d)}{KM}. \tag{9}$$

Strictly below the boundary, $\mathbb{E}\sum_{t\ge 0}\sum_{i>1} p_{i,t}^2 < \infty$. At the boundary, $\mathbb{E}\sum_{t<T} q_t^2 = O(\log(e + T))$; both below and at it, $\sum_t q_t^2 < \infty$ a.s. and $p_{1,t} \to 1$ a.s. For any fixed $\eta > \eta_c$ and $0 < \gamma < 1 - \eta_c/\eta$, an explicit bounded IID instance has $R_T \ge c_\gamma T^\gamma$ for all sufficiently large T. The hard instance may depend on γ.

We prove the upper bound by an exact binary extremum and a common deformation of subset potentials. We prove the lower bound by a normalized change of measure, using a standard SGB only after a zero-reward embedding.

A.1 THE BINARY EXPONENTIAL EXTREMUM

Strict convexity follows from $\psi_d''(s) = (2-d)e^s/2 + de^{-s}/2 > 0$. Together with $\psi_d(0) = 0$, $\psi_d'(0) = -d < 0$, and divergence at infinity, this proves the unique positive root and the sign of ψ_d on either side. If $Z_1, Z_2 \in [-M, M]$ have mean gap at least Δ, then, for $p \in (0,1)$, $q = 1 - p$ and $a > 0$,

$$\frac{\mathbb{E}e^{-aqZ_1} - 1}{q} + \frac{\mathbb{E}e^{apZ_2} - 1}{p} \le \psi_d(aM). \tag{10}$$

To prove this, use the convex chord bound $\mathbb{E}e^{vX} \le \cosh v + (\mathbb{E}X)\sinh v$ for $X \in [-1, 1]$. The mean-dependent objective has a negative coefficient for the first mean and a positive coefficient for the second. Its maximum occurs at gap d and one of the two endpoints $(1, 1-d)$, $(-1+d, -1)$; their objectives are interchanged by swapping p, q. At $(1, 1-d)$ the exponential expansion is

$$-ds + \sum_{n\ge 2} \frac{s^n}{n!}\{(-1)^n q^{n-1} + a_n p^{n-1}\}, \quad s = aM, \quad a_n = \begin{cases} 1 & n \text{ even}, \\ 1-d & n \text{ odd}. \end{cases}$$

The even coefficient is at most 1, and the odd coefficient is at most $1 - d$, using $d \le 1$. This proves (10). Taking $p \uparrow 1$ at that endpoint attains the supremum, with passage justified by the absolutely uniformly convergent exponential series. Conditioning on the current history preserves the argument; a probability mixture of tail-arm residuals retains the support and mean-gap conditions.

A.2 TWO-DIMENSIONAL FACES AND EXACT CROSS-LAYER CANCELLATION

Let $m = K - 1$, $\mathcal{I} = \{2, \dots, K\}$, and $z \in [0, 1/4]$. For each nonempty $A \subseteq \mathcal{I}$ set

$$h_{A,z} = e^{z\theta_1} \prod_{i\in A} e^{\theta_i - \theta_1}.$$

The same factor $e^{z\theta_1}$ is used at every layer. Fix a history, put $s = |A|$, $d_A = s - z$, $C = \mathcal{I} \setminus A$, $a = P_A$, $c = P_C$, and $p = p_1 = 1 - a - c$. Let M_1, M_A, M_C be the residual MGFs of the best arm and the within-group mixtures at this history. Empty-group MGFs can be filled with any legal tail

law since their multiplier is zero. Hold these distributions fixed when differentiating the artificial variables a, c. Direct substitution of the score gives $\mathbb{E}_t \Delta h_{A,z} = h_{A,z} f(a,c)$ with

$$f(a,c) = pM_1(-\eta[(d_A+1)a + d_A c]) \\ + aM_A(\eta[(d_A+1)p + c]) + cM_C(\eta[d_A p - a]) - 1.$$

The exponents and their first derivatives are bounded by K before the factor η. For affine w, ℓ,

$$\partial_{ac}[wM(\eta\ell)] = \eta(w_a\ell_c + w_c\ell_a)M'(\eta\ell) + w\eta^2\ell_a\ell_c M''(\eta\ell).$$

Thus on the whole probability triangle,

$$|f_{ac}| \le \mathcal{B}, \qquad \mathcal{B} = e^{K\eta_+ M}(6K\eta_+ M + 3K^2\eta_+^2 M^2).$$

The rectangle $[0,a] \times [0,c]$ is inside the triangle and $f(0,0) = 0$. The fundamental theorem of calculus and (10) on its edges give

$$f(a,c) = f(a,0) + f(0,c) + \int_0^a \int_0^c f_{ac}(v,w)\, dw\, dv,$$
$$f(a,0) \le a(1-a)\psi_d((s+1-z)\eta M),$$
$$f(0,c) \le c(1-c)\psi_d((s-z)\eta M).$$

For $\eta_+ \le \eta_c$, take $z = 0$ in the strict case and $z > 0$ at equality. Define the positive layer margins

$$u_s(z) = \min_{\eta \in [\eta_-, \eta_+]} [-\psi_d((s+1-z)\eta M)] = \min_{\eta \in \{\eta_-, \eta_+\}} [-\psi_d((s+1-z)\eta M)].$$

The equality uses concavity of $-\psi_d$. The other edge is nonpositive; using $1 - a \ge p_1$ proves

$$\mathbb{E}_t \Delta h_{A,z} \le h_{A,z}\{-u_s(z)p_1 P_A + \mathcal{B} P_A P_C\}. \tag{11}$$

Set $w_1(z) = 1$ and $w_r(z) = 2\mathcal{B}(r-1)w_{r-1}(z)/u_r(z)$ for $2 \le r \le m$. Let $V_z = \sum_{A \ne \varnothing} w_{|A|}(z) h_{A,z}$. For $i \in A$, $j \notin A$, $j > 1$, the exact identity

$$h_{A,z} p_i p_j = h_{A \cup \{j\},z} p_1 p_i \tag{12}$$

holds because the common deformation cancels. A fixed size-r set and i in that set have exactly $r-1$ predecessors. Their positive coefficient is absorbed by half the current negative coefficient: $\mathcal{B}(r-1)w_{r-1} - u_r w_r = -u_r w_r/2$. Retaining only singleton losses yields

$$\mathbb{E}_t V_z(\theta_{t+1}) \le V_z(\theta_t) - u_1(z) e^{z\theta_{1,t}} \sum_{i>1} p_{i,t}^2, \qquad \mathbb{E} \sum_{t \ge 0} e^{z\theta_{1,t}} \sum_{i>1} p_{i,t}^2 \le \frac{V_z(\theta_0)}{u_1(z)}. \tag{13}$$

Finite-time integrability follows from bounded logit increments. The last inequality follows by finite summation, dropping a nonnegative endpoint, and monotone convergence. At $z = 0$ in the strict case this proves integrable infinite square occupancy and hence $p_1 \to 1$ a.s.

A.3 The critical deformation loses only one factor of $1/z$

At $\eta_+ = \eta_c$, set $z_0 = 1/4$. For $r < m$ define

$$\underline{u}_r = \min_{\eta \in \{\eta_-, \eta_+\},\ z \in \{0, z_0\}} [-\psi_d((r+1-z)\eta M)] > 0.$$

Concavity separately in η, z gives $u_r(z) \ge \underline{u}_r$. At the top layer, concavity between 0 and z_0 gives $u_m(z) \ge c_* z$, where $c_* = u_m(z_0)/z_0 > 0$. For $m \ge 2$, define $\bar{w}_1 = 1$, $\bar{w}_r = 2\mathcal{B}(r-1)\bar{w}_{r-1}/\underline{u}_r$ for $2 \le r < m$, and $\bar{w}_m = 2\mathcal{B}(m-1)\bar{w}_{m-1}/c_*$. Then $w_r \le \bar{w}_r$ for $r < m$ and $w_m \le \bar{w}_m/z$. With $H_r^0 = \sum_{|A|=r} h_{A,0}(\theta_0)$ and $E_0 = e^{z_0(\theta_{1,0})_+}$, take

$$C_0 = E_0 \left(z_0 \sum_{r<m} \bar{w}_r H_r^0 + \bar{w}_m H_m^0 \right), \qquad C_1 = C_0/\underline{u}_1.$$

For $m = 1$ there is no recursion: take $C_0 = z_0 E_0 H_1^0$ and $C_1 = E_0 H_1^0/c_*$. In both cases, for $0 < z \le z_0$,

$$V_z(\theta_0) \le C_0/z, \qquad V_z(\theta_0)/u_1(z) \le C_1/z. \tag{14}$$

Using a common smallest margin at every layer instead would incorrectly introduce multiple powers of $1/z$ and fail to prove the desired logarithm.

Let $\Omega = q/p_1$. Coordinate conservation and AM–GM give

$$\theta_1 \geq S_0/K + (m/K)\log(m/\Omega).$$

Put $c_z = e^{zS_0/K} m^{zm/K}$, $b_z = zm/K$, $a_z = 1 - b_z > 0$. Then $e^{z\theta_1} \geq c_z \Omega^{-b_z}$ and $V_z \geq c_z \Omega^{a_z}$. Stopping the nonnegative supermartingale at a threshold, first at finite time and then passing to the limit, yields

$$\mathbb{P}(\sup_{t\geq 0} \Omega_t \geq H) \leq V_z(\theta_0)/(c_z H^{a_z}).$$

If the supremum is not attained, approximate the threshold from below. On $\Omega_t \leq H$, $q_t^2 \leq mH^{b_z} e^{z\theta_{1,t}} \sum_{i>1} p_{i,t}^2/c_z$. Therefore

$$\mathbb{E}\sum_{t<T} q_t^2 \leq \frac{mH^{b_z} V_z(\theta_0)}{c_z u_1(z)} + \frac{TV_z(\theta_0)}{c_z H^{a_z}}. \tag{15}$$

Choose $U = e^4 + T$, $z = 1/\log U$, and $H = U^2$ only in the analysis. Then $H^{b_z} = e^{2m/K}$, $T/H^{a_z} \leq e^{2m/K}/U$, and c_z has a uniform positive lower bound. Equations (14)–(15) give expected finite-horizon square occupancy $O(\log(e+T))$. For the a.s. conclusion, instead fix one $z > 0$. The maximal inequality gives $\Omega_{\max} < \infty$ a.s.; thus $e^{z\theta_{1,t}}$ has a pathwise positive lower bound. Equation (13) gives unweighted square summability a.s. and $p_1 \to 1$. The reciprocal random lower bound is *not* integrated; this does not prove integrable infinite critical occupancy.

A.4 Closing expected regret

For $\Psi(v) = 2\sum_{n\geq 0} v^n/(n+2)!$, the exponential series gives $\mathbb{E}_t e^{vZ} \leq 1 + v\mathbb{E}_t Z + v^2 M^2 \Psi(\eta_+ M)/2$ for $|v| \leq \eta_+$. Applied to $x_i = e^{-\theta_i}$, baseline cancellation in the linear term, $|\mu_i - \mu\cdot p| \leq 2M$, and the score variance $p_i(1-p_i) \leq p_i$ yield $\mathbb{E}_t x_{i,t+1} \leq x_{i,t} + \mathfrak{a}$, where

$$\mathfrak{a} = \frac{e^{-S_0/K}}{K}\left(2M\eta_+ + \tfrac{1}{2}\eta_+^2 M^2 \Psi(\eta_+ M)\right).$$

Jensen and conservation imply $\mathbb{E}(\theta_{1,T} - \theta_{1,0}) \leq \sum_{i>1} \log(1 + \mathfrak{a} e^{\theta_{i,0}} T)$. After taking expectation, the best-coordinate martingale has mean zero, so this same increment equals $\mathbb{E}\sum_{t<T} \eta_t p_{1,t} r_t$. Finally $r_t \leq p_{1,t} r_t + 2Mq_t^2$, giving

$$R_T \leq \eta_-^{-1} \sum_{i>1} \log(1 + \mathfrak{a} e^{\theta_{i,0}} T) + 2M\mathbb{E}\sum_{t<T} q_t^2. \tag{16}$$

The strict and critical occupation bounds prove the upper implication of Theorem 5 uniformly in the declared class.

A.5 A clock lemma and cumulative allocation

Lemma 6 (Predictable clock)**.** *Let $M_n = \sum_{t<n} D_t$ be a martingale. If predictable $a_t \in [0, a_{\max}]$ satisfies $\mathbb{E}_t D_t^2 \leq C a_t$, and $A_n = \sum_{t<n} a_t$, then M_n converges on $\{A_\infty < \infty\}$ and $M_n/A_n \to 0$ a.s. on $\{A_\infty = \infty\}$.*

Proof. Since A_{t+1} is pre-action measurable and $\sum_t a_t/(1+A_{t+1})^2 \leq \int_0^\infty (1+x)^{-2} dx = 1$, the martingale series $\sum_t D_t/(1+A_{t+1})$ is bounded in L^2 and converges. Pathwise Kronecker's lemma proves the infinite-clock claim. Stopping at the first $A_n > j$ bounds second moments by $C(j + a_{\max})$. For each integer j, martingale convergence and restriction to $\{A_\infty < j\}$ prove the finite-clock claim. □

Lemma 7 (Cumulative soft-min allocation)**.** *Suppose $p_t = \mathrm{softmax}(\theta_t)$, the coordinate sum is constant, and a vector ν has unique maximizer a, with $g_i = \nu_a - \nu_i > 0$. Suppose $0 < v_- \leq v_t \leq v_+$ is predictable and converges a.s. to a deterministic $v_\infty > 0$. Assume*

$$\mathbb{E}_t \Delta\theta_i = v_t p_i(\nu_i - \nu\cdot p), \qquad \mathbb{E}_t(\Delta M_i)^2 \leq C(1-p_a), \qquad p_{a,t} \to 1 \text{ a.s.},$$

where ΔM_i is the centered coordinate increment. Then $\sum_{t<T} p_{i,t}/\log T \to 1/(v_\infty K g_i)$ a.s. for every $i \neq a$. The actual sampling distribution need not equal the reference p.

Proof. Let $m = K-1$, $d_i(T) = g_i \sum_{t<T} v_t p_{i,t}$, $d = \sum_{i \neq a} d_i$, $r_t = \sum_{i \neq a} g_i p_{i,t}$ and $q_t = 1 - p_{a,t}$. Exact compensation gives

$$\theta_{a,T} = \theta_{a,0} + d - H_T + M_{a,T}, \quad \theta_{i,T} = \theta_{i,0} - d_i + H_{i,T} + M_{i,T},$$

where $H_T = \sum v_t q_t r_t$ and $H_{i,T} = \sum v_t p_{i,t} r_t \leq H_T$. The clock increment $\delta d_t = v_t r_t$ is bounded and is at least $v_- g_{\min} q_t$. If $d_\infty < \infty$, Lemma 6 makes every coordinate converge to a finite value, contradicting $p_a \to 1$. Thus $d \to \infty$; weighted Cesaro and the clock lemma give $H = o(d)$, $M_i = o(d)$, and $\theta_a = d + o(d)$, $\theta_i = -d_i + o(d)$.

The allocation weights are exactly

$$w_i = \frac{\delta d_i}{\delta d} = \frac{e^{-d_i + e_i}}{\sum_{j \neq a} e^{-d_j + e_j}}, \quad e_i = \log g_i + \theta_{i,0} + H_i + M_i = o(d).$$

For every $\epsilon > 0$, eventually $|e_i| \leq \epsilon d$. The Gibbs entropy identity, with $u_i = d_i - e_i$, gives $\sum_i w_i u_i \leq \log m + \min_i u_i$, hence $\sum_i w_i d_i \leq d/m + 2\epsilon d + \log m$. For $F = \sum_i (d_i - d/m)^2$, exact expansion yields

$$\delta F \leq 4\epsilon d\, \delta d + 2 \log m\, \delta d + (\delta d)^2.$$

Since $\sum d\, \delta d \leq d(T)^2/2$ and $\sum (\delta d)^2 \leq C d(T)$, we get $\limsup F/d^2 \leq 2\epsilon$. Letting rational $\epsilon \downarrow 0$ proves $d_i/d \to 1/m$. Thus $\theta_a - \theta_i = (K/m)d + o(d)$. With $c_K = K/m$, the softmax ratios and eventual $p_a \geq 1/2$ imply

$$c_\epsilon e^{-(c_K + \epsilon) d(t)} \leq \delta d_t \leq C_\epsilon e^{-(c_K - \epsilon) d(t)}.$$

Applying $e^x - 1 \geq x$ to $e^{(c_K+\epsilon)d}$ gives linear lower growth; applying $e^x - 1 \leq x e^x$ and the bounded clock increments to $e^{(c_K-\epsilon)d}$ gives linear upper growth. Consequently $d(T)/\log T \to m/K$ and $d_i(T)/\log T \to 1/K$. Weighted Cesaro and $v_t \to v_\infty$ complete the conclusion. No expected-regret assumption or L^1 upgrade was used. □

A.6 A VALID ZERO-REWARD EMBEDDING FOR THE LOWER BOUND

The sole external convergence input is Robertson et al. (2025, Theorem 4.2): for standard zero-baseline SGB with finite K, bounded fixed IID arm laws, fixed $\eta > 0$ and finite initial logits, the total probability of the optimal set converges a.s. to 1. In the construction below the optimal arm is unique. We apply that theorem to the full standard process, not directly to a reward-tilted action distribution.

For bounded positive weights $w_i(y)$, put $c_i = \int w_i \, dP_i$, $C = \max_i c_i$ and $W(p) = \sum_i p_i c_i$. Consider the joint kernel

$$\mathcal{Q}_p(i, dy) = p_i w_i(y) P_i(dy)/W(p),$$

with the *original* score $e_i - p$ and fixed scalar. Define genuine IID reward distributions for a second, standard SGB:

$$P_i'(dy) = C^{-1} w_i(y) P_i(dy) + (1 - c_i/C)\delta_0(dy).$$

Mark the first component as accepted and the added zero as rejected. A rejection leaves the logits unchanged; the acceptance probability $W(p)/C$ is uniformly bounded away from zero. Summing the geometric waiting law shows that the accepted update subsequence has exactly kernel $\mathcal{Q}$. The argument also distinguishes genuine zero rewards using the latent mark. There are infinitely many accepted updates a.s. If $\nu_i = \int y w_i(y) P_i(dy)$ has unique maximizer a, the imported convergence theorem applies to P_i' (whose means are ν_i/C), and then along the accepted subsequence gives $p_a \to 1$ under $\mathcal{Q}$. Direct calculation gives

$$\mathbb{E}_{\mathcal{Q},t} \Delta\theta_i = \frac{\eta}{W(p_t)} p_i(\nu_i - \nu \cdot p), \qquad \mathbb{E}_{\mathcal{Q},t} (\Delta M_i)^2 \leq C_V p_i(1 - p_i) \leq C_V(1 - p_a).$$

If $c_a = 1$, the effective scalar $v_t = \eta/W(p_t)$ converges to η, lies in a fixed positive interval, and Lemma 7 applies.

A.7 NORMALIZED LIKELIHOOD COST AND THE EXPLICIT HARD INSTANCE

Suppose arm 1 is uniquely optimal under the original laws, arm 2 is strictly better than arms $j \geq 3$, and some $\lambda > 0$ satisfies

$$\mathbb{E}e^{-\lambda Y^{(1)}} - 1 + \lambda\mu_2 = 0, \qquad g := \mu_2 - \mathbb{E}[Y^{(1)}e^{-\lambda Y^{(1)}}] > 0. \tag{17}$$

Take $w_1(y) = e^{-\lambda y}$ and $w_j = 1$ for $j \geq 2$. The tilted means have unique best arm 2 and $c_2 = 1$; the preceding embedding and allocation lemma give $\sum_{t<T} p_{1,t}/\log T \to 1/(\eta K g)$, $\mathbb{Q}$-a.s. The finite-history log density ratio is exactly

$$\mathcal{L}_T = \log \frac{d\mathbb{Q}^T}{d\mathbb{P}^T} = -\lambda \sum_{t<T} Y_t \mathbf{1}_{\{I_t=1\}} - \sum_{t<T} \log W(p_t), \qquad W(p) = 1 + (c_1 - 1)p_1.$$

Here $c_1 - 1 = -\lambda\mu_2$. Expanding only in the eventual region $p_1 \to 0$ gives conditional mean increment $\lambda g p_1 + O(p_1^2)$, with uniform remainder there. Weighted Cesaro gives $\sum p_1^2 = o(\log T)$. The centered increment variance is $O(p_1)$ since only action 1 contributes reward randomness and rewards/weights are bounded. Lemma 6 therefore implies $\mathcal{L}_T/\log T \to \lambda/(\eta K)$ a.s. The normalization term has not been discarded. All horizons here count transitions of the tilted chain, not the extra calendar steps of its embedding.

The actual tilted action-2 probability is $p_2/W \to 1$; the bounded-increment martingale law of large numbers gives $N_2(T)/T \to 1$. For $\beta \in (\lambda/(\eta K), 1)$, the event $E_T = \{N_2(T) \geq T/2, \mathcal{L}_T \leq \beta \log T\}$ has $\mathbb{Q}(E_T) \to 1$. At sufficiently large T,

$$\mathbb{P}(E_T) = \mathbb{E}_{\mathbb{Q}}[e^{-\mathcal{L}_T}\mathbf{1}_{E_T}] \geq \tfrac{1}{2}T^{-\beta}, \qquad R_T \geq \tfrac{1}{4}\Delta_2 T^{1-\beta}.$$

Only fixed finite-horizon density identities are used.

To instantiate this bound, fix $\eta > \eta_c$ and $0 < \gamma < 1 - \eta_c/\eta$. Let

$$s_\gamma = K\eta M(1-\gamma) > s_*, \qquad e = \frac{1}{2}\min\left\{1, \frac{\psi_d(s_\gamma)}{\sinh s_\gamma - s_\gamma}\right\}, \qquad \varepsilon = eM.$$

Set $\mathbb{P}(Y^{(1)} = M) = (d+e)/2$, $\mathbb{P}(Y^{(1)} = -M) = 1 - (d+e)/2$, $Y^{(2)} = -M + \varepsilon$, and $Y^{(j)} = -M$ for $j \geq 3$, with baseline 0. The smallest gap is exactly Δ, and all rewards lie in $[-M, M]$. The strictly convex function $f_e(s) = \psi_d(s) + e(s - \sinh s)$ has derivative $-d$ at zero, diverges at infinity, and satisfies $f_e(s_\gamma) \geq \psi_d(s_\gamma)/2 > 0$. Its positive root $s_e < s_\gamma$ has $f_e'(s_e) > 0$. With $\lambda = s_e/M$, these are exactly (17), and $\lambda/(\eta K) < 1 - \gamma$. Choose $\beta = [\lambda/(\eta K) + 1 - \gamma]/2$. The preceding lower bound gives $R_T \geq (\Delta/4)T^{1-\beta} \geq (\Delta/4)T^{\gamma}$ eventually. The positive ε ensures that the auxiliary best arm 2 is unique; it cannot be deleted when $K > 2$. Taking the admissible constant rule $\eta_t = \eta_+$ proves the lower implication of (9).

A.8 SCOPE AND A PATHWISE COEFFICIENT COROLLARY

For fixed $0 < \eta \leq \eta_c$, the established $p_1 \to 1$, the original mean drift and bounded residual variance verify Lemma 7 with $v_t = \eta$, $\nu = \mu$. Writing $Q_i(T) = \sum_{t<T} p_{i,t}$ and applying Lemma 6 to actual counts gives

$$\frac{Q_i(T)}{\log T}, \frac{N_i(T)}{\log T} \to \frac{1}{\eta K \Delta_i} \quad \text{a.s.}, \qquad \frac{\widehat{R}_T}{\log T} \to \frac{K-1}{\eta K} \quad \text{a.s.}$$

At the critical point these are not automatically L^1 statements. Also $s_*(d) = 2d + O(d^2)$ as $d \downarrow 0$, so the safety scale is $2\Delta/(KM^2)(1 + O(\Delta/M))$. If raw rewards satisfy $|Y| \leq R$ and $|b_t| \leq B$, substituting $M = R + B$ is a valid sufficient embedding when $\Delta \leq R + B$; for $B > 0$ it is not an equality of instance classes and does not establish necessity for the smaller separately constrained class. A simultaneous reward/baseline translation preserves residuals, updates and regret.

B CUMULATIVE ALLOCATION UNDER RAW AND RESIDUAL BUDGETS

This appendix first proves a raw-budget allocation theorem, then transfers its expectation argument to the strict residual-safe region. For the raw-budget result, rewards are independent pull-indexed

IID with $|Y^{(i)}| \le R$, $|b_t| \le B$, unique best arm 1, finite deterministic θ_0, and a fixed scalar $\eta > 0$. Write $M = R + B$, $\Delta_{\min} = \min_{i>1} \Delta_i$, $\Delta_{\max} = \max_{i>1} \Delta_i$, and

$$\Psi(x) = 2 \sum_{n \ge 0} \frac{x^n}{(n+2)!}, \quad \Gamma_j = M^2 \Psi(j\eta M), \quad c = \Delta_{\min} - \tfrac{1}{2} K \eta \Gamma_K > 0.$$

Here R is a reward budget; R_T remains expected pseudo-regret. Constants in the raw-budget proof may depend on K, R, B, η, the arm means and θ_0, but not on T or the predictable baseline rule. Set $Q_i(T) = \sum_{t<T} p_{i,t}$ and $\mathcal{E}_i(T) = \eta \sum_{t<T} p_{1,t} p_{i,t}$.

Theorem 8 (Cumulative allocation and expectation)**.** *Under the preceding assumptions, for every $i > 1$,*

$$\frac{Q_i(T)}{\log T}, \ \frac{N_i(T)}{\log T} \longrightarrow \frac{1}{\eta K \Delta_i}, \tag{18}$$

$$\frac{\mathcal{E}_i(T)}{\log T} \longrightarrow \frac{1}{K \Delta_i}, \qquad \frac{\widehat{R}_T}{\log T} \longrightarrow \frac{K-1}{\eta K}. \tag{19}$$

All these limits hold almost surely and in L^1; consequently $R_T / \log T \to (K-1)/(\eta K)$. Almost surely, $\theta_{1,T} / \log T \to (K-1)/K$, $\theta_{i,T} / \log T \to -1/K$, and $\log p_{i,T} / \log T \to -1$. With the ratio defined as zero when the denominator is zero,

$$\frac{N_i(T)}{\sum_{j>1} N_j(T)} \longrightarrow \frac{\Delta_i^{-1}}{\sum_{j>1} \Delta_j^{-1}} \quad \textit{almost surely and in } L^1.$$

To prove the theorem, it is enough to establish the pathwise clock coefficients and uniform integrability, then bound the count-compensator and exposure differences. The first two subsections give the integrable square-occupancy and moment bounds needed for these steps; the last two establish the clock limits and justify passage to expectation.

B.1 An integrable square-occupancy bound

For $J(p) = \mathrm{diag}(p) - pp^\top$ and a deterministic vector v whose coordinate oscillation is at most j, conditional Taylor expansion gives

$$\mathbb{E}_t e^{v^\top \theta_{t+1}} - e^{v^\top \theta_t} \le e^{v^\top \theta_t} \left\{ \eta v^\top J(p_t) \mu + \frac{\eta^2 \Gamma_j}{2} v^\top J(p_t) v \right\}. \tag{20}$$

Indeed the scalar increment is $u = \eta (Y_t - b_t) v^\top (e_{I_t} - p_t)$, $|u| \le j\eta M$, and $e^u \le 1 + u + \frac{1}{2} u^2 \Psi(j\eta M)$ follows term by term from the power series. Conditional centering cancels the common baseline in the first moment, and $\mathbb{E}_t u^2 \le \eta^2 M^2 v^\top J(p_t) v$.

Let $I = \{2, \dots, K\}$, $b = \Delta_{\max} + \eta \Gamma_K / 2$, $C = b/c$, $h_A = \exp(\sum_{i \in A} (\theta_i - \theta_1))$, and $V = \sum_{\varnothing \ne A \subseteq I} C^{|A|-1} h_A$. Apply (20) with $v_1 = -s$, $v_i = 1$ for $i \in A$, and $v_i = 0$ otherwise, where $s = |A|$. The pairwise formulas $v^\top J \mu = \sum_{i<j} p_i p_j (v_i - v_j)(\mu_i - \mu_j)$ and $v^\top J v = \sum_{i<j} p_i p_j (v_i - v_j)^2$ give

$$\mathbb{E}_t \Delta h_A \le \eta h_A \left[-c p_1 \left((s+1) \sum_{i \in A} p_i + s \sum_{j \in I \setminus A} p_j \right) + b \sum_{i \in A, j \in I \setminus A} p_i p_j \right].$$

For each D of size $r \ge 2$, the identity $h_{D \setminus \{j\}} p_i p_j = h_D p_1 p_i$ makes the total positive coefficient $bC^{r-2}(r-1)$, dominated by $cC^{r-1}(r+1)$ from D itself. The unused negative terms can be discarded. The singleton terms are $-2c\eta h_{\{i\}} p_1 p_i = -2c\eta p_i^2$. Thus, by telescoping nonnegative V and then monotone convergence,

$$\mathbb{E} \sum_{t \ge 0} q_t^2 \le (K-1) \mathbb{E} \sum_{t \ge 0} \sum_{i > 1} p_{i,t}^2 \le \frac{(K-1) V(\theta_0)}{2c\eta} < \infty. \tag{21}$$

In particular $q_t \to 0$ almost surely.

B.2 A LOGARITHMIC EXPECTATION BOUND WITHOUT AN ONSET TIME

The sum of logits is the constant $S_0 = \sum_i \theta_{i,0}$. Using (20) with $v = -e_i$, the drift and variance terms are bounded above by $p_i\Delta_{\max}$ and p_i, respectively. Since $e^{-\theta_i}p_i = (\sum_j e^{\theta_j})^{-1} \le e^{-S_0/K}/K$,

$$\mathbb{E}_t e^{-\theta_{i,t+1}} - e^{-\theta_{i,t}} \le A_0, \qquad A_0 = \frac{e^{-S_0/K}}{K}(\eta\Delta_{\max} + \eta^2\Gamma_1/2).$$

Consequently $\mathbb{E}e^{-\theta_{i,T}} \le e^{-\theta_{i,0}} + A_0T$ and Jensen gives $\mathbb{E}(\theta_{i,0} - \theta_{i,T}) \le \log(1 + A_0e^{\theta_{i,0}}T)$. Writing $r_t = \sum_{i>1} \Delta_i p_{i,t}$, we have $\mathbb{E}_t\Delta\theta_1 = \eta p_1 r$ and $r \le p_1 r + \Delta_{\max}q^2$. Hence

$$R_T \le \eta^{-1}\sum_{i>1}\log(1 + A_0e^{\theta_{i,0}}T) + \Delta_{\max}\mathbb{E}\sum_{t\ge0} q_t^2 = O(\log(e + T)). \tag{22}$$

This also bounds $\mathbb{E}\sum_{t<T} q_t$ by $O(\log(e + T))$.

B.3 CLOCK BALANCING AND THE PATHWISE LIMIT

We use the clock martingale Lemma 6, whose finite-clock and infinite-clock alternatives require only a predictable bounded clock increment and a conditional variance bounded by a constant times that increment. Define, only in this proof,

$$a_i(T) = \eta\Delta_i Q_i(T), \quad a(T) = \sum_{i>1} a_i(T), \quad H(T) = \eta\sum_{t<T} q_t r_t, \quad H_i(T) = \eta\sum_{t<T} p_{i,t}r_t.$$

Explicitly, define the centered coordinate martingale by

$$\mathcal{M}_j(T) = \sum_{t<T}\left[(\theta_{j,t+1} - \theta_{j,t}) - \mathbb{E}_t(\theta_{j,t+1} - \theta_{j,t})\right].$$

Here $\mathbb{E}_t$ is conditional expectation given $\mathcal{F}_t$; $\Delta Z_t = Z_{t+1} - Z_t$ denotes a one-step increment when applied to a process, whereas Δ_i denotes a fixed mean gap. The compensator identities are

$$\theta_{1,T} = \theta_{1,0} + a(T) - H(T) + \mathcal{M}_1(T),$$
$$\theta_{i,T} = \theta_{i,0} - a_i(T) + H_i(T) + \mathcal{M}_i(T).$$

All H, H_i are bounded by the integrable random variable $\eta\Delta_{\max}\sum_{t\ge0} q_t^2$. The centered logit martingales satisfy

$$\mathbb{E}_t(\Delta\mathcal{M}_j)^2 \le \eta^2M^2p_j(1 - p_j) \le \eta^2M^2q_t \le \frac{\eta M^2}{\Delta_{\min}}\Delta a_t.$$

If $a(\infty) < \infty$, the finite-clock alternative makes every logit converge to a finite value, contradicting $p_1 \to 1$. Thus $a \to \infty$ and the infinite-clock alternative gives $\mathcal{M}_j/a \to 0$ almost surely.

For completeness, the deterministic balancing argument is explicit. Suppose m nondecreasing clocks satisfy $a = \sum a_i \to \infty$, bounded increments $d_t = \Delta a_t > 0$, and

$$\frac{\Delta a_{i,t}}{d_t} = \frac{\exp(-a_i(t) + e_i(t))}{\sum_j \exp(-a_j(t) + e_j(t))}, \qquad \max_i |e_i(t)| = o(a(t)).$$

For each $\varepsilon > 0$, eventually $|e_i| \le \varepsilon a$. The entropy identity for these weights w_i implies $\sum_i w_i a_i \le \min_i a_i + 2\varepsilon a + \log m \le a/m + 2\varepsilon a + \log m$. For $F = \sum_i (a_i - a/m)^2$, expansion gives

$$\Delta F \le 4\varepsilon a d_t + 2(\log m)d_t + d_t^2.$$

Summing, using $\sum a d_t \le a(T)^2/2$ and $\sum d_t^2 \le (\sup d_t)a(T)$, yields $\limsup F/a(T)^2 \le 2\varepsilon$. Intersecting the probability-one events for positive rational ε proves $a_i/a \to 1/m$.

Apply this with $m = K - 1$ and $e_i = \log\Delta_i + \theta_{i,0} + H_i + \mathcal{M}_i$; the softmax denominator cancels from $\Delta a_i/\Delta a$. Thus $\theta_1 - \theta_i = [K/(K-1)]a + o(a)$. Put $c_K = K/(K-1)$. For every sufficiently small $\varepsilon > 0$, eventually

$$c_\varepsilon e^{-(c_K+\varepsilon)a(t)} \le \Delta a_t \le C_\varepsilon e^{-(c_K-\varepsilon)a(t)},$$

because $p_1 \geq 1/2$ and $\Delta a = \eta \sum_{i>1} \Delta_i p_i$. The inequalities $e^u - 1 \geq u$ and $e^u - 1 \leq u e^u$, together with bounded Δa, make $e^{(c_K+\varepsilon)a(t)}$ grow at least linearly and $e^{(c_K-\varepsilon)a(t)}$ at most linearly. Dividing their logarithms by $\log T$ and then sending $\varepsilon \downarrow 0$ gives

$$\frac{a(T)}{\log T} \to \frac{K-1}{K}, \qquad \frac{a_i(T)}{\log T} \to \frac{1}{K} \quad \text{almost surely.} \tag{23}$$

The compensator identities now imply all asserted logit and logarithmic probability limits and the almost-sure Q_i limits.

B.4 Uniform integrability and passage to expectation

Let $X_i(T) = (-\theta_{i,T})_+$. The inverse-logit estimate gives

$$\mathbb{P}(X_i(T) > x) \leq \min\{1, C_i(1+T)e^{-x}\} \quad (x \geq 0)$$

for a fixed $C_i > 0$. With $u_T = \log \max\{1, C_i(1+T)\}$, tail integration gives $\mathbb{E}X_i(T)^2 \leq u_T^2 + 2u_T + 2 = O(\log^2(e+T))$. Moreover $(\theta_{1,T})_+ \leq |S_0| + \sum_{i>1} X_i(T)$. Put $U(T) = \eta \sum_{t<T} p_{1,t} r_t = a(T) - H(T) \geq 0$. The first compensator identity bounds it by $(\theta_{1,T})_+ + |\theta_{1,0}| + |\mathcal{M}_1(T)|$, while orthogonality and (22) give $\mathbb{E}\mathcal{M}_1(T)^2 \leq \eta^2 M^2 \mathbb{E}\sum_{t<T} q_t = O(\log(e+T))$. It follows that $\sup_{T\geq 2} \mathbb{E}[U(T)/\log T]^2 < \infty$. Thus $U(T)/\log T$ is uniformly integrable and converges in L^1 to $(K-1)/K$ by (23) and $H(\infty) < \infty$ a.s. Also $\mathbb{E}H(T)/\log T \to 0$ by (21), so $a(T)/\log T$ converges in L^1 and is uniformly integrable. The domination $0 \leq a_i \leq a$ and the a.s. limits give L^1 convergence for every a_i. No integrable random convergence-onset time, nor a second moment of $H(\infty)$, has been assumed.

Finally $J_i(T) = N_i(T) - Q_i(T)$ has centered increments with conditional variance at most $p_{i,t}$. Since $Q_i \to \infty$, Lemma 6 gives $J_i/Q_i \to 0$ almost surely. Cauchy–Schwarz and (22) give $\mathbb{E}|J_i(T)| \leq \sqrt{\mathbb{E}Q_i(T)} = O(\sqrt{\log(e+T)})$, so its normalized error also vanishes in L^1. Likewise

$$0 \leq \eta Q_i(T) - \mathcal{E}_i(T) = \eta \sum_{t<T} q_t p_{i,t} \leq \eta \sum_{t\geq 0} q_t^2$$

vanishes after division by $\log T$, both a.s. and in L^1. Finite sums prove the regret conclusion. The tail-count ratio has the claimed a.s. limit and lies in $[0,1]$, so bounded convergence proves its L^1 limit. This completes the theorem.

Why cumulative allocation is the stated object. The theorem does not assert $tp_{i,t} \to 1/(\eta K \Delta_i)$. For example, take $K = 3$, deterministic rewards $(1, 0.6, 0.2)$, baseline zero and $\eta = 10^{-3}$. The strict condition holds because $\Psi(0.003) \leq e^{0.003}$ and $c \geq 0.4 - 0.0015e^{0.003} > 0$. Every tail arm is pulled infinitely often by (18). Along its pull times, direct substitution in the score update and $p_1 \to 1$ gives $p_{i,t+1}/p_{i,t} \to e^{2\eta\mu_i} > 1$. A finite positive limit of $tp_{i,t}$ would force this ratio to tend to one. A limit of zero or infinity would force $Q_i(T)/\log T$ to tend to zero or infinity by comparison with the harmonic series, also a contradiction. Thus even an extended nonnegative pointwise limit need not exist. The cumulative and expectation conclusions require no such limit.

B.5 Allocation throughout the strict residual-safe region

Corollary 9 (Strict residual-safe allocation)**.** *Fix $K \geq 2$, $M > 0$, $0 < \Delta \leq M$, a finite deterministic initialization, and an instance in $\mathcal{C}_K(\Delta, M)$ with an admissible common predictable baseline. For a constant $0 < \eta < \eta_c = s_*(\Delta/M)/(KM)$, the count and exposure limits in* (18)–(19) *hold almost surely and in L^1. Moreover $\widehat{R}_T/\log T \to (K-1)/(K\eta)$ in both modes, and $R_T/\log T \to (K-1)/(K\eta)$. All parameters are fixed as $T \to \infty$; no uniformity as $\eta \uparrow \eta_c$ is asserted.*

Proof. It suffices to establish uniform integrability of the normalized gap-weighted clocks and then transfer their limits to counts and exposure. The strict part of Appendix A, its inverse-logit estimate, and its pathwise coefficient corollary give

$$\mathbb{E}\sum_{t\geq 0} q_t^2 < \infty, \qquad R_T = O(\log(e+T)), \qquad \mathbb{E}e^{-\theta_{i,T}} \leq C_i(1+T), \qquad \frac{a_i(T)}{\log T} \to \frac{1}{K} \quad \text{a.s.},$$

where $a_i(T) = \eta\Delta_i Q_i(T)$ and $a(T) = \sum_{i>1} a_i(T)$. Constants below are finite for the fixed model parameters and fixed $\eta < \eta_c$; they need not remain bounded at the endpoint.

The common residual bound implies $|\mu_i - b_t| \le M$, hence $\Delta_i \le 2M$ and $r_t \le 2Mq_t$. Set

$$\mathcal{D}_T = \eta \sum_{t<T} q_t r_t, \qquad \mathcal{U}_T = a(T) - \mathcal{D}_T = \eta \sum_{t<T} p_{1,t} r_t.$$

Then $0 \le \mathcal{D}_T \le \mathcal{D}_\infty \le 2\eta M \sum_{t\ge 0} q_t^2$ and $\mathbb{E}\mathcal{D}_\infty < \infty$. Consequently $\mathcal{D}_T/\log T \to 0$ a.s. and in L^1, the latter by domination by $\mathcal{D}_\infty/\log 2$ for integer $T \ge 2$.

Markov's inequality for the inverse-logit estimate gives

$$\mathbb{P}((-\theta_{i,T})_+ > u) \le \min\{1, C_i(1+T)e^{-u}\}.$$

Writing $v_{i,T} = \log\max\{1, C_i(1+T)\}$, tail integration bounds $\mathbb{E}[(-\theta_{i,T})_+^2]$ by $v_{i,T}^2 + 2v_{i,T} + 2$. Since $S_0 = \sum_i \theta_{i,0}$ is conserved, $(\theta_{1,T})_+ \le |S_0| + \sum_{i>1}(-\theta_{i,T})_+$, so its second moment is $O(\log^2(e+T))$.

The conditional score drift and its centered martingale give $\theta_{1,T} = \theta_{1,0} + \mathcal{U}_T + \mathcal{M}_{1,T}$, with

$$\mathbb{E}\mathcal{M}_{1,T}^2 \le \eta^2 M^2 \mathbb{E}\sum_{t<T} p_{1,t} q_t \le \frac{\eta^2 M^2}{\Delta} R_T = O(\log(e+T)).$$

Here bounded residual increments justify orthogonality and finite second moments, and $r_t \ge \Delta q_t$ gives the last inequality. The bound $0 \le \mathcal{U}_T \le (\theta_{1,T})_+ + |\theta_{1,0}| + |\mathcal{M}_{1,T}|$ now proves

$$\sup_{T\ge 2} \mathbb{E}\left[\left(\frac{\mathcal{U}_T}{\log T}\right)^2\right] < \infty.$$

Thus $\mathcal{U}_T/\log T$ is uniformly integrable. Its a.s. limit is $(K-1)/K$, so the uniform-integrability convergence theorem gives its L^1 limit. Adding $\mathcal{D}_T/\log T$ proves L^1 convergence and uniform integrability of $a(T)/\log T$. Domination $0 \le a_i \le a$, together with the individual a.s. limits, proves the L^1 clock limits.

Finally $N_i - Q_i$ is a centered counting martingale with conditional variance bounded by $p_{i,t}$. Since $\mathbb{E}Q_i \le R_T/\Delta_i = O(\log(e+T))$, $\mathbb{E}|N_i - Q_i| \le \sqrt{\mathbb{E}Q_i} = O(\sqrt{\log(e+T)})$. The a.s. count transfer follows from Lemma 6, since $Q_i/\log T \to 1/(K\eta\Delta_i) > 0$. The remaining exposure error obeys

$$0 \le \eta Q_i(T) - \mathcal{E}_i(T) = \eta \sum_{t<T} q_t p_{i,t} \le \eta \sum_{t\ge 0} q_t^2.$$

Its normalization vanishes a.s. and in L^1 by the same integrable domination. Finite summation over the arms proves the regret conclusions. □

This result uses the strict residual-safe occupancy estimate in place of the sufficient condition $c > 0$. The latter remains useful on its own raw reward/baseline domain, including gaps larger than $M = R + B$. At the critical endpoint the signed two-arm family in Appendix D exhibits a positive expectation defect, so the strict inequality above is essential.

C Separate reward and baseline budgets for two arms

Fix $R > 0$, $B \ge 0$, $0 < \Delta \le 2R$, finite deterministic initialization, and $0 < \eta_- \le \eta_+ < \infty$. Consider two independent IID reward laws supported on $[-R, R]$, with mean gap $\Delta_\nu = \mu_1 - \mu_2 \ge \Delta$, and all predictable common baselines in $[-B, B]$ and rates in $[\eta_-, \eta_+]$. Define, with $\lambda = 2\eta$,

$$\begin{aligned} J_-(\eta, b) &= e^{\lambda b}[\cosh(\lambda R) + (1 - \Delta/R)\sinh(\lambda R)] - 1 - \lambda(R + b), \\ J_+(\eta, b) &= e^{-\lambda(R-b)} - 1 + \lambda(R - \Delta - b), \\ \Phi_{R,B,\Delta}(\eta) &= \max_{b\in\{-B,B\}} \max\{J_-(\eta, b), J_+(\eta, b)\}. \end{aligned} \tag{24}$$

Repeated corners when $B = 0$ cause no change. Here R is the raw reward budget, while R_T is expected pseudo-regret.

Theorem 10 (Separate-budget criterion)**.** *The preceding class admits a uniform bound $R_T \le C\log(e+T)$ if and only if $\Phi_{R,B,\Delta}(\eta_+) \le 0$. The constant depends only on the displayed budgets, rate interval and initialization. Under strict inequality, $\mathbb{E}\sum_{t\ge 0} q_t^2 < \infty$ uniformly. Under equality, $\mathbb{E}\sum_{t<T} q_t^2 = O(\log(e+T))$ uniformly. If the criterion is positive, one fixed-baseline instance within these same budgets, at rate η_+, has $R_T \ge cT^\gamma$ eventually for some $c, \gamma > 0$.*

For sufficiency, we first compute the attained corner supremum and then turn its sign into a global drift bound, including the equality case. The log-odds identity converts the resulting occupation estimate to regret. For necessity, a violating corner supplies an original instance within the same budgets, to which the normalized likelihood lower bound applies.

C.1 The exact corner maximization

For fixed laws, baseline b and rate η, let

$$\phi_{\eta,b}(r) = \mathbb{E}e^{-2r\eta(Y^{(1)}-b)} - 1 + 2r\eta(\mu_2 - b).$$

Convexity of the exponential in the reward bounds its expectation by the chord joining the values at $-R$ and R:

$$\mathbb{E}e^{-\lambda Y^{(1)}} \le \cosh(\lambda R) - \frac{\mu_1}{R}\sinh(\lambda R).$$

Equality is attained by a two-point law with that mean. The resulting expression for $\phi(1)$ is affine in (μ_1, μ_2); it increases with μ_2 and decreases with μ_1. Its maximum over $-R \le \mu_2 \le \mu_1 - \Delta \le R - \Delta$ lies on $\mu_2 = \mu_1 - \Delta$, and then at an endpoint of this line segment. The two endpoint means are $(-R+\Delta, -R)$ and $(R, R-\Delta)$, giving J_- and J_+, respectively. Both have realizations inside the raw reward budget. Convexity in b gives the maximum at $b = \pm B$. Thus (24) is an attained supremum of $\phi_{\eta,b}(1)$.

Assume for now that $\Phi(\eta_+) \le 0$ and set $M_0 = R + B$. For fixed laws and b, convexity in η and the value zero at $\eta = 0$ imply, for every $\eta \in [\eta_-, \eta_+]$,

$$\phi_{\eta,b}(1) \le (\eta/\eta_+)\phi_{\eta_+,b}(1) \le (\eta/\eta_+)\Phi(\eta_+) \le 0. \tag{25}$$

All bounds below hold conditionally for the chosen predictable pair (η_t, b_t), so they remain valid for arbitrary admissible controllers.

C.2 From the corner to a global potential

Write $X = e^{\theta_2 - \theta_1}$, $p = (1+X)^{-1}$ and $q = X/(1+X)$. For $0 < r \le 1$, the exact conditional drift is

$$\begin{aligned}\mathbb{E}_t\Delta X^r &= X^r pq F_r(p),\\ F_r(p) &= \frac{\mathbb{E}e^{-2r\eta q(Y^{(1)}-b)} - 1}{q} + \frac{\mathbb{E}e^{2r\eta p(Y^{(2)}-b)} - 1}{p}.\end{aligned} \tag{26}$$

The endpoint values of these quotients are defined by continuity; $F_r(0) = \phi_{\eta,b}(r)$. To bound derivatives, for bounded Z use $(\mathbb{E}e^{auZ} - 1)/u = \mathbb{E}[aZ\int_0^1 e^{auZs}\,ds]$. Differentiation under this bounded integral shows that the absolute derivative of either term in (26) is at most $(2\eta_+ M_0)^2 e^{2\eta_+ M_0}/2$. Therefore

$$|F_r(p) - \phi_{\eta,b}(r)| \le C_F p, \qquad C_F = (2\eta_+ M_0)^2 e^{2\eta_+ M_0}, \quad 0 < r \le 1. \tag{27}$$

Choose

$$r_0 = \min\left\{\frac{1}{4}, \frac{\Delta}{4\eta_+ M_0^2 e^{2\eta_+ M_0}}\right\}, \qquad c_0 = r_0\eta_-\Delta > 0.$$

Taylor expansion with the nonnegative remainder bound $\Psi(u) \le e^u$ for $u \ge 0$ gives, at every p,

$$F_{r_0}(p) \le -2r_0\eta[\Delta - r_0\eta M_0^2\Psi(2r_0\eta M_0)] \le -c_0.$$

Indeed the linear terms sum to $-2r_0\eta\Delta_\nu$. The quadratic terms are bounded by

$$2r_0^2\eta^2 M_0^2\Psi(2r_0\eta M_0)(p+q) = 2r_0^2\eta^2 M_0^2\Psi(2r_0\eta M_0).$$

This also proves $\phi_{\eta,b}(r_0) \le -c_0$. For $r_0 < r \le 1$, $pX^{r-r_0} \le 1$. Combining this inequality with (27) and the negative drift for X^{r_0} gives

$$V_r(X) = X^r + \frac{C_F}{c_0} X^{r_0}, \qquad \mathbb{E}_t \Delta V_r \le \phi_{\eta,b}(r) X^r pq. \tag{28}$$

This is a global estimate, including p near one.

C.3 The closed endpoint and expected regret

Convexity in r, (25), and $\phi(r_0) \le -c_0$ imply for $r \in (r_0, 1)$ that

$$\phi(r) \le -\frac{c_0(1-r)}{1-r_0}, \qquad \mathbb{E} \sum_{t<T} X_t^r p_t q_t \le \frac{(1-r_0)V_r(x_0)}{c_0(1-r)}, \quad x_0 = X_0.$$

At $r = 1$, $V_1(X_t)$ is a nonnegative supermartingale dominating X_t. Optional stopping at the first crossing before a finite horizon and then monotone limits give $\mathbb{P}(\sup_{t\ge 0} X_t \ge H) \le V_1(x_0)/H$. Since $q^2 = Xpq$, splitting at H yields

$$\mathbb{E} \sum_{t<T} q_t^2 \le H^{1-r} \frac{(1-r_0)V_r(x_0)}{c_0(1-r)} + T \frac{V_1(x_0)}{H}.$$

Take $U = e^2 + T$, $r = 1 - 1/\log U \ge 1/2 > r_0$, and $H = U^2$. Then $H^{1-r} = e^2$, and $V_r(x_0)$ is bounded uniformly in this range. This proves $\mathbb{E}\sum_{t<T} q_t^2 = O(\log(e+T))$. If $\Phi(\eta_+) < 0$, (25) gives the uniform negative margin $-(\eta_-/\eta_+)\Phi(\eta_+) > 0$ at $r = 1$; telescoping (28) then gives an integrable infinite square occupancy.

It remains to control the logarithmic logit term. Let $\Lambda = 1/X$ and $L = \log \Lambda$. Conditional on action 1 its multiplicative increment is $e^{2\eta q(Y^{(1)}-b)}$; on action 2 it is $e^{-2\eta p(Y^{(2)}-b)}$. Using $|e^u - 1| \le |u|e^{|u|}$ and $\Lambda pq = p^2 \le 1$ gives $\mathbb{E}_t|\Delta\Lambda| \le C_\Lambda = 4\eta_+ M_0 e^{2\eta_+ M_0}$. Jensen thus yields $\mathbb{E}(L_T - L_0) \le \log(1 + C_\Lambda x_0 T)$. On the other hand $\mathbb{E}_t \Delta L = 2\eta_t \Delta_\nu p_t q_t$ exactly. As $q = pq + q^2$ and $\Delta_\nu \le 2R$,

$$R_T \le \frac{\mathbb{E}(L_T - L_0)}{2\eta_-} + 2R\mathbb{E} \sum_{t<T} q_t^2 = O(\log(e+T)).$$

Every displayed constant is uniform over the stated class.

C.4 Necessity inside the same raw-budget class

If $\Phi(\eta_+) > 0$, choose a maximizing endpoint baseline and a maximizing pair of reward laws constructed in the corner calculation. Keep that baseline fixed and set $Z_i = Y^{(i)} - b$, with means m_i. Their gap is exactly Δ, and the algorithm is precisely a zero-baseline SGB with rewards Z_i. The convex function

$$f(\lambda) = \mathbb{E}e^{-\lambda Z_1} - 1 + \lambda m_2$$

has $f(0) = 0$, $f'(0) = -\Delta < 0$, and $f(2\eta_+) > 0$. It is strictly convex unless $Z_1 = 0$ a.s.; in that exceptional case it is linear and cannot have these signs. Thus there is a unique root $\lambda_* \in (0, 2\eta_+)$ with $f'(\lambda_*) > 0$. These are exactly the two conditions in (17), applied to Z_i: $g = m_2 - \mathbb{E}[Z_1 e^{-\lambda_* Z_1}] > 0$.

For clarity, every interface of that lower-bound argument is valid here. There are only two arms, so the auxiliary best arm 2 is unique without additional perturbations. The shifted laws are bounded, initialization is finite, the rate is the fixed η_+, and the exponential weights are positive and bounded. The zero-reward embedding of Appendix A supplies genuine IID auxiliary laws and permits the imported convergence theorem. Its normalized likelihood calculation gives $\mathcal{L}_T/\log T \to \lambda_*/(2\eta_+) < 1$ under that auxiliary law and $N_2(T)/T \to 1$. For any $\beta \in (\lambda_*/(2\eta_+), 1)$, the same finite-history change of measure gives $R_T \ge (\Delta/4)T^{1-\beta}$ eventually. The auxiliary laws are used only for a proof change of measure; the bad *original* instance is the previously constructed law on $[-R, R]$ with $b \in \{-B, B\}$. It therefore obeys exactly the budgets being classified.

Fixed-law version. The same argument, without maximizing over reward laws, proves that for any fixed bounded two-arm IID laws with positive gap, finite deterministic initialization, a finite baseline interval $[b_-, b_+]$, and a fixed rate interval $0 < \eta_- \le \eta_+ < \infty$, the exact criterion for a bound uniform over the predictable common controllers is

$$\max_{b\in\{b_-,b_+\}} \{\mathbb{E}e^{-2\eta_+(Y^{(1)}-b)} - 1 + 2\eta_+(\mu_2 - b)\} \le 0.$$

The constants may now depend on the fixed laws. In the sufficient direction, replace Δ by their gap and M_0 by a bound on their residuals; convexity in baseline and rate, the potential proof and the logit bound remain unchanged. In the necessary direction a violating endpoint baseline gives the same strictly convex f, now with $f'(0) = -(\mu_1 - \mu_2) < 0$, and the same lower-bound interface.

Relation to the residual budget. When $B = 0$ and $0 < \Delta \le R$, $J_-(\eta, 0) = \psi_{\Delta/R}(2\eta R)$ and

$$J_-(\eta, 0) - J_+(\eta, 0) = (2 - \Delta/R)[\sinh(2\eta R) - 2\eta R] \ge 0.$$

The criterion therefore recovers the two-arm residual threshold in this range. The equality of the two formulations is not assumed in general. For instance, when $\Delta = 2R$ and $B \le R$, the raw-budget laws must be the deterministic pair $(R, -R)$ and both corner functions are strictly negative at every finite $\eta > 0$. This is compatible with the theorem: there need not be a finite unsafe ceiling for a separately constrained class.

D CRITICAL EXPECTATION DEFECTS AND FINITE-HORIZON CROSSOVER

This section concerns a specific two-arm endpoint family, not every IID law or adaptive baseline. Fix $M > 0$, $d \in (0, 1]$, independent rewards

$$\mathbb{P}(Y^{(1)} = M) = d/2, \qquad \mathbb{P}(Y^{(1)} = -M) = 1 - d/2, \qquad Y^{(2)} = -M,$$

baseline zero, and a fixed scalar $\eta > 0$. The mean gap is $\Delta = dM$. Let $s = 2\eta M$, $L_t = \theta_{1,t} - \theta_{2,t}$, $X_t = e^{-L_t}$, $X_0 = x > 0$, $p_t = (1 + X_t)^{-1}$, $q_t = 1 - p_t$, and

$$Q_T = \sum_{t<T} q_t, \quad N_T = \sum_{t<T} \mathbf{1}_{\{I_t=2\}}, \quad H_T = \sum_{t<T} q_t^2, \quad G_T = sd \sum_{t<T} p_t q_t.$$

Thus $\widehat{R}_T = \Delta N_T$ and $R_T = \mathbb{E}\widehat{R}_T$. Write $A_d(z) = \cosh z + (1-d)\sinh z$, $\psi_d(z) = A_d(z) - 1 - z$, $\kappa = s_*/s$, $\epsilon = 1 - \kappa$, and $v_* = s_*\psi_d'(s_*) > 0$, where s_* is the positive root. Each s specifies one fixed-scalar infinite process; a row indexed by T with parameter s_T is a different process, not within-run adaptation.

Theorem 11 (Critical defect)**.** *There is a continuous, strictly positive harmonic function $h_s(x) = \lim_n \mathbb{E}_{s,x} X_n^\kappa$, normalized by $h_s(x) \to 0$ as $x \downarrow 0$ and $h_s(x)/x^\kappa \to 1$ as $x \to \infty$. For every fixed $s > 0$, $X_t \to 0$, $H_\infty < \infty$ a.s., and $L_T/\log T, G_T/\log T \to 1$ a.s. and in L^1; $Q_T/\log T, N_T/\log T \to 1/(sd)$ and $\widehat{R}_T/\log T \to M/s$ a.s. Moreover:*

1. *If $s < s_*$, $\mathbb{E}H_\infty \le x/[-\psi_d(s)]$, $R_T = (M/s)\log T + O(1)$, and the count/regret limits hold in L^1.*
2. *If $s = s_*$, with $a_*(x) = h_{s_*}(x)/v_* > 0$,*

$$\mathbb{E}H_T = a_*(x)\log T + O(1), \qquad R_T = [M/s_* + \Delta a_*(x)]\log T + O(1).$$

The normalized counts are not uniformly integrable, and $\mathbb{E}|N_T/\log T - 1/(s_*d)| \to a_*(x)$. The analogous absolute-defect statement holds for Q_T and for realized pseudo-regret $\widehat{R}_T = \Delta N_T$, whose limiting absolute defect is $\Delta a_*(x)$. The limits hold in L^p for $0 < p < 1$, whereas $\mathbb{E}(N_T/\log T)^p \to \infty$ for $p > 1$.

3. *If $s > s_*$, $\mathbb{E}H_T = \Theta(T^{1-\kappa})$ and $R_T = \Theta(T^{1-\kappa})$.*

Constants in fixed-parameter statements may depend on M, d, s, x.

Theorem 12 (Uniform near-critical crossover)**.** *Fix M, d, x. Put $F_\epsilon(T) = (T^\epsilon - 1)/\epsilon$ for $\epsilon \neq 0$ and $F_0(T) = \log T$, and define $a_s(x) = \epsilon h_s(x)/\psi_d(s)$ off the root and $a_{s_*} = a_*$. There is a finite constant uniform over $7/8 \le \kappa(s) \le 9/8$, $T \ge 2$, such that*

$$|\mathbb{E}H_T - a_s(x)F_\epsilon(T)| \le C[1 + T^\epsilon + T^{-1/3}F_\epsilon(T)], \tag{29}$$

$$\left|\frac{\mathbb{E}H_T}{F_\epsilon(T)} - a_s(x)\right| \le C[|\epsilon| + 1/\log T + T^{-1/3}], \tag{30}$$

$$\left|\frac{R_T - (M/s)\log T}{\Delta F_\epsilon(T)} - a_s(x)\right| \le C[|\epsilon| + 1/\log T + T^{-1/3}]. \tag{31}$$

In particular for any $s_T \to s_*$, $\mathbb{E}_{s_T,x}H_T \sim a_*(x)F_{\epsilon_T}(T)$. No assumption that $\epsilon_T \log T$ converges or remains bounded is needed.

D.1 Exact kernel and globally signed power drift

At $X_t = u$, write $p = (1+u)^{-1}$, $q = 1 - p$. The three outcomes are

$$X_{t+1} = \begin{cases} ue^{sq}, & (1 - d/2)p, \\ ue^{-sq}, & (d/2)p, \\ ue^{-sp}, & q. \end{cases} \tag{32}$$

The right column gives probabilities. Every finite-time power moment exists because $xe^{-st} \le X_t \le xe^{st}$. Let $D_s = P_s - I$. Define

$$\mathcal{F}_z(p) = \frac{A_d(zq) - 1}{q} + \frac{e^{-zp} - 1}{p}, \qquad B_z(p) = \psi_d(z) - \mathcal{F}_z(p),$$

$$f_r(u) = \frac{u^{r+1}}{(1+u)^2}, \qquad g_{s,r}(u) = B_{sr}(p(u))f_r(u).$$

Difference quotients extend continuously to the endpoints. Direct evaluation of (32) gives the exact identity

$$D_s(u^r) = \psi_d(sr)f_r(u) - g_{s,r}(u). \tag{33}$$

For $z \ge 0$, $B_z(0) = 0$ and

$$0 \le B_z(p) \le z^2 e^z p, \qquad |\partial_z B_z(p)| \le Cp \tag{34}$$

on each compact z interval. To check the sign, expand

$$B_z(p) = \sum_{n \ge 2} \frac{z^n}{n!}\{a_n(1 - q^{n-1}) - (-1)^n p^{n-1}\}, \quad a_n = 1 \ (n \text{ even}), \quad a_n = 1 - d \ (n \text{ odd}).$$

Every even coefficient is $1 - q^{n-1} - p^{n-1} \ge 0$ and every odd coefficient is $(1-d)(1-q^{n-1}) + p^{n-1} \ge 0$. For the quantitative bound, differentiate the integral difference-quotient representation to obtain

$$\partial_p \mathcal{F}_z(p) = z^2 \int_0^1 v[e^{-zpv} - A_d(zqv)]\, dv.$$

Its absolute value is at most $z^2 e^z$, and the absolute value of its z derivative is at most $e^z(2z + z^2/3)$. Integrate in p from zero; the series and derivatives are uniformly convergent on compact sets. Thus X_t^r is a nonnegative supermartingale for $0 < r \le \kappa$, and for $0 < r < \kappa$, finite summation and monotone convergence yield

$$\mathbb{E}_{s,u} \sum_{t \ge 0} f_r(X_t) \le u^r / c_{s,r}, \qquad c_{s,r} = -\psi_d(sr) > 0. \tag{35}$$

D.2 Fractional moments and the weighted linear scale

For $0 < r < \kappa$, let $a = \min\{r, \kappa - r\}$ and $A = \max\{1, x\}$. Set $m_r(t) = \mathbb{E}X_t^r$ and split it into $X_t \le 1$ and $X_t > 1$ pieces. If the low piece is at least $m_r/2$, Jensen/Hölder gives $\mathbb{E}f_r(X_t) \ge$

$\frac{1}{4}(m_r/2)^{1+1/r}$. If the high piece is at least $m_r/2$, interpolate powers $r-1$ and κ on $X>1$ (so negative powers cause no origin integrability issue) to get $\mathbb{E}f_r(X_t) \geq \frac{1}{4}(m_r/2)^{1+1/(\kappa-r)}A^{-\kappa/(\kappa-r)}$. Since $m_r \leq A^r$, both imply

$$\mathbb{E}f_r(X_t) \geq \frac{1}{4A}\left(\frac{m_r(t)}{2A^r}\right)^{1+1/a}.$$

Indeed the exponents $1+1/r$ and $1+1/(\kappa-r)$ are at most $1+1/a$; after normalization the respective A factors are at least $1/A$. By (33), $m_r(t+1) \leq m_r(t) - bm_r(t)^{1+1/a}$, where

$$b = \min\left\{\frac{c_{s,r}}{4A(2A^r)^{1+1/a}}, \frac{1}{2}x^{-r/a}\right\} > 0.$$

The inequality $(1-v)^{-1/a} \geq 1 + v/a$ implies

$$m_r(t) \leq (x^{-r/a} + bt/a)^{-a} = O((1+t)^{-a}). \tag{36}$$

This also proves uniformity on compact parameter ranges used below.

For $\mathcal{I} = \{s : 7/8 \leq \kappa(s) \leq 9/8\}$, fix $\rho = 1/4$ and define the finite measure $\mu_{s,T}(E) = \mathbb{E}_{s,x}[X_T^\kappa \mathbf{1}_{\{X_T/T \in E\}}]$. The low power $r = \kappa - \rho$ in (36) gives $\mathbb{E}X_T^{\kappa-\rho} \leq CT^{-\rho}$ uniformly. For a higher power, (33) and $pq \leq 1/u$ give $D_s(u^{\kappa+1}) \leq \psi_d(s(\kappa+1))u^\kappa$. Hence $\mathbb{E}X_T^{\kappa+1} \leq x^{\kappa+1} + Cx^\kappa T$. Interpolating with $\mathbb{E}X_T^\kappa \leq x^\kappa$ yields $\mathbb{E}X_T^{\kappa+\rho} \leq CT^\rho$. Therefore

$$\int (y^\rho + y^{-\rho})\, d\mu_{s,T} \leq C, \quad \left|\int \log y\, d\mu_{s,T}\right| \leq C, \quad \left|\int (y^v - 1)\, d\mu_{s,T}\right| \leq C|v| \quad (|v| \leq 1/8). \tag{37}$$

The last two statements use $|\log y| \leq C_\rho(y^\rho + y^{-\rho})$ and $|y^v - 1| \leq |v||\log y|e^{|v||\log y|}$. This is control under the power-weighted measure, not a claim that X_T/T has a nondegenerate limit under the original probability measure.

D.3 Positive harmonic mass, normalization and an error certificate

At $r = \kappa$, the first term in (33) vanishes. Thus

$$h_s(x) = x^\kappa - \mathbb{E}_{s,x}\sum_{t\geq 0} g_{s,\kappa}(X_t) = \lim_n \mathbb{E}_{s,x}X_n^\kappa \geq 0 \tag{38}$$

is well-defined by nonnegative loss summation. Choose $\max\{0, \kappa-1\} < r < \kappa$. Since $p(u)u^{\kappa-r} \leq 1$, equations (34) and (35) give

$$x^\kappa - C_{s,r}x^r \leq h_s(x) \leq x^\kappa, \qquad C_{s,r} = s_*^2 e^{s_*}/[-\psi_d(sr)]. \tag{39}$$

It follows that $h_s(x)/x^\kappa \to 1$ at infinity and $h_s(x) \to 0$ at zero. The finite three-point kernel permits termwise passage through the decreasing sequence $P_s^n(u^\kappa)$, giving $P_s h_s = h_s$. For a sufficiently large fixed H, (39) makes $h_s(u) > 0$ for $u \geq H$. From any $x > 0$, a finite natural sequence of best-arm pulls with reward $-M$ increases X by a multiplier at least $e^{sq(x)} > 1$ at each step. This sequence has positive probability and reaches H in finite time. Harmonicity therefore gives $h_s(x) > 0$ for every finite x. More explicitly, take $H \geq \max\{x, (2C_{s,r})^{1/(\kappa-r)}\}$ and iterate $u_{j+1} = u_j e^{sq(u_j)}$ until $u_n \geq H$; then

$$h_s(x) \geq \tfrac{1}{2}u_n^\kappa \prod_{j<n}\frac{1-d/2}{1+u_j} > 0.$$

Also the $n=3$ term of the nonnegative series gives $B_{s_*}(p) \geq s_*^3 p^2/6 > 0$. The loss at time zero proves

$$0 < h_s(x) \leq x^\kappa - g_{s,\kappa}(x) \leq x^\kappa - \tfrac{1}{6}s_*^3 x^\kappa p(x)^3 q(x) < x^\kappa. \tag{40}$$

In particular, the critical harmonic mass cannot be replaced by initial odds.

Let $h_{s,n}(x) = P_s^n(u^\kappa)(x)$. The Markov property, harmonicity and (39) imply

$$0 \leq h_{s,n}(x) - h_s(x) \leq C_{s,r}\mathbb{E}X_n^r \leq C_{s,r}(x^{-r/a} + bn/a)^{-a}. \tag{41}$$

For $s \in \mathcal{I}$ one may take $r = 1/3$ and $a = 1/3$, obtaining a uniform $C(1+n)^{-1/3}$ bound, also locally uniform in $x > 0$. Every finite recursion is continuous in s, x; the local uniform limit makes $h_s(x)$ jointly continuous near criticality. For other fixed s, choose an admissible r in (39) to get continuity in x.

For completeness, normalization uniquely determines the continuous harmonic function. If g is another such function, stop at first exit from (a, b). Equation (35) and $\inf_{[a,b]} f_r > 0$ imply a finite mean exit time. Multipliers lie in $[e^{-s}, e^s]$, so the stopped state is in the compact interval $[ae^{-s}, be^s]$. Bounded convergence justifies optional stopping of $g - h_s$. At upper exit its expected absolute value is at most εx^κ for sufficiently large b, since X^κ is a nonnegative supermartingale. At lower exit it is bounded by $\sup_{0<u\le a} |g(u) - h_s(u)| \to 0$. First let $a \downarrow 0$, then $\varepsilon \downarrow 0$. Thus $g = h_s$.

D.4 Typical paths and expected logits without count UI

Choose $0 < r < \min\{1, \kappa\}$. Supermartingale convergence and (35) imply $X_t \to 0$ a.s.: a positive finite limit would make $\sum f_r(X_t)$ diverge. Eventually $X_t \le 1$ and $q_t^2 = X_t^{1-r} f_r(X_t) \le f_r(X_t)$, so $H_\infty < \infty$ a.s. The exact logit decomposition is $L_T = L_0 + G_T + M_T$ with $\mathbb{E}_t \Delta L_t = sdp_t q_t$ and $\mathbb{E}_t(\Delta M_t)^2 \le s^2 p_t q_t = (s/d)\Delta G_t$. Lemma 6 gives $G_T \to \infty$ and $L_T/G_T \to 1$: a finite clock would force a finite logit limit, contradicting $X_t \to 0$. Eventually, for every small $\varepsilon > 0$,

$$(sd/4)e^{-(1+\varepsilon)G_t} \le \Delta G_t \le sde^{-(1-\varepsilon)G_t}.$$

For $a = 1 + \varepsilon$, the inequality $e^y - 1 \ge y$ gives, eventually,

$$e^{aG_{t+1}} - e^{aG_t} \ge ae^{aG_t}\Delta G_t \ge asd/4.$$

For $b = 1 - \varepsilon > 0$, use $e^y - 1 \le ye^y$ and $0 \le \Delta G_t = sdp_t q_t \le sd/4$ to obtain

$$e^{bG_{t+1}} - e^{bG_t} \le be^{bG_t}\Delta G_t e^{b\Delta G_t} \le bsde^{bsd/4}.$$

Summing from the finite path-dependent time at which the bounds hold yields $\liminf G_T/\log T \ge 1/(1+\varepsilon)$ and $\limsup G_T/\log T \le 1/(1-\varepsilon)$. Take a countable sequence $\varepsilon \downarrow 0$; hence $G_T/\log T \to 1$ and, using $L_T/G_T \to 1$, $L_T/\log T \to 1$ a.s. The exact identity $Q_T = G_T/(sd) + H_T$ gives the predictable count limit. The count martingale $N_T - Q_T$ has variance clock $\sum pq = G_T/(sd)$; Lemma 6 proves the actual count and a.s. regret limits.

For expectation, choose $0 < r < \kappa/2$ in (36). Log Jensen gives $\mathbb{E}L_T \ge -(1/r)\log \mathbb{E}X_T^r \ge \log T - C$. For $\Lambda_t = e^{L_t} = 1/X_t$, the kernel gives $\mathbb{E}_t|\Delta L_t| = 2sp_t q_t$, $|\Delta L_t| \le s$, and hence $\mathbb{E}_t|\Delta\Lambda_t| \le 2se^s \Lambda_t p_t q_t = 2se^s p_t^2 \le 2se^s$. Thus $\mathbb{E}\Lambda_T \le x^{-1} + 2se^s T$, and a second Jensen inequality gives

$$\mathbb{E}L_T = \log T + O(1). \tag{42}$$

On $\mathcal{I}$, $r = 1/3$ makes the constants uniform. Nonnegativity, $\mathbb{E}G_T = \mathbb{E}L_T - L_0$, and Scheffe's lemma give $G_T/\log T \to 1$ in L^1. Also $\mathbb{E}M_T^2 \le (s/d)\mathbb{E}G_T = O(\log T)$ gives the L^1 logit limit. Crucially,

$$R_T = (M/s)(\mathbb{E}L_T - L_0) + \Delta\mathbb{E}H_T. \tag{43}$$

This concentrates the expected/typical discrepancy in square occupancy; it does not assume count uniform integrability.

D.5 Moment subtraction cancels the critical singularity

For $s \in \mathcal{I}$ and r between 1 and $\kappa(s)$, set $r_0 = 1/3$. Equation (34) and its derivative give

$$|\partial_r g_{s,r}(u)| \le Cp(u)u^r p(u)q(u)(1 + |\log u|) \le Cf_{r_0}(u).$$

The second bound holds since $13/24 \le r - r_0 \le 19/24$ and $u^{r-r_0}(1+|\log u|)/(1+u)$ is uniformly bounded; splitting at 1 and using $\sup_{z\ge 0} ze^{-az} = 1/(ea)$ verifies it. Consequently, for $J_{s,r}(T) = \mathbb{E}\sum_{t<T} g_{s,r}(X_t)$,

$$|J_{s,1}(T) - J_{s,\kappa}(T)| \le C|1 - \kappa| \tag{44}$$

uniformly in T, by the integral mean formula and (35). This also bounds the infinite loss series. Finite summation of (33), with $m_r(T) = \mathbb{E}X_T^r$, gives

$$m_1(T) = x + \psi_d(s)\mathbb{E}H_T - J_{s,1}(T), \qquad m_\kappa(T) = x^\kappa - J_{s,\kappa}(T).$$

Subtracting first, rather than dividing either equation by a small margin, gives the exact identity

$$\psi_d(s)\mathbb{E}H_T - [m_1(T) - m_\kappa(T)] = x^\kappa - x + J_{s,1}(T) - J_{s,\kappa}(T).$$

For fixed $x > 0$, the mean-value theorem bounds the initial term by

$$|x^\kappa - x| \le C_x|\kappa - 1|, \qquad C_x = |\log x| \max_{v\in[7/8,9/8]} x^v < \infty.$$

Together with (44), this proves

$$|\psi_d(s)\mathbb{E}H_T - [m_1(T) - m_\kappa(T)]| \le C|\epsilon|.$$

By the weighted measure in (37), $m_1(T) = T^\epsilon \int y^\epsilon \, d\mu_{s,T}$ and $m_\kappa(T) = \int 1 \, d\mu_{s,T}$, so

$$|m_1(T) - T^\epsilon m_\kappa(T)| \le C|\epsilon|T^\epsilon.$$

Combining the two inequalities yields

$$|\psi_d(s)\mathbb{E}H_T - m_\kappa(T)(T^\epsilon - 1)| \le C|\epsilon|(1 + T^\epsilon). \tag{45}$$

Since $s = s_*/(1-\epsilon)$, $\psi_d(s) = v_*\epsilon + O(\epsilon^2)$; the ratio $\epsilon/\psi_d(s)$ extends continuously to a strictly positive, bounded function on $\mathcal{I}$, with value $1/v_*$ at criticality. For $\epsilon \neq 0$, division of (45) and replacement of $m_\kappa(T)$ by $h_s(x)$ using (41) prove (29). For each finite T, the kernel, expectation and harmonic coefficient are continuous in s, so the same uniform inequality extends to $s = s_*$ without exchanging the time and parameter limits. Finally,

$$\frac{1 + T^\epsilon}{F_\epsilon(T)} = |\epsilon| \coth(|\epsilon| \log T/2) \le |\epsilon| + 2/\log T,$$

with its continuous value at zero. This proves (30); (42)–(43) and $F_\epsilon(T)^{-1} \le |\epsilon| + 1/\log T$ prove (31).

The critical coefficient can also be checked directly, without parameter limits. Differentiate (33) at $s = s_*, r = 1$:

$$D_{s_*}(u \log u) = v_* q(u)^2 - \partial_r g_{s_*,r}(u)\big|_{r=1}.$$

Differentiation and expectation commute at finite T because of the deterministic bounds $xe^{-sT} \le X_T \le xe^{sT}$. After summing, the derivative-loss term is uniformly bounded by the proof of (44). Equation (37) gives

$$\mathbb{E}[X_T \log X_T] = (\mathbb{E}X_T)\log T + \mathbb{E}[X_T \log(X_T/T)] = h_*(x)\log T + O(1),$$

using $|\mathbb{E}X_T - h_*(x)| \le CT^{-1/3}$. Therefore $\mathbb{E}H_T = (h_*(x)/v_*)\log T + O(1)$, confirming the same coefficient. Despite $X_T \to 0$ a.s., its mean tends to $h_*(x) > 0$.

D.6 UI, SUPERCRITICAL POWER ORDER, AND LIMIT ORDER

For $s < s_*$, apply (33) at $r = 1$ to get $\mathbb{E}H_\infty \le x/[-\psi_d(s)]$. Equations (42) and (43) prove the expected coefficient; the a.s. count limit and nonnegative Scheffe argument give L^1 convergence and UI. At $s = s_*$, let $Z_T = N_T/\log T$ and $c = 1/(s_* d)$. Then $Z_T \to c$ a.s. but $\mathbb{E}Z_T \to c + a_*(x) > c$. Bounded convergence for $\min(Z_T, c)$ proves $\mathbb{E}|Z_T - c| = \mathbb{E}Z_T + c - 2\mathbb{E}\min(Z_T, c) \to a_*(x)$. For $0 < p < 1$, truncate $|Z_T - c|^p$ at a fixed bound A; the tail is at most $A^{p-1}\sup_T \mathbb{E}|Z_T - c|$. Let $T \to \infty$, then $A \to \infty$, to get L^p convergence. For $p > 1$ and fixed $A > c$, bounded convergence gives $\mathbb{E}[Z_T \mathbf{1}_{\{Z_T \le A\}}] \to c$, so the tail first moment tends to $a_*(x)$. Thus $\liminf_T \mathbb{E}Z_T^p \ge A^{p-1}a_*(x)$; let $A \to \infty$. The same arguments apply to Q_T and scaled regret. Monotone convergence of H_T gives $\mathbb{E}H_\infty = \infty$ at criticality.

For fixed $s > s_*$, $0 < \kappa < 1$. Choose $0 < \rho < \kappa/2$. Equation (36) gives $\mathbb{E}X_T^{\kappa-\rho} \le CT^{-\rho}$. For sufficiently small fixed $a > 0$, $\mathbb{E}[X_T^\kappa \mathbf{1}_{\{X_T \le aT\}}] \le Ca^\rho \le h_s(x)/2$. Therefore $\mathbb{E}X_T \ge (aT)^{1-\kappa}h_s(x)/2$. The high-power argument gives $\mathbb{E}X_T^{\kappa+1} \le C(1+T)$; interpolation with $\mathbb{E}X_T^\kappa \le x^\kappa$ yields $\mathbb{E}X_T \le C(1+T)^{1-\kappa}$. Choose $0 < r < \min\{1, \kappa\}$; then $g_{s,1}(u) \le s^2 e^s p(u) u^{1-r} f_r(u) \le s^2 e^s f_r(u)$, so $J_{s,1}(\infty) < \infty$. The exact first-moment identity $\psi_d(s)\mathbb{E}H_T = \mathbb{E}X_T - x + J_{s,1}(T)$ proves the claimed power order, and (43) transfers it to regret.

For any near-critical rows $s_T \to s_*$, continuity gives $a_{s_T}(x) \to a_*(x)$; (30) proves the uniform relative asymptotic. If $\epsilon_T \log T \to z \in \mathbb{R}$, then

$$\frac{R_T(s_T)}{\log T} \to \frac{M}{s_*} + \Delta a_*(x)\frac{e^z - 1}{z},$$

where the last ratio equals 1 at zero. If $\epsilon_T \log T \to -\infty$, $\mathbb{E}H_T \sim a_*/|\epsilon_T|$ and $R_T/\log T \to M/s_*$. If it tends to $+\infty$, $R_T \sim \Delta a_* T^{\epsilon_T}/\epsilon_T$. Taking $T \to \infty$ first on the subcritical side and then $s \uparrow s_*$ also yields $(\kappa - 1)\mathbb{E}H_\infty \to a_*(x)$. The limiting expected logarithmic coefficient jumps from M/s_* to $M/s_* + \Delta a_*(x)$ at the critical point although each finite-horizon expectation is continuous. As $x \to \infty$, after taking the time limit at each fixed x, this coefficient is asymptotic to $\Delta x/v_*$. No uniform claim for simultaneous $d \downarrow 0$ or $x \to \infty$ with T has been made.

D.7 What the harmonic recursion certifies

Start with $h_{s,0}(u) = u^\kappa$ and apply the three-point kernel (32) recursively. Equation (41) supplies an explicit interval

$$\max\{0, h_{s,n}(x) - C_{s,r}(x^{-r/a} + bn/a)^{-a}\} \le h_s(x) \le h_{s,n}(x).$$

This proves computability in principle, independently of the target regret limit. Full history expansion has exponential cost. Interpolation, a grid, or Monte Carlo does not inherit this certificate without its own error analysis. The theorem is an endpoint-family result, not an all-K critical coefficient, an adaptive-scalar crossover theorem, or a numerical robustness theorem. Fixed supercritical power order does not assert convergence to an explicit leading constant.

E Experimental definitions and reproducibility protocol

This paper selects complete safety and criticality suites from the retained archive. The selection follows the paper's mathematical scope, not the sign of observed effects. The original design was committed locally before pilot execution; it was not publicly preregistered. No new scientific simulation was run for this manuscript. All 88,064 selected original-law trajectories and all 14,336 proposal trajectories reached their registered horizons. The complete provenance archive also retains the certification, numerical stress and efficiency-diagnostic suites, including all 384 recorded failures. Those other suites are not evidence for a new theorem in this paper. Rewards are independent across arms and indexed by pull count. Each trajectory is stationary. A checkpoint is not an independent replicate.

Table 4: Complete selected original-law suites. Proposal observations are kept separate: 58 cells, 14,336 paths and 2,861,563,904 scheduled updates.

Suite	Cells	Trajectories	Scheduled updates
Safety	240	30,720	2,013,265,920
Criticality	58	57,344	11,446,255,616

E.1 Stationary laws and parameter sweeps

Safety. We use $M = 1$, $K \in \{2, 3, 5, 10\}$, $d \in \{1/8, 1/2, 1\}$, $\eta/\eta_c \in \{0.5, 0.9, 1, 1.1, 1.5\}$, and initial logits either all zero or $(0, 1, 0, \ldots, 0)$. Each cell has 128 trajectories of 65,536 updates. The signed hard family has best reward in $\{-1, 1\}$ with probability $(d+e)/2$ of $+1$, second-arm reward $-1 + e$, and all other rewards -1. Specifically, with $\gamma = (1 - 1/1.1)/2$ and $s_\gamma = 1.1 s_*(d)(1 - \gamma)$, we set

$$e = \frac{1}{2}\min\left\{1, \frac{\psi_d(s_\gamma)}{\sinh(s_\gamma) - s_\gamma}\right\}.$$

The construction fixes e once per (K, d); that same law is used at every rate ratio. Thus varying the rate does not also vary the reward instance. The deterministic comparison has best reward 1 and all tail rewards $1 - d$. Both have zero baseline. A class-level bad-instance theorem does not say every supercritical cell of either family must fail.

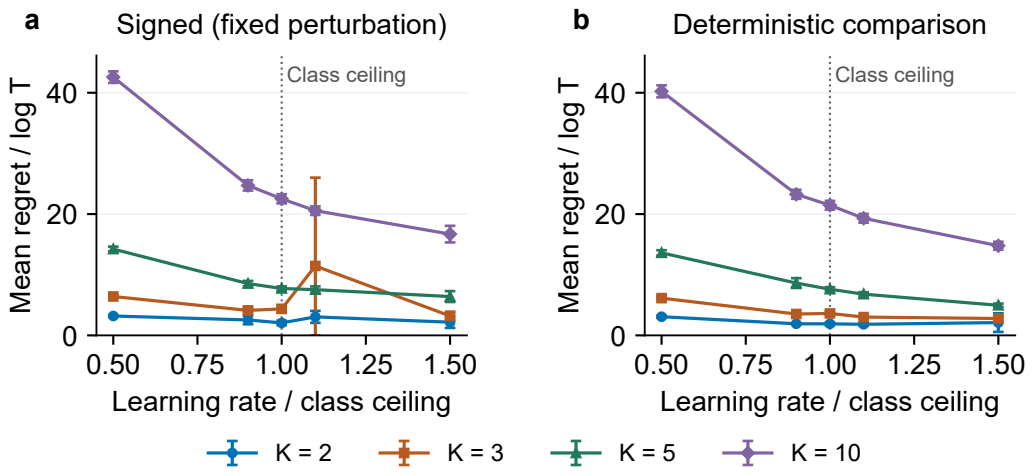


Figure 3: Crossing the class ceiling need not visibly raise regret for a given instance and horizon. Safety sweep at $d = 0.5$, logits $(0, 1, 0, \ldots, 0)$ and $T = 65,536$: 128 independent paths per point. Means are divided by $\log T$; bars are descriptive 1.96-SE intervals truncated at zero. The vertical line marks the analytic class ceiling $\eta/\eta_c = 1$. At $K = 3$, $\eta/\eta_c = 1.1$ in (a), the largest two of 128 paths contribute 68.6% of that cell's regret.

Criticality. The fixed-rate endpoint cells use $K = 2$, best reward ± 1 with probability $d/2$ of $+1$, and tail reward -1; $d \in \{1/4, 3/4\}$, $x \in \{0.1, 1, 10\}$, and $\eta/\eta_c \in \{0.75, 1, 1.25\}$. Logits start at $(-\log(x)/2, \log(x)/2)$. Each of these 18 cells has 2,048 trajectories of 262,144 updates. The 40 crossover cells use $d = 1/2$, $x \in \{1, 10\}$, terminal horizons $T \in \{4096, 16384, 65536, 262144\}$, and $z = \epsilon \log T \in \{-1, -0.5, 0, 0.5, 1\}$. Each has 512 trajectories. We set $s = s_*/(1 - z/\log T)$ before a trajectory and keep it constant. Crossover comparisons use terminal rows only, not intermediate checkpoints as if their rate had been selected for a different horizon.

E.2 Finite numerical laws

The CPU simulator evaluates subtract-maximum softmax in binary64 and uses a finite-bit dyadic categorical law in both action sampling and the score. Reference-target cells use 52 bits. Consequently even a nominal exact-target cell is a finite approximation, not a literal exact-real sample or an infinite-history relative-error certificate. Bernoulli draws use 53-bit thresholds. The artifact records the resulting exact rational means and a bound for the discrepancy of the nominal and executed regret definitions. Nominal-critical scalars retain their high-precision root bracket, binary64 rounding side and actual executed argument.

The 42 retained stress-suite trace samples (2,730 states) provide a separate finite arithmetic diagnostic. Their checks are available in the full replay but do not certify all time steps of the selected safety or criticality trajectories. No infinite-history full-support guarantee is inferred.

E.3 Rare-event measurement and uncertainty

For positive target and proposal branch laws $P_j(x)$ and $Q_j(x)$, put $L_0 = 1$ and $L_{t+1} = L_t P_{J_t}(X_t)/Q_{J_t}(X_t)$. At each finite horizon, the prefix density identity gives

$$\mathbb{E}_Q \sum_{t<T} L_t q(X_t)^2 = \mathbb{E}_P \sum_{t<T} q(X_t)^2 = \mathbb{E}_P H_T.$$

This is an exact-arithmetic finite-history identity, not an exchange of an asymptotic limit with expectation. The target update does not change. We neither self-normalize by terminal likelihoods nor pool proposal paths with original-law samples. The stored binary64 probability representation has

a separately reported nonzero discrepancy from the integer draw law; the numerical estimates are therefore not arithmetic certificates of unbiasedness.

The likelihood proposal adds 512 paths per fixed criticality cell and 128 paths per crossover cell: 14,336 paths and 2,861,563,904 scheduled updates, separate from the original-law evaluation. The target transition and its score remain unchanged. The proposal uses integer 62-bit categorical weights proportional to the target branch probability times $X^{\kappa}_{t+1}/X^{\kappa}_{t}$. Unnormalized prefix likelihoods, not terminal normalized weights, estimate finite-horizon square occupancy and predictable regret. Stored binary64 likelihood arithmetic approximates the exact Radon–Nikodym identity. Complete short trees check the implemented cancellation; independent integer-kernel probes quantify the separate probability-representation discrepancy. We retain weight variation, support and numerical-failure diagnostics. Neither a finite harmonic moment nor a small reported standard error certifies the limiting harmonic mass.

At each registered checkpoint, we report means, sample standard errors (sample variance denominator $n-1$), type-7 empirical quantiles 0.1/0.5/0.9/0.99, extrema, and the share of total observed regret contributed by the largest $\lceil 0.01n \rceil$ trajectories. Winsorized normalized regret uses $\min\{\widehat{R}_T/\log T, c\}$ at fixed $c \in \{2, 10, 100\}$. Normal 1.96-SE bands are descriptive. A separate Hoeffding interval uses the bounded original-law range $[0, \Delta_{\max} T]$ with confidence 0.95; that bound is not applied to unbounded likelihood-weighted estimators. No ordinary-MC interval rules out unobserved rare mass.

Each trajectory has domain-separated action, posterior and per-arm reward streams; proposal paths use a separate domain. Recognized numerical boundaries are terminal recorded outcomes, not replacement seeds. After a failure, checkpoint means are withheld whenever not all scheduled paths reached that checkpoint, instead of publishing an unconditional survivor mean. Raw transaction logs preserve resumptions and exact keys. Pilot domains are distinct and are used only for runtime and code checking.

E.4 Reading the class-ceiling sweep

Figure 3 compares the fixed-perturbation signed family with a deterministic comparison at the same displayed budgets. For $K = 10$, $d = 0.5$, increasing η/η_c from 1 to 1.5 lowers the signed-family mean regret$/\log T$ from 22.52 to 16.70. This finite-instance observation illustrates the distinction between a class guarantee and a visible transition at one horizon. Full grids follow in Appendix F.

F Complete finite-sample views and numerical diagnostics

The following views retain every registered terminal design stratum; all checkpoint rows remain in the source tables. The statistical unit is an independent trajectory, never a checkpoint. Mean intervals are descriptive 1.96-SE intervals, truncated only at the known nonnegative-regret boundary. They do not certify unobserved tail mass. No hypothesis tests or multiplicity-adjusted claims are made. Original-law and likelihood-proposal samples are always separate.

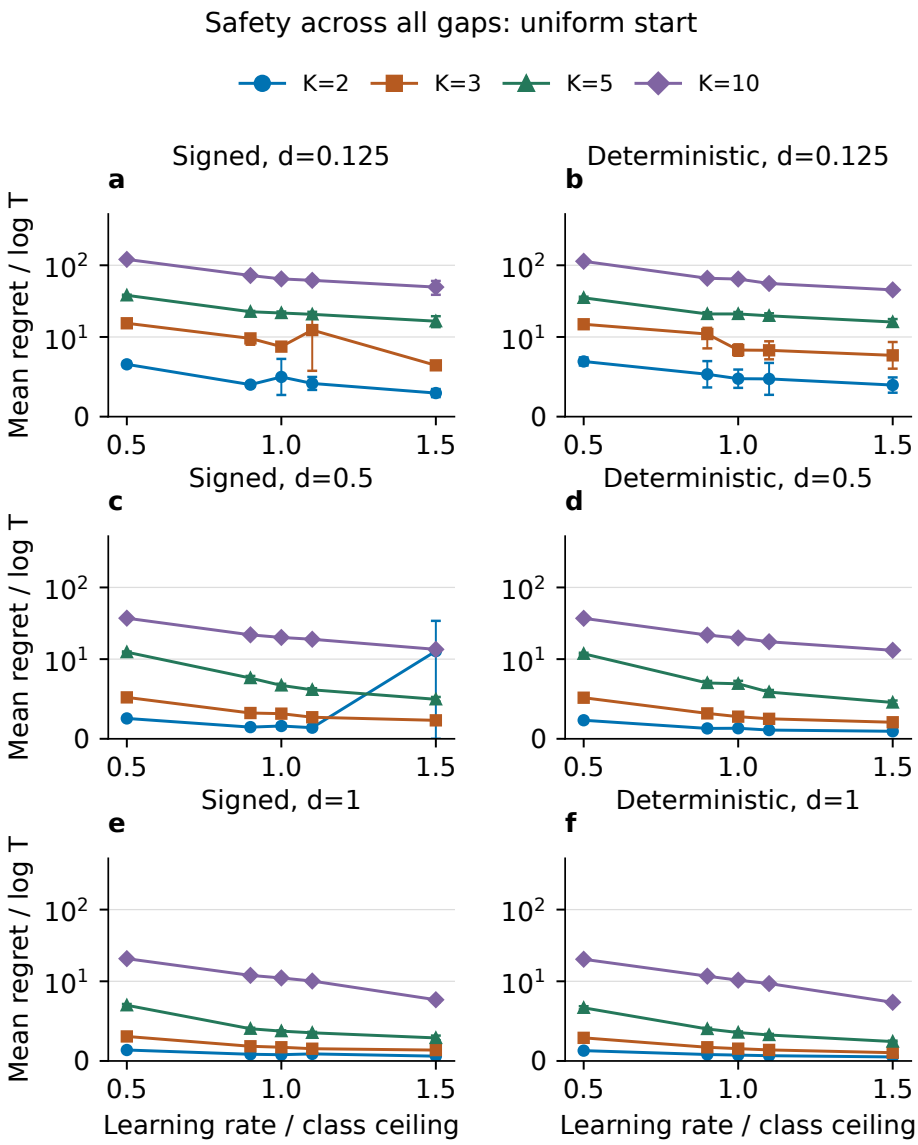


Figure 4: Complete safety grid, uniform logits. Each of 120 conditions has 128 independent original-law paths, $T = 65,536$. Points are mean pseudo-regret divided by $\log T$; bars are 1.96 trajectory-SE, truncated at zero. All four arm counts and five rate ratios are retained. The shared symlog axis has linear threshold 10 in these normalized units. Signed and deterministic laws remain distinct; finite costs are not an empirical certificate of the class-uniform safety iff.

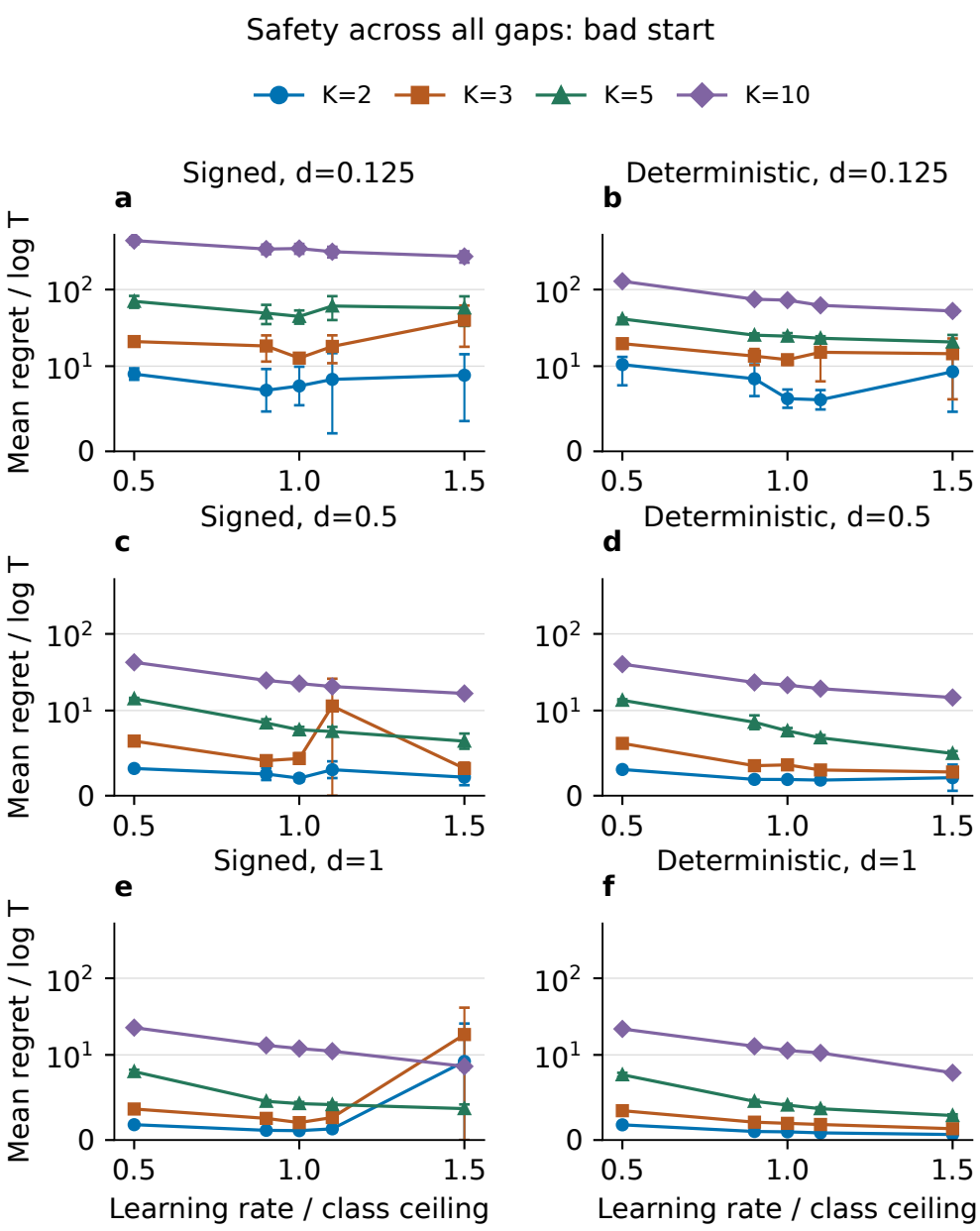


Figure 5: Complete safety grid, initial logits $(0, 1, 0, \ldots, 0)$. Each of 120 conditions has 128 independent original-law paths, $T = 65,536$. Points are mean pseudo-regret divided by $\log T$; bars are 1.96 trajectory-SE, truncated at zero. All four arm counts and five rate ratios are retained. The shared symlog axis has linear threshold 10 in these normalized units. Signed and deterministic laws remain distinct; finite costs are not an empirical certificate of the class-uniform safety iff.

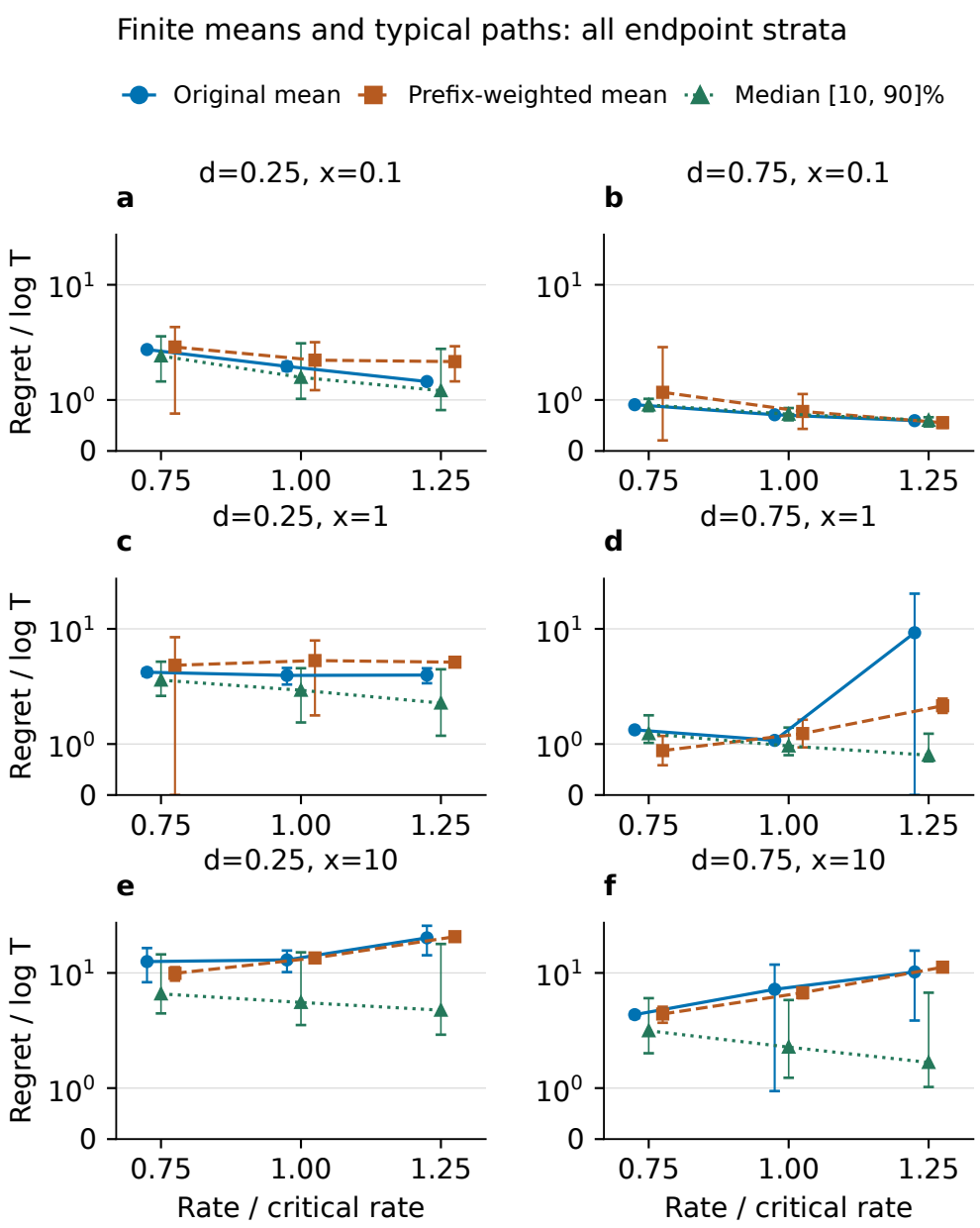


Figure 6: All six fixed (d, x) endpoint strata and three rate ratios at $T = 262,144$. Blue circles: original-law realized pseudo-regret mean (2,048 paths per condition); orange squares: separate prefix-likelihood-weighted predictable-regret mean (512 proposal paths). Both estimate finite expected regret under the target law, subject to the disclosed finite arithmetic approximation, and are not pooled. Bars for means are 1.96-SE; green triangles and whiskers are the original median and 10th–90th path quantiles, not a median confidence interval. Horizontal offsets are only for visibility; the shared symlog axis has linear threshold 2. No finite estimate is substituted for the limiting harmonic defect.

Table 5: Terminal square occupancy and weight diagnostics. The $H_T/\log T$ columns show means $\pm$ one descriptive SE across independent trajectories: 2,048 original-law paths and 512 separate proposal paths per condition. Original-law SEs are recomputed from the retained logs; proposal SEs use the existing prefix-weighted summaries. All 18 fixed conditions are retained. Terminal-weight ESS describes terminal weights, not the prefix-sum estimator.

d	x	Rate ratio	Original $H/\log T$	Weighted $H/\log T$	Terminal-weight ESS
0.25	0.1	0.75	0.242 ± 0.013	0.258 ± 0.004	1.00
0.25	0.1	1	0.553 ± 0.154	0.452 ± 0.005	1.80
0.25	0.1	1.25	0.445 ± 0.057	1.331 ± 0.032	3.56
0.25	1	0.75	3.553 ± 0.571	2.940 ± 0.063	57.56
0.25	1	1	4.782 ± 1.291	5.366 ± 0.080	1.00
0.25	1	1.25	6.200 ± 1.179	11.850 ± 0.227	247.48
0.25	10	0.75	41.011 ± 10.963	31.434 ± 0.958	33.03
0.25	10	1	46.056 ± 7.452	51.667 ± 1.001	61.69
0.25	10	1.25	88.722 ± 17.403	93.251 ± 1.690	159.35
0.75	0.1	0.75	0.023 ± 0.001	0.023 ± 0.000	1.43
0.75	0.1	1	0.022 ± 0.001	0.027 ± 0.000	3.39
0.75	0.1	1.25	0.030 ± 0.004	0.060 ± 0.003	4.06
0.75	1	0.75	0.247 ± 0.009	0.291 ± 0.004	1.00
0.75	1	1	0.304 ± 0.022	0.537 ± 0.010	2.13
0.75	1	1.25	11.176 ± 10.268	1.585 ± 0.048	2.93
0.75	10	0.75	2.949 ± 0.196	3.515 ± 0.089	1.08
0.75	10	1	7.514 ± 3.868	6.502 ± 0.115	18.33
0.75	10	1.25	12.583 ± 4.914	14.676 ± 0.313	1.01

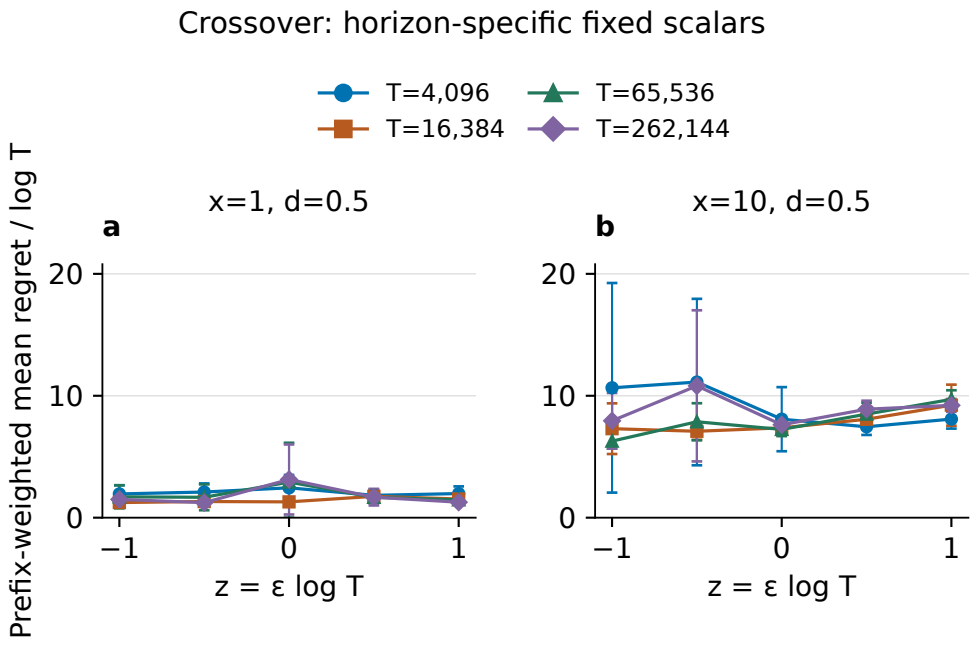


Figure 7: All 40 registered crossover terminal rows: both initializations, all four horizons and all five $z = \epsilon \log T$ values. Each point uses 128 independent proposal paths; bars are 1.96-SE of the unnormalized prefix-weighted predictable-regret estimator, divided by $\log T$ and truncated at zero. Each (T, z) condition fixes its scalar before simulation. Connecting points only guides the eye across distinct conditions, not a time-varying-scalar trajectory. Original-law estimates (512 paths per row) remain separately available in the full source table. No limiting h line, fitted limit, or normalized-to-zero ratio is used.

G HISTORICAL CUMULATIVE-ALLOCATION DIAGNOSTIC

This post hoc reanalysis uses existing original-law safety logs; no new trajectories were simulated. Selection was fixed before extracting the allocation outcomes: the signed safety family with $d = 0.5$, initial logits $(0, 1, 0, \ldots, 0)$, all four registered arm counts, and the smallest registered rate ratio $\eta/\eta_c = 0.5$. The selected cells contain 128 independent trajectories each, observed at all nine registered checkpoints from 256 to 65,536. All 512 trajectories are complete.

For these cells $R = 1$, $B = 0$, $\Delta_{\min} = 0.5$, and $K\eta = 0.4158152445$. The allocation theorem's sufficient margin is $c = 0.26001$ (rounded), positive using both nominal means and the archived finite-bit executed reward means. This verifies that the nominal parameter choices lie inside the stated sufficient region. It does not remove the binary64 state and finite-bit action approximations described in Appendix E.

Table 6 reports the normalized count and predictable mass for every suboptimal arm at the terminal horizon. Both exact-model limits equal one. Values above one at this finite horizon are retained; the table is an illustration, not an estimate certified to equal the asymptotic limit or a test of L^1 convergence. Arms sharing a trajectory are not treated as independent replicates. Standard errors are descriptive, computed across the 128 independent paths separately for each arm. The artifact includes every arm at every checkpoint and the eligibility audit for all original-law E1/E2 fixed-rate cells.

Table 6: Cumulative-allocation diagnostics at $T = 65,536$. Columns show means $\pm$ one descriptive SE. Each displayed quantity has exact-model asymptotic reference one.

K	Arm i	Δ_i	$\eta K \Delta_i N_i / \log T$	$\eta K \Delta_i Q_i / \log T$
2	2	0.5000	1.328 ± 0.034	1.313 ± 0.039
3	2	0.5000	1.631 ± 0.044	1.634 ± 0.051
3	3	0.6177	1.035 ± 0.017	1.015 ± 0.026
5	2	0.5000	2.232 ± 0.050	2.213 ± 0.058
5	3	0.6177	1.229 ± 0.018	1.214 ± 0.026
5	4	0.6177	1.220 ± 0.016	1.242 ± 0.025
5	5	0.6177	1.237 ± 0.020	1.233 ± 0.030
10	2	0.5000	4.000 ± 0.090	3.983 ± 0.099
10	3	0.6177	1.706 ± 0.020	1.673 ± 0.026
10	4	0.6177	1.716 ± 0.023	1.738 ± 0.033
10	5	0.6177	1.715 ± 0.020	1.712 ± 0.031
10	6	0.6177	1.696 ± 0.019	1.674 ± 0.030
10	7	0.6177	1.725 ± 0.024	1.741 ± 0.035
10	8	0.6177	1.711 ± 0.018	1.712 ± 0.026
10	9	0.6177	1.709 ± 0.021	1.726 ± 0.030
10	10	0.6177	1.728 ± 0.021	1.733 ± 0.030

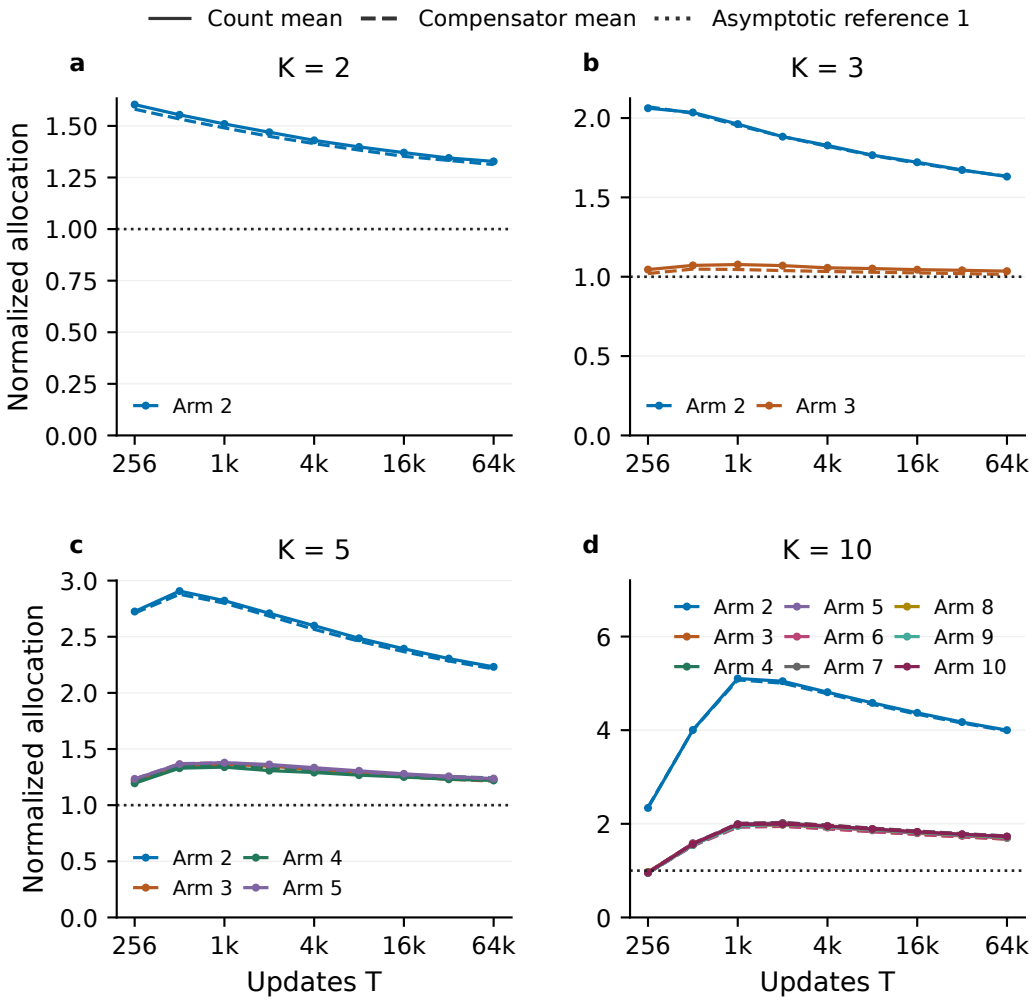


Figure 8: All nine checkpoints from the same four selected cells and all 16 suboptimal arms. Colors identify arms within each panel; solid lines show $\eta K \Delta_i \mathbb{E}_{\mathrm{MC}} N_i(T)/\log T$ and dashed lines the corresponding compensator means. Each mean uses 128 independent paths. The horizontal line is the exact-model asymptotic reference one. Finite-horizon deviations, including the larger $K = 10$, arm-2 count, are retained. Checkpoints share trajectories; no curve fitting or independent checkpoint inference is used. Descriptive SEs for every point are supplied in the accompanying CSV, and terminal SEs appear in Table 6.