CONSTANT-RANK CONFIDENCE AMPLIFICATION FOR PARAMETER-FREE HEAVY-TAILED GRAPH-UNIMODAL BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study stochastic graph-unimodal bandits with arm-wise symmetric rewards and unknown finite centered $(1 + a_i)$-th moments. Fixed sample coverage alone does not control the decision cost of adaptive cache reuse, so we charge each fixed-prefix failure by the maximum number of decisions its cached endpoint can govern. Amplified prefix reuse with localization (APR-LOCAL) pairs constant-rank resampled-median-of-means (RMM) endpoints with dyadic median-of-means (MoM) leader scores; independent-base amplification makes the confidence dependence logarithmic, and prefix epoch budgets keep the expected total bad-cache decision charge finite. The charge bounds suboptimal local choices in the leader's own clock, and periodic leader pulls with summable MoM score tails make suboptimal leader occurrences finite in expectation. Both the certified and practical settings attain expected regret $O(\sum_{j\in\mathcal{N}(i^\star)}[\Delta_j + M_j^{1/a_j}\Delta_j^{-1/a_j}]\log T)$ plus a finite term. Under a global clock, APR-UCB attains the all-arm $O(\log T)$ guarantee, improving the published RMM-UCB horizon factor from $\log^2 T$. A symmetric Gaussian-subfamily lower bound matches the horizon order and optimal-neighbor support, with equal moment-gap powers at $a_j = 1$. Across 30 paired trials on two path instances of a infinite variance, matched comparisons place most of the certified-practical regret gap before cache activation and show the MoM leader raising the optimal-neighborhood pull share.

1 INTRODUCTION

In heavy-tailed stochastic bandits, rare rewards make empirical averages unreliable even when means exist. Classical robust indices require a known tail exponent and moment bound (Bubeck et al., 2013); recent work identifies barriers to adaptation (Genalti et al., 2024; Genalti & Metelli, 2025) and parameter-free polynomial guarantees for finite $(1 + \epsilon)$-moment classes (Genalti & Metelli, 2026). We focus on the logarithmic regime enabled by exact arm-wise symmetry and a mean vector unimodal on a known finite graph. Our anytime policy takes neither moment parameters nor the horizon as input and confines growing exploration cost to optimal-arm neighbors.

The main difficulty is controlling endpoint reuse under a leader-local clock. RMM gives exact fixed-sample coverage (Tamás et al., 2024), while OSUB selects a neighborhood through its current leader (Combes & Proutiere, 2014). Because a cached endpoint can govern many actions that determine subsequent leaders and cache queries, a valid analysis has to evaluate failure probability and reuse count together.

Local selection counts connect reuse cost to leader ascent. Our pathwise bound splits suboptimal local choices into a sublinear function of leader occurrences and an integrable error term. If local best neighbors receive too few samples, this error term must grow linearly with leader occurrences. If they receive enough samples, a suboptimal leader can persist only through score errors. Periodic forcing and summable dyadic score tails bound both cases in expectation, leaving only optimal-arm neighbors in the growing regret term.

APR-LOCAL uses fixed-rank RMM endpoints and dyadic median-of-means (MoM) leader scores. Tolerant bad-block bounds hold the fixed rank at four in the certified setting and eight in the practical

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.

one. The practical proof separates early cache activation from the later sample threshold needed for accuracy. Where the published RMM-UCB bound pays $\log^2 T$ (Tamás et al., 2025), a global clock recovers $\log T$.

Paired experiments compare global and graph-local policies on symmetric Pareto and Student-$t_{1.5}$ rewards. Matched-budget and empirical-leader ablations measure endpoint, budget, and leader effects; asymmetric perturbations test sensitivity to symmetry.

Our contributions are:

1. An ascent argument connects pathwise local selection counts to finite expected suboptimal-leader occurrences through conditions on endpoint errors and leader scores.
2. A parameter-free policy for symmetric heavy tails achieves neighbor-local logarithmic regret. A Gaussian-subfamily lower bound identifies necessary optimal-neighbor exploration; the global-clock specialization gives the corresponding all-arm upper bound.
3. Upper-tail certificates cover both endpoint settings. Paired ablations quantify their activation costs and the effect of robust leaders.

2 Related Work

Heavy-tailed and adaptive bandits. Truncated-mean and MoM UCB methods attain finite-$(1+\epsilon)$-moment rates when the moment order and scale are known (Bubeck et al., 2013). KL-based indices can attain the first-order heavy-tail lower bound under a known moment class, albeit at higher computational cost (Agrawal et al., 2021). APE^2 removes the moment-scale input while retaining the moment order (Lee et al., 2020); data-driven Catoni estimators adapt the scale in fixed-confidence tasks (Bhatt et al., 2022). AdaR-UCB adapts both heavy-tail parameters under a truncated non-positivity condition on the optimal reward; the same work establishes minimax adaptation barriers over unrestricted classes (Genalti et al., 2024). The associated open problem (Genalti & Metelli, 2025) subsequently led to a parameter-free polynomial-rate frontier whose unknown tail-order guarantee is pointwise for every fixed $\epsilon > 0$, but not uniform over $\epsilon \in (0, 1]$ (Genalti & Metelli, 2026). Under a truncated non-negativity condition on the optimal arm's loss, uniINF provides stochastic–adversarial best-of-both-worlds guarantees without known tail parameters (Chen et al., 2025). Median clipping and Clipped-SGD-UCB exploit symmetric noise with horizon- and envelope-dependent tuning (Kornilov et al., 2024; Dorn et al., 2025). Under exact centered symmetry, we obtain an anytime, instance-dependent logarithmic guarantee with graph localization. Table 10 in section L compares their inputs and reward conditions.

Symmetric mean estimation and RMM. Under a finite variance bound, MoM and related robust location estimators provide sub-Gaussian-type deviation bounds (Lugosi & Mendelson, 2019). Time-uniform robust confidence sequences likewise use a known moment order and bound (Wang & Ramdas, 2023). RMM takes a different route: sign-resampled MoM statistics give exact fixed-sample coverage for a symmetric distribution without moment inputs (Tamás et al., 2024), and the one-sided construction supports the anytime parameter-free RMM-UCB policy (Tamás et al., 2025). We amplify independent bases at fixed ranks four and eight, certify both settings with tolerant bad-block bounds, and bound excess local selections by allocating failure budgets through maximum cache reuse. For compactly supported symmetric laws, adaptive location estimators can attain faster-than-root-n rates (Kao et al., 2024).

Graph-unimodal bandits. For Bernoulli rewards, OSUB uses a graph leader and a leader-occurrence clock to match a lower bound whose growing term includes only optimal-arm neighbors (Combes & Proutiere, 2014). Unimodal Thompson sampling provides another asymptotically optimal local mechanism for Bernoulli rank-one bandits (Trinh et al., 2020). Heavy-tailed graphical bandits exploit side observations from a feedback graph (Gou et al., 2023); our learner observes only the pulled arm, and the graph constrains the mean vector instead. We retain OSUB's local clock, neighborhood restriction, and periodic leader pulls. Our ascent argument combines a pathwise local-count bound with score tails summable under dyadic reuse. Each statistic has a distinct role: cached endpoints limit local misselections, and MoM scores rule out persistent suboptimal leaders once better neighbors have enough samples.

3 SETTING AND CONFIDENCE CONSTRUCTION

3.1 GRAPH-UNIMODAL HEAVY-TAILED BANDITS

Let $G = (V, E)$ be a known finite undirected graph with $K = |V|$ and maximum degree γ. Arm i has i.i.d. rewards $(X_{i,n})_{n\geq 1}$ of mean μ_i; $N_i(t)$ is its pull count through t. At time t, the policy selects I_t and incurs pseudo-regret $\Delta_{I_t} = \mu_{i^\star} - \mu_{I_t}$, where $i^\star$ is the unique optimum; $R_T = \sum_{t=1}^{T} \Delta_{I_t}$ is cumulative pseudo-regret. For unknown tail exponents $a_i \in (0, 1]$ and centered-moment bounds $M_i < \infty$,

$$X_{i,1} - \mu_i \text{ is symmetric}, \qquad \mathbb{E}|X_{i,1} - \mu_i|^{1+a_i} \leq M_i. \tag{1}$$

The mean vector is graph-unimodal: every $i \neq i^\star$ has a neighbor with strictly larger mean. Write $\mathcal{N}(i)$ for the neighbors of i and $\mathcal{N}^+(i) = \{i\} \cup \mathcal{N}(i)$. The algorithm knows only G. Unsubscripted logarithms are natural.

3.2 A CONSTANT-RANK ONE-SIDED ENDPOINT

Split an i.i.d. sample $X_{1:n}$ into odd k nearly equal groups G_ℓ and let $\widehat{\mu}_0$ be its lower-median MoM estimate. For $j \in \{1, \ldots, m-1\}$ draw independent Rademacher signs $\alpha_{h,j} \in \{-1, 1\}$; barred terms below are group averages of $\alpha_{h,j}$ and $\alpha_{h,j}X_h$ over G_ℓ. Form the extended-real intersection

$$V_{\ell,j} = \frac{\widehat{\mu}_0 - \overline{\alpha X}_{\ell,j}}{1 - \overline{\alpha}_{\ell,j}}. \tag{2}$$

Let $U^{(j)} = \operatorname{med}^-_{1\leq \ell \leq k} V_{\ell,j}$. Zero denominators give signed infinities, with $0/0$ resolved by the shared random permutation (equation (26)). Here m is the RMM rank (one original and $m-1$ resampled components), while r_0 is its rank-tail parameter and gives undercoverage r_0/m. The endpoint U_{m,r_0} is the $(m - r_0)$-th ordered component endpoint, with ties defined in section A.1. Its fixed-sample coverage guarantee is $\mathbb{P}(U_{m,r_0} < \mu) \leq r_0/m$ for every symmetric distribution.

For group size s, define

$$\zeta_s = \mathbb{P}\{\mathrm{Bin}(s, 1/2) < s/4\}, \quad \bar{\zeta}_g = \max\{\zeta_g, \zeta_{g+1}\}, \quad q_g = \tfrac{1}{4} + \tfrac{3}{4}\bar{\zeta}_g, \tag{3}$$

and, for odd k, let $B_k(u) = \mathbb{P}\{\mathrm{Bin}(k, u) \geq (k+1)/2\}$.

Lemma 1 (Tolerant bad-block upper tail)**.** *Suppose $\mathbb{E}|X - \mu|^{1+a} \leq M$ and every MoM group has size g or $g+1$. With $\eta_g = (12M)^{1/(1+a)} g^{-a/(1+a)}$,*

$$\mathbb{P}\{U_{m,r_0} - \mu > 4\eta_g\} \leq (m - r_0)\{B_k(1/4) + B_k(q_g)\}. \tag{4}$$

For $(m, r_0, k, g) = (4, 1, 21, 3)$, $q_3 = 11/32$, and the right-hand side equals $0.2248967607690244\ldots < 1/4$.

For amplification at all group sizes $g \geq 3$, corollary 13 uses the sharper $1/6$ moment bound to give the uniform base upper tail $3\{B_{21}(1/6) + B_{21}(31/96)\} = 0.1348306174\ldots < 1/4$.

The fixed-design search in section H places this tuple on the sample–sign frontier of the stated search box.

The gain over an all-block analysis is a sharper event inclusion. Within a signed group, sign symmetry makes the signed mean independent of the number of negative signs. When the ordinary MoM center is accurate, the signed mean exceeds its lower tolerance, and at least one quarter of the signs are negative, $V_{\ell,j} - \mu \leq 4\eta_g$. A component can therefore be too large only when the ordinary center fails or a strict majority of signed groups lies in the *union* of the other two failure events. This union produces q_g without requiring every group to be balanced. Section A proves the inclusions and the order-statistic factor.

3.3 EXACT-BINOMIAL AMPLIFICATION

Partition a deterministic n-prefix into odd L disjoint bases, compute independent rank-four endpoints $U_1, \ldots, U_L$, and return $U^{\mathrm{CA}} = \operatorname{med}^-(U_1, \ldots, U_L)$. For $p \in (0, 1)$ and $\beta \in (0, 1/2)$, define

$$L(p, \beta) = \min\{L \in 2\mathbb{N} + 1 : B_L(\beta) \leq p\}. \tag{5}$$

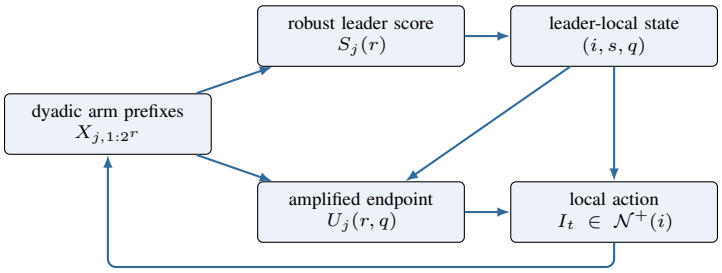

Figure 1: Coupled adaptive state of APR-LOCAL. The same dyadic prefixes generate candidate endpoints and robust leader scores. The leader-local state determines which caches are queried, and each action updates both statistics through the next prefix.

For $x > 0$, define the proof-only accuracy scale

$$h_i(x) = \max\left\{3, \left\lceil \left(\frac{12\, 4^{1+a_i} M_i}{x^{1+a_i}}\right)^{1/a_i} \right\rceil \right\}, \tag{6}$$

and set $L = L(p, 1/4)$. If every within-base MoM group contains at least $h_i(x)$ observations, then

$$\mathbb{P}\{U^{\mathrm{CA}} < \mu_i\} \le p, \qquad \mathbb{P}\{U^{\mathrm{CA}} - \mu_i > x\} \le p. \tag{7}$$

Each failure requires a strict majority of independent base failures, and $L(p, 1/4) = O(\log(1/p))$. Since the L bases partition one prefix and rank four uses three sign vectors per observation, the total sign count is exactly $3n$, independent of p. The unknown quantity $h_i(x)$ appears only in the analysis.

4 Adaptive Prefix Reuse and Graph-Local Learning

4.1 Decision-weighted reuse validity

Fixed-sample coverage concerns one deterministic prefix, whereas regret counts every decision its cache influences. We bound this exposure by weighting each failure by maximum reuse. If arm j has $N_j(t)$ observations, let $r = \lfloor \log_2 N_j(t) \rfloor$ be its sample epoch. The global policy uses $q = \lfloor \log_2 t \rfloor$; the graph policy uses $q = \lfloor \log_2 s \rfloor$ when its current leader has appeared s times. For every integer $z \ge 0$, set

$$w(z) = (z+1)[1 + \log(z+1)]^2, \qquad p_{r,q} = \frac{1}{2^{q+1} w(r) w(q)}. \tag{8}$$

The index $U_j(r, q)$ uses only the first 2^r observations of arm j and is cached for (r, q). A key is accessed at most 2^q times in a global time epoch and at most 2^q times in a leader-local epoch. Arm j belongs to at most $\gamma + 1$ closed neighborhoods, so the corresponding expected bad-cache decision charges are bounded by

$$K \sum_{r,q \ge 0} 2^q\, 2p_{r,q} < \infty, \qquad K(\gamma+1) \sum_{r,q \ge 0} 2^q\, 2p_{r,q} < \infty. \tag{9}$$

The first sum is global and the second graph-local. The charge bounds cache reuse on every trajectory. Taking expectations therefore uses unconditional fixed-key failure probabilities, even though the policy chooses which keys to query adaptively.

The graph-local state computes its robust leader score $S_i(r)$ from the same dyadic prefix 2^r. This score is the MoM with the smallest feasible odd block count at least $8 \log(2/\delta_r)$, where $\delta_r = 2^{-4r}$. If g_r is the minimum group size, then every sufficiently large epoch satisfies

$$\mathbb{P}\{|S_i(r) - \mu_i| > b_i(r)\} \le 2^{-4r}, \qquad b_i(r) = \left(\frac{8M_i}{g_r^{a_i}}\right)^{1/(1+a_i)} \to 0. \tag{10}$$

Only r enters the score construction; a_i, M_i, and $b_i(r)$ are proof quantities.

4.2 GRAPH-LOCAL POLICY AND REGRET

For $x > 0$ and $Q \geq 0$, let $u_j(x, Q)$ be the smallest dyadic threshold beyond which

$$2^r \geq 21 h_j(x) L(p_{r,Q}, 1/4) \tag{11}$$

continues to hold. The estimate $L(p_{r,Q}, 1/4) = O(Q + 1 + \log(e + r))$ gives

$$u_j(x, Q) = O(h_j(x)\,[Q + 1 + \log\{e + h_j(x)(Q + 1)\}])\,. \tag{12}$$

Lemma 18 proves the inversion and persistence across epochs.

APR-LOCAL. Initialize scores at $-\infty$ and pull each arm once. Scores update at counts 2^r to the lower median of interleaved group means, where k_r is the smallest odd integer at least $8\log(2/2^{-4r})$, capped by the largest feasible odd count. Score ties are broken first by pull count and then by arm identifier. After leader i is selected, its local occurrence count s is incremented. Occurrences $s \equiv 1 \pmod{\gamma+1}$ force i; every other occurrence maximizes cached indices over $\mathcal{N}^+(i)$ with $q = \lfloor \log_2 s \rfloor$. Prefixes shorter than $63L(p_{r,q}, 1/4)$ receive index $+\infty$; index ties use the same deterministic rule. The cache key and full specification appear in section D. For node i, define $\mathcal{B}_i = \arg\max_{h \in \mathcal{N}^+(i)} \mu_h$ and $\mu_i^{\mathrm{loc}} = \max_{h \in \mathcal{N}^+(i)} \mu_h$. If $j \notin \mathcal{B}_i$, let $\delta_{i,j} = \mu_i^{\mathrm{loc}} - \mu_j$.

The first counting argument is local. Fix a leader i, choose any proof-only $b_i^{\mathrm{loc}} \in \mathcal{B}_i$, and let $Q_{i,j}(S)$ be the number of nonforced selections of $j \in \mathcal{N}^+(i) \setminus \mathcal{B}_i$ during the first S occurrences of leader i.

Lemma 2 (Adaptive-count-safe local selection). *There is a nondecreasing random process $Z_{i,j}$ such that*

$$Q_{i,j}(S) \leq u_j(\delta_{i,j}/2, \lfloor \log_2 S \rfloor) + Z_{i,j}(S), \qquad \mathbb{E}Z_{i,j}(\infty) < \infty. \tag{13}$$

Once j crosses its threshold, excluding the two cache failures gives the strict comparison

$$U_{b_i^{\mathrm{loc}}} \geq \mu_i^{\mathrm{loc}} > \mu_j + \delta_{i,j}/2 \geq U_j. \tag{14}$$

This rules out selecting j. Each pre-threshold selection raises its sample count. Later selections are charged to $Z_{i,j}$, the 2^q-weighted sum of undercoverage and accuracy-eligible upper deviations over fixed keys. Its nondecreasing envelope has finite expectation by equation (9) and applies pathwise at every adaptive count.

The local-count bound also limits how often a suboptimal arm can remain the leader.

Lemma 3 (Robust ascent). *If $i \neq i^\star$ and $\ell_i(T)$ is its number of leader occurrences by T, then*

$$\sup_{T \geq 1} \mathbb{E}\ell_i(T) < \infty. \tag{15}$$

From local counts to ascent. Proposition 22 separates three requirements: periodic leader pulls, a pathwise bound $\sum_{j \in \mathcal{N}^+(i) \setminus \mathcal{B}_i} Q_{i,j}(s) \leq f_i(s) + Y_i$ with $f_i(s) = o(s)$ and $\mathbb{E}Y_i < \infty$, and vanishing score radii whose error probabilities are summable after weighting epoch r by 2^r. For APR-LOCAL, lemma 2 gives $f_i(s) = O(\log s)$ and $Y_i = \sum_j Z_{i,j}(\infty)$, while equation (10) gives 2^{-4r} score tails. With $\eta = \gamma/[2(\gamma + 1)]$, split occurrences by whether locally best arms have at least ηs total pulls. Above this threshold, both a better neighbor and the forced leader have $\Omega(s)$ samples; weighted score tails bound these occurrences. Below it, at least $\eta s - 2$ nonforced choices are locally suboptimal, forcing $Y_i \geq \eta s/2$ eventually. Integrability bounds this count. The count must be sublinear in s. A global-time bound $O(\log \tau_{i,s})$ need not be $o(s)$, since the time $\tau_{i,s}$ of the s-th occurrence can grow much faster than s. This is why the argument uses a leader-local clock (section E).

Theorem 4 (Neighbor-localized parameter-free regret). *Under equation* (1) *and graph unimodality, for every fixed finite graph and every $T \geq K$,* APR-LOCAL *satisfies*

$$\mathbb{E}R_T \leq C_{\mathrm{init}} + C^{\mathrm{asc}} + \sum_{j \in \mathcal{N}(i^\star)} \Delta_j \left[u_j(\Delta_j/2, \lfloor \log_2 T \rfloor) + C_j^{\mathrm{loc}} \right], \tag{16}$$

where $C_{\mathrm{init}} = \sum_{i \neq i^\star} \Delta_i$ is initialization regret; C^{asc} and C_j^{loc} are the finite, horizon-independent suboptimal-leader regret and neighbor-j bad-cache charge. Consequently,

$$\mathbb{E}R_T = O\left(\sum_{j \in \mathcal{N}(i^\star)} [\Delta_j + M_j^{1/a_j} \Delta_j^{-1/a_j}] \log T \right) + C_{\mathrm{climb}}. \tag{17}$$

Only neighbors of $i^\star$ occur in the growing term; $C_{\mathrm{climb}} < \infty$ collects the remaining finite charges.

For each fixed instance, C_{climb} is independent of T; K, the graph, gaps, and moment parameters enter through the finite sums in equation (176).

Proof sketch. The pathwise decomposition is

$$R_T \le C_{\mathrm{init}} + \Delta_{\max} \sum_{i \neq i^\star} \ell_i(T) + \sum_{j \in \mathcal{N}(i^\star)} \Delta_j Q_{i^\star, j}(\ell_{i^\star}(T)).$$

Here $\Delta_{\max} = \max_i \Delta_i$. The robust-ascent lemma makes the middle term finite in expectation. For the last term, $\mathcal{B}_{i^\star} = \{i^\star\}$, and lemma 2 applies to each suboptimal neighbor in the closed candidate set. This proves equation (16). Finally, equations (6) and (12) imply $\Delta_j h_j(\Delta_j/2) = O(\Delta_j + M_j^{1/a_j} \Delta_j^{-1/a_j})$, which yields equation (17). Section F supplies all constants and the complete regret decomposition. □

Practical configuration. The practical tuple $(m, r_0, k, g_{\min}) = (8, 1, 9, 1)$ also satisfies equation (17). Its runtime minimum group size $g_{\min}$ is one; the upper-tail analysis requires at least six observations per group. Symmetry gives one-sided group tails $1/12$ and, for $g \ge 6$,

$$\mathbb{P}(U_{8,1} - \mu > 4\eta_g) \le 7\{B_9(1/12) + B_9(47/256)\} < 1/8. \tag{18}$$

With $c = 1$ or $K(\gamma{+}1)$ and $h_j^{\mathrm{p}}(x) = \max\{6, h_j(x)\}$, prefixes of length at least $9h_j^{\mathrm{p}}(x)L(p_{r,q}/c, 1/8)$ meet the two-tail interface. Selections before this accuracy threshold enter the deterministic sample-count term, covering the implemented $g_{\min} = 1$ configuration. Section I proves the finite-time and global-clock bounds.

Corollary 5 (Degree-local leading term)**.** *Let*

$$d_\star := |\mathcal{N}(i^\star)|, \qquad A_\star := \max_{j \in \mathcal{N}(i^\star)} \left[\Delta_j + M_j^{1/a_j} \Delta_j^{-1/a_j}\right].$$

Under the conditions of theorem 4,

$$\mathbb{E}R_T = O(d_\star A_\star \log T) + C_{\mathrm{climb}}. \tag{19}$$

For a path, $d_\star \le 2$, so the coefficient of the horizon-dependent term has no explicit factor in the total number of arms K.

Global-clock specialization: APR-UCB**.** Replacing the leader-local state by the global epoch $q = \lfloor \log_2 t \rfloor$, with $r_j = \lfloor \log_2 N_j(t) \rfloor$, and maximizing $U_j(r_j, q)$ over all arms gives APR-UCB. An insufficient prefix receives index $+\infty$; ties use least pull count and then arm identifier.

Theorem 6 (Unstructured all-arm parameter-free regret)**.** *Under equation* (1) *and a unique optimal arm, with no graph assumption, for every $T \ge K$,* APR-UCB *satisfies*

$$\mathbb{E}R_T \le \sum_{j \neq i^\star} \Delta_j \left[u_j(\Delta_j/2, \lfloor \log_2 T \rfloor) + C_j^{\mathrm{glob}}\right], \qquad C_j^{\mathrm{glob}} < \infty. \tag{20}$$

Consequently,

$$\mathbb{E}R_T = O\left(\sum_{j \neq i^\star} [\Delta_j + M_j^{1/a_j} \Delta_j^{-1/a_j}] \log T\right) + C_{\mathrm{glob}}. \tag{21}$$

Here $C_{\mathrm{glob}} < \infty$ collects finite cache and activation charges.

Proof sketch. For a fixed global epoch q, any cache key is accessed on at most 2^q rounds, so the first sum in equation (9) makes its expected decision charge finite. Set $u_j^T = u_j(\Delta_j/2, \lfloor \log_2 T \rfloor)$, and let $Z_j^{\mathrm{glob}}(T)$ be the 2^q-weighted sum of optimal-arm undercoverage events and accuracy-eligible j-arm upper deviations over fixed keys with $q \le \lfloor \log_2 T \rfloor$. This nondecreasing envelope bounds all post-threshold pulls, giving

$$N_j(T) \le u_j^T + Z_j^{\mathrm{glob}}(T), \qquad \sup_T \mathbb{E}Z_j^{\mathrm{glob}}(T) \le \mathbb{E}Z_j^{\mathrm{glob}}(\infty) < \infty.$$

If neither event occurs, the j-index is strictly below the optimal-arm index. Taking expectations and applying equations (6) and (12) proves the theorem. Section C.3 gives the complete pathwise proof. □

4.3 COMPUTATIONAL COST

A newly constructed cache on $n = 2^r$ observations uses $3n$ signs for the certified setting and $7n$ for the practical setting, with $O(n + L \log L)$ time in either case. Dyadic reuse stores $O(K \log^2 T)$ cached endpoints and gives an $O(T \log T)$ sign-operation bound. The implementation also retains $O(T)$ reward observations to reconstruct prefixes. Implementation details and matched-budget diagnostics appear in figure 15 and table 16.

5 SYMMETRY-PRESERVING INFORMATION BENCHMARK

For $p \in (1, 2]$, let $\mathcal{S}_p(M)$ contain distributions symmetric about their means with centered p-th moment at most M. Within this class, a low-frequency characteristic-function argument gives, for $P, Q \in \mathcal{S}_p(M)$ with center separation $\delta \le (8M)^{1/p}$,

$$\mathrm{TV}(P, Q) \ge \frac{3}{4\pi} \frac{\delta}{(8M)^{1/p}}, \qquad \mathrm{KL}(P, Q) \ge \frac{9}{8\pi^2 8^{2/p}} \frac{\delta^2}{M^{2/p}}. \tag{22}$$

Equal-variance Gaussian shifts attain this quadratic order while preserving symmetry, giving the following neighbor lower bound.

Theorem 7 (Gaussian-subfamily neighbor lower bound)**.** *Let $p_j = 1 + a_j \in (1, 2]$, $c_p = \mathbb{E}|Z|^p$ for $Z \sim \mathcal{N}(0, 1)$, and $\sigma_j^2 = (M_j / c_{p_j})^{2/p_j}$. On the graph-unimodal Gaussian subfamily, every uniformly efficient policy (defined in section G) with arm $j \sim \mathcal{N}(\mu_j, \sigma_j^2)$ satisfies*

$$\liminf_{T \to \infty} \frac{\mathbb{E}R_T}{\log T} \ge \sum_{j \in \mathcal{N}(i^\star)} \frac{2\sigma_j^2}{\Delta_j} = \sum_{j \in \mathcal{N}(i^\star)} 2c_{p_j}^{-2/p_j} \frac{M_j^{2/p_j}}{\Delta_j}. \tag{23}$$

Proof sketch. Neighbors with $M_j = 0$ contribute zero; fix j with $M_j > 0$. Raise only neighbor j to mean $\mu_{i^\star} + \varepsilon$. Truncating increasing paths at j, or appending $(i^\star, j)$, preserves graph unimodality with optimum j. For arm laws P_j, Q_j^ε and induced history laws $\mathbb{P}^T, \mathbb{Q}^T$, the trajectory KL chain rule gives

$$\mathrm{KL}(\mathbb{P}^T, \mathbb{Q}^T) = \mathbb{E}N_j(T)\, \mathrm{KL}(P_j, Q_j^\varepsilon). \tag{24}$$

Uniform efficiency makes $\{N_j(T) < T/2\}$ have probability tending to one on the original instance and polynomially small on the alternative. Binary data processing bounds equation (24) below by $(1 - o(1)) \log T$. Substitute Gaussian KL $(\Delta_j + \varepsilon)^2/(2\sigma_j^2)$, let $\varepsilon \downarrow 0$, and sum the gap-weighted bounds. Section G gives the details. □

The Gaussian instance belongs to the assumed moment class for every $p_j = 1 + a_j$, so the bound matches both the logarithmic horizon order and the optimal-neighbor support. The moment–gap powers agree at $a_j = 1$; section G.5 records the gap for $a_j < 1$.

Table 1 compares the growing regret terms. RMM-UCB uses rank $\Theta(p_t^{-1})$ and $n(m_t - 1)$ signs; our amplification uses $\Theta(\log p_t^{-1})$ fixed-rank bases and $3n$ or $7n$ signs. APR-UCB needs no graph, while APR-LOCAL localizes exploration to optimal-arm neighbors.

Table 1: Regret guarantees and exploration sets. Coefficients suppress constants and finite remainders.

Method	Additional structure	Growing arms	Horizon	Per-arm coefficient
RMM-UCB	None	All $j \ne i^\star$	$\log^2 T$	$\Delta_j + M_j^{1/a_j} \Delta_j^{-1/a_j}$
APR-UCB	None	All $j \ne i^\star$	$\log T$	$\Delta_j + M_j^{1/a_j} \Delta_j^{-1/a_j}$
APR-LOCAL	Known graph unimodality	$j \in \mathcal{N}(i^\star)$	$\log T$	$\Delta_j + M_j^{1/a_j} \Delta_j^{-1/a_j}$
Gaussian neighbor LB	Gaussian subfamily	$j \in \mathcal{N}(i^\star)$	$\log T$	$M_j^{2/(1+a_j)} \Delta_j^{-1}$

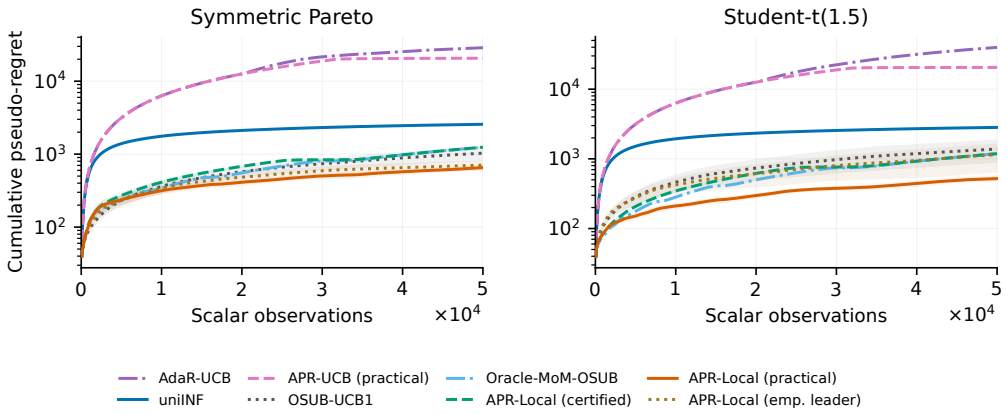

Figure 2: Mean cumulative pseudo-regret on the 63-arm path with pointwise 95% Student-t intervals over 30 paired trials. All eight methods are shown on a logarithmic ordinate; uniINF receives the horizon.

Table 2: Final pseudo-regret on the 63-arm path at $T = 50{,}000$ (mean $\pm$ 95% Student-t half-width) and amplified-cache activation over 30 trials per law. A dash marks methods without this cache; lower regret is better.

Method	Graph	Pareto	Student-$t_{1.5}$	Activation
AdaR-UCB	No	28767.59±526.24	39790.32±299.30	–
uniINF	No	2566.09±46.81	2825.17±56.05	–
Practical APR-UCB	No	20647.23±44.92	20515.69±38.76	30/30, 30/30
OSUB-UCB1	Yes	1029.56±228.37	1383.04±508.69	–
Oracle-MoM-OSUB	Yes	1244.87±28.03	1189.16±18.54	–
APR-LOCAL$_{\text{cert}}$	Yes	1244.85±31.40	1184.87±14.89	30/30, 30/30
APR-LOCAL$_{\text{pract}}$	Yes	**654.35±59.33**	**520.61±53.54**	30/30, 30/30
APR-LOCAL$_{\text{emp}}$	Yes	706.83±106.82	1147.78±509.92	30/30, 30/30

6 EXPERIMENTS

6.1 PROTOCOL

The main study uses a 63-arm path with $\mu_i = -0.04|i - 31|$, $i = 0, \ldots, 62$, and symmetric Pareto (shape 1.5) and Student-$t_{1.5}$ rewards. Both have infinite variance and finite 1.4-th centered moments. Each environment uses 30 independent trials, $T = 50{,}000$, paired pre-generated arm streams, and each policy uses a separate random number stream. Shaded bands show 95% Student-t confidence intervals for the mean pseudo-regret across 30 trials. Paired contrasts use 20,000-resampled bootstrap intervals and Holm-corrected Wilcoxon tests. AdaR-UCB receives the same scalar-observation budget, and the optimal reward satisfies its truncated non-positivity condition. The graph-free baseline uniINF (Chen et al., 2025) receives T and losses $-X$; symmetry at $\mu_{i^\star} = 0$ satisfies its optimal loss condition. Its published constants are used throughout. Full 31 and 15 arm comparisons appear in sections J and K. These nested paths share the optimal neighborhood and noise streams; larger sizes add more distant arms. Distributions and baseline settings are specified in section L.

The certified and practical settings share the leader, cache, and action rules. APR-LOCAL$_{\text{cert}}$ uses $(m, r_0, k, g_{\min}) = (4, 1, 21, 3)$ with budget $p_{r,q}$; APR-LOCAL$_{\text{pract}}$ uses $(8, 1, 9, 1)$ with $p_{r,q}/\lceil K(\gamma + 1)\rceil$. APR-LOCAL$_{\text{emp}}$ changes only the latter's leader score. The matched-budget study in section J.4 runs the practical tuple with $p_{r,q}$.

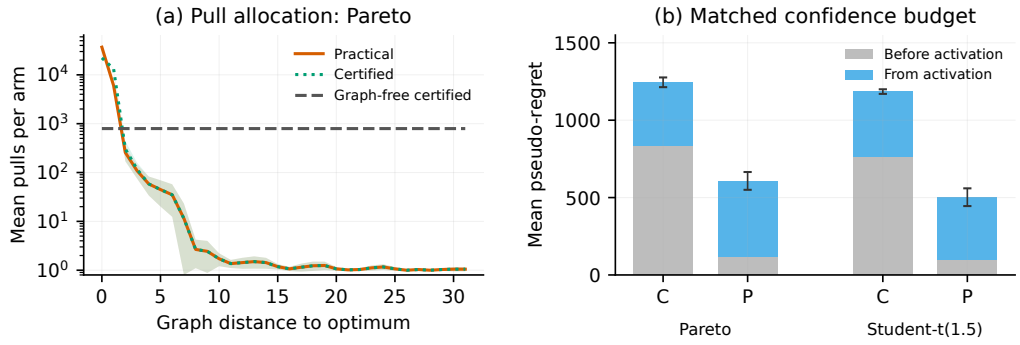

Figure 3: Pull allocation and cache activation on the 63-arm path. (a) Mean pulls per arm at each distance from the optimum, with 95% Student-t intervals over 30 Pareto trials. (b) Regret before and from each policy's first finite endpoint at the same budget $p_{r,q}$. C and P denote certified and practical tuples; error bars give 95% Student-t intervals for total regret.

6.2 Findings

APR-Local$_{\text{pract}}$ has the lowest sample mean among the eight methods on the 63-arm path: 654.35 on Pareto and 520.61 on Student-t (table 2). Relative to uniINF, the mean reductions are 1911.75 (paired bootstrap 95% interval $[1844.17, 1984.70]$) and 2304.55 ($[2238.15, 2374.92]$). Regret outside the optimal closed neighborhood is 171.39 versus 2495.20 on Pareto and 81.42 versus 2747.28 on Student-t. The corresponding neighbor costs are 482.96 versus 70.90 and 439.19 versus 77.89 (table 8).

Replacing the MoM leader by an empirical mean gives a paired regret difference of 52.49 on Pareto ($[-49.00, 154.69]$) and 627.17 on Student-t ($[240.76, 1164.22]$). The optimal neighborhood pull share falls from 97.8% to 95.9% and from 99.1% to 89.0%. Table 6 reports the paired contrasts for the methods under 31 and 63 arms.

6.3 Cache activation and finite-horizon regret

At the same budget $p_{r,q}$, mean regret before first activation is 831.84 versus 118.91 on Pareto and 759.93 versus 99.93 on Student-t, for certified versus practical constants. Most of the observed total gap precedes activation (section J.4).

In figure 3a, the certified and practical local policies keep 97.6–97.8% of Pareto pulls in the optimal closed neighborhood; certified global pulls are nearly uniform. Panel (b) separates the endpoint settings' activation costs from subsequent regret. The 15 arm tree, long horizon, and asymmetric perturbation studies are in sections L.10, L.13 and M.4.

7 Conclusion

Our analysis accounts for how often each cached endpoint is reused. Together with periodic leader pulls and concentration bounds for the dyadic MoM scores, this bounds the expected number of rounds with a suboptimal leader and yields logarithmic regret for both endpoint settings, with leading terms involving only neighbors of the optimal arm. In the paired experiments, most of the regret difference between the certified and practical settings occurs before their first finite endpoints. A separate ablation compares the MoM leader with an empirical-mean leader.

AI USE STATEMENT

During manuscript preparation, the authors used GPT-6 for language polishing, grammar correction, and enhancing textual readability. For the expanded path experiments, GPT-6 also assisted with execution scripts, statistical analysis, and figure preparation. The authors assume full responsibility for the content of this submission.

REPRODUCIBILITY STATEMENT

The appendix specifies every distribution, parameter, confidence budget, baseline, and statistical test. The supplementary code records seeds, checkpoints, pull counts, and package versions. Unit tests, independent rank and FTRL oracles, deterministic replay checks, and fresh-directory source builds verify the RMM constants, leader schedule, controlled graph generators, baseline formulas, cache diagnostics, and reported artifact structure.

REFERENCES

Shubhada Agrawal, Sandeep K. Juneja, and Wouter M. Koolen. Regret minimization in heavy-tailed bandits. In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 26–62. PMLR, 2021. URL https://proceedings.mlr.press/v134/agrawal21a.html.

Sujay Bhatt, Guanhua Fang, Ping Li, and Gennady Samorodnitsky. Nearly optimal Catoni's m-estimator for infinite variance. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1925–1944. PMLR, 2022. URL https://proceedings.mlr.press/v162/bhatt22b.html.

Sébastien Bubeck, Nicolò Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013. doi: 10.1109/TIT.2013.2277869. URL https://arxiv.org/abs/1209.1727.

Yu Chen, Jiatai Huang, Yan Dai, and Longbo Huang. uniINF: Best-of-both-worlds algorithm for parameter-free heavy-tailed MABs. In *International Conference on Learning Representations*, volume 2025, pp. 71762–71798, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/b2c4b7d34b3d96b9dc12f7bce424b7ae-Paper-Conference.pdf.

Richard Combes and Alexandre Proutiere. Unimodal bandits: Regret lower bounds and optimal algorithms. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 521–529. PMLR, 2014. URL https://proceedings.mlr.press/v32/combes14.html.

Yuriy Dorn, Aleksandr Katrutsa, Ilgam Latypov, and Andrey Pudovikov. Fast UCB-type algorithms for stochastic bandits with heavy and super heavy symmetric noise. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, pp. 649–657. International Foundation for Autonomous Agents and Multiagent Systems, 2025. doi: 10.65109/RRLV7904. URL https://www.ifaamas.org/Proceedings/aamas2025/pdfs/p649.pdf.

Gianmarco Genalti and Alberto Maria Metelli. Open problem: Regret minimization in heavy-tailed bandits with unknown distributional parameters. In *Proceedings of the 38th Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 1–5. PMLR, 2025. URL https://proceedings.mlr.press/v291/genalti25a.html.

Gianmarco Genalti and Alberto Maria Metelli. Parameter-free heavy-tailed bandits. *arXiv preprint arXiv:2607.29460*, 2026. doi: 10.48550/arXiv.2607.29460. URL https://arxiv.org/abs/2607.29460.

Gianmarco Genalti, Lupo Marsigli, Nicola Gatti, and Alberto Maria Metelli. (ϵ, u)-adaptive regret minimization in heavy-tailed bandits. In *Proceedings of the 37th Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 1882–1915. PMLR, 2024. URL https://proceedings.mlr.press/v247/genalti24a.html.

Yutian Gou, Jinfeng Yi, and Lijun Zhang. Stochastic graphical bandits with heavy-tailed rewards. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 734–744. PMLR, 2023. URL https://proceedings.mlr.press/v216/gou23a.html.

Yu-Chun Kao, Min Xu, and Cun-Hui Zhang. Choosing the p in l_p loss: Adaptive rates for symmetric mean estimation. In *Proceedings of the 37th Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 2795–2839. PMLR, 2024. URL https://proceedings.mlr.press/v247/kao24a.html.

Nikita Kornilov, Yuriy Dorn, Aleksandr Lobanov, Nikolay Kutuzov, Innokentiy Shibaev, Eduard Gorbunov, Alexander Nazin, and Alexander Gasnikov. Median clipping for zeroth-order non-smooth convex optimization and multi-armed bandit problem with heavy-tailed symmetric noise. *arXiv preprint arXiv:2402.02461*, 2024. doi: 10.48550/arXiv.2402.02461. URL https://arxiv.org/abs/2402.02461.

Kyungjae Lee, Hongjun Yang, Sungbin Lim, and Songhwai Oh. Optimal algorithms for stochastic multi-armed bandits with heavy tailed rewards. In *Advances in Neural Information Processing Systems*, volume 33, pp. 8452–8462. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/607bc9ebe4abfcd65181bfbef6252830-Abstract.html.

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019. doi: 10.1007/s10208-019-09427-x.

Ambrus Tamás, Szabolcs Szentpéteri, and Balázs Csáji. Data-driven confidence intervals with optimal rates for the mean of heavy-tailed distributions. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3439–3447. PMLR, 2024. URL https://proceedings.mlr.press/v238/tamas24a.html.

Ambrus Tamás, Szabolcs Szentpéteri, and Balázs Csáji. Data-driven upper confidence bounds with near-optimal regret for heavy-tailed bandits. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2116–2124. PMLR, 2025. URL https://proceedings.mlr.press/v258/tamas25a.html.

Cindy Trinh, Emilie Kaufmann, Claire Vernade, and Richard Combes. Solving bernoulli rank-one bandits with unimodal thompson sampling. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117 of *Proceedings of Machine Learning Research*, pp. 862–889. PMLR, 2020. URL https://proceedings.mlr.press/v117/trinh20a.html.

Hongjian Wang and Aaditya Ramdas. Huber-robust confidence sequences. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 9662–9679. PMLR, 2023. URL https://proceedings.mlr.press/v206/wang23p.html.

A Complete Confidence-Amplification Proofs

This appendix states the fixed-sample coverage guarantee for the one-sided RMM endpoint and proves the upper-tail and amplification bounds used by the bandit analysis. Throughout, $\log$ is the natural logarithm and $\log_2$ has base two.

A.1 Fixed-sample one-sided RMM coverage

For an i.i.d. sample $D = (X_1, \dots, X_n)$ of deterministic size n from a distribution symmetric about μ, let $R(\theta)$ denote the randomized rank statistic of Tamás et al. (2025). We fix its tie convention explicitly. For labeled values A_j, A_h and an independent uniform permutation π of $\{0, \dots, m-1\}$, write $A_j \prec_\pi A_h$ if $A_j < A_h$, or if $A_j = A_h$ and $\pi(j) > \pi(h)$. Reversing the permutation converts this to the increasing-permutation convention and preserves its uniform law. The accepted confidence set and its upper endpoint are

$$\Theta_n = \{\theta : R(\theta) \le m - r\}, \qquad U = \sup \Theta_n = U_{(m-r)}, \tag{25}$$

where $U_1, \dots, U_{m-1}$ are the component endpoints and $U_{(1)} \prec_\pi \cdots \prec_\pi U_{(m-1)}$ is their order under the independent tie-breaking permutation π. We write the source rank parameter r as r_0, its component U_j as $U^{(j)}$, and its endpoint $U_{(m-r)}$ as $U_{m,r_0}(D)$. Thus our construction in section 3 is the same fixed-sample confidence endpoint with the equivalent tie convention above.

Extended-real intersections. For each group ℓ and component j, put

$$A_{\ell,j} = \widehat{\mu}_0 - \overline{\alpha X}_{\ell,j}, \qquad D_{\ell,j} = 1 - \overline{\alpha}_{\ell,j} \ge 0.$$

The ratio in equation (2) has the following complete definition:

$$V_{\ell,j} = \begin{cases} A_{\ell,j}/D_{\ell,j}, & D_{\ell,j} > 0, \\ +\infty, & D_{\ell,j} = 0,\ A_{\ell,j} > 0, \\ -\infty, & D_{\ell,j} = 0,\ A_{\ell,j} < 0, \\ +\infty, & D_{\ell,j} = A_{\ell,j} = 0,\ \pi(j) > \pi(0), \\ -\infty, & D_{\ell,j} = A_{\ell,j} = 0,\ \pi(j) < \pi(0). \end{cases} \tag{26}$$

To check the boundary interpretation, define $S_0(\theta) = \widehat{\mu}_0 - \theta$ and $S_{\ell,j}(\theta) = \overline{\alpha X}_{\ell,j} - \overline{\alpha}_{\ell,j}\theta$. Their difference is $S_0(\theta) - S_{\ell,j}(\theta) = A_{\ell,j} - D_{\ell,j}\theta$. When $D_{\ell,j} > 0$, the set $\{\theta : S_{\ell,j}(\theta) \prec_\pi S_0(\theta)\}$ is a lower half-line ending at $A_{\ell,j}/D_{\ell,j}$. When $D_{\ell,j} = 0$, it is all of $\mathbb{R}$ if $A_{\ell,j} > 0$, or if $A_{\ell,j} = 0$ and $\pi(j) > \pi(0)$; otherwise it is empty. These are precisely the infinite endpoints in equation (26). Each $U^{(j)}$ is the lower median over $\ell = 1, \dots, k$, using the same permutation for all groups in that base. The component order statistic then yields equation (25), with $\sup \varnothing = -\infty$ and $\sup \mathbb{R} = +\infty$.

Proposition 8 (Fixed-sample RMM undercoverage (Tamás et al., 2025))**.** *For integers $1 \le r_0 < m$ and every deterministic sample size n,*

$$\mathbb{P}\{U_{m,r_0}(D) < \mu\} \le \frac{r_0}{m}. \tag{27}$$

The probability is over both the sample and the auxiliary RMM randomization.

The cited rank theorem gives $\mathbb{P}\{\mu \in \Theta_n\} = 1 - r_0/m$. The set Θ_n equals either $(-\infty, U_{m,r_0})$ or $(-\infty, U_{m,r_0}]$, as determined by π, with the empty-set and full-line interpretations at infinite endpoints. Hence

$$\{U_{m,r_0} < \mu\} \subseteq \{\mu \notin \Theta_n\}, \tag{28}$$

and equation (27) follows. Every later invocation fixes an arm and a deterministic prefix length, exactly matching this fixed-sample statement.

A.2 Moment and sign preliminaries

Lemma 9 (Symmetric sign independence)**.** *Let W be symmetric about zero and let α be an independent Rademacher sign. Then $\alpha W \overset{d}{=} W$ and αW is independent of α. If $(W_h)_{h=1}^s$ and $(\alpha_h)_{h=1}^s$ are independent coordinatewise copies, then the vector $(\alpha_h W_h)_{h=1}^s$ is independent of the entire sign vector $(\alpha_h)_{h=1}^s$.*

Proof. For a Borel set A and $u \in \{-1, 1\}$,

$$\mathbb{P}(\alpha W \in A, \alpha = u) = \mathbb{P}(\alpha = u)\mathbb{P}(uW \in A) \tag{29}$$

$$= \mathbb{P}(\alpha = u)\mathbb{P}(W \in A), \tag{30}$$

where the second equality uses $W \overset{d}{=} -W$. Summing equation (30) over u gives $\alpha W \overset{d}{=} W$, while the factorization itself proves independence. For vectors, apply the scalar factorization to a rectangle $A_1 \times \cdots \times A_s$ and a fixed sign vector. Independence across coordinates makes the joint probability the product of the scalar probabilities. A monotone-class argument extends the result from rectangles to the product Borel sigma-field. □

Lemma 10 (One-group moment bound)**.** *Let $Y_1, \ldots, Y_s$ be independent, mean-zero random variables with $\mathbb{E}|Y_h|^{1+a} \le M$, where $a \in (0, 1]$. Then*

$$\mathbb{E}\left|\frac{1}{s}\sum_{h=1}^{s} Y_h\right|^{1+a} \le \frac{2M}{s^a}. \tag{31}$$

Consequently, for $\eta_g = (12M)^{1/(1+a)} g^{-a/(1+a)}$ and every $s \ge g$,

$$\mathbb{P}\left\{\frac{1}{s}\sum_{h=1}^{s} Y_h > \eta_g\right\} \le \mathbb{P}\left\{\left|\frac{1}{s}\sum_{h=1}^{s} Y_h\right| > \eta_g\right\} \le \frac{1}{6} \le \frac{1}{4}. \tag{32}$$

Proof. Put $p = 1 + a \in (1, 2]$. The von Bahr–Esseen inequality for independent mean-zero variables gives

$$\mathbb{E}\left|\sum_{h=1}^{s} Y_h\right|^{p} \le 2\sum_{h=1}^{s} \mathbb{E}|Y_h|^p \le 2sM. \tag{33}$$

Dividing equation (33) by $s^p = s^{1+a}$ proves equation (31). Markov's inequality and $s \ge g$ then give

$$\mathbb{P}\left\{\left|s^{-1}\sum_{h=1}^{s} Y_h\right| > \eta_g\right\} \le \frac{2M/s^a}{(12M)g^{-a}} = \frac{1}{6}\left(\frac{g}{s}\right)^a \le \frac{1}{6}. \tag{34}$$

□

Lemma 11 (Median events and binomial domination)**.** *Let k be odd and write $m_k = (k+1)/2$. For real numbers $z_1, \ldots, z_k$ and a threshold c,*

$$\mathrm{med}^-(z_1, \ldots, z_k) > c \Longrightarrow \sum_{\ell=1}^{k} \mathbf{1}\{z_\ell > c\} \ge m_k, \tag{35}$$

$$\mathrm{med}^-(z_1, \ldots, z_k) < c \Longrightarrow \sum_{\ell=1}^{k} \mathbf{1}\{z_\ell < c\} \ge m_k. \tag{36}$$

Moreover, if $B_1, \ldots, B_k$ are independent Bernoulli variables with $\mathbb{P}(B_\ell = 1) \le \beta$, then, for every integer v,

$$\mathbb{P}\left\{\sum_{\ell=1}^{k} B_\ell \ge v\right\} \le \mathbb{P}\{\mathrm{Bin}(k, \beta) \ge v\}. \tag{37}$$

Proof. Let $z_{(1)} \le \cdots \le z_{(k)}$ be the order statistics. Because $\mathrm{med}^-(z_1, \ldots, z_k) = z_{(m_k)}$, the inequality $z_{(m_k)} > c$ leaves the $k - m_k + 1 = m_k$ entries $z_{(m_k)}, \ldots, z_{(k)}$ above c, proving equation (35). The lower-tail statement follows in the same way from the first m_k entries.

For equation (37), let $\pi_\ell = \mathbb{P}(B_\ell = 1)$ and take independent $V_1, \ldots, V_k \sim \mathrm{Unif}[0, 1]$. The coupled variables $\widetilde{B}_\ell = \mathbf{1}\{V_\ell \le \pi_\ell\}$ have the same joint law as $(B_\ell)_\ell$, whereas $B'_\ell = \mathbf{1}\{V_\ell \le \beta\}$ are i.i.d. Bernoulli(β). Since $\pi_\ell \le \beta$, $\widetilde{B}_\ell \le B'_\ell$ for every ℓ on every coupled outcome. Therefore $\sum_\ell \widetilde{B}_\ell \le \sum_\ell B'_\ell$ pathwise, and taking upper-tail probabilities proves the claim. □

Lemma 12 (Ordinary and signed MoM tails)**.** *Split a sample into an odd number k of independent groups, each of size g or $g+1$. Under the assumptions of lemma 10, the lower median $\widehat{\mu}_0$ of the ordinary group means satisfies*

$$\mathbb{P}\{\widehat{\mu}_0 - \mu > \eta_g\} \le B_k(1/4). \tag{38}$$

The same bound holds for the lower median of group means formed from independent signed centered observations $\alpha_h(X_h - \mu)$.

Proof. For each group, equation (32) bounds the relevant one-sided failure probability by $1/4$. The group events are independent because the data groups are disjoint. By equations (35) and (37), a median failure therefore has probability at most $B_k(1/4)$, proving equation (38). For signed observations, lemma 9 preserves the centered distribution and the moment bound, so the same argument applies. □

A.3 Detailed tolerant bad-block proof

Proof of lemma 1. Step 1: exact groupwise algebra. Fix one component j and one MoM group G_ℓ of size $s_\ell \in \{g, g+1\}$. Let

$$Z_{\ell,j} = \sum_{h \in G_\ell} \mathbf{1}\{\alpha_{h,j} = -1\}. \tag{39}$$

Because $\overline{\alpha}_{\ell,j} = 1 - 2Z_{\ell,j}/s_\ell$, the denominator in equation (2) is

$$1 - \overline{\alpha}_{\ell,j} = \frac{2Z_{\ell,j}}{s_\ell}. \tag{40}$$

Here $\overline{\alpha}_{\ell,j} = s_\ell^{-1} \sum_{h \in G_\ell} \alpha_{h,j}$ and $\overline{\alpha X}_{\ell,j} = s_\ell^{-1} \sum_{h \in G_\ell} \alpha_{h,j} X_h$, which makes the barred notation in equation (2) explicit. Centering the numerator at μ gives the exact identity

$$V_{\ell,j} - \mu = \frac{(\widehat{\mu}_0 - \mu) - \overline{\alpha(X-\mu)}_{\ell,j}}{1 - \overline{\alpha}_{\ell,j}}. \tag{41}$$

Define the common ordinary-center failure

$$C = \{\widehat{\mu}_0 - \mu > \eta_g\}, \tag{42}$$

and, for the signed group, define

$$D_{\ell,j} = \{\overline{\alpha(X-\mu)}_{\ell,j} < -\eta_g\}, \qquad H_{\ell,j} = \{Z_{\ell,j} < s_\ell/4\}. \tag{43}$$

On $C^c \cap D_{\ell,j}^c \cap H_{\ell,j}^c$, the numerator in equation (41) is at most $2\eta_g$, while equation (40) is at least $1/2$. Therefore

$$V_{\ell,j} - \mu \le 4\eta_g. \tag{44}$$

The extended-real convention when the denominator vanishes is irrelevant on this event because $H_{\ell,j}^c$ makes the denominator positive.

Step 2: probability of one bad signed group. By lemmas 9 and 10, $\mathbb{P}(D_{\ell,j}) \le 1/4$. The signed-observation vector is also independent of the sign vector, so $D_{\ell,j}$ is independent of $H_{\ell,j}$. Since $\mathbb{P}(H_{\ell,j}) \le \bar{\zeta}_g$, we obtain

$$\mathbb{P}(D_{\ell,j} \cup H_{\ell,j}) = \mathbb{P}(D_{\ell,j}) + \{1 - \mathbb{P}(D_{\ell,j})\}\mathbb{P}(H_{\ell,j}) \tag{45}$$

$$\le \bar{\zeta}_g + \mathbb{P}(D_{\ell,j})(1 - \bar{\zeta}_g) \tag{46}$$

$$\le \bar{\zeta}_g + \frac{1}{4}(1 - \bar{\zeta}_g) = \frac{1}{4} + \frac{3}{4}\bar{\zeta}_g = q_g. \tag{47}$$

The first inequality substitutes the upper bound on $\mathbb{P}(H_{\ell,j})$. The resulting affine function of $\mathbb{P}(D_{\ell,j})$ has slope $1 - \bar{\zeta}_g \ge 0$, so its maximum over $[0, 1/4]$ is attained at $1/4$. Specifically, $D_{\ell,j}$ is measurable with respect to $\sigma(\alpha_{h,j}(X_h - \mu) : h \in G_\ell)$, while $H_{\ell,j}$ is measurable with respect to $\sigma(\alpha_{h,j} : h \in G_\ell)$. Lemma 9 makes these sigma-fields independent. For different groups, both sigma-field pairs use

disjoint observations and disjoint auxiliary signs. Hence the indicators $\mathbf{1}\{D_{\ell,j} \cup H_{\ell,j}\}$ are independent across ℓ, although their success probabilities can differ when group sizes are g and $g+1$.

Step 3: from groups to one component. If C^c holds and the component endpoint $U^{(j)} = \mathrm{med}^-(V_{1,j}, \ldots, V_{k,j})$ exceeds $\mu + 4\eta_g$, then equation (44) fails in a strict majority of groups. Hence

$$\{U^{(j)} - \mu > 4\eta_g\} \subseteq C \cup \left\{ \sum_{\ell=1}^{k} \mathbf{1}\{D_{\ell,j} \cup H_{\ell,j}\} \geq \frac{k+1}{2} \right\}. \tag{48}$$

Using lemma 12 and equation (47) together with the binomial domination in lemma 11 in equation (48) yields

$$\mathbb{P}\{U^{(j)} - \mu > 4\eta_g\} \leq B_k(1/4) + B_k(q_g). \tag{49}$$

Step 4: from component endpoints to the RMM order statistic. There are $m-1$ component endpoints. Fix any deterministic subset $J \subseteq \{1, \ldots, m-1\}$ with $|J| = m - r_0$. If the $(m-r_0)$-th ordered component endpoint is larger than $\mu + 4\eta_g$, at most $m - r_0 - 1$ components are at most that threshold. Consequently, every subset of $m - r_0$ components, including J, contains at least one component above the threshold. Thus

$$\{U_{m,r_0} - \mu > 4\eta_g\} \subseteq \bigcup_{j \in J} \{U^{(j)} - \mu > 4\eta_g\}. \tag{50}$$

The union bound and equation (49) prove equation (4). Independence across components is neither assumed nor needed. The randomized tie permutation used by RMM also causes no extra term: all inclusions above use the strict event ">", so a component equal to the threshold is never counted as an upper-tail failure, regardless of its tie rank. □

Corollary 13 (Certified base arithmetic)**.** *For* $(m, r_0, k, g) = (4, 1, 21, 3)$*,*

$$q_3 = \frac{11}{32}, \qquad 3\{B_{21}(1/4) + B_{21}(11/32)\} = 0.2248967607690244\ldots < \frac{1}{4}. \tag{51}$$

Moreover, for $(m, r_0, k) = (4, 1, 21)$*, any integer* $g \geq 3$*, and independent MoM groups whose sizes are all at least g,*

$$\mathbb{P}\{U_{4,1} - \mu > 4\eta_g\} \leq 3\{B_{21}(1/6) + B_{21}(31/96)\} = 0.1348306174115538\ldots < \frac{1}{4}. \tag{52}$$

Proof. Step 1: verify the design-size calculation. For $Z \sim \mathrm{Bin}(3, 1/2)$, $\mathbb{P}(Z < 3/4) = \mathbb{P}(Z = 0) = 1/8$. For $Z' \sim \mathrm{Bin}(4, 1/2)$, $\mathbb{P}(Z' < 1) = 1/16$. Therefore $\bar{\zeta}_3 = 1/8$ and $q_3 = 1/4 + (3/4)(1/8) = 11/32$. Substitution into

$$B_{21}(u) = \sum_{v=11}^{21} \binom{21}{v} u^v (1-u)^{21-v} \tag{53}$$

gives equation (51). This calculation concerns group sizes three and four. The uniform claim below uses the sharper moment bound rather than monotonicity in the group size.

Step 2: control sign imbalance at every size. For an integer $s \geq 3$, the strict inequality in the definition of ζ_s gives the exact formula

$$\zeta_s = 2^{-s} \sum_{v=0}^{\lceil s/4 \rceil - 1} \binom{s}{v}. \tag{54}$$

For $3 \leq s \leq 13$, exact evaluation yields

$$(\zeta_3, \ldots, \zeta_8) = \left(\frac{1}{8}, \frac{1}{16}, \frac{3}{16}, \frac{7}{64}, \frac{1}{16}, \frac{9}{256} \right),$$
$$(\zeta_9, \ldots, \zeta_{13}) = \left(\frac{23}{256}, \frac{7}{128}, \frac{67}{2048}, \frac{79}{4096}, \frac{189}{4096} \right).$$

Every entry is at most $3/16$. For $s \geq 14$, Hoeffding's inequality for $Z \sim \mathrm{Bin}(s, 1/2)$ gives

$$\zeta_s \leq \mathbb{P}\{Z - s/2 \leq -s/4\} \leq \exp(-s/8) \leq \exp(-7/4) < \frac{3}{16}. \tag{55}$$

Consequently, $\zeta_s \le 3/16$ for every integer $s \ge 3$.

Step 3: retain the sharper group moment bound. Fix a group with size $s_\ell \ge g \ge 3$ and use the events $C, D_{\ell,j}, H_{\ell,j}$ from equations (42) and (43). By equation (32), its ordinary and signed one-sided moment failures each have probability at most $1/6$ at tolerance η_g. The independence of the ordinary groups and lemma 11 therefore give

$$\mathbb{P}(C) \le B_{21}(1/6). \tag{56}$$

By lemma 9, $D_{\ell,j}$ is independent of $H_{\ell,j}$. Writing $d = \mathbb{P}(D_{\ell,j}) \le 1/6$ and $z = \mathbb{P}(H_{\ell,j}) = \zeta_{s_\ell} \le 3/16$, we obtain

$$\mathbb{P}(D_{\ell,j} \cup H_{\ell,j}) = d + (1-d)z \le \frac{1}{6} + \frac{5}{6}\frac{3}{16} = \frac{31}{96}. \tag{57}$$

Here $d + (1-d)z$ is nondecreasing in each argument on $[0,1]^2$. For a fixed component j, the union indicators remain independent across groups, since the groups use disjoint data and auxiliary signs. The component inclusion equation (48) and binomial domination then imply

$$\mathbb{P}\{U^{(j)} - \mu > 4\eta_g\} \le B_{21}(1/6) + B_{21}(31/96). \tag{58}$$

The rank-four union bound equation (50) contributes the factor $m - r_0 = 3$. Finally, exact evaluation of equation (53) at the two rational arguments proves equation (52). □

A.4 Detailed exact-binomial amplification proof

Lemma 14 (Fixed-prefix confidence guarantee)**.** *Fix an arm i, a deterministic prefix, $x > 0$, and $p \in (0,1)$. Split the prefix into $L = L(p, 1/4)$ disjoint base samples. If every within-base MoM group contains at least $h_i(x)$ observations, then equation* (7) *holds.*

Proof. Step 1: verify both one-base tails. For a base block, proposition 8 with $(m, r_0) = (4, 1)$ gives

$$\mathbb{P}(U_\ell < \mu_i) \le \frac{1}{4}. \tag{59}$$

Next, equation (6) implies

$$h_i(x)^{a_i} \ge \frac{12\, 4^{1+a_i} M_i}{x^{1+a_i}}. \tag{60}$$

Raising $4\eta_{h_i(x)}$ to power $1 + a_i$ and using equation (60) yields

$$\{4\eta_{h_i(x)}\}^{1+a_i} = \frac{4^{1+a_i} 12 M_i}{h_i(x)^{a_i}} \le x^{1+a_i}, \tag{61}$$

so $4\eta_{h_i(x)} \le x$. Every group has size at least $h_i(x) \ge 3$, so equation (52) applies with $g = h_i(x)$ and gives

$$\mathbb{P}(U_\ell - \mu_i > x) \le \mathbb{P}\{U_\ell - \mu_i > 4\eta_{h_i(x)}\} \le 0.1348306174115538\ldots < \frac{1}{4}. \tag{62}$$

Step 2: verify independence of the amplified bases. The L base endpoints are independent because their observations, signs, and tie permutations belong to disjoint blocks and independent auxiliary streams. This independence concerns a deterministic prefix partition, with no conditioning on an action time or adaptive arm count.

Step 3: amplify each side by the exact binomial tail. For odd L, equation (36) shows that the event $\mathrm{med}^-(U_1, \ldots, U_L) < \mu_i$ requires at least $(L+1)/2$ events from equation (59). Applying equation (37) to their independent indicators gives

$$\mathbb{P}\{U^{\mathrm{CA}} < \mu_i\} \le B_L(1/4) \le p. \tag{63}$$

Similarly, equation (35) shows that $U^{\mathrm{CA}} - \mu_i > x$ requires a strict majority of the independent events in equation (62). A second application of equation (37) yields

$$\mathbb{P}\{U^{\mathrm{CA}} - \mu_i > x\} \le B_L(1/4) \le p, \tag{64}$$

which completes equation (7). □

Corollary 15 (Two-sided fixed-prefix interval)**.** *Fix a deterministic prefix and construct amplified endpoints $U^+ = U^{\mathrm{CA}}(X; p/2)$ and $U^- = U^{\mathrm{CA}}(-X; p/2)$ with the same certified base design and independent auxiliary randomizations. Define*

$$I_p = [-U^-, U^+]. \tag{65}$$

If every within-base group has at least $h_i(x)$ observations for each construction, then

$$\mathbb{P}\{\mu_i \notin I_p\} \le p, \tag{66}$$

$$\mathbb{P}\{I_p \not\subseteq [\mu_i - x, \mu_i + x]\} \le p. \tag{67}$$

The two endpoints together generate exactly $6n$ Rademacher signs on a prefix of length n.

Proof. Applying lemma 14 to X gives $\mathbb{P}(U^+ < \mu_i) \le p/2$ and $\mathbb{P}(U^+ - \mu_i > x) \le p/2$. The centered variable $-X - (-\mu_i)$ has the same moment bound and is symmetric, so the same lemma applied to $-X$ yields

$$\mathbb{P}(-U^- > \mu_i) \le p/2, \qquad \mathbb{P}(\mu_i + U^- > x) \le p/2. \tag{68}$$

The first event from each construction covers the two ways that μ_i can leave I_p; the second event from each construction covers the two ways that I_p can extend beyond $[\mu_i - x, \mu_i + x]$. Two union bounds prove equations (66) and (67). Each endpoint uses $3n$ signs by lemma 17, giving the stated total. □

Remark 16 (Generic amplification scheme). Suppose a one-sided base endpoint on a deterministic sample block obeys

$$\mathbb{P}(U < \mu) \le \beta_-, \qquad \mathbb{P}(U - \mu > x) \le \beta_+, \qquad \max\{\beta_-, \beta_+\} < \tfrac{1}{2}. \tag{69}$$

For independent bases, their lower median reaches error probability p on both sides with any odd

$$L \ge \max\{L(p, \beta_-), L(p, \beta_+)\}. \tag{70}$$

The two binomial-dominance steps in lemma 14 give the same conclusion. A base using $m - 1$ sign transforms per observation produces $(m - 1)n$ signs after disjoint-block amplification. This generic form gives the tuple trade-off audited in section H.

Lemma 17 (Amplification size and randomization cost)**.** *For $p \in (0, 1)$,*

$$L(p, 1/4) \le \max\{1, 8\log(1/p) + 3\}. \tag{71}$$

If a prefix of length n is partitioned among the L bases, the amplified rank-four construction generates exactly $3n$ Rademacher signs.

Proof. If $Y \sim \mathrm{Bin}(L, 1/4)$ and L is odd, Hoeffding's inequality gives

$$B_L(1/4) = \mathbb{P}\{Y \ge (L+1)/2\} \le \mathbb{P}\{Y - L/4 \ge L/4\} \le e^{-L/8}. \tag{72}$$

Choose the smallest odd integer no smaller than $\max\{1, 8\log(1/p)\}$. It is at most the real threshold plus two and satisfies equation (72), proving equation (71). Rank four uses $m - 1 = 3$ independent sign transforms for each observation in its base sample. Since the base samples partition the prefix, the count is $3\sum_{\ell=1}^{L} n_\ell = 3n$. □

The floor operations preserve the required group size. A nearly equal split of n observations into L bases has minimum base size $\lfloor n/L \rfloor$. If $n \ge 21 h_i(x) L$, then $\lfloor n/L \rfloor \ge 21 h_i(x)$. Splitting any such base into 21 nearly equal groups has minimum group size

$$\left\lfloor \frac{\lfloor n/L \rfloor}{21} \right\rfloor \ge h_i(x). \tag{73}$$

Thus the accuracy condition used in lemma 14 holds for every base and every one of its 21 groups.

B ASSUMPTION–USE MAP

Table 3 identifies where each assumption first enters the proof and records its downstream role. Statistical symmetry and graph unimodality enter different branches of the analysis: symmetry constructs the index, while unimodality localizes its growing regret term.

Table 3: Assumptions and their exact proof roles.

Assumption	First use	Downstream role
I.i.d. rewards within each arm; arm streams mutually independent	Lemmas 10 and 12	Independent MoM groups, disjoint amplification bases, cache-event sums, and the bandit trajectory KL chain rule.
Exact symmetry about each arm mean	Proposition 8 and lemma 9	Fixed-prefix undercoverage, independence of signed values and signs, tolerant upper tails, and the two-sided corollary.
Finite centered $(1 + a_i)$-moment for some $a_i \in (0, 1]$, $M_i < \infty$	Lemma 10	Accuracy scale $h_i(x)$, dyadic activation, robust-leader concentration, and the moment–gap coefficient in the upper bound.
Independent auxiliary signs and tie permutations	Proposition 8 and lemma 14	Exact rank coverage, independent base failures, deterministic cache coupling, and exact randomization accounting.
Deterministic dyadic arm prefixes and epoch-indexed caches	Lemma 14 and section C	Fixed-sample validity under adaptive interaction and reuse-weighted summability of decision charges.
Unique optimal arm	Theorem 6	Positive global gaps, strict post-threshold index comparisons, and the regret decomposition.
Known finite undirected graph and graph-unimodal mean vector	Lemmas 2 and 3	Closed-neighborhood candidate sets, robust ascent, optimal-neighbor localization, and graph-valid lower-bound alternatives.
Uniform efficiency of the evaluated policy	Section G	Change-of-measure probability separation used only for the asymptotic neighbor lower bound.
Gaussian rewards with calibrated variances	Theorem 7	A graph-unimodal subfamily attaining the quadratic symmetric information order and the explicit lower-bound coefficient.

C SEQUENTIAL VALIDITY AND LOCAL COUNTS

C.1 A PERSISTENT DYADIC ACTIVATION THRESHOLD

Lemma 18 (Dyadic activation rate)**.** *Fix $x > 0$, an arm j, and $Q \geq 0$. Let $u_j(x, Q)$ be the smallest dyadic number $2^{r_{\mathrm{act}}}$ such that*

$$2^r \geq 21 h_j(x) L(p_{r,Q}, 1/4) \quad \textit{for every } r \geq r_{\mathrm{act}}. \tag{74}$$

Then the explicit universal choice $C = 2^{21}$ satisfies

$$u_j(x, Q) \leq C h_j(x) \left[Q + 1 + \log\{e + h_j(x)(Q + 1)\}\right]. \tag{75}$$

Proof. Write $H = h_j(x) \geq 3$ and $A = Q + 1 \geq 1$. From equation (8),

$$\log \frac{1}{p_{r,Q}} = A \log 2 + \log w(r) + \log w(Q). \tag{76}$$

For every integer $z \geq 0$,

$$\log w(z) = \log(z + 1) + 2 \log\{1 + \log(z + 1)\} \leq 3 \log(e + z). \tag{77}$$

Since $\log(e + Q) \leq Q + 1 = A$ and $\log 2 < 1$, we have

$$\log \frac{1}{p_{r,Q}} \leq A + 3 \log(e + r) + 3A \tag{78}$$

$$= 4A + 3 \log(e + r). \tag{79}$$

Using equation (71) and $A \geq 1$ therefore yields the explicit bound

$$L(p_{r,Q}, 1/4) \leq 8\{4A + 3 \log(e + r)\} + 3 \leq 35\{A + \log(e + r)\}. \tag{80}$$

Define

$$D = A + \log(e + HA), \qquad n_0 = 2^{20}HD, \qquad r_{\mathrm{act}} = \lceil \log_2 n_0 \rceil. \tag{81}$$

Because $A \geq 1$ and $H \geq 3$, we have $D \geq 1$ and $A \leq D$. Moreover,

$$\log H \leq \log(HA) \leq \log(e + HA) \leq D, \qquad \log D \leq D - 1 \leq D. \tag{82}$$

Hence, using $1/\log 2 < 3/2$ and $D \geq 1$,

$$r_{\mathrm{act}} \leq 1 + \log_2 n_0 = 21 + \frac{\log H + \log D}{\log 2} \leq 21D + \frac{2D}{\log 2} < 24D \leq 25D, \tag{83}$$

$$\log(e + r_{\mathrm{act}}) \leq \log\{(e + 25)D\} = \log(e + 25) + \log D < 4D + D = 5D. \tag{84}$$

For the second line, $e + r_{\mathrm{act}} \leq e + 25D \leq (e + 25)D$ and $\log(e + 25) < 4 \leq 4D$. In particular, $A + \log(e + r_{\mathrm{act}}) \leq 6D$, and equation (80) gives the base case

$$21HL(p_{r_{\mathrm{act}},Q}, 1/4) \leq 21 \cdot 35H\{A + \log(e + r_{\mathrm{act}})\} \tag{85}$$

$$\leq 21 \cdot 35 \cdot 6\, HD = 4410HD \tag{86}$$

$$< n_0 \leq 2^{r_{\mathrm{act}}}, \tag{87}$$

where $2^{20} > 4410$.

For persistence, consider the function on nonnegative integers

$$f(r) = \frac{2^r}{A + \log(e + r)} \tag{88}$$

and put $D_r = A + \log(e + r) \geq 1$. Then

$$\frac{f(r+1)}{f(r)} = \frac{2D_r}{D_r + \log(1 + 1/(e + r))} > 1, \tag{89}$$

because $\log(1 + 1/(e + r)) < \log 2 < 1 \leq D_r$. Thus f is strictly increasing. The inequality chain ending at equation (87) gives $f(r_{\mathrm{act}}) \geq 21 \cdot 35H$, and hence $f(r) \geq 21 \cdot 35H$ for every $r \geq r_{\mathrm{act}}$. Equivalently,

$$2^r \geq 21 \cdot 35H\{A + \log(e + r)\} \tag{90}$$

holds for every $r \geq r_{\mathrm{act}}$. Combining equations (80) and (90) proves equation (74). Finally,

$$u_j(x, Q) \leq 2^{r_{\mathrm{act}}} < 2n_0 = 2^{21}H\{A + \log(e + HA)\}, \tag{91}$$

which is equation (75) with $C = 2^{21}$. □

C.2 Reuse-weighted cache failures

For a fixed arm j, sample epoch r, and clock epoch q, the cache $U_j(r, q)$ is a function only of the deterministic prefix $X_{j,1:2^r}$ and auxiliary randomness assigned to that key. At an unavailable key set $U_j(r, q) = +\infty$, as in the policy. Define

$$E^-_{j,r,q} = \{U_j(r, q) < \mu_j\}. \tag{92}$$

In particular, an unavailable key has an empty undercoverage event. For a fixed accuracy $x > 0$, write $\mathbb{N}_0 = \{0, 1, \ldots\}$ and define the deterministic, proof-only key set

$$\mathcal{A}_j(x) = \{(r, q) \in \mathbb{N}_0^2 : 2^r \geq 21h_j(x)L(p_{r,q}, 1/4)\}. \tag{93}$$

Membership depends only on (r, q) and the fixed instance quantities, not on a terminal horizon, a realized sample count, or cache access. By equation (73), every key in $\mathcal{A}_j(x)$ is available and has at least $h_j(x)$ observations in every group. Set

$$E^+_{j,r,q}(x) = \begin{cases} \{U_j(r, q) > \mu_j + x\}, & (r, q) \in \mathcal{A}_j(x), \\ \varnothing, & (r, q) \notin \mathcal{A}_j(x). \end{cases} \tag{94}$$

Exact amplified coverage gives $\mathbb{P}(E^-_{j,r,q}) \leq p_{r,q}$ at every available key, and the bound is automatic at unavailable keys. The accuracy part of lemma 14 gives $\mathbb{P}(E^+_{j,r,q}(x)) \leq p_{r,q}$ at every eligible key;

our empty-event convention gives the same bound at every other key. These are marginal bounds for deterministic keys, not bounds conditional on adaptive access.

For any fixed comparator arm b and candidate arm j, introduce the error envelope

$$W_{b,j}(x;Q) = \sum_{q=0}^{Q} \sum_{r\geq 0} 2^q \left[\mathbf{1}\{E^-_{b,r,q}\} + \mathbf{1}\{E^+_{j,r,q}(x)\}\right], \qquad Q \in \mathbb{N}_0, \tag{95}$$

and let $W_{b,j}(x) = \lim_{Q\to\infty} W_{b,j}(x;Q)$. The full envelope sums undercoverage over all keys and upper deviations only over the fixed set $\mathcal{A}_j(x)$. All summands are nonnegative, so the partial envelopes are nondecreasing in Q.

Lemma 19 (Finite expected cache impact)**.** *For a fixed finite graph, the expected number of policy decisions affected by any event in equations* (92) *and* (94) *is finite.*

Proof. Fix a leader i with $j \in \mathcal{N}^+(i)$. During local-clock epoch q, leader i has between 2^q and $2^{q+1}-1$ occurrences, so one cache key can be used on at most 2^q of its decisions. Arm j belongs to the closed neighborhood of at most $\gamma+1$ possible leaders. Therefore the total reuse multiplicity of (j,r,q) is at most $(\gamma+1)2^q$.

Let $Z(x)$ count affected decisions, not merely distinct bad caches. Upper deviations are counted only at the fixed eligible keys in equations (93) and (94). The deterministic reuse bound gives the pathwise inequality

$$Z(x) \leq \sum_{j=1}^{K} \sum_{r,q\geq 0} (\gamma+1)2^q \left[\mathbf{1}\{E^-_{j,r,q}\} + \mathbf{1}\{E^+_{j,r,q}(x)\}\right]. \tag{96}$$

Taking expectations in equation (96) and applying Tonelli's theorem to the nonnegative series and the two fixed-prefix tail bounds gives

$$\begin{aligned}
\mathbb{E}Z(x) &\leq \sum_{j=1}^{K} \sum_{r,q\geq 0} (\gamma+1)2^q \{\mathbb{P}(E^-_{j,r,q}) + \mathbb{P}(E^+_{j,r,q}(x))\} && (97)\\
&\leq K(\gamma+1) \sum_{r,q\geq 0} 2^q\, 2p_{r,q} && (98)\\
&= K(\gamma+1) \left(\sum_{r\geq 0} \frac{1}{w(r)}\right)\left(\sum_{q\geq 0} \frac{1}{w(q)}\right). && (99)
\end{aligned}$$

The equality uses $2^q \cdot 2p_{r,q} = 1/[w(r)w(q)]$. Define $f(x) = 1/((x+1)[1+\log(x+1)]^2)$ for $x \geq 0$. Its derivative is

$$f'(x) = -\frac{3+\log(x+1)}{(x+1)^2[1+\log(x+1)]^3} < 0, \tag{100}$$

so f is positive and decreasing. With the substitution $u = 1+\log(x+1)$,

$$\int_0^\infty \frac{dx}{(x+1)[1+\log(x+1)]^2} = \int_1^\infty \frac{du}{u^2} = 1. \tag{101}$$

The integral test now gives the discrete bound

$$\begin{aligned}
\sum_{n=0}^{\infty} \frac{1}{w(n)} &= f(0) + \sum_{n=1}^{\infty} f(n) && (102)\\
&\leq f(0) + \int_0^\infty f(x)\,dx = 2, && (103)
\end{aligned}$$

where $f(n) \leq \int_{n-1}^{n} f(x)\,dx$ for every $n \geq 1$ follows from monotonicity. Therefore the right-hand side of equation (99) is at most $4K(\gamma+1)$ and is finite. The expectation bound follows from the deterministic reuse multiplicities and fixed-prefix marginal bounds, allowing dependence between cache events.

The same calculation for a fixed pair (b, j) gives

$$\mathbb{E}W_{b,j}(x) \le \sum_{r,q\ge 0} 2^q\, 2p_{r,q} \tag{104}$$

$$= \left(\sum_{r\ge 0} \frac{1}{w(r)}\right)\left(\sum_{q\ge 0} \frac{1}{w(q)}\right) \le 4. \tag{105}$$

Thus $W_{b,j}(x)$ is finite almost surely. It dominates all its partial envelopes on the same probability space, simultaneously in Q. □

C.3 Global-time policy and complete unstructured regret proof

The unstructured policy uses the same fixed-prefix caches without a graph or a leader. Its complete operational specification is as follows.

1. Pull each arm once. Before every later round t, set $q_t = \lfloor \log_2 t \rfloor$.
2. For every arm j, set $r_j(t) = \lfloor \log_2 N_j(t-1) \rfloor$ and $L = L(p_{r_j(t),q_t}, 1/4)$. The accessed triple $(j, r_j(t), q_t)$ is selected adaptively from the deterministic cache-key space.
3. If $\lfloor 2^{r_j(t)}/L \rfloor < 21 \cdot 3$, set the index to $+\infty$. Otherwise partition the first $2^{r_j(t)}$ observations into L nearly equal bases, compute $U_{m=4,r_0=1}$ with 21 MoM groups on each base, and cache the lower median of the L endpoints.
4. Pull an index maximizer over all K arms, breaking ties by least pull count and then arm identifier.

As for the graph-local policy, all signs and tie permutations are assigned to deterministic cache keys before interaction. The algorithm uses neither the horizon nor any a_j or M_j.

Proof of theorem 6. Fix $T \ge K$ and write

$$Q_T = \lfloor \log_2 T \rfloor, \qquad x_j = \Delta_j/2, \qquad u_j^T = u_j(x_j, Q_T) \quad (j \ne i^\star). \tag{106}$$

First, u_j^T is a valid threshold for every cache accessed before T. At round $t \le T$, $q_t \le Q_T$. Both 2^{q+1} and $w(q)$ are nondecreasing, so $p_{r,q_t} \ge p_{r,Q_T}$ and hence

$$L(p_{r,q_t}, 1/4) \le L(p_{r,Q_T}, 1/4). \tag{107}$$

If $N_j(t-1) \ge u_j^T = 2^{r_{\mathrm{act}}}$, then $r_j(t) \ge r_{\mathrm{act}}$ and $2^{r_j(t)} \ge u_j^T$. The persistence clause in equation (74), together with equation (107), therefore guarantees at least $h_j(x_j)$ observations in every within-base MoM group. Consequently, every accessed post-threshold candidate key belongs to $\mathcal{A}_j(x_j)$. Its upper deviation can therefore be charged to the deterministic-key event in equation (94).

Consider a post-initialization round $t \le T$ on which arm $j \ne i^\star$ is selected and $N_j(t-1) \ge u_j^T$. Index maximization gives

$$U_j(r_j(t), q_t) \ge U_{i^\star}(r_{i^\star}(t), q_t). \tag{108}$$

If neither

$$E^-_{i^\star, r_{i^\star}(t), q_t} \quad \text{nor} \quad E^+_{j, r_j(t), q_t}(x_j) \tag{109}$$

occurs, then

$$U_{i^\star}(r_{i^\star}(t), q_t) \ge \mu_{i^\star} > \mu_j + \Delta_j/2 \ge U_j(r_j(t), q_t), \tag{110}$$

contradicting equation (108). Thus every post-threshold pull of j is charged to at least one event in equation (109).

Define $Z_j^{\mathrm{glob}}(0) = 0$ and, for $T \ge 1$, set

$$Z_j^{\mathrm{glob}}(T) := W_{i^\star,j}(x_j; \lfloor \log_2 T \rfloor), \qquad Z_j^{\mathrm{glob}}(\infty) := W_{i^\star,j}(x_j). \tag{111}$$

This process is nondecreasing: increasing T only adds nonnegative summands from the next global-clock epoch. Its definition is independent of the moving threshold u_j^T.

Before the threshold, every pull increments N_j by one, and the initialization pull is included among the at most u_j^T pulls made while the pre-decision count is below u_j^T. Every remaining selected-j

round has an eligible candidate key and satisfies equation (109). A global epoch q contains at most 2^q rounds, so any fixed comparator or candidate key can be charged on at most 2^q such rounds. Grouping those decision charges by their keys therefore bounds their total by $Z_j^{\mathrm{glob}}(T)$, giving

$$N_j(T) \le u_j^T + Z_j^{\mathrm{glob}}(T). \tag{112}$$

The same full envelope dominates these charges for every T:

$$Z_j^{\mathrm{glob}}(\infty) = \sum_{r,q\ge 0} 2^q \left[\mathbf{1}\{E^-_{i^\star,r,q}\} + \mathbf{1}\{E^+_{j,r,q}(x_j)\}\right]. \tag{113}$$

In particular, no decision can be removed from the envelope when the terminal threshold increases. Tonelli's theorem and the deterministic-key bounds give

$$C_j^{\mathrm{glob}} := \mathbb{E}Z_j^{\mathrm{glob}}(\infty) \le \sum_{r,q\ge 0} 2^q\, 2p_{r,q} \tag{114}$$

$$= \left(\sum_{r\ge 0} \frac{1}{w(r)}\right)\left(\sum_{q\ge 0} \frac{1}{w(q)}\right) \le 4 < \infty. \tag{115}$$

Since $R_T = \sum_{j\neq i^\star} \Delta_j N_j(T)$, taking expectations in equation (112) and summing over suboptimal arms gives equation (20).

For the displayed rate, write $a = a_j$, $M = M_j$, and $\Delta = \Delta_j$. The definition of h_j and $\max\{3, \lceil y\rceil\} \le 3 + y$ give

$$\Delta h_j(\Delta/2) \le 3\Delta + \left[12\, 8^{1+a}\right]^{1/a} M^{1/a}\Delta^{-1/a}. \tag{116}$$

Section F expands the same substitution of half the gap into equation (6). Let $A_T = 1 + \lfloor \log_2 T\rfloor$. Since $A_T \ge 1$,

$$\log\{e + h_j(\Delta/2)A_T\} \le \log\{e + h_j(\Delta/2)\} + \log(e + A_T) \tag{117}$$

$$\le \log\{e + h_j(\Delta/2)\} + 2A_T. \tag{118}$$

Insert equation (118) into the explicit activation bound equation (91) and multiply by Δ_j:

$$\Delta_j u_j^T \le 2^{21}\Delta_j h_j(\Delta_j/2)\left[3A_T + \log\{e + h_j(\Delta_j/2)\}\right]. \tag{119}$$

Combining equations (116) and (119) and $A_T \le 1 + \log_2 T$ yields

$$\Delta_j u_j^T = O\Big([\Delta_j + M_j^{1/a_j}\Delta_j^{-1/a_j}]\log T\Big) + C'_j, \tag{120}$$

where the horizon-independent term can be chosen as $C'_j = 2^{21}\Delta_j h_j(\Delta_j/2)\log\{e + h_j(\Delta_j/2)\}$. Thus $C_{\mathrm{glob}} = \sum_{j\neq i^\star}(\Delta_j C_j^{\mathrm{glob}} + C'_j) < \infty$, which proves equation (21). This constant is independent of T for each fixed instance; its sum runs over all suboptimal arms, and its summands depend on their gaps and moment parameters. □

C.4 Explicit parameter-free leader concentration

For epoch r, let $\delta_r = 2^{-4r}$ and let $\bar{k}_r$ be the smallest odd integer no smaller than $8\log(2/\delta_r)$. If $\bar{k}_r \le 2^r$, set $k_r = \bar{k}_r$; otherwise use the largest feasible odd block count no larger than 2^r. Split the prefix into k_r nearly equal groups and let $g_r = \lfloor 2^r/k_r\rfloor$.

Lemma 20 (Parameter-free leader score)**.** *For every arm i, every epoch $r \ge 8$ satisfies*

$$\mathbb{P}\{|S_i(r) - \mu_i| > b_i(r)\} \le 2^{-4r}, \qquad b_i(r) = \left(\frac{8M_i}{g_r^{a_i}}\right)^{1/(1+a_i)}, \tag{121}$$

and $b_i(r) \to 0$. The implementation of $S_i(r)$ uses only r.

Proof. Step 1: one group. For one group of size $s \geq g_r$, equation (31) and Markov's inequality imply

$$\mathbb{P}\left\{\left|s^{-1}\sum_{h=1}^{s}(X_{i,h}-\mu_i)\right| > b_i(r)\right\} \leq \frac{2M_i/s^{a_i}}{8M_i/g_r^{a_i}} \leq \frac{1}{4}. \tag{122}$$

Step 2: the median of groups. If the median of the k_r independent group means misses by more than $b_i(r)$, at least $(k_r+1)/2$ groups satisfy the event in equation (122). Applying lemma 11 and then Hoeffding's inequality gives

$$\mathbb{P}\{|S_i(r)-\mu_i| > b_i(r)\} \leq B_{k_r}(1/4) \leq e^{-k_r/8}. \tag{123}$$

Step 3: feasibility and the requested confidence level. Because $\delta_r = 2^{-4r}$, the smallest odd integer above the requested threshold obeys

$$\bar{k}_r \leq 8\log(2/\delta_r) + 2 \tag{124}$$

$$= 8(1+4r)\log 2 + 2 \leq 25r + 8. \tag{125}$$

For $r = 8$, $2^r = 256 \geq 208 = 25r+8$, and the ratio $2^{r+1}/(25(r+1)+8)$ divided by $2^r/(25r+8)$ is $2(25r+8)/(25r+33) > 1$ for $r \geq 1$. Thus $\bar{k}_r \leq 2^r$ for all $r \geq 8$, so $k_r = \bar{k}_r$ at every such epoch. Substituting the defining lower bound on k_r into equation (123) gives

$$e^{-k_r/8} \leq e^{-\log(2/\delta_r)} = \delta_r/2 = 2^{-4r-1} \leq 2^{-4r}. \tag{126}$$

Step 4: vanishing proof radius. For $r \geq 8$, $2^r/k_r \geq 1$, hence $\lfloor 2^r/k_r \rfloor \geq 2^{r-1}/k_r$. Combining this with equation (125) yields

$$g_r \geq \frac{2^{r-1}}{25r+8}, \qquad b_i(r) \leq \left[8M_i\left(\frac{25r+8}{2^{r-1}}\right)^{a_i}\right]^{1/(1+a_i)}. \tag{127}$$

The right-hand side tends to zero because an exponential dominates the linear numerator. Neither k_r nor the computed median uses a_i, M_i, or $b_i(r)$. □

C.5 Adaptive-count-safe local selection bound

Fix a leader i and choose one proof-only arm $b_i^{\mathrm{loc}} \in \mathcal{B}_i$. Let $Q_{i,j}(S)$ count nonforced selections of $j \in \mathcal{N}^+(i) \setminus \mathcal{B}_i$ during the first S occurrences of leader i. Set $x_{i,j} = \delta_{i,j}/2$.

Lemma 21 (Local count)**.** *There is a nondecreasing random process $Z_{i,j}(S)$ such that*

$$Q_{i,j}(S) \leq u_j(x_{i,j}, \lfloor \log_2 S \rfloor) + Z_{i,j}(S), \qquad \mathbb{E}Z_{i,j}(\infty) < \infty. \tag{128}$$

Proof. Let $u = u_j(x_{i,j}, \lfloor \log_2 S \rfloor)$. Before the dyadic prefix of arm j reaches u, every selection of j increases its total sample count by one. Samples collected while another arm is leader only make the crossing earlier. More precisely, if n_j^0 is the count of arm j before the first of these S leader occurrences, then at most $(u-n_j^0)_+ \leq u$ counted selections can occur while $N_j < u$. This proves the deterministic pre-threshold part of the bound.

At any one of the first S leader-i occurrences, the current local-clock epoch q satisfies $q \leq Q := \lfloor \log_2 S \rfloor$. Both 2^{q+1} and $w(q)$ are nondecreasing, so $p_{r,q} \geq p_{r,Q}$. The defining binomial tail condition for $p_{r,Q}$ is therefore also feasible for the larger level $p_{r,q}$; minimality in equation (5) gives $L(p_{r,q}, 1/4) \leq L(p_{r,Q}, 1/4)$. The persistent threshold defining $u_j(x_{i,j}, Q)$ therefore guarantees the required group size for the actual cache at every post-threshold occurrence, not only at epoch Q. Equivalently, the candidate key at every such occurrence belongs to the fixed set $\mathcal{A}_j(x_{i,j})$ in equation (93).

Index the post-threshold nonforced occurrences that select j by $t \in \mathcal{T}_{i,j}(S)$. At such a time let (r_t, q_t) be the cache epochs actually used for j, and let (r_t^b, q_t) be those used for the fixed comparator b_i^{loc}. If neither

$$U_{b_i^{\mathrm{loc}}}(r_t^b, q_t) < \mu_i^{\mathrm{loc}} \quad \text{nor} \quad U_j(r_t, q_t) > \mu_j + \delta_{i,j}/2 \tag{129}$$

occurs, then

$$U_{b_i^{\mathrm{loc}}} \geq \mu_i^{\mathrm{loc}} > \mu_j + \delta_{i,j}/2 \geq U_j. \tag{130}$$

The middle inequality is strict because $\delta_{i,j} > 0$. Hence j cannot maximize the local indices, including under deterministic tie breaking. Therefore the pathwise implication

$$\mathbf{1}\{t \in \mathcal{T}_{i,j}(S)\} \leq \mathbf{1}\{E^-_{b_i^{\mathrm{loc}}, r_t^b, q_t}\} + \mathbf{1}\{E^+_{j,r_t,q_t}(x_{i,j})\} \tag{131}$$

holds for every post-threshold selection of j.

Define $Z_{i,j}(0) = 0$ and, for $S \geq 1$, set

$$Z_{i,j}(S) := W_{b_i^{\mathrm{loc}},j}(x_{i,j}; \lfloor \log_2 S \rfloor), \qquad Z_{i,j}(\infty) := W_{b_i^{\mathrm{loc}},j}(x_{i,j}). \tag{132}$$

This definition includes all undercoverage keys and all accuracy-eligible upper-deviation keys in the indicated local-clock epochs. These key sets are fixed independently of the terminal threshold u. Increasing S only adds nonnegative terms, so $Z_{i,j}(S)$ is nondecreasing.

During local-clock epoch q, leader i has at most 2^q decisions. Thus each comparator or candidate key can be charged at most 2^q times by equation (131). Every charged candidate key belongs to $\mathcal{A}_j(x_{i,j})$, and every charged epoch satisfies $q \leq \lfloor \log_2 S \rfloor$. Summing that implication over $\mathcal{T}_{i,j}(S)$ and grouping each kind of indicator by its key therefore gives

$$|\mathcal{T}_{i,j}(S)| \leq Z_{i,j}(S).$$

Together with the pre-threshold count, this proves $Q_{i,j}(S) \leq u + Z_{i,j}(S)$ for every $S \geq 1$ on the same sample path. The full envelope is

$$Z_{i,j}(\infty) = \sum_{r,q \geq 0} 2^q \left[\mathbf{1}\{E^-_{b_i^{\mathrm{loc}},r,q}\} + \mathbf{1}\{E^+_{j,r,q}(x_{i,j})\} \right]. \tag{133}$$

Our definition of E^+ omits precisely the candidate keys outside the fixed accuracy set; an unavailable comparator has $E^- = \varnothing$. Tonelli's theorem and the deterministic-key bounds yield

$$\mathbb{E} Z_{i,j}(\infty) \leq \sum_{r,q \geq 0} 2^q \, 2p_{r,q} \tag{134}$$

$$= \left(\sum_{r \geq 0} \frac{1}{w(r)} \right) \left(\sum_{q \geq 0} \frac{1}{w(q)} \right) \leq 4 < \infty. \tag{135}$$

Both the count bound and domination by $Z_{i,j}(\infty)$ hold simultaneously for all S. They can therefore be evaluated at a random leader count without applying any probability bound at that random count. This proves equation (128). □

D Complete Operational Specification

We use r_0 for the RMM rank-tail parameter, r for the sample epoch, and r_{act} for the proof-only activation epoch.

Certified APR-LOCAL

1. Initialize every robust score to $-\infty$ and pull every arm once. Whenever arm j reaches count 2^r, set $\delta_r = 2^{-4r}$, take the smallest odd $k_r \geq 8\log(2/\delta_r)$ capped by the largest feasible odd count, split the prefix interleavingly into k_r groups, and store their lower-median mean as $S_j(r)$.
2. At a new round, choose as leader i the arm with largest current robust score; break ties by least pull count and then arm identifier. Increment its occurrence count to s.
3. If $(s-1) \bmod (\gamma+1) = 0$, pull leader i.
4. Otherwise, for each $j \in \mathcal{N}^+(i)$, set $r = \lfloor \log_2 N_j \rfloor$, $q = \lfloor \log_2 s \rfloor$, and $L = L(p_{r,q}, 1/4)$.
5. If $\lfloor 2^r/L \rfloor < 21 \cdot 3$, set $U_j(r,q) = +\infty$. Otherwise partition the first 2^r observations into L nearly equal base samples; on each base compute $U_{m=4,r_0=1}$ with $k = 21$ MoM groups, and cache the lower median of the L base endpoints under the key (j, r, q).
6. Pull an index maximizer in $\mathcal{N}^+(i)$, breaking ties by least pull count and then arm identifier.
7. After observing the reward, apply the power-of-two score update in Step 1 to the selected arm; all other scores and caches remain unchanged.

Figure 4: Complete specification of the rank-four configuration.

All signs and tie permutations can be coupled as an infinite array sampled before interaction. Cached values are therefore fixed random variables indexed by deterministic arm prefixes, even though the policy accesses them adaptively.

E COMPLETE ROBUST-ASCENT PROOF

Proof of lemma 3. Let $\tau_{i,s}$ be the decision time at which arm i becomes leader for the s-th time, if that occurrence exists, and let $\ell_i(T)$ be the number of its leader occurrences by horizon T. Fix $i \neq i^\star$. Write $N_h(\tau_{i,s}^-) := N_h(\tau_{i,s} - 1)$ for the pull count immediately before that decision, and let $r_h(\tau_{i,s}^-) = \lfloor \log_2 N_h(\tau_{i,s}^-) \rfloor$ be the score epoch available to the leader rule. Graph unimodality gives $\mu_i^{\mathrm{loc}} > \mu_i$. Hence $i \notin \mathcal{B}_i$ and, with the following definitions,

$$d_i := |\mathcal{B}_i| \in \{1, \ldots, \gamma\}, \qquad g_i := \mu_i^{\mathrm{loc}} - \mu_i > 0. \tag{136}$$

Set

$$\eta = \frac{\gamma}{2(\gamma+1)}. \tag{137}$$

Partition the positive integers corresponding to realized leader occurrences into the random sets

$$\mathcal{A}_i = \left\{ s : \sum_{b \in \mathcal{B}_i} N_b(\tau_{i,s}^-) \geq \eta s \right\}, \tag{138}$$

$$\mathcal{C}_i = \left\{ s : \sum_{b \in \mathcal{B}_i} N_b(\tau_{i,s}^-) < \eta s \right\}. \tag{139}$$

High-count occurrences. Put $c = \gamma + 1$. The rule $(s-1) \bmod c = 0$ forces leader i at occurrence numbers $1, 1+c, 1+2c, \ldots$. Before decision $\tau_{i,s}$, the first $s-1$ leader occurrences have already produced

$$F_{s-1} = \begin{cases} 0, & s = 1, \\ 1 + \lfloor (s-2)/c \rfloor, & s \geq 2 \end{cases} \tag{140}$$

forced pulls of arm i. Initialization contributes one additional pull. Thus $N_i(\tau_{i,s}^-) \geq 1 + F_{s-1}$, and for every $s \geq 1$ this implies

$$N_i(\tau_{i,s}^-) \geq \frac{s}{2(\gamma+1)}. \tag{141}$$

For $s \geq 2$, the inequality follows from $1 + F_{s-1} \geq 1 + (s-2)/c \geq s/(2c)$; the case $s = 1$ is immediate. If $s \in \mathcal{A}_i$, the pigeonhole principle gives an arm $b \in \mathcal{B}_i$ with

$$N_b(\tau_{i,s}^-) \geq \frac{\eta s}{d_i}. \tag{142}$$

From equations (141) and (142), the current score epochs obey

$$r_i(\tau_{i,s}^-) \geq r_i^-(s) := \left\lfloor \log_2 \frac{s}{2(\gamma+1)} \right\rfloor, \tag{143}$$

$$r_b(\tau_{i,s}^-) \geq r_b^-(s) := \left\lfloor \log_2 \frac{\eta s}{d_i} \right\rfloor. \tag{144}$$

Because every $b \in \mathcal{B}_i$ has mean μ_i^{loc} and the radii in equation (127) converge to zero, choose a deterministic epoch $R_i \geq 8$ such that $b_h(r) \leq g_i/4$ for every $h \in \{i\} \cup \mathcal{B}_i$ and every $r \geq R_i$. The lower epoch bounds in equations (143) and (144) diverge with s, so choose s_i^A large enough that both are at least R_i whenever $s \geq s_i^A$. If the current scores of i and the pigeonholed arm b are both good, then

$$S_b \geq \mu_i^{\mathrm{loc}} - g_i/4 = \mu_i + 3g_i/4 > \mu_i + g_i/4 \geq S_i, \tag{145}$$

contradicting that i is the global score maximizer.

To handle the random epochs in equation (145), take the union of deterministic bad-score events above the corresponding epoch lower bounds. With $F_{h,r} = \{|S_h(r) - \mu_h| > b_h(r)\}$,

$$\{s \in \mathcal{A}_i\} \subseteq \bigcup_{r \geq r_i^-(s)} F_{i,r} \cup \bigcup_{b \in \mathcal{B}_i} \bigcup_{r \geq r_b^-(s)} F_{b,r} \qquad (s \geq s_i^A). \tag{146}$$

By lemma 20 and a union bound,

$$\mathbb{P}(s \in \mathcal{A}_i) \leq \sum_{r \geq r_i^-(s)} 2^{-4r} + d_i \sum_{r \geq r_b^-(s)} 2^{-4r} \tag{147}$$

$$= \frac{2^{-4r_i^-(s)} + d_i 2^{-4r_b^-(s)}}{1 - 2^{-4}} \leq C_i s^{-4} \tag{148}$$

for $s \geq s_i^A$. To verify the last step, if $c_0 s \geq 1$, then

$$2^{-4\lfloor \log_2(c_0 s) \rfloor} \leq \frac{16}{c_0^4 s^4}. \tag{149}$$

Apply equation (149) separately to the two terms in equation (148): use $c_0 = 1/[2(\gamma+1)]$ for the score of arm i, and $c_0 = \eta/d_i$ for each of the d_i locally best arms. This gives the explicit finite choice

$$C_i = \frac{16}{1 - 2^{-4}} \left([2(\gamma+1)]^4 + d_i (d_i/\eta)^4 \right). \tag{150}$$

Therefore, by Tonelli applied to the membership indicators,

$$\mathbb{E}|\mathcal{A}_i| = \sum_{s \geq 1} \mathbb{P}(s \in \mathcal{A}_i) < \infty. \tag{151}$$

Low-count occurrences. Among the first s occurrences of leader i, at most $\lceil s/(\gamma+1) \rceil$ decisions are forced. Thus the number of nonforced decisions is at least

$$s - \left\lceil \frac{s}{\gamma+1} \right\rceil \geq \frac{\gamma s}{\gamma+1} - 1. \tag{152}$$

If $s \in \mathcal{C}_i$, fewer than ηs observations of arms in $\mathcal{B}_i$ are present immediately before the s-th decision. Consequently, among the first $s-1$ leader-i decisions, fewer than ηs can have selected an arm in $\mathcal{B}_i$. The current s-th decision may select one additional arm in $\mathcal{B}_i$, so at most $\eta s + 1$ of the first s nonforced decisions need be excluded. Combining this off-by-one bound with equations (137) and (152), at least

$$\frac{\gamma s}{\gamma+1} - 1 - (\eta s + 1) = \eta s - 2 \tag{153}$$

nonforced decisions select arms in $\mathcal{N}^+(i) \setminus \mathcal{B}_i$.

Sum equation (128) over these at most $\gamma + 1$ arms and set $Z_i(s) = \sum_j Z_{i,j}(s)$, where the sum is over $j \in \mathcal{N}^+(i) \setminus \mathcal{B}_i$. This finite sum is nondecreasing and

$$\mathbb{E}Z_i(\infty) = \sum_{j \in \mathcal{N}^+(i) \setminus \mathcal{B}_i} \mathbb{E}Z_{i,j}(\infty) < \infty \tag{154}$$

by lemma 21. On $s \in \mathcal{C}_i$,

$$\eta s - 2 \le \sum_{j \in \mathcal{N}^+(i) \setminus \mathcal{B}_i} u_j(\delta_{i,j}/2, \lfloor \log_2 s \rfloor) + Z_i(s). \tag{155}$$

The set $\mathcal{N}^+(i) \setminus \mathcal{B}_i$ is finite, and for each of its arms the quantities $h_j(\delta_{i,j}/2)$ are fixed instance constants. Since $\lfloor \log_2 s \rfloor + 1 \le 1 + \log_2 s$, the explicit activation bound equation (91) yields a finite D_i such that

$$\sum_{j \in \mathcal{N}^+(i) \setminus \mathcal{B}_i} u_j(\delta_{i,j}/2, \lfloor \log_2 s \rfloor) \le D_i(1 + \log s) \tag{156}$$

for all $s \ge 1$. Because $(1 + \log s)/s \to 0$, choose a deterministic s_i^C such that, for every $s \ge s_i^C$, that sum plus two is at most $\eta s/2$. Then equation (155) implies

$$Z_i(\infty) \ge Z_i(s) \ge \frac{\eta s}{2} \quad \text{whenever } s \ge s_i^C \text{ and } s \in \mathcal{C}_i. \tag{157}$$

For a realized nonnegative value $Z_i(\infty)$, at most $2Z_i(\infty)/\eta$ positive integers can satisfy equation (157). Therefore

$$\mathbb{E}|\mathcal{C}_i| \le s_i^C + \frac{2}{\eta}\mathbb{E}Z_i(\infty) < \infty, \tag{158}$$

where finiteness follows from equation (154).

Completion. Every realized occurrence belongs to exactly one of $\mathcal{A}_i$ and $\mathcal{C}_i$. Hence, for all T,

$$\ell_i(T) \le |\mathcal{A}_i| + |\mathcal{C}_i|. \tag{159}$$

Taking expectations and applying equations (151) and (158) proves

$$\sup_{T \ge 1} \mathbb{E}\ell_i(T) < \infty \qquad (i \ne i^\star). \tag{160}$$

□

E.1 Sufficient conditions for ascent

The preceding proof depends on the endpoint construction through its pathwise local-selection bound. The following proposition states the conditions on that bound and on the leader scores separately.

Proposition 22 (Local counts imply finite expected ascent)**.** *Consider a finite graph with maximum degree $\gamma \ge 1$ and a graph-unimodal mean vector. After sampling every arm once, a policy chooses a leader maximizing $S_h(\lfloor \log_2 N_h(t-1) \rfloor)$ and selects an arm in that leader's closed neighborhood. Each $S_h(r)$ is a score defined on the deterministic prefix of length 2^r, including prefixes never reached by the policy. Occurrences $s \equiv 1 \pmod{\gamma + 1}$ force a pull of the leader; all other occurrences are nonforced.*

For each $i \ne i^\star$, let $Q_i(s) = \sum_{j \in \mathcal{N}^+(i) \setminus \mathcal{B}_i} Q_{i,j}(s)$ count nonforced selections outside the locally best set among its first s leader occurrences. Suppose there are a deterministic nonnegative function $f_i(s) = o(s)$ and a nonnegative random variable Y_i such that

$$Q_i(s) \le f_i(s) + Y_i \quad \textit{simultaneously at every realized } s, \qquad \mathbb{E}Y_i < \infty. \tag{161}$$

For every arm h, suppose deterministic radii $\rho_h(r) \ge 0$ and probability bounds $\varepsilon_h(r) \in [0,1]$ satisfy

$$\rho_h(r) \longrightarrow 0, \qquad \mathbb{P}\{|S_h(r) - \mu_h| > \rho_h(r)\} \le \varepsilon_h(r), \qquad \sum_{r \ge 0} 2^r \varepsilon_h(r) < \infty. \tag{162}$$

Then $\mathbb{E}\ell_i(\infty) < \infty$ for every $i \ne i^\star$, where $\ell_i(\infty) = \lim_{T \to \infty} \ell_i(T)$.

Proof. Fix $i \neq i^\star$, put $d_i = |\mathcal{B}_i|$, and set $\eta = \gamma/[2(\gamma+1)]$ and $c_i = \min\{1/[2(\gamma+1)], \eta/d_i\} > 0$. Define $\tau_{i,s} = \infty$ when the s-th leader-i occurrence does not exist. Counts below are evaluated only on $\tau_{i,s} < \infty$ and immediately before that decision. Partition realized occurrences into $\mathcal{A}_i$, where $\sum_{b\in\mathcal{B}_i} N_b(\tau_{i,s}-1) \geq \eta s$, and $\mathcal{C}_i$, where this sum is smaller.

Initialization and the forced schedule give $N_i(\tau_{i,s}-1) \geq s/[2(\gamma+1)]$, as in equations (140) and (141). On $s \in \mathcal{A}_i$, some $b \in \mathcal{B}_i$ also satisfies $N_b(\tau_{i,s}-1) \geq \eta s/d_i$. Both score epochs are therefore at least $q_i(s) := \max\{0, \lfloor \log_2(c_i s) \rfloor\}$. Graph unimodality gives $g_i = \mu_i^{\mathrm{loc}} - \mu_i > 0$. Choose R_i such that $\rho_h(r) \leq g_i/4$ for all $h \in \{i\} \cup \mathcal{B}_i$ and all $r \geq R_i$, then choose s_i^A such that $q_i(s) \geq R_i$ for $s \geq s_i^A$. If both current score errors are within their radii, then $S_b \geq \mu_i + 3g_i/4 > \mu_i + g_i/4 \geq S_i$, contrary to the leader rule. Consequently

$$\{s \in \mathcal{A}_i\} \subseteq \bigcup_{h\in\{i\}\cup\mathcal{B}_i} \bigcup_{r\geq q_i(s)} \{|S_h(r) - \mu_h| > \rho_h(r)\} \qquad (s \geq s_i^A). \tag{163}$$

For each $r \geq 0$, the inequality $q_i(s) \leq r$ implies $s < 2^{r+1}/c_i$. Tonelli and a union bound thus give

$$\begin{aligned}\sum_{s\geq s_i^A} \mathbb{P}(s \in \mathcal{A}_i) &\leq \sum_{h\in\{i\}\cup\mathcal{B}_i} \sum_{r\geq 0} \varepsilon_h(r)\, |\{s \geq s_i^A : q_i(s) \leq r\}| \\ &\leq \frac{2}{c_i} \sum_{h\in\{i\}\cup\mathcal{B}_i} \sum_{r\geq 0} 2^r \varepsilon_h(r) < \infty. \end{aligned} \tag{164}$$

Hence $\mathbb{E}|\mathcal{A}_i| < \infty$.

Among the first s occurrences, at least $\gamma s/(\gamma+1) - 1 = 2\eta s - 1$ decisions are nonforced. On $s \in \mathcal{C}_i$, fewer than ηs of the first $s-1$ decisions selected an arm in $\mathcal{B}_i$, and the current decision adds at most one. Thus

$$\eta s - 2 \leq Q_i(s) \leq f_i(s) + Y_i. \tag{165}$$

Since $f_i(s) = o(s)$, choose s_i^C such that $f_i(s) + 2 \leq \eta s/2$ for all $s \geq s_i^C$. Every such $s \in \mathcal{C}_i$ obeys $s \leq 2Y_i/\eta$, giving the pathwise bound $|\mathcal{C}_i| \leq s_i^C - 1 + 2Y_i/\eta$ and therefore $\mathbb{E}|\mathcal{C}_i| < \infty$. Finally, $\ell_i(\infty) = |\mathcal{A}_i| + |\mathcal{C}_i|$ proves the claim. □

Verification for APR-LOCAL**.** Take

$$f_i(s) = \sum_{j\in\mathcal{N}^+(i)\setminus\mathcal{B}_i} u_j(\delta_{i,j}/2, \lfloor \log_2 s \rfloor), \qquad Y_i = \sum_{j\in\mathcal{N}^+(i)\setminus\mathcal{B}_i} Z_{i,j}(\infty). \tag{166}$$

Lemma 21 gives the simultaneous pathwise inequality and integrability in equation (161). Equation (91) gives $f_i(s) = O(\log(es)) = o(s)$. For $r \geq 8$, set $\rho_h(r) = b_h(r)$ from equation (127) and $\varepsilon_h(r) = 2^{-4r}$. For $r < 8$, set $\rho_h(r) = 0$ and $\varepsilon_h(r) = 1$. Lemma 20 verifies equation (162), since the radii vanish and $\sum_{r\geq 8} 2^{-3r} < \infty$. The factor 2^r counts how many leader occurrences can expose an error at epoch r once the relevant sample counts grow linearly. The local bound is sublinear in those occurrences, even when their global decision times are far apart.

F Detailed Regret Proof

Proof of theorem 4. Let $\Delta_{\max} = \max_i \Delta_i$. Every round has exactly one leader. After the one-pull-per-arm initialization, split the regret according to whether the current leader equals $i^\star$. On a round with leader $i \neq i^\star$, the one-step regret is at most $\Delta_{\max}$, so equation (160) and monotone convergence give

$$\mathbb{E}\ell_i(\infty) := \mathbb{E}\Big[\lim_{T\to\infty} \ell_i(T)\Big] = \lim_{T\to\infty} \mathbb{E}\ell_i(T) < \infty. \tag{167}$$

Therefore

$$C^{\mathrm{asc}} := \Delta_{\max} \sum_{i\neq i^\star} \mathbb{E}\ell_i(\infty) < \infty. \tag{168}$$

This term includes forced and nonforced choices made while a suboptimal arm is leader.

On rounds whose leader is $i^\star$, graph unimodality and uniqueness of the optimum give $\mathcal{B}_{i^\star} = \{i^\star\}$. Forced leader pulls incur zero regret. Let $S_T = \ell_{i^\star}(T) \leq T$. The number of nonforced pulls

of $j \in \mathcal{N}(i^\star)$ on these rounds is exactly $Q_{i^\star,j}(S_T)$. If $S_T = 0$ this count is zero. Otherwise, monotonicity of u_j in its second argument and of $Z_{i^\star,j}$ gives the pathwise bound

$$Q_{i^\star,j}(S_T) \le u_j(\Delta_j/2, \lfloor \log_2 T \rfloor) + Z_{i^\star,j}(\infty). \tag{169}$$

The regret therefore satisfies the pathwise decomposition

$$R_T \le C_{\mathrm{init}} + \Delta_{\max} \sum_{i \neq i^\star} \ell_i(T) + \sum_{j \in \mathcal{N}(i^\star)} \Delta_j Q_{i^\star,j}(S_T), \tag{170}$$

$$C_{\mathrm{init}} = \sum_{i \neq i^\star} \Delta_i. \tag{171}$$

Take expectations, insert equation (169), and use Tonelli for the finite nonnegative sums. This proves equation (16) with $C_j^{\mathrm{loc}} = \mathbb{E} Z_{i^\star,j}(\infty)$.

For the rate, write $a = a_j$, $M = M_j$, and $\Delta = \Delta_j$. From $\max\{3, \lceil y \rceil\} \le 3 + y$ for $y \ge 0$ and equation (6) with $x = \Delta/2$,

$$\Delta h_j(\Delta/2) \le 3\Delta + \left[12\, 8^{1+a}\right]^{1/a} M^{1/a} \Delta^{-1/a}. \tag{172}$$

This constant follows from $4^{1+a}/(\Delta/2)^{1+a} = 8^{1+a}/\Delta^{1+a}$, which gives the displayed constant after taking the $1/a$ power and multiplying by Δ.

Set $Q = \lfloor \log_2 T \rfloor$ and $A_T = 1 + \lfloor \log_2 T \rfloor$. Since $A_T \ge 1$,

$$\log\{e + h_j(\Delta/2) A_T\} \le \log\{e + h_j(\Delta/2)\} + \log(e + A_T) \le \log\{e + h_j(\Delta/2)\} + 2A_T. \tag{173}$$

Combining the explicit activation bound equation (91) with equation (173) first gives

$$\Delta_j u_j(\Delta_j/2, Q) \le 2^{21} \Delta_j h_j(\Delta_j/2) \left[3A_T + \log\{e + h_j(\Delta_j/2)\}\right]. \tag{174}$$

Using equation (172) and $A_T \le 1 + \log_2 T$ gives

$$\Delta_j u_j(\Delta_j/2, Q) = O\Big(\left[\Delta_j + M_j^{1/a_j} \Delta_j^{-1/a_j}\right] \log T\Big) + C_j', \tag{175}$$

where $C_j' = 2^{21} \Delta_j h_j(\Delta_j/2) \log\{e + h_j(\Delta_j/2)\}$ is independent of T. Summing equation (175) gives the rate form in theorem 4; one admissible horizon-independent remainder is

$$C_{\mathrm{climb}} = C_{\mathrm{init}} + C^{\mathrm{asc}} + \sum_{j \in \mathcal{N}(i^\star)} \left[\Delta_j \mathbb{E} Z_{i^\star,j}(\infty) + C_j'\right] < \infty. \tag{176}$$

For a fixed instance, C_{climb} is independent of T. It can depend on the number of arms, the graph, the gaps and the moment parameters through initialization, the ascent thresholds and the finite activation terms. Neighbor localization applies to the coefficient of $\log T$; the ascent sum in equation (168) ranges over all suboptimal arms. □

The Δ_j term in equation (172) is the cost of the certified minimum group size $h_j \ge 3$. In the moment-dominated heavy-tail regime, $M_j^{1/a_j} \Delta_j^{-1/a_j} \ge \Delta_j$, the leading coefficient reduces to the moment–gap term emphasized in the main text.

G Complete Symmetry-Preserving Lower-Bound Proof

G.1 Low-frequency separation of symmetric centers

Let $\mathcal{S}_p(M)$ be the class defined in section 5, with $M > 0$.

Lemma 23 (Center separation). *If $P, Q \in \mathcal{S}_p(M)$ have means μ_P, μ_Q and $\delta = |\mu_P - \mu_Q| \le (8M)^{1/p}$, then*

$$\mathrm{TV}(P, Q) \ge \frac{3}{4\pi} \frac{\delta}{(8M)^{1/p}}. \tag{177}$$

Proof. For $X \sim P$, symmetry of $X - \mu_P$ gives

$$\varphi_P(t) = \mathbb{E}e^{itX} = e^{it\mu_P}\rho_P(t), \qquad \rho_P(t) = \mathbb{E}\cos\{t(X - \mu_P)\} \in \mathbb{R}. \tag{178}$$

The sine term vanishes because it is odd under a symmetric distribution. For $p \in (1, 2]$ and all real u,

$$1 - \cos u \le 2|u|^p. \tag{179}$$

For $|u| \le 1$, this follows from $1 - \cos u \le u^2/2 \le |u|^p/2$; for $|u| > 1$, use $1 - \cos u \le 2 \le 2|u|^p$. Set $t = (8M)^{-1/p}$. Then

$$1 - \rho_P(t) = \mathbb{E}[1 - \cos\{t(X - \mu_P)\}] \tag{180}$$

$$\le 2t^p\mathbb{E}|X - \mu_P|^p \le \frac{1}{4}, \tag{181}$$

so $\rho_P(t) \ge 3/4$; the same holds for Q.

Assume without loss of generality that $\mu_Q - \mu_P = \delta$ and rotate both characteristic functions by $e^{-it(\mu_P+\mu_Q)/2}$. Their imaginary parts are $-\rho_P(t)\sin(t\delta/2)$ and $+\rho_Q(t)\sin(t\delta/2)$. Therefore

$$|\varphi_P(t) - \varphi_Q(t)| \ge \{\rho_P(t) + \rho_Q(t)\}|\sin(t\delta/2)| \tag{182}$$

$$\ge \frac{3}{2}|\sin(t\delta/2)|. \tag{183}$$

Because $t\delta \le 1$, we have $t\delta/2 \in [0, 1/2] \subset [0, \pi/2]$. The chord bound $\sin y \ge 2y/\pi$ gives

$$|\varphi_P(t) - \varphi_Q(t)| \ge \frac{3t\delta}{2\pi}. \tag{184}$$

For every measurable complex-valued f with $|f| \le 1$, $|\mathbb{E}_P f - \mathbb{E}_Q f| \le 2\,\mathrm{TV}(P, Q)$. Apply this to $f(x) = e^{itx}$ and combine with equation (184) to obtain equation (177). □

Proposition 24 (Symmetric information modulus)**.** *Under the conditions of lemma 23,*

$$\mathrm{KL}(P, Q) \ge \frac{9}{8\pi^2 8^{2/p}} \frac{\delta^2}{M^{2/p}}. \tag{185}$$

Conversely, if

$$c_p = \mathbb{E}|Z|^p = \frac{2^{p/2}\Gamma((p+1)/2)}{\sqrt{\pi}}, \quad Z \sim \mathcal{N}(0, 1), \tag{186}$$

and $\sigma = (M/c_p)^{1/p}$, then $P = \mathcal{N}(\mu, \sigma^2)$ and $Q = \mathcal{N}(\mu + \delta, \sigma^2)$ belong to $\mathcal{S}_p(M)$ and

$$\mathrm{KL}(P, Q) = \frac{c_p^{2/p}}{2} \frac{\delta^2}{M^{2/p}}. \tag{187}$$

Proof. Pinsker's inequality is $\mathrm{TV}(P, Q)^2 \le \mathrm{KL}(P, Q)/2$. Substituting equation (177) and multiplying by two gives equation (185). For the Gaussian pair, $X - \mathbb{E}X \overset{d}{=} \sigma Z$, and hence

$$\mathbb{E}|X - \mathbb{E}X|^p = \sigma^p\mathbb{E}|Z|^p = \sigma^p \frac{2^{p/2}\Gamma((p+1)/2)}{\sqrt{\pi}} = \sigma^p c_p = M. \tag{188}$$

Thus both distributions lie in the class and are symmetric about their respective means. The equal-variance Gaussian divergence is

$$\mathrm{KL}\{\mathcal{N}(\mu, \sigma^2), \mathcal{N}(\mu + \delta, \sigma^2)\} = \frac{\delta^2}{2\sigma^2}, \tag{189}$$

which becomes equation (187) after substituting σ. □

G.2 A GRAPH-VALID ALTERNATIVE

Lemma 25 (Raising an optimal neighbor preserves graph unimodality)**.** *Let μ be graph-unimodal with unique optimum $i^\star$, and fix $j \in \mathcal{N}(i^\star)$. For $\varepsilon > 0$, change only coordinate j to $\mu'_j = \mu_{i^\star} + \varepsilon$. Then μ' is graph-unimodal with unique optimum j.*

Proof. The new coordinate is strictly larger than every unchanged mean, so j is the unique optimum. Fix a starting node v. Under μ, choose a strictly increasing graph path from v to $i^\star$. If this path meets j, truncate it at the first visit to j. Every earlier edge remains strictly increasing because only the terminal value at j was raised. If the path does not meet j, append the edge $(i^\star, j)$; the original path is unchanged and $\mu'_{i^\star} < \mu'_j$. For $v = j$, the length-zero path suffices. Thus every node has a strictly increasing path to j. □

G.3 TRAJECTORY CHANGE OF MEASURE

Call a policy uniformly efficient on the Gaussian graph-unimodal family if $\mathbb{E}_\theta R_T = o(T^\alpha)$ for every instance θ and every $\alpha > 0$.

Lemma 26 (Neighbor sampling constraint)**.** *Let θ be an instance with optimum $i^\star$. Fix $j \in \mathcal{N}(i^\star)$ and an alternative $\theta^{j,\varepsilon}$ that changes only arm j to have mean $\mu_{i^\star} + \varepsilon$. For every uniformly efficient policy,*

$$\liminf_{T\to\infty} \frac{\mathbb{E}_\theta N_j(T)}{\log T} \operatorname{KL}(P_j, Q_j^\varepsilon) \geq 1. \tag{190}$$

Proof. Let $\mathbb{P}_\theta^T$ and $\mathbb{P}_{\theta^{j,\varepsilon}}^T$ denote the laws of the complete interaction history through time T, including the policy's internal randomness. The policy uses the same action kernel under both instances, and only observations from arm j change distribution. Conditional on the past history H_{t-1} and action I_t, the one-step observation divergence is zero when $I_t \neq j$ and is $\operatorname{KL}(P_j, Q_j^\varepsilon)$ when $I_t = j$. The action kernel and the internal-randomness kernel contribute zero divergence because they are the same under both environments. The KL chain rule therefore gives, step by step,

$$\operatorname{KL}(\mathbb{P}_\theta^T, \mathbb{P}_{\theta^{j,\varepsilon}}^T) = \sum_{t=1}^{T} \mathbb{E}_\theta\big[\mathbf{1}\{I_t = j\} \operatorname{KL}(P_j, Q_j^\varepsilon)\big] \tag{191}$$

$$= \mathbb{E}_\theta N_j(T) \operatorname{KL}(P_j, Q_j^\varepsilon). \tag{192}$$

Define $A_T = \{N_j(T) < T/2\}$. Under the original instance, $\Delta_j \mathbb{E}_\theta N_j(T) \leq \mathbb{E}_\theta R_T = o(T^\alpha)$ for every $\alpha > 0$. Markov's inequality implies

$$\mathbb{P}_\theta^T(A_T^c) \leq \frac{2\mathbb{E}_\theta N_j(T)}{T} = o(T^{\alpha-1}). \tag{193}$$

Under the alternative, arm j is uniquely optimal and its gap to the old optimum is ε. Thus $\varepsilon \mathbb{E}_{\theta^{j,\varepsilon}}[T - N_j(T)] \leq \mathbb{E}_{\theta^{j,\varepsilon}} R_T = o(T^\alpha)$, so

$$\mathbb{P}_{\theta^{j,\varepsilon}}^T(A_T) \leq \frac{2\mathbb{E}_{\theta^{j,\varepsilon}}[T - N_j(T)]}{T} = o(T^{\alpha-1}). \tag{194}$$

Let $p_T = \mathbb{P}_\theta^T(A_T)$ and $q_T = \mathbb{P}_{\theta^{j,\varepsilon}}^T(A_T)$. Data processing through the indicator of A_T gives

$$\operatorname{KL}(\mathbb{P}_\theta^T, \mathbb{P}_{\theta^{j,\varepsilon}}^T) \geq \operatorname{kl}(p_T, q_T). \tag{195}$$

For $p, q \in (0,1)$, binary relative entropy can be written as

$$\operatorname{kl}(p,q) = p \log \frac{1}{q} + (1-p) \log \frac{1}{1-q} - H_{\mathrm{bin}}(p), \qquad H_{\mathrm{bin}}(p) = -p \log p - (1-p) \log(1-p). \tag{196}$$

Since $(1-p)\log(1/(1-q)) \geq 0$ and $H_{\mathrm{bin}}(p) \leq \log 2$, it follows that

$$\operatorname{kl}(p,q) \geq p \log \frac{1}{q} - \log 2. \tag{197}$$

The same inequality holds at the boundary by the extended-real convention; in particular, if $q = 0 < p$, both $\operatorname{kl}(p,q)$ and $p\log(1/q)$ are $+\infty$. By equation (193), $p_T \to 1$. Fix any $\alpha \in (0,1)$. Equation (194) implies $q_T = o(T^{\alpha-1})$. Thus there is a deterministic sequence $\varepsilon_T \downarrow 0$ with $q_T = T^{\alpha-1}\varepsilon_T$ whenever $q_T > 0$ (if $q_T = 0$, the following logarithm is $+\infty$). Consequently,

$$\log(1/q_T) = (1-\alpha)\log T + \log(1/\varepsilon_T), \tag{198}$$

and hence

$$\liminf_{T\to\infty} \frac{\log(1/q_T)}{\log T} \geq 1 - \alpha. \tag{199}$$

Combining equations (192), (195), (197) and (199) and dividing by $\log T$ gives

$$\liminf_{T\to\infty} \frac{\mathbb{E}_\theta N_j(T) \operatorname{KL}(P_j, Q_j^\varepsilon)}{\log T} \tag{200}$$

$$\geq \liminf_{T\to\infty} \left[p_T \frac{\log(1/q_T)}{\log T} - \frac{\log 2}{\log T} \right] \geq 1 - \alpha, \tag{201}$$

because $p_T \to 1$. The left-hand side does not depend on the freely chosen $\alpha \in (0,1)$. Taking the supremum of the right-hand side over such α, equivalently letting $\alpha \downarrow 0$, proves equation (190). □

G.4 Proof of the neighbor lower bound

Proof of theorem 7. A neighbor with $M_j = 0$ has $\sigma_j^2 = 0$ and contributes zero to the lower bound. Fix $j \in \mathcal{N}(i^\star)$ with $M_j > 0$ and $\varepsilon > 0$. In the Gaussian subfamily, change only the mean of arm j from μ_j to $\mu_{i^\star} + \varepsilon$, keeping variance σ_j^2 fixed. Translation preserves symmetry and the centered p_j-th moment, and lemma 25 proves that the alternative remains in the graph-unimodal family. Its one-arm divergence is

$$\mathrm{KL}(P_j, Q_j^\varepsilon) = \frac{(\Delta_j + \varepsilon)^2}{2\sigma_j^2}. \tag{202}$$

By lemma 26,

$$\liminf_{T\to\infty} \frac{\mathbb{E}N_j(T)}{\log T} \geq \frac{2\sigma_j^2}{(\Delta_j + \varepsilon)^2}. \tag{203}$$

The inequality holds for every $\varepsilon > 0$, so letting $\varepsilon \downarrow 0$ gives

$$\liminf_{T\to\infty} \frac{\mathbb{E}N_j(T)}{\log T} \geq \frac{2\sigma_j^2}{\Delta_j^2}. \tag{204}$$

Because the neighbor set is finite and all regret contributions are nonnegative,

$$\liminf_{T\to\infty} \frac{\mathbb{E}R_T}{\log T} = \liminf_{T\to\infty} \sum_{i\neq i^\star} \Delta_i \frac{\mathbb{E}N_i(T)}{\log T} \tag{205}$$

$$\geq \sum_{j\in\mathcal{N}(i^\star)} \Delta_j \liminf_{T\to\infty} \frac{\mathbb{E}N_j(T)}{\log T} \tag{206}$$

$$\geq \sum_{j\in\mathcal{N}(i^\star)} \frac{2\sigma_j^2}{\Delta_j}. \tag{207}$$

Substituting $\sigma_j^2 = (M_j/c_{p_j})^{2/p_j}$ gives equation (23). □

G.5 Exact upper–lower comparison

For $M_j > 0$, write $p_j = 1 + a_j$. The moment-dependent part of the upper coefficient in equation (172) divided by the lower coefficient in equation (23) is, up to constants depending only on a_j,

$$\frac{M_j^{1/a_j} \Delta_j^{-1/a_j}}{M_j^{2/(1+a_j)} \Delta_j^{-1}} = \left(\frac{M_j^{1/(1+a_j)}}{\Delta_j} \right)^{(1-a_j)/a_j}. \tag{208}$$

When the proof-only minimum group size dominates, the corresponding ratio is of order

$$\frac{\Delta_j}{M_j^{2/(1+a_j)} \Delta_j^{-1}} = \frac{\Delta_j^2}{M_j^{2/(1+a_j)}}. \tag{209}$$

At $a_j = 1$, the moment-dependent powers match. For $a_j < 1$, the ratios in equations (208) and (209) depend on the instance scale and gap and can grow without a class-independent bound. In this regime, the Gaussian benchmark matches the horizon order and optimal-neighbor support.

H Base-Design Constant Calculation

For any tuple (m, r_0, k, g), define its failure bound at design group size g (with balanced groups of size g or $g + 1$)

$$\beta(m, r_0, k, g) = \max\left\{ \frac{r_0}{m}, (m - r_0)\{B_k(1/4) + B_k(q_g)\} \right\}. \tag{210}$$

Whenever $\beta < 1/2$, the generic amplification result in remark 16, with exactly kg observations per base, gives the fixed-design sample and sign costs

$$n_p = kgL(p, \beta), \qquad s_p = (m - 1)n_p. \tag{211}$$

These are design-size cost proxies: q_g need not decrease with g, so equation (210) does not by itself certify all larger group sizes. The adaptive policy uses the uniform bound in equation (52), whose two-sided base bound stays at $\max\{1/4, 0.1348306174\ldots\} = 1/4$ and therefore leaves the chosen tuple's costs and $L(p, 1/4)$ unchanged. The rank-eight practical setting has a separate uniform certificate for group sizes at least six (lemma 27).

We evaluated every integer tuple with $3 \le m \le 64$, $1 \le r_0 < m/2$, odd $1 \le k \le 199$, and $1 \le g \le 99$. Binomial tails and the odd amplification inverse were evaluated directly. The search contains 2970 best-by-rank rows over $p \in \{10^{-2}, 10^{-6}, 10^{-9}\}$ and produces the 34-point sample–sign Pareto table shipped with the artifact.

Table 4: Representative optima of the design-size cost proxy. "Chosen" denotes the certified policy's constants, whose uniform certificate is equation (52); all optima refer to the stated proxy and integer box.

p	Selection	(m, r_0, k, g)	β	L	n_p	s_p
10^{-6}	Sign minimum	$(3, 1, 11, 3)$	0.345343	225	7425	14850
	Chosen	$(4, 1, 21, 3)$	0.250000	79	4977	14931
	Rank-five	$(5, 1, 27, 3)$	0.200000	51	4131	16524
	Sample minimum	$(42, 1, 103, 3)$	0.023810	9	2781	114021
10^{-9}	Chosen / sign minimum	$(4, 1, 21, 3)$	0.250000	125	7875	23625
	Rank-three	$(3, 1, 11, 3)$	0.345343	359	11847	23694
	Rank-five	$(5, 1, 27, 3)$	0.200000	81	6561	26244
	Sample minimum	$(56, 1, 113, 3)$	0.017857	13	4407	242385

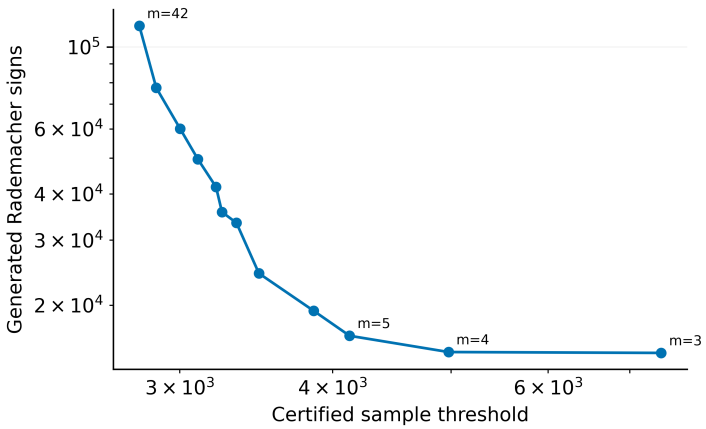

Figure 5: Complete sample–sign Pareto frontier of the design-size cost proxy at $p = 10^{-6}$ within the declared search box. The accompanying CSV contains all target probabilities and tuple fields.

Within this design-size proxy, at $p = 10^{-6}$ the chosen tuple uses 0.55% more signs than the sign minimum and 20.5% more samples than the rank-five design; it saves 9.6% of the rank-five signs. At $p = 10^{-9}$ it is sign-optimal. At $p = 10^{-2}$, $(4, 1, 21, 3)$ uses $L = 19$, $n_p = 1197$, and $s_p = 3591$, tying the sign optimum. These fixed-design operating points motivate rank four when randomization is the primary cost, rank five when activation samples are primary, and larger m only when sample threshold dominates a much larger sign budget.

For the earlier-certificate comparison at $p = 10^{-9}$, $L(p, 1/4) = 125$. The all-block tuple $(4, 1, 11, 39)$ requires

$$11 \cdot 39 \cdot 125 = 53{,}625 \tag{212}$$

observations. The tolerant certificate uses $(m, r_0, k, g) = (4, 1, 21, 3)$ and requires

$$21 \cdot 3 \cdot 125 = 7{,}875. \tag{213}$$

The raw threshold decreases by a factor 6.81; the first sufficient dyadic prefix changes from $65{,}536$ to $8{,}192$. At their respective raw thresholds, the sign counts are $160{,}875$ and $23{,}625$.

I Regret Guarantees for the Practical Configuration

The experimental setting $(m, r_0, k, g_{\min}) = (8, 1, 9, 1)$ attains neighbor-local logarithmic regret. Its endpoint uses seven sign-resampled components, nine groups per base, and the amplification count $L(p, 1/8)$. The policy, leader score, forced-pull schedule, and cache keys are those in section D. The runtime construction requires one observation per group. Its accuracy analysis below applies once each group contains at least six observations.

For a fixed budget scale $c \geq 1$, write

$$p_{r,q}^{(c)} = \frac{p_{r,q}}{c} = \frac{1}{c\, 2^{q+1} w(r) w(q)}, \qquad L_{r,q}^{(c)} = L(p_{r,q}^{(c)}, 1/8). \tag{214}$$

The matched-budget experiments use $c = 1$, and the scope-normalized graph-local experiments use $c = K(\gamma + 1)$. At key (j, r, q), partition the first 2^r observations of arm j into $L_{r,q}^{(c)}$ nearly equal bases, and each base into nine nearly equal groups. If $2^r < 9L_{r,q}^{(c)}$, set the index to $+\infty$. Otherwise, all groups are nonempty; take the lower median of the base endpoints, with independent signs and tie permutations across bases. Denote this cached index by $U_j^{\mathrm{p}}(r, q; c)$. All fixed-key variables can be defined on the underlying reward streams and random seeds, including keys the policy never visits.

I.1 A uniform rank-eight base certificate

Lemma 27 (Practical base upper tail)**.** *Suppose $X - \mu$ is symmetric and $\mathbb{E}|X - \mu|^{1+a} \leq M$ for $a \in (0, 1]$. Let a rank-eight base endpoint $U_{8,1}$ use nine disjoint groups, each containing at least $g \geq 6$ observations. With $\eta_g = (12M)^{1/(1+a)} g^{-a/(1+a)}$,*

$$\mathbb{P}(U_{8,1} - \mu > 4\eta_g) \leq \beta_{\mathrm{p}} := 7\{B_9(1/12) + B_9(47/256)\} = 0.09789276331140953\ldots < \frac{1}{8}. \tag{215}$$

The bound holds with the endpoint and tie conventions of section A.1.

Proof. Put $Y = X - \mu$. For a group of size s, lemma 10 gives $\mathbb{E}|\bar{Y}_s|^{1+a} \leq 2M/s^a$. The mean $\bar{Y}_s$ is symmetric, so for every $t > 0$,

$$\mathbb{P}(\bar{Y}_s > t) = \mathbb{P}(\bar{Y}_s < -t) = \tfrac{1}{2}\mathbb{P}(|\bar{Y}_s| > t) \leq \frac{M}{s^a t^{1+a}}. \tag{216}$$

This identity also holds for distributions with atoms. For $M > 0$ and $s \geq g$, substituting $t = \eta_g$ bounds each one-sided probability by $1/12$. A signed centered mean has the same distribution, and satisfies the same bound. If $M = 0$, the centered observations vanish almost surely; the strict deviation events at $\eta_g = 0$ are empty.

For a group of size s, let Z_s count its negative Rademacher signs. Then

$$\zeta_s = \mathbb{P}(Z_s < s/4) = 2^{-s} \sum_{v=0}^{\lceil s/4 \rceil - 1} \binom{s}{v}. \tag{217}$$

Exact summation for $6 \leq s \leq 17$ yields

$$\begin{aligned}(\zeta_6, \ldots, \zeta_{11}) &= \left(\frac{7}{64}, \frac{1}{16}, \frac{9}{256}, \frac{23}{256}, \frac{7}{128}, \frac{67}{2048}\right), \\ (\zeta_{12}, \ldots, \zeta_{17}) &= \left(\frac{79}{4096}, \frac{189}{4096}, \frac{235}{8192}, \frac{9}{512}, \frac{697}{65536}, \frac{1607}{65536}\right).\end{aligned} \tag{218}$$

Each value is at most $7/64$. For $s \geq 18$, Hoeffding's inequality gives $\zeta_s \leq e^{-s/8} \leq e^{-9/4} < 7/64$. Hence

$$\sup_{s \geq 6} \zeta_s \leq 7/64. \tag{219}$$

Let $\widehat{\mu}_0$ be the ordinary nine-group MoM center. For group ℓ of size s_ℓ and component $j \in \{1, \ldots, 7\}$, define

$$\begin{aligned} C &= \{\widehat{\mu}_0 - \mu > \eta_g\}, \\ D_{\ell,j} &= \{\overline{\alpha(X - \mu)}_{\ell,j} < -\eta_g\}, \qquad H_{\ell,j} = \{Z_{\ell,j} < s_\ell/4\}. \end{aligned} \tag{220}$$

On $C^c \cap D_{\ell,j}^c \cap H_{\ell,j}^c$, the centered intersection has numerator at most $2\eta_g$ and denominator at least $1/2$:

$$V_{\ell,j} - \mu = \frac{(\widehat{\mu}_0 - \mu) - \overline{\alpha(X-\mu)}_{\ell,j}}{1 - \bar{\alpha}_{\ell,j}} \le 4\eta_g. \tag{221}$$

Every zero denominator belongs to $H_{\ell,j}$, so signed infinities and the randomized $0/0$ convention are included in this event bound.

By symmetry, the signed centered data vector is independent of its sign vector (lemma 9). In particular, $D_{\ell,j}$ and $H_{\ell,j}$ are independent. Equations (216) and (219) therefore give

$$\mathbb{P}(D_{\ell,j} \cup H_{\ell,j}) \le \frac{1}{12} + \frac{11}{12}\frac{7}{64} = \frac{47}{256}. \tag{222}$$

The ordinary groups are independent, so $\mathbb{P}(C) \le B_9(1/12)$. For a fixed component j, the events $D_{\ell,j} \cup H_{\ell,j}$ are also independent across groups. On C^c, a component median above $\mu + 4\eta_g$ requires at least five of these group failures. The binomial domination in lemma 11 gives

$$\mathbb{P}(U^{(j)} - \mu > 4\eta_g) \le B_9(1/12) + B_9(47/256). \tag{223}$$

For $(m, r_0) = (8, 1)$, the base endpoint is the maximum of the seven component endpoints. A union bound proves equation (215); it requires neither independence between components nor conditioning on the shared ordinary center. The exact rational value is

$$\beta_{\mathrm{p}} = \frac{45495827913291159369435 05}{46475169741161613370589184} < \frac{1}{8}. \tag{224}$$

□

I.2 Fixed-prefix amplification and persistent accuracy

For $x > 0$, define the proof quantity

$$h_j^{\mathrm{p}}(x) = \max\left\{6, \left\lceil \left(\frac{12\, 4^{1+a_j} M_j}{x^{1+a_j}} \right)^{1/a_j} \right\rceil \right\}. \tag{225}$$

Lemma 28 (Practical amplification)**.** *For a deterministic prefix and $L = L(p, 1/8)$ independent rank-eight bases, the practical amplified endpoint U^{p} satisfies*

$$\mathbb{P}(U^{\mathrm{p}} < \mu_j) \le p. \tag{226}$$

If the prefix size is $n \ge 9h_j^{\mathrm{p}}(x)L$, it also satisfies

$$\mathbb{P}(U^{\mathrm{p}} > \mu_j + x) \le p. \tag{227}$$

When the endpoint is constructed, its randomization cost is exactly $7n$ Rademacher signs.

Proof. The rank coverage in proposition 8 bounds each base undercoverage probability by $1/8$ for every admissible fixed size. If a base cannot be constructed, the policy's $+\infty$ index has no undercoverage event. Otherwise, disjoint observations and independent base randomizations give independent base failures; a median undercoverage needs at least $(L+1)/2$ failures. Its probability is at most $B_L(1/8) \le p$.

Under the stated sample condition, the smallest within-base group has size

$$\left\lfloor \frac{\lfloor n/L \rfloor}{9} \right\rfloor \ge h_j^{\mathrm{p}}(x). \tag{228}$$

Equation (225) ensures $4\eta_{h_j^{\mathrm{p}}(x)} \le x$, including when $M_j = 0$. Lemma 27 bounds each base upper failure by $\beta_{\mathrm{p}} < 1/8$, so the same majority argument proves equation (227). Each observation receives seven signs within its unique base, giving $7\sum_{\ell=1}^{L} n_\ell = 7n$ signs in total. □

For an integer $Q \geq 0$, let $u_j^{\mathrm{p}}(x, Q; c)$ be the smallest dyadic number 2^{r_*} such that

$$2^r \geq 9h_j^{\mathrm{p}}(x)L(p_{r,Q}^{(c)}, 1/8) \quad \text{for every integer } r \geq r_*. \tag{229}$$

This threshold is nondecreasing in Q: the budgets decrease with Q, so the required amplification counts increase.

Lemma 29 (Practical accuracy threshold)**.** *With $H = h_j^{\mathrm{p}}(x)$ and $A = Q + 1 + \log c$,*

$$u_j^{\mathrm{p}}(x, Q; c) \leq 2^{21}H\{A + \log(e + HA)\}. \tag{230}$$

In particular, for a fixed instance and fixed c, $u_j^{\mathrm{p}}(x, \lfloor \log_2 s \rfloor; c) = O(\log(es))$.

Proof. The estimate $\log w(z) \leq 3\log(e + z)$ from equation (77) yields

$$\log(1/p_{r,Q}^{(c)}) \leq 4A + 3\log(e + r). \tag{231}$$

Since $B_v(1/8) \leq B_v(1/4)$ for every odd v, lemma 17 gives

$$L(p_{r,Q}^{(c)}, 1/8) \leq L(p_{r,Q}^{(c)}, 1/4) \leq 35\{A + \log(e + r)\}. \tag{232}$$

Put $D = A + \log(e + HA)$, $n_0 = 2^{20}HD$, and $r_*^0 = \lceil \log_2 n_0 \rceil$. Since $A \geq 1$ and $D \geq 1$, $\log H \leq D$ and $\log D \leq D$, whence $r_*^0 < 25D$ and $\log(e + r_*^0) < 5D$. Consequently,

$$9HL(p_{r_*^0,Q}^{(c)}, 1/8) \leq 9 \cdot 35 \cdot 6HD < n_0 \leq 2^{r_*^0}. \tag{233}$$

To verify persistence, set $D_r = A + \log(e + r)$. The ratio of consecutive values of $2^r/D_r$ is

$$\frac{2D_r}{D_r + \log(1 + 1/(e + r))} > 1, \tag{234}$$

because $D_r \geq 1$ and the added logarithm is smaller than one. Thus the stronger sufficient inequality $2^r \geq 9 \cdot 35H\{A + \log(e + r)\}$ persists for all $r \geq r_*^0$. It implies equation (229), and $2^{r_*^0} < 2n_0$ proves equation (230). □

I.3 Adaptive local counts and ascent

Define fixed-key events

$$E_{j,r,q}^{-} = \{U_j^{\mathrm{p}}(r, q; c) < \mu_j\},$$
$$E_{j,r,q}^{+}(x) = \begin{cases} \{U_j^{\mathrm{p}}(r, q; c) > \mu_j + x\}, & 2^r \geq 9h_j^{\mathrm{p}}(x)L_{r,q}^{(c)}, \\ \varnothing, & \text{otherwise.} \end{cases} \tag{235}$$

Lemma 28 bounds both event probabilities by $p_{r,q}^{(c)}$.

Lemma 30 (Practical local selection counts)**.** *Fix a leader i and $j \in \mathcal{N}^+(i) \setminus \mathcal{B}_i$. There is a nondecreasing process $Z_{i,j}^{\mathrm{p}}$ such that, simultaneously at every realized leader count $S \geq 1$,*

$$Q_{i,j}(S) \leq u_j^{\mathrm{p}}(\delta_{i,j}/2, \lfloor \log_2 S \rfloor; c) + Z_{i,j}^{\mathrm{p}}(S), \qquad \mathbb{E}Z_{i,j}^{\mathrm{p}}(\infty) \leq 4/c. \tag{236}$$

Proof. Fix $b_i^{\mathrm{loc}} \in \mathcal{B}_i$ and put $x = \delta_{i,j}/2$. For $S \geq 1$, define

$$Z_{i,j}^{\mathrm{p}}(S) = \sum_{q=0}^{\lfloor \log_2 S \rfloor} \sum_{r \geq 0} 2^q \left[\mathbf{1}\{E_{b_i^{\mathrm{loc}},r,q}^{-}\} + \mathbf{1}\{E_{j,r,q}^{+}(x)\} \right], \qquad Z_{i,j}^{\mathrm{p}}(0) = 0. \tag{237}$$

It is nondecreasing, and Tonelli's theorem gives

$$\mathbb{E}Z_{i,j}^{\mathrm{p}}(\infty) \leq \sum_{r,q \geq 0} 2^q \, 2p_{r,q}^{(c)} = \frac{1}{c} \left(\sum_{r \geq 0} \frac{1}{w(r)} \right)^2 \leq \frac{4}{c}, \tag{238}$$

using equation (103). No independence between cache events is needed in this expectation.

For a realized S, set $u = u_j^{\mathrm{p}}(x, \lfloor \log_2 S \rfloor; c)$. At most u selections of j occur with pre-decision count $N_j < u$, because each such selection increments that count. When $N_j \geq u$, its dyadic prefix is at least u. Equation (229) and the monotonicity of L in the leader epoch imply accuracy eligibility at every current $q \leq \lfloor \log_2 S \rfloor$. If neither the comparator's undercoverage nor the candidate's upper failure occurs, then their actual cached indices satisfy

$$U^{\mathrm{p}}_{b_i^{\mathrm{loc}}} \geq \mu_i^{\mathrm{loc}} > \mu_j + \delta_{i,j}/2 \geq U_j^{\mathrm{p}}. \tag{239}$$

Thus a nonforced selection of j after the threshold requires at least one of these failures. Each key can govern at most 2^q decisions within leader i's epoch q, so all such selections are bounded by equation (237). This argument holds on each trajectory for every realized S, proving equation (236).

Candidate selections using groups of sizes one through five occur before the accuracy threshold and enter the same sample-count term. A key can be shared across at most $\gamma + 1$ leaders; summing the fixed-leader bound over the finite graph preserves integrability. □

Lemma 31 (Practical ascent)**.** *Under equation* (1) *and graph unimodality,* $\mathbb{E}\ell_i(\infty) < \infty$ *for every suboptimal arm* i *under the practical graph-local policy.*

Proof. For $K = 1$, there are no suboptimal arms. For $K \geq 2$, graph unimodality implies $\gamma \geq 1$, and we verify the conditions of proposition 22. The practical policy initializes each arm once, maximizes the same dyadic MoM leader scores, selects only within the leader's closed neighborhood, and forces the leader at occurrences $s \equiv 1 \pmod{\gamma + 1}$.

For each $i \neq i^\star$, set

$$f_i^{\mathrm{p}}(s) = \sum_{j \in \mathcal{N}^+(i) \setminus \mathcal{B}_i} u_j^{\mathrm{p}}(\delta_{i,j}/2, \lfloor \log_2 s \rfloor; c),$$
$$Y_i^{\mathrm{p}} = \sum_{j \in \mathcal{N}^+(i) \setminus \mathcal{B}_i} Z_{i,j}^{\mathrm{p}}(\infty). \tag{240}$$

Lemma 30 gives the simultaneous pathwise bound $\sum_j Q_{i,j}(s) \leq f_i^{\mathrm{p}}(s) + Y_i^{\mathrm{p}}$ and $\mathbb{E}Y_i^{\mathrm{p}} \leq 4|\mathcal{N}^+(i) \setminus \mathcal{B}_i|/c$. Lemma 29 gives $f_i^{\mathrm{p}}(s) = O(\log(es)) = o(s)$. For the scores, take $\rho_h(r) = b_h(r)$ and $\varepsilon_h(r) = 2^{-4r}$ for $r \geq 8$, using lemma 20; take $\rho_h(r) = 0$ and $\varepsilon_h(r) = 1$ for $r < 8$. Then

$$\rho_h(r) \to 0, \qquad \mathbb{P}(|S_h(r) - \mu_h| > \rho_h(r)) \leq \varepsilon_h(r), \qquad \sum_{r \geq 0} 2^r \varepsilon_h(r) = \sum_{r=0}^{7} 2^r + \sum_{r \geq 8} 2^{-3r} < \infty. \tag{241}$$

All premises of proposition 22 now hold.

Its proof also gives a finite upper bound for the ascent constant. Specifically, put $d_i = |\mathcal{B}_i|$, $\eta = \gamma/[2(\gamma + 1)]$, and $c_i = \min\{1/[2(\gamma + 1)], \eta/d_i\}$. Choose s_i^A so that the score radii are below $(\mu_i^{\mathrm{loc}} - \mu_i)/4$ at every epoch at least $\max\{0, \lfloor \log_2(c_i s) \rfloor\}$ for $s \geq s_i^A$. Choose s_i^C so that $f_i^{\mathrm{p}}(s) + 2 \leq \eta s/2$ for $s \geq s_i^C$. Both choices are finite by the properties just proved. The high-count and low-count bounds (164) and (165) yield

$$\mathbb{E}\ell_i(\infty) \leq s_i^A + s_i^C + \frac{2}{c_i} \sum_{h \in \{i\} \cup \mathcal{B}_i} \sum_{r \geq 0} 2^r \varepsilon_h(r) + \frac{2}{\eta} \mathbb{E}Y_i^{\mathrm{p}} < \infty. \tag{242}$$

□

I.4 Complete local and global regret bounds

Theorem 32 (Practical neighbor-local regret)**.** *Assume equation* (1)*, graph unimodality, and a unique optimal arm. The practical* APR-Local *configuration* $(8, 1, 9, 1)$ *with amplification* $L(p, 1/8)$ *and any fixed budget scale* $c \geq 1$ *satisfies, for every* $T \geq K$*,*

$$\mathbb{E}R_T \leq C_{\mathrm{init}} + C_{\mathrm{p}}^{\mathrm{asc}} + \sum_{j \in \mathcal{N}(i^\star)} \Delta_j \left[u_j^{\mathrm{p}}(\Delta_j/2, \lfloor \log_2 T \rfloor; c) + \frac{4}{c} \right], \tag{243}$$

where $C_{\text{init}} = \sum_{i \neq i^\star} \Delta_i$ *and*

$$C_{\text{p}}^{\text{asc}} = \Delta_{\max} \sum_{i \neq i^\star} \mathbb{E}\ell_i(\infty) < \infty, \qquad \Delta_{\max} = \max_i \Delta_i. \tag{244}$$

For each fixed instance and fixed finite graph, this gives

$$\mathbb{E}R_T = O\left(\sum_{j \in \mathcal{N}(i^\star)} [\Delta_j + M_j^{1/a_j} \Delta_j^{-1/a_j}] \log T \right) + C_{\text{p}}, \qquad C_{\text{p}} < \infty. \tag{245}$$

The constant C_{p} *is independent of* T*, and the numerical constants in the* O *notation may depend on the fixed tail exponents. Both* $c = 1$ *and* $c = K(\gamma + 1)$ *are covered.*

Proof. When $K = 1$, every term is zero. Otherwise, initialization incurs C_{init}. Every subsequent decision with a suboptimal leader has regret at most $\Delta_{\max}$. The total contribution of those decisions is bounded pathwise by $\Delta_{\max} \sum_{i \neq i^\star} \ell_i(T)$. Lemma 31 and monotone convergence give the finite expectation in equation (244).

For the optimal leader, $\mathcal{B}_{i^\star} = \{i^\star\}$ and forced pulls incur no regret. Put $S_T = \ell_{i^\star}(T) \leq T$. If $S_T = 0$, its neighbor selection counts vanish. Otherwise, the simultaneous bound in lemma 30, evaluated on the realized trajectory, and monotonicity in Q give

$$Q_{i^\star,j}(S_T) \leq u_j^{\text{p}}(\Delta_j/2, \lfloor \log_2 T \rfloor; c) + Z_{i^\star,j}^{\text{p}}(\infty). \tag{246}$$

Thus

$$\begin{aligned} R_T \leq C_{\text{init}} &+ \Delta_{\max} \sum_{i \neq i^\star} \ell_i(T) \\ &+ \sum_{j \in \mathcal{N}(i^\star)} \Delta_j \left[u_j^{\text{p}}(\Delta_j/2, \lfloor \log_2 T \rfloor; c) + Z_{i^\star,j}^{\text{p}}(\infty) \right]. \end{aligned} \tag{247}$$

Taking expectations and applying equation (238) proves equation (243).

For the rate, let $H_j = h_j^{\text{p}}(\Delta_j/2)$ and $A_T = \lfloor \log_2 T \rfloor + 1 + \log c$. Since $\max\{6, \lceil y \rceil\} \leq 6 + y$ for $y \geq 0$,

$$\Delta_j H_j \leq 6\Delta_j + \left(12\, 8^{1+a_j}\right)^{1/a_j} M_j^{1/a_j} \Delta_j^{-1/a_j}. \tag{248}$$

Because $A_T \geq 1$, $\log(e + H_j A_T) \leq \log(e + H_j) + \log A_T \leq \log(e + H_j) + A_T$. Hence lemma 29 implies

$$\Delta_j u_j^{\text{p}}(\Delta_j/2, \lfloor \log_2 T \rfloor; c) \leq 2^{21} \Delta_j H_j \{2A_T + \log(e + H_j)\}. \tag{249}$$

Combining equations (248) and (249) proves equation (245). For example, one may take

$$C_{\text{p}} = C_{\text{init}} + C_{\text{p}}^{\text{asc}} + \sum_{j \in \mathcal{N}(i^\star)} \Delta_j \left[\frac{4}{c} + 2^{21} H_j \{2(1 + \log c) + \log(e + H_j)\} \right]. \tag{250}$$

The growing sum ranges only over optimal-arm neighbors. The finite remainder includes ascent from all suboptimal leaders and can depend on the full graph and instance. □

Corollary 33 (Practical global-clock regret)**.** *Assume equation* (1) *and a unique optimal arm. After one initial pull per arm, let the global policy maximize the practical cached indices over all arms, with* $q = \lfloor \log_2 t \rfloor$ *and budget* $p_{r,q}^{(c)}$*. For any fixed* $c \geq 1$ *and* $T \geq K$*,*

$$\mathbb{E}R_T \leq C_{\text{init}} + \sum_{j \neq i^\star} \Delta_j \left[u_j^{\text{p}}(\Delta_j/2, \lfloor \log_2 T \rfloor; c) + \frac{4}{c} \right]. \tag{251}$$

In particular, the experimental scope budget $c = K$ *gives*

$$\mathbb{E}R_T = O\left(\sum_{j \neq i^\star} [\Delta_j + M_j^{1/a_j} \Delta_j^{-1/a_j}] \log T \right) + C_{\text{p}}^{\text{glob}}, \qquad C_{\text{p}}^{\text{glob}} < \infty. \tag{252}$$

Proof. For each suboptimal j, fix $x = \Delta_j/2$ and $u = u_j^{\mathrm{p}}(x, \lfloor \log_2 T \rfloor; c)$. At most u of its post-initialization pulls have pre-decision count below u. After this threshold, the candidate key is accuracy-eligible. If the optimal arm's index does not under-cover and arm j's index does not exceed $\mu_j + x$, then $U_{i^\star}^{\mathrm{p}} \geq \mu_{i^\star} > \mu_j + x \geq U_j^{\mathrm{p}}$, which excludes selecting j. Define

$$Z_j^{\mathrm{glob}}(T) = \sum_{q=0}^{\lfloor \log_2 T \rfloor} \sum_{r \geq 0} 2^q \left[\mathbf{1}\{E_{i^\star,r,q}^-\} + \mathbf{1}\{E_{j,r,q}^+(x)\} \right]. \tag{253}$$

A global epoch contains at most 2^q decisions, so $N_j(T) \leq 1 + u + Z_j^{\mathrm{glob}}(T)$ pathwise. The same series calculation as equation (238) gives $\mathbb{E}Z_j^{\mathrm{glob}}(\infty) \leq 4/c$. Multiplying by gaps and summing proves equation (251). Equations (248) and (249) then give equation (252), with a finite remainder obtained from equation (250) by omitting ascent and summing over all suboptimal arms. □

J Complete path-size comparisons

J.1 Instances, algorithms, and pairing

We compare the same eight main-table methods on 15-, 31-, and 63-arm paths. Arm i has coordinate $x = i - (K-1)/2$ and mean $\mu_i = -0.04|x|$. Adjacent coordinates share an edge, so the maximum degree remains two. The optimum and its two neighbors have means $0, -0.04, -0.04$ at every size. Rewards follow equations (255) and (256) with unchanged noise scales. Larger paths add more distant arms: the largest gap is 0.28, 0.60, and 1.24, respectively. One-pull-per-arm initialization costs 2.24, 9.60, and 39.68 regret.

Each size and reward law uses 30 independent trials and $T = 50{,}000$ scalar observations. All methods share each arm's pre-generated reward stream within a trial. The 15-arm benchmark retains its original data. Coordinates $-7, \ldots, 7$ reuse the historical noise draws and seed tree. The archived simulator generates streams with peak mean 0.8; subtracting 0.8 centers them at the means above. Each added coordinate uses NumPy's `SeedSequence` with `[20260803, 20260921, scenario, trial, coordinate+31]`. Scenario identifiers are zero for Pareto and one for Student-t, and trial identifiers range from zero to 29. Common coordinates have identical reward streams across path sizes.

Policies retain the original implementations and parameter settings. Their seeds are children of the original trial-specific policy seed root: OSUB-UCB1 uses child 1, Oracle-MoM-OSUB child 3, practical APR-UCB child 4, APR-LCB$_{\mathrm{pract}}$ child 5, APR-LCB$_{\mathrm{cert}}$ child 9, APR-LCB$_{\mathrm{emp}}$ child 10, AdaR-UCB child 11, and uniINF child 12. Child indices start at zero. The root follows `SeedSequence(20260803, spawn_key=(scenario, trial))`, whose first child generates reward streams and second child generates policy seeds. Seeds are retained across sizes.

Two additional settings supply the mechanism comparisons. Certified APR-UCB uses child 7, tuple $(4, 1, 21, 3)$, and budget $p_{r,q}$. The matched-budget practical local setting uses child 5 and tuple $(8, 1, 9, 1)$ with budget $p_{r,q}$. The main practical local and global settings retain budgets $p_{r,q}/(3K)$ and $p_{r,q}/K$. Oracle methods receive the centered-moment bounds in section L. AdaR-UCB consumes 25,000 two-reward decisions, and uniINF consumes 50,000 single-reward decisions with $S_1 = 4$ and the supplied horizon. Both retain their original initialization rules.

The path extension holds the optimal neighborhood, edge gap, and noise law fixed while changing distant-arm gaps, initialization cost, and the algorithms' existing K-dependent budget and update terms. The complete 15-arm benchmark and its mechanism studies appear in sections K to M.

J.2 Statistical analysis

Mean intervals use Student-t quantiles across the 30 trials. For the 31- and 63-arm comparisons, paired percentile-bootstrap intervals use 20,000 resamples with analysis seed 2026092103. The 28 pairwise contrasts among the eight main methods are evaluated in each of four size–law combinations. Two-sided Wilcoxon signed-rank tests use Holm correction jointly over these 112 contrasts. The matched-budget family contains eight tests: terminal and preactivation regret at both sizes under both laws. It receives a separate Holm correction. Bootstrap intervals are pointwise; adjusted test

probabilities appear in the released results. The path extensions are exploratory, and all runs are retained. The 15-arm tables preserve their original analysis protocols and intervals.

J.3 Full comparisons at 31 and 63 arms

Table 5: Final pseudo-regret on the 31-arm path at $T = 50{,}000$ (mean $\pm$ 95% Student-t half-width) and amplified-cache activation over 30 trials per law. A dash marks methods without this cache; lower regret is better.

Method	Graph	Pareto	Student-$t_{1.5}$	Activation
AdaR-UCB	No	4449.10±126.59	19078.55±490.33	–
uniINF	No	824.96±20.93	882.73±24.68	–
Practical APR-UCB	No	5488.64±63.67	5270.16±73.84	30/30, 30/30
OSUB-UCB1	Yes	997.07±228.70	1350.08±509.11	–
Oracle-MoM-OSUB	Yes	1213.73±27.73	1156.55±17.47	–
APR-LCB$_{\text{cert}}$	Yes	1213.41±31.05	1151.42±13.46	30/30, 30/30
APR-LCB$_{\text{pract}}$	Yes	**585.55±60.17**	**475.77±57.27**	30/30, 30/30
APR-LCB$_{\text{emp}}$	Yes	694.68±99.83	1079.83±514.27	30/30, 30/30

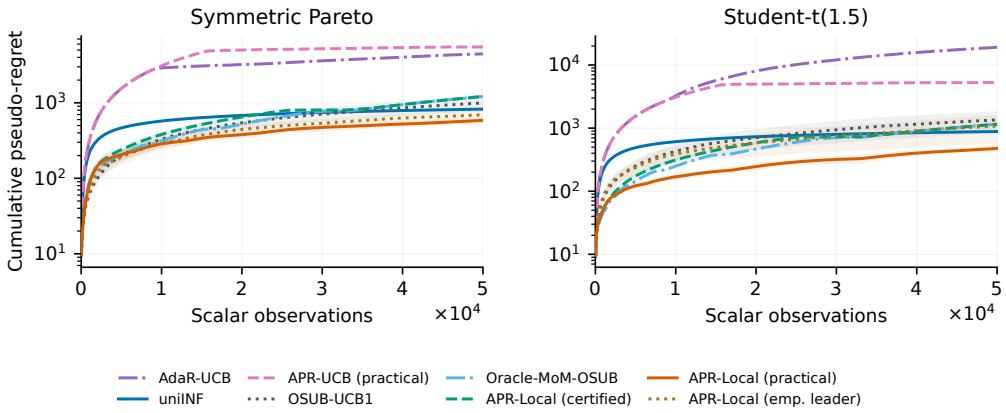

Figure 6: The complete eight-method comparison on the 31-arm path. Curves are mean cumulative pseudo-regret with pointwise 95% Student-t intervals over 30 trials. The ordinate is logarithmic.

Table 6: Comparator minus APR-LOCAL$_{\mathrm{pract}}$ terminal regret. Positive values favor APR-LOCAL$_{\mathrm{pract}}$. Entries give mean paired differences, followed by pointwise 95% bootstrap intervals. Each size and law has 30 paired trials.

K	Comparator	Pareto	Student-$t_{1.5}$
31	AdaR-UCB	3863.55	18602.78
		[3727.64, 4016.30]	[18121.85, 19080.16]
31	uniINF	239.41	406.96
		[179.16, 298.70]	[361.10, 453.54]
31	Practical APR-UCB	4903.09	4794.38
		[4840.65, 4964.11]	[4730.06, 4861.61]
31	OSUB-UCB1	411.52	874.31
		[195.56, 649.25]	[447.43, 1418.33]
31	Oracle-MoM-OSUB	628.18	680.77
		[584.02, 673.25]	[629.34, 732.52]
31	APR-LOCAL$_{\mathrm{cert}}$	627.86	675.65
		[588.73, 667.97]	[623.28, 729.15]
31	APR-LOCAL$_{\mathrm{emp}}$	109.13	604.06
		[9.95, 204.60]	[202.27, 1144.43]
63	AdaR-UCB	28113.25	39269.70
		[27610.92, 28599.76]	[38990.80, 39544.42]
63	uniINF	1911.75	2304.55
		[1844.17, 1984.70]	[2238.15, 2374.92]
63	Practical APR-UCB	19992.89	19995.08
		[19925.59, 20059.35]	[19947.77, 20041.92]
63	OSUB-UCB1	375.21	862.43
		[163.92, 603.54]	[443.80, 1382.67]
63	Oracle-MoM-OSUB	590.53	668.55
		[544.87, 638.67]	[615.71, 719.47]
63	APR-LOCAL$_{\mathrm{cert}}$	590.50	664.26
		[546.26, 635.42]	[612.50, 715.46]
63	APR-LOCAL$_{\mathrm{emp}}$	52.49	627.17
		[−49.00, 154.69]	[240.76, 1164.22]

J.4 MATCHING ENDPOINT BUDGETS

The certified and practical local policies use different endpoint tuples. We compare them at the same budget $p_{r,q}$, retaining their leader, cache, and action rules. Each trajectory is split at its own first finite amplified endpoint: preactivation regret includes rounds before that endpoint is first queried, and the remainder includes that round onward. Table 7 reports both parts. The practical instance- and scope-budget settings share their initial policy seed.

Table 7: Matched-budget local comparison at $p_{r,q}$. C and P denote certified and practical endpoint tuples. Pre and Post split each policy at its first finite endpoint. Total entries are means ± 95% Student-t half-widths over 30 trials.

K	Law	Tuple	Pre	Post	Total
31	Pareto	C	799.48	413.93	1213.41 ± 31.05
31	Pareto	P	86.93	488.52	575.44 ± 57.56
31	Student-t	C	725.53	425.89	1151.42 ± 13.46
31	Student-t	P	65.53	402.57	468.10 ± 57.04
63	Pareto	C	831.84	413.01	1244.85 ± 31.40
63	Pareto	P	118.91	488.75	607.66 ± 57.56
63	Student-t	C	759.93	424.95	1184.87 ± 14.89
63	Student-t	P	99.93	402.48	502.40 ± 57.31

J.5 Neighbor and distant-arm regret

Table 8: Mean regret from the optimal-arm neighbors and from arms outside the optimal closed neighborhood. The optimal arm contributes zero. Components sum to total regret before rounding.

		Optimal-arm neighbors		Outside the closed neighborhood	
Law	K	APR-LOCAL$_{\text{pract}}$	uniINF	APR-LOCAL$_{\text{pract}}$	uniINF
Pareto	15	458.37	62.80	120.45	328.94
Pareto	31	448.93	63.74	136.62	761.22
Pareto	63	482.96	70.90	171.39	2495.20
Student-t	15	412.26	68.11	33.02	355.38
Student-t	31	428.66	68.91	47.12	813.82
Student-t	63	439.19	77.89	81.42	2747.28

J.6 Execution and consistency checks

The 31- and 63-arm experiments comprise 1,200 policy runs across eight main-table methods and two mechanism settings, each with 50,000 scalar observations. The code records seeds, 90 regret checkpoints, and final pull counts for each run. APR runs also record first-finite-endpoint counts and endpoint-build statistics. Every run passes pull-count conservation, regret reconstruction, curve monotonicity, and regional decomposition checks; APR runs also pass activation-split checks. Historical replay checks under both reward laws verify the seed mapping and the unchanged policy implementations.

The full three-size practical–uniINF diagnostic retains 360 action traces. Its 15-arm replay gives the same pull counts as the archived 120 policy runs. With both policies split at the practical policy's first finite endpoint, the practical-minus-uniINF Pareto regret is -24.69 before activation and 211.77 from activation onward. Neighbor choices with a correct leader and finite candidate endpoints account for 86.84% of practical post-activation Pareto regret. The comparison tables use trials as sampling units.

K 15-arm benchmark and mechanism studies

K.1 Protocol

The 15-arm study uses a path with $\mu_i = -0.04|i - 7|$, $i = 0, \ldots, 14$, under the same reward laws, trial count, pairing, inference, and endpoint tuples as the main protocol (section 6). Full distributions, baselines, statistics, and integrity checks appear in section L.

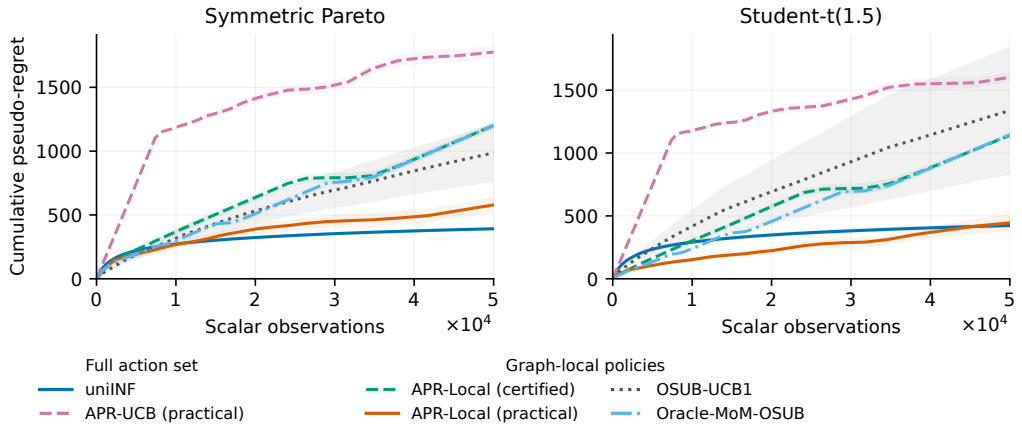

Figure 7: Mean cumulative pseudo-regret with 95% Student-t intervals over 30 paired trials on the 15-arm path. Legends distinguish full-set and graph-local policies; uniINF receives the horizon.

Table 9: Final pseudo-regret on the 15-arm path at $T = 50{,}000$ (mean $\pm$ 95% half-width) and amplified-cache activation. A dash marks methods without this cache; lower regret is better.

Method	Graph	Pareto	Student-$t_{1.5}$	Activation
AdaR-UCB	No	1659.19 ± 34.59	8826.74 ± 692.21	–
uniINF	No	$\mathbf{391.74 \pm 13.37}$	$\mathbf{423.49 \pm 19.97}$	–
Practical APR-UCB	No	1775.98 ± 44.92	1602.35 ± 37.96	$30/30, 30/30$
OSUB-UCB1	Yes	986.41 ± 227.78	1338.55 ± 509.81	–
Oracle-MoM-OSUB	Yes	1200.81 ± 26.24	1146.20 ± 17.62	–
APR-LOCAL$_{\text{cert}}$	Yes	1201.73 ± 30.20	1140.75 ± 13.54	$30/30, 30/30$
APR-LOCAL$_{\text{pract}}$	Yes	578.82 ± 62.08	445.28 ± 71.68	$30/30, 30/30$
APR-LOCAL$_{\text{emp}}$	Yes	688.47 ± 95.77	1055.30 ± 517.17	$30/30, 30/30$

K.2 FINDINGS

uniINF has the lowest sample mean in both environments. Relative to uniINF, APR-LOCAL$_{\text{pract}}$ incurs 187.08 more regret on Pareto (paired bootstrap 95% interval $[127.07, 245.91]$) and 21.79 on Student-t ($[-37.97, 84.35]$). The Student-t interval includes zero.

Replacing only the dyadic MoM leader by an empirical mean raises paired regret by 109.65 on Pareto (bootstrap 95% interval $[24.21, 199.10]$) and 610.02 on Student-t ($[207.44, 1173.41]$). The optimal-neighborhood pull share falls from 98.1% to 95.9% and from 99.3% to 89.2%.

Against AdaR-UCB, practical APR-UCB loses 116.78 on Pareto and wins 7224.39 on Student-t; both paired intervals exclude zero (section M.1). OSUB-UCB1 has lower mean Pareto regret than APR-LOCAL$_{\text{cert}}$, with a paired difference interval $[-427.90, 17.44]$. Its empirical 90th percentile is 2028.77, versus 1278.34 for APR-LOCAL$_{\text{cert}}$, showing greater dispersion across these 30 runs (figure 19).

K.3 CACHE ACTIVATION AND FINITE-HORIZON REGRET

At the same budget $p_{r,q}$, mean regret before first activation is 787.01 versus 74.81 on Pareto and 714.19 versus 54.19 on Student-t, for certified versus practical constants. Most of the observed total gap precedes activation (section M.3).

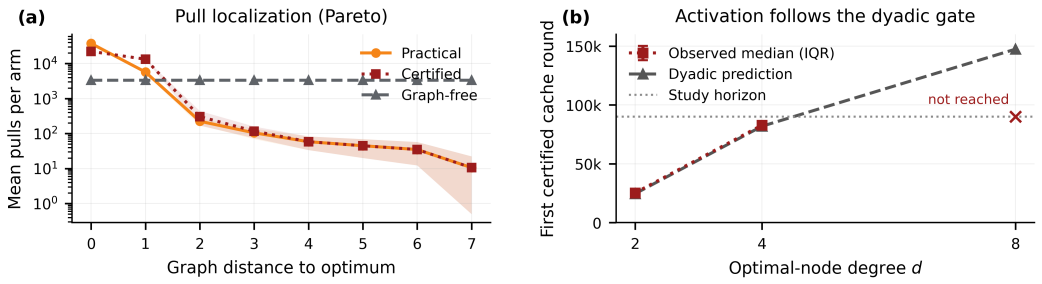

Figure 8: Pull allocation and cache activation on 15-arm graphs. (a) Mean pulls per arm with 95% Student-t intervals over 30 Pareto trials. (b) Certified first-cache median and IQR (interquartile range) over 30 trials on 15-arm trees with root degree $d \in \{2, 4, 8\}$, edge gap 0.04, and symmetric Pareto rewards. The cross marks no activation by $T = 90{,}000$; branch lengths are given in section L.13.

In figure 8a, robust leaders keep 97.7–99.3% of pulls in the optimal closed neighborhood; certified global pulls are nearly uniform.

In figure 8b, predicted first-cache times track medians at degrees two and four and exceed the horizon at degree eight. Median leader resolution occurs at least ten times earlier. All 30 certified regret trajectories per law stay constant from $70{,}861$ to $200{,}000$ in the paired extension (figure 14).

Under mean-preserving asymmetric perturbations with probability 0.15, practical Student-t regret rises by 101.68 (paired interval $[47.84, 157.17]$), while Pareto regret changes by 0.27 ($[-46.07, 44.56]$). Section M.4 gives the law and diagnostics.

L Complete Experimental Protocol and Results

This appendix compares leader scores, cache rules, confidence budgets, and rank-amplification choices, and measures cache activation and graph localization. The 15-arm regret and two-arm comparisons use 30 paired trials; the controlled graph and parameter sweeps use 12 paired trials per configuration. The controlled-degree study uses 30 trials. Each table states its inferential convention.

L.1 Position among parameter-adaptive methods

Table 10 expands the related-work comparison by separating run-time inputs, reward assumptions, and rate objectives. In particular, it distinguishes the unrestricted polynomial adaptation frontier from the logarithmic regime enabled by exact symmetry.

Table 10: Information and assumption map for closely related adaptive heavy-tailed bandit methods.

Method	Run-time inputs	Reward structure	Main objective
APE^2 (Lee et al., 2020)	Tail order; moment scale unknown	Stochastic rewards with a finite specified-order moment	Gap-dependent heavy-tail regret with scale adaptation
Median clipping (Kornilov et al., 2024)	Horizon and clipping/noise tuning	Symmetric heavy-tailed noise	Problem-independent bandit/zeroth-order guarantees
uniINF (Chen et al., 2025)	Horizon; tail order and scale unknown	Optimal-arm truncated non-negative loss condition	Stochastic–adversarial best of both worlds
Parameter-Free HTB (Genalti & Metelli, 2026)	Horizon; tail order and scale unknown	Union of finite-$(1+\epsilon)$ moment classes; pointwise in fixed ϵ	Parameter-free polynomial-rate frontier
RMM-UCB (Tamás et al., 2025)	Moment parameters and horizon unknown	Exact arm-wise symmetry; finite centered $(1+a_i)$-th moments	Anytime, all-suboptimal-arm $\log^2 T$ guarantee
APR-UCB/ APR-Local	Moment parameters and horizon unknown	Exact symmetry and finite centered $(1+a_i)$-th moments; known unimodal graph for APR-Local	Anytime all-arm $\log T$; optimal-neighbor $\log T$ with graph

L.2 15-arm benchmark environments

The benchmark graph is a 15-arm path indexed by $i = 0, \dots, 14$, with unique optimum $i^\star = 7$ and

$$\mu_i = -0.04|i - 7|. \tag{254}$$

We use two symmetric infinite-variance reward families.

1. **Symmetric Pareto:**

$$X_i = \mu_i + 0.12S(1+Y), \qquad S \sim \mathrm{Rad}(1/2), \quad \mathbb{P}(Y > y) = (1+y)^{-1.5},\ y \geq 0. \tag{255}$$

   Here S is a Rademacher sign, Y is independent of S, and 0.12 is the noise scale; $\mathrm{Rad}(1/2)$ assigns probabilities $1/2$ to each of -1 and $+1$.

2. **Student-t:**

$$X_i = \mu_i + 0.18W, \qquad W \sim t_{1.5}. \tag{256}$$

   Here $t_{1.5}$ denotes the centered Student distribution with 1.5 degrees of freedom, and 0.18 is the noise scale.

Both families have finite centered 1.4-th moments and infinite variance. The oracle MoM baselines receive $a = 0.4$ and the exact centered-moment bounds

$$M_{\mathrm{P}} = 15(0.12)^{1.4}, \qquad M_{\mathrm{Stu}} = (0.18)^{1.4}(1.5)^{0.7}\frac{\Gamma(1.2)\Gamma(0.05)}{\sqrt{\pi}\,\Gamma(0.75)},$$

where Γ is Euler's gamma function (numerically, $M_{\mathrm{P}} \approx 0.770806$ and $M_{\mathrm{Stu}} \approx 0.991008$); neither APR policy receives a, M_{P}, or M_{Stu}.

The optimal reward is symmetric about zero. Hence, for every $u > 0$,

$$\mathbb{E}[X_{i^\star}\mathbf{1}\{|X_{i^\star}| > u\}] = 0, \tag{257}$$

which satisfies the truncated non-positivity condition used by AdaR-UCB (Genalti et al., 2024). For losses $\ell = -X$, it also gives $\mathbb{E}[\ell_{i^\star}\mathbf{1}\{|\ell_{i^\star}| > u\}] = 0$, satisfying uniINF's truncated non-negativity condition (Chen et al., 2025). The historical reward arrays used optimal mean 0.8; subtracting 0.8 gives the means in equation (254). A translation-equivariance audit replayed the first three original trials per environment under this shift. All recorded index, sign-generation, and leader-score diagnostics agreed; the largest terminal pseudo-regret difference was 4.6×10^{-13}.

L.3 Trials, pairing, and inference

Each 15-arm environment uses 30 independent trials of horizon $T = 50{,}000$, starting from master seed 20260803. Within one trial, all policies receive the same pre-generated reward stream for each arm. Each randomized policy receives an independent child seed, so environmental noise is paired without forcing the policies to make identical randomized choices. Cumulative pseudo-regret uses the known mean vector and is stored at 90 logarithmic checkpoints. We also record arm pulls, finite-index builds, first activation time, random-sign counts, leader occurrences, and wall-clock time.

Pointwise bands are 95% Student-t intervals across the 30 independent trials. The six original ablations, four AdaR contrasts, and matched leader ablation were specified before their corresponding batches. Paired effects use 20,000-resample percentile bootstrap intervals and two-sided Wilcoxon signed-rank tests, with Holm correction within each contrast family across the two environments. Activation rates use exact Clopper–Pearson intervals. Wall-clock time is reported only as an execution diagnostic.

L.4 Baselines and implementation scope

The comparison set varies parameter access, adaptive state, and leader estimation.

- UCB1 and OSUB-UCB1 are empirical-mean references whose classical guarantees use light-tailed rewards.
- Oracle-MoM-UCB and Oracle-MoM-OSUB receive the true moment order and a valid moment bound.
- APR-Local$_{\mathrm{cert}}$ uses $(m, r_0, k, g_{\min}) = (4, 1, 21, 3)$ with budget $p_{r,q}$. APR-Local$_{\mathrm{pract}}$ uses $(m, r_0, k, g_{\min}) = (8, 1, 9, 1)$ with the scope-normalized budget $p_{r,q}/[K(\gamma + 1)]$, where $g_{\min}$ is the minimum allowed group size. The matched-budget comparison in section M.3 uses $p_{r,q}$ for both tuples. Their amplification counts are $L(p, 1/4)$ and $L(p, 1/8)$, respectively, where p is the assigned cache budget. Both satisfy the regret bounds in theorems 4 and 32.
- APR-Local$_{\mathrm{emp}}$ shares every practical-setting component except the leader score, which is the empirical mean at the same dyadic epochs.
- The all-block certificate uses $(m, r_0, k, g) = (4, 1, 11, 39)$; the tolerant certificate uses $(4, 1, 21, 3)$. Instance-budget policies use $p_{r,q}$ from equation (8); scope-normalized policies use $p_{r,q}/K$ globally and $p_{r,q}/[K(\gamma + 1)]$ locally. The certified global policy uses the instance budget $p_{r,q}$.
- AdaR-UCB follows the published two-sample-per-decision algorithm and receives 50,000 scalar rewards, equal to every other policy's total observation budget.
- uniINF is an independent implementation of the published Algorithm 1 of Chen et al. (2025). It receives T, uses the full action set and one scalar loss per decision, and uses the published constants without tuning; section M.2 specifies its updates.
- Direct RMM-UCB follows its published rank schedule in a separate two-arm bounded-horizon functional and cost diagnostic. Regret and sign counts through $T = 160$ audit the faithful schedule; its terminal rank is then extrapolated algebraically to the main horizon.

L.5 Complete 15-arm regret table

The graph-local configuration reduces regret under both reward families. Relative to practical APR-UCB, APR-Local$_{\mathrm{pract}}$ reduces mean final pseudo-regret by 1197.16 on Pareto (bootstrap 95%

Table 11: Final pseudo-regret at $T = 50{,}000$: mean $\pm$ 95% Student-t half-width over 30 trials. Lower is better.

Algorithm	Symmetric Pareto	Student-$t_{1.5}$
UCB1	2114.17 ± 268.79	1985.32 ± 249.71
OSUB-UCB1	986.41 ± 227.78	1338.55 ± 509.81
Oracle-MoM-UCB	6398.72 ± 4.55	6399.85 ± 5.00
Oracle-MoM-OSUB	1200.81 ± 26.24	1146.20 ± 17.62
AdaR-UCB	1659.19 ± 34.59	8826.74 ± 692.21
uniINF	$\mathbf{391.74 \pm 13.37}$	$\mathbf{423.49 \pm 19.97}$
Practical APR-UCB	1775.98 ± 44.92	1602.35 ± 37.96
APR-LOCAL$_{\text{pract}}$	578.82 ± 62.08	445.28 ± 71.68
APR-LOCAL$_{\text{emp}}$	688.47 ± 95.77	1055.30 ± 517.17
All-block APR-LOCAL	1437.64 ± 31.69	1377.74 ± 28.55
Certified APR-UCB	7466.92 ± 0.00	7466.92 ± 0.00
APR-LOCAL (scope)	1434.19 ± 33.03	1364.77 ± 30.77
APR-LOCAL$_{\text{cert}}$	1201.73 ± 30.20	1140.75 ± 13.54

interval $[1142.28, 1250.20]$) and 1157.07 on Student-t ($[1091.88, 1218.60]$), winning all 30 pairs. The instance-budget certified APR-UCBnever builds a finite index and rotates across all arms. Its certified APR-Local counterpart reduces regret by 6265.19 and 6326.17, with both Holm-adjusted Wilcoxon values below 6×10^{-6}.

The matched leader comparison holds amplified endpoints, prefix caches, local clocks, neighborhoods, and forced sampling fixed. Relative to APR-LOCAL$_{\text{emp}}$, APR-LOCAL$_{\text{pract}}$ reduces mean paired regret by 109.65 on Pareto (bootstrap 95% interval $[24.21, 199.10]$) and 610.02 on Student-t ($[207.44, 1173.41]$). The corresponding optimal-neighborhood pull shares rise from 95.9% to 98.1% and from 89.2% to 99.3%.

L.6 ACTIVATION AND CERTIFIED-CONSTANT ABLATIONS

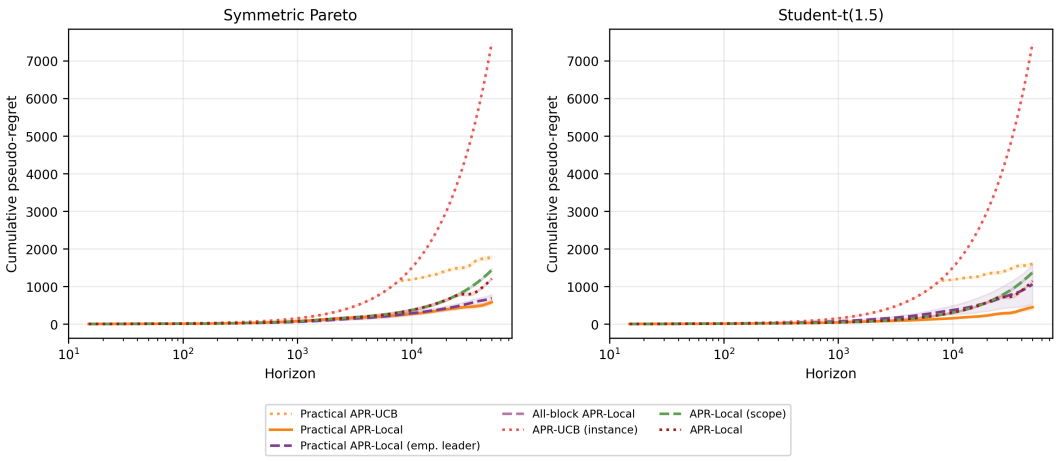

Figure 9: Endpoint-constant, confidence-budget, and leader ablations. The tolerant constants form finite indices within the evaluated horizon, and the matched empirical-leader curve uses the practical endpoint.

Holding the instance budget fixed, the tolerant certificate reduces mean regret relative to the all-block certificate by 235.91 on Pareto (bootstrap 95% interval $[229.39, 244.16]$) and 237.00 on Student-t ($[225.29, 254.16]$). It wins all paired trials, with adjusted Wilcoxon $p < 6 \times 10^{-6}$. The 30/30 activation proportion has exact 95% interval $[0.884, 1]$, while a 0/30 result has interval $[0, 0.116]$.

Table 12: Index activation over 30 trials. Parentheses give the conditional median first activation round; a dash means no activation.

Algorithm	Symmetric Pareto	Student-$t_{1.5}$
Practical APR-UCB	30/30 (7667)	30/30 (7667)
APR-LOCAL$_{\text{pract}}$	30/30 (1589)	30/30 (1613)
APR-LOCAL$_{\text{emp}}$	30/30 (1628)	30/30 (1803)
All-block APR-LOCAL	0/30 (−)	0/30 (−)
Certified APR-UCB	0/30 (−)	0/30 (−)
APR-LOCAL (scope)	13/30 (49654)	26/30 (49270)
APR-LOCAL$_{\text{cert}}$	**30/30 (25564)**	**30/30 (24714)**

The confidence budget produces a second finite-horizon difference. Instance accounting activates in all trials, compared with 13/30 and 26/30 for scope normalization. The corresponding exploratory mean regret reductions are 232.46 and 224.02. Instance accounting places K and γ in the finite cache-failure remainder (equation (99)) and activates earlier in table 12.

L.7 PULL LOCALIZATION AND ENDPOINT-CONSTANT EFFECT

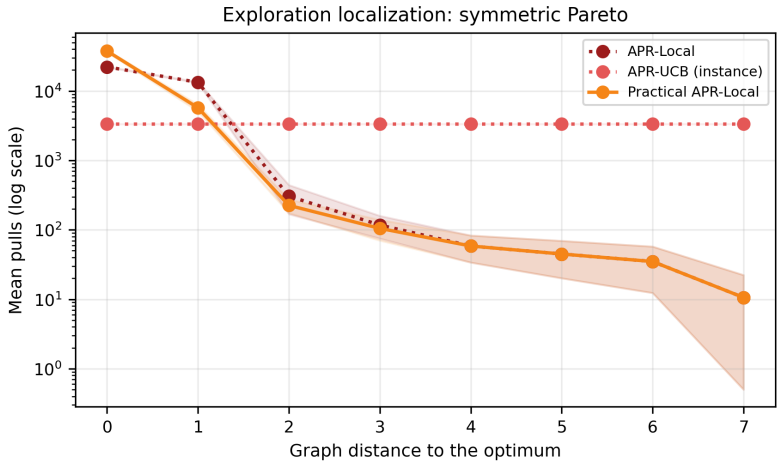

Figure 10: Mean pulls by graph distance on symmetric Pareto, with 95% intervals. The certified APR-UCB remains nearly uniform, while the robust-leader settings concentrate observations near the optimum.

The instance-budget certified APR-UCB assigns approximately one fifteenth of its pulls to each arm, leaving every arm prefix below its sample threshold. The instance-budget graph policy places 97.7% of pulls in the optimal closed neighborhood on Pareto and 98.5% on Student-t. APR-LOCAL$_{\text{pract}}$ places 98.1% and 99.3%, while APR-LOCAL$_{\text{emp}}$ places 95.9% and 89.2%.

The certified and practical settings share the graph-local leader and action rules and differ in endpoint tuple and confidence-budget normalization. The practical setting has lower final regret than APR-LOCAL$_{\text{cert}}$ by 622.91 on Pareto (bootstrap 95% interval $[574.81, 672.61]$) and 695.47 on Student-t ($[625.47, 762.95]$), winning every pair. This difference quantifies the combined finite-horizon effect of these configuration changes. Both configurations have neighbor-local regret guarantees (theorems 4 and 32); section M.3 separates endpoint and budget effects empirically.

L.8 ADAR-UCB AND DIRECT RMM-UCB DIAGNOSTICS

Relative to APR-LOCAL$_{\text{pract}}$, the paired AdaR-UCB regret excess is 1080.38 on Pareto (bootstrap 95% interval $[1020.80, 1139.05]$) and 8381.46 on Student-t ($[7682.17, 8992.54]$). Relative to APR-LOCAL$_{\text{cert}}$, the excesses are 457.46 ($[406.01, 502.74]$) and 7685.99 ($[6990.04, 8287.11]$). All eight

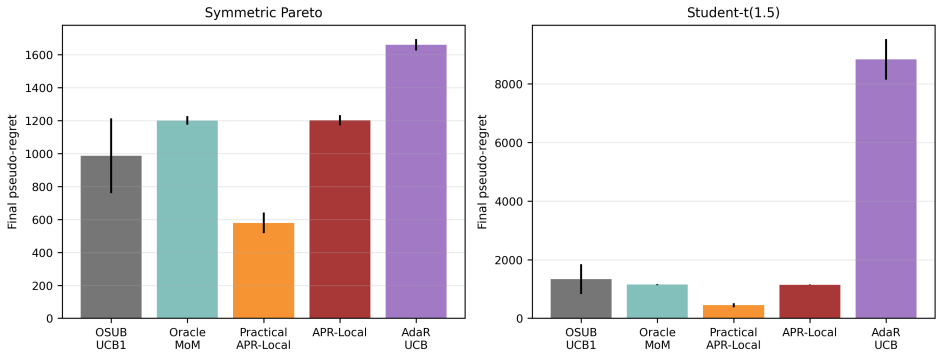

Figure 11: Final pseudo-regret for AdaR-UCB and four graph policies. Bars show means and 95% Student-t intervals over 30 trials. AdaR-UCB uses 50,000 scalar observations, corresponding to 25,000 two-sample decisions.

AdaR contrast intervals exclude zero, and the Holm-adjusted Wilcoxon values are at most 6×10^{-6}. AdaR-UCB's mean optimal-arm pull share is 46.9% on Pareto and 2.3% on Student-t. Comparisons between full-set policies appear in section M.1.

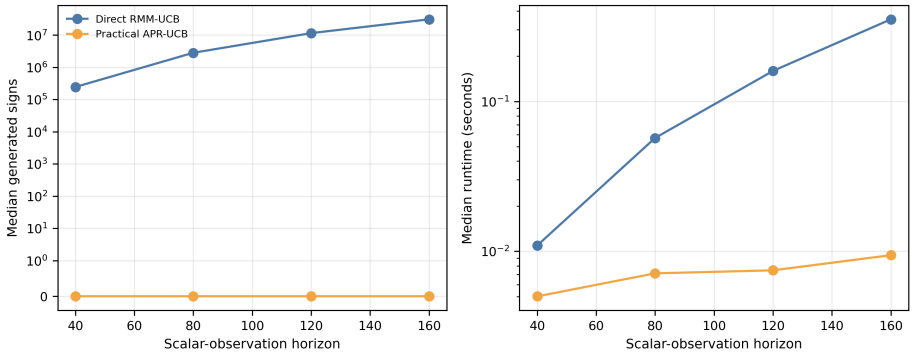

Figure 12: Computational diagnostic for the published direct RMM-UCB rank schedule. Sign generation grows rapidly even at bounded horizons.

Table 13: Bounded-horizon functional check for faithful direct RMM-UCB. Regret is mean $\pm$ 95% Student-t half-width over 10 trials on a two-arm instance. The last two columns are deterministic schedule costs per trajectory and are identical across reward families.

T	Pareto regret	Student-$t_{1.5}$ regret	Signs	Terminal rank
40	1.53 ± 1.06	1.68 ± 1.17	245,382	546
80	4.67 ± 2.63	1.72 ± 1.98	2,830,140	1,538
120	5.57 ± 3.95	4.09 ± 3.72	11,525,561	2,752
160	7.55 ± 5.43	5.95 ± 4.79	30,915,696	4,123

The faithful direct-RMM diagnostic contains 10 paired trials at horizons up to 160. At $T = 120$, one trajectory generates approximately 11.53 million signs; at $T = 160$, approximately 30.92 million. Extrapolating the formula, the terminal rank at $T = 50{,}000$ is 5,853,382, before accounting for repeated index construction; table 13 reports the resulting regret estimates. Figure 15 supplies the moderate-horizon sign-matched control.

L.9 GRAPH-FREE TWO-ARM MEDIUM/LONG-HORIZON CHECK

The two-arm experiment evaluates the certified and practical APR-UCB settings covered by theorem 6 and corollary 33. The two arms have means $(0, 0.04)$, matching the nearest-neighbor gap in the 15-arm path. We use the same symmetric Pareto and Student-$t_{1.5}$ noise families in 30 paired trials with $T = 200{,}000$. Certified APR-UCB uses $(m, r_0, k, g_{\min}) = (4, 1, 21, 3)$ and the instance confidence budget; the practical global variant uses $(8, 1, 9, 1)$ under the same budget. Bands are two-sided 95% Student-t intervals across trials.

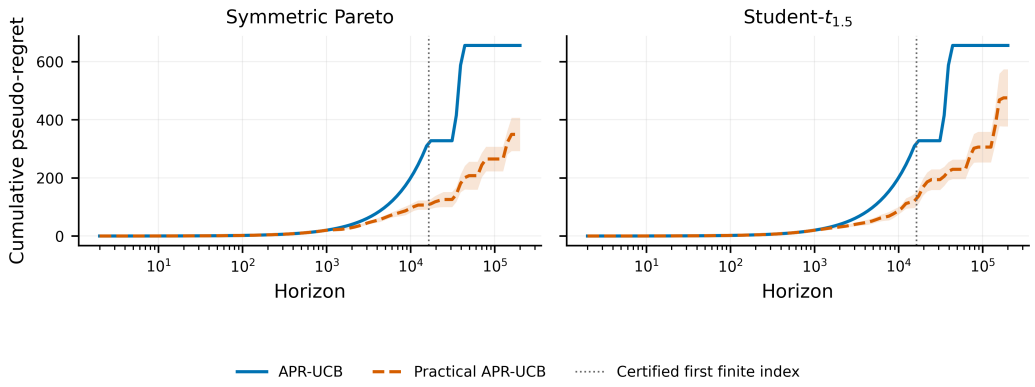

Figure 13: Graph-free two-arm cumulative pseudo-regret to $T = 200{,}000$. Curves are means with 95% Student-t intervals over 30 paired trials; the dotted line marks the certified setting's first finite index.

Table 14: Two-arm graph-free final regret and finite-index activation. Parentheses give the median first finite-index round; lower regret is better.

Method	Pareto regret	Student-$t_{1.5}$ regret	Activation
Certified APR-UCB	655.36 ± 0.00	655.36 ± 0.00	30/30 (16,384)
Practical APR-UCB	349.53 ± 57.33	475.14 ± 98.19	30/30 (1,024)

Before finite indices exist, least-pulled tie breaking alternates the two arms. The certified setting then incurs one further dyadic exploration phase and plateaus by roughly 50,000 rounds; every trajectory ends with exactly 16,384 suboptimal pulls. These equal pull counts produce the zero half-width across stochastic reward streams. Its median sign cost is 1,277,952 under either family. The practical variant activates earlier and has lower mean regret, but wider cross-trial variation; its median sign costs are 2,992,640 on Pareto and 3,082,240 on Student-t.

L.10 GRAPH-LOCAL LONG-HORIZON TRANSITION

We extend the 15-arm path experiment to $T = 200{,}000$ with 30 paired trials under each reward law. The comparison retains OSUB-UCB1, Oracle-MoM-OSUB, and the certified and practical APR-Local settings. Reward streams are shared within each trial, policy randomization uses independent child seeds, and all parameters match the 15-arm benchmark.

Table 15: Graph-local final regret at $T = 200{,}000$. Values are means with 95% half-widths; parentheses give Pareto / Student-$t_{1.5}$ median first-cache rounds.

Method	Pareto	Student-$t_{1.5}$	Activation
OSUB-UCB1	1331.10 ± 528.80	2254.99 ± 937.60	–
Oracle-MoM-OSUB	4600.32 ± 36.89	4473.23 ± 4.18	–
APR-LOCAL$_{\text{cert}}$	1516.86 ± 51.72	1339.79 ± 10.49	30/30 (25,905/24,698)
APR-LOCAL$_{\text{pract}}$	$\mathbf{1000.47 \pm 100.37}$	$\mathbf{1020.33 \pm 187.79}$	30/30 (1,628/1,640)

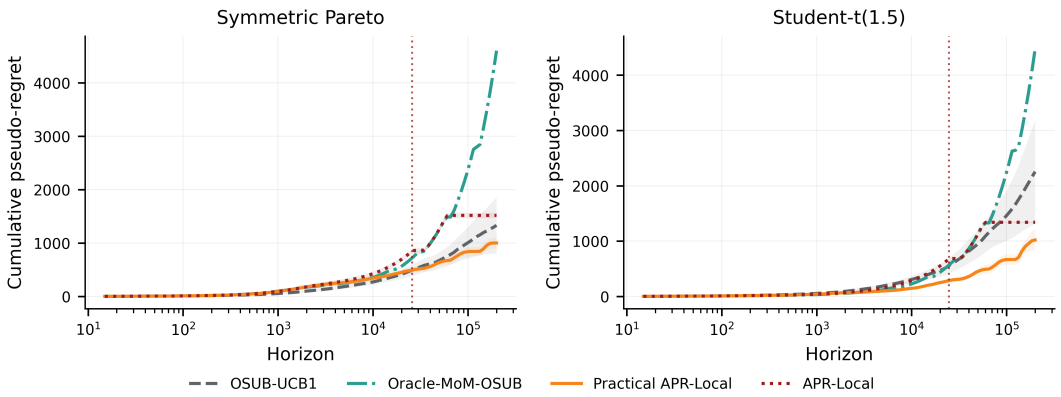

Figure 14: Graph-local cumulative pseudo-regret to $T = 200{,}000$. Curves are means with 95% Student-t intervals over 30 paired trials; the vertical dotted line marks the certified median first-cache round.

The certified transition is visible after its first cache: mean regret rises from 858.46 to 1516.86 on Pareto and from 683.73 to 1339.79 on Student-t between the first post-activation checkpoint and the end of the run. This increment is completed by checkpoint $70{,}861$. Every certified trajectory under both reward laws has identical pseudo-regret at $70{,}861$ and $200{,}000$, establishing a trajectory-level plateau. The practical policy has the lowest mean regret among the four policies in this experiment.

L.11 Sign-matched rank-four RMM diagnostic

The faithful RMM-UCB schedule in table 13 grows its rank with the confidence target. A second diagnostic fixes $m = 4$, uses the same dyadic (arm, sample epoch, global epoch) cache as APR-UCB, and applies the published RMM intersection rule once per prefix. Every new endpoint therefore generates exactly $3n$ signs on an n-observation prefix, matching the certified CA endpoint budget before cache reuse. This diagnostic compares a single rank-four RMM endpoint with the amplified endpoint at the same per-build sign budget. The published RMM-UCB rank schedule is evaluated separately in table 13.

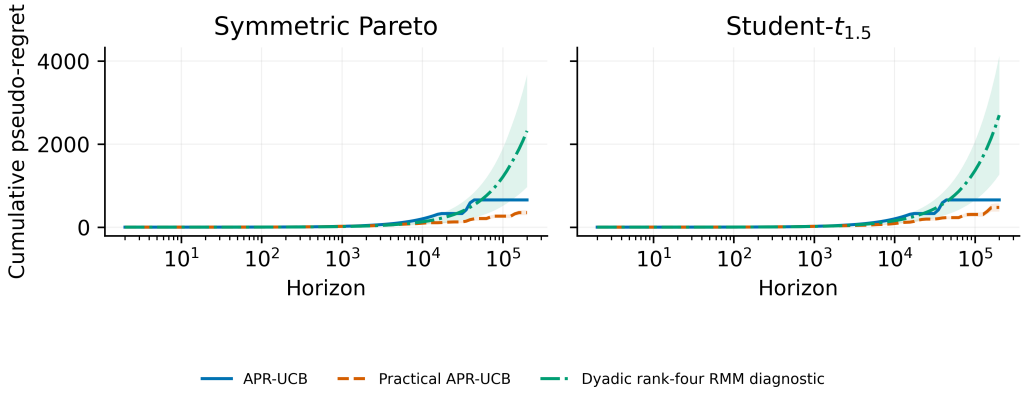

Figure 15: Graph-free two-arm regret at matched per-endpoint random-sign budget. Curves show means and 95% Student-t intervals over 30 paired trials.

The rank-four diagnostic reaches a finite randomized endpoint immediately but exhibits high cross-trial regret variation. Certified amplification forms at round $16{,}384$ and lowers mean regret by 1662.27 on Pareto and 2041.00 on Student-t while using a similar median sign total. With the per-endpoint sign budget matched, this is the effect of the outer median of independent constant-rank bases.

Table 16: Budget-matched two-arm diagnostic at $T = 200{,}000$. Signs and first-finite rounds are conditional medians; parenthesized pairs give Pareto / Student-$t_{1.5}$.

Method	Pareto regret	Student-t regret	Signs	First finite
Certified APR-UCB	655.36 ± 0.00	655.36 ± 0.00	(1,277,952/1,277,952)	(16,384/16,384)
Practical APR-UCB	349.53 ± 57.33	475.14 ± 98.19	(2,992,640/3,082,240)	(1,024/1,024)
Dyadic rank-four RMM	2317.63 ± 1348.72	2696.36 ± 1423.89	(1,179,843/1,180,152)	(3/4)

L.12 Topology, arm-count, gap, and tail sensitivity

The exploratory study uses $T = 30{,}000$ and 12 independent symmetric-Pareto trials per configuration. Each configuration includes OSUB-UCB1, Oracle-MoM-OSUB, APR-Local$_{\text{pract}}$, and APR-Local$_{\text{cert}}$. Topology means decrease linearly with graph distance from a unique optimum, with edge gap 0.04. The five graphs are a path, a star, a balanced binary tree, a 4×4 grid, and a fixed irregular graph. Separate path scans vary $K \in \{7, 15, 31\}$, edge gap in $\{0.02, 0.04, 0.08\}$, and Pareto tail index in $\{1.2, 1.5, 1.8\}$. Each scan reports mean regret with 95% Student-t intervals over the 12 paired trials.

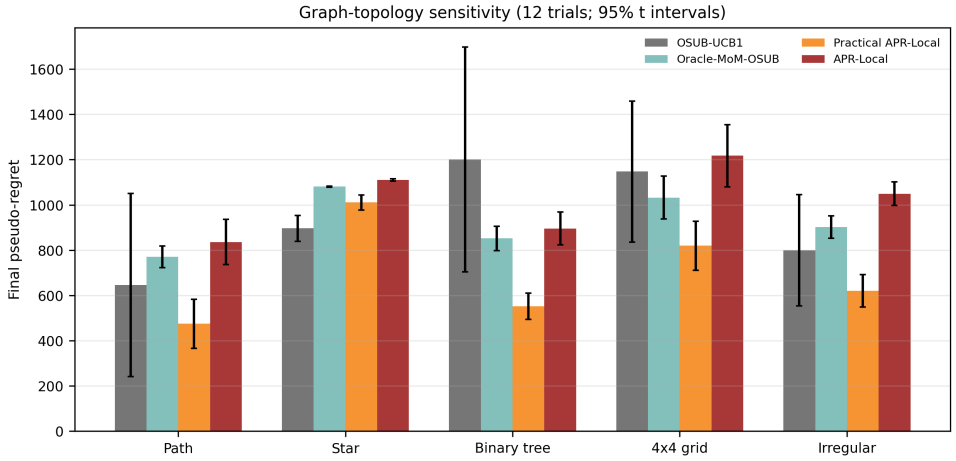

Figure 16: Exploratory topology sensitivity at $T = 30{,}000$ over 12 trials. Certified activation occurs in all path and binary-tree trials but in none of the star, grid, or irregular-graph trials.

Table 17: Topology sensitivity for the two APR-Local configurations. Regret is mean $\pm$ 95% half-width. The final column also gives certified activation and conditional median first round.

Graph	K	γ	Practical regret	Certified regret / activation
Path	15	2	474.30 ± 108.51	835.72 ± 99.55 / 12/12 (25922)
Star	15	14	1010.31 ± 32.80	1110.13 ± 4.99 / 0/12
Binary tree	15	3	551.59 ± 58.05	895.33 ± 72.29 / 12/12 (26985)
4×4 grid	16	4	819.74 ± 108.44	1217.28 ± 137.25 / 0/12
Irregular	15	3	620.05 ± 72.29	1049.07 ± 51.84 / 0/12

The certified setting activates in every path and tree trial and in no star, grid, or irregular-graph trial. The tree and irregular graph share $\gamma = 3$ but have different activation outcomes, showing that maximum degree alone does not determine activation; candidate allocation also matters. The practical APR-Local setting activates in every topology trial. On the star, the optimal closed neighborhood contains all 15 arms, and the practical setting's mean regret is more than twice that on the path.

Increasing the path from 15 to 31 arms changes mean regret from 526.62 to 533.80 for the practical setting and from 848.86 to 875.11 for the certified setting. The overlapping intervals and near-flat estimates contrast with a linear-in-K increase over the tested range. Increasing the edge gap from 0.02 to 0.08 raises certified regret from 510.39 to 1413.67: conservative pre-activation pulls cost

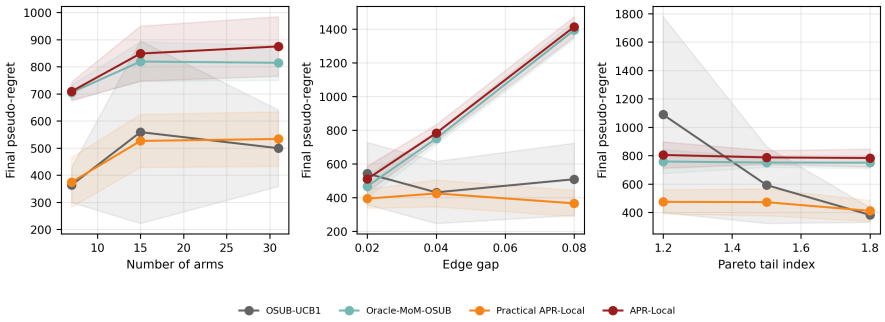

Figure 17: Exploratory arm-count, edge-gap, and Pareto-tail sensitivity over 12 trials per point. Shading gives 95% Student-t intervals.

more before the larger separation can be exploited. As the Pareto tail index rises from 1.2 to 1.8, practical regret falls from 474.57 to 411.19, whereas certified regret changes from 804.70 to 783.21 with overlapping intervals.

L.13 Cache formation and robust-leader resolution

We separate three events. *First cache* is the first round at which one finite CA endpoint is constructed. *Candidate ready* is the first round at which every endpoint in the current leader's closed neighborhood is finite. *Leader resolution* is one plus the final round whose robust leader is suboptimal; trajectories unresolved at the horizon are right-censored. This last quantity measures gap resolution by the parameter-free leader score and is distinct from certified-index formation.

The controlled graphs are trees with balanced branch lengths, $K = 15$, optimum at the center, edge gap 0.04, symmetric-Pareto tail index 1.5, and center degree $d \in \{2, 4, 8\}$. The branch lengths are $(7, 7)$, $(4, 4, 3, 3)$, and $(2, 2, 2, 2, 2, 1, 1)$, respectively; means are -0.04 times distance from the root and the Pareto scale is 0.12. Edge and optimal-neighbor gaps stay fixed; the complete arm-gap multiset varies with topology. Thirty paired trials run to $T = 90{,}000$. Under balanced pulls over the optimal closed neighborhood, candidate prefix epoch r corresponds to round $(d+1)2^r$. Substituting its clock epoch $q = \lfloor \log_2((d+1)2^r) \rfloor$ into the exact block count and solving

$$\left\lfloor \frac{2^r}{L(p_{r,q}, 1/4)} \right\rfloor \geq 63 \tag{258}$$

gives the predictions in table 18.

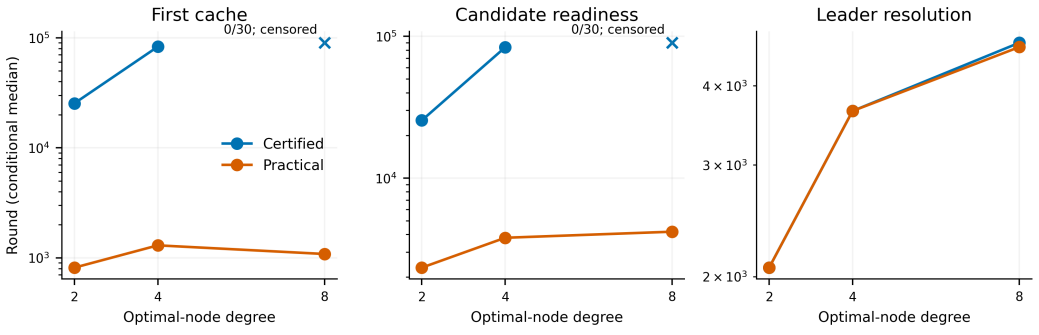

Figure 18: Controlled-degree activation decomposition over 30 trials. Left: first-cache medians and the balanced dyadic prediction. Center: candidate readiness. Right: retrospective robust-leader resolution within the recorded horizon. Bars show interquartile ranges.

Table 18: Certified controlled-degree results over 30 trials. Medians are conditional on observation; "$>$ 90k" records the predicted threshold beyond the experimental horizon.

d	Balanced prediction	First cache	Candidate ready	Leader resolution	Regret
2	24,576	30/30 (25,327.5)	30/30 (25,490)	30/30 (2066.5)	1437.26 ± 40.43
4	81,920	30/30 (82,886.5)	30/30 (83,571)	30/30 (3649.5)	2770.94 ± 28.09
8	$147{,}456 > 90$k	0/30	0/30	30/30 (4676)	3248.18 ± 11.68

The observed cache medians follow the dyadic threshold: degree two uses prefix 8192 with $L = 127$, while degree four reaches a later clock epoch and uses prefix $16{,}384$ with $L = 139$. Degree eight would require the same prefix shared across nine local candidates, producing the $147{,}456$ prediction. Leader resolution occurs one to two orders of magnitude earlier. Thus, in these runs, certified cache formation occurs well after leader resolution and determines the length of the preactivation period.

Table 19: Certified pre-activation allocation. Entries are across-trial medians of per-arm pulls at first cache; degree eight uses the $T = 90{,}000$ censoring snapshot. Neighbor ranges span the individual optimal neighbors, and the final column spans individual arms outside the optimal closed neighborhood.

d	Snapshot	Optimal center	Each optimal neighbor	Each farther arm
2	First cache	8192	8185–8186	1–128
4	First cache	16,384	16,275.5–16,276.5	2–255.5
8	Horizon	10,078.5	9851.5–9852.5	58–191

The center and its optimal neighbors are nearly balanced at the cache snapshot. The remaining pulls record transient suboptimal leaders and their candidate neighborhoods, accounting for the small excess over the balanced prediction at degrees two and four. The artifact contains all 2700 trial–algorithm–arm rows with graph distances and candidate flags.

Table 20: Certified path activation sensitivity at $T = 30{,}000$ over 12 trials. Parentheses give the conditional median first-cache round.

Dimension	Value	Activation	Mean regret
Edge gap	0.02	11/12 (26,691)	510.39
	0.04	12/12 (25,714)	782.93
	0.08	12/12 (24,884)	1413.67
Pareto tail index	1.2	12/12 (25,254)	804.70
	1.5	12/12 (25,648)	787.44
	1.8	12/12 (25,444)	783.21

Cache formation remains concentrated near the same dyadic threshold across the tested gap and tail ranges. Gap affects regret during the pre-activation phase through its per-pull cost, while the certified group-size threshold is distribution free.

L.14 Integrity and deterministic replay

The original 15-arm benchmark contains 660 terminal rows, 59,400 curve rows, and 9,900 pull-count rows. All pass unique-key, nonnegativity, curve-monotonicity, endpoint-equality, and pull-sum checks. The submission-stage additions contain 60 AdaR rows, 240 direct-RMM diagnostic rows, 672 sensitivity rows, and 60 mean-shift audit rows; all pass their corresponding uniqueness and conservation checks. The graph-free two-arm batch including the sign-matched diagnostic contains 16,920 regret-curve rows and 180 terminal rows. The controlled-degree batch contains 180 terminal rows and 2700 arm-level activation snapshots; the constant search adds 2970 best-by-rank rows. These additions pass unique-key, curve-monotonicity, endpoint-equality, pull-sum, dyadic-prefix, and declared- search-box checks. The graph-local long-horizon extension adds 28,800 regret-curve rows and 240 terminal rows; it passes the same unique-key, finite-value, endpoint, and curve-monotonicity checks.

The original eighteen unit tests verify the tolerant constant through exact rational binomial sums, the exact binomial block count, RMM sign accounting, leader block schedule, AdaR formula and two-sample update, direct-RMM rank schedule, controlled graph generators, the sign-matched cache, matched empirical-leader update, and checkpoint logic. A fresh-directory rerun of the first three AdaR trials, first two direct-RMM trials at every horizon, and first sensitivity trial in every configuration matches 6, 48, and 56 terminal rows field by field after excluding wall-clock time. Twenty mean-shift audit rows also match exactly. An additional fresh-directory deterministic replay reproduces 5,400 regret-curve rows, 60 terminal rows, and 900 pull rows. A separate subset replay of the two-arm study reproduces 1,128 curve rows and 12 terminal rows exactly after excluding wall-clock time. Six additional tests check the extended-real rank convention and the uniINF optimizer and updates, bringing the total to 24. The new experiments add 60 uniINF trials, extend the controlled-degree study to 180 terminal rows, and run 720 budget/perturbation configurations, including 180 historical replays. The latter reproduce activation rounds exactly and terminal regret within 3.84×10^{-9}. Another 4800 independent replications evaluate fixed-prefix endpoints. Every new run passes pull-count and regret conservation checks.

M Additional comparisons and diagnostics

M.1 Paired comparisons and run-level distributions

We reanalyse all 30 original 15-arm trials per environment at $T = 50{,}000$. Trial identifiers align across policies, and all runs are included. The additional paired contrasts use 20,000-resample percentile bootstrap intervals, with analysis seed 2026091601; these are pointwise exploratory intervals. Algorithm means retain the 95% Student-t convention of section L.

Table 21 separates graph-free comparisons with AdaR-UCB from graph-local comparisons with OSUB-UCB1. On Pareto rewards, practical APR-UCB has higher mean regret than AdaR-UCB by 116.78; on Student-t rewards, it has lower mean regret by 7224.39. Both sides of that comparison evaluate complete confidence-index implementations under equal scalar-observation budgets.

Table 21: Additional paired contrasts at $T = 50{,}000$. Entries are mean regret differences with 95% paired bootstrap intervals; positive values favor the second algorithm. Each column uses 30 paired trials.

First − second algorithm	Symmetric Pareto	Student-$t_{1.5}$
AdaR-UCB − Certified APR-UCB	-5807.73	1359.82
	$[-5841.10, -5776.04]$	$[667.78, 1973.07]$
AdaR-UCB − Practical APR-UCB	-116.78	7224.39
	$[-159.95, -73.86]$	$[6524.78, 7835.07]$
OSUB-UCB1 − APR-Local$_{\mathrm{cert}}$	-215.32	197.81
	$[-427.90, 17.44]$	$[-224.89, 726.16]$
OSUB-UCB1 − APR-Local$_{\mathrm{pract}}$	407.59	893.27
	$[202.89, 635.95]$	$[472.55, 1415.36]$

OSUB-UCB1 has lower mean regret than APR-Local$_{\mathrm{cert}}$ on Pareto and higher mean regret on Student-t; both paired intervals include zero. OSUB-UCB1 has a lower median and a higher empirical 90th percentile than APR-Local$_{\mathrm{cert}}$ in both environments (table 22). Its Student-t 90th-percentile interval is wide because it is estimated from 30 runs. Figure 19 displays every trial through the empirical cumulative distribution functions. In these instances, $0 \le R_T \le 0.28T$, so fixed-horizon pseudo-regret has finite variance under both reward laws.

Table 22: Terminal pseudo-regret distributions at $T = 50{,}000$ over all 30 runs. The 90th percentiles use linear empirical quantiles; brackets give descriptive 95% percentile-bootstrap intervals with 20,000 resamples.

Environment	Algorithm	Median	90th percentile [95% interval]	Maximum
Pareto	OSUB-UCB1	767.36	2028.77 [1370.56, 2219.32]	2265.80
	APR-LOCAL$_{\text{cert}}$	1186.96	1278.34 [1240.76, 1321.12]	1516.64
	APR-LOCAL$_{\text{pract}}$	587.94	804.68 [710.96, 822.81]	859.96
Student-t	OSUB-UCB1	957.80	2425.84 [1628.24, 5289.11]	6117.04
	APR-LOCAL$_{\text{cert}}$	1126.98	1176.26 [1140.86, 1235.33]	1262.80
	APR-LOCAL$_{\text{pract}}$	463.46	658.51 [537.66, 808.73]	956.40

Paired differences are computed before rounding. For AdaR-UCB and APR-LOCAL$_{\text{pract}}$ on Pareto the unrounded means $1659.194666\ldots$ and $578.817333\ldots$ differ by 1080.38, whereas the displayed two-decimal means differ by 1080.37.

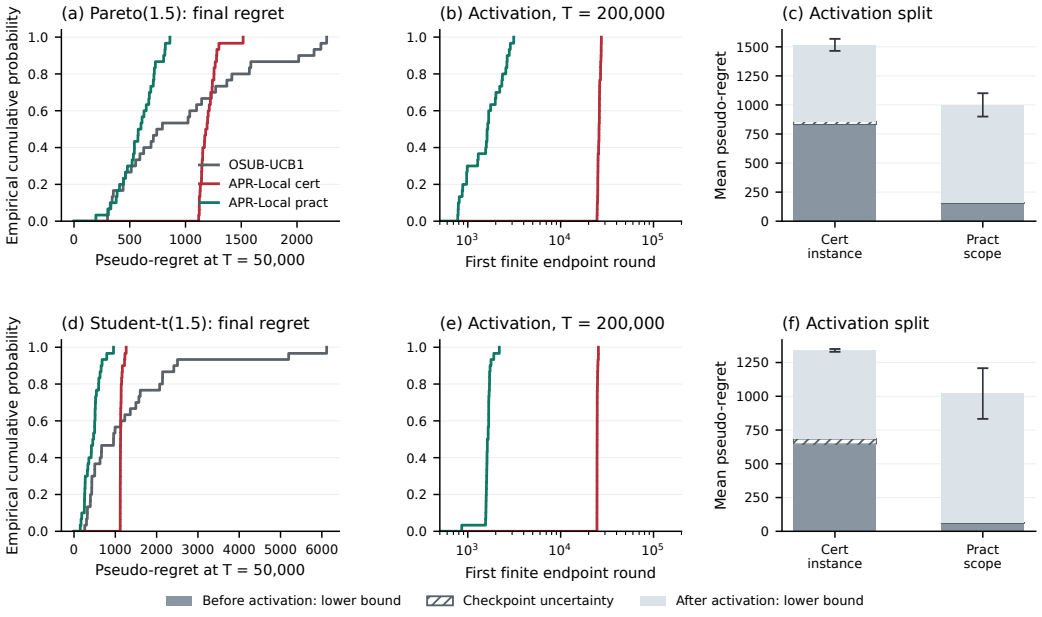

Figure 19: Run-level distributions and activation diagnostics on the 15-arm path. (a,d) Terminal pseudo-regret ECDFs at $T = 50{,}000$. (b,e) First-finite-endpoint times in the separate 30-trial study at $T = 200{,}000$. (c,f) Mean regret split at the first finite endpoint: the hatched segment brackets the split between saved checkpoints, and error bars give 95% Student-t intervals for total mean regret. The certified and practical configurations use instance and scope-normalized budgets, respectively.

The activation panels use the existing $T = 200{,}000$ study. All 30 runs of both policies activate in each environment. The median first-finite-endpoint rounds are $25{,}904.5$ for APR-LOCAL$_{\text{cert}}$ and 1628 for APR-LOCAL$_{\text{pract}}$ on Pareto, and $24{,}697.5$ and 1640 on Student-t. If τ is the first such round, we bracket $R_{\tau-1}$ using the adjacent recorded checkpoints and the per-round gap bound 0.28. The resulting mean preactivation-regret bounds are $[831.79, 858.46]$ and $[152.76, 159.58]$ on Pareto, and $[646.06, 683.73]$ and $[59.34, 61.84]$ on Student-t, for certified and practical settings respectively. These bounds quantify checkpoint resolution, not sampling uncertainty. The matched-budget experiment in section M.3 separates the effects of endpoint constants from budget normalization.

M.2 A KNOWN-HORIZON PARAMETER-FREE BASELINE: UNIINF

We implement Algorithm 1 of Chen et al. (2025) independently, using its published constants and the full action set. uniINF receives T and one observed loss per round. Rewards are converted to losses by $\ell_t = -X_{I_t,N_{I_t}(t)}$, where I_t is the sampled arm; the implementation receives neither the graph nor moment parameters. The mean vector in equation (254) has a unique optimum with mean zero. Its optimal-arm loss is symmetric and integrable, so

$$\mathbb{E}[\ell_{i^\star}\mathbf{1}\{|\ell_{i^\star}| > u\}] = 0 \qquad (u \geq 0).$$

This satisfies the optimal-arm truncated non-negativity condition of uniINF. Both reward laws also have finite absolute 1.4-th moments.

Let $L_{t,i}$ denote cumulative estimated loss for arm i after round t, and let $\mathcal{S}_K$ be the probability simplex on K arms. With $L_{0,i} = 0$ and $S_1 = 4$, the sampling distribution is

$$\pi_t = \operatorname*{arg\,min}_{\pi \in \mathcal{S}_K} \left\{ \sum_{i=1}^{K} L_{t-1,i}\pi_i - S_t \sum_{i=1}^{K} \log \pi_i \right\}, \qquad I_t \sim \pi_t. \tag{259}$$

For the observed loss, define the threshold $C_t = S_t/[4(1 - \pi_{t,I_t})]$ and the clipped loss $c_t = \min\{C_t, \max\{-C_t, \ell_t\}\}$. The updates are

$$L_{t,i} = L_{t-1,i} + \frac{\ell_t \mathbf{1}\{i = I_t\}\mathbf{1}\{|\ell_t| < C_t\}}{\pi_{t,I_t}}, \tag{260}$$

$$S_{t+1}^2 = S_t^2 + \frac{c_t^2(1 - \pi_{t,I_t})^2}{K \log T}. \tag{261}$$

Thus the recursive scale starts at $S_1^2 = 16$, and a loss exactly at the threshold is skipped. There are no forced initialization pulls or tuned constants.

For numerical solution of equation (259), set $d_i = (L_{t-1,i} - \min_j L_{t-1,j})/S_t$. The Karush–Kuhn–Tucker conditions give $\pi_{t,i} = 1/(d_i + z)$, where the unique normalization root satisfies $\sum_i 1/(d_i + z) = 1$ and $z \in [1, K]$. Safeguarded Newton iterations solve this scalar equation with simplex residual tolerance 5×10^{-14}. The implementation was checked against an independent Brent root solver, constrained sequential least-squares optimization, threshold-boundary updates, and a 2000-round trace computed directly from the published equations.

We run uniINF on the original 30 environmental seeds per law, for 60 new trials in total, with $K = 15$ and $T = 50{,}000$ scalar observations. The original reward arrays are shifted by -0.8 to obtain the centered environments, and uniINF receives an independent policy seed. Mean terminal pseudo-regret is 391.74 ± 13.37 on Pareto and 423.49 ± 19.97 on Student-t, where the half-widths are 95% Student-t intervals. Table 23 reports the corresponding paired comparisons, using 20,000 bootstrap resamples and seed 20283834. uniINF has lower sample mean regret than every listed comparator on both instances. Its improvement over APR-LOCAL$_{\text{pract}}$ is 187.08 on Pareto; on Student-t, the difference is 21.79 with an interval spanning zero. These results compare finite-horizon performance under the stated input conditions: uniINF uses the known horizon and all arms, whereas APR-LOCAL uses the graph and is anytime.

Table 23: Comparator minus uniINF terminal pseudo-regret at $T = 50{,}000$. Positive entries favor uniINF. Each environment has 30 paired trials; brackets are pointwise 95% percentile-bootstrap intervals.

Comparator	Symmetric Pareto	Student-$t_{1.5}$
AdaR-UCB	1267.46	8403.25
	[1233.37, 1301.12]	[7714.30, 9003.99]
Practical APR-UCB	1384.24	1178.86
	[1342.76, 1430.36]	[1140.14, 1213.34]
Certified APR-UCB	7075.18	7043.43
	[7062.63, 7087.67]	[7024.23, 7062.06]
APR-LOCAL$_{\text{cert}}$	809.99	717.26
	[783.56, 841.55]	[700.09, 734.04]
APR-LOCAL$_{\text{pract}}$	187.08	21.79
	[127.07, 245.91]	[−37.97, 84.35]

M.3 MATCHING THE CONFIDENCE BUDGET

The original comparison changes both the endpoint tuple and the confidence budget: the certified policy uses $p_{r,q}$ from (8), whereas the practical policy uses $p_{r,q}/[K(\gamma+1)]$. We add practical APR-LOCAL with the instance budget $p_{r,q}$ to separate these changes. The certified and practical tuples are $(4, 1, 21, 3)$ and $(8, 1, 9, 1)$, respectively. Their leader rule, cache mechanism, and forced-pull schedule are unchanged.

We use the 15-arm path in (254), $T = 50{,}000$, and 30 paired trials per reward family. The reward and policy seed trees are replayed from the 15-arm benchmark (master seed 20260803). The two practical configurations share their initial policy seed; their subsequent random number consumption can diverge with their actions. Replayed configurations are counted once. All 180 historical policy–trial outcomes agree with the original results to numerical summation precision, including identical activation rounds. Mean intervals are Student-t intervals; paired contrasts use 10,000 percentile bootstrap resamples. The intervals in this study are pointwise 95% intervals.

At the instance budget, practical APR-LOCAL has mean regret 563.06 on Pareto and 456.84 on Student-t. Relative to the original scope budget, the paired changes are -15.76 $[-57.42, 27.93]$ and 11.56 $[-56.21, 82.93]$. The certified-minus-practical differences at the same instance budget are 638.67 $[593.99, 684.34]$ and 683.91 $[630.24, 736.86]$. The certified-minus-practical regret difference persists after budget matching.

Let τ be the first round at which any candidate endpoint is finite. We record the pull counts immediately before that round and compute $R_{<\tau} = \sum_{t<\tau} \Delta_{I_t}$ and $R_{\geq\tau} = \sum_{t=\tau}^{T} \Delta_{I_t}$ exactly from those counts. All APR-Local runs activate by T. Table 24 and figure 20 show that most of the observed total gap occurs before activation. Activation rounds and post-activation durations differ across configurations.

Table 24: Exact activation decomposition under the unperturbed reward laws, with 30 paired trials per family. The activation round is a median; regret columns are means. Regret components sum to the total before rounding. Every configuration activates in all 30 trials.

Noise	Tuple / budget	Median τ	$R_{<\tau}$	$R_{\geq\tau}$	R_T
Pareto	Certified / instance	25,564	787.01	414.72	1201.73
	Practical / instance	881.5	74.81	488.25	563.06
	Practical / scope	1588.5	100.00	478.81	578.82
Student-t	Certified / instance	24,714	714.19	426.55	1140.75
	Practical / instance	1587.5	54.19	402.65	456.84
	Practical / scope	1612.5	61.49	383.79	445.28

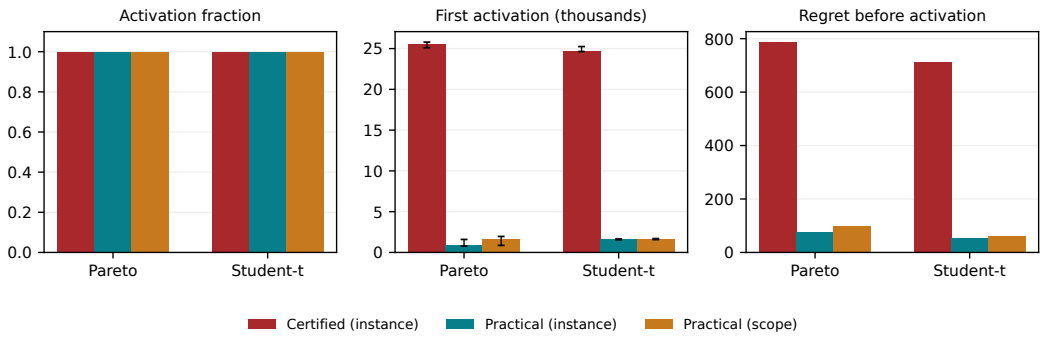

Figure 20: Activation under the unperturbed reward laws in 30 paired trials. Left: fraction producing a finite endpoint by $T = 50{,}000$. Middle: median first-activation round with interquartile ranges. Right: mean regret before activation.

M.4 Sensitivity to asymmetric perturbations

We perturb each original reward X by an independent bounded random variable:

$$X^{(\rho)} = X + sIB, \qquad I \sim \text{Bernoulli}(\rho), \qquad \mathbb{P}(B = 2) = \tfrac{1}{3}, \quad \mathbb{P}(B = -1) = \tfrac{2}{3}, \tag{262}$$

where I and B are independent, and s is the original noise scale (0.12 for Pareto and 0.18 for Student-t). The added noise has mean zero and variance $2\rho s^2$. Arm means, gaps, graph, and tail exponent remain fixed, while asymmetry and central noise scale change together. We use $\rho \in \{0, 0.05, 0.15\}$, the same base reward streams across policies and levels, and nested Bernoulli masks generated from a separate seed 20260916. The 30 trials per family yield 720 policy runs; the $\rho = 0$ rows are the replays described above.

Table 25 and figure 21 report every cell. On Student-t, the original scope-budget practical policy rises from 445.28 to 546.96 at $\rho = 0.15$: the paired increase is 101.68 $[47.84, 157.17]$. Its instance-budget counterpart rises from 456.84 to 490.86, a change of 34.03 $[-12.82, 80.38]$. Certified regret changes by 0.32 $[-11.92, 12.27]$. On Pareto, the scope-budget practical policy is nearly unchanged (578.82 versus 579.09), while certified regret decreases by 33.36, with interval $[-56.29, -14.66]$ for the perturbed-minus-original contrast. The response to this perturbation therefore depends on both the noise family and the budget.

The logged number of suboptimal-leader rounds provides a second view of the Student-t change. At $\rho = 0$ and 0.15, its mean is 820.67 and 1102.57 for practical / scope, 889.43 and 1079.50 for practical / instance, and 1499.80 and 1260.10 for certified / instance. These counts exclude the 15 initialization pulls.

Table 25: Mean pseudo-regret $\pm$ 95% Student-t half-width, with $T = 50{,}000$ and 30 paired trials in every cell. Instance and scope refer to $p_{r,q}$ and $p_{r,q}/[K(\gamma + 1)]$, respectively.

Noise	ρ	Certified instance	Practical instance	Practical scope	OSUB-UCB1
Pareto	0	1201.73 ± 30.20	563.06 ± 57.33	578.82 ± 62.08	986.41 ± 227.78
	.05	1179.31 ± 26.16	529.65 ± 62.16	556.83 ± 45.44	979.29 ± 226.51
	.15	1168.37 ± 16.72	540.55 ± 49.70	579.09 ± 49.65	974.33 ± 223.78
Student-t	0	1140.75 ± 13.54	456.84 ± 57.12	445.28 ± 71.68	1338.55 ± 509.81
	.05	1137.81 ± 16.21	431.53 ± 57.63	485.64 ± 54.11	1323.93 ± 513.22
	.15	1141.07 ± 16.67	490.86 ± 54.48	546.96 ± 45.02	1327.29 ± 513.05

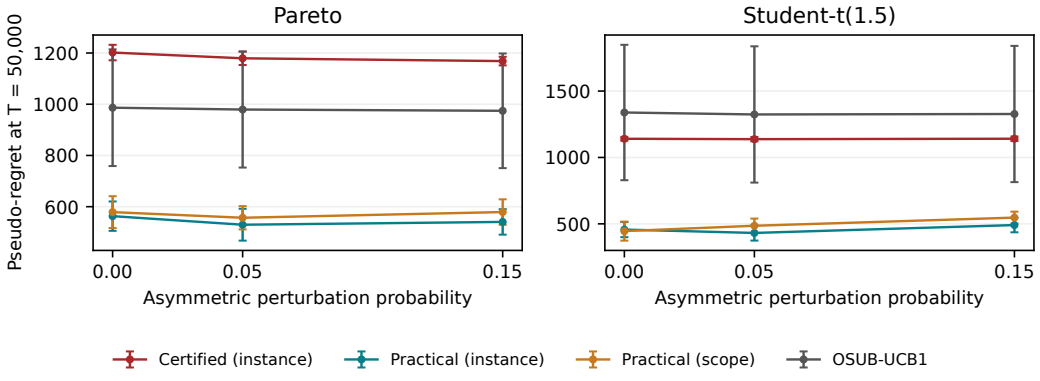

Figure 21: Regret under (262), with a 15-arm path, $T = 50{,}000$, and 30 paired trials. Error bars are 95% Student-t intervals for the mean.

Fixed-prefix endpoints. For each family, ρ, and endpoint tuple, we generate 200 independent reward/sign replications at $(n, p) = (2048, 0.05)$ and $(4096, 0.001)$. Table 26 reports all 24 cells. Every endpoint is finite. At $p = 0.05$, undercoverage ranges from $6/200$ to $13/200$, and every

exact Clopper–Pearson interval contains 0.05. At $p = 0.001$, eleven cells have zero failures and the symmetric Pareto practical cell has one. The corresponding upper confidence limits are 0.0183 and 0.0275, so 200 replications do not resolve nominal undercoverage of 0.001.

We also evaluate the upper-deviation threshold

$$x_g = 4(12M_\rho)^{1/(1+a)} g^{-a/(1+a)}, \quad g = \left\lfloor \frac{\lfloor n/L \rfloor}{k} \right\rfloor, \quad a = 0.4, \tag{263}$$

where $L = L(p, 1/m)$ and a valid moment bound follows from Minkowski:

$$M_\rho^{1/(1+a)} = M_0^{1/(1+a)} + s\big[\rho\big(\tfrac{1}{3}2^{1+a} + \tfrac{2}{3}\big)\big]^{1/(1+a)}. \tag{264}$$

Here M_0 is the original moment bound for the chosen family. No replication exceeds x_g, but these thresholds (5.7071–15.7490) are much larger than the observed endpoint 95th percentiles (0.0525–0.0899). Under exact symmetry, both tuples have formal upper-tail certificates (corollary 13 and lemma 27); the practical bound requires group size at least six. The perturbed-law results measure finite-sample sensitivity to departures from symmetry.

Table 26: Fixed-prefix undercoverage and endpoint scale. C and P denote the certified and practical tuples. Each row uses 200 independent replications; intervals are exact two-sided 95% Clopper–Pearson intervals for the undercoverage probability. $Q_{.95}$ is the empirical endpoint 95th percentile about the true center zero. All finite-endpoint rates are 1.

Noise	ρ	Tuple	(n, p)	Failures	95% interval	$Q_{.95}$
Pareto	0	C	$(2048, .05)$	13/200	$[.0351, .1086]$	.0557
	0	P	$(2048, .05)$	13/200	$[.0351, .1086]$	.0736
	.05	C	$(2048, .05)$	10/200	$[.0242, .0900]$	.0560
	.05	P	$(2048, .05)$	13/200	$[.0351, .1086]$	.0749
	.15	C	$(2048, .05)$	9/200	$[.0208, .0837]$	.0554
	.15	P	$(2048, .05)$	13/200	$[.0351, .1086]$	.0764
	0	C	$(4096, .001)$	0/200	$[0, .0183]$	.0529
	0	P	$(4096, .001)$	1/200	$[.0001, .0275]$	.0743
	.05	C	$(4096, .001)$	0/200	$[0, .0183]$	.0525
	.05	P	$(4096, .001)$	0/200	$[0, .0183]$	.0722
	.15	C	$(4096, .001)$	0/200	$[0, .0183]$	.0533
	.15	P	$(4096, .001)$	0/200	$[0, .0183]$	.0717
Student-t	0	C	$(2048, .05)$	6/200	$[.0111, .0642]$	.0609
	0	P	$(2048, .05)$	7/200	$[.0142, .0708]$	.0899
	.05	C	$(2048, .05)$	11/200	$[.0278, .0963]$	.0582
	.05	P	$(2048, .05)$	8/200	$[.0174, .0773]$	.0895
	.15	C	$(2048, .05)$	10/200	$[.0242, .0900]$	.0582
	.15	P	$(2048, .05)$	10/200	$[.0242, .0900]$	.0887
	0	C	$(4096, .001)$	0/200	$[0, .0183]$	.0544
	0	P	$(4096, .001)$	0/200	$[0, .0183]$	.0827
	.05	C	$(4096, .001)$	0/200	$[0, .0183]$	.0548
	.05	P	$(4096, .001)$	0/200	$[0, .0183]$	.0818
	.15	C	$(4096, .001)$	0/200	$[0, .0183]$	.0555
	.15	P	$(4096, .001)$	0/200	$[0, .0183]$	.0851