LEARNING FROM ILL-BEHAVED REWARDS VIA POPULATION-RELATIVE FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

Heavy-tailed real-valued feedback can make numerical reward statistics unstable or even undefined. Rather than robustly estimating such statistics, we change the representation of the feedback itself. Each reward is mapped to its rank under the uniform mixture of arm distributions, producing a bounded population-relative score equal to the probability of beating an independently sampled, uniformly chosen opponent. Thus a single scalar observation yields implicit comparison feedback without pairwise queries, reward moments, tail parameters, or truncation thresholds. The reference distribution is unknown and must itself be learned online from adaptively sampled arms. We develop RANK-UCB1, which estimates it using uniformly weighted per-arm empirical CDFs, compresses them with Greenwald–Khanna summaries, and controls the accumulated error of historical scores computed under earlier reference estimates. The resulting algorithm achieves $O(T^{2/3})$ expected rank-regret on every fixed finite-arm instance with positive gaps, together with a uniform sublinear finite-horizon guarantee. An exact variant without continued forced exploration has almost-sure logarithmic rank-regret and sublinear expected regret. Under first-order stochastic dominance, population-relative scores preserve parameter, quantile, and finite-mean orderings. Experiments across heavy-tailed and infinite-mean settings show effective allocation and negligible loss from streaming compression.

1 INTRODUCTION

Stochastic multi-armed bandits (MABs) typically learn from numerical reward magnitude. Classical algorithms such as UCB estimate arm means under bounded or sufficiently well-behaved rewards (Auer et al., 2002). Heavy-tailed feedback makes this approach difficult: rare extremes can dominate estimates, and the population mean may be infinite or undefined. Examples include Pareto-like delays in distributed systems (Wang et al., 2015) and bursty network traffic (Crovella & Bestavros, 1997). LLM tool selection, retrieval, and model routing are important potential applications, where long-tailed usage and heterogeneous costs or utilities motivate methods that tolerate ill-behaved real-valued feedback (Kandpal et al., 2023; Su et al., 2025; Poon et al., 2026).

Consider a Pareto reward $X \geq 1$ with $\Pr(X > x) = x^{-a}$. Its mean is $a/(a-1)$ for $a > 1$ and infinite for $a \leq 1$. Even at $a = 1.1$, the mean is 11 while the median is below 2, so most of the expectation is carried by rare observations. Robust estimators can mitigate this instability when suitable moments exist, but they retain the numerical reward scale and typically require moment assumptions, tail information, tuning, or a particular target such as a quantile. We ask a different question: *must a learner preserve reward magnitude at all?*

Our answer is no. We change the representation before learning by mapping each reward to its rank under the uniform mixture of arm distributions, yielding bounded *population-relative feedback*. The expected score is the probability of beating an independent reward from a uniformly chosen arm. Every observation contributes comparison information without requiring reward moments, tail parameters, or a truncation threshold. Although the objective is Borda-like, the learner receives a potentially unbounded scalar reward, rather than a pairwise comparison, and constructs its relative position against a reference learned online.

Learning that reference is the central challenge. It is shared across arms but estimated from adaptively and unevenly sampled observations. Moreover, scores are retained when generated: an arm's

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.

average contains rewards evaluated under different, earlier reference estimates. Accuracy of the current reference alone therefore does not control the scores used for decisions. Our analysis tracks this accumulated *historical plug-in error*, connecting reference learning to the behavior of the implemented algorithm.

We develop RANK-UCB1, which combines uniformly weighted per-arm CDF estimates, deterministic forced exploration, and Greenwald–Khanna compression (Greenwald & Khanna, 2001). Forced exploration refines the shared reference, while compression enables streaming updates without storing the full reward history. With horizon-dependent precision, the algorithm achieves $O(T^{2/3})$ expected rank-regret on every fixed finite-arm instance with positive suboptimality gaps; a complementary gap-free analysis gives a uniform sublinear finite-horizon guarantee. For the exact variant without continued forced exploration, we also establish almost-sure logarithmic rank-regret and sublinear expected regret (Section 6).

Under first-order stochastic dominance, rank scores preserve the direction of parameter, quantile, and finite-mean orderings, so useful arm orderings can be learned without estimating reward magnitude. Six benchmark settings and an infinite-mean extension examine both this calibration and the distinct preferences induced by crossing distributions. On the Pareto and exponential benchmarks, where distinct objectives share the same optimal arm, RANK-UCB1 achieves higher best-arm pull fractions than all four extreme-bandit baselines. It also allocates 95.95% of pulls to the rank-optimal arm in the infinite-mean setting. Ablations show that forced exploration reduces allocation variability and GK compression closely preserves exact eCDF performance with a much smaller reference representation. These results support population-relative feedback as a preprocessing layer for learning from ill-behaved real-valued feedback.

2 RELATED WORK

Heavy-tailed bandits are often handled by robustifying the mean. Classical approaches replace the sample mean with truncated means, median-of-means, or Catoni-type estimators, obtaining regret bounds under weak moment assumptions but typically depending on moment bounds or tail parameters (Bubeck et al., 2013). More recent work studies adaptation to unknown distributional parameters through adaptive trimmed-mean estimators and related data-driven procedures with $1 + \epsilon$ reward-moment assumptions (Chen et al., 2025; Genalti & Metelli, 2026). A different line optimizes alternative targets such as quantiles or qualitative/ordinal criteria (Szörényi et al., 2015); these formulations are robust, but they deliberately focus on selected parts of the distribution.

Closely related are extreme or max K-armed bandits, which optimize tail-dominance or largest-observed-reward criteria rather than the mean. Existing methods include truncation-based UCB variants such as ThresholdAscent (Streeter & Smith, 2006), quantile- or median-based algorithms such as MaxMedian and QoMax (Bhatt et al., 2023; Baudry et al., 2022), and parametric or semi-parametric approaches such as ExtremeHunter that estimate tail indices or distributional parameters (Carpentier & Valko, 2014). Our work is methodologically adjacent to these lines, but differs in objective: we keep a bounded stochastic-bandit style learning problem without explicit truncation and without committing to a single quantile or extreme-value criterion.

The win-probability objective is related to the Borda criterion in dueling and preference bandits (Xu et al., 2019; Bengs et al., 2021), and to learning from ranking feedback (Maran et al., 2024). These models receive comparisons or rankings as feedback. Our setting instead allows arbitrary real-valued, potentially heavy-tailed or infinite-mean rewards, and estimates the comparison population online through per-arm empirical CDFs. Our analysis controls this coupling under adaptive sampling and streaming compression, yielding guarantees for the implemented learner. Categorized bandits exploit a known partition of arms with ordered categories under sub-Gaussian mean feedback (Jedor et al., 2019); our ordering result instead calibrates a learned rank objective without requiring such a partition or reward moments.

3 SETUP AND OBJECTIVE

We consider a stochastic K-armed bandit with arms $i \in [K] \triangleq \{1, \dots, K\}$. Arm i has an i.i.d. reward stream $R_{i,1}, R_{i,2}, \dots$ with unknown CDF F_i; streams are independent across arms. Rewards are finite

real values, with no boundedness or moment assumptions. In particular, their means may be infinite or undefined. At round t, a policy chooses I_t from past observations and receives $r_t = R_{I_t, N_{I_t}(t)}$, where $N_i(t)$ counts pulls through round t. We write $n_i(t) = N_i(t-1)$ for the count available at selection time.

Oracle mixture reference and rank surrogates. Define the uniform mixture reference and the oracle surrogate by

$$G^\star(r) = \frac{1}{K}\sum_{j=1}^{K} F_j(r), \qquad X_{i,n} = G^\star(R_{i,n}) \in [0,1]. \tag{1}$$

Boundedness gives Hoeffding concentration for the oracle scores regardless of the raw reward tails (Hoeffding, 1963). Each observation contributes its relative position in the reference distribution, with no truncation threshold or selected quantile.

Win-probability interpretation. The score of arm i is

$$\mu_i = \mathbb{E}[G^\star(R_{i,1})] = \frac{1}{K}\sum_{j=1}^{K} \mathbb{P}(R_i \geq R'_j) = \mathbb{P}(R_i \geq R'_J), \tag{2}$$

where $R'_j \sim F_j$ and $J \sim \mathrm{Unif}([K])$ are independent of R_i. This follows by conditioning on R_i and using $F_j(r) = \mathbb{P}(R'_j \leq r)$. We use right-continuous CDFs throughout, so a tie receives full credit and self-comparisons use an independent draw. For continuous rewards, ties vanish and the self-comparison contributes $1/2$. The identity also holds for distributions with atoms under the stated convention. It connects our objective to Borda-type scores while requiring only scalar reward feedback (Bengs et al., 2021).

Regret under rank objectives. Let $\mu_\star = \max_i \mu_i$, choose $i^\star \in \arg\max_i \mu_i$, and define $\Delta_i = \mu_\star - \mu_i$. The random rank pseudo-regret and its expectation are

$$R_T = \sum_{t=1}^{T} \Delta_{I_t} = \sum_i \Delta_i N_i(T), \qquad \mathcal{R}^{\mathrm{rank}}(T) = \mathbb{E}R_T = \sum_i \Delta_i \mathbb{E}N_i(T). \tag{3}$$

These quantities are well-defined for every reward distribution considered here, with $0 \leq R_T \leq T$. Multiple optimal arms are allowed. When suboptimal arms exist, write $\Delta_{\min} = \min_{i:\Delta_i>0} \Delta_i > 0$.

A scale-independent comparison. The objective depends on order rather than numerical units. For any common strictly increasing transformation h, the transformed CDFs satisfy $F_j^h(h(r)) = F_j(r)$, so every oracle rank and every μ_i is unchanged. The same identity holds for exact empirical ranks. Consequently, changing units or applying a common logarithmic transformation does not change which arm is preferred. This invariance makes the reference distribution an interpretable comparison population: the question is how often an outcome exceeds an independently drawn alternative, rather than how large its value is in a particular coordinate system.

4 Rank surrogates in practice

4.1 From the oracle reference to an online plug-in

The oracle reference depends on unknown CDFs. For each arm j, let $\widehat{F}_{j,n}$ denote the exact eCDF of its first n rewards and let $\widetilde{F}_{j,n}$ denote the CDF query supplied by its data structure. An empty structure returns zero. At round t, we first insert r_t into the selected arm's structure and then form

$$\widetilde{G}_t(r) = \frac{1}{K}\sum_{j=1}^{K} \widetilde{F}_{j,N_j(t)}(r), \qquad Z_t = \widetilde{G}_t(r_t) \in [0,1]. \tag{4}$$

Thus all CDF counts in the score are *post-update*. The score Z_t is retained as computed; subsequent reference updates do not rescore earlier rewards. Uniform weighting keeps the reference aligned with $G^\star$ even when arms have very different sample counts.

Pooling all observed rewards into a single eCDF would instead weight arm j by $N_j(t)/t$. That reference changes its population weights as the policy changes its allocations. Averaging the per-arm eCDFs preserves the intended uniform opponent distribution, separating the definition of the objective from the acquisition of evidence about it. This distinction is especially useful late in learning, when one arm receives most of the budget but the other arms still define meaningful alternatives. A scalar observation can be queried against every maintained CDF using the same reward draw; the construction therefore obtains reference-relative feedback without requesting an additional comparison or reward observation.

4.2 Maintaining empirical CDFs: exact eCDF vs. streaming sketches

The exact eCDF is $\widehat{F}_{j,n}(r) = n^{-1} \sum_{m=1}^{n} \mathbf{1}\{R_{j,m} \le r\}$. It admits uniform concentration through the DKW–Massart inequality (Massart, 1990), but requires storing all rewards. We use a GK summary per arm (Greenwald & Khanna, 2001) to compress these observations while controlling rank error. Specifically, the CDF interface must return values in $[0, 1]$ and satisfy, for every arm, sample prefix, and query,

$$\|\widetilde{F}_{j,n} - \widehat{F}_{j,n}\|_\infty \le \eta + \frac{\kappa}{n}, \qquad n \ge 1. \tag{5}$$

Here η is the chosen precision and κ is a fixed rank-rounding constant. Exact eCDFs have $\eta = \kappa = 0$. Appendix C specifies how rank intervals produce this CDF contract, including duplicate values and buffered observations.

4.3 Stability under exploration imbalance

A rarely sampled arm still contributes $1/K$ to the reference for every other arm. FE therefore serves a second role beyond searching for a good arm: it learns the common frame in which all arms are compared. Bounded oracle scores control sampling fluctuations, while uniform eCDF bounds control the reference error. Section 6 combines these two sources of uncertainty for the historical scores actually used by the algorithm.

Retaining scores makes the update inexpensive: each arm needs only its cumulative surrogate sum and count in addition to its CDF representation. It also defines the statistical challenge. The current reference may be accurate even while an arm's average still contains scores generated under less accurate earlier references. Our historical-error argument follows those actual scoring times, so the same analysis applies to the streaming implementation without revisiting stored observations.

5 Algorithm: Rank-UCB1

For each arm, Rank-UCB1 maintains a pull count n_i, a cumulative score S_i, and a CDF structure $\mathcal{S}_i$. The main algorithm uses GK summaries and cumulative round-robin FE. For $0 < \alpha < 1$, the target number of forced pulls is $M(t) = \max\{K, \lceil t^\alpha \rceil\}$. All other rounds maximize

$$U_i(t) = \widehat{\mu}_i(t) + \sqrt{\frac{\rho \log t}{n_i(t)}}, \qquad \widehat{\mu}_i(t) = \frac{S_i(t)}{n_i(t)}. \tag{6}$$

The first K rounds initialize all arms, so every denominator in this index is positive. We use $\alpha = 2/3$ and $\rho = 1/6$; the FE guarantees below hold for every fixed $\rho \ge 0$. For a prescribed horizon $T \ge K$, the default GK precision is $\eta = T^{-1/3}$.

The exact-eCDF version changes only the CDF structures. The no-FE ablation forces the first K pulls and uses (6) thereafter. Index ties may be resolved by any rule based on past information. Each round uses exactly one raw reward. The score calculation, index, and proofs all use the same update order.

6 Analysis of Rank-UCB1

We first separate what bounded oracle scores provide from what must be learned about the reference. Complete proofs are in Appendix A.

Algorithm 1 RANK-UCB1 (GK FE; post-update scoring)

Require: K, horizon $T \geq K$, $\alpha = 2/3$, $\rho = 1/6$, GK precision $\eta = T^{-1/3}$.

For each i, set $n_i \leftarrow 0$, $S_i \leftarrow 0$, and initialize an empty GK structure $\mathcal{S}_i$ satisfying (5).
Set cumulative forced-pull count $f \leftarrow 0$.
for $t = 1, \ldots, T$ **do**
  if $f < \max\{K, \lceil t^{\alpha} \rceil\}$ **then**
    $I_t \leftarrow 1 + (f \bmod K)$; $f \leftarrow f + 1$.
  else
    $I_t \in \arg\max_i \{S_i/n_i + \sqrt{\rho \log t / n_i}\}$.
  end if
  Pull I_t once and observe r_t.
  Insert r_t into $\mathcal{S}_{I_t}$; set $n_{I_t} \leftarrow n_{I_t} + 1$.
  $Z_t \leftarrow K^{-1} \sum_{j=1}^{K} \mathrm{CDF}(\mathcal{S}_j, r_t)$, using the updated structures.
  $S_{I_t} \leftarrow S_{I_t} + Z_t$.
end for

6.1 ORACLE OBSERVATION

If $G^\star$ is known, applying classical UCB1 to $X_{i,n}$ with $\rho = 2$, one initial pull per arm, and no continued FE gives (Auer et al., 2002)

$$\mathbb{E}R_T \leq \sum_{i:\Delta_i>0} \left(\frac{8 \log T}{\Delta_i} + \left(1 + \frac{\pi^2}{3}\right) \Delta_i \right). \tag{7}$$

No reward moments enter this observation: the concentration applies to ranks. The practical problem is to learn $G^\star$ while retaining scores formed from earlier estimates.

6.2 LEARNING WITHOUT CONTINUED FORCED EXPLORATION

Theorem 6.1 (Exact-eCDF without continued FE)**.** *For the exact-eCDF variant initialized once per arm, with any fixed $\rho > 0$, every fixed finite-arm instance satisfies*

$$\frac{N_i(T)}{\log T} \longrightarrow \frac{\rho}{\Delta_i^2} \quad (\Delta_i > 0), \qquad \frac{R_T}{\log T} \longrightarrow \rho \sum_{i:\Delta_i>0} \frac{1}{\Delta_i} \quad \textit{almost surely}, \tag{8}$$

and $\mathbb{E}R_T = o(T)$, without reward moment assumptions.

The index eventually samples every arm often enough for the reference to become consistent. Cesàro averaging then makes each historical score mean consistent as well. The pathwise logarithmic limit does not by itself imply logarithmic *expected* regret. Appendix A.8 gives an explicit expected sublinear rate and a stronger, instance-dependent sufficient condition for $O(\log T)$ expectation. Whether that condition holds for a particular Pareto or lognormal instance requires a separate argument. FE gives the main algorithm an expected polynomial-rate guarantee without this additional condition.

6.3 HISTORICAL ERROR AND THE MAIN FIXED-INSTANCE GUARANTEE

Write $\tau_{i,m}$ for the global time of arm i's mth pull. The reference used in $Z_{\tau_{i,m}}$ depends on that time, so concentration must control the average of these historical errors.

Lemma 6.2 (Historical surrogate error)**.** *Fix $T \geq K$ and $\delta \in (0, 1)$. Under Algorithm 1, set*

$$L = \log(4KT/\delta), \quad \beta = \alpha/2, \quad C = \sqrt{KL} + \kappa\sqrt{2K}, \quad A = C/(1-\beta).$$

With probability at least $1 - \delta$, simultaneously for all rounds $s \leq T$ and all observed arm prefixes of length $n \geq 1$,

$$|Z_s - X_{I_s, N_{I_s}(s)}| \leq C s^{-\beta} + \eta, \tag{9}$$

$$\left| \frac{1}{n} \sum_{m=1}^{n} Z_{\tau_{i,m}} - \mu_i \right| \leq \sqrt{\frac{L}{2n}} + A n^{-\beta} + \eta. \tag{10}$$

The proof first controls all deterministic sample prefixes using DKW–Massart and Hoeffding, then substitutes the adaptive counts. Uniformity over queries covers post-update scoring even though the current reward is included in its own eCDF. The factor $n^{-\beta}$ in (10) accounts for the retained earlier scores.

Let $F_{\mathrm{FE}}(T)$ be the number of forced pulls, including initialization, and let $f_i(T)$ be the forced count for arm i. Define

$$R_T^{\mathrm{FE}} = \sum_i \Delta_i f_i(T), \qquad \overline{\Delta} = \frac{1}{K}\sum_i \Delta_i.$$

Theorem 6.3 (Fixed-instance expected rank-regret)**.** *For exact eCDFs, any fixed $0 < \alpha < 1$, any fixed $\rho \geq 0$, and every fixed finite-arm instance with suboptimal arms, there is a finite constant $B = B(K, \Delta_{\min}, \alpha, \rho)$, independent of T, such that*

$$0 \leq \mathbb{E}R_T - R_T^{\mathrm{FE}} \leq B. \tag{11}$$

Moreover,

$$\begin{aligned} F_{\mathrm{FE}}(T) &= \min\{T, \max\{K, \lceil T^{\alpha} \rceil\}\}, \\ \mathbb{E}R_T &= \overline{\Delta} F_{\mathrm{FE}}(T) + O(1). \end{aligned} \tag{12}$$

Thus $\alpha = 2/3$ gives $\mathbb{E}R_T = O(T^{2/3})$ on each fixed instance. If every arm is optimal, $R_T = 0$.

Outside the prescribed FE rounds, expected cumulative regret is bounded independently of the horizon. The historical-error bound makes nonforced selections of a positive-gap arm eventually require a concentration failure; summable failure probabilities leave reference exploration as the leading cost. The constant depends on the instance gaps. The following result gives a complementary guarantee uniform over instances.

Corollary 6.4 (Uniform finite-horizon expected rank-regret)**.** *For Algorithm 1 with $\alpha = 2/3$, any fixed $\rho \geq 0$, and a CDF interface satisfying* (5) *with fixed κ, every $T \geq \max\{K, 2\}$ satisfies*

$$\mathbb{E}R_T = \widetilde{O}\Big(K + T^{2/3} + K^{5/6}T^{7/9} + \sqrt{\rho K T} + \eta T\Big). \tag{13}$$

The bound is capped at T; hidden constants depend only on κ and hidden factors are logarithmic in K, T, with no dependence on suboptimality gaps or reward moments. For fixed K, ρ, both exact eCDFs and GK with $\eta = T^{-1/3}$ yield $\mathbb{E}R_T = \widetilde{O}(T^{7/9})$.

This guarantee remains uniform when arm gaps become arbitrarily small, whereas Theorem 6.3 gives the sharper $T^{2/3}$ rate on each fixed positive-gap instance. The $T^{7/9}$ term accounts for errors retained from earlier reference estimates; sketch error contributes ηT, which is lower order at the default precision. Thus the compressed algorithm learns sublinearly without a prescribed gap lower bound. Appendix A.4 proves the explicit high-probability bound and derives this expectation result.

6.4 GK COMPRESSION PRESERVES THE FIXED-INSTANCE RATE

Corollary 6.5 (GK FE)**.** *Suppose the GK CDF interface satisfies* (5) *with fixed κ. At fixed precision η,* (11) *holds whenever $2\eta < \Delta_{\min}$, with B also depending on κ and $\Delta_{\min} - 2\eta$. With $\alpha = 2/3$ and $\eta_T = T^{-1/3}$ chosen separately for each horizon-T run, every fixed positive-gap instance satisfies*

$$\mathbb{E}R_T = \overline{\Delta} F_{\mathrm{FE}}(T) + O(1) = O(T^{2/3}) \quad \text{as } T \to \infty. \tag{14}$$

The precision schedule requires no gap or tail information. It makes the persistent CDF error smaller than any fixed positive gap for sufficiently large horizons. The statistical and sketch errors enter the same historical-error bound.

6.5 CALIBRATION UNDER STOCHASTIC ORDERING

Suppose the arm distributions belong to an FSD-ordered family: $\theta_a > \theta_b$ implies $F_{\theta_a}(r) \leq F_{\theta_b}(r)$ for all r (Shaked & Shanthikumar, 2007). Since $G^{\star}$ is nondecreasing, the rank map preserves this order.

Proposition 6.6 (FSD calibration). *Under this ordering, $\mu_{\theta_a} \geq \mu_{\theta_b}$, all fixed quantiles are nondecreasing in θ, and finite means are nondecreasing in θ. If a monotone coupling also satisfies $\mathbb{P}(G^\star(R_{\theta_a}) > G^\star(R_{\theta_b})) > 0$, the rank inequality is strict. Thus an arm with largest parameter is optimal for all three criteria; with strict rank and mean orderings, their unique optimal arms coincide.*

6.5.1 Prevalence of FSD in heavy-tailed families and implications

FSD captures a common mechanism: actions change outcome magnitude in a consistent direction. Examples include increasing a positive scale or threshold, decreasing a Pareto tail index at fixed minimum, increasing a Lomax scale, and increasing lognormal location at fixed dispersion. These cover both power-law and multiplicative-growth models; Appendix B gives the precise directions and couplings. FSD is a calibration condition, while all regret guarantees apply without it.

Within these families, rank scores, quantiles, and finite means share a parameter ordering. The methodological implication is that learning a useful arm ordering need not require estimating its potentially unstable reward magnitude. RANK-UCB1 makes every observation contribute to a bounded score and learns the reference from the same observations. Thus an ordering that is meaningful for several objectives can be learned without supplying a moment bound or selecting a single quantile. FSD provides a practical criterion for choosing the surrogate: when actions shift outcomes in a consistently favorable direction, rank learning can identify an arm that is also optimal for finite mean and every fixed quantile. When distributions cross, the relevant question is whether success means outperforming a random alternative or maximizing reward magnitude. Rank surrogates directly address the former, including when the latter has no finite expectation.

7 Practical performance

7.1 Computational complexity

Each round uses one reward, one CDF insertion, K CDF queries, and at most K index evaluations. FE adds a counter and a cyclic pointer. Exact eCDFs use $O(T + K)$ storage; order-statistic trees give $O(KT \log(T + 1))$ total time. A GK summary with precision η stores $O(\eta^{-1} \log(2 + \eta n_i))$ entries per arm (Greenwald & Khanna, 2001), giving

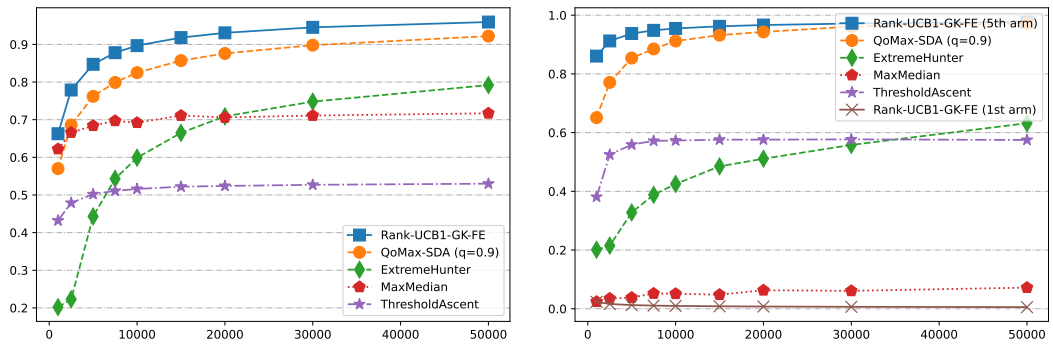
$$O\big(K\eta^{-1} \log(2 + \eta T)\big) = O(KT^{1/3} \log(T + 1)) \quad \text{when } \eta = T^{-1/3}. \tag{15}$$

If insertion costs $U(M)$ amortized time and a CDF query costs $Q(M)$ for a representation of at most M entries, total time is $O(T[U(M) + KQ(M) + K])$. The buffered adapter has logarithmic queries and a conservative total bound $O(T(M + K) \log(M + 1))$, including compression and buffer maintenance (Appendix C). On the two main instances, GK uses about 134 peak entries per arm, compared with roughly $48{,}000$ for exact FE, while closely preserving its allocation behavior.

7.2 Empirical performance

We compare with QoMax-SDA, ExtremeHunter, MaxMedian, and ThresholdAscent (Baudry et al., 2022; Carpentier & Valko, 2014; Bhatt et al., 2023; Streeter & Smith, 2006). Their benchmark configurations require no oracle reward-moment bounds. RANK-UCB1 uses the same fixed settings in every environment: $\rho = 1/6$, $\alpha = 2/3$, and $\eta = 50{,}000^{-1/3} \simeq 0.02714$. Each curve contains nine checkpoints of a 50,000-round trajectory, averaged over 1,000 independent trials. Appendix D provides all six benchmark configurations, full results, seeds, sensitivity studies, and robust-mean comparisons.

Why these benchmarks and metrics? The synthetic families are controlled abstractions of mechanisms documented in real heavy-tailed data: Pareto-type straggler delays in distributed computing (Wang et al., 2015), heavy-tailed web traffic (Crovella & Bestavros, 1997), and power-law or lognormal workloads associated with multiplicative growth (Mitzenmacher, 2004). They isolate tail behavior, stochastic ordering, and moment existence while making population objectives exactly known. This permits a reproducible assessment of how an algorithm responds to the statistical mechanisms that create difficult real-valued feedback.



(a) Pareto: common optimal arm

(b) Lognormal: different objectives

Figure 1: Best-arm pull fractions versus rounds, averaged over 1,000 trials. In (a), every method is evaluated on the same dominant arm. In (b), the extreme-bandit baselines target arm 1; RANK-UCB1 reports allocation to both the rank-optimal arm 5 and arm 1. The horizontal axis is the observation budget and the vertical axis is pull fraction.

Our primary metric is the *best-arm pull fraction*, $N_{i^\star}(T)/T$. It directly measures how quickly the policy concentrates its budget on the desired arm. Unlike statistics formed from realized reward magnitudes, this bounded allocation statistic cannot be dominated by a single extreme observation; it remains interpretable even when means are infinite. To also assess regret on shared objectives, Appendix D reports rank pseudo-regret and, on finite-mean FSD instances, mean pseudo-regret computed from analytic population scores and recorded pull counts. Thus regret evaluation itself introduces no noisy estimate of the ground-truth means.

Pareto: efficient learning under a common ordering. The five unit-minimum Pareto arms have tail indices $(1.1, 1.3, 1.9, 2.1, 2.3)$. Arm 1 is stochastically dominant, so rank, finite-mean, and extreme objectives agree. Figure 1(a) shows that RANK-UCB1 reaches a best-arm pull fraction of 0.9594 at $T = 50{,}000$, versus 0.922, 0.792, 0.717, and 0.530 for QoMax-SDA, ExtremeHunter, MaxMedian, and ThresholdAscent. It already reaches 0.8968 at $T = 10{,}000$. The gain illustrates the value of turning the full reward stream into bounded comparison information: this ordering can be learned effectively despite an optimal arm with infinite variance and without estimating its tail index.

The common-ordering results extend beyond the five-arm Pareto example. On seven-arm Pareto, RANK-UCB1 reaches a best-arm pull fraction of 0.971, compared with 0.865 for the strongest extreme-bandit baseline; on ten-arm exponential rewards, the corresponding fractions are 0.936 and 0.803. These comparisons evaluate all methods on the same optimal arm, demonstrating effective allocation across different arm counts and tail regimes (Appendix D.2).

Lognormal: what the objective selects. The five arms have log-locations $(1, 1.5, 2, 3, 3.5)$ and dispersions $(4, 3, 2, 1, 0.5)$. Arm 1 has the largest mean and asymptotically heaviest tail; arm 5 maximizes the rank score. In Figure 1(b), RANK-UCB1 allocates 97.66% of pulls to arm 5 and 0.55% to arm 1. The population win scores are $\mu_5 = 0.68325$ and $\mu_1 = 0.39062$: a higher expectation driven by rare extremes need not imply a higher probability of superiority. The plot displays this difference in learning targets, with the baseline curves retaining their extreme-optimal target. Rank learning favors the arm with stronger relative outcomes across repeated decisions.

The full suite in Appendix D.2 also includes censored Gaussian rewards with ties and a Dirac-Pareto mixture, which respectively test discrete probability mass and disagreement between rank and extreme objectives.

Infinite-mean stress test. We additionally consider five Pareto arms for which every reward mean is infinite. The same GK FE configuration reaches a rank-optimal pull fraction of 0.9595 at $T = 50{,}000$, without changing the algorithm or supplying tail information (Appendix D.6). This verifies experimentally that the moment-free formulation remains operational in the regime where a finite-mean objective ceases to exist.

Table 1: Two ablations at $T = 50{,}000$: exact no-FE versus exact FE isolates exploration; exact FE versus GK FE isolates compression. Entries are mean $\pm$ standard deviation over 1,000 trials. $f_\star$ is rank-optimal pull fraction and $W_T = T^{-1}\sum_i N_i(T)\mu_i$ is average population win-probability.

	Pareto		Lognormal	
Variant	$f_\star$	W_T	$f_\star$	W_T
Exact, no FE	0.9738 ± 0.0120	0.59991 ± 0.00055	0.9953 ± 0.0017	0.68250 ± 0.00024
Exact, FE	0.9598 ± 0.0092	0.59785 ± 0.00038	0.9767 ± 0.0008	0.67804 ± 0.00010
GK, FE	0.9594 ± 0.0093	0.59784 ± 0.00038	0.9766 ± 0.0008	0.67804 ± 0.00010

7.3 Forced exploration and compression

Table 1 isolates the two mechanisms using paired reward streams. FE spends some pulls on reference coverage: it slightly reduces average win-probability, while reducing the standard deviation of optimal-arm allocation by 23% on Pareto and 53% on lognormal. This measured tradeoff accompanies its expected-regret guarantee for every fixed instance. GK changes the mean FE pull fraction by less than 0.0004 in either setting, while reducing peak per-arm entry counts by more than a factor of 350. Appendix D.4 further varies GK precision and the exploration coefficient.

Broader evidence. Appendix D.2 covers all six extreme-bandit benchmarks, including seven-arm Pareto, exponential, censored Gaussian with ties, and a Dirac–Pareto mixture. Appendix D.5 compares exact FE and no-FE with truncated-mean and median-of-means UCB on both mean and rank pseudo-regret. Despite oracle moment information, 64 of 76 tested radius configurations have pilot mean pseudo-regret above RANK-UCB1's formal-evaluation average. Selected configurations outperform it on fresh trials; RANK-UCB1 requires neither this extensive instance-specific tuning nor moment bounds. Appendix D.5 reports the full grid and separate evaluation protocols.

8 Discussion and conclusion

A defining feature of our approach is the win-probability score $\mu_i = \mathbb{E}[G^\star(R_i)]$. It remains meaningful when the reward mean is infinite or undefined and captures the probability of obtaining a stronger outcome than a random reference arm. FSD calibration explains when this ordering also identifies the best tail parameter, quantile, and finite mean. With crossing distributions, the objective should follow the decision-maker's utility: consistent superiority, mean magnitude, and extreme outcomes encode different preferences, as the lognormal and mixture experiments illustrate.

The uniform mixture provides a reference independent of sampling frequencies. Its meaning depends on the candidate arm set: adding or removing an arm can alter the scores and potentially their maximizer. For a fixed decision problem, this makes the comparison population explicit. Task-driven references, including a fixed external population or a weighted set of alternatives, offer a way to align rank learning with particular applications. Learning and adapting such references is a promising direction for future work.

More broadly, we view population-relative feedback as a potential *preprocessing module for heavy-tailed rewards in reinforcement learning*: real-valued feedback is mapped to a bounded relative-position representation before it reaches a learning algorithm. This suggests a new design pattern for RL applications where heterogeneous utilities or costs can make absolute reward scale unreliable. Extending the present stochastic-bandit analysis to state-dependent references and temporally dependent feedback is an open direction. Under calibrated orderings, it can also be understood as *using rank to infer mean ordering*, without having to estimate the (sometimes fragile) mean itself.

RANK-UCB1 makes this viewpoint concrete. It learns a shared reference from the observations already collected, controls the historical error that this learning creates, and compresses the reference with deterministic rank guarantees. The resulting GK FE algorithm has fixed-instance $O(T^{2/3})$ expected rank-regret without reward moment assumptions, together with strong allocation performance and a small empirical memory footprint. Rank surrogates therefore provide both a practical method for heavy-tailed bandits and a route to rethinking the representation of difficult reward signals.

AI USE STATEMENT

In this work, we used generative AI tools for <formulate mathematical claims, assist in the writing of proofs, design or provide feedback on research methodology or experiments, implement methods, assist with translation, interpret results>. We have not used generative AI tools for <develop theoretical models or conceptual frameworks, help provide critical ingredients for proving mathematical claims, propose or refine hypotheses>, and <generate synthetic data sets, clean and reformat dataset, support qualitative and thematic data analysis> are not applicable to this work. Additionally, we used generative AI tools for <brainstorming, edit a research paper to improve readability, identify relevant literature>. We also used generative AI tools to organize and refactor our experimental code to improve its structure and reproducibility.

We have reviewed all AI-assisted work. We manually verified the authenticity and relevance of literature references suggested by the LLM. We also carefully reviewed, tested, and validated all LLM-assisted code modifications for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Appendix A gives complete proofs for the theoretical claims; Appendix C specifies the CDF interface and update order. Appendix D provides distributions, analytic evaluation formulas, all tuning settings, seeds, and commands. The *submitted supplementary experiment package* includes source code, trial-level results, configuration records, and plotting scripts.

REFERENCES

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

Dorian Baudry, Yoan Russac, and Emilie Kaufmann. Efficient algorithms for extreme bandits. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 2210–2248. PMLR, 2022.

Viktor Bengs, Róbert Busa-Fekete, Adil El Mesaoudi-Paul, and Eyke Hüllermeier. Preference-based online learning with dueling bandits: A survey. *Journal of Machine Learning Research*, 22(7): 1–108, 2021.

Sujay Bhatt, Ping Li, and Gennady Samorodnitsky. Extreme bandits using robust statistics. *IEEE Transactions on Information Theory*, 69(3):1761–1776, 2023.

Sébastien Bubeck, Nicolò Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

Alexandra Carpentier and Michal Valko. Extreme bandits. In *Advances in Neural Information Processing Systems 27*, pp. 1089–1097, 2014.

Yu Chen, Jiatai Huang, Yan Dai, and Longbo Huang. uniINF: Best-of-both-worlds algorithm for parameter-free heavy-tailed MABs. In *The Thirteenth International Conference on Learning Representations*, 2025.

Mark E. Crovella and Azer Bestavros. Self-similarity in world wide web traffic: evidence and possible causes. *IEEE/ACM Transactions on Networking*, 5(6):835–846, 1997.

Gianmarco Genalti and Alberto Maria Metelli. Parameter-free heavy-tailed bandits. arXiv preprint arXiv:2607.29460, 2026.

Michael Greenwald and Sanjeev Khanna. Space-efficient online computation of quantile summaries. In *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*, pp. 58–66. ACM, 2001.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Matthieu Jedor, Vianney Perchet, and Jonathan Louëdec. Categorized bandits. In *Advances in Neural Information Processing Systems*, volume 32, pp. 14399–14409, 2019.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 15696–15707. PMLR, 2023.

Davide Maran, Francesco Bacchiocchi, Francesco Emanuele Stradi, Matteo Castiglioni, Nicola Gatti, and Marcello Restelli. Bandits with ranking feedback. In *Advances in Neural Information Processing Systems 37*, pp. 80567–80608, 2024.

Pascal Massart. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990.

Michael Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2):226–251, 2004.

Manhin Poon, Xiangxiang Dai, Xutong Liu, Fang Kong, John C. S. Lui, and Jinhang Zuo. Online multi-LLM selection via contextual bandits under unstructured context evolution. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(29):24855–24863, 2026.

Moshe Shaked and J. George Shanthikumar. *Stochastic Orders*. Springer Series in Statistics. Springer, 2007. ISBN 9780387329154.

Matthew J. Streeter and Stephen F. Smith. A simple distribution-free approach to the max K-armed bandit problem. In *Principles and Practice of Constraint Programming – CP 2006*, volume 4204 of *Lecture Notes in Computer Science*, pp. 560–574. Springer, 2006.

Yunyue Su, Jinshuai Zhang, Bowen Fang, Wen Ye, Jinghao Zhang, Bowen Song, Weiqiang Wang, Qiang Liu, and Liang Wang. Toolscaler: Scalable generative tool calling via structure-aware semantic tokenization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 556–578. Association for Computational Linguistics, 2025.

Balázs Szörényi, Róbert Busa-Fekete, Paul Weng, and Eyke Hüllermeier. Qualitative multi-armed bandits: A quantile-based approach. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1660–1668. PMLR, 2015.

Da Wang, Gauri Joshi, and Gregory Wornell. Using straggler replication to reduce latency in large-scale parallel computing. *ACM SIGMETRICS Performance Evaluation Review*, 43(3):7–11, 2015.

Liyuan Xu, Junya Honda, and Masashi Sugiyama. Dueling bandits with qualitative feedback. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):5549–5556, 2019.

A Detailed proofs

All probabilities refer to the independent, pre-generated reward streams from Section 3 and any nonanticipating policy randomization. The rank pseudo-regret is random; expectation is written explicitly. The arguments allow atoms and multiple optimal arms.

A.1 Simultaneous concentration over sample prefixes

For fixed $T \geq K$ and $L > 0$, define $\mathcal{E}_T(L)$ as the event that, for every $i \in [K]$ and every deterministic $1 \leq n \leq T$,

$$\|\widehat{F}_{i,n} - F_i\|_\infty \leq \sqrt{\frac{L}{2n}}, \qquad \left|\frac{1}{n}\sum_{m=1}^{n} X_{i,m} - \mu_i\right| \leq \sqrt{\frac{L}{2n}}. \tag{16}$$

The DKW–Massart inequality bounds the failure probability of the first inequality by $2e^{-L}$ (Massart, 1990). Since $X_{i,m}$ is an i.i.d. bounded sequence for each arm, Hoeffding gives the same bound for the second (Hoeffding, 1963). A union bound over $2KT$ events yields

$$\mathbb{P}(\mathcal{E}_T(L)^c) \leq 4KTe^{-L}.$$

Taking $L = \log(4KT/\delta)$ gives probability at least $1 - \delta$. On this event, the inequalities can be evaluated at random counts $N_i(s)$. This avoids conditioning an i.i.d. claim on an adaptively chosen count. Uniformity of the CDF bound over its argument also permits querying the current reward after it has been inserted.

A.2 Counting forced exploration

Let $F_{\mathrm{FE}}(s)$ denote the cumulative forced count after round s. The first K rounds are forced. Subsequently $M(s) = \max\{K, \lceil s^\alpha \rceil\}$ is at most s and increases by at most one per round, since $0 < \alpha < 1$. Induction on the update rule gives

$$F_{\mathrm{FE}}(s) = \min\{s, M(s)\}.$$

The cyclic allocation supplies at least $\lfloor F_{\mathrm{FE}}(s)/K \rfloor$ observations to each arm. For $s \geq K$, use $\lfloor x \rfloor \geq x/2$ for $x \geq 1$ to obtain

$$N_j(s) \geq \frac{s^\alpha}{2K}, \qquad j \in [K]. \tag{17}$$

Round-robin counts differ from $F_{\mathrm{FE}}(T)/K$ by at most one. Consequently,

$$\left|R_T^{\mathrm{FE}} - \overline{\Delta} F_{\mathrm{FE}}(T)\right| \leq K. \tag{18}$$

These are deterministic statements about the schedule, with no assumption on how the index allocates its other pulls.

A.3 Proof of Lemma 6.2

On $\mathcal{E}_T(L)$, (17) bounds each arm's eCDF error at round $s \geq K$ by $\sqrt{KL}s^{-\beta}$, with $\beta = \alpha/2$. Also,

$$\frac{\kappa}{N_j(s)} \leq \frac{\kappa}{\sqrt{N_j(s)}} \leq \kappa\sqrt{2K}s^{-\beta}.$$

Combining these bounds with (5) and averaging across arms gives (9). For $s < K$, both the actual and oracle scores belong to $[0, 1]$, while $Cs^{-\beta} \geq 1$, so the same bound holds during initialization.

For the mth pull of arm i, the global time satisfies $\tau_{i,m} \geq m$. Thus

$$\begin{aligned}\left|\frac{1}{n}\sum_{m=1}^{n} Z_{\tau_{i,m}} - \frac{1}{n}\sum_{m=1}^{n} X_{i,m}\right| &\leq \frac{C}{n}\sum_{m=1}^{n} \tau_{i,m}^{-\beta} + \eta \\ &\leq \frac{C}{n}\sum_{m=1}^{n} m^{-\beta} + \eta \leq \frac{C}{1-\beta} n^{-\beta} + \eta.\end{aligned}$$

The last inequality follows by integration of the decreasing function $x^{-\beta}$. Adding the oracle-mean bound in (16) proves (10). In particular, on any nonforced round t, writing $s = t - 1 \geq K$, all arms satisfy

$$|\widehat{\mu}_i(t) - \mu_i| \leq E_t(L) \triangleq \sqrt{KL}s^{-\beta} + A(2K)^{\beta}s^{-\alpha\beta} + \eta. \tag{19}$$

The exponent $\alpha\beta = \alpha^2/2$ accounts for averaging older scores and then converting per-arm sample counts into global time.

A.4 Proof of Corollary 6.4: a uniform finite-horizon bound

Theorem A.1 (Uniform bound with FE)**.** *Under the assumptions of Lemma 6.2, for any fixed $\rho \geq 0$, with probability at least $1 - \delta$,*

$$R_T \leq \min\Bigg\{T,\ F_{\mathrm{FE}}(T) + \frac{2\sqrt{KL}}{1-\beta}T^{1-\beta} + \frac{2A(2K)^{\beta}}{1-\alpha\beta}T^{1-\alpha\beta} + 2\sqrt{\rho KT\log T} + 2\eta T\Bigg\}. \tag{20}$$

Proof. Write $b_i(t) = \sqrt{\rho \log t / n_i(t)}$. On a nonforced round, the index comparison with any optimal arm gives

$$\begin{aligned}\Delta_{I_t} &\leq |\mu_\star - \widehat{\mu}_{i^\star}(t)| + |\widehat{\mu}_{I_t}(t) - \mu_{I_t}| + b_{I_t}(t) - b_{i^\star}(t)\\ &\leq 2E_t(L) + b_{I_t}(t).\end{aligned}$$

This uses nonnegativity of the bonus, without requiring the index itself to be a confidence upper bound. Forced rounds contribute at most $F_{\mathrm{FE}}(T)$. For each arm, the pre-pull counts encountered on its nonforced selections are distinct, so

$$\sum_{t:\text{nonforced}} b_{I_t}(t) \leq \sqrt{\rho\log T}\sum_i \sum_{n=1}^{N_i(T)} n^{-1/2} \leq 2\sqrt{\rho KT\log T}.$$

Sum (19) over nonforced rounds and upper-bound by the corresponding full sums. The bound $\sum_{s=1}^{T} s^{-q} \leq T^{1-q}/(1-q)$ for $0 < q < 1$ proves (20); $R_T \leq T$ gives the minimum. □

For fixed α and bounded κ, the bound is

$$\widetilde{O}\Big(K + T^{\alpha} + K^{(1+\alpha)/2}T^{1-\alpha^2/2} + \sqrt{\rho KT} + \eta T\Big),$$

capped at T, where the hidden logarithms include the confidence parameter. Choosing $\delta = 1/T$ adds at most one to the expected regret bound. At $\alpha = 2/3$ and $\eta = T^{-1/3}$, this becomes

$$\mathbb{E}R_T = \widetilde{O}\Big(K + T^{2/3} + K^{5/6}T^{7/9} + \sqrt{\rho KT}\Big).$$

This proves Corollary 6.4; exact eCDFs follow by setting $\eta = \kappa = 0$. The exponent $7/9$ is an upper bound from this analysis; no matching lower bound is claimed.

A.5 Proof of Theorem 6.3 and Corollary 6.5

For each time $t > K$, apply the prefix event with horizon t and $L_t = \log(4Kt^4)$. Its failure probability is at most t^{-3}. Let $p = \alpha^2/2$ and

$$D = \sqrt{K}\left[1 + \frac{(1+\sqrt{2}\kappa)(2K)^{\beta}}{1-\beta}\right].$$

Since $L_t \geq 1$ and $s^{-\beta} \leq s^{-p}$ for $s = t - 1 \geq 1$, (19) gives

$$E_t(L_t) \leq D\sqrt{L_t}(t-1)^{-p} + \eta.$$

Moreover, (17) gives $b_{I_t}(t) \leq \sqrt{2\rho K\log t}(t-1)^{-\beta}$. On this event, any nonforced selection satisfies

$$\Delta_{I_t} \leq 2D\sqrt{\log(4Kt^4)}(t-1)^{-p} + \sqrt{2\rho K\log t}(t-1)^{-\beta} + 2\eta. \tag{21}$$

For exact eCDFs, take $\eta = \kappa = 0$. More generally, assume $\gamma = \Delta_{\min} - 2\eta > 0$. The first two terms on the right converge to zero. Hence there is a deterministic finite t_0, depending only on $K, \alpha, \rho, \kappa, \gamma$, after which their sum is at most $\gamma/2$. On the good event the right side is then strictly less than $\Delta_{\min}$, so no suboptimal arm can be selected on a nonforced round. Therefore

$$\mathbb{E} \sum_{t \le T : \text{nonforced}} \Delta_{I_t} \le (t_0 - 1) + \sum_{t=t_0}^{T} t^{-3} \le t_0 + 1.$$

This proves (11) with a horizon-independent constant. Combining it with (18) proves (12). By the first Borel–Cantelli lemma, suboptimal nonforced selections also occur only finitely often almost surely. Thus for exact eCDFs, or fixed-precision sketches with $2\eta < \Delta_{\min}$,

$$\frac{R_T}{T^\alpha} \longrightarrow \overline{\Delta} \quad \text{almost surely,} \qquad \frac{\mathbb{E}R_T}{T^\alpha} \longrightarrow \overline{\Delta}.$$

For the horizon-dependent precision $\eta_T = T^{-1/3}$, all $T \ge (4/\Delta_{\min})^3$ satisfy $\gamma \ge \Delta_{\min}/2$. A single t_0 therefore works for all sufficiently large horizons, proving (14). Runs with different precisions are different policies; the expectation statement does not assert almost-sure convergence along a single anytime GK run. Constants can grow substantially as the gap shrinks; Corollary 6.4 gives the separate gap-free guarantee.

A.6 Proof of the oracle bound

For known $G^\star$, each arm supplies i.i.d. observations in $[0, 1]$. Initialize each arm once and use bonus $\sqrt{2 \log t / n_i(t)}$. If a suboptimal arm has at least $8 \log T / \Delta_i^2$ observations, a subsequent selection requires an upper deviation of its empirical oracle mean or a lower deviation of an optimal arm's empirical oracle mean at some prefix. Hoeffding bounds each deviation at round t and fixed prefix by t^{-4}. The usual union bound over the candidate prefixes and summation over rounds gives

$$\mathbb{E}N_i(T) \le \frac{8 \log T}{\Delta_i^2} + 1 + \frac{\pi^2}{3}$$

as in the UCB1 analysis of Auer et al. (2002). Multiplication by Δ_i and summation prove (7).

A.7 Proof of Theorem 6.1

Deterministic logarithmic coverage. Fix $0 < a < \rho$. We show that there is a deterministic T_a such that every possible trajectory satisfies

$$\min_i N_i(T) \ge a \log T, \qquad T \ge T_a. \tag{22}$$

If $N_i(T) < a \log T$, throughout the latter half of this horizon, arm i has index at least $\sqrt{\rho \log(T/2)/(a \log T)}$. Choose $d > 0$ with $1 + d < \sqrt{\rho/a}$. For large T, this index exceeds $1 + d$. Since each historical score mean is at most one, any selected arm j must have pre-pull count less than $\rho \log T / d^2$. Each arm can therefore be selected at most $\lfloor \rho \log T / d^2 \rfloor + 1$ times in that interval. The resulting $O(K \log T)$ selections cannot fill an interval of length at least $T/2$. This proves (22) for all sufficiently large T.

Consistency of the retained scores. For every arm, the eCDF converges uniformly almost surely. This follows, for example, from DKW–Massart and Borel–Cantelli at deterministic sample prefixes. Because every arm's count tends to infinity, on the same probability-one event

$$e_s = \max_j \|\widehat{F}_{j,N_j(s)} - F_j\|_\infty \longrightarrow 0.$$

For all post-initialization pulls, $|Z_{\tau_{i,m}} - X_{i,m}| \le e_{\tau_{i,m}}$. Cesàro convergence and the strong law for the bounded i.i.d. oracle scores yield $n^{-1} \sum_{m=1}^{n} Z_{\tau_{i,m}} \to \mu_i$ simultaneously for all arms. Substituting the diverging $n_i(t)$ gives

$$\widehat{\mu}_i(t) \longrightarrow \mu_i \quad \text{almost surely.} \tag{23}$$

No independence of the retained scores is asserted.

Upper logarithmic limit. Fix a trajectory satisfying (23), a suboptimal arm i, and $0 < \varepsilon < \Delta_i/2$. After a finite random time t_ε, all empirical score errors are below ε. If i is then selected,

$$\mu_i + \varepsilon + \sqrt{\rho \log t / n_i(t)} \geq U_i(t) \geq U_{i^\star}(t) \geq \mu_\star - \varepsilon.$$

It follows that

$$N_i(T) \leq \max\left\{ N_i(t_\varepsilon),\, 1 + \frac{\rho \log T}{(\Delta_i - 2\varepsilon)^2} \right\}.$$

Divide by $\log T$, let $T \to \infty$, then let $\varepsilon \downarrow 0$ to obtain the required upper limit. This step does not take the expectation of the random transient.

Lower logarithmic limit. Fix $0 < c < \rho/\Delta_i^2$ and choose $\varepsilon > 0$ such that $\sqrt{\rho/c} > \Delta_i + 3\varepsilon$. If $N_i(T) < c \log T$ for an arbitrarily large T, then for every round u in its latter half,

$$U_i(u) \geq \mu_i - \varepsilon + \sqrt{\rho \log(T/2)/(c \log T)} > \mu_\star + 2\varepsilon.$$

The selected arm j has empirical score at most $\mu_\star + \varepsilon$, so its bonus must exceed ε. Thus $n_j(u) < \rho \log T/\varepsilon^2$. As in the coverage argument, the interval can contain only $O(K \log T)$ selections, a contradiction. Letting $c \uparrow \rho/\Delta_i^2$ proves the lower limit. Summing over the finite set of suboptimal arms gives the limit for $R_T/\log T$. Finally, $R_T/T \to 0$ almost surely and $0 \leq R_T/T \leq 1$, so bounded convergence gives $\mathbb{E}R_T = o(T)$.

A.8 Expected rates without continued FE

The difference between pathwise and expected regret is governed by the tails of the historical estimation error. A sufficient condition for expected logarithmic regret on a fixed instance is: for some $0 < \varepsilon < \Delta_{\min}/2$,

$$\sum_{t=K+1}^{\infty} \mathbb{P}\Big(\max_i |\widehat{\mu}_i(t) - \mu_i| > \varepsilon \Big) < \infty. \tag{24}$$

On every complementary round, selection of suboptimal i requires $n_i(t) \leq \rho \log t/(\Delta_i - 2\varepsilon)^2$. Counting these selections and bounding each exceptional round's regret by one gives

$$\mathbb{E}R_T \leq K + \sum_{i:\Delta_i > 0} \Delta_i \left(1 + \frac{\rho \log T}{(\Delta_i - 2\varepsilon)^2} \right) + \sum_{t>K} \mathbb{P}\Big(\max_i |\widehat{\mu}_i(t) - \mu_i| > \varepsilon \Big).$$

Hence (24) implies $O(\log T)$ expectation. This is an additional instance-dependent condition; pointwise almost-sure consistency is insufficient to establish it.

A general explicit bound follows without this condition. Write $\Delta = \Delta_{\min}$. For every $0 < \zeta < \rho\Delta^2/2$,

$$\mathbb{E}R_T = O_{K,\rho,\Delta,\zeta}\big(T/(\log T)^\zeta \big). \tag{25}$$

To prove it, choose $0 < a < \rho$ and positive γ, r, u with $v = \gamma + r + u < \Delta/2$. For large t, let $m = \lfloor \gamma a \log(t-1) \rfloor$. Summing DKW over all deterministic prefixes $n \geq \lceil a \log m \rceil$ gives

$$\mathbb{P}\left(\exists j, n \geq \lceil a \log m \rceil : \|\widehat{F}_{j,n} - F_j\|_\infty > r \right) \leq \frac{2K}{1 - e^{-2r^2}} m^{-2ar^2}.$$

Outside this event, (22) makes every score generated at a global time $s \geq m$ accurate within r of its oracle value. Among the $n_i(t) \geq a \log(t-1)$ retained scores of an arm, at most m were generated earlier. Their average contribution to the error is at most $m/n_i(t) \leq \gamma$. Similarly, a Hoeffding sum over all $n \geq \lceil a \log(t-1) \rceil$ bounds the probability of an oracle-mean error exceeding u by

$$\frac{2K}{1 - e^{-2u^2}} (t-1)^{-2au^2}.$$

Thus, for deterministic constants C_1, C_2,

$$\mathbb{P}\left(\max_i |\widehat{\mu}_i(t) - \mu_i| > v \right) \leq C_1 (\log t)^{-2ar^2} + C_2 t^{-2au^2}.$$

The same selection count as above, followed by summation over t, gives $\mathbb{E}R_T = O(T/(\log T)^{2ar^2})$. Taking a arbitrarily close to ρ and r arbitrarily close to $\Delta/2$, while keeping γ, u positive, proves (25).

The essential issue is the shared reference: an arm with a low score can still determine the ranking of other arms, and old reference errors persist in their stored scores. Therefore the sampling rate suggested by an arm's own gap need not supply the information needed to rank the other arms. The above results leave room for stronger expected rates on specific Pareto or lognormal instances; they do not assert those rates from finite experimental curves.

B STOCHASTICALLY ORDERED FAMILIES

B.1 PROOF OF FSD CALIBRATION

Let $Q_\theta(u) = \inf\{x : F_\theta(x) \geq u\}$ be the generalized quantile function. If $F_{\theta_a} \leq F_{\theta_b}$, then $Q_{\theta_a}(u) \geq Q_{\theta_b}(u)$ for every $u \in (0,1)$. Couple the rewards using a common uniform U: $R_{\theta_a} = Q_{\theta_a}(U)$ and $R_{\theta_b} = Q_{\theta_b}(U)$. Monotonicity of $G^\star$ gives

$$G^\star(R_{\theta_a}) \geq G^\star(R_{\theta_b}) \quad \text{almost surely.}$$

Expectations preserve this inequality because both scores are bounded. It is strict if the two coupled scores differ with positive probability. Quantile monotonicity was established directly; finite-mean monotonicity follows by taking expectations of the coupled rewards. This proves Proposition 6.6. Weak ordering alone permits ties, so coincidence of unique optimizers is asserted under the stated strictness conditions.

B.2 PARAMETER DIRECTIONS AND CONCRETE FAMILIES

A sufficient condition for FSD is a representation $R_\theta = h(\theta, Z)$ with $h(\cdot, z)$ nondecreasing for every fixed z. In particular, location increases preserve FSD, and positive scale increases do so for a nonnegative base variable. The following formulas make the directions explicit; $q \in (0,1)$.

Pareto Type I. For minimum $x_m > 0$ and tail index $a > 0$,

$$F(r) = 1 - (x_m/r)^a \quad (r \geq x_m), \qquad Q(q) = x_m(1-q)^{-1/a}.$$

Increasing x_m at fixed a, or decreasing a at fixed x_m, increases every quantile. The median is $x_m 2^{1/a}$ and the finite mean is $ax_m/(a-1)$ for $a > 1$, with the same directions. For $a \leq 1$, the positive mean is infinite while the rank objective remains bounded.

Lomax / Pareto Type II. For $F(r) = 1 - (1 + r/\lambda)^{-a}$ on $r \geq 0$,

$$Q(q) = \lambda((1-q)^{-1/a} - 1), \qquad \mathbb{E}R = \lambda/(a-1) \quad (a > 1).$$

Increasing λ or decreasing a gives FSD and increases the median and all quantiles.

Lognormal with fixed dispersion. If $R = \exp(m + \sigma Z)$ for $Z \sim \mathcal{N}(0,1)$ and fixed σ, increasing m gives a monotone coupling. In this case

$$Q(q) = \exp(m + \sigma\Phi^{-1}(q)), \quad \mathrm{Med}(R) = e^m, \quad \mathbb{E}R = e^{m+\sigma^2/2}.$$

Changing σ at fixed m changes dispersion and produces crossing CDFs. For independent lognormal arms, the difference of their log-rewards is normal, yielding the pairwise probability formula used in Section 7.

Other positive scale families. For $R_\lambda = \lambda Z$ with $Z \geq 0$, increasing λ gives FSD. This includes Fréchet and fixed-shape Weibull families; Weibull is included as a scale-family example, with its tail classification depending on its shape. Generalized Pareto exceedances with fixed shape $\xi > 0$ have

$$Q(q) = \frac{\sigma}{\xi}((1-q)^{-\xi} - 1), \qquad \mathbb{E}R = \frac{\sigma}{1-\xi} \quad (0 < \xi < 1),$$

so increasing σ gives FSD; adding an increasing threshold preserves it. In contrast, increasing the scale of a symmetric real-valued distribution generally spreads both tails and need not give FSD.

C IMPLEMENTATION AND COMPUTATIONAL DETAILS

C.1 EXACT ECDF AND UPDATE ORDER

An exact structure stores the multiset of rewards of one arm, augmented with subtree counts if logarithmic rank queries are desired. Its CDF query is the number of observations less than or equal to the query, divided by the arm count. Algorithm 1 inserts the observed reward and increments its count before querying any CDF. Empty structures return zero during initialization. Only the selected arm's score sum is changed. Earlier surrogate scores need not be stored individually: their cumulative sum and count suffice.

C.2 GK CDF QUERY CONTRACT

GK maintains ordered retained values and intervals enclosing their ranks (Greenwald & Khanna, 2001). A CDF interface must enclose the empirical *upper rank* $r_n(x) = \#\{m \le n : R_m \le x\}$, including all ties. Suppose its interval $[\ell(x), u(x)]$ contains $r_n(x)$ and has width at most $2\eta n + 2\kappa$. The clipped midpoint query

$$\widetilde{F}_n(x) = \left[\frac{\ell(x) + u(x)}{2n}\right]_{[0,1]}$$

then has error at most $\eta + \kappa/n$. Clipping cannot increase the error because the true CDF lies in $[0, 1]$. The analysis uses this deterministic contract at every query and update.

For an ordered summary, one construction uses the last retained value at most x to lower-bound the upper rank and the first retained value strictly greater than x to upper-bound it, subtracting one from the successor's rank upper bound. Queries below the minimum and at or above the maximum can be handled exactly. Duplicate values must be treated by upper-rank semantics, for example by ordering occurrences within equal values and locating the successor after all equal keys. The rounding constant κ accommodates integer endpoints; it is fixed independently of sample count and horizon.

If n_s points have been summarized and b pending points are stored exactly, query the pending buffer exactly and add its rank to the summarized rank estimate before dividing by $n = n_s + b$. An error bound $\eta n_s + \kappa$ in the summarized rank gives total CDF error at most $\eta + \kappa/n$. The buffer and any search index must be included in the reported storage. The standard GK space bound is used with a CDF adapter maintaining these interval invariants; a quantile-query interface or empirical accuracy test alone is not a substitute for the CDF contract.

C.3 TIES, ORDER, AND NUMERICAL REPRESENTATION

Right-continuous queries give full credit to ties, as in (2). In particular, an atomic distribution's independent self-comparison need not equal $1/2$. Giving half credit instead would require replacing both the population CDF and every empirical query by their mid-CDF versions. The present manuscript consistently uses full credit. Index tie-breaking is a separate algorithmic choice and may be any nonanticipating rule.

The mathematical interface accepts arbitrary finite real rewards. A numeric implementation must preserve their ordering, including negative values if present. The supplied implementation uses an order-preserving integer encoding of nonnegative IEEE-754 rewards, which covers every experimental distribution. General real inputs require a signed order-preserving adapter; the mathematical CDF interface and proofs apply to arbitrary finite real rewards. Strictly increasing transformations preserve exact ranks and the population objective; they do not impose reward moment conditions.

C.4 STORAGE, OPERATION COUNTS, AND RUNTIME REPORTING

The learner stores two scalars and one CDF structure per arm, plus the FE counter. It does not store histories of reference distributions or rescore rewards. At each round, one structure receives an insertion and all K structures receive a query. Let M upper-bound all live entries in any arm's representation. Binary searches of the finalized summary and pending buffer give $Q(M) = O(\log(M + 1))$. In the map-and-heap GK update, a compression candidate is either removed or reinserted with a threshold above the current threshold; hence each of the at most M candidates is processed at most once in

that update. Map and heap operations cost $O(\log(M+1))$, giving a conservative $O(M\log(M+1))$ bound per summarized observation. There are at most T such observations. Sorted-buffer insertion and summary finalization add at most $O(TM)$ total work, while the KT queries and index scans add $O(KT\log(M+1))$. Together these yield the stated $O(T(M+K)\log(M+1))$ bound. It is a worst-case accounting bound, separate from the measured runtime below.

The implementation buffers $\max\{1, \lfloor(2\eta)^{-1}\rfloor\}$ exact observations between GK finalizations. Both the finalized summary and pending buffer support binary-search CDF queries. The pending array incurs linear-time insertion in its buffer size, and finalization reconstructs the summary; those costs are included in runtime. Measured storage counts all live representations: GK tuples, finalized entries, compression-heap entries, and the exact buffer. The reported maximum is the peak for a single arm during a trial, not bytes or the sum across arms. Table 2 reports this count and total simulation time. The latter includes reward generation and all policy operations.

Table 2: Implementation measurements at $T = 50{,}000$ over 1,000 trials (mean $\pm$ standard deviation). Storage is peak entries per arm; time is milliseconds per complete trial.

Environment	Variant	Peak entries/arm	Time (ms)
Pareto-5	Exact, FE	47990.7 ± 458.9	57.76 ± 7.62
Pareto-5	GK, FE	133.6 ± 3.9	31.47 ± 3.74
Lognormal	Exact, FE	48832.5 ± 39.0	50.69 ± 7.31
Lognormal	GK, FE	133.5 ± 3.8	28.87 ± 3.77

All policies consume one observation per round. For every plotted trajectory, GK precision remains fixed at $\eta = 50{,}000^{-1/3}$, including its earlier checkpoints. The horizon-dependent guarantee in Corollary 6.5 describes the family of runs indexed by their prescribed terminal horizon. The finite-horizon theorem also applies to any earlier checkpoint with that run's fixed precision.

D Complete experiments

D.1 Protocol, metrics, and analytic objectives

All formal configurations use 1,000 trials and a terminal horizon of $50{,}000$. Checkpoints are $1000, 2500, 5000, 7500, 10000, 15000, 20000, 30000, 50000$. Rank variants use the same per-trial, per-arm reward streams within each environment; ties between equal indices use a separate random stream. This pairing isolates policy changes from changes to the underlying reward sequences. Unless varied explicitly, $\rho = 1/6$, FE has exponent $2/3$, and GK precision is $\eta_0 = 50{,}000^{-1/3}$. Every round, including initialization and FE, consumes one reward observation and counts against the same budget.

For a unique rank-optimal arm, write $f_\star(T) = N_{i^\star}(T)/T$. Besides this allocation metric, we report

$$W_T = \frac{1}{T}\sum_i N_i(T)\mu_i, \qquad R_T = T(\mu_\star - W_T), \qquad R_T^{\mathrm{mean}} = \sum_i N_i(T)(m_\star - m_i), \tag{26}$$

where $m_i = \mathbb{E}R_i$ is used only in the finite-mean Pareto and exponential comparisons. These metrics use analytic population scores and recorded pull counts, rather than realized reward sums. For a positive-gap instance,

$$\Delta_{\min}(1 - f_\star(T)) \le R_T/T \le \Delta_{\max}(1 - f_\star(T)),$$

which connects the primary allocation metric directly to normalized rank-regret. Every $\pm$ in the tables denotes the sample standard deviation across independent trials, not a standard error.

All rank scores are computed as $\mu_i = K^{-1}\sum_j p_{ij}$ from the following pairwise probabilities, using independent draws. For Pareto arms with index a_i and minimum b_i,

$$p_{ij} = \begin{cases} 1 - \dfrac{a_i}{a_i + a_j}\left(\dfrac{b_j}{b_i}\right)^{a_j}, & b_i \ge b_j, \\ \dfrac{a_j}{a_i + a_j}\left(\dfrac{b_i}{b_j}\right)^{a_i}, & b_i < b_j. \end{cases}$$

For exponential rates λ_i, $p_{ij} = \lambda_j/(\lambda_i + \lambda_j)$. For lognormal log-locations ℓ_i and dispersions σ_i, $p_{ij} = \Phi((\ell_i - \ell_j)/\sqrt{\sigma_i^2 + \sigma_j^2})$. For censored Gaussian rewards $1 + \sigma_i \max\{Z, 0\}$ with $Z \sim \mathcal{N}(0, 1)$, $p_{ij} = 1/2 + (2\pi)^{-1} \arctan(\sigma_i/\sigma_j)$. The latter includes full credit to ties at one, so $p_{ii} = 5/8$. For a mixture with mass z_i at zero and otherwise a unit-minimum Pareto reward of index a_i, $p_{ij} = z_j + (1 - z_i)(1 - z_j)a_j/(a_i + a_j)$. Direct conditioning gives these formulas and covers the atoms explicitly. Finite means for the robust comparisons are $m_i = a_i/(a_i - 1)$ and $m_i = 1/\lambda_i$, respectively.

D.2 Six benchmark settings and complete comparisons

We use the benchmark families summarized by Baudry et al. (2022), including instances studied by Bhatt et al. (2023); Carpentier & Valko (2014). The four extreme-bandit baseline series are the fixed nine-checkpoint comparison data supplied with the experiment package, following those benchmark settings. They are reused without rerunning these four algorithms. All Rank-UCB1 and robust-mean results are computed from the supplied trial-level observations.

The distributions and arm numbering used by the simulator are:

1. **Pareto-5.** Unit-minimum Pareto arms with indices $(1.1, 1.3, 1.9, 2.1, 2.3)$. Arm 1 is stochastically dominant; it is the target for every method.
2. **Pareto-7.** Indices $(2.5, 2.8, 4, 3, 1.4, 1.4, 1.9)$ and minima $(1, 1, 1, 1, 1.1, 1, 1)$. Arm 5 is stochastically dominant and is the shared target.
3. **Exponential.** Ten rates $(2.1, 2.4, 1.9, 1.3, 1.1, 2.9, 1.5, 2.2, 2.6, 1.4)$. Arm 5, with the smallest rate, is the shared target. This light-tailed setting checks whether the learning mechanism also applies outside extreme heavy-tail regimes.
4. **Censored Gaussian.** Twenty arms $\max\{1, Z_i\}$ with $Z_i \sim \mathcal{N}(1, \sigma_i^2)$ and
$$\sigma = (3.35, 2.29, 1.79, 2.67, 1.70, 1.36, 1.90, 2.19, 0.80, 0.12,\\ 1.65, 1.19, 1.88, 0.89, 1.64, 1.50, 2.22, 3.03, 1.08, 0.48).$$
The equivalent representation $1 + \sigma_i \max\{Z, 0\}$ gives an FSD coupling; arm 1 is the shared target. The atom at one exercises the full-tie scoring rule.
5. **Dirac–Pareto.** Arms 1 and 2 are unit-minimum Pareto with indices 1.5 and 3. Arm 3 is zero with probability 0.8 and otherwise Pareto with index 1.1. Arm 3 dominates the far tail; arm 1 is rank-optimal. The extreme-bandit comparison panel uses arm 3 for all curves, while Table 3 additionally reports the rank-optimal allocation.
6. **Lognormal.** Log-locations $(1, 1.5, 2, 3, 3.5)$ and dispersions $(4, 3, 2, 1, 0.5)$. Arm 1 maximizes the mean and the asymptotic tail, while arm 5 maximizes rank. Figure 1(b) includes both Rank-UCB1 allocations and the baselines' allocations to arm 1.

Table 3: GK FE results at $T = 50{,}000$ (mean $\pm$ standard deviation, 1,000 trials). "Panel target" is the target used for the Rank-UCB1 curve in each comparison panel. It differs from the rank-optimal arm in the Dirac–Pareto panel.

Environment	Panel-target fraction	Rank-optimal fraction	Rank-regret
Pareto-5	0.9594 ± 0.0093	0.9594 ± 0.0093	185.34 ± 19.05
Pareto-7	0.9705 ± 0.0026	0.9705 ± 0.0026	274.09 ± 10.55
Exponential	0.9362 ± 0.0121	0.9362 ± 0.0121	287.09 ± 29.43
Censored Gaussian	0.4481 ± 0.0636	0.4481 ± 0.0636	679.76 ± 58.70
Dirac–Pareto	0.0091 ± 0.0000	0.9812 ± 0.0004	205.93 ± 2.63
Lognormal	0.9766 ± 0.0008	0.9766 ± 0.0008	260.35 ± 4.88
Infinite-mean Pareto	0.9595 ± 0.0078	0.9595 ± 0.0078	166.55 ± 19.64

On Pareto-7, GK FE reaches 0.9705 optimal-arm allocation compared with 0.865 for the strongest extreme-bandit baseline, MaxMedian. On exponential arms it reaches 0.9362, compared with 0.803 for QoMax-SDA. These results extend the common-ordering gains beyond the five-arm main example. The censored Gaussian setting has more arms and smaller rank differences: GK FE reaches 0.4481,

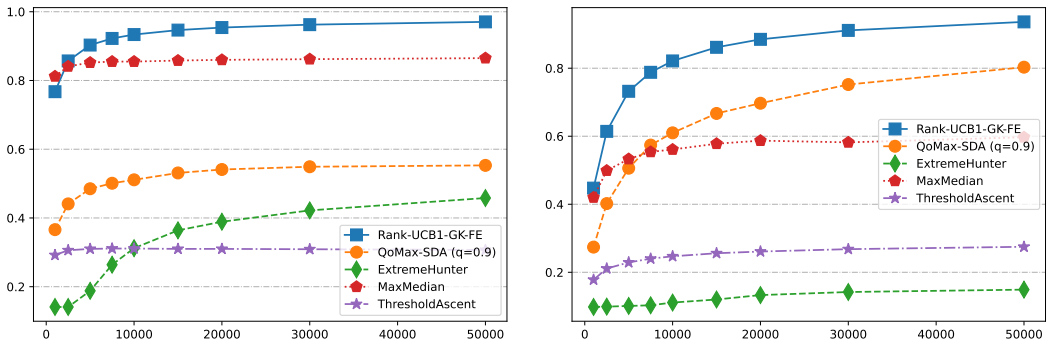


Figure 2: Best-arm pull fractions versus rounds on Pareto-7 (left) and exponential arms (right). Each panel compares all policies on its shared dominant arm. The horizontal axis is rounds and the vertical axis is pull fraction.

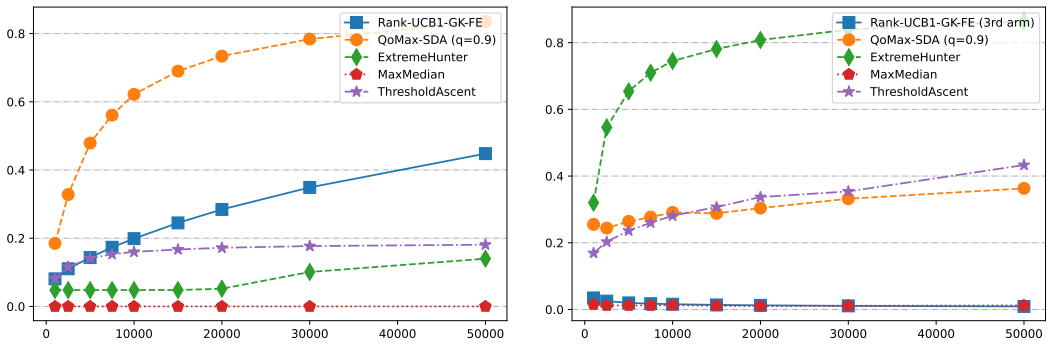


Figure 3: Pull fractions versus rounds on censored Gaussian arms (left, shared dominant arm 1) and the Dirac–Pareto mixture (right, extreme-optimal arm 3). The mixture panel exposes different objectives; RANK-UCB1's rank-optimal allocation to arm 1 is reported separately in Table 3.

while QoMax-SDA reaches 0.836 on this light-tailed extreme-bandit benchmark. The complete suite therefore identifies where bounded rank learning is most useful, with the strongest common-objective gains on the Pareto and exponential instances.

In the Dirac–Pareto instance, GK FE allocates 0.9812 of its pulls to rank-optimal arm 1 and 0.0091 to extreme-optimal arm 3. Arm 3's occasional large outcomes give it the heaviest tail but a low probability of superiority. Together with the lognormal example, this demonstrates how the rank objective distinguishes repeated relative performance from rare-event magnitude. Curves targeting different arms are used to explain this objective choice rather than to rank methods on a common utility.

D.3 FE AND COMPRESSION ABLATIONS

Table 4 shows three checkpoints for the same paired trials as Table 1. FE improves reference coverage while assigning a prescribed fraction of pulls to every arm. At $50{,}000$ rounds, the schedule makes $\lceil 50{,}000^{2/3} \rceil = 1358$ forced pulls in total, including initialization. The resulting reduction in the across-trial standard deviation accompanies a modest change in mean allocation and population win-probability. Compression has a much smaller effect: at the final horizon the GK/exact FE differences in mean optimal-arm allocation are -0.0003653 for Pareto and -0.0000161 for lognormal.

Table 4: Rank-optimal pull fractions at three checkpoints; mean $\pm$ standard deviation over the same 1,000 paired trials. Precision stays fixed throughout each trajectory.

Environment	T	Exact, no-FE	Exact, FE	GK, FE
Pareto-5	1,000	0.6747 ± 0.1447	0.6630 ± 0.1229	0.6627 ± 0.1230
Pareto-5	10,000	0.9135 ± 0.0405	0.8976 ± 0.0320	0.8968 ± 0.0323
Pareto-5	50,000	0.9738 ± 0.0120	0.9598 ± 0.0092	0.9594 ± 0.0093
Lognormal	1,000	0.9012 ± 0.0348	0.8610 ± 0.0266	0.8606 ± 0.0267
Lognormal	10,000	0.9820 ± 0.0064	0.9548 ± 0.0038	0.9548 ± 0.0039
Lognormal	50,000	0.9953 ± 0.0017	0.9767 ± 0.0008	0.9766 ± 0.0008

D.4 SENSITIVITY TO GK PRECISION AND INDEX EXPLORATION

We vary one factor at a time on Pareto-5 and lognormal, using 1,000 paired trials per configuration, the same formal seed set, and $T = 50{,}000$. The GK precision sweep uses $\eta \in \{\eta_0/2, \eta_0, 2\eta_0, 4\eta_0\}$ at fixed $\rho = 1/6$ and FE exponent $2/3$. The coefficient sweep uses $\rho \in \{1/12, 1/6, 1/3, 2/3\}$ with $\eta = \eta_0$ and the same FE schedule. The default configuration is shared with the main experiment; the sweeps introduce 12 additional configurations. No best setting is selected from these sweeps for the main comparisons.

Table 5: GK precision sensitivity. Each entry is mean $\pm$ standard deviation; storage is peak internal CDF entries per arm.

	Pareto-5		Lognormal	
Precision	$f_\star$	Peak entries	$f_\star$	Peak entries
$0.5\eta_0$	0.9597 ± 0.0092	261.4 ± 5.1	0.9766 ± 0.0008	260.0 ± 4.7
$1\eta_0$	0.9594 ± 0.0093	133.6 ± 3.9	0.9766 ± 0.0008	133.5 ± 3.8
$2\eta_0$	0.9590 ± 0.0095	67.7 ± 2.9	0.9766 ± 0.0008	67.7 ± 2.8
$4\eta_0$	0.9580 ± 0.0099	33.5 ± 2.0	0.9766 ± 0.0008	33.5 ± 2.0

Across the eightfold range from $\eta_0/2$ to $4\eta_0$, the Pareto allocation changes by less than 0.0017, and lognormal by less than 0.0001, while peak entry counts fall from about 260 to 34. These measurements show that the observed arm ordering can be stable across a substantial range of CDF precisions. They complement the deterministic query contract: the theorem provides an explicit sufficient accuracy condition, while the sweep measures finite-horizon behavior on these instances.

Smaller coefficients allocate more of the nonforced budget to the best arm in these two environments. The default $1/6$ is fixed across the suite, and its FE guarantee does not require it to be an oracle

Table 6: Exploration-coefficient sensitivity with GK and FE fixed. Entries are rank-optimal pull fractions, mean ± standard deviation.

ρ	Pareto-5	Lognormal
$1/12$	0.9667 ± 0.0088	0.9775 ± 0.0005
$1/6$	0.9594 ± 0.0093	0.9766 ± 0.0008
$1/3$	0.9451 ± 0.0113	0.9747 ± 0.0012
$2/3$	0.9197 ± 0.0135	0.9703 ± 0.0020

confidence radius. Table 6 quantifies the finite-budget effect of additional index exploration without introducing a tail-dependent parameter choice.

D.5 Robust-mean baselines and complete tuning results

We compare exact FE and no-FE with Truncated-Mean UCB and Median-of-Means (MoM) UCB, using the robust estimators and radii of Bubeck et al. (2013) with explicit empirical radius multipliers. Pareto-5 and exponential arms provide shared rank and mean optimizers. Both mean and rank pseudo-regret are reported, computed by (26).

Estimators and information supplied. At global time t, an arm's rth observation contributes to the truncated estimate only if

$$X_r \le \left(\frac{ur}{2\log(\max\{2,t\})} \right)^{1/(1+\epsilon)}.$$

The sum of included observations is divided by the total arm count n, and the radius is

$$c\,4u^{1/(1+\epsilon)} \left(\frac{2\log(\max\{2,t\})}{n} \right)^{\epsilon/(1+\epsilon)}.$$

For MoM, let $L_t = 1/8 + 2\log(\max\{2,t\})$ and $k = \max\{1, \lfloor \min\{8L_t, n/2\} \rfloor\}$. Split the first $k\lfloor n/k \rfloor$ samples into k consecutive equal-sized blocks and take the median of their means. Its radius is

$$c\,(12v)^{1/(1+\epsilon)} (16L_t/n)^{\epsilon/(1+\epsilon)}.$$

The index is infinite at zero samples for truncation and at $n < 2$ for MoM. At an even number of blocks the implementation averages the two central block means. Equal indices use the separate tie-breaking stream.

The robust baselines receive distribution-derived moment bounds. For Pareto-5 we use $(\epsilon, u) = (0.05, 22)$ and $(\epsilon, v) = (0.05, 27.3262809401)$; for exponential arms we use $(1, 2/1.1^2)$ and $(1, 1/1.1^2)$. Here u bounds the raw $(1+\epsilon)$ moment and v bounds the centered moment. These values are not supplied to RANK-UCB1. Each radius multiplier c is selected separately by minimizing mean pseudo-regret at 50,000 on 100 pilot trials, using the complete 19-value grid in Table 9. Thus four method–instance pairs require 76 pilot configurations and 7,600 pilot trials. Formal evaluation uses 1,000 fresh trials with disjoint seeds.

Table 7: Selected robust configurations and the cost of unscaled radii. Pilot regret is mean pseudo-regret at $T = 50{,}000$, averaged over 100 trials.

Environment	Estimator	Selected c	Selected pilot regret	$c = 1$ pilot regret
Pareto-5	Truncated	0.20	2366.9	199976.4
Pareto-5	MoM	0.04	3487.0	318673.9
Exponential	Truncated	0.20	329.6	6941.2
Exponential	MoM	0.075	299.7	12820.2

The selected robust methods achieve the smallest average regret in these oracle-informed comparisons. Their tuning is substantial: for example, unscaled truncated UCB has pilot mean regret about 200,000

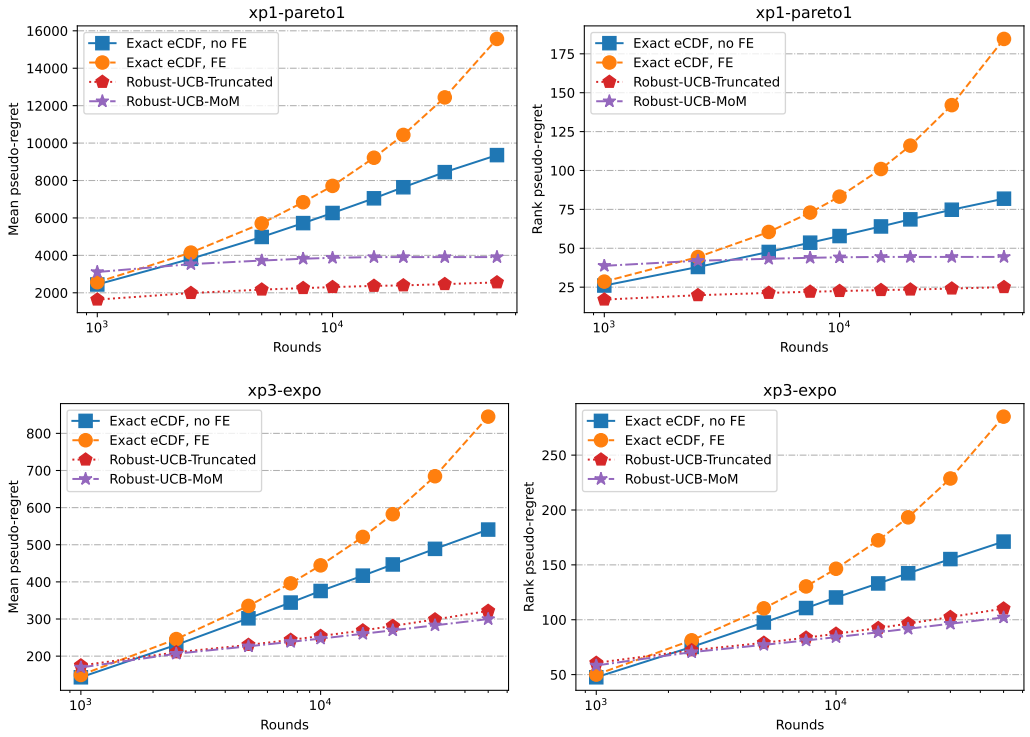


Figure 4: Robust-mean comparisons on shared objectives: Pareto-5 (top) and exponential arms (bottom), with mean pseudo-regret (left) and rank pseudo-regret (right). Each curve averages 1,000 formal trials. Robust baselines use oracle moment bounds and separately tuned radii; rank variants share a fixed $\rho = 1/6$.

Table 8: Shared-objective regret at $T = 50{,}000$, mean $\pm$ standard deviation over 1,000 formal trials.

Environment	Algorithm	Mean pseudo-regret	Rank pseudo-regret
Pareto-5	Exact, no FE	9357.3 ± 4045.3	81.88 ± 27.49
Pareto-5	Exact, FE	15566.2 ± 3061.7	184.57 ± 18.76
Pareto-5	Truncated UCB	2555.3 ± 2663.1	25.00 ± 18.60
Pareto-5	MoM UCB	3908.1 ± 5547.3	44.44 ± 38.66
Exponential	Exact, no FE	540.9 ± 127.0	171.21 ± 38.20
Exponential	Exact, FE	845.1 ± 99.1	285.08 ± 28.89
Exponential	Truncated UCB	321.7 ± 77.7	110.16 ± 22.75
Exponential	MoM UCB	299.8 ± 72.1	102.09 ± 20.90

on Pareto-5, compared with 2,367 at the selected scale. The selected multipliers are below one and are empirical settings rather than theorem-certified confidence radii. The rank methods use the same settings across all distributions and require neither these moment values nor an instance-specific search. Exact no-FE reduces the finite-horizon allocation cost relative to FE; FE supplies the general expected-regret guarantee and lower across-trial variability documented by the ablations. Reporting both variants and the full tuning effort makes their statistical and information requirements explicit.

Table 9: Complete radius-multiplier grid: mean pseudo-regret at $T = 50{,}000$ on 100 independent pilot trials per configuration. The formal trials use disjoint seeds.

	Pareto-5		Exponential	
c	Truncated	MoM	Truncated	MoM
0.0003	300341.9	122542.6	19397.9	5150.9
0.001	300341.7	122542.5	19355.1	4905.5
0.003	295727.1	110344.7	19147.0	4600.2
0.01	290086.1	77045.9	18527.2	3587.8
0.02	269160.7	20934.0	18557.5	2384.3
0.03	244479.0	7164.4	15306.3	1283.7
0.04	198708.4	3487.0	12494.2	699.2
0.05	192962.4	4299.1	9928.5	421.2
0.075	165507.9	21573.0	8714.5	299.7
0.1	140442.3	67575.1	5114.2	509.9
0.15	5146.6	152942.9	1314.7	1243.2
0.2	2366.9	206355.9	329.6	2234.6
0.25	7186.0	237420.3	594.2	3338.0
0.3	16122.9	257932.5	938.1	4516.6
0.35	28132.1	272105.1	1318.1	5645.7
0.4	43740.3	281851.0	1725.4	6661.2
0.5	76556.7	294804.6	2585.3	8427.7
0.75	151258.4	311128.9	4843.8	11267.6
1	199976.4	318673.9	6941.2	12820.2

D.6 Infinite-mean rewards

All five arms are unit-minimum Pareto with indices $(0.5, 0.6, 0.7, 0.8, 1.0)$, so every arm has infinite mean. Arm 1 is stochastically dominant and maximizes the rank score. No finite-mean regret is defined or used. Table 10 and Figure 5 report allocation, population win-probability, and rank-regret. GK FE allocates 95.95% of the budget to the best arm and closely matches exact FE. This shows that the method's absence of reward moment assumptions is operational: learning continues in a regime where a finite mean cannot be the target.

Table 10: Infinite-mean Pareto results at $T = 50{,}000$; mean $\pm$ standard deviation over 1,000 trials.

Variant	Rank-optimal fraction	Population win-probability	Rank-regret
Exact, no FE	0.9725 ± 0.0104	0.58034 ± 0.00057	91.50 ± 28.29
Exact, FE	0.9598 ± 0.0077	0.57885 ± 0.00039	165.82 ± 19.40
GK, FE	0.9595 ± 0.0078	0.57884 ± 0.00039	166.55 ± 19.64

D.7 Reproduction and computation records

The supplementary package contains a self-contained `experiments/` directory with the C++ simulator, vendored GK implementation, runner, plotting and verification scripts, and compressed trial-level data. The exact eCDF uses an order-statistic tree and the GK adapter uses a deterministic summary with an exact insertion buffer. GK's external implementation and license are recorded in `third_party/UPSTREAM.md`; its adapter's score semantics are given in Appendix C.

The formal base seed is 20260730, and the pilot base seed is 30260730. Trial r, numbered from zero, starts from $\mathrm{splitmix64}(\text{base seed} + r)$; deterministic arm-specific seeds and a separate tie-breaking

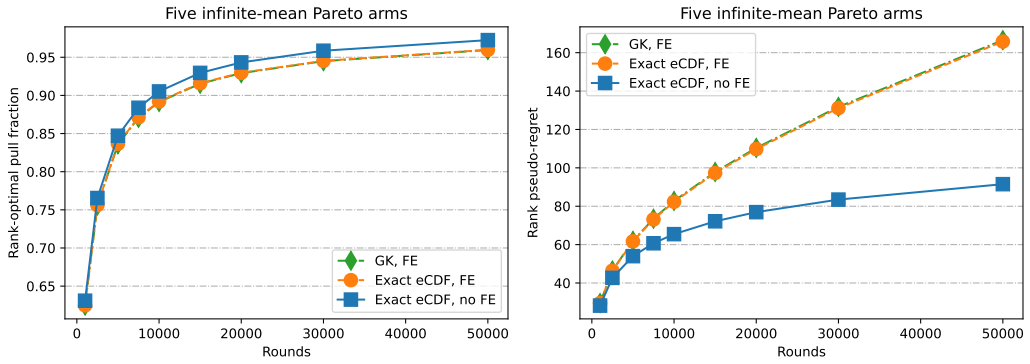


Figure 5: Learning with infinite-mean Pareto rewards: rank-optimal pull fraction and rank-regret. Each point averages 1,000 trials; all three methods receive the same per-arm reward streams.

seed are derived by the supplied simulator. Checkpoint rows contain every arm's pull count, so all reported objectives can be recomputed independently of the learner's surrogate scores. Configuration manifests record source hashes and experimental settings. The 19 main configurations, 76 pilot configurations, and 12 sensitivity configurations together comprise 1.93 billion one-observation rounds. The latter two types serve different purposes: only the robust pilot configurations select parameters for formal evaluation.

The recorded simulation environment is Windows with 12 logical CPUs, Python 3.14.6, GCC 16.1.0 with GNU libstdc++, and 10 concurrent main-suite jobs. Sensitivity studies use six concurrent jobs. Plot generation uses Python 3.11.4, NumPy 1.24.3, and Matplotlib 3.7.1. Timing is total simulation time per trial on the recorded host, so concurrent workload can affect it. Entry counts include internal data-structure copies and are not byte measurements.

From the supplementary `experiments/` directory, the commands are:

```
python run.py --suite all --jobs 10
python summarize.py
python plot.py
python run_sensitivity.py
g++ -std=c++17 -O2 verify.cpp -o verify
./verify
```

On Windows, the last command is `.\verify.exe`. The runner resumes complete trials; for a fresh reproduction, start from a copy without the generated `data/raw`, `data/pilot`, and `data/sensitivity` directories. Individual main suites are selected with `-suite benchmarks`, `ablations`, `robust`, or `infinite`. The sensitivity runner fixes the grids and validates agreement with default runs before changing one parameter at a time. Verification checks the schedule, CDF queries including duplicate values, and analytic population objectives; numerical checks supplement the proofs rather than replace them.