# THE PRICE OF TAIL SAFETY IN HEAVY-TAILED BANDITS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Traditional bandit algorithms focus on minimizing expected regret to achieve good average performance. However, low expected regret can mask tail risk—the possibility of exceptionally large regret on individual runs, which can be unacceptable in risk-sensitive applications. Under heavy-tailed rewards, extreme observations can distort reward estimates and lead to large regret. Additional exploration can make these estimates more reliable and reduce this risk, but at the cost of higher expected regret. We study this trade off under a known bound on the centered $(1+\varepsilon)$-th moment. Our target is pseudo-regret, which measures decision error: realized payoff shortfall also contains reward noise that can cause large losses even for an oracle. We establish matching upper and lower bounds relating a worst-case expected-regret budget to the pseudo-regret tail probability. For fixed $\varepsilon$ and a fixed number of arms, a budget of order $\rho$ times the minimax regret scale yields a $\rho$-fold improvement in the negative-log tail rate, up to constants. We therefore propose Budget-Calibrated Robust Racing (BCRR), a novel algorithm that jointly calibrates exploration and robust arm elimination to the expected-regret budget, attaining the optimal trade off between average performance and tail protection. Numerical experiments demonstrate the algorithm's effectiveness in controlling the risk of large decision errors.

## 1 INTRODUCTION

In repeated financial decisions, a small average shortfall need not rule out a bad run with a large cumulative loss (Huo & Fu, 2017). A trading desk choosing among execution or investment strategies may therefore care both about expected performance and about the probability of a severe shortfall. Stochastic multi-armed bandits provide a basic model of this repeated choice, but their standard expected-regret criterion does not characterize the upper tail of the regret distribution.

For bounded or sub-Gaussian rewards, regret-distribution results show how minimax-order expected regret can coexist with a rapidly decaying probability of linear regret (Simchi-Levi et al., 2022; 2025; Lee & Oh, 2026). Financial returns and trading costs, however, can have heavy tails (Cont, 2001). Under only finite moments, rare large observations affect both realized payoff and the reliability of the estimates used to choose actions.

These two effects must be separated. Realized shortfall can be large because the learner chooses suboptimal arms, or because reward noise is extreme even when it chooses an optimal arm. The latter effect already occurs for an oracle, preventing a uniformly light realized-shortfall tail over a finite-moment class. We therefore first control *pseudo-regret*, the loss from suboptimal-arm selection, and then include reward noise when translating the guarantee back to realized payoff loss.

Controlling decision errors also has a sampling cost. A finite-moment robust estimate supports an exponentially small failure probability only when its sample size is large enough. Eliminating uncertain arms early saves exploration but risks discarding the best arm; waiting for stronger evidence uses more of the regret budget. This leads to our question:

*Given a worst-case expected-regret budget, how small can the probability of a large decision error be under heavy-tailed rewards?*



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.

Our main result is a matched budget–tail frontier. Let $R_T^\star$ denote the heavy-tailed minimax expected-regret scale. For each fixed moment exponent $\varepsilon$ and number of arms $K$, an allowance of order $\rho R_T^\star$ is sufficient and necessary, up to constants, for a $\rho$-fold improvement in the negative-log tail rate at thresholds sufficiently above the budget and below a fixed fraction of the horizon. The threshold dependence has power $1/\varepsilon$. The result holds nonasymptotically under a known bound on the centered $(1+\varepsilon)$-th moment. We develop it in three contributions.

**Separating reward noise from decision errors.** We show that even always choosing the best arm can incur large losses because of reward noise. We quantify this risk and separate it from the losses caused by choosing suboptimal arms, identifying the part that exploration can control (Section 4).

**Matching bounds and an optimal policy.** We establish an achievable tail bound and a matching converse under a worst-case expected-regret budget, compared at a common absolute threshold (Section 5). We propose Budget-Calibrated Robust Racing (BCRR), a novel algorithm which attains this frontier by jointly calibrating its elimination radius and confidence parameter to the regret budget (Section 6). We also extend BCRR to unknown horizons while preserving the orders of both the expected-regret and tail guarantees.

**Quantile guarantees and efficiency limits.** We derive quantile guarantees and realized-shortfall bounds from the frontier (Section 7). We also prove that uniform tail protection cannot coexist with logarithmic regret on every fixed separated instance, establishing a limit on instance-dependent efficiency.

## 2 Related Work

Existing work supplies two sides of our question. Regret-distribution results study how average performance constrains tail behavior under light-tailed rewards, while heavy-tailed bandit results identify the expected-regret scale under finite moments. We connect these lines by asking for the finite-horizon regret-distribution frontier when robust estimation itself has a nonnegligible sampling cost. Appendix F compares the closest guarantees in detail.

**Regret-distribution frontiers.** Classical stochastic bandit theory is built around expected regret and high-probability regret bounds (Auer et al., 2002; Audibert & Bubeck, 2009; Bubeck & Cesa-Bianchi, 2012; Slivkins, 2019; Lattimore & Szepesvári, 2020). A distributional view instead asks for the probability of exceeding each regret threshold; early work studied deviations and typical behavior of standard policies (Salomon & Audibert, 2011; Fan & Glynn, 2022). Most closely related, Simchi-Levi et al. (2022; 2025) characterize tradeoffs between expected regret and regret-tail decay under light-tailed rewards, including the incompatibility between logarithmic instance-dependent regret and uniformly light tails. Related all-confidence guarantees have also been developed for bandits and reinforcement learning under a sub-exponential concentration assumption (Lee & Oh, 2026). These results provide the light-tailed analogue of our question. Under only a finite $(1+\varepsilon)$ moment, however, uniform empirical-regret bounds require a polynomial noise term, and an exponential-confidence robust estimate is available only when its confidence budget is supported by the sample size. The relevant frontier must therefore separate pseudo-regret from reward noise and incorporate this finite-moment sampling constraint.

**Finite-moment bandits and the frontier scale.** Heavy-tailed bandit work provides the robust estimation tools and the natural expected-regret scale for our frontier. Early methods establish minimax rates under finite moments (Bubeck et al., 2013; Wei & Srivastava, 2020), and later work gives instance-dependent or adaptive guarantees (Agrawal et al., 2021; Genalti et al., 2024; Tamás et al., 2024). These results identify the minimax expected-regret scale at which worst-case finite-moment exploration becomes possible. Expected-regret guarantees leave the survival curve associated with a larger or smaller allowance unspecified. Recent work on regret tails for generic reward classes instead targets asymptotic, instance-dependent KL optimality (Panda & Agrawal, 2026). Together these works supply estimation tools, expected-regret benchmarks, and fixed-instance tail results; they do not determine the finite-horizon minimax tail probability under a prescribed worst-case budget.

**Exploration and regret-tail control.** A common lesson across robust and tail-aware policy design is that additional exploration can prevent rare but persistent decision errors. This appears in work on fragility under risk-parameter misspecification (Fan & Glynn, 2025), in elimination and UCB-style designs built from confidence bounds (Auer, 2002; Even-Dar et al., 2006; Simchi-Levi et al., 2022; 2025), and in variance-inflated Thompson sampling for safety and robustness (Thompson, 1933; Agrawal & Goyal, 2012; Kaufmann et al., 2012; Zhu & Simchi-Levi, 2025). These policies motivate exploration as the resource used to control regret tails, but do not quantify its finite-moment minimax price. Our contribution is the matching dependence on the budget and regret threshold under a centered finite-moment bound, with the valid sample size for robust estimation built into the policy.

## 3 Problem Formulation and Tail Objectives

There are $K \geq 2$ arms and horizon $T \geq K$. Let $\mathcal{F}_{t-1}$ denote the observed history before round $t$. The learner chooses $A_t \in [K]$ and receives a fresh sample with conditional law $Y_t \mid (\mathcal{F}_{t-1}, A_t) \sim \nu_{A_t}$. Let $\mu_i = \mathbb{E}_{\nu_i} Y$, $\mu^\star = \max_i \mu_i$, $\Delta_i = \mu^\star - \mu_i$, and $N_i(T) = \sum_{t=1}^T \mathbf{1}\{A_t = i\}$. A policy may randomize and may use the known horizon and moment parameters.

We assume $\mu_i \in [0,1]$, but do not bound the rewards themselves. Instead, the centered $(1+\varepsilon)$ moment is uniformly bounded. This allows rare observations with unbounded magnitude while controlling their $(1+\varepsilon)$ moments uniformly.

**Assumption 1** (Moment-bounded heavy tails)**.** *For known $\varepsilon \in (0,1]$ and $v \geq 1$,*

$$\nu = (\nu_1, \ldots, \nu_K) \in \mathcal{H}_K(v, \varepsilon) \quad \Longleftrightarrow \quad \mu_i \in [0,1] \text{ and } \mathbb{E}_{\nu_i}\left[|Y - \mu_i|^{1+\varepsilon}\right] \leq v \quad \forall i \in [K].$$

The class includes both light-tailed and heavy-tailed distributions; no exponential-moment assumption is imposed. For $v < 1$, using $1 \vee v$ preserves the upper guarantees but loses the refinement available at small noise scales. We use the same class definition for $K = 1$ in the oracle construction of Section 4.

The natural benchmark $T\mu^\star$ is the cumulative return if every round yielded the mean of an optimal arm. Two ways of measuring shortfall relative to this benchmark will be used. Pseudo-regret records only decision error, the loss caused by the chosen actions.

$$\mathsf{Reg}_T = \sum_{t=1}^T (\mu^\star - \mu_{A_t}) = \sum_{i=1}^K N_i(T)\Delta_i.$$

Empirical regret, the realized mean-benchmark shortfall, also contains payoff noise.

$$\widehat{\mathsf{Reg}}_T = T\mu^\star - \sum_{t=1}^T Y_t = \mathsf{Reg}_T - M_T, \qquad M_T = \sum_{t=1}^T (Y_t - \mu_{A_t}).$$

Under the present moment assumption, $M_T$ may be heavy-tailed and need not satisfy a uniform exponential tail bound. We therefore target $\mathsf{Reg}_T$. When a realized-payoff guarantee is required, $M_T$ is included as a separate noise term. Empirical regret is not the coupled realized regret that subtracts a simultaneously sampled best-arm reward sequence, under which the oracle's noise can cancel.

Our primary object is the upper tail of pseudo-regret. For a constant $\gamma \in (0,1)$, the event $\{\mathsf{Reg}_T > \gamma T\}$ means that decision errors have accumulated a constant fraction of the largest possible regret scale. More generally, we study the worst-case survival probability

$$\sup_{\nu \in \mathcal{H}_K(v,\varepsilon)} \mathbb{P}_\nu(\mathsf{Reg}_T > x)$$

for every threshold $x$. Equivalently, for a failure probability $\delta \in (0,1)$, one may ask for the $(1-\delta)$-quantile $Q_{1-\delta}(\mathsf{Reg}_T)$.

The natural budget unit is the heavy-tailed minimax expected-regret scale before saturation at $T$ (Bubeck et al., 2013; Wei & Srivastava, 2020):

$$\alpha_\varepsilon = \frac{\varepsilon}{1+\varepsilon}, \qquad R_T^\star(v, \varepsilon, K) = v^{1/(1+\varepsilon)} K^{\alpha_\varepsilon} T^{1/(1+\varepsilon)}.$$

This scale corresponds to the balance between exploration and finite-moment estimation error. If each arm is sampled about $T/K$ times, finite-moment uncertainty is of order $v^{1/(1+\varepsilon)}(K/T)^{\varepsilon/(1+\varepsilon)}$; multiplying this unresolved gap by the horizon gives $R_T^\star$. Since $\mathsf{Reg}_T \le T$, the corresponding minimax scale over all parameter regimes is $T \wedge R_T^\star$; our frontier concerns the nontrivial regime $R_T^\star \ll T$. We write $B_T$ for a policy's worst-case expected pseudo-regret allowance, meaning

$$\sup_{\nu \in \mathcal{H}_K(v,\varepsilon)} \mathbb{E}_\nu[\mathsf{Reg}_T] \le B_T.$$

The budget–tail question is how small the survival probability can be when $B_T$ is fixed. For fixed $K$, we use $\rho \ge 1$ to index allowances of order $B_T \asymp \rho R_T^\star$. Section 5 gives the additional logarithmic term when $K$ grows.

## 4 Unavoidable Empirical-Regret Noise

After defining the tail objectives, we first show that the realized shortfall in Section 3 cannot itself have a uniformly light tail. The obstruction already appears in the oracle case: one extreme negative reward can dominate the realized payoff over the remaining trajectory.

**Theorem 1** (Oracle lower bound for empirical regret)**.** *Fix $\varepsilon \in (0,1]$ and $v \ge 1$. There are universal constants $c, C > 0$ such that for every integer $T \ge 1$ and every $x \ge Cv^{1/(1+\varepsilon)}T^{1/(1+\varepsilon)}$, one can construct a one-arm instance $\nu \in \mathcal{H}_1(v,\varepsilon)$ for which the oracle policy satisfies*

$$\mathbb{P}_\nu(\widehat{\mathsf{Reg}}_T > x) = \mathbb{P}_\nu(-M_T > x) \ge c\,\frac{vT}{x^{1+\varepsilon}}.$$

*Consequently no bandit policy can guarantee an exponential empirical-regret tail uniformly over $\mathcal{H}_K(v,\varepsilon)$.*

The proof is in Appendix A; identical copies embed the one-arm construction into any $K$-arm class. The instance may depend on $(T,x)$: the theorem rules out a uniform exponential bound. It concerns the mean-benchmark shortfall in Section 3, for which the oracle still incurs noise.

This obstruction does not preclude meaningful pseudo-regret-tail guarantees. The following decomposition shows that pseudo-regret is the controllable component. Its tail transfers to empirical regret after including the same polynomial noise term.

**Proposition 1** (From pseudo-regret to realized shortfall)**.** *For any $\nu \in \mathcal{H}_K(v,\varepsilon)$, any adaptive policy, and any $x > 0$,*

$$\mathbb{P}_\nu(\widehat{\mathsf{Reg}}_T > x) \le \mathbb{P}_\nu(\mathsf{Reg}_T > x/2) + C_\varepsilon\,\frac{vT}{x^{1+\varepsilon}}.$$

The proof is given in Appendix A.

The noise term has the same order as the oracle lower bound in its stated range. We next characterize the pseudo-regret frontier, then use this decomposition in Section 7 to recover realized-shortfall guarantees.

## 5 The Budget–Tail Frontier

We give the tail guarantee and its expected-regret cost for BCRR, whose construction appears in Section 6, and then a converse for any policy with a prescribed budget. To compare the bounds, we distinguish an excess threshold $y$ above BCRR's exploration baseline from an absolute regret threshold $z$. The match holds at thresholds sufficiently above the budget and below a fixed fraction of $T$.

**Theorem 2** (Budget-calibrated pseudo-regret tail)**.** *Fix $T \ge K \ge 2$, $v \ge 1$, $\varepsilon \in (0,1]$, $\rho \ge 1$, and $\tau > 2$. Let $\pi_{\rho,\tau}$ be the BCRR policy in Algorithm 1, using known $(T,v,\varepsilon)$. There are constants $c_{\varepsilon,\tau}, C_{\varepsilon,\tau} > 0$, depending only on $(\varepsilon,\tau)$, such that, for every $y > 0$,*

$$\sup_{\nu \in \mathcal{H}_K(v,\varepsilon)} \mathbb{P}_\nu\big(\mathsf{Reg}_T^{\pi_{\rho,\tau}} > C_{\varepsilon,\tau}\rho R_T^\star + y\big) \le C_{\varepsilon,\tau} K \exp\left\{-c_{\varepsilon,\tau}\rho\left(\frac{y}{R_T^\star}\right)^{1/\varepsilon}\right\}.$$

The proof is in Appendix B. The bound separates the exploration baseline $C_{\varepsilon,\tau}\rho R_T^\star$ from deviations beyond it. Increasing $\rho$ enlarges this baseline and strengthens the bound at a fixed excess threshold $y$. It does not assert that tail probabilities at a fixed absolute threshold decrease monotonically with $\rho$ on every instance.

**Theorem 3** (Expected-regret cost of tail protection)**.** *Under the assumptions of Theorem 2, the same policy satisfies*

$$\sup_{\nu\in\mathcal{H}_K(v,\varepsilon)} \mathbb{E}_\nu[\mathsf{Reg}_T^{\pi_{\rho,\tau}}] \le C_{\varepsilon,\tau}\left(\rho R_T^\star + K\right) + C_{\varepsilon,\tau} R_T^\star \rho^{-\varepsilon}\left[1+\log^\varepsilon(C_{\varepsilon,\tau}K)\right].$$

This follows by integrating the tail bound; see Appendix B. The first term pays for exploration, and the second bounds the expected contribution of residual deviations. For fixed $K$, the total is $O_{\varepsilon,\tau,K}(\rho R_T^\star)$, so a constant $\rho$ gives a minimax-order expected-regret guarantee. For growing $K$, the same budget order holds when $\rho^{1+\varepsilon} \gtrsim 1+\log^\varepsilon K$; otherwise the displayed logarithmic term remains. Here $K \le R_T^\star$ because $T \ge K$ and $v \ge 1$.

The converse shows that this dependence on the budget is unavoidable: any policy with a prescribed expected-regret allowance faces a corresponding limit on its worst-case tail protection.

**Theorem 4** (Converse under an expected-regret budget)**.** *Fix $T \ge K \ge 2$, $v \ge 1$, and $\varepsilon \in (0,1]$. Let $\pi$ be any policy such that $\sup_{\nu\in\mathcal{H}_K(v,\varepsilon)} \mathbb{E}_\nu[\mathsf{Reg}_T^\pi] \le B_T$ for some $B_T \ge R_T^\star$. There are constants $c_\varepsilon, C_\varepsilon > 0$, depending only on $\varepsilon$, such that, for every $z \in [C_\varepsilon B_T, c_\varepsilon T]$,*

$$\sup_{\nu\in\mathcal{H}_K(v,\varepsilon)} \mathbb{P}_\nu(\mathsf{Reg}_T^\pi > z) \ge \exp\left\{-C_\varepsilon \frac{B_T}{R_T^\star}\left(\frac{z}{R_T^\star}\right)^{1/\varepsilon}\right\}.$$

**Proof intuition.** A small expected-regret budget leaves some suboptimal arm lightly sampled on average. The moment constraint permits making this arm optimal by increasing the frequency of rare, large rewards, while keeping the two instances difficult to distinguish. Limited exploration then leaves the policy vulnerable to persistently missing the best arm. Choosing the gap appropriately turns this failure into regret exceeding $z$. A larger budget buys more observations and hence more evidence, explaining its linear contribution to the tail exponent. Appendix C.2 formalizes this argument; the least favorable instance may depend on $z$.

**Comparing the bounds at the same threshold.** For $z \ge 2C_{\varepsilon,\tau}\rho R_T^\star$, set $y = z - C_{\varepsilon,\tau}\rho R_T^\star \ge z/2$ in Theorem 2. After adjusting constants, this gives

$$\sup_{\nu\in\mathcal{H}_K(v,\varepsilon)} \mathbb{P}_\nu(\mathsf{Reg}_T^{\pi_{\rho,\tau}} > z) \le C_{\varepsilon,\tau} K \exp\left\{-c_{\varepsilon,\tau}\rho\left(\frac{z}{R_T^\star}\right)^{1/\varepsilon}\right\}. \tag{1}$$

For each fixed $\varepsilon$ and $K$, Theorem 3 permits a budget $B_T$ of order $\rho R_T^\star$. Substituting that order into Theorem 4 gives the same exponent as (1), up to constants. Thus the bounds match in their dependence on $T$, $v$, and the budget multiplier over their common range $C\rho R_T^\star \le z \le cT$. The constants may depend on $\varepsilon$; the comparison is not uniform as $\varepsilon \downarrow 0$. It is informative only when this threshold range is nonempty.

**Linear regret.** For any fixed $\gamma \in (0,1)$ with $2C_{\varepsilon,\tau}\rho R_T^\star \le \gamma T$, taking $z = \gamma T$ in (1) yields the rate $\exp\{-c_{\varepsilon,\tau,\gamma}\rho(T/R_T^\star)^{1/\varepsilon}\}$, up to the prefactor. At fixed $(K, v, \varepsilon, \rho, \tau)$ and sufficiently large $T$, the converse matches this rate for every sufficiently small fixed $\gamma > 0$. In particular, at a constant minimax budget the optimal worst-case linear regret probability is $\exp\{-\Theta(T^{1/(1+\varepsilon)})\}$ in that range.

The dependence on $\varepsilon$ should be read together with the budget scale. As $\varepsilon$ decreases, the moment assumption becomes weaker and the permitted minimax-order budget grows faster with $T$. Keeping $\rho$ fixed therefore allows more exploration under the weaker assumption. The faster power of $T$ in the linear-regret tail rate is obtained alongside this larger expected-regret allowance. Comparisons across moment exponents consequently describe different exploration budgets; they do not show that heavier-tailed rewards are intrinsically easier to learn from.

## 6 BUDGET-CALIBRATED ROBUST RACING

BCRR constructs the policy used in Section 5. It samples surviving arms in phases and eliminates an arm only when its robust estimate is sufficiently below the active leader. The design has two requirements: keep each arm's exploration cost within its regret budget, and wait until the sample size supports the confidence level used for elimination.

The per-arm regret scale is

$$q_T = R_T^\star/K = v^{1/(1+\varepsilon)}(T/K)^{1/(1+\varepsilon)}.$$

For an arm with gap $\Delta$, sampling on the scale $\rho q_T/\Delta$ incurs regret on the scale $\rho q_T$. This motivates an elimination radius proportional to $\rho q_T/n$ after $n$ samples.

**Robust estimation.** BCRR uses a median-of-means estimate $\widehat{\mu}_{n,u}$, where $u$ controls the failure probability. The following standard finite-moment bound is proved with explicit constants in Appendix B.

**Proposition 2** (Phase-tuned robust mean)**.** *Let $p = 1+\varepsilon$. There are constants $c_\varepsilon, C_\varepsilon, \bar{c}_\varepsilon > 0$ and a median-of-means estimator $\widehat{\mu}_{n,u}$ such that, for i.i.d. samples with mean $\mu$ and centered pth moment at most $v$, whenever $1 \le u \le \bar{c}_\varepsilon n$,*

$$\mathbb{P}\left(|\widehat{\mu}_{n,u} - \mu| > C_\varepsilon v^{1/p}\left(\frac{u}{n}\right)^{(p-1)/p}\right) \le 2e^{-c_\varepsilon u}.$$

*One admissible construction uses $\bar{c}_\varepsilon = 1/32$ and $C_\varepsilon = (16\,18^{p-1})^{1/p}$.*

For fixed $n$, increasing $u$ lowers the failure bound and enlarges the estimation radius. The condition $u \le \bar{c}_\varepsilon n$ gives the sample size needed to use the chosen confidence parameter.

**Choice of the confidence parameter.** We choose the confidence parameter by matching the estimation radius to $\rho q_T/n$. The following proposition gives its expression and the sample size required for its validity.

**Proposition 3** (Confidence parameter and sample size)**.** *Let $p = 1+\varepsilon$, $\alpha = (p-1)/p$, and $r(n,u) = v^{1/p}(u/n)^\alpha$. Solving $r(n,u) = \rho q_T/n$ gives*

$$u_{\rho,T}(n) = \rho^{(1+\varepsilon)/\varepsilon}\left(\frac{T}{Kn}\right)^{1/\varepsilon}.$$

*The validity condition $u_{\rho,T}(n) \le \bar{c}_\varepsilon n$ is equivalent to*

$$n \ge \bar{c}_\varepsilon^{-\varepsilon/(1+\varepsilon)}\rho(T/K)^{1/(1+\varepsilon)}.$$

The proof is given in Appendix B.2. Set $A_\varepsilon = C_\varepsilon$ from Proposition 2 and $a_{\rho,T}(n) = A_\varepsilon \rho q_T/n$. Then $a_{\rho,T}(n) = C_\varepsilon v^{1/p}(u_{\rho,T}(n)/n)^\alpha$ is exactly the radius supported by that proposition. BCRR samples without elimination until the sample-size condition above holds. It then clips estimates to $[0,1]$ and removes arms more than $\tau a_{\rho,T}(n)$ below the estimated leader, with $\tau > 2$.

The confidence parameter $u_{\rho,T}(n)$ decreases with $n$ because elimination errors become less costly in later phases. Early mistakes can leave the policy selecting a clearly suboptimal arm for many rounds. Provided earlier elimination steps are accurate, the surviving gaps shrink as sampling proceeds, so later errors can only affect choices among increasingly near-optimal arms. The algorithm can therefore relax its confidence requirement as the potential regret cost declines. Meanwhile, the radius $a_{\rho,T}(n)$ keeps shrinking: the weaker probability guarantee applies to an increasingly stringent accuracy target. The stronger early confidence requirements must still be supported by the sample size, which is why BCRR postpones elimination while a phase is invalid.

**Dependence on the regret budget.** On accurate phases, an optimal arm survives and every arm with $\Delta > (\tau+2)a_{\rho,T}(n)$ is removed. For gaps large enough to permit elimination within the horizon, the elimination sample size is on the scale $n_\rho(\Delta) \asymp_{\varepsilon,\tau} \rho q_T/\Delta$. Substituting this scale into the confidence parameter gives

$$u_{\rho,T}(n_\rho(\Delta)) \asymp_{\varepsilon,\tau} \rho^{(1+\varepsilon)/\varepsilon}\left(\frac{T\Delta}{K\rho q_T}\right)^{1/\varepsilon} = \rho\left(\frac{T\Delta}{R_T^\star}\right)^{1/\varepsilon}.$$

Thus both the exploration cost and the tail exponent grow linearly with $\rho$, up to constants. Arms too close to optimal for this elimination argument contribute at most order $\rho R_T^\star$ over the horizon on the good event. To control an additional regret threshold $y$, the proof tracks accurate phases until all remaining gaps are of order $y/T$. Summing the earlier dyadic phase failures gives Theorem 2; integrating its bound gives Theorem 3.

Algorithm 1 summarizes the sampling and elimination steps.

**Algorithm 1** Budget-Calibrated Robust Racing (BCRR)

**Input:** horizon $T$, number of arms $K$, moment parameters $(v, \varepsilon)$, budget parameter $\rho \geq 1$, threshold $\tau > 2$
**Constants:** let $\bar{c}_\varepsilon, C_\varepsilon$ be as in Proposition 2; set $A_\varepsilon = C_\varepsilon$
$q_T \leftarrow v^{1/(1+\varepsilon)}(T/K)^{1/(1+\varepsilon)}$
$L_T \leftarrow \lfloor T/K \rfloor$, $A \leftarrow [K]$, $n_0 \leftarrow 1$, $s \leftarrow 0$
Pull every arm once, unless the horizon ends
**if** the horizon has ended **then**
    stop
**end if**
**while** time remains, $|A| > 1$, and $n_s \leq L_T$ **do**
    $u_s \leftarrow \rho^{(1+\varepsilon)/\varepsilon}(T/(Kn_s))^{1/\varepsilon}$, $a_s \leftarrow A_\varepsilon \rho q_T / n_s$
    **if** $u_s \leq \bar{c}_\varepsilon n_s$ **then**
        **for** each active arm $i \in A$ **do**
            compute $\widehat{\mu}_{i,s} \leftarrow \Pi_{[0,1]}(\widehat{\mu}_{n_s,u_s})$ from the first $n_s$ rewards of arm $i$
        **end for**
        $M_s \leftarrow \max_{j \in A} \widehat{\mu}_{j,s}$
        $A \leftarrow \{i \in A : \widehat{\mu}_{i,s} \geq M_s - \tau a_s\}$
    **end if**
    **if** $n_s = L_T$ **then**
        **break**
    **end if**
    $n_{s+1} \leftarrow \min\{2n_s, L_T\}$
    Pull each active arm $i \in A$ until $N_i = n_{s+1}$, unless the horizon ends
    $s \leftarrow s + 1$
**end while**
**for** all remaining rounds **do**
    pull an arm in $A$ with largest last robust estimate, or an arbitrary arm in $A$ if no valid estimate has been computed
**end for**

For $\varepsilon = 1$, the radius is proportional to $\rho\sqrt{vT/K}/n$, the finite-variance analogue of the early radius in Simchi-Levi et al. (2022).

**Unknown-horizon extension.** For an unknown horizon, we restart BCRR in epochs of lengths $K, 2K, 4K, \ldots$. Each epoch uses its own length as the local horizon and starts with all arms active and no previous observations. The resulting policy retains the expected-regret and tail guarantees, up to constants.

**Corollary 1** (Anytime BCRR)**.** *Fix $K \geq 2$, known $v \geq 1$ and $\varepsilon \in (0, 1]$, $\rho \geq 1$, and $\tau > 2$. The restarted policy $\pi^{\mathrm{any}}_{\rho,\tau}$ does not use the terminal horizon. For every deterministic $T \geq K$, it satisfies the tail and expectation bounds of Theorems 2 and 3, with constants depending only on $(\varepsilon, \tau)$ and no additional $\log T$ factor.*

**Remark.** With geometrically increasing epoch lengths, the total regret budget remains comparable to that of the latest epoch. Each round adds nonnegative pseudo-regret, so stopping partway through an epoch cannot exceed the regret of its full run. The active epoch's horizon is within a constant factor of $T$, which preserves the regret scale at intermediate terminal times. In the tail analysis, earlier epochs receive smaller shares of the excess-regret threshold, chosen to be increasingly generous

relative to their own regret scales. Their contributions to the tail bound then decay fast enough to avoid an extra $\log T$ factor. Appendix D.1 gives the pseudocode and proof.

## 7 Consequences and Limits of Tail Safety

The frontier gives a family of policies indexed by an exploration allowance. We now explain how its guarantee reads at a chosen confidence level, how it translates to realized payoff loss, and which form of instance-dependent efficiency is incompatible with uniform tail protection.

**Quantiles and the choice of $\rho$.** Inverting Theorem 2 gives, for each fixed $\rho \geq 1$ and every $\delta \in (0,1)$,

$$\sup_{\nu \in \mathcal{H}_K(v,\varepsilon)} Q_{1-\delta}(\mathsf{Reg}_T^{\pi_{\rho,\tau}}) \lesssim_{\varepsilon,\tau} R_T^\star \left[\rho + \rho^{-\varepsilon} \log^{\varepsilon}(K/\delta)\right].$$

The first term is the exploration baseline; the second is the additional allowance for rare decision errors. Increasing $\rho$ raises the first and lowers the second, so it need not improve the total quantile bound. For a prespecified target $\delta$, balancing them suggests $\rho \asymp \max\{1, \log^{\varepsilon/(1+\varepsilon)}(K/\delta)\}$, when allowed by the expected-regret budget and the nontrivial threshold range. This selects a policy for that target; the displayed bound holds for all $\delta$ once $\rho$ is fixed. Corollary 5 in Appendix D gives the formal bounds and a matching quantile converse at a constant minimax budget.

**Using the frontier.** With a fixed expected-regret allowance, use Theorem 3 to identify admissible values of $\rho$ and compare their quantile bounds at the confidence level of interest. If instead the aim is to minimize the bound for a prespecified $\delta$, start with the balancing choice above and evaluate its expected-regret cost. This makes the exploration cost of the chosen protection explicit. In either case, choose $\rho$ before the run; $\delta$ guides tuning and evaluation but is not an input to BCRR.

The preceding quantile concerns action selection. Realized shortfall also contains the martingale noise $M_T$. Combining Theorem 2 with Proposition 1 yields the following guarantee.

**Corollary 2** (Realized mean-benchmark shortfall tail)**.** *Under the assumptions of Theorem 2, for every $z > 0$,*

$$\sup_{\nu \in \mathcal{H}_K(v,\varepsilon)} \mathbb{P}_\nu(\widehat{\mathsf{Reg}}_T > z) \leq C_{\varepsilon,\tau} K \exp\left\{-c_{\varepsilon,\tau}\rho\left(\frac{(z - C_{\varepsilon,\tau}\rho R_T^\star)_+}{R_T^\star}\right)^{1/\varepsilon}\right\} + C_{\varepsilon,\tau}\frac{vT}{z^{1+\varepsilon}}.$$

Thus a realized-shortfall quantile bound adds a noise allowance of order $(vT/\delta)^{1/(1+\varepsilon)}$ to the pseudo-regret bound. This term can dominate at small $\delta$ and cannot be removed uniformly over the moment class, by Theorem 1. The conclusion concerns worst-case guarantees; it does not assert that every instance has polynomial payoff tails.

**The cost in instance-dependent efficiency.** The budget guarantee in Section 5 is worst-case. It does not promise logarithmic regret on each fixed instance with separated means. The following consequence of a separate change-of-measure argument shows why that stronger requirement cannot accompany uniform tail protection.

**Corollary 3** (Incompatibility with minimax tail safety)**.** *Fix $K \geq 2$, $v \geq 1$, and $\varepsilon \in (0,1]$. No policy family can satisfy both:*

*(i) $\mathbb{E}_\nu[\mathsf{Reg}_T] = O_\nu(\log T)$ on every fixed $\nu \in \mathcal{H}_K(v,\varepsilon)$ with a unique optimal arm;*

*(ii) for constants $c, C, c_0, C_0 > 0$ independent of $\nu$ and $T$,*

$$\sup_{\nu \in \mathcal{H}_K(v,\varepsilon)} \mathbb{P}_\nu(\mathsf{Reg}_T > z) \leq C\exp\left\{-c(z/R_T^\star)^{1/\varepsilon}\right\}$$

*for every $z \in [C_0 R_T^\star, c_0 T]$ and all sufficiently large $T$.*

The proof is in Appendix D. Its key step, Theorem 5, constructs two fixed instances differing in one arm. Logarithmic sampling of that arm on the base instance leaves only $O(\log T)$ KL information. On the alternative, the probability of linear pseudo-regret is then at least $cT^{-C}$, which eventually

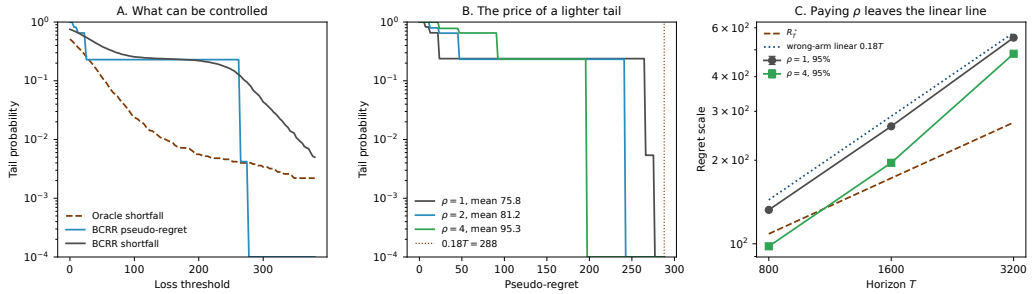



Figure 1: Two-arm instance with gap $0.18$. Panel A shows survival functions of oracle empirical regret, BCRR$(\rho = 1)$ pseudo-regret, and BCRR empirical regret on a common reward table. Panel B shows raw pseudo-regret survival for $\rho \in \{1, 2, 4\}$, with mean cost in the legend. Panel C compares the $95\%$ pseudo-regret quantile with $R_T^\star$ and the line $0.18T$.

exceeds any bound of order $\exp\{-c'T^{1/(1+\varepsilon)}\}$. The same qualitative incompatibility is known for light-tailed rewards (Simchi-Levi et al., 2022; 2025). BCRR avoids the conflict by allowing more exploration even on separated instances. Its minimax budget and uniform tail guarantee therefore do not imply logarithmic instance-dependent regret.

## 8 Experiments

The experiments illustrate three finite-horizon observations: realized shortfall retains an oracle noise tail; on this instance, increasing $\rho$ reduces the far upper tail of pseudo-regret at a higher mean cost; and a high pseudo-regret quantile can be compared with $R_T^\star$ and the lock-on line $0.18T$.

The instance has two arms with means $(.58, .40)$ and symmetric Pareto noise of index $1.7$. A run that persistently selects the suboptimal arm accumulates regret at rate $0.18$ per round. The main horizon is $T = 1600$, with $\varepsilon = .5$ and $v = 1$. Panels A–B use $5{,}000$ coupled replications; panel C uses $2{,}500$ replications for each $T \in \{800, 1600, 3200\}$. Random seeds are shared across $\rho$ and, in panel A, across the oracle and BCRR on the same reward table. The implementation uses radius multiplier $.08$, validity multiplier $1$, and $\tau = 2.2$.

Figure 1 supports three observations. First, Panel A serves as a control: the oracle already has a heavy realized tail, so large realized shortfall does not by itself imply a decision error. BCRR improves the pseudo-regret tail, but it cannot remove the reward-noise tail. Second, Panel B varies $\rho$ on this instance. As $\rho$ increases from $1$ to $4$, the mean pseudo-regret rises from $75.8$ to $95.3$, and the upper tail sits farther below the lock-on line $0.18T$. Third, Panel C tracks the $95\%$ quantile against $T$: larger $\rho$ keeps it farther from the lock-on line $0.18T$. Appendix G reports a budget-matched UCB comparison, moment-proxy sensitivity, and a check using empirical Fama–French residuals in place of Pareto shocks.

## 9 Conclusion and Limitations

We characterize how a worst-case expected-regret allowance limits the achievable upper tail of decision-induced regret under finite-moment rewards. BCRR attains the upper bound by coupling its elimination radius to a sample-supported confidence budget. The converse matches the exponent at absolute regret thresholds sufficiently above the exploration allowance and below a fixed fraction of the horizon.

The match holds for fixed $\varepsilon$ and $K$ in the stated budget and threshold range; growing $K$ can add a logarithmic term to the expected-regret bound. The lower-bound instance may depend on the threshold. Geometric restarts remove the need for a known horizon, while moment parameters remain known; the conditional extension in Appendix E does not remove this latter requirement on the full moment class. Realized shortfall retains a worst-case noise floor, and uniform tail protection precludes logarithmic regret on every fixed separated instance.

## REFERENCES


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

Shubhada Agrawal, Sandeep K. Juneja, and Wouter M. Koolen. Regret minimization in heavy-tailed bandits. In Mikhail Belkin and Samory Kpotufe (eds.), *Proceedings of the Thirty Fourth 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.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Annual Conference on Learning Theory*, pp. 217–226, 2009. URL https://www.learningtheory.org/colt2009/papers/022.pdf.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002. URL https://www.jmlr.org/papers/v3/auer02a.html.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256, 2002. doi: 10.1023/A:1013689704352.

Lilian Besson and Émilie Kaufmann. What doubling tricks can and can't do for multi-armed bandits, 2018. URL https://arxiv.org/abs/1803.06971.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012. doi: 10.1561/2200000024.

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.

Rama Cont. Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236, 2001. doi: 10.1080/713665670.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7:1079–1105, 2006. URL https://www.jmlr.org/papers/v7/evendar06a.html.

Lin Fan and Peter W. Glynn. The typical behavior of bandit algorithms, 2022. URL https://arxiv.org/abs/2210.05660.

Lin Fan and Peter W. Glynn. The fragility of optimized bandit algorithms. *Operations Research*, 73(6):3173–3198, 2025. doi: 10.1287/opre.2023.0283. URL https://arxiv.org/abs/2109.13595.

Kenneth R. French. Kenneth r. french data library. https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html, 2026. Daily portfolio return data; website accessed September 26, 2026.

Gianmarco Genalti, Lupo Marsigli, Nicola Gatti, and Alberto Maria Metelli. $(\epsilon, u)$-adaptive regret minimization in heavy-tailed bandits. In *Proceedings of the Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 1882–1915. PMLR, 2024. URL https://arxiv.org/abs/2310.02975.

Xiaoguang Huo and Feng Fu. Risk-aware multi-armed bandit problem with application to portfolio selection. *Royal Society Open Science*, 4(11):171377, 2017. doi: 10.1098/rsos.171377.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On bayesian upper confidence bounds for bandit problems. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pp. 592–600. PMLR, 2012. URL https://proceedings.mlr.press/v22/kaufmann12.html.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

Harin Lee and Min-hwan Oh. Unified framework of distributional regret in multi-armed bandits and reinforcement learning. In *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 4520–4584. PMLR, 2026. URL https://arxiv.org/abs/2605.05102.

Subhodip Panda and Shubhada Agrawal. Regret tail characterization of optimal bandit algorithms with generic rewards. In *2026 IEEE International Symposium on Information Theory*, 2026. URL https://arxiv.org/abs/2604.14876.

Antoine Salomon and Jean-Yves Audibert. Deviations of stochastic bandit regret. In *Algorithmic Learning Theory*, pp. 159–173. Springer, 2011. doi: 10.1007/978-3-642-24412-4_15.

David Simchi-Levi, Zeyu Zheng, and Feng Zhu. A simple and optimal policy design for online learning with safety against heavy-tailed risk. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=VK9jfSPnnb.

David Simchi-Levi, Zeyu Zheng, and Feng Zhu. Regret distribution in stochastic bandits: Optimal trade-off between expectation and tail risk, 2025. URL https://arxiv.org/abs/2304.04341v2.

Aleksandrs Slivkins. *Introduction to Multi-Armed Bandits*. Now Publishers, 2019. doi: 10.1561/2200000068.

Ambrus Tamás, Szabolcs Szentpéteri, and Balázs Csanád Csáji. Data-driven upper confidence bounds with near-optimal regret for heavy-tailed bandits, 2024. URL https://arxiv.org/abs/2406.05710.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933. doi: 10.1093/biomet/25.3-4.285.

Bengt von Bahr and Carl-Gustav Esseen. Inequalities for the $r$th absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, 36(1):299–303, 1965. doi: 10.1214/aoms/1177700291.

Lai Wei and Vaibhav Srivastava. Minimax policy for heavy-tailed bandits, 2020. URL https://arxiv.org/abs/2007.10493.

Feng Zhu and David Simchi-Levi. Adaptive variance inflation in thompson sampling: Efficiency, safety, robustness, and beyond. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Proofs of the reward-noise results, BCRR upper bounds, and lower bounds are in Appendices A, B, and C. Appendix D contains the corollaries and unknown-horizon extension, and Appendix E treats unknown moment parameters under an additional confidence-interval assumption. Appendices F and G contain literature comparisons and supplementary experiments.

# A Oracle Noise and Realized Payoff Shortfall

## A.1 Oracle noise floor

*Proof of Theorem 1.* Put $p = 1 + \varepsilon$ and fix $x \geq Cv^{1/p}T^{1/p}$ with $C$ large. Let $\mu = 1/2$ and let $Z$ be a centered three-point random variable satisfying

$$\mathbb{P}(Z = -2x) = \theta, \qquad \mathbb{P}(Z = 2x) = \theta, \qquad \mathbb{P}(Z = 0) = 1 - 2\theta, \qquad \theta = c_0 v(2x)^{-p},$$

where $c_0 > 0$ is small enough that $2\theta \leq 1/2$ and $\mathbb{E}|Z|^p \leq v$. Set $Y = \mu + Z$. Then $\mathbb{E}Y = \mu \in [0, 1]$ and the centered $p$th moment is at most $v$, so this is a valid one-arm instance.

For the oracle policy, $M_T = \sum_{t=1}^{T} Z_t$. On the event that exactly one sample equals $-2x$ and all other samples equal 0, we have $-M_T = 2x > x$. The assumption $x \geq Cv^{1/p}T^{1/p}$ gives $T\theta \leq c_0/(2^p C^p)$. Choose $C$ large enough that $2T\theta \leq 1/4$. Then $(1 - 2\theta)^{T-1} \geq c$ for a universal constant $c > 0$; for instance, use $(1-u)^m \geq \exp(-2um)$ when $u \leq 1/2$ and $um \leq 1/4$. Therefore

$$\mathbb{P}(-M_T > x) \geq T\theta(1 - 2\theta)^{T-1} \geq c\,\frac{vT}{x^p}.$$

For any $K \geq 2$, placing identical copies of $\nu$ on all arms gives the same selected-reward law under every adaptive policy, so the construction embeds into $\mathcal{H}_K(v, \varepsilon)$. □

## A.2 Shortfall decomposition

*Proof of Proposition 1.* Put $p = 1 + \varepsilon$. Since $\widehat{\mathsf{Reg}}_T = \mathsf{Reg}_T - M_T$,

$$\{\widehat{\mathsf{Reg}}_T > x\} \subseteq \{\mathsf{Reg}_T > x/2\} \cup \{-M_T > x/2\}.$$

Conditional on the history, $Y_t - \mu_{A_t}$ is a martingale difference with conditional $p$th moment at most $v$. For $1 < p \leq 2$, the scalar inequality $|a + b|^p \leq |a|^p + p|a|^{p-1}\operatorname{sgn}(a)b + C_p|b|^p$, followed by conditional expectation and iteration, gives the martingale von Bahr–Esseen bound

$$\mathbb{E}|M_T|^p \leq C_p \sum_{t=1}^{T} \mathbb{E}|Y_t - \mu_{A_t}|^p \leq C_p vT.$$

Markov's inequality therefore yields $\mathbb{P}(-M_T > x/2) \leq C_\varepsilon vT/x^p$. □

# B Proofs of the Pseudo-Regret-Tail Bounds

Proposition 2 bounds the estimation error at each valid phase. Lemmas 1 and 2 then bound regret when these estimates are accurate, including samples collected during invalid phases. Summing the phase failure probabilities gives Theorem 2; integrating this tail bound gives Theorem 3.

Throughout this section set

$$p = 1 + \varepsilon, \qquad \alpha = \frac{p-1}{p}, \qquad \beta = \frac{p}{p-1}, \qquad q_T = v^{1/p}\left(\frac{T}{K}\right)^{1/p}, \qquad R_T^\star = Kq_T.$$

Fix $\rho \geq 1$ and $\tau > 2$ throughout the upper-bound proof. Constants denoted by $c_{\varepsilon,\tau}, C_{\varepsilon,\tau}$ may change from line to line. We put $\kappa_\varepsilon = \bar{c}_\varepsilon^{\alpha}$, and prove all statements for the nontrivial case

$$\rho q_T \leq \kappa_\varepsilon v^{1/p} L, \qquad L = \lfloor T/K \rfloor.$$

If this fails, then either $L = 0$, in which case $T < K$ and $R_T^\star \geq T$ because $v \geq 1$, or $L \geq 1$ and

$$T \leq 2KL < \frac{2}{\kappa_\varepsilon v^{1/p}} \rho K q_T \leq C_\varepsilon \rho R_T^\star.$$

Thus $\mathsf{Reg}_T \leq T \leq C_\varepsilon \rho R_T^\star$, which proves both bounds in this case. Henceforth assume $\rho q_T \leq \kappa_\varepsilon v^{1/p} L$.

### B.1 PHASE-TUNED ROBUST MEAN

*Proof of Proposition 2.* Choose $\bar{c}_\varepsilon = 1/32$. For $1 \le u \le \bar{c}_\varepsilon n$, split the sample into $m = \lceil 8u \rceil$ consecutive blocks of common size $b = \lfloor n/m \rfloor$, discarding at most $m-1$ observations. Since $u \ge 1$ and $\bar{c}_\varepsilon$ is small, $8u \le m \le 9u$ and $b \ge n/(18u)$. Let $\bar{X}_j$ be the mean in block $j$ and let $\widehat{\mu}_{n,u}$ be the median of the block means.

For every block, the von Bahr–Esseen inequality (von Bahr & Esseen, 1965) gives

$$\mathbb{E}|\bar{X}_j - \mu|^p \le 2vb^{-(p-1)} \le 2\,18^{p-1} v \left(\frac{u}{n}\right)^{p-1}.$$

Set $C_\varepsilon = (16\,18^{p-1})^{1/p}$. Markov's inequality then yields

$$\mathbb{P}\left(|\bar{X}_j - \mu| > C_\varepsilon v^{1/p} \left(\frac{u}{n}\right)^{(p-1)/p}\right) \le \frac{1}{8}.$$

The block events are independent. A Chernoff bound therefore shows that the probability that at least half of the blocks are bad is at most $e^{-c_\varepsilon m} \le e^{-c_\varepsilon u}$. If fewer than half are bad, their median is within the displayed radius. Enlarging constants gives the stated bound. Because $\mu \in [0,1]$, projection onto $[0,1]$ cannot increase the estimation error, so the same event controls the clipped estimates used by BCRR. □

### B.2 RISK CALIBRATION

*Proof of Proposition 3.* Solving $v^{1/p}(u/n)^{(p-1)/p} = v^{1/p}(T/K)^{1/p}/n$ gives

$$u = n\left(\frac{(T/K)^{1/p}}{n}\right)^{p/(p-1)} = \left(\frac{T}{Kn}\right)^{1/(p-1)} = \left(\frac{T}{Kn}\right)^{1/\varepsilon}.$$

The condition $u_T(n) \lesssim n$ is equivalent to $T/K \lesssim n^{1+\varepsilon}$, or $n \gtrsim (T/K)^{1/(1+\varepsilon)}$. With the radius multiplied by $\rho$, the same calculation gives

$$u_{\rho,T}(n) = \rho^{p/(p-1)} u_T(n).$$

Consequently $u_{\rho,T}(n) \lesssim n$ is equivalent, up to constants, to $\rho^p T/K \lesssim n^p$, or

$$n \gtrsim \rho(T/K)^{1/p} = \rho q_T / v^{1/p}.$$

Keeping the constant in $u_{\rho,T}(n) \le \bar{c}_\varepsilon n$ gives the exact condition $n \ge \bar{c}_\varepsilon^{-(p-1)/p}\rho(T/K)^{1/p}$ stated in the proposition. □

### B.3 PHASE NOTATION AND DETERMINISTIC ELIMINATION

Let $L = \lfloor T/K \rfloor$. We use the deterministic phase grid

$$n_0 = 1, \qquad n_s = \min\{2^s, L\} \quad (s \ge 0), \qquad u_s = \rho^\beta \left(\frac{T}{Kn_s}\right)^{1/(p-1)}, \qquad a_s = A_\varepsilon \rho q_T / n_s.$$

Only distinct values of $n_s$ are used; denote the terminal grid index by $S$, so $n_S = L$. A phase is called *valid* if $u_s \le \bar{c}_\varepsilon n_s$.

If only one arm remains, we continue the phase grid for analysis, computing estimates at valid targets from that arm's subsequent pulls. This leaves the actions and regret unchanged: the only arm cannot be eliminated by comparison with itself. At a target $n \le L$, active arms have at most $n$ pulls and eliminated arms were removed at earlier targets. Every phase, including $S$, can therefore be completed within $Kn \le KL \le T$ total pulls.

The calibration identity and $A_\varepsilon = C_\varepsilon$ give

$$a_s = C_\varepsilon v^{1/p} \left(\frac{u_s}{n_s}\right)^{(p-1)/p}.$$

The nontrivial condition above is equivalent to $u_S \le \bar{c}_\varepsilon n_S$, so the final phase is valid. On invalid phases BCRR performs no elimination.

For $\eta > 0$, define the phase-good event

$$\mathcal{G}_s(\eta) = \{|\widehat{\mu}_{i,s} - \mu_i| \le \eta \text{ for every arm estimated at the valid phase } s\}.$$

Proposition 2 and a union bound give, for every valid phase $s$,

$$\mathbb{P}(\mathcal{G}_s(a_s)^c) \le 2Ke^{-c_\varepsilon u_s} \le C_\varepsilon K \exp\left\{-c_\varepsilon n_s \left(\frac{a_s}{v^{1/p}}\right)^\beta\right\}. \tag{B.1}$$

**Lemma 1** (One-step elimination)**.** *Fix a valid phase $s$. Suppose $\mathcal{G}_s(a_s)$ holds and that an optimal arm $i^\star$ is active when the phase-$s$ elimination step is applied. Then after the elimination step:*

*(i) $i^\star$ remains active;*

*(ii) every arm with gap larger than $(\tau+2)a_s$ is eliminated.*

*Proof.* On $\mathcal{G}_s(a_s)$,

$$\widehat{\mu}_{i^\star,s} \ge \mu^\star - a_s, \qquad M_s \le \mu^\star + a_s.$$

Thus $\widehat{\mu}_{i^\star,s} \ge M_s - 2a_s > M_s - \tau a_s$, so the optimal arm is retained because $\tau > 2$. If an active arm $i$ has $\Delta_i > (\tau+2)a_s$, then

$$\widehat{\mu}_{i,s} \le \mu^\star - \Delta_i + a_s < \mu^\star - (\tau+1)a_s.$$

Since $M_s \ge \widehat{\mu}_{i^\star,s} \ge \mu^\star - a_s$, we have $M_s - \tau a_s \ge \mu^\star - (\tau+1)a_s$, and therefore $\widehat{\mu}_{i,s} < M_s - \tau a_s$. Such an arm is eliminated. □

**Lemma 2** (Good phases imply small regret)**.** *Fix $x > 0$ and put $\eta = x/(64T)$. Let $s_x$ be the first valid phase such that $(\tau+2)a_{s_x} \le \eta$, if such a phase exists. If $s_x$ exists and $\mathcal{G}_s(a_s)$ holds for every valid phase $s \le s_x$, then*

$$\mathsf{Reg}_T \le C_{\varepsilon,\tau}\rho R_T^\star + x/4.$$

*Proof.* Lemma 1 implies by induction that the optimal arm is active after every good valid phase $s \le s_x$, and that after such a phase all active arms with gap larger than $(\tau+2)a_s$ have been eliminated. Invalid phases do not change the active set, and all arms are initially active.

Consider an arm with $\Delta_i > \eta$. Let $s(i)$ be the first phase such that the phase is valid and $(\tau+2)a_{s(i)} < \Delta_i$. This phase occurs no later than $s_x$, because $(\tau+2)a_{s_x} \le \eta < \Delta_i$. If $s(i) = 0$, then $n_{s(i)} = 1 \le \rho q_T/\Delta_i$ because $q_T \ge 1$ and $\Delta_i \le 1$. If $s(i) > 0$ and the preceding distinct phase is invalid, then $u_{s(i)-1} > \bar{c}_\varepsilon n_{s(i)-1}$. Since $u_s/n_s = (\rho q_T/(v^{1/p}n_s))^\beta$, dyadicity gives

$$n_{s(i)} \le \frac{2\rho q_T}{\kappa_\varepsilon v^{1/p}} \le C_\varepsilon \frac{\rho q_T}{\Delta_i},$$

where the last step uses $\Delta_i \le 1$ and $v \ge 1$. Finally, if $s(i) > 0$ and the preceding phase is valid, that phase does not yet satisfy the elimination inequality, so $(\tau+2)a_{s(i)-1} \ge \Delta_i$. Again by dyadicity,

$$n_{s(i)} \le 2n_{s(i)-1} \le \frac{2(\tau+2)A_\varepsilon \rho q_T}{\Delta_i}.$$

In all cases, $n_{s(i)} \le C_{\varepsilon,\tau}\rho q_T/\Delta_i$. Arm $i$ is eliminated by phase $s(i)$, after at most $n_{s(i)}$ pulls, so its regret is at most

$$\Delta_i n_{s(i)} \le C_{\varepsilon,\tau}\rho q_T.$$

Summing over the at most $K$ arms with gap above $\eta$ gives $C_{\varepsilon,\tau}\rho K q_T = C_{\varepsilon,\tau}\rho R_T^\star$.

After phase $s_x$, all active arms have gap at most $\eta$. Later estimation errors cannot reactivate eliminated arms. The total regret from arms with gap at most $\eta$ is bounded by

$$\sum_{t:A_t \text{ has } \Delta_{A_t} \le \eta} \Delta_{A_t} \le \eta T = x/64,$$

which is absorbed into $x/4$. Samples from invalid phases are already included in these counts. □

### B.4 TAIL BOUND

*Proof of Theorem 2.* Write $x$ for the excess threshold $y$ in the theorem statement. If $C_{\varepsilon,\tau}\rho R_T^\star + x \geq T$, then the event is empty after increasing $C_{\varepsilon,\tau}$, because $\mathsf{Reg}_T \leq T$. Hence assume $C_{\varepsilon,\tau}\rho R_T^\star + x < T$. Put

$$y = \max\{x, D_{\varepsilon,\tau}\rho R_T^\star\}, \qquad \eta = \frac{y}{64T}.$$

Here $D_{\varepsilon,\tau}$ is a sufficiently large fixed constant. If $y \geq T$, then the target event is empty after increasing the leading constant in the theorem, so assume $y < T$. To verify that $s_y$ in Lemma 2 exists, use $n_S = L$ and $T \leq 2KL$ to obtain

$$(\tau+2)a_S = \frac{(\tau+2)A_\varepsilon \rho q_T}{L} \leq \frac{C_{\varepsilon,\tau}^{(0)}\rho R_T^\star}{T}.$$

Choose $D_{\varepsilon,\tau} \geq 64 C_{\varepsilon,\tau}^{(0)}$. Since $y \geq D_{\varepsilon,\tau}\rho R_T^\star$, this gives $(\tau+2)a_S \leq y/(64T) = \eta$. Thus $s_y$ exists and satisfies $(\tau+2)a_{s_y} \leq \eta$ and, unless $s_y$ is the first valid phase, $\eta < (\tau+2)a_{s_y-1}$.

The inequality $(\tau+2)a_{s_y} \leq \eta$ gives $n_{s_y} \geq (\tau+2)A_\varepsilon \rho q_T/\eta$. If the preceding distinct phase is valid and does not yet satisfy the target inequality, dyadicity gives $n_{s_y} \leq C_{\varepsilon,\tau}\rho q_T/\eta$. If $s_y$ is the first valid phase, then either $s_y = 0$, in which case $n_{s_y} = 1 \leq \rho q_T/v^{1/p}$ because $T \geq K$, or the preceding distinct phase is invalid. In the latter case, $u_{s_y-1}/n_{s_y-1} > \bar{c}_\varepsilon$ and dyadicity give $n_{s_y} \leq 2\rho q_T/(\kappa_\varepsilon v^{1/p})$. Since $y < T$ implies $\eta < 1$, and $v \geq 1$, this is again bounded by $C_{\varepsilon,\tau}\rho q_T/\eta$. Therefore

$$c_{\varepsilon,\tau}\frac{\rho q_T}{\eta} \leq n_{s_y} \leq C_{\varepsilon,\tau}\frac{\rho q_T}{\eta}.$$

Since $a_{s_y} = A_\varepsilon \rho q_T/n_{s_y}$,

$$\begin{aligned} n_{s_y}\left(\frac{a_{s_y}}{v^{1/p}}\right)^{p/(p-1)} &= n_{s_y}\left(\frac{A_\varepsilon \rho q_T}{n_{s_y} v^{1/p}}\right)^{p/(p-1)} \\ &= A_\varepsilon^{p/(p-1)}\rho^{p/(p-1)}\left(\frac{T}{K n_{s_y}}\right)^{1/(p-1)}. \end{aligned}$$

The upper bound on $n_{s_y}$ gives $T/(Kn_{s_y}) \geq c_{\varepsilon,\tau}T\eta/(K\rho q_T) = c_{\varepsilon,\tau}y/(\rho R_T^\star)$. Hence

$$n_{s_y}\left(\frac{a_{s_y}}{v^{1/p}}\right)^{p/(p-1)} \geq c_{\varepsilon,\tau}\rho\left(\frac{y}{R_T^\star}\right)^{1/\varepsilon} \geq c_{\varepsilon,\tau}\rho\left(\frac{x}{R_T^\star}\right)^{1/\varepsilon}. \tag{B.2}$$

For valid $s \leq s_y$, write $m = s_y - s$. Apart from a possible capped terminal step, the phase sizes are dyadic, so $n_s \leq 2^{1-m}n_{s_y}$. The displayed identity shows that the exponent $n_s(a_s/v^{1/p})^{p/(p-1)}$ is at least $2^{m/\varepsilon}$ times the value in (B.2), up to constants. Therefore, using (B.1) and summing a geometric series over valid phases,

$$\mathbb{P}\left(\bigcup_{\substack{s \leq s_y \\ s \text{ valid}}} \mathcal{G}_s(a_s)^c\right) \leq C_{\varepsilon,\tau}K\exp\left\{-c_{\varepsilon,\tau}\rho\left(\frac{x}{R_T^\star}\right)^{1/\varepsilon}\right\}.$$

On the complement, Lemma 2, applied with $y$, gives $\mathsf{Reg}_T \leq C_{\varepsilon,\tau}\rho R_T^\star + y/4$. If $y = x$ this is already of the desired form; if $y = D_{\varepsilon,\tau}\rho R_T^\star$, the additional $y/4$ is absorbed by increasing the leading constant multiplying $\rho R_T^\star$. Therefore

$$\{\mathsf{Reg}_T > C'_{\varepsilon,\tau}\rho R_T^\star + x\} \subseteq \bigcup_{\substack{s \leq s_y \\ s \text{ valid}}} \mathcal{G}_s(a_s)^c$$

for a sufficiently large constant $C'_{\varepsilon,\tau}$. Renaming constants gives the first display of the theorem.

For the linear-regret consequence in Section 5, fix $\gamma \in (0,1)$. If $\gamma T \leq C_{\varepsilon,\tau}\rho R_T^\star$, then the bound below with the positive part is trivial after increasing $C_{\varepsilon,\tau,\gamma}$. Otherwise, apply the first display with $x = \gamma T - C_{\varepsilon,\tau}\rho R_T^\star > 0$ to obtain

$$\mathbb{P}_\nu(\mathsf{Reg}_T > \gamma T) \leq C_{\varepsilon,\tau}K\exp\left\{-c_{\varepsilon,\tau}\rho\left(\frac{\gamma T - C_{\varepsilon,\tau}\rho R_T^\star}{R_T^\star}\right)^{1/\varepsilon}\right\},$$

which also gives a valid bound with the positive part after absorbing constants. If in addition $C_{\varepsilon,\tau}\rho R_T^\star \leq \gamma T/2$, then $\gamma T - C_{\varepsilon,\tau}\rho R_T^\star \geq \gamma T/2$, giving the simplified nontrivial-regime bound. □

### B.5 Expected regret

Theorem 3 follows from Theorem 2 and the following integration bound.

**Lemma 3** (Tail integration with a budget multiplier)**.** *Let $Z$ be a nonnegative random variable. Fix $R > 0$, $a > 0$, $\rho \geq 1$, $b \geq 1$, $c > 0$, and $\varepsilon \in (0, 1]$. Suppose that, for every $y > 0$,*

$$\mathbb{P}(Z > a\rho R + y) \leq b \exp\left\{-c\rho \left(\frac{y}{R}\right)^{1/\varepsilon}\right\}.$$

*Then there is a constant $C_{\varepsilon,c} > 0$ such that*

$$\mathbb{E}Z \leq a\rho R + C_{\varepsilon,c} R \rho^{-\varepsilon} \left(1 + \log^{\varepsilon}(eb)\right).$$

*Proof.* Using the layer-cake identity for nonnegative random variables,

$$\mathbb{E}Z = \int_0^\infty \mathbb{P}(Z > t)\, dt.$$

Split the integral at $a\rho R$ and then set $t = a\rho R + y$ in the tail part:

$$\begin{aligned}\mathbb{E}Z &\leq a\rho R + \int_0^\infty \min\left\{1, b \exp\left[-c\rho \left(\frac{y}{R}\right)^{1/\varepsilon}\right]\right\} dy \\ &= a\rho R + R \int_0^\infty \min\left\{1, b \exp\{-c\rho u^{1/\varepsilon}\}\right\} du,\end{aligned}$$

where $u = y/R$. Define

$$u_0 = \left(\frac{2\log(eb)}{c\rho}\right)^{\varepsilon}.$$

For $0 \leq u \leq u_0$, the integrand is at most one, hence this part of the integral is at most $u_0$. For $u > u_0$, we have $c\rho u^{1/\varepsilon} > 2\log(eb)$, and therefore

$$be^{-c\rho u^{1/\varepsilon}} = \exp\{-c\rho u^{1/\varepsilon} + \log b\} \leq \exp\{-c\rho u^{1/\varepsilon}/2\}.$$

Consequently,

$$\begin{aligned}\int_{u_0}^\infty \min\left\{1, be^{-c\rho u^{1/\varepsilon}}\right\} du &\leq \int_0^\infty \exp\{-c\rho u^{1/\varepsilon}/2\}\, du \\ &= \varepsilon(2/c)^{\varepsilon}\Gamma(\varepsilon)\rho^{-\varepsilon} = C_{\varepsilon,c}\rho^{-\varepsilon}.\end{aligned}$$

Combining the two pieces gives

$$\mathbb{E}Z \leq a\rho R + R\rho^{-\varepsilon}\left[(2/c)^{\varepsilon}\log^{\varepsilon}(eb) + C_{\varepsilon,c}\right],$$

which is the claimed bound after increasing $C_{\varepsilon,c}$. □

*Proof of Theorem 3.* Apply Theorem 2 with

$$Z = \mathsf{Reg}_T, \qquad R = R_T^\star, \qquad b = C_{\varepsilon,\tau}K,$$

and with the theorem's leading constant absorbed into the parameter $a$ of Lemma 3. The lemma yields

$$\mathbb{E}\mathsf{Reg}_T \leq C_{\varepsilon,\tau}\rho R_T^\star + C_{\varepsilon,\tau}R_T^\star\rho^{-\varepsilon}\left[1 + \log^{\varepsilon}(C_{\varepsilon,\tau}K)\right].$$

This implies the stated bound, which includes an additional $C_{\varepsilon,\tau}K$ term. □

## C Proofs of the Budget Frontier and Instance Barrier

Both lower bounds use the following binary KL inequality and rare-spike family.

### C.1 CHANGE-OF-MEASURE LEMMAS

**Lemma 4** (Binary KL inversion)**.** *If an event $G$ satisfies $\mathbb{P}(G) \geq 3/4$ under a law $P$, and $\mathrm{KL}(P, Q) \leq d$, then*

$$Q(G) \geq ce^{-Cd}$$

*for universal constants $c, C > 0$.*

*Proof.* By data processing, $d \geq \mathrm{kl}(P(G), Q(G))$. For $a \in [3/4, 1]$ and $b \in (0, 1)$,

$$\mathrm{kl}(a, b) = a \log(a/b) + (1 - a) \log((1 - a)/(1 - b)) \geq \frac{3}{4} \log(1/b) - \log 2.$$

Thus $b \geq \exp\{-(4/3)(d + \log 2)\}$, which proves the claim. □

**Lemma 5** (Rare-spike moment family)**.** *There is a constant $\bar{\Delta}_\varepsilon > 0$ such that, for every $v \geq 1$ and every $\Delta \in (0, \bar{\Delta}_\varepsilon]$, there exist distributions $P_0, P_1, P_2$ with means $0, \Delta, 2\Delta$, respectively, such that, with $p = 1 + \varepsilon$, each has centered pth moment at most $v$ and*

$$\mathrm{KL}(P_0, P_2) \leq C_\varepsilon \left( \frac{\Delta}{v^{1/p}} \right)^{p/(p-1)}.$$

*Proof.* Let $h = a_\varepsilon (v/\Delta)^{1/(p-1)}$ and $\theta = \Delta/h$, where $a_\varepsilon > 0$ is a sufficiently small constant. Choose $\bar{\Delta}_\varepsilon \leq 1/2$ small enough that, for every $\Delta \leq \bar{\Delta}_\varepsilon$ and every $v \geq 1$, we have $3\theta \leq 1/2$ and $\theta \leq 1/6$. The first inequality makes all Bernoulli parameters below valid, and the second is the small-parameter range used in the KL bound. The same choice also ensures the means $0, \Delta, 2\Delta$ lie in $[0, 1]$. Let $Q_a$ be the two-point distribution on $\{0, h\}$ with mass $(a + 1)\theta$ at $h$, for $a \in \{0, 1, 2\}$. Its mean is $(a + 1)\Delta$. Let $P_a$ be the law of $X - \Delta$ when $X \sim Q_a$. Then the means of $P_0, P_1, P_2$ are $0, \Delta, 2\Delta$, respectively.

For $X \sim Q_a$,

$$\mathbb{E}|X - \mathbb{E}X|^p \leq C_p(a + 1)\theta h^p = C_p(a + 1)\Delta h^{p-1} = C_p(a + 1)a_\varepsilon^{p-1} v.$$

Choosing $a_\varepsilon$ small enough makes this bound at most $v$. Shifting by $-\Delta$ preserves centered moments, and the KL divergence is

$$\mathrm{KL}(P_0, P_2) = \mathrm{kl}(\theta, 3\theta) \leq C\theta,$$

where the last inequality holds for $\theta \leq 1/6$. Finally,

$$\theta = \frac{\Delta}{h} = C_\varepsilon \left( \frac{\Delta}{v^{1/p}} \right)^{p/(p-1)}.$$

□

### C.2 BUDGET FRONTIER

*Proof of Theorem 4.* Write $x$ for the absolute threshold $z$ in the theorem statement and put $p = 1+\varepsilon$. Let

$$m = \lfloor K/2 \rfloor, \qquad \Delta = \frac{4x}{T}.$$

Choose the constant $c_\varepsilon$ in the theorem small enough that $\Delta \leq \bar{\Delta}_\varepsilon$, the admissible range in Lemma 5. Construct the base instance $\nu^0$ by putting $P_1$ on arm 1, putting $P_0$ on candidate arms $j \in \{2, \ldots, m+1\}$, and filling all remaining arms with copies of $P_0$. For each candidate $j$, let $\nu^j$ be the alternative that replaces only arm $j$ by $P_2$.

On $\nu^0$, arm 1 is optimal and every candidate arm has gap $\Delta$. Therefore the candidate arms alone contribute

$$\Delta \sum_{j=2}^{m+1} N_j(T)$$

to the pseudo-regret. Since the policy has worst-case expected regret at most $B_T$,

$$\Delta \sum_{j=2}^{m+1} \mathbb{E}_0[N_j(T)] \le \mathbb{E}_0[\mathsf{Reg}_T^\pi] \le B_T.$$

Averaging over the $m$ candidate arms gives a candidate $j^\star$ such that

$$\mathbb{E}_0[N_{j^\star}(T)] \le \frac{B_T}{m\Delta}.$$

Let $\mathbb{P}_0$ and $\mathbb{P}_{j^\star}$ be the trajectory laws on the base and this chosen alternative. The two induced laws differ only through observations collected when the learner pulls arm $j^\star$. Thus the adaptive chain rule for KL divergence gives

$$\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_{j^\star}) = \mathbb{E}_0[N_{j^\star}(T)]\mathrm{KL}(P_0, P_2) \le C_\varepsilon \frac{B_T}{m\Delta} \left(\frac{\Delta}{v^{1/p}}\right)^{p/(p-1)}.$$

Using $m \asymp K$ and $(R_T^\star)^{p/(p-1)} = v^{1/(p-1)} K T^{1/(p-1)}$, the last display is bounded by

$$\begin{aligned} C_\varepsilon \frac{B_T}{K\Delta} \left(\frac{\Delta}{v^{1/p}}\right)^{p/(p-1)} &= C_\varepsilon \frac{B_T \Delta^{1/(p-1)}}{K v^{1/(p-1)}} \\ &= C_\varepsilon \frac{B_T}{R_T^\star} \left(\frac{T\Delta}{R_T^\star}\right)^{1/(p-1)} \\ &\le C_\varepsilon \frac{B_T}{R_T^\star} \left(\frac{x}{R_T^\star}\right)^{1/\varepsilon}, \end{aligned}$$

where the last inequality absorbs the constant $4^{1/\varepsilon}$ from $T\Delta = 4x$.

Now set

$$G = \{N_{j^\star}(T) \le T/2\}.$$

On the base instance, Markov's inequality and the averaging bound give

$$\mathbb{P}_0(N_{j^\star}(T) > T/2) \le \frac{2B_T}{m\Delta T} \le \frac{1}{4},$$

after increasing the lower-end constant in $x \ge C_\varepsilon B_T$, because $\Delta T = 4x$ and $m \ge 1$. Hence $\mathbb{P}_0(G) \ge 3/4$. Lemma 4 gives

$$\mathbb{P}_{j^\star}(G) \ge c \exp\{-C\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_{j^\star})\}.$$

Under the alternative $\nu^{j^\star}$, arm $j^\star$ has mean $2\Delta$, arm 1 has mean $\Delta$, and all $P_0$ arms have mean 0. Therefore every pull of an arm different from $j^\star$ incurs regret at least $\Delta$. On $G$, at least $T/2$ pulls are different from $j^\star$, so

$$\mathsf{Reg}_T^\pi \ge \Delta(T - N_{j^\star}(T)) \ge \frac{\Delta T}{2} = 2x > x.$$

Thus $G \subseteq \{\mathsf{Reg}_T^\pi > x\}$ under $\nu^{j^\star}$, and

$$\mathbb{P}_{j^\star}(\mathsf{Reg}_T^\pi > x) \ge c \exp\left\{-C_\varepsilon \frac{B_T}{R_T^\star} \left(\frac{x}{R_T^\star}\right)^{1/\varepsilon}\right\}.$$

Since $B_T \ge R_T^\star$ and $x \ge C_\varepsilon B_T$, the exponent $E$ in the last display satisfies

$$E = \frac{B_T}{R_T^\star} \left(\frac{x}{R_T^\star}\right)^{1/\varepsilon} \ge C_\varepsilon^{1/\varepsilon} \left(\frac{B_T}{R_T^\star}\right)^{1+1/\varepsilon} \ge C_\varepsilon^{1/\varepsilon}.$$

Hence $ce^{-CE} \ge e^{-C'E}$ for a larger $C'$, which absorbs the leading constant and gives the stated lower bound. □

### C.3 Instance-dependent barrier

**Theorem 5** (Logarithmic-regret policies have polynomial linear tails)**.** *Fix $K \geq 2$, $\varepsilon \in (0, 1]$, and $v \geq 1$. There is a constant $\bar{\Delta}_\varepsilon > 0$ such that the following holds for every $\Delta_0 \in (0, \bar{\Delta}_\varepsilon]$. Let $P_0, P_1, P_2$ be the distributions given by Lemma 5 with $\Delta = \Delta_0$, so their means are $0, \Delta_0, 2\Delta_0$, respectively. Let $\pi$ be any $K$-arm policy. On the base instance $\nu^0 = (P_1, P_0, \ldots, P_0)$, suppose the policy is logarithmic on the designated suboptimal arm 2:*

$$\mathbb{E}_{\nu^0}[N_2(T)] \leq a \log T \qquad \textit{for all sufficiently large } T.$$

*Let $\nu^1 = (P_1, P_2, P_0, \ldots, P_0)$ be the neighboring instance obtained by replacing arm 2 by a distribution of mean $2\Delta_0$. Then there are constants $c, C > 0$, depending only on $(\varepsilon, v, a, \Delta_0)$, such that*

$$\mathbb{P}_{\nu^1}\left(\mathsf{Reg}_T^\pi > \frac{\Delta_0 T}{4}\right) \geq cT^{-C} \qquad \textit{for all sufficiently large } T.$$

*Proof of Theorem 5.* Let $G = \{N_2(T) \leq T/2\}$. On the base instance $\nu^0 = (P_1, P_0, \ldots, P_0)$, arm 2 is suboptimal and the hypothesis gives $\mathbb{E}_{\nu^0}[N_2(T)] \leq a \log T$. Therefore Markov's inequality yields

$$\mathbb{P}_{\nu^0}(G^c) = \mathbb{P}_{\nu^0}(N_2(T) > T/2) \leq \frac{2\mathbb{E}_{\nu^0}[N_2(T)]}{T} \leq \frac{2a \log T}{T},$$

and hence, for all sufficiently large $T$,

$$\mathbb{P}_{\nu^0}(G) \geq \frac{3}{4}.$$

Let $\mathbb{P}_{\nu^0}$ and $\mathbb{P}_{\nu^1}$ denote the full trajectory laws. The two instances differ only in the distribution of arm 2, which changes from $P_0$ to $P_2$. The adaptive KL chain rule gives

$$\mathrm{KL}(\mathbb{P}_{\nu^0}, \mathbb{P}_{\nu^1}) = \mathbb{E}_{\nu^0}[N_2(T)]\mathrm{KL}(P_0, P_2).$$

For the fixed gap $\Delta_0$, Lemma 5 gives a finite constant $D_{\varepsilon,v,\Delta_0}$ such that $\mathrm{KL}(P_0, P_2) \leq D_{\varepsilon,v,\Delta_0}$. Consequently

$$\mathrm{KL}(\mathbb{P}_{\nu^0}, \mathbb{P}_{\nu^1}) \leq aD_{\varepsilon,v,\Delta_0} \log T = C_{\varepsilon,v,\Delta_0,a} \log T.$$

Applying Lemma 4 to the event $G$ gives

$$\mathbb{P}_{\nu^1}(G) \geq c \exp\{-C\mathrm{KL}(\mathbb{P}_{\nu^0}, \mathbb{P}_{\nu^1})\} \geq cT^{-C_{\varepsilon,v,\Delta_0,a}}.$$

Under $\nu^1 = (P_1, P_2, P_0, \ldots, P_0)$, arm 2 is optimal with mean $2\Delta_0$, arm 1 has mean $\Delta_0$, and every remaining arm has mean 0. Thus every pull outside arm 2 incurs regret at least $\Delta_0$. On $G$, at least $T/2$ pulls are not arm 2, so

$$\mathsf{Reg}_T^\pi \geq \Delta_0(T - N_2(T)) \geq \frac{\Delta_0 T}{2} > \frac{\Delta_0 T}{4}.$$

Therefore

$$\mathbb{P}_{\nu^1}\left(\mathsf{Reg}_T^\pi > \frac{\Delta_0 T}{4}\right) \geq \mathbb{P}_{\nu^1}(G) \geq cT^{-C_{\varepsilon,v,\Delta_0,a}}.$$

□

## D Corollaries and the Unknown-Horizon Extension

### D.1 Unknown-horizon extension

**Restart schedule.** Epoch $j \geq 0$ starts after $S_j = K(2^j - 1)$ rounds and lasts $H_j = K2^j$ rounds. At its start, initialize BCRR with all $K$ arms active, empty reward histories, local horizon $H_j$, and the fixed parameters $(v, \varepsilon, \rho, \tau)$. Run this copy for $H_j$ rounds, then discard it and start the next epoch. This geometric restart schedule (Besson & Kaufmann, 2018) requires no evaluation horizon. An incorrect elimination lasts at most until the next restart, when all arms become active again. Algorithm 2 runs each copy one pull at a time, allowing interaction to end partway through an epoch.

**Algorithm 2** Anytime BCRR via geometric restarts

**Input:** number of arms $K \geq 2$, known moment parameters $v \geq 1$, $\varepsilon \in (0, 1]$, budget parameter $\rho \geq 1$, threshold $\tau > 2$

$j \leftarrow 0$
**while** interaction continues **do**
  $H_j \leftarrow K2^j$
  Initialize a fresh BCRR copy $\mathcal{B}_j$ (Algorithm 1) with inputs $(H_j, K, v, \varepsilon, \rho, \tau)$
    All arms are active; pull counts are zero and reward histories are empty
  **for** $\ell = 1, \ldots, H_j$ **do**
    **if** the interaction has ended **then stop**
    $A \leftarrow$ next arm proposed by $\mathcal{B}_j$
    Pull arm $A$ and observe reward $Y$
    Supply $(A, Y)$ to $\mathcal{B}_j$ and advance its local state by one pull
  **end for**
  Discard $\mathcal{B}_j$ and all of its state and observations
  $j \leftarrow j + 1$
**end while**

*Proof of Corollary 1.* Fix an instance $\nu \in \mathcal{H}_K(v, \varepsilon)$ and a deterministic $T \geq K$. Write $a = 1/(1 + \varepsilon)$ and $R(h) = v^a K^{1-a} h^a$, so that $R(T) = R^\star_T$. Let $J$ be the epoch containing round $T$, with $S_J < T \leq S_J + H_J$. Since $S_J = H_J - K$, we have $H_J < T + K \leq 2T$. Let $X_j$ be the pseudo-regret of the full run in epoch $j$, for $0 \leq j \leq J$. If the last epoch is unfinished, continue that copy on fresh reward samples to define $X_J$. Every pseudo-regret increment is nonnegative, hence, pathwise on this enlarged probability space,

$$\mathsf{Reg}_T^{\pi^{\mathrm{any}}_{\rho,\tau}} \leq \sum_{j=0}^{J} X_j.$$

Fresh observations in each epoch give the same BCRR guarantee, conditional on the history at its start. No independence between the $X_j$ is needed for the union bound below. By Theorem 2, there are $c_0, C_0 > 0$, depending only on $(\varepsilon, \tau)$, such that

$$\mathbb{P}_\nu\big(X_j > C_0 \rho R(H_j) + u\big) \leq C_0 K \exp\left\{-c_0 \rho \left(\frac{u}{R(H_j)}\right)^{1/\varepsilon}\right\} \qquad (u > 0).$$

The total exploration baseline is bounded by a geometric series:

$$\sum_{j=0}^{J} R(H_j) = R(H_J) \sum_{m=0}^{J} 2^{-am} \leq \frac{2^a}{1 - 2^{-a}} R(T) = D_\varepsilon R(T). \tag{D.1}$$

Fix $y > 0$ and assign epoch $J - m$ the deviation threshold

$$y_m = (1 - 2^{-a/2})\, y\, 2^{-am/2}, \qquad 0 \leq m \leq J.$$

These thresholds sum to at most $y$. The union bound gives

$$\begin{aligned}\mathbb{P}_\nu\big(\mathsf{Reg}_T^{\pi^{\mathrm{any}}_{\rho,\tau}} > C_0 D_\varepsilon \rho R(T) + y\big) \\ \leq \sum_{m=0}^{J} \mathbb{P}_\nu\big(X_{J-m} > C_0 \rho R(H_{J-m}) + y_m\big).\end{aligned}$$

Since $R(H_{J-m}) = R(H_J) 2^{-am} \leq 2^a R(T) 2^{-am}$,

$$\frac{y_m}{R(H_{J-m})} \geq \frac{1 - 2^{-a/2}}{2^a} \frac{y}{R(T)} 2^{am/2}.$$

Set

$$d_\varepsilon = \left(\frac{1 - 2^{-a/2}}{2^a}\right)^{1/\varepsilon}, \quad A = c_0 d_\varepsilon \rho \left(\frac{y}{R(T)}\right)^{1/\varepsilon}, \quad \lambda = 2^{a/(2\varepsilon)} > 1.$$

The preceding probability is at most $C_0 K \sum_{m=0}^{J} e^{-A\lambda^m}$. If $A \geq 1$, then $\lambda^m \geq 1 + m(\lambda - 1)$ gives

$$\sum_{m=0}^{J} e^{-A\lambda^m} \leq \frac{e^{-A}}{1 - e^{-A(\lambda-1)}} \leq \frac{e^{-A}}{1 - e^{-(\lambda-1)}}.$$

If $A < 1$, the trivial probability bound $1 \leq e\, e^{-A}$ suffices. After enlarging the prefactor and baseline constants, both cases give

$$\sup_{\nu \in \mathcal{H}_K(v,\varepsilon)} \mathbb{P}_\nu\Big(\mathsf{Reg}_T^{\pi^{\mathrm{any}}_{\rho,\tau}} > C_{\varepsilon,\tau}\rho R_T^\star + y\Big) \leq C_{\varepsilon,\tau} K \exp\left\{-c_{\varepsilon,\tau}\rho\left(\frac{y}{R_T^\star}\right)^{1/\varepsilon}\right\}.$$

All constants are independent of $T, K, v, \rho$ and $\nu$. Lemma 3 now gives

$$\sup_{\nu \in \mathcal{H}_K(v,\varepsilon)} \mathbb{E}_\nu\big[\mathsf{Reg}_T^{\pi^{\mathrm{any}}_{\rho,\tau}}\big] \leq C_{\varepsilon,\tau}\rho R_T^\star + C_{\varepsilon,\tau} R_T^\star \rho^{-\varepsilon}\big[1 + \log^\varepsilon(C_{\varepsilon,\tau} K)\big],$$

with the same policy and constants for every deterministic $T \geq K$. □

**Remark.** For fixed $K$, the expected regret is of order $\rho R_T^\star$. Together with Theorem 4, this gives the same frontier on its stated threshold range, up to constants. The logarithmic instance-dependent requirement in Corollary 3 is a separate constraint (Simchi-Levi et al., 2025). The bounds hold separately at each deterministic horizon; they do not give a bound uniform over time or at arbitrary stopping times. The construction still requires known moment parameters and discards data at each restart.

### D.2 Minimax tails and regret quantiles

**Corollary 4** (Minimax pseudo-regret-tail lower bound)**.** *Fix $\varepsilon \in (0, 1]$, $K \geq 2$, and $v \geq 1$. Let $\pi$ be any policy satisfying*

$$\sup_{\nu \in \mathcal{H}_K(v,\varepsilon)} \mathbb{E}_\nu[\mathsf{Reg}_T^\pi] \leq A R_T^\star(v, \varepsilon, K)$$

*for a constant $A > 0$. Then there exist constants $c_{\varepsilon,A}, C_{\varepsilon,A} > 0$ such that, for every $x \in [C_{\varepsilon,A} R_T^\star, c_{\varepsilon,A} T]$,*

$$\sup_{\nu \in \mathcal{H}_K(v,\varepsilon)} \mathbb{P}_\nu(\mathsf{Reg}_T^\pi > x) \geq \exp\left[-C_{\varepsilon,A}\left(\frac{x}{R_T^\star}\right)^{1/\varepsilon}\right].$$

*Proof.* Apply Theorem 4 with $B_T = (A \vee 1) R_T^\star$. Its hypothesis holds because $\sup_\nu \mathbb{E}_\nu \mathsf{Reg}_T^\pi \leq A R_T^\star \leq B_T$. The lower bound on the upper-tail probability gives the claim after absorbing $A \vee 1$ into $c_{\varepsilon,A}$ and $C_{\varepsilon,A}$. □

**Corollary 5** (Regret quantiles and confidence levels)**.** *For every $\delta \in (0, 1)$, BCRR$(\rho, \tau)$ satisfies*

$$\sup_{\nu \in \mathcal{H}_K(v,\varepsilon)} \mathbb{P}_\nu\left(\mathsf{Reg}_T > C_{\varepsilon,\tau}\rho R_T^\star + C_{\varepsilon,\tau} R_T^\star \rho^{-\varepsilon} \log^\varepsilon\left(\frac{C_{\varepsilon,\tau} K}{\delta}\right)\right) \leq \delta.$$

*Conversely, for any policy with $\sup_{\nu \in \mathcal{H}_K(v,\varepsilon)} \mathbb{E}_\nu[\mathsf{Reg}_T] \leq A R_T^\star$, there are constants $c_{\varepsilon,A}, C_{\varepsilon,A} > 0$ such that, for all $\delta \in [\exp\{-c_{\varepsilon,A}(T/R_T^\star)^{1/\varepsilon}\}, c_{\varepsilon,A}]$,*

$$\sup_{\nu \in \mathcal{H}_K(v,\varepsilon)} Q_{1-\delta}(\mathsf{Reg}_T) \geq c_{\varepsilon,A} R_T^\star \log^\varepsilon(1/\delta),$$

*where $Q_{1-\delta}(Z)$ denotes the $(1-\delta)$-quantile of $Z$. For empirical regret,*

$$\sup_{\nu \in \mathcal{H}_K(v,\varepsilon)} Q_{1-\delta}(\widehat{\mathsf{Reg}}_T) \leq C_{\varepsilon,\tau}\rho R_T^\star + C_{\varepsilon,\tau} R_T^\star \rho^{-\varepsilon} \log^\varepsilon\left(\frac{C_{\varepsilon,\tau} K}{\delta}\right) + C_{\varepsilon,\tau}\left(\frac{vT}{\delta}\right)^{1/(1+\varepsilon)},$$

*and, for all sufficiently small $\delta$, the order of the final reward-noise term cannot be improved uniformly over $\mathcal{H}_K(v, \varepsilon)$.*

*Proof of Corollary 5.* Let

$$B = C_{\varepsilon,\tau} K \quad \text{and} \quad x_\delta = C_{\varepsilon,\tau} R_T^\star \rho^{-\varepsilon} \log^\varepsilon(B/\delta).$$

Taking $C_{\varepsilon,\tau}$ large enough ensures $B/\delta \geq e$ and

$$c_{\varepsilon,\tau} \rho \left( \frac{x_\delta}{R_T^\star} \right)^{1/\varepsilon} \geq \log(B/\delta).$$

Theorem 2 then gives

$$\mathbb{P}_\nu(\mathsf{Reg}_T > C_{\varepsilon,\tau} \rho R_T^\star + x_\delta) \leq B \exp\{-\log(B/\delta)\} = \delta.$$

For the converse, put

$$x_\delta = c_{\varepsilon,A} R_T^\star \log^\varepsilon(1/\delta).$$

The stated range of $\delta$ is chosen so that $x_\delta \in [C_{\varepsilon,A} R_T^\star, c_{\varepsilon,A} T]$, the validity range of Corollary 4. That corollary gives an instance with

$$\mathbb{P}_\nu(\mathsf{Reg}_T > x_\delta) \geq \exp\left\{ -C_{\varepsilon,A} \left( \frac{x_\delta}{R_T^\star} \right)^{1/\varepsilon} \right\}.$$

Choosing the constant $c_{\varepsilon,A}$ small enough makes the right-hand side at least $2\delta$ (and shrinking the upper endpoint of the stated $\delta$ range if necessary). Hence $\mathbb{P}_\nu(\mathsf{Reg}_T > x_\delta) > \delta$ for the constructed instance, which implies $Q_{1-\delta}(\mathsf{Reg}_T) \geq x_\delta$ under the usual lower-quantile convention.

For realized shortfall, use

$$\{\widehat{\mathsf{Reg}}_T > x_1 + x_2\} \subseteq \{\mathsf{Reg}_T > x_1\} \cup \{-M_T > x_2\}.$$

Choose

$$x_1 = C_{\varepsilon,\tau} \rho R_T^\star + C_{\varepsilon,\tau} R_T^\star \rho^{-\varepsilon} \log^\varepsilon\left( \frac{C_{\varepsilon,\tau} K}{\delta} \right) \quad \text{and} \quad x_2 = C_{\varepsilon,\tau} \left( \frac{vT}{\delta} \right)^{1/(1+\varepsilon)}.$$

The first part of the proof gives $\mathbb{P}(\mathsf{Reg}_T > x_1) \leq \delta/2$ after increasing constants, while Markov's inequality applied to $\mathbb{E}|M_T|^{1+\varepsilon} \leq C_\varepsilon vT$ gives $\mathbb{P}(-M_T > x_2) \leq \delta/2$. The union bound proves the shortfall claim. For the lower bound on the reward-noise term, apply the oracle construction of Theorem 1, where $\mathsf{Reg}_T = 0$. For sufficiently small $\delta$, the choice $x = c_\varepsilon (vT/\delta)^{1/(1+\varepsilon)}$ gives an instance for which $\mathbb{P}(\widehat{\mathsf{Reg}}_T > x) > \delta$ after adjusting constants. □

### D.3 Realized payoff shortfall

*Proof of Corollary 2.* For any $x > 0$, Proposition 1 gives

$$\mathbb{P}_\nu(\widehat{\mathsf{Reg}}_T > x) \leq \mathbb{P}_\nu(\mathsf{Reg}_T > x/2) + C_\varepsilon \frac{vT}{x^{1+\varepsilon}}.$$

Apply Theorem 2 to the first probability. If $x/2 \leq C_{\varepsilon,\tau} \rho R_T^\star$, then

$$(x - C_{\varepsilon,\tau} \rho R_T^\star)_+ = 0$$

after increasing $C_{\varepsilon,\tau}$, and the exponential term in the corollary is at least a positive constant times $K$, hence it dominates the trivial upper bound one. If $x/2 > C_{\varepsilon,\tau} \rho R_T^\star$, set $y = x/2 - C_{\varepsilon,\tau} \rho R_T^\star > 0$. Theorem 2 yields

$$\mathbb{P}_\nu(\mathsf{Reg}_T > x/2) \leq C_{\varepsilon,\tau} K \exp\left\{ -c_{\varepsilon,\tau} \rho \left( \frac{x - C_{\varepsilon,\tau} \rho R_T^\star}{R_T^\star} \right)^{1/\varepsilon} \right\},$$

after relabeling constants. Combining the two cases gives the displayed upper bound. The lower bound of Theorem 1 applies to the oracle policy, for which $\mathsf{Reg}_T = 0$ and $\widehat{\mathsf{Reg}}_T = -M_T$; therefore the polynomial term cannot be removed uniformly. □

### D.4 Incompatibility with logarithmic instance-dependent regret

*Proof of Corollary 3.* Assume, toward a contradiction, that a policy satisfies both properties. Fix one of the embedded base instances in Theorem 5, with $\Delta_0$ small enough that $\Delta_0 T/4$ lies in the constant-fraction range where property (ii) is assumed to hold. On the base instance, every suboptimal arm has gap $\Delta_0$, so property (i) implies $\mathbb{E}_{\nu^0} N_2(T) \le \mathbb{E}_{\nu^0} \mathsf{Reg}_T/\Delta_0 = O(\log T)$. Also, $C_0 R_T^\star < \Delta_0 T/4$ for sufficiently large $T$. The theorem therefore applies and produces a neighboring finite-moment instance $\nu^1$ such that, for all sufficiently large $T$,

$$\mathbb{P}_{\nu^1}\left(\mathsf{Reg}_T^\pi > \frac{\Delta_0 T}{4}\right) \ge cT^{-C}.$$

Property (ii), applied with $z = \Delta_0 T/4$, gives on the same instance

$$\mathbb{P}_{\nu^1}\left(\mathsf{Reg}_T^\pi > \frac{\Delta_0 T}{4}\right) \le C\exp\left\{-c\left(\frac{\Delta_0 T}{4R_T^\star}\right)^{1/\varepsilon}\right\}.$$

For fixed $K, \Delta_0, v, \varepsilon$, we have $R_T^\star = v^{1/(1+\varepsilon)} K^{\varepsilon/(1+\varepsilon)} T^{1/(1+\varepsilon)}$, so the exponent is a positive constant times $T^{1/(1+\varepsilon)}$. Thus the upper bound is $C\exp\{-c'T^{1/(1+\varepsilon)}\}$, which is eventually smaller than $cT^{-C}$. □

## E A Conditional Transfer for Unknown Moment Parameters

Suppose a confidence interval can be computed without knowing $(v, \varepsilon)$ and satisfies the coverage and width bounds below. Proposition 4 gives the resulting regret guarantee. The existence of such an interval over the full finite-moment class is an additional assumption.

**Adaptive confidence intervals.** Fix a coarse parameter range

$$\varepsilon \in [\varepsilon_{\min}, 1], \qquad 1 \le v \le V_{\max},$$

where $\varepsilon_{\min} > 0$ is known. With $L = \lfloor T/K \rfloor$, let $\mathcal{N}_T = \{2^s : 2^s \le L\} \cup \{L\}$ be the dyadic sample-size grid with the possibly capped terminal phase, and write

$$q_T(v, \varepsilon, K) = v^{1/(1+\varepsilon)}(T/K)^{1/(1+\varepsilon)}, \qquad R_T^\star(v, \varepsilon, K) = K q_T(v, \varepsilon, K).$$

For each $n \in \mathcal{N}_T$ and $\rho \ge 1$, the interval procedure returns an estimate-radius pair $(\widetilde{\mu}_{n,\rho}, \widetilde{r}_{n,\rho})$ without using the true $(v, \varepsilon)$.

**Assumption 2** (Honest adaptive robust confidence)**.** *There are constants $c_A, C_A > 0$, depending only on $\varepsilon_{\min}$, and complexity factors $\mathsf{L}_T, \mathsf{H}_T \ge 1$ that are polylogarithmic in $(T, K, V_{\max}, 1/\varepsilon_{\min})$, such that the following holds. For every distribution with mean $\mu$ and centered $(1+\varepsilon)$ moment at most $v$, with $\varepsilon \in [\varepsilon_{\min}, 1]$ and $1 \le v \le V_{\max}$, the primitive satisfies, for every $n \in \mathcal{N}_T$ and every $\rho \ge 1$,*

$$\mathbb{P}\left(|\widetilde{\mu}_{n,\rho} - \mu| > \widetilde{r}_{n,\rho} \text{ or } \widetilde{r}_{n,\rho} > C_A \mathsf{L}_T \rho \frac{q_T(v, \varepsilon, K)}{n}\right)$$
$$\le C_A \mathsf{H}_T \exp\left\{-c_A \rho^{(1+\varepsilon)/\varepsilon}\left(\frac{T}{Kn}\right)^{1/\varepsilon}\right\}.$$

Coverage prevents elimination of the optimal arm; the width bound controls how many samples are needed to eliminate suboptimal arms.

**Elimination rule.** The wrapper uses the dyadic phases of Algorithm 1 and the continuation convention in Appendix B.3 when one arm remains. At phase $s$, compute $(\widetilde{\mu}_{i,s,\rho}, \widetilde{r}_{i,s,\rho})$ for each active arm using the fixed $\rho$. Eliminate arm $i$ if

$$\widetilde{\mu}_{i,s,\rho} + \tau \widetilde{r}_{i,s,\rho} < \max_{j \in A_s}\left\{\widetilde{\mu}_{j,s,\rho} - \tau \widetilde{r}_{j,s,\rho}\right\}, \qquad \tau > 1.$$

**Proposition 4** (Conditional adaptive transfer)**.** *Suppose Assumption 2 holds. Fix $\varepsilon_{\min} > 0$, $\rho \geq 1$, and $\tau > 1$. For every true instance $\nu \in \mathcal{H}_K(v, \varepsilon)$ with unknown $\varepsilon \in [\varepsilon_{\min}, 1]$ and $v \leq V_{\max}$, the wrapper satisfies, for all $x > 0$,*

$$\mathbb{P}_\nu(\mathsf{Reg}_T > C_{\varepsilon_{\min},\tau}\rho \mathsf{L}_T R_T^\star(v,\varepsilon,K) + x) \leq C_{\varepsilon_{\min},\tau} K \mathsf{H}_T \exp\left\{-c_{\varepsilon_{\min},\tau}\rho\left(\frac{x}{\mathsf{L}_T R_T^\star(v,\varepsilon,K)}\right)^{1/\varepsilon}\right\}.$$

*Consequently,*

$$\mathbb{E}_\nu \mathsf{Reg}_T \leq C_{\varepsilon_{\min},\tau}\big(\rho \mathsf{L}_T R_T^\star(v,\varepsilon,K) + K\big) + C_{\varepsilon_{\min},\tau}\mathsf{L}_T R_T^\star(v,\varepsilon,K)\rho^{-\varepsilon}\left[1 + \log^{\varepsilon}(C_{\varepsilon_{\min},\tau}K\mathsf{H}_T)\right].$$

*Proof.* Put

$$q_T = q_T(v,\varepsilon,K), \qquad R_T^\star = Kq_T, \qquad b_s = C_A \mathsf{L}_T \rho q_T / n_s.$$

Fix a sufficiently large constant $D_{\varepsilon_{\min},\tau}$. If $D_{\varepsilon_{\min},\tau}\rho\mathsf{L}_T R_T^\star \geq T$, the result is trivial after enlarging the leading constant because $\mathsf{Reg}_T \leq T$. Assume henceforth that $D_{\varepsilon_{\min},\tau}\rho\mathsf{L}_T R_T^\star < T$.

For a phase $s$, let $\mathcal{E}_s$ be the event that, for every arm estimated at that phase,

$$|\widetilde{\mu}_{i,s,\rho} - \mu_i| \leq \widetilde{r}_{i,s,\rho} \quad \text{and} \quad \widetilde{r}_{i,s,\rho} \leq b_s.$$

By Assumption 2 and a union bound over arms,

$$\mathbb{P}(\mathcal{E}_s^c) \leq C_{\varepsilon_{\min}} K \mathsf{H}_T \exp\left\{-c_{\varepsilon_{\min}}\rho^{(1+\varepsilon)/\varepsilon}\left(\frac{T}{Kn_s}\right)^{1/\varepsilon}\right\}. \tag{E.1}$$

On $\mathcal{E}_s$, an optimal arm is retained. Moreover, every active arm with $\Delta_i > 2(\tau+1)b_s$ is eliminated, because

$$\widetilde{\mu}_{i,s,\rho} + \tau\widetilde{r}_{i,s,\rho} \leq \mu^\star - \Delta_i + (\tau+1)b_s < \mu^\star - (\tau+1)b_s \leq \widetilde{\mu}_{i^\star,s,\rho} - \tau\widetilde{r}_{i^\star,s,\rho}.$$

Fix $y > 0$ and set $\eta = y/(64T)$. Let $s_y$ be the first phase such that $2(\tau+1)b_{s_y} \leq \eta$. As in the proof of Lemma 2, if $\mathcal{E}_s$ holds for every phase $s \leq s_y$, then every arm with $\Delta_i > \eta$ is eliminated after at most $C_{\varepsilon_{\min},\tau}\mathsf{L}_T\rho q_T/\Delta_i$ pulls, and arms with $\Delta_i \leq \eta$ contribute at most $\eta T$. Therefore

$$\mathsf{Reg}_T \leq C_{\varepsilon_{\min},\tau}\rho\mathsf{L}_T R_T^\star + y/4. \tag{E.2}$$

Take $y = \max\{x, D_{\varepsilon_{\min},\tau}\rho\mathsf{L}_T R_T^\star\}$. If $x \geq T$, the target regret event is empty, so assume $x < T$. The nontrivial condition above gives $y < T$, and a large enough $D_{\varepsilon_{\min},\tau}$ ensures that $s_y$ exists by the final phase. Dyadicity then gives

$$n_{s_y} \asymp_{\varepsilon_{\min},\tau} \frac{\rho\mathsf{L}_T q_T}{\eta} = \frac{\rho\mathsf{L}_T R_T^\star}{y}\,\frac{T}{K}. \tag{E.3}$$

For $s \leq s_y$, write $m = s_y - s$. The possible capped terminal phase costs at most a factor two, so $n_s \leq 2^{1-m}n_{s_y}$. Substitution of (E.3) into (E.1) yields

$$\mathbb{P}(\mathcal{E}_s^c) \leq C_{\varepsilon_{\min}} K \mathsf{H}_T \exp\left\{-c_{\varepsilon_{\min},\tau}2^{m/\varepsilon}\rho\left(\frac{y}{\mathsf{L}_T R_T^\star}\right)^{1/\varepsilon}\right\}.$$

Summing over $m \geq 0$ yields

$$\mathbb{P}\left(\bigcup_{s \leq s_y}\mathcal{E}_s^c\right) \leq C_{\varepsilon_{\min},\tau} K \mathsf{H}_T \exp\left\{-c_{\varepsilon_{\min},\tau}\rho\left(\frac{y}{\mathsf{L}_T R_T^\star}\right)^{1/\varepsilon}\right\}. \tag{E.4}$$

Equations (E.2)–(E.4), together with $y \geq x$ and $y \geq D_{\varepsilon_{\min},\tau}\rho\mathsf{L}_T R_T^\star$, prove the tail claim after adjusting constants. Applying Lemma 3 with $R = \mathsf{L}_T R_T^\star$ and $b = C_{\varepsilon_{\min},\tau}K\mathsf{H}_T$, then adding the offset and the initial $K$ pulls, proves the expectation claim. □

Constructing intervals that satisfy Assumption 2 with polylogarithmic losses remains an estimation problem. Genalti et al. (2024) study the limitations of unknown-moment adaptation, while Tamás et al. (2024) obtain parameter-free robust UCBs under additional distributional assumptions. The budget lower bound also applies to adaptive policies.

## F Detailed Comparison with Related Work

Table 1 compares the reward assumptions and regret criteria of related results. For fixed $K$, BCRR has expected pseudo-regret of order $\rho R_T^\star$ and satisfies

$$\mathbb{P}(\mathsf{Reg}_T > C\rho R_T^\star + x) \le CK \exp\left\{-c\rho \left(\frac{x}{R_T^\star}\right)^{1/\varepsilon}\right\},$$

and Theorem 4 gives the converse for policies with worst-case expected pseudo-regret at most $B_T$, on its stated threshold range:

$$\sup_{\nu \in \mathcal{H}_K(v,\varepsilon)} \mathbb{P}_\nu(\mathsf{Reg}_T^\pi > x) \ge \exp\left\{-C\frac{B_T}{R_T^\star}\left(\frac{x}{R_T^\star}\right)^{1/\varepsilon}\right\}.$$

The coefficient of the negative-log tail rate is therefore $B_T/R_T^\star$, up to constants.

| Work | Setting and object | Main guarantee |
|---|---|---|
| Simchi-Levi et al. | Regret-distribution tradeoffs for sub-Gaussian bandits | Tradeoffs between expected regret and regret tails under light-tailed rewards |
| Lee and Oh | All-confidence distributional regret for bandits and RL under sub-exponential concentration | Regret bounds at all confidence levels under exponential concentration |
| Heavy-tailed expected-regret work | Expected regret under finite moments | Expected-regret bounds at the finite-moment minimax scale $R_T^\star$ |
| Panda and Agrawal | Asymptotic, instance-dependent regret tails for broad reward classes | Asymptotic tail rates on fixed instances |
| This paper | Finite centered $(1+\varepsilon)$ moment; nonasymptotic worst-case guarantees | Matching pseudo-regret tail bounds for a prescribed expected-regret budget |

Table 1: Reward assumptions and regret guarantees in related work.

**Light-tailed regret distributions.** Simchi-Levi et al. (2022; 2025) and Lee & Oh (2026) derive distributional regret guarantees under light-tailed reward assumptions. Exponential concentration makes empirical means reliable at a logarithmic confidence cost. Under a finite-moment assumption, even an oracle can incur the polynomial shortfall of Theorem 1:

$$\mathbb{P}(-M_T > x) \gtrsim \frac{vT}{x^{1+\varepsilon}} \qquad \text{for } x \gtrsim v^{1/(1+\varepsilon)} T^{1/(1+\varepsilon)}.$$

This rules out a uniform stretched-exponential tail for realized shortfall. The decomposition

$$\widehat{\mathsf{Reg}}_T = \mathsf{Reg}_T - M_T$$

separates decision regret from reward noise.

The estimation radius also changes. A sub-Gaussian mean estimate has radius of order $\sqrt{u/n}$ at failure probability $e^{-u}$. Under a finite $p = 1+\varepsilon$ moment, the robust estimate satisfies

$$r(n,u) = v^{1/p}\left(\frac{u}{n}\right)^{(p-1)/p}, \qquad \mathbb{P}(|\widehat{\mu}_{n,u} - \mu| > Cr(n,u)) \le 2e^{-cu},$$

but only while $u \lesssim n$. Matching this radius to the BCRR elimination scale gives

$$r(n,u) \asymp \rho\frac{q_T}{n} \quad \Longrightarrow \quad u = \rho^{(1+\varepsilon)/\varepsilon}\left(\frac{T}{Kn}\right)^{1/\varepsilon},$$

and the validity condition $u \lesssim n$ forces the exploration floor

$$n \gtrsim \rho\left(\frac{T}{K}\right)^{1/(1+\varepsilon)}.$$

BCRR delays elimination until this sample requirement is met.

**Heavy-tailed expected-regret bandits.** Heavy-tailed bandit algorithms attain the expected-regret scale $R_T^\star$ from Section 3, up to constants and, for some algorithms, logarithmic factors. Since regret is nonnegative,

$$\mathbb{E}[\mathsf{Reg}_T] = \int_0^\infty \mathbb{P}(\mathsf{Reg}_T > x)\, dx.$$

An expectation bound restricts this integral but leaves the tail shape undetermined. For a budget $B_T \asymp \rho R_T^\star$, Theorems 2 and 4 identify the negative-log tail rate on the stated threshold range:

$$\rho \left(\frac{x}{R_T^\star}\right)^{1/\varepsilon} = \frac{B_T}{R_T^\star}\left(\frac{x}{R_T^\star}\right)^{1/\varepsilon} \quad \text{up to constants.}$$

Improving this rate by a factor $\rho$ requires an expected-regret budget of order $\rho R_T^\star$.

**Asymptotic and instance-dependent tail results.** Panda & Agrawal (2026) study asymptotic regret tails on fixed instances, including moment-bounded models. Our bounds hold at finite horizons, uniformly over the moment class, for a prescribed expected-regret budget.

The incompatibility between logarithmic instance-dependent regret and uniform tail protection also occurs under light-tailed rewards. In the finite-moment setting, Theorem 5 shows that if the base instance satisfies

$$\mathbb{E}_{\nu^0}[N_2(T)] \le a \log T,$$

then a neighboring instance satisfies

$$\mathbb{P}_{\nu^1}\left(\mathsf{Reg}_T^\pi > \frac{\Delta_0 T}{4}\right) \ge cT^{-C}.$$

This polynomial lower bound exceeds the minimax tail rate at linear thresholds, which for fixed $K$ and $v$ is

$$\exp\left\{-c\left(\frac{T}{R_T^\star}\right)^{1/\varepsilon}\right\} = \exp\left\{-cT^{1/(1+\varepsilon)}\right\}.$$

## G Additional Numerical Sensitivity

The main figure reports raw pseudo-regret tails. Here, panels labeled *excess* subtract the sample median of pseudo-regret within each policy configuration; their $99\%$ excess quantile is the sample $99\%$ quantile minus that median. Other pseudo-regret quantiles use the raw scale.

Figure 2 compares algorithms at similar mean pseudo-regret and examines sensitivity to BCRR's tuning parameters. Panel A plots mean pseudo-regret against the $99\%$ excess quantile, varying $\rho \in \{1, 1.5, 2, 3, 4\}$ for BCRR and the radius multipliers for MoM-UCB and Mean-UCB. Panel B shows excess tails at the default radii; panels C–D vary the BCRR radius multiplier and elimination threshold. Default-radius comparisons use $1{,}000$ replications, and the sweep in panel A uses $800$ replications with common seeds.

Figure 3 uses empirical financial residuals in place of synthetic Pareto noise. We center five daily momentum-portfolio series from the Kenneth French Data Library (French, 2026), pool their residuals to estimate a robust MAD scale, and rescale them to the noise level of the main simulation. This experiment uses five arms with means $(.50, .46, .43, .40, .36)$. Each episode samples historical dates with replacement and uses the same date across arms to preserve cross-sectional market dependence. In panels B–C, larger $\rho$ increases mean pseudo-regret and reduces the median-centered upper tail. Panel D compares BCRR with UCB baselines at similar mean pseudo-regret.

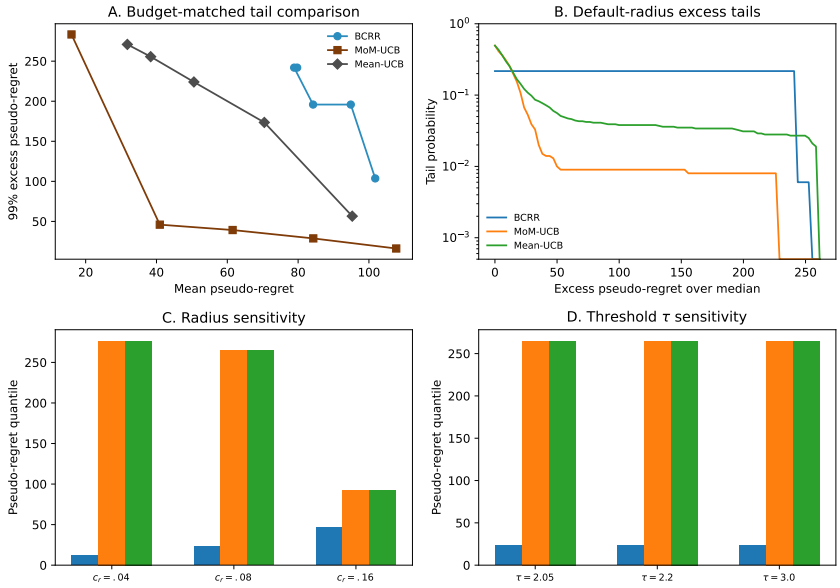



Figure 2: Budget-matched and tuning diagnostics under the main Pareto setting. A: mean pseudo-regret versus the 99% excess quantile for BCRR and radius-inflated UCB baselines. B: default-radius excess tails. C: sensitivity to the BCRR radius multiplier. D: sensitivity to the elimination threshold $\tau$.

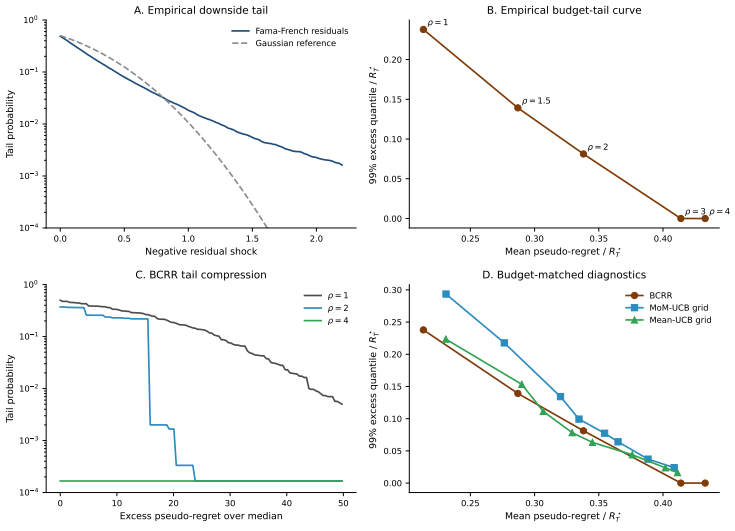



Figure 3: Budget–tail pattern under empirical financial residuals. A: downside tail of Fama–French momentum-portfolio residuals, compared with a Gaussian reference of the same standard deviation. B: mean pseudo-regret versus the 99% excess quantile as $\rho$ varies. C: median-centered BCRR pseudo-regret tails for $\rho \in \{1, 2, 4\}$. D: budget-matched UCB diagnostic on the same residual episodes.

Figure 4 varies the moment parameters supplied to BCRR while holding the environment fixed. In panels A–B, understating $v$ narrows the racing radius and produces a heavier excess tail; overstating $v$ widens the radius and reduces that tail in this simulation. Panels C–D vary $\varepsilon$: smaller values assume heavier-tailed rewards and give more conservative elimination. Each configuration uses 2,000 replications with common random seeds across parameter values. The quantile panels report the 50%, 90%, and 95% levels.

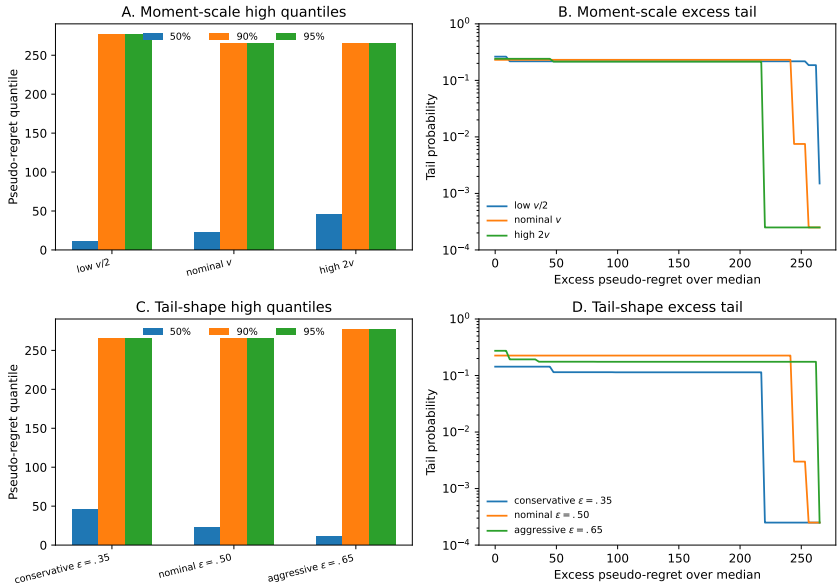



Figure 4: Sensitivity to the moment parameters supplied to BCRR. A–B: pseudo-regret quantiles and excess tails as $v$ varies. C–D: the corresponding plots as $\varepsilon$ varies. The reward distribution is fixed across configurations.

## H AI USE STATEMENT

Generative AI tools assisted with editing and manuscript organization. The authors independently wrote the proofs and take responsibility for the manuscript.