A Sharp Adaptation Frontier for Heavy-Tailed Bandits with Unknown Moments

Pahan Dewasurendra
Johns Hopkins University

Abstract

Heavy-tailed bandit algorithms attain optimal regret when given a finite moment order and its bound. Without either parameter, optimal adaptation is impossible unless one restricts the reward distributions. What can be achieved under the moment condition alone has remained open. We give a sharp answer in the form of a Pareto frontier.

Our algorithm takes only an exploration exponent $\rho \in (0, 1)$. It combines deterministic forced exploration with a median-of-means empirical leader and uses neither the moment order, its bound, nor a tail-sign condition. If the unknown rewards have $(1+\epsilon)$th absolute moments bounded by u, its worst-case regret is, up to explicit arm factors and an arbitrarily slow factor,

$$u^{1/(1+\epsilon)} T^{1-\rho\epsilon/(1+\epsilon)},$$

while its regret on every fixed instance is $O(T^\rho)$. A matching lower bound shows that improving either polynomial exponent worsens the other. Most notably, every fixed $\rho \le 2/3$ attains this frontier simultaneously for all $\epsilon \in (0, 1]$. Thus at any point on this common segment, unknown tail order has no further polynomial price beyond unknown scale. This answers the assumption-free rate question posed at COLT 2025 on the maximal common segment and gives the first matching algorithm for the unknown-scale lower tradeoff.

1 Introduction

Heavy-tailed multi-armed bandits model sequential decisions when rewards may have infinite variance. In the standard formulation, every arm satisfies

$$\mathbb{E}|X|^{1+\epsilon} \le u, \qquad \epsilon \in (0, 1], \quad u > 0. \tag{1}$$

When ϵ and u are known, robust UCB and MOSS variants achieve the optimal worst-case rate $u^{1/(1+\epsilon)} T^{1/(1+\epsilon)}$, with the usual $K^{\epsilon/(1+\epsilon)}$ factor (Bubeck et al., 2013). Wei and Srivastava (2021) sharpen the corresponding minimax construction. Both parameters determine how much of an extreme observation should be trusted.

In practice neither parameter is generally available. Recent work shows a sharp qualitative obstruction. No algorithm can recover the known-parameter rate uniformly over u, or uniformly over ϵ, without further distributional assumptions (Genalti et al., 2024). Algorithms that do recover it assume, for example, a favorable sign for every truncated tail of an optimal arm, as in Huang et al. (2022), Genalti et al. (2024), and Chen et al. (2025). Another route assumes symmetry (Tamás et al., 2024). Such conditions exclude simple one-sided heavy tails.

This leaves a basic question posed explicitly by Genalti and Metelli (2025a): what are the best regret rates when both parameters are unknown and no assumption beyond (1) is imposed? The question has two parts. One needs a lower bound that quantifies the cost and an algorithm that attains it. A recent technical note gives an unknown-u tradeoff between worst-case and fixed-instance rates (Genalti and Metelli, 2025b), but no matching algorithm was known.

We close this gap. The answer is not one rate. It is a Pareto frontier indexed by the desired fixed-instance exponent ρ. Write $p = 1 + \epsilon$. Ignoring slowly varying factors, the frontier is

$$\boxed{\begin{array}{c}\text{fixed instance: } T^\rho, \\ \text{worst case: } u^{1/p} T^{1-\rho(p-1)/p}\end{array}} \tag{2}$$

The two exponents satisfy the exact lower relation

$$\rho + \frac{p}{p-1}\left(1 - \rho\frac{p-1}{p}\right) = \frac{p}{p-1}. \tag{3}$$

Algorithmic idea. The algorithm is deliberately simple. It reserves $K^{1-\rho} t^\rho$ cyclic pulls for exploration

Table 1: Parameter adaptation under a finite pth absolute moment. "Tail sign" denotes truncated nonnegativity for losses or its reward analogue. PF means parameter-free in both p and u.

Method	PF	Extra assumption	Guarantee
Robust UCB	no	none	known-param. optimal
AdaTINF	yes	tail sign	worst-case optimal
AdaR-UCB	yes	tail sign	both near-optimal
uniINF	yes	tail sign	best of both worlds
RMM-UCB	yes	symmetry	both near-optimal
ForceMoM	yes	none	sharp tradeoff

by time t. On all other rounds it chooses the largest median-of-means estimate computed only from these dedicated samples. The number of median blocks grows arbitrarily slowly. This construction avoids the central difficulty faced by adaptive UCB methods. It never needs to turn an unknown moment condition into a numerical confidence radius.

Two properties produce the frontier. On each exploitation round, the expected gap of the empirical leader is at most its mean-estimation error, which is

$$u^{1/p}\left(\frac{K^\rho L(t)}{t^\rho}\right)^{(p-1)/p}$$

for an arbitrarily slow L. Summing gives the worst-case exponent in (2). On a fixed instance, the median block count tends to infinity. The probability of choosing any suboptimal leader is then summable, even though neither p nor u is known. Only forced pulls remain, giving T^ρ regret.

Contributions.

- We give the first assumption-free algorithm with a uniform moment-free finite-time regret bound when both heavy-tail parameters are unknown.
- We prove worst-case and fixed-instance bounds for the same algorithm. Together with the information-theoretic lower bound, they determine the polynomial unknown-scale adaptation frontier.
- One implementation with any $\rho \le 2/3$ is frontier-optimal for every $p \in (1, 2]$ simultaneously. This is the largest segment shared by all moment orders and requires no knowledge of p.

2 Setting and Adaptation Criteria

There are $K \ge 2$ arms. Pulls from arm i are i.i.d. with law ν_i and mean μ_i. At round t, a policy chooses I_t from past actions and observations, then observes an independent draw from ν_{I_t}. Let

$$\mu_* = \max_i \mu_i, \qquad \Delta_i = \mu_* - \mu_i.$$

The expected cumulative regret is

$$R_T(\nu) = \mathbb{E}_\nu \sum_{t=1}^{T} \Delta_{I_t}.$$

For $p \in (1, 2]$ and $u > 0$, let

$$\mathcal{H}_{p,u} = \{\nu : \mathbb{E}_{\nu_i}|X|^p \le u \text{ for all } i\}.$$

The policy is not given p, u, or an upper bound on either. Jensen's inequality gives $|\mu_i| \le u^{1/p}$ and $\Delta_i \le 2u^{1/p}$.

Following Hadiji and Stoltz (2023) and Genalti and Metelli (2025b), a moment-free distribution-free bound Φ_p satisfies

$$\sup_{\nu \in \mathcal{H}_{p,u}} R_T(\nu) \le u^{1/p}\Phi_p(T) \quad \text{for every } u, T. \tag{4}$$

A distribution-dependent rate Ψ satisfies

$$\limsup_{T\to\infty} \frac{R_T(\nu)}{\Psi(T)} < \infty \quad \text{for every fixed } \nu. \tag{5}$$

Constants in (5) may depend on the instance. This distinction is essential. Uniformly logarithmic instance-dependent regret is impossible when the moment scale is unrestricted. This appears both in eventual consistency results (Ashutosh et al., 2021) and in the sharp lower tradeoff (Genalti and Metelli, 2025b).

3 A Parameter-Free Empirical Leader

Fix $\rho \in (0, 1)$. Let $\ell : \mathbb{N} \to [1, \infty)$ be nondecreasing, $\ell(n) \to \infty$, $\ell(n+1) - \ell(n) \le 1$, and $\log \ell(n) = o(\log n)$. The concrete choice

$$\ell(n) = \log\log(e^e + n) \tag{6}$$

is sufficient. Define

$$m_0 = \lceil 16 \log(2K + e) \rceil .$$

The first Km_0 rounds pull every arm m_0 times. These and later forced pulls form a dedicated exploration stream. Exploitation rewards are not used for estimation.

After initialization, set the target total number of exploration pulls to

$$D_t = \max\left\{Km_0, \left\lceil K^{1-\rho} t^\rho \right\rceil\right\}. \tag{7}$$

At round t, force the next arm in cyclic order if fewer than D_t exploration pulls have occurred. Otherwise exploit. The target rises by at most one per round. Indeed, the derivative of $K^{1-\rho}t^\rho$ is at most ρ once $t \geq K$. Induction from the end of initialization shows that the target is met exactly. Cyclic assignment makes arm exploration counts differ by at most one. Consequently, each arm has at least

$$n_t = \lfloor D_{t-1}/K \rfloor \gtrsim (t/K)^\rho \tag{8}$$

dedicated samples before an exploitation decision.

For $n \geq m_0$, let $q(n)$ be the largest odd integer no larger than

$$\min\{n,\, 2\lceil \log(2K) + \ell(n) \rceil + 1\}. \tag{9}$$

Split the first $q\lfloor n/q \rfloor$ samples into q disjoint blocks, average within each block, and take their median. Denote the resulting arm estimate by $M_i(n)$. On an exploitation round, ForceMoM chooses an arm with largest $M_i(n_i)$.

The algorithm is anytime. It needs no numerical tail information. A direct implementation is polynomial. Store arm-wise prefix sums and cache estimates between exploration updates. Recomputing the $q(n)$ block means and their median after a new dedicated sample costs $O(q(n))$ with linear-time median selection. Through n dedicated samples per arm, this is $O(Kn\{\log K + \ell(n)\})$ work and $O(Kn)$ storage. Each leader scan costs $O(K)$.

3.1 A parameter-free MoM bound

The same estimator works for every unknown p because its tuning is only a block count.

Lemma 1 (Median-of-means envelope)**.** *Let $X_1, \dots, X_n$ be i.i.d. with mean μ and $\mathbb{E}|X|^p \leq u$ for some $p \in (1, 2]$. Form a median of $q \geq 3$ odd block means, where $q \leq n$, and write $k = (q+1)/2$. Then, for every $x > 0$,*

$$\Pr\{|M_{n,q} - \mu| > x\} \leq \min\left\{1, \left(\frac{64u(q/n)^{p-1}}{x^p}\right)^k\right\}. \tag{10}$$

Moreover,

$$\mathbb{E}|M_{n,q} - \mu| \leq Cu^{1/p}(q/n)^{(p-1)/p} \tag{11}$$

for a universal constant C. If K arm estimates use block counts with $k \geq \log(2K)$ and common upper bounds q and n^{-1}, then

$$\mathbb{E}\max_{i \leq K} |M_i - \mu_i| \leq Cu^{1/p}(q/n)^{(p-1)/p}. \tag{12}$$

Proof. Jensen and convexity give

$$\mathbb{E}|X - \mu|^p \leq 2^{p-1}\{\mathbb{E}|X|^p + |\mu|^p\} \leq 2^p u.$$

If a block has size $m = \lfloor n/q \rfloor$, the von Bahr–Esseen inequality (von Bahr and Esseen, 1965) yields

$$\mathbb{E}|\bar{X} - \mu|^p \leq 2^{p+1} u m^{1-p} \leq 16u(q/n)^{p-1}.$$

The last step uses $m \geq n/(2q)$ and $p \leq 2$. By Markov's inequality, one block is bad at radius x with probability at most $16u(q/n)^{p-1}x^{-p}$. A bad median requires a set of k bad blocks. Independence of the disjoint blocks and a union bound over at most 2^q sets give (10), since $q = 2k - 1$ and $2^q \leq 4^k$.

Let $x_0 = 64^{1/p}u^{1/p}(q/n)^{(p-1)/p}$. Integrating (10) gives

$$\mathbb{E}|M_{n,q} - \mu| \leq x_0 + \int_{x_0}^\infty (x_0/x)^{pk}\, dx \leq 2x_0,$$

because $q \geq 3$ implies $pk - 1 > 1$. This proves (11).

For $Z = \max_i |M_i - \mu_i|$, the union bound replaces the tail crossover x_0 by at most $K^{1/(pk)}x_0$. Integrating again gives the same factor. It is at most e when $k \geq \log(2K)$. This proves (12) without needing independence across arms. □

4 Worst-Case Regret

The leader rule converts estimation error directly into regret. If J_t is the empirical leader and i_* is optimal, then pointwise

$$\begin{aligned}\Delta_{J_t} &= \mu_* - \mu_{J_t} \\ &\leq |M_{i_*} - \mu_*| + |M_{J_t} - \mu_{J_t}| \\ &\leq 2\max_i |M_i - \mu_i|. \end{aligned} \tag{13}$$

No confidence level is chosen by the algorithm.

Theorem 2 (Finite-time parameter-free regret)**.** *There is a universal constant C such that* ForceMoM*, for every $\rho \in (0,1)$, $p \in (1,2]$, $u > 0$, and $T \geq 1$, satisfies, with $\gamma_p = (p-1)/p$,*

$$\begin{aligned}\sup_{\nu \in \mathcal{H}_{p,u}} R_T(\nu) \leq Cu^{1/p}\big\{&K\log(eK) \\ &+ K^{1-\rho}T^\rho \\ &+ K^{\rho\gamma_p}T^{1-\rho\gamma_p}L_T^{\gamma_p}\big\},\end{aligned} \tag{14}$$

where $L_T = 1 + \log(2K) + \ell(T)$. The policy is independent of p and u.

Proof. There are at most D_T exploration rounds after accounting for the initial constant target. Since every gap is at most $2u^{1/p}$, their contribution is bounded by

$$R_T^E \leq Cu^{1/p}\{K\log(eK) + K^{1-\rho}T^\rho\}. \tag{15}$$

On an exploitation round, all arm sample counts differ by at most one. Equations (9) and (8) therefore permit the common bounds

$$q_i \le CL_t, \qquad n_i \ge cK^{-\rho}t^{\rho}.$$

The initialization guarantees the block-count condition in (12). Combining that equation with (13) gives

$$\mathbb{E}\Delta_{J_t} \le Cu^{1/p}K^{\rho(p-1)/p}L_t^{(p-1)/p}t^{-\rho(p-1)/p}. \quad (16)$$

Put $a = \rho(p-1)/p$. Since $a < 1/2$, monotonicity of L_t and integral comparison imply

$$\sum_{t=1}^{T} L_t^{(p-1)/p}t^{-a} \le 2L_T^{(p-1)/p}T^{1-a}.$$

Summing (16) and adding (15) proves (14). The same enlarged constant covers horizons ending during initialization. □

For $\rho \le 2/3$, the exploitation exponent dominates the exploration exponent for every $p \in (1, 2]$ because

$$\rho \le 1 - \rho(p-1)/p.$$

Thus, up to K and the arbitrarily slow L_T factor, the distribution-free exponent is

$$a_p(\rho) = 1 - \rho\frac{p-1}{p}. \quad (17)$$

5 Fixed Instances and the Sharp Frontier

The slowly increasing block count, unnecessary for the expectation bound, is what makes exploitation mistakes summable on every fixed instance.

Theorem 3 (Fixed-instance regret)**.** *For every fixed bandit instance with a finite pth absolute moment for some $p > 1$,* ForceMoM *satisfies*

$$\sum_{t=1}^{\infty} \Pr\{t \text{ is exploit and } I_t \notin \arg\max_i \mu_i\} < \infty. \quad (18)$$

Consequently,

$$\limsup_{T\to\infty} \frac{R_T(\nu)}{T^{\rho}} \le K^{-\rho}\sum_{i=1}^{K}\Delta_i. \quad (19)$$

Proof. If a suboptimal arm i with gap Δ_i is the empirical leader, either its estimate overshoots by $\Delta_i/2$ or an optimal estimate undershoots by that amount. Lemma 1 bounds each probability by

$$2\left[\frac{Cu\overline{q}_t^{\,p-1}}{\Delta_i^p n_t^{p-1}}\right]^{(\underline{q}_t+1)/2}. \quad (20)$$

Here $n_t \gtrsim t^{\rho}$, the smaller block count $\underline{q}_t \to \infty$, and the larger count satisfies $\overline{q}_t \le \underline{q}_t + 2$ and $\log \overline{q}_t = o(\log n_t)$. For fixed p, u, Δ_i, the logarithm of the bracket in (20) is eventually at most $-(p-1)\log n_t/2$. Since $n_t \gtrsim t^{\rho}$, it is then at most $-(p-1)\rho\log t/4$. Finally, $\underline{q}_t \to \infty$, so the complete right side is at most $2t^{-3}$ for all sufficiently large t. Summing over time and over the finitely many suboptimal arms proves (18).

It remains to count forced pulls. Cyclic assignment gives each arm $K^{-1}D_T + O(1)$ such pulls. Since $D_T = K^{1-\rho}T^{\rho} + o(T^{\rho})$, their regret divided by T^{ρ} converges to $K^{-\rho}\sum_i \Delta_i$. The exploitation regret is bounded because (18) holds and all gaps are fixed. This proves (19). □

We now compare the two upper bounds with the information-theoretic constraint. The following form is due to Genalti and Metelli (2025b), adapting the range argument of Hadiji and Stoltz (2023). We include a proof in the supplement because it is central to the conclusion.

Theorem 4 (Unknown-scale lower tradeoff)**.** *Fix $p \in (1, 2]$. Suppose a policy, without knowing u, has a moment-free distribution-free bound $\Phi_p(T) = o(T)$ as in (4). Then on every fixed instance ν with at least one suboptimal arm,*

$$\liminf_{T\to\infty} \frac{R_T(\nu)}{(T/\Phi_p(T))^{p/(p-1)}} \ge c_p \sum_{i:\Delta_i>0}\Delta_i, \quad (21)$$

where $c_p > 0$ depends only on p. Hence polynomial exponents a and b in $\Phi_p(T) = T^{a+o(1)}$ and $\Psi(T) = T^{b+o(1)}$ must satisfy

$$b + \frac{p}{p-1}a \ge \frac{p}{p-1}. \quad (22)$$

Proof. Fix a suboptimal arm i with gap Δ_i. For $\beta \in (0, 1/2)$, replace its law by

$$\nu'_{i,\beta} = (1-\beta)\nu_i + \beta\delta_{\mu_i+2\Delta_i/\beta}. \quad (23)$$

Its new mean is $\mu_* + \Delta_i$. Let v be a finite common pth-moment bound for the original instance and take $v'_\beta = v + (1-\beta)\mathbb{E}_{\nu_i}|X|^p + \beta|\mu_i + 2\Delta_i/\beta|^p$. This bounds every arm of the alternative and satisfies

$$(v'_\beta)^{1/p} = 2\Delta_i\beta^{-(p-1)/p}\{1 + o(1)\} \quad (\beta \downarrow 0). \quad (24)$$

Moreover, $\mathrm{KL}(\nu_i, \nu'_{i,\beta}) \le \log(1/(1-\beta)) \le 2\beta\log 2$.

Let P, P' denote the two history laws and let $N_i(T)$ count pulls of arm i. Data processing applied to the action at an independent uniform round, followed by $\mathrm{kl}(x, y) \ge (1-x)\log(1/(1-y)) - \log 2$, gives

$$\left(1 - \frac{v^{1/p}\Phi_p(T)}{T\Delta_i}\right)\log\frac{T\Delta_i}{(v'_\beta)^{1/p}\Phi_p(T)} - \log 2 \le 2\beta\log 2\,\mathbb{E}_P N_i(T). \quad (25)$$

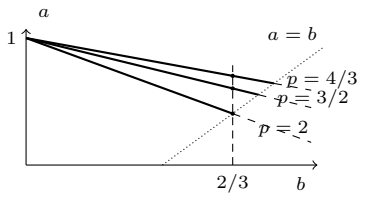

Figure 1: Exponent geometry. The lower bound requires $a \geq 1-b(p-1)/p$. Thick segments are matched by ForceMoM until each line meets $a = b$. The vertical line at $b = 2/3$ marks the endpoint of the largest segment shared by every $p \in (1, 2]$. Its dots show the three orders displayed.

Here the regret guarantee bounds $\mathbb{E}_P N_i(T)$ on the original instance and $T - \mathbb{E}_{P'} N_i(T)$ on the alternative.

Put $r = p/(p-1)$, $\alpha = 8^{-r}$, and $\beta_T = \alpha^{-1}(\Phi_p(T)/T)^r$. By (24), the denominator inside the logarithm in (25) is $\Delta_i T\{1 + o(1)\}/4$. Its left side therefore tends to $\log 2$, while $\beta_T \to 0$. It follows that

$$\liminf_{T\to\infty} \frac{\mathbb{E}_P N_i(T)}{(T/\Phi_p(T))^r} \geq \frac{1}{2\,8^r}.$$

Multiplying by Δ_i and summing over suboptimal arms proves (21). The exponent relation follows by taking powers of T. □

Substituting (17) into (22) forces $b \geq \rho$. Theorem 3 attains $b = \rho$. We have therefore proved the main result.

Corollary 5 (Simultaneous sharp frontier)**.** *Fix any $\rho \in (0, 2/3]$. One policy, independent of p and u, attains for every $p \in (1, 2]$ the polynomial exponents*

$$(a_p, b) = \left(1 - \rho\frac{p-1}{p}, \rho\right).$$

These exponents satisfy the lower tradeoff with equality for every p. The slowly varying gap in (14) *can be made arbitrarily small by the choice of ℓ.*

The same result also closes the entire relevant unknown-scale tradeoff when the order may be used only to choose the design point.

Corollary 6 (Order-aware unknown-scale frontier)**.** *Fix $p \in (1, 2]$ and $\rho \leq p/(2p-1)$. Then* ForceMoM *with design exponent ρ uses no scale information and attains*

$$(a_p, b) = \left(1 - \rho\frac{p-1}{p}, \rho\right),$$

which meets (22) *with equality. The balanced choice $\rho_p^* = p/(2p-1)$ gives, up to L_T,*

$$R_T \leq C u^{1/p} K^{(p-1)/(2p-1)} T^{p/(2p-1)} L_T^{(p-1)/p} \qquad (26)$$

Table 2: Polynomial exponent pairs (a, b) for worst-case and fixed-instance regret. Logarithmic fixed-instance regret is recorded as $b = 0$. The last column uses one policy with $\rho = 2/3$ for every unknown p.

p	known p, u	unknown u, p-aware	both unknown
2	$(1/2, 0)$	$(2/3, 2/3)$	$(2/3, 2/3)$
3/2	$(2/3, 0)$	$(3/4, 3/4)$	$(7/9, 2/3)$
4/3	$(3/4, 0)$	$(4/5, 4/5)$	$(5/6, 2/3)$
$p \downarrow 1$	$(1, 0)$	$(1, 1)$	$(1, 2/3)$

apart from the additive initialization term, and has the same fixed-instance polynomial exponent.

Proof. The restriction on ρ is exactly $\rho \leq 1 - \rho(p-1)/p$, so the exploitation term in Theorem 2 dominates. Substitution in Theorem 4 proves sharpness. At $\rho = p/(2p-1)$, the exploration and exploitation powers of both K and T agree, giving (26). □

If p were known but u were not, balancing the two exponents would use $\rho = p/(2p-1)$ and produce $T^{p/(2p-1)}$. No single fixed ρ of this form works for all p. Corollary 5 instead identifies the entire common frontier. Its largest point is $\rho = 2/3$, where

$$a_p(2/3) = 1 - \frac{2(p-1)}{3p}.$$

For finite variance this gives the sharp balanced $T^{2/3}$ rate. For heavier tails the worst-case rate approaches linear smoothly.

Table 2 separates two claims that can otherwise be confused. Knowing p permits selecting a different balanced point for each order. Without p, the choice $\rho = 2/3$ instead fixes the same instance exponent for all orders and is sharp at that point for every p. The latter does not claim to reproduce the p-aware balanced point when $p < 2$.

There is no contradiction with the separate lower bound for adaptation to p at fixed u (Genalti et al., 2024). That benchmark is the faster known-scale rate $T^{1/p}$. Here the unknown scale has already enlarged the frontier. The same algorithm shows that learning p requires no additional polynomial loss along its common portion.

6 Related Work and Discussion

Known parameters. Bubeck et al. (2013) introduced robust UCB based on truncation, median-of-means, and Catoni estimators. Later algorithms sharpened the worst-case logarithmic factors (Wei and Srivastava, 2021). An optimal index policy was developed by Agrawal et al. (2021), and a randomized policy by

Park et al. (2026). These methods tune their estimators or indices using the moment parameters. In particular, H-BE requires both p and u despite its closed-form action probabilities (Park et al., 2026). Recent adversarial best-of-both-worlds methods can remove the tail-sign condition, but their clipping and learning rates explicitly use the moment order and scale (Zhao et al., 2025).

Parameter-free methods with assumptions. AdaTINF and AdaR-UCB adapt to both parameters under truncated tail-sign conditions in Huang et al. (2022) and Genalti et al. (2024). uniINF obtains a best-of-both-worlds guarantee under the same condition (Chen et al., 2025). RMM-UCB instead assumes symmetry (Tamás et al., 2024). These results seek the known-parameter optimum. Our goal is different. We impose no shape or sign condition and characterize the unavoidable loss.

Unrestricted adaptation. R-UCB-MoM is parameter-oblivious and eventually consistent over a broad class, but its guarantee starts after an instance-dependent threshold and does not provide a finite-time moment-free worst-case rate (Ashutosh et al., 2021). The forced-exploration architecture itself has been studied for sub-Gaussian rewards (Han et al., 2024). Replacing the empirical mean is not enough for our result. The diverging MoM block count and dedicated exploration stream are what yield both a uniform finite-time bound and summable fixed-instance errors under every unknown $p > 1$.

Scope. The frontier is polynomial. Arbitrarily slow factors remain between the upper bound and (21). Removing them while retaining one algorithm for all p is open. It is also open whether one parameter-free policy can adapt among the order-specific frontier points beyond the common range $\rho \leq 2/3$. Our result concerns stochastic finite-armed bandits. Adversarial, contextual, and linear variants require different tools.

7 Conclusion

Unknown heavy-tail parameters do not make regret minimization hopeless, but they change its objective. There is an exact tradeoff between protection against an unrestricted moment scale and fast learning on fixed instances. A forced-exploration median leader attains that tradeoff without estimating either tail parameter. The same policy is exponent-optimal for every finite moment order, resolving the assumption-free adaptation question on the common Pareto frontier.

References

Agrawal, S., Juneja, S. K., and Koolen, W. M. (2021). Regret minimization in heavy-tailed bandits. In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 26–62. PMLR.

Ashutosh, K., Nair, J., Kagrecha, A., and Jagannathan, K. (2021). Bandit algorithms: Letting go of logarithmic regret for statistical robustness. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 622–630. PMLR.

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

Chen, Y., Huang, J., Dai, Y., and Huang, L. (2025). uniINF: Best-of-both-worlds algorithm for parameter-free heavy-tailed MABs. In *International Conference on Learning Representations*.

Genalti, G., Marsigli, L., Gatti, N., and Metelli, A. M. (2024). (ϵ, u)-adaptive regret minimization in heavy-tailed bandits. In *Proceedings of the 37th Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 1882–1915. PMLR.

Genalti, G. and Metelli, A. M. (2025a). Open problem: Regret minimization in heavy-tailed bandits with unknown distributional parameters. In *Proceedings of the 38th Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 1–5. PMLR.

Genalti, G. and Metelli, A. M. (2025b). A regret lower bound for u-adaptive heavy-tailed bandits. Technical report, Politecnico di Milano.

Hadiji, H. and Stoltz, G. (2023). Adaptation to the range in k-armed bandits. *Journal of Machine Learning Research*, 24(13):1–33.

Han, Q., Zhu, L., and Guo, F. (2024). Forced exploration in bandit problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12270–12277.

Huang, J., Dai, Y., and Huang, L. (2022). Adaptive best-of-both-worlds algorithm for heavy-tailed multi-armed bandits. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9173–9200. PMLR.

Park, H.-j., Cho, Y.-S., and Lee, K. (2026). Boltzmann exploration for heavy-tailed bandits. In *Proceedings of the 29th International Conference on Artificial*

Intelligence and Statistics, Proceedings of Machine Learning Research. PMLR.

Tamás, A., Szentpéteri, S., and Csáji, B. C. (2024). Data-driven upper confidence bounds with near-optimal regret for heavy-tailed bandits. *arXiv preprint arXiv:2406.05710*.

von Bahr, B. and Esseen, C.-G. (1965). Inequalities for the rth absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, 36(1):299–303.

Wei, L. and Srivastava, V. (2021). Minimax policy for heavy-tailed bandits. *IEEE Control Systems Letters*, 5(4):1423–1428.

Zhao, C., Ito, S., and Li, S. (2025). Heavy-tailed linear bandits: Adversarial robustness, best-of-both-worlds, and beyond. *arXiv preprint arXiv:2508.13679*.

This supplement gives complete proofs of all results in the main paper. It also reproves the unknown-scale lower tradeoff with a corrected constant.

A Algorithm and Exploration Counts

Fix $K \geq 2$ and $\rho \in (0,1)$. Let

$$m_0 = \lceil 16 \log(2K + e) \rceil, \qquad t_0 = K m_0.$$

The algorithm first pulls each arm m_0 times in cyclic order. These pulls are marked as exploration. For $t > t_0$, define

$$D_t = \max\{t_0, \lceil K^{1-\rho} t^{\rho} \rceil\}.$$

If the exploration count before round t is smaller than D_t, the algorithm explores the next arm in cyclic order. Otherwise it exploits.

Let F_t be the total number of exploration pulls through round t, and let $N_i^E(t)$ be the number assigned to arm i. The initialization has $F_{t_0} = D_{t_0} = t_0$.

Lemma 7 (Exact tracking)**.** *For every $t \geq t_0$, $F_t = D_t$. Moreover,*

$$\left\lfloor \frac{D_t}{K} \right\rfloor \leq N_i^E(t) \leq \left\lceil \frac{D_t}{K} \right\rceil \tag{27}$$

for every arm i.

Proof. Let $g(t) = K^{1-\rho} t^{\rho}$. For $t \geq K$,

$$g'(t) = \rho K^{1-\rho} t^{\rho-1} \leq \rho < 1.$$

Hence $g(t) - g(t-1) < 1$ and $\lceil g(t) \rceil - \lceil g(t-1) \rceil \in \{0, 1\}$. Taking the maximum with the constant t_0 preserves this property. Since $g(t_0) = K m_0^{\rho} \leq K m_0 = t_0$, we have $D_{t_0} = t_0$.

Inductively, suppose $F_{t-1} = D_{t-1}$. If $D_t = D_{t-1}$, the condition for forced exploration is false and $F_t = D_t$. If $D_t = D_{t-1} + 1$, the condition is true, so exactly one exploration pull occurs and again $F_t = D_t$. This proves exact tracking. Cyclic assignment makes the arm counts differ by at most one, which gives (27). □

Before any exploitation decision at $t > t_0$, every arm therefore has at least

$$n_t := \lfloor D_{t-1}/K \rfloor \geq \left\lfloor K^{-\rho} (t-1)^{\rho} \right\rfloor. \tag{28}$$

Since $t_0/K = m_0 \geq 31$, the quantity inside the last floor is at least $m_0^{\rho} \geq 1$. Since $\lfloor x \rfloor \geq x/2$ for every $x \geq 1$, we have uniformly for $t > t_0$ that

$$n_t \geq \frac{1}{2} K^{-\rho} (t-1)^{\rho}. \tag{29}$$

The exploration regret admits both uniform and instance-specific bounds. Jensen's inequality implies $\Delta_i \leq 2u^{1/p}$, so

$$R_T^E \leq 2u^{1/p} D_T \leq 2u^{1/p} \{K m_0 + K^{1-\rho} T^{\rho} + 1\}. \tag{30}$$

On a fixed instance, cyclic assignment and Lemma 7 give

$$\lim_{T \to \infty} \frac{R_T^E}{T^{\rho}} = K^{-\rho} \sum_{i=1}^{K} \Delta_i. \tag{31}$$

B Median-of-Means Bounds

We prove the estimator lemma uniformly for all $p \in (1, 2]$. Let $X_1, \ldots, X_n$ be i.i.d., let $\mu = \mathbb{E}X$, and assume $\mathbb{E}|X|^p \leq u$. Let $q \leq n$ be odd and put $k = (q+1)/2$ and $m = \lfloor n/q \rfloor$. Form q block averages $\bar{X}_1, \ldots, \bar{X}_q$ from qm samples and let $M_{n,q}$ be their median.

Lemma 8 (One block)**.** *For every block j,*

$$\mathbb{E}|\bar{X}_j - \mu|^p \leq 16u(q/n)^{p-1}. \tag{32}$$

Proof. By Jensen, $|\mu|^p \leq u$. Convexity gives

$$\mathbb{E}|X - \mu|^p \leq 2^{p-1} \{\mathbb{E}|X|^p + |\mu|^p\} \leq 2^p u.$$

The von Bahr–Esseen inequality (von Bahr and Esseen, 1965) yields

$$\begin{aligned} \mathbb{E}|\bar{X}_j - \mu|^p &\leq m^{-p} \, 2m \mathbb{E}|X - \mu|^p \\ &\leq 2^{p+1} u m^{1-p}. \end{aligned}$$

Since $n/q \geq 1$, $m = \lfloor n/q \rfloor \geq n/(2q)$. Also $2^{p+1} 2^{p-1} = 2^{2p} \leq 16$. Substitution proves (32). □

Lemma 9 (MoM tail and mean)**.** *For every $x > 0$,*

$$\Pr\{|M_{n,q} - \mu| > x\} \le \min\left\{1, \left(\frac{64u(q/n)^{p-1}}{x^p}\right)^k\right\}. \tag{33}$$

If $q \ge 3$, then

$$\mathbb{E}|M_{n,q} - \mu| \le 128u^{1/p}(q/n)^{(p-1)/p}. \tag{34}$$

Proof. By Lemma 8 and Markov's inequality, the probability that one block differs from μ by more than x is at most

$$\theta_x = \min\{1, 16u(q/n)^{p-1}x^{-p}\}.$$

If the median differs by more than x, at least k blocks do. Their independence and a union bound over subsets of size k give

$$\Pr\{|M_{n,q} - \mu| > x\} \le 2^q \theta_x^k.$$

Since $q = 2k - 1$, $2^q \le 4^k$. This proves (33).

Set

$$x_0 = 64^{1/p}u^{1/p}(q/n)^{(p-1)/p}.$$

Integrating (33) gives

$$\begin{aligned}\mathbb{E}|M_{n,q} - \mu| &\le x_0 + \int_{x_0}^{\infty} (x_0/x)^{pk}\,dx \\ &= x_0\left(1 + \frac{1}{pk-1}\right).\end{aligned}$$

For $q \ge 3$, $k \ge 2$ and $pk - 1 > 1$. Hence the last display is at most $2x_0 \le 128u^{1/p}(q/n)^{(p-1)/p}$. □

Lemma 10 (Maximum over arms)**.** *Suppose K arm estimators satisfy* (33) *with arm-specific q_i, n_i, k_i, where $q_i \le q$, $n_i \ge n$, and $k_i \ge \log(2K)$. Then*

$$\mathbb{E}\max_{i \le K}|M_i - \mu_i| \le 128e\,u^{1/p}(q/n)^{(p-1)/p}. \tag{35}$$

No independence across arms is required.

Proof. Let $k_0 = \min_i k_i$ and $A = 64u(q/n)^{p-1}$. The union bound and (33) imply

$$\Pr\{\max_i |M_i - \mu_i| > x\} \le \min\{1, K(A/x^p)^{k_0}\}$$

whenever $x^p \ge A$. Let $x_0 = A^{1/p}K^{1/(pk_0)}$. Tail integration as in the previous proof gives an expectation at most $2x_0$. Since $k_0 \ge \log(2K)$,

$$K^{1/(pk_0)} \le \exp\{\log K/\log(2K)\} \le e.$$

Also $2A^{1/p} \le 128u^{1/p}(q/n)^{(p-1)/p}$. □

For the algorithm, let $\ell : \mathbb{N} \to [1, \infty)$ be nondecreasing with $\ell(n) \to \infty$, $\ell(n+1) - \ell(n) \le 1$, and $\log \ell(n) = o(\log n)$. Define

$$\begin{aligned}q(n) = \max\{r : r \le n,\ r \le 2\lceil \log(2K) + \ell(n)\rceil + 1,\\ r \text{ is odd}\}.\end{aligned}$$

The assumed schedule is nondecreasing and satisfies $\ell(n+1) - \ell(n) \le 1$. Hence $q(n+1) - q(n) \le 2$. This is the only regularity property of the block schedule needed below. For $n \ge m_0$, both arguments of the minimum exceed $2\log(2K)+1$. Thus $q(n) \ge 2\log(2K) - 1$, and $k(n) = (q(n)+1)/2 \ge \log(2K)$. Also

$$q(n) \le 3 + 2\log(2K) + 2\ell(n). \tag{36}$$

C Proof of the Finite-Time Upper Bound

Let J_t be the arm chosen on an exploitation round and fix an optimal arm i_*. Since $M_{J_t} \ge M_{i_*}$,

$$\begin{aligned}\Delta_{J_t} &= \mu_* - \mu_{J_t} \\ &\le |M_{i_*} - \mu_*| + |M_{J_t} - \mu_{J_t}| \\ &\le 2\max_i |M_i - \mu_i|.\end{aligned} \tag{37}$$

The exploration counts of any two arms differ by at most one. Applying Lemma 10, (29), and (36), we obtain for all exploitation rounds outside a finite initial interval

$$\begin{aligned}\mathbb{E}\Delta_{J_t} &\le Cu^{1/p}\left(\frac{1 + \log(2K) + \ell(t)}{K^{-\rho}t^{\rho}}\right)^{(p-1)/p} \\ &= Cu^{1/p}K^{\rho(p-1)/p}L_t^{(p-1)/p}t^{-\rho(p-1)/p},\end{aligned} \tag{38}$$

where $L_t = 1 + \log(2K) + \ell(t)$.

Put $a = \rho(p-1)/p$. Since $p \le 2$ and $\rho < 1$, $a < 1/2$. Monotonicity of L_t and integral comparison give

$$\sum_{t=1}^{T} L_t^{(p-1)/p}t^{-a} \le 2L_T^{(p-1)/p}T^{1-a}. \tag{39}$$

Combining (30), (38), and (39), and enlarging the universal constant to cover $T \le t_0$, proves

$$\begin{aligned}R_T \le Cu^{1/p}\big\{&K\log(eK) + K^{1-\rho}T^{\rho} \\ &+ K^{\rho(p-1)/p}T^{1-\rho(p-1)/p}L_T^{(p-1)/p}\big\}.\end{aligned}$$

This is Theorem 2 of the main paper.

If $\rho \le 2/3$, then for every $p \in (1, 2]$,

$$1 - \rho(p-1)/p - \rho = 1 - \rho(2p-1)/p \ge 1 - 3\rho/2 \ge 0.$$

Thus the exploitation term has the larger or equal exponent.

D Proof of the Fixed-Instance Bound

Fix an instance in $\mathcal{H}_{p,u}$ and a suboptimal arm i with $\Delta_i > 0$. If i is selected on an exploitation round, then

$$M_i \geq M_{i_*}.$$

This implies either $M_i - \mu_i \geq \Delta_i/2$ or $\mu_* - M_{i_*} \geq \Delta_i/2$. Let $\underline{q}_t$ and $\overline{q}_t$ be the smaller and larger block counts among the two estimates. The arm exploration counts differ by at most one, so $\overline{q}_t \leq \underline{q}_t + 2$. By Lemma 9, for a constant C and all sufficiently large t,

$$\Pr\{J_t = i\} \leq 2\left[\frac{Cu\overline{q}_t^{p-1}}{\Delta_i^p n_t^{p-1}}\right]^{(\underline{q}_t+1)/2}. \tag{40}$$

Here n_t is the smaller sample count.

We verify summability. Since $\log \ell(n) = o(\log n)$,

$$\log \overline{q}_t = o(\log n_t).$$

Hence, for every fixed p, u, Δ_i, the logarithm of the quantity inside the brackets in (40) is eventually at most

$$-\frac{p-1}{2}\log n_t.$$

By (29), it is then at most $-(p-1)\rho \log t/4$ after increasing the instance-dependent threshold. Since $\underline{q}_t \to \infty$, eventually

$$\frac{(p-1)\rho(\underline{q}_t+1)}{8} \geq 3.$$

Therefore $\Pr\{J_t = i\} \leq 2t^{-3}$ eventually. Summing over time and over the finitely many suboptimal arms proves

$$\sum_{t=1}^{\infty} \Pr\{t \text{ is exploit and } J_t \text{ is suboptimal}\} < \infty.$$

The expected total regret due to exploitation is finite. Equation (31) now proves

$$\limsup_{T\to\infty} \frac{R_T}{T^\rho} \leq K^{-\rho} \sum_i \Delta_i.$$

E The Unknown-Scale Lower Tradeoff

For completeness, we prove the exponent lower bound used in the main paper. The argument follows Genalti and Metelli (2025b), which adapts Hadiji and Stoltz (2023). Our statement uses a p-dependent constant. This is all that is needed for the polynomial frontier.

Theorem 11 (Lower tradeoff)**.** *Fix $p \in (1, 2]$. Suppose a policy does not know u and, for a function $\Phi(T) = o(T)$, satisfies*

$$R_T(\eta) \leq v^{1/p}\Phi(T) \tag{41}$$

for every $v > 0$, every $\eta \in \mathcal{H}_{p,v}$, and every T. Then every fixed instance ν with at least one suboptimal arm satisfies

$$\liminf_{T\to\infty} \frac{R_T(\nu)}{(T/\Phi(T))^{p/(p-1)}} \geq c_p \sum_{i:\Delta_i>0} \Delta_i, \tag{42}$$

where one may take

$$c_p = \frac{1}{2\,8^{p/(p-1)}}. \tag{43}$$

Proof. Fix a suboptimal arm a of ν, with law ν_a, mean μ_a, and gap $\Delta_a > 0$. Let v be any finite common upper bound on the arms' pth absolute moments. For $\beta \in (0, 1/2)$, define an alternative law

$$\nu'_{a,\beta} = (1-\beta)\nu_a + \beta\delta_{\mu_a + 2\Delta_a/\beta}. \tag{44}$$

Its mean is $\mu_a + 2\Delta_a = \mu_* + \Delta_a$, so arm a is uniquely better than every original optimal arm by at least Δ_a. Let v'_β be a common pth-moment bound for the alternative bandit. We may take

$$v'_\beta = v + (1-\beta)\mathbb{E}_{\nu_a}|X|^p + \beta|\mu_a + 2\Delta_a/\beta|^p. \tag{45}$$

The redundant v term safely covers all unchanged arms.

Write $N_a(T)$ for the number of pulls of arm a, and let P and P' be the history laws under the original and alternative instances. Mixture domination gives

$$\mathrm{KL}(\nu_a, \nu'_{a,\beta}) \leq \log\frac{1}{1-\beta} \leq 2\beta\log 2.$$

The standard bandit change-of-measure identity and data processing give

$$\mathrm{kl}\left(\frac{\mathbb{E}_P N_a(T)}{T}, \frac{\mathbb{E}_{P'} N_a(T)}{T}\right) \leq 2\beta\log 2\,\mathbb{E}_P N_a(T). \tag{46}$$

For $x, y \in [0,1]$, binary relative entropy obeys

$$\mathrm{kl}(x,y) \geq (1-x)\log\frac{1}{1-y} - \log 2. \tag{47}$$

The moment-free guarantee implies

$$\mathbb{E}_P N_a(T) \leq \frac{v^{1/p}\Phi(T)}{\Delta_a}, \tag{48}$$

$$T - \mathbb{E}_{P'} N_a(T) \leq \frac{(v'_\beta)^{1/p}\Phi(T)}{\Delta_a}. \tag{49}$$

Substituting (48) and (49) into (46) through (47) yields

$$\left(1 - \frac{v^{1/p}\Phi(T)}{T\Delta_a}\right)\log\left(\frac{T\Delta_a}{(v'_\beta)^{1/p}\Phi(T)}\right) - \log 2 \leq 2\beta\log 2\,\mathbb{E}_P N_a(T). \tag{50}$$

Put $r = p/(p-1)$ and choose

$$\alpha = 8^{-r}, \qquad \beta_T = \alpha^{-1}\left(\frac{\Phi(T)}{T}\right)^r. \tag{51}$$

Since $\Phi(T) = o(T)$, $\beta_T \to 0$. Equation (45) gives

$$(v'_{\beta_T})^{1/p}\Phi(T) = 2\Delta_a \alpha^{(p-1)/p} T\{1 + o(1)\}. \tag{52}$$

The right side equals $\Delta_a T\{1+o(1)\}/4$. The first factor on the left of (50) tends to one. The logarithm tends to $\log 4$, so the entire left side tends to $\log 2$. Using (51) on the right gives

$$\begin{aligned}\liminf_{T\to\infty} \frac{\mathbb{E}_P N_a(T)}{(T/\Phi(T))^r} &\geq \frac{\alpha}{2\log 2}\log 2 \\ &= \frac{1}{2\,8^r} = c_p.\end{aligned}$$

Multiplying by Δ_a and summing this inequality over all suboptimal arms proves (42). □

If $\Phi(T) = T^{a+o(1)}$ and a fixed-instance rate is $T^{b+o(1)}$, Theorem 11 implies

$$b \geq \frac{p}{p-1}(1-a),$$

or equivalently

$$b + \frac{p}{p-1}a \geq \frac{p}{p-1}.$$

For the algorithmic exponent $a = 1 - \rho(p-1)/p$, the right side is exactly ρ. This completes the proof of the simultaneous sharp frontier.