Boltzmann Exploration for Heavy-Tailed Bandits

Hyeon-jun Park[1] **Yoon-sik Cho**[1,*] **Kyungjae Lee**[2,*]
[1]Department of Artificial Intelligence, Chung-Ang University
{hjp202, yoonsik}@cau.ac.kr
[2]Department of Statistics, Korea University
kyungjae_lee@korea.ac.kr

Abstract

We study the stochastic multi-armed bandit problem with heavy-tailed rewards, assuming only that each arm's reward distribution has a finite p-th moment for $p \in (1, 2]$. Although prior work has proposed algorithms that are robust to heavy-tailed rewards, these methods do not admit closed-form action-selection probabilities. This hinders efficient offline evaluation and can introduce bias in inverse propensity weighting (IPW) estimators. We propose heavy Boltzmann exploration (H-BE), a Boltzmann-style randomized policy whose action-selection probabilities remain available in closed form under heavy-tailed noise. Theoretically, we show that H-BE achieves the minimax-optimal gap-independent regret bound $O(\nu^{\frac{1}{p}} K^{1-\frac{1}{p}} T^{\frac{1}{p}})$. It also attains the gap-dependent regret bound $O(\sum_{i:\Delta_i>0} \log(T\Delta_i^{\frac{p}{p-1}}/K)/\Delta_i^{\frac{1}{p-1}})$, where ν bounds the p-th moment, K is the number of arms, T is the horizon, and Δ_i is the suboptimality gap of arm i. Empirically, H-BE attains competitive cumulative regret relative to state-of-the-art baselines, while its explicit propensities enable more stable and efficient offline evaluation.

1 INTRODUCTION

Decision-making systems, such as recommendation platforms, online advertising markets, and financial allocation, frequently confront reward distributions with heavy-tails, where rare but extreme outcomes dominate empirical averages (Clauset et al., 2009). In such environments, classical analyses based on sub-Gaussian noise become unreliable: a small number of outliers can both destabilize learning dynamics and obscure genuine arm comparisons. This paper studies the stochastic multi-armed bandit problem under heavy-tailed noise, assuming that rewards have a finite p-th moment for $p \in (1, 2]$.

When rare, extreme outcomes can cause large regret spikes before learning stabilizes, deploying an untested policy online is costly and risky (Zhu et al., 2024). In such cases, it is preferable to use offline evaluation on logged interaction data before online use. The reliability of offline evaluation depends on the logging policy $\pi^{\log}$ providing positive support for all actions the target policy π^{tar} may take; i.e., $\text{supp}(\pi^{\text{tar}}) \subseteq \text{supp}(\pi^{\log})$ with $\pi^{\log} > 0$ for relevant actions, and on accurate access to those probabilities. Deterministic policies such as upper-confidence-bound (UCB) algorithms (Auer et al., 2002) typically collapse propensities to $\{0, 1\}$ after an exploration phase. Some actions receive zero probability and become unobservable in the logs. Randomized policies, by contrast, maintain positive support with strictly positive selection probabilities $\pi \in (0, 1)$, so actions that a target policy may choose remain represented in the logged data. Since inverse probability weighting (IPW) (Horvitz and Thompson, 1952) and doubly-robust (Robins and Rotnitzky, 1995) estimators require nonzero support to control variance, randomized policies have a structural advantage for offline evaluation. Moreover, when the randomized policy provides closed-form probabilities, offline evaluation can use the exact propensities (rather than Monte Carlo (MC) approximations), eliminating an avoidable error source and yielding more reliable assessments before online operation (Maillard, 2016).

Despite this advantage, existing randomized algorithms that offer closed-form action-selection probabilities have been developed under bounded or sub-Gaussian assumptions. Maillard-style sampling (Maillard, 2011; Bian and Jun, 2022) and their KL-based refinement (Qin et al., 2023) specify action-selection probabilities

*Corresponding authors

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

analytically and are therefore well suited to offline evaluation. However, their analysis and parameterization rely on light-tail concentration and do not address the heavy-tailed rewards setting. Conversely, in the heavy-tailed setting there exist randomized exploration strategies, such as perturbation-based (Lee et al., 2020; Lee and Lim, 2022) or posterior-sampling–based variants (Dubey and Pentland, 2019) equipped with robust estimators that stabilize learning under outliers, but they typically do not provide closed-form propensities. In such cases, the action-selection probabilities must be estimated, for example via MC simulation, which introduces additional computational cost and potential bias. We close the above gap by proposing a randomized bandit algorithm for heavy-tailed rewards that retains closed-form action-selection probabilities. We highlight our contributions as follows:

- We introduce heavy Boltzmann exploration (H-BE) which is the first randomized bandit algorithm under heavy-tailed rewards whose action-selection propensities admit a closed form. We show that H-BE can achieve the minimax-optimal gap-independent regret bound $O(\nu^{\frac{1}{p}} T^{\frac{1}{p}} K^{1-\frac{1}{p}})$ and a gap-dependent regret bound $O(\sum_{i:\Delta_i>0} \log(T\Delta_i^{\frac{p}{p-1}}/K)/\Delta_i^{\frac{1}{p-1}})$.
- We derive a non-asymptotic ℓ_1 loss bound for the IPW estimator under heavy tails, explicitly isolating the propensity-estimation error. H-BE avoids this error via closed-form propensities, enabling reliable offline evaluation, whereas methods relying on Monte Carlo estimates incur additional bias and variance.
- On synthetic datasets, the proposed method attains empirical regret comparable to state-of-the-art baselines. In IPW-based offline evaluation, it achieves lower bias than approaches that estimate propensities via Monte Carlo simulation. Additional experimental results are provided in Appendix H.

2 PRELIMINARIES

We consider the stochastic multi-armed bandit problem with heavy-tailed rewards. Let $\mathcal{A} := \{1, 2, \ldots, K\}$ be the set of K actions (arms), and let $\{\mu_1, \ldots, \mu_K\}$ denote the mean rewards corresponding to each action. Without loss of generality, we assume that $\mu_1 > \mu_2 > \cdots > \mu_K$. We define the suboptimality gap as $\Delta_i := \mu_1 - \mu_i$ for each $i \in [K]$. At each round $t \in \{1, \ldots, T\}$, the learner selects an action $i \in [K]$ and observes a stochastic reward $y_t = \mu_i + \zeta_{ti}$, where μ_i is the unknown (but fixed) mean reward of arm i, and ζ_{ti} is a zero-mean noise variable for all $t \in [T]$ and $i \in [K]$. To model the heavy-tailed noise, we adopt the following assumption:

Assumption 1 (Heavy-tailed noise (Bubeck et al., 2013))**.** *Let $p \in (1, 2]$ be arbitrary and let $\nu > 0$. We assume that for each arm $i \in [K]$, the distribution $\mathcal{D}_i$ satisfies $\mathbb{E}_{X\sim\mathcal{D}_i}[|X|^p] \le \nu$.*

This setting extends the classical sub-Gaussian bandit model to heavy-tailed rewards, allowing unbounded noise with possibly infinite variance. This assumption is standard in the literature on heavy-tailed bandits; see, e.g., (Medina and Yang, 2016; Shao et al., 2018; Chowdhury and Gopalan, 2019; Xue et al., 2020). The learner aims to minimize the cumulative regret $R_T = T\mu_1 - \sum_{t=1}^{T} \mathbb{E}[\mu_{i_t}]$ where i_t denotes the selected arm at time step t. Therefore, a smaller regret implies better performance.

3 RELATED WORK

We review (i) sub-Gaussian and bounded randomized methods and (ii) heavy-tailed bandits.

Randomized Exploration in Sub-Gaussian Bandits. In the sub-Gaussian bandit setting, a broad family of algorithms has been developed, including upper confidence bound (UCB) algorithms (Auer et al., 2002), Thompson sampling (TS) (Thompson, 1933), and Boltzmann-Gumbel exploration (BGE) (Cesa-Bianchi et al., 2017). A common limitation of these methods is the lack of closed-form action-selection probabilities, which are crucial for offline evaluation via inverse probability weighting (IPW) (Horvitz and Thompson, 1952) or doubly robust estimators (Robins and Rotnitzky, 1995). For example, TS is widely used due to its Bayesian simplicity and strong empirical performance (Chapelle and Li, 2011), yet its propensities are not available analytically and are typically approximated by Monte Carlo (MC). As a result, while TS can support offline evaluation, it is often inefficient when accurate propensities are required. Notably, Maillard sampling (MS) (Maillard, 2011) offers a randomized strategy with closed-form action-selection probabilities and thus serves as an effective alternative to TS for offline evaluation. MS can be viewed as a special case of Boltzmann exploration (Cesa-Bianchi et al., 2017): at time $t \in [T]$, it selects arm i with probability proportional to $p_{ti} \propto \exp(\eta_t \hat{\Delta}^2_{t-1,i})$, where $\hat{\mu}_{t-1,i}$ is the empirical mean of the rewards from arm i up to time step $t-1$, $\hat{\Delta}_{t-1,i} = \max_{j\in[K]} \hat{\mu}_{t-1,j} - \hat{\mu}_{t-1,i}$ and η_t is a learning rate. Maillard (2011) proved that MS is asymptotically optimal and can achieve a gap-independent regret bound $O(\sqrt{K}T^{3/4})$. A refined analysis by Bian and Jun (2022) showed that a variant of MS satisfies the sub-UCB criterion and attains gap-independent

Table 1: Comparison of Regret Bounds for Heavy-Tailed Algorithms under Assumption 1

Algorithm	**Gap-Dependent Bound** $O(\cdot)$	**Gap-Independent Bound** $\Theta(\cdot)$	**Closed-Form Prob.**
Robust UCB (Bubeck et al., 2013)	$\sum_{i:\Delta_i>0} \log(T)/\Delta_i^{1/(p-1)}$	$(K\log T)^{1-\frac{1}{p}}T^{\frac{1}{p}}$	×
Robust MOSS (Wei and Srivastava, 2020)	$\sum_{i:\Delta_i>0} \log(T\Delta_i^{p/(p-1)}/K)/\Delta_i^{1/(p-1)}$	$K^{1-\frac{1}{p}}T^{\frac{1}{p}}$	×
MR-UCB (Lee and Lim, 2022)	$\sum_{i:\Delta_i>0} \log(T\Delta_i^{p/(p-1)}/K)^{p/(p-1)}/\Delta_i^{1/(p-1)}$	$K^{1-\frac{1}{p}}T^{\frac{1}{p}}$	×
APE^2 (Lee et al., 2020)	$\sum_{i:\Delta_i>0} \log(T\Delta_i^{p/(p-1)})^{p/(p-1)}/\Delta_i^{1/(p-1)}$	$K^{1-\frac{1}{p}}T^{\frac{1}{p}}\log K$	×
MR-APE^2 (Lee and Lim, 2022)	$\sum_{i:\Delta_i>0} \log(T\Delta_i^{p/(p-1)}/K)^{p/(p-1)}/\Delta_i^{1/(p-1)}$	$K^{1-\frac{1}{p}}T^{\frac{1}{p}}$	×
H-BE (ours)	$\sum_{i:\Delta_i>0} \log(T\Delta_i^{p/(p-1)}/K)/\Delta_i^{1/(p-1)}$	$K^{1-\frac{1}{p}}T^{\frac{1}{p}}$	✓

regret bound $O(\sqrt{KT\log K})$. More recently, Kullback-Leibler MS (KL-MS) extended MS to bounded reward settings by replacing the squared reward gap in exponential weight with the binary KL divergence (Qin et al., 2023). KL-MS achieves asymptotic optimality in Bernoulli bandits and an adaptive worst-case regret bound $O(\sqrt{\mu^*(1-\mu^*)KT\log K})$, where μ^* denotes the expected reward of the optimal arm. Meanwhile, Cesa-Bianchi et al. (2017) showed that standard Boltzmann exploration with monotone learning-rate schedules can suffer from poor theoretical guarantees and even linear regret. Their remedy, BGE, employs arm-specific temperatures tied to uncertainty and samples actions via Gumbel perturbations, but still lacks closed-form selection probabilities. Finally, the Minimum Empirical Divergence (MED) algorithm (Honda and Takemura, 2011) also specifies an explicit sampling distribution and attains asymptotic optimality under bounded-support reward distributions. For linear bandits, Balagopalan and Jun (2024) recently extended the MED principle to the sub-Gaussian setting, building on the confidence-bound framework of Abbasi-yadkori et al. (2011). In summary, among algorithms with closed-form propensities, existing analyses are restricted to sub-Gaussian or bounded-support rewards. To the best of our knowledge, no theoretical analysis is available for randomized heavy-tailed bandit algorithms with closed-form action-selection probabilities.

Bandits with Heavy-Tailed Rewards. The stochastic multi-armed bandit problem under heavy-tailed noise has received growing attention in recent years (Bubeck et al., 2013; Medina and Yang, 2016; Shao et al., 2018; Chowdhury and Gopalan, 2019; Lee et al., 2020; Agrawal et al., 2021; Lee and Lim, 2022; Huang et al., 2022; Genalti et al., 2024). In a seminal work, Bubeck et al. (2013) initiated the formal study of this setting by introducing the robust UCB algorithm based on robust mean estimators, such as the trimmed mean, median-of-means, and Catoni's estimator. Under Assumption 1, they showed that Robust UCB attains the gap-dependent regret bound $O(\sum_{i:\Delta_i>0}(\log T/\Delta_i^{1/(p-1)}))$ and the gap-independent regret bound $O(T^{\frac{1}{p}}(K\log T)^{1-\frac{1}{p}})$. While the gap-dependent bound aligns with the lower bound $\Omega(\sum_{i:\Delta_i>0}(\log T/\Delta_i^{1/(p-1)}))$, the gap-independent bound fails to match the lower bound $\Omega(K^{1-\frac{1}{p}}T^{\frac{1}{p}})$ by a factor of $(\log T)^{1-\frac{1}{p}}$. Subsequent work has advanced these ideas by developing adaptive algorithms that remove the need for prior knowledge of moment bounds or tail indices (Wei and Srivastava, 2020; Lee et al., 2020; Lee and Lim, 2022; Genalti et al., 2024). Other contributions improved the regret guarantee of robust UCB (Lee et al., 2020; Wei and Srivastava, 2020; Lee and Lim, 2022) or extended the analysis to more complex frameworks, such as linear (Medina and Yang, 2016; Shao et al., 2018; Xue et al., 2020; Wang et al., 2025), generalized linear (Xue et al., 2023), nonlinear bandits (Chowdhury and Gopalan, 2019), and best-of-both-worlds guarantees (Huang et al., 2022).

Although algorithms with randomized policies have been introduced for heavy-tailed bandits, closed-form action-selection probabilities remain unavailable. Lee and Lim (2022) introduced a generalized Cesa estimator (Cesa-Bianchi et al., 2017), termed the p-robust estimator, which removes the dependence on the moment bound ν, and proposed perturbation-based randomized algorithms (MR-UCB and MR-APE). These methods inject noise from heavy-tailed perturbations (e.g., Weibull, Pareto, Fréchet) into robust mean estimates to enable stochastic action selection, and they achieve the minimax-optimal gap-independent regret bound $O(K^{1-\frac{1}{p}}T^{\frac{1}{p}})$ though their gap-dependent bounds remain above the known lower bound $\Omega(\sum_{i:\Delta_i>0}\log(T)/\Delta_i^{1/(p-1)})$ (see Table 1). Another line of work extends Thompson sampling to symmetric α-stable rewards with $\alpha \in (1,2)$ (Dubey and Pentland, 2019). Addressing the intractability of the α-stable likelihood, the authors adopt a Gaussian scale-mixture auxiliary-variable formulation and derive two variants of TS, α-TS and robust α-TS (robust TS) that enable approximate posterior inference and arm selection under α-stable noise. To improve robustness in heavy-tailed regimes, robust TS employs a truncated-mean estimator and attains a Bayesian regret bound $O((KT)^{1/(1+\rho)})$ for $\rho \in (0, \alpha-1)$. However, despite their randomized exploration, algorithms such as MR-APE, MR-UCB, and robust TS lack closed-form

Algorithm 1 Heavy Boltzmann Exploration (H-BE)

Require: T, K, p, ν, c

Initialize: Pull each arm once
for $t = K+1$ to T **do**
  for each arm $i \in [K]$ **do**

$$b_{ti} \leftarrow \left(\nu n_i(t)/2\log_+(T/Kn_i(t))\right)^{\frac{1}{p}}$$

  Compute truncated mean:

$$\hat{\mu}_{ti} \leftarrow \frac{1}{n_i(t)} \sum_{s=1}^{n_i(t)} y_{s,i} \cdot \mathbf{1}\left[|y_{s,i}| \le b_{ti}\right]$$

  end for
  for each arm $i \in [K]$ **do**
    $\omega_{ti} \leftarrow \exp\left(\frac{t(\hat{\mu}_{ti}+\beta_{ti})}{c^{p/(p-1)}\nu^{1/p}}\right)$
  end for
  Normalize: $\pi_{ti} \leftarrow \omega_{ti}/\sum_{j=1}^{K}\omega_{tj}$
  Sample $I_t \sim \text{Categorical}(\pi_{t1}, \ldots, \pi_{tK})$
  Observe y_t from arm I_t
  Set $n_{I_t}(t) \leftarrow n_{I_t}(t)+1$ and $y_{n_{I_t}(t),I_t} \leftarrow y_t$
end for

action-selection probabilities. Accordingly, when these algorithms are used for offline evaluation, their action-selection probabilities must be approximated, which incurs computational overhead and can introduce bias.

4 PROPOSED METHOD

In this section, we introduce heavy Boltzmann exploration (H-BE) (Algorithm 1), a Boltzmann-style randomized policy for heavy-tailed rewards with closed-form action-selection probabilities. We then develop a non-asymptotic ℓ_1 bound for an inverse-probability-weighted (IPW) estimator to analyze how Monte Carlo (MC) propensity estimation affects IPW estimator. Finally, we establish gap-independent and gap-dependent regret bounds for H-BE and show that they match known lower bounds in the heavy-tailed setting.

4.1 Algorithm

At each round $t \in [T]$, H-BE computes robust mean estimates $\hat{\mu}_{ti}$ for each arm $i \in [K]$. As the robust mean estimator, we employ the truncation estimator, defined as follows:

Definition 1 (Truncation estimator (Bubeck et al., 2013))**.** *Given i.i.d. random variables $X_1, X_2, \ldots, X_n$ satisfying Assumption 1 and a truncation threshold b_{ti}, the truncation estimator is defined as $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \mathbf{1}\{|X_i| \le b_{ti}\}$ where $\mathbf{1}\{\cdot\}$ is an indicator function.*

Intuitively, the truncation estimator reduces the influence of heavy-tailed outliers by discarding samples with $|X_i| > b_{ti}$ where $b_{ti} := \left(\frac{\nu n_i(t)}{2\log_+(T/Kn_i(t))}\right)^{1/p}$, $n_i(t)$ is the number of times arm i is selected up to time t, and $\log_+(u) := \max\{1, \log u\}$. By filtering out such outliers, it yields a more stable estimate of the mean even when higher-order moments may not exist. Under Assumption 1, the truncation estimator admits an exponentially decaying tail bound for its estimation error.

Theorem 2. *Let $X_1, X_2, \ldots X_n$ be independent real-valued random variables satisfying Assumption 1 with common mean μ. Let $\epsilon > 0$ be arbitrary. Denote $\hat{\mu}_n$ as the truncation estimator. Define $n_0 := \frac{C_p \nu^{1/(p-1)}}{\epsilon^{p/(p-1)}} \log_+\left(\frac{T}{Kn}\right)$ with some constant $C(p)$ depending only on p. Then, there exists a constant $c(p) > 0$ depending only on p for which, for every $n \ge n_0$,*

$$\mathbb{P}\{\mu - \hat{\mu}_n \ge \epsilon\} \le \exp\left(-c(p)\nu^{-\frac{1}{p}} n^{1-\frac{1}{p}} \epsilon\right). \quad (1)$$

The full proof is deferred to Appendix B. This theorem serves as the heavy-tailed analogue of Hoeffding's and Bernstein's inequalities, delivering the corresponding exponentially decaying error guarantees under heavy-tailed noise.

After computing the robust means, H-BE assigns each arm $i \in [K]$ a weight

$$\omega_{ti} = \exp\left(\frac{t(\hat{\mu}_{ti}+\beta_{ti})}{c^{p/(p-1)}\nu^{1/p}}\right) \quad (2)$$

where $\beta_{ti} := \left(c(2\nu)^{1/(p-1)}\log_+(T/Kn_i(t))/n_i(t)\right)^{1-\frac{1}{p}}$ is a confidence bonus and t denotes the time step. These weights are normalized to a probability vector $\pi_{ti} := \omega_{ti}/\sum_{j=1}^{K}\omega_{tj}$ and the next action is drawn as $I_t \sim \text{Categorical}(\pi_{t1}, \ldots, \pi_{tK})$. This sampling strategy makes arms with larger robust scores $\hat{\mu}_{ti} + \beta_{ti}$ more likely to be selected while maintaining exploration through the confidence bonus. Specifically, the confidence bonus β_{ti} depends on $n_i(t)$ and shrinks as $n_i(t)$ grows. Thus, β_{ti} promotes exploration in early time steps, and the algorithm gradually shifts from exploration to exploitation, ultimately favoring the arm with the highest robust mean $\hat{\mu}_{ti}$. A detailed description and pseudo-code are provided in Algorithm 1. Unlike other randomized algorithms for heavy-tailed rewards (e.g., robust TS (Dubey and Pentland, 2019) and perturbation-based methods (Lee et al., 2020; Lee and Lim, 2022)), H-BE admits a closed-form expression for its action-selection probabilities (2). In the next section, we present the theoretical guarantees of H-BE and theoretically show that having closed-form action-selection probabilities benefits offline evaluation.

4.2 Main Results

Our main contributions are threefold: (*i*) an ℓ_1 loss analysis of the inverse probability weighting (IPW) estimator under heavy-tailed noise, (*ii*) a gap-independent regret bound, and (*iii*) a gap-dependent regret bound for H-BE.

We begin with the ℓ_1 loss of IPW estimator where $\ell_1(\hat{\mu}, \mu) := \mathbb{E}|\hat{\mu} - \mu|$. This risk is tail-robust and remains well-posed under a first-moment assumption, making it suitable for heavy-tailed rewards. Formally, we observe logged data $(I_t, p^{\log}_{t,I_t}, r_t)_{t=1}^T$ where I_t is the arm selected at time t, $p^{\log}_{t,I_t}$ is the logging propensity of choosing arm I_t at time t, and r_t is the corresponding reward. For a target policy with propensities p^{tar}_{t,I_t}, the IPW estimator is defined as

$$\hat{\mu}_{\text{IPW}} := \frac{1}{T}\sum_{t=1}^{T} \frac{p^{\text{tar}}_{t,I_t}}{p^{\log}_{t,I_t}} r_t. \tag{3}$$

Thus, the IPW estimator requires access to the true propensities $p^{\log}_{t,I_t}$. When an algorithm does not provide closed-form sampling probabilities, these must be approximated (e.g., MC), and replacing $p^{\log}_{t,I_t}$ with estimate $\hat{p}^{\log}_{t,I_t}$ introduces additional error. Our analysis decomposes the ℓ_1 loss into a time-dependent component and a MC component arising from propensity approximation.

Theorem 3. *Fix $\delta \in (0,1)$. Consider the uniform target policy. Let $\hat{\mu}$ denote the IPW estimator with estimated logging propensities $\hat{p}^{log}_{t,I_t}$, defined by $\hat{\mu} := \frac{1}{T}\sum_{t=1}^{T} \frac{r_t}{K\hat{p}^{log}_{t,I_t}}$. Assume that there exists $d \in (0,1)$ such that $p^{log}_{t,I_t} \in [d,1]$ for all $t \in [T]$. Then, with probability at least $1-\delta$, the expected ℓ_1 loss of the IPW estimator satisfies*

$$\mathbb{E}|\hat{\mu} - \mu| \le \frac{c_1 \nu^{1/p}}{KdT^{1-1/p}} + \frac{c_2 \nu^{1/p}}{Kd^2\sqrt{M}}, \tag{4}$$

where c_1, c_2 are some constants depending only on p and M denotes the Monte Carlo samples with $M \ge M_0$ for some $M_0 := M_0(d, \delta)$.

The full proof is deferred to Appendix C. The bound decomposes the expected ℓ_1 loss of the IPW estimator into two parts: (*i*) time-dependent error and (*ii*) MC probability estimation error. The time-dependent error decays at rate $T^{-(1-1/p)}$, which recovers $T^{-1/2}$ rate when rewards have a finite variance ($p = 2$), and slows monotonically as the tail becomes heavier ($p \to 1$). The second term captures the bias due to MC estimation of the action-selection probabilities. Deterministic UCB-style policies and randomized methods without closed-form $p^{\log}_{t,I_t}$ therefore require estimating $\hat{p}^{\log}_{t,I_t}$ and incur $M^{-1/2}$. In contrast, H-BE provides closed-form probabilities (2). Thus, we can directly plug the true propensity $p^{\log}_{t,I_t}$ into IPW estimator and remove this MC error term, leaving only the time-dependent error rate $T^{-(1-1/p)}$. In section 6, we empirically validate these theoretical results. Next, we introduce the gap-independent regret bound of H-BE algorithm.

Theorem 4 (Gap-independent Regret Bound)**.** *The gap-independent regret bound of Algorithm 1 is bounded by*

$$R(T) \le O\Big(\nu^{\frac{1}{p}} T^{\frac{1}{p}} K^{1-\frac{1}{p}} + \sum_{i=1}^{K} \Delta_i\Big). \tag{5}$$

The full proof is deferred to Appendix E. The term $\sum_{i=1}^{K} \Delta_i$ in the regret bound arises from the initialization, where the algorithm pulls each arm once during the first $t \le K$ rounds. In (Bubeck et al., 2013), the authors showed that the gap-independent lower bound under heavy-tailed noise (Assumption 1) satisfies $\Omega(T^{\frac{1}{p}} K^{1-\frac{1}{p}})$. For notational simplicity, assume that $\nu = 1$. Then, Assumption 1 implies that $|\mu_i| \le \nu^{1/p}$, and thus, $\Delta_i \le 2\nu^{1/p} \le 2$ for each $i \in [K]$. The regret bound of Algorithm 1 becomes $O(\nu^{\frac{1}{p}} T^{\frac{1}{p}} K^{1-\frac{1}{p}} + 2K)$ which matches the lower bound and minimax-optimal. Under Assumption 1, several algorithms attain the minimax-optimal regret bound. For instance, MR-APE (Lee and Lim, 2022) achieves $O(e^{\nu} \nu^{\frac{1}{p}} T^{\frac{1}{p}} K^{1-\frac{1}{p}})$, and Robust MOSS (Wei and Srivastava, 2020) achieves $O(\nu^{\frac{1}{p}} T^{\frac{1}{p}} K^{1-\frac{1}{p}})$, the same order as H-BE. However, both are UCB-style methods and therefore lack closed-form action-selection probabilities, which is impractical for offline evaluation that relies on IPW. Now, we present the gap-dependent regret bound of H-BE.

Theorem 5 (Gap-dependent Regret Bound)**.** *The gap-dependent regret of Algorithm 1 is bounded by*

$$R(T) \le O\left(\sum_{i:\Delta_i > 0} \frac{\nu^{\frac{1}{p-1}} \log(T\Delta_i^{\frac{p}{p-1}}/K) + K}{\Delta_i^{1/(p-1)}}\right). \tag{6}$$

The entire proof is deferred to Appendix F. We compare our method with algorithms that achieve the minimax-optimal gap-independent regret bound. Robust MOSS attains $\sum_{i:\Delta_i>0} O\Big(\frac{\log(T\Delta_i^{p/(p-1)}/K)}{\Delta_i^{1/(p-1)}}\Big)$ which matches the order achieved by H-BE. However, Robust MOSS is a UCB-style exploration algorithm. In contrast, MR-UCB and MR-APE yield the gap-dependent bound $O\Big(\sum_{i:\Delta_i>0} \frac{\log(T\Delta_i^{p/(p-1)}/K)^{p/(p-1)}}{\Delta_i^{1/(p-1)}}\Big)$, incurring an additional logarithmic power $p/(p-1) \ge 2$. Under Assumption 1, any algorithm satisfies the gap-dependent lower bound $\Omega(\sum_{i:\Delta_i>0} \frac{\log(T)}{\Delta_i^{1/(p-1)}})$ which does not include the $\Delta_i^{p/(p-1)}$ factor inside the logarithm, indicating a mild logarithmic mismatch between the lower bound and our upper bound (Bubeck et al., 2013).

5 REGRET ANALYSIS

In this section, we provide a proof sketch of gap-independent regret bound (Theorem 4). Before the proof sketch, we introduce several key tools for the regret analysis. We start with the Gumbel-max trick.

Gumbel-Max Trick. The Gumbel-max trick samples from a categorical distribution by selecting the maximizer of Gumbel-perturbed logits (Huijben et al., 2022). Given log-scores $\{\log w_{ti}\}_{i\in D}$ with $w_{ti} > 0$, draw $g_{ti} \sim G(0,\beta)$ i.i.d., and set

$$I_t = \arg\max_{i\in D}(\log w_{ti} + g_{ti}),\ \mathbb{P}(I_t = i) = \frac{w_{ti}^{1/\beta}}{\sum_{j\in D} w_{tj}^{1/\beta}}, \tag{7}$$

where $\beta > 0$ controls the exploration level (smaller β yields greedier choices). In H-BE, this rule is instantiated with $g_{ti} \sim G(0, \alpha/t)$ and $w_{ti} = \exp(\hat{\mu}_{ti} + \beta_{ti})$, where $\alpha = c^{p/(p-1)}\nu^{1/p}$. Observe that (7) expresses the decision score as the sum of a robust mean estimate $\hat{\mu}_{ti}$, a confidence bonus β_{ti}, and Gumbel noise g_{ti}. Leveraging the Gumbel-max trick, H-BE admits an equivalent follow-the-perturbed-leader formulation, which enables a frequentist regret analysis (Lattimore and Szepesvári, 2020). In particular, this allows us to apply the following arm decomposition in our regret analysis.

Lemma 6 (Arm decomposition (Lattimore and Szepesvári, 2020))**.** *Assume that arm 1 is optimal. Let $i > 1$ be an action and $\epsilon \in \mathbb{R}$ be arbitrary. Denote $s = n_i(t)$ as the number of times arm i is played up to time step t. Then the expected number of times Algorithm 1 plays action i is bounded by*

$$\begin{aligned}\mathbb{E}[n_i(T)] &\le \mathbb{E}\left[\sum_{s=1}^{T}\mathbf{1}\{A_t(i), E_i(t)\}\right] + \mathbb{E}\left[\sum_{s=1}^{T}\mathbf{1}\{A_t(i), E_i^c(t)\}\right] \\ &\le 1 + \mathbb{E}\left[\sum_{s=1}^{T}\left(\frac{1}{G_{1s}} - 1\right)\right] + \mathbb{E}\left[\sum_{s=1}^{T}\mathbf{1}\{G_{is} > \frac{1}{T}\}\right]\end{aligned} \tag{8}$$

where $A_t(i) = \{I_t = i\}$, $E_i(t) := \{\theta_i(t) \le \mu_1 - \epsilon\}$, $G_{is} = 1 - F_{is}(\mu_1 - \epsilon)$, F_{is} is a CDF of $G(\hat{\mu}_{ti} + \beta_{ti}, \alpha/t)$, and $\theta_i(t) \sim G(\hat{\mu}_{ti} + \beta_{ti}, \alpha/t)$.

Additionally, we introduce a lemma that serves as a key component in the regret analysis.

Lemma 7. *Let $X_1, X_2, \ldots$ be i.i.d. with mean μ and satisfy Assumption 1. Define $\beta_s := (c(2\nu)^{\frac{1}{p-1}}\log_+(T/Ks)/s)^{1-\frac{1}{p}}$. Denote $\hat{\mu}_s$ as a truncation estimator. Then, for any $x > 0$, we have*

$$\mathbb{P}\{\exists s \in [T] : \hat{\mu}_s + \beta_s - \mu + x \le 0\} \le O\Big(\frac{K\nu^{1/(p-1)}}{Tx^{p/(p-1)}}\Big). \tag{9}$$

The proof of Lemma 7 is given in Appendix D. Intuitively, this lemma controls the probability that the truncation mean estimator underestimates the true mean μ even after adding a confidence bonus β_s. It ensures that $\hat{\mu}_s + \beta_s$ exceeds $\mu - x$ with high probability, uniformly over time $s \in [T]$. We now present a proof sketch of gap-independent regret bound.

5.1 Proof Sketch of Theorem 4

proof sketch of Theorem 4. For any horizon T, the regret $R(T)$ can be decomposed as follows:

$$R(T) = \sum_{i:\Delta_i>0}\Delta_i\mathbb{E}[n_i(T)] \tag{10}$$

$$\le \mathbb{E}[2\Delta T] + T\left(\frac{eK}{T}\right)^{1-\frac{1}{p}} + \sum_{i\in S}\Delta_i\mathbb{E}[n_i(T)] \tag{11}$$

where $\Delta := \mu_1 - \min_{1\le s\le T}(\hat{\mu}_{s1} + \beta_{s1})$ and $S := \{i \in [K] : \Delta_i > \max(2\Delta, (eK/T)^{1-1/p})\}$. Thus, the regret can be decomposed into three components: (i) the regret incurred when the optimal arm is underestimated ($\mathbb{E}[2\Delta T]$), (ii) the regret from suboptimal arms with gaps smaller than $(eK/T)^{1-1/p}$, and (iii) the regret from suboptimal arms with gaps larger than $(eK/T)^{1-1/p}$. For case (i), applying Lemma 7 gives the bound $O(\nu^{\frac{1}{p}}T^{\frac{1}{p}}K^{1-\frac{1}{p}})$ immediately. Additionally, the second term is directly bounded by $O(T^{\frac{1}{p}}K^{1-\frac{1}{p}})$. Then, we need to bound the third term $\sum_{i\in S}\Delta_i\mathbb{E}[n_i(T)]$. To this end, we choose $\ell > 0$ such that $\beta_{si} \le \Delta_i/8$ for all $s \ge \ell$ with $s = n_i(t)$. This choice ensures the preconditions of Theorem 2 and thus allows us to apply it for all $s \ge \ell$. For $i \in S$,

$$\begin{aligned}&\Delta_i\mathbb{E}[n_i(T)] \\ &\le \Delta_i(\ell+1) + \Delta_i\mathbb{E}\left[\sum_{t=\ell}^{T}\mathbf{1}\{I_t = i, E_i^c(t), n_i(t) > \ell\}\right] \\ &+ \Delta_i\mathbb{E}\left[\sum_{t=\ell}^{T}\mathbf{1}\{I_t = i, E_i(t), n_i(t) > \ell\}\right].\end{aligned} \tag{12}$$

where $E_i(t) = \{\theta_i(t) \le \mu_1 - \epsilon\}$. Since $i \in S$, using $\Delta_i > (eK/T)^{1-1/p}$ and the definition of ℓ, we can derive $\sum_{i\in S}\Delta_i(\ell+1) \le O(T^{\frac{1}{p}}K^{1-\frac{1}{p}})$. The second term of (12) corresponds to cases where the log-score of a suboptimal arm i is relatively large due to sampling process, i.e., $E_i^c(t) = \{\theta_i(t) > \mu_1 - \epsilon\}$ with $\theta_i(t) = \hat{\mu}_{ti} + \beta_{ti} + g_{ti}$ and actually the suboptimal arm i is selected at time t. This can be attributed to the following events: (i) the score $\hat{\mu}_{ti} + \beta_{ti}$ of the suboptimal arm i is small but the Gumbel noise g_{ti} is large; and (ii) $\hat{\mu}_{ti}$ is overestimated. The first case can be bounded using the properties of the CDF of the Gumbel distribution. The second case can be bounded by applying the tail bound of the truncated estimator (Theorem 2). Finally, by Lemma 6, the

third term of (12) is bounded by $\mathbb{E}[\sum_{s=1}^{T}(1/G_{1s}(\epsilon)-1)]$ where $G_{1s}(\epsilon) = 1 - F_{1s}(\mu_1 - \epsilon)$ and F_{1s} is a CDF of $G(\hat{\mu}_{1s} + \beta_{1s}, \alpha/t)$. This quantity captures the regret incurred by underestimation of the optimal arm's mean. Importantly, it is distinct from the underestimation of score of the optimal arm, i.e., the first term of (11). This term can be controlled using the concentration properties of the truncated estimator (Theorem 2) and the distributional characteristics of the Gumbel noise (Lemma 12) and bounded by $O(K^{1-\frac{1}{p}}T^{\frac{1}{p}})$. Taking all components into account, we obtain the final regret bound $O(\nu^{\frac{1}{p}}K^{1-\frac{1}{p}}T^{\frac{1}{p}})$. □

Remark. Recall that the cumulative regret has the standard decomposition $R_T = \sum_{i\in[K]}\Delta_i\mathbb{E}[n_i(T)]$. Hence, the rate of the regret bound is determined by how sharply we control $\mathbb{E}[n_i(T)]$. Leveraging the Gumbel–max trick in the regret analysis allows us to employ Lemma 6 to obtain the arm decomposition, which in turn yields a minimax-optimal gap-independent bound. For comparison, consider Maillard Sampling (MS), a Boltzmann-style policy with closed-form action probabilities under sub-Gaussian noise. Its regret analysis also proceeds via an arm decomposition but only after a short burn-in phase u, and this burn-in phase u ultimately dominates the regret bound:

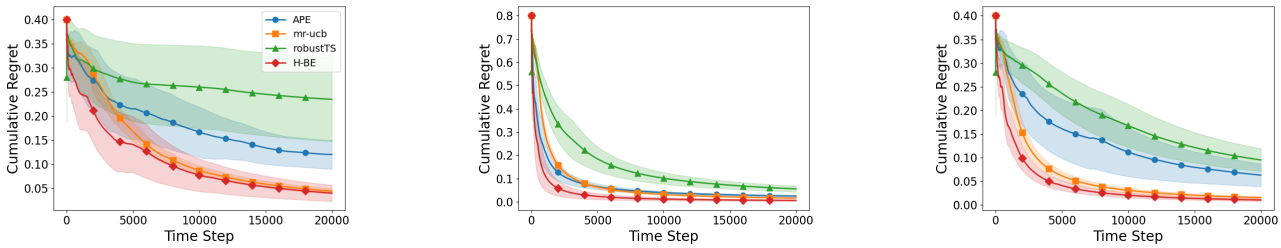
$$\mathbb{E}[n_i(T)] \le u + \mathbb{E}\left[\sum_{t=K+1+u}^{T} \mathbf{1}\{I_t = i,\ n_i(t) \ge u\}\right]. \tag{13}$$

MS analyzes the event $\{I_t = i, n_i(t) \ge u\}$ by splitting into three cases: (i) both the optimal and suboptimal arms are well estimated; (ii) the suboptimal arm is overestimated; and (iii) the optimal arm is underestimated. The first case is straightforward to bound. The second is handled by concentration tail bounds for the empirical mean. The third is controlled by combining the choice of the burn-in parameter u, the MS policy design, and the empirical-mean tail inequality. In particular, when rewards are heavy-tailed, replacing the empirical mean in MS with a robust mean estimator admitting exponentially decaying tail bound allows the same proof scheme to extend to this setting. However, this MS-style regret analysis incurs an additional $\log T$ factor stemming from the definition of the burn-in phase u. Algorithms that rely on a similar proof strategy, such as MS$^+$ (Bian and Jun, 2022) with u, also incur an additional $\log K$ term.

Using the Gumbel–Max trick to interpret the sampling rule of H-BE, we apply Lemma 6 to leverage the arm decomposition. This removes the need for a burn-in parameter u to handle the bad event (case (iii) in MS) in which the optimal arm is underestimated. In our arm decomposition, we directly bound the bad event using Lemma 7 (the first term of (11)). This arm-decomposition technique is inspired by MOSS (Audibert and Bubeck, 2009). In order to control the regret incurred by the bad events, this technique requires a maximal inequality. MOSS derives such an inequality by using exponential submartingales under sub-Gaussian rewards (see Theorem 9.2 and Lemma 9.3 (Lattimore and Szepesvári, 2020)) whose derivation relies on moment-generating function (MGF) of sub-Gaussian random variables. However, under Assumption 1 the noise is heavy-tailed and no direct MGF control is available. We circumvent this by using a truncated-mean estimator and by constructing a local MGF for the truncated variables. We defer the full proof of Lemma 7 to Appendix D.

6 EXPERIMENTS

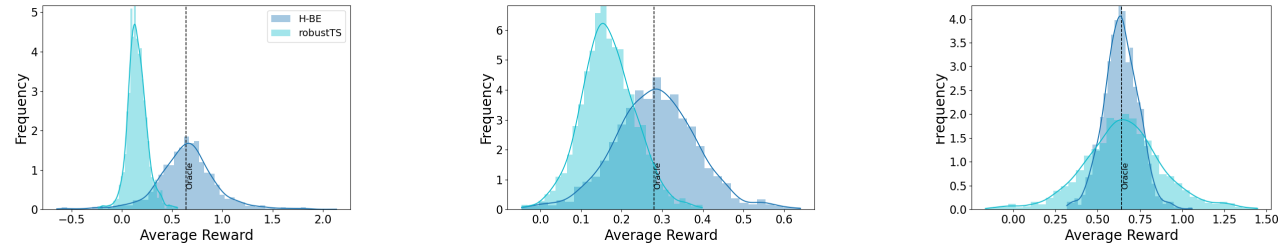
In this section, we present experimental results. We conduct (i) synthetic experiments to verify that the theoretical regret guarantees (Theorems 4 and 5) align with practical performance, and (ii) an offline evaluation testing the implication of Theorem 3, namely that algorithms with closed-form action-selection probabilities enjoy an advantage in offline evaluation. On synthetic data, we benchmark H-BE against randomized algorithms under heavy-tailed rewards, including APE (Lee et al., 2020), MR-UCB (Lee and Lim, 2022), and robust TS (Dubey and Pentland, 2019). For offline evaluation, we include robust TS and estimate its action-selection probabilities by Monte Carlo (MC) sampling. We generate heavy-tailed noise from Pareto and symmetric α-stable families so that only a finite p-th moment is guaranteed for some $p \in (1, 2]$. Rewards are modeled as $r_{ti} = \mu_i + \zeta_{ti}$ where ζ_{ti} are $i.i.d.$ across (t, i) and independent of μ_i. We draw $\zeta_{ti} \sim \text{Pareto}(\alpha, x_m)$ with shape parameter $\alpha > 0$ and scale $x_m > 0$. For a given moment order $p \in (1, 2]$, the p-th moment exists if and only if $\alpha > p$. In this setting, we have $\nu = \mathbb{E}|\zeta|^p = \frac{\alpha x_m^p}{\alpha - p}$. Throughout the experiments we set $x_m = 1$ and choose $\alpha = p + 0.1$ to guarantee $\alpha > p$. For α-stable distribution, we implement the Chambers–Mallows–Stuck sampler (Chambers et al., 1976) with stability parameter $\alpha_{st} = p + 0.1$ so that $\alpha_{st} > p$, skewness $\beta = 0$, scale $\sigma = 1$, and location $\mu = 0$. The p-th moment exists for $p < \alpha_{st}$ and equals $\nu = \mathbb{E}|\zeta|^p = \sigma^p\Gamma(1 - p/\alpha_{st})$. We conduct experiments with $p \in \{1.4, 1.8\}$ and suboptimality gap $\Delta_i \in \{0.4, 0.8\}$. The true mean of the optimal arm is 1, and the means of the suboptimal arms are computed by subtracting the suboptimality gap from an one-hot vector. In the synthetic-dataset experiments, we use $K = 10$, $T = 20000$, and 10 random seeds. For offline evaluation, we use $K = 10$, $T = 1000$, $M = 1000$, and 2000 random seeds. Here, M denotes the MC sam-



(a) Pareto, $p = 1.4$, $\Delta_i = 0.4$ (b) Pareto, $p = 1.8$, $\Delta_i = 0.8$ (c) Pareto, $p = 1.8$, $\Delta_i = 0.4$

(d) α-stable, $p = 1.4$, $\Delta_i = 0.4$ (e) α-stable, $p = 1.8$, $\Delta_i = 0.8$ (f) α-stable, $p = 1.8$, $\Delta_i = 0.4$

Figure 1: Cumulative Regret for Synthetic Datasets. Each curve reports the time-averaged cumulative regret $R(T)/T$ with shaded regions showing the empirical 50% confidence interval across random seeds. The legend is shown only in (a) and shared across all subfigures for clarity.



(a) Pareto, $p = 1.4$, $\Delta_i = 0.4$ (b) Pareto, $p = 1.8$, $\Delta_i = 0.8$ (c) Pareto, $p = 1.8$, $\Delta_i = 0.4$

(d) α-stable, $p = 1.4$, $\Delta_i = 0.4$ (e) α-stable, $p = 1.8$, $\Delta_i = 0.8$ (f) α-stable, $p = 1.8$, $\Delta_i = 0.4$

Figure 2: Offline evaluation results (histograms with KDE) for all synthetic settings. Each panel shows the distribution of average rewards compared with the oracle (dashed vertical line).

ples. Additional experimental results are provided in Appendix H.

Synthetic Results. The synthetic results are shown in Figure 1. H-BE consistently achieves the fastest decay and lowest steady-state regret, demonstrating robustness to heavy-tailed noise. Under Pareto noise (Figures 1a–1c), H-BE dominates throughout, while APE and MR-UCB plateau slightly higher, and Robust TS lags due to its α-stable assumptions being mismatched to Pareto tails. Under symmetric α-stable noise (Figures 1d–1f), Robust TS improves markedly and approaches H-BE, though H-BE still achieves the lowest long-run regret.

Offline Evaluation Results. We now interpret the offline-evaluation results (histograms with KDE). Figure 2 shows that the distribution of H-BE's average reward concentrates tightly around the oracle line, while robust TS appears more dispersed and slightly left-shifted. This pattern matches Theorem 3. H-BE uses closed-form action probabilities, so the IPW weights are exact and the bias is lower. Robust TS relies on MC estimates $\hat{p}_{t,I_t}$ which add estimation noise and can

induce bias.

7 CONCLUSION

We introduced heavy Boltzmann exploration (H-BE), a randomized algorithm for heavy-tailed bandits that admits closed-form action-selection probabilities. Our analysis establishes minimax-optimal gap-independent regret and near-optimal gap-dependent regret bound (up to a logarithmic factor) for H-BE, while the availability of exact propensities eliminates Monte Carlo–induced error in IPW-based offline evaluation. Experiments support the theory, demonstrating competitive regret and more reliable offline evaluation.

ACKNOWLEDGEMENT

This research was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) under Grant RS-2023-00211357. This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University)]

References

Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Shubhada Agrawal, Sandeep K Juneja, and Wouter M Koolen. Regret minimization in heavy-tailed bandits. In *Conference on Learning Theory*, pages 26–62. PMLR, 2021.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, pages 217–226, 2009.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 05 2002. doi: 10.1023/A:1013689704352.

Kapilan Balagopalan and Kwang-Sung Jun. Minimum empirical divergence for sub-gaussian linear bandits. *arXiv preprint arXiv:2411.00229*, 2024.

Jie Bian and Kwang-Sung Jun. Maillard sampling: Boltzmann exploration done optimally. In *International Conference on Artificial Intelligence and Statistics*, pages 54–72. PMLR, 2022.

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

Nicolò Cesa-Bianchi, Claudio Gentile, Gábor Lugosi, and Gergely Neu. Boltzmann exploration done right. *Advances in neural information processing systems*, 30, 2017.

John M Chambers, Colin L Mallows, and BW4159820341 Stuck. A method for simulating stable random variables. *Journal of the american statistical association*, 71(354):340–344, 1976.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24, 2011.

Kwok Pui Choi. *Some sharp inequalities for martingale transforms*. University of Illinois at Urbana-Champaign, 1987.

Sayak Ray Chowdhury and Aditya Gopalan. Bayesian optimization under heavy-tailed payoffs. In *Annual Conference on Neural Information Processing Systems*, pages 13790–13801, 2019.

Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

Abhimanyu Dubey and Alex Pentland. Thompson sampling on symmetric alpha-stable bandits. pages 5715–5721, 08 2019. doi: 10.24963/ijcai.2019/792.

Gianmarco Genalti, Lupo Marsigli, Nicola Gatti, and Alberto Maria Metelli. (ε, u)-adaptive regret minimization in heavy-tailed bandits. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 1882–1915. PMLR, 2024.

Junya Honda and Akimichi Takemura. An asymptotically optimal policy for finite support models in the multiarmed bandit problem. *Machine Learning*, 85: 361–391, 2011.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

Jiatai Huang, Yan Dai, and Longbo Huang. Adaptive best-of-both-worlds algorithm for heavy-tailed multi-armed bandits. In *International Conference on Machine Learning*, pages 9173–9200. PMLR, 2022.

Iris AM Huijben, Wouter Kool, Max B Paulus, and Ruud JG Van Sloun. A review of the gumbel-max trick and its extensions for discrete stochasticity in machine learning. *IEEE transactions on pattern analysis and machine intelligence*, 45(2):1353–1371, 2022.

Samuel Kotz and Saralees Nadarajah. *Extreme value distributions: theory and applications*. world scientific, 2000.

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

Kyungjae Lee and Sungbin Lim. Minimax optimal bandits for heavy tail rewards. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Kyungjae Lee, Hongjun Yang, Sungbin Lim, and Songhwai Oh. Optimal algorithms for stochastic multi-armed bandits with heavy tailed rewards. In *Advances in Neural Information Processing Systems*, 2020.

Odalric-Ambrym Maillard. *APPRENTISSAGE SÉQUENTIEL: Bandits, Statistique et Renforcement.* PhD thesis, Université des Sciences et Technologie de Lille-Lille I, 2011.

Odalric-Ambrym Maillard. Self-normalization techniques for streaming confident regression. 2016.

Andres Muñoz Medina and Scott Yang. No-regret algorithms for heavy-tailed linear bandits. In *Proceedings of the 33nd International Conference on Machine Learning*, volume 48, pages 1642–1650. JMLR.org, June 2016.

Joaquim Ortega-Cerdà and Jordi Saludes. Marcinkiewicz–zygmund inequalities. *Journal of approximation theory*, 145(2):237–252, 2007.

Hao Qin, Kwang-Sung Jun, and Chicheng Zhang. Kullback-leibler maillard sampling for multi-armed bandits with bounded rewards. *Advances in Neural Information Processing Systems*, 36:60514–60526, 2023.

James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

Han Shao, Xiaotian Yu, Irwin King, and Michael R. Lyu. Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. In *Annual Conference on Neural Information Processing Systems*, pages 8430–8439, December 2018.

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.

Jean Ville. *Etude critique de la notion de collectif*, volume 3. Gauthier-Villars Paris, 1939.

Bengt von Bahr and Carl-Gustav Esseen. Inequalities for the rth absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, pages 299–303, 1965.

Jing Wang, Yu-Jie Zhang, Peng Zhao, and Zhi-Hua Zhou. Heavy-tailed linear bandits: Huber regression with one-pass update. *arXiv preprint arXiv:2503.00419*, 2025.

Lai Wei and Vaibhav Srivastava. Minimax policy for heavy-tailed bandits. *IEEE Control Systems Letters*, 5(4):1423–1428, 2020.

Bo Xue, Guanghui Wang, Yimu Wang, and Lijun Zhang. Nearly optimal regret for stochastic linear bandits with heavy-tailed payoffs. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 2936–2942. ijcai.org, 2020.

Bo Xue, Yimu Wang, Yuanyu Wan, Jinfeng Yi, and Lijun Zhang. Efficient algorithms for generalized linear bandits with heavy-tailed rewards. *Advances in Neural Information Processing Systems*, 36:70880–70891, 2023.

Jin Zhu, Runzhe Wan, Zhengling Qi, Shikai Luo, and Chengchun Shi. Robust offline reinforcement learning with heavy-tailed rewards. In *International Conference on Artificial Intelligence and Statistics*, pages 541–549. PMLR, 2024.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes. Section 2 and Section 4]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes. Section 5]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes. Section 2 and Section 4]
   (b) Complete proofs of all theoretical results. [Yes. We provide complete proofs of all theoretical results in the Appendix.]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [No]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [No]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes. Section 6]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [No]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Not Applicable]

   (b) The license information of the assets, if applicable. [Not Applicable]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Boltzmann Exploration for Heavy-Tailed Bandits: Supplementary Materials

A Auxiliary Lemmas

In this section, we collect several auxiliary results used throughout the appendix. We first recall standard tools, including the tail formula for the Gumbel distribution, the Marcinkiewicz–Zygmund inequality, the von Bahr–Esseen inequality, and Ville's inequality. For completeness, we state these results in the form needed for our analysis. We then present two additional technical lemmas, Lemmas 12 and 13, whose proofs are deferred to later subsections.

Lemma 8 (Gumbel Tail Bound (Kotz and Nadarajah, 2000))**.** *Let $X \sim \mathrm{Gumbel}(\mu, \beta)$ be a Gumbel distributed random variable with location parameter $\mu \in \mathbb{R}$ and scale parameter $\beta > 0$. Then for any $x \in \mathbb{R}$, the upper and lower tail probabilities satisfy:*

$$\mathbb{P}(X \geq x) = 1 - \exp\left(-\exp\left(-\frac{x-\mu}{\beta}\right)\right) \quad \text{and} \quad \mathbb{P}(X \leq x) = \exp\left(-\exp\left(-\frac{x-\mu}{\beta}\right)\right). \tag{14}$$

In particular, for $t > 0$,

$$\mathbb{P}(X \geq \mu - t) = 1 - \exp\left(-\exp\left(\frac{t}{\beta}\right)\right), \quad \mathbb{P}(X \leq \mu + t) = \exp\left(-\exp\left(-\frac{t}{\beta}\right)\right). \tag{15}$$

Lemma 9 (Marcinkiewicz–Zygmund inequality (Ortega-Cerdà and Saludes, 2007))**.** *Let $Y_1, \ldots, Y_T$ be independent mean-zero random variables with $\mathbb{E}|Y_t|^q < \infty$ for some $1 \leq q < +\infty$. Then there exists a constant C_q depending only on q such that*

$$\mathbb{E}\left[\Big|\sum_{t=1}^{T} Y_t\Big|^q\right] \leq C_q\, \mathbb{E}\left[\Big(\sum_{t=1}^{T} |Y_t|^2\Big)^{q/2}\right]. \tag{16}$$

Lemma 10 (von Bahr–Esseen inequality (von Bahr and Esseen, 1965))**.** *Let $X_1, \ldots, X_n$ be random variables with $\mathbb{E}|X_i|^r < \infty$ for some $1 \leq r \leq 2$, and define partial sums $S_m := \sum_{\nu=1}^{m} X_\nu$ for $1 \leq m \leq n$. Assume the martingale–difference condition*

$$\mathbb{E}[X_{m+1} \mid S_m] = 0 \quad a.s. \qquad (1 \leq m \leq n-1). \tag{17}$$

Then, we have

$$\mathbb{E}\big|S_n\big|^r \;\leq\; 2\sum_{\nu=1}^{n} \mathbb{E}|X_\nu|^r. \tag{18}$$

Lemma 11 (Ville's inequality Ville (1939); Choi (1987))**.** *Let $X_0, X_1, X_2, \ldots$ be a non-negative supermartingale. Then, for any real number $a > 0$,*

$$\mathbb{P}\{\sup_{n \geq 0} X_n \geq a\} \leq \frac{\mathbb{E}[X_0]}{a}. \tag{19}$$

Lemma 12. *Fix an arbitrary $\epsilon > 0$. Let $L' := \max\left(\left\lceil (4c\nu^{1/p}\log 2/\epsilon)^{\frac{p}{p-1}} + 1\right\rceil, \left\lceil \frac{4\alpha}{\epsilon}\log 2 + 1\right\rceil\right)$ for $\alpha = c^{\frac{p}{p-1}}\nu^{\frac{1}{p}}$. Suppose that for $s \geq L'$, the precondition of Theorem 2 holds. Then, we have*

$$\mathbb{E}\left[\sum_{s=L'}^{T}\left(\frac{1}{G_{1s}(\epsilon)} - 1\right)\right] \leq O\left(\frac{\nu^{1/(p-1)}}{\epsilon^{p/(p-1)}} + \frac{\nu^{1/p}}{\epsilon}\right) \tag{20}$$

where s denotes $n_i(t)$.

Proof. The proof is given in subsection G.1 □

Lemma 13. *Let $g : [0,\infty) \to [0,\infty)$ be measurable, bounded, and unimodal. Suppose that there exists $y^* \geq 0$ such that g is non-decreasing on $[0, y^*]$ and non-increasing on $[y^*, \infty)$. Then, for every $\gamma > 0$, we have*

$$\sum_{j\geq 1} g(j/\gamma) \;\leq\; \gamma \int_0^\infty g(y)\, dy \;+\; \sup_{y\geq 0} g(y) \tag{21}$$

Proof. The proof is given in subsection G.2. □

B Proof of Theorem 2

The tail behavior of the truncation estimator was analyzed by Bubeck et al. (2013). In their original result, the guarantee was given in the form of a high-probability bound: with probability at least $1-\delta$, the estimation error lies inside a confidence interval, so the result is expressed as a confidence interval for fixed confidence level δ (Assumption 1 in (Bubeck et al., 2013)). In contrast, Theorem 2 describes how the error probability decays for a fixed error threshold ϵ. In the sub-Gaussian case, one can easily convert between the two viewpoints (confidence level δ versus error threshold ϵ). However, in the heavy-tailed setting a naive conversion of the high-probability bound of Bubeck et al. (2013) into an ϵ-based statement would force the truncation threshold b_{ti} to depend on ϵ. Since later in the bandit setting the error threshold ϵ is identified with the unknown suboptimality gap Δ, such ϵ-dependence would translate into gap-dependence, contradicting the standard assumption that no prior knowledge of the gaps is available. Our Theorem 2 avoids this problem: by introducing a lower bound on ϵ (equivalently, on samples n), we construct a truncation estimator whose threshold b_{ti} is independent of ϵ, while retaining the exponential error bound. This ϵ–independence is crucial for applying the result in bandit algorithms.

Proof. We divide the argument into three steps. First, we decompose the truncation estimator into a centered empirical term and a truncation-bias term. Second, we bound the truncation bias using the finite p-th moment assumption. Third, we apply an exponential-moment argument to the centered term and choose the exponential parameter appropriately. Fix $s \geq 1$. Define

$$b_s := \left(\frac{\nu s}{2\log_+(T/(Ks))}\right)^{1/p}, \qquad Z_i := X_i \mathbf{1}\{|X_i| \leq b_s\}, \qquad \hat{\mu}_s := \frac{1}{s}\sum_{i=1}^{s} Z_i. \tag{22}$$

Thus, $\hat{\mu}_s$ is the truncation estimator with truncation level b_s. Next define

$$U_i := Z_i - \mathbb{E}Z_i, \tag{23}$$

$$B'_s := \mu - \frac{1}{s}\sum_{i=1}^{s} \mathbb{E}Z_i, \tag{24}$$

$$B_s := \frac{1}{s}\sum_{i=1}^{s} \mathbb{E}[|X_i|\mathbf{1}\{|X_i| > b_s\}]\,. \tag{25}$$

Here, U_i is the centered truncated variable, B'_s is the actual bias induced by truncation, and B_s is a simpler upper bound on that bias. By adding and subtracting $\frac{1}{s}\sum_{i=1}^{s}\mathbb{E}Z_i$, we obtain the decomposition

$$\hat{\mu}_s - \mu = \frac{1}{s}\sum_{i=1}^{s}(Z_i - \mathbb{E}Z_i) - \left(\mu - \frac{1}{s}\sum_{i=1}^{s}\mathbb{E}Z_i\right) = \frac{1}{s}\sum_{i=1}^{s} U_i - B'_s. \tag{26}$$

We now bound the truncation bias. Since

$$X_i\mathbf{1}\{|X_i| > b_s\} \leq |X_i|\mathbf{1}\{|X_i| > b_s\}, \tag{27}$$

taking expectations yields $B'_s \leq B_s$. Moreover, using the elementary inequality

$$\mathbb{E}[|X|\mathbf{1}\{|X| > b\}] \leq \frac{\mathbb{E}|X|^p}{b^{p-1}}, \tag{28}$$

we obtain

$$B_s = \frac{1}{s}\sum_{i=1}^{s}\mathbb{E}[|X_i|\mathbf{1}\{|X_i| > b_s\}] \tag{29}$$

$$\leq \frac{1}{s}\sum_{i=1}^{s}\frac{\mathbb{E}|X_i|^p}{b_s^{p-1}} \tag{30}$$

$$\leq \frac{\nu}{b_s^{p-1}} \tag{31}$$

$$= 2^{\,1-\frac{1}{p}}\nu^{1/p}\left(\frac{\log_+(T/(Ks))}{s}\right)^{1-\frac{1}{p}}. \tag{32}$$

Thus the truncation bias decays as $s^{-(1-1/p)}$ up to the logarithmic factor. We next turn to the centered term $\sum_{i=1}^{s} U_i$. Since $|Z_i| \leq b_s$ and

$$|\mathbb{E}Z_i| \leq \mathbb{E}|Z_i| \leq b_s, \tag{33}$$

we have

$$|U_i| = |Z_i - \mathbb{E}Z_i| \leq 2b_s. \tag{34}$$

Therefore, if we define

$$L_s := \frac{1}{2b_s} = 2^{\frac{1}{p}-1}\nu^{-1/p}s^{-1/p}\left(\log_+\left(\frac{T}{Ks}\right)\right)^{1/p}, \tag{35}$$

then $|\lambda U_i| \leq 1$ for every $\lambda \in [0, L_s]$. For $|y| \leq 1$ and $p \in (1, 2]$, we use

$$e^{-y} \leq 1 - y + \frac{e}{2}|y|^2 \leq 1 - y + \frac{e}{2}|y|^p. \tag{36}$$

Applying this with $y = \lambda U_i$, using $\mathbb{E}U_i = 0$, and then using $1 + z \leq e^z$, we get for every $\lambda \in [0, L_s]$,

$$\mathbb{E}e^{-\lambda U_i} \leq \exp\Big(\frac{e}{2}\lambda^p\mathbb{E}|U_i|^p\Big). \tag{37}$$

To bound $\mathbb{E}|U_i|^p$, use

$$|a - b|^p \leq 2^{p-1}(|a|^p + |b|^p), \tag{38}$$

which is valid for $p \in (1, 2]$. Since $\mathbb{E}|Z_i|^p \leq \mathbb{E}|X_i|^p \leq \nu$, we obtain

$$\mathbb{E}|U_i|^p = \mathbb{E}|Z_i - \mathbb{E}Z_i|^p \tag{39}$$

$$\leq 2^{p-1}\Big(\mathbb{E}|Z_i|^p + |\mathbb{E}Z_i|^p\Big) \tag{40}$$

$$\leq 2^{p-1}\Big(\mathbb{E}|Z_i|^p + \mathbb{E}|Z_i|^p\Big) \tag{41}$$

$$\leq 2^p\nu. \tag{42}$$

Hence, by independence of the U_i's,

$$\mathbb{E}\exp\left(-\lambda\sum_{i=1}^{s}U_i\right) = \prod_{i=1}^{s}\mathbb{E}e^{-\lambda U_i} \leq \exp(c_0\nu s\lambda^p), \tag{43}$$

where $c_0 := e\,2^{p-1}$. Now fix $\epsilon > 0$ and define $t := (\epsilon - B_s)_+$, where $(x)_+ := \max\{x, 0\}$. Since $B_s' \leq B_s$, the event $\{\hat{\mu}_s - \mu \leq -\epsilon\}$ implies

$$\frac{1}{s}\sum_{i=1}^{s}U_i \leq -(\epsilon - B_s') \leq -(\epsilon - B_s)_+ = -t. \tag{44}$$

Therefore,

$$\{\hat{\mu}_s - \mu \leq -\epsilon\} \subseteq \left\{ \frac{1}{s} \sum_{i=1}^{s} U_i \leq -t \right\}. \tag{45}$$

Applying Markov's inequality to

$$\exp\left(-\lambda \sum_{i=1}^{s} U_i \right), \tag{46}$$

we obtain, for any $\lambda \in [0, L_s]$,

$$\mathbb{P}\left\{ \frac{1}{s} \sum_{i=1}^{s} U_i \leq -t \right\} = \mathbb{P}\left\{ e^{-\lambda \sum_{i=1}^{s} U_i} \geq e^{\lambda s t} \right\} \tag{47}$$

$$\leq e^{-\lambda s t} \mathbb{E} \exp\left(-\lambda \sum_{i=1}^{s} U_i \right) \tag{48}$$

$$\leq \exp(-\lambda s t + c_0 \nu s \lambda^p). \tag{49}$$

Let

$$\phi(\lambda) := -\lambda s t + c_0 \nu s \lambda^p. \tag{50}$$

We now impose a lower bound on s so that the bias term is not too large. Assume $\epsilon \geq 2B_s$. Then

$$t = (\epsilon - B_s)_+ \geq \epsilon/2. \tag{51}$$

Since

$$B_s = 2^{\,1-\frac{1}{p}} \nu^{1/p} \left(\frac{\log_+(T/(Ks))}{s} \right)^{1-\frac{1}{p}}, \tag{52}$$

a sufficient condition for $\epsilon \geq 2B_s$ is

$$\frac{\log_+(T/(Ks))}{s} \leq \left(\frac{\epsilon}{2^{\,2-\frac{1}{p}} \nu^{1/p}} \right)^{\frac{p}{p-1}}, \tag{53}$$

that is,

$$s \geq \left(\frac{2^{\,2-\frac{1}{p}} \nu^{1/p}}{\epsilon} \right)^{\frac{p}{p-1}} \log_+\left(\frac{T}{Ks} \right). \tag{54}$$

Next consider the unconstrained minimizer of ϕ:

$$\lambda_* = \left(\frac{t}{p c_0 \nu} \right)^{1/(p-1)}. \tag{55}$$

If $\lambda_* > L_s$, then ϕ is decreasing throughout the admissible interval $[0, L_s]$, so the best choice is the boundary point $\lambda = L_s$. Using $t \geq \epsilon/2$, a sufficient condition for $\lambda_* > L_s$ is

$$\left(\frac{\epsilon}{2pc_0\nu} \right)^{1/(p-1)} > 2^{\frac{1}{p}-1} \nu^{-1/p} s^{-1/p} \left(\log_+\left(\frac{T}{Ks} \right) \right)^{1/p}, \tag{56}$$

which is equivalent to

$$s \geq 2^{\,1-p} (2pc_0)^{\frac{p}{p-1}} \nu^{1/(p-1)} \epsilon^{-p/(p-1)} \log_+\left(\frac{T}{Ks} \right). \tag{57}$$

Combining (54) and (57), we can choose a constant $C(p) > 0$, depending only on p, such that whenever

$$s \geq C(p)\, \nu^{1/(p-1)} \epsilon^{-p/(p-1)} \log_+\left(\frac{T}{Ks}\right), \tag{58}$$

both conditions hold. In particular, we then have $t \geq \epsilon/2$ and $\lambda_* > L_s$. Since

$$\phi'(\lambda) = -st + c_0 \nu s p \lambda^{p-1}, \qquad \phi''(\lambda) = c_0 \nu s p (p-1) \lambda^{p-2} > 0, \tag{59}$$

the function ϕ is convex and ϕ' is increasing. Because $\lambda_* > L_s$, we have $\phi'(L_s) < 0$, and therefore ϕ is strictly decreasing on $[0, L_s]$. Hence the minimum over the admissible interval is attained at $\lambda = L_s$:

$$\min_{\lambda \in [0, L_s]} \phi(\lambda) = \phi(L_s). \tag{60}$$

Furthermore, from $t > p c_0 \nu L_s^{p-1}$ we obtain

$$L_s t - c_0 \nu L_s^p \geq \left(1 - \frac{1}{p}\right) L_s t. \tag{61}$$

Substituting $\lambda = L_s$ and using $t \geq \epsilon/2$, we conclude

$$\mathbb{P}\left\{\frac{1}{s}\sum_{i=1}^{s} U_i \leq -t\right\} \leq \exp\bigl(-s\bigl(L_s t - c_0 \nu L_s^p\bigr)\bigr) \tag{62}$$

$$\leq \exp\left(-\frac{p-1}{p} 2^{\frac{1}{p}-2} \nu^{-1/p} s^{1-\frac{1}{p}} \left(\log_+\left(\frac{T}{Ks}\right)\right)^{1/p} \epsilon\right). \tag{63}$$

Finally, since $\log_+(T/(Ks)) \geq 1$, this implies

$$\mathbb{P}\{\hat{\mu}_s - \mu \leq -\epsilon\} \leq \exp\Bigl(-c(p)\, \nu^{-1/p} s^{1-\frac{1}{p}} \epsilon\Bigr) \tag{64}$$

for some constant $c(p) > 0$ depending only on p, provided that

$$s \geq C(p)\, \nu^{1/(p-1)} \epsilon^{-p/(p-1)} \log_+\left(\frac{T}{Ks}\right). \tag{65}$$

This proves the claim. □

C Proof of Theorem 3

Proof. We split the proof into two parts. First, we control the bias introduced by replacing the true logging propensity p_{t,I_t} with its Monte Carlo estimate $\hat{p}_{t,I_t}$. Second, we control the fluctuation of the resulting IPW estimator around its mean. This yields the desired decomposition into a time-dependent term and a Monte Carlo term.

Recall that

$$\hat{\mu} := \frac{1}{T}\sum_{t=1}^{T} \frac{r_t}{K \hat{p}_{t,I_t}}, \qquad \mu := \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\frac{r_t}{K p_{t,I_t}}\right]. \tag{66}$$

By the triangle inequality,

$$\mathbb{E}|\hat{\mu} - \mu| \leq \mathbb{E}|\hat{\mu} - \mathbb{E}\hat{\mu}| + |\mathbb{E}\hat{\mu} - \mu|. \tag{67}$$

We bound the two terms on the right-hand side separately.

Step 1: Bias due to propensity estimation. Let

$$\hat{X}_t := \frac{r_t}{K\hat{p}_{t,I_t}}, \qquad X_t := \frac{r_t}{Kp_{t,I_t}}. \tag{68}$$

Then

$$\mathbb{E}\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\hat{X}_t, \qquad \mu = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}X_t, \tag{69}$$

and hence

$$|\mathbb{E}\hat{\mu} - \mu| = \frac{1}{T}\left|\sum_{t=1}^{T}\mathbb{E}(\hat{X}_t - X_t)\right|. \tag{70}$$

Using the definitions of $\hat{X}_t$ and X_t,

$$|\mathbb{E}\hat{\mu} - \mu| = \frac{1}{TK}\left|\sum_{t=1}^{T}\mathbb{E}\left[\left(\frac{1}{\hat{p}_{t,I_t}} - \frac{1}{p_{t,I_t}}\right)r_t\right]\right| \le \frac{1}{T}\sum_{t=1}^{T}|\mathbb{E}(\hat{X}_t - X_t)|. \tag{71}$$

Let $q := \frac{p}{p-1}$ be the conjugate exponent of p. By Hölder's inequality,

$$|\mathbb{E}(\hat{X}_t - X_t)| \le \frac{1}{K}\left(\mathbb{E}\left|\frac{1}{\hat{p}_{t,I_t}} - \frac{1}{p_{t,I_t}}\right|^q\right)^{1/q}(\mathbb{E}|r_t|^p)^{1/p}. \tag{72}$$

Since the rewards satisfy the finite-pth moment assumption,

$$(\mathbb{E}|r_t|^p)^{1/p} \le \nu^{1/p}. \tag{73}$$

Now assume that there exists a constant $d \in (0,1)$ such that

$$p_{t,I_t} \in [d,1] \qquad \text{for all } t \in [T]. \tag{74}$$

For each fixed t, the Monte Carlo estimator $\hat{p}_{t,I_t}$ is given by

$$\hat{p}_{t,I_t} = \frac{1}{M}\sum_{i=1}^{M}\mathbf{1}\{I_t = \arg\max_{j\in[K]}\theta_{ij}\}, \tag{75}$$

where θ_{ij} denotes the sampled score of arm j in the ith Monte Carlo draw. Thus, if we define

$$Z_{ti} := \mathbf{1}\{I_t = \arg\max_{j\in[K]}\theta_{ij}\}, \tag{76}$$

then $Z_{ti} \sim \text{Bernoulli}(p_{t,I_t})$ and

$$\hat{p}_{t,I_t} = \frac{1}{M}\sum_{i=1}^{M}Z_{ti}. \tag{77}$$

By Bernstein's inequality, for each fixed t, if

$$M \ge \frac{\log(T/\delta)\big(2p_{t,I_t}(1-p_{t,I_t}) + \frac{2}{3}(p_{t,I_t} - d)\big)}{(p_{t,I_t} - d)^2}, \tag{78}$$

then

$$\mathbb{P}(\hat{p}_{t,I_t} > d) \ge 1 - \delta/T. \tag{79}$$

Applying a union bound over $t = 1, \dots, T$, we obtain that the event

$$E := \left\{ \min_{t \in [T]} \hat{p}_{t,I_t} \geq d \right\} \tag{80}$$

holds with probability at least $1 - \delta$. On the event E, both p_{t,I_t} and $\hat{p}_{t,I_t}$ lie in $[d, 1]$, and the map $x \mapsto 1/x$ is d^{-2}-Lipschitz on this interval. Therefore,

$$\left| \frac{1}{\hat{p}_{t,I_t}} - \frac{1}{p_{t,I_t}} \right| \leq \frac{1}{d^2} |\hat{p}_{t,I_t} - p_{t,I_t}|, \tag{81}$$

and hence

$$\mathbb{E} \left| \frac{1}{\hat{p}_{t,I_t}} - \frac{1}{p_{t,I_t}} \right|^q \leq \frac{1}{d^{2q}} \mathbb{E} |\hat{p}_{t,I_t} - p_{t,I_t}|^q. \tag{82}$$

To bound the right-hand side, define

$$Y_{ti} := Z_{ti} - p_{t,I_t}. \tag{83}$$

Then the Y_{ti} are i.i.d., mean zero, and bounded by 1 in absolute value, and

$$\hat{p}_{t,I_t} - p_{t,I_t} = \frac{1}{M} \sum_{i=1}^{M} Y_{ti}. \tag{84}$$

Applying Lemma 9 (Marcinkiewicz–Zygmund inequality),

$$\mathbb{E} |\hat{p}_{t,I_t} - p_{t,I_t}|^q = \mathbb{E} \left| \frac{1}{M} \sum_{i=1}^{M} Y_{ti} \right|^q \tag{85}$$

$$\leq \frac{C_q}{M^q} \mathbb{E} \left[\left(\sum_{i=1}^{M} |Y_{ti}|^2 \right)^{q/2} \right] \tag{86}$$

$$\leq \frac{C_q}{M^q} \cdot M^{q/2} = \frac{C_q}{M^{q/2}}, \tag{87}$$

where the second inequality uses $|Y_{ti}| \leq 1$.

Substituting this into (82) and then into (72), we obtain on E that

$$|\mathbb{E}(\hat{X}_t - X_t)| \leq \frac{C_q^{1/q}}{K d^2 M^{1/2}} \nu^{1/p}. \tag{88}$$

Averaging over t yields

$$|\mathbb{E}\hat{\mu} - \mu| \leq \frac{C_q^{1/q}}{K d^2 M^{1/2}} \nu^{1/p}. \tag{89}$$

Step 2: Fluctuation around the mean. We now bound $\mathbb{E}|\hat{\mu} - \mathbb{E}\hat{\mu}|$. Let $\{\mathcal{F}_t\}_{t=0}^T$ be the filtration containing all information up to time t, including the randomness used to construct $\hat{p}_{t,I_t}$. Define

$$W_t := \hat{X}_t - \mathbb{E}[\hat{X}_t \mid \mathcal{F}_{t-1}], \tag{90}$$

$$S_m := \sum_{t=1}^{m} W_t, \tag{91}$$

$$A_T := \sum_{t=1}^{T} \Big(\mathbb{E}[\hat{X}_t \mid \mathcal{F}_{t-1}] - \mathbb{E}\hat{X}_t \Big). \tag{92}$$

Then

$$\hat{\mu} - \mathbb{E}\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T}(\hat{X}_t - \mathbb{E}\hat{X}_t) = \frac{1}{T}(S_T + A_T). \tag{93}$$

Therefore, by Lyapunov's inequality,

$$\mathbb{E}|\hat{\mu} - \mathbb{E}\hat{\mu}| \le \frac{1}{T}(\mathbb{E}|S_T|^p)^{1/p} + \frac{1}{T}(\mathbb{E}|A_T|^p)^{1/p}. \tag{94}$$

We first handle the martingale-difference part S_T. Since $\mathbb{E}[W_t \mid \mathcal{F}_{t-1}] = 0$ a.s., and S_m is $\mathcal{F}_m$-measurable,

$$\mathbb{E}[W_{m+1} \mid S_m] = \mathbb{E}[\mathbb{E}[W_{m+1} \mid \mathcal{F}_m] \mid S_m] = 0 \qquad \text{a.s.} \tag{95}$$

Hence the hypothesis of Lemma 10 applies with $X_t = W_t$, $n = T$, and $r = p$. We obtain

$$(\mathbb{E}|S_T|^p)^{1/p} \le \left(2\sum_{t=1}^{T}\mathbb{E}|W_t|^p\right)^{1/p}. \tag{96}$$

Next we bound $\mathbb{E}|W_t|^p$. Using $|a-b|^p \le 2^{p-1}(|a|^p + |b|^p)$ and Jensen's inequality,

$$\mathbb{E}|W_t|^p = \mathbb{E}\left|\hat{X}_t - \mathbb{E}[\hat{X}_t \mid \mathcal{F}_{t-1}]\right|^p \tag{97}$$

$$\le 2^{p-1}\left(\mathbb{E}|\hat{X}_t|^p + \mathbb{E}\left|\mathbb{E}[\hat{X}_t \mid \mathcal{F}_{t-1}]\right|^p\right) \tag{98}$$

$$\le 2^p \mathbb{E}|\hat{X}_t|^p. \tag{99}$$

On the event E, we have $\hat{p}_{t,I_t} \ge d$, so

$$\mathbb{E}|\hat{X}_t|^p = \mathbb{E}\left|\frac{r_t}{K\hat{p}_{t,I_t}}\right|^p \le \left(\frac{1}{Kd}\right)^p \mathbb{E}|r_t|^p \le \left(\frac{1}{Kd}\right)^p \nu. \tag{100}$$

Substituting into (96), we obtain

$$\frac{1}{T}(\mathbb{E}|S_T|^p)^{1/p} \le \frac{2^{(p+1)/p}}{Kd}\nu^{1/p}T^{-(1-1/p)}. \tag{101}$$

We now turn to the predictable term A_T. By Minkowski's inequality and then Jensen's inequality,

$$(\mathbb{E}|A_T|^p)^{1/p} \le \sum_{t=1}^{T}\left(\mathbb{E}\left|\mathbb{E}[\hat{X}_t \mid \mathcal{F}_{t-1}] - \mathbb{E}\hat{X}_t\right|^p\right)^{1/p} \tag{102}$$

$$\le 2\sum_{t=1}^{T}\left(\mathbb{E}|\hat{X}_t|^p\right)^{1/p} \tag{103}$$

$$\le 2T\left(\frac{\nu}{(Kd)^p}\right)^{1/p}. \tag{104}$$

Therefore,

$$\frac{1}{T}(\mathbb{E}|A_T|^p)^{1/p} \le \frac{2}{Kd}\nu^{1/p}. \tag{105}$$

This bound is valid but too crude to capture the averaging effect in T. A sharper estimate is obtained by applying the standard inequality

$$\left|\sum_{t=1}^{T}a_t\right|^p \le T^{p-1}\sum_{t=1}^{T}|a_t|^p, \tag{106}$$

which yields

$$(\mathbb{E}|A_T|^p)^{1/p} \le T^{1-\frac{1}{p}} \left(\sum_{t=1}^{T} \mathbb{E} \left| \mathbb{E}[\hat{X}_t \mid \mathcal{F}_{t-1}] - \mathbb{E}\hat{X}_t \right|^p \right)^{1/p} \tag{107}$$

$$\le 2T^{1-\frac{1}{p}} \left(\sum_{t=1}^{T} \mathbb{E}|\hat{X}_t|^p \right)^{1/p}. \tag{108}$$

Using (100), we conclude

$$\frac{1}{T}(\mathbb{E}|A_T|^p)^{1/p} \le \frac{2}{Kd} \nu^{1/p} T^{-(1-1/p)}. \tag{109}$$

Combining (94), (101), and (109), we get

$$\mathbb{E}|\hat{\mu} - \mathbb{E}\hat{\mu}| \le \frac{2^{(p+1)/p}}{Kd} \nu^{1/p} T^{-(1-1/p)} + \frac{2}{Kd} \nu^{1/p} T^{-(1-1/p)}. \tag{110}$$

Hence,

$$\mathbb{E}|\hat{\mu} - \mathbb{E}\hat{\mu}| = \mathcal{O}\left(\frac{\nu^{1/p}}{Kd\, T^{1-1/p}} \right). \tag{111}$$

Step 3: Final combination. On the event E, combining (89) and (111) with (67), we obtain

$$\mathbb{E}|\hat{\mu} - \mu| \le \frac{c_1 \nu^{1/p}}{Kd\, T^{1-1/p}} + \frac{c_2 \nu^{1/p}}{Kd^2 \sqrt{M}}, \tag{112}$$

where $c_1, c_2 > 0$ depend only on p. Since $\mathbb{P}(E) \ge 1 - \delta$, the above bound holds with probability at least $1 - \delta$. This proves the claim. □

D Proof of Lemma 7

Proof. We divide the proof into four steps. First, we rewrite the event $\{\hat{\mu}_s + \beta_s \le \mu - x\}$ in terms of the centered truncated sum. Second, we construct an exponential supermartingale for this centered sum and derive a maximal inequality over an arbitrary interval of sample sizes. Third, we apply a peeling argument to partition the index set $\{1, \dots, T\}$ into short blocks. Finally, we sum the resulting blockwise bounds in two regimes, depending on whether the unconstrained maximizer of the exponential bound lies inside the admissible interval or not.

Step 1: Reduction to a centered lower-tail event. Fix $s \ge 1$. Set

$$b_s := \left(\frac{\nu\, s}{2 \log_+(T/Ks)} \right)^{\frac{1}{p}}. \tag{113}$$

Let

$$Z_i := X_i \mathbf{1}\{|X_i| \le b_s\}, \qquad U_i := Z_i - \mathbb{E}Z_i, \qquad B'_s := \mu - \frac{1}{s} \sum_{i=1}^{s} \mathbb{E}Z_i. \tag{114}$$

Then

$$\hat{\mu}_s - \mu = \frac{1}{s} \sum_{i=1}^{s} Z_i - \mu \tag{115}$$

$$= \frac{1}{s} \sum_{i=1}^{s} (Z_i - \mathbb{E}Z_i) + \frac{1}{s} \sum_{i=1}^{s} \mathbb{E}Z_i - \mu \tag{116}$$

$$= \frac{1}{s} \sum_{i=1}^{s} U_i - \left(\mu - \frac{1}{s} \sum_{i=1}^{s} \mathbb{E}Z_i \right) \tag{117}$$

$$= \frac{1}{s} \sum_{i=1}^{s} U_i - B'_s. \tag{118}$$

Thus, we have the basic decomposition

$$\hat{\mu}_s - \mu = \frac{1}{s}\sum_{i=1}^{s} U_i - B'_s. \tag{119}$$

As in the proof of Theorem 2, we upper-bound the truncation bias term. Since

$$X_i \mathbf{1}\{|X_i| > b_s\} \le |X_i| \mathbf{1}\{|X_i| > b_s\}, \tag{120}$$

we have

$$B'_s \le B_s := \frac{1}{s}\sum_{i=1}^{s} \mathbb{E}[|X_i| \mathbf{1}\{|X_i| > b_s\}]. \tag{121}$$

Using the pointwise bound $|X_i| \mathbf{1}\{|X_i| > b_s\} \le |X_i|^p / b_s^{p-1}$,

$$B_s = \frac{1}{s}\sum_{i=1}^{s} \mathbb{E}\big[|X_i| \, \mathbf{1}\{|X_i| > b_s\}\big] \le \frac{1}{s}\sum_{i=1}^{s} \frac{\mathbb{E}|X_i|^p}{b_s^{p-1}} = \frac{\nu}{b_s^{p-1}} = 2^{1-\frac{1}{p}} \nu^{\frac{1}{p}} \Big(\frac{\log_+(T/Ks)}{s}\Big)^{1-\frac{1}{p}}. \tag{122}$$

Next we compare the confidence bonus β_s with the truncation bias B_s. By definition,

$$\frac{\beta_s}{2B_s} \ge \frac{c^{1-\frac{1}{p}} (2\nu)^{\frac{1}{p}} (\log_+(T/Ks)/s)^{1-\frac{1}{p}}}{2^{2-\frac{1}{p}} \nu^{\frac{1}{p}} (\log_+(T/Ks)/s)^{1-\frac{1}{p}}}. \tag{123}$$

Hence, by choosing c large enough, we may ensure that

$$\beta_s \ge 2B_s \tag{124}$$

for all relevant s. Therefore, on the event $\{\hat{\mu}_s + \beta_s \le \mu - x\}$,

$$\frac{1}{s}\sum_{i=1}^{s} U_i = \hat{\mu}_s - \mu + B'_s \le -(x + \beta_s) + B_s \le -\frac{x + \beta_s}{2}. \tag{125}$$

Thus,

$$\{\hat{\mu}_s + \beta_s \le \mu - x\} \subseteq \left\{ \frac{1}{s}\sum_{i=1}^{s} U_i \le -\frac{x + \beta_s}{2} \right\}. \tag{126}$$

Step 2: Exponential supermartingale bound on an arbitrary interval. We now derive a uniform lower-tail bound for the centered sum over an interval I of sample sizes. Since $|Z_i| \le b_s$ and $|\mathbb{E}Z_i| \le \mathbb{E}|Z_i| \le b_s$, we have $|U_i| \le 2b_s$. Let

$$b^* = \max_{s \in I} b_s \qquad \text{and} \qquad L := \frac{1}{2b^*}. \tag{127}$$

Then, for any $\lambda \in [0, L]$, we have $|\lambda U_i| \le 1$. The local Taylor inequality (valid for $|y| \le 1$) gives

$$e^{-y} \le 1 - y + \frac{e}{2}|y|^2 \le 1 - y + \frac{e}{2}|y|^p. \tag{128}$$

Let $y = \lambda U_i$ with $\lambda \in [0, L]$. Since $\mathbb{E}[U_i] = 0$ and $1 + a \le e^a$ for $a \in \mathbb{R}$, we obtain

$$\mathbb{E}[\exp(-\lambda U_i)] \le 1 + \frac{e}{2}|\lambda|^p \mathbb{E}|U_i|^p \le \exp\Big(\frac{e}{2}|\lambda|^p \mathbb{E}|U_i|^p\Big). \tag{129}$$

In addition,

$$|U_i|^p = |Z_i - \mathbb{E}Z_i|^p \le 2^{p-1}(|Z_i|^p + |\mathbb{E}Z_i|^p) \tag{130}$$

implies

$$\mathbb{E}|U_i|^p \le 2^p \mathbb{E}|Z_i|^p \le 2^p \nu. \tag{131}$$

Using independence and $\mathbb{E}|X_i|^p \le \nu$,

$$\mathbb{E}[\exp(-\lambda \sum_{i=1}^{s} U_i)] \le \exp(c_1(p)\nu s \lambda^p), \tag{132}$$

where $c_1(p)$ is a constant depending only on p. Hence,

$$M_s(\lambda) := \exp\left(-\lambda \sum_{i=1}^{s} U_i - c_1 \nu s \lambda^p\right), \qquad \lambda \in [0, L], \tag{133}$$

is a nonnegative supermartingale for $s \ge 0$. Now fix an arbitrary $\epsilon > 0$. If

$$\frac{1}{s}\sum_{i=1}^{s} U_i \le -\epsilon, \tag{134}$$

then

$$-\lambda \sum_{i=1}^{s} U_i - c_1 \nu s \lambda^p \ge s(\lambda \epsilon - c_1 \nu \lambda^p). \tag{135}$$

Therefore, for any interval $I \subseteq \{1, \dots, T\}$,

$$\left\{\exists s \in I : \frac{1}{s}\sum_{i=1}^{s} U_i \le -\epsilon\right\} \subseteq \left\{\exists s \in I : M_s(\lambda) \ge \exp(s(\lambda\epsilon - c_1\nu\lambda^p))\right\}. \tag{136}$$

Let $f(\lambda) := \lambda\epsilon - c_1\nu\lambda^p$. Then, the unconstrained maximizer is

$$\lambda^* = \left(\frac{\epsilon}{c_1 \nu p}\right)^{\frac{1}{p-1}}, \tag{137}$$

and

$$\sup_{\lambda \ge 0} f(\lambda) = \frac{p-1}{p} \cdot \frac{\epsilon^{\frac{p}{p-1}}}{(c_1\nu)^{1/(p-1)}} =: c_2(p) \frac{\epsilon^{\frac{p}{p-1}}}{\nu^{1/(p-1)}}. \tag{138}$$

The interval where $f(\lambda)$ is strictly positive contains the unconstrained maximizer. Indeed, $f(\lambda) > 0$ if and only if

$$0 \le \lambda < \left(\frac{\epsilon}{c_1\nu}\right)^{\frac{1}{p-1}}, \tag{139}$$

and

$$\lambda^* < p^{\frac{1}{p-1}} \lambda^* = \left(\frac{\epsilon}{c_1\nu}\right)^{\frac{1}{p-1}}. \tag{140}$$

Thus, the positivity region strictly contains λ^* for all $p > 1$. Pick λ so that $f(\lambda) > 0$. Then, for every $s \in I$, we have

$$\exp(s f(\lambda)) \ge \exp(s_0 f(\lambda)), \tag{141}$$

where $s_0 := \min I$. Since $\mathbb{E}[M_0(\lambda)] = 1$, Ville's inequality (Lemma 11) yields

$$\mathbb{P}\left\{\exists s \in I : \frac{1}{s}\sum_{i=1}^{s} U_i \le -\epsilon\right\} \le \exp\left(-s_0 \sup_{\lambda \in [0,L]} (\lambda\epsilon - c_1\nu\lambda^p)\right). \tag{142}$$

We now distinguish two cases. If $\lambda^* \le L$, then

$$\sup_{\lambda\in[0,L]} f(\lambda) = c_2(p)\frac{\epsilon^{\frac{p}{p-1}}}{\nu^{1/(p-1)}}. \tag{143}$$

Otherwise, if $\lambda^* > L$, then concavity of f gives

$$\sup_{\lambda\in[0,L]} f(\lambda) = f(L) = \epsilon L - c_1\nu L^p, \tag{144}$$

and $\lambda^* > L \iff \epsilon > c_1\nu p L^{p-1}$ implies $c_1\nu L^p < (\epsilon/p)L$. Thus,

$$f(L) \ge \left(1-\frac{1}{p}\right)\epsilon L. \tag{145}$$

Therefore, for every $\epsilon > 0$,

$$\sup_{\lambda\in[0,L]} f(\lambda) \ge \min\left\{c_2(p)\frac{\epsilon^{\frac{p}{p-1}}}{\nu^{1/(p-1)}}, \left(1-\frac{1}{p}\right)\epsilon L\right\}. \tag{146}$$

Consequently,

$$\mathbb{P}\{\exists s \in I : \hat{\mu}_s + \beta_s \le \mu - x\} \le \exp\left(-s_0 \min\left\{c_2(p)\frac{\epsilon^{\frac{p}{p-1}}}{\nu^{1/(p-1)}}, \left(1-\frac{1}{p}\right)\epsilon L\right\}\right). \tag{147}$$

Step 3: Peeling. We now partition the full index set into short blocks and apply the previous interval bound blockwise. Let $I = \{1,\dots,T\}$. Partition I into peeling blocks

$$I_j := \left\{s \in \mathbb{Z}_+ : \left\lceil\frac{j}{\gamma}\right\rceil \le s < \left\lceil\frac{j+1}{\gamma}\right\rceil\right\}, \qquad j = 1,2,\dots, \tag{148}$$

with

$$\gamma = \frac{c_\gamma(p)x^{\frac{p}{p-1}}}{(2\nu)^{1/(p-1)}}. \tag{149}$$

For a fixed block I_j, set

$$s_{j0} := \min I_j, \qquad s_{j*} := \max I_j, \qquad \epsilon_j := \frac{x + \beta_{s_{j*}}}{2}. \tag{150}$$

For notational convenience, within each block we write

$$s_0 := s_{j0}, \qquad s_* := s_{j*}, \qquad \epsilon := \epsilon_j. \tag{151}$$

Since β_s is nonincreasing in s, replacing β_s by β_{s_*} gives a valid lower bound throughout the block.

Also, since $(a+b)^q \ge a^q + b^q$ for any $q > 1$ and $a, b \ge 0$, we have

$$\epsilon^{p/(p-1)} \ge 2^{-p/(p-1)}\left(x^{p/(p-1)} + \beta_{s_*}^{p/(p-1)}\right). \tag{152}$$

We now consider the two cases separately.

Case 1: $\lambda^* \in [0, L]$. In this case, the unconstrained maximizer is feasible, so the blockwise bound is governed by the term

$$c_2(p)\frac{\epsilon^{p/(p-1)}}{\nu^{1/(p-1)}}. \tag{153}$$

Using the lower bound on $\epsilon^{p/(p-1)}$,

$$s_0 c_2 \frac{\epsilon^{\frac{p}{p-1}}}{\nu^{1/(p-1)}} \geq \frac{c_2 s_0}{2^{p/(p-1)}\nu^{1/(p-1)}} \left(x^{\frac{p}{p-1}} + \beta_{s_*}^{\frac{p}{p-1}} \right) = \frac{c_2 s_0 x^{\frac{p}{p-1}}}{2^{p/(p-1)}\nu^{1/(p-1)}} + \frac{c_2 s_0 \beta_{s_*}^{\frac{p}{p-1}}}{2^{p/(p-1)}\nu^{1/(p-1)}}. \tag{154}$$

Hence,

$$\mathbb{P}\{\exists s \in I : \hat{\mu}_s + \beta_s \leq \mu - x\} \leq \sum_{j\geq 0} \mathbb{P}\{\exists s \in I_j : \hat{\mu}_s + \beta_s \leq \mu - x\} \tag{155}$$

$$\leq \sum_{j\geq 0} \exp\left(-c_2 s_0 \frac{\epsilon^{\frac{p}{p-1}}}{\nu^{1/(p-1)}} \right) \tag{156}$$

$$\leq \sum_{j\geq 0} \exp\left(-\frac{c_2 s_0 x^{\frac{p}{p-1}}}{2^{p/(p-1)}\nu^{1/(p-1)}} \right) \exp\left(-\frac{c_2 s_0 \beta_{s_*}^{\frac{p}{p-1}}}{2^{p/(p-1)}\nu^{1/(p-1)}} \right). \tag{157}$$

The first exponential gives the desired decay in x, while the second exponential is used to recover the factor K/T through the specific form of β_s. More precisely,

$$\mathbb{P}\{\exists s \in I : \hat{\mu}_s + \beta_s \leq \mu - x\} \leq \frac{K}{T\gamma} \sum_{j\geq 0} (j+1) \exp\left(-\frac{c_2 (j/\gamma) x^{\frac{p}{p-1}}}{2^{p/(p-1)}\nu^{1/(p-1)}} \right) \tag{158}$$

$$\leq \frac{K}{T\gamma} \sum_{j=0}^{\infty} (j+1) \exp(-j) \tag{159}$$

$$\leq O\left(\frac{K\nu^{1/(p-1)}}{T x^{p/(p-1)}} \right), \tag{160}$$

where (158) follows by setting $c \geq \frac{2}{c_2}$, and (159) holds by choosing $c_\gamma = c_2/2$. The last bound uses the finiteness of $\sum_{j\geq 0}(j+1)e^{-j}$ together with the definition of γ.

Case 2: $\lambda^* \notin [0, L]$. In this case, the admissible interval truncates the unconstrained maximizer, so the blockwise bound is governed by the boundary value $f(L)$. We keep the same

$$\epsilon = \frac{x + \beta_{s_*}}{2}, \qquad L = \frac{1}{2b^*}, \tag{161}$$

where $b^* = \max_{s\in I_j} b_s$. We first compare s_* and s_0 within the same block. Since

$$s_* \leq \left\lceil \frac{j+1}{\gamma} \right\rceil - 1 < \frac{j+1}{\gamma} \qquad \text{and} \qquad s_0 \geq \frac{j}{\gamma}, \tag{162}$$

for $j \geq 1$ we have

$$\frac{s_*}{s_0} < \frac{(j+1)/\gamma}{j/\gamma} = \frac{j+1}{j} \leq 2. \tag{163}$$

Since $s_* \leq 2s_0$, it follows that

$$\frac{\log_+(T/Ks_0)}{\log_+(T/Ks_*)} \leq \max\{1 + \log 2, (1 - \log 2)^{-1}\} =: M, \tag{164}$$

and therefore

$$\frac{b^*}{b_{s_0}} = \left(\frac{s_*}{s_0} \cdot \frac{\log_+(T/Ks_0)}{\log_+(T/Ks_*)} \right)^{1/p} \leq (2M)^{1/p}. \tag{165}$$

Hence,

$$L = \frac{1}{2b^*} \geq \frac{1}{2(2M)^{1/p} b_{s_0}}. \tag{166}$$

Similarly,

$$\frac{\beta_{s_*}}{\beta_{s_0}} = \left(\frac{s_0}{s_*} \cdot \frac{\log_+(T/Ks_*)}{\log_+(T/Ks_0)}\right)^{1-\frac{1}{p}} \geq \left(\frac{1}{2M}\right)^{1-\frac{1}{p}}. \tag{167}$$

Now split the exponent linearly:

$$s_0\epsilon\left(1-\frac{1}{p}\right)L = s_0\left(1-\frac{1}{p}\right)\frac{x}{4b^*} + s_0\left(1-\frac{1}{p}\right)\frac{\beta_{s_*}}{4b^*}. \tag{168}$$

We treat the two terms separately.

For the second term,

$$\frac{s_0\beta_{s_*}}{4b^*} \geq \frac{s_0\beta_{s_0}}{4(2M)^{1/p}b_{s_0}}\left(\frac{1}{2M}\right)^{1-\frac{1}{p}} = \frac{s_0}{2M}\cdot\frac{c^{1-\frac{1}{p}}(2\nu)^{\frac{1}{p}}(\log_+(T/Ks_0))^{1-\frac{1}{p}}s_0^{\frac{1}{p}-1}}{\nu^{\frac{1}{p}}s_0^{\frac{1}{p}}2^{-\frac{1}{p}}(\log_+(T/Ks_0))^{-\frac{1}{p}}} = M^{-1}c^{1-\frac{1}{p}}2^{\frac{2}{p}-3}\log_+(T/Ks_0). \tag{169}$$

Choosing c such that

$$c^{1-\frac{1}{p}} \geq M2^{3-\frac{2}{p}}(1-1/p)^{-1}, \tag{170}$$

we obtain

$$\exp\left(-s_0\left(1-\frac{1}{p}\right)\frac{\beta_{s_*}}{4b^*}\right) \leq \exp\left(-\log_+\left(\frac{T}{Ks_0}\right)\right) \leq \frac{Ks_0}{T}. \tag{171}$$

For the first term, since $\frac{1}{2b^*} \geq \frac{1}{2(2M)^{1/p}b_{s_0}}$,

$$s_0\left(1-\frac{1}{p}\right)\frac{x}{4b^*} \geq s_0\left(1-\frac{1}{p}\right)\frac{x}{4(2M)^{1/p}b_{s_0}}. \tag{172}$$

By definition,

$$\frac{s_0}{b_{s_0}} = 2^{\frac{1}{p}}\nu^{-\frac{1}{p}}s_0^{1-\frac{1}{p}}\left(\log_+\left(\frac{T}{Ks_0}\right)\right)^{\frac{1}{p}}. \tag{173}$$

Hence,

$$\left(1-\frac{1}{p}\right)\frac{s_0x}{4(2M)^{1/p}b_{s_0}} = \left(1-\frac{1}{p}\right)\frac{x}{4(2M)^{1/p}}2^{\frac{1}{p}}\nu^{-\frac{1}{p}}s_0^{1-\frac{1}{p}}\left(\log_+\left(\frac{T}{Ks_0}\right)\right)^{\frac{1}{p}} \geq \left(1-\frac{1}{p}\right)\frac{x}{4M^{1/p}}\nu^{-\frac{1}{p}}s_0^{1-\frac{1}{p}}, \tag{174}$$

where the last inequality uses $\log_+(\cdot) \geq 1$. Thus,

$$\exp\left(-s_0\left(1-\frac{1}{p}\right)\frac{x}{4b^*}\right) \leq \exp\left(-\left(1-\frac{1}{p}\right)\frac{x}{4M^{1/p}}\nu^{-\frac{1}{p}}s_0^{1-\frac{1}{p}}\right). \tag{175}$$

Therefore, when $\lambda^* \notin [0, L]$,

$$\begin{aligned}
\mathbb{P}\{\exists s \in I_j : \hat{\mu}_s + \beta_s \leq \mu - x\} &\leq \exp\left(-s_0(1-1/p)\epsilon L\right) && (176)\\
&\leq \exp\left(-s_0\left(1-\frac{1}{p}\right)\frac{x}{4b^*}\right)\exp\left(-s_0\left(1-\frac{1}{p}\right)\frac{\beta_{s_*}}{4b^*}\right) && (177)\\
&\leq \frac{Ks_0}{T}\exp(-c_3xs_0^{1-\frac{1}{p}}) && (178)
\end{aligned}$$

with

$$c \geq \left(\frac{pM}{p-1}\cdot 2^{3-\frac{2}{p}}\right)^{\frac{p}{p-1}} \qquad \text{and} \qquad c_3 := \left(1-\frac{1}{p}\right)\frac{1}{4M^{1/p}}\nu^{-\frac{1}{p}}. \tag{179}$$

Let

$$\psi(z) := z \exp(-c_3 x z^{1-\frac{1}{p}}). \tag{180}$$

Then one can check that $\psi(z)$ is unimodal for $z \geq 0$. By Lemma 13, we obtain

$$\begin{aligned}
\mathbb{P}\{\exists s \in I : \hat{\mu}_s + \beta_s \leq \mu - x\} &\leq \sum_{j \geq 0} \mathbb{P}\{\exists s \in I_j : \hat{\mu}_s + \beta_s \leq \mu - x\} && (181)\\
&\leq \sum_{j \geq 0} \frac{K s_0}{T} \exp(-c_3 x s_0^{1-\frac{1}{p}}) && (182)\\
&= \sum_{j \geq 0} \frac{K}{T} \psi(s_0) && (183)\\
&\leq \frac{K}{T} \left(\gamma \int_0^\infty y \exp(-c_3 x y^{1-\frac{1}{p}}) dy + \sup_{y \geq 0} \psi(y) \right). && (184)
\end{aligned}$$

Substituting $u = c_3 x y^q$ with $q = 1 - \frac{1}{p}$, we have

$$y = \left(\frac{u}{c_3 x} \right)^{1/q}, \qquad dy = \frac{1}{q} (c_3 x)^{-\frac{1}{q}} u^{\frac{1}{q}-1} du. \tag{185}$$

Then

$$\begin{aligned}
\int_0^\infty y \exp(-c_3 x y^q) dy &= \int_0^\infty \left(\frac{u}{c_3 x} \right)^{\frac{1}{q}} \exp(-u) \frac{1}{q} (c_3 x)^{-\frac{1}{q}} u^{\frac{1}{q}-1} du && (186)\\
&= \frac{1}{q} (c_3 x)^{-\frac{2}{q}} \int_0^\infty u^{\frac{2}{q}-1} \exp(-u) du && (187)\\
&= \frac{1}{q} (c_3 x)^{-\frac{2}{q}} \Gamma\left(\frac{2}{q} \right). && (188)
\end{aligned}$$

Also, solving $\psi'(y) = 0$ gives the maximizer

$$y^* := (c_3 x q)^{-1/q}, \tag{189}$$

and hence

$$\sup_{y \geq 0} \psi(y) = \psi(y^*) = (c_3 x q)^{-1/q} e^{-1/q}. \tag{190}$$

Therefore,

$$\begin{aligned}
\mathbb{P}\{\exists s \in I : \hat{\mu}_s + \beta_s \leq \mu - x\} &\leq \frac{K}{T} \left(\gamma \frac{1}{q} (c_3 x)^{-\frac{2}{q}} \Gamma\left(\frac{2}{q} \right) + (c_3 x q)^{-1/q} e^{-1/q} \right) && (191)\\
&\leq O\left(\frac{K \nu^{1/(p-1)}}{T x^{p/(p-1)}} \right), && (192)
\end{aligned}$$

where the last inequality follows from the definition of

$$\gamma = \frac{c_\gamma(p) x^{\frac{p}{p-1}}}{(2\nu)^{1/(p-1)}}. \tag{193}$$

Combining (160) and (192), we conclude that

$$\mathbb{P}\{\exists s \in [T] : \hat{\mu}_s + \beta_s \leq \mu - x\} \leq O\left(\frac{K \nu^{1/(p-1)}}{T x^{p/(p-1)}} \right), \tag{194}$$

which completes the proof. □

E Proof of Theorem 4

Proof. We prove the gap-dependent regret bound by separating arms according to whether their gaps are small or large. For arms with very small gaps, a trivial bound is sufficient. For arms with larger gaps, we show that once such an arm has been sampled sufficiently many times, selecting it can happen only if at least one of several unfavorable events occurs. We then bound the probability of each such event and sum the resulting contributions.

Step 1: Regret decomposition. We start from the standard regret decomposition

$$R(T) = \sum_{i:\Delta_i>0} \mathbb{E}[n_i(T)]\Delta_i. \tag{195}$$

We split the arms according to whether their gaps are below or above the threshold $(eK/T)^{1-1/p}$:

$$R(T) = \sum_{i:\Delta_i>0} \mathbb{E}[n_i(T)]\Delta_i \le \sum_{i:\Delta_i\le(eK/T)^{1-1/p}} T\Delta_i + \sum_{i:\Delta_i>(eK/T)^{1-1/p}} \mathbb{E}[n_i(T)]\Delta_i \tag{196}$$

$$\le \sum_{i:\Delta_i\le(eK/T)^{1-1/p}} \left(\frac{eK}{\Delta_i^{\frac{p}{p-1}}}\right)\cdot\Delta_i + \sum_{i:\Delta_i>(eK/T)^{1-1/p}} \mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{I_t=i\}\Delta_i\right]. \tag{197}$$

The second inequality uses the condition $\Delta_i \le (eK/T)^{1-1/p}$, which implies

$$T \le \frac{eK}{\Delta_i^{p/(p-1)}}. \tag{198}$$

Thus, the first term already has the desired order, and it remains to control the second term corresponding to arms with sufficiently large gaps.

Step 2: A decomposition for large-gap arms. Recall that the sampling rule of Algorithm 1 can be interpreted as

$$I_t = \arg\max_{j\in[K]}(\hat{\mu}_{tj} + \beta_{tj} + g_{tj}), \tag{199}$$

where $g_{tj} \sim G(0, \alpha/t)$ and $\alpha = c^{\frac{p}{p-1}}\nu^{\frac{1}{p}}$. Let

$$\theta_j(t) = \hat{\mu}_{tj} + \beta_{tj} + g_{tj}, \tag{200}$$

for all $j \in [K]$ and $t \ge 1$. Then, for an arbitrary $\ell > 0$ and $i \in [K]$, we have

$$\mathbb{E}\left[n_i(T)\right] = \mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{I_t=i\}\right] \le \ell + \sum_{t=K+1}^{T}\mathbb{P}\{\theta_i(t) > \theta_1(t),\, n_i(t) > \ell\}. \tag{201}$$

In words, the expected number of pulls of arm i is bounded by the first ℓ pulls, plus the probability that arm i is selected after it has already been sampled more than ℓ times.

We now consider the case $\Delta_i > (eK/T)^{1-1/p}$. For such an arm, we show that if arm i is selected at time t, then at least one of several unfavorable events must occur.

Define

$$E_1 := \{\hat{\mu}_{t1} + \beta_{t1} \le \mu_1 - \tfrac{\Delta_i}{4}\} \tag{202}$$

$$E_2 := \{\hat{\mu}_{ti} - \beta_{ti} > \mu_i + \tfrac{\Delta_i}{4}\} \tag{203}$$

$$E_3 := \{g_{ti} - g_{t1} > \tfrac{\Delta_i}{4}\} \tag{204}$$

$$E_4 := \{\beta_{ti} > \tfrac{\Delta_i}{8}\}, \tag{205}$$

where

$$\beta_{ti} := \left(\frac{c(2\nu)^{1/(p-1)} \log_+(T/Kn_i(t))}{n_i(t)} \right)^{1-\frac{1}{p}}. \tag{206}$$

These events correspond, respectively, to: - underestimation of the optimal arm, - overestimation of arm i, - an unusually favorable perturbation difference for arm i, - and a confidence bonus for arm i that is still too large.

We claim that if all four events are false, then $\theta_1(t) > \theta_i(t)$ must hold. Indeed, suppose that all the events are false. Then, by E_1^c,

$$\theta_1(t) = \hat{\mu}_{t1} + \beta_{t1} + g_{t1} > \mu_1 - \frac{\Delta_i}{4} + g_{t1}. \tag{207}$$

Also, by E_2^c and E_4^c,

$$\theta_i(t) = \hat{\mu}_{ti} + \beta_{ti} + g_{ti} \leq \mu_i + 2\beta_{ti} + \frac{\Delta_i}{4} + g_{ti} \leq \mu_i + \frac{2}{4}\Delta_i + g_{ti}. \tag{208}$$

Hence,

$$\begin{aligned}
\theta_i(t) - \theta_1(t) &< \left(\mu_i + \frac{2}{4}\Delta_i + g_{ti}\right) - \left(\mu_1 - \frac{\Delta_i}{4} + g_{t1}\right) && (209)\\
&= -\Delta_i + g_{ti} + \frac{3}{4}\Delta_i - g_{t1} && (210)\\
&= -\frac{\Delta_i}{4} + g_{ti} - g_{t1} && (211)\\
&\leq 0, && (212)
\end{aligned}$$

where the last inequality follows from E_3^c. Therefore,

$$\{\theta_i(t) > \theta_1(t),\, n_i(t) > \ell\} \subseteq E_1 \cup E_2 \cup E_3 \cup (E_4 \cap \{n_i(t) > \ell\}). \tag{213}$$

Step 3: Choice of ℓ and probability bounds. We now choose ℓ so that E_4 cannot occur once arm i has been sampled more than ℓ times. Let

$$\ell := \left\lceil \frac{c8^{\frac{p}{p-1}}(2\nu)^{\frac{1}{p-1}} \log((T\Delta_i^{\frac{p}{p-1}})/K)}{\Delta_i^{\frac{p}{p-1}}} \right\rceil. \tag{214}$$

Then, for $\Delta_i > (eK/T)^{1-1/p}$, the event E_4 does not hold whenever $n_i(t) > \ell$. Thus,

$$\begin{aligned}
\mathbb{P}\{\theta_i(t) > \theta_1(t), n_i(t) > \ell\} &\leq \mathbb{P}\{E_1\} + \mathbb{P}\{E_2\} + \mathbb{P}\{E_3\} + \mathbb{P}\{E_4, n_i(t) > \ell\} && (215)\\
&\leq \mathbb{P}\{E_1\} + \mathbb{P}\{E_2\} + \mathbb{P}\{E_3\}. && (216)
\end{aligned}$$

Now, both E_1 and E_2 can be controlled using Lemma 7. Indeed,

$$E_1 \subseteq \left\{\exists\, s \in [T] :\ \hat{\mu}_{s1} + \beta_{s1} \leq \mu_1 - \tfrac{\Delta_i}{4}\right\},$$

and an analogous statement holds for E_2. Therefore, there exists $c' > 0$ such that

$$\mathbb{P}\{E_1\} + \mathbb{P}\{E_2\} \leq \frac{Kc'(2\nu)^{\frac{1}{p-1}} 4^{\frac{p}{p-1}}}{T\Delta_i^{\frac{p}{p-1}}}. \tag{217}$$

Hence,

$$\mathbb{P}\{\theta_i(t) > \theta_1(t), n_i(t) > \ell\} \leq \frac{Kc'(2\nu)^{\frac{1}{p-1}} 4^{\frac{p}{p-1}}}{T\Delta_i^{\frac{p}{p-1}}} + \mathbb{P}\{E_3\}. \tag{218}$$

It remains to bound E_3. Since $g_{t1}, g_{ti} \sim G(0, \alpha/t)$, we know that

$$g_{ti} - g_{t1} \sim \text{Logistic}(0, \alpha/t). \tag{219}$$

Thus, for all $t \geq 1$,

$$\mathbb{P}\{E_3\} = \mathbb{P}\{g_{ti} - g_{t1} > \frac{\Delta_i}{4}\} = 1 - F_{\text{Logistic}}(\Delta_i/4) = \frac{1}{1+\exp((t/\alpha)\cdot(\Delta_i/4))} \leq \exp\Big(-\frac{t}{\alpha}\cdot\frac{\Delta_i}{4}\Big), \tag{220}$$

where $F_{\text{Logistic}}(x) := \frac{1}{1+\exp(-(t/\alpha)x)}$ denotes the CDF of Logistic$(0, \alpha/t)$. Therefore,

$$\sum_{t=K+1}^{T} \mathbb{P}\{E_3\} \leq \sum_{t=K+1}^{T} \exp\left(-\frac{t}{\alpha}\cdot\frac{\Delta_i}{4}\right) \leq \frac{4\alpha}{\Delta_i}. \tag{221}$$

Step 4: Final summation. We now combine the above bounds. Using the decomposition of $\mathbb{E}[n_i(T)]$ and multiplying by Δ_i, we obtain

$$R(T) = \sum_{i:\Delta_i>0} \mathbb{E}[n_i(T)]\Delta_i \leq \sum_{i:\Delta_i>0} \frac{Ke}{\Delta_i^{1/(p-1)}} + \sum_{i:\Delta_i>0} \Delta_i \left[\ell + \sum_{t=\ell}^{T} \mathbb{P}\{\theta_i(t) > \theta_1(t), n_i(t) > \ell\}\right]. \tag{222}$$

Substituting the bound on ℓ and the estimate (218), together with the bound on the sum of $\mathbb{P}(E_3)$, yields

$$R(T) \leq \sum_{i:\Delta_i>0} \frac{Ke}{\Delta_i^{1/(p-1)}} + \sum_{i:\Delta_i>0} \left[\frac{c8^{\frac{p}{p-1}}(2\nu)^{\frac{1}{p-1}}\log(T\Delta_i^{\frac{p}{p-1}}/K)}{\Delta_i^{1/(p-1)}} + \frac{Kc'(2\nu)^{\frac{1}{p-1}}4^{\frac{p}{p-1}}}{\Delta_i^{1/(p-1)}} + 4\alpha\right]. \tag{223}$$

Finally, recalling that $\alpha = c^{p/(p-1)}\nu^{1/p}$ and absorbing constants into the big-O notation, we conclude that

$$R(T) \leq O\left(\sum_{i:\Delta_i>0} \frac{\nu^{\frac{1}{p-1}}\log(T\Delta_i^{\frac{p}{p-1}}/K) + K}{\Delta_i^{1/(p-1)}}\right). \tag{224}$$

This completes the proof. □

F Proof of Theorem 5

Proof. We prove the gap-independent regret bound by decomposing the regret into three parts. The first part corresponds to possible underestimation of the optimal arm. The second part comes from arms with very small gaps, for which a trivial regret bound is sufficient. The third part corresponds to arms with sufficiently large gaps, whose total contribution can be controlled uniformly. We treat these three terms in turn.

Step 1: Regret decomposition. We define

$$\Delta := \mu_1 - \min_{1\leq s\leq T}(\hat{\mu}_{s1} + \beta_{s1}), \tag{225}$$

and

$$S = \{i : \Delta_i > \max\{2\Delta, (eK/T)^{1-1/p}\}\}. \tag{226}$$

Here, Δ measures the amount by which the optimistic estimate of the optimal arm may fall below its true mean over the whole time horizon. Using this random quantity, we decompose the regret of Algorithm 1 as follows:

$$R(T) = \sum_{i:\Delta_i>0} \Delta_i \mathbb{E}[n_i(T)] \tag{227}$$

$$\leq \mathbb{E}[2T\Delta] + \mathbb{E}[\sum_{i:\Delta_i>2\Delta} \Delta_i n_i(T)] \tag{228}$$

$$\leq \mathbb{E}[2T\Delta] + T\cdot\left(\frac{eK}{T}\right)^{1-\frac{1}{p}} + \mathbb{E}[\sum_{i\in S} \Delta_i n_i(T)]. \tag{229}$$

The second term above comes from arms with gaps at most $(eK/T)^{1-1/p}$, which are controlled simply by the bound $n_i(T) \leq T$.

Step 2: Bounding the underestimation term $\mathbb{E}[2T\Delta]$. By the definition of Δ, for any $x > 0$,

$$\mathbb{P}\{\Delta \geq x\} = \mathbb{P}\{\mu_1 - \min_{1\leq s\leq T}(\hat{\mu}_{s1} + \beta_{s1}) - x \geq 0\} \tag{230}$$

$$= \mathbb{P}\{\exists s \in [T] : \mu_1 - \hat{\mu}_{s1} - \beta_{s1} - x \geq 0\}. \tag{231}$$

Thus, Lemma 7 implies that there exists $c' > 0$ such that

$$\mathbb{E}[2T\Delta] = 2T \int_{x=0}^{\infty} \mathbb{P}\{\Delta \geq x\} dx \leq 2T \int_{x=0}^{\infty} \min\left(1, \frac{Kc'(2\nu)^{1/(p-1)}}{Tx^{p/(p-1)}}\right) dx. \tag{232}$$

Let

$$x_0 := \left(\frac{Kc'(2\nu)^{1/(p-1)}}{T}\right)^{\frac{p-1}{p}}. \tag{233}$$

Splitting the integral at x_0, we obtain

$$\int_{x=0}^{\infty} \min\left(1, \frac{Kc'(2\nu)^{1/(p-1)}}{Tx^{p/(p-1)}}\right) dx = \int_{x=0}^{x_0} 1\, dx + \int_{x_0}^{\infty} \frac{Kc'(2\nu)^{1/(p-1)}}{Tx^{p/(p-1)}} dx \tag{234}$$

$$= x_0 + \frac{Kc'(2\nu)^{1/(p-1)}}{T} \int_{x_0}^{\infty} x^{-p/(p-1)} dx. \tag{235}$$

Now,

$$\int_{x_0}^{\infty} x^{-p/(p-1)} dx = \left[\frac{x^{1-p/(p-1)}}{1-p/(p-1)}\right]_{x_0}^{\infty} = (p-1)\left(\frac{Kc'(2\nu)^{\frac{1}{p-1}}}{T}\right)^{-\frac{1}{p}}. \tag{236}$$

Substituting this into (232), we obtain

$$\mathbb{E}[2T\Delta] \leq 2T \int_{x=0}^{\infty} \min\left(1, \frac{Kc'(2\nu)^{\frac{1}{p-1}}}{Tx^{\frac{p}{p-1}}}\right) dx \tag{237}$$

$$\leq 2Tx_0 + 2T\frac{Kc'(2\nu)^{\frac{1}{p-1}}}{T}(p-1)\left(\frac{Kc'(2\nu)^{\frac{1}{p-1}}}{T}\right)^{-\frac{1}{p}} \tag{238}$$

$$= 2T^{\frac{1}{p}}K^{1-\frac{1}{p}}(c')^{1-\frac{1}{p}}(2\nu)^{\frac{1}{p}} + 2(p-1)(c')^{1-\frac{1}{p}}(2\nu)^{\frac{1}{p}}K^{1-\frac{1}{p}}T^{\frac{1}{p}} \tag{239}$$

$$= O(\nu^{\frac{1}{p}}T^{\frac{1}{p}}K^{1-\frac{1}{p}}). \tag{240}$$

Step 3: Bounding the contribution of arms in S. We now turn to the third term of (229), namely the contribution of arms whose gaps are large enough relative to both the random quantity Δ and the threshold $(eK/T)^{1-1/p}$. Let $\epsilon = \Delta_i/2$ and choose

$$\ell := \left\lceil \frac{c(2\nu)^{\frac{1}{p-1}}8^{\frac{p}{p-1}}\log_+(T\Delta_i^{\frac{p}{p-1}}/K)}{\Delta_i^{p/(p-1)}} \right\rceil. \tag{241}$$

For $i \in S$, we have

$$\ell \geq c(2\nu)^{\frac{1}{p-1}}8^{\frac{p}{p-1}}\left(\frac{1}{\Delta_i}\right)^{\frac{p}{p-1}}\log_+\left(\frac{T}{K}\cdot\frac{eK}{T}\right) \geq c(2\nu)^{\frac{1}{p-1}}8^{\frac{p}{p-1}}\left(\frac{1}{\Delta_i}\right)^{\frac{p}{p-1}}. \tag{242}$$

Take $c > 0$ such that $c > \frac{1}{(2\nu)^{1/(p-1)}8^{p/(p-1)}}$. Then, for all $s \geq \ell$,

$$\beta_{si} \leq \beta_{\ell i} \leq \left\{\frac{c(2\nu)^{\frac{1}{p-1}}\log_+(T/Ks)}{\ell}\right\}^{1-\frac{1}{p}} \leq \left\{\frac{c(2\nu)^{\frac{1}{p-1}}\log_+\left(\frac{T}{K}\cdot(\Delta_i^{\frac{p}{p-1}})\right)}{\ell}\right\}^{1-\frac{1}{p}} \leq \frac{\Delta_i}{8}. \tag{243}$$

Thus, once arm i has been pulled more than ℓ times, its confidence bonus is already small compared to the gap Δ_i. For $i \in S$, we decompose $\Delta_i \mathbb{E}[n_i(T)]$ as

$$\Delta_i \mathbb{E}[n_i(T)] \leq \Delta_i(\ell+1) + \Delta_i \mathbb{E}\left[\sum_{t=\ell}^{T} \mathbf{1}\{I_t = i, E_i^c(t), n_i(t) > \ell\}\right] + \Delta_i \mathbb{E}\left[\sum_{t=\ell}^{T} \mathbf{1}\{I_t = i, E_i(t), n_i(t) > \ell\}\right] \tag{244}$$

where $E_i(t) = \{\theta_i(t) \leq \mu_1 - \epsilon\}$. The first term accounts for the first ℓ pulls of arm i. The second term corresponds to the event that arm i is selected while its perturbed score is unusually large. The third term corresponds to the event that arm i is selected even though its perturbed score is below $\mu_1 - \epsilon$.

Step 3a: Bounding the $E_i^c(t)$ term. Recall that $E_i^c(t) = \{\theta_i(t) > \mu_1 - \frac{\Delta_i}{2}\}$, $\alpha = c^{\frac{p}{p-1}} \nu^{\frac{1}{p}}$, and $\theta_i(t) \sim G(\hat{\mu}_{ti} + \beta_{ti}, \frac{\alpha}{t})$. Therefore,

$$\Delta_i \mathbb{E}\left[\sum_{t=\ell}^{T} \mathbf{1}\{I_t = i, E_i^c(t), n_i(t) > \ell\}\right] = \Delta_i \sum_{t=\ell}^{T} \mathbb{P}\{I_t = i, \theta_i(t) > \mu_1 - \frac{\Delta_i}{2}, n_i(t) > \ell\}. \tag{245}$$

We bound the probability on the right-hand side by splitting according to whether the empirical estimate of arm i is too optimistic:

$$\mathbb{P}\{I_t = i, \theta_i(t) > \mu_1 - \frac{\Delta_i}{2}\} \leq \mathbb{P}\{\theta_i(t) > \mu_1 - \frac{\Delta_i}{2}, \mu_i - \hat{\mu}_{ti} > \beta_{ti} - \frac{\Delta_i}{4}\} \tag{246}$$

$$+ \mathbb{P}\{I_t = i, \mu_i - \hat{\mu}_{ti} < \beta_{ti} - \frac{\Delta_i}{4}\}. \tag{247}$$

For the first term, since $\theta_i(t) = \hat{\mu}_{ti} + \beta_{ti} + g_{ti}$ and $\theta_i(t) > \mu_1 - \frac{\Delta_i}{2}$, we have

$$g_{ti} > \mu_1 - \frac{\Delta_i}{2} - \hat{\mu}_{ti} - \beta_{ti} \tag{248}$$

$$= \mu_1 - \mu_i + \mu_i - \frac{\Delta_i}{2} - \hat{\mu}_{ti} - \beta_{ti} \tag{249}$$

$$= \Delta_i + \mu_i - \frac{\Delta_i}{2} - \hat{\mu}_{ti} - \beta_{ti} \tag{250}$$

$$= \frac{\Delta_i}{2} + \mu_i - \hat{\mu}_{ti} - \beta_{ti} \tag{251}$$

$$> \frac{\Delta_i}{4}, \tag{252}$$

where the last inequality uses the condition $\mu_i - \hat{\mu}_{ti} > \beta_{ti} - \frac{\Delta_i}{4}$. For the second term, since $\beta_{ti} \leq \frac{\Delta_i}{8}$ when $n_i(t) \geq \ell$,

$$\hat{\mu}_{ti} - \mu_i > \frac{\Delta_i}{4} - \beta_{ti} > \frac{\Delta_i}{8}. \tag{253}$$

Therefore,

$$\mathbb{P}\{I_t = i, \theta_i(t) > \mu_1 - \frac{\Delta_i}{2}, n_i(t) > \ell\} \leq \mathbb{P}\{g_{ti} > \frac{\Delta_i}{4}\} + \mathbb{P}\{I_t = i,\ \hat{\mu}_{ti} - \mu_i > \frac{\Delta_i}{8}\}. \tag{254}$$

Let $\tau_a(k)$ denote the time step at which arm a is pulled for the k-th time. Then

$$\sum_{t=\ell}^{T} \mathbb{P}\{I_t = i,\ \theta_i(t) > \mu_1 - \frac{\Delta_i}{2}, n_i(t) > \ell\} \tag{255}$$

$$\leq \sum_{t=\ell}^{T} \exp\left(-t \cdot \frac{\Delta_i}{4\alpha}\right) + \mathbb{E}\left[\sum_{k=\ell}^{T} \sum_{t=\tau_i(k)}^{\tau_i(k+1)-1} \mathbf{1}\{I_t = i\} \cdot \mathbf{1}\{\hat{\mu}_{ti} - \mu_i > \frac{\Delta_i}{8}\}\right] \tag{256}$$

$$\leq \sum_{t=\ell}^{T} \exp\left(-t \cdot \frac{\Delta_i}{4\alpha}\right) + \mathbb{E}\left[\sum_{k=\ell}^{T} \mathbf{1}\{\hat{\mu}_{\tau_i(k),i} - \mu_i \geq \frac{\Delta_i}{8}\} \sum_{t=\tau_i(k)}^{\tau_i(k+1)-1} \mathbf{1}\{I_t = i\}\right] \tag{257}$$

$$\leq \sum_{t=1}^{\infty} \exp\left(-t \cdot \frac{\Delta_i}{4\alpha}\right) + \sum_{k\geq\ell} \mathbb{P}\{\hat{\mu}_{\tau_i(k),i} - \mu_i > \frac{\Delta_i}{8}\} \tag{258}$$

$$\leq \frac{1}{\exp\left(\Delta_i/4\alpha\right) - 1} + \sum_{k\geq 1} \exp\left(-\frac{k^{1-\frac{1}{p}}}{c\nu^{1/p}} \cdot \frac{\Delta_i}{8}\right) \tag{259}$$

$$\leq \frac{4\alpha}{\Delta_i} + 1 + c_0 \nu^{\frac{1}{p-1}} \left(\frac{8}{\Delta_i}\right)^{\frac{p}{p-1}} \tag{260}$$

where (259) follows by Theorem 2, and the last inequality uses $1 + x \leq e^x$ for all $x \in \mathbb{R}$. Here, c_0 is a constant depending only on p. Combining (245) and (260), we obtain

$$\Delta_i \mathbb{E}\left[\sum_{t=\ell}^{T} \mathbf{1}\{I_t = i, E_i^c(t), n_i(t) > \ell\}\right] \leq \Delta_i \left(\frac{4\alpha}{\Delta_i} + c\nu^{\frac{1}{p-1}} \left(\frac{8}{\Delta_i}\right)^{\frac{p}{p-1}}\right) \tag{261}$$

$$\leq O\left(\nu^{\frac{1}{p-1}} \Delta_i^{-\frac{1}{p-1}}\right) \tag{262}$$

$$\leq O\left(\nu^{\frac{1}{p-1}} \left(\frac{T}{K}\right)^{\frac{1}{p}}\right). \tag{263}$$

Step 3b: Bounding the $E_i(t)$ term. We now consider the third term in (244). By the argument in the proof of Theorem 36.2 in (Lattimore and Szepesvári, 2020), we know that

$$\mathbb{E}\left[\sum_{t=\ell}^{T} \mathbf{1}\{I_t = i, E_i(t), n_i(t) > \ell\}\right] \leq \mathbb{E}\left[\sum_{s=\ell}^{T} \left(\frac{1}{G_{1s}(\epsilon)} - 1\right)\right]. \tag{264}$$

In Lemma 12, we defined

$$L' := \max\left(\left\lceil \left(\frac{4c\nu^{1/p} \log 2}{\epsilon}\right)^{\frac{p}{p-1}} + 1 \right\rceil, \left\lceil \frac{4\alpha}{\epsilon} \log 2 + 1 \right\rceil\right). \tag{265}$$

If $L' < \ell$, then Lemma 12 applies directly for all $s \geq \ell$, and therefore, for $\Delta_i > (eK/T)^{1-1/p}$,

$$\Delta_i \mathbb{E}\left[\sum_{s=\ell}^{T} \left(\frac{1}{G_{1s}(\Delta_i/2)} - 1\right)\right] \leq \Delta_i O\left(\frac{\nu^{1/(p-1)}}{\Delta_i^{p/(p-1)}} + \frac{\nu^{1/p}}{\Delta_i}\right) = O\left(\nu^{\frac{1}{p-1}} T^{\frac{1}{p}} K^{-\frac{1}{p}}\right). \tag{266}$$

Now suppose that $L' > \ell$. Then we split the sum at L':

$$\mathbb{E}\left[\sum_{t=\ell}^{T} \mathbf{1}\{I_t = i, E_i(t), n_i(t) > \ell\}\right] \leq \mathbb{E}\left[\sum_{t=1}^{L'} \mathbf{1}\{I_t = i, E_i(t), n_i(t) > \ell\}\right] + \mathbb{E}\left[\sum_{t=L'+1}^{T} \mathbf{1}\{I_t = i, E_i(t), n_i(t) > \ell\}\right] \tag{267}$$

$$\leq \sum_{t=1}^{L'} \mathbb{P}\{I_t = i, E_i(t), n_i(t) > \ell\} + \mathbb{E}\left[\sum_{s=L'+1}^{T} \left(\frac{1}{G_{1s}(\epsilon)} - 1\right)\right] \tag{268}$$

$$\leq L' + O\left(\frac{\nu^{1/(p-1)}}{\Delta_i^{p/(p-1)}} + \frac{\nu^{1/p}}{\Delta_i}\right) \tag{269}$$

$$\leq O\left(\frac{\nu^{1/(p-1)}}{\Delta_i^{p/(p-1)}} + \frac{\nu^{1/p}}{\Delta_i}\right). \tag{270}$$

Hence,

$$\Delta_i \mathbb{E}\left[\sum_{t=\ell}^{T} \mathbf{1}\{I_t = i, E_i(t), n_i(t) > \ell\}\right] \leq O(\nu^{\frac{1}{p-1}} T^{\frac{1}{p}} K^{-\frac{1}{p}}). \tag{271}$$

Thus, in both cases, the third term in (244) is of the desired order.

Step 4: Final combination. It remains to control the first term $\Delta_i \ell$. It can be verified that $\Delta_i \cdot \ell$ attains its maximum

$$c(2\nu)^{\frac{1}{p-1}} 8^{\frac{p}{p-1}} e^{-\frac{1}{p}} \left(\frac{T}{K}\right)^{1/p} \tag{272}$$

for $\Delta_i > (eK/T)^{1-1/p}$. Therefore,

$$R(T) \leq \mathbb{E}[2T\Delta] + T \cdot \left(\frac{eK}{T}\right)^{1-\frac{1}{p}} + \mathbb{E}[\sum_{i \in S} \Delta_i n_i(T)] \tag{273}$$

$$\leq O(\nu^{\frac{1}{p}} T^{\frac{1}{p}} K^{1-\frac{1}{p}}) + eT^{\frac{1}{p}} K^{1-\frac{1}{p}} \tag{274}$$

$$+ K \cdot \left(O(\nu^{\frac{1}{p-1}} T^{\frac{1}{p}} K^{-\frac{1}{p}}) + O(\nu^{\frac{1}{p-1}} T^{\frac{1}{p}} K^{-\frac{1}{p}})\right) + \sum_{i \in S} \Delta_i \cdot \ell + \sum_{i=1}^{K} \Delta_i \tag{275}$$

$$\leq O(\nu^{\frac{1}{p}} T^{\frac{1}{p}} K^{1-\frac{1}{p}} + \sum_{i=1}^{K} \Delta_i). \tag{276}$$

This completes the proof. □

G Proofs for Auxiliary Lemmas

G.1 Proof of Lemma 12

Proof. Recall that we have defined $\alpha = c^{\frac{p}{p-1}} \nu^{\frac{1}{p}}$. Then, the sampling rule of Algorithm 1 can be written as:

$$I_t = \arg\max_{i \in [K]} (\hat{\mu}_{ti} + \beta_{ti} + g_{ti}), \quad \text{where} \quad g_{ti} \sim G(0, \alpha/t). \tag{277}$$

Fix $s \in \mathbb{N}$ with $s = n_i(t)$ for some $t \geq 1$, and assume s satisfies the precondition of Theorem 2. Define E_s to be the event that $\hat{\mu}_{s1} + \beta_{s1} - \frac{\epsilon}{4} \geq \mu_1 - \frac{\epsilon}{2}$. In addition, we define X_{s1} as a random variable which is sampled from $G(\hat{\mu}_{s1} + \beta_{s1}, \alpha/t)$. Let $A_s := \frac{s}{\alpha} \cdot \frac{\epsilon}{4}$, $B_s := \frac{s^{1-\frac{1}{p}}}{c\nu^{1/p}} \cdot \frac{\epsilon}{4}$, $L_A := \frac{4\alpha}{\epsilon} \log 2$, and $L_B := \left(\frac{4c\nu^{1/p} \log 2}{\epsilon}\right)^{\frac{p}{p-1}}$. Since $\beta_{s1} \geq 0$, we have

$$\mathbb{P}(E_s) = \mathbb{P}\left\{\hat{\mu}_{s1} + \beta_{s1} \geq \mu_1 - \tfrac{\epsilon}{4}\right\} \geq \mathbb{P}\left\{\hat{\mu}_{s1} \geq \mu_1 - \tfrac{\epsilon}{4}\right\} \geq 1 - e^{-B_s} \tag{278}$$

where the last inequality holds by Theorem 2. Moreover, using $t \geq s$ and the elementary inequality $e^{-e^a} \leq e^{-a}$ for $a \geq 0$,

$$\mathbb{P}\{X_{s1} > \mu_1 - \epsilon | E_s\} = \mathbb{P}\{X_{s1} + \frac{\epsilon}{2} > \mu_1 - \frac{\epsilon}{2} | E_s\} \tag{279}$$

$$\geq \mathbb{P}\{X_{s1} \geq \hat{\mu}_{s1} + \beta_{s1} - \frac{3}{4}\epsilon | E_s\} \tag{280}$$

$$\geq \mathbb{P}\{X_{s1} \geq \hat{\mu}_{s1} + \beta_{s1} - \frac{\epsilon}{4} | E_s\} \tag{281}$$

$$\geq 1 - \exp\Big(- \exp\Big(\frac{s}{\alpha} \cdot \frac{\epsilon}{4}\Big)\Big) \tag{282}$$

$$\geq 1 - e^{-A_s}. \tag{283}$$

For any events A, E, the inequality $\mathbb{P}(A) \geq \mathbb{P}(A \mid E)\mathbb{P}(E)$ implies

$$\frac{1}{\mathbb{P}(A)} - 1 \;\leq\; \frac{1}{\mathbb{P}(A \mid E)\,\mathbb{P}(E)} - 1. \tag{284}$$

Applying this with $A = \{X_{s1} > \mu_1 - \epsilon\}$ and $E = E_s$ yields

$$\mathbb{E}\Big[\frac{1}{G_{1s}(\epsilon)} - 1\Big] \leq \mathbb{E}\Big[\frac{1}{(1 - e^{-A_s})(1 - e^{-B_s})} - 1\Big]. \tag{285}$$

Let $m_s := \min\{A_s, B_s\}$. Then

$$(1 - e^{-A_s})(1 - e^{-B_s}) \geq (1 - e^{-m_s})^2 \quad \Rightarrow \quad \frac{1}{(1 - e^{-A_s})(1 - e^{-B_s})} - 1 \leq \frac{1}{(1 - e^{-m_s})^2} - 1. \tag{286}$$

Choose L' so that $e^{-A_s} \leq \frac{1}{2}$ and $e^{-B_s} \leq \frac{1}{2}$ hold for all $s \geq L'$, i.e.,

$$s \geq L_A = \frac{4\alpha}{\epsilon} \log 2, \qquad s \geq L_B = \left(\frac{4c\,\nu^{1/p} \log 2}{\epsilon}\right)^{\frac{p}{p-1}}, \qquad L' := \max\{L_A, L_B\}. \tag{287}$$

For $y \in (0, \frac{1}{2}]$ one has $\frac{1}{(1-y)^2} - 1 \leq 8y$, so for all $s \geq L'$,

$$\frac{1}{(1 - e^{-A_s})(1 - e^{-B_s})} - 1 \leq 8e^{-m_s} \leq 8\big(e^{-A_s} + e^{-B_s}\big), \tag{288}$$

where we used $e^{-\min(a,b)} \leq e^{-a} + e^{-b}$. Summing and bounding each series separately,

$$\sum_{s=L'}^{\infty} \exp(-A_s) \leq \int_{L'-1}^{\infty} \exp\Big(-\frac{\epsilon}{4\alpha}x\Big)\,dx = \frac{4\alpha}{\epsilon} \exp\Big(-\frac{\epsilon}{4\alpha}(L' - 1)\Big) \leq \frac{4\alpha}{\epsilon}, \tag{289}$$

and with $q := 1 - \frac{1}{p} \in (0, 1)$ and $\kappa := \frac{\epsilon}{4c\,\nu^{1/p}}$,

$$\sum_{s=L'}^{\infty} \exp(-B_s) = \sum_{s=L'}^{\infty} \exp(-\kappa s^q) \leq \int_{L'-1}^{\infty} \exp(-\kappa x^q)\,dx \leq \frac{1}{q}\,\Gamma\Big(\frac{1}{q}\Big)\,\kappa^{-1/q}. \tag{290}$$

Combining the bounds,

$$\sum_{s=L'}^{\infty} \left(\frac{1}{(1 - e^{-A_s})(1 - e^{-B_s})} - 1\right) \leq 8\left(\frac{4\alpha}{\epsilon} + \frac{1}{q}\Gamma\Big(\frac{1}{q}\Big)\,\kappa^{-1/q}\right) \leq C\left(\frac{\nu^{1/p}}{\epsilon} + \frac{\nu^{1/(p-1)}}{\epsilon^{p/(p-1)}}\right), \tag{291}$$

for a constant $C = C(p)$. Replacing the upper limit ∞ by T only decreases the left-hand side, which proves the claim. □

G.2 Proof of Lemma 13

Proof. Let $h := 1/\gamma$ and $J^* := \lceil y^*/h \rceil$. Then

$$\sum_{j\geq 1} g(jh) = \sum_{j=1}^{J^*-1} g(jh) \;+\; g(J^*h) \;+\; \sum_{j\geq J^*+1} g(jh). \tag{292}$$

Decreasing part. Since g is non-increasing on $[y^*, \infty)$, for every $j \geq J^*$,

$$h\, g((j+1)h) \leq \int_{jh}^{(j+1)h} g(y)\, dy. \tag{293}$$

Summing over $j \geq J^*$ yields

$$\sum_{j\geq J^*} g((j+1)h) \;=\; \sum_{j\geq J^*+1} g(jh) \;\leq\; \frac{1}{h}\int_{J^*h}^{\infty} g(y)\, dy. \tag{294}$$

Increasing part. Since g is non-decreasing on $[0, y^*]$, for $j = 1, \ldots, J^* - 1$,

$$h\, g(jh) \;\leq\; \int_{jh}^{(j+1)h} g(y)\, dy. \tag{295}$$

Summing over $j = 1, \ldots, J^* - 1$ yields

$$\sum_{j=1}^{J^*-1} g(jh) \;\leq\; \frac{1}{h}\int_{h}^{J^*h} g(y)\, dy \;\leq\; \frac{1}{h}\int_{0}^{J^*h} g(y)\, dy. \tag{296}$$

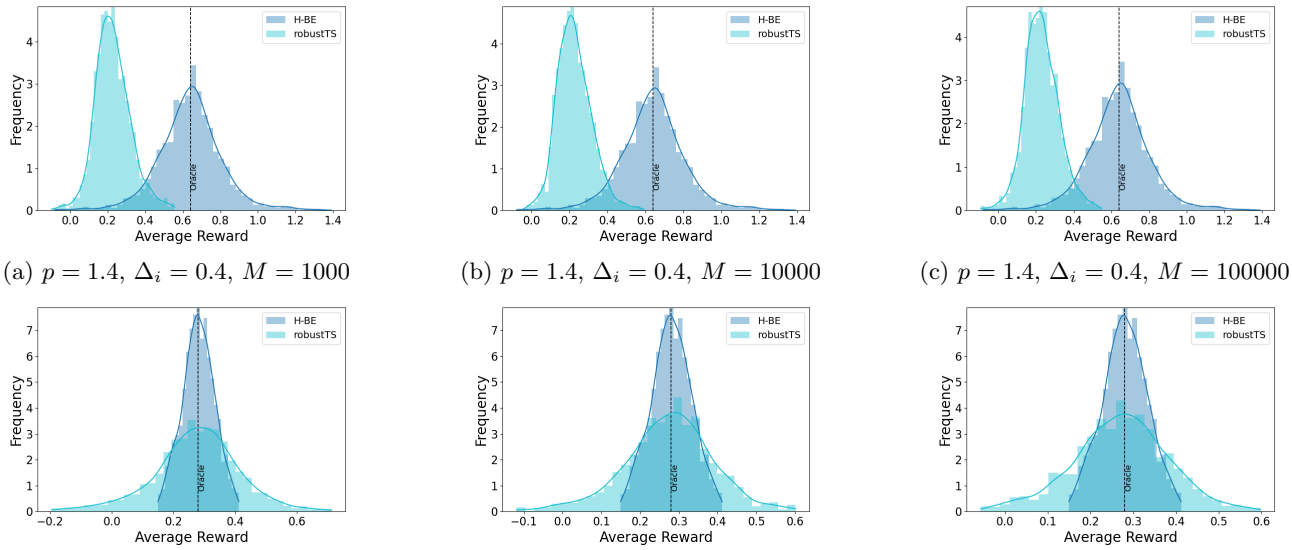
Combining the two parts and using $g(J^*h) \leq \sup_{y\geq 0} g(y)$,

$$\sum_{j\geq 1} g(jh) \;\leq\; \frac{1}{h}\int_{0}^{J^*h} g(y)\, dy \;+\; \frac{1}{h}\int_{J^*h}^{\infty} g(y)\, dy \;+\; \sup_{y\geq 0} g(y) \;\leq\; \frac{1}{h}\int_{0}^{\infty} g(y)\, dy \;+\; \sup_{y\geq 0} g(y). \tag{297}$$

Since $h = 1/\gamma$, the claim follows. □

H Additioanl Experiments

H.1 Effect of Monte Carlo Sample Size on Offline Evaluation



(a) $p = 1.4$, $\Delta_i = 0.4$, $M = 1000$ (b) $p = 1.4$, $\Delta_i = 0.4$, $M = 10000$ (c) $p = 1.4$, $\Delta_i = 0.4$, $M = 100000$

(d) $p = 1.8$, $\Delta_i = 0.8$, $M = 1000$ (e) $p = 1.8$, $\Delta_i = 0.8$, $M = 10000$ (f) $p = 1.8$, $\Delta_i = 0.8$, $M = 100000$

Figure 3: Empirical distributions of the IPW estimator over repeated runs for H-BE and robustTS under α-stable noise setting. For robustTS, the Monte Carlo sample size varies over $M \in \{1000, 10000, 100000\}$. The dashed vertical line indicates the oracle value.

Algorithm	M	**MSE**	**Bias**	**Variance**	**MSE**	**Bias**	**Variance**
		α-*stable*, $p = 1.4$, $\Delta_i = 0.4$			α-*stable*, $p = 1.8$, $\Delta_i = 0.8$		
H-BE	–	**0.003735**	**0.0008**	**0.0037**	**0.130859**	**0.0027**	0.1309
robustTS	1000	0.028073	-0.0062	0.0280	0.196847	-0.4111	0.0279
robustTS	10000	0.018354	-0.0075	0.0183	0.186895	-0.4135	**0.0159**
robustTS	100000	0.019026	-0.0093	0.0189	0.196081	-0.4101	0.0279

Table 2: Effect of the Monte Carlo sample size M on offline evaluation in two representative α-stable settings. We report the MSE, bias, and variance of the IPW estimator.

To further investigate how the Monte Carlo sample size M affects offline evaluation, we visualize the empirical distribution of the IPW estimator over repeated runs for the two settings in which robustTS deviated the most from H-BE in our original experiments. Specifically, the first setting corresponds to the α-stable environment in which the variance gap was the largest, while the second corresponds to the setting in which the bias gap was the most pronounced. Figure 3 compares H-BE and robustTS for $M \in \{1000, 10000, 100000\}$.

In the first setting ($p = 1.4$ and $\Delta_i = 0.4$), the H-BE estimator is concentrated around the oracle, whereas the robustTS estimator is clearly shifted to the left and exhibits substantially larger dispersion. Increasing M from 1000 to 10000 makes the robustTS distribution somewhat tighter, but the gap from H-BE remains substantial. Increasing M further to 100000 yields almost no visible improvement, which is consistent with the variance results reported in Table 2.

In the second ($p = 1.8$ and $\Delta_i = 0.8$), more challenging setting, the contrast is even sharper. The H-BE estimator remains tightly concentrated near the oracle, while the robustTS estimator is much more diffuse and continues to exhibit a noticeable mismatch relative to the oracle even as M increases. Although moving from $M = 1000$ to $M = 10000$ reduces part of the spread, the improvement quickly saturates, and at $M = 100000$ the distribution remains substantially broader than that of H-BE. This is again consistent with the quantitative results in Table 2, where the variance decreases from $M = 1000$ to $M = 10000$ but does not continue to improve at $M = 100000$.

Overall, these histograms show that increasing M can mitigate some Monte Carlo noise in propensity estimation, but only up to a moderate scale. Beyond that point, the remaining error is dominated by instability in the estimated propensities themselves, especially under heavy-tailed rewards. As a result, larger M does not continue to improve offline evaluation reliably and may even slightly worsen it. In contrast, H-BE uses exact closed-form action-selection probabilities, so its IPW estimates remain much more tightly concentrated around the oracle across all settings.

H.2 Real-World Dataset Experiment

We additionally evaluate our method on a real-world dataset constructed from daily stock prices of NASDAQ-listed companies. This setup reflects a realistic sequential decision-making scenario in which an investor selects one stock each day so as to maximize daily return. A similar formulation has also been considered in prior work on heavy-tailed bandits Chowdhury and Gopalan (2019).

We use daily data from January 2010 to December 2021, which yields roughly 2000 trading days. The entire period is treated as the interaction horizon of the bandit problem. The dataset contains the rates of change in open, high, low, and close prices, as well as trading volume, from one day to the next. For each stock i, we define the reward at day t as the relative close-price return

$$r_{t,i} := \frac{C_{t+1,i} - C_{t,i}}{C_{t,i}},$$

where $C_{t,i}$ denotes the closing price of stock i on day t. At each trading day t, the learner observes the historical data and selects one stock among ten candidates. If the learner selects stock i at day t, the instantaneous regret is defined as

$$\max_{j \in [K]} r_{t,j} - r_{t,i},$$

which measures the gap between the best achievable next-day return among the ten stocks and the return obtained by the selected stock.

We compare H-BE against several established heavy-tailed baselines, namely robustTS Dubey and Pentland (2019), MR-UCB Lee and Lim (2022), and APE Lee et al. (2020). All algorithms are run on the same sequence of stock returns over the full horizon, using 10 random seeds. Table 3 reports the cumulative regret and the corresponding standard deviation. H-BE achieves the smallest mean regret among the compared methods, while remaining competitive in variability. This suggests that the proposed method is not only theoretically well justified but also practically effective on a real-world heavy-tailed dataset.

Algorithm	**Mean**	**Std**
robustTS	58.9560	0.5765
MR-UCB	58.1860	0.0000
APE	57.9855	0.7151
H-BE	**57.8702**	0.4585

Table 3: Cumulative regret on the NASDAQ stock dataset over 10 random seeds. We report the mean and standard deviation of cumulative regret.

For offline evaluation, we consider a setting in which the target policy is the uniform policy over the ten stocks, while each bandit algorithm serves as the logging policy. Given a logged trajectory $\{(I_t, r_{t,I_t})\}_{t=1}^{T}$ generated by a logging policy, we compute the inverse propensity weighting (IPW) estimator as

$$\hat{V}_{\mathrm{IPW}} = \frac{1}{T} \sum_{t=1}^{T} \frac{\pi_{\mathrm{tar}}(I_t)}{\pi_{\mathrm{log}}(I_t)} r_{t,I_t},$$

where π_{tar} and π_{log} denote the action-selection probabilities of the target and logging policies, respectively.

As in the synthetic offline-evaluation experiment, the logging propensities of robustTS are approximated by Monte Carlo simulation because they are not available in closed form. To obtain an oracle benchmark, we compute the true value of the uniform policy directly from the full sequence of observed returns.

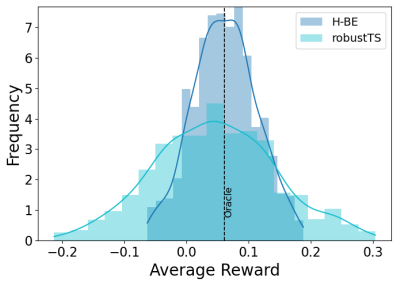


Figure 4: Empirical distributions of the IPW estimator on the NASDAQ stock dataset. The dashed vertical line indicates the oracle value.

We repeat the offline evaluation over 1000 random seeds and report the MSE, bias, and variance in Table 4. The results show that H-BE yields substantially lower MSE and variance than robustTS, while also having negligible bias. This is consistent with our synthetic experiments: because H-BE provides exact closed-form action-selection probabilities, it avoids the additional approximation noise introduced by Monte Carlo propensity estimation and thereby yields more reliable offline evaluation.

Algorithm	MSE	Bias	Variance
robustTS	0.011419	-0.0154	0.0112
H-BE	**0.002935**	**0.0001**	**0.0029**

Table 4: Offline evaluation on the NASDAQ stock dataset. We report the MSE, bias, and variance of the IPW estimator over 1000 random seeds.

H.3 Performance in a Light-Tailed Stochastic Regime

Our theoretical analysis is derived under the finite p-th moment assumption (Assumption 1), which includes the sub-Gaussian case when $p = 2$. In this regime, the gap-independent regret bound in Theorem 4 reduces to the standard $\sqrt{KT}$ scaling, matching the minimax order of UCB up to logarithmic factors.

To complement this observation empirically, we additionally evaluate H-BE under Gaussian noise with mean 0 and variance 1. We consider two settings with suboptimality gaps $\Delta_i \in \{0.4, 0.8\}$. All other experimental settings are the same as in the main paper: 10 arms, horizon $T = 20000$, and 10 random seeds. Table 5 reports the cumulative regret.

Gap	Algorithm	Mean $\pm$ Std
0.8	UCB	0.0095 ± 0.0013
0.8	H-BE	$\mathbf{0.0014 \pm 0.0006}$
0.4	UCB	0.0182 ± 0.0028
0.4	H-BE	$\mathbf{0.0077 \pm 0.0023}$

Table 5: Cumulative regret under Gaussian noise over 10 random seeds.

In both cases, H-BE performs slightly better than the classical UCB baseline. This is consistent with the theoretical prediction that H-BE does not incur any order-level regret penalty in light-tailed environments. Taken together, these results indicate that when the noise regime is unknown a priori, H-BE remains competitive in sub-Gaussian settings while simultaneously retaining robustness to heavier tails.

H.4 Additional Synthetic Baseline: BGE-I

We also include an additional synthetic-data baseline based on Boltzmann-Gumbel exploration. In particular, we consider the heavy-tailed variant of the Boltzmann–Gumbel exploration policy described in Cesa-Bianchi et al.

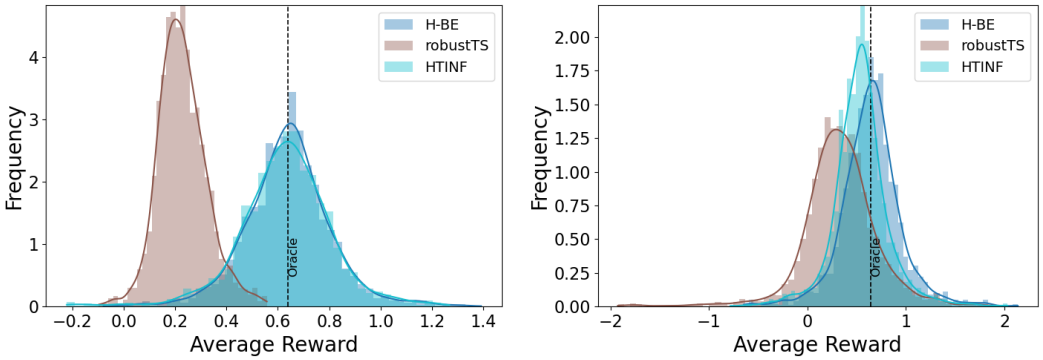
Algorithm	Mean	Std	Mean	Std
	Pareto, $p = 1.4$, $\Delta_i = 0.4$		α-*stable*, $p = 1.8$, $\Delta_i = 0.8$	
APE	0.1201	0.0605	0.0928	0.0583
robustTS	0.2345	0.1747	0.0224	0.0218
MR-UCB	0.0440	0.0100	0.0473	0.0135
BGE-I	0.1105	0.1298	0.0923	0.1028
HTINF	0.2894	0.0418	0.2865	0.0150
H-BE	**0.0401**	0.0334	**0.0204**	0.0159

Table 6: Cumulative regret on two challenging synthetic instances over 10 random seeds.

(2017), where the empirical mean is replaced by an influence-function–based robust estimator. We refer to this method as BGE-I. We report cumulative regret in Table 6.

Overall, BGE-I consistently underperforms H-BE. From a theoretical perspective, the gap-dependent regret bound for BGE-I follows the form given in Theorem 5 of Cesa-Bianchi et al. (2017), where the variance parameter enters through an exponential factor. Although this term is asymptotically dominated by the leading contribution, it can still be non-negligible unless the horizon is very large, especially when the tail is sufficiently heavy. Consequently, in moderate-horizon regimes such as those considered here, this dependence can deteriorate empirical performance in a way that is not visible from the asymptotic expression alone. Moreover, the available analysis of BGE-I applies to the finite-variance setting, whereas our theory covers the more general finite p-th moment regime with $p \in (1, 2]$. This mismatch also contributes to its weaker empirical performance in our heavier-tailed synthetic environments.

H.5 Additional Offline-Evaluation Baseline: HTINF



(a) $p = 1.4$, $\Delta_i = 0.4$, $M = 1000$ (b) $p = 1.4$, $\Delta_i = 0.4$, $M = 10000$

We further include HTINF (Huang et al., 2022) as an additional baseline for offline evaluation. Since a public implementation of HTINF is not available, we followed the procedure described in the original paper and solved the corresponding optimization equations using Newton's method with a fixed error tolerance. We use the same experimental setup as in our original synthetic experiments.

Because HTINF is formulated in terms of losses, we convert rewards to losses by the transformation specified in the original method. However, the theoretical guarantees of HTINF rely on Assumption 3.6 of (Huang et al., 2022), which requires that the truncated loss of some optimal arm remain nonnegative almost surely. This condition is violated in our synthetic construction for sufficiently large realizations of the heavy-tailed noise. Thus, HTINF is evaluated here in a setting where its own theoretical assumptions do not apply. By contrast, H-BE remains theoretically valid under the finite p-th moment assumptions used throughout our experiments.

Table 7 reports the offline-evaluation results for the two representative synthetic settings considered above. Across

Algorithm	**MSE**	**Bias**	**Variance**	**MSE**	**Bias**	**Variance**
	Pareto, $p = 1.4$, $\Delta_i = 0.4$			α-*stable*, $p = 1.8$, $\Delta_i = 0.8$		
H-BE	**0.713018**	**0.0343**	**0.7118**	**0.130859**	**0.0027**	**0.1309**
HTINF	0.809679	-0.0868	0.8021	0.182187	-0.0236	0.1816
robustTS	0.926443	-0.3266	0.8198	0.196847	-0.4111	0.0279

Table 7: Offline evaluation on two representative synthetic settings. We report the MSE, bias, and variance of the IPW estimator.

both environments, H-BE and HTINF outperform robustTS, which relies on Monte Carlo propensity estimation. Moreover, H-BE achieves the lowest MSE in every configuration.