Tight Regret Bounds in Multi-Armed Bandits with Heterogeneous Variances

Anonymous authors
Paper under double-blind review

Abstract

We study stochastic multi-armed bandits with heterogeneous reward variances. In the known-variance setting, we propose a variance-aware MOSS algorithm that achieves minimax-optimal regret matching an information-theoretic lower bound up to constants. For the unknown-variance case, we construct high-probability variance upper confidence bounds and show that the resulting algorithm attains the same minimax rate up to a logarithmic factor. Our analysis establishes sharp worst-case guarantees that explicitly capture the variance structure of the problem.

1 Introduction

Stochastic multi-armed bandit (MAB) regret minimization is an online learning and sequential decision-making problem in which a learner interacts with an environment over a sequence of rounds, choosing one out of K potential arms and receiving a stochastic reward sampled from an unknown distribution every round. The goal is to minimize cumulative regret—that is, the difference in performance compared to always pulling the optimal arm.

In most real-world decision-making systems, actions (or "arms") often yield rewards with inherently different levels of uncertainty. A motivating example can be found in online advertising on streaming platforms or TV. Different ads may exhibit highly variable click-through rates due to demographic differences, seasonal trends, or delivery platforms. Some ads return consistent engagement, while others may show erratic performance. A variance-aware algorithm can quickly discount noisy or unreliable arms and focus exploration on more stable options—leading to faster and more efficient ad selection.

When the reward variances of arms are known in advance, they can be directly incorporated into confidence-based exploration strategies, allowing the learner to reduce uncertainty more efficiently. High-variance arms can be pulled more frequently to obtain more reliable estimates, while low-variance arms can be sufficiently tested by fewer pulls. In practice, however, reward variances are unknown and must be estimated online. This significantly increases the complexity of the learning problem, as the learner must explore not only to estimate the means of the arms, but also their variances. This added uncertainty leads to additional regret and makes the design of effective algorithms more challenging.

In this work, we propose two algorithms for regret minimization in bandits with heterogeneous reward variances. Our first algorithm, VarUCB-Known, assumes the reward variances are known and incorporates them into the confidence bounds used to guide exploration. Our second algorithm, VarUCB-Unknown, estimates variances via high-probability confidence bounds and adjusts exploration accordingly. While the known-variance setting yields tighter confidence intervals and faster convergence, the unknown-variance setting introduces ambiguity and requires more conservative balancing between exploration and estimation. In either case, our goal is to develop algorithms that achieve minimax regret and adapt to the statistical complexity of the environment.

Related Works Best-arm identification (BAI) in stochastic multi-armed bandits has been extensively explored under both the fixed-budget (Bubeck et al., 2010; Audibert & Bubeck, 2010) and fixed-confidence (Even-Dar et al., 2006) frameworks. Within the fixed-budget setting, Gabillon et al. (2011) introduced the

GapE-V algorithm, a variance-adaptive strategy that achieves exponentially decreasing error rates with increasing budget. However, GapE-V requires knowledge of a complexity parameter that depends on unknown problem-specific quantities such as reward variances and gaps. To mitigate this limitation, an adaptive variant, A-GapE-V, was also proposed, which estimates the parameter on the fly but lacks theoretical guarantees. Faella et al. (2020) later extended the successive rejects algorithm (Audibert & Bubeck, 2010) into a variance-adaptive version and demonstrated comparable error bounds to those achieved by sequential halving (Karnin et al., 2013). While the majority of subsequent research on BAI with heterogeneous variances has focused on the fixed-confidence setting (Lu et al., 2021; Zhou & Tian, 2022; Jourdan et al., 2023), the fixed-budget case remains relatively underexplored.

More recently, Lalitha et al. (2023) proposed SHVar and SHAdaVar, two algorithms tailored for fixed-budget BAI when reward variances are known and unknown, respectively. The core idea behind these methods is to allocate more sampling budget to arms with higher variances. Notably, SHAdaVar adopts a greedy approach that overestimates unknown variances to guide allocation. Their analysis introduces new lower bounds on sample complexity that do not require solving the allocation problem in closed form, and experimental results show superior performance across diverse settings.

To build on this foundation, several new works have recently advanced the study of variance-aware bandits. Saha & Kveton (2023) developed a Bayesian Thompson Sampling approach for Gaussian bandits with unknown heterogeneous variances, demonstrating prior-dependent Bayes regret bounds that improved as variance uncertainty decreased. In contrast, their bounds depend on Bayesian priors and hold only in an instance-dependent sense, whereas ours are prior-free and valid in the worst case. Kato (2024) presented a minimax optimal algorithm for fixed-budget BAI under the small-gap regime, showing tight matching bounds between upper and lower errors. Their metric is terminal mis-identification probability, whereas we track cumulative regret throughout the horizon. Kato (2025) proposed a Generalized Neyman Allocation strategy that achieved asymptotic worst-case optimality in identifying the best arm when variances are known. The result assumes known variances and optimises sample allocation, while our work quantifies the extra regret cost when variances are unknown. Lastly, Zhang et al. (2025) introduced SHRVar, a multi-metric fixed-budget BAI framework that generalizes sequential halving using a novel variance-aware sampling and elimination strategy, achieving exponential error decay and enhanced empirical performance. In contrast, we consider a single-metric setting and deliver tight upper-lower bounds on regret rather than on final error probability.

Our Contribution. We study stochastic bandits with heterogeneous reward variances in both known- and unknown-variance regimes.

For the known-variance setting, we design a variance-aware MOSS-style algorithm and prove a minimax-optimal regret bound of order

$$\Theta\left(\sqrt{T\sum_{i=1}^{K}\sigma_i^2}\right),$$

matching a corresponding information-theoretic lower bound. Our analysis provides both gap-dependent and gap-independent guarantees that explicitly capture the variance structure.

For the unknown-variance case, we construct high-probability upper confidence bounds on empirical variances and show that the resulting algorithm achieves the same minimax rate up to a logarithmic factor.

Finally, we establish matching lower bounds via KL-divergence arguments, demonstrating the optimality of the known-variance algorithm and clarifying the fundamental cost of variance estimation.

2 Preliminary

We consider the standard stochastic MAB problem with K arms labeled $a_1, a_2, ...a_K$. Each arm a_i is associated with an unknown reward distribution which is σ_i^2-sub-Gaussian, with true mean μ_i and variance σ_i^2 (which is known in the known-variance setting). We denote the horizon as T. At each round $t = 1, 2, \ldots, T$, the learner selects one arm and receives a reward r_t drawn from its distribution. Let $n_i(t)$ represent the number of times that arm a_i has been pulled up to time t, and let $\widehat{\mu}_i(t)$ denote the empirical mean of arm a_i after $n_i(t)$ pulls.

The goal of the learner is to minimize cumulative regret, which measures the difference between the expected reward from always pulling the best arm and the expected reward obtained by the learner. Let $\mu^\star$ be the mean of the optimal arm, i.e. $\mu^\star = \max_i \mu_i$. Then, the expected regret after T rounds is defined as:

$$\mathbb{E}[R_T] = \sum_{t=1}^{T} (\mu^* - \mathbb{E}[X_{i_t}])$$

where i_t is the arm pulled at round t. This can be rewritten by decomposing the regret as follows:

$$\mathbb{E}[R_T] = \sum_{i=1}^{K} \Delta_i \mathbb{E}[n_i(T)] \tag{1}$$

where $\Delta_i = \mu^\star - \mu_i$ is the suboptimality gap between the optimal arm and arm i, and $n_i(T)$ is the total number of times arm i was pulled over the entire horizon, T. We focus on the known-variance setting in our initial analysis, where the variances $\sigma_1^2, ..., \sigma_K^2$ are known in advance. Later, we extend our analysis to the more realistic unknown-variance setting.

In Lai & Robbins (1985) where the rewards are assumed to be 1-sub-Gaussian, they propose the UCB (upper confidence bound) index, defined as follows:

$$\mathrm{UCB}_i(t) = \begin{cases} \infty & \text{if } n_i(t) = 0 \\ \hat{\mu}_i(t) + \sqrt{\frac{4 \log T}{n_i(t)}} & \text{otherwise} \end{cases}$$

along with the UCB algorithm, which pulls the arm with the highest UCB index at every round (breaking ties arbitrarily).

The core idea of this is *optimism in the face of uncertainty*. It can be proved through standard Hoeffding-style concentration inequalities that the UCB index is a high-probability upper confidence bound on the true mean. This type of algorithm has been applied to various other bandit settings, such as latent Maillard & Mannor (2014), contextual Chu et al. (2011), cascading Kveton et al. (2015), combinatorial Cesa-Bianchi & Lugosi (2012), Lipschitz Magureanu et al. (2014) and corrupted Lykouris et al. (2018) bandits. In all of these works, the rewards are assumed to be bounded or 1-sub-Gaussian, effectively normalizing variances to ≤ 1.

3 Main Results

3.1 Minimax Regret with Known Variances

Recall that our goal is to minimize the cumulative regret

$$\mathcal{R}_T = \sum_{i=1}^{K} \Delta_i \mathbb{E}[n_i(T)], \qquad \Delta_i := \mu^\star - \mu_i,$$

where $i^\star = \arg\max_i \mu_i$.

We now consider a *variance-aware minimax optimal* UCB algorithm. Assume that rewards from arm i are σ_i-sub-Gaussian and that σ_i^2 is known.

Unlike classical UCB, which uses a $\log T$ exploration term, the minimax MOSS strategy replaces $\log T$ with $\log_+\left(\frac{T}{K n_i(t)}\right)$, which shrinks automatically once an arm has been sampled sufficiently often. This modification removes the extra $\sqrt{\log T}$ factor in worst-case regret.

We define the index

$$\mathrm{UCB}_i(t) = \begin{cases} \infty, & \text{if } n_i(t) = 0, \\ \hat{\mu}_i(t) + \sigma_i \sqrt{\frac{4}{n_i(t)} \log_+\left(\frac{T}{K\, n_i(t)}\right)}, & \text{otherwise.} \end{cases} \tag{2}$$

At each round the algorithm selects

$$i_t = \arg\max_{i \in [K]} \mathrm{UCB}_i(t)$$

and updates the empirical mean and pull counts.

Algorithm 1 Variance-Aware MOSS (known σ_i^2)

Input: horizon T, arms K, known variances $\{\sigma_i^2\}$
Pull each arm once
for $t = K+1, \ldots, T$ **do**
  for each $i \in [K]$ **do**
    Compute

$$\mathrm{UCB}_i(t) = \hat{\mu}_i(t) + \sigma_i \sqrt{\frac{4}{n_i(t)} \log_+\left(\frac{T}{K\, n_i(t)}\right)}$$

  end for
  Select $i_t \leftarrow \arg\max_i \mathrm{UCB}_i(t)$
  Observe reward and update $n_{i_t}(t+1)$, $\hat{\mu}_{i_t}(t+1)$
end for

Theorem 1 (Minimax regret with known variances) *The cumulative regret of Algorithm 1 satisfies*

$$\mathcal{R}_T \;\le\; C\sqrt{T \sum_{i=1}^{K} \sigma_i^2} + \sum_{i=1}^{K} \Delta_i,$$

for a universal constant C. In particular,

$$\mathcal{R}_T = O\left(\sqrt{T \sum_{i=1}^{K} \sigma_i^2}\right).$$

Observe that when $\sigma_i \equiv 1$ for all i, the bound reduces to the classical minimax rate $O(\sqrt{KT})$. Compared to the previous $\sqrt{T \log T \sum_i \sigma_i^2}$ bound Audibert et al. (2009), this removes the extra $\sqrt{\log T}$ factor and is therefore minimax optimal.

Proof sketch. The proof follows the minimax MOSS strategy rather than the classical "good event" stopping-time argument used in gap-dependent analyses. The key idea is to control how much the optimal arm can fail to be optimistic, and then use this control to bound the number of pulls of suboptimal arms.

We begin by defining the *optimism deficit* of the optimal arm,

$$\Delta = \left(\mu^\star - \min_{1 \le s \le T} \left[\hat{\mu}_{i^\star,s} + \sigma_{i^\star}\sqrt{\frac{4}{s}\log_+\left(\frac{T}{Ks}\right)}\right]\right)^+.$$

This random variable measures how far below its true mean the optimal arm's index can fall at any time. Using sub-Gaussian concentration together with a dyadic peeling argument, one shows that the tail of Δ decays at rate $O(\sigma_{i^\star}^2 K/(Tx^2))$. Integrating this tail bound yields

$$\mathbb{E}[\Delta] \le C_1 \sigma_{i^\star} \sqrt{\frac{K}{T}},$$

so the expected optimism deficit is of order $\sqrt{K/T}$.

We then decompose the regret according to whether the gap of an arm is small or large relative to Δ. For arms with $\Delta_i \le 2\Delta$, a crude bound $n_i(T) \le T$ suffices. Their total contribution is therefore at most $2T\Delta$, whose expectation is of order $\sigma_{i^\star}\sqrt{KT}$.

For arms with $\Delta_i > 2\Delta$, selection of such an arm requires its index to exceed $\mu_i + \Delta_i/2$. Sub-Gaussian concentration implies that this event can occur only about $O(\sigma_i^2/\Delta_i^2)$ times. Because the MOSS exploration term $\log_+(T/(Kn))$ shrinks as n grows, this yields a per-arm regret contribution of order $O(\sqrt{T\sigma_i^2/K})$. Summing over all large-gap arms and combining the two regimes gives

$$\mathcal{R}_T \leq C\sqrt{T\sum_{i=1}^{K}\sigma_i^2} + \sum_i \Delta_i.$$

The crucial feature of the MOSS exploration term is that it automatically decreases with the number of pulls, preventing the extra logarithmic factors that arise in standard UCB analyses and yielding a minimax-optimal rate.

3.2 Minimax Regret when the Variances are Unknown

We now consider the setting in which the reward variances are unknown and must be estimated online. Assume rewards from arm i are Gaussian with unknown mean μ_i and unknown variance σ_i^2.

In contrast to the known-variance case, the exploration term must account for uncertainty in both the mean and the variance. We therefore construct an upper confidence bound on the variance using a χ^2 tail bound.

Variance confidence bound. Fix arm i and let $n_i(t)$ denote the number of pulls up to time t. Define the unbiased sample variance

$$\hat{\sigma}_i^2(t) = \frac{\sum_{s\leq t:a_s=i}(r_s - \hat{\mu}_i(t))^2}{n_i(t)-1}.$$

If rewards are Gaussian, then

$$\frac{(n_i(t)-1)\hat{\sigma}_i^2(t)}{\sigma_i^2} \sim \chi^2_{n_i(t)-1}.$$

Applying a χ^2 tail bound (e.g., Lalitha et al. (2023)), for $n_i(t) \geq 4\log(1/\delta)+1$, with probability at least $1-\delta$,

$$\sigma_i^2 \leq \frac{\hat{\sigma}_i^2(t)}{1-2\sqrt{\frac{\log(1/\delta)}{n_i(t)-1}}}.$$

Setting $\delta = T^{-2}$ gives a uniform high-probability bound over all arms and times.

Variance-aware minimax index. Motivated by the known-variance MOSS algorithm, we replace σ_i in the minimax index with its upper confidence bound estimate.

We define

$$\mathrm{UCB}_i(t) = \begin{cases} \infty, & n_i(t) = 0, \\ \hat{\mu}_i(t) + \hat{\sigma}_i(t)\sqrt{\frac{4}{n_i(t)\left(1-2\sqrt{\frac{2\log T}{n_i(t)}}\right)}\log_+\left(\frac{T}{K\,n_i(t)}\right)}, & \text{otherwise.} \end{cases} \tag{3}$$

The additional denominator term $1-2\sqrt{\frac{2\log T}{n_i(t)}}$ inflates the exploration bonus to account for uncertainty in variance estimation.

Algorithm 2 Variance-MOSS-Unknown (Gaussian rewards, unknown σ_i^2)

Require: horizon T, number of arms K

Pull each arm twice (to define $\hat{\sigma}_i^2$)
for $t = 2K+1, \ldots, T$ **do**
  for each $i \in [K]$ **do**
    Let $n_i(t)$ be pulls of arm i so far, and compute empirical mean $\hat{\mu}_i(t)$
    Compute unbiased empirical variance

$$\hat{\sigma}_i^2(t) = \frac{1}{n_i(t)-1} \sum_{s \le t:\, a_s = i} \big(r_s - \hat{\mu}_i(t)\big)^2$$

    Define the variance upper confidence bound

$$\hat{\sigma}_i^{2,\mathrm{UCB}}(t) := \frac{\hat{\sigma}_i^2(t)}{1 - 2\sqrt{\frac{2\log T}{n_i(t)-1}}} \quad \text{(when } n_i(t) \ge 8\log T + 1\text{; otherwise set it to } +\infty\text{)}$$

    Set

$$\mathrm{UCB}_i(t) = \hat{\mu}_i(t) + \sqrt{\hat{\sigma}_i^{2,\mathrm{UCB}}(t)}\sqrt{\frac{4}{n_i(t)} \log_+\!\left(\frac{T}{K\, n_i(t)}\right)}$$

  end for
  Play $i_t \in \arg\max_{i \in [K]} \mathrm{UCB}_i(t)$ and observe r_t
end for

Theorem 2 (Minimax regret with unknown variances) *Suppose rewards are Gaussian with unknown heterogeneous variances. Then the cumulative regret of Algorithm 2 satisfies*

$$\mathcal{R}_T \le C \sqrt{T \sum_{i=1}^{K} \sigma_i^2 \, \log T}$$

for a universal constant C.

Compared to the known-variance case, an additional $\log T$ factor appears. This reflects the extra uncertainty introduced by estimating variances online.

When $\sigma_i \equiv 1$, the regret becomes

$$O(\sqrt{KT} \log T),$$

which is larger than classical minimax MOSS by a factor $\log T$.

This gap is unavoidable in general, since variance estimation requires additional exploration. A matching lower bound is given in the next section.

Proof sketch. The proof follows the minimax MOSS analysis used in the known-variance case, with one additional ingredient to control the unknown variances. The main difficulty is that the exploration bonus now depends on an empirical estimate of σ_i^2, so we must first ensure that this estimate is uniformly reliable over all arms and all times.

We begin by constructing a high-probability upper confidence bound for each variance using a χ^2 tail inequality. By choosing the failure probability to be of order T^{-2} and taking a union bound over all arms and all rounds, we obtain an event on which the empirical variance-based plug-in $\hat{\sigma}_i^{2,\mathrm{UCB}}(t)$ dominates the true variance σ_i^2 simultaneously for all i and $t \le T$. On this event, the exploration bonus used by the algorithm is at least as large as the bonus in the known-variance MOSS algorithm, so the index of the optimal arm remains sufficiently optimistic.

As in the known-variance analysis, we define the *optimism deficit*

$$\Delta = \left(\mu^\star - \min_{1\le s\le T}[\hat{\mu}_{i^\star,s} + \text{bonus}_{i^\star}(s)]\right)^+.$$

Using the same dyadic peeling argument as in the MOSS proof, one shows that the tail of Δ decays at rate $O(\sigma_{i^\star}^2 K/(Tx^2))$, which implies

$$\mathbb{E}[\Delta] = O\left(\sigma_{i^\star}\sqrt{\frac{K}{T}}\right).$$

Thus, the expected deficit of the optimal arm has the same order as in the known-variance case.

We then decompose the regret into contributions from arms whose gaps are small relative to Δ and those whose gaps are larger. For arms with $\Delta_i \le 2\Delta$, a crude bound $T_i(T) \le T$ suffices, and their total contribution is controlled by $T\mathbb{E}[\Delta]$, which yields a term of order $\sqrt{KT}$.

For arms with $\Delta_i > 2\Delta$, selection of such an arm requires its index to exceed $\mu_i + \Delta_i/2$. This leads to a counting argument bounding the number of such "optimistic" events using sub-Gaussian concentration. The only difference from the known-variance case is that the variance plug-in term inflates the effective confidence radius. Because this variance control must hold uniformly over arms and time, an additional logarithmic factor appears in the bound on the expected number of pulls of suboptimal arms.

Summing the contributions of all arms and combining the two regimes yields

$$R_T \le C\sqrt{T\sum_{i=1}^{K}\sigma_i^2}\,\log T + \sum_{i=1}^{K}\Delta_i.$$

The overall structure of the proof is therefore identical to the known-variance minimax argument; the only source of the extra $\log T$ factor is the uniform high-probability control of the unknown variances.

4 Lower Bounds

In this section, we construct a lower bound in a known heterogeneous variance setting that matches our upper bound up to some factor of log. We achieve a lower bound similar to that of Chapter 15 in Lattimore & Szepesvári (2020). More specifically, we construct two bandit environments such that any policy is guaranteed to do badly on at least one of them. Suppose our environment has K arms $a_1, ..., a_K$ with variances given by σ_i^2. We then have the following result.

Theorem 3 *There exists an environment in which the regret is lower bounded by*

$$R_T = \Omega\left(\sqrt{T\sum_j \sigma_j^2}\right) \tag{4}$$

We defer the details of the proof to the appendix, but present the key ideas here. To formalize this, we start with the following key lemmas:

Lemma 4 (Bretagnolle-Huber inequality) *Let P and Q be probability measures on the same measurable space $(\Omega, \mathcal{F})$, and let $A \in \mathcal{F}$ be an arbitrary event. Then,*

$$P(A) + Q(A^c) \ge \frac{1}{2}\exp(-\mathrm{D}(P,Q))$$

where $A^c = \Omega\backslash A$ is the complement of A.

This lemma allows us to lower bound the probability of A and its complement A^c across different distributions. A is chosen to be a "bad" event, where on this event the same algorithm π would achieve poor performance in that environment. Setting the distributions as in the bandit environment, we have the following:

Lemma 5 (Divergence decomposition) *Let $\nu = (P_1, \ldots, P_k)$ be the reward distributions associated with one k-armed bandit, and let $\nu' = (P'_1, \ldots, P'_k)$ be the reward distributions associated with another k-armed bandit. Fix some policy π and let $\mathbb{P}_\nu = \mathbb{P}_{\nu\pi}$ and $\mathbb{P}_{\nu'} = \mathbb{P}_{\nu'\pi}$ be the probability measures on the canonical bandit model (Section 4.6) induced by the n-round interconnection of π and ν (respectively, π and ν'). Then,*

$$\mathrm{D}\left(\mathbb{P}_\nu, \mathbb{P}_{\nu'}\right) = \sum_{i=1}^{k} \mathbb{E}_\nu\left[T_i(n)\right] \mathrm{D}\left(P_i, P'_i\right)$$

Given that a bandit environment is complex, this allows us to compute the divergence between two bandit environments by computing the pairwise divergence between the rewards of corresponding arms.

Now it remains to construct the two environments. We let the first environment ν be a Gaussian bandit with the following means:

$$\mu = (\Delta, 0, 0, \ldots, 0). \tag{5}$$

Let arm i be an arm that is pulled at most $\frac{\sigma_i^2 T}{\sum_j \sigma_j^2}$ times under policy π. Such an arm must exist; otherwise, the total number of pulls would exceed the horizon T.

We now define a second environment ν', which differs from ν only at arm i. The new mean vector is given by

$$\mu' = (\Delta, 0, 0, \ldots, 0, 2\Delta, 0, \ldots, 0), \tag{6}$$

where the mean of arm i is set to 2Δ (the rest of the arms remain unchanged).

We now calculate the regrets under both environments. For the first environment ν, we have

$$R_T(\pi, \nu) \geq \mathbb{P}_\mu(n_1(T) \leq T/2) \cdot \frac{T\Delta}{2},$$

and for the second environment ν', we have

$$R_T(\pi, \nu') > \mathbb{P}_{\mu'}(n_1(T) > T/2) \cdot \frac{T\Delta}{2}.$$

The construction of μ' is novel in that we select an arm that is pulled relatively few times under ν—accounting for the variance structure—and penalize the algorithm by increasing the mean of that same arm in ν'. This ensures that at least one of the two environments induces high regret.

We now optimize over Δ to maximize the lower bound. Substituting

$$\Delta = \sqrt{\frac{\sum_j \sigma_j^2}{4T}},$$

into the combined regret expression yields

$$R_T(\pi, \nu) + R_T(\pi, \nu') \geq C\sqrt{T \sum_j \sigma_j^2},$$

for some constant $C > 0$. Therefore, there exists an environment ν (or ν') such that

$$R_T \geq c\sqrt{T \sum_j \sigma_j^2},$$

for some constant $c > 0$, establishing a minimax lower bound in the known-variance setting.

5 Experiments

We evaluate VarUCB-Known, VarUCB-Unknown, and the standard UCB algorithm (variance-agnostic) over a horizon of $T = 1{,}000{,}000$ rounds. We consider $K \in \{10, 50\}$ arms and two variance regimes: low ($[1, 5]$) and high ($[20, 50]$). For each setting, we run 50 independent trials. In every trial, the arm reward means μ_i are drawn uniformly from $[0, 1]$ and the arm reward variances σ_i^2 are drawn uniformly from the specified variance range.

For each algorithm, we compute the mean cumulative regret across trials and plot it along with a shaded region representing ± 1 standard deviation. Figure 1 shows results for the low variance regimes for $K = 10$ and $K = 50$ arms while Figure 2 shows results for high variance arms.

As expected, VarUCB-Unknown consistently incurs higher regret than VarUCB-Known. This is likely because VarUCB-Unknown must allocate additional rounds to estimating each arm's variance, which increases uncertainty and delays convergence to the optimal arm.

Low-variance setting (Figure 1): The standard UCB algorithm achieves lower cumulative regret than both VarUCB variants. This is consistent with our theoretical intuition: when variances are small, the "good event" (where the true mean is below the UCB index) occurs more frequently, allowing UCB to identify and commit to the optimal arm more quickly. In contrast, the VarUCB algorithms devote extra exploration to estimating arm variances; in this regime, the variance differences are too small to provide an advantage, so this extra exploration results in slightly higher regret.

High-variance setting (Figure 2): Here, the standard UCB algorithm exhibits roughly linear cumulative regret, while the VarUCB variants achieve logarithmic-like growth. UCB is also noticeably less stable, with a rapidly widening ± 1 standard deviation band, whereas the VarUCB variants maintain much narrower uncertainty bands. This performance gap likely occurs because standard UCB depends on the good event to achieve sub-linear regret; with high variance, the empirical mean estimates are noisier, increasing the chance that the vanilla UCB index poorly estimates the true mean of the optimal arm, thereby preventing the good event from occurring. In essence the variance-aware algorithms ensure that the good event occurs with higher probability, at the expense of incurring higher regret on the good events. When the bad event occurs, the algorithm is likely to achieve linear regret.

In the $K = 10$ setting, the mean regret of standard UCB across rounds eventually surpasses that of both VarUCB variants. We hypothesize that the same crossover would occur in the $K = 50$ case given a sufficiently large horizon T. However, even with $T = 1{,}000{,}000$, standard UCB's regret remains below that of the VarUCB variants for $K = 50$. This is likely because, with more arms but the same fixed reward range $[0, 1]$, under the bad event G^c where the wrong arm is committed to, there are more arms whose means are close to the optimal arm, reducing the regret gap for standard UCB in the early stages.

6 Conclusion

We have studied stochastic multi-armed bandit problems in which the reward variances differ across arms, a setting motivated by numerous real-world applications where actions yield outcomes with varying uncertainty. Our work introduces and analyzes two variance-aware UCB algorithms: VarUCB-Known, which assumes prior knowledge of variances, and VarUCB-Unknown, which estimates them online using empirical confidence bounds.

For both settings, we derived tight regret bounds—both gap-dependent and gap-independent—and showed that our algorithms adapt to the statistical difficulty of each arm. In particular, we established that the unknown-variance setting incurs only a logarithmic overhead in regret compared to the known-variance case. We also derived a matching lower bound in the known-variance regime, demonstrating the near-optimality of our upper bounds.

Our empirical evaluations confirmed the theoretical predictions: in high-variance regimes, the proposed algorithms outperform classical UCB by better managing exploration under uncertainty; in low-variance

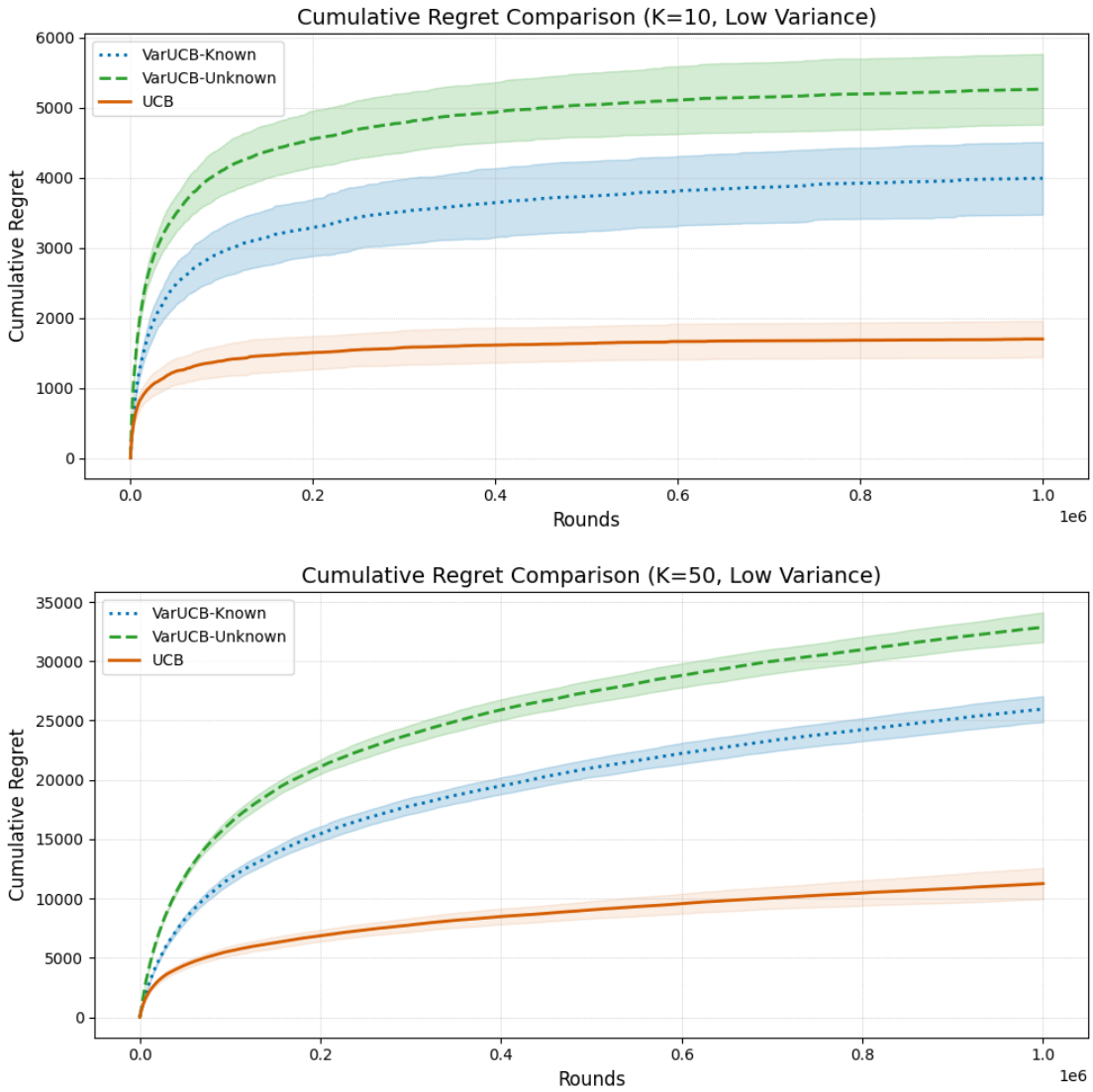

Figure 1: These plots show the regret in the low-variance regime for the standard UCB algorithm, as well as for VarUCB-Known and VarUCB-Unknown. The top figure corresponds to $K = 10$ arms, while the bottom figure shows results for $K = 50$ arms. In this regime, the standard UCB algorithm outperforms both VarUCB-Known and VarUCB-Unknown, which were proposed in this paper. This is expected: when the good event holds, algorithms that engage in less exploration tend to perform better empirically.

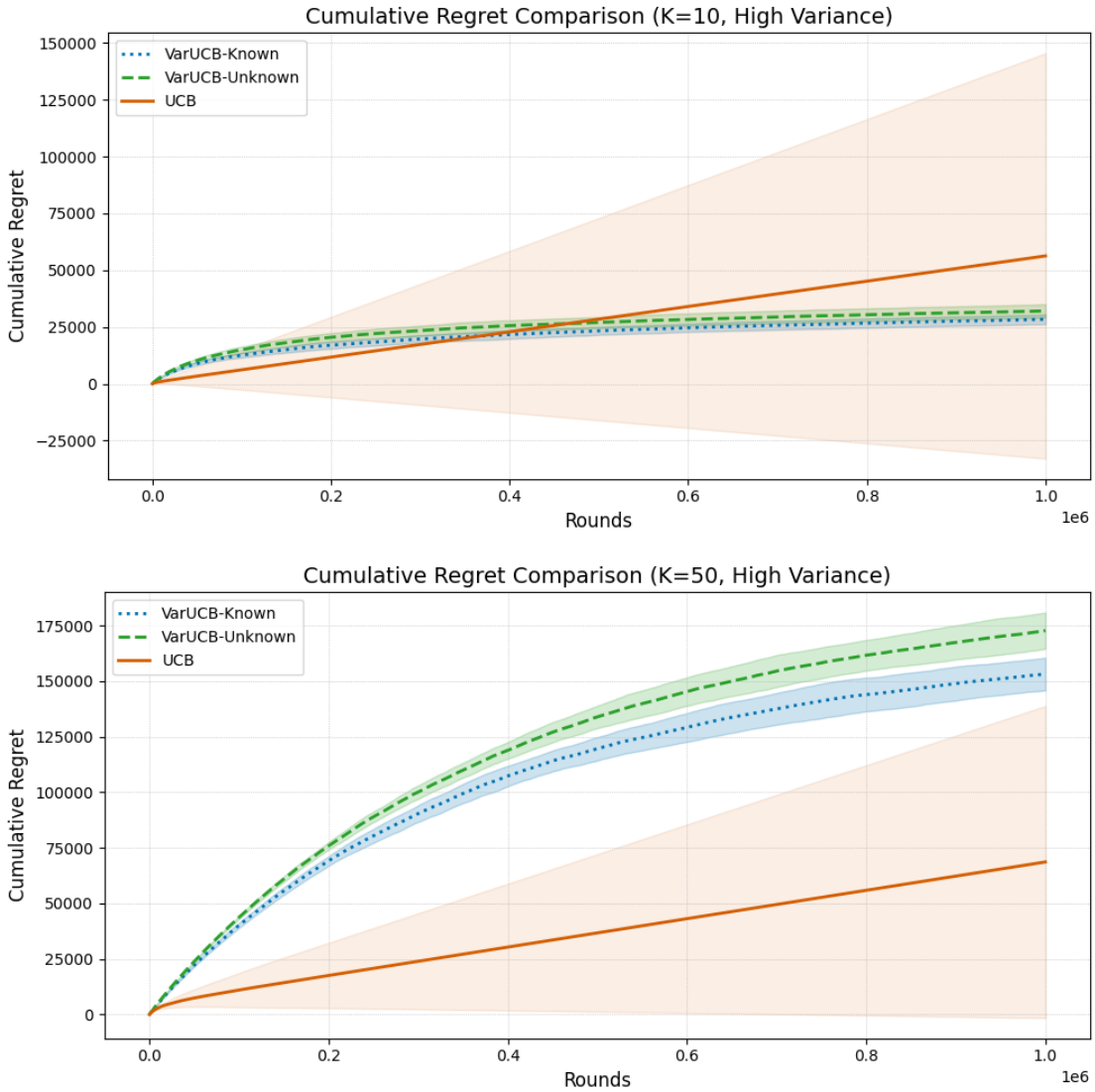

Figure 2: These plots show the regret in the high-variance regime for the standard UCB algorithm, as well as for VarUCB-Known and VarUCB-Unknown. The top figure corresponds to $K = 10$ arms, while the bottom figure shows results for $K = 50$ arms. Contrary to the low variance case, in this regime, the standard UCB algorithm performs worse than both VarUCB-Known and VarUCB-Unknown, which were proposed in this paper. This is expected: when the true mean isn't below the UCB index, the algorithm is likely to achieve linear regret and misidentify the optimal arm. The large shaded region also suggests that vanilla UCB is extremely unstable in this setting. The algorithms all perform better in the case of fewer arms, since they can explore each arm more thoroughly.

settings, classical UCB is more effective due to reduced estimation overhead. These results highlight the importance of accounting for reward variance structure in bandit algorithm design.

It would be interesting, as future work, to derive a sharper lower bound in the case where the variances are unknown, or to develop a min-max algorithm that removes these factors of $\log T$ in both the known and unknown variance case that matches the lower bound.

References

Jean-Yves Audibert and Sébastien Bubeck. Best arm identification in multi-armed bandits. In *COLT-23th Conference on learning theory-2010*, pp. 13–p, 2010.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration for multi-armed bandit problems, 2010. URL `https://arxiv.org/abs/0802.2655`.

Nicolo Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 208–214. JMLR Workshop and Conference Proceedings, 2011.

Eyal Even-Dar, Shie Mannor, Yishay Mansour, and Sridhar Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6), 2006.

Marco Faella, Alberto Finzi, and Luigi Sauro. Rapidly finding the best arm using variance. In *ECAI 2020*, pp. 2585–2591. IOS Press, 2020.

Victor Gabillon, Mohammad Ghavamzadeh, Alessandro Lazaric, and Sébastien Bubeck. Multi-bandit best arm identification. *Advances in Neural Information Processing Systems*, 24, 2011.

Marc Jourdan, Degenne Rémy, and Kaufmann Emilie. Dealing with unknown variances in best-arm identification. In *International Conference on Algorithmic Learning Theory*, pp. 776–849. PMLR, 2023.

Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *International conference on machine learning*, pp. 1238–1246. PMLR, 2013.

Masahiro Kato. Worst-case optimal multi-armed gaussian best arm identification with a fixed budget, 2024. URL `https://arxiv.org/abs/2310.19788`.

Masahiro Kato. Generalized neyman allocation for locally minimax optimal best-arm identification, 2025. URL `https://arxiv.org/abs/2405.19317`.

Branislav Kveton, Csaba Szepesvari, Zheng Wen, and Azin Ashkan. Cascading bandits: Learning to rank in the cascade model. In *International conference on machine learning*, pp. 767–776. PMLR, 2015.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Anusha Lalitha Lalitha, Kousha Kalantari, Yifei Ma, Anoop Deoras, and Branislav Kveton. Fixed-budget best-arm identification with heterogeneous reward variances. In *Uncertainty in Artificial Intelligence*, pp. 1164–1173. PMLR, 2023.

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

Pinyan Lu, Chao Tao, and Xiaojin Zhang. Variance-dependent best arm identification. In *Uncertainty in Artificial Intelligence*, pp. 1120–1129. PMLR, 2021.

Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 114–122, 2018.

Stefan Magureanu, Richard Combes, and Alexandre Proutiere. Lipschitz bandits: Regret lower bound and optimal algorithms. In *Conference on Learning Theory*, pp. 975–999. PMLR, 2014.

Odalric-Ambrym Maillard and Shie Mannor. Latent bandits. In *International Conference on Machine Learning*, pp. 136–144. PMLR, 2014.

Aadirupa Saha and Branislav Kveton. Only pay for what is uncertain: Variance-adaptive thompson sampling, 2023. URL `https://arxiv.org/abs/2303.09033`.

Qining Zhang, Tanner Fiez, Yi Liu, and Wenyang Liu. Multi-metric adaptive experimental design under fixed budget with validation, 2025. URL `https://arxiv.org/abs/2506.03062`.

Ruida Zhou and Chao Tian. Approximate top-m arm identification with heterogeneous reward variances. In *International Conference on Artificial Intelligence and Statistics*, pp. 7483–7504. PMLR, 2022.

A Key Lemmas

Lemma 6 *Assume that $X_i - \mu$ are independent, σ-subgaussian random variables. Then for any $\varepsilon \geq 0$,*

$$\mathbb{P}(\hat{\mu} \geq \mu + \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right) \quad \textit{and} \quad \mathbb{P}(\hat{\mu} \leq \mu - \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right),$$

where $\hat{\mu} = \frac{1}{n}\sum_{t=1}^{n} X_t$.

We present the proof of Theorem 1.

Proof: Without loss of generality assume arm 1 is optimal, so $\mu_1 = \mu^\star$.

Step 1: Anytime concentration. If $X_1, \ldots, X_s$ are independent σ-sub-Gaussian with mean μ, then

$$\mathbb{P}\left(\exists s \leq T : |\hat{\mu}_s - \mu| \geq \sigma\sqrt{\frac{2x}{s}}\right) \leq e^{-x}.$$

Step 2: Define the optimism deficit of the optimal arm. Define the random variable

$$\Delta := \left(\mu^\star - \min_{1\leq s\leq T}\left[\hat{\mu}_{1,s} + \sigma_1\sqrt{\frac{4}{s}\log_+\left(\frac{T}{Ks}\right)}\right]\right)^+.$$

We now bound $\mathbb{P}(\Delta \geq x)$ using a dyadic peeling argument.

Observe that the event $\{\Delta \geq x\}$ implies that there exists $1 \leq s \leq T$ such that

$$\hat{\mu}_{1,s} - \mu^\star \leq -\sigma_1\sqrt{\frac{4}{s}\log_+\left(\frac{T}{Ks}\right)} - x.$$

Equivalently, writing $S_s := \sum_{t=1}^{s}(r_t - \mu^\star)$,

$$\exists s \leq T : \quad S_s \leq -sx - \sqrt{4s\log_+\left(\frac{T}{Ks}\right)}.$$

Partition $\{1, \ldots, T\}$ into dyadic blocks

$$\mathcal{I}_j := \{2^j, \ldots, 2^{j+1} - 1\}, \qquad j = 0, 1, \ldots, \lfloor \log_2 T \rfloor.$$

If the above event occurs for some $s \in \mathcal{I}_j$, then necessarily

$$\exists s \le 2^{j+1} : \quad S_s \le -2^j x - \sqrt{4 \cdot 2^j \log_+\left(\frac{T}{K2^{j+1}}\right)}. \tag{7}$$

This is the key peeling step: within each dyadic block, we replace s by its minimum 2^j inside the boundary and by its maximum 2^{j+1} inside the supremum.

By the maximal inequality for sub-Gaussian partial sums, for any $a > 0$ and any m,

$$\mathbb{P}(\exists s \le m : S_s \le -a) \le \exp\left(-\frac{a^2}{2m\sigma_1^2}\right).$$

Applying this to $m = 2^{j+1}$ and the threshold

$$a = 2^j x + \sqrt{4 \cdot 2^j \log_+\left(\frac{T}{K2^{j+1}}\right)},$$

and summing over j, we obtain

$$\mathbb{P}(\Delta \ge x) \le \sum_{j=0}^{\lfloor \log_2 T \rfloor} \exp\left(-\frac{\left(2^j x + \sqrt{4 \cdot 2^j \log_+\left(\frac{T}{K2^{j+1}}\right)}\right)^2}{2^{j+2}\sigma_1^2}\right). \tag{8}$$

Using $(a+b)^2 \ge a^2 + b^2$ and simplifying, the series is bounded by

$$\mathbb{P}(\Delta \ge x) \le \min\left\{1, \frac{C_1 \sigma_1^2 K}{T x^2}\right\}.$$

Integrating the tail,

$$\mathbb{E}[\Delta] \le C_2 \sigma_1 \sqrt{\frac{K}{T}}.$$

Hence

$$\mathbb{E}[T\Delta] \le C_2 \sigma_1 \sqrt{KT}.$$

Step 3: Regret Decomposition Recall the following regret decomposition.

$$R_T = \sum_{i:\Delta_i > 0} \Delta_i \mathbb{E}[T_i(T)].$$

Split arms according to whether $\Delta_i \le 2\Delta$ or $\Delta_i > 2\Delta$.

Step 4: Small-gap arms. If $\Delta_i \le 2\Delta$, then

$$\Delta_i T_i(T) \le 2T\Delta.$$

Summing and taking expectation gives

$$\mathbb{E}[\text{small-gap regret}] \le 2T\mathbb{E}[\Delta] \le C_3 \sigma_1 \sqrt{KT}.$$

Step 5: Large-gap arms. Fix a suboptimal arm i with $\Delta_i > 0$. Define

$$\kappa_i = \sum_{s=1}^{T} \mathbf{1}\left(\hat{\mu}_{i,s} + \sigma_i \sqrt{\frac{4}{s}\log_+\left(\frac{T}{Ks}\right)} \geq \mu_i + \frac{\Delta_i}{2}\right).$$

If $\Delta_i > 2\Delta$, then whenever arm i is selected, its index must exceed $\mu_i + \Delta_i/2$. Hence

$$T_i(T) \leq \kappa_i.$$

Using sub-Gaussian concentration,

$$\mathbb{E}[\kappa_i] \leq C_4\left(\frac{\sigma_i^2}{\Delta_i^2} + \sqrt{\frac{T}{K}}\right).$$

Therefore

$$\Delta_i \mathbb{E}[T_i(T)] \leq C_4\left(\frac{\sigma_i^2}{\Delta_i} + \Delta_i\sqrt{\frac{T}{K}}\right).$$

For arms with $\Delta_i \geq C_5\sigma_i\sqrt{\frac{K}{T}}$, the second term dominates and summing over such arms gives

$$\sum_{i:\text{large-gap}} \Delta_i \mathbb{E}[T_i(T)] \leq C_6\sqrt{T\sum_{i=1}^{K}\sigma_i^2} + \sum_{i=1}^{K}\Delta_i.$$

Step 6: Combine bounds. Adding small-gap and large-gap contributions,

$$R_T \leq C\sqrt{T\sum_{i=1}^{K}\sigma_i^2} + \sum_{i=1}^{K}\Delta_i.$$

This completes the proof. □

B VarUCB-Unknown Regret Proof

We present the proof of Theorem 2.

Proof: Assume arm 1 is optimal, i.e. $\mu_1 = \mu^\star$, and define $\Delta_i = \mu^\star - \mu_i$.

Step 1: A uniform "variance-good" event. Fix $\delta = T^{-2}$. For Gaussian rewards,

$$\frac{(s-1)\hat{\sigma}_i^2(s)}{\sigma_i^2} \sim \chi^2_{s-1}.$$

A standard χ^2 tail inequality implies that for all $s \geq 8\log T + 1$,

$$\mathbb{P}\left(\sigma_i^2 > \frac{\hat{\sigma}_i^2(s)}{1 - 2\sqrt{\frac{2\log T}{s-1}}}\right) \leq \frac{1}{T^3}.$$

Taking a union bound over $i \in [K]$ and $s \in \{8\log T + 1, \ldots, T\}$ yields an event $\mathcal{G}_\sigma$ such that

$$\mathcal{G}_\sigma := \bigcap_{i=1}^{K}\bigcap_{s=8\log T+1}^{T}\left\{\sigma_i^2 \leq \hat{\sigma}_i^{2,\text{UCB}}(s)\right\}, \qquad \mathbb{P}(\mathcal{G}_\sigma) \geq 1 - \frac{1}{T^2}. \tag{9}$$

On $\mathcal{G}_\sigma$, for every arm and every sample size in the valid range, the algorithm's variance plug-in satisfies $\sqrt{\hat{\sigma}_i^{2,\text{UCB}}(s)} \geq \sigma_i$.

Step 2: Anytime concentration for the means. Conditioned on the sampling times, Gaussian rewards are σ_i-sub-Gaussian. Thus for any fixed arm i and any $x > 0$,

$$\mathbb{P}\left(\exists s \le T : |\hat{\mu}_{i,s} - \mu_i| \ge \sigma_i \sqrt{\frac{2x}{s}}\right) \le e^{-x}.$$

(We only need this for x of order $\log(\cdot)$ below.)

Step 3: Define the optimism deficit of the optimal arm. Define

$$\Delta := \left(\mu^\star - \min_{1\le s\le T}\left[\hat{\mu}_{1,s} + \sqrt{\hat{\sigma}_1^{2,\mathrm{UCB}}(s)}\sqrt{\frac{4}{s}\log_+\left(\frac{T}{Ks}\right)}\right]\right)^+.$$

On $\mathcal{G}_\sigma$, we may lower bound the index of arm 1 by the known-variance MOSS index (since $\sqrt{\hat{\sigma}_1^{2,\mathrm{UCB}}(s)} \ge \sigma_1$), hence

$$\Delta \le \left(\mu^\star - \min_{1\le s\le T}\left[\hat{\mu}_{1,s} + \sigma_1\sqrt{\frac{4}{s}\log_+\left(\frac{T}{Ks}\right)}\right]\right)^+.$$

Therefore, the same peeling argument equation (8) as in the known-variance MOSS proof applies *on* $\mathcal{G}_\sigma$. It yields the tail bound

$$\mathbb{P}(\Delta \ge x \mid \mathcal{G}_\sigma) \;\le\; \min\left\{1, \frac{C_1\sigma_1^2 K}{Tx^2}\right\}.$$

Unconditioning and using $\mathbb{P}(\mathcal{G}_\sigma^c) \le T^{-2}$ implies

$$\mathbb{E}[\Delta] \le C_2\sigma_1\sqrt{\frac{K}{T}} + \underbrace{\mathbb{E}[\Delta\,\mathbf{1}\{\mathcal{G}_\sigma^c\}]}_{\le \mathbb{P}(\mathcal{G}_\sigma^c)} \;\le\; C_2\sigma_1\sqrt{\frac{K}{T}} + \frac{1}{T^2}.$$

Up to constants, this matches the known-variance control of $\mathbb{E}[\Delta]$.

Step 4: Regret decomposition and small-gap arms. Write

$$R_T = \sum_{i:\Delta_i>0} \Delta_i\,\mathbb{E}[T_i(T)].$$

Split arms into those with $\Delta_i \le 2\Delta$ and those with $\Delta_i > 2\Delta$. For the small-gap arms, since $T_i(T) \le T$,

$$\sum_{i:\Delta_i\le 2\Delta} \Delta_i\,T_i(T) \le 2T\Delta.$$

Taking expectations gives

$$\mathbb{E}[\text{small-gap regret}] \le 2T\,\mathbb{E}[\Delta] \le C_3\sigma_1\sqrt{KT}.$$

Step 5: Large-gap arms and a κ_i argument. Fix a suboptimal arm i and define

$$\kappa_i := \sum_{s=1} T\mathbf{1}\left(\hat{\mu}_{i,s} + \sqrt{\hat{\sigma}_i^{2,\mathrm{UCB}}(s)}\sqrt{\frac{4}{s}\log_+\left(\frac{T}{Ks}\right)} \ge \mu_i + \frac{\Delta_i}{2}\right).$$

On the event $\{\Delta_i > 2\Delta\} \cap \mathcal{G}_\sigma$, the optimal arm's index is always larger than $\mu^\star - \Delta > \mu^\star - \Delta_i/2 = \mu_i + \Delta_i/2$, so if arm i is selected then its index must be at least $\mu_i + \Delta_i/2$. Hence,

$$T_i(T) \le \kappa_i \qquad \text{on } \{\Delta_i > 2\Delta\} \cap \mathcal{G}_\sigma. \tag{10}$$

We now bound $\mathbb{E}[\kappa_i]$. Let $s_0 := \lceil 16 \log T \rceil$ so that the variance-UCB is finite. For $s \geq s_0$, on $\mathcal{G}_\sigma$ we have $\hat{\sigma}_i^{2,\mathrm{UCB}}(s) \leq \frac{\hat{\sigma}_i^2(s)}{1-2\sqrt{2\log T/(s-1)}}$ and in particular the inflation factor satisfies

$$\frac{1}{\sqrt{1-2\sqrt{\frac{2\log T}{s-1}}}} \leq 1 + c\sqrt{\frac{\log T}{s}} \quad \text{for an absolute constant } c \text{ and all } s \geq s_0.$$

Using the Gaussian concentration of $\hat{\mu}_{i,s} - \mu_i$ and summing tail probabilities over s yields

$$\mathbb{E}[\kappa_i] \leq C_4 \left(\frac{\sigma_i^2}{\Delta_i^2} \log T + \sqrt{\frac{T}{K}} \log T \right).$$

(The $\log T$ arises from the uniform variance control and the resulting inflation in the mean confidence radius.) Multiplying by Δ_i gives

$$\Delta_i \, \mathbb{E}[T_i(T)] \leq \Delta_i \, \mathbb{E}[\kappa_i] + \Delta_i \cdot T\mathbb{P}(\mathcal{G}_\sigma^c) \leq C_4 \left(\frac{\sigma_i^2}{\Delta_i} \log T + \Delta_i \sqrt{\frac{T}{K}} \log T \right) + O\left(\frac{1}{T}\right).$$

Step 6: Sum over large-gap arms and conclude. Summing the above bound over arms with $\Delta_i > 2\Delta$ and using Cauchy–Schwarz for the second term yields

$$\mathbb{E}[\text{large-gap regret}] \leq C_5 \sqrt{T \sum_{i=1}^{K} \sigma_i^2} \, \log T + \sum_{i=1}^{K} \Delta_i.$$

Combining small-gap and large-gap contributions proves the theorem. □

C Known Variances Lower Bound

Lemma 7 *The divergence between two Gaussian distributions distributions with means μ_i and μ_i' (and both with variance σ_i^2)*

$$D(N(\mu_i, \sigma_i), N(\mu_i, \sigma_i')) = \frac{(\mu_i - \mu_i')^2}{2\sigma_i^2}$$

This is a routine computation,

Proof:

$$D(P||Q) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}} \log\left(\frac{\frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}}{\frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x-\mu_i')^2}{2\sigma_i^2}}}\right) dx \tag{11}$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}} \left(-\frac{(x-\mu_i)^2}{2\sigma_i^2} + \frac{(x-\mu_i')^2}{2\sigma_i^2}\right) dx \tag{12}$$

$$= \int_{-\infty}^{\infty} \frac{1}{2\sigma_i^2\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}} \left(2(\mu_i - \mu_i')x - \mu_i^2 + (\mu_i')^2\right) dx \tag{13}$$

$$= \frac{\mu_i - \mu_i'}{\sigma_i^2} \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}} x dx}_{=\mu_i} + \frac{-\mu_i^2 + (\mu_i')^2}{2\sigma_i^2} \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}} dx}_{=1} \tag{14}$$

$$= \frac{\mu_i(\mu_i - \mu_i')}{\sigma_i^2} + \frac{-\mu_i^2 + (\mu_i')^2}{2\sigma_i^2} \tag{15}$$

$$= \frac{(\mu_i - \mu_i')^2}{2\sigma_i^2} \tag{16}$$

□

We now present the proof of Theorem 4

Proof: We defer the details of the proof to the appendix, but present the key ideas here. We let the first environment ν environment, be a Gaussian distribution with the following means.

$$\mu = (\Delta, 0, 0, ..., 0) \tag{17}$$

Let arm i be an arm that's pulled at most $\frac{\sigma_i^2 T}{\sum_j \sigma_j^2}$ times. This must exist otherwise the total number of pulls exceeds T. On the other hand, we construct the other ν' environment with a Gaussian bandit with mean

$$\mu' = (\Delta, 0, 0, ..., 0, 2\Delta, 0, ..., 0), \tag{18}$$

where μ' at arm i equal to 2Δ. Then we calculate,

$$R_T(\pi, \nu_\mu) \geq \mathbb{P}_\mu(n_1(T) \leq n/2)\frac{T\Delta}{2}$$

and

$$R_T(\pi, \nu_\mu') > \mathbb{P}_{\mu'}(n_1(T) > T/2)\frac{T\Delta}{2}.$$

By lemma 4, we apply the Bretagnolle-Huber inequality

$$R_T(\pi, \nu_\mu) + R_n(\pi, \nu_\mu') > \frac{T\Delta}{2}(\mathbb{P}_\mu(n_1(n) \leq T/2) + \mathbb{P}_{\mu'}(n_1(T) > T/2)) \geq \frac{T\Delta}{4}\exp(-D(\mathbb{P}_\nu, \mathbb{P}_\nu'))$$

By lemma 5,

$$\mathrm{D}\left(\mathbb{P}_\nu || \mathbb{P}_{\nu'}\right) = \mathbb{E}_\nu\left[n_i(T)\right] \cdot D(N(\Delta, \sigma_i^2), N(2\Delta, \sigma_i^2))$$

Then

$$= \frac{\Delta^2}{2\sigma_i^2}\mathbb{E}_\nu\left[T_i(n)\right]$$

Since for arm i, $\mathbb{E}_\nu\left[n_i(T)\right] \leq \frac{T\sigma_i^2}{\sum_j \sigma_j^2}$. Therefore,

$$\mathrm{D}\left(\mathbb{P}_\nu || \mathbb{P}_{\nu'}\right) \leq \frac{\Delta^2}{2\sigma_i^2} \cdot \frac{T\sigma_i^2}{\sum_j \sigma_j^2} = \frac{T\Delta^2}{2\sum_j \sigma_j^2}$$

We plug this back again into lemma 5,

$$R_n(\pi, \nu_\mu) + R_n(\pi, \nu'_\mu) \geq \frac{T\Delta}{4} \exp\left(-\frac{T\Delta^2}{2\sum_j \sigma_j^2}\right)$$

We optimize over Δ to maximize the RHS of the above. In particular, we pick $\Delta = \sqrt{\frac{\sum_j \sigma_j^2}{4T}}$ into the lower bound, we get:

$$R_T(\pi, \nu_\mu) + R_T(\pi, \nu'_\mu) \geq C\sqrt{T\sum_j \sigma_j^2}$$

Therefore, there exists an environment for which we have

$$R_T \geq c\sqrt{T\sum_j \sigma_j^2}$$

for some constant $c > 0$ □