A Simple and Optimal Policy Design with Safety against Heavy-tailed Risk for Stochastic Bandits

David Simchi-Levi
Institute for Data, Systems, and Society, Massachusetts Institute of Technology, MA 02139, dslevi@mit.edu

Zeyu Zheng
Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA 94720, zyzheng@berkeley.edu

Feng Zhu
Institute for Data, Systems, and Society, Massachusetts Institute of Technology, MA 02139, fengzhu@mit.edu

We study the stochastic multi-armed bandit problem and design new policies that enjoy both worst-case optimality for expected regret and light-tailed risk for regret distribution. Starting from the two-armed bandit setting with time horizon $T$, we propose a simple policy and prove that the policy (i) enjoys the worst-case optimality for the expected regret at order $O(\sqrt{T\ln T})$ and (ii) has the worst-case tail probability of incurring a linear regret decay at an exponential rate $\exp(-\Omega(\sqrt{T}))$, a rate that we prove to be best achievable for all worst-case optimal policies. Briefly, our proposed policy achieves a delicate balance between doing more exploration at the beginning of the time horizon and doing more exploitation when approaching the end, compared to the standard Successive Elimination policy and Upper Confidence Bound policy. We then improve the policy design and analysis to work for the general $K$-armed bandit setting. Specifically, the worst-case probability of incurring a regret larger than any $x > 0$ is upper bounded by $\exp(-\Omega(x/\sqrt{KT}))$. We then enhance the policy design to accommodate the “any-time” setting where $T$ is not known a priori, and prove equivalently desired policy performances as compared to the “fixed-time” setting with known $T$. A brief account of numerical experiments is conducted to illustrate the theoretical findings. We conclude by extending our proposed policy design to the general stochastic linear bandit setting and proving that the policy leads to both worst-case optimality in terms of expected regret order and light-tailed risk on the regret distribution.

Key words: stochastic bandits, worst-case optimality, instance-dependent consistency, heavy-tailed risk

1. Introduction

The stochastic multi-armed bandit (MAB) problem is a widely studied problem in the domain of sequential decision-making under uncertainty, with applications such as online advertising, recommendation systems, digital clinical trials, financial portfolio design, etc. In a standard MAB problem, whose formulation will be formally discussed later in the Section 2, the decision maker sequentially chooses one of many not fully known arms (designs) in each of many time periods. The objective of decision maker is typically to maximize the expectation of the sum of rewards accumulated across all time period. The MAB problem and its associated policy design also provide
notable theoretical insights exhibiting the exploration-exploitation trade-off, where the decision maker objectively needs to balance the exploration of arms whose reward distributions are relatively unknown and the exploitation of arms whose expected rewards are high and relatively more known. There is a vast of literature on MAB problem, and we refer to Slivkins et al. (2019) among others for a review.

For policy design and analysis, much of the MAB literature uses the metric of maximizing the expected cumulative reward, or equivalently minimizing the expected regret (where regret refers to the difference between the cumulative reward obtained by always pulling the best arm and by executing a policy that does not a priori know the reward distributions). The optimality of a policy is often characterized through its expected regret’s rate (order of dependence) on the experiment horizon $T$.

However, if an MAB policy design only focuses on optimizing the expected regret, the policy design may be exposed to risks that can arise in other aspects. As recently documented in Fan and Glynn (2022) about the standard Upper Confidence Bound (UCB) policy (Auer et al. 2002), as well as will be extended in our work about the Successive Elimination (SE) policy (Even-Dar et al. 2006) and the Thompson Sampling (TS) policy (Russo et al. 2018), these renowned policies, despite of enjoying optimality on expected regret, can incur a “heavy-tailed risk”. That is, the distribution of the regret has a heavy tail — the probability of incurring a linear regret slowly decays at a polynomial rate $\Omega(\text{poly}(1/T))$ as $T$ tends to infinity. In contrast, a “light-tailed” risk in this MAB setting means that the probability of a policy incurring a linear regret decays at an exponential rate $\exp(-\Omega(T^\beta))$ for some $\beta > 0$. The heavy-tailed risk can be undesired when an MAB policy is used in applications that are sensitive to tail risks, including but not limited to finance, healthcare, supply chain, etc. In fact, understanding heavy-tailed risks and their associated disruptions in the aforementioned applications have been a keen focus in the operational research literature; see Bouchaud and Georges (1990), Bouchaud et al. (2000), Chopra and Sodhi (2004), Embrechts et al. (2013), Sodhi and Tang (2021) for example.

Noting that the renowned policies may incur a heavy-tailed risk on the regret distribution, when achieving the optimality on the rate of expected regret, our work is primarily motivated by an attempt to answer the following question. *Is it possible to design a policy that on one hand enjoys the classical notion of optimality regarding expected regret, whereas on the other hand enjoys light-tailed risk for the regret distribution?* If the answer is yes, then that policy design would enjoy both optimality and safety against heavy-tailed risk. Motivated by this question, we summarize our contributions in Section 1.1.

To facilitate describing the results on regret orders and function orders, we adopt the family of Bachmann–Landau notation. That is, we use $O(\cdot)$ ($\tilde{O}(\cdot)$) and $\Omega(\cdot)$ ($\tilde{\Omega}(\cdot)$) to present upper and
lower bounds on the growth rate up to constant (logarithmic) factors, respectively, and \( \Theta(\cdot) (\tilde{\Theta}(\cdot)) \) to characterize the rate when the upper and lower bounds match up to constant (logarithmic) factors. We use \( o(\cdot) \) and \( \omega(\cdot) \) to present strictly dominating upper bounds and strictly dominated lower bounds, respectively.

### 1.1. Our Contributions

1. We first argue that instant-dependent consistency (formal definition to be discussed in Section 2) and light-tailed risk are incompatible. Recently, Fan and Glynn (2022) showed that information-theoretically optimized bandit policies as well as general UCB policies suffer from heavy-tailed risk. Built upon their analysis and results, we find that any instant-dependent consistent policy cannot avoid heavy-tailed risk: if an instant-dependent consistent policy has the probability of incurring a linear regret decay as \( \exp(-f(T)) \), then \( f(T) \) must be \( o(T^\beta) \) as \( T \to +\infty \) for any \( \beta > 0 \). Moreover, any policy that has instance-dependent \( O(\ln T) \) expected regret, including the standard UCB and SE policy, and the TS policy, incurs (i) a linear regret with probability \( \Omega(\text{poly}(1/T)) \) and (ii) an expected regret that is almost linear in \( T \) if the risk parameter is severely misspecified. The implication is that if we want to find a policy design that avoids the heavy-tailed risk on regret distribution, we shall explore policies that are different different from policies that have instance-dependent \( O(\ln T) \) expected regret.

2. We show that worst-case optimality and light-tailed risk can co-exist for policy design. Starting from the two-armed bandit setting, we provide a new policy design and prove that it enjoys both the worst-case optimality \( \tilde{O}(\sqrt{T}) \) for the expected regret and the light-tailed risk \( \exp(-\Omega(\sqrt{T})) \) for the regret distribution. We also prove that such exponential decaying rate of the tail probability is the best achievable within the class of worst-case optimal policies, as a “lower bound” result. Our policy design builds upon the idea of confidence bounds, and constructs different bonus terms compared to the standard ones to ensure safety against heavy-tailed risk.

3. We extend our results from the two-armed bandit to the general \( K \)-armed bandit and characterize the tail probability bound for any regret threshold in an explicit form and through non-asymptotic lens. By further improving our policy design, we show that the worst-case probability of incurring a regret larger than \( x \) is bounded approximately by \( \exp(-\Omega(x/\sqrt{KT})) \) for any \( x > 0 \). We then enhance the policy design to accommodate the “any-time” setting where \( T \) is not known a priori, as a more challenging setting compared to the “fixed-time” setting where \( T \) is known a priori. We design a policy for the “any-time” setting and prove that the policy enjoys an equivalently desired exponential decaying tail and optimal expected regret as in the “fixed-time” setting. Despite of the simplicity of our proposed policy design, the associated proof techniques are novel
and may be useful for broader analysis on regret distribution and tail risk. Our result also partially answers an open question raised in Lattimore and Szepesvári (2020) for the stochastic MAB problem. A brief account of experiments are conducted to illustrate our theoretical findings.

4. We finally extend the idea of our policy design to apply to the linear bandit setting, a setting that sits in the broad stochastic bandit problems and deviates from the MAB setting. In the linear bandit setting, the decision maker chooses an action in each period from a potentially time-varying continuous action set, instead of from \( K \) discrete arms (see, e.g., Abbasi-Yadkori et al. 2011 for reference). We prove that our simple policy design can be integrated to classical linear bandit algorithms and lead to both worst-case optimality in terms of expected regret and light-tailed risk on the regret distribution.

1.2. Related Work

Our work builds upon the vast literature of designing and analyzing policies for the stochastic MAB problem and its various extensions. Comprehensive reviews can be found in Bubeck and Cesa-Bianchi (2012), Russo et al. (2018), Slivkins et al. (2019), Lattimore and Szepesvári (2020). A standard paradigm for obtaining a near-optimal regret is to first fix some confidence parameter \( \delta > 0 \). Then a “good event” is defined such that good properties are retained conditioned on the event (for example, in the stochastic MAB problem, the good event is such that the mean of each arm always lies in the confidence bound). Then one can obtain both high-probability and worst-case expected regret bounds through careful analysis on the good event. It is known that the stochastic MAB problem has the following regret bound: for any fixed \( \delta \in (0, 1) \), the regret bound of UCB is bounded by \( O(\sqrt{KT\ln(T/\delta)}) \) with probability at least \( 1 - \delta \). Or equivalently speaking, the probability of incurring a \( \Omega(\sqrt{KT\ln(T/\delta)}) \) regret is bounded by \( \delta \). However, the parameter \( \delta \) must be an input parameter for the policy. We will discuss this issue in more details in Section 3. In Section 17.1 of Lattimore and Szepesvári (2020), an open question is asked: is it possible to design a single policy such that the worst-case probability of incurring a \( \Omega(\sqrt{KT\ln(1/\delta)}) \) regret is bounded by \( \delta \) for any \( \delta > 0 \) and any \( K \)-armed bandit problem with 1-subgaussian stochastic rewards? We partially answer this question by designing a policy such that for any \( \delta > 0 \), the probability of incurring a

\[
\Omega\left(\frac{\sqrt{KT\ln(T/\delta)}}{\sqrt{\ln T}}\right)
\]

regret is bounded by \( \delta \). We note that there has been a related result in the adversarial bandit setting (see, e.g., Neu 2015, Lattimore and Szepesvári 2020). It is shown that for the \( K \)-armed bandit problem with adversarial rewards uniformly in \([0, 1]\), there exists a single policy EXP3-IX such that the worst-case probability of incurring a

\[
\Omega\left(\frac{\sqrt{KT\ln(K/\delta)}}{\sqrt{\ln K}}\right)
\]
regret is bounded by $\delta$ for any $\delta > 0$. The difference between this result and ours are two-folds. From the policy design prospective, the idea behind EXP3-IX is to use exponential weight, while the idea behind our policy is to use a modified confidence bound designed to handle the stochastic setting. From the model setting perspective, in the adversarial setting, rewards are assumed to be uniformly bounded, and the bound is an input to the policy. While in the stochastic setting, the magnitude of a single reward is uncontrollable. In fact, naively reducing the bound under the adversarial setting into one under the stochastic setting not only makes the new bound sub-optimal on the order of $\ln(1/\delta)$ but also requires knowing the confidence parameter $\delta$ in advance. This point is discussed in detail in Section 4.

There has been not much work on understanding the tail risk of stochastic bandit algorithms. Two earlier works are Audibert et al. (2009), Salomon and Audibert (2011) and they studied the concentration properties of the regret around the instance-dependent mean $O(\ln T)$. They showed that in general the regret of the policies concentrate only at a polynomial rate. That is, the probability of incurring a regret of $c(\ln T)^p$ (with $c > 0$ and $p > 1$ fixed) is approximately polynomially decaying with $T$. Different from our work, the concentration in their work is under an instance-dependent environment, and so such polynomial rate might be different across different instances. Nevertheless, their results indicate that standard bandit algorithms generally have undesirable concentration properties. Recently, Ashutosh et al. (2021) showed that an online learning policy with the goal of obtaining logarithmic regret can be fragile, in the sense that a mis-specified risk parameter (e.g., the parameter for subgaussian noises) in the policy can incur an instance-dependent expected regret of $\omega(\ln T)$. They then developed robust algorithms to circumvent the issue. Note that their goal is to handle mis-specification related with risk, but still the task is to minimize the expected regret.

Our work is inspired by Fan and Glynn (2022), who first provided a rigorous formulation to analyze heavy-tailed risk for bandit algorithms and showed that for any information-theoretically optimized bandit policy, the probability of incurring a linear regret is very heavy-tailed: at least $\Omega(1/T)$. They additionally showed that optimized UCB bandit designs are fragile to mis-specifications and they modified UCB algorithms to ensure a desired polynomial rate of tail risk, which makes the algorithms more robust to mis-specifications. Built upon their analysis, we show an incompatibility result. That is, a large family of policies — all policies that are consistent — suffer from heavy-tailed risk (see Section 1.1). Further, we propose a simple and new policy design that leads to both light-tailed risk (tail bound exponentially decaying with $\sqrt{T}$) and worst-case optimality (expected regret bounded by $\tilde{O}(\sqrt{T})$). We then show that our proposed simple policy design naturally extends to the general $K$-armed MAB setting and the linear bandit setting, and prove the compatibility between worst-case optimality and light-tailed risk.
Recently, there is an increasing line of works analyzing the limiting behaviour of standard UCB and TS policies by considering the diffusion approximations (see, e.g., Araman and Caldentey 2021, Wager and Xu 2021, Fan and Glynn 2021, Kalvit and Zeevi 2021). These works typically consider asymptotic limiting regimes that are set such that the gaps between arm means shrink with the total time horizon. We do not consider limiting regimes but instead consider the original problem setting with general parameters (e.g., gaps). We study how the tail probability decays with $T$ under original environments without taking the gaps to zero. Another line of works closely related with ours involve solving risk-averse formulations of the stochastic MAB problem (see, e.g., Sani et al. 2012, Galichet et al. 2013, Maillard 2013, Zimin et al. 2014, Vakili and Zhao 2016, Cassel et al. 2018, Tamkin et al. 2019, Prashanth et al. 2020, Zhu and Tan 2020, Baudry et al. 2021, Khajonchotpanya et al. 2021, Chen and Yang 2022, Chang and Tan 2022). Compared to standard stochastic MAB problems, the main difference in their works is that arm optimality is defined using formulations other than the expected value, such as mean-variance criteria and (conditional) value-at-risk. These formulations consider some single metric that is different compared to the expected regret. From the formulation perspective, our work is different in the sense that we develop policies that on one hand maintain the classical worst-case optimal expected regret, whereas simultaneously achieve light-tailed risk bound. The policy design and analysis in our work are therefore also different from the literature and might be of independent interest.

1.3. Organization and Notation

The rest of the paper is organized as follows. In Section 2, we discuss the setup and introduce the key concepts: light-tailed risk, instance-dependent consistency, worst-case optimality. In Section 3, we show results on the incompatibility between light-tailed risk and consistency, and show the compatibility between light-tailed risk and worst-case optimality via a new policy design. In Section 4, we consider the general $K$-armed bandit model and show a precise regret tail bound for our new policy design. We detail the proof road-map and how to further improve the design. In Section 5, we present numerical experiments. In Section 6, we show how to extend our policy design into the general linear bandit setting and obtain similar light-tailed regret bound as in the MAB case. Finally, we conclude in Section 7. All detailed proofs are left to the supplementary material.

Before proceeding, we introduce some notation. For any $a, b \in \mathbb{R}$, $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. For any $a \in \mathbb{R}$, $a_+ = \max\{a, 0\}$. We denote $[N] = \{1, \cdots, N\}$ for any positive integer $N$.

2. The Setup

In this section, we first discuss the model setup. We then formally define three terms that appeared in the introduction and will appear in the rest of this work: light-tailed risk, instance-dependent consistency, and worst-case optimality.
Fix a time horizon of $T$ and the number of arms as $K$. Throughout the paper, we assume that $T \geq 3$, $K \geq 2$, and $T \geq K$. In each time $t \in [T]$, based on all the information prior to time $t$, the decision maker (DM) pulls an arm $a_t \in [K]$ and receives a reward $r_{t,a_t}$. More specifically, let $H_t = \{a_1, r_{1,a_1}, \ldots, a_{t-1}, r_{t-1,a_{t-1}}\}$ be the history prior to time $t$. When $t = 1$, $H_1 = \emptyset$. At time $t$, the DM is free to adopt any admissible policy $\pi_t : H_t \rightarrow a_t$ that maps the history $H_t$ to an action $a_t$ that may be realized from a discrete probability distribution on $[K]$. The environment then independently samples a reward $r_{t,a_t} = \theta_{a_t} + \epsilon_{t,a_t}$ and reveals it to the DM. Here, $\theta_{a_t}$ is the mean reward of arm $a_t$, and $\epsilon_{t,a_t}$ is an independent zero-mean noise term. We assume that $\epsilon_{t,a_t}$ is $\sigma$-subgaussian. That is, there exists a $\sigma > 0$ such that for any time $t$ and arm $k$,

$$\max \{\mathbb{P}(\epsilon_{t,k} \geq x), \mathbb{P}(\epsilon_{t,k} \leq -x)\} \leq \exp(-x^2/(2\sigma^2)).$$

Let $\theta = (\theta_1, \ldots, \theta_K)$ be the mean vector. Let $\theta_\ast = \max\{\theta_1, \ldots, \theta_K\}$ be the optimal mean reward among the $K$ arms. Note that DM does not know $\theta$ at the beginning, except that $\theta \in [0,1]^K$. The empirical regret of the policy $\pi = (\pi_1, \ldots, \pi_T)$ under the mean vector $\theta$ and the noise parameter $\sigma$ over a time horizon of $T$ is defined as

$$\hat{R}_{\theta,\sigma}^\pi(T) = \theta_\ast \cdot T - \sum_{t=1}^{T} (\theta_{a_t} + \epsilon_{t,a_t}).$$

Let $\Delta_k = \theta_\ast - \theta_k$ be the gap between the optimal arm and the $k$th arm. Let $n_{t,k}$ be the number of times arm $k$ has been pulled up to time $t$. That is, $n_{t,k} = \sum_{s=1}^{t} \mathbb{1}\{a_s = k\}$. For simplicity, we will also use $n_k = n_{T,k}$ to denote the total number of times arm $k$ is pulled throughout the whole time horizon. We define $\tau_k(n)$ as the time period that arm $k$ is pulled for the $n$th time. Define the pseudo regret as

$$\hat{R}_{\theta,\sigma}^\pi(T) = \sum_{k=1}^{K} n_k \Delta_k$$

and the genuine noise as

$$N_{\pi}(T) = \sum_{t=1}^{T} \epsilon_{t,a_t} = \sum_{k=1}^{K} \sum_{m=1}^{n_k} \epsilon_{t,m},k.$$

Then the empirical regret can also be written as $\hat{R}_{\theta,\sigma}^\pi(T) = R_{\theta,\sigma}^\pi(T) - N_{\pi}(T)$. We note that for most cases considered in this paper, the environment admits $\sigma$-subgaussian noises by default, and we will write $\hat{R}_{\theta}^\pi(T)$ instead of $\hat{R}_{\theta,\sigma}^\pi(T)$ and $R_{\theta}^\pi(T)$ instead of $R_{\theta,\sigma}^\pi(T)$ unless otherwise specified.

The following simple lemma gives the mean and the tail probability of the genuine noise $N_{\pi}(T)$. Intuitively, it shows when bounding the mean or the tail probability of the empirical regret, one only need to consider the pseudo regret term. We will make it more precise when we discuss the proof of main theorems.

**Lemma 1.** We have $\mathbb{E}[N_{\pi}(T)] = 0$ and

$$\max \{\mathbb{P}(N_{\pi}(T) \geq x), \mathbb{P}(N_{\pi}(T) \leq -x)\} \leq \exp\left(\frac{-x^2}{2\sigma^2 T}\right).$$
2.1. Light-tailed Risk, Instance-dependent Consistency, Worst-case Optimality

Now we describe concepts that are needed to formalize the policy design and analysis.

1. Light-tailed risk. A policy is called light-tailed, if for any constant $c > 0$, there exists some $\beta > 0$ and constant $C > 0$ such that

$$\limsup_{T \to +\infty} \ln \left\{ \sup_{\theta} \mathbb{P} \left( \hat{R}_\theta^* (T) > cT \right) \right\} \leq -C.$$  

Note that here, we allow $\beta$ and $C$ to be dependent on $c$. In brief, a policy has light-tailed risk if the probability of incurring a linear regret can be bounded by an exponential term of polynomial $T$:

$$\sup_{\theta} \mathbb{P} (\hat{R}_\theta^* (T) \geq cT) = \exp(\Omega(T^\beta))$$

for some $\beta > 0$. If a policy is not light-tailed, we say it is heavy-tailed.

Remark 1. We clarify that conventionally, a distribution is called “lighted-tailed” when its moment generating function is finite around a neighborhood of zero. Our definition of “light-tailed” emphasizes the boundary between heavy and light to separate polynomial rate of decaying versus exponential-polynomial rate of decaying, which is aligned with but technically different from the conventional definition of “lighted-tailed”. For example, for regret random variables $R(T)$ indexed by $T$, when $T$ is large, if $\mathbb{P}(R(T) > T/2) \sim T^{-\beta}$ for some positive $\beta$, then its distribution is heavy-tailed in both our definition and the conventional definition. If $\mathbb{P}(R(T) > T/2) \sim \exp(-T^{\beta})$ for $\beta \in (0, 1)$, then its distribution is lighted-tailed in our definition and is heavy-tailed in the conventional definition. If $\mathbb{P}(R(T) > T/2) \sim \exp(-T^{\beta})$ for $\beta \geq 1$, then its distribution is lighted-tailed in both our definition and the conventional definition. Therefore, when we claim safety against heavy-tailed risk, it indicates tail distribution that is lighter than any polynomial rate of decay.

2. Instance-dependent consistency. A policy is called consistent or instance-dependent consistent, if for any underlying true mean vector $\theta$, the policy has that

$$\limsup_{T \to +\infty} \frac{\mathbb{E} \left[ \hat{R}_\theta^* (T) \right]}{T^\beta} = 0$$

holds for any $\beta > 0$. In brief, a policy is consistent if the expected regret can never be polynomially growing in $T$ for any fixed instance.

3. Worst-case optimality. A policy is said to be worst-case optimal, if for any $\beta > 0$, the policy has that

$$\limsup_{T \to +\infty} \frac{\sup_{\theta} \mathbb{E} \left[ \hat{R}_\theta^* (T) \right]}{T^{1/2 + \beta}} = 0.$$
In brief, a policy is worst-case optimal if the worst-case expected regret can never be growing in a polynomial rate faster than $T^{1/2}$. Note that here we adopt a relaxed definition of optimality, in the sense that we do not clarify how the regret scale with the number of arms $K$ compared to that in literature. The notion of worst-case optimality in this work focuses on the dependence on $T$. For example, a policy with worst-case regret $O(poly(K)\sqrt{T} \cdot poly(lnT))$ is also optimal by our definition.

It is well known that for the stochastic MAB problem, one can design algorithms to achieve both instance-dependent consistency and worst-case optimality. Among them, two types of policies are of prominent interest: Successive Elimination (SE) and Upper Confidence Bound (UCB). We list the algorithm paradigms in Algorithm 1 and 2. The bonus term $\text{rad}(n)$ is typically set as

$$\text{rad}(n) = \sigma \sqrt{\frac{\eta \ln T}{n}}$$

with $\eta > 0$ being some tuning parameter.

**Algorithm 1 Successive Elimination**

1: $\mathcal{A} = [K]$. $t \leftarrow 0$.
2: while $t < T$ do
3: \hspace{1em} Pull each arm in $\mathcal{A}$ once. $t \leftarrow t + |\mathcal{A}|$.
4: \hspace{1em} Eliminate any $k \in \mathcal{A}$ from $\mathcal{A}$ if
5: \hspace{1em} \exists k': \hat{\mu}_{t,k'} - \text{rad}(n_{t,k'}) > \hat{\mu}_{t,k} + \text{rad}(n_{t,k})$

5: end while

**Algorithm 2 Upper Confidence Bound**

1: $\mathcal{A} = [K]$. $t \leftarrow 0$.
2: while $t < T$ do
3: \hspace{1em} $t \leftarrow t + 1$.
4: \hspace{1em} Pull the arm with the highest UCB: $\arg \max_k \{\hat{\mu}_{t-1,k} + \text{rad}(n_{t-1,k})\}$.
5: end while

SE maintains an active action set, and for each arm in the action set, it maintains a confidence interval. After pulling each arm in the action set, SE updates the action set by eliminating any arm whose confidence interval is strictly dominated by others. As a comparison, UCB does not shrink the active action set: it always pulls the arm with the highest upper confidence bound. These two algorithms share similar structure, in the sense that they both utilize confidence intervals to guide the actions.

### 3. The Basic Case: Two-armed Bandit

We start from the simple two-armed bandit setting. The general multi-armed setting is deferred to the next section. We first show that all policies that enjoy instant-dependent consistency (to be formally defined, which encompasses a range of widely discussed policies in the literature) are heavy-tailed in terms of regret distribution. The result reveals an incompatibility between instant-dependent consistency and light-tailed risk. Then we show how to add a simple twist to standard confidence bound based policies to obtain light-tailed risk. Moreover, we show that our design leads to an optimal tail decaying rate for all policies that enjoy worst-case optimal order of expected regret.
3.1. Instance-dependent Consistency Causes Heavy-tailed Risk

**Theorem 1.** If a policy \( \pi \) is instance-dependent consistent, then it can never be light-tailed. Moreover, if \( \pi \) satisfies

\[
\limsup_{T \to +\infty} \frac{E\left[\hat{R}_{\pi,\theta}(T)\right]}{\ln T} < +\infty
\]

for any \( \theta \), then we have the following stronger argument. For any \( c \in (0, 1/2) \), there exists \( C_\pi > 0 \) such that

\[
\liminf_{T \to +\infty} \frac{\ln \left\{ \sup_\theta P\left(\hat{R}_{\theta,\sigma_0}(T) > cT\right)\right\}}{\ln T} \geq -C_\pi \frac{\sigma^2}{\sigma_0^2}
\]

for any \( \sigma_0 \geq \sigma \).

Theorem 1 suggests that a consistent policy must have a risk tail heavier than an exponential one. The proof of Theorem 1 adapts a change of measure argument appeared in Fan and Glynn (2022). Intuitively speaking, if we want a policy to be adaptive enough to handle different instances, then the cost we have to pay is heavy-tailed risk. Moreover, if the policy achieves \( O(\ln T) \) regret for any fixed instance \( \theta \) (the constant is typically dependent on \( \theta \)), then the probability of incurring a linear regret becomes \( \exp(-O(\ln T)) = \Omega(\text{poly}(1/T)) \). To make things worse, if such a policy that achieves instance-dependent \( O(\ln T) \) regret under \( \sigma \)-subgaussian noises is used in an environment where the true risk parameter \( \sigma_0 \) is much larger than \( \sigma \), then the probability of incurring a linear regret becomes \( \Omega(1/T^\varepsilon) \), where \( \varepsilon > 0 \) can be arbitrarily close to 0 as \( \sigma_0 \) increases. As a result, the worst-case expected regret scales almost linearly in \( T \). Our argument resonates with Theorem 1 in Ashutosh et al. (2021) and Corollary 2 in Fan and Glynn (2022). Ashutosh et al. (2021) showed that a policy cannot achieve \( O(\ln T) \) regret if the policy is always consistent for all environments regardless of the value of the risk parameter \( \sigma \), and Fan and Glynn (2022) showed that \( \pi = \text{UCB} \) optimized for i.i.d Gaussian rewards with variance \( \sigma^2 \) satisfies (3) for \( C_\pi = 1 \).

Many standard policies are known to achieve instance-dependent \( O(\ln T) \) regret. One special case is the family of confidence bound related policies (SE and UCB). From Theorem 1, the standard bonus term (1) will always lead to a tail polynomially dependent on \( T \). Another example is the Thompson Sampling (TS) policy. It has been established that \( \pi = \text{TS} \) with Beta or Gaussian priors has the property (2) (see, e.g., Theorem 1 and 2 in Agrawal and Goyal 2012, proof of Theorem 1.3 in Agrawal and Goyal 2017). Theorem 1 then suggests that (3) also holds for \( \pi = \text{TS} \).

We need to remark on the difference between Theorem 1 and high-probability bounds in the stochastic MAB literature. It has been well-established that UCB or SE with

\[
\text{rad}(n) = \sigma \sqrt{\frac{\eta \ln(1/\delta)}{n}}
\]

for
achieves $\tilde{O}(\sqrt{T \cdot \text{polylog}(T/\delta)})$ regret with probability at least $1 - \delta$ (see, e.g., Section 1.3 in Slivkins et al. (2019), Section 7.1 in Lattimore and Szepesvári (2020)). Such design also leads to a consistent policy. However, the bound holds only for fixed $\delta$. In fact, one can see that the bonus design is dependent on the confidence parameter $\delta$. If $\delta = \exp(-\Omega(T^\beta))$ with $\beta > 0$, then the scaling speed of the regret with respect to $T$ can only be greater than $1/2$, which is sub-optimal. As a comparison, in our problem, ideally we seek to find a single policy such that it achieves $\tilde{O}(\sqrt{T \cdot \text{polylog}(T/\delta)})$ regret for any $\delta > 0$.

Up till now, we have made two observations. First, from standard stochastic MAB results, consistency and optimality can hold simultaneously. Second, from Theorem 1, consistency and light-tailed risk are always incompatible. Then a natural question arises: Can we design a policy that enjoys both optimality and light-tailed risk? If we can, then can we make the tail risk decay with $T$ in an optimal rate? We answer these two questions with an affirmative “yes” in the next section.

### 3.2. Worst-case Optimality Permits Light-tailed Risk

In this section, we propose a new policy design that achieves both light-tailed risk and worst-case optimality. The design is very simple. We still use the idea of confidence bounds, but instead of setting the bonus as (1), we set

$$\text{rad}(n) = \sigma \frac{\sqrt{\eta T \ln T}}{n}$$

with $\eta > 0$ being a tuning parameter. Theorem 2 gives performance guarantees for the mean and the tail probability of the empirical regret when $\pi = \text{SE}$.

**Theorem 2.** For the two-armed bandit problem, the SE policy with $\eta \geq 4$ and the bonus term being (4) satisfies the following two properties.

1. $\sup_\theta \mathbb{E}[\hat{R}_\pi(T)] = O(\sqrt{T \ln T})$.
2. For any $c > 0$ and any $\alpha \in (1/2, 1]$, we have

$$\sup_\theta \mathbb{P}(\hat{R}_\pi(T) \geq cT^\alpha) = \exp(-\Omega(T^{\alpha - 1/2})).$$

The first item in Theorem 2 means that with the modified bonus term, the worst case regret is still bounded by $O(\sqrt{T \ln T})$, which is the same as the regret bounds for SE and UCB with the standard bonus term (1). The second item shows that the tail probability of incurring a $\Omega(T^\alpha)$ regret ($\alpha > 1/2$) is exponentially decaying in $\Omega(T^{\alpha - 1/2})$, and thus the policy is light-tailed. The detailed proof of Theorem 2 is provided in the supplementary material. The illustrative road-map of the proof is delegated to Section 4, where we provide the proof idea for Theorem 4 that is a strict generalization of Theorem 2. In Theorem 4, we also demonstrate that the new bonus design
Simchi-Levi, Zheng and Zhu: Optimal Stochastic Bandit Policies with Light-tailed Risk

(4) allows robustness and gives exponential decaying tail risk even under an misspecified deviation parameter, avoiding (3). Here, we give some intuition on the new bonus design. Our new bonus term inflates the standard one by a factor of $\sqrt{T/n}$. This means our policy is more conservative than the traditional confidence bound methods, especially at the beginning. In fact, one can observe that for the first $\Theta(\sqrt{T})$ time periods, our policy consistently explores between arm 1 and 2, regardless of the environment. A naturally corollary is that our policy is never “consistent”, following Theorem 1. However, the bonus term (4) decays at a faster rate on the number of pulling times $n$ compared to (1). This means as the experiment goes on, the policy leans towards exploitation. We note that this is not the same as the explore-then-commit policy, which is well-known to achieve a sub-optimal $\Theta(T^{2/3})$ order of expected regret.

The following theorem shows that the risk tail in Theorem 2 is not improvable in term of order on $T$. That is, if the policy $\pi$ is worst-case optimal, then for fixed $\alpha \in (1/2, 1]$, the exponent of $\alpha - 1/2$ is tight.

**Theorem 3.** Let $c \in (0, 1/2)$. Consider the 2-armed bandit problem with gaussian noise. Let $\pi$ be a worst-case optimal policy. That is, for any $\alpha > 1/2$,

$$\limsup_T \sup_\theta \frac{E[\tilde{R}_\theta^\pi(T)]}{T^\alpha} = 0.$$ 

Then for any $\alpha \in (1/2, 1]$,

$$\liminf_T \frac{\ln \left\{ \sup_\theta P(\tilde{R}_\theta^\pi(T) \geq cT^\alpha) \right\}}{T^\beta} = 0$$

holds for any $\beta > \alpha - 1/2$.

Theorem 3 also relies on the change of measure argument appeared in the proof of Theorem 1. However, there are two notable differences: we only have worst-case optimality rather than consistency, and the regret threshold $cT^\alpha$ is in general not linear in $T$. Therefore, we need to take care of constructing the specific “hard” instance when doing the change of measure. The detailed proof is delegated to the supplementary material.

**4. The General Case: Multi-armed Bandit**

In this section, we provide step-by-step extensions to our previous results in Section 3 to the general multi-armed bandit setting. We first give a direct extension where the bonus term is set as (4). It turns out that such bonus design only yields a $\tilde{O}(K\sqrt{T})$ expected regret, which has a linear dependence on $K$, and so we study how to achieve the optimal dependence on both $K$ and $T$ by slightly modifying the design. Finally, we relax the assumption of knowing $T$ a priori and give an any-time policy that enjoys an equivalent tail probability bound as compared to the fixed-time case.
4.1. The Direct Extension

We first provide a generalization of our previous tail probability bound in Theorem 2 from the following aspects: (a) a general $K$-armed bandit setting; (b) an analysis for UCB aside from SE; (c) a detailed characterization of the tail bound for any fixed regret threshold.

**Theorem 4.** For the $K$-armed bandit problem, both the policy $\pi = \text{SE}$ and the policy $\pi = \text{UCB}$ with

$$\text{rad}(n) = \sigma \sqrt{\frac{\eta T \ln T}{n}}$$

satisfy the following two properties.

1. If $\eta \geq 4$, then $\sup_\theta \mathbb{E} \left[ \hat{R}_\theta^\pi(T) \right] \leq 4K + 4K\sigma \sqrt{\eta T \ln T}$.

2. If $\eta > 0$, then for any $x > 0$, we have

$$\sup_\theta \mathbb{P}(\hat{R}_\theta^\pi(T) \geq x) \leq \exp \left( -\frac{x^2}{2K\sigma^2T} \right) + 2K \exp \left( -\frac{(x - 2K - 4K\sigma \sqrt{\eta T \ln T})^2}{32\sigma^2K^2T} \right) + K^2T \exp \left( -\frac{x \sqrt{\eta \ln T}}{8\sigma K \sqrt{T}} \right).$$

**Proof Idea.** We provide a road-map of proving Theorem 4. The expected regret bound is proved using standard techniques. That is, we define “the good event” to be such that the mean of each arm always lies in the confidence bounds throughout the whole time horizon. Conditioned on the good event, the regret of each arm is bounded by $O(\sqrt{T \ln T})$, and thus the total expected regret is $O(K \sqrt{T \ln T})$.

The proof of the tail bound requires additional efforts. Without loss of generality, we assume arm 1 is optimal. We first illustrate the proof for $\pi = \text{SE}$.

1. We use

$$\sup_\theta \mathbb{P}(\hat{R}_\theta^\pi(T) \geq x) \leq \mathbb{P} \left( N^\pi(T) \leq -x/\sqrt{K} \right) + \sup_\theta \mathbb{P} \left( R_\theta^\pi(T) \geq x(1 - 1/\sqrt{K}) \right)$$

The term with the genuine noise can be easily bounded using Lemma 1. We are left to bound the tail of the pseudo regret. By a union bound, we observe that

$$\mathbb{P} \left( R_\theta^\pi(T) \geq x(1 - 1/\sqrt{K}) \right) \leq \sum_{k \neq 1} \mathbb{P} \left( n_k \Delta_k \geq x/(K + \sqrt{K}) \right) \leq \sum_{k \neq 1} \mathbb{P} \left( n_k \Delta_k \geq x/(2K) \right)$$

Thus, we reduce bounding the sum of the regret incurred by different arms to bounding that by a single sub-optimal arm.

2. For any $k \neq 1$, we define

$$S_k = \{ \text{Arm 1 is not eliminated before arm } k \}.$$
With a slight abuse of notation, we let $n_0 = \lceil x/(2K \Delta_k) \rceil - 1$. Consider the case when both $n_k \Delta_k \geq x/(2K)$ and $S_k$ happen. This corresponds to the risk of spending too much time before correctly discarding a sub-optimal arm. Then arm 1 and $k$ are both not eliminated after each of them being pulled $n_0$ times. This indicates

$$\hat{\mu}_{1(n_0),1} - \frac{\sigma \sqrt{\eta T \ln T}}{n_0} \leq \hat{\mu}_{1(n_0),k} + \frac{\sigma \sqrt{\eta T \ln T}}{n_0}.$$ 

The probability of this event can be bounded using concentration of subgaussian variables, which yields the second term in the tail probability bound in Theorem 4. We note that the choice of $n_0$ is important. Also, at this step, even if we replace our new bonus term by the standard one, the bound still holds.

3. Now consider the situation when both $n_k \Delta_k \geq x/(2K)$ and $\bar{S}_k$ happen. This corresponds to the risk of wrongly discarding the optimal arm. Then after some phase $n$, the optimal arm 1 is eliminated by some arm $k'$, while arm $k$ is not eliminated. Note that $k = k'$ does not necessarily hold when $K > 2$. As a consequence, we have the following two events hold simultaneously:

$$\hat{\mu}_{1(n),k'} - \frac{\sigma \sqrt{\eta T \ln T}}{n} \geq \hat{\mu}_{1(n),1} + \frac{\sigma \sqrt{\eta T \ln T}}{n} \quad \text{and} \quad \hat{\mu}_{1(n),k} + \frac{\sigma \sqrt{\eta T \ln T}}{n} \geq \hat{\mu}_{1(n),1} + \frac{\sigma \sqrt{\eta T \ln T}}{n}.$$ 

The first event leads to

$$\text{Mean of some noise terms} \geq \frac{2\sigma \sqrt{\eta T \ln T}}{n} + \Delta_{k'} \geq \frac{x}{2K T}.$$ 

The second inequality leads to

$$\text{Mean of some noise terms} \geq \Delta_k \geq \frac{x}{2K T}.$$ 

Now comes the trick to deal with an arbitrary $n$. We bound the probabilities of the two events separately and take the minimum of the two probabilities $(\mathbb{P}(A \cap B) \leq \min\{\mathbb{P}(A), \mathbb{P}(B)\}) \ (\forall A, B))$. Then such minimum can be further bounded using the basic inequality $\min\{a, b\} \leq \sqrt{ab} \ (\forall a, b \geq 0)$. This makes the probability bounded by $\exp(-\Omega(\Delta_k \sqrt{T}))$, which yields the last term in Theorem 4. We note that at this step, the $\sqrt{T}/n$ design in our new bonus term plays a crucial role. The standard bonus term (1) does not suffice to get an exponential bound.

We next illustrate the proof for $\pi = \text{UCB}$, enlightened by our proof for $\pi = \text{SE}$. The proof is in fact simpler. We use the first step in the proof for $\pi = \text{SE}$. For fixed $k$, we also take the same $n_0 = \lceil x/(2K \Delta_k) \rceil - 1$. The difference here is that we do not need to define the event $S_k$. When arm $k$ is pulled for the $(n_0 + 1)$th time, by the design of the UCB policy, there exists some $n$ such that

$$\mu_1 + \frac{\sum_{m=1}^{n_0} \epsilon_{1(m),1} + \sigma \sqrt{\eta T \ln T}}{n} \leq \mu_k + \frac{\sum_{m=1}^{n_0} \epsilon_{k(m),k} + \sigma \sqrt{\eta T \ln T}}{n_0}.$$
Now comes the trick. The event is included by a union of two events described as follows:

\[
\sum_{m=1}^{n_0} \epsilon_{k(m), k} + \sigma \sqrt{\eta T \ln T} \geq \Delta_k / 2 \quad \text{and} \quad \exists n: \sum_{m=1}^{n_0} \epsilon_{t_1(m), 1} + \sigma \sqrt{\eta T \ln T} \leq -\Delta_k / 2.
\]

The probability of each of the two events can be bounded using similar techniques adopted when \( \pi = \text{SE} \). In fact, it is implicitly shown in our proof that UCB can yield better constants than SE. We still need to emphasize that when bounding the second event, similar to the argument for \( \pi = \text{SE} \), the choice of our new bonus term is crucial.

**Remarks.** Some remarks for Theorem 4 are as follows.

1. For the regret bound, we note that compared to the optimal \( \tilde{\Theta}(\sqrt{KT}) \) bound, we have an additional \( \sqrt{K} \) term. We should point out that the additional \( \sqrt{K} \) term is not surprising under the bonus term (4). An intuitive explanation is as follows. Compared to the bonus term (1), we widen the bonus term by a factor of \( \sqrt{T/n} \). Among the \( K-1 \) arms, there must exist an arm such that it is pulled no more than \( T/K \) times throughout the whole time horizon. That is, the bonus term of this arm is always inflated by a factor of at least \( \sqrt{K} \). The standard regret bound analysis will, as a result, lead to an additional \( \sqrt{K} \) factor compared to the optimal regret bound \( \tilde{\Theta}(\sqrt{KT}) \).

2. For the tail bound, from our proof road-map, one can see that the tail bound in Theorem 4 is also valid for the pseudo regret \( \sup_{\theta} \mathbb{P}(R_\theta^*(T) \geq x) \). To get a neat form of the tail bound, one can notice that the last term in the bound can be written as

\[
K \exp \left( -\frac{x \sqrt{\eta \ln T} - 8\sigma K \sqrt{T \ln(KT)}}{8\sigma K \sqrt{T}} \right) \leq K \exp \left( -\frac{(x - 16K\sigma \sqrt{1/\eta \cdot T \ln T}) \sqrt{\eta \ln T}}{8\sigma K \sqrt{T}} \right).
\]

Since the tail probability has a trivial upper bound of 1, the last term can be replaced by

\[
K \exp \left( -\frac{(x - 16K\sigma \sqrt{1/\eta \cdot T \ln T}) \sqrt{\eta \ln T}}{8\sigma K \sqrt{T}} \right). \quad \text{Therefore, if we let}
\]

\[
y = \frac{x - 2K - 16\sigma K \sqrt{\eta \ln T}}{8\sigma K \sqrt{T}} \quad \text{then for any } x \geq 0, \text{ we get a neat form}
\]

\[
\sup_{\theta} \mathbb{P}(R_\theta^*(T) \geq x) \leq \exp (-y^2) + K \exp (-y^2) + K \exp \left( -y \sqrt{\eta \ln T} \right) \leq 4K \exp \left( -(y^2 \wedge y \sqrt{\eta \ln T}) \right).
\]

One can observe that for any \( \eta > 0 \), our policy always yields a \( \tilde{O}(\sqrt{T}) \) expected regret (although with a constant larger than that in the first result in Theorem 4). In fact, notice that for any \( x > 0 \)

\[
\mathbb{E}[\hat{R}_\theta^*(T)] = \mathbb{E}[R_\theta^*(T)] \leq x + \mathbb{P}(R_\theta^*(T) \geq x) \cdot T.
\]
If we let \( x = 2K + C\sigma K \sqrt{(\eta \vee 1/\eta)T \ln T} \) with the absolute constant \( C \) being moderately large, then \( \mathbb{P}(\hat{R}_{\theta}(T) \geq x) \cdot T = O(1) \). As a result, the worst-case regret becomes

\[
O \left( K\sigma \sqrt{\eta \vee 1/\eta} T \ln T \right).
\]

This observation has two implications.

(a) First, our policy design is not sensitive to the parameter \( \eta \) regarding the growth rate on \( T \), as opposed to the standard UCB policy with (1), where a very small \( \eta \) can possibly make the UCB policy no longer enjoy a \( \tilde{O}(\sqrt{T}) \) worst-case regret. For completeness, we summarize this point with a proof in the supplementary material. In fact, we show that when \( \eta \) is very small, the regret for either SE or UCB is lower bounded by \( \tilde{\Omega}(T^{1-2\eta}) \), the order of which can be arbitrarily close to 1.

(b) Second, our policy design is not that sensitive to the risk parameter \( \sigma \) regarding the growth rate on \( T \). More concretely, if our policy uses a misspecified risk parameter \( \sigma' < \sigma \), then note that

\[
\sigma' \sqrt{\eta} = \sigma \sqrt{\eta \frac{\sigma'^2}{\sigma^2}} \triangleq \sigma \sqrt{\eta'},
\]

we can treat our policy as if we are using the true risk parameter \( \sigma \) but with a scaled tuning parameter \( \eta' > 0 \). The tail probability of incurring a linear regret still decays at an \( \exp(-\Omega(\sqrt{T})) \) rate, and moreover, the expected regret still grows at a \( \sqrt{T} \) rate. This is in contrast with standard UCB or TS policies, where a specified risk parameter smaller than the true one will, by Theorem 1, possibly make the expected regret scale at an order larger than 1/2 (see also, e.g., Corollary 2 in Fan and Glynn 2022).

We note that the two implications above also hold for refined policies discussed in the following sections.

4.2. Optimal Expected Regret

A natural question is whether we can improve the regret bound in Theorem 4 to \( \tilde{\Theta}(\sqrt{KT}) \) and get a probability bound of

\[
\ln \left\{ \sup_{\theta} \mathbb{P}(\hat{R}_{\theta}(T) \geq x) \right\} = -\Omega \left( \frac{x}{\sqrt{KT}} \right)
\]

for large \( x \). By slightly modifying the bonus term (4), we give a “yes” answer to this question.

**Theorem 5.** For the \( K \)-armed bandit problem, both the policy \( \pi = SE \) and the policy \( \pi = UCB \) with

\[
\text{rad}(n) = \sigma \sqrt{\frac{\ln T}{n}} \cdot \max \left\{ \sqrt{\frac{\eta_1 T}{nK}}, \sqrt{\eta_2} \right\}
\]

(6)

satisfy the following two properties.
1. If \( \eta_1, \eta_2 \geq 4 \), then \( \sup_{\theta} \mathbb{E} [R_\theta^T] \leq 4K + 8\sigma \sqrt{\max\{\eta_1, \eta_2\} KT \ln T} \).

2. If \( \eta_1 > 0, \eta_2 \geq 0 \), then for any \( x > 0 \), we have
\[
\sup_{\theta} \mathbb{P}(\hat{R}_\theta^T(T) \geq x) \leq \exp \left( -\frac{x^2}{8K \sigma^2 T} \right) + 4K \exp \left( -\frac{(x - 2K - 8\sigma (\eta_1 \vee \eta_2) KT \ln T)^2}{128\sigma^2 KT} \right) + 2K^2 T \exp \left( -\frac{(x - 2K)_{+} \ln T}{16\sigma \sqrt{KT}} \right).
\]

Proof Idea. We provide a brief road-map for proving Theorem 5. The basic idea is similar to that for Theorem 4, but directly applying the analysis will still lead to the same bound in Theorem 4. Here we emphasize the key difference in the analysis. Roughly speaking, the main challenge is to reduce the \( K \) factor into a \( \sqrt{K} \) factor. To address the challenge, we define the (random) arm set as
\[
A_0 = \{ k \neq 1 : n_k \leq 1 + T/K \}.
\]

The bound for the expected regret is then proved using standard techniques, but by considering arms in or not in \( A_0 \) separately. We stress that the standard techniques are feasible only when \( \eta_1 \) and \( \eta_2 \) are both not too small. Otherwise, it is not valid to show that the good event (the mean of each arm always lies in the confidence bounds throughout the whole time horizon) happens with high probability. To obtain better bounds when \( \eta_1 \) and \( \eta_2 \) are small, we should resort to the tail bound.

To prove the tail bound, instead of using (5), we take
\[
\mathbb{P} \left( \hat{R}_\theta^T(T) \geq x(1 - 1/2\sqrt{K}) \right) = \mathbb{P} \left( \sum_{k \in A_0} n_k \Delta_k + \sum_{k \notin A_0} n_k \Delta_k \geq x(1 - 1/2\sqrt{K}) \right) \leq \mathbb{P} \left( \sum_{k \in A_0} (n_k - 1) \Delta_k + \sum_{k \notin A_0} (n_k - 1) \Delta_k \geq x(1 - 1/2\sqrt{K}) - K \right) \leq \mathbb{P} \left( \bigcup_{k \in A_0} \left\{ n_k \Delta_k \geq \frac{x - 2K}{4K} \right\} \right) \bigcup \left( \bigcup_{k \notin A_0} \left\{ (n_k - 1) \Delta_k \geq \frac{(n_k - 1)(x - 2K)}{4T} \right\} \right) \leq \sum_{k \neq 1} \mathbb{P} \left( n_k \Delta_k \geq \frac{x - 2K}{4K}, \ k \in A_0 \right) + \mathbb{P} \left( (n_k - 1) \Delta_k \geq \frac{(n_k - 1)(x - 2K)}{4T}, \ k \notin A_0 \right) \leq \sum_{k \neq 1} \mathbb{P} \left( n_k \Delta_k \geq \frac{x - 2K}{4K}, \ k \in A_0 \right) + \sum_{k \neq 1} \mathbb{P} \left( \Delta_k \geq \frac{x - 2K}{4T}, \ k \notin A_0 \right) (7)
\]

That is, when \( k \in A_0 \), we consider the event that \( n_k \Delta_k = \Omega(x/K) \); when \( k \notin A_0 \), we consider the event that \( \Delta_k = \Omega(x/T) \). In this way, in each event we consider, \( \Delta_k \) is guaranteed to be \( \Omega(x/T) \),
and when \( k \in A_0 \), \( \Delta_k \) enjoys a possibly better lower bound. Combined with the bonus design of \( \sqrt{T/K}/n \), we can get an exponential \(-\Omega(x/\sqrt{KT})\) term for the tail probability. Detailed derivation are left to the supplementary material. We note that if we do as in (5), we can only have \( \Delta_k = \Omega(x/KT) \), yielding an exponential \(-\Omega(x/K\sqrt{T})\) term.

**Remarks.** Some remarks are as follows. For the tail bound, we can do a similar thing to get a neat form as in Theorem 4. If we let \( \eta_1 = \eta > 0 \), \( \eta_2 \in [0, \eta] \) and

\[
y = \left( x - 2K - 32\sigma \sqrt{(\eta \lor 1/\eta)KT \ln T} \right)/16\sigma \sqrt{KT},
\]

then for any \( x \geq 0 \), we have

\[
\sup_{\theta} \Pr(\hat{R}_\theta^*(T) \geq x) \leq 8K \exp \left( -y^2 \land y \sqrt{\eta \ln T} \right).
\]

There are two observations:

1. For any \( \eta > 0 \), our policy always yields a \( O(\sqrt{KT \ln T}) \) expected regret (although with a constant larger than that in the first result in Theorem 5). In fact, notice that for any \( x > 0 \)

\[
\mathbb{E}[\hat{R}_\theta^*(T)] = \mathbb{E}[R_\theta^*(T)] \leq x + \Pr(R_\theta^*(T) \geq x) \cdot T.
\]

If we let \( x = 2K + C\sigma \sqrt{(\eta \lor 1/\eta)KT \ln T} \) with the absolute constant \( C \) being moderately large, then \( \Pr(R_\theta^*(T) \geq x) \cdot T = O(1) \). As a result, the worst-case regret becomes

\[
O \left( \sigma \sqrt{(\eta \lor 1/\eta)KT \ln T} \right).
\]

We need to stress that \( \eta_2 \geq 0 \) does not have much effect in obtaining the light tail, and it is OK to take \( \eta_2 = 0 \). Nevertheless, the bonus design (6) indeed incorporates our new design with the standard one (1).

2. If we set \( \eta = 1 \) and

\[
\delta = 8K \exp \left( -(y - \sqrt{\ln T})_+ \sqrt{\ln T} \right) \geq 8K \exp \left( -y^2 \land y \sqrt{\ln T} \right)
\]

Then one can see that for any \( \delta > 0 \), with probability at least \( 1 - \delta \), the regret of our policy is bounded by

\[
O \left( \sigma \sqrt{KT} \left( \sqrt{\ln T} + \frac{\ln(8K/\delta)}{\sqrt{\ln T}} \right) \right) = O \left( \sigma \sqrt{KT} \frac{\ln(T/\delta)}{\sqrt{\ln T}} \right).
\]

This partially answers the open question in Section 17.1 of Lattimore and Szepesvári (2020) (see Section 1.2) up to a logarithmic factor.
We note that a naive way to reduce the high-probability bound in the literature under the adversarial setting (see Section 1.2) into one under the stochastic setting is as follows. For simplicity, we assume \( \sigma = 1 \). First, a simple union bound suggests that with probability at least \( 1 - 2T \exp(-C^2/2) \), all the rewards are bounded within \([-C, 1 + C]\). Then applying the known result under the adversarial setting, one knows that for any \( \delta' > 0 \), with probability at least \( 1 - \delta' - T \exp(-C^2/2) \), the regret in the stochastic setting is bounded by

\[
O \left( C \sqrt{KT \ln(K/\delta')} \right).
\]

Now let \( \delta = \delta' + 2T \exp(-C^2/2) \), then with probability at least \( 1 - \delta \) the regret is bounded by

\[
O \left( \sqrt{\ln \left( \frac{T}{\delta - \delta'} \right)} \sqrt{KT \ln(K/\delta')} \right).
\]

However, such bound has a dependence of approximately \( \ln(1/\delta)^{3/2} \) on \( \delta \), which is sub-optimal. Moreover, \( C \) is dependent on the confidence parameter \( \delta \), meaning a direct application of adversarial bandit policy requires knowing \( \delta \) in advance (which is not desirable for obtaining a light-tailed policy).

### 4.3. From Fixed-time to Any-time

Finally, we enhance the policy design to accommodate the “any-time” setting where \( T \) is not known a priori, as a more challenging setting compared to the “fixed-time” setting where \( T \) is known a priori. We design a policy for the “any-time” setting and prove that the policy enjoys an equivalently desired exponential decaying tail and optimal expected regret as in the “fixed-time” setting. That is, our any-time policy enjoys a similar tail bound in Theorem 5. In the following, we use \( \text{rad}_t(n) \) to denote the bonus term at time \( t \).

**Theorem 6.** For the \( K \)-armed bandit problem, \( \pi = UCB \) with

\[
\text{rad}_t(n) = \sigma \sqrt{\frac{n t (1 \lor \ln(Kt))}{n \sqrt{K}}}
\]

satisfies the following property: fix any \( \eta > 0 \), for any \( x > 0 \), we have

\[
\sup_{\delta} \mathbb{P}(\hat{R}_\eta^T(T) \geq x) \leq \exp \left( -\frac{x^2}{8K \sigma^2 T} \right) + 2KT^2 \exp \left( -\frac{(x - 2K - 16\sigma \sqrt{2\eta KT \ln T})^2}{512 \sigma^2 KT} \right) + 2KT^3 \exp \left( -\frac{(x - 2K + \sqrt{\eta \ln T})^2}{16 \sigma \sqrt{KT}} \right).
\]

It is clear that for any \( \eta > 0 \), the policy in Theorem 6 always yields an expected regret of

\[
O(\sigma \sqrt{(\eta \lor 1/\eta)KT \ln T}).
\]
The reason is the same as that for Theorem 5. Another remark is that Theorem 6 only involves the UCB policy. In fact, the SE policy can always fail under an any-time bonus design. This is because SE will never pull an arm if this arm was eliminated previously. Therefore, even in the basic 2-armed setting, at the beginning when \( t \) is small compared to \( T \), the behaviour of SE with \( rad_t(n) \) can be nearly as worse as that of SE with (1): it may eliminate the optimal arm with a probability heavy-tailed in \( T \). On the contrary, in UCB, arms are always activated, and so the gradually time-increasing nominator in the bonus term will take effect and prevents the optimal arm from being discarded forever.

The bonus design \( rad_t(n) \) in Theorem 6 can be approximately regarded as replacing the \( T \) term in (6) with \( t \). We use \( 1 \lor \ln(Kt) \) instead of \( \ln t \) primarily out of convenience for analysis. The basic idea of proving Theorem 6 is also similar to that for Theorem 5, but requires more delicate formulas. The main challenge here stems from the unfixed \( t \) in the bonus term. In the proof of Theorem 5, in each event we consider, \( \Delta_k = \Omega(x/T) \). However, such lower bound may not be large enough, and the tail probability of wrongly discarding the optimal arm can only be bounded by

\[
\exp\left(-\Omega(\Delta_k \sqrt{t/K})\right),
\]

which is not an informative bound with an uncontrolled \( t \). Also, it is not clear whether a \( \sqrt{\ln T} \) term can be produced in the last term of the tail bound (the probability of wrongly discarding the optimal arm) under an any-time bonus design, which is essential to obtain an expected \( O(\sqrt{T \ln T}) \) regret bound. Both issues show that we need to rectify the set \( A_0 \) and the formula (7) such that \( \Delta_k \) enjoys a possibly better bound depending on \( t_k \), and that \( \ln t_k \) is connected with \( \ln T \) in the analysis. This involves several tricks, as we will discuss next. Fix a time horizon of \( T \), we define \( t_k = t_k(n_{T,k}) \) as the last time period we pull arm \( k \), and define

\[
A_1 = \left\{ k \neq 1 : n_k \leq 1 + \frac{t_k^{3/4} T^{1/4}}{K} \right\}
\]

to replace \( A_0 \), and instead of (7), we take

\[
\mathbb{P}\left(R_{\theta}^*(T) \geq x(1 - 1/2\sqrt{K})\right)
\]

\[
= \mathbb{P}\left(\sum_{k \in A_1} n_k \Delta_k + \sum_{k \notin A_1} n_k \Delta_k \geq x(1 - 1/2\sqrt{K})\right)
\]

\[
\leq \mathbb{P}\left(\sum_{k \in A_1} (n_k - 1) \Delta_k + \sum_{k \notin A_1} (n_k - 1) \Delta_k \geq x(1 - 1/2\sqrt{K}) - K\right)
\]

\[
\leq \mathbb{P}\left(\bigcup_{k \in A_1} \left\{ (n_k - 1) \Delta_k \geq \frac{x - 2K}{4K} \right\}\bigcup \left(\bigcup_{k \notin A_1} \left\{ (n_k - 1) \Delta_k \geq \frac{(n_k - 1)(x - 2K)}{4\sqrt{t_k} T} \right\}\right)\right)
\]
\[ \leq \sum_{k \neq 1} \left( P \left( (n_k - 1) \Delta_k \geq \frac{x - 2K}{4K}, \ k \in A_1 \right) + P \left( (n_k - 1) \Delta_k \geq \frac{(n_k - 1)(x - 2K)}{4\sqrt{t_k T}}, \ k \notin A_1 \right) \right) \]
\[ \leq \sum_{k \neq 1} \left( P \left( (n_k - 1) \Delta_k \geq \frac{x - 2K}{4K}, \ k \in A_1 \right) + P \left( \Delta_k \geq \frac{x - 2K}{4\sqrt{t_k T}}, \ k \notin A_1 \right) \right) \]

The correctness of the second inequality above stems from the fact that
\[ \sum_k \frac{n_k}{\sqrt{t_k}} = O(\sqrt{T}). \]

The specific form in \( A_1 \) allows us to ensure \( \Delta_k = \Omega\left(\frac{x}{\sqrt{t_k T}}\right) \), and meanwhile derive the additional \( \sqrt{\ln T} \) factor in the the last term of the tail bound. Details are all left to the supplementary material. We also need to stress that since \( n_k \) and \( t_k \) are both random variables, when bounding the probabilities, we must use a union bound to cover through all possible pairs \((n_k, t_k)\). This is the reason why we have an additional \( T^2 \) factor before the exponential tail.

5. Numerical Illustration

In this section, we provide a brief account of numerical experiments results to illustrate our theoretical findings. We first consider a two armed-bandit problem with \( \theta = (0.2, 0.8), \sigma = 1, T = 500 \) and Gaussian noise. We test five confidence bound type policies: SE and UCB with the classical bonus design described in (1), SE\_new and UCB\_new with the proposed new bonus design in (4), and UCB\_any with the bonus design \( \text{rad}_i(n) \) in Theorem 6. We let \( \kappa \triangleq \sigma \sqrt{\eta} \). The tuning parameter \( \kappa \) has 4 choices: \( \kappa \in \{0.1, 0.2, 0.4, 0.8\} \). We also consider the TS policy assuming the mean reward of each arm \( i \) following the prior \( \mathcal{N}(0,1) \) and the sample from each arm \( i \) following \( \mathcal{N}(\theta_i, \kappa^2) \). That said, we evaluate TS under mis-specified risk parameters. For each policy and \( \kappa \), we run 5000 simulation paths and for each path we collect the cumulative reward. We provide the empirical mean for the cumulative reward in Table 1. We also plot the empirical distribution (histogram) for a policy’s cumulative reward in Figure 1.

| Policy  | \( \kappa \) | 0.1  | 0.2  | 0.4  | 0.8  |
|--------|-------------|-----|-----|-----|-----|
| SE     |             | 311.60 | 336.46 | 375.53 | 374.69 |
| UCB    |             | 349.68 | 359.68 | 377.17 | 390.23 |
| TS     |             | 351.00 | 360.71 | 377.94 | 390.32 |
| SE\_new|             | 351.00 | 360.71 | 377.94 | 390.32 |
| UCB\_new|           | 351.00 | 360.71 | 377.94 | 390.32 |
| UCB\_any|          | 351.00 | 360.71 | 377.94 | 390.32 |

Table 1: Empirical mean for the cumulative reward

Table 1 shows that, SE\_new (or UCB\_new) achieves empirical mean for the cumulative reward as high as that SE (or UCB) can achieve. The highest empirical mean for the cumulative reward
that can be achieved by SE_new (or UCB_new) with various choices of $\kappa$ is comparable to the highest empirical mean that can be achieved by SE (or UCB). We note that there is no direct implication by comparing all different algorithms at the same value of $\kappa$, because the algorithms use the parameter $\kappa$ in different ways. For example, for some value of $\kappa$, SE has a higher empirical mean for the cumulative reward compared to SE_new, whereas for some other value of $\kappa$, SE has a smaller empirical mean compared to SE_new. There is no direct implication by fixing a value of $\kappa$ and comparing different algorithms. Nevertheless, we do observe that TS performs similarly to UCB, from both Figure 1 and Table 1, and so we put our discussion on confidence bound policies. Figure 1 shows that, compared to SE, SE_new has much lower probability of incurring a low cumulative reward. The implication is that (i) in terms of the empirical mean of cumulative reward, SE_new is as good as SE; (ii) in terms of the risk of incurring a low cumulative reward, SE_new is much better (i.e., lower risk) than SE. The same implication holds analogously for the comparison between UCB_new and UCB. Indeed, one can observe that for both SE and UCB with (1) and TS, there is a significant part of distribution around 100, indicating a significant risk of incurring a linear regret, especially when $\kappa$ is small. In contrast, with the new design (4), the reward is highly concentrated for every $\kappa > 0$ with almost no tail risk of getting a low total reward. Particularly, when $\kappa = 0.1$ or $\kappa = 0.2$, UCB_new achieves both high empirical mean and light-tailed distribution.

Next, we consider a 4-armed bandit problem with $\theta = (0.2, 0.4, 0.6, 0.8)$, $\sigma = 1$, $T = 500$ and Gaussian noise. Same as in the two-armed case, we test six policies: SE, UCB, TS, SE_new, UCB_new, UCB_any. The tuning parameter has 4 choices: $\kappa \in \{0.1, 0.2, 0.4, 0.8\}$. For each policy and $\kappa$, we run 5000 simulation paths and for each path we collect the cumulative reward. We plot the empirical distribution (histogram) for a policy’s cumulative reward in Figure 2. We also report the empirical mean in Table 2.
Indeed, one can observe that for both SE and UCB with (1), there is a significant part of
distribution around 200 and 300, which means that with an non-negligible probability the two
policies always pull arm 2 or 3, incurring a linear regret. Such phenomenon becomes more significant
when $\kappa$ becomes smaller. In contrast, with the new design (4), the reward is highly concentrated for
every $\kappa > 0$. Particularly, when $\kappa = 0.1$, either SE_new or UCB_new achieves both high empirical
mean and light-tailed distribution. When $\kappa$ is relatively large, e.g., $\kappa = 0.8$, the empirical mean is
not very satisfactory. This is consistent with Theorem 4, which indicates an additional $\sqrt{K}$ factor
compared to the optimal $\tilde{O}(\sqrt{KT})$ expected regret, if $\kappa$ is not scaled by a factor of $\sqrt{K}$ as in (6).

6. Extension to Linear Bandits
In this section, we further extend our policy design to the setting of linear bandits. We show that
the simple policy design that leads to both optimality and light-tailed risk for the multi-armed
bandit setting can be naturally extended to the linear bandit setting. We briefly review the setting
of linear bandits as follows (see, e.g., Dani et al. 2008, Abbasi-Yadkori et al. 2011, for reference of
more details). In each time period $t$, the decision maker (DM) is given an action set $\mathcal{A}_t \subseteq \mathbb{R}^d$
from which the DM needs to select one action $a_t \in \mathcal{A}_t$ to take for the time period $t$. Subsequently a reward

| Policy   | $\kappa$ | 0.1   | 0.2   | 0.4   | 0.8   |
|----------|----------|-------|-------|-------|-------|
| SE       | 293.11   | 311.74| 351.81| 316.64|
| UCB      | 339.41   | 348.52| 360.26| 369.25|
| TS       | 341.05   | 349.86| 359.82| 365.26|
| SE_new   | 361.93   | 334.18| 283.69| 251.52|
| UCB_new  | 371.10   | 361.13| 339.29| 309.71|
| UCB_any  | 368.86   | 359.87| 335.68| 305.88|

Table 2  Empirical mean for the cumulative reward

Figure 2  Empirical distribution for the cumulative reward; bottom two are new proposed policies
of $r_t = \theta^\top a_t + \epsilon_{t,a_t}$ is collected, where $\theta \in \mathbb{R}^d$ is an unknown parameter and $\epsilon_{t,a_t}$ is an independent $\sigma$-subgaussian mean-zero noise. More specifically, let $H_t = \{a_1, r_1, a_1, \ldots, a_{t-1}, r_{t-1}, a_{t-1}\}$ be the history prior to time $t$. When $t = 1$, $H_1 = \emptyset$. At time $t$, the DM adopts a policy $\pi_t : H_t \mapsto a_t$ that maps the history $H_t$ to an action $a_t$, where $a_t$ may be realized from some probability distribution on $A_t$.

Adopting the standard assumptions in the linear bandits literature, we presume that $\|\theta\|_\infty \leq 1$ and $\|a\|_2 \leq 1$ for any $a \in A_t$ and any $t$. Let $a^*_t = \arg\max_{a \in A_t} \theta^\top a$. The empirical regret is defined as

$$\hat{R}_\theta(T) = \sum_{t=1}^T \theta^\top a^*_t - \sum_{t=1}^T r_{t,a_t} = \sum_{t=1}^T \theta^\top (a^*_t - a_t) - \sum_{t,a_t} \epsilon_{t,a_t} \triangleq R_\theta^*(T) - N^*(T).$$

Same as in the MAB setting, $N^*(T)$ also enjoys the fast concentration property in Lemma 1. We provide the Linear UCB policy (UCB-L) in Algorithm 3 and show in Theorem 7 an explicit exponentially decaying regret tail bound under a carefully specified bonus term $\text{rad}_t(z)$. Note that in standard bonus design, $\text{rad}_t(z) \propto \sqrt{z}$ (see, e.g., the OFUL policy in Abbasi-Yadkori et al. 2011). We need to emphasize that the additional $\sqrt{dz}$ term in (8) is necessary, without which the policy may be lack of exploration at the very beginning and then always stick to a suboptimal action for a small $\sigma$.

**Algorithm 3 Linear UCB (UCB-L)**

1: $t \leftarrow 0$, $V_0 = I$, $\hat{\theta}_0 = 0$.

2: while $t < T$ do

3: \quad $t \leftarrow t + 1$. Observe $A_t$.

4: \quad Take the action with the highest UCB:

$$a_t = \arg\max_{a \in A_t} \left\{ \hat{\theta}_{t-1}^\top a + \text{rad}_t(a^\top V_{t-1}^{-1} a) \right\}.$$

5: \quad $V_t = V_{t-1} + a_t a_t^\top$, $\hat{\theta}_t = V_{t-1}^{-1} (\sum_{s \leq t} a_s r_s)$.

6: end while

**Theorem 7.** Let $T \geq d$. For the stochastic linear bandit problem, $\pi = \text{UCB-L}$ with

$$\text{rad}_t(z) = z \sigma \sqrt{\frac{\eta T}{d}} + \sqrt{dz}$$

satisfies the following property: for any fixed $\eta > 0$, for any $x > 0$, we have

$$\sup_{\theta} \mathbb{P}(\hat{R}_\theta(T) \geq x) \leq \exp\left(-\frac{x^2}{2\sigma^2 d^2 T}\right) + 2d(T/d)^{2d+1} \exp\left(-\frac{(x - 4\sqrt{d} - 32d \sqrt{T \ln T - 16\sigma \sqrt{\eta} d T \ln T})^2}{512\sigma^2 d T \ln^2 T}\right) + 2d(T/d)^{2d+1} \exp\left(-\frac{(x - 4\sqrt{d} + \sqrt{\eta})}{8\sigma \sqrt{d T \ln T}}\right).$$
The main technical challenge to prove Theorem 7 is that the analysis for the MAB setting is not directly applicable — the estimation of the unknown parameter $\theta$ is entangled with uncontrollable arm feature vectors in the linear bandit setting. It is also not straightforward how the equation (5) can be adapted to accommodate the linear bandit setting and prevent the appearance of $K$ (the number of arms may even be infinite). The main idea to resolve these two obstacles is by noticing that $a_t^\top V_{t-1}^{-1}a_t$ can be regarded as a counterpart of $n_i$ in the MAB setting (though they are not equivalent) and that the pseudo regret $R^\pi_\theta(T)$ can be written as

$$R^\pi_\theta(T) = \sum_t \theta^\top (a_t^* - a_t) = \sum_t \frac{\theta^\top (a_t^* - a_t)}{a_t^\top V_{t-1}^{-1}a_t} \cdot a_t^\top V_{t-1}^{-1}a_t.$$ 

Another important fact is that

$$\sum_t a_t^\top V_{t-1}^{-1}a_t = O(d \ln T) = \tilde{O}(d).$$

With the observations in hand, the tail bound is then obtained by further adapting and refining the analysis for Theorem 5. Details are provided in the appendix.

Some interpretation and remarks for Theorem 7 are as follows.

1. To have a more interpretable representation of the upper bound derived in Theorem 7, we can do a similar change-of-variable trick as in Theorem 4 to get a more neat form. If we denote a variable $y$ as

$$y = \frac{\left( x - 4\sqrt{d} - 32d\sqrt{T \ln T} - 16\sqrt{\eta \sqrt{1/\eta d \sqrt{T \ln T}} T} \right)}{32\sqrt{dT \ln T}},$$

then for any $x \geq 0$, we have

$$\sup_\theta \mathbb{P}(R^\pi_\theta(T) \geq x) \leq 8d(T/d)^{2d+1} \exp \left(-y^2 \land y \sqrt{\eta}\right) \leq 8T^{3d} \exp \left(-y^2 \land y \sqrt{\eta}\right).$$

One can observe that for any $\eta > 0$, our policy always yields a $\tilde{O}(\sqrt{T})$ expected regret, achieving the optimal order. In fact, notice that for any $x > 0$

$$\mathbb{E}[\hat{R}^\pi_\theta(T)] = \mathbb{E}[R^\pi_\theta(T)] \leq x + \mathbb{P}(R^\pi_\theta(T) \geq x) \cdot \sqrt{dT}.$$ 

If we let

$$x = 4\sqrt{d} + 32d\sqrt{T \ln T} + C\sigma \sqrt{\eta \sqrt{1/\eta d \sqrt{T \ln T}}} T$$

with the absolute constant $C$ being moderately large, then $\mathbb{P}(R^\pi_\theta(T) \geq x) \cdot T = O(1)$. As a result, the worst-case expected regret becomes controlled by the order of

$$O\left(d\sqrt{T \ln T} + \sigma \sqrt{\eta \sqrt{1/\eta d \sqrt{T \ln T}}} T\right).$$
We have to point out that compared to the $O(d\sqrt{T}\ln T)$ regret in previous linear bandits literature (see, e.g., Abbasi-Yadkori et al. 2011), our result has an additional factor of $\sqrt{d\ln T}$. Even though this additional factor does not affect the optimal $\tilde{O}(\sqrt{T})$ expected regret order achieved by our policy design on the linear bandit setting, we think as future work it might be useful to see whether our analysis can be refined to improve on the additional factor.

2. We would also like to discuss a bit more about the computation aspect. Apparently, the main computation step is Line 4 in Algorithm 3, where the objective function is a convex function of $a$. When $|A_t|$ is finite and small (e.g., $|A_t| = O(\text{poly}(d))$), we can enumerate all choices for $a$. However, in general, even when $A_t$ is a convex set, maximizing a convex function can be NP-Hard. A similar discussion can be found in Dani et al. (2008), and in the future work it might be interesting to see whether the maximization problem is efficiently solvable under other special cases.

7. Conclusion

In this work, we consider the stochastic multi-armed bandit problem with a joint goal of minimizing the worst-case expected regret and obtaining light-tailed probability bound of the regret distribution. We demonstrate that light-tailed risk and instance-dependent consistency are incompatible, and show that light-tailed risk and worst-case optimality can co-exist through a simple new policy design. We also discuss generalizations of our results and show how to achieve the optimal rate dependence on both the number of arms $K$ and the time horizon $T$ with or without knowing $T$. We extend the simple and optimal policy design to the linear bandit setting.

There are several prospective future directions. Technically, one direction is to improve our policy design for linear bandits on the tail bound and the computation efficiency. Empirically, it would be interesting to see how the policy design works in various practical settings. Methodologically, it is tempting to see whether our policy design can be integrated into more complex settings such as reinforcement learning. We hope our results and analysis in this paper may bring about new insights on understanding and alleviating the tail risk of learning algorithms under a stochastic environment.

References

Abbasi-Yadkori Y, Pál D, Szepesvári C (2011) Improved algorithms for linear stochastic bandits. Advances in neural information processing systems 24.

Agrawal S, Goyal N (2012) Analysis of Thompson sampling for the multi-armed bandit problem. Conference on learning theory, 39–1 (JMLR Workshop and Conference Proceedings).

Agrawal S, Goyal N (2017) Near-optimal regret bounds for thompson sampling. Journal of the ACM (JACM) 64(5):1–24.
Araman VF, Caldentey R (2021) Diffusion approximations for a class of sequential testing problems. arXiv preprint arXiv:2102.07030.

Ashutosh K, Nair J, Kagrecha A, Jagannathan K (2021) Bandit algorithms: Letting go of logarithmic regret for statistical robustness. International Conference on Artificial Intelligence and Statistics, 622–630 (PMLR).

Audibert JY, Munos R, Szepesvári C (2009) Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. Theoretical Computer Science 410(19):1876–1902.

Auer P, Cesa-Bianchi N, Fischer P (2002) Finite-time analysis of the multiarmed bandit problem. Machine learning 47(2):235–256.

Baudry D, Gautron R, Kaufmann E, Maillard O (2021) Optimal thompson sampling strategies for support-aware cvar bandits. International Conference on Machine Learning, 716–726 (PMLR).

Bouchaud JP, Georges A (1990) Anomalous diffusion in disordered media: statistical mechanisms, models and physical applications. Physics reports 195(4-5):127–293.

Bouchaud JP, Potters M, et al. (2000) Theory of financial risks: from statistical physics to risk management (Cambridge University Press).

Bubeck S, Cesa-Bianchi N (2012) Regret analysis of stochastic and nonstochastic multi-armed bandit problems. arXiv preprint arXiv:1204.5721.

Cassel A, Mannor S, Zeevi A (2018) A general approach to multi-armed bandits under risk criteria. Conference On Learning Theory, 1295–1306 (PMLR).

Chang JQ, Tan VY (2022) A unifying theory of thompson sampling for continuous risk-averse bandits. Proc. of the 36th AAAI Conference on Artificial Intelligence. AAAI Press.

Chen N, Yang S (2022) Bridging adversarial and nonstationary multi-armed bandit. arXiv preprint arXiv:2201.01628.

Chopra S, Sodhi M (2004) Supply-chain breakdown. MIT Sloan management review 46(1):53–61.

Dani V, Hayes TP, Kakade SM (2008) Stochastic linear optimization under bandit feedback. Proceedings of the 21st Annual Conference on Learning Theory 355–366.

Embrechts P, Klüppelberg C, Mikosch T (2013) Modelling extremal events: for insurance and finance, volume 33 (Springer Science & Business Media).

Even-Dar E, Mannor S, Mansour Y, Mahadevan S (2006) Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. Journal of machine learning research 7(6).

Fan L, Glynn PW (2021) Diffusion approximations for thompson sampling. arXiv preprint arXiv:2105.09232.

Fan L, Glynn PW (2022) The fragility of optimized bandit algorithms. arXiv preprint arXiv:2109.13595.
Galichet N, Sebag M, Teytaud O (2013) Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. Asian Conference on Machine Learning, 245–260 (PMLR).

Kalvit A, Zeevi A (2021) A closer look at the worst-case behavior of multi-armed bandit algorithms. Advances in Neural Information Processing Systems 34.

Khajonchotpanya N, Xue Y, Rujeerapaiboon N (2021) A revised approach for risk-averse multi-armed bandits under cvar criterion. Operations Research Letters 49(4):465–472.

Lattimore T, Szepesvári C (2020) Bandit algorithms (Cambridge University Press).

Maillard OA (2013) Robust risk-averse stochastic multi-armed bandits. International Conference on Algorithmic Learning Theory, 218–233 (Springer).

Neu G (2015) Explore no more: Improved high-probability regret bounds for non-stochastic bandits. Advances in Neural Information Processing Systems 28.

Prashanth L, Jagannathan K, Kollà RK (2020) Concentration bounds for cvar estimation: The cases of light-tailed and heavy-tailed distributions. Proceedings of the 37th International Conference on Machine Learning, 5577–5586.

Russo DJ, Van Roy B, Kazerouni A, Osband I, Wen Z, et al. (2018) A tutorial on thompson sampling. Foundations and Trends® in Machine Learning 11(1):1–96.

Salomon A, Audibert JY (2011) Deviations of stochastic bandit regret. International Conference on Algorithmic Learning Theory, 159–173 (Springer).

Sani A, Lazaric A, Munos R (2012) Risk-aversion in multi-armed bandits. Advances in Neural Information Processing Systems 25.

Slivkins A, et al. (2019) Introduction to multi-armed bandits. Foundations and Trends® in Machine Learning 12(1-2):1–286.

Sodhi MS, Tang CS (2021) Supply chain management for extreme conditions: research opportunities. Journal of Supply Chain Management 57(1):7–16.

Tamkin A, Keramati R, Dann C, Brunskill E (2019) Distributionally-aware exploration for cvar bandits. NeurIPS 2019 Workshop on Safety and Robustness on Decision Making.

Vakili S, Zhao Q (2016) Risk-averse multi-armed bandit problems under mean-variance measure. IEEE Journal of Selected Topics in Signal Processing 10(6):1093–1111.

Wager S, Xu K (2021) Diffusion asymptotics for sequential experiments. arXiv preprint arXiv:2101.09855 .

Zhu Q, Tan V (2020) Thompson sampling algorithms for mean-variance bandits. International Conference on Machine Learning, 11599–11608 (PMLR).

Zimin A, Ibsen-Jensen R, Chatterjee K (2014) Generalized risk-aversion in stochastic multi-armed bandits. arXiv preprint arXiv:1405.0833 .
Supplementary Material

To prove Theorem 1, we need the following lemma.

**Lemma 2.** Consider the two-armed bandit problem with \( \sigma \)-Gaussian noise. Let \( \pi \) be a policy such that for any true mean vector \( \theta \),

\[
\limsup_{T \to +\infty} \frac{\mathbb{E}[\tilde{R}_\theta(T)]}{T} = 0.
\]

That is, the expected regret under \( \pi \) is always sub-linear in \( T \). Then for any \( \bar{\theta} = (\bar{\theta}_1, \bar{\theta}_2) \) where \( \bar{\theta}_1 > \bar{\theta}_2 \), and any \( \delta > 0 \), we have

\[
\limsup_{T \to +\infty} \mathbb{P}_\theta(\tilde{\mu}_{T,2} - \bar{\theta}_2) \geq \delta = 0.
\]

**Proof of Lemma 2.** Define

\[
E_T = \{ \tilde{\mu}_{T,2} - \bar{\theta}_2 \leq \delta \}.
\]

Fix any positive integer \( N \), we have

\[
\mathbb{P}_\theta^\pi(E_T) = \mathbb{P}_\theta^\pi(E_T; n_{T,2} < N) + \mathbb{P}_\theta^\pi(E_T; n_{T,2} \geq N) \\
\leq \mathbb{P}_\theta^\pi(n_{T,2} < N) + \sum_{n=N}^{+\infty} \mathbb{P}_\theta^\pi(E_T; n_{T,2} = n) \\
\leq \mathbb{P}_\theta^\pi(n_{T,2} < N) + \sum_{n=N}^{+\infty} 2\exp(-\frac{n\delta^2}{2\sigma^2}).
\]

Thus,

\[
\limsup_{T \to +\infty} \mathbb{P}_\theta^\pi(E_T) \leq \limsup_{T \to +\infty} \mathbb{P}_\theta^\pi(n_{T,2} < N) + \sum_{n=N}^{+\infty} 2\exp(-\frac{n\delta^2}{2\sigma^2})
\]

holds for any \( N \). Note that the last term converges to 0 as \( N \to +\infty \). It suffices to show \( \mathbb{P}_\theta^\pi(n_{T,2} < N) \to 0 \) as \( T \to +\infty \) for any fixed \( N \). Suppose this does not hold, then we can find \( p > 0 \) and a sequence \( \{T(m)\}_{m=1}^{+\infty} \) such that

\[
\mathbb{P}_\theta^\pi(n_{T(m),2} < N) > p.
\]

Let \( M \) be some large number such that \( q \triangleq p - N \exp\left(-\frac{M^2}{2\sigma^2}\right) > 0 \). Consider an alternative environment \( \theta = (\theta_1, \theta_2) \) where \( \theta_2 > \theta_1 = \bar{\theta}_1 \). Using the change of measure argument, we have

\[
\mathbb{P}_\theta^\pi(n_{T(m),2} < N) \\
= \mathbb{E}_\theta^\pi[1\{n_{T(m),2} < N\}] \\
= \mathbb{E}_\theta^\pi \left[ \exp \left( \sum_{n=1}^{n_{T(m),2}} \frac{(X_{t_2(n,2)} - \bar{\theta}_2)^2 - (X_{t_2(n,2)} - \theta_2)^2}{2\sigma^2} \right) 1\{n_{T(m),2} < N\} \right] \\
= \mathbb{E}_\theta^\pi \left[ \exp \left( n_{T(m),2} \frac{(\bar{\theta}_2 - \theta_2)(\theta_2 - \bar{\theta}_2)}{2\sigma^2} + \frac{(\theta_2 - \bar{\theta}_2)(\bar{\theta}_2 - \theta_2)}{\sigma^2} \right) 1\{n_{T(m),2} < N\} \right] \\
\geq \mathbb{E}_\theta^\pi \left[ \exp \left( n_{T(m),2} \frac{(\bar{\theta}_2 - \theta_2)^2}{2\sigma^2} + \frac{(\theta_2 - \bar{\theta}_2)^2}{\sigma^2} \right) 1\{\hat{\theta}_{T(m),2} > \bar{\theta}_2 - M, n_{T(m),2} < N\} \right].
\]
\[
\begin{align*}
&\geq \mathbb{E}_{\theta}^\pi \left[ \exp \left( N \left( -\frac{(\hat{\theta}_2 - \theta_2)^2}{2\sigma^2} - \frac{M(\theta_2 - \hat{\theta}_2)}{\sigma^2} \right) \right) \right] 1\{\hat{\theta}_{T(m),2} > \hat{\theta}_2 - M, n_{T(m),2} < N \} \\
&= \exp \left( N \left( -\frac{(\hat{\theta}_2 - \theta_2)^2}{2\sigma^2} - \frac{M(\theta_2 - \hat{\theta}_2)}{\sigma^2} \right) \right) \mathbb{P}_{\theta}^\pi(\hat{\theta}_{T(m),2} > \hat{\theta}_2 - M, n_{T(m),2} < N).
\end{align*}
\]

Note that
\[
\mathbb{P}_{\theta}^\pi(\hat{\theta}_{T(m),2} > \hat{\theta}_2 - M, n_{T(m),2} < N) = p - \sum_{n=1}^{N-1} \mathbb{P}_{\theta}^\pi(\hat{\theta}_{T(m),2} \leq \hat{\theta}_2 - M, n_{T(m),2} = n) \geq p - \sum_{n=1}^{N-1} \exp \left( -\frac{nM^2}{2\sigma^2} \right) \geq p - N \exp \left( -\frac{M^2}{2\sigma^2} \right) = q > 0.
\]

Therefore, there exists a constant positive probability such that \( \pi \) pulls arm 2 no more than \( N \) times under \( \theta \). As a result, \( \pi \) incurs a linear expected regret under \( \theta \), leading to a contradiction.

**Proof of Theorem 1.**

To prove the first statement, we consider the environment where the noise \( \epsilon \) is gaussian with standard deviation \( \sigma \). Let \( \theta_1 = 1/2 \). Let \( \theta = (\theta_1, \theta_2) \) and \( \hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2) \), where \( \theta_2 = \theta_1 + \frac{1}{2} \) and \( \hat{\theta}_2 = \theta_1 - \frac{1}{2} \). Let \( c' \in (c, 1/2) \).

Define
\[
E_T = \{ |\hat{\theta}_{T,2} - \hat{\theta}_2| \leq \delta \}
\]
where \( \delta > 0 \) is a small number, and
\[
F_T = \{ n_2 \leq f(T) \}.
\]

Here, \( f(T) > 0 \) is a strictly increasing function such that
\[
\limsup_T \frac{f(T)}{T} < 1 - 2c'.
\]

We will detail how \( f(T) \) should be chosen under different conditions in the last step of the proof. Then there exists \( T_0 \) such that \( f(T) < (1 - 2c')T \) for any \( T > T_0 \). Now we fix any \( T > T_0 \). Under the environment \( \hat{\theta} \), we have
\[
\mathbb{P}_{\theta}^\pi(\hat{F}_T) = \mathbb{P}_{\theta}^\pi(n_2 > f(T)) \leq \frac{\mathbb{E}_{\theta}^\pi[n_2]}{f(T)} \leq \frac{2\mathbb{E}[R^\pi_\theta(T)]}{f(T)} = \frac{2\mathbb{E}[[\hat{R}^\pi_\theta(T)]]}{f(T)}.
\]

Combined with the weak law of large numbers, we have
\[
\liminf_T \mathbb{P}_{\theta}^\pi(E_T, F_T) \geq 1 - \limsup_T \frac{2\mathbb{E}[\hat{R}^\pi_\theta(T)]}{f(T)}.
\]

Notice that
\[
\begin{align*}
\mathbb{P} \left( \hat{R}^\pi_\theta(T) \geq c'T \right) &\geq \mathbb{P} \left( R^\pi_\theta(T) \geq c'T, -N^\pi(T) \geq -(c' - c)T \right) \\
&= \mathbb{P} \left( R^\pi_\theta(T) \geq c'T \right) - \mathbb{P} \left( R^\pi_\theta(T) \geq c'T, N^\pi(T) > (c' - c)T \right)
\end{align*}
\]
\[ \geq P (R^n_\theta (T) \geq c'T) - P (N^n (T) > (c' - c)T) \]
\[ \geq P (R^n_\theta (T) \geq c'T) - \exp \left( -\frac{(c' - c)^2T}{2\sigma^2} \right) \]

The last inequality holds from Lemma 1. Now

\[ P (R^n_\theta (T) \geq c'T) \]
\[ \geq P_\theta (n_1 \geq 2c'T) \]
\[ \geq P_\theta (n_2 \leq (1 - 2c')T) \]
\[ \geq P_\theta (n_2 \leq f(T)) \]
\[ \geq P_\theta (E_T, F_T) \]
\[ = E_\theta^\pi \left[ \mathbb{1}_{\{E_T, F_T\}} \right] \]
\[ = E_\theta^\pi \left[ \exp \left( \sum_{n=1}^{n_2} \frac{(X_{12(n),2} - \hat{\theta}_2)^2 - (X_{12(n),2} - \theta_2)^2}{2\sigma^2} \right) \mathbb{1}_{\{E_T, F_T\}} \right] \]
\[ = E_\theta^\pi \left[ \exp \left( n_2 - \frac{(\hat{\theta}_2 - \theta_2)^2}{2\sigma^2} \right) + \frac{(\theta_2 - \hat{\theta}_2)(\theta_2 - \delta)}{\sigma^2} \right) \mathbb{1}_{\{E_T, F_T\}} \]
\[ \geq E_\theta^\pi \left[ \exp \left( n_2 - \frac{(\hat{\theta}_2 - \theta_2)^2}{2\sigma^2} - \frac{\delta(\theta_2 - \hat{\theta}_2)}{\sigma^2} \right) \mathbb{1}_{\{E_T, F_T\}} \right] \]
\[ = E_\theta^\pi \left[ \exp \left( f(T) - \frac{(\hat{\theta}_2 - \theta_2)^2}{2\sigma^2} - \frac{\delta(\theta_2 - \hat{\theta}_2)}{\sigma^2} \right) \mathbb{1}_{\{E_T, F_T\}} \right] \]
\[ = \exp (-f(T)(1/2\sigma^2 + \delta/\sigma^2))P_\theta^\pi (E_T, F_T). \]

Therefore,

\[ \liminf_T \ln \left\{ \sup_{\omega'} P \left( \hat{R}_\theta^\omega (T) \geq c'T \right) \right\} \]
\[ \geq \liminf_T \ln \left\{ \exp (-f(T)(1/2\sigma^2 + \delta/\sigma^2))P_\theta^\pi (E_T, F_T) - \exp \left( -\frac{(c' - c)^2T}{2\sigma^2} \right) \right\} \] (10)

Now assume that \( \pi \) is consistent. Then we set \( f(T) = T^\beta \) with \( \beta \in (0, 1) \). From (9), we have

\[ \liminf_T \mathbb{P}_\theta (E_T, F_T) \geq 1 - \limsup_T \frac{2E[\hat{R}_\theta^\pi (T)]}{T^\beta} = 1. \]

Then from (10), we have

\[ \liminf_T \ln \left\{ \sup_{\omega'} \mathbb{P}(\hat{R}_\theta^\omega (T) \geq c'T) \right\} \]
\[ \geq \frac{-1/2\sigma^2 + \delta/\sigma^2}{T^\beta}. \]

Since \( \delta > 0 \) is arbitrary, we have

\[ \liminf_T \frac{\ln \left\{ \sup_{\omega'} \mathbb{P}(\hat{R}_\theta^\omega (T) \geq c'T) \right\}}{T^\beta} \geq -1/2\sigma^2. \]
Note again that $\beta > 0$ is arbitrary. Now let $0 < \beta' < \beta$, we have
\[
\liminf_t \frac{\ln \left( \sup_{\beta'} \mathbb{P}(\tilde{R}_\pi^e(T) \geq cT) \right)}{T^{\beta'}} = \liminf_t \frac{\ln \left( \sup_{\beta'} \mathbb{P}(\tilde{R}_\pi^e(T) \geq cT) \right)}{T^{\beta'}} \cdot \liminf_t T^{\beta' - \beta} \geq -1/2\sigma^2 \cdot 0 = 0.
\]

To prove (3), we consider two environments. The first environment is $\theta = (1/2, 1)$ where the noise $\epsilon$ is gaussian with standard deviation $\sigma$ for the first arm and $\sigma_0 > \sigma$ for the second arm. The second environment is $\bar{\theta} = (1/2, 0)$ where the noise $\epsilon$ is gaussian with standard deviation $\sigma$. To simplify notations, we will write $\tilde{R}_\pi^e(T)$ instead of $\tilde{R}_{\bar{\theta}, \sigma_0}(T)$ and $R_\pi^e(T)$ instead of $R_{\bar{\theta}, \sigma_0}(T)$. One need to keep in mind that under the environment with mean vector $\theta$, the second arm has a larger deviation. We assume that $\pi$ satisfies
\[
\limsup_t \frac{\mathbb{E}[\tilde{R}_\pi^e(T)]}{\ln T} = c_\pi \sigma^2 < +\infty.
\]

Let $c' \in (c, 1/2)$. Define
\[
E_T = \left\{ |\hat{\theta}_{T,2} - \tilde{\theta}_2| \leq \delta \right\}
\]
where $\delta > 0$ is a small number, and
\[
F_T = \{ n_2 \leq 4c_\pi \sigma^2 \ln T \}.
\]

Then there exists $T_0$ such that $4c_\pi \sigma^2 \ln T < (1 - 2c')T$ for any $T > T_0$. Now we fix any $T > T_0$. Under the environment $\bar{\theta}$, we have
\[
\mathbb{P}_\bar{\theta}(\tilde{F}_T) = \mathbb{P}_\bar{\theta}(n_2 > 4c_\pi \sigma^2 \ln T) \leq \frac{\mathbb{E}_\bar{\theta}[n_2]}{4c_\pi \sigma^2 \ln T} \leq \frac{2\mathbb{E}[R_\pi^e(T)]}{4c_\pi \sigma^2 \ln T} = \frac{2\mathbb{E}[\tilde{R}_\pi^e(T)]}{4c_\pi \sigma^2 \ln T}.
\]

Combined with the weak law of large numbers, we have
\[
\liminf_T \mathbb{P}_\bar{\theta}(E_T, F_T) \geq 1 - \limsup_T \frac{2\mathbb{E}[\tilde{R}_\pi^e(T)]}{4c_\pi \sigma^2 \ln T} \geq 1/2.
\]

Notice that
\[
\begin{align*}
\mathbb{P}(\tilde{R}_\pi^e(T) \geq cT) & \geq \mathbb{P}(R_\pi^e(T) \geq cT, N^e(T) \geq -(c' - c)T) \\
& = \mathbb{P}(R_\pi^e(T) \geq cT) - \mathbb{P}(R_\pi^e(T) \geq cT, N^e(T) > (c' - c)T) \\
& \geq \mathbb{P}(R_\pi^e(T) \geq cT) - \mathbb{P}(N^e(T) > (c' - c)T) \\
& \geq \mathbb{P}(R_\pi^e(T) \geq cT) - \exp \left( -\frac{(c' - c)^2 T}{2\sigma^2} \right)
\end{align*}
\]

The last inequality holds from Lemma 1. Now
\[
\begin{align*}
\mathbb{P}(R_\pi^e(T) \geq cT) & \geq \mathbb{P}_\pi^e(n_1 \geq 2c'T) \\
& \geq \mathbb{P}_\pi^e(n_2 \leq (1 - 2c')T) \\
& \geq \mathbb{P}_\pi^e(n_2 \leq 4c_\pi \sigma^2 \ln T) \\
& \geq \mathbb{P}_\pi^e(E_T, F_T)
\end{align*}
\]
\[ \begin{aligned}
\theta &> \frac{\theta_2 - \theta_2}{2\sigma_0^2} + \frac{\theta_2 - \theta_2}{\sigma_0^2} ) \right) \{ E_T F_T \} \\
&= \frac{\exp \left( \frac{1}{2} \sum_{n=1}^{n_2} \left( X_{t_2(n),2} - \hat{\theta}_2 \right)^2 - \frac{\left( X_{t_2(n),2} - \theta_2 \right)^2}{2\sigma_0^2} \right) \right) \{ E_T F_T \} \\
&= \frac{\exp \left( \frac{1}{2} \sum_{n=1}^{n_2} \left( X_{t_2(n),2} - \theta_2 \right)^2 - \frac{\left( X_{t_2(n),2} - \theta_2 \right)^2}{2\sigma_0^2} \right) \right) \{ E_T F_T \} \\
&= \frac{\exp \left( \frac{1}{2} \sum_{n=1}^{n_2} \left( X_{t_2(n),2} - \theta_2 \right)^2 - \frac{\left( X_{t_2(n),2} - \theta_2 \right)^2}{2\sigma_0^2} \right) \right) \{ E_T F_T \} \\
&= \text{exp}(-4c_\pi \sigma^2 \ln T \left( 1/2 \sigma_0^2 + \frac{\delta}{\sigma_0^2} \right) P_{\theta}^\pi (E_T, F_T).
\end{aligned} \]

Therefore,
\[
\liminf_T \frac{\ln \left\{ \sup_{\theta'} \mathbb{P} \left( \hat{R}_{\theta',\sigma_0}^\pi (T) \geq cT \right) \right\}}{\ln T} \geq \liminf_T \frac{\ln \left\{ \exp(-4c_\pi \sigma^2 \ln T \left( 1/2 \sigma_0^2 + \frac{\delta}{\sigma_0^2} \right) P_{\theta}^\pi (E_T, F_T) - \exp \left( - \left( \frac{c' - c}{2\sigma^2} \right) T \right) \right\}}{\ln T} \geq -(2c_\pi \sigma^2 / \sigma_0^2 + 4c_\pi \delta / \sigma_0^2).
\]

Since \( \delta > 0 \) is arbitrary, we have
\[
\liminf_T \frac{\ln \left\{ \sup_{\theta'} \mathbb{P} \left( \hat{R}_{\theta',\sigma_0}^\pi (T) \geq cT \right) \right\}}{\ln T} \geq -2c_\pi \sigma^2 / \sigma_0^2.
\]

Let \( C_\pi = 2c_\pi \), we have
\[
\liminf_T \frac{\ln \left\{ \sup_{\theta'} \mathbb{P} \left( \hat{R}_{\theta',\sigma_0}^\pi (T) \geq cT \right) \right\}}{\ln T} \geq -C_\pi \sigma^2 / \sigma_0^2.
\]

**Proof of Theorem 2.** Without loss of generality, we assume \( \theta_1 > \theta_2 \). We prove the results one by one. Since the environment \( \theta \) is fixed, we will write \( \mathbb{P} (E) \) instead of \( \mathbb{P}_{\theta}^\pi (E_T) \).

1. From Lemma 1,
\[ E[\hat{R}_{\theta}^\pi (T)] = E[R_{\theta}^\pi (T)] = E[n_2] \cdot \Delta_2. \]

Let \( G \) be the event such that
\[ G = \{ \mu_k \in CI_{t,k}, \ \forall (t,k) \}. \]

Then
\[ \mathbb{P}(\bar{G}) \leq \sum_{(t,k)} \mathbb{P}(\mu_k \notin CI_{t,k}) \leq 2 \sum_{n=1}^{T} 2 \exp(-2\eta T \ln T/n) \leq 4T^{1-2\eta}. \]
Thus,
\[
E[n_2] = E[n_2|G]P(G) + E[n_2|\bar{G}]P(\bar{G}) \\
\leq E[n_2|G] + T \cdot P(\bar{G}) \\
\leq E[n_2|G] + 4T^{2-2\eta} \\
\leq E[n_2|G] + 4.
\]

With a slight abuse of notation, we let \( t \) be the largest time period such that arm 2 is pulled but subsequently not eliminated from \( A \). Then under \( G \), we have
\[
\mu_1 - 2\sigma \frac{\sqrt{\eta T \ln T}}{n_{t,2} - 1} \leq \mu_2 + 2\sigma \frac{\sqrt{\eta T \ln T}}{n_{t,2} - 1}.
\]
Therefore,
\[
n_{t,2} \leq 1 + 4\sigma \frac{\sqrt{\eta T \ln T}}{\Delta_2}
\]
and thus,
\[
n_2 \leq 2 + 4\sigma \frac{\sqrt{\eta T \ln T}}{\Delta_2}.
\]
As a result,
\[
E[R^\alpha_2(T)] \leq 2\Delta_2 + 4\sigma \sqrt{\eta T \ln T} + 4 = O(\sqrt{T \ln T}).
\]

2. We have
\[
P(\bar{R}^\alpha_2(T) \geq cT^\alpha) \leq P(R^\alpha_2(T) \geq cT^\alpha/2) + P(N^\alpha_2(T) \leq -cT^\alpha/2)
\]
From Lemma 1, the second term can be bounded as
\[
P(N^\alpha_2(T) \leq -cT^\alpha/2) \leq \exp \left( -\frac{c^2T^{2\alpha}}{2\sigma^2T} \right) \leq \exp \left( -\frac{c^2T^{2\alpha-1}}{2\sigma^2} \right).
\]
(11)
We are left to bound \( P(R^\alpha_2(T) \geq cT^\alpha/2) \). Let \( S \) be the event defined as
\[
S = \{ \text{Arm 1 is never eliminated throughout the whole time horizon} \}.
\]
Then
\[
\bar{S} = \{ \exists t \text{ such that arm 1 is eliminated at time } t \}.
\]
So
\[
P(R^\alpha_2(T) \geq cT^\alpha/2) = P(R^\alpha_2(T) \geq cT^\alpha/2, S) + P(R^\alpha_2(T) \geq cT^\alpha/2, \bar{S}).
\]
Let \( T \) be such that
\[
cT^\alpha \geq \max\{4, 16\sigma \sqrt{\eta T \ln T}\}.
\]
Let \( n_0 = \lceil cT^\alpha/2\Delta_2 \rceil - 1 \), then
\[
n_0 \geq cT^\alpha/4\Delta_2.
\]
Also, if \(R^n_0(T) \geq cT^n_0/2\), we must have
\[
T \geq cT^n_0/2\Delta_2,
\]
which means \(\Delta_2 \geq cT^{n-1}_0/2\). We have
\[
\mathbb{P}(R^n_0(T) \geq cT^n_0/2, S) \\
= \mathbb{P}(n_2 \geq cT^n_0/2\Delta_2, S) \\
\leq \mathbb{P}(n_2 \geq n_0 + 1, \text{ arm 1 and 2 are pulled in turn for } (n_0 + 1) \text{ times}) \\
\leq \mathbb{P}(\text{arm 1 and 2 are pulled in turn for } n_0 \text{ times and arm 1 and 2 are both not eliminated}) \\
\leq \mathbb{P}\left(\mu_1 - \frac{\sum_{m=1}^{n_0} \epsilon_{t_1(m),1} + \sigma\sqrt{T\ln T}}{n_0} \leq \mu_2 + \frac{\sum_{m=1}^{n_0} \epsilon_{t_2(m),2} + \sigma\sqrt{T\ln T}}{n_0}\right) \\
= \mathbb{P}\left(\sum_{m=1}^{n_0} \epsilon_{t_1(m),1} \geq \Delta_2 - \frac{2\sigma\sqrt{T\ln T}}{n_0}\right) + \mathbb{P}\left(\sum_{m=1}^{n_0} \epsilon_{t_2(m),2} \geq \Delta_2 - \frac{2\sigma\sqrt{T\ln T}}{n_0}\right) \\
\leq 2\exp\left(-n_0 \left(\frac{\Delta_2}{2} - \frac{\sigma\sqrt{T\ln T}}{n_0}\right)^2 / 2\sigma^2\right) \\
= 2\exp\left(-n_0 \Delta_2^2 \left(1 - \frac{8\sigma\sqrt{T\ln T}}{cT^n_0}\right)^2 / 2\sigma^2\right) \\
\leq 2\exp\left(-\frac{n_0 \Delta_2^2}{8\sigma^2}\right) \\
\leq 2\exp\left(-\frac{cT^n_0 \cdot cT^{n-1}}{128\sigma^2}\right) \\
= 2\exp\left(-\frac{c^2T^{2n-1}}{128\sigma^2}\right). \tag{12}\]

Meanwhile,
\[
\mathbb{P}(R^n_0(T) \geq cT^n_0/2, S) \\
\leq \mathbb{P}\left(\exists n \leq T/2 : \tilde{\mu}_{t_1(n)} + \frac{\sigma\sqrt{T\ln T}}{n} < \tilde{\mu}_{t_2(n),2} - \frac{\sigma\sqrt{T\ln T}}{n}\right) \\
= \mathbb{P}\left(\exists n \leq T/2 : \mu_1 + \frac{\sum_{m=1}^{n} \epsilon_{t_1(m),1} + \sigma\sqrt{T\ln T}}{n} < \mu_2 + \frac{\sum_{m=1}^{n} \epsilon_{t_2(m),2} - \sigma\sqrt{T\ln T}}{n}\right) \\
\leq \sum_{n=1}^{\lceil T/2 \rceil} \mathbb{P}\left(\sum_{m=1}^{n} \epsilon_{t_1(m),1} - \frac{\epsilon_{t_1(m),1}}{n} \geq \Delta_2 + \frac{2\sigma\sqrt{T\ln T}}{n}\right) \\
\leq \sum_{n=1}^{\lceil T/2 \rceil} \mathbb{P}\left(\sum_{m=1}^{n} \epsilon_{t_2(m),2} - \frac{\epsilon_{t_1(m),1}}{n} \geq \Delta_2 + \frac{\sigma\sqrt{T\ln T}}{n}\right) + \mathbb{P}\left(\sum_{m=1}^{n} \epsilon_{t_1(m),1} \frac{\epsilon_{t_1(m),1}}{n} \geq \Delta_2 + \frac{\sigma\sqrt{T\ln T}}{n}\right) \\
\leq 2\sum_{n=1}^{\lceil T/2 \rceil} \exp\left(-n \left(\frac{\Delta_2}{2} + \frac{\sigma\sqrt{T\ln T}}{n}\right)^2 / 2\sigma^2\right) \\
\leq T \exp\left(-2n \frac{\Delta_2 \sigma\sqrt{T\ln T}}{n} / 2\sigma^2\right) \\
\leq T \exp(-\sigma \cdot cT^{n-1} \cdot \sqrt{T\ln T} / 2\sigma^2)
Then under the environment \( \gamma \)

\[
\alpha > \frac{1}{2} \quad \text{always exists because} \quad \frac{1}{2} - \frac{1}{2} \gamma = 3/2 - \alpha \quad \text{and} \quad 3/2 - \alpha < 3/2 - 1/2 = 1.
\]

For notational simplicity, we will write \( \theta \) (\( \tilde{\theta} \)) instead of \( \theta(T) \) (\( \tilde{\theta}(T) \)), but we must keep in mind that \( \theta \) (\( \tilde{\theta} \)) is dependent on \( T \). Define

\[
E_T = \{ |\tilde{\mu}_{T,2} - \tilde{\theta}_2| \leq \delta \}
\]

where \( \delta > 0 \) is a small number, and

\[
F_T = \{ n_2 \leq T' \}.
\]

Then under the environment \( \tilde{\theta} \), we have

\[
P^\gamma_{\tilde{\theta}}(F_T) = P^\gamma_{\tilde{\theta}}(n_2 > T') \leq \frac{E^\gamma_{\tilde{\theta}}[n_2]}{T'} \leq \frac{E[R^\gamma_\tilde{\theta}(T)]}{T'^{\gamma + \alpha - 1}} \leq \frac{\sup_\theta E[R^\gamma_\theta(T)]}{T'^{\gamma + \alpha - 1}} \to 0
\]

as \( T \to +\infty \). Combined with the weak law of large numbers, we have

\[
\liminf_T P^\gamma_{\tilde{\theta}}(E_T, F_T) = 1.
\]

Let \( c' \in (c, 1/2) \). There exists \( T_0 \) such that \( (1 - 2c')T > T_0 \) for any \( T > T_0 \). Fix \( T > T_0 \). Notice that

\[
P \left( R^\gamma_\tilde{\theta}(T) \geq c'T^\gamma \right)
\]

\[
\geq P \left( R^\gamma_\tilde{\theta}(T) \geq c'T^\gamma, N^\gamma(T) \leq -(c' - c)T^\gamma \right)
\]

\[
= P \left( R^\gamma_\tilde{\theta}(T) \geq c'T^\gamma \right) - P \left( R^\gamma_\tilde{\theta}(T) \geq c'T^\gamma, N^\gamma(T) < -(c' - c)T^\gamma \right)
\]

\[
\geq P \left( R^\gamma_\tilde{\theta}(T) \geq c'T^\gamma \right) - \exp \left( -\frac{(c' - c)^2T^{2\alpha - 1}}{2\sigma^2} \right)
\]

The last inequality holds from Lemma 1. Now

\[
P \left( R^\gamma_\theta(T) \geq c'T^\gamma \right)
\]
\[
\begin{align*}
&\geq \mathbb{P}_\theta(n_1 \geq 2c' T) \\
&\geq \mathbb{P}_\theta(n_2 \leq (1 - 2c') T) \\
&\geq \mathbb{P}_\theta(n_2 \leq T^\gamma) \\
&\geq \mathbb{P}_\theta(E_T, F_T) \\
&= \mathbb{E}_\theta^* \left[ 1 \{ E_T F_T \} \right] \\
&= \mathbb{E}_\theta^* \left[ \exp \left( \sum_{n=1}^{n_2} (X_{i2(n)_2} - \tilde{\theta}_2)^2 - (X_{i2(n)_2} - \theta_2)^2 \right) \frac{1}{2\sigma^2} \right] 1 \{ E_T F_T \} \\
&\geq \mathbb{E}_\theta^* \left[ \exp \left( n_2 \left( \frac{\tilde{\theta}_2 - \theta_2}{2\sigma^2} + \frac{(\theta_2 - \tilde{\theta}_2) \tilde{\theta}_2}{\sigma^2} \right) \right) \frac{1}{2\sigma^2} \right] 1 \{ E_T F_T \} \\
&\geq \mathbb{E}_\theta^* \left[ \exp \left( n_2 \left( \frac{\tilde{\theta}_2 - \theta_2}{2\sigma^2} + \frac{(\theta_2 - \tilde{\theta}_2) \tilde{\theta}_2}{\sigma^2} \right) \right) \frac{1}{2\sigma^2} \right] 1 \{ E_T F_T \} \\
&\geq \mathbb{E}_\theta^* \left[ \exp \left( T^\gamma \left( - \frac{\tilde{\theta}_2 - \theta_2}{2\sigma^2} - \frac{\delta (\theta_2 - \tilde{\theta}_2)}{\sigma^2} \right) \right) \frac{1}{2\sigma^2} \right] 1 \{ E_T F_T \} \\
&= \exp(-T^\gamma + 2\alpha - 2/2\sigma^2 - \delta T^{\gamma + \alpha - 1}/\sigma^2)\mathbb{P}_\theta^* (E_T, F_T).
\end{align*}
\]

Notice that

\[ \gamma + 2\alpha - 2 < 2\alpha - 1, \quad \gamma + 2\alpha - 2 \leq \gamma + \alpha - 1, \]

and \( \delta > 0 \) can be arbitrary. Therefore,

\[ \liminf_{t} \ln \left\{ \sup_{\theta} \mathbb{P}(\hat{R}_\theta^* (T) \geq cT^\alpha) \right\} \geq \liminf_{t} \frac{-T^{\gamma + 2\alpha - 2/2\sigma^2}}{T^\beta} = 0. \]

Since \( \ln \{ \sup_{\theta} \mathbb{P}(R_\theta^* (T) \geq cT^\alpha) \} \leq 0 \) always holds, we obtain the result. \( \square \)

**Proof of Theorem 4.** Without loss of generality, we assume \( \theta_1 = \theta_* \). We prove the results one by one.

1. From Lemma 1,

\[ \mathbb{E}[R_\theta^*] = \mathbb{E}[\hat{R}_\theta^*] = \sum_{k=2}^{K} \mathbb{E}[n_k] \cdot \Delta_k. \]

Let \( G \) be the event such that

\[ G = \{ \mu_k \in \text{CI}_{t,k}, \ \forall (t,k) \}. \]

Then

\[ \mathbb{P}(\bar{G}) \leq \sum_{(t,k)} \mathbb{P}(\mu_k \notin \text{CI}_{t,k}) \leq K \sum_{n=1}^{T} 2 \exp(-\eta \ln T) \leq 2KT^{1-\eta/2}. \]

Thus,

\[ \mathbb{E}[R_\theta^*] = \sum_{k=2}^{K} (\mathbb{E}[n_k | G] \mathbb{P}(G) + \mathbb{E}[n_k | \bar{G}] \mathbb{P}(\bar{G})) \Delta_k. \]
\[
\leq \sum_{k=2}^{K} \mathbb{E}[n_k[G]] + T \cdot \mathbb{P}(G) \\
\leq \sum_{k=2}^{K} \mathbb{E}[n_k[G]] + 2KT^{2-\eta/2} \\
\leq \sum_{k=2}^{K} \mathbb{E}[n_k[G]] + 2K.
\]

(a) Let \( \pi = \text{SE} \). Fix any arm \( k \neq 1 \). We let \( t'_k \) be the largest time period such that we have traversed all the arms in \( A \), and meanwhile arm \( k \) is not eliminated from \( A \). Then \( n_k = n_{t'_k,k} + 1 \). When doing the elimination after \( t_k \), arm 1 and \( k \) are both pulled \( n_{t'_k,k} \) times. Under \( G \), we have
\[
\mu_1 - 2\sigma \sqrt{\eta T \ln T} / n_{t'_k,k} \leq \mu_1 - 2\sigma \sqrt{\eta T \ln T} / n_{t'_k,k} - 1 \leq \hat{\mu}_{t'_k,k} - \sigma \sqrt{\eta T \ln T} / n_{t'_k,k} \leq \mu_k + 2\sigma \sqrt{\eta T \ln T} / n_{t'_k,k}.
\]
Therefore,
\[
n_{t'_k,k} \leq 1 + 4\sigma \sqrt{\eta T \ln T} / \Delta_k
\]
and thus,
\[
n_k \leq 2 + 4\sigma \sqrt{\eta T \ln T} / \Delta_k.
\]
As a result,
\[
\mathbb{E}[\hat{R}^\pi_\theta(T)] \leq 2 \sum_{k=2}^{K} \Delta_k + 4 \sum_{k=2}^{K} \sigma \sqrt{\eta T \ln T} + 2K \leq 4K + 4K\sigma \sqrt{\eta T \ln T}.
\]

(b) Let \( \pi = \text{UCB} \). Fix any arm \( k \neq 1 \). We let \( t_k \) be the largest time period such that arm \( k \) is pulled. Then \( n_k = n_{t_k,k} = n_{t_{k-1},k} + 1 \). Under \( G \), we have
\[
\mu_1 \leq \hat{\mu}_{t_{k-1},1} + \sigma \sqrt{\eta T \ln T} / n_{t_{k-1},1} \leq \hat{\mu}_{t_{k-1},k} + \sigma \sqrt{\eta T \ln T} / n_{t_{k-1},k} \leq \mu_k + 2\sigma \sqrt{\eta T \ln T} / n_{t_{k-1},k}
\]
Therefore,
\[
n_{t_{k-1},k} \leq 2\sigma \sqrt{\eta T \ln T} / \Delta_k
\]
and thus,
\[
n_k \leq 1 + 2\sigma \sqrt{\eta T \ln T} / \Delta_k.
\]
As a result,
\[
\mathbb{E}[\hat{R}^\pi_\theta(T)] \leq \sum_{k=2}^{K} \Delta_k + 2 \sum_{k=2}^{K} \sigma \sqrt{\eta T \ln T} + 2K \leq 3K + 2K \sigma \sqrt{\eta T \ln T}.
\]

2. We have
\[
\mathbb{P}(\hat{R}^\pi_\theta(T) \geq x) \leq \mathbb{P} \left( R^\pi_\theta(T) \geq x(1 - 1/\sqrt{K}) \right) + \mathbb{P} \left( N^\pi_\theta(T) \leq -x/\sqrt{K} \right)
\]
From Lemma 1, the second term can be bounded as
\[
\mathbb{P} \left( N^\pi_\theta(T) \leq -x/K \right) \leq \exp \left( -\frac{x^2}{2K\sigma^2T} \right). \tag{14}
\]
We are left to bound \( \mathbb{P} \left( R^\pi_\theta(T) \geq x(1 - 1/\sqrt{K}) \right) \).
(a) Let $\pi = \text{SE}$. For any $k \neq 1$, let $S_k$ be the event defined as

$$S_k = \{\text{Arm 1 is not eliminated before arm } k\}.$$ 

Then

$$\bar{S}_k = \{\text{Arm 1 is eliminated before arm } k\}.$$ 

So

$$\mathbb{P}\left(R_k^* (T) \geq x(1 - 1/\sqrt{K})\right) \leq \sum_{k=2}^{K} \mathbb{P}\left(n_k \Delta_k \geq x/(K + \sqrt{K})\right) = \sum_{k=2}^{K} \mathbb{P}(n_k \Delta_k \geq x/2K, S_k) + \mathbb{P}(n_k \Delta_k \geq x/2K, \bar{S}_k).$$

Let $x > 0$. Fix any $k \neq 1$. With a slight abuse of notation, we let $n_0 = \lfloor x/2K\Delta_k \rfloor - 1$, then

$$n_0 \geq x/2K\Delta_k - 1 \geq (x - 2K)/2K\Delta_k.$$ 

Also, if $n_k \Delta_k \geq x/2K$, we must have

$$T \geq x/2K\Delta_k,$$

which means $\Delta_k \geq x/2KT$. By the definition of $n_0$, arm $k$ is not eliminated after being pulled $n_0$ times. So under $S_k$, after arm 1 being pulled $n_0$ times, it is still in the active set. We have

$$\mathbb{P}(n_k \Delta_k \geq x/2K, S_k) = \mathbb{P}(n_k \geq x/2K\Delta_k, S_k) \leq \mathbb{P}(\text{arm 1 and } k \text{ are both not eliminated after each of them being pulled } n_0 \text{ times})$$

$$\leq \mathbb{P}\left(\hat{\mu}_1(n_0) - \sigma\sqrt{n_0 \ln T} n_0 \geq \hat{\mu}_k(n_0) + \frac{\sigma\sqrt{n_0 \ln T} n_0}{n_0}\right)$$

$$= \mathbb{P}\left(\hat{\mu}_1 - \sum_{m=1}^{n_0} \frac{\epsilon_{t_1(m),1}}{n_0} + \frac{\sigma\sqrt{n_0 \ln T} n_0}{n_0} \leq \hat{\mu}_k + \sum_{m=1}^{n_0} \frac{\epsilon_{t_k(m),k}}{n_0} + \frac{2\sigma\sqrt{n_0 \ln T} n_0}{n_0}\right)$$

$$= \mathbb{P}\left(\sum_{m=1}^{n_0} \frac{\epsilon_{t_1(m),1} - \epsilon_{t_k(m),k}}{n_0} \geq \frac{\Delta_k}{n_0} - \frac{2\sigma\sqrt{n_0 \ln T}}{n_0}\right)$$

$$\leq 2 \exp\left(-n_0\left(\frac{\Delta_k}{2} - \frac{\sigma\sqrt{n_0 \ln T}}{n_0}\right)^2/2\sigma^2\right)$$

$$= 2 \exp\left(-n_0\Delta_k^2\left(1 - \frac{2\sigma\sqrt{n_0 \ln T}}{n_0\Delta_k}\right)^2/8\sigma^2\right)$$

$$\leq 2 \exp\left(-\frac{x(x - 2K)_+}{4KT} \left(1 - \frac{4K^2\sigma\sqrt{n_0 \ln T} x - 2K}{x - 2K}\right)^2/8\sigma^2\right)$$

$$\leq 2 \exp\left(-\frac{(x - 2K - 4K\sigma\sqrt{n_0 \ln T})^2}{32\sigma^2 K^2 T}\right).$$
In the following, we bound $P(n_k \Delta_k \geq x/2K, \bar{S}_k)$. Suppose that after $n$ phases, arm 1 is eliminated by arm $k'$ ($k'$ is not necessarily $k$). By the definition of $\bar{S}_k$, arm $k$ is not eliminated. Therefore, we have

$$
\hat{\mu}_{t_k(n),k'} - \frac{\sigma \sqrt{T \ln T}}{n} \geq \hat{\mu}_{t_1(n),1} + \frac{\sigma \sqrt{T \ln T}}{n}, \quad \hat{\mu}_{t_k(n),k} + \frac{\sigma \sqrt{T \ln T}}{n} \geq \hat{\mu}_{t_1(n),1} + \frac{\sigma \sqrt{T \ln T}}{n}
$$

(16)

holds simultaneously. The first inequality holds because arm 1 is eliminated. The second inequality holds because arm $k$ is not eliminated. Now for fixed $n$,

$$
P \left( 16 \text{ happens; } \Delta_k \geq \frac{x}{2KT} \right) \leq \mathbb{P} \left( \exists k': \hat{\mu}_{t_k(n),k'} - \frac{\sigma \sqrt{T \ln T}}{n} \geq \hat{\mu}_{t_1(n),1} + \frac{\sigma \sqrt{T \ln T}}{n} \right)
$$

$$\land \mathbb{P} \left( \hat{\mu}_{t_k(n),k} + \frac{\sigma \sqrt{T \ln T}}{n} \geq \hat{\mu}_{t_1(n),1} + \frac{\sigma \sqrt{T \ln T}}{n}; \Delta_k \geq x/2KT \right)$$

$$\leq \sum_{k' \neq 1} \mathbb{P} \left( \frac{\sum_{m=1}^{n} \left( \epsilon_{t_k(m),k'} - \epsilon_{t_1(m),1} \right)}{n} \geq \frac{2 \sigma \sqrt{T \ln T}}{n} \right) \land \mathbb{P} \left( \frac{\sum_{m=1}^{n} \epsilon_{t_1(m),1}}{n} \geq x/2KT \right)$$

$$\leq 2K \exp \left( -\frac{\eta T \ln T}{2n} \right) \land 2K \exp \left( -\frac{n x^2}{32 \sigma^2 K^2 T^2} \right)
$$

$$= 2K \exp \left( -\frac{\eta T \ln T}{2n} \lor \frac{n x^2}{32 \sigma^2 K^2 T^2} \right)
$$

$$\leq 2K \exp \left( -\frac{x \sqrt{\eta T \ln T}}{8 \sigma K \sqrt{T}} \right)
$$

Therefore,

$$P(n_k \Delta_k \geq x/2K, \bar{S})
$$

$$= P(\exists n \leq T/2: (16) \text{ happens; } n_k \Delta_k \geq x/2K)
$$

$$\leq \sum_{n=1}^{\lfloor T/2 \rfloor} P \left( (16) \text{ happens; } \Delta_k \geq \frac{x}{2KT} \right)
$$

$$\leq KT \exp \left( -\frac{x \sqrt{\eta \ln T}}{8 \sigma K \sqrt{T}} \right).$$

(17)

Note that the equations above hold for any instance $\theta$. Combining (14), (15), (17) yields

$$\sup_{\theta} P(\hat{R}_{\theta}^n(T) \geq x)
$$

$$\leq \exp \left( -\frac{x^2}{2K \sigma^2 T} \right) + 2K \exp \left( -\frac{(x - 2K - 4K \sigma \sqrt{T \ln T})^2}{32 \sigma^2 K^2 T^2} \right) + K^2 T \exp \left( -\frac{x \sqrt{\eta \ln T}}{8 \sigma K \sqrt{T}} \right)
$$

(b) Let $\pi = \text{UCB}$. From (a), we know that

$$P \left( R_{\theta}^n(T) \geq x(1 - 1/\sqrt{K}) \right) \leq \sum_{k=2}^{K} P(n_k \Delta_k \geq x/(K + \sqrt{K})) \leq \sum_{k=2}^{K} P(n_k \Delta_k \geq x/2K).$$
Let $x > 0$. Fix $k \neq 1$. With a slight abuse of notation, we let $n_0 = \lceil x/2K \Delta_k \rceil - 1$. Remember that $t_k(n_0 + 1)$ is the time period that arm $k$ is pulled for the $(n_0 + 1)$th time. We emphasize again that $\Delta_k \geq x/(2KT)$. Then

$$
P(n_k \Delta_k \geq x/2K) = P(n_k \geq x/2K \Delta_k)$$

$$\leq P \left( \mu_{t_k(n_0+1)-1,k} + \frac{\sigma \sqrt{\eta T \ln T}}{n_k(n_0+1)-1,k} \leq \mu_{t_k(n_0+1)-1,k} + \frac{\sigma \sqrt{\eta T \ln T}}{n_0} \right)$$

$$= P \left( \mu_1 + \sum_{m=1}^{n_0} \epsilon_{t_k(n_0+1)-1,k} + \frac{\sigma \sqrt{\eta T \ln T}}{n_0} \right) \leq \mu_1 + \sum_{m=1}^{n_0} \epsilon_{t_k(n_0+1)-1,k} + \frac{\sigma \sqrt{\eta T \ln T}}{n_0} \right)$$

$$\leq P \left( \sum_{m=1}^{n_0} \epsilon_{t_k(n_0+1)-1,k} + \frac{\sigma \sqrt{\eta T \ln T}}{n_0} \right) \geq \Delta_k \right)$$

$$\leq P \left( \sum_{m=1}^{n_0} \epsilon_{t_k(n_0+1)-1,k} + \frac{\sigma \sqrt{\eta T \ln T}}{n_0} \right) \geq \Delta_k \right) + \sum_{m=1}^{n_0} \epsilon_{t_k(n_0+1)-1,k} + \frac{\sigma \sqrt{\eta T \ln T}}{n_0} \right) \geq \Delta_k \right) + \sum_{m=1}^{n_0} \epsilon_{t_k(n_0+1)-1,k} + \frac{\sigma \sqrt{\eta T \ln T}}{n_0} \right) \geq \Delta_k \right) + \sum_{m=1}^{n_0} \epsilon_{t_k(n_0+1)-1,k} + \frac{\sigma \sqrt{\eta T \ln T}}{n_0} \right) \geq \Delta_k \right) + \sum_{m=1}^{n_0} \epsilon_{t_k(n_0+1)-1,k} + \frac{\sigma \sqrt{\eta T \ln T}}{n_0} \right) \geq \Delta_k \right) + \sum_{m=1}^{n_0} \epsilon_{t_k(n_0+1)-1,k} + \frac{\sigma \sqrt{\eta T \ln T}}{n_0} \right) \geq \Delta_k \right)

$$\leq \exp \left( - \frac{(x - 2K - 4K\sigma \sqrt{\eta T \ln T})^2}{2\sigma^2 K T} \right) + T \exp \left( - \frac{x \sqrt{\eta T \ln T}}{8\sigma K T} \right).$$

The last inequality holds from (15) and concentration of subgaussian variables. Note that the equations above hold for any instance $\theta$. Combining (14), (18) yields

$$\sup_{\theta} P(\tilde{R}_\theta(T) \geq x) \leq \exp \left( - \frac{x^2}{2K \sigma^2 T} \right) + K \exp \left( - \frac{(x - 2K - 4K\sigma \sqrt{\eta T \ln T})^2}{2\sigma^2 K T} \right) + K^2 T \exp \left( - \frac{x \sqrt{\eta T \ln T}}{8\sigma K T} \right).$$

Remark: SE or UCB with (1) may lead to a sub-optimal regret when $\eta$ is too small.

Let $\theta = (1, 0)$ and $\sigma = 1$ with independent Gaussian noise. We first consider $\pi = \text{SE}$. The probability that arm 1 is eliminated after the first phase is

$$P \left( \mu_{1,1} + \sqrt{\eta \ln T} < \mu_{2,2} - \sqrt{\eta \ln T} \right) = P \left( \epsilon_{1,1} - \epsilon_{2,2} < -1 - 2\sqrt{\eta \ln T} \right) \geq \frac{1}{\sqrt{2\pi}\sqrt{\eta \ln T} \left( 1 + \sqrt{\eta \ln T} \right)^2} \exp \left( - \left( 1 + \sqrt{2\eta \ln T} \right)^2 / 2 \right) \Theta \left( \frac{T^{-\eta}}{\sqrt{\ln T}} \right)$$

The inequality holds because for a standard normal variable $X$, it is established that

$$P(X > t) \geq \frac{1}{\sqrt{2\pi}} \cdot \frac{t}{1 + t^2} \exp(-t^2/2)$$

Therefore, the expected regret is at least

$$\Theta \left( \frac{T^{-\eta}}{\sqrt{\ln T}} \right) \cdot (T - 2) = \Theta \left( \frac{T^{1-\eta}}{\sqrt{\ln T}} \right).$$
If \( \eta \) is very small, then apparently the regret is sub-optimal.

Now we consider \( \pi = \text{UCB} \). The probability that arm 1 is pulled only once is

\[
P \left( \forall 2 \leq t \leq T : \hat{\mu}_{1,1} + \sqrt{\eta \ln T} < \hat{\mu}_{t,2} \right)
\geq P \left( \forall 2 \leq t \leq T : \hat{\mu}_{1,1} + \sqrt{\eta \ln T} < \hat{\mu}_{t,2} \right)
\geq P \left( \{ \hat{\mu}_{1,1} < -1 - 2\sqrt{\eta \ln T} \} \cap \left\{ \forall 1 \leq t < T : \sum_{s=2}^{t+1} \epsilon_{s,2} > -1 - t\sqrt{\eta \ln T} \right\} \right)
= P \left( \left\{ \epsilon_{1,1} < -2 - 2\sqrt{\eta \ln T} \right\} \right) \cdot \mathbb{P} \left( \left\{ \forall 1 \leq t < T : \sum_{s=2}^{t+1} \epsilon_{s,2} > -1 - t\sqrt{\eta \ln T} \right\} \right)
\]

We have

\[
P \left( \left\{ \epsilon_{1,1} < -2 - 2\sqrt{\eta \ln T} \right\} \right)
\geq \frac{1}{\sqrt{2\pi}} \frac{2 + 2\sqrt{\eta \ln T}}{1 + (2 + 2\sqrt{\eta \ln T})^2} \exp(- (2 + 2\sqrt{\eta \ln T})^2 / 2)
= \Theta \left( \frac{T^{-2\eta}}{\sqrt{\ln T}} \right)
\]

We use a martingale argument and the optional sampling theorem to bound the second probability. Define \( Z_t = \sum_{s=2}^{t+1} \epsilon_{s,2} \). Define the stopping time

\[\tau = \inf \{ Z_t \leq -1 - t\sqrt{\eta \ln T} \}\]

Then

\[
P \left( \left\{ \forall 1 \leq t < T : \sum_{s=2}^{t+1} \epsilon_{s,2} > -1 - t\sqrt{\eta \ln T} \right\} \right) = \mathbb{P} (\tau \geq T)
\]

For fixed \( T \), \( \tau \wedge (T - 1) \) is finite. Notice that

\[
\exp(-2\sqrt{\eta \ln T} Z_t - 2\eta \ln T \cdot t)
\]

is a martingale with mean 1. By the optional sampling theorem, we have

\[
1 = \mathbb{E} \left[ \exp(-2\sqrt{\eta \ln T} Z_{\tau \wedge (T - 1)} - 2\eta \ln T \cdot (\tau \wedge (T - 1))) \right]
\geq \mathbb{E} \left[ \exp(-2\sqrt{\eta \ln T} Z_{\tau} - 2\eta \ln T \cdot \tau) \mathbb{I} \{ \tau < T \} \right]
\geq \exp(2\sqrt{\eta \ln T}) \mathbb{P} (\tau < T)
\]

Therefore, the second probability is bounded by

\[
1 - \exp(-2\sqrt{\eta \ln T}).
\]

The expected regret is at least

\[
\Theta \left( \frac{T^{-2\eta}}{\sqrt{\ln T}} \right) \cdot \left( 1 - \exp(-2\sqrt{\eta \ln T}) \right) \cdot (T - 2) = \Omega \left( \frac{T^{1-2\eta}}{\sqrt{\ln T}} \right).
\]

\[\square\]

**Proof of Theorem 5.** Without loss of generality, we assume \( \theta_1 = \theta_* \). We prove the results one by one.
From Lemma 1, 
\[ \mathbb{E}[R^*_\pi] = \mathbb{E}[\hat{R}^*_\pi] = \sum_{k=2}^{K} \mathbb{E}[n_k] \cdot \Delta_k. \]

Let \( G \) be the event such that 
\[ G = \{ \mu_k \in \text{CI}_{t,k}, \ \forall (t,k) \}. \]

Then 
\[ \mathbb{P}(\bar{G}) \leq \sum_{(t,k)} \mathbb{P}(\mu_k \notin \text{CI}_{t,k}) \leq K \sum_{n=1}^{T} 2 \exp(-\frac{\eta T \ln T}{2n}) \leq 2KT^{1-\eta/2}. \]

Thus, 
\[ \mathbb{E}[R^*_\pi] = \sum_{k=2}^{K} (\mathbb{E}[n_k | G] \mathbb{P}(G) + \mathbb{E}[n_k | \bar{G}] \mathbb{P}(\bar{G})) \Delta_k \]
\[ \leq \sum_{k=2}^{K} \mathbb{E}[n_k | G] \Delta_k + T \cdot \mathbb{P}(\bar{G}) \]
\[ \leq \sum_{k=2}^{K} \mathbb{E}[n_k | G] \Delta_k + 2KT^{1-\eta/2} \]
\[ \leq \sum_{k=2}^{K} \mathbb{E}[n_k | G] \Delta_k + 2K. \]

Define the (random) arm set 
\[ \mathcal{A}_0 = \{ k \neq 1 : n_k \leq 1 + \frac{T}{K} \} \]
as the set of arms that are pulled no more than \( 1 + \frac{T}{K} \) times. Then 
\[ \mathbb{E}[R^*_\pi] \leq \mathbb{E} \left[ \sum_{k \in \mathcal{A}_0} n_k \Delta_k | G \right] + \mathbb{E} \left[ \sum_{k \notin \mathcal{A}_0} n_k \Delta_k | G \right] + 2K \]

(a) Let \( \pi = \text{SE} \). Fix any \( k \neq 1 \). We let \( t'_k \) be the largest time period such that we have traversed all the arms in \( \mathcal{A} \), and meanwhile arm \( k \) is not eliminated from \( \mathcal{A} \). Then \( n_k = n_{t'_k,k} \) or \( n_k = n_{t'_k,k} + 1 \). When doing the elimination after \( t_k \), arm 1 and \( k \) are both pulled \( n_{t'_k,k} \) times. Under \( G \), we have 
\[ \mu_1 - 2\text{rad}(n_{t'_k,k}) \leq \hat{\mu}_{t'_k,1} - \text{rad}(n_{t'_k,k}) \leq \mu_k + \text{rad}(n_{t'_k,k}) \leq \mu_k + 2\text{rad}(n_{t'_k,k}). \]

--- Fix any \( k \in \mathcal{A}_0 \). Then under \( G \), we have 
\[ \Delta_k \leq 4\text{rad}(n_{t'_k,k}) \leq 4\sigma \sqrt{\left( \eta_1 \lor \eta_2 \right) T \ln T} / \sqrt{Kn_{t'_k,k}}. \]

Therefore, 
\[ n_{t'_k,k} \leq 4\sigma \sqrt{\left( \eta_1 \lor \eta_2 \right) T \ln T} / \sqrt{K\Delta_k} \]
and thus, 
\[ n_k \leq 2 + 4\sigma \sqrt{\left( \eta_1 \lor \eta_2 \right) T \ln T} / \sqrt{K\Delta_k}. \]
As a result,
\[
\mathbb{E}\left[ \sum_{k \in A_0} n_k \Delta_k \mid G \right] \leq 2 \mathbb{E}\left[ \sum_{k \in A_0} \Delta_k \right] + 4 \mathbb{E}\left[ \sum_{k \in A_0} \sigma \sqrt{(\eta_1 \lor \eta_2)T \ln T} \right] \\
\leq 2 \mathbb{E}[|S|] + 4 \sigma \frac{\mathbb{E}[|S|]}{\sqrt{K}} \sqrt{(\eta_1 \lor \eta_2)T \ln T}.
\]

— Fix any \( k \notin A_0 \). Then under \( G \), we have
\[
\Delta_k \leq 4 \text{rad}(n_{t_k,k}') \leq 4 \sigma \frac{\sqrt{(\eta_1 \lor \eta_2) \ln T}}{\sqrt{n_{t_k,k}'}}.
\]

Therefore,
\[
\Delta_k \leq 4 \sigma \frac{\sqrt{(\eta_1 \lor \eta_2) \ln T}}{\sqrt{n_{t_k,k}'}}
\]

and thus,
\[
\Delta_k \leq 4 \sigma \frac{\sqrt{(\eta_1 \lor \eta_2) \ln T}}{\max\{n_k - 2, 0\}}.
\]

As a result,
\[
\mathbb{E}\left[ \sum_{k \notin A_0} n_k \Delta_k \mid G \right] \leq 2 \mathbb{E}\left[ \sum_{k \notin A_0} \Delta_k \right] + 4 \mathbb{E}\left[ \sum_{k \notin A_0} \sigma \sqrt{(\eta_1 \lor \eta_2) \max\{n_k - 2, 0\} \ln T} \right] \\
\leq 2(K - \mathbb{E}[|S|]) + 4 \mathbb{E}\left[ \sum_{k \notin A_0} \sigma \sqrt{(\eta_1 \lor \eta_2)n_k \ln T} \right] \\
\leq 2(K - \mathbb{E}[|S|]) + 4 \sigma \sqrt{(\eta_1 \lor \eta_2)(K - \mathbb{E}[|S|])T \ln T}.
\]

Now we have
\[
\mathbb{E}[R^n_\ast] \leq \mathbb{E}\left[ \sum_{k \in S} n_k \Delta_k \mid G \right] + \mathbb{E}\left[ \sum_{k \notin A_0} n_k \Delta_k \mid G \right] + 2K \\
\leq 4K + 4 \sigma \sqrt{T \ln T} \left( \sqrt{(\eta_1 \lor \eta_2) \frac{\mathbb{E}[|S|]}{\sqrt{K}}} + \sqrt{(\eta_1 \lor \eta_2)(K - \mathbb{E}[|S|])} \right) \\
\leq 4K + 8 \sigma \sqrt{(\eta_1 \lor \eta_2)KT \ln T}.
\]

(b) Let \( \pi = \text{UCB} \). With a slight abuse of notation, we let \( t_k \) be the largest time period such that arm \( k \) is pulled. Then \( n_k = n_{t_k,k} = 1 + n_{t_k-1,k} \). Under \( G \), we have
\[
\mu_1 \leq \hat{\mu}_1 + \text{rad}(n_{t_k-1,k}) \leq \hat{\mu}_k + \text{rad}(n_{t_k-1,k}) \leq \mu_k + 2 \text{rad}(n_{t_k-1,k}).
\]

— Fix any \( k \in A_0 \). Then under \( G \), we have
\[
\mu_1 \leq \mu_k + 2 \sigma \frac{\sqrt{(\eta_1 \lor \eta_2)T \ln T}}{\sqrt{K(n_{t_k,k} - 1)}}.
\]

Therefore,
\[
n_k = n_{t_k,k} \leq 1 + 2 \sigma \frac{\sqrt{(\eta_1 \lor \eta_2)T \ln T}}{\sqrt{K \Delta_k}}.
\]

As a result,
\[
\mathbb{E}\left[ \sum_{k \in A_0} n_k \Delta_k \mid G \right] \leq \mathbb{E}\left[ \sum_{k \in A_0} \Delta_k \right] + 2 \mathbb{E}\left[ \sum_{k \in A_0} \sigma \sqrt{(\eta_1 \lor \eta_2)T \ln T} \right] \\
\leq \mathbb{E}[|S|] + 2 \sigma \frac{\mathbb{E}[|S|]}{\sqrt{K}} \sqrt{(\eta_1 \lor \eta_2)T \ln T}.
\]
— Fix any $k \notin A_0$. Then under $G$, we have
\[
\mu_1 \leq \mu_k + 2\sigma \frac{\sqrt{(\eta_1 \lor \eta_2) \ln T}}{\sqrt{n_{t_k,k} - 1}}.
\]
Therefore,
\[
\Delta_k \leq 2\sigma \frac{\sqrt{(\eta_1 \lor \eta_2) \ln T}}{\sqrt{n_{t_k,k} - 1}}
\]
and thus,
\[
\Delta_k \leq 2\sigma \frac{\sqrt{(\eta_1 \lor \eta_2) \ln T}}{\sqrt{\max\{n_k - 1, 0\}}}
\]
As a result,
\[
\mathbb{E} \left[ \sum_{k \notin A_0} n_k \Delta_k \bigg| G \right] \leq \mathbb{E} \left[ \sum_{k \notin A_0} \Delta_k \right] + 2\mathbb{E} \left[ \sum_{k \notin A_0} \sigma \sqrt{(\eta_1 \lor \eta_2) \max\{n_k - 2, 0\} \ln T} \right] \leq (K - \mathbb{E}[|S|]) + 2\mathbb{E} \left[ \sum_{k \notin A_0} \sigma \sqrt{(\eta_1 \lor \eta_2) n_k \ln T} \right] \leq (K - \mathbb{E}[|S|]) + 2\sigma \sqrt{(\eta_1 \lor \eta_2) (K - \mathbb{E}[|S|]) T \ln T}.
\]
Now we have
\[
\mathbb{E}[R_0^\pi] \leq \mathbb{E} \left[ \sum_{k \in S} n_k \Delta_k \bigg| G \right] + \mathbb{E} \left[ \sum_{k \notin A_0} n_k \Delta_k \bigg| G \right] + 2K \leq 3K + 2\sigma \sqrt{T \ln T} \left( \sqrt{(\eta_1 \lor \eta_2) \mathbb{E}[|S|]} + \sqrt{(\eta_1 \lor \eta_2) (K - \mathbb{E}[|S|])} \right) \leq 4K + 4\sigma \sqrt{(\eta_1 \lor \eta_2) K T \ln T}.
\]
2. Let $x \geq 2K$. We have
\[
\mathbb{P}(\hat{R}_0^\pi(T) \geq x) \leq \mathbb{P} \left( R_0^\pi(T) \geq x(1 - 1/2\sqrt{K}) \right) + \mathbb{P} \left( N_0^\pi(T) \leq -x/2\sqrt{K} \right)
\]
From Lemma 1, the second term can be bounded as
\[
\mathbb{P} \left( N_0^\pi(T) \leq -x/2\sqrt{K} \right) \leq \exp \left( -\frac{x^2}{8K\sigma^2 T} \right).
\]
We are left to bound $\mathbb{P} \left( R_0^\pi(T) \geq x(1 - 1/2\sqrt{K}) \right)$.
(a) Let $\pi = \text{SE}$. For any $k \neq 1$, let $S_k$ be the event defined as
\[
S_k = \{ \text{Arm 1 is not eliminated before arm } k \}.
\]
Then
\[
\bar{S}_k = \{ \text{Arm 1 is eliminated before arm } k \}.
\]
So
\[
\mathbb{P} \left( R_0^\pi(T) \geq x(1 - 1/2\sqrt{K}) \right)
\]
\[
\begin{align*}
&= \Pr \left( \sum_{k \in A_0} n_k \Delta_k + \sum_{k \notin A_0} n_k \Delta_k \geq x(1-1/2\sqrt{K}) \right) \\
&\leq \Pr \left( \sum_{k \in A_0} (n_k-1) \Delta_k + \sum_{k \notin A_0} (n_k-1) \Delta_k \geq x(1-1/2\sqrt{K}) - K \right) \\
&\leq \Pr \left( \left( \bigcup_{k \in A_0} \{ (n_k-1) \Delta_k \geq \frac{x-2K}{4K} \} \right) \cup \left( \bigcup_{k \notin A_0} \{ (n_k-1) \Delta_k \geq \frac{(n_k-1)(x-2K)}{4T} \} \right) \right) \\
&\leq \sum_{k \neq 1} \Pr \left( (n_k-1) \Delta_k \geq \frac{x-2K}{4K}, k \in A_0 \right) + \Pr \left( (n_k-1) \Delta_k \geq \frac{(n_k-1)(x-2K)}{4T}, k \notin A_0 \right) \\
&\leq \sum_{k \neq 1} \Pr \left( (n_k-1) \Delta_k \geq \frac{x-2K}{4K}, k \in A_0 \right) + \Pr \left( \Delta_k \geq \frac{x-2K}{4T}, k \notin A_0 \right)
\end{align*}
\]

The reason that the first inequality holds is as follows. To prove it, we only need to show that the following cannot holds:

\[
(n_k-1) \Delta_k < \frac{x-2K}{4K}, \quad \forall k \in A_0; \quad (n_k-1) \Delta_k < \frac{(n_k-1)(x-2K)}{4T}, \quad \forall k \notin A_0.
\]

If not, then we have

\[
\sum_{k \neq 1} (n_k-1) \Delta_k = \sum_{k \in A_0} (n_k-1) \Delta_k + \sum_{k \notin A_0} (n_k-1) \Delta_k < \frac{(x-2K)|A_0|}{4K} + \frac{x-2K}{4} \\
< \frac{x-2K}{4} + \frac{x-2K}{4} \\
= \frac{x-2K}{2} \leq x(1-1/2\sqrt{K}) - K.
\]

Therefore,

\[
\begin{align*}
&\Pr \left( R_T^n \geq x(1-1/\sqrt{K}) \right) \\
&\leq \sum_{k \neq 1} \Pr \left( (n_k-1) \Delta_k \geq \frac{x-2K}{4K}, k \in A_0 \right) + \Pr \left( (n_k-1) \Delta_k \geq \frac{(n_k-1)(x-2K)}{4T}, k \notin A_0 \right) \\
&= \sum_{k \neq 1} \Pr \left( (n_k-1) \Delta_k \geq \frac{x-2K}{4K}, k \in A_0, S_k \right) + \sum_{k \neq 1} \Pr \left( (n_k-1) \Delta_k \geq \frac{x-2K}{4K}, k \in A_0, \bar{S}_k \right) \\
&\quad + \sum_{k \neq 1} \Pr \left( (n_k-1) \Delta_k \geq \frac{(n_k-1)(x-2K)}{4T}, k \notin A_0, S_k \right) + \sum_{k \neq 1} \Pr \left( \Delta_k \geq \frac{x-2K}{4T}, k \notin A_0, \bar{S}_k \right)
\end{align*}
\]

Fix \( k \neq 1 \). Now for each \( k \), we consider bounding the four terms separately.

\( -k \in A_0 \). With a slight abuse of notation, we let \( n_0 = \lceil \frac{x-2K}{4K \Delta_k} \rceil \leq n_k - 1 \). Also,

\[
\Delta_k \geq \frac{x-2K}{4K(n_k-1)} \geq \frac{(x-2K)K}{4KT} = \frac{x-2K}{4T}.
\]

We have

\[
\begin{align*}
&\Pr \left( (n_k-1) \Delta_k \geq \frac{x-2K}{4K}, k \in A_0, S_k \right) \\
&\leq \Pr \left( \hat{\mu}_{t_1(n_0),1} - \text{rad}(n_0) \leq \hat{\mu}_{t_1(n_0),k} + \text{rad}(n_0) \right) I \left\{ \Delta_k \geq \frac{x-2K}{4T} \right\}
\end{align*}
\]
\[
\begin{align*}
\mathbb{P} \left( \mu_1 - \frac{\sum_{m=1}^{n_0} \epsilon_{t_1(m),1}}{n_0} - \text{rad}(n_0) \leq \mu_k + \frac{\sum_{m=1}^{n_0} \epsilon_{t_k(m),k}}{n_0} + \text{rad}(n_0) \right) & \cdot \left\{ \Delta_k \geq \frac{x - 2K}{4T} \right\} \\
\leq & \mathbb{P} \left( \sum_{m=1}^{n_0} \epsilon_{t_1(m),1} - \epsilon_{t_k(m),k} \geq \frac{\Delta_k}{2} - \text{rad}(n_0) \right) \cdot \left\{ \Delta_k \geq \frac{x - 2K}{4T} \right\} \\
& + \mathbb{P} \left( \sum_{m=1}^{n_0} \epsilon_{t_k(m),k} \geq \frac{\Delta_k}{2} - \text{rad}(n_0) \right) \cdot \left\{ \Delta_k \geq \frac{x - 2K}{4T} \right\} \\
\leq & 2 \exp \left( -n_0 \left( \frac{\Delta_k}{2} - \text{rad}(n_0) \right)^2 / 2\sigma^2 \right) \cdot \left\{ \Delta_k \geq \frac{x - 2K}{4T} \right\} \\
& \cdot \frac{1}{2} \mathbb{P} \left( \left( \sum_{m=1}^{n_0} \epsilon_{t_1(m),1} - \epsilon_{t_k(m),k} \right) \geq \frac{\Delta_k}{2} - \text{rad}(n_0) \right) \cdot \left\{ \Delta_k \geq \frac{x - 2K}{4T} \right\} \\
= & 2 \exp \left( -n_0 \left( \frac{\Delta_k}{2} - \frac{2\sqrt{\eta_1 \eta_2 T \ln T}}{n_0 K^{1/2}} \right)^2 / 2\sigma^2 \right) \cdot \left\{ \Delta_k \geq \frac{x - 2K}{4T} \right\} \\
= & 2 \exp \left( -n_0 \Delta_k^2 \left( 1 - \frac{2\sqrt{\eta_1 \eta_2 T \ln T}}{n_0 K^{1/2}} \right)^2 / 8\sigma^2 \right) \cdot \left\{ \Delta_k \geq \frac{x - 2K}{4T} \right\} \\
\leq & 2 \exp \left( -\frac{(x - 2K)^2}{16KT} \left( 1 - \frac{8\sigma \sqrt{\eta_1 \eta_2 T \ln T}}{x - 2K} \right) \right) \cdot \left\{ \Delta_k \geq \frac{x - 2K}{4T} \right\} \\
\leq & 2 \exp \left( -\frac{(x - 2K - 8\sigma \sqrt{\eta_1 \eta_2 T \ln T})^2}{128\sigma^2 KT} \right), \quad (20)
\end{align*}
\]

Then we bound \( \mathbb{P}(n_k \Delta_k \geq (x - 2K)/4K, k \in \mathcal{A}_0, \tilde{S}_k) \). Suppose that after \( n \) phases, arm 1 is eliminated by arm \( k' \) (\( k' \) is not necessarily \( k \)). By the definition of \( \tilde{S}_k \), arm \( k \) is not eliminated. Therefore, we have

\[
\hat{\mu}_{t_{k'}(n),k'} - \frac{\sigma \sqrt{\eta_1 T \ln T}}{n \sqrt{K}} \geq \hat{\mu}_{t_1(n),1} + \frac{\sigma \sqrt{\eta_1 T \ln T}}{n \sqrt{K}}, \quad \hat{\mu}_{t_k(n),k} + \frac{\sigma \sqrt{\eta_1 T \ln T}}{n \sqrt{K}} \geq \hat{\mu}_{t_1(n),1} + \frac{\sigma \sqrt{\eta_1 T \ln T}}{n \sqrt{K}} \quad (21)
\]

holds simultaneously. The first inequality holds because arm 1 is eliminated. The second inequality holds because arm \( k \) is not eliminated. Now for fixed \( n \),

\[
\begin{align*}
& \mathbb{P} \left( \text{(21) happens; } \Delta_k \geq \frac{x - 2K}{4T} \right) \\
\leq & \mathbb{P} \left( \exists k': \hat{\mu}_{t_{k'}(n),k'} - \frac{\sigma \sqrt{\eta_1 T \ln T}}{n \sqrt{K}} \geq \hat{\mu}_{t_1(n),1} + \frac{\sigma \sqrt{\eta_1 T \ln T}}{n \sqrt{K}} \right) \\
& \land \mathbb{P} \left( \hat{\mu}_{t_k(n),k} + \frac{\sigma \sqrt{\eta_1 T \ln T}}{n \sqrt{K}} \geq \hat{\mu}_{t_1(n),1} + \frac{\sigma \sqrt{\eta_1 T \ln T}}{n \sqrt{K}}; \Delta_k \geq \frac{x - 2K}{4T} \right) \\
\leq & \mathbb{P} \left( \exists k': \sum_{m=1}^{n} \left( \frac{\epsilon_{t_{k'}(m),k'}}{n} - \frac{\epsilon_{t_1(m),1}}{n} \right) \geq \frac{2\sigma \sqrt{\eta_1 T \ln T}}{n \sqrt{K}} \right) \\
& \land \mathbb{P} \left( \sum_{m=1}^{n} \left( \frac{\epsilon_{t_k(m),k}}{n} - \frac{\epsilon_{t_1(m),1}}{n} \right) \geq \Delta_k; \Delta_k \geq \frac{x - 2K}{4T} \right) \\
\leq & \sum_{k' \neq 1} \left( \mathbb{P} \left( \sum_{m=1}^{n} \frac{\epsilon_{t_{k'}(m),k'}}{n} \geq \frac{\sigma \sqrt{\eta_1 T \ln T}}{n \sqrt{K}} \right) + \mathbb{P} \left( \sum_{m=1}^{n} \frac{\epsilon_{t_1(m),1}}{n} \leq -\frac{\sigma \sqrt{\eta_1 T \ln T}}{n \sqrt{K}} \right) \right) \\
& \land \left( \mathbb{P} \left( \sum_{m=1}^{n} \frac{\epsilon_{t_k(m),k}}{n} \geq \frac{(x - 2K)_+}{8T} \right) + \mathbb{P} \left( \sum_{m=1}^{n} \frac{\epsilon_{t_1(m),1}}{n} \leq -\frac{(x - 2K)_+}{8T} \right) \right) \\
\leq & 2K \exp \left( -\frac{\eta_1 T \ln T}{2nK} \right) \land 2K \exp \left( -\frac{n(x - 2K)_+^2}{128\sigma^2 T^2} \right) \\
= & 2K \exp \left( -\frac{\eta_1 T \ln T}{2nK} \lor \frac{n(x - 2K)_+^2}{128\sigma^2 T^2} \right)
\end{align*}
\]
\[
\leq 2K \exp \left( -\frac{(x - 2K) + \sqrt{\eta_1 \ln T}}{16\sigma \sqrt{KT}} \right)
\]

Therefore,

\[
P(n_k \Delta_k \geq (x - 2K) / 4K, \bar{S}_k, k \in A_0)
= P(\exists n \leq T/2: (21) \text{ happens}; n_k \Delta_k \geq (x - 2K) / 4K, k \in A_0)
\leq \sum_{n=1}^{\lceil T/2 \rceil} P \left( (21) \text{ happens}; \Delta_k \geq \frac{x - 2K}{4T} \right)
\leq KT \exp \left( -\frac{(x - 2K) + \sqrt{\eta_1 \ln T}}{16\sigma \sqrt{KT}} \right).
\]

\[\quad - k \notin A_0. \text{ With a slight abuse of notation, we let } n_0 = \left\lceil \frac{T}{K} \right\rceil \leq n_k - 1. \text{ Also, }\]
\[
\Delta_k \geq \frac{x - 2K}{4T}.
\]

We have

\[
P \left( (n_k - 1) \Delta_k \geq \frac{(n_k - 1)(x - 2K)}{4T}, k \notin A_0, S_k \right)
\leq P \left( \hat{\mu}_{t_1(n), 1} - \text{rad}(n_0) \leq \hat{\mu}_{t_k(n), k} + \text{rad}(n_0) \right)
\leq P \left( \mu_1 - \sum_{m=1}^{n_0} \epsilon_{t_1(m), 1} \leq \mu_k + \sum_{m=1}^{n_0} \epsilon_{t_k(m), k} \leq \text{rad}(n_0) \right)
\leq P \left( \sum_{m=1}^{n_0} \epsilon_{t_1(m), 1} \geq \frac{\Delta_k}{2} - \text{rad}(n_0) \right) + P \left( \sum_{m=1}^{n_0} \epsilon_{t_k(m), k} \geq \frac{\Delta_k}{2} - \text{rad}(n_0) \right)
\leq 2 \exp \left( -n_0 \left( \frac{\Delta_k}{2} - \text{rad}(n_0) \right)^2 / 2\sigma^2 \right)
= 2 \exp \left( -\left( \frac{\sqrt{n_0} \Delta_k}{2} - \sigma \sqrt{(\eta_1 \lor \eta_2) \ln T} \right)^2 / 2\sigma^2 \right)
\leq 2 \exp \left( -\frac{(x - 2K) - \sigma \sqrt{(\eta_1 \lor \eta_2) \ln T} \ln T}{8\sqrt{KT}} \right) + 2\sigma^2
\leq 2 \exp \left( -\frac{(x - 2K) - \sigma \sqrt{(\eta_1 \lor \eta_2) \ln T} \ln T}{128\sigma^2 KT} \right).
\]

Then we bound \(P((n_k - 1) \Delta_k \geq (n_k - 1)(x - 2K) / 4T, k \notin A_0, \bar{S}_k)\). The procedure is nearly the same as in the case where \(k \in A_0\). Suppose that after \(n\) phases, arm 1 is eliminated by arm \(k'\) (\(k'\) is not necessarily \(k\)).

By the definition of \(\bar{S}_k\), arm \(k\) is not eliminated. Therefore, we have

\[
\hat{\mu}_{t_1(n), 1} = \frac{\sqrt{n \eta T} \ln T}{n \sqrt{K}} \geq \hat{\mu}_{t_1(n), 1} + \frac{\sigma \sqrt{n \eta T} \ln T}{n \sqrt{K}}, \quad \hat{\mu}_{t_k(n), k} + \frac{\sigma \sqrt{n \eta T} \ln T}{n \sqrt{K}} \geq \hat{\mu}_{t_1(n), 1} + \frac{\sigma \sqrt{n \eta T} \ln T}{n \sqrt{K}}
\]
holds simultaneously. The first inequality holds because arm 1 is eliminated. The second inequality holds because arm \(k\) is not eliminated. Now for fixed \(n\),

\[
P \left( (24) \text{ happens}; \Delta_k \geq \frac{x - 2K}{4T}, k \notin A_0 \right)
\]
\[
\begin{align*}
&\leq \mathbb{P} \left( \exists k': \mu_{tk'(n),k'} - \frac{\sigma \sqrt{\eta_1 T \ln T}}{n \sqrt{K}} \geq \mu_{t_1(n),1} + \frac{\sigma \sqrt{\eta_1 T \ln T}}{n \sqrt{K}} \right) \\
&\quad \land \mathbb{P} \left( \mu_{tk(n),k} + \frac{\sigma \sqrt{\eta_1 T \ln T}}{n \sqrt{K}} \geq \mu_{t_1(n),1} + \frac{\sigma \sqrt{\eta_1 T \ln T}}{n \sqrt{K}} ; \Delta_k \geq x - 2K \right) \\
&\leq \mathbb{P} \left( \exists k': \sum_{m=1}^{n} \left( \epsilon_{tk'(m),k'} - \epsilon_{t_1(m),1} \right) \geq 2\sigma \sqrt{\eta_1 T \ln T} \right) \\
&\quad \land \mathbb{P} \left( \sum_{m=1}^{n} \left( \epsilon_{tk(m),k} - \epsilon_{t_1(m),1} \right) \geq \Delta_k ; \Delta_k \geq x - 2K \right) \\
&\leq \sum_{k' \neq 1} \left( \mathbb{P} \left( \sum_{m=1}^{n} \epsilon_{tk'(m),k'} \geq \frac{\sigma \sqrt{\eta_1 T \ln T}}{n \sqrt{K}} \right) + \mathbb{P} \left( \sum_{m=1}^{n} \epsilon_{t_1(m),1} \leq -\frac{\sigma \sqrt{\eta_1 T \ln T}}{n \sqrt{K}} \right) \right) \\
&\quad \land \left( \mathbb{P} \left( \sum_{m=1}^{n} \epsilon_{tk(m),k} \geq \frac{(x - 2K)}{2T} \right) + \mathbb{P} \left( \sum_{m=1}^{n} \epsilon_{t_1(m),1} \leq -\frac{(x - 2K)}{2T} \right) \right) \\
&\leq 2K \exp \left( -\frac{\eta_1 T \ln T}{2nk} \right) \land 2K \exp \left( -\frac{n(x - 2K)}{128\sigma^2 T^2} \right) \\
&= 2K \exp \left( -\frac{\eta_1 T \ln T}{2nk} \lor \frac{n(x - 2K)}{128\sigma^2 T^2} \right) \\
&\leq 2K \exp \left( -\frac{(x - 2K)}{2T} \frac{\sqrt{\eta_1 T}}{16\sigma \sqrt{KT}} \right)
\end{align*}
\]

Therefore,
\[
\mathbb{P}((n_k - 1)\Delta_k \geq (n_k - 1)(x - 2K)/4T, \tilde{S}_k, k \notin A_0) \\
= \mathbb{P} (\exists n \leq T/2 : (24) \text{ happens}; (n_k - 1)\Delta_k \geq (n_k - 1)(x - 2K)/4T, k \notin A_0) \\
\leq \sum_{n=1}^{T/2} \mathbb{P} \left( (24) \text{ happens}; \Delta_k \geq \frac{x - 2K}{4T} \right) \\
\leq KT \exp \left( -\frac{(x - 2K)}{16\sigma \sqrt{KT}} \right) . \tag{25}
\]

Note that the equations above hold for any instance \( \theta \). Combining (19), (20), (22), (23), (25) yields
\[
\sup_\theta \mathbb{P}(\hat{R}_\theta^n(T) \geq x) \\
\leq \exp \left( -\frac{x^2}{8K\sigma^2 T} \right) + 4K \exp \left( -\frac{(x - 2K - 8\sigma \sqrt{(\eta_1 \lor \eta_2)KT \ln T})^2}{128\sigma^2 KT} \right) + 2K^2 T \exp \left( -\frac{(x - 2K)}{16\sigma \sqrt{KT}} \right)
\]

(b) From (a), we have
\[
\mathbb{P} \left( R_\theta^n(T) \geq x(1 - 1/\sqrt{K}) \right) \\
\leq \sum_{k \neq 1} \left( \mathbb{P} \left( (n_k - 1)\Delta_k \geq \frac{x - 2K}{4K}, k \in A_0 \right) + \mathbb{P} \left( (n_k - 1)\Delta_k \geq \frac{(n_k - 1)(x - 2K)}{4T}, k \notin A_0 \right) \right) .
\]

Fix \( k \neq 1 \). Now for each \( k \), we consider bounding the two terms separately.

\(-k \in A_0\). With a slight abuse of notation, we let \( n_0 = \left\lceil \frac{x - 2K}{4K\Delta_k} \right\rceil \leq n_k - 1 \). Remember that \( t_k(n_0 + 1) \) is the time period that arm \( k \) is pulled for the \( (n_0 + 1) \)th time. We have
\[
\mathbb{P} \left( (n_k - 1)\Delta_k \geq \frac{x - 2K}{4K}, k \in A_0 \right) \\
= \mathbb{P} \left( n_k \geq 1 + \frac{x - 2K}{4K\Delta_k}, k \in A_0 \right)
\]

\(\sum_{k \neq 1} \left( \mathbb{P} \left( (n_k - 1)\Delta_k \geq \frac{x - 2K}{4K}, k \in A_0 \right) + \mathbb{P} \left( (n_k - 1)\Delta_k \geq \frac{(n_k - 1)(x - 2K)}{4T}, k \notin A_0 \right) \right) .
\]
\[
\theta 
= \frac{50}{\text{Simchi-Levi, Zheng and Zhu: Optimal Stochastic Bandit Policies with Light-tailed Risk}}
\]

\[
\begin{align*}
\leq & \ P\left(\hat{\mu}_{t_k(n_0+1)-1.1} + \text{rad}(n_{t_k(n_0+1)-1.1}) \leq \hat{\mu}_{t_k(n_0+1)-1,k} + \text{rad}(n_0)\right) \\
= & \ P\left(\mu_1 + \sum_{m=1}^{n_{t_k(n_0+1)-1.1}} \frac{\epsilon_{t_k,m,1}}{n_{t_k(n_0+1)-1.1}} + \text{rad}(n_{t_k(n_0+1)-1.1}) \leq \mu_k + \sum_{m=1}^{n_0} \frac{\epsilon_{t_k,m,k}}{n_0} + \text{rad}(n_0)\right) \\
\leq & \ P\left(\exists n \in [T]: \mu_1 + \sum_{m=1}^{n} \frac{\epsilon_{t_k,m,1}}{n} + \text{rad}(n) \leq \mu_k + \sum_{m=1}^{n_0} \frac{\epsilon_{t_k,m,k}}{n_0} + \text{rad}(n_0)\right) \\
\leq & \ P\left(\exists n \in [T]: \left(\sum_{m=1}^{n_0} \frac{\epsilon_{t_k,m,k}}{n_0} + \text{rad}(n_0)\right) - \left(\sum_{m=1}^{n} \frac{\epsilon_{t_k,m,1}}{n} + \text{rad}(n)\right) \geq \Delta_k\right) \\
\leq & \ P\left(\sum_{m=1}^{n_0} \frac{\epsilon_{t_k,m,k}}{n_0} + \text{rad}(n_0) \geq \frac{\Delta_k}{2}\right) + \frac{1}{2} \sum_{n=1}^{T} \ P\left(\sum_{m=1}^{n} \frac{\epsilon_{t_k,m,1}}{n} \geq \frac{\Delta_k}{2}\right) \\
\leq & \ P\left(\sum_{m=1}^{n} \frac{\epsilon_{t_k,m,1}}{n} \geq \frac{\Delta_k}{2}\right) \\
& \quad + \sum_{n=1}^{T} \ P\left(\sum_{m=1}^{n} \frac{\epsilon_{t_k,m,1}}{n} \geq \frac{(x-2K) + \sqrt{\eta_1 T \ln T}}{128\sigma^2 K T}\right) \\
\leq & \ \exp\left(-\frac{(x-2K) + \sqrt{\eta_1 T \ln T}}{128\sigma^2 K T}\right) + T \exp\left(-\frac{(x-2K) + \sqrt{\eta_1 T \ln T}}{16\sigma \sqrt{KT}}\right).
\end{align*}
\]

The last inequality holds from (20) and (22).

\( -k \notin \mathcal{A}_0 \). With a slight abuse of notation, we let \( n_0 = \lceil \frac{T}{K} \rceil \leq n_k - 1 \). Remember that \( t_k(n_0+1) \) is the time period that arm \( k \) is pulled for the \( (n_0+1) \)th time. We have

\[
\begin{align*}
\P \left( (n_k-1)\Delta_k \geq \frac{(n_k-1)(x-2K)}{4T}, \ k \notin \mathcal{A}_0 \right) \\
= & \ P\left(\Delta_k \geq \frac{x-2K}{4T}, \ k \notin \mathcal{A}_0 \right) \\
\leq & \ P\left(\hat{\mu}_{t_k(n_0+1)-1.1} + \text{rad}(n_{t_k(n_0+1)-1.1}) \leq \hat{\mu}_{t_k(n_0+1)-1,k} + \text{rad}(n_0)\right) \\
= & \ P\left(\mu_1 + \sum_{m=1}^{n_{t_k(n_0+1)-1.1}} \frac{\epsilon_{t_k,m,1}}{n_{t_k(n_0+1)-1.1}} + \text{rad}(n_{t_k(n_0+1)-1.1}) \leq \mu_k + \sum_{m=1}^{n_0} \frac{\epsilon_{t_k,m,k}}{n_0} + \text{rad}(n_0)\right) \\
\leq & \ P\left(\exists n \in [T]: \mu_1 + \sum_{m=1}^{n} \frac{\epsilon_{t_k,m,1}}{n} + \text{rad}(n) \leq \mu_k + \sum_{m=1}^{n_0} \frac{\epsilon_{t_k,m,k}}{n_0} + \text{rad}(n_0)\right) \\
\leq & \ P\left(\exists n \in [T]: \left(\sum_{m=1}^{n_0} \frac{\epsilon_{t_k,m,k}}{n_0} + \text{rad}(n_0)\right) - \left(\sum_{m=1}^{n} \frac{\epsilon_{t_k,m,1}}{n} + \text{rad}(n)\right) \geq \Delta_k\right) \\
\leq & \ P\left(\sum_{m=1}^{n_0} \frac{\epsilon_{t_k,m,k}}{n_0} + \text{rad}(n_0) \geq \frac{\Delta_k}{2}\right) + \frac{1}{2} \sum_{n=1}^{T} \ P\left(\sum_{m=1}^{n} \frac{\epsilon_{t_k,m,1}}{n} \geq \frac{\Delta_k}{2}\right) \\
\leq & \ \exp\left(-\frac{(x-2K) + \sqrt{\eta_1 T \ln T}}{128\sigma^2 K T}\right) + T \exp\left(-\frac{(x-2K) + \sqrt{\eta_1 T \ln T}}{16\sigma \sqrt{KT}}\right).
\end{align*}
\]

The last inequality holds from (23) and (25). Note that the equations above hold for any instance \( \theta \). Combining (19), (26), (27) yields

\[
\sup_{\theta} \P\left(\tilde{R}_n^\theta(T) \geq x\right) \\
\leq \ \exp\left(-\frac{x^2}{8K\sigma^2 T}\right) + 4K \exp\left(-\frac{(x-2K) + \sqrt{\eta_1 T \ln T}}{128\sigma^2 K T}\right) + 2K^2 T \exp\left(-\frac{(x-2K) + \sqrt{\eta_1 T \ln T}}{16\sigma \sqrt{KT}}\right).
\]
Proof of Theorem 6. Fix a time horizon of $T$. We write $t_k = t_k(n_{T,k})$ as the last time that arm $k$ is pulled throughout the $T$ time periods. By the definition of $n_k$ and $t_k$, the following event happens w.p. 1:

$$
\hat{\mu}_{t_k-1,1} + \text{rad}_k(n_{t_k-1,1}) \leq \hat{\mu}_{t_k-1,k} + \text{rad}_k(n_{t_k-1,k})
$$

Define

$$A_1 = \left\{ k \neq 1 : n_k \leq 1 + \frac{t_k^{3/4}T^{1/4}}{K}\right\}.$$

Fix $x \geq 2K$. We have

$$
P \left( R_k^*(T) \geq x(1-1/2\sqrt{K}) \right)
= P \left( \sum_{k \in A_1} n_k \Delta_k + \sum_{k \notin A_1} n_k \Delta_k \geq x(1-1/2\sqrt{K}) \right)
\leq P \left( \left( \bigcup_{k \in A_1} \{ n_k \Delta_k \geq \frac{x-2K}{4K} \} \right) \bigcup \left( \bigcup_{k \notin A_1} \{ n_k \Delta_k \geq \frac{(n_k-1)(x-2K)}{4\sqrt{T}t_k} \} \right) \right)
\leq \sum_{k \neq 1} \left( P \left( n_k \Delta_k \geq \frac{x-2K}{4K}, k \in A_1 \right) + P \left( n_k \Delta_k \geq \frac{(n_k-1)(x-2K)}{4\sqrt{T}t_k}, k \notin A_1 \right) \right)
\leq \sum_{k \neq 1} \left( P \left( n_k \Delta_k \geq \frac{x-2K}{4K}, k \in A_1 \right) + P \left( \Delta_k \geq \frac{x-2K}{4\sqrt{T}t_k}, k \notin A_1 \right) \right)

The reason that the second inequality holds is as follows. To prove it, we only need to show that the following cannot holds:

$$
(n_k-1)\Delta_k < \frac{x-2K}{4K}, \quad \forall k \in A_1; \quad (n_k-1)\Delta_k < \frac{(n_k-1)(x-2K)}{8\sqrt{T}t_k}, \quad \forall k \notin A_1.
$$

If not, then we have

$$
\sum_{k \neq 1} (n_k-1)\Delta_k = \sum_{k \in A_1} (n_k-1)\Delta_k + \sum_{k \notin A_1} (n_k-1)\Delta_k
\leq \frac{(x-2K)|A_1|}{4K} + \frac{x-2K}{8\sqrt{T}} \sum_{k \notin A_1} n_k \sqrt{\frac{4K}{t_k}}
\leq \frac{x-2K}{4} + \frac{x-2K}{8\sqrt{T}} \sum_{k \notin A_1} \frac{n_k}{\sqrt{t_k}}
\leq \frac{x-2K}{2} \leq x(1-1/2\sqrt{K}) - K.
$$

In fact, to bound $\sum_{k \notin A_1} \frac{n_k}{\sqrt{t_k}}$, we can assume $0 = t_{k_0} < t_{k_1} < t_{k_2} < \cdots$. Then we have

$$
t_{k_i} \geq n_{k_1} + \cdots + n_{k_i},
$$

because before up to time $t_{k_i}$, arms $k_1, \cdots, k_i$ have been pulled completely, and after time $t_{k_i}$ none of them will be pulled. Thus,

$$
\sum_{k \notin A_1} \frac{n_k}{\sqrt{t_k}} = \sum_{i=1}^{|A_1|} \frac{t_{k_i} - t_{k_{i-1}}}{\sqrt{t_{k_i}}} \leq \sum_{i=1}^{|A_1|} \frac{t_{k_i} - t_{k_{i-1}}}{\sqrt{t_{k_i}} + \sqrt{t_{k_{i-1}}}} = 2 \sum_{i=1}^{|A_1|} \frac{t_{k_i} - t_{k_{i-1}}}{\sqrt{t_{k_i}} + \sqrt{t_{k_{i-1}}}} \leq 2\sqrt{T}.
$$

Now fix $k \neq 1$. For each $k$, we consider bounding the two terms separately.
• $k \in A_1$. Remember that $n_k$ is the last time period that arm $k$ is pulled. Then from $k \in A_1$, we know

$$\frac{t_k^{3/4}T^{1/4}}{K} \geq n_k - 1 \geq \frac{x - 2K}{4K\Delta_k}$$

Thus,

$$\Delta_k \geq \frac{x - 2K}{4t_k^{3/4}T^{1/4}}.$$ 

We have

$$P\left( (n_k - 1)\Delta_k \geq \frac{x - 2K}{4K}, \ k \in A_1 \right)$$

$$= P\left( n_k \geq 1 + \frac{x - 2K}{4K\Delta_k}, \ k \in A_1 \right)$$

$$= P\left( \hat{\mu}_{k-1,1} + \text{rad}_{t_k}(n_{t_k-1,1}) \leq \mu_{t_k-1,k} + \text{rad}_{t_k}(n_{t_k-1,k}); \ \frac{t_k^{3/4}T^{1/4}}{K} \geq n_k - 1 \geq \frac{x - 2K}{4K\Delta_k} \right)$$

$$= P\left( \mu_k + \frac{\sum_{n=1}^{n_k-1} \epsilon_{t_k(m),1}}{n_{t_k-1,1}} + \text{rad}_{t_k}(n_{t_k-1,1}) \leq \mu_k + \frac{\sum_{n=1}^{n_k-1} \epsilon_{t_k(m),k}}{n_k - 1} + \text{rad}_{t_k}(n_k - 1); \ \frac{t_k^{3/4}T^{1/4}}{K} \geq n_k - 1 \geq \frac{x - 2K}{4K\Delta_k} \right)$$

$$\leq P\left( \exists n \in [T]: \mu_k + \frac{\sum_{m=1}^{n_k} \epsilon_{t_k(m),1}}{n} + \text{rad}_{t_k}(n) \leq \mu_k + \frac{\sum_{m=1}^{n_k} \epsilon_{t_k(m),k}}{n_k - 1} + \text{rad}_{t_k}(n_k - 1); \ \frac{t_k^{3/4}T^{1/4}}{K} \geq n_k - 1 \geq \frac{x - 2K}{4K\Delta_k} \right)$$

$$\leq \sum_{K_n \leq t_k^{3/4}T^{1/4}} \text{Pr}\left( \exists n \in [T]: \left( \frac{\sum_{m=1}^{n_k} \epsilon_{t_k(m),k}}{n_0} + \text{rad}_{t_k}(n_0) \right) - \left( \frac{\sum_{m=1}^{n_k} \epsilon_{t_k(m),1}}{n} + \text{rad}_{t_k}(n) \right) \geq \frac{x - 2K}{4Kn_0} \right)$$

Note that here $n_k$ and $t_k$ are both random variables, so we need to decompose the probability by $n_k - 1 = n_0$ and $t_k = t$ through all possible $(n_0, t)$. Now for any $n_0$ and $t$ such that $K_n \leq t^{3/4}T^{1/4}$, we have

$$P\left( \exists n \in [T]: \left( \frac{\sum_{m=1}^{n_k} \epsilon_{t_k(m),k}}{n_0} + \text{rad}_{t_k}(n_0) \right) - \left( \frac{\sum_{m=1}^{n_k} \epsilon_{t_k(m),1}}{n} + \text{rad}_{t_k}(n) \right) \geq \frac{x - 2K}{4Kn_0} \right)$$

$$\leq P\left( \frac{\sum_{m=1}^{n_k} \epsilon_{t_k(m),k}}{n_0} + \text{rad}_{t_k}(n_0) \geq \frac{x - 2K}{8Kn_0} \right) + P\left( \exists n \in [T]: \frac{\sum_{m=1}^{n_k} \epsilon_{t_k(m),1}}{n} + \text{rad}_{t_k}(n) \leq \frac{x - 2K}{8Kn_0} \right)$$

$$\leq P\left( \frac{\sum_{m=1}^{n_k} \epsilon_{t_k(m),k}}{n_0} + \frac{\sigma \sqrt{2\eta(1 + \ln K t)}}{n_0 \sqrt{K}} \geq \frac{x - 2K}{8Kn_0} \right) + \sum_{n=1}^{T} P\left( \frac{\sum_{m=1}^{n_k} \epsilon_{t_k(m),1}}{n_0} \geq \frac{x - 2K}{8Kn_0} + \frac{\sigma \sqrt{\eta \ln(K t)}}{8Kn_0} \right) + \frac{\sqrt{\eta \ln(2t)}}{8K\sigma \sqrt{t^{1/4}T^{1/4}}}$$

$$\leq \exp\left( -n_0 \left( \frac{x - 2K}{8Kn_0} - \frac{\sigma \sqrt{\ln T}}{n_0 \sqrt{K}} \right)^2 / 2\sigma^2 \right) + \sum_{n=1}^{T} \exp\left( -\frac{(x - 2K) + \sqrt{\ln T}}{8\sigma \sqrt{K T}} \right)$$

$$\leq \exp\left( -\frac{1}{n_0} \left( \frac{x - 2K}{8K} - \frac{\sigma \sqrt{\ln T}}{\sqrt{K}} \right)^2 / 2\sigma^2 \right) + T \exp\left( -\frac{(x - 2K) + \sqrt{\ln T}}{8\sigma \sqrt{K T}} \right)$$

$$\leq \exp\left( -\frac{K}{T} \left( \frac{x - 2K}{8K} - \frac{\sigma \sqrt{\ln T}}{\sqrt{K}} \right)^2 / 2\sigma^2 \right) + T \exp\left( -\frac{(x - 2K) + \sqrt{\ln T}}{8\sigma \sqrt{K T}} \right)$$
\[ \leq \exp\left(-\frac{(x-2K-8\sigma\sqrt{2\gamma KT\ln T})^2}{128\sigma^2 KT}\right) + T \exp\left(-\frac{(x-2K+\sqrt{\gamma\ln T})}{8\sigma\sqrt{KT}}\right). \]

Note that here we use the fact that for any \(1 \leq t \leq T\),
\[ \frac{\ln(2t)}{\sqrt{t}} \geq \frac{1}{4} \frac{\ln T}{\sqrt{T}}. \]

So we have
\[
\mathbb{P}\left((n_k - 1)\Delta_k \geq \frac{x-2K}{4K}, \ k \in A_1\right) \\
\leq \sum_{K_{n_0} \leq t^{3/4}T^{1/4}} \mathbb{P}\left(\exists n \in [T]: \left(\sum_{m=1}^{n_{k_{n_0}}} \epsilon_{k_{m}}(m) + \text{rad}_T(n_0)\right) - \left(\sum_{m=1}^{n_{k_{n_0}}} \epsilon_{k_{m}}(m) + \text{rad}_{K_{n_0}}(n)\right) \geq \frac{x-2K}{4K_{n_0}}\right) \\
\leq T^2 \exp\left(-\frac{(x-2K-8\sigma\sqrt{2\gamma KT\ln T})^2}{128\sigma^2 KT}\right) + T^3 \exp\left(-\frac{(x-2K+\sqrt{\gamma\ln T})}{8\sigma\sqrt{KT}}\right) \tag{28}
\]

\( \bullet \ k \notin A_1. \) Remember that \(t_k\) is the last time period that arm \(k\) is pulled. Then from \(k \notin A_1\), we know
\[ t_k \geq n_k \geq 1 + \frac{t_k^{3/4}T^{1/4}}{K}. \]

We have
\[
\mathbb{P}\left((n_k - 1)\Delta_k \geq \frac{(n_k - 1)(x-2K)}{8\sqrt{t_kT}}, \ k \notin A_k\right) \\
= \mathbb{P}\left(\Delta_k \geq \frac{x-2K}{8\sqrt{t_kT}}, \ n_k \geq 1 + \frac{t_k^{3/4}T^{1/4}}{K}\right) \\
= \mathbb{P}\left(\hat{\mu}_{t_k-1,1} + \text{rad}_{t_k}(n_{t_k-1,1}) \leq \hat{\mu}_{t_k-1,k} + \text{rad}_{t_k}(n_k-1); \Delta_k \geq \frac{x-2K}{8\sqrt{t_kT}}, \ n_k \geq 1 + \frac{t_k^{3/4}T^{1/4}}{K}\right) \\
= \mathbb{P}\left(\mu_1 + \sum_{m=1}^{n_{t_k-1,1}} \frac{\epsilon_{t_k(m),1}}{n_{t_k-1,1}} + \text{rad}_{t_k}(n_{t_k-1,1}) \leq \mu_k + \sum_{m=1}^{n_k-1} \frac{\epsilon_{t_k(m),k}}{n_k-1} + \text{rad}_{t_k}(n_k-1); \right. \\
\left. \Delta_k \geq \frac{x-2K}{8\sqrt{t_kT}}, \ n_k \geq 1 + \frac{t_k^{3/4}T^{1/4}}{K}\right) \\
\leq \mathbb{P}\left(\exists n \in [T]: \mu_1 + \sum_{m=1}^{n} \frac{\epsilon_{t_k(m),1}}{n} + \text{rad}_T(n) \leq \mu_k + \sum_{m=1}^{n_k-1} \frac{\epsilon_{t_k(m),k}}{n_k-1} + \text{rad}_T(n_k-1); \right. \\
\left. \Delta_k \geq \frac{x-2K}{8\sqrt{t_kT}}, \ n_k-1 \geq \frac{t_k^{3/4}T^{1/4}}{K}\right) \\
\leq \sum_{K_{n_0} \leq t^{3/4}T^{1/4}} \mathbb{P}\left(\exists n \in [T]: \left(\sum_{m=1}^{n_{0}} \frac{\epsilon_{t_k(m),k}}{n_0} + \text{rad}_T(n_0)\right) - \left(\sum_{m=1}^{n_{0}} \frac{\epsilon_{t_k(m),1}}{n} + \text{rad}_T(n)\right) \geq \frac{x-2K}{8\sqrt{T}}\right)
\]

Note that here \(n_k\) and \(t_k\) are both random variables, so we need to decompose the probability by \(n_k - 1 = n_0\) and \(t_k = t\) through all possible \((n_0,t)\). Now for any \((n_0,t)\) such that \(t \geq n_0\) and \(K_{n_0} \geq t^{3/4}T^{1/4}\), we know that
\[ Kt \geq K\mu_{n_0} \geq T^{1/4}, \]
and so
\[ \ln(Kt) \geq \frac{1}{4} \ln T. \]
We have

\[
\mathbb{P}\left( \exists n \in [T] : \left( \sum_{m=1}^{n_0} \frac{\epsilon_{\ell,m,k}}{n_0} \right) + \text{rad}_i(n_0) \right) - \left( \sum_{m=1}^{n_0} \frac{\epsilon_{\ell,m,1}}{n} + \text{rad}_i(n) \right) \geq \frac{x - 2K}{8 \sqrt{tT}} \right) 
\leq \mathbb{P}\left( \sum_{m=1}^{n_0} \frac{\epsilon_{\ell,m,k}}{n_0} + \text{rad}_i(n_0) \geq \frac{x - 2K}{16 \sqrt{tT}} \right) + \mathbb{P}\left( \exists n \in [T] : \sum_{m=1}^{n_0} \frac{\epsilon_{\ell,m,1}}{n} + \text{rad}_i(n) \leq -\frac{x - 2K}{16 \sqrt{tT}} \right) 
\leq \mathbb{P}\left( \sum_{m=1}^{n_0} \frac{\epsilon_{\ell,m,k}}{n_0} + \frac{\sigma \sqrt{\eta \ln(KT)}}{n_0 \sqrt{K}} \geq x - 2K + \frac{\sigma \sqrt{\eta \ln(KT)}}{16 \sqrt{tT}} \right) + \sum_{n=1}^{T} \mathbb{P}\left( \sum_{m=1}^{n_0} \frac{-\epsilon_{\ell,m,1}}{n} \geq \frac{x - 2K}{8 n \sqrt{K T}} \right) 
\leq \exp\left( -n_0 \left( \frac{x - 2K}{16 \sqrt{tT}} - \frac{\sqrt{\eta \ln(KT)}}{n_0 \sqrt{K}} \right)^2 / 2 \sigma^2 \right) + \sum_{n=1}^{T} \exp\left( -\frac{(x - 2K) + \sqrt{\eta \ln(KT)}}{16 \sigma \sqrt{K T}} \right) 
\leq \exp\left( -\frac{(x - 2K - 16 \sigma \sqrt{\eta KT})^2}{512 \sigma^2 KT} \right) + T \exp\left( -\frac{(x - 2K) + \sqrt{\eta \ln(KT)}}{16 \sigma \sqrt{K T}} \right) 
\]

So we have

\[
\mathbb{P}\left( (n_k - 1) \Delta_k \geq \frac{(n_k - 1)(x - 2K)}{8 \sqrt{t_k T}}, \ k \notin A_1 \right) 
\leq \sum_{n_0, n_0 \geq 1/K} \mathbb{P}\left( \exists n \in [T] : \left( \sum_{m=1}^{n_0} \frac{\epsilon_{\ell,m,k}}{n_0} \right) + \text{rad}_i(n_0) \right) - \left( \sum_{m=1}^{n_0} \frac{\epsilon_{\ell,m,1}}{n} + \text{rad}_i(n) \right) \geq \frac{x - 2K}{8 \sqrt{tT}} \right) 
\leq T^2 \exp\left( -\frac{(x - 2K - 16 \sigma \sqrt{\eta KT})^2}{512 \sigma^2 KT} \right) + T^3 \exp\left( -\frac{(x - 2K) + \sqrt{\eta \ln(KT)}}{16 \sigma \sqrt{K T}} \right) 
\]

(29)

Note that all the equations above hold for any instance \( \theta \). Combining (19), (28), (29) yields

\[
\sup_{\theta} \mathbb{P}(R^2_{\theta}(T) \geq x) 
\leq \exp\left( -\frac{x^2}{8 \sigma^2 KT} \right) + 2 KT^2 \exp\left( -\frac{(x - 2K - 16 \sigma \sqrt{\eta KT \ln(T)})^2}{512 \sigma^2 KT} \right) + 2 KT^3 \exp\left( -\frac{(x - 2K) + \sqrt{\eta \ln(T)}}{16 \sigma \sqrt{K T}} \right). 
\]

\[\square\]

**Proof of Theorem 7.** To simplify notations, we write \( \Delta_i = \theta^\top (a_i^* - a_i) \). Also, we write

\[
A_i = [a_1, \cdots, a_t], \quad R_t = [r_1, \cdots, r_t]^\top, \quad \E_i = [\epsilon_{1,a_1}, \cdots, \epsilon_{t,a_t}]^\top.
\]

Then

\[
\hat{\theta}_t = V_t^{-1} A_t R_t = V_t^{-1} A_t (A_t^\top \theta + \E_t) = \theta - V_t^{-1} \theta + V_t^{-1} A_t \E_t.
\]

Note that

\[
R^2_{\theta}(T) = \sum_i \Delta_i = \sum_i \frac{\Delta_i}{a_i^\top V_{i-1} a_i} \cdot a_i^\top V_{i-1} a_i
\]

and from Lemma 11 in Abbasi-Yadkori et al. (2011),

\[
\sum_i a_i^\top V_{i-1} a_i \leq 2 \ln \det V_{T-1} - 2 \ln \det V_1
\]
shown that for any $\delta > 0$, w.p. at least $1 - \delta$, the following holds:

$$(A_{t-1}E_{t-1})^T V_{t-1}^{-1} A_{t-1} E_{t-1} \leq 2\sigma^2 \log \left( \frac{\det(V_{t-1})/\det(V_0)}{\delta} \right) \leq 2\sigma^2 \log \left( \frac{(T/d)^{2d}}{\delta} \right)$$

Thus, for any $y \geq 0$, we have

$$\mathbb{P} \left( \sqrt{(A_{t-1}E_{t-1})^T V_{t-1}^{-1} A_{t-1} E_{t-1}} \geq y \right) \leq (T/d)^{2d} \exp \left( -\frac{x^2}{2\sigma^2} \right)$$

We have

$$\sup_{\theta} \mathbb{P}(\hat{R}_\theta^\ast(T) \geq x) \leq \mathbb{P}(N^\ast(T) \leq -x/d) + \sup_{\theta} \mathbb{P}(R_\theta^\ast(T) \geq x(1 - 1/d))$$

$$\leq \exp \left( -\frac{x^2}{2\sigma^2 d^2} \right) + \sup_{\theta} \mathbb{P}(R_\theta^\ast(T) \geq x/2)$$

Also, for any $\theta$,

$$\mathbb{P}(R_\theta^\ast(T) \geq x/2) \leq \mathbb{P} \left( \bigcup_{t \geq 2} \left\{ \frac{\Delta_t}{a_t^\top V_{t-1}^{-1} a_t} \geq \frac{x - 4\sqrt{d}}{8d \ln T}, a_t^\top V_{t-1}^{-1} a_t > d/t \right\} \right)$$

The reason that (30) holds is as follows. To prove it, we only need to show that the following events cannot hold simultaneously:

$$\Delta_t < \frac{x - 4\sqrt{d}}{4t \ln T}, \text{ if } a_t^\top V_{t-1}^{-1} a_t \leq d/t; \quad \frac{\Delta_t}{a_t^\top V_{t-1}^{-1} a_t} < \frac{x - 4\sqrt{d}}{8d \ln T}, \text{ if } a_t^\top V_{t-1}^{-1} a_t > d/t.$$ 

If not, then

$$R_\theta^\ast(T) = \theta^\top (a_1^* - a_1) + \sum_{t \geq 2} \Delta_t \mathbb{1} \{ a_t^\top V_{t-1}^{-1} a_t \leq d/t \}$$

$$+ \frac{\Delta_t}{a_t^\top V_{t-1}^{-1} a_t} \mathbb{1} \{ a_t^\top V_{t-1}^{-1} a_t \geq d/t \}$$

$$< 2\sqrt{d} + \frac{x - 4\sqrt{d}}{4t \ln T} + \frac{x - 4\sqrt{d}}{8d \ln T} a_t^\top V_{t-1}^{-1} a_t$$

$$\leq 2\sqrt{d} + \frac{x}{4} - \sqrt{d} + \frac{x}{4} - \sqrt{d} = x/2.$$ 

This is a contradiction. At time $t$, the policy takes action $a_t$, which means

$$\hat{\theta}_{t-1}^\top a_t + (a_t^\top V_{t-1}^{-1} a_t)\sigma \sqrt{\frac{nt}{d}} + \sqrt{d(a_t^\top V_{t-1}^{-1} a_t)} \geq$$

$$\hat{\theta}_{t-1}^\top a_t^* + (a_t^*^\top V_{t-1}^{-1} a_t^*)\sigma \sqrt{\frac{nt}{d}} + \sqrt{d(a_t^*^\top V_{t-1}^{-1} a_t^*)}$$

$$\Leftrightarrow \theta^\top a_t - \theta^\top V_{t-1}^{-1} a_t + (V_{t-1}^{-1} A_{t-1} E_{t-1})^\top a_t + (a_t^\top V_{t-1}^{-1} a_t)\sigma \sqrt{\frac{nt}{d}} + \sqrt{d(a_t^\top V_{t-1}^{-1} a_t)} \geq$$
\[ \theta^\top a_t^* - \theta^\top V_{t-1}^{-1}a_t + (V_{t-1}^{-1}A_{t-1}E_{t-1})^\top a_t^* + (a_t^\top V_{t-1}^{-1}a_t^*)\sigma \sqrt{\frac{nt}{d}} + \sqrt{d(a_t^\top V_{t-1}^{-1}a_t^*)} \]

\[ \Rightarrow (V_{t-1}^{-1}A_{t-1}E_{t-1})^\top a_t + (a_t^\top V_{t-1}^{-1}a_t)\sigma \sqrt{\frac{nt}{d}} + \sqrt{d(a_t^\top V_{t-1}^{-1}a_t)} - \theta^\top V_{t-1}^{-1}a_t \geq \]

\[ \Delta_t + (V_{t-1}^{-1}A_{t-1}E_{t-1})^\top a_t + (a_t^\top V_{t-1}^{-1}a_t)\sigma \sqrt{\frac{nt}{d}} + \sqrt{d(a_t^\top V_{t-1}^{-1}a_t)} - \theta^\top V_{t-1}^{-1}a_t, \]

\[ \Rightarrow a_t^\top V_{t-1}^{-1}A_{t-1}E_{t-1} + (a_t^\top V_{t-1}^{-1}a_t)\sigma \sqrt{\frac{nt}{d}} + 2\sqrt{d(a_t^\top V_{t-1}^{-1}a_t)} \geq \]

\[ \Delta_t + a_t^\top V_{t-1}^{-1}A_{t-1}E_{t-1} + (a_t^\top V_{t-1}^{-1}a_t)\sigma \sqrt{\frac{nt}{d}} \]

\[ \Rightarrow a_t^\top V_{t-1}^{-1}A_{t-1}E_{t-1} \geq \frac{\Delta_t}{2} - (a_t^\top V_{t-1}^{-1}a_t)\sigma \sqrt{\frac{nt}{d}} - 2\sqrt{d(a_t^\top V_{t-1}^{-1}a_t)} \text{ or } \]

\[ - a_t^\top V_{t-1}^{-1}A_{t-1}E_{t-1} \geq \frac{\Delta_t}{2} + (a_t^\top V_{t-1}^{-1}a_t)\sigma \sqrt{\frac{nt}{d}} \]

Note that in (31) we use the following inequality: for any \( a \in A_t, \)

\[ |\theta^\top V_{t-1}^{-1}a| \leq \sqrt{\theta^\top V_{t-1}^{-1}\theta} \sqrt{a^\top V_{t-1}^{-1}a} \leq \sqrt{d(a^\top V_{t-1}^{-1}a)} \]

Combining (30) and (31) yields

\[ P(R^*_t(T) \geq x/2) \]

\[ \leq \sum_t P \left( \Delta_t \geq \frac{x - 4\sqrt{d}}{4t \ln T}, a_t^\top V_{t-1}^{-1}a_t \leq d/t, a_t^\top V_{t-1}^{-1}A_{t-1}E_{t-1} \geq \frac{\Delta_t}{2} - (a_t^\top V_{t-1}^{-1}a_t)\sigma \sqrt{\frac{nt}{d}} - 2\sqrt{d(a_t^\top V_{t-1}^{-1}a_t)} \right) \]

\[ + \sum_t P \left( \Delta_t \geq \frac{x - 4\sqrt{d}}{4t \ln T}, a_t^\top V_{t-1}^{-1}a_t \leq d/t, -a_t^\top V_{t-1}^{-1}A_{t-1}E_{t-1} \geq \frac{\Delta_t}{2} + (a_t^\top V_{t-1}^{-1}a_t)\sigma \sqrt{\frac{nt}{d}} \right) \]

\[ + \sum_t P \left( \frac{\Delta_t}{a_t^\top V_{t-1}^{-1}a_t} \geq \frac{x - 4\sqrt{d}}{8t \ln T}, a_t^\top V_{t-1}^{-1}a_t > \frac{d}{t}, \right. \]

\[ \left. -a_t^\top V_{t-1}^{-1}A_{t-1}E_{t-1} \geq \frac{\Delta_t}{2} - (a_t^\top V_{t-1}^{-1}a_t)\sigma \sqrt{\frac{nt}{d}} - 2\sqrt{d(a_t^\top V_{t-1}^{-1}a_t)} \right) \]

\[ + \sum_t P \left( \frac{\Delta_t}{a_t^\top V_{t-1}^{-1}a_t} \geq \frac{x - 4\sqrt{d}}{8t \ln T}, a_t^\top V_{t-1}^{-1}a_t > \frac{d}{t}, -a_t^\top V_{t-1}^{-1}A_{t-1}E_{t-1} \geq \frac{\Delta_t}{2} + (a_t^\top V_{t-1}^{-1}a_t)\sigma \sqrt{\frac{nt}{d}} \right) \]

We bound each term separately. We have

\[ P \left( \Delta_t \geq \frac{x - 4\sqrt{d}}{4t \ln T}, a_t^\top V_{t-1}^{-1}a_t \leq d/t, a_t^\top V_{t-1}^{-1}A_{t-1}E_{t-1} \geq \frac{\Delta_t}{2} - (a_t^\top V_{t-1}^{-1}a_t)\sigma \sqrt{\frac{nt}{d}} - 2\sqrt{d(a_t^\top V_{t-1}^{-1}a_t)} \right) \]

\[ \leq P \left( a_t^\top V_{t-1}^{-1}a_t \leq d/t, a_t^\top V_{t-1}^{-1}A_{t-1}E_{t-1} \geq \frac{x - 4\sqrt{d}}{8t \ln T} - \sigma \sqrt{\eta d/t} - 2d/\sqrt{t} \right) \]

\[ \leq P \left( a_t^\top V_{t-1}^{-1}a_t \leq d/t, \frac{(x - 4\sqrt{d})^2} {8t \ln T} - \sigma \sqrt{\eta d/t} - 2d/\sqrt{t} \geq \frac{x - 4\sqrt{d}}{8t \ln T} \right) \]

\[ \leq P \left( (A_{t-1}E_{t-1})^\top V_{t-1}^{-1}A_{t-1}E_{t-1} \geq \frac{(x - 4\sqrt{d} - 16d\sqrt{T \ln T} - 8\sigma \sqrt{\eta d T \ln T})^2}{8\sqrt{d T \ln T}} \right) \]

\[ \leq (T/d)^{2d} \exp \left( -\frac{(x - 4\sqrt{d} - 16d\sqrt{T \ln T} - 8\sigma \sqrt{\eta d T \ln T})^2}{128\sigma^2 d T \ln^2 T} \right) \]
and

$$
P\left( \Delta_t \geq \frac{x - 4\sqrt{d}}{4t \ln T}, \ a_t^\top V_{t-1}^{-1} a_t \leq d/t, \ -a_t^\top V_{t-1}^{-1} A_t^\top \epsilon_{t-1} \geq \frac{\Delta_t}{2} + (a_t^\top V_{t-1}^{-1} a_t) \sigma \sqrt{\frac{\eta t}{d}} \right) \\
\leq P \left( -a_t^\top V_{t-1}^{-1} A_t^\top \epsilon_{t-1} \geq \sqrt{x - 4\sqrt{d}} \frac{1}{4t \ln T} (a_t^\top V_{t-1}^{-1} a_t) \sigma \sqrt{\frac{\eta t}{d}} \right) \\
\leq P \left( \frac{|a_t^\top V_{t-1}^{-1} A_t^\top \epsilon_{t-1}|}{\sqrt{a_t^\top V_{t-1}^{-1} a_t}} \geq \sqrt{x - 4\sqrt{d}} \frac{\sigma \sqrt{\eta t}}{2 \sqrt{dT \ln T}} \right) \\
\leq P \left( (A_t^\top \epsilon_{t-1})^\top V_{t-1}^{-1} A_t^\top \epsilon_{t-1} \geq \frac{(x - 4\sqrt{d}) \sigma \sqrt{\eta t}}{2 \sqrt{dT \ln T}} \right) \\
= \frac{(T/d)^{2\eta} \exp \left( -\frac{(x - 4\sqrt{d}) \sigma \sqrt{\eta t}}{4 \sigma \sqrt{dT \ln T}} \right)}
$$

and

$$
P \left( \Delta_t \geq \frac{x - 4\sqrt{d}}{8d \ln T}, \ a_t^\top V_{t-1}^{-1} a_t > \frac{d}{t}, \ a_t^\top V_{t-1}^{-1} A_t^\top \epsilon_{t-1} \geq \frac{\Delta_t}{2} - (a_t^\top V_{t-1}^{-1} a_t) \sigma \sqrt{\frac{\eta t}{d}} - 2\sqrt{d(a_t^\top V_{t-1}^{-1} a_t)} \right) \\
\leq P \left( \frac{\Delta_t}{a_t^\top V_{t-1}^{-1} a_t} \geq \frac{x - 4\sqrt{d}}{8d \ln T}, \ a_t^\top V_{t-1}^{-1} a_t > \frac{d}{t}, \ a_t V_{t-1}^{-1} A_t^\top \epsilon_{t-1} \geq \frac{\Delta_t}{2a_t^\top V_{t-1}^{-1} a_t} - \sigma \sqrt{\frac{\eta t}{d}} - 2\sqrt{t} \right) \\
\leq P \left( a_t^\top V_{t-1}^{-1} a_t > \frac{d}{t}, \ a_t^\top V_{t-1}^{-1} A_t^\top \epsilon_{t-1} \geq \frac{x - 4\sqrt{d}}{16d \ln T} - \sigma \sqrt{\frac{\eta t}{d}} - 2\sqrt{t} \right) \\
\leq P \left( \frac{|a_t^\top V_{t-1}^{-1} A_t^\top \epsilon_{t-1}|}{\sqrt{a_t^\top V_{t-1}^{-1} a_t}} \geq \sqrt{x - 4\sqrt{d} - 32d \sqrt{T \ln T} - 16\sigma \sqrt{\eta} d \ln T} \right) + \sqrt{\frac{\eta t}{d}} \right) \\
\leq P \left( \frac{(A_t^\top \epsilon_{t-1})^\top V_{t-1}^{-1} A_t^\top \epsilon_{t-1} \geq \frac{x - 4\sqrt{d} - 32d \sqrt{T \ln T} - 16\sigma \sqrt{\eta} d \ln T}{16 \sqrt{dT \ln T}} \right) \\
\leq (T/d)^{2\eta} \exp \left( -\frac{(x - 4\sqrt{d} - 32d \sqrt{T \ln T} - 16\sigma \sqrt{\eta} d \ln T)^2}{512 \sigma ^2 d T \ln ^2 T} \right)
$$

and

$$
P \left( \Delta_t \geq \frac{x - 4\sqrt{d}}{8d \ln T}, \ a_t^\top V_{t-1}^{-1} a_t > \frac{d}{t}, \ -a_t^\top V_{t-1}^{-1} A_t^\top \epsilon_{t-1} \geq \frac{\Delta_t}{2} + (a_t^\top V_{t-1}^{-1} a_t) \sigma \sqrt{\frac{\eta t}{d}} \right) \\
\leq P \left( \frac{\Delta_t}{a_t^\top V_{t-1}^{-1} a_t} \geq \frac{x - 4\sqrt{d}}{8d \ln T}, \ a_t^\top V_{t-1}^{-1} a_t > \frac{d}{t}, \ -a_t^\top V_{t-1}^{-1} A_t^\top \epsilon_{t-1} \geq \frac{\Delta_t}{2} - (a_t^\top V_{t-1}^{-1} a_t) \sigma \sqrt{\frac{\eta t}{d}} \right) \\
\leq P \left( \frac{|a_t^\top V_{t-1}^{-1} A_t^\top \epsilon_{t-1}|}{\sqrt{a_t^\top V_{t-1}^{-1} a_t}} \geq \frac{(x - 4\sqrt{d}) \sqrt{\eta}}{4 \sqrt{dT \ln T}} \right) \\
\leq P \left( (A_t^\top \epsilon_{t-1})^\top V_{t-1}^{-1} A_t^\top \epsilon_{t-1} \geq \frac{(x - 4\sqrt{d}) \sigma \sqrt{\eta}}{4 \sqrt{dT \ln T}} \right) \\
$$
\[
\leq (T/d)^{2d} \exp \left( -\frac{(x - 4\sqrt{d} + \sqrt{\eta})}{8\sigma \sqrt{dT \ln T}} \right).
\]

Plugging the four bounds above into (30) yields the final result

\[
\sup_\theta \mathbb{P} (\hat{R}_\theta (T) \geq x) \leq \exp \left( -\frac{x^2}{2\sigma^2 T^2} \right) + 2d(T/d)^{2d+1} \exp \left( -\frac{(x - 4\sqrt{d} - 32d\sqrt{T \ln T} - 16\sqrt{\eta T \ln T})^2}{512\sigma^2 T \ln^2 T} \right)
+ 2d(T/d)^{2d+1} \exp \left( -\frac{(x - 4\sqrt{d} + \sqrt{\eta})}{8\sigma \sqrt{dT \ln T}} \right).
\]