Decision Variance in Online Learning

Sattar Vakili, Alexis Boukouvalas
Prowler.io, Cambridge, UK
{sattar,alexis@prowler.io}

Abstract
Online learning has classically focused on the expected behaviour of learning policies. Recently, risk-averse online learning has gained much attention. In this paper, a risk-averse multi-armed bandit problem where the performance of policies is measured based on the mean-variance of the rewards is studied. The variance of the rewards depends on the variance of the underlying processes as well as the variance of the player's decisions. The performance of two existing policies is analyzed and new fundamental limitations on risk-averse learning is established. In particular, it is shown that although an $O(\log T)$ distribution-dependent regret in time $T$ is achievable (similar to the risk-neutral setting), the worst-case (i.e. minimax) regret is lower bounded by $\Omega(T)$ (in contrast to the $\Omega(\sqrt{T})$ lower bound in the risk-neutral setting). The lower bound results are even stronger in the sense that they are proven for the case of online learning with full feedback.

1 Introduction
Multi-armed bandit (MAB) is a class of online learning and decision making problems. The origins of the problem can be traced back to the works by Thompson [1] for clinical trial applications. In this classic problem, there are $K$ arms and a single player who plays the arms sequentially. Playing each arm $k$ at each time $t$ results in a reward $X_{k,t}$ drawn from an unknown distribution. The player observes only the reward of the selected arm. In the classic formulation of the problem the objective is to maximize the cumulative reward over time or equivalently to minimize regret that is defined as the cumulative loss in comparison to the best arm selection policy under known distribution model (i.e., always playing the arm with the highest expected reward). The crux of the classic MAB problem is in the trade-off between exploration, learning the statistics of the rewards for each arm, and exploitation, capitalizing on the gathered information to make the optimal decision at each time.

The classic MAB mainly focuses on the expected regret of arm selection policies. Motivated by emerging engineering and financial applications, there recently has been increased attention to MAB under risk measures. In this paper, we study a MAB problem under a mean-variance measure which is a
common risk measure in modern portfolio selection [2]. The mean-variance of a random variable \( X \) is defined as, \( \text{MV}(X) = \sigma^2(X) - \lambda \mu(X) \), a linear combination of its mean \( \mu(X) \) and its variance \( \sigma^2(X) \) [3]. The higher the value of \( \lambda \) is, the more risk tolerant the measure is. In mean-variance portfolio optimization, the objective is to maximize the expected return for a given level of variance or to minimize the variance for a given expected return. The parameter \( \lambda \) can be interpreted as the Lagrangian multiplier in this constraint optimization problem.

Let \( \{ \pi_t \}_{t=1}^T \) denote the arm played by an arm selection policy \( \pi \) at time \( t \) and \( X_{\pi_t,t} \) denote the reward obtained at time \( t \) under policy \( \pi \). We define the cumulative mean-variance of observations as

\[
\text{MV}_\pi(T) = \sum_{t=1}^{T} \text{MV}(X_{\pi_t,t}).
\]  

(1)

Similar to the risk-neutral MAB, we set our objective to minimize the cumulative mean-variance of the rewards or equivalently minimize the regret defined as the excess in cumulative mean-variance in comparison to the optimal arm selection policy under known distribution model:

\[
\mathcal{R}_\pi(T) = \text{MV}_\pi(T) - \text{MV}_{\pi^*}(T),
\]  

(2)

(\( \pi^* \) is the optimal policy under known distribution model).

The regret definition in risk-averse MAB is similar to the one in risk-neutral MAB except for that the measure of expected value is replaced with the measure of mean-variance. However, the performance of a policy is not merely determined by the mean-variance of the rewards of the selected arms but also, as we shall see in Sec. 2 by the variance in the decisions; hence, the title of the paper.

1.1 Contribution

In the risk-neutral MAB, an \( \Omega(\log T) \) lower bound on distribution-dependent regret was shown in the seminal work by Lai and Robbins [4]. The lower bound on the worst case (minimax) regret can be concluded to be \( \Omega(\sqrt{T}) \) from [5], as well as the lower bound results for non-stochastic MAB in [6]. In Sec. 3 we establish fundamental limits on performance of policies under the risk measure. Specifically, we show that the \( \Omega(\log T) \) lower bound on distribution-dependent regret holds also under the mean-variance risk measure. However, the variance in the decisions makes an \( \Omega(T) \) worst case regret inevitable. This lower bound result is even stronger in the sense that it is established under the case of full feedback setting where the player observes the reward of all arms at each time; in contrast to observing only the selected arm.

We also analyze the performance of Upper Confidence Bound (UCB)-type policies under the risk-averse measure. In particular we study a modification of classic UCB introduced in [7] for risk-neutral MAB, referred to as Mean Variance Lower Confidence Bound (MV-LCB). We also study a policy based on arm eliminations introduced in [8] for risk-neutral MAB referred to as Confidence
Bounds based Arm Eliminations (CB-AE). We show that, while an $O(\log T)$ distribution-dependent regret is achievable, both MV-LCB and CB-AE have a linear worst-case regret in time. We also provide simulation examples that show CB-AE outperforms MV-LCB.

1.2 Related Work

The risk-neutral MAB problem has been extensively studied for various applications including clinical trials, internet advertisement, web search, target tracking (see [9] and references therein), as well as various financial and economical applications (see [10] and references therein).

The MAB problem has been studied much less extensively under the measure of mean-variance. In [11] and [12], the empirical mean-variance of the observed reward process was considered as a risk measure (which is commonly referred to as volatility) and a modification of UCB and a modification of DSEE (introduced in [13] for risk neutral MAB) were shown to achieve $O(\log T)$ distribution dependent regret and $O(T^{\frac{2}{3}})$ worst-case regret, respectively. The empirical mean-variance calculated using the whole reward process is different than the actual instantaneous mean-variance considered in this paper where the variance in the decisions causes a dominant term in regret. An $\Omega(T^{\frac{2}{3}})$ lower bound on worst case regret was established in [11] which showed the order optimality of the modification of the DSEE policy. Under a non-stochastic and full feedback setting, [14] considered a linear combination of mean and standard deviation (in contrast to mean-variance) and established a negative result showing the infeasibility of sublinear regret. In [15], the quality of an arm was measured by a general function of the mean and the variance of the random variable. This study, however, is closer to the risk-neutral MAB problems than to the problem studied in this paper in that the variance in the decisions does not effect the regret in [15] as it dominantly does in our results.

In [16, 17], MAB under the measure of value at risk was studied. In [17], learning policies using the measure of conditional value at risk were developed. However, the performance guarantees were still within the risk-neutral MAB framework (in terms of the loss in the expected total reward) under the assumption that the best arm in terms of the mean value is also the best arm in terms of the conditional value at risk. Another risk measure for MAB problems was considered in [18] in which the logarithm of moment generating function was used as a risk measure and high probability bounds on regret were obtained.

2 Problem Formulation

Consider a stochastic MAB problem with a discrete set $[K] = \{1, 2, \ldots, K\}$ of actions. At each time $t$, a learner chooses an action $k \in [K]$ and receives the corresponding reward $X_{k,t}$, drawn from an unknown distribution $f_k$. The rewards are independent over $k$, and i.i.d. over $t$. Let $\mathcal{F} = \{f_k\}_{k=1}^K$ denote the set of distributions. We use $\mathbb{E}_\mathcal{F}$ and $\mathbb{P}_\mathcal{F}$ to denote the expectation and
probability with respect to $\mathcal{F}$ and drop the subscript $\mathcal{F}$ when it is clear from the context.

An arm selection policy $\pi$ specifies a sequence of mappings $\{\pi_t\}_{t \geq 1}$ from the history of observations to the arm to play at each time $t$: $\pi_t = [K]^{t-1} \times \mathbb{R}^{t-1} \to [K]$. We use $\{X_{\pi_t,t}\}_{t \geq 1}$ to denote the random reward sequence under policy $\pi$.

The mean-variance of a random variable $X$ is defined as, $\text{MV}(X) = \sigma^2(X) - \lambda \mu(X)$, a linear combination of its mean $\mu(X)$ and variance $\sigma^2(X)$. We also use the notation $\mu_k = \mu(X_{k,1})$, $\sigma^2_k = \sigma^2(X_{k,1})$ and $\text{MV}_k = \text{MV}(X_{k,1})$.

The objective is to design an arm selection policy $\pi$ which minimizes the total mean-variance of the observations or equivalently minimizes regret defined as

$$R^\pi(T) = \sum_{t=1}^T \text{MV}(X_{\pi_t,t}) - \sum_{t=1}^T \text{MV}(X_{\pi^*_t,t}),$$

(3)

where $\pi^*_t = k^* \overset{\Delta}{=} \arg \min_{k \in [K]} \text{MV}_k$, for all $t$, is the optimal policy under a known model. Unlike the risk-neutral MAB, it is not obvious that the optimal policy under known model is a single arm policy. We shall show this using Lemma 1.

Let $\mathbb{1}[.]$ denote the indicator function, meaning, for an event $\mathcal{E}$, $\mathbb{1}[\mathcal{E}] = 1$ if and only if $\mathcal{E}$ is true and $\mathbb{1}[^{\mathcal{E}}] = 0$ otherwise. Let $\tau_{k,t} = \sum_{s=1}^t \mathbb{1}[\pi_s = k]$ denote the number of times that arm $k$ has been played until time $t$.

The risk-neutral regret $\mathbb{E}[\sum_{t=1}^T (X_{k_{\text{max}},t} - X_{\pi_t,t})]$, where $k_{\text{max}} = \arg \max_{k \in [K]} \mu_k$, can be expressed as a weighted sum of $\mathbb{E}[\tau_{k,T}]$ with weights $\mu_{k_{\text{max}}} - \mu_k$. In contrast, the risk-averse regret $R^\pi(T)$ given in (3) depends also on the variance of the decisions through time. Lemma 1 provides an expression of regret which is used throughout the paper to analyze the performance of the policies. Let $\Gamma_k = \text{MV}_k - \text{MV}_{k^*}$ and $\Delta_k = \mu_k - \mu_{k^*}$.

**Lemma 1** The regret of a policy $\pi$ under the measure of total mean-variance of rewards satisfies

$$R^\pi(T) = \sum_{k=1}^K \mathbb{E}[\tau_{k,T}] \Gamma_k + \sum_{t=1}^T \mathbb{E} \left[ \left( \sum_{k \in [K] \setminus k^*} (\mathbb{1}[\pi_t = k] - \mathbb{P}[\pi_t = k]) \Delta_k \right)^2 \right].$$

(4)

**Proof.** See Appendix A.

The regret expression given in Lemma 1 shows that $R^\pi(T) \geq 0$ for any policy $\pi$, and $R_{\pi^*}(T) = 0$, which proves that the optimal single arm policy is the optimal policy under the risk-averse measure. The second term in the regret expression in (4) captures the variance of the decisions.

**Concentration Inequalities.** We assume the random variable $(X_{k,1} - \mu_k)^2 - \sigma^2_k$, for all $k$, is Sub-Gaussian with parameter $b^2$, i.e., its moment generating function is bounded by that of a Gaussian distribution with variance $b^2$:

$$\mathbb{E} \left[ \exp \left( u ((X_{k,1} - \mu_k)^2 - \sigma^2_k) \right) \right] \leq \exp \left( \frac{u^2 b^2}{2} \right).$$

4
As a result of the Chernoff-Hoeffding bound \cite{21}, we have the concentration inequalities on the sample mean and the sample mean-variance given in Lemma 2. The sample mean, the sample variance and the sample mean-variance of each arm \(k\) up to time \(t\) are, respectively, defined as 
\[
\bar{\mu}_{k,t} = \frac{1}{\tau_{k,t}} \sum_{s=1}^{t} \mathbb{I}[\pi_s = k] X_{k,s},
\]
\[
\bar{\sigma}^2_{k,t} = \frac{1}{\tau_{k,t}} \sum_{s=1}^{t} \mathbb{I}[\pi_s = k](X_{k,s} - \bar{\mu}_{k,t})^2
\]
and 
\[
\bar{\text{MV}}_{k,t} = \bar{\sigma}^2_{k,t} - \lambda \bar{\mu}_{k,t}.\]
To keep the notation uncluttered we drop the specification of the policy from \(\tau_{k,t}, \bar{\mu}_{k,t}, \bar{\sigma}^2_{k,t}\) and \(\bar{\text{MV}}_{k,t}\).

**Lemma 2 (Lemma 1 in \cite{11})**  
Let \(\bar{\text{MV}}_t\) be the sample mean-variance of a random variable \(X\) obtained from \(t\) i.i.d. observations. Let \(\mu = \mathbb{E}[X]\), \(\sigma^2 = \mathbb{E}[(X - \mu)^2]\), and assume that \((X - \mu)^2 - \sigma^2\) has a sub-Gaussian distribution, i.e.,
\[
\mathbb{E}[e^{u((X - \mu)^2 - \sigma^2)}] \leq e^{\zeta_1 u^2/2}
\]
for some constant \(\zeta_1 > 0\). As a result \(X - \mu\) has a sub-Gaussian distribution, i.e.,
\[
\mathbb{E}[e^{u(X - \mu)}] \leq e^{\zeta_0 u^2/2}.
\]
Let \(\zeta = \max\{\zeta_0, \zeta_1\}\). We have, for all constants \(\alpha \in (0, \frac{1}{2\zeta}]\) and \(\delta \in (0, 2 + \lambda]\),
\[
\begin{align*}
\mathbb{P}[\bar{\text{MV}}_t - \text{MV}(X) > \delta] &\leq 2 \exp(-\frac{\alpha \delta^2}{(2+\lambda)^2}), \\
\mathbb{P}[\bar{\text{MV}}_t - \text{MV}(X) < -\delta] &\leq 2 \exp(-\frac{\alpha \delta^2}{(2+\lambda)^2}).
\end{align*}
\]

### 3 The Proof of Lower Bounds

The regret expression given in Lemma 1 consists of two terms. The first term comes from playing suboptimal arms and the second term corresponds to the variance in the decisions. Since the second term is always positive, an \(\Omega(\log T)\) lower bound for distribution-dependent regret can be established following the similar lines as in the proof of the lower bound results for risk-neutral MAB provided in \cite{4, 5}. We thus focus on the lower bound for worst case regret which is significantly different than its counterpart in risk-neutral MAB due to dominant effect of the second term in risk averse regret under the worst case setting.

For the results presented in the rest of this section, we assume the player observes the rewards of all arms at each time \(t\), in contrast to observing only the reward of the chosen arm. Following the terminology in the literature, we refer to this case as full feedback setting. The lower bounds established under full feedback hold in the bandit setting as well (since extra information cannot worsen the performance). The following lemma is used in establishing the lower bound for worst case regret under risk-averse setting.
Lemma 3 Let $\nu$ and $\nu'$ be two probability distributions supported on some set $X$ with $\nu'$ absolutely continuous with respect to $\nu$. For any measurable function $\phi : X \to \{0, 1\}$, we have

$$\Pr_{\nu}(\phi(X) = 1) + \Pr_{\nu'}(\phi(X) = 0) \geq \frac{1}{2} \exp(-\text{KL}(\nu, \nu')).$$ (5)

Notations $\Pr_{\nu}(\cdot)$ denotes the probability measure with respect to $\nu$. Lemma 3 was used in [5] to establish a lower bound on the risk-neutral bandit with side information.

Our lower bound proof is based on a coupling argument in a 2-armed bandit. Let $F = (f_1, f_2)$ and $F' = (f_1', f_2')$ denote two different distribution assignments for a 2-armed bandit. Let $f_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, a normal distribution with mean $\mu_1 = \frac{3}{4}$ and variance $\sigma_1^2 = \frac{3}{16} - 4\Gamma^2$. Also, let $f_2 \sim \mathcal{B}(p)$, a Bernoulli distribution with $p = 1/4 + 2\Gamma$ and $f_2' \sim \mathcal{B}(q)$ with $q = 1/4 - 2\Gamma$, $\Gamma \in (0, \frac{1}{8})$. For simplicity of presentation let us assume $\lambda = 0$. Note that for the difference between the variance of the above distributions we have $\sigma_2^2 - \sigma_1^2 = \Gamma$ and $\sigma_1^2 - \sigma_2'^2 = \Gamma$, where $\sigma_k^2$ and $\sigma_k'^2$ are the variance of arm $k \in \{0, 1\}$ under $F$ and $F'$, respectively. For any arm selection policy $\pi$, we prove that, under at least one of the two systems, the number of times the suboptimal arm is played is high in expectation. Specifically, we show that $\{E_{F}[\tau_{2,T}] \lor E_{F'}[\tau_{1,T}] \} \geq \{0.01 \Gamma^2 \land \frac{T}{2e}\}$.

Lemma 4 For any arm selection policy $\pi$ with full information and any parameter $\Gamma > 0$, in the 2-armed bandit with number of rounds $T \geq 100$,

$$\{E_{F}[\tau_{2,T}] \lor E_{F'}[\tau_{1,T}] \} \geq \left\{ 0.01 \Gamma^2 \land \frac{T}{2e} \right\}.$$ (6)

Proof. See Appendix B.

Using Lemma 4 we establish a lower bound on the worst case regret performance of any policy $\pi$.

Theorem 1 For any arm selection policy $\pi$ with full information, there exists a distribution assignment $F$ to a 2-armed bandit where

$$R_{F}^{\pi}(T) \geq \frac{T}{4e}.$$ (7)

Proof. See Appendix C.

4 Risk-averse Policies

In this section, we introduce and analyze the performance of Mean-Variance Lower Confidence Bound (MV-LCB) policy and Confidence Bounds based Arm Elimination (CB-AE) policy.
4.1 MV-LCB

MV-LCB is a modification of the classic UCB policy first introduced in [7] for risk-neutral MAB and then adopted for risk-averse MABs in [12, 11]. At each time $t$, MV-LCB plays the arm with the smallest lower confidence bound on mean-variance:

$$\pi_{t}^{MV-LCB} = \arg\min_{k} \bar{\mu}_{k,t} - \sqrt{\frac{c \log t}{\tau_{k,t}}},$$

where $c$ is a constant that depends on the distribution class parameters $\alpha_2, \alpha_3$.

**Algorithm 1** MV-LCB Policy.

1: Initialization: $T \in \mathbb{N}, [K]$.
2: for $t = 1, 2, \ldots, T$ do
3: Play $\pi_{t}^{MV-LCB} = \arg\min_{k} \bar{\mu}_{k,t} - \sqrt{\frac{c \log t}{\tau_{k,t}}}$
4: end for

**Theorem 2** When there is a positive gap in the mean-variances of the best and the second best arms, for $c \geq \frac{3(2+\lambda)^2}{\alpha}$, the regret of MV-LCB policy satisfies

$$R^{\pi_{MV-LCB}}(T) \leq \sum_{k \in [K] \setminus k^*} \left( \frac{4c \log T}{\Gamma_k^2} + 5 \wedge T \right) \left( \Gamma_k + \frac{(K-1)\Delta_k^2}{4} \right).$$

Proof. See Appendix D. Theorem 2 shows a logarithmic upper bound on the distribution-dependent regret of LCB.

4.2 CB-AE

The CB-AE policy is a modification of Improved UCB introduced in [8] which proceeds in steps $n = 0, 1, 2, \ldots$. At each step $n$, a set of actions $\mathcal{K}_n$, initialized at $\mathcal{K}_0 = [K]$, are played, each $u_n = \lceil C \log T \rceil$ times where $\Gamma_n = \Gamma_0 2^{-n}$ is initialized at $\Gamma_0 > 0$ and $C > 0$ is a constant that depends only on the distribution class parameter $\alpha$. At each step, a number of actions are potentially removed from $\mathcal{K}_n$ based on upper and lower confidence bounds on their mean-variance, respectively, in the from of $\bar{\mu}_{k|u_n} + \frac{\Gamma_n}{4}$ and $\bar{\mu}_{j|u_n} - \frac{\Gamma_n}{4}$, where $\bar{\mu}_{k|u_n}$ is the sample mean-variance obtained from the $u_n$ observations at step $n$. If the lower confidence bound of arm $k$ is bigger than the minimum of the upper confidence bounds of all other remaining arms, arm $k$ is removed $\mathcal{K}_{n+1} = \mathcal{K}_n \setminus \{k\};$ see lines 6-10 in Algorithm 2.

$^1\alpha$ is the distribution class parameter specified in concentration inequalities in Lemma 2.
Algorithm 2 CB-AE Policy.

1: Initialization: $\hat{\Gamma}_0 = 1$, $n = 0$, $T \in \mathbb{N}$, $K_0 = [K]$.
2: while time is left do
3: $K_{n+1} = K_n$
4: $u_n = \lceil \frac{C \log T}{\hat{\Gamma}_n^2} \rceil$.
5: Choose each arm $k \in K_n$ for $u_n$ times.
6: for $k \in K_n$ do
7: if $\bar{MV}_k[u_n] - \hat{\Gamma}_n > \min_{j \in K_n} \bar{MV}_j[u_n] + \hat{\Gamma}_n$ then
8: Remove arm $k$: $K_{n+1} \leftarrow K_{n+1} \setminus \{k\}$.
9: end if
10: end for
11: $n = n + 1$
12: $\hat{\Gamma}_{n+1} = \frac{\hat{\Gamma}_n}{2}$
13: end while

Let $n_k = \min\{n : \hat{\Gamma}_n \leq \Gamma_k\}$ and $n_{\text{max}}$ be the number of steps taken in CB-AE. Let $\Delta_{\text{max}} = \max_{k \in [K]\setminus*} |\Delta_k|$.

**Theorem 3** The risk-averse regret performance of CB-AE policy, for $C \geq \frac{64}{\alpha}$, satisfies

$$R^{\pi_{CB-AE}}(T) \leq \sum_{k \in [K]\setminus*} \left( \frac{4C \log T}{\Gamma_k^2} + \log_2 \left( \frac{1}{\Gamma_k} \right) + \frac{K \log_2 T + 2}{T^3} \wedge T \right) \Gamma_k + \frac{1}{2} \log_2 T \Delta_{\text{max}}^2 \sum_{k \in [K]\setminus*} \left( \left( \frac{C \log T}{\Gamma_k^2} + 1 \right) \mathbb{I}[n_k \leq n_{\text{max}}] \right)$$

$$+ \left( \frac{C \log T}{\Gamma_k^2} + 1 \right) \mathbb{I}[n_k - 1 \leq n_{\text{max}}] + \left( \frac{K \log_2 T + 2}{T^4} + \frac{K \log_2 T}{T} \right) \left( \frac{(K - 1)^2 T \Delta_{\text{max}}^2}{4} \right) \right)$$ (10)

Theorem 2 shows a logarithmic upper bound on the distribution-dependent regret of CB-AE.

5 Simulations

We compare the performance of MV-LCB and CB-AE policies and experimentally verify the theoretical predictions discussed above. We simulate $K = 4$ arms, each Bernoulli distributed on a long time horizon ($T = 10000$ and repeat the experiment $M = 1000$ times; see the appendix for further details on the exact simulation setup. We modify the MV gap $\Gamma$ of the optimal and all other arms to simulate different settings where the optimal arm is quite easy to identify $\Gamma \gg 0$ or alternatively where the arms become indistinguishable as $\Gamma \to 0$. 

8
We show the performance of the two policies in terms of regret in Figure 1. As it is expected, CB-AE shows a better regret performance in the simulations in comparison to MV-LCB. The reason is that CB-AE, by fixing the arm elimination structure, reduces the variance in the decisions.

While both policies show a linear worst case regret performance, it is worth mentioning that MV-LCB has a linear regret performance for all the settings where there exists a $k \neq *$ with $\Gamma_k = O\left(\frac{1}{\sqrt{T}}\right)$ and $\Delta_k >> 0$ (for example $\Delta_k = 1$). On the other hand, CB-AE, as it can be seen from the upper bound in Theorem 3, has a linear regret for the particular case of $\Gamma_k = \Theta\left(\frac{1}{\sqrt{T}}\right)$ and $\Delta_k >> 0$. The CB-AE policy recovers the sublinear regret for the smaller values of $\Gamma_k$, that is almost equivalent good arms do not cause a linear regret in contrast to the MV-LCB case, which is a useful property from a practical perspective.

6 Conclusion

In this paper, we studied MAB problems under a mean-variance measure. We showed that a dominant term in risk-averse regret comes from the variance in the decisions. We established fundamental limits on learning policies; while a logarithmic distribution-dependent regret is achievable by UCB type policies, similar to the risk-neutral MAB, an $\Omega(T)$ worst case regret is inevitable in contrast to the $\Omega(\sqrt{T})$ counterpart lower bound in the risk-neutral setting.

Appendix A

Proof 1 (Proof of Lemma 1) We analyze the mean and the variance of the observed reward at time $t$ under policy $\pi$. For the $E[X_{\pi,t}]$ we have:

\[
E[X_{\pi,t}] = E\left[\sum_{k=1}^{K} I[\pi_t = k]X_{k,t}\right]
\]

\[
= \sum_{k=1}^{K} E[I[\pi_t = k]X_{k,t}]
\]

\[
= \sum_{k=1}^{K} E\left[ E\left[I[\pi_t = k]X_{k,t} \mid I[\pi_t = k]\right]\right]
\]

\[
= \sum_{k=1}^{K} E\left[I[\pi_t = k]E\left[X_{k,t} \mid I[\pi_t = k]\right]\right]
\]

\[
= \sum_{k=1}^{K} E\left[I[\pi_t = k]\mu_k\right]
\]

\[
= \sum_{k=1}^{K} Pr[\pi_t = k]\mu_k.
\]
Figure 1: Comparison of the performance of MV-LCB and CB-AE. In all examples MV-LCB was run with $c = 1$ and CB-AE with $\Gamma_0 = 1$ and $C = 16$. In the caption $\Gamma$, the difference in MV between the optimal and other arms, is shown. The experiment was run with a time horizon $T = 10000$ using $M = 1000$ repetitions.
Equation (11) comes from the linearity of the expectation and equation (12) is a result of the property of the conditional expectation that for two random variables $Y$ and $Z$: $E[YZ] = E[E[YZ|Z]]$.

For the variance of $X_{\pi,t}$, we have

$$
E\left[ \left( X_{\pi,t} - E[X_{\pi,t}] \right)^2 \right]
= E\left[ \left( \sum_{k=1}^{K} I[\pi_t = k]X_{k,t} - \sum_{k=1}^{K} I[\pi_t = k]E[X_{k,t}] \right)^2 \right]
= E\left[ \left( \sum_{k=1}^{K} I[\pi_t = k]X_k - \sum_{k=1}^{K} I[\pi_t = k]\mu_k + \sum_{k=1}^{K} I[\pi_t = k]\mu_k - \sum_{k=1}^{K} \Pr[\pi_t = k]\mu_k \right)^2 \right]
= E\left[ \left( \sum_{k=1}^{K} I[\pi_t = k](X_k - \mu_k) + \sum_{k=1}^{K} (I[\pi_t = k] - \Pr[\pi_t = k])\mu_k \right)^2 \right]
= E\left[ \left( \sum_{k=1}^{K} I[\pi_t = k](X_k - \mu_k) \right)^2 \right] + 2E\left[ \left( \sum_{k=1}^{K} I[\pi_t = k](X_k - \mu_k) \right)\left( \sum_{k=1}^{K} (I[\pi_t = k] - \Pr[\pi_t = k])\mu_k \right) \right] + 2\sum_{k=1}^{K} \Pr[\pi_t = k](X_k - \mu_k)^2
$$

We analyze the three term in (14) separately.
The first term:
\[
E \left[ \sum_{k=1}^{K} \mathbb{I}[\pi_t = k] (X_{k,t} - \mu_k) \right]^2
\]

\[
= E \left[ \left( \sum_{j=1}^{K} \mathbb{I}[\pi_t = j] (X_{j,t} - \mu_j) \right) \left( \sum_{k=1}^{K} \mathbb{I}[\pi_t = k] (X_{k,t} - \mu_k) \right) \right]
\]

\[
= \sum_{j=1}^{K} \sum_{k=1}^{K} E \left[ \mathbb{I}[\pi_t = j] \mathbb{I}[\pi_t = k] (X_{j,t} - \mu_j) (X_{k,t} - \mu_k) \right]
\]

\[
= \sum_{k=1}^{K} \left[ E \left[ \mathbb{I}[\pi_t = k] (X_{k,t} - \mu_k)^2 \right] + \sum_{j=1}^{K} \sum_{k \neq j} E \left[ \mathbb{I}[\pi_t = j] \mathbb{I}[\pi_t = k] (X_{j,t} - \mu_j)(X_{k,t} - \mu_k) \right] \right]
\]

\[
= \sum_{k=1}^{K} \text{Pr}[\pi_t = k] \sigma_k^2. \quad (15)
\]

The last equality is proven similar to (13).

The second term:
\[
E \left[ \left( \sum_{k=1}^{K} (\mathbb{I}[\pi_t = k] - \text{Pr}[\pi_t = k]) \mu_k \right)^2 \right]
\]

\[
= E \left[ \left( \sum_{k=1}^{K} (\mathbb{I}[\pi_t = k] - \text{Pr}[\pi_t = k]) \mu_k + (\mathbb{I}[\pi_t = k^*] - \text{Pr}[\pi_t = k^*]) \mu_{k^*} \right)^2 \right]
\]

\[
= E \left[ \left( \sum_{k \neq k^*} \mathbb{I}[\pi_t = k] - \text{Pr}[\pi_t = k] \right) \mu_k + \left( 1 - \sum_{k=1}^{K} \mathbb{I}[\pi_t = k] - (1 - \sum_{k \neq k^*} \text{Pr}[\pi_t = k]) \right) \mu_{k^*} \right)^2 \quad (16)
\]

\[
= E \left[ \left( \sum_{k \neq k^*} \mathbb{I}[\pi_t = k] - \text{Pr}[\pi_t = k] \right) \Delta_k \right]^2. \quad (17)
\]

The equation (16) holds because \( \sum_{k=1}^{K} \mathbb{I}[\pi_t = k] = 1 \) and \( \sum_{k=1}^{K} \text{Pr}[\pi_t = k] = 1 \).
The third term:
\[
\mathbb{E} \left[ \left( \sum_{k=1}^{K} \mathbb{I}[\pi_t = k] (X_k - \mu_k) \right) \left( \sum_{k=1}^{K} (\mathbb{I}[\pi_t = k] - \mathbb{P}[\pi_t = k]) \mu_k \right) \right] = \mathbb{E} \left[ \left( \sum_{k=1}^{K} \mathbb{I}[\pi_t = k] (X_k - \mu_k) \right) \left( \sum_{k=1}^{K} (\mathbb{I}[\pi_t = k] - \mathbb{P}[\pi_t = k]) \mu_k \right) \right] = 0.
\]
(18)

Combining (13), (14), (15), (17), (18), we have
\[
MV(X_{\pi_t}, t) = \sum_{k=1}^{K} \mathbb{P}[\pi_t = k] MV_k + \mathbb{E} \left[ \left( \sum_{k=1 \atop k \neq k^*}^{K} (\mathbb{I}[\pi_t = k] - \mathbb{P}[\pi_t = k]) \Delta_k \right)^2 \right].
\]
(19)

Summing up the mean variance of observations at each time and subtracting that of the optimal single arm strategy we arrive at (4).

Appendix B

Proof 2 (Proof of Lemma 4) For the KL divergence between \( f_2 \) and \( f'_2 \), we have
\[
\text{KL}(f_2, f'_2) = p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q}
\]
\[
= -\left( \frac{1}{4} + 2\Gamma \right) \log \frac{1}{4} - 2\Gamma - \left( \frac{3}{4} - 2\Gamma \right) \log \frac{3}{4} - 2\Gamma
\]
\[
= -\left( \frac{1}{4} + 2\Gamma \right) \log(1 - \frac{4\Gamma}{4 + 2\Gamma}) - \left( \frac{3}{4} - 2\Gamma \right) \log(1 + \frac{4\Gamma}{4 + 2\Gamma})
\]
\[
\leq -\left( \frac{1}{4} + 2\Gamma \right) \left( -\frac{4\Gamma}{4 + 2\Gamma} - \frac{1}{2} \left( \frac{4\Gamma}{4 + 2\Gamma} \right)^2 - \frac{1}{3} \left( \frac{4\Gamma}{4 + 2\Gamma} \right)^3 \right)
\]
\[
\leq -\left( \frac{1}{4} + 2\Gamma \right) \left( \frac{4\Gamma}{4 - 2\Gamma} + \frac{1}{2} \left( \frac{4\Gamma}{4 - 2\Gamma} \right)^2 + \frac{1}{3} \left( \frac{4\Gamma}{4 - 2\Gamma} \right)^3 \right)
\]
\[
\leq \Gamma^2 \left( \frac{8}{\frac{1}{4} + 2\Gamma} + \frac{64\Gamma}{3(\frac{1}{4} + 2\Gamma)^2} - \frac{8}{\frac{1}{4} - 2\Gamma} - \frac{64\Gamma}{3(1 + \frac{1}{8})^3(\frac{1}{4} - 2\Gamma)^2} \right)
\]
\[
\leq 22\Gamma^2.
\]
(20)

Inequality (20) is obtained based on truncated Taylor expansion of \( \log(1 + x) \) for \( x \in (-1, 1) \) and the last inequality holds for all \( \Gamma \in (0, \frac{1}{4}) \).

Let \( f_k^{(1)}(x_{k,1}, x_{k,2}, \ldots, x_{k,t}) = \prod_{s=1}^{t} f_k(x_{k,s}) \) denote the joint distribution of
the samples drawn from \( f_k \).

\[
\{ \mathbb{E}_F[\tau_2, T] \} \vee \mathbb{E}_F[\tau_1, T] \geq \frac{1}{2} (\mathbb{E}_F[\tau_2, T] + \mathbb{E}_F[\tau_1, T]) = \frac{1}{2} \sum_{t=1}^{T} \left( \mathbb{P}_F[\pi_t = 2] + \mathbb{P}_F[\pi_t = 1] \right) \\
\geq \frac{1}{2} \sum_{t=1}^{T} \exp(-\text{KL}(f_2, f'_2)) \\
= \frac{1}{2} \sum_{t=1}^{T} \exp(-\sum_{s=1}^{t} \text{KL}(f_2, f'_2)) \geq \frac{1}{2} \sum_{t=1}^{T} \exp(-22t\Gamma^2). \tag{24}
\]

Inequality (22) is obtained by Lemma 3. Inequality (23) is based on the assumption of i.i.d. samples for each arm over time, and (24) is obtained by replacing the upper bound on the \( \text{KL}(f_2, f'_2) \) from (21). To derive the desired lower bound in (6) we consider 2 cases for \( \Gamma \) as follows.

**Case 1** If \( \Gamma \leq \frac{1}{\sqrt{2T}} \), then

\[
\frac{1}{2} \sum_{t=1}^{T} \exp(-22t\Gamma^2) \geq \frac{1}{2e} T. \tag{25}
\]

**Case 2** If \( \Gamma > \frac{1}{\sqrt{2T}} \), then

\[
\frac{1}{2} \sum_{t=1}^{T} \exp(-22t\Gamma^2) \geq \frac{1}{2} \int_{x=1}^{T} \exp(-22x\Gamma^2)dx \\
= \frac{1}{44\Gamma^2} (\exp(-22\Gamma^2) - \exp(-22T\Gamma^2)) \\
= \frac{1}{44\Gamma^2} \exp(-22\Gamma^2) (1 - \exp(-22(T - 1)\Gamma^2)) \\
\geq \frac{\exp(-\frac{22}{44})}{44\Gamma^2} (1 - \exp(-\frac{T - 1}{T})) \\
\geq \frac{\exp(-\frac{99}{100})}{44\Gamma^2} (1 - \exp(-\frac{99}{100})) \geq 0.01 \frac{T}{\Gamma^2}. \tag{26}
\]

Inequality (26) holds for \( T \geq 100 \).

Combining (24), (25) and (27), we arrive at the theorem.
Appendix C

Proof 3 (Proof of Theorem 1) Let $|T| = \{1, 2, \ldots , T\}$ denote the set of time instances. For each $S \subseteq [T]$ and any policy $\pi$ in a 2-armed bandit, we construct a new policy $\pi^S$, based on $\pi$, that is obtained by altering the decision of policy $\pi$ on set $S$. In particular,

$$\begin{align*}
\pi^S_t &= \pi_t, \quad \text{if } t \notin S \\
\pi^S_t &= 3 - \pi_t, \quad \text{if } t \in S.
\end{align*}$$

(28)

In a 2-armed bandit, let $\Delta = \Delta_k$ where $k \in \{1, 2\}$ and $k \neq \ast$. In the second term in regret expression given in (4), we have

$$\begin{align*}
\mathbb{E}_F \left[ \left( \sum_{k=1}^{K} (\mathbb{I}[\pi_t = k] - \mathbb{P}_F[\pi_t = k])\Delta_k \right)^2 \right] &= \mathbb{E}_F \left[ \left( (\mathbb{I}[\pi_t \neq \ast] - \mathbb{P}_F[\pi_t \neq \ast])\Delta \right)^2 \right] \\
&= \mathbb{P}_F[\pi_t \neq \ast](1 - \mathbb{P}_F[\pi_t \neq \ast])\Delta^2.
\end{align*}$$

(29)

The first term in the regret expression given in (4), is always positive. Thus

$$R^\pi_F(T) \geq \sum_{t=1}^{T} \mathbb{P}_F[\pi_t \neq \ast](1 - \mathbb{P}_F[\pi_t \neq \ast])\Delta^2. \quad (30)$$

Since $\pi^S_t = \pi_t$ or $1 - \pi_t$, we have, for all $S \subseteq [T]$

$$\begin{align*}
\mathbb{P}_F[\pi^S_t \neq \ast](1 - \mathbb{P}_F[\pi^S_t \neq \ast])\Delta^2 &= \mathbb{P}_F[\pi_t \neq \ast](1 - \mathbb{P}_F[\pi_t \neq \ast])\Delta^2.
\end{align*}$$

(31)

By construction of $\{\pi^S\}_{S \subseteq [T]}$, there exists a $S_0 \subseteq [T]$ that $\mathbb{P}_F[\pi^S_{t_0} \neq \ast] \leq \frac{1}{2}$ for all $t \in [T]$. For $S_0$, we have

$$\begin{align*}
\sum_{t=1}^{T} \mathbb{P}_F[\pi^S_{t} \neq \ast](1 - \mathbb{P}_F[\pi^S_{t} \neq \ast])\Delta^2 \\
&\geq \frac{1}{2} \sum_{t=1}^{T} \mathbb{P}_F[\pi^S_{t} \neq \ast].
\end{align*}$$

(32)

From Theorem 4, there exists a distributing $F$ for a 2-armed bandit where

$$\sum_{t=1}^{T} \mathbb{P}_F[\pi^S_{t} \neq \ast] \geq \left( \frac{0.01}{T^2} \wedge \frac{T}{2e} \right).$$

(33)
Thus, combining (29), (30), (31) and (32), there exists a distribution $F$ for a 2-armed bandit where

$$R^\pi_F(T) \geq \sum_{t=1}^{T} \Pr_{F}[\pi_t \neq \star] (1 - \Pr_{F}[\pi_t \neq \star]) \Delta^2$$

$$= \sum_{t=1}^{T} \Pr_{F}[\pi_t S_0 \neq \star] (1 - \Pr_{F}[\pi_t S_0 \neq \star]) \Delta^2$$

$$\geq \frac{1}{2} \sum_{t=1}^{T} \Pr_{F}[\pi_t S_0 \neq \star] \Delta^2$$

$$\geq \left\{ \frac{0.005}{\Gamma^2} \wedge \frac{T}{4e} \right\} \Delta^2.$$

Choosing $\Gamma = \sqrt{\frac{0.02e}{T}}$, we have

$$R^\pi_F(T) \geq \frac{T}{4e}. \quad \text{(33)}$$

**Appendix D**

**Proof 4 (Proof of Theorem 2)** From the regret expression given in (4), we have

$$R^\pi(T) = \sum_{k=1}^{K} E[\tau_k,T] \Delta_k + \sum_{k=1}^{K} E\left[ \left( \sum_{t=1}^{T} (I[\pi_t = k] - \Pr[\pi_t = k]) \Delta_k \right)^2 \right]$$

$$\leq \sum_{k=1}^{K} E[\tau_k,T] \Delta_k + (K - 1) \sum_{t=1}^{T} \sum_{k \in [K] \setminus k^*} E\left[ (I[\pi_t = k] - \Pr[\pi_t = k])^2 \right] \Delta_k^2$$

$$= \sum_{k=1}^{K} E[\tau_k,T] \Delta_k + (K - 1) \sum_{k \in [K] \setminus k^*} \sum_{t=1}^{T} \Pr[\pi_t = k] (1 - \Pr[\pi_t = k]) \Delta_k^2$$

$$= \sum_{k=1}^{K} E[\tau_k,T] \Delta_k + (K - 1) \sum_{k \in [K] \setminus k^*} \sum_{t=1}^{T} \Pr[\pi_t = k] \wedge \frac{1}{4} \Delta_k^2. \quad \text{(34)}$$

Following the similar line in the analysis of the performance of UCB in [7].
and mean-variance UCB in [11] let $b_k = \frac{4c\log T}{\Gamma_k^2}$. We have

$$\bar{\mathcal{MV}}_{k, t} - \sqrt{c\log t} \tau_{k, t} - (\bar{\mathcal{MV}}_{*, t} - \sqrt{c\log t} \tau_{*, t})$$

$$= (\bar{\mathcal{MV}}_{k, t} + \sqrt{c\log t} \tau_{k, t} - MV_k) - (\bar{\mathcal{MV}}_{*, t} - \sqrt{c\log t} \tau_{*, t} - MV_*)$$

$$+ (MV_k - MV_*) - 2\sqrt{c\log t} \tau_{*, t})$$

(35)

For $\tau_{k, t} \geq b_k$, the third term in (35) is positive. Thus, when $\tau_{k, t} \geq b_k$,

$$\Pr[\pi_t = k] = \Pr[\bar{\mathcal{MV}}_{k, t} - \sqrt{c\log t} \tau_{k, t} - (\bar{\mathcal{MV}}_{*, t} - \sqrt{c\log t} \tau_{*, t}) \leq 0]$$

$$\leq \Pr[\bar{\mathcal{MV}}_{k, t} + \sqrt{c\log t} \tau_{k, t} - MV_k \leq 0] + \Pr[\bar{\mathcal{MV}}_{*, t} - \sqrt{c\log t} \tau_{*, t} - MV_* \geq 0]$$

$$\leq 4\exp(-\frac{\alpha c\log t}{(2+\rho)^2})$$

$$\leq 4t^{-3}.$$  

Where the last inequality is obtained by Lemma 2. We thus have

$$E[\tau_{k, T}] \leq b_k + \sum_{t=b_k+1}^{T} 4t^{-3}$$

$$\leq \frac{4c\log T}{\Gamma_k^2} + 5$$

(36)

In the second term in (34), we have

$$\sum_{t=1}^{T} \{\Pr[\pi_t = k] \wedge \frac{1}{4}\} \leq \frac{1}{4} b_k + \sum_{t=b_k+1}^{T} 4t^{-3}$$

$$\leq \frac{c\log T}{\Gamma_k^2} + 5$$

(37)

Combining (34), (36), and (37), we arrive at the theorem.

Appendix E

Proof 5 (Proof of Theorem 3) To analyze the performance of CB-AE policy, we establish the following three facts:
Fact 1. The probability that the best arm is eliminated at a step $n$ by a suboptimal arm is upper bounded by $\frac{K}{T^4}$: for $k \neq \ast$,

$$
\Pr \left[ \bar{\mathcal{M}}_\ast[u_n] - \frac{\hat{\Gamma}_n}{4} > \min_{j \in \mathbb{K}_n} \bar{\mathcal{M}}_j[u_n] + \frac{\hat{\Gamma}_n}{4} \right] 
\leq \sum_{k \in \mathbb{K}_n \setminus k} \Pr \left[ \bar{\mathcal{M}}_\ast[u_n] - \frac{\hat{\Gamma}_n}{4} > \bar{\mathcal{M}}_k[u_n] + \frac{\hat{\Gamma}_n}{4} \right] 
\leq \sum_{k \in \mathbb{K}_n \setminus k} \Pr \left[ \bar{\mathcal{M}}_\ast[u_n] - \mathcal{M}_\ast > \frac{\hat{\Gamma}_n}{4} \text{ or } \bar{\mathcal{M}}_k[u_n] - \mathcal{M}_k < -\frac{\hat{\Gamma}_n}{4} \right] 
\leq \Pr \left[ \bar{\mathcal{M}}_\ast[u_n] - \mathcal{M}_\ast > \frac{\hat{\Gamma}_n}{4} \right] + \sum_{k \in \mathbb{K}_n \setminus k} \Pr \left[ \bar{\mathcal{M}}_k[u_n] - \mathcal{M}_k < -\frac{\hat{\Gamma}_n}{4} \right] 
\leq K \exp \left( -\frac{\alpha u_n \hat{\Gamma}_n^2}{16} \right) 
\leq K \frac{1}{T^4}.
$$

Fact 2. Conditioned on the probability that the optimal arm is not eliminated, the probability that a suboptimal arm $k$ is not eliminated at step $n$ where $\hat{\Gamma}_n < \Gamma_k$ is also upper bounded by $\frac{2}{T^4}$: for $k \neq \ast$,

$$
\Pr \left[ \bar{\mathcal{M}}_k[u_n] - \frac{\hat{\Gamma}_n}{4} < \min_{j \in \mathbb{K}_n} \bar{\mathcal{M}}_j[u_n] + \frac{\hat{\Gamma}_n}{4} \right] 
\leq \Pr \left[ \bar{\mathcal{M}}_k[u_n] - \frac{\hat{\Gamma}_n}{4} < \bar{\mathcal{M}}_\ast[u_n] + \frac{\hat{\Gamma}_n}{4} \right] 
\leq \Pr \left[ \bar{\mathcal{M}}_\ast[u_n] - \mathcal{M}_\ast > \frac{\hat{\Gamma}_n}{4} \text{ or } \bar{\mathcal{M}}_k[u_n] - \mathcal{M}_k < -\frac{\hat{\Gamma}_n}{4} \right] 
\leq \Pr \left[ \bar{\mathcal{M}}_\ast[u_n] - \mathcal{M}_\ast > \frac{\hat{\Gamma}_n}{4} \right] + \Pr \left[ \bar{\mathcal{M}}_k[u_n] - \mathcal{M}_k < -\frac{\hat{\Gamma}_n}{4} \right] 
\leq 2 \exp \left( -\frac{\alpha u_n \hat{\Gamma}_n^2}{16} \right) 
\leq 2 \frac{1}{T^4}.
$$

Fact 3. Conditioned on the probability that the optimal arm is not eliminated, the probability that a suboptimal arm is eliminated at a step $n$ where $\hat{\Gamma}_n > 4\Gamma_k$
is upper bounded by $\frac{K}{T}$: for $k \neq \ast$,

$$
\Pr \left[ \tilde{\mathcal{M}}_k[u_n] - \frac{\hat{\Gamma}_n}{4} > \min_{j \in K_n} \tilde{\mathcal{M}}_j[u_n] + \frac{\hat{\Gamma}_n}{4} \right] 
\leq \sum_{j \in K_n \setminus k} \Pr \left[ \tilde{\mathcal{M}}_k[u_n] - \frac{\hat{\Gamma}_n}{4} > \tilde{\mathcal{M}}_j[u_n] + \frac{\hat{\Gamma}_n}{4} \right] 
\leq \sum_{j \in K_n \setminus k} \Pr \left[ \tilde{\mathcal{M}}_k[u_n] - \mathcal{M}_k > \frac{\hat{\Gamma}_n}{8} \text{ or } \tilde{\mathcal{M}}_j[u_n] - \mathcal{M}_j < -\frac{\hat{\Gamma}_n}{8} \right] 
\leq \Pr \left[ \tilde{\mathcal{M}}_k[u_n] - \mathcal{M}_k > \frac{\hat{\Gamma}_n}{8} \right] + \sum_{j \in K_n \setminus k} \Pr \left[ \tilde{\mathcal{M}}_j[u_n] - \mathcal{M}_j < -\frac{\hat{\Gamma}_n}{8} \right] 
\leq K \exp \left( -\frac{\alpha u_n \hat{\Gamma}^2}{64} \right) 
\leq \frac{K}{T}.
$$

Let $n_k = \min\{n : \hat{\Gamma}_n \leq \Gamma_k\}$ and let $n_{\text{max}}$ be the total number of steps at time $T$. Clearly, $n_{\text{max}} \leq \log_2 T$. Using Facts 1 and 2, we have, for $k \neq \ast$,

$$
\mathbb{E}[\tau_{k,T}] 
\leq \sum_{m=1}^{n_k} u_n + (\Pr[\text{the best arm is eliminated by a suboptimal arm}])T 
+ (\Pr[\text{arm } k \text{ is not eliminated at (or before) step } n_k])T 
\leq \sum_{m=1}^{n_k} \left( \frac{C \log T}{\Gamma^2_k} + 1 \right) + \frac{Kn_{\text{max}} + 2}{T^3} 
\leq \sum_{m=1}^{n_k} \frac{C \log T}{\Gamma^2_k} \left( \frac{1}{4} \right)^{m-1} + \frac{K \log_2 T + 2}{T^3} 
\leq \frac{4C}{3} \frac{\log T}{\Gamma^2_k} + \log_2 \left( \frac{1}{\Gamma_k} \right) + \frac{K \log_2 T + 2}{T^3}. 
$$

The first term in regret expression given in (4) is thus upper bounded by

$$
\sum_{k \in |K| \setminus k^*} \left( \frac{4C}{3} \log \frac{T}{\Gamma^2_k} + \log_2 \left( \frac{1}{\Gamma_k} \right) + \frac{K \log_2 T + 2}{T^3} \right) \Gamma_k 
\leq \frac{4C}{3} \frac{\log T}{\Gamma^2_k} + \log_2 \left( \frac{1}{\Gamma_k} \right) + \frac{K \log_2 T + 2}{T^3} \wedge T \Gamma_k 
$$

(38)

The second term of the upper bound on regret given in (34) is upper bounded
by

$$(K - 1) \sum_{k \in [K] \setminus k^*} \sum_{t=1}^{T} \{ \Pr[\pi_t = k] \land \frac{1}{4} \} \Delta_k^2$$

$$\leq (K - 1) \sum_{k \in [K] \setminus k^*} u_{nk} - 1 \mathbb{I}[n_k - 1 \leq n_{\max}] (n_{\max} - n_k + 1) \frac{2\Delta_{\max}^2}{4}$$

$$+ (K - 1) \sum_{k \in [K] \setminus k^*} u_{nk} \geq 2 \mathbb{I}[n_k - 2 \leq n_{\max}] (n_{\max} - n_k + 2) \frac{2\Delta_{\max}^2}{4}$$

$$+ \left( \Pr[\text{the best arm is eliminated by a suboptimal arm}] + \Pr[\text{arm } k \text{ is not eliminated at (or before) step } n_k] + \Pr[\text{arm } k \text{ is eliminated at (or before) step } n_k - 3] \right) \frac{T\Delta_{\max}^2}{4}$$

$$= \frac{1}{2} \log_2 T \Delta_{\max}^2 \sum_{k \in [K] \setminus k^*} \left( \left( \frac{C \log T}{\Gamma_k^2} + 1 \right) \mathbb{I}[n_k \leq n_{\max}] \right)$$

$$+ \left( \frac{C \log T}{\Gamma_k^2} + 1 \right) \mathbb{I}[n_k - 1 \leq n_{\max}]$$

$$+ \left( \frac{K \log_2 T + 2}{T^4} + \frac{K \log_2 T}{T} \right) (K - 1)^2 \frac{T\Delta_{\max}^2}{4}.$$  (39)

Combining (38) and (39), we arrive at the theorem.

Appendix F

We provide further detail on the simulation experiment presented in Section 5. The $K = 4$ arms are Bernoulli distributed with mean $\mu_1 = 1$ for the optimal arm and $\mu_k = 2$ for the other 3 arms. The variance of the former is $\sigma_1^2 = 1$ and for latter is changed for each experiment to be \{2.5, 2.2, 2.1, 2.05, 2.01, 2.0\} simulating different $\Gamma$ values.

References

[1] W. Thompson, “On the likelihood that one unknown probability exceeds another in view of the evidence of two samples,” *Biometrika* 25, pp. 285–294, 1933.

[2] M. C. Steinbach, “Markowitz Revisited: Mean-variance Models in Financial Portfolio Analysis,” *SIAM Review* vol. 43, no. 1, pp. 31-85, 2001.

[3] H. Markowitz, “Portfolio selection,” *The Journal of Finance*, vol. 7, no. 1 pp. 77-91, 1952.
[4] T. Lai, H. Robbins, “Asymptotically Efficient Adaptive Allocation Rules,” 
Advances in Applied Mathematics, vol. 6, no. 1, pp. 4-22, 1985.

[5] S. Bubeck, V. Perchet, P. Rigollet, “Bounded Regret in Stochastic Multi-
armed Bandits,” available at http://arxiv.org/abs/1302.1611, 2013.

[6] P. Auer, N. Cesa-Bianchi, Y. Freund, R. E. Schapire, “The Non-stochastic 
Multi-armed Bandit Problem,” SIAM Journal on Computing, Vol. 32, pp. 48-
77, 2003.

[7] P. Auer, N. Cesa-Bianchi, P. Fischer, “Finite-time Analysis of the Multi-
armed Bandit Problem,” Machine Learning, vol. 47, pp. 235-256, 2002.

[8] P. Auer, R. Ortner, “UCB revisited: Improved Regret Bounds for the 
Stochastic Multi-armed Bandit Problem,” Periodica Mathematica Hunga-
rica, vol. 61, no. 1-2, pp. 55-65, September, 2010.

[9] S. Bubeck, N. Cesa-Bianchi, “Regret Analysis of Stochastic and Nonstochas-
tic Multi-armed Bandit Problems,” Foundations and Trends in Machine 
Learning, vol. 5, 2012.

[10] D. Bergemann, J. Valimaki, “Bandit Problems”, Cowles Foundation Dis-
ussion Paper, No. 1551, 2006.

[11] S. Vakili, Q. Zhao, “Risk-Averse Multi-Armed Bandit Problems under 
Mean-Variance Measure,” IEEE Journal of Selected Topics in Signal Pro-
cessing (JSTSP): Special Issue on Financial Signal Processing and Machine 
Learning for Electronic Trading, vol. 10, no. 6, pp. 1093-1111, 2016.

[12] A. Sani, A Lazaric, R Munos, “Risk Aversion in Multi-armed Bandits,” 
Neural Information Processing Systems (NIPS), 2012.

[13] S. Vakili, K. Liu, Q. Zhao, “Deterministic Sequencing of Exploration and 
Exploitation for Multi-Armed Bandit Problems,” IEEE Journal of Selected 
Topics in Signal Processing, vol. 7, no. 5, pp. 759 - 767, 2013.

[14] E. Even-Dar, M. Kearns, J. Wortman, “Risk-sensitive Online Learning,” 
17th international conference on Algorithmic Learning Theory (ALT-06), 
pp. 199-213, 2006.

[15] A. Zimin, R. Ibsen-Jensen, K. Chatterjee, “Generalized Risk-
Aversion in Stochastic Multi-Armed Bandits,” available at 
http://arxiv.org/abs/1405.0833, 2014.

[16] S. Vakili, Q. Zhao, “Mean-Variance and Value at Risk in Multi-Armed Ban-
dit Problems,” 53rd Annual Allerton Conference on Communication, Control, 
and Computing, 2015.

[17] N. Galichet, M. Sebag, O. Teytaud, “Exploration vs Exploitation vs Safety: 
Risk-averse Multi-Armed Bandits,” Asian Conference on Machine Learning, 
2013.
[18] O. Maillard, “Robsut Risk-Averse Stochastic Multi-armed Bandits,” Algorithmic Learning Theory, vol 8139, pp. 218-233.

[19] J. Y. Yu, E. Nikolova, “Sample Complexity of Risk-averse Bandit-arm Selection,” 23rd international joint conference on Artificial Intelligence, pp. 2576-2582, 2013.

[20] L. Tran-Thanh, J. Y. Yu, “Functional Badits,” available at http://arxiv.org/abs/1405.2432, 2014.

[21] R. G. Antonioni, Y. Kozachenko, A. Volodin, “Convergence of series of dependent $\phi$-subgaussian random variables,” Journal of Mathematical Analysis and Applications, vol. 338, no. 2, pp. 1188-1203, 2008.

[22] P. Chareka, O. Chareka, S. Kennendy, “Locally Sub-Gaussian Random Variable and the Strong Law of Large Numbers,” Atlantic Electronic Journal of Mathematics, vol. 1, no. 1, pp. 75-81, 2006.

[23] R. Vershynin, “Introduction to the Non-Asymptotic Analysis of Random Matrices,” available at http://arxiv.org/abs/1011.3027v6, 2011.