This item was submitted to [Loughborough's Research Repository](#) by the author.
Items in Figshare are protected by copyright, with all rights reserved, unless otherwise indicated.

Risk-aware multi-armed bandits with refined upper confidence bounds

PLEASE CITE THE PUBLISHED VERSION

<https://doi.org/10.1109/lsp.2020.3047725>

PUBLISHER

Institute of Electrical and Electronics Engineers (IEEE)

VERSION

AM (Accepted Manuscript)

PUBLISHER STATEMENT

Personal use is permitted, but republication/redistribution requires IEEE permission. See
http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

LICENCE

All Rights Reserved

REPOSITORY RECORD

Liu, Xingchi, Mahsa Derakhshani, Sangarapillai Lambotharan, and Mihaela Van der Schaar. 2020. "Risk-aware Multi-armed Bandits with Refined Upper Confidence Bounds". Loughborough University.
<https://hdl.handle.net/2134/13516655.v1>.

Risk-Aware Multi-Armed Bandits with Refined Upper Confidence Bounds

Xingchi Liu, Mahsa Derakhshani, *Senior Member, IEEE*, Sangarapillai Lambotharan, *Senior Member, IEEE*, and Mihaela van der Schaar, *Fellow, IEEE*

Abstract—The classical multi-armed bandit (MAB) framework studies the exploration-exploitation dilemma of the decision-making problem and always treats the arm with the highest expected reward as the optimal choice. However, in some applications, an arm with a high expected reward can be risky to play if the variance is high. Hence, the variation of the reward should be considered to make the arm-selection process risk-aware. In this paper, the mean-variance metric is investigated to measure the uncertainty of the received rewards. We first study a risk-aware MAB problem when the reward follows a Gaussian distribution, and a concentration inequality on the variance is developed to design a Gaussian risk aware-upper confidence bound algorithm. Furthermore, we extend this algorithm to a novel asymptotic risk aware-upper confidence bound algorithm by developing an upper confidence bound of the variance based on the asymptotic distribution of the sample variance. Theoretical analysis proves that both proposed algorithms achieve the $\mathcal{O}(\log(T))$ regret. Finally, numerical results demonstrate that our algorithms outperform several risk-aware MAB algorithms.

Index Terms—Multi-armed bandit, risk-aware bandit, exploration and exploitation

I. INTRODUCTION

The multi-armed bandit (MAB) models the online sequential decision making problem for various applications including financial portfolio design, online recommendations, and crowdsourcing systems. In a standard MAB, a random reward of an arm can be observed once played by an agent, and the objective is to maximize the cumulative rewards in a certain number of plays. The main challenge is how to balance the tradeoff between exploiting the existing knowledge of arms to get a considerable cumulative reward and further exploring arms to gain more information for potentially higher revenue. The standard MAB always treats the arm with the highest expected reward as the optimal choice. However, in many applications, not only the expected rewards of arms but also the uncertainty of the rewards imposed by the variations are important. For instance, in clinical trials, instead of choosing a treatment which reaches the best average therapeutic result but may occasionally lead to unacceptably poor results, a treatment that works consistently well for every patient is more reliable and hence desirable. Therefore, for such applications, the tradeoff between the expected rewards and the variances of arms should be considered in a risk-aware MAB framework.

This work was supported by the Royal Academy of Engineering under the Leverhulme Trust Research Fellowship scheme (Derakhshani-LTRF1920\16\67). X. Liu, M. Derakhshani and S. Lambotharan are with Signal Processing and Networks Research Group, Wolfson School of Mechanical, Electrical and Manufacturing Engineering, Loughborough University, U.K (email:X.Liu3@lboro.ac.uk; M.Derakhshani@lboro.ac.uk; S.Lambotharan@lboro.ac.uk). M. van der Schaar is with the Faculty of Engineering, University of California at Los Angeles, U.S and the Faculty of Engineering, University of Cambridge, U.K (email:mihaela@ee.ucla.edu).

One of the mainstream measures to balance the tradeoff between maximizing the expected reward and minimizing the uncertainty of the reward is the mean-variance (MV) [1]. This paradigm considers a linear combination of the mean and the variance of the reward when determining the optimal arm. In [2], an algorithm is designed based on the MV paradigm to minimize the cumulative learning regret by deriving a lower confidence bound of the MV. However, their proposed MV lower confidence bound (MV-LCB) algorithm achieves a $\mathcal{O}(\log^2 T)$ learning regret which is worse than the classical risk-neutral MAB algorithms. In [3]–[5], finer analyses of the theoretical performance of MV-LCB are presented, and a new definition of the cumulative learning regret is derived for the MV measurement. By extending MV-LCB, a new algorithm is designed to reach a $\mathcal{O}(\log(T))$ regret. However, the regret bound only holds for a limited class of reward distributions. Another way to measure the risk is to use conditional value at risk level (CVaR) [6]–[9]. In [6], it is proved that the learning regret of the proposed CVaR-based scheme is $\mathcal{O}(\log(T^2))$.

In this paper, we study the risk-aware MAB problem based on the MV paradigm and propose two novel risk-aware MAB algorithms. The key difference of the proposed algorithms from the existing ones (which utilize Hoeffding inequality to drive the confidence bound of the reward variance) is that a finer confidence bound is derived by determining the distribution of the sample variance. Focusing on the MABs with continuous reward distributions, we first build a finer upper confidence bound (UCB) of the MV with a Gaussian reward assumption and design a Gaussian risk aware-upper confidence bound (GRA-UCB) algorithm to solve the risk-aware MAB. We prove that GRA-UCB can reach a $\mathcal{O}(\log(T))$ regret. Next, utilizing the asymptotic distribution of the empirical variance and by extending the GRA-UCB algorithm, a novel asymptotic risk aware-upper confidence bound (ARA-UCB) algorithm is generally designed for sub-Gaussian reward distributions and also proved to achieve $\mathcal{O}(\log(T))$ learning regret. Both proposed algorithms perform numerically well.

II. PROBLEM FORMULATION

Consider a risk-aware MAB problem with K arms, i.e., $\mathcal{K} = \{1, 2, \dots, K\}$. Playing each of the arms yields a continuous random reward sampled from an independent distribution. In risk-aware MABs, the risk of receiving a very low reward is considered and the agent prefers playing the arm with a higher mean and lower uncertainty. To measure the risk, in [2], the MV of an arm i is defined as

$$\eta_i = \sigma_i^2 - \rho\mu_i, \quad (1)$$

where μ_i and σ_i^2 are the mean and variance of the reward of arm i respectively. In (1), $\rho \geq 0$ is a risk-tolerance factor introduced to balance reward-risk tradeoff. As $\rho \rightarrow \infty$, the risk-aware MAB problem degenerates to a risk-neutral MAB, and when $\rho = 0$, the problem aims to find the arm with the lowest risk. Defining the arm played at t following the arm-selection policy π as $\pi(t)$, the observed reward is denoted as $r_{\pi(t)}(t)$. After playing arms for T rounds, the cumulative MV following the policy π can be calculated as

$$\eta_\pi(T) = \mathbb{E} \left[\sum_{t=1}^T \left((r_{\pi(t)}(t) - \frac{1}{T} \sum_{t=1}^T r_{\pi(t)}(t))^2 - \rho r_{\pi(t)}(t) \right) \right], \quad (2)$$

The objective of the risk-aware MAB problem is to minimize $\eta_\pi(T)$ under the given risk-tolerance factor ρ . In order to demonstrate the performance of an arm-selection policy, the optimal policy π^* which has the full knowledge of arms is used as the benchmark to calculate the performance gap between the optimal policy and a proposed policy π . The performance gap is referred to as the learning regret defined as

$$\text{Reg}_\pi(T) = \eta_\pi(T) - \eta_{\pi^*}(T). \quad (3)$$

In risk-aware MAB, as proved in [5], simply playing the arm with the lowest MV may not be the optimal policy but only a proxy of the optimal policy, that is in general intractable.

III. PROPOSED ALGORITHM

Let us first define some notations used in the proposed algorithm. Denoting $r_i(t)$ as the received reward by playing arm i at round t , the empirical mean $\bar{\mu}_i(t)$ and the empirical variance $s_i^2(t)$ of arm i until round t are defined as

$$\bar{\mu}_i(t) = \frac{1}{\tau_i(t)} \sum_{d=1}^{\tau_i(t)} r_i(t_i(d)), \quad (4)$$

$$s_i^2(t) = \frac{1}{\tau_i(t) - 1} \sum_{d=1}^{\tau_i(t)} (r_i(t_i(d)) - \bar{\mu}_i(t))^2, \quad (5)$$

where $t_i(d)$ represents the round when the d^{th} reward of arm i is observed by the agent and $\tau_i(t)$ denotes the number of times arm i has been played till round t . Accordingly, the empirical MV can be calculated as $\bar{\eta}_i = s_i^2(t) - \rho \bar{\mu}_i(t)$.

We design two risk-aware bandit algorithms, namely, GRA-UCB and ARA-UCB, depending on whether the reward distribution knowledge is available or not. Both algorithms are index-based policies which assign different indexes to each arm. The indexes represent the estimation of the UCB of the MV based on the historical observations, and the arm with the lowest index is played. The key difference is that for GRA-UCB, the reward is assumed to follow Gaussian distribution so the UCB is specifically derived for this case. However, in ARA-UCB, an asymptotic UCB of the variance is derived in the absence of the reward distributions knowledge.

A. Gaussian risk aware-upper confidence bound (GRA-UCB)

Suppose that the rewards of arms follow a Gaussian distribution with different means and variances. We utilize the following facts to design an arm-selection algorithm.

Fact 1. (Hoeffding inequality) Let $(X - \mu)$ be a sub-Gaussian random variable. Define the empirical mean over n samples

as $\bar{\mu}_n$ and $\mu = \mathbb{E}[X]$, then

$$\Pr \{ \bar{\mu}_n \geq \mu + \delta \} \leq e^{-2n\delta^2}, \Pr \{ \bar{\mu}_n \leq \mu - \delta \} \leq e^{-2n\delta^2}. \quad (6)$$

Fact 1 states that the deviation δ of the empirical mean from the true mean after n samples is bounded by $e^{-2n\delta^2}$. Here, setting $\delta = \sqrt{\frac{\log t}{\tau_i(t)}}$, then we have

$$\Pr \{ \bar{\mu}_n \geq \mu + \sqrt{\frac{\log t}{\tau_i(t)}} \} \leq \frac{1}{t^2}, \Pr \{ \bar{\mu}_n \leq \mu - \sqrt{\frac{\log t}{\tau_i(t)}} \} \leq \frac{1}{t^2}. \quad (7)$$

Fact 2. Let X be a Gaussian random variable with variance σ^2 . Define the empirical variance over n samples as s_n^2 , based on [10] we have

$$\Pr \{ \sigma^2 \leq \frac{(n-1)s_n^2}{\chi_{1-\alpha, n-1}^2} \} = 1 - \alpha, \quad (8)$$

where $\chi_{1-\alpha, n-1}^2$ is the upper 100α percentage points of the chi-square distribution with $(n-1)$ degrees of freedom.

Fact 2 gives a definition of the confidence interval of the variance of a random variable when the variable follows the Gaussian distribution. It implies that there is a probability of $100(1-\alpha)\%$ that the constructed confidence interval based on the sample variance will contain the true value of σ^2 .

In GRA-UCB algorithm, each arm is assigned with an index which is the UCB of the MV. Based on Facts 1 and 2, the index of arm i is defined as

$$B_i(t) = \frac{(\tau_i(t) - 1)s_i^2(t)}{\chi_{1-\alpha, \tau_i(t)-1}^2} - \rho \left(\bar{\mu}_i(t) + \sqrt{\frac{\log t}{\tau_i(t)}} \right), \quad (9)$$

where the first and second terms represent the UCBs of the variance and the mean based on (8) and (7), respectively. In each round, after calculating indexes, the arm with the lowest $B_i(t)$ will be played and the number of plays of this arm are updated as $\tau_i(t+1) = \tau_i(t) + 1$. Subsequently, the empirical mean and variance are updated based on (4) and (5).

B. Asymptotic risk aware-upper confidence bound (ARA-UCB)

GRA-UCB is designed under the assumption that the reward follows an independent Gaussian distribution. However, when the reward distribution is unknown, the confidence interval presented in (8) is not pertinent. Following the idea of GRA-UCB, and in this section, we utilize the asymptotic distribution of the empirical variance to drive a confidence interval without any prior assumption of the reward distribution.

Fact 3. Let X be a continuous random variable with mean μ , variance σ^2 , and $\mu_4 = \mathbb{E}[(X - \mu)^4]$. According to [11], the asymptotic distribution of the empirical variance is

$$\sqrt{n}(s_n^2 - \sigma^2) \xrightarrow{d} \mathcal{N}(0, \mu_4 - \sigma^4). \quad (10)$$

Based on Fact 3, in the following lemma, we develop an asymptotic UCB on the variance.

Lemma 1. Applying Fact 3, for a sufficiently large n , an asymptotic confidence interval of the variance can be derived as

$$\Pr \{ \sigma^2 \geq v_n^u \} \leq \alpha, \quad (11)$$

where $v_n^u = \frac{ns_n^2 + \sqrt{n\chi_{\alpha,1}^2(\mu_4 - s_n^4) + (\chi_{\alpha,1}^2)^2\mu_4}}{n + \chi_{\alpha,1}^2}$ and s_n^4 is the empirical estimate of σ^4 .

Proof: The distribution presented in Fact 3 can be reformulated as a standard Gaussian distribution as

$$\sqrt{n} (s_n^2 - \sigma^2) / \sqrt{\mu_4 - \sigma^4} \xrightarrow{d} \mathcal{N}(0, 1). \quad (12)$$

Therefore we have $n (s_n^2 - \sigma^2)^2 / (\mu_4 - \sigma^4) \xrightarrow{d} \chi_1^2$. Consequently, the one-sided confidence interval is defined as

$$\Pr\left\{\frac{n (s_n^2 - \sigma^2)^2}{\mu_4 - \sigma^4} \leq \chi_{\alpha,1}^2\right\} = 1 - \alpha. \quad (13)$$

Consequently, the $(1 - \alpha)$ asymptotic confidence interval of the variance σ^2 is established as

$$\Pr\{v_n^l \leq \sigma^2 \leq v_n^u\} = 1 - \alpha, \quad (14)$$

where $v_n^l = \frac{ns_n^2 - \sqrt{n\chi_{\alpha,1}^2(\mu_4 - s_n^4) + (\chi_{\alpha,1}^2)^2 \mu_4}}{n + \chi_{\alpha,1}^2}$. From (14), obviously we have $\Pr\{\sigma^2 \geq v_n^u\} \leq \alpha$, which completes the proof.

Since μ_4 is unknown, an estimate of μ_4 is required. We use $\bar{\mu}_4 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{\mu})^4$ in the asymptotic UCB provided in Lemma 1 to define a refined bound of the MV compared with the GRA-UCB policy. Therefore, an alternative index will be assigned to each arm to determine arm-selection. The index of arm i is defined as

$$D_i(t) = v'_{\tau_i(t)} - \rho \left(\bar{\mu}_i(t) + \sqrt{\log t / \tau_i(t)} \right), \quad (15)$$

where $v'_{\tau_i(t)} = \frac{ns_{\tau_i(t)}^2 + \sqrt{\tau_i(t)\chi_{\alpha,1}^2(\bar{\mu}_4 - s_{\tau_i(t)}^4) + (\chi_{\alpha,1}^2)^2 \bar{\mu}_4}}{\tau_i(t) + \chi_{\alpha,1}^2}$. Similar to GRA-UCB algorithm, the agent chooses the arm with the smallest index to play.

IV. THEORETICAL ANALYSIS

In this section, we study the regret analysis of the proposed algorithms. We provide a complete analysis for the learning regret of the proposed GRA-UCB algorithm. For the ARA-UCB algorithm, Fact 2 is replaced with Lemma 1 to drive the bound of learning regret. The analysis is similar to the GRA-UCB case and the theoretical bound still holds. In Section II, we mentioned that the optimal arm-selection policy is intractable in risk-aware bandit. Therefore, we measure the learning regret by using an approximated policy $\tilde{\pi}^*$ which keeps playing the arm with the smallest MV.

A. Learning regret of GRA-UCB

Lemma 2. (Lemma 2 of [5]) *The learning regret of a policy π with respect to the approximated policy $\tilde{\pi}^*$ is bounded as*

$$\widetilde{Reg}_\pi(T) \leq \sum_{i=1}^K \mathbb{E}[\tau_i] (\Delta_{i,i^*} + \Gamma_{i,i^*}^2) + \sigma_{i^*}^2,$$

where $i^* = \operatorname{argmin}_i \eta_i$, $\Delta_{i,i^*} = \eta_i - \eta_{i^*}$ and $\Gamma_{i,i^*} = \mu_i - \mu_{i^*}$.

Lemma 3. (Theorem 1 of [5]) *With $\sigma_{\max}^2 = \max_{i \in \mathcal{K}} \sigma_i^2$, the learning regret difference can be bounded as*

$$Reg_\pi(T) - \widetilde{Reg}_\pi(T) \leq \sigma_{\max}^2 \left(\sum_{i \neq i^*}^K \Gamma_{i,i^*}^2 / \Delta_{i,i^*} + 1 \right), \quad (16)$$

Lemma 4. *The expected number of plays of any suboptimal arm $i \neq i^*$ in GRA-UCB can be upper bounded by*

$$\mathbb{E}[\tau_i(T)] \leq (4\rho^2 / \Delta_{i,i^*}^2 + C) \log T + 3. \quad (17)$$

Proof: First, we upper bound $\tau_i(T)$ as

$$\tau_i(T) = \sum_{t=1}^T \{I_t = i\} = l_i + \sum_{t=l_i+1}^T \{I_t = i, \tau_i(t) \geq l_i\} \quad (18)$$

where I_t represents the index of the arm played at round t , $\{A\}$ is an indicator function which is equal one if A is true and zero otherwise, and l_i is an arbitrary positive integer. Thus,

$$\tau_i(T) \leq l_i + \sum_{t=l_i+1}^T \{B_i(t) - B_{i^*}(t) \leq 0, \tau_i(t) \geq l_i\}. \quad (19)$$

Based on (9), we can write $B_i(t) - B_{i^*}(t)$ as

$$\begin{aligned} B_i(t) - B_{i^*}(t) &= (\hat{s}_{t,\tau_i(t)}^2 - \sigma_i^2 - \rho(\bar{\mu}_i(t) - \mu_i - c_t)) \\ &\quad - (\hat{s}_{t,\tau_{i^*}(t)}^2 - \rho(\bar{\mu}_{i^*}(t) - \mu_{i^*} + c_t^*)) \\ &\quad + (\Delta_{i,i^*} - 2\rho c_t + \sigma_{i^*}^2), \end{aligned} \quad (20)$$

where $c_t = \sqrt{\log t / \tau_i(t)}$, $c_t^* = \sqrt{\log t / \tau_{i^*}(t)}$, and $\hat{s}_{t,\tau_i(t)}^2 = (\tau_i(t) - 1)s_{\tau_i(t)}^2 / \chi_{1-\alpha,\tau_i(t)-1}^2$.

For $\tau_i(t) \geq l_i = \left\lceil \frac{(4\rho^2 \log T)}{\Delta_{i,i^*}^2} \right\rceil$, the last term of (20) will be always positive since $\Delta_{i,i^*} - 2\rho c_t \geq 0$. Consequently, to ensure $B_i(t) - B_{i^*}(t) \leq 0$, the first term of (20) needs to be negative and the second term of (20) needs to be positive. Therefore, continuing from (19) and (20), we have

$$\begin{aligned} \tau_i(T) &< l_i + \sum_{t=l_i+1}^T \{\hat{s}_{t,\tau_i(t)}^2 \leq \sigma_i^2\} + \\ &\quad \sum_{t=l_i+1}^T \{\bar{\mu}_i(t) \geq \mu_i + c_t\} + \sum_{t=l_i+1}^T \{\bar{\mu}_{i^*}(t) \leq \mu_{i^*} - c_t^*\}. \end{aligned} \quad (21)$$

Applying Facts 1 and 2, $\mathbb{E}[\tau_i(T)]$ can be written as

$$\begin{aligned} \mathbb{E}[\tau_i(T)] &\leq l_i + \sum_{t=l_i+1}^T \alpha + 2 \sum_{t=l_i+1}^T t^{-2} \\ &\leq (4\rho^2 \log T) / \Delta_{i,i^*}^2 + 1 + \int_{l_i}^T \alpha dt + 2 \int_{l_i}^T t^{-2} dt. \end{aligned} \quad (22)$$

By setting $\alpha = C/t$, $C > 0$, $\mathbb{E}[\tau_i(T)]$ is upper bounded as

$$\begin{aligned} \mathbb{E}[\tau_i(T)] &\leq (4\rho^2 \log T) / \Delta_{i,i^*}^2 + 1 + C \log T + 2l_i^{-1} \\ &\leq (4\rho^2 / \Delta_{i,i^*}^2 + C) \log T + 3, \end{aligned} \quad (23)$$

which completes the proof.

Theorem 1. *Combining Lemmas 2, 3, and 4, the learning regret of the GRA-UCB algorithm can be upper bounded as*

$$\begin{aligned} Reg_\pi(T) &\leq \sum_{i=1}^K \left(\left(\frac{4\rho^2}{\Delta_{i,i^*}^2} + C \right) \log T + 3 \right) (\Delta_{i,i^*} + \Gamma_{i,i^*}^2) \\ &\quad + \sigma_{i^*}^2 + \sigma_{\max}^2 \left(\sum_{i \neq i^*}^K \Gamma_{i,i^*}^2 / \Delta_{i,i^*} + 1 \right). \end{aligned} \quad (24)$$

B. Learning regret of ARA-UCB

In ARA-UCB, similar to the analysis of GRA-UCB, we need to bound the number of plays for sub-optimal arms. The following lemma provides an upper bound on $\mathbb{E}[\tau_i(T)]$.

Lemma 5. *The expected number of plays of any suboptimal arm $i \neq i^*$ in ARA-UCB for sub-Gaussian rewards can be upper bounded by*

$$\mathbb{E}[\tau_i(T)] \leq ((4\rho^2 / \Delta_{i,i^*}^2) + C) \log T + \beta \log L_i + 3. \quad (25)$$

Figure 1: The learning regret with Gaussian reward

Proof: According to (15) and (19), we have

$$\begin{aligned} D_i(t) - D_{i^*}(t) &= (v'_{\tau_i(t)} - \sigma_i^2 - \rho(\bar{\mu}_i(t) - \mu_i - c_t)) \\ &\quad - (v'_{\tau_{i^*}(t)} - \rho(\bar{\mu}_{i^*}(t) - \mu_{i^*} + c_t^*)) \\ &\quad + (\Delta_{i,i^*} - 2\rho c_t + \sigma_{i^*}^2). \end{aligned} \quad (26)$$

For $\tau_i(t) \geq l_i = \left\lceil \frac{(4\rho^2 \log T)}{\Delta_{i,i^*}^2} \right\rceil$, $\Delta_{i,i^*} - 2\rho c_t \geq 0$, and hence, the last term of (26) will always be positive. To ensure $D_i(t) - D_{i^*}(t) \leq 0$, we need the first term of (26) to be negative and the second term of (26) to be positive. Consequently, continuing from (26), we have

$$\begin{aligned} \tau_i(T) &\leq l_i + \sum_{t=l_i+1}^T \{v'_{\tau_i(t)} \leq \sigma_i^2\} + \\ &\quad \sum_{t=l_i+1}^T \{\bar{\mu}_i(t) \geq \mu_i + c_t\} + \sum_{t=l_i+1}^T \{\bar{\mu}_{i^*}(t) \leq \mu_{i^*} - c_t^*\}. \end{aligned} \quad (27)$$

According to Lemma 1, we can conclude that

$$\Pr\{v'_{\tau_i(t)} \leq \sigma_i^2\} \leq \beta\alpha, \quad (28)$$

where $\beta > 1$ is a scalar. Given a sufficiently large number of samples (L_i) of arm i , Lemma 1 holds when $\beta = 1$. However, when $\tau_i(t) < L_i$ which means $\tau_i(t)$ is not sufficiently large, the actual confidence bound in (14) is smaller than the expected bound. Since the reward is sub-Gaussian, β cannot go to infinity and will decrease when more samples are collected.

Applying Fact 1 and (28), by setting $\alpha = Ct^{-1}$ with $C > 0$, the expected value of $\tau_i(T)$ in (27) can be written as

$$\begin{aligned} \mathbb{E}[\tau_i(T)] &\leq l_i + \sum_{t=l_i+1}^{L_i} (\beta - 1)\alpha + \sum_{t=l_i+1}^T \alpha + 2 \sum_{t=l_i+1}^T t^{-2}, \\ &\leq \left\lceil \frac{(4\rho^2 \log T)}{\Delta_{i,i^*}^2} \right\rceil + \int_{l_i}^{L_i} \beta \frac{C}{t} dt + \int_{l_i}^T \frac{C}{t} dt + 2 \int_{l_i}^T t^{-2} dt, \\ &\leq (4\rho^2 / \Delta_{i,i^*}^2 + C) \log T + \beta \log L_i + 3, \end{aligned} \quad (29)$$

which completes the proof.

Theorem 2. Combining Lemmas 2, 3, and 5, the learning regret of ARA-UCB algorithm for sub-Gaussian rewards can be upper bounded as

$$\begin{aligned} \text{Reg}_\pi(T) &\leq \sum_{i=1}^K \left(\left(\frac{4\rho^2}{\Delta_{i,i^*}^2} + C \right) \log T + \beta \log L_i + 3 \right) \times \\ &\quad (\Delta_{i,i^*} + \Gamma_{i,i^*}^2) + \sigma_{i^*}^2 + \sigma_{\max}^2 \left(\sum_{i \neq i^*}^K \Gamma_{i,i^*}^2 / \Delta_{i,i^*} + 1 \right). \end{aligned} \quad (30)$$

Figure 2: The learning regret with truncated Gamma reward

V. NUMERICAL RESULTS

In this section, we simulate the proposed algorithms and compare them with several benchmarks. Two reward distributions are considered which are Gaussian distribution and truncated Gamma distribution with the range $[0, 10]$. Since different reward distributions with different parameters are considered, different risk tolerance factors ($\rho = 1$ and $\rho = 2.5$, respectively) are used to ensure the optimal arm is risk sensitive with relatively high mean and low variance. Moreover, the arm number in both experiments is set to be $K = 10$. To evaluate the performance of GRA-UCB and ARA-UCB in comparison with the existing schemes, we simulate three algorithms, namely, MV-LCB [2], MV-UCB [5], and CVaR [6]. Finally, we calculate the learning regret by running each simulation for 100 times and taking the average results.

In Figure 1, we observe that both proposed algorithms reach low cumulative learning regret, because GRA-UCB is designed for the Gaussian reward and ARA-UCB is model agnostic. In addition, although MV-UCB performs relatively well, the performance is sensitive to its parameter b which needs to be adjusted for better performance and no clear hint is discussed in the literature. Moreover, our proposed schemes outperform the MV-LCB because the confidence bound used in MV-LCB is not tight enough.

In Figure 2, benefiting from the asymptotic upper confidence bound, ARA-UCB reaches the lowest learning regret. Additionally, we should also note that, because truncated Gamma distribution is used to generate reward while MV-UCB is based on an assumption of symmetric reward distribution, the derived confidence bound of MV-UCB is inaccurate.

VI. CONCLUSION

In this paper, we studied a different type of MAB problem which is referred to as the risk-aware MAB. Instead of purely considering the expected reward when making decision, the variance of each arm is also taken into account to reduce the risk of receiving low rewards. The key novelty of this paper is building the confidence bound of the MV with the help of identifying the sampled reward variance distribution and designing a risk-aware MAB algorithm when the reward distribution knowledge is unknown. The proposed algorithms achieve logarithmic learning regret and performs numerically well as compared to several benchmarks.

REFERENCES

- [1] H. Markowitz, "Portfolio selection," *The journal of finance*, vol. 7, no. 1, pp. 77–91, 1952.
- [2] A. Sani, A. Lazaric, and R. Munos, "Risk-aversion in multi-armed bandits," in *Advances in Neural Information Processing Systems*, 2012, pp. 3275–3283.
- [3] S. Vakili and Q. Zhao, "Mean-variance and value at risk in multi-armed bandit problems," in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sep. 2015, pp. 1330–1335.
- [4] —, "Risk-averse online learning under mean-variance measures," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 1911–1915.
- [5] —, "Risk-averse multi-armed bandit problems under mean-variance measure," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 6, pp. 1093–1111, Sep. 2016.
- [6] N. Galichet, M. Sebag, and O. Teytaud, "Exploration vs exploitation vs safety: Risk-aware multi-armed bandits," in *Asian Conference on Machine Learning*, 2013, pp. 245–260.
- [7] Y. David, B. Szörényi, M. Ghavamzadeh, S. Mannor, and N. Shimkin, "Pac bandits with risk constraints," in *ISAIM*, 2018.
- [8] A. Prashanth L, K. Jagannathan, and R. K. Kolla, "Concentration bounds for CVaR estimation: The cases of light-tailed and heavy-tailed distributions," *arXiv preprint arXiv:1901.00997*, 2019.
- [9] A. Kagrecha, J. Nair, and K. Jagannathan, "Distribution oblivious, risk-aware algorithms for multi-armed bandits with unbounded rewards," in *Advances in Neural Information Processing Systems 32*, 2019, pp. 11 272–11 281.
- [10] D. C. Montgomery and G. C. Runger, *Applied statistics and probability for engineers*. John Wiley and Sons, 2014.
- [11] A. DasGupta, *Asymptotic theory of statistics and probability*. Springer Science & Business Media, 2008.