CVaR-Regret Bounds for Multi-Armed Bandits

Chenmien Tan CHENMIENTAN@OUTLOOK.COM
University of Nottingham Ningbo China

Paul Weng PAUL.WENG@SJTU.EDU.CN
Shanghai Jiao Tong University

Editors: Emtiyaz Khan and Mehmet Gönen

Abstract

In contrast to risk-averse multi-armed bandit (MAB), where one aims for a best risk-sensitive arm while having a risk-neutral attitude when running the risk-averse MAB algorithm, in this paper, we aim for a best arm with respect to the mean like in the standard MAB, but we adopt a risk-averse attitude when running a standard MAB algorithm. Conditional value-at-risk (CVaR) of the regret is adopted as the metric to evaluate the performance of algorithms, which is an extension of the traditional expected regret minimization framework. For this new problem, we revisit several classic algorithms for stochastic and non-stochastic bandits, UCB, MOSS, and Exp3-IX with its variants and propose parameters with good theoretically guaranteed CVaR-regret, which match the results of the expected regret and achieve (nearly-)optimality up to constant. In the non-stochastic setting, we show that implicit exploration achieves a trade-off between the variability of the regret and the regret in expectation. Numerical experiments are conducted to validate our results.

Keywords: multi-armed bandit; conditional value-at-risk

1. Introduction

Multi-armed bandit (MAB) is a fundamental model for studying optimal learning in sequential decision-making. It has a wide variety of applications, e.g., portfolio optimization (Shen et al., 2015) or web optimization (White, 2012). In MAB, an agent (or player) faces a repeated game (i.e., decision-making problem) where it has to choose an action called arm, which would yield a reward. To maximize the total gain, the agent needs to balance the exploration of all possible arms and the exploitation of the seemingly most profitable one.

Running a MAB algorithm to solve this trade-off results in observing a stochastic regret, i.e., difference between reward of a posteriori known best arm and reward of chosen arm. Traditionally, the quality of MAB algorithms is measured in terms of their expected total regrets (Auer et al., 2002; Audibert and Bubeck, 2009). In practice, a MAB algorithm is generally run only a few times (even possibly once), whereas the empirical average of the observed regret is only guaranteed to concentrate around its mean after a large number of runs. Thus, the regret in expectation fails to satisfactorily account for the risk sensitivity of the player. Indeed in many scenarios such as medical treatment or financial investment, we may be interested in the regret in extreme circumstances rather than on average. In this regard, a regret bound in probability may be a more desirable choice (Audibert et al., 2009; Auer et al., 1995), however such bound still fails to describe the tail risk beyond a confidence level.

As an illustration, consider the following situation where the best arm is underestimated while some suboptimal arms are over-estimated, which would lead the player to keep drawing the latter arms until the over-estimation and under-estimation are corrected, which may take some time and lead to a high total regret. Although this situation may happen rarely, a risk-sensitive player would prefer to avoid it by finding a better trade-off between exploration and exploitation. Intuitively, more exploration than usual is required to control the probability of such adverse events, even if a potential price on the expected regret has to be paid.

Risk sensitivity has been investigated in MABs. In these risk-sensitive MAB studies, the goal is to find a risk-sensitive best arm, but algorithms are still analyzed in terms of regret bounds in expectation or probability. Various risk measures have been considered, such as mean-variance (Sani et al., 2012; Vakili and Zhao, 2016; Zhu and Tan, 2020) and conditional value-at-risk (Maillard, 2013; Galichet et al., 2013; L.A. et al., 2020; Baudry et al., 2021). We refer readers to Tan et al. (2022) for a comprehensive survey. In contrast to risk-sensitive MABs, in this paper, we study the conditional value-at-risk (CVaR) of the regret (CVaR-regret) while arms are still evaluated in terms of their means. To the best of our knowledge, no previous work has analyzed bounds on CVaR-regret.

The remaining of the paper is organized as follows. In Section 2, after recalling the necessary background, we formally state the problem studied in this paper. In Section 3, we then state our main results regarding CVaR-regret in various MAB settings. In Section 4, we conduct some experiments to verify our analyses. We conclude the paper in Section 5.

2. Background and Problem Formulation

Let $K \geq 2$ be the number of arms and $T \geq K$ be the time horizon. In each round $t \in [T] = \{1, \dots, T\}$ of the iterated game, the agent chooses a distribution $\boldsymbol{p}_t$ over $[K]$ to sample an action a_t and the environment reveals the gain g_{t,a_t} (or equivalently, the loss ℓ_{t,a_t}) associated with the arm. Depending on how the reward signals are generated, MAB can be divided into stochastic and non-stochastic ones. The objective of the agent is to maximize the total gain $\sum_{t=1}^{T} g_{t,a_t}$, or equivalently, minimize the regret in the corresponding setting. To achieve the goal, the agent needs to make decisions based on rational policy $\boldsymbol{\pi} = (\pi_1, \dots, \pi_T)$, where $\pi_t : (a_1, g_1, \dots, a_{t-1}, g_{t-1}) \mapsto \boldsymbol{p}_t$ maps the interaction history to a distribution over $[K]$. It is worth noting that the agent may have no prior knowledge of the horizon. Any algorithm that does not rely on the knowledge of the horizon is called anytime.

2.1. Stochastic MAB

A stochastic bandit is defined by a collection of distributions $\boldsymbol{\nu} = (\nu_1, \dots, \nu_K)$ from which the reward signals are sampled from. With given bandit instance $\boldsymbol{\nu}$ and policy $\boldsymbol{\pi}$, the regret is a well-defined random variable, which can be used to evaluate the quality of the algorithm. However, for technical reasons (Bubeck and Cesa-Bianchi, 2012), the regret of stochastic bandit is too difficult to bound. A reasonable alternative is the pseudo-regret, which is a random variable with respect to the stochastic choices of arms:

$$\overline{R}_T = \sum_{k=1}^{K} \Delta_k N_k(T)$$

where $\mu_k = \mathbb{E}[\nu_k]$ is the mean return of the k-th arm, $\Delta_k = \max_{s\in[K]} \mu_s - \mu_k$ is its sub-optimality gap, $\mathbb{I}\{\cdot\}$ is the indicator function, and $N_k(t) = \sum_{s=1}^{t} \mathbb{I}\{a_t = k\}$ is the pull time of the k-th arm in the first t rounds.

For the sake of control the tail probability, a common assumption is that the gain of each arm obeys σ-subgaussian distribution, i.e., $\mathbb{E}[\exp(\lambda(\nu_k - \mu_k))] \leq \exp(\sigma^2\lambda^2/2)$ for any $\lambda \in \mathbb{R}$. The assumption covers a large collection of distributions such as the ones with bounded support and provides the Chernoff-Hoeffding bound, i.e., $\mathbb{P}\{\hat{\mu}_{k,s} - \mu_k \geq \varepsilon\} \leq \exp(-\frac{s\varepsilon^2}{2\sigma^2})$, where $\hat{\mu}_{k,s}$ is the empirical mean of the k-th arm in its s trials.

To minimize the regret, one may consider how the regret increases with respect to the horizon with fixed bandit instance, where Lai and Robbins (1985) show that every consistent policy suffers $\Omega(\log T)$ expected regret. Another approach is to consider the worst instance the agent possibly confront for fixed horizon, where Auer et al. (1995) prove that every algorithm has $\Omega(\sqrt{KT})$ expected regret. Both of above two scenarios are extensively investigated, where UCB (Auer et al., 2002) and MOSS (Audibert and Bubeck, 2009) are probably the most fundamental algorithms that match the order of respective lower bounds.

2.2. Non-stochastic MAB

Stochastic bandit assumes that rewards are sampled from a stationary distribution, which limits its practical usage. In contrast, an adversarial bandit is defined by losses $\boldsymbol{\ell} \in [0,1]^{T\times K}$ secretly selected by an adversary, where the (total) regret is defined as: (Auer et al., 1995):

$$R_T = \sum_{t=1}^{T} \ell_{t,a_t} - \min_{k\in[K]} \sum_{t=1}^{T} \ell_{t,k}$$

Note this total regret is again a random variable with the respect to the stochastic choices of the arms. As adversarial bandit contains no statistical assumption, we only concern the worst instance the agent may confront for fixed horizon. The most basic approach is probably Exp3 (Auer et al., 1995), which adopts exponentially weighted forecaster $p_{t,k} \propto \exp(-\eta_t \sum_{s=1}^{t-1} \widehat{\ell}_{s,k})$ with the importance-weighted estimator:

$$\widehat{\ell}_{t,k} = \frac{\ell_{t,a_t}\mathbb{I}\{a_t = k\}}{p_{t,k}}$$

where $p_{t,k} = \mathbb{P}\{a_t = k|\mathcal{F}_{t-1}\}$ is the probability of drawing the k-th arm conditioned on the interaction history up to the end of round $t-1$. Despite that Exp3 enjoys a good expected regret (Bubeck and Cesa-Bianchi, 2012), the fluctuation of the estimator brings difficulty to risk analysis. Therefore, we consider instead Exp3-IX (Kocák et al., 2014) with the following estimator:

$$\widetilde{\ell}_{t,k} = \frac{\ell_{t,a_t}\mathbb{I}\{a_t = k\}}{p_{t,k} + \gamma_t}$$

In this paper, we also consider two variants of adversarial bandits. The first one is bandit with expert advice, where the player has access to N "experts" to help choose the best arm, which are represented by N distributions $\boldsymbol{\xi}_t(1), \ldots, \boldsymbol{\xi}_t(N)$ over $[K]$ in each round

$t \in [T]$ (Auer et al., 1995). The objective of the learner is to minimize the regret against the best expert:

$$R_T = \sum_{t=1}^{T} \ell_{t,a_t} - \min_{n\in[N]} \sum_{t=1}^{T} \boldsymbol{\xi}_t(n)^T \boldsymbol{\ell}_t$$

Regret R_T is again a random variable since the a_t's are random.

The other variant is non-stationary bandits (Herbster and Warmuth, 1998). Beyond chasing the best single arm, a class of sequences that switch at most S time between arms is considered. Denote $C(S) \subseteq [K]^T$ as the set of sequences with no more than S switches. The objective of the learner is to minimize the (total) regret against the best sequence:

$$R_T = \sum_{t=1}^{T} \ell_{t,a_t} - \min_{(J_t)_{t=1}^T \in C(S)} \sum_{t=1}^{T} \ell_{t,J_t}$$

2.3. Risk Measures

Let X be a real random variable in terms of loss, for any $\alpha \in (0, 1]$, the conditional value-at-risk (CVaR) of X at level $1 - \alpha$ is defined as (Artzner et al., 1999)

$$\mathrm{CVaR}_{1-\alpha}(X) = \inf_{\lambda\in\mathbb{R}} \left\{ \lambda + \frac{1}{\alpha}\mathbb{E}[(X-\lambda)^+] \right\}$$

where $(X)^+ = \max\{0, X\}$. Let $F(x) = \mathbb{P}\{X \le x\}$ and $F^{-1}(x) = \inf\{y \in \mathbb{R} : F(y) \ge x\}$, CVaR can be equivalently defined as $\mathrm{CVaR}_{1-\alpha}(X) = \frac{1}{\alpha}\int_0^\alpha F^{-1}(1-x)\mathrm{d}x$. From this definition, one can easily show that $\mathbb{P}\{X > f(\gamma)\} \le \gamma$ for any $\gamma \in (0, 1)$ implies $\mathrm{CVaR}_{1-\alpha}(X) \le \frac{1}{\alpha}\int_0^\alpha f(\gamma)\mathrm{d}\gamma$ for any $\alpha \in (0, 1]$. Informally speaking, CVaR represents the conditional expectation of X over the worst α-fraction. As a coherent risk measure (Artzner et al., 1999), CVaR enjoys several desirable properties including translation invariance, positive homogeneity, and sub-additivity.

Since it is used in our proofs, we also mention another coherent risk measure called entropic value-at-risk (EVaR), which is defined as:

$$\mathrm{EVaR}_{1-\alpha}(X) = \inf_{\lambda>0} \left\{ \frac{1}{\lambda} \log \frac{\mathbb{E}[\exp(\lambda X)]}{\alpha} \right\}$$

and serves as an upper bound of CVaR (Ahmadi-Javid, 2012).

2.4. Problem Statement

In this paper, we aim at bounding the *CVaR-regret*, which is the metric we use to evaluate a bandit algorithm. It is defined as the CVaR of the pseudo-regret $\mathrm{CVaR}_{1-\alpha}\left(\overline{R}_T\right)$ for stochastic MAB and the CVaR of the regret $\mathrm{CVaR}_{1-\alpha}(R_T)$ for non-stochastic MAB. Since $\mathrm{CVaR}_0(X) = \mathbb{E}[X]$, the traditional expected regret is a special case of the CVaR-regret.

3. Main Results

We now present our theoretical analyses where we revisit the following existing algorithms: UCB, MOSS, and Exp3-IX with some variants, Exp4-IX and Exp3-SIX.

3.1. Stochastic Bandits

We firstly revisit UCB (Auer et al., 2002), as shown in Algorithm 1. Our result extends the traditional result for the expected regret and shows that UCB with sufficient exploration rate enjoys logarithmic CVaR-regret under the assumption of subgaussian reward distributions, where the leading factor is independent of α. Remarkably, additional exploration beyond the threshold is poisonous, regardless of the value of α.

Algorithm 1: UCB

Input: $\rho \geq 0$

Pull each arm once

for $t = K, \ldots, T-1$ **do**

  $a_{t+1} \leftarrow \arg\max_{k \in [K]} \left(\hat{\mu}_{k, N_k(t)} + \sqrt{\frac{\rho \log T}{N_k(t)}} \right)$

  Sample a_{t+1}

end

Theorem 1 *Suppose that $\nu_k - \mu_k$ is a 1-subgaussian distribution for any $k \in [K]$. Then, for Algorithm 1 with $\rho \geq 4$, for any $\alpha \in (0, 1]$*

$$\mathrm{CVaR}_{1-\alpha}(\overline{R}_T) \leq 4 \left(\rho \log T + \sqrt{2\pi\rho \left(2 \log^2 \frac{1}{\alpha} + 2 \log \frac{1}{\alpha} + 1 \right) \log T} + 2 \log \frac{e}{\alpha} \right) \sum_{k=1}^{K} \frac{1}{\Delta_k} + \sqrt{\frac{32}{\alpha}} + \sum_{k=1}^{K} \Delta_k$$

Proof Denote $*$ as the optimal arm. A sub-optimal arm k will be pulled in round $t > K$ only if $\hat{\mu}_{*, N_*(t)} \leq \mu_* - \frac{\Delta_k}{2}$ or $\hat{\mu}_{k, N_k(t)} \geq \mu_k + \frac{\Delta_k}{2}$. Thus, the pseudo-regret can be decomposed as the corresponding two terms, which are caused by underestimation and overestimation, respectively.

Step I: bounding the regret caused by underestimation

Define $\Delta = \max \left\{ 0, \mu_* - \min_{t \in [T]} \left(\hat{\mu}_{*, N_*(t)} + \sqrt{\frac{\rho \log T}{N_*(t)}} \right) \right\}$. For any $\varepsilon > 0$, there is

$$\begin{aligned} \mathbb{P}\{\Delta > \varepsilon\} &\leq \mathbb{P} \left\{ \exists s \in \mathbb{N}_+ : \hat{\mu}_{*,s} + \sqrt{\frac{\rho \log T}{s}} < \mu_* - \varepsilon \right\} \\ &\leq \sum_{s=1}^{\infty} \exp\left(-\frac{s(\sqrt{\frac{\rho \log T}{s}} + \varepsilon)^2}{2} \right) \leq T^{-\rho/2} \sum_{s=1}^{\infty} \exp\left(-\frac{\varepsilon^2 s}{2} \right) = T^{-\rho/2} \frac{\exp(-\frac{\varepsilon^2}{2})}{1 - \exp(-\frac{\varepsilon^2}{2})} \end{aligned}$$

By letting $\gamma = T^{-\rho/2} \frac{\exp(-\frac{\varepsilon^2}{2})}{1 - \exp(-\frac{\varepsilon^2}{2})}$ we obtain $\varepsilon = \sqrt{2 \log(T^{-\rho/2} \gamma^{-1} + 1)}$. Since $x \geq \log(x + 1), \forall x \in (-1, \infty)$, for any $\gamma \in (0, 1)$

$$\mathbb{P} \left\{ \Delta > \sqrt{2 T^{-\rho/2} \gamma^{-1}} \right\} \leq \mathbb{P} \left\{ \Delta > \sqrt{2 \log \left(T^{-\rho/2} \gamma^{-1} + 1 \right)} \right\} \leq \gamma$$

Hence we yield $\mathrm{CVaR}_{1-\alpha}(\Delta) \leq \sqrt{\frac{8}{\alpha}} T^{-\rho/4}$.

Step II: bounding the regret caused by overestimation

Now suppose k satisfies $\Delta_k > 2\Delta$. For such an arm k and any $s \in \mathbb{N}_+$

$$\mathbb{P}\{N_k(T) > s\} \leq \mathbb{P}\left\{\hat{\mu}_{k,s} + \sqrt{\frac{\rho \log T}{s}} \geq \mu_k + \frac{\Delta_k}{2}\right\} \leq \exp\left(-\frac{sc^2\Delta_k^2}{8}\right) \leq T^{-\frac{\rho c^2}{2(1-c)^2}}$$

The second inequality sign holds by assuming u is large sufficiently to satisfy $\frac{\Delta_k}{2} - \sqrt{\frac{\rho \log T}{s}} \geq \frac{c\Delta_k}{2}$ for some $c \in (0,1)$, which leads to $\frac{s\Delta_k^2}{8} \geq \frac{\rho \log T}{2(1-c)^2}$ and the last inequality sign. Recall that the inequality holds for any integer such that $s \geq \frac{4\rho \log T}{(1-c)^2\Delta_k^2}$. By solving $\gamma = T^{-\frac{\rho c^2}{2(1-c)^2}}$ we obtain

$$\mathbb{P}\left\{N_k(T) > \frac{4}{\Delta_k^2}\left(\sqrt{2\log\frac{1}{\gamma}} + \sqrt{\rho \log T}\right)^2 + 1\right\} \leq \gamma$$

Hence, we yield

$$\mathrm{CVaR}_{1-\alpha}(N_k(T)) \leq \frac{4}{\Delta_k^2}\left(\rho \log T + \sqrt{2\pi\rho\left(2\log^2\frac{1}{\alpha} + 2\log\frac{1}{\alpha} + 1\right)\log T} + 2\log\frac{e}{\alpha}\right) + 1$$

Little calculus needs to be done to see the inequality above. Let $I = \int_0^\alpha \sqrt{\log\frac{1}{\gamma}}\mathrm{d}\gamma = 2\int_{\sqrt{\log\frac{1}{\alpha}}}^\infty \gamma^2 \exp(-\gamma^2)\mathrm{d}\gamma$. Then,

$$I^2 \leq 4\int_0^{\frac{\pi}{2}} \cos^2\theta \sin^2\theta \mathrm{d}\theta \int_{\sqrt{2\log\frac{1}{\alpha}}}^\infty r^5 \exp(-r^2)\mathrm{d}r = \frac{\alpha^2\pi}{4}\left(2\log^2\frac{1}{\alpha} + 2\log\frac{1}{\alpha} + 1\right)$$

Step III: summing

The pseudo-regret can be decomposed as follows

$$\overline{R}_T = \sum_{k=1}^K \Delta_k N_k(T)\mathbb{I}\{\Delta_k \leq 2\Delta\} + \sum_{k=1}^K \Delta_k N_k(T)\mathbb{I}\{\Delta_k > 2\Delta\}$$

Then, by the sub-additivity of CVaR, we have

$$\mathrm{CVaR}_{1-\alpha}\left(\overline{R}_T\right) \leq 2T\mathrm{CVaR}_{1-\alpha}(\Delta) + \sum_{k=1}^K \Delta_k \mathrm{CVaR}_{1-\alpha}(N_k(T)\mathbb{I}\{\Delta_k > 2\Delta\})$$

By substituting the values we complete the proof. ■

Now we turn to MOSS (Audibert and Bubeck, 2009), given in Algorithm 2. Just as before, sufficient exploration guarantees that the instance-independent CVaR-regret grows at the square root rate. However, additional exploration may be applied to deal with the worst bandit instance, which makes MOSS more difficult to tune. For simplicity, in the following content, we denote $\log^+ x = \max\{0, \log x\}$.

Algorithm 2: MOSS

Input: $\rho \geq 0$

Pull each arm once

for $t = K, \ldots, T-1$ **do**

  $a_{t+1} \leftarrow$

  $\arg\max_{k\in[K]}\left(\hat{\mu}_{k,N_k(t)} + \sqrt{\frac{\rho}{N_k(t)}\log^+\frac{T}{KN_k(t)}}\right)$

  Sample a_{t+1}

end

Theorem 2 *Suppose that $\nu_k - \mu_k$ is a 1-subgaussian distribution for any $k \in [K]$. Then, for Algorithm 2 with $\rho = 3$, for any $\alpha \in (0, 1]$,*

$$\mathrm{CVaR}_{1-\alpha}\left(\overline{R}_T\right) \leq \left(\frac{15}{\sqrt{\alpha}} + \sqrt{10\left(2\log^2\frac{1}{\alpha} + 2\log\frac{1}{\alpha} + 1\right)} + 2\log\frac{1}{\alpha} + 10\right)\sqrt{KT} + \sum_{k=1}^{K}\Delta_k$$

Proof Again, we denote $*$ as the optimal arm and decompose the pseudo-regret as underestimation and overestimation terms.

Step I: bounding the regret caused by underestimation

Define $\Delta = \max\{0, \mu_* - \min_{t\in[T]}\left(\hat{\mu}_{*,N_*(t)} + \sqrt{\frac{\rho}{N_*(t)}\log^+\frac{T}{KN_*(t)}}\right)\}$. By denoting $S_s = s(\hat{\mu}_{*,s} - \mu_*)$, for any $\varepsilon > 0$ and $\rho > 2$, there is

$$\begin{aligned}
\mathbb{P}\{\Delta > \varepsilon\} &\leq \mathbb{P}\left\{\exists s \in \mathbb{N}_+ : \hat{\mu}_{*,s} + \sqrt{\frac{\rho}{s}\log^+\frac{T}{Ks}} < \mu_* - \varepsilon\right\} \\
&= \mathbb{P}\left\{\exists s \in \mathbb{N}_+ : S_s + \sqrt{\rho s\log^+\frac{T}{Ks}} + s\varepsilon < 0\right\} \\
&\leq \sum_{n=0}^{\infty}\mathbb{P}\left\{\exists\left(\frac{\rho}{2}\right)^n \leq s \leq \left(\frac{\rho}{2}\right)^{n+1} : S_s + \sqrt{\rho s\log^+\frac{T}{Ks}} + s\varepsilon < 0\right\} \\
&\leq \sum_{n=0}^{\infty}\mathbb{P}\left\{\exists s \leq \left(\frac{\rho}{2}\right)^{n+1} : S_s + \sqrt{\frac{\rho^{n+1}}{2^n}\log^+\frac{2^{n+1}T}{\rho^{n+1}K}} + \frac{\rho^n}{2^n}\varepsilon < 0\right\} \\
&\leq \sum_{n=0}^{\infty}\exp(-\frac{2^n(\sqrt{\frac{\rho^{n+1}}{2^n}\log^+\frac{2^{n+1}T}{\rho^{n+1}K}} + \frac{\rho^n}{2^n}\varepsilon)^2}{\rho^{n+1}}) \leq \frac{K}{T}\sum_{n=0}^{\infty}\left(\frac{\rho}{2}\right)^{n+1}\exp(-2^{-n}\rho^{n-1}\varepsilon^2) \\
&\leq \frac{K}{T}\left(\frac{\rho^2}{2e\varepsilon^2} + \int_0^{\infty}\left(\frac{\rho}{2}\right)^{x+1}\exp(-2^{-x}\rho^{x-1}\varepsilon^2)\mathrm{d}x\right) \leq \left(\frac{1}{e} + \frac{1}{\log\frac{\rho}{2}}\right)\frac{\rho^2 K}{2\varepsilon^2 T}
\end{aligned}$$

The fourth and fifth inequality signs follow inequality 2.17 of Hoeffding (1963) and $(x+y)^2 \geq x^2 + y^2, \forall x, y > 0$. The sixth one holds by observing that $f(x) = \left(\frac{\rho}{2}\right)^{x+1}\exp(-\varepsilon^2 2^{-x}\rho^{x-1})$ is at most unimodal. For such a function there is $\sum_{x=a}^{b} f(x) \leq \max_{x\in[a,b]} f(x) + \int_a^b f(x)\mathrm{d}x$. By a similar argument as in Theorem 1 we bound $\mathrm{CVaR}_{1-\alpha}(\Delta)$.

Step II: bounding the regret caused by overestimation

Now we suppose k satisfies $\Delta_k > \max\{2\Delta, \sqrt{4e\rho K/T}\}$. For such an arm k and any $s \in \mathbb{N}_+$

$$\begin{aligned}
\mathbb{P}\{N_k(T) > s\} &\leq \mathbb{P}\left\{\hat{\mu}_{k,s} + \sqrt{\frac{\rho}{s}\log^+\frac{T}{Ks}} \geq \mu_k + \frac{\Delta_k}{2}\right\} \\
&\leq \mathbb{P}\left\{\hat{\mu}_{k,s} + \sqrt{\frac{\rho}{s}\log\frac{\Delta_k^2 T}{4\rho K}} \geq \mu_k + \frac{\Delta_k}{2}\right\} \leq \exp\left(-\frac{sc^2\Delta_k^2}{8}\right) \leq \left(\frac{\Delta_k^2 T}{4\rho K}\right)^{-\frac{\rho c^2}{2(1-c)^2}}
\end{aligned}$$

The third inequality sign holds by assuming s is large sufficiently to satisfy $\frac{\Delta_k}{2} - \sqrt{\frac{\rho}{s}\log\frac{\Delta_k^2 T}{4\rho K}} \geq \frac{c\Delta_k}{2}$ for some $c \in (0,1)$, which in turn guarantees the second inequality sign as $s \geq \frac{4\rho}{(1-c)^2\Delta_k^2}\log\frac{\Delta_k^2 T}{4\rho K} > \frac{4\rho}{\Delta_k^2}$. The last inequality follows by recalling that $\frac{s\Delta_k^2}{8} \geq \frac{\rho}{2(1-c)^2}\log\frac{\Delta_k^2 T}{4\rho K}$. By a similar argument as in Theorem 1, we obtain

$$\begin{aligned}\mathrm{CVaR}_{1-\alpha}(N_k(T)) &\leq \frac{4}{\Delta_k^2}\left(\rho\log\frac{\Delta_k^2 T}{4\rho K} + \sqrt{2\pi\rho\left(2\log^2\frac{1}{\alpha} + 2\log\frac{1}{\alpha} + 1\right)\log\frac{\Delta_k^2 T}{4\rho K}} + 2\log\frac{e}{\alpha}\right) + 1 \\ &\leq \frac{4}{\Delta_k}\sqrt{\frac{T}{K}}\left(\frac{\log\frac{e}{\alpha}}{\sqrt{e\rho}} + \sqrt{\frac{\pi}{2e}\left(2\log^2\frac{1}{\alpha} + 2\log\frac{1}{\alpha} + 1\right)} + \frac{\sqrt{\rho}}{e}\right) + 1\end{aligned}$$

The second inequality sign follows by $\log x \leq x/e, \forall x > 0$.

Step III: summing

The pseudo-regret can be decomposed as follows

$$\begin{aligned}\overline{R}_T \leq \sum_{k=1}^{K}\Delta_k N_k(T)\mathbb{I}\{\Delta_k \leq 2\Delta\} + \sum_{k=1}^{K}\Delta_k N_k(T)\mathbb{I}\left\{\Delta_k \leq \sqrt{4e\rho K/T}\right\} \\ + \sum_{k=1}^{K}\Delta_k N_k(T)\mathbb{I}\left\{\Delta_k > \max\{2\Delta, \sqrt{4e\rho K/T}\}\right\}\end{aligned}$$

Then, by the sub-additivity of CVaR, we have

$$\begin{aligned}\mathrm{CVaR}_{1-\alpha}(\overline{R}_T) \leq 2T\mathrm{CVaR}_{1-\alpha}(\Delta) + \sqrt{4e\rho KT} \\ + \sum_{k=1}^{K}\Delta_k \mathrm{CVaR}_{1-\alpha}\left(N_k(T)\mathbb{I}\left\{\Delta_k > \max\{2\Delta, \sqrt{4e\rho K/T}\}\right\}\right)\end{aligned}$$

By substituting the values we complete the proof. ■

3.2. Non-stochastic Bandits

In the adversarial setting, we consider the Exp3-IX, as illustrated in Algorithm 3. Regarding of the algorithm, Neu (2015) recommends $\gamma_t = \eta_t/2$ for achieving a good high probability bound. With some further investigation, we consider that the CVaR-regret is a more natural metric for revealing the nature of implicit exploration, where the rate γ_t is associated with α. By letting $\alpha = 1$ we yield $\gamma_t = 0$ and $\mathrm{CVaR}_0(R_T) = \mathbb{E}[R_T] \leq \sqrt{2KT\log K}$ and $2\sqrt{KT\log K}$ in the horizon-known and anytime cases, where Exp3-IX boils down to Exp3 and our bounds match the result of Bubeck and Cesa-Bianchi (2012). In this regard, implicit exploration is not beneficial for minimizing the expected regret but reducing the variability of the regret.

Algorithm 3: Exp3-IX

Input: $\{\eta_t\}_{t=1}^T, \{\gamma_t\}_{t=1}^T$

Initialization: $\widetilde{L}_{0,k} = 0, \forall k \in [K]$

for $t = 1, \ldots, T$ **do**

  $p_{t,k} \leftarrow \frac{\exp(-\eta_t \widetilde{L}_{t-1,k})}{\sum_{k=1}^{K}\exp(-\eta_t \widetilde{L}_{t-1,k})}, \forall k \in [K]$

  Sample $a_t \sim \boldsymbol{p}_t = (p_{t,1}, \ldots, p_{t,K})$

  $\widetilde{L}_{t,k} \leftarrow \widetilde{L}_{t-1,k} + \widetilde{\ell}_{t,k}, \forall k \in [K]$

end

Theorem 3 *For Algorithm 3 with* $\eta_t = \sqrt{\frac{2\log K}{KT}}$ *and* $\gamma_t = \sqrt{\frac{\log\frac{1}{\alpha}}{2KT}}$, *we have:*

$$\mathrm{CVaR}_{1-\alpha}(R_T) \le \sqrt{2KT\log K} + \sqrt{2KT\log\frac{1}{\alpha}} + \frac{1}{2}\sqrt{\log K\log\frac{1}{\alpha}} + \frac{1}{2}\log\frac{1}{\alpha}$$

For Algorithm 3 with $\eta_t = \sqrt{\frac{\log K}{Kt}}$ *and* $\gamma_t = \frac{1}{2}\sqrt{\frac{\log\frac{1}{\alpha}}{Kt}}$, *we have:*

$$\mathrm{CVaR}_{1-\alpha}(R_T) \le 2\sqrt{KT\log K} + 2\sqrt{KT\log\frac{1}{\alpha}} + \frac{1}{2}\sqrt{\log K\log\frac{1}{\alpha}} + \frac{1}{2}\log\frac{1}{\alpha}$$

Proof Denote $*$ as the optimal arm.

Step I: bounding the random regret stochastically

By the standard analysis for Exp3 of Bubeck and Cesa-Bianchi (2012), for any non-increasing sequence $\{\eta_t\}_{t=1}^T$ we have

$$\begin{aligned}\sum_{t=1}^T\sum_{k=1}^K p_{t,k}\widetilde{\ell}_{t,k} &\le \sum_{t=1}^T\frac{\eta_t}{2}\sum_{k=1}^K p_{t,k}\widetilde{\ell}_{t,k}^2 + \sum_{t=1}^T\widetilde{\ell}_{t,*} + \frac{\log K}{\eta_T}\\ &\le \sum_{t=1}^T\frac{\eta_t}{2}\sum_{k=1}^K\widetilde{\ell}_{t,k} + \sum_{t=1}^T\widetilde{\ell}_{t,*} + \frac{\log K}{\eta_T}\end{aligned}$$

Recall that $\ell_{t,a_t} = \sum_{k=1}^K(p_{t,k}+\gamma_t)\widetilde{\ell}_{t,k}$. Then,

$$\begin{aligned}R_T = \sum_{t=1}^T\ell_{t,a_t} - \sum_{t=1}^T\ell_{t,*} &= \sum_{t=1}^T\sum_{k=1}^K(p_{t,k}+\gamma_t)\widetilde{\ell}_{t,k} - \sum_{t=1}^T\ell_{t,*}\\ &\le \sum_{t=1}^T\left(\frac{\eta_t}{2}+\gamma_t\right)\sum_{k=1}^K\widetilde{\ell}_{t,k} + \sum_{t=1}^T\left(\widetilde{\ell}_{t,*}-\ell_{t,*}\right) + \frac{\log K}{\eta_T}\end{aligned}$$

Hence, to bound R_T, it suffices to bound $\sum_{t=1}^T(\frac{\eta_t}{2}+\gamma_t)\sum_{k=1}^K\widetilde{\ell}_{t,k}$ and $\sum_{t=1}^T(\widetilde{\ell}_{t,*}-\ell_{t,*})$.

Step II: bounding the CVaR of the two terms respectively

Suppose $0\le\alpha_{t,k}\le 1$ for any $t\in[T], k\in[K]$. Since $\exp(x/(1+\lambda))\le 1+x, \forall 0\le x\le 2\lambda$ and $0\le 2\gamma_t\alpha_{t,k}\ell_{t,a_t}\mathbb{I}\{a_t=k\}/p_{t,k}\le 2\gamma_t/p_{t,k}$, for any $t\in[T], k\in[K]$

$$\begin{aligned}\exp\left(2\gamma_t\alpha_{t,k}\widetilde{\ell}_{t,k}\right) &= \exp\left(2\gamma_t\alpha_{t,k}\frac{\ell_{t,a_t}\mathbb{I}\{a_t=k\}}{p_{t,k}(1+\gamma_t/p_{t,k})}\right)\\ &\le 1 + 2\gamma_t\alpha_{t,k}\frac{\ell_{t,a_t}\mathbb{I}\{a_t=k\}}{p_{t,k}} = 1+2\gamma_t\alpha_{t,k}\widehat{\ell}_{t,k}\end{aligned}$$

Using the inequality, one can further derive

$$\begin{aligned}\exp\left(2\gamma_T\alpha_{t,k}\widetilde{\ell}_{t,k}\right) &= \exp\left[\frac{\gamma_T}{\gamma_t}\log\left(\exp(2\gamma_t\alpha_{t,k}\widetilde{\ell}_{t,k})\right)\right]\\ &\le \exp\left(\frac{\gamma_T}{\gamma_t}\log(1+2\gamma_t\alpha_{t,k}\widehat{\ell}_{t,k})\right) \le 1+2\gamma_T\alpha_{t,k}\widehat{\ell}_{t,k}\end{aligned}$$

The last inequality sign follows by $x\log(1+y) \le \log(1+xy), \forall x \in [0,1], y \in [0,\infty)$. Now multiplying the above two inequalities with respect to $k \in [K]$ yields

$$\exp\left(2\gamma_t \sum_{k=1}^{K} \widetilde{\ell}_{t,k}\right) \le \prod_{k=1}^{K}(1+2\gamma_t \widehat{\ell}_{t,k}) = 1 + 2\gamma_t \sum_{k=1}^{K} \widehat{\ell}_{t,k}$$

$$\exp\left(2\gamma_T \sum_{k=1}^{K} \alpha_{t,k}\widetilde{\ell}_{t,k}\right) \le \prod_{k=1}^{K}(1+2\gamma_T \alpha_{t,k}\widehat{\ell}_{t,k}) = 1 + 2\gamma_T \sum_{k=1}^{K} \alpha_{t,k}\widehat{\ell}_{t,k}$$

where the first inequality assumes $\alpha_{t,k} \equiv 1$ and both equality signs hold as $\widehat{\ell}_{t,k} \neq 0$ iff $a_t = k$. Taking the expectation of above two inequalities we obtain

$$\mathbb{E}\left[\exp\left(2\gamma_t \sum_{k=1}^{K} \widetilde{\ell}_{t,k}\right) \middle| \mathcal{F}_{t-1}\right] \le 1 + 2\gamma_t \sum_{k=1}^{K} \ell_{t,k} \le \exp\left(2\gamma_t \sum_{k=1}^{K} \ell_{t,k}\right)$$

$$\mathbb{E}\left[\exp\left(2\gamma_T \sum_{k=1}^{K} \alpha_{t,k}\widetilde{\ell}_{t,k}\right) \middle| \mathcal{F}_{t-1}\right] \le 1 + 2\gamma_T \sum_{k=1}^{K} \alpha_{t,k}\ell_{t,k} \le \exp\left(2\gamma_T \sum_{k=1}^{K} \alpha_{t,k}\ell_{t,k}\right)$$

Multiplying them with respect to $t \in [T]$ yields

$$\mathrm{EVaR}_{1-\alpha}\left(2\sum_{t=1}^{T}\gamma_t \sum_{k=1}^{K}\left(\widetilde{\ell}_{t.k} - \ell_{t,k}\right)\right) \le \log\frac{1}{\alpha}$$

$$\mathrm{EVaR}_{1-\alpha}\left(2\gamma_T \sum_{t=1}^{T}\sum_{k=1}^{K}\alpha_{t,k}\left(\widetilde{\ell}_{t,k} - \ell_{t,k}\right)\right) \le \log\frac{1}{\alpha}$$

Here by letting $\alpha_{t,k} = \mathbb{I}\{k = *\}$ we obtain $\mathrm{EVaR}_{1-\alpha}(2\gamma_T \sum_{t=1}^{T}(\widetilde{\ell}_{t,*} - \ell_{t,*})) \le \log\frac{1}{\alpha}$. Recall that EVaR is an upper bound of CVaR (Ahmadi-Javid, 2012), we yield bounds for the CVaR of the two terms.

Step III: summing

In the horizon known case, by letting $\eta_t = \eta$ and $\gamma_t = \gamma$ we yield

$$\begin{aligned}\mathrm{CVaR}_{1-\alpha}(R_T) &\le \left(\frac{\eta}{4\gamma} + \frac{1}{2}\right)\left(2\gamma KT + \log\frac{1}{\alpha}\right) + \frac{1}{2\gamma}\log\frac{1}{\alpha} + \frac{\log K}{\eta} \\ &= \frac{\eta KT}{2} + \frac{\log K}{\eta} + \gamma KT + \frac{1}{2\gamma}\log\frac{1}{\alpha} + \left(\frac{\eta}{4\gamma} + \frac{1}{2}\right)\log\frac{1}{\alpha}\end{aligned}$$

In the anytime case, by letting $\eta_t = \eta_1/\sqrt{t}$ and $\gamma_t = \gamma_1/\sqrt{t}$ and following $\sum_{t=1}^{T} 1/\sqrt{t} \le \int_0^T 1/\sqrt{t}\mathrm{d}t = 2\sqrt{T}$ we yield

$$\begin{aligned}\mathrm{CVaR}_{1-\alpha}(R_T) &\le \left(\frac{\eta_1}{4\gamma_1} + \frac{1}{2}\right)\left(4\gamma_1 K\sqrt{T} + \log\frac{1}{\alpha}\right) + \frac{\sqrt{T}}{2\gamma_1}\log\frac{1}{\alpha} + \frac{\sqrt{T}\log K}{\eta_1} \\ &= \eta_1 K\sqrt{T} + \frac{\sqrt{T}\log K}{\eta_1} + 2\gamma_1 K\sqrt{T} + \frac{\sqrt{T}}{2\gamma_1}\log\frac{1}{\alpha} + \left(\frac{\eta_1}{4\gamma_1} + \frac{1}{2}\right)\log\frac{1}{\alpha}\end{aligned}$$

By substituting the values we complete the proof. ■

The result for the adversarial setting can be generalised to the contextual and non-stationary cases with little effort. In the setting of bandit with expert advice, the modified version of Exp4 (Auer et al., 1995) that includes implicit exploration yields the Exp4-IX algorithm (see Algorithm 4). Similar to the adversarial case, Exp4-IX matches the expected regret $\sqrt{2KT\log N}$ and $2\sqrt{KT\log N}$ of Exp4 in the horizon-known and anytime cases (Bubeck and Cesa-Bianchi, 2012), respectively.

Algorithm 4: Exp4-IX

Input: $\{\eta_t\}_{t=1}^T, \{\gamma_t\}_{t=1}^T$
Initialization: $\widetilde{L}_0(n) = 0, \forall n \in [N]$
for $t = 1, \dots, T$ **do**
  $q_t(n) \leftarrow \frac{\exp(-\eta_t \widetilde{L}_{t-1}(n))}{\sum_{n=1}^N \exp(-\eta_t \widetilde{L}_{t-1}(n))}, \forall n \in [N]$
  $p_{t,k} \leftarrow \sum_{n=1}^N \xi_{t,k}(n) q_t(n), \forall k \in [K]$
  Sample $a_t \sim \boldsymbol{p}_t = (p_{t,1}, \dots, p_{t,K})$
  $\widetilde{L}_t(n) \leftarrow \widetilde{L}_{t-1}(n) + \boldsymbol{\xi}_t(n)^T \widetilde{\boldsymbol{\ell}}_t, \forall n \in [N]$
end

Theorem 4 *For Algorithm 4 with* $\eta_t = \sqrt{\frac{2\log N}{KT}}$ *and* $\gamma_t = \sqrt{\frac{\log\frac{1}{\alpha}}{2KT}}$, *we have:*

$$\mathrm{CVaR}_{1-\alpha}(R_T) \le \sqrt{2KT\log N} + \sqrt{2KT\log\frac{1}{\alpha}} + \frac{1}{2}\sqrt{\log N\log\frac{1}{\alpha}} + \frac{1}{2}\log\frac{1}{\alpha}$$

For Algorithm 4 with $\eta_t = \sqrt{\frac{\log N}{Kt}}$ *and* $\gamma_t = \frac{1}{2}\sqrt{\frac{\log\frac{1}{\alpha}}{Kt}}$, *we have:*

$$\mathrm{CVaR}_{1-\alpha}(R_T) \le 2\sqrt{KT\log N} + 2\sqrt{KT\log\frac{1}{\alpha}} + \frac{1}{2}\sqrt{\log N\log\frac{1}{\alpha}} + \frac{1}{2}\log\frac{1}{\alpha}$$

Proof Denote $*$ as the optimal expert. By the standard analysis for Exp3 of Bubeck and Cesa-Bianchi (2012), for any non-increasing sequence $\{\eta_t\}_{t=1}^T$ we have

$$\sum_{t=1}^T\sum_{k=1}^K p_{t,k}\widetilde{\ell}_{t,k} \le \sum_{t=1}^T \frac{\eta_t}{2}\sum_{n=1}^N q_t(n)\left(\sum_{k=1}^K \xi_{t,k}(n)\widetilde{\ell}_{t,k}\right)^2 + \sum_{t=1}^T\sum_{k=1}^K \xi_{t,k}(*)\widetilde{\ell}_{t,k} + \frac{\log N}{\eta_T}$$

Following Jensen's inequality

$$\sum_{n=1}^N q_t(n)\left(\sum_{k=1}^K \xi_{t,k}(n)\widetilde{\ell}_{t,k}\right)^2 \le \sum_{n=1}^N q_t(n)\sum_{k=1}^K \xi_{t,k}(n)\widetilde{\ell}_{t,k}^2 = \sum_{k=1}^K p_{t,k}\widetilde{\ell}_{t,k}^2 \le \sum_{k=1}^K \widetilde{\ell}_{t,k}$$

Recall that $\ell_{t,a_t} = \sum_{k=1}^K (p_{t,k} + \gamma_t)\widetilde{\ell}_{t,k}$. Then

$$\begin{aligned} R_T &= \sum_{t=1}^T \ell_{t,a_t} - \sum_{t=1}^T\sum_{k=1}^K \xi_{t,k}(*)\ell_{t,k} = \sum_{t=1}^T\sum_{k=1}^K (p_{t,k}+\gamma_t)\widetilde{\ell}_{t,k} - \sum_{t=1}^T\sum_{k=1}^K \xi_{t,k}(*)\ell_{t,k} \\ &\le \sum_{t=1}^T\left(\frac{\eta_t}{2}+\gamma_t\right)\sum_{k=1}^K \widetilde{\ell}_{t,k} + \sum_{t=1}^T\sum_{k=1}^K \xi_{t,k}(*)\left(\widetilde{\ell}_{t,k} - \ell_{t,k}\right) + \frac{\log N}{\eta_T} \end{aligned}$$

By a similar argument as in Theorem 3 where $\alpha_{t,k} = \xi_{t,k}(*)$ and substituting the values we complete the proof. ■

In the setting of tracking the best sequence, the modified version of the shared algorithm in Cesa-Bianchi et al. (2012) that includes implicit exploration yields Exp3-SIX (see Algorithm 5). It is worth noting that explicit exploration is also included in the algorithm due to technical need. This scenario is more complicated so that additional logarithmic term occur in the bound. Moreover, proposing anytime parameters with good theoretical guarantees remains open.

Algorithm 5: Exp3-SIX

Input: $\{\eta_t\}_{t=1}^T, \{\gamma_t\}_{t=1}^T, \{\beta_t\}_{t=1}^T$
Initialization: $p_{1,k} = 1/K, \forall k \in [K]$
for $t = 1, \dots, T$ **do**
  Sample $a_t \sim \boldsymbol{p}_t = (p_{t,1}, \dots, p_{t,K})$
  $v_{t+1,k} \leftarrow \frac{p_{t,k}\exp\left(-\eta_t \widetilde{\ell}_{t,k}\right)}{\sum_{k=1}^K p_{t,k}\exp\left(-\eta_t \widetilde{\ell}_{t,k}\right)}, \forall k \in [K]$
  $p_{t+1,k} \leftarrow (1-\beta_t)v_{t+1,k} + \beta_t/K, \forall k \in [K]$
end

Theorem 5 *For Algorithm 5 with* $\beta_t = \frac{S}{T-1}$, $\eta_t = \sqrt{\frac{2\overline{S}}{KT}\log\frac{eKT}{S}}$, *and* $\gamma_t = \sqrt{\frac{\log\frac{1}{\alpha}}{2KT}}$, *where* $\overline{S} = S + 1$, *we have:*

$$\mathrm{CVaR}_{1-\alpha}(R_T) \le \sqrt{2KT\overline{S}\log\frac{eKT}{S}} + \sqrt{2KT\log\frac{1}{\alpha}} + \frac{1}{2}\sqrt{\overline{S}\log\frac{eKT}{S}\log\frac{1}{\alpha}} + \frac{1}{2}\log\frac{1}{\alpha}$$

Proof Since $\exp(-x) \le x^2/2 - x + 1, \forall x \ge 0$, we have

$$\begin{aligned}\sum_{k=1}^K p_{t,k}\widetilde{\ell}_{t,k} &\le -\frac{1}{\eta_t}\left(\sum_{k=1}^K p_{t,k}\exp\left(-\eta_t\widetilde{\ell}_{t,k}\right) - 1\right) + \frac{\eta_t}{2}\sum_{k=1}^K p_{t,k}\widetilde{\ell}_{t,k}^2 \\ &\le -\frac{1}{\eta_t}\log\left(\sum_{k=1}^K p_{t,k}\exp\left(-\eta_t\widetilde{\ell}_{t,k}\right)\right) + \frac{\eta_t}{2}\sum_{k=1}^K \widetilde{\ell}_{t,k}\end{aligned}$$

The second inequality sign holds by $\log x \le x - 1, \forall x > 0$. One may find, for any $k \in [K]$

$$-\frac{1}{\eta_t}\log\left(\sum_{k=1}^K p_{t,k}\exp\left(-\eta_t\widetilde{\ell}_{t,k}\right)\right) = -\frac{1}{\eta_t}\log\frac{p_{t,k}\exp\left(-\eta_t\widetilde{\ell}_{t,k}\right)}{v_{t+1,k}} = \widetilde{\ell}_{t,k} - \frac{1}{\eta_t}\log\frac{p_{t,k}}{v_{t+1,k}}$$

where the RHS is seemingly dependent on k but essentially not. Thus, for any distribution $\boldsymbol{q}_t = (q_{t,1}, \dots, q_{t,K})$ over $[K]$, we have

$$\sum_{k=1}^K p_{t,k}\widetilde{\ell}_{t,k} - \sum_{k=1}^K q_{t,k}\widetilde{\ell}_{t,k} \le -\frac{1}{\eta_t}\sum_{k=1}^K q_{t,k}\log\frac{p_{t,k}}{v_{t+1,k}} + \frac{\eta_t}{2}\sum_{k=1}^K \widetilde{\ell}_{t,k}$$

Denote J_t^* as the best sequence. Then, by the analysis of Theorem 2 in Cesa-Bianchi et al. (2012), we obtain:

$$\begin{aligned}\sum_{t=1}^T\sum_{k=1}^K p_{t,k}\widetilde{\ell}_{t,k} - \sum_{t=1}^T \widetilde{\ell}_{t,J_t^*} \le \sum_{t=1}^T \frac{\eta_t}{2}\sum_{k=1}^K \widetilde{\ell}_{t,k} + \sum_{t=2}^T \frac{1}{\eta_{t-1}}\log\frac{1}{1-\beta_t} \\ + \frac{S}{\eta_T}\log\frac{K(1-\beta_T)}{\beta_T} + \frac{1}{\eta_T}\log K\end{aligned}$$

Recall that $\ell_{t,a_t} = \sum_{k=1}^{K}(p_{t,k} + \gamma_t)\widetilde{\ell}_{t,k}$. Then

$$\begin{aligned} R_T &= \sum_{t=1}^{T} \ell_{t,a_t} - \sum_{t=1}^{T} \ell_{t,J_t^*} = \sum_{t=1}^{T}\sum_{k=1}^{K}(p_{t,k} + \gamma_t)\widetilde{\ell}_{t,k} - \sum_{t=1}^{T} \ell_{t,J_t^*} \\ &\leq \sum_{t=1}^{T}\left(\frac{\eta_t}{2} + \gamma_t\right)\sum_{k=1}^{K}\widetilde{\ell}_{t,k} + \sum_{t=1}^{T}\left(\widetilde{\ell}_{t,J_t^*} - \ell_{t,J_t^*}\right) \\ &+ \sum_{t=2}^{T}\frac{1}{\eta_{t-1}}\log\frac{1}{1-\beta_t} + \frac{S}{\eta_T}\log\frac{K(1-\beta_T)}{\beta_T} + \frac{1}{\eta_T}\log K \end{aligned}$$

By a similar argument as in Theorem 3 where $\alpha_{t,k} = \mathbb{I}\{J_t^* = k\}$, substituting the values, and $-x\log x - (1-x)\log(1-x) \leq x\log(e/x), \forall x \in (0,1)$ we complete the proof. ■

4. Empirical Evaluation

We validate our analyses in two scenarios. In the following content, without special specification, CVaR-regret refers to the CVaR of regret at level 0.95. In each setting of Scenario 1 and 2, we repeat for 500 and 100 times and adopt bootstrapping with size 5,000 and 1,000 respectively to estimate the expected regret and CVaR-regret with their standard errors.

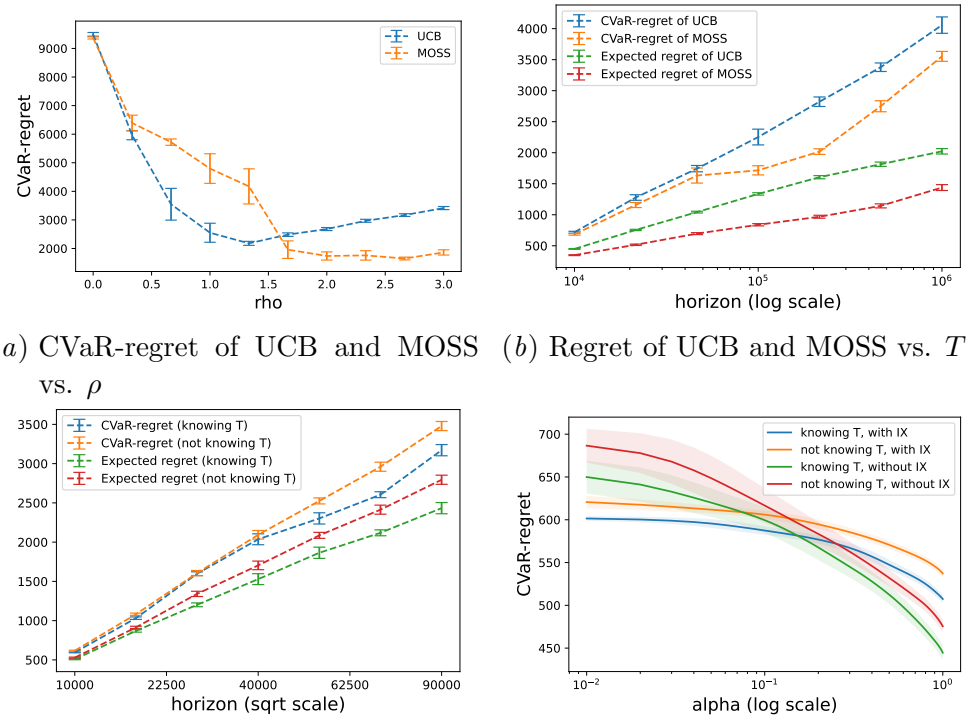
4.1. Scenario 1: Bandit with Low Gains

We set $K = 10$ and generate the gains of all arms from independent Bernoulli trials, which is a frequent scenario in Internet advertising. The mean of the first arm is set as 0.1 and the other 9 arms are evenly divided into 3 groups with means 0.05, 0.02, and 0.01, which is the same as Scenario 2 of Garivier and Cappé (2011). We firstly tuned the exploration rate ρ for UCB and MOSS in the case $T = 10^5$, as illustrated in Figure 1(*a*). In the Bernoulli scenario, the CVaR-regret of UCB is guaranteed to be logarithmic when $\rho \geq 1$. Empirically the CVaR-regret sharply drops before ρ reaches the value around the threshold and then slowly rises. MOSS shows a similar pattern while the CVaR-regret is more stable with large exploration rate. We then test tuned UCB, MOSS, and Exp3-IX with its anytime version under different horizons, where the results are shown in Figures 1(*b*) and 1(*c*). The parameters η_t and γ_t of Exp3-IX in the experiments are as proposed in Theorem 3. It is in line with our expectations that CVaR-regret and expected regret are of $\Theta(\log T)$ and $\Theta(\sqrt{T})$ in the stochastic and adversarial settings and the CVaR-regret suffers a greater factor.

We also consider the case $T = 10^4$ to illustrate how CVaR-regret varies with respect to α in the adversarial case, as demonstrated in Figure 1(*d*). Here without implicit exploration (IX) refers to setting $\gamma_t = 0$, where the Exp3-IX boils down to Exp3. Recall that $\text{CVaR}_0(X) = \mathbb{E}[X]$, thus the CVaR-regret corresponding to $\alpha = 1$ is the expected regret. It can be observed that Exp3 enjoys a lower expected regret while suffering a higher variability.

4.2. Scenario 2: Non-stationary Bandit

Inspired by Neu (2015), we again generate the losses of all arms from independent Bernoulli trials but reduce the number of arms to 2. The means of the first and second arms are set



(a) CVaR-regret of UCB and MOSS vs. ρ

(b) Regret of UCB and MOSS vs. T

(c) Regret of Exp3-IX vs. T

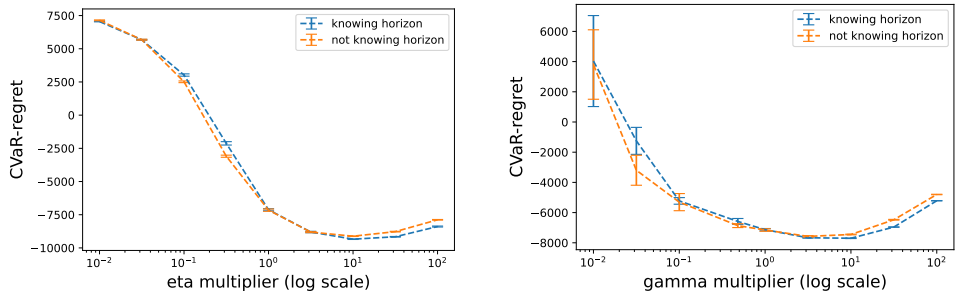
(d) CVaR-regret of Exp3-IX vs. α

Figure 1: Experiments in the Scenario 1

as $0.5+\Delta$ and $0.5-\Delta$ for rounds $t \leq \frac{T}{2}$ and $0.5-4\Delta$ and $0.5+\Delta$ for rounds $t > \frac{T}{2}$, which guarantees the second arm to be the best one up to the first half of the game while the first arm eventually becomes the leader. We set $T = 10^6, \Delta = 0.1$ and test the performance of Exp3-IX with different parameters, where η and γ are initially set as proposed in Theorem 3 and the multiplier is varied between 10^{-2} and 10^2. The results are presented in Figures 2(a) and 2(b), where the parameters we propose almost minimize the CVaR-regret. Remarkably, the data point closest to the left of base parameter in Figure 2(b) adopts the parameter proposed by Neu (2015), which is inferior to our choice. Empirically larger values of η_t and γ_t lead to a better performance while the gap shrinks with respect to the horizon.

5. Conclusion

We presented the analyses of the newly-introduced CVaR-regret for stochastic and non-stochastic MAB algorithms, which match the results of the traditional expected regret framework and yield bounds (nearly-)optimal up to constant. For UCB, the exploration that guarantees a logarithmic growing expected regret is sufficient for governing the tail risk. The case for MOSS is more complicated, where extra exploration needs possibly to be implemented to cope with the worst instance. For Exp3-IX, we recommend the implicit exploration rate γ_t to be associated with the selected level α and reveal that the implicit exploration controls the tail risk at the cost of suffering a larger expected regret.



(a) CVaR-regret vs. η-multiplier (b) CVaR-regret vs. γ-multiplier

Figure 2: Experiments in the Scenario 2

Acknowledgments

This work is supported in part by the program of National Natural Science Foundation of China (No. 62176154) and the program of the Shanghai NSF (No. 19ZR1426700).

References

Amir Ahmadi-Javid. Entropic value-at-risk: A new coherent risk measure. *Journal of Optimization Theory and Applications*, 155(3):1105–1123, 2012.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematics Finance*, 9(3):203–228, 1999.

Jean Yves Audibert and Sbastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, 2009.

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

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *FOCS*, 1995.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine Learning*, 47:235–256, 2002.

Dorian Baudry, Romain Gautron, Emilie Kaufmann, and Odalric-Ambrym Maillard. Optimal thompson sampling strategies for support-aware CVaR bandits. In *ICML*, 2021.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in ML*, 5(1):1–122, 2012.

Nicolò Cesa-Bianchi, Pierre Gaillard, Gábor Lugosi, and Gilles Stoltz. Mirror descent meets fixed share (and feels no regret). In *NeurIPS*, 2012.

Nicolas Galichet, Michèle Sebag, and Olivier Teytaud. Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. In *ACML*, 2013.

Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *COLT*, 2011.

Mark Herbster and Manfred K. Warmuth. Tracking the best expert. *Machine Learning*, 32:151–178, 1998.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

Tomáš Kocák, Gergely Neu, Michal Valko, and Rémi Munos. Efficient learning by implicit exploration in bandit problems with side observations. In *NeurIPS*, 2014.

Prashanth L.A., Krishna Jagannathan, and Ravi Kolla. Concentration bounds for CVaR estimation: The cases of light-tailed and heavy-tailed distributions. In *ICML*, 2020.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

Odalric-Ambrym Maillard. Robust risk-averse stochastic multi-armed bandits. In *ALT*, 2013.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *NeurIPS*, 2015.

Amir Sani, Alessandro Lazaric, and Rémi Munos. Risk-aversion in multi-armed bandits. In *NeurIPS*, 2012.

Weiwei Shen, Jun Wang, Yu-Gang Jiang, and Hongyuan Zha. Portfolio choices with orthogonal bandit learning. In *IJCAI*, 2015.

Vincent Y. F. Tan, Prashanth L. A., and Krishna Jagannathan. A survey of risk-aware multi-armed bandits. In *IJCAI*, 2022.

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

John White. *Bandit algorithms for website optimization*. O'Reilly Media, Inc., 2012.

Qiuyu Zhu and Vincent Y. F. Tan. Thompson sampling algorithms for mean-variance bandits. In *ICML*, 2020.