A One-Size-Fits-All Solution to Conservative Bandit Problems

Yihan Du,[1] Siwei Wang,[1] Longbo Huang[1]
[1] Tsinghua University
duyh18@mails.tsinghua.edu.cn, {wangsw2020, longbohuang}@tsinghua.edu.cn

arXiv:2012.07341v3 [cs.LG] 16 Dec 2020

Abstract

In this paper, we study a family of conservative bandit problems (CBPs) with *sample-path* reward constraints, i.e., the learner's reward performance must be at least as well as a given baseline at any time. We propose a general one-size-fits-all solution to CBPs and present its applications to three encompassed problems, i.e., conservative multi-armed bandits (CMAB), conservative linear bandits (CLB) and conservative contextual combinatorial bandits (CCCB). Different from previous works which consider high probability constraints on the expected reward, our algorithms guarantee sample-path constraints on the actual received reward, and achieve better theoretical guarantees (T-independent additive regrets instead of T-dependent) and empirical performance. Furthermore, we extend the results and consider a novel conservative *mean-variance* bandit problem (MV-CBP), which measures the learning performance in both the expected reward and variability. We design a novel algorithm with $O(1/T)$ normalized additive regrets (T-independent in the cumulative form) and validate this result through empirical evaluation.

1 Introduction

The multi-armed bandit (MAB) problem (Thompson 1933; Auer, Cesa-Bianchi, and Fischer 2002) is a classic online learning model that characterizes the exploration-exploitation trade-off in sequential decision making. While existing bandit algorithms achieve satisfactory regret bounds over the whole learning processes, they can perform wildly and lose much in the initial exploratory phase. This limitation has hindered their applications in real-world scenarios such as health sciences, marketing and finance, where it is important to guarantee safe and smooth algorithm behavior in initialization. Hence, studying bandit problems with safe (conservative) exploration contributes to solving this issue.

In this paper, we study the conservative bandit problems (CBPs) with *sample-path* reward constraints. Specifically, a learner is given a set of regular arms and a default arm. At each timestep, the learner chooses a regular arm or the default arm to play and receives a reward according to the played arm. The learning's objective is to minimize the expected cumulative regret (equivalently, maximize the expected cumulative reward), while ensuring that the received cumulative reward must stay above a fixed percentage of what one can obtain by always playing the default arm.

CBPs have extensive real-world applications including recommendation systems, company operation and finance. For instance, in finance, investors are offered various financial products including the fixed-income security such as bank deposit (default arm), and the fluctuating equity securities such as stocks (regular arms). While the fixed-income security is a safe and reasonable option, investors want to find better choices to earn higher returns. Meanwhile, compared to the returns they can obtain by simply depositing the money, investors do not want to lose too much when exploring other investment choices. CBPs provide an effective model for such exploration-exploitation trade-off with the safe exploration guarantees.

We propose a general one-size-fits-all solution GenCB for CBPs, and present its applications to three important CBP problems, i.e., conservative multi-armed bandits (CMAB), conservative linear bandits (CLB) and conservative contextual combinatorial bandits (CCCB). We provide theoretical analysis and empirical evaluations for these algorithms, and show that our algorithms outperform existing ones both theoretically and empirically. Table 1 presents the comparison of regret bounds between our algorithms and existing ones. In the table, each regret term contains two components, the first component incurred by regular arms and the second term due to playing the default arm. One can see that our algorithms possess better regret guarantees. Moreover, unlike existing algorithms that only provide high probability bounds with T-dependent conservative regrets, we not only obtain expected bounds but also have T-independent conservative regrets.

Our work distinguishes itself from previous conservative bandit works, e.g., (Wu et al. 2016; Kazerouni et al. 2017; Garcelon et al. 2020; Zhang, Li, and Liu 2019) in two aspects: (i) Previous works consider high probability guarantees on the expected reward. Such models cannot directly handle many risk-adverse tasks, e.g., a start-up does not wish to tolerate any failure probability to reach the basic earning under the debt, or an asset management company must perform better than the promised return. While one can choose a very small δ in previous algorithms to provide high-probability guarantees, the $\ln(1/\delta)$-dependent regrets will boost accordingly. Instead, we focus on a certainty

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Table 1: Comparison of regret bounds for CBPs. "Type" refers to the type of regret bounds. "E" and "H" denote the expected and high probability bounds, respectively. Here $H = \sum_{i=1}^{K} \Delta_i^{-1}$. d is the dimension in CLB and CCCB. For high probability bounds, the convention in the bandit literature is to choose $\delta = 1/T$. Note that our formulation focuses on a sample-path reward constraint, while the other results consider the constraints on the expected reward.

Problem	Algorithm	Regret bound	Type
CMAB	GenCB-CMAB (ours)	$O\big(H \ln T + \frac{H}{\alpha}[\ln(\frac{H}{\alpha})]^2\big)$	E
CMAB	ConUCB (Wu et al. 2016)	$O\big(H \ln(\frac{T}{\delta}) + \sum_{i=1}^{K} \frac{1}{\alpha \max\{\Delta_i, \Delta_0 - \Delta_i\}} \ln(\boldsymbol{T}/\delta)\big)$	H
CMAB	Lower Bound (Wu et al. 2016)	$O\big(\max\{\frac{K}{\alpha}, \sqrt{KT}\}\big)$	E
CLB	GenCB-CLB (ours)	$O\big(d \ln(T)\sqrt{T} + \frac{d^2}{\alpha}[\ln(\frac{d}{\alpha})]^2\big)$	E
CLB	CLUCB (Kazerouni et al. 2017)	$O\big(d \ln(\frac{T}{\delta})\sqrt{T} + \frac{d^2}{\alpha}[\ln(\frac{d}{\alpha\boldsymbol{\delta}})]^2\big)$	H
CLB	CLUCB2 (Garcelon et al. 2020)	$O\big(d \ln(\frac{T}{\delta})\sqrt{T} + \frac{d^2}{\alpha^2}[\ln(\frac{d}{\alpha\boldsymbol{\delta}})]^2\big)$	H
CCCB	GenCB-CCCB (ours)	$O\big(d \ln(KT)\sqrt{T} + \frac{(K+d)^2}{\alpha}[\ln(\frac{K+d}{\alpha})]^2\big)$	E
CCCB	C3UCB (Zhang, Li, and Liu 2019)	$O\big(d \ln(KT)\sqrt{T} + \frac{d}{\alpha}\sqrt{\frac{d}{K}\ln(\frac{K}{\delta}\boldsymbol{T})}\big)$	H

(sample-path) guarantee on the actual empirical reward. Doing so ensures safe exploration (our regret bounds do not contain δ) and better suits such tasks. (ii) Our problem formulation, solution and analysis offer a general framework for studying a family of CBPs, including CMAB (Wu et al. 2016), CLB (Kazerouni et al. 2017; Garcelon et al. 2020) and CCCB (Zhang, Li, and Liu 2019). Moreover, our algorithms achieve better theoretical and empirical performance than previous schemes.

We also extend our results to the mean-variance setting (Markowitz et al. 1952; Sani, Lazaric, and Munos 2012), called *conservative mean-variance bandit problem* (MV-CBP), which focuses on the balance between the expected reward and variability with safe exploration. Different from the typical CBPs which only consider the expected reward into learning performance, MV-CBP takes into account both the mean and variance of the arms, and is more suitable for practical tasks that are sensitive to reward fluctuations, e.g., clinical trials and finance. For example, many risk-adverse investors prefer stable assets (e.g., bonds) with satisfactory returns than volatile assets (e.g., derivatives) with high returns, and they do not want to suffer wild fluctuations when exploring different financial products.

Note that the mean-variance regret in MV-CBP (formally defined in Eq. (4) in next section) consists not only the gap of *mean-variance* (a combination of both measures) between the played arms and the optimal arm, but also an additional variance for playing arms with different means, called *exploration risk*, which requires alternative techniques beyond those in typical CBPs. To tackle this issue, we carefully adapt our solution and analysis for the CBPs and make non-trivial extensions. Our results offer new insight into algorithm design for mean-variance bandit problems.

Our contributions are summarized as follows.

- We study a family of CBPs with sample-path reward constraints, which encompasses previously studied CMAB (Wu et al. 2016), CLB (Kazerouni et al. 2017; Garcelon et al. 2020) and CCCB (Kazerouni et al. 2017). We propose a general one-size-fits-all solution GenCB for CBPs, which can translate a standard bandit algorithm into a conservative bandit algorithm and achieve better (T-independent conservative regret rather than T-dependent) theoretical regret bounds than previous works in the three specific problems.
- We extend the conservative bandit formulation to a novel conservative mean-variance bandit setting, which characterizes the trade-off between the expected reward and variability. We propose an algorithm, MV-CUCB, and prove that it achieves an $O(1/T)$ normalized additive regret for the extended problem.
- We conduct extensive experiments for the considered problems. The results match our theoretical bounds and demonstrate that our algorithms achieve the performance superiority compared to existing algorithms.

1.1 Related Work

Conservative Bandit Literature. Recently, there are several works (Wu et al. 2016; Kazerouni et al. 2017; Zhang, Li, and Liu 2019; Garcelon et al. 2020) studying bandit problems with conservative exploration constraints. Under the constraints on the expected rewards, (Wu et al. 2016) propose an algorithm ConUCB for CMAB. (Kazerouni et al. 2017) design an algorithm CLUCB for CLB and (Garcelon et al. 2020) further propose an improved algorithm CLUCB2. (Zhang, Li, and Liu 2019) present an algorithm C3UCB for CCCB. Under the stage-wise constraints, (Khezeli and Bitar 2020) restrict the expected reward at any timestep to stay above a given baseline. (Amani, Alizadeh, and Thrampoulidis 2019) confine the played arm at any timestep to stay in a given safe set. Under the interleaving constraint, (Katariya et al. 2019) require the chosen action at any timestep to perform better than the default ac-

tion when interleaving in the combinatorial semi-bandit setting. (Bubeck, Perchet, and Rigollet 2013) study the standard K-armed bandit problem with knowledge of the highest expected reward and the smallest gap, (Locatelli, Gutzeit, and Carpentier 2016) consider the thresholding pure exploration problem, and the settings and methods in both works are different from ours.

Mean-variance Bandit Literature. (Sani, Lazaric, and Munos 2012) open the mean-variance bandit literature which considers both the expected reward and variability into performance measures, and a series of follow-ups (Maillard 2013; Vakili, Boukouvalas, and Zhao 2019; Cardoso and Xu 2019) have emerged recently. To our best knowledge, this paper is the first to study the mean-variance bandit problem with conservative exploration.

2 Problem Formulation

In this section, we first review previous standard (non-conservative) bandit problems (SBPs) and then give the formulation of the Conservative Bandit Problems (CBPs).

Standard Bandit Problems (SBPs). In a standard bandit problem, a learner is given a set of arms $\mathcal{X}$, where each arm $x \in \mathcal{X}$ has an unknown reward distribution in $[0, 1]$ with mean of μ_x. Each arm x at timestep t has a random reward $r_{t,x} = \mu_x + \eta_{t,x}$, where $\eta_{t,x}$ is an independent random noise with respect to t. At each timestep t, the learner plays an arm x_t and only observes the reward r_{t,x_t} of the chosen arm. Let $x_* = \operatorname{argmax}_{x\in\mathcal{X}} \mu_x$ denote the optimal arm. The learning performance over a time horizon T is measured by *expected cumulative regret*

$$\mathbb{E}[\mathcal{R}_T] = \mu_{x_*}T - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{x_t}\right] = \sum_{x\neq x_*} \mathbb{E}[N_x(T)]\Delta_x, \quad (1)$$

where $\Delta_x = \mu_{x_*} - \mu_x$ and $N_x(T)$ is the number of times arm x was played over time T. The regret characterizes the loss due to not always playing the optimal arm. The goal of standard bandit algorithms is to minimize Eq. (1).

Conservative Bandit Problems (CBPs). The CBPs provide an alternative default arm x_0 to play. In this case, since playing x_0 is a default (baseline) policy that the learner is familiar with, for ease of analysis we assume that x_0 has a known constant reward $0 < \mu_0 < \mu_{x_*}$ as previous works (Wu et al. 2016; Kazerouni et al. 2017; Zhang, Li, and Liu 2019) do.[1]

Then, during the learning process, the learner is required to ensure that the cumulative reward under the chosen policy is lower bounded by a fraction of the reward from always pulling the default arm. Specifically, given a parameter $\alpha \in (0, 1)$, for any timestep t, the learner's cumulative empirical reward should be least $1-\alpha$ fraction of the reward of always playing x_0, i.e.,

$$\sum_{s=1}^{t} r_{s,x_s} \geq (1-\alpha)\mu_0 t, \quad \forall t \in \{1, \ldots, T\}. \quad (2)$$

Here α controls the strictness of the constraint, i.e., how conservative we want the leaner to behave, and can be viewed as the weight we place on safety in exploration. The goal of conservative bandit algorithms is to minimize the expected cumulative regret (Eq. (1)) while satisfying the reward constraint (Eq. (2)).

We note that constraint (2) is a *sample-path* reward constraint, which is different from the high-probability constraints on the expected reward in prior works (Wu et al. 2016; Kazerouni et al. 2017; Garcelon et al. 2020; Zhang, Li, and Liu 2019). This setting is particularly useful when the practical tasks cannot tolerate higher losses than the baseline with certainty, e.g., health care and investment. On the other hand, it also imposes new challenges in algorithm design and regret analysis.

Our formulation is a general framework which encompasses various bandit problems from the prospective of conservative exploration. For example, in CMAB which studies a conservative version of the classic K-armed bandit problem (Thompson 1933; Auer, Cesa-Bianchi, and Fischer 2002; Agrawal and Goyal 2012), $\mathcal{X} = [K]$ and μ_x is an arbitrary value.[2] In CLB which considers the linear bandit problem (Dani, Hayes, and Kakade 2008; Abbasi-yadkori, Pál, and Szepesvári 2011) with conservative exploration, $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$ and each arm $x \in \mathbb{R}^d$ has an expected reward $\mu_x = x^\top\theta^*$, where $\theta^* \in \mathbb{R}^d$ is an unknown parameter. In CCCB which investigates the contextual combinatorial bandit problem (Qin, Chen, and Zhu 2014) with the safe exploration requirement, there is a set of base arms $[K]$ and $\mathcal{X}$ is a collection of subsets of base arms, which represents certain combinatorial structure (e.g., matchings and paths). For each $x \in \mathcal{X}$, μ_x is associated with the expected rewards of its containing base arms. We will analyze the CBPs under specific bandit settings in the next section.

3 A General Solution to Conservative Bandits

In this section, we first present a general solution for CBPs, and its regret analysis. Then, we present its applications to three specific problems, i.e., CMAB, CLB and CCCB, and show that in all three cases, our algorithm achieves tighter bounds than existing algorithms.

Algorithm 1 illustrates the proposed solution to CBPs, called GenCB, which offers a general scheme for translating a standard non-conservative bandit algorithm $\mathcal{A}_S$ into a conservative bandit algorithm. In the algorithm, m denotes the time horizon of $\mathcal{A}_S$, and the number of times we play the regular arms, $r_S(t)$ denotes the cumulative reward from sampling regular arms, and $N_0(t)$ denotes the number of times x_0 is played up to time t.

The main idea of GenCB is to play regular arms as much as possible while ensuring the sample-path reward constraint

[1] This assumption can be relaxed to that x_0 has a random reward within a known interval $[r_0^\ell, r_0^h]$ ($r_0^\ell > 0$) by sightly changing the right-hand-side of the `if` statements in our algorithms, and our analysis procedure still works. While previous works can remove this assumption by estimating μ_0, this is due to that their constraints are imposed on the expected reward.

[2] $[K] \stackrel{\text{def}}{=} \{1, \ldots, K\}$.

Algorithm 1: General Solution to Conservative Bandits (GenCB)

Input: Standard bandit problem and algorithm $\mathcal{A}_S$, regular arms $\mathcal{X}$, default arm x_0 with reward μ_0, parameter α.

$\forall t \geq 0, N_0(t) \leftarrow 0, r_S(t) \leftarrow 0.\ m \leftarrow 0;$
for $t = 1, 2, \ldots$ **do**
  if $r_S(t-1) + N_0(t-1)\mu_0 \geq (1-\alpha)\mu_0 t$ **then**
    $m \leftarrow m + 1;$
    Play an arm x_t according to $\mathcal{A}_S$, observe r_{t,x_t} and update the statistical information;
    $N_0(t) \leftarrow N_0(t-1);$
    $r_S(t) \leftarrow r_S(t-1) + r_{t,x_t};$
  else
    Play x_0 and receive reward μ_0;
    $N_0(t) \leftarrow N_0(t-1) + 1;$
    $r_S(t) \leftarrow r_S(t-1);$

in the worst case, since playing the default arm cannot provide any information for identifying the optimal arm. At each time, GenCB checks if playing a regular arm can satisfy the sample-path reward constraint in the worst case (this pull feedbacks zero reward). If it can, we play a regular arm x_t according to $\mathcal{A}_S$, observe reward r_{t,x_t} and update the statistical information. Otherwise, we choose the default arm.

Different from previous conservative algorithms (Wu et al. 2016; Kazerouni et al. 2017; Zhang, Li, and Liu 2019), GenCB guarantees the constraint with certainty rather than with high probability, and GenCB uses the received cumulative reward rather than the lower confidence bound to check the constraint. Doing so makes our algorithm less conservative and boosts its empirical performance significantly (see Section 5 for empirical comparisons).

Next, we present the regret analysis for GenCB. Note that, the regret for CBPs can be decomposed into (i) the regret incurred by regular arms, and (ii) the regret due to playing the default arm, i.e., conservative regret. Since the analysis of the former is similar to that in SBPs, as in the conservative bandit literature (Wu et al. 2016; Kazerouni et al. 2017; Garcelon et al. 2020; Zhang, Li, and Liu 2019), we mainly focus the conservative regret. We remark that our analysis is different from those in prior works, and can be applied to several specific CBPs including CMAB, CLB and CCCB. We give the regret bound of GenCB as follows.

Theorem 1. *Given a standard bandit problem and a corresponding algorithm $\mathcal{A}_S$ with regret $\mathbb{E}[\mathcal{R}_T(\mathcal{A}_S)] \leq B(T)$, GenCB (Algorithm 1) guarantees the sample-path reward constraint Eq.* (2) *and achieves a regret bound*

$$\mathbb{E}[\mathcal{R}_T(\textit{GenCB})] \leq B(T) + C\Delta_0,$$

where C is a problem-specific constant independent of T and $\Delta_0 = \mu_{x_*} - \mu_{x_0}$.

Proof. First, it can be seen from the algorithm that the sample-path reward constraint Eq. (2) can be guaranteed. Next, we prove the regret bound of GenCB. We use $\mathcal{S}_t$ to denote the set of timesteps up to time t during which we play regular arms and use m_t to denote its size. Let τ denote the last timestep we play x_0, i.e., τ is the last timestep such that $r_S(\tau-1) + N_0(\tau-1)\mu_0 < (1-\alpha)\mu_0\tau$ holds. Rearranging the terms, and subtracting $(1-\alpha)\mu_0 N_0(\tau-1)$ from both sides (note that $\tau = N_0(\tau-1) + m_{\tau-1} + 1$), we have

$$\begin{aligned}\alpha\mu_0 N_0(\tau-1) <& (1-\alpha)\mu_0(m_{\tau-1}+1) - r_S(\tau-1)\\ =& (1-\alpha)\mu_0(m_{\tau-1}+1)\\ &+ \sum_{t\in\mathcal{S}_{\tau-1}}(\mu_{x_t} - r_{t,x_t}) - \sum_{t\in\mathcal{S}_{\tau-1}}\mu_{x_t}. \quad (3)\end{aligned}$$

$\sum_{t\in\mathcal{S}_{\tau-1}}(\mu_{x_t} - r_{t,x_t})$ is the deviation between the sum of $m_{\tau-1}$ samples and their means. Using the Azuma-Hoeffding inequality, it can be upper bounded by $F\sqrt{m_{\tau-1}\ln(1/\delta)}$ with probability at least $1-\delta$, for some constant F that varies in different settings. By setting $\delta = 1/\tau$, we can obtain an expected upper bound as $F\sqrt{m_{\tau-1}\ln\tau} + 1$. Taking expectation on both sides of Eq. (3), setting $m = \mathbb{E}[m_{\tau-1}+1]$ and replacing $\mathbb{E}[\sum_{t\in\mathcal{S}_{\tau-1}}\mu_{x_t}]$ with $\mu_{x_*}\mathbb{E}[m_{\tau-1}] - \mathbb{E}[\mathcal{R}_{m_{\tau-1}}(\mathcal{A}_S)]$, we have

$$\begin{aligned}\alpha\mu_0\mathbb{E}[N_0(\tau-1)] <& -(\Delta_0 + \alpha\mu_0)m + \mathbb{E}[B(m_{\tau-1})]\\ &+ \mathbb{E}[F\sqrt{m_{\tau-1}\ln(\tau)}] + 1\\ \overset{(a)}{<}& -(\Delta_0 + \alpha\mu_0)m + B(m) + 2\\ &+ F\sqrt{m\ln(\mathbb{E}[N_0(\tau-1)] + m)},\end{aligned}$$

where (a) comes from Jensen's inequality. Note that since $B(m)$ and $F\sqrt{m\ln(\mathbb{E}[N_0(\tau-1)] + m)}$ are sublinear with respect to m, for any $m \geq 2$, the right-hand-side can be upper bounded by $G[\ln(\sqrt{\mathbb{E}[N_0(\tau-1)]})]^2$ where G is a constant factor that only depends on problem parameters. Then, we obtain $\mathbb{E}[N_0(\tau-1)] \leq \frac{G}{\alpha\mu_0}[\ln(\frac{G}{\alpha\mu_0})]^2$. Thus, $\mathbb{E}[N_0(T)] = \mathbb{E}[N_0(\tau)] = \mathbb{E}[N_0(\tau-1)] + 1 \leq C$, where $C \triangleq \frac{G}{\alpha\mu_0}[\ln(\frac{G}{\alpha\mu_0})]^2 + 1$ is independent of T.

Combining the regrets for $\mathcal{A}_S$ and x_0, we obtain that $\mathbb{E}[\mathcal{R}_T(\textit{GenCB})] \leq B(T) + C\Delta_0$. □

Remark 1. Theorem 1 shows that GenCB provides a general algorithmic and analytical framework for translating a standard bandit problem into a conservative bandit algorithm, and only generate an additional T-independent regret due to the reward constraint. To the best of our knowledge, this is the first general analysis procedure which works for a family of CBPs with sample-path reward constraints, and it provides an expected regret bound (rather than high probability bounds in (Wu et al. 2016; Kazerouni et al. 2017; Garcelon et al. 2020; Zhang, Li, and Liu 2019)) with T-independent conservative regret.

Below, we apply GenCB to three widely studied CBPs, i.e., CMAB, CLB and CCCB. Due to space limit, we only present the main theorems, and defer the algorithm pseudocodes and proofs to the supplementary materials.

3.1 Application to Conservative Multi-Armed Bandits (CMAB)

The conservative multi-armed bandit (CMAB) problem is a variation of the classic K-armed bandit model with con-

servative exploration (Wu et al. 2016), which has extensive applications including clinical trials, online advertising and wireless network. In CMAB, $\mathcal{X} = [K]$ and μ_i $(1 \le i \le K)$ can be an arbitrary value. Without lose of generality, we assume $\mu_1 \ge \mu_2 \ge \cdots \ge \mu_K$ and denote $\mu_* \triangleq \mu_1$.

We apply the GenCB algorithm with the UCB algorithm (Auer, Cesa-Bianchi, and Fischer 2002) to this setting, by replacing Line 5 in Algorithm 1 with $x_t \leftarrow \operatorname{argmax}_{i\in[K]} \left(\hat{\mu}_i + \sqrt{2\ln m/N_i(t-1)}\right)$, where $\hat{\mu}_i$ is the reward empirical mean for arm i, and name this version of the algorithm GenCB-CMAB.

The main idea of GenCB-CMAB is to play the arm with the maximum upper confidence bound whenever the reward constraint is satisfied (otherwise we play the default arm). The regret bound for GenCB-CMAB is summarized below.

Theorem 2. *For the conservative multi-armed bandit problem, GenCB-CMAB guarantees the sample-path reward constraint Eq.* (2) *and achieves the regret bound*

$$O\left(H\ln T + \frac{H\Delta_0}{\alpha\mu_0(\Delta_0+\alpha\mu_0)}\left[\ln\left(\frac{H}{\alpha\mu_0(\Delta_0+\alpha\mu_0)}\right)\right]^2\right),$$

where $H = \sum_{i>1}\Delta_i^{-1}$.

Remark 2. The first term owes to playing the regular arms, which is similar to the result in standard MAB (Auer, Cesa-Bianchi, and Fischer 2002), and the second term is caused by the default arm, i.e., the conservative regret, which is the main focus in conservative bandit study. Compared to the existing algorithm ConUCB (Wu et al. 2016), GenCB-CMAB only incurs a T-independent conservative regret rather than $\ln T$ (see Table 1). Our result also matches the regret lower bound derived in (Wu et al. 2016) for CMAB with expected reward constraints, which also holds for our sample-path reward constraint setting.

3.2 Application to Conservative Linear Bandits (CLB)

The conservative linear bandit (CLB) (Kazerouni et al. 2017; Garcelon et al. 2020) problem considers the linear bandit problem (Dani, Hayes, and Kakade 2008; Abbasi-yadkori, Pál, and Szepesvári 2011) with safe exploration. In CLB where there is a linear structure among arms, $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$ and $\mu_x = x^\top\theta^*$, where $\theta^* \in \mathbb{R}^d$ is an unknown parameter. We make the common assumptions, i.e., $\|x\|_2 \le L, \forall x \in \mathcal{X}$ and $\|\theta^*\|_2 \le S$, as previous linear bandit papers (Dani, Hayes, and Kakade 2008; Abbasi-yadkori, Pál, and Szepesvári 2011; Kazerouni et al. 2017) do.

For CLB, we apply GenCB with the LinUCB algorithm (Abbasi-yadkori, Pál, and Szepesvári 2011) by replacing Line 5 in Algorithm 1 with $(x_t, \tilde{\theta}_t) \leftarrow \operatorname{argmax}_{(x,\theta)\in\mathcal{X}\times\mathcal{C}_t} x^\top\theta$. Here $\mathcal{C}_t = \{\theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_{t-1}\|_{V_{t-1}} \le \sqrt{d\ln(2m^2(1+mL^2/\lambda))} + \sqrt{\lambda}S\}$ is a confidence ellipsoid that contains θ^* with high probability, and we define $\hat{\theta}_t = V_t^{-1}b_t$, $V_t = \lambda I + \sum_{s=1}^t x_s x_s^\top$, $b_t = \sum_{s=1}^t r_{s,x_s}x_s$ and $\lambda \ge \max\{1, L^2\}$.[3] We name this version of the algorithm GenCB-CLB, whose key idea is to play a regular arm according to the optimism in the face of uncertainty principle while ensuring the sample-path reward constraint. Below, we have the regret bound of GenCB-CLB.

Theorem 3. *For the conservative linear bandit problem, GenCB-CLB guarantees the sample-path reward constraint Eq.* (2) *and has the regret bound*

$$O\left(d\ln\left(\frac{LT}{\lambda}\right)\sqrt{T} + \frac{d^2S^2\lambda\Delta_0}{\alpha\mu_0\tilde{\Delta}_0}\left[\ln\left(\frac{dS\sqrt{\lambda}}{\alpha\mu_0\tilde{\Delta}_0}\right)\right]^2\right),$$

where $\tilde{\Delta}_0 = \Delta_0 + \alpha\mu_0$.

Remark 3. Similarly, the first term is aligned with the result in standard linear bandits (Dani, Hayes, and Kakade 2008; Abbasi-yadkori, Pál, and Szepesvári 2011), and the second term is the conservative regret due to the default arm. While the existing algorithms CLUCB (Kazerouni et al. 2017) and CLUCB2 (Garcelon et al. 2020) have $\ln(1/\delta)$-dependent conservative regrets with high probability (do not contain T either), these results are of $\ln T$ order when making the convention $\delta = 1/T$. In contrast, we provide an expected bound with a T-independent conservative regret.

3.3 Application to Conservative Contextual Combinatorial Bandits (CCCB)

The conservative contextual combinatorial bandit (CCCB) problem (Zhang, Li, and Liu 2019) investigates the contextual combinatorial bandit problem under the safe exploration requirement. In CCCB, $\mathcal{X}$ is a collection of subsets of *base arms* $x_1, \ldots, x_K \in \mathbb{R}^d$ and generated from certain combinatorial structure (e.g., matchings and paths). The learner plays a *super arm* (subset of base arms) $A_t \in \mathcal{X}$ or the default arm x_0 at each timestep. The expected reward of base arm x_e is $w_e^* = x_e^\top\theta^*$ and that of super arm A is $f(A, \boldsymbol{w}^*)$, where θ^* is an unknown parameter and f satisfies two mild assumptions, i.e., monotonicity and Lipschitz continuous with parameter P (Qin, Chen, and Zhu 2014; Zhang, Li, and Liu 2019). Similar to CLB, we assume $\|x\|_2 \le L, \forall x \in \mathcal{X}$ and $\|\theta^*\|_2 \le S$. At timestep t, the random reward of a base arm x_e and a super arm A are $w_{t,e} = w_e^* + \eta_{t,e} \in [0,1]$ and $r_{t,A} = f(A, \boldsymbol{w}^*) + \eta_{t,A} \in [0, K]$, respectively. After pulling super arm A_t, we receive the random reward r_{t,A_t} and observe a semi-bandit feedback, i.e., $w_{t,e}$ for each $e \in A_t$.

For CCCB, we apply GenCB with the C2UCB algorithm (Qin, Chen, and Zhu 2014), by replacing Line 5 in Algorithm 1 with $A_t \leftarrow \operatorname{argmax}_{A\in\mathcal{X}} f(A, \bar{\boldsymbol{w}}_t)$. Here $\bar{w}_{t,e} = x_e^\top\hat{\theta}_{t-1} + (\sqrt{d\ln(2m^2(1+mKL^2/\lambda))} + \sqrt{\lambda}S)\|x_e\|_{V_{t-1}^{-1}}$ is the upper confidence bound of w_e^*, and we define $\hat{\theta}_t = V_t^{-1}b_t$, $V_t = \lambda I + \sum_{s=1}^t\sum_{e\in A_s} x_e x_e^\top$, $b_t = \sum_{s=1}^t\sum_{e\in A_s} w_{s,e}x_e$ and $\lambda \ge \max\{1, L^2\}$. The key idea here is to play a super arm with the maximum upper confidence bound according to the historical observations on base arms. Theorem 4 below gives the regret bound of GenCB-CCCB.

Theorem 4. *For the contextual combinatorial bandit problem, GenCB-CCCB ensures the sample-path reward con-*

[3] $\|x\|_V \stackrel{\text{def}}{=} \sqrt{x^\top Vx}, \forall x \in \mathbb{R}^d, \forall V \in \mathbb{R}^{d\times d}$.

Algorithm 2: MV-CUCB

Input: Reugular arms $[K]$, default arm x_0 with $\mathrm{MV}_0 = \rho\mu_0$, parameters α, $\rho > \frac{2}{\alpha\mu_0}$.

$\forall t \geq 0, \forall 0 \leq i \leq K, N_i(t) \leftarrow 0.\ m \leftarrow 0.$
$\widehat{\mathrm{MV}}_0(\mathcal{A}) \leftarrow 0;$
for $t = 1, 2, \ldots$ **do**
  if $(t-1)\widehat{\mathrm{MV}}_{t-1}(\mathcal{A}) - 2 \geq (1-\alpha)\mathrm{MV}_0 t$ **then**
    $m \leftarrow m + 1;$
    $x_t \leftarrow \operatorname{argmax}_{i\in[K]} \left(\widehat{\mathrm{MV}}_i + (5+\rho)\sqrt{\frac{\ln(12Km^3)}{2N_i(t-1)}}\right);$
    Pull arm x_t, observe the random reward r_{t,x_t} and update $\widehat{\mathrm{MV}}_{x_t}$;
    $N_{x_t}(t) \leftarrow N_{x_t}(t-1) + 1$ and $\forall 0 \leq i \leq K, i \neq x_t, N_i(t) \leftarrow N_i(t-1);$
  else
    Play x_0 and receive reward μ_0;
    $N_0(t) \leftarrow N_0(t-1) + 1$ and $\forall 1 \leq i \leq K, N_i(t) \leftarrow N_i(t-1);$

straint Eq. (2) *and achieves the regret bound*

$$O\left(Pd\ln\left(\frac{KLT}{\lambda}\right)\sqrt{T} + \frac{D^2}{\alpha\mu_0\tilde{\Delta}_0}\left[\ln\left(\frac{D}{\alpha\mu_0\tilde{\Delta}_0}\right)\right]^2\right),$$

where $D = K + P\sqrt{\lambda}Sd$ *and* $\tilde{\Delta}_0 = \Delta_0 + \alpha\mu_0$.

Remark 4. The first term is consistent with the result in standard contextual combinatorial bandits (Qin, Chen, and Zhu 2014), and the second conservative regret term is due to playing the default arm. Compared to the state-of-the-art algorithm C3UCB (Zhang, Li, and Liu 2019), GenCB-CCCB provides a T-independent conservative regret, while C3UCB incurs a $\ln T$ regret (see Table 1).

4 Conservative Mean-Variance Bandits

We now extend CBPs to the mean-variance (Sani, Lazaric, and Munos 2012; Maillard 2013; Cardoso and Xu 2019) setting (MV-CBP), which focuses on finding arms that achieve effective trade-off between the expected reward and variability. MV-CBP increments the typical conservative bandit model and better suits the tasks emphasizing on reward fluctuations. It also brings additional complications for algorithm design and regret analysis beyond GenCB.

4.1 Problem Formulation for MV-CBP

To introduce our MV-CBP formulation, we first review the standard mean-variance bandit setting (Sani, Lazaric, and Munos 2012). Each arm $x \in [K]$ is associated with a measure mean-variance, which is formally defined as $\mathrm{MV}_x = \rho\mu_x - \sigma_x^2$, where σ_x^2 is the reward variance and ρ is a weight parameter. Let $x_*^{\mathrm{MV}} = \operatorname{argmax}_{x\in[K]} \mathrm{MV}_x$ denote the mean-variance optimal arm. Given i.i.d. reward samples $\{Z_{x,s}\}_{s=1}^t$ of arm x, we define the *empirical mean-variance* $\widehat{\mathrm{MV}}_{x,t} = \rho\hat{\mu}_{x,t} - \hat{\sigma}_{x,t}^2$, where $\hat{\mu}_{x,t} = \frac{1}{t}\sum_{s=1}^t Z_{x,s}$ and $\hat{\sigma}_{x,t}^2 = \frac{1}{t}\sum_{s=1}^t (Z_{x,s} - \hat{\mu}_{x,t})^2$.

For an algorithm $\mathcal{A}$ and its sample path $\{r_{t,x_t}\}_{t=1}^T$ over time horizon T, we define the empirical mean-variance $\widehat{\mathrm{MV}}_T(\mathcal{A}) = \rho\hat{\mu}_T(\mathcal{A}) - \hat{\sigma}_T^2(\mathcal{A})$, where $\hat{\mu}_T(\mathcal{A}) = \frac{1}{T}\sum_{t=1}^T r_{t,x_t}$ and $\hat{\sigma}_T^2(\mathcal{A}) = \frac{1}{T}\sum_{t=1}^T (r_{t,x_t} - \hat{\mu}_T(\mathcal{A}))^2$. Naturally, for algorithm $\mathcal{A}$ over time T, we define the *mean-variance regret* $\mathcal{R}_T^{\mathrm{MV}}(\mathcal{A}) = \widehat{\mathrm{MV}}_{x_*,T} - \widehat{\mathrm{MV}}_T(\mathcal{A})$, which is the difference of the mean-variance performance between $\mathcal{A}$ and what we could have achieved by always playing x_*^{MV}.

Due to the difficulty of the $\mathcal{R}_T^{\mathrm{MV}}(\mathcal{A})$ metric, we follow the mean-variance bandit literature and use a more tractable mesure *mean-variance pseudo-regret* (Sani, Lazaric, and Munos 2012) defined as:

$$\widetilde{\mathcal{R}}_T^{\mathrm{MV}}(\mathcal{A}) = \frac{1}{T}\sum_{x\neq x_*} N_{x,T}\Delta_x^{\mathrm{MV}} + \frac{2}{T^2}\sum_{x\in\mathcal{X}}\sum_{y\neq x} N_{x,T}N_{y,T}\Gamma_{x,y}^2, \quad (4)$$

where $N_{x,T}$ is a shorthand for $N_x(T)$, $\Delta_x^{\mathrm{MV}} = \widehat{\mathrm{MV}}_{x_*} - \widehat{\mathrm{MV}}_x$ and $\Gamma_{x,y} = \mu_x - \mu_y$. It has been shown that any bound on $\widetilde{\mathcal{R}}_T^{\mathrm{MV}}(\mathcal{A})$ immediately translates into an bound on $\mathcal{R}_T^{\mathrm{MV}}(\mathcal{A})$ (Lemma 1 in (Sani, Lazaric, and Munos 2012)). Thus, most theoretical analysis (Sani, Lazaric, and Munos 2012; Maillard 2013; Cardoso and Xu 2019) on mean-variance bandits has been done via $\widetilde{\mathcal{R}}_T^{\mathrm{MV}}(\mathcal{A})$. Note that, in MV-CBP the measures $\widehat{\mathrm{MV}}_T(\mathcal{A})$ and $\mathcal{R}_T^{\mathrm{MV}}(\mathcal{A})$ are both *normalized* quantities over T.

In addition to minimizing the regret, the learner is also required to guarantee the following mean-variance constraint:

$$\widehat{\mathrm{MV}}_t(\mathcal{A}) \geq (1-\alpha)\mathrm{MV}_0, \quad \forall t \in \{1, \ldots, T\}. \quad (5)$$

Here MV_0 denotes the mean-variance of our default arm x_0 with known constant reward μ_0 and zero variance. The goal in MV-CBP is to minimize Eq. (4) while satisfying Eq. (5).

4.2 Algorithm for MV-CBP

We propose a novel algorithm named MV-CUCB for MV-CBP (illustrated in Algorithm 2). The main idea is to compute the upper confidence bound of mean-variance for each arm and select one according to the optimism principle whenever the constraint is not violated. Theorem 5 summarizes the performance results of MV-CUCB. Its proof is deferred to the supplementary materials due to space limit.

Theorem 5. *For the conservative mean-variance multi-armed bandit problem with* $\alpha\mathrm{MV}_0 > 2$*, MV-CUCB (Algorithm 2) ensures the mean-variance constraint Eq.* (5) *and achieves the following regret bound:*[4]

$$\tilde{O}\left(\frac{\rho^2\ln(KT)}{T}\left(H_1 + H_2 + \frac{\rho^2\ln(KT)}{T}H_3\right) + \frac{\rho^3\sqrt{K}(H_1^{\mathrm{MV}} + 4H_2^{\mathrm{MV}}) + (\rho^4 KH_3^{\mathrm{MV}} + \rho K)\tilde{\Delta}_0^{\mathrm{MV}}}{(\alpha\mathrm{MV}_0 - 2)\tilde{\Delta}_0^{\mathrm{MV}}T}\cdot\Delta_0^{\mathrm{MV}}\right),$$

where $H_1^{\mathrm{MV}} = \sum_{i>1}(\Delta_i^{\mathrm{MV}})^{-1}$, $H_2^{\mathrm{MV}} = \sum_{i>1}(\Delta_i^{\mathrm{MV}})^{-2}$, $H_3^{\mathrm{MV}} = \sum_{i>1}\sum_{j>1,j\neq i}(\Delta_i^{\mathrm{MV}}\Delta_j^{\mathrm{MV}})^{-2}$ *and* $\tilde{\Delta}_0^{\mathrm{MV}} = \Delta_0^{\mathrm{MV}} + \alpha\mathrm{MV}_0$.

[4] $\tilde{O}$ omits the logarithmic terms that are independent of T.

Remark 5. Since a pull of x_0 not only accumulates MV_0 but also causes an exploration risk (bounded by 2 for reward distributions in $[0,1]$) due to the switch between different arms, we need the mild assumption $\alpha\mathrm{MV}_0 > 2$ to guarantee that a pull of x_0 can earn a positive budget against the baseline $(1-\alpha)\mathrm{MV}_0$. Recall that the result is a normalized regret over T, the first term of $O(\ln T/T)$ order owes to regular arms, which agrees with the previous non-conservative mean-variance result (Sani, Lazaric, and Munos 2012). The second term is the conservative regret for satisfying the constraint, which is of only $O(1/T)$ order and independent of T in the cumulative form. To our best knowledge, Theorem 5 is the first result for conservative bandits with mean-variance objectives.

5 Experiments

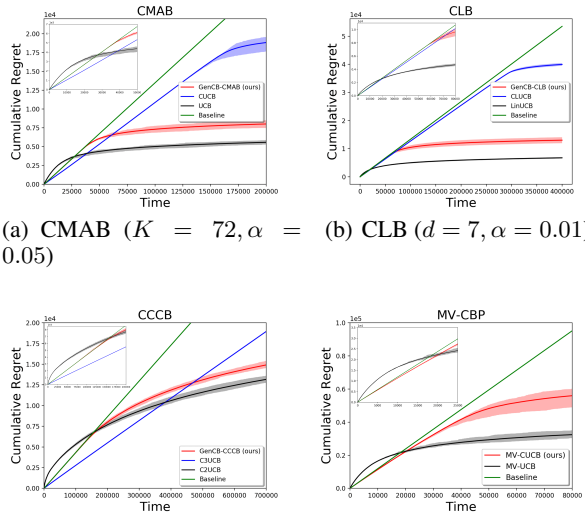
We conduct experiments for our algorithms in four problems, i.e., CMAB, CLB, CCCB and MV-CBP, with a wide range of parameter settings. Due to space limit, we only present part of the experimental results here. Please see full results in the supplementary materials.

In all experiments, we assume the rewards to take i.i.d. Bernoulli values. For CMAB, we set $K \in \{24, 72, 144\}$, $\alpha \in \{0.05, 0.1, 0.15\}$, $\mu_0 = 0.7$ and $\mu_1, \ldots, \mu_K$ as an arithmetic sequence from 0.8 to 0.2. For CLB and CCCB, we set $d \in \{5, 7, 9\}$, $\alpha \in \{0.01, 0.02, 0.03\}$, $K = 2d$ and $f(A, \boldsymbol{w}^*) = \sum_{e \in A} w_e^*$. For MV-CBP, we use the same parameter settings as CMAB and additionally set $\rho \in \{10, 30, 60\}$. For each algorithm, we perform 50 independent runs and present the average (middle curve), maximum (upper curve) and minimum (bottom curve) cumulative regrets across runs. For each figure, we also zoom in the initial exploratory phase in the upper left sub-figure to compare algorithm performance in this phase.

Experiments for CBPs. In the experiments for CMAB (Figure 1(a)), CLB (Figure 1(b)) and CCCB (Figure 1(c)), we compare GenCB-CMAB, GenCB-CLB and GenCB-CCCB to previous CBP algorithms CUCB (Wu et al. 2016), CLUCB (Kazerouni et al. 2017) and C3UCB (Zhang, Li, and Liu 2019), the standard bandit algorithms UCB (Auer, Cesa-Bianchi, and Fischer 2002), LinUCB (Abbasi-yadkori, Pál, and Szepesvári 2011) and C2UCB (Qin, Chen, and Zhu 2014), and the conservative baseline $(1-\alpha)\mu_0$, respectively.

We see that, in the exploration phase, existing non-conservative algorithms suffer higher losses than the baseline, while our algorithms and previous CBP algorithms achieve similar performance as (or better than) the baseline due to the conservative constraints. However, since previous CBP algorithms use lower confidence bounds (rather than the empirical rewards in ours) to check the constraints, they are forced to play the default arm more and act more conservatively compared to ours.

In the exploitation phase, when compared to non-conservative algorithms, our algorithms have additional regrets that keep constant as T increases, which matches our T-independent conservative regret bounds. Compared to previous CBP algorithms, our schemes achieve significantly better performance, since we play the default arm less and enjoy a lower conservative regret.



(a) CMAB ($K = 72, \alpha = 0.05$) (b) CLB ($d = 7, \alpha = 0.01$)

(c) CCCB ($d = 7, \alpha = 0.01$) (d) MV-CBP ($K = 24, \alpha = 0.05, \rho = 10$)

Figure 1: Experiments for the studied problems, i.e., CMAB, CLB, CCCB and MV-CBP.

Experiments for MV-CBP. In the experiments for MV-CBP (Figure 1(d)), we present the mean-variance regret in the cumulative form $T \cdot \widetilde{\mathcal{R}}_T^{\mathrm{MV}}$ for clarity of comparison. Since MV-CUCB is the first algorithm for MV-CBP, we compare it with the standard mean-variance bandit algorithm MV-UCB and the baseline $(1-\alpha)\mathrm{MV}_0 T$. We can see that, in the exploration phase, MV-UCB suffers from a higher regret than the baseline while MV-CUCB follows the baseline closely. One also sees that MV-CUCB achieves this with only an additional constant overall regret compared to MV-UCB, which matches our T-independent bound of conservative regret.

6 Conclusion and Future Works

In this paper, we propose a general solution to a family of conservative bandit problems (CBPs) with sample-path reward constraints, and present its applications to three encompassed problems, i.e., conservative multi-armed bandits (CMAB), conservative linear bandits (CLB) and conservative contextual combinatorial bandits (CCCB). We show that our algorithms outperform existing schemes both theoretically (incurs a T-independent conservative regret rather than a T-dependent) and empirically. Moreover, we study a novel extension of the conservative bandit problem to the mean-variance setting (MV-CBP) and develop an algorithm with $O(1/T)$ normalized conservative regret (T-independent in the cumulative form). We also validate this result through empirical evaluation.

There are several directions worth further investigation. One is to consider more general conservative mean-variance bandits other than the K-armed setting, e.g., a contextual extension. Another direction is to consider other practical conservative constraints which capture the safe exploration requirement in real-world applications.

7 Ethical Impact

In this paper, we study a family of conservative bandit problems and present algorithms with theoretical guarantees and experimental results. While our work mainly focuses on the theoretical analysis, it may have potential social impacts on the applications including finance and clinical trials. For example, our algorithms may help risk-adverse investors choose financial products, with the objective of obtaining high cumulative returns while guaranteeing a certain baseline during exploration. We believe that this work does not involve any ethical issue.

Acknowledgement

This work is supported in part by the National Natural Science Foundation of China Grant 61672316, the Zhongguancun Haihua Institute for Frontier Information Technology and the Turing AI Institute of Nanjing.

References

Abbasi-yadkori, Y.; Pál, D.; and Szepesvári, C. 2011. Improved Algorithms for Linear Stochastic Bandits. In *Advances in Neural Information Processing Systems*, 2312–2320.

Agrawal, S.; and Goyal, N. 2012. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, 39–1.

Amani, S.; Alizadeh, M.; and Thrampoulidis, C. 2019. Linear Stochastic Bandits Under Safety Constraints. In *Advances in Neural Information Processing Systems*, 9256–9266.

Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine Learning* 47(2-3): 235–256.

Bubeck, S.; Perchet, V.; and Rigollet, P. 2013. Bounded regret in stochastic multi-armed bandits. In *Conference on Learning Theory*, 122–134.

Cardoso, A. R.; and Xu, H. 2019. Risk-averse stochastic convex bandit. In *International Conference on Artificial Intelligence and Statistics*, 39–47.

Dani, V.; Hayes, T.; and Kakade, S. M. 2008. Stochastic Linear Optimization under Bandit Feedback. In *Conference on Learning Theory*.

Garcelon, E.; Ghavamzadeh, M.; Lazaric, A.; and Pirotta, M. 2020. Improved Algorithms for Conservative Exploration in Bandits. In *AAAI Conference on Artificial Intelligence*.

Katariya, S.; Kveton, B.; Wen, Z.; and Potluru, V. 2019. Conservative Exploration using Interleaving. In *International Conference on Artificial Intelligence and Statistics*.

Kazerouni, A.; Ghavamzadeh, M.; Yadkori, Y. A.; and Van Roy, B. 2017. Conservative contextual linear bandits. In *Advances in Neural Information Processing Systems*, 3910–3919.

Khezeli, K.; and Bitar, E. 2020. Safe Linear Stochastic Bandits. In *AAAI Conference on Artificial Intelligence*.

Locatelli, A.; Gutzeit, M.; and Carpentier, A. 2016. An optimal algorithm for the Thresholding Bandit Problem. In *International Conference on Machine Learning*, 1690–1698.

Maillard, O.-A. 2013. Robust risk-averse stochastic multi-armed bandits. In *International Conference on Algorithmic Learning Theory*, 218–233. Springer.

Markowitz, H. M.; et al. 1952. Portfolio Selection. *Journal of Finance* 7(1): 77–91.

Qin, L.; Chen, S.; and Zhu, X. 2014. Contextual Combinatorial Bandit and its Application on Diversified Online Recommendation. In *International Conference on Data Mining*, 461–469.

Sani, A.; Lazaric, A.; and Munos, R. 2012. Risk-aversion in multi-armed bandits. In *Advances in Neural Information Processing Systems*, 3275–3283.

Thompson, W. R. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25(3/4): 285–294.

Vakili, S.; Boukouvalas, A.; and Zhao, Q. 2019. Decision variance in risk-averse online learning. In *Conference on Decision and Control*, 2738–2744. IEEE.

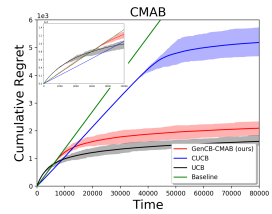
Wu, Y.; Shariff, R.; Lattimore, T.; and Szepesvári, C. 2016. Conservative bandits. In *International Conference on Machine Learning*, 1254–1262.

Zhang, X.; Li, S.; and Liu, W. 2019. Contextual Combinatorial Conservative Bandits. *arXiv preprint:1911.11337* .

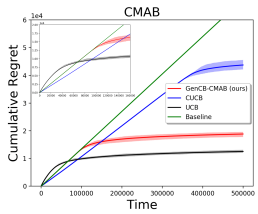
Supplementary Material

A More Experimental Results

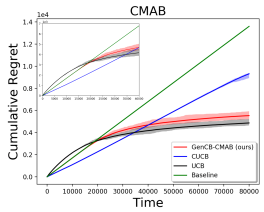
In this section, we present more experimental results for CMAB, CLB, CCCB and MV-CPB, which are shown in Figure 2, 3, 4 and 5, respectively. The parameter settings of our experiments are described in Section 5 of the main paper.



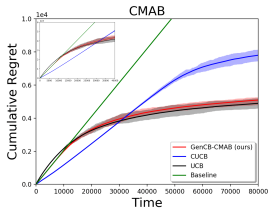
(a) $K = 24, \alpha = 0.05$



(b) $K = 144, \alpha = 0.05$

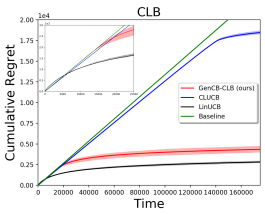


(c) $K = 72, \alpha = 0.1$

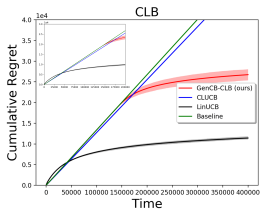


(d) $K = 72, \alpha = 0.15$

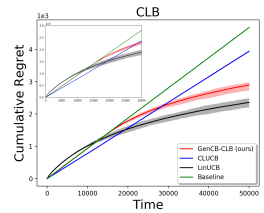
Figure 2: Experiments for CMAB.



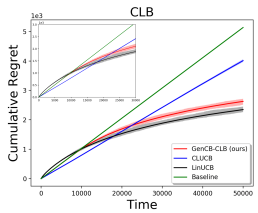
(a) $d = 5, \alpha = 0.01$



(b) $d = 9, \alpha = 0.01$



(c) $d = 5, \alpha = 0.02$



(d) $d = 5, \alpha = 0.03$

Figure 3: Experiments for CLB.

B Technical Tools

We present some technical tools (Lemmas 1,2 and Facts 1,2) below.

Lemma 1. *For $m \geq 2$, $c_1 = 2$, $c_2 = 2\mathbb{E}[N_0(\tau - 1)] \geq 4$, $c_3 = \Delta_0 + \alpha\mu_0 \in (0, 1)$, $c_4 = 8H$ where $H \geq 2$, define function $g_1(m) = -c_3 m + c_1\sqrt{m}\ln(c_2 m) + c_4 \ln(m)$. Then, $g_1(m)$ can be upper bounded by*

$$g_1(m) \leq \frac{132 c_1 c_4}{c_3}\left[\ln\left(\frac{10\sqrt{c_2 c_4}}{c_3}\right)\right]^2.$$

Proof. Taking the derivative of $g_1(m)$, we obtain

$$g_1'(m) = -c_3 + \frac{c_1 \ln(c_2 m) + 2c_1}{2\sqrt{m}} + \frac{c_4}{m}$$

Let $\tilde{m}_1 = \frac{c_4}{c_3}, \tilde{m}_2 = \frac{100 c_4 [\ln(c_2 c_4/c_3^2)]^2}{c_3^2}$. Then, we have

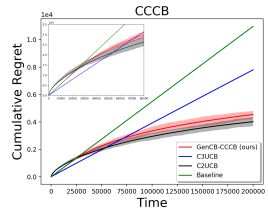
$$g_1'(\tilde{m}_1) = \frac{c_1 \ln(c_2 \tilde{m}_1) + 2c_1}{2\sqrt{\tilde{m}_1}} > 0$$

and

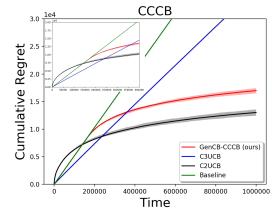
$$g_1'(\tilde{m}_2) = -c_3 + \frac{c_1 \ln(\frac{100 c_2 c_4 [\ln(c_2 c_4/c_3^2)]^2}{c_3^2}) + 2c_1}{10\sqrt{c_4}\ln(c_2 c_4/c_3^2)} \cdot \frac{c_3}{2} + \frac{c_3}{100[\ln(c_2 c_4/c_3^2)]^2} \cdot c_3.$$

Since

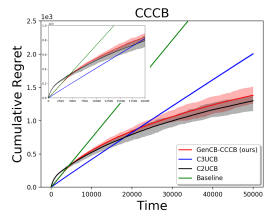
$$\begin{aligned}
& c_1 \ln(\frac{100 c_2 c_4 [\ln(c_2 c_4/c_3^2)]^2}{c_3^2}) + 2c_1 \\
=& c_1 \ln(\frac{100 c_2 c_4}{c_3^2}) + 2c_1 \ln(\ln(\frac{c_2 c_4}{c_3^2})) + 2c_1 \\
\leq& c_1 \ln(\frac{100 c_2 c_4}{c_3^2}) + 2c_1 \ln(\frac{c_2 c_4}{c_3^2}) + 2c_1 \\
\leq& 3c_1 \ln(\frac{100 c_2 c_4}{c_3^2}) + 2c_1 \\
=& 6\ln(\frac{c_2 c_4}{c_3^2}) + 6\ln(100) + 4 \\
<& 10\sqrt{c_4}\ln(\frac{c_2 c_4}{c_3^2}),
\end{aligned}$$



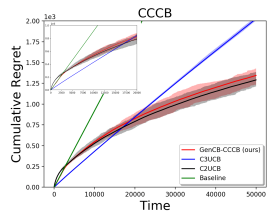
(a) $d = 5, \alpha = 0.01$



(b) $d = 9, \alpha = 0.01$

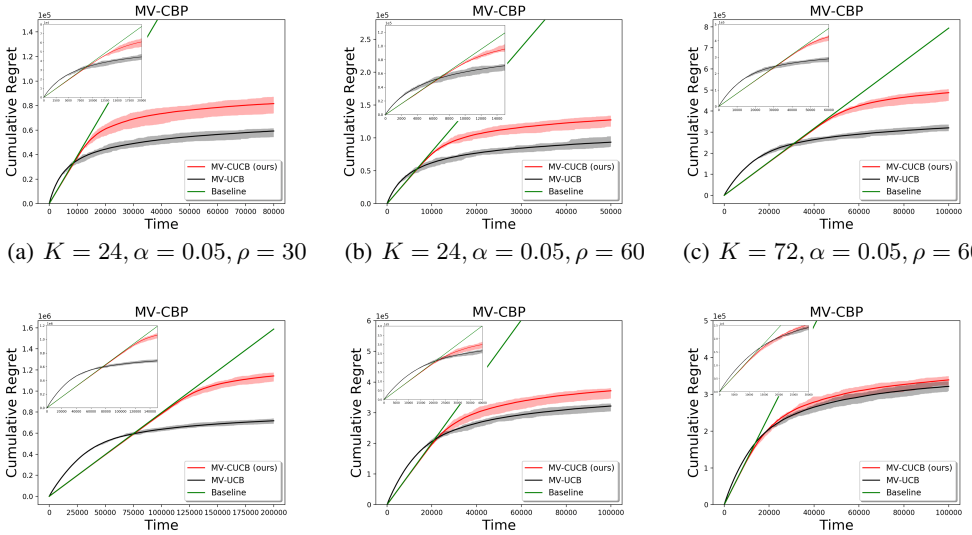


(c) $d = 5, \alpha = 0.02$



(d) $d = 5, \alpha = 0.03$

Figure 4: Experiments for CCCB.



(a) $K = 24, \alpha = 0.05, \rho = 30$ (b) $K = 24, \alpha = 0.05, \rho = 60$ (c) $K = 72, \alpha = 0.05, \rho = 60$

(d) $K = 144, \alpha = 0.05, \rho = 60$ (e) $K = 72, \alpha = 0.1, \rho = 60$ (f) $K = 72, \alpha = 0.15, \rho = 60$

Figure 5: Experiments for MV-CBP.

we have

$$\frac{c_1 \ln(\frac{100c_2c_4[\ln(c_2c_4/c_3^2)]^2}{c_3^2}) + 2c_1}{10\sqrt{c_4}\ln(c_2c_4/c_3^2)} < 1.$$

In addition, it is clear that $\frac{c_3}{100[\ln(c_2c_4/c_3^2)]^2} < \frac{1}{2}$. Thus, we have

$$g_1'(\tilde{m}_2) = -c_3 + \frac{c_1 \ln(\frac{100c_2c_4[\ln(c_2c_4/c_3^2)]^2}{c_3^2}) + 2c_1}{10\sqrt{c_4}\ln(c_2c_4/c_3^2)} \cdot \frac{c_3}{2} + \frac{c_3}{100[\ln(c_2c_4/c_3^2)]^2} \cdot c_3 < 0.$$

Thus,

$$\begin{aligned} g_1(m) \leq & -c_3\tilde{m}_1 + c_1\sqrt{\tilde{m}_2}\ln(c_2\tilde{m}_2) + c_4\ln(\tilde{m}_2) \\ \leq & (c_1\sqrt{\tilde{m}_2} + c_4)\ln(c_2\tilde{m}_2) \\ \leq & (\frac{10c_1\sqrt{c_4}}{c_3}\ln(\frac{c_2c_4}{c_3^2}) + c_4)\cdot \\ & \ln(\frac{100c_2c_4[\ln(c_2c_4/c_3^2)]^2}{c_3^2}) \\ \leq & \frac{10c_1\sqrt{c_4} + c_4}{c_3} \cdot \ln(\frac{c_2c_4}{c_3^2}) \cdot 3\ln(\frac{100c_2c_4}{c_3^2}) \\ \leq & \frac{33c_1c_4}{c_3}\left[\ln\left(\frac{100c_2c_4}{c_3^2}\right)\right]^2 \\ = & \frac{132c_1c_4}{c_3}\left[\ln\left(\frac{10\sqrt{c_2c_4}}{c_3}\right)\right]^2. \end{aligned}$$

□

Lemma 2. *For $m \geq 2$, $c_1 = 2(5+\rho)\sqrt{2K}$, $c_2 = 6K\mathbb{E}[N_0(\tau-1)]$, $c_3 = \Delta_0 + \alpha\mathrm{MV}_0 \in (2, \rho)$, $c_4 = 8\sqrt{2}K + 12(5+\rho)^2(H_1^{\mathrm{MV}} + 4H_2^{\mathrm{MV}}) > 8\sqrt{2}K + 12(5+\rho)(K-1) + 48(K-1)$ where $\rho > \frac{2}{\alpha\mu_0} > 2$, $c_4 > 3c_1$, $c_4 > 12c_3$, define function $g_2(m) = -c_3m + c_1\sqrt{m}\ln(c_2m) + c_4\ln(m)$. Then, $g_2(m)$ can be upper bounded by*

$$g_2(m) \leq \frac{48c_1c_4}{c_3}\left[\ln(\frac{3\sqrt{c_2}c_4}{c_3})\right]^2.$$

Proof. Taking the derivative of $g_2(m)$, we obtain

$$g_2'(m) = -c_3 + \frac{c_1\ln(c_2m) + 2c_1}{2\sqrt{m}} + \frac{c_4}{m}$$

Let $\tilde{m}_1 = \frac{c_4}{c_3}$, $\tilde{m}_2 = \frac{9c_4^2[\ln(c_2c_4^2/c_3^2)]^2}{c_3^2}$. Then, we have

$$g_2'(\tilde{m}_1) = \frac{c_1\ln(c_2\tilde{m}_1) + 2c_1}{2\sqrt{\tilde{m}_1}} > 0$$

and

$$g_2'(\tilde{m}_2) = -c_3 + \frac{c_1\ln(\frac{9c_2c_4^2[\ln(c_2c_4^2/c_3^2)]^2}{c_3^2}) + 2c_1}{3c_4\ln(c_2c_4^2/c_3^2)} \cdot \frac{c_3}{2} + \frac{c_3}{9c_4[\ln(c_2c_4^2/c_3^2)]^2} \cdot c_3.$$

Since

$$\begin{aligned} & c_1\ln(\frac{9c_2c_4^2[\ln(c_2c_4^2/c_3^2)]^2}{c_3^2}) + 2c_1 \\ = & c_1\ln(\frac{c_2c_4^2}{c_3^2}) + c_1\ln(9) + 2c_1\ln(\ln(\frac{c_2c_4^2}{c_3^2})) + 2c_1 \\ \leq & c_1\ln(\frac{c_2c_4^2}{c_3^2}) + c_1\ln(9) + 2c_1\ln(\frac{c_2c_4^2}{c_3^2}) + 2c_1 \\ = & 3c_1\ln(\frac{c_2c_4^2}{c_3^2}) + (2+\ln(9))c_1 \end{aligned}$$

$$<9c_1 \ln(\frac{c_2 c_4^2}{c_3^2})$$
$$<3c_4 \ln(\frac{c_2 c_4^2}{c_3^2}),$$

we have

$$\frac{c_1 \ln(\frac{9c_2c_4^2[\ln(c_2c_4^2/c_3^2)]^2}{c_3^2}) + 2c_1}{3c_4 \ln(c_2c_4^2/c_3^2)} < 1.$$

In addition, it is clear that $\frac{c_3}{9c_4[\ln(c_2c_4^2/c_3^2)]^2} < \frac{1}{2}$. Thus, we have

$$g_2'(\tilde{m}_2) = -c_3 + \frac{c_1 \ln(\frac{9c_2c_4^2[\ln(c_2c_4^2/c_3^2)]^2}{c_3^2}) + 2c_1}{3c_4 \ln(c_2c_4^2/c_3^2)} \cdot \frac{c_3}{2} + \frac{c_3}{9c_4[\ln(c_2c_4^2/c_3^2)]^2} \cdot c_3 < 0.$$

Thus,

$$\begin{aligned} g_2(m) \leq & -c_3\tilde{m}_1 + c_1\sqrt{\tilde{m}_2}\ln(c_2\tilde{m}_2) + c_4\ln(\tilde{m}_2) \\ \leq & (c_1\sqrt{\tilde{m}_2} + c_4)\ln(c_2\tilde{m}_2) \\ \leq & (\frac{3c_1c_4}{c_3}\ln(\frac{c_2c_4^2}{c_3^2}) + c_4)\ln(\frac{9c_2c_4^2[\ln(c_2c_4^2/c_3^2)]^2}{c_3^2}) \\ \leq & \frac{4c_1c_4}{c_3}\ln(\frac{c_2c_4^2}{c_3^2}) \cdot 3\ln(\frac{9c_2c_4^2}{c_3^2}) \\ \leq & \frac{12c_1c_4}{c_3}\left[\ln(\frac{9c_2c_4^2}{c_3^2})\right]^2 \\ = & \frac{48c_1c_4}{c_3}\left[\ln(\frac{3\sqrt{c_2}c_4}{c_3})\right]^2. \end{aligned}$$

□

Fact 1 (Lemma 9 in (Kazerouni et al. 2017))**.** *For any $m \geq 2$ and $c_1, c_2, c_3 > 0$, the following holds*

$$-c_3m + c_1\sqrt{m}\ln(c_2m) \leq \frac{16c_1^2}{9c_3}\left[\ln\left(\frac{2c_1\sqrt{c_2}e}{c_3}\right)\right]^2.$$

Fact 2 (Lemma 10 in (Kazerouni et al. 2017))**.** *Let c_1 and c_2 be two positive constants such that $\ln(c_1c_2) \geq 1$. Then, any $z > 0$ satisfying $z \leq c_1\ln(c_2z)$ also satisfies $z \leq 2c_1\ln(c_1c_2)$.*

C Algorithm Pseudo-code and Proof for CMAB

Algorithm 3 presents the algorithm pseudo-code of GenCB-CMAB for CMAB, and we give the detailed proof of Theorem 2 in the following.

Proof. First, we prove that GenCB-CMAB satisfies the sample-path reward constraint Eq. (2) by induction. At timestep $t = 1$, since the LHS of the if statement (in Line 3 of Algorithm 3) is zero and RHS is positive, GenCB-CMAB will pull the default arm x_0 and receive reward $\mu_0 \geq (1-\alpha)\mu_0$, which satisfies the constraint. Suppose that the sample-path reward constraint holds at timestep $t-1$. At time step t, if GenCB-CMAB plays x_0, it is clear that the constraint still holds for t. If GenCB-CMAB plays a regular arm x_t, which implies $r_S(t-1)+N_0(t-1)\mu_0 \geq (1-\alpha)\mu_0 t$, the received cumulative reward is $r_S(t-1)+N_0(t-1)\mu_0 + r_{t,x_t} \geq r_S(t-1) + N_0(t-1)\mu_0 \geq (1-\alpha)\mu_0 t$, and thus the constraint still holds for t.

Algorithm 3: GenCB-CMAB

Input: Reugular arms $[K]$, default arm x_0 with reward μ_0, parameter α.

1 $\forall t \geq 0, \forall 0 \leq i \leq K, N_i(t) \leftarrow 0. \forall t \geq 0, r_S(t) \leftarrow 0.$ $m \leftarrow 0$;
2 **for** $t = 1, 2, \ldots$ **do**
3 **if** $r_S(t-1) + N_0(t-1)\mu_0 \geq (1-\alpha)\mu_0 t$ **then**
4 $m \leftarrow m + 1$;
5 $x_t \leftarrow \text{argmax}_{i\in[K]}\left(\hat{\mu}_i + \sqrt{\frac{2\ln m}{N_i(t-1)}}\right)$;
6 Play arm x_t, observe the random reward r_{t,x_t} and update the empirical mean $\hat{\mu}_{x_t}$;
7 $N_{x_t}(t) \leftarrow N_{x_t}(t-1) + 1$ and $\forall 0 \leq i \leq K, i \neq x_t, N_i(t) \leftarrow N_i(t-1)$;
8 $r_S(t) \leftarrow r_S(t-1) + r_{t,x_t}$;
9 **else**
10 Play x_0 and receive reward μ_0;
11 $N_0(t) \leftarrow N_0(t-1) + 1$;
12 $r_S(t) \leftarrow r_S(t-1)$;

Recall that m_t denotes the number of times we played regular arms up to t, and $\mathcal{S}_t$ denotes the set of timesteps when we played regular arms up to t.

Fix a time horizon T. Let $\tau \leq T$ denote the last timestep when GenCB-CMAB played arm x_0. For timestep $1 \leq n \leq \tau$, define event

$$\mathcal{E}_n := \left\{\sum_{t=1}^{n} r_{t,x_t} \geq \sum_{t=1}^{n} \mu_{x_t} - 2\sqrt{n\ln(\tau)}\right\}.$$

According to the Azuma-Hoeffding inequality, we have

$$\Pr[\bar{\mathcal{E}}_n] \leq \frac{1}{\tau^2}.$$

We further define event

$$\mathcal{E} := \bigcap_{n=1}^{\tau} \mathcal{E}_n.$$

By a union bound over $1 \leq n \leq \tau$, we have

$$\begin{aligned} \Pr[\bar{\mathcal{E}}] \leq & \sum_{n=1}^{\tau} \frac{1}{\tau^2} \\ \leq & \tau \cdot \frac{1}{\tau^2} \\ \leq & \frac{1}{\tau}. \end{aligned}$$

Thus, $\Pr[\mathcal{E}] > 1 - \frac{1}{\tau}$. In other words, with probability at least $1 - \frac{1}{\tau}$, we have that for any $1 \leq n \leq \tau$, $\sum_{t=1}^{n} r_{t,x_t} \geq \sum_{t=1}^{n} \mu_{x_t} - 2\sqrt{n\ln(\tau)}$.

At the timestep τ, we have the following three equivalent inequalities:

$$\sum_{t\in\mathcal{S}_{\tau-1}} r_t + N_0(\tau-1)\mu_0 < (1-\alpha)\mu_0\tau$$
$$\sum_{t\in\mathcal{S}_{\tau-1}} r_t + N_0(\tau-1)\mu_0 < (1-\alpha)\mu_0(N_0(\tau-1)+m_{\tau-1}+1)$$
$$\alpha\mu_0 N_0(\tau-1) < (1-\alpha)\mu_0(m_{\tau-1}+1) - \sum_{t\in\mathcal{S}_{\tau-1}} r_t$$

In the standard multi-armed bandit problem, we define the pseudo-regret of the well-known UCB (Auer, Cesa-Bianchi, and Fischer 2002) algorithm for any time m as

$$\tilde{\mathcal{R}}_m(\mathsf{UCB}) = \mu_* m - \sum_{t=1}^{m} \mu_{x_t}.$$

Thus, we have

$$\sum_{t=1}^{m} \mu_{x_t} = \mu_* m - \tilde{\mathcal{R}}_m(\mathsf{UCB}).$$

Then, we have

$$\begin{aligned}
\alpha\mu_0 N_0(\tau-1) &< (1-\alpha)\mu_0(m_{\tau-1}+1) - \sum_{t\in\mathcal{S}_{\tau-1}} r_t \\
&\quad - \tilde{\mathcal{R}}_{m_{\tau-1}}(\mathsf{UCB}) + \tilde{\mathcal{R}}_{m_{\tau-1}}(\mathsf{UCB}) \\
&= (1-\alpha)\mu_0(m_{\tau-1}+1) - \sum_{t\in\mathcal{S}_{\tau-1}} r_t - \mu_* m_{\tau-1} \\
&\quad + \sum_{t\in\mathcal{S}_{\tau-1}} \mu_{x_t} + \tilde{\mathcal{R}}_{m_{\tau-1}}(\mathsf{UCB}) \\
&= -(\mu_* - (1-\alpha)\mu_0)(m_{\tau-1}+1) + \sum_{t\in\mathcal{S}_{\tau-1}} \mu_{x_t} \\
&\quad - \sum_{t\in\mathcal{S}_{\tau-1}} r_t + \tilde{\mathcal{R}}_{m_{\tau-1}}(\mathsf{UCB}) + \mu_* \\
&= -(\Delta_0 + \alpha\mu_0)(m_{\tau-1}+1) + \sum_{t\in\mathcal{S}_{\tau-1}} \mu_{x_t} \\
&\quad - \sum_{t\in\mathcal{S}_{\tau-1}} r_t + \tilde{\mathcal{R}}_{m_{\tau-1}}(\mathsf{UCB}) + \mu_* \qquad (1)
\end{aligned}$$

In the following analysis, we assume $m_{\tau-1} \geq 1, N_0(\tau-1) \geq 2$, since otherwise the theorem trivially holds. Since $\tau = m_{\tau-1} + N_0(\tau-1) + 1 > 2$, we have $\mathbb{E}[m_{\tau-1}|\mathcal{E}] \leq \frac{\mathbb{E}[m_{\tau-1}]}{\Pr[\mathcal{E}]} \leq \frac{\mathbb{E}[m_{\tau-1}]}{1-\frac{1}{\tau}} < 2\mathbb{E}[m_{\tau-1}]$ and $\mathbb{E}[N_0(\tau-1)|\mathcal{E}] \leq \frac{\mathbb{E}[N_0(\tau-1)]}{\Pr[\mathcal{E}]} \leq \frac{\mathbb{E}[N_0(\tau-1)]}{1-\frac{1}{\tau}} < 2\mathbb{E}[N_0(\tau-1)]$. Taking expectation of both sides in (1), from $\mathbb{E}[\tilde{\mathcal{R}}_{m_{\tau-1}}(\mathsf{UCB})] \leq \mathbb{E}[8H\ln(m_{\tau-1}+1)+5K]$ and Jensen's inequality, we have

$$\begin{aligned}
\alpha\mu_0\mathbb{E}[N_0(\tau-1)] &< -(\Delta_0+\alpha\mu_0)\mathbb{E}[m_{\tau-1}+1] \\
&\quad + \mathbb{E}\left[\sum_{t\in\mathcal{S}_{\tau-1}} \mu_{x_t} - \sum_{t\in\mathcal{S}_{\tau-1}} r_t|\mathcal{E}\right]\Pr[\mathcal{E}] \\
&\quad + \mathbb{E}\left[\sum_{t\in\mathcal{S}_{\tau-1}} \mu_{x_t} - \sum_{t\in\mathcal{S}_{\tau-1}} r_t|\bar{\mathcal{E}}\right]\Pr[\bar{\mathcal{E}}] \\
&\quad + \mathbb{E}[\tilde{\mathcal{R}}_{m_{\tau-1}}(\mathsf{UCB})] + \mu_* \\
&\leq -(\Delta_0+\alpha\mu_0)\mathbb{E}[m_{\tau-1}+1] \\
&\quad + \mathbb{E}\left[2\sqrt{m_{\tau-1}\ln(\tau)}|\mathcal{E}\right] + \tau\cdot\frac{1}{\tau} \\
&\quad + \tilde{\mathcal{R}}(\mathbb{E}[m_{\tau-1}]) + \mu_* \\
&< -(\Delta_0+\alpha\mu_0)\mathbb{E}[m_{\tau-1}+1] \\
&\quad + 2\sqrt{2\mathbb{E}[m_{\tau-1}+1]}\ln(2\mathbb{E}[N_0(\tau-1)]\cdot \\
&\quad \mathbb{E}[m_{\tau-1}+1]) + 8H\ln(\mathbb{E}[m_{\tau-1}+1]) \\
&\quad + 5K + 2 \qquad (2)
\end{aligned}$$

Let $m = \mathbb{E}[m_{\tau-1}+1] \geq 2$, $c_1 = 2\sqrt{2}$, $c_2 = 2\mathbb{E}[N_0(\tau-1)] \geq 2$, $c_3 = \Delta_0 + \alpha\mu_0 \in (0,1)$, $c_4 = 8H$ where $H \geq 2$. The RHS of Eq. (2) can be written as a constant term plus

$$g_1(m) = -c_3 m + c_1\sqrt{m}\ln(c_2 m) + c_4\ln(m).$$

According to Lemma 1, we have

$$g_1(m) \leq \frac{132c_1c_4}{c_3}\left[\ln\left(\frac{10\sqrt{c_2c_4}}{c_3}\right)\right]^2.$$

Then, we have

$$\begin{aligned}
&\alpha\mu_0\mathbb{E}[N_0(\tau-1)] \\
&< \frac{2112\sqrt{2}H}{\Delta_0+\alpha\mu_0}\left[\ln\left(\frac{40\sqrt{H\mathbb{E}[N_0(\tau-1)]}}{\Delta_0+\alpha\mu_0}\right)\right]^2 + 5K + 2 \\
&< \frac{2994H}{\Delta_0+\alpha\mu_0}\left[\ln\left(\frac{40\sqrt{H\mathbb{E}[N_0(\tau-1)]}}{\Delta_0+\alpha\mu_0}\right)\right]^2.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\mathbb{E}[N_0(\tau-1)] &< \frac{2994H}{\alpha\mu_0(\Delta_0+\alpha\mu_0)}\cdot \\
&\qquad \left[\ln\left(\frac{40\sqrt{H\mathbb{E}[N_0(\tau-1)]}}{\Delta_0+\alpha\mu_0}\right)\right]^2 \\
\sqrt{\mathbb{E}[N_0(\tau-1)]} &< \sqrt{\frac{2994H}{\alpha\mu_0(\Delta_0+\alpha\mu_0)}}\cdot \\
&\qquad \ln\left(\frac{40\sqrt{H\mathbb{E}[N_0(\tau-1)]}}{\Delta_0+\alpha\mu_0}\right)
\end{aligned}$$

According to Fact 2 (set $z = \sqrt{\mathbb{E}[N_0(\tau-1)]}, c_1 = \sqrt{\frac{2994H}{\alpha\mu_0(\Delta_0+\alpha\mu_0)}}, c_2 = \frac{40\sqrt{H}}{\Delta_0+\alpha\mu_0}$),

$$\begin{aligned}
\sqrt{\mathbb{E}[N_0(\tau-1)]} &\leq 2\sqrt{\frac{2994H}{\alpha\mu_0(\Delta_0+\alpha\mu_0)}}\cdot \\
&\qquad \ln\left(\sqrt{\frac{2994H}{\alpha\mu_0(\Delta_0+\alpha\mu_0)}}\cdot\frac{40\sqrt{H}}{\Delta_0+\alpha\mu_0}\right)
\end{aligned}$$

Algorithm 4: GenCB-CLB

Input: Reugular arms $\mathcal{X} \subseteq \mathbb{R}^d$, default arm x_0 with reward μ_0, parameter α, L, S, $\lambda \geq \max\{1, L^2\}$.

$\forall t \geq 0, N_0(t) \leftarrow 0, r_S(t) \leftarrow 0$. $m \leftarrow 0$. $V_0 \leftarrow \lambda I$. $b_0 \leftarrow \mathbf{0}^d$;

for $t = 1, 2, \ldots$ **do**

  if $r_S(t-1) + N_0(t-1)\mu_0 \geq (1-\alpha)\mu_0 t$ **then**

    $m \leftarrow m + 1$;

    $\mathcal{C}_t \leftarrow \{\theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_{t-1}\|_{V_{t-1}} \leq \sqrt{d\ln(2m^2(1+mL^2/\lambda))} + \sqrt{\lambda}S\}$;

    $(x_t, \tilde{\theta}_t) \leftarrow \text{argmax}_{(x,\theta)\in\mathcal{X}\times\mathcal{C}_t} x^\top\theta$;

    Play arm x_t and observe the random reward r_{t,x_t};

    $V_t \leftarrow V_{t-1} + x_t x_t^\top$, $b_t \leftarrow b_{t-1} + r_{t,x_t}x_t$;

    $\hat{\theta}_t \leftarrow V_t^{-1} b_t$;

    $r_S(t) \leftarrow r_S(t-1) + r_{t,x_t}$;

  else

    Play x_0 and receive reward μ_0;

    $N_0(t) \leftarrow N_0(t-1) + 1$;

    $r_S(t) \leftarrow r_S(t-1)$;

$$\mathbb{E}[N_0(\tau-1)] \leq \frac{11976H}{\alpha\mu_0(\Delta_0+\alpha\mu_0)} \cdot \left[\ln\left(\frac{2189H}{(\Delta_0+\alpha\mu_0)\sqrt{\alpha\mu_0(\Delta_0+\alpha\mu_0)}}\right)\right]^2$$

Thus,

$$\begin{aligned}\mathbb{E}[N_0(T)] =& \mathbb{E}[N_0(\tau)] \\ =& \mathbb{E}[N_0(\tau-1)] + 1 \\ =& O\left(\frac{H}{\alpha\mu_0(\Delta_0+\alpha\mu_0)}\left[\ln\left(\frac{H}{\alpha\mu_0(\Delta_0+\alpha\mu_0)}\right)\right]^2\right).\end{aligned}$$

Theorem 2 follows from $\mathbb{E}[\mathcal{R}_T(\text{GenCB-CMAB})] \leq \mathbb{E}[\mathcal{R}_T(\text{UCB})] + \mathbb{E}[N_0(T)]\Delta_0$. □

D Algorithm Pseudo-code and Proof for CLB

Algorithm 4 presents the algorithm pseudo-code of GenCB-CLB for CLB, and we give the detailed proof of Theorem 3 in the following.

Proof. Since the proof of satisfaction on the performance constraint Eq. (2) is the same to Theorem 2, we mainly give the proof of regret bound here.

For timestep $1 \leq n \leq \tau$, define event

$$\mathcal{E}_n := \left\{\sum_{t=1}^n r_{x_t} \geq \sum_{t=1}^n \mu_{x_t} - 2\sqrt{n\ln(\tau)}\right\}.$$

According to the Azuma-Hoeffding inequality, we have

$$\Pr[\bar{\mathcal{E}}_n] \leq \frac{1}{\tau^2}.$$

We further define event

$$\mathcal{E} := \bigcap_{n=1}^{\tau} \mathcal{E}_n.$$

By a union bound over $1 \leq n \leq \tau$, we have

$$\begin{aligned}\Pr[\bar{\mathcal{E}}] \leq& \sum_{n=1}^{\tau}\frac{1}{\tau^2} \\ \leq& \tau \cdot \frac{1}{\tau^2} \\ \leq& \frac{1}{\tau}.\end{aligned}$$

Thus, $\Pr[\mathcal{E}] > 1 - \frac{1}{\tau}$. In other words, with probability at least $1 - \frac{1}{\tau}$, we have that for any $1 \leq n \leq \tau$, $\sum_{t=1}^n r_{x_t} \geq \sum_{t=1}^n x_t^\top\theta^* - 2\sqrt{n\ln(\tau)}$.

For the confidence ellipsoid in the LinUCB (Abbasi-yadkori, Pál, and Szepesvári 2011) algorithm, we set the confidence parameter $\delta_t = 1/(2t^2)$ and the confidence ellipsoid for timestep t is

$$\mathcal{C}_t = \{\theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_{t-1}\|_{V_{t-1}} \leq \sqrt{d\ln(2m_t^2(1+m_tL^2/\lambda))} + \sqrt{\lambda}S\}.$$

Then, we can obtain

$$\begin{aligned}&\mathbb{E}[\tilde{\mathcal{R}}_{m_{\tau-1}}(\text{LinUCB})] \\ =& \mathbb{E}\left[\mu_* m_{\tau-1} - \sum_{t=1}^{m_{\tau-1}} \mu_{x_t}\right] \\ \leq& \mathbb{E}\left[4\sqrt{m_{\tau-1}d\ln\left(1+\frac{m_{\tau-1}L^2}{\lambda d}\right)} \cdot \left(\sqrt{\lambda}S + \sqrt{d\ln\left(2m_{\tau-1}^2 \cdot \left(1+\frac{m_{\tau-1}L^2}{\lambda}\right)\right)}\right) + \Delta_{\max}\right].\end{aligned}$$

Fix time horizon T. Recall that $\tau \leq T$ is the last timestep when we played x_0. Similar to the analysis in CMAB (Eq. (1)), at timestep τ, we have

$$\begin{aligned}\alpha\mu_0 N_0(\tau-1) <& -(\Delta_0+\alpha\mu_0)(m_{\tau-1}+1) + \sum_{t\in\mathcal{S}_{\tau-1}} \mu_{x_t} \\ &- \sum_{t\in\mathcal{S}_{\tau-1}} r_t + \tilde{\mathcal{R}}_{m_{\tau-1}}(\text{LinUCB}) + \mu_*.\end{aligned}$$

Taking expectation of both sides, we have

$$\begin{aligned}\alpha\mu_0\mathbb{E}[N_0(\tau-1)] <& -(\Delta_0+\alpha\mu_0)\mathbb{E}[m_{\tau-1}+1] \\ &+ \mathbb{E}\left[\sum_{t\in\mathcal{S}_{\tau-1}} \mu_{x_t} - \sum_{t\in\mathcal{S}_{\tau-1}} r_t | \mathcal{E}\right]\Pr[\mathcal{E}] \\ &+ \mathbb{E}\left[\sum_{t\in\mathcal{S}_{\tau-1}} \mu_{x_t} - \sum_{t\in\mathcal{S}_{\tau-1}} r_t | \bar{\mathcal{E}}\right]\Pr[\bar{\mathcal{E}}] \\ &+ \mathbb{E}[\tilde{\mathcal{R}}_{m_{\tau-1}}(\text{LinUCB})] + \mu_*\end{aligned}$$

$$
\begin{aligned}
&\leq -(\Delta_0+\alpha\mu_0)\mathbb{E}[m_{\tau-1}+1] \\
&\quad + \mathbb{E}\left[2\sqrt{m_{\tau-1}\ln(\tau)}|\mathcal{E}\right] + \tau\cdot\frac{1}{\tau} \\
&\quad + \mathbb{E}[\tilde{\mathcal{R}}_{m_{\tau-1}}(\mathsf{LinUCB})] + \mu_* \\
&< -(\Delta_0+\alpha\mu_0)\mathbb{E}[m_{\tau-1}+1] \\
&\quad +2\sqrt{2\mathbb{E}[m_{\tau-1}]\ln(2\mathbb{E}[N_0(\tau-1)]\mathbb{E}[m_{\tau-1}+1])} \\
&\quad + 4\sqrt{\mathbb{E}[m_{\tau-1}]d\ln\left(1+\frac{\mathbb{E}[m_{\tau-1}]L^2}{\lambda d}\right)}\cdot \\
&\left(\sqrt{\lambda}S+\sqrt{d\ln\left(2\mathbb{E}[m_{\tau-1}]^2\cdot\left(1+\frac{\mathbb{E}[m_{\tau-1}]L^2}{\lambda}\right)\right)}\right) \\
&\quad + \Delta_{\max} + 2 \\
&< -(\Delta_0+\alpha\mu_0)\mathbb{E}[m_{\tau-1}+1] \\
&\quad + 19d\sqrt{\lambda}S\sqrt{(\mathbb{E}[m_{\tau-1}+1])}\cdot \\
&\quad \ln\left(2\mathbb{E}[N_0(\tau-1)]\cdot\mathbb{E}[m_{\tau-1}+1]\right)+3,
\end{aligned}
$$

where $\Delta_{\max} = \max_{x\in\mathcal{X}}\Delta_x$.

Let $m = \mathbb{E}[m_{\tau-1}+1]$, $c_1 = 19d\sqrt{\lambda}S$, $c_2 = 2\mathbb{E}[N_0(\tau-1)]$, $c_3 = \Delta_0+\alpha\mu_0$. According to Fact 1, we have

$$
\alpha\mu_0\mathbb{E}[N_0(\tau-1)] < \frac{5776d^2S^2\lambda}{9(\Delta_0+\alpha\mu_0)}\cdot \left[\ln\left(\frac{147d\sqrt{\lambda}S\sqrt{\mathbb{E}[N_0(\tau-1)]}}{\Delta_0+\alpha\mu_0}\right)\right]^2+3.
$$

Thus, we have

$$
\begin{aligned}
\mathbb{E}[N_0(\tau-1)] &< \frac{645d^2S^2\lambda}{\alpha\mu_0(\Delta_0+\alpha\mu_0)}\cdot \left[\ln\left(\frac{147d\sqrt{\lambda}S\sqrt{\mathbb{E}[N_0(\tau-1)]}}{\Delta_0+\alpha\mu_0}\right)\right]^2 \\
\sqrt{\mathbb{E}[N_0(\tau-1)]} &< \frac{26d\sqrt{\lambda}S}{\sqrt{\alpha\mu_0(\Delta_0+\alpha\mu_0)}}\cdot \ln\left(\frac{147d\sqrt{\lambda}S\sqrt{\mathbb{E}[N_0(\tau-1)]}}{\Delta_0+\alpha\mu_0}\right)
\end{aligned}
$$

According to Fact 2 (set $z = \sqrt{\mathbb{E}[N_0(\tau-1)]}, c_1 = \frac{26d\sqrt{\lambda}S}{\sqrt{\alpha\mu_0(\Delta_0+\alpha\mu_0)}}, c_2 = \frac{147d\sqrt{\lambda}S}{\Delta_0+\alpha\mu_0}$),

$$
\begin{aligned}
\sqrt{\mathbb{E}[N_0(\tau-1)]} &\leq \frac{52d\sqrt{\lambda}S}{\sqrt{\alpha\mu_0(\Delta_0+\alpha\mu_0)}}\cdot \ln\left(\frac{3822d^2S^2\lambda}{(\Delta_0+\alpha\mu_0)^{\frac{3}{2}}\sqrt{\alpha\mu_0}}\right) \\
\mathbb{E}[N_0(\tau-1)] &\leq \frac{2704d^2S^2\lambda}{\alpha\mu_0(\Delta_0+\alpha\mu_0)}\cdot \left[\ln\left(\frac{3822d^2S^2\lambda}{(\Delta_0+\alpha\mu_0)^{\frac{3}{2}}\sqrt{\alpha\mu_0}}\right)\right]^2
\end{aligned}
$$

Algorithm 5: GenCB-CCCB

Input: Reugular arms (decision class) $\mathcal{X}$, base arms $x_1,\ldots,x_K\in\mathbb{R}^d$, default arm x_0 with reward μ_0, parameter α, L, S, $\lambda\geq\max\{1,L^2\}$.

$\forall t\geq 0, N_0(t)\leftarrow 0, r_S(t)\leftarrow 0$. $m\leftarrow 0$. $V_0\leftarrow\lambda I$, $b_0\leftarrow\mathbf{0}^d$;
for $t=1,2,\ldots$ **do**
  if $r_S(t-1)+N_0(t-1)\mu_0\geq(1-\alpha)\mu_0 t$ **then**
    $m\leftarrow m+1$;
    $\hat{w}_{t,e}\leftarrow x_e^\top\hat{\theta}_{t-1}, \forall e\in[K]$;
    $\bar{w}_{t,e}\leftarrow\hat{w}_{t,e}+(\sqrt{d\ln(2m^2(1+mKL^2/\lambda))}$
      $+\sqrt{\lambda}S)\|x_e\|_{V_{t-1}^{-1}}, \forall e\in[K]$;
    $A_t\leftarrow\operatorname{argmax}_{A\in\mathcal{X}}\bar{f}(A,\bar{\boldsymbol{w}})$;
    Play arm A_t and observe the random reward $w_{t,e}$ for all $e\in A_t$;
    $V_t\leftarrow\lambda I+\sum_{s=1}^t\sum_{e\in A_s}x_ex_e^\top$;
    $b_t\leftarrow\sum_{s=1}^t\sum_{e\in A_s}w_{s,e}x_e$;
    $\hat{\theta}_t\leftarrow V_t^{-1}b_t$;
    $r_S(t)\leftarrow r_S(t-1)+r_{t,A_t}$;
  else
    Play x_0 and receive reward μ_0;
    $N_0(t)\leftarrow N_0(t-1)+1$;
    $r_S(t)\leftarrow r_S(t-1)$;

Thus,

$$
\begin{aligned}
&\mathbb{E}[N_0(T)] \\
=&\mathbb{E}[N_0(\tau)] \\
=&\mathbb{E}[N_0(\tau-1)]+1 \\
=&O\left(\frac{d^2S^2\lambda}{\alpha\mu_0(\Delta_0+\alpha\mu_0)}\left[\ln\left(\frac{dS\sqrt{\lambda}}{\alpha\mu_0(\Delta_0+\alpha\mu_0)}\right)\right]^2\right).
\end{aligned}
$$

Theorem 3 follows from $\mathbb{E}[\mathcal{R}_T(\mathsf{GenCB\text{-}CLB})] \leq \mathbb{E}[\mathcal{R}_T(\mathsf{LinUCB})]+\mathbb{E}[N_0(T)]\Delta_0$. □

E Algorithm Pseudo-code and Proof for CCCB

Algorithm 5 presents the algorithm pseudo-code of GenCB-CCCB for CCCB, and we give the detailed proof of Theorem 4 in the following.

Proof. Since the proof of satisfaction on the performance constraint Eq. (2) is the same to Theorem 2, we mainly give the proof of regret bound here.

For timestep $1\leq n\leq\tau$, define event

$$
\mathcal{E}_n := \left\{\sum_{t=1}^n r_{t,A_t}\geq\sum_{t=1}^n f(A_t,\boldsymbol{w}^*)-2K\sqrt{n\ln(\tau)}\right\}.
$$

According to the Azuma-Hoeffding inequality, we have

$$
\Pr[\bar{\mathcal{E}}_n]\leq\frac{1}{\tau^2}.
$$

We further define event

$$\mathcal{E} := \bigcap_{n=1}^{\tau} \mathcal{E}_n.$$

By a union bound over $1 \le n \le T$, we have

$$\begin{aligned}\Pr[\bar{\mathcal{E}}] \le & \sum_{n=1}^{\tau} \frac{1}{\tau^2} \\ \le & \tau \cdot \frac{1}{\tau^2} \\ \le & \frac{1}{\tau}.\end{aligned}$$

Thus, $\Pr[\bar{\mathcal{E}}] > 1 - \frac{1}{\tau}$. In other words, with probability at least $1 - \frac{1}{\tau}$, we have that for any $1 \le n \le \tau$, $\sum_{t=1}^{n} r_{t,A_t} \ge \sum_{t=1}^{n} f(A_t, \boldsymbol{w}^*) - 2K\sqrt{n \ln(\tau)}$.

For the confidence ellipsoid in the C2UCB (Qin, Chen, and Zhu 2014) algorithm, we set the confidence parameter $\delta_t = 1/(2t^2)$ and the confidence ellipsoid for timestep t is

$$\begin{aligned}\mathcal{C}_t = \big\{\theta \in \mathbb{R}^d : & \|\theta - \hat{\theta}_{t-1}\|_{V_{t-1}} \le \\ & \sqrt{d \ln(2m^2(1 + mKL^2/\lambda))} + \sqrt{\lambda} S\big\}.\end{aligned}$$

Then, we can obtain

$$\begin{aligned}&\mathbb{E}[\mathcal{R}_{m_{\tau-1}}(\textsf{C2UCB})] \\ \le & \mathbb{E}\Bigg[2P\sqrt{2dm_{\tau-1}\ln\left(1 + \frac{m_{\tau-1}KL^2}{\lambda d}\right)} \cdot \\ & \left(\sqrt{\lambda}S + \sqrt{d\ln\left(2m_{\tau-1}^2 \cdot \left(1 + \frac{m_{\tau-1}KL^2}{\lambda}\right)\right)}\right) + \Delta_{\max}\Bigg],\end{aligned}$$

where $\Delta_{\max} = \max_{A \in \mathcal{X}}(f(A_*, \boldsymbol{w}^*) - f(A, \boldsymbol{w}^*))$.

Fix time horizon T. Recall that $\tau \le T$ is the last timestep when we played x_0. Similar to the conservative multi-armed bandit case (Eq. (1)), at timestep τ, we have

$$\begin{aligned}\alpha\mu_0 N_0(\tau - 1) < & -(\Delta_0 + \alpha\mu_0)(m_{\tau-1} + 1) \\ & + \sum_{t \in S_{\tau-1}} f(A_t, \boldsymbol{w}^*) - \sum_{t \in S_{\tau-1}} r_{t,A_t} \\ & + \tilde{\mathcal{R}}(m_{\tau-1}) + \mu_*.\end{aligned}$$

Taking expectation of both sides, we have

$$\begin{aligned}&\alpha\mu_0\mathbb{E}[N_0(\tau - 1)] \\ < & -(\Delta_0 + \alpha\mu_0)\mathbb{E}[m_{\tau-1} + 1] \\ & + \mathbb{E}\left[\sum_{t \in S_{\tau-1}} f(A_t, \boldsymbol{w}^*) - \sum_{t \in S_{\tau-1}} r_{t,A_t} | \mathcal{E}\right] \Pr[\mathcal{E}] \\ & + \mathbb{E}\left[\sum_{t \in S_{\tau-1}} f(A_t, \boldsymbol{w}^*) - \sum_{t \in S_{\tau-1}} r_{t,A_t} | \bar{\mathcal{E}}\right] \Pr[\bar{\mathcal{E}}] \\ & + \mathbb{E}[\tilde{\mathcal{R}}(m_{\tau-1})] + \mu_* \\ \le & -(\Delta_0 + \alpha\mu_0)\mathbb{E}[m_{\tau-1} + 1] + \mathbb{E}\left[2K\sqrt{m_{\tau-1}\ln(\tau)} | \mathcal{E}\right] \\ & + \tau \cdot \frac{1}{\tau} + \mathbb{E}[\tilde{\mathcal{R}}(m_{\tau-1})] + \mu_* \\ < & -(\Delta_0 + \alpha\mu_0)\mathbb{E}[m_{\tau-1} + 1] \\ & + 2K\sqrt{2\mathbb{E}[m_{\tau-1}]\ln(2\mathbb{E}[N_0(\tau - 1)]\mathbb{E}[m_{\tau-1} + 1])} \\ & + 2P\sqrt{2d\mathbb{E}[m_{\tau-1}]\ln\left(1 + \frac{\mathbb{E}[m_{\tau-1}]KL^2}{\lambda d}\right)} \cdot \\ & \left(\sqrt{\lambda}S + \sqrt{d\ln\left(2\mathbb{E}[m_{\tau-1}]^2 \cdot \left(1 + \frac{\mathbb{E}[m_{\tau-1}]KL^2}{\lambda}\right)\right)}\right) \\ & + \Delta_{\max} + 2 \\ < & -(\Delta_0 + \alpha\mu_0)\mathbb{E}[m_{\tau-1} + 1] + (2\sqrt{2}K + 10P\sqrt{\lambda}Sd) \cdot \\ & \sqrt{(\mathbb{E}[m_{\tau-1} + 1])}\ln(2K\mathbb{E}[N_0(\tau - 1)] \cdot \mathbb{E}[m_{\tau-1} + 1]) + 3.\end{aligned}$$

Let $m = \mathbb{E}[m_{\tau-1} + 1]$, $c_1 = 2\sqrt{2}K + 10P\sqrt{\lambda}Sd$, $c_2 = 2K\mathbb{E}[N_0(\tau - 1)]$, $c_3 = \Delta_0 + \alpha\mu_0$. According to Fact 1, we have

$$\begin{aligned}&\alpha\mu_0\mathbb{E}[N_0(\tau - 1)] < \frac{16(2\sqrt{2}K + 10P\sqrt{\lambda}Sd)^2}{9(\Delta_0 + \alpha\mu_0)} \cdot \\ & \left[\ln\left(\frac{8(2\sqrt{2}K + 10P\sqrt{\lambda}Sd)\sqrt{K\mathbb{E}[N_0(\tau - 1)]}}{\Delta_0 + \alpha\mu_0}\right)\right]^2 + 3.\end{aligned}$$

Thus, we have

$$\begin{aligned}&\mathbb{E}[N_0(\tau - 1)] < \frac{5(2\sqrt{2}K + 10P\sqrt{\lambda}Sd)^2}{\alpha\mu_0(\Delta_0 + \alpha\mu_0)} \cdot \\ & \left[\ln\left(\frac{8K(2\sqrt{2}K + 10P\sqrt{\lambda}Sd)\sqrt{\mathbb{E}[N_0(\tau - 1)]}}{\Delta_0 + \alpha\mu_0}\right)\right]^2 \\ & \sqrt{\mathbb{E}[N_0(\tau - 1)]} < \frac{3(2\sqrt{2}K + 10P\sqrt{\lambda}Sd)}{\sqrt{\alpha\mu_0(\Delta_0 + \alpha\mu_0)}} \cdot \\ & \ln\left(\frac{8K(2\sqrt{2}K + 10P\sqrt{\lambda}Sd)\sqrt{\mathbb{E}[N_0(\tau - 1)]}}{\Delta_0 + \alpha\mu_0}\right)\end{aligned}$$

According to Fact 2 (set $z = \sqrt{\mathbb{E}[N_0(\tau - 1)]}, c_1 = \frac{3(2\sqrt{2}K + 10P\sqrt{\lambda}Sd)}{\sqrt{\alpha\mu_0(\Delta_0 + \alpha\mu_0)}}, c_2 = \frac{8K(2\sqrt{2}K + 10P\sqrt{\lambda}Sd)}{\Delta_0 + \alpha\mu_0}$),

$$\begin{aligned}\sqrt{\mathbb{E}[N_0(\tau - 1)]} \le & \frac{12(2\sqrt{2}K + 10P\sqrt{\lambda}Sd)}{\sqrt{\alpha\mu_0(\Delta_0 + \alpha\mu_0)}} \cdot \\ & \ln\left(\frac{5K(2\sqrt{2}K + 10P\sqrt{\lambda}Sd)}{(\Delta_0 + \alpha\mu_0)\sqrt{\alpha\mu_0(\Delta_0 + \alpha\mu_0)}}\right) \\ \mathbb{E}[N_0(\tau - 1)] \le & \frac{144(2\sqrt{2}K + 10P\sqrt{\lambda}Sd)^2}{\alpha\mu_0(\Delta_0 + \alpha\mu_0)} \cdot \\ & \left[\ln\left(\frac{5K(2\sqrt{2}K + 10P\sqrt{\lambda}Sd)}{(\Delta_0 + \alpha\mu_0)\sqrt{\alpha\mu_0(\Delta_0 + \alpha\mu_0)}}\right)\right]^2\end{aligned}$$

Thus,

$$\begin{aligned}&\mathbb{E}[N_0(T)] \\ = & \mathbb{E}[N_0(\tau)]\end{aligned}$$

$$
\begin{aligned}
&=\mathbb{E}[N_0(\tau-1)]+1\\
&=O\left(\frac{(K+P\sqrt{\lambda}Sd)^2}{\alpha\mu_0(\Delta_0+\alpha\mu_0)}\left[\ln\left(\frac{K+P\sqrt{\lambda}Sd}{\alpha\mu_0(\Delta_0+\alpha\mu_0)}\right)\right]^2\right).
\end{aligned}
$$

Theorem 4 follows from $\mathbb{E}[\mathcal{R}_T(\textsf{GenCB-CCCB})] \leq \mathbb{E}[\mathcal{R}_T(\textsf{C2UCB})]+\mathbb{E}[N_0(T)]\Delta_0$. □

F Proof for MV-CBP

We give the detailed proof of Theorem 5 below.

Proof. In order to prove that MV-CUCB satisfies the sample-path reward constraint Eq. (5), we give the following inequalities first. For any time horizon T,

$$
\begin{aligned}
&T\cdot\widehat{\mathrm{MV}}_T(\mathcal{A})\\
=&T\cdot\left(\rho\hat{\mu}_T(\mathcal{A})-\hat{\sigma}_T^2(\mathcal{A})\right)\\
=&T\cdot\frac{\rho}{T}\sum_{t=1}^{T}r_{t,x_t}-T\cdot\frac{1}{T}\sum_{t=1}^{T}r_{t,x_t}^2+T\cdot\left(\frac{\sum_{t=1}^{T}r_{t,x_t}}{T}\right)^2\\
=&\rho\sum_{t=1}^{T-1}r_{t,x_t}+\rho r_{T,x_T}-\sum_{t=1}^{T-1}r_{t,x_t}^2-r_{T,x_T}^2\\
&+\frac{1}{T}\left(\left(\sum_{t=1}^{T-1}r_{t,x_t}\right)^2+r_{T,x_T}^2+2\left(\sum_{t=1}^{T-1}r_{t,x_t}\right)r_{T,x_T}\right)\\
=&\rho\sum_{t=1}^{T-1}r_{t,x_t}-\sum_{t=1}^{T-1}r_{t,x_t}^2+\frac{1}{T-1}\left(\sum_{t=1}^{T-1}r_{t,x_t}\right)^2\\
&-\frac{1}{T-1}\left(\sum_{t=1}^{T-1}r_{t,x_t}\right)^2+\rho r_{T,x_T}-r_{T,x_T}^2\\
&+\frac{1}{T}\left(\left(\sum_{t=1}^{T-1}r_{t,x_t}\right)^2+r_{T,x_T}^2+2\left(\sum_{t=1}^{T-1}r_{t,x_t}\right)r_{T,x_T}\right)\\
\geq&(T-1)\cdot\widehat{\mathrm{MV}}_{T-1}(\mathcal{A})-1-\frac{1}{T(T-1)}\left(\sum_{t=1}^{T-1}r_{t,x_t}\right)^2\\
\geq&(T-1)\cdot\widehat{\mathrm{MV}}_{T-1}(\mathcal{A})-2 \qquad (3)
\end{aligned}
$$

Now we prove that MV-CUCB satisfies the sample-path reward constraint Eq. (5) by induction. At timestep $t=1$, since the LHS of the if statement (in Line 3 of Algorithm 3) is -2 and RHS is positive, MV-CUCB will pull the default arm x_0 and receive reward μ_0. Then, we have $\widehat{\mathrm{MV}}_1(\mathcal{A})=\mathrm{MV}_0\geq(1-\alpha)\mathrm{MV}_0$, which satisfies the constraint. Suppose that the sample-path reward constraint holds at timestep $t-1$. At time step t, if MV-CUCB plays x_0, since the exploration risk caused by one pull is bounded by 2 and $\alpha\mathrm{MV}_0>2$, the constraint still holds for t. If MV-CUCB plays a regular arm x_t, which implies $(t-1)\widehat{\mathrm{MV}}_{t-1}(\mathcal{A})-2\geq(1-\alpha)\mathrm{MV}_0 t$, then from Eq. (3) we have $t\widehat{\mathrm{MV}}_t(\mathcal{A})\geq(t-1)\widehat{\mathrm{MV}}_{t-1}(\mathcal{A})-2\geq(1-\alpha)\mathrm{MV}_0 t$, and thus the constraint still holds for t.

Next, we prove the regret bound of the MV-CUCB algorithm. Fix a time horizon T. Let $\tau\leq T$ denote the last timestep when the algorithm pulled arm x_0. For ease of notation, we use $N_{i,\tau-1}$ as a shorthand for $N_i(\tau-1)$, $\forall 0\leq i\leq K$. Define event

$$
\mathcal{F}:=\Bigg\{\forall i=1,\dots,K,\forall 1\leq n\leq\tau,
$$

$$
|\hat{\mu}_{i,n}-\mu_i|\leq\sqrt{\frac{\ln(6K\tau^2)}{2n}}\text{ and }|\hat{\sigma}_{i,n}^2-\sigma_i^2|\leq 5\sqrt{\frac{\ln(6K\tau^2)}{2n}}\Bigg\}.
$$

Similar to Lemma 1 in (Sani, Lazaric, and Munos 2012), using Chernoff-Hoeffding inequality and a union bound over arms and observations, we have $\Pr[\bar{\mathcal{F}}]\leq\frac{1}{\tau}$. Thus, $\Pr[\mathcal{F}]>1-\frac{1}{\tau}$.

Conditioning on $\mathcal{F}$, we have

$$
\begin{aligned}
&\sum_{i=1}^{K}N_{i,\tau-1}\mathrm{MV}_i-\frac{2}{\tau-1}\sum_{i=1}^{K}\sum_{\substack{j\neq i\\ j\neq 0}}N_{i,\tau-1}N_{j,\tau-1}\Gamma_{i,j}^2\\
\leq&\sum_{i=1}^{K}N_{i,\tau-1}\left(\widehat{\mathrm{MV}}_i+2(5+\rho)\sqrt{\frac{\log(6K\tau^2)}{2N_{i,\tau-1}}}\right)\\
&-\Bigg(\frac{2}{\tau-1}\sum_{i=1}^{K}\sum_{\substack{j\neq i\\ j\neq 0}}N_{i,\tau-1}N_{j,\tau-1}\Gamma_{i,j}^2\\
&+\frac{2\sqrt{2}}{\tau-1}\sum_{i=1}^{K}\sum_{\substack{j\neq i\\ j\neq 0}}N_{j,\tau-1}\log(6K\tau^2)\\
&+\frac{2\sqrt{2}}{\tau-1}\sum_{i=1}^{K}\sum_{\substack{j\neq i\\ j\neq 0}}N_{i,\tau-1}\log(6K\tau^2)\Bigg)\\
&+\frac{2\sqrt{2}}{\tau-1}\sum_{i=1}^{K}\sum_{\substack{j\neq i\\ j\neq 0}}N_{j,\tau-1}\log(6K\tau^2)\\
&+\frac{2\sqrt{2}}{\tau-1}\sum_{i=1}^{K}\sum_{\substack{j\neq i\\ j\neq 0}}N_{i,\tau-1}\log(6K\tau^2)\\
\leq&\sum_{i=1}^{K}N_{i,\tau-1}\widehat{\mathrm{MV}}_i+(5+\rho)\sum_{i=1}^{K}\sqrt{2N_{i,\tau-1}\log(6K\tau^2)}\\
&-\frac{1}{\tau-1}\sum_{i=1}^{K}\sum_{\substack{j\neq i\\ j\neq 0}}N_{i,\tau-1}N_{j,\tau-1}\cdot\\
&\left(|\Gamma_{i,j}|+\sqrt{\frac{\log(6K\tau^2)}{2N_{i,\tau-1}}}+\sqrt{\frac{\log(6K\tau^2)}{2N_{j,\tau-1}}}\right)^2\\
&+4\sqrt{2}K\log(6K\tau^2)\\
\leq&\sum_{i=1}^{K}N_{i,\tau-1}\widehat{\mathrm{MV}}_i-\frac{1}{\tau-1}\sum_{i=1}^{K}\sum_{\substack{j\neq i\\ j\neq 0}}N_{i,\tau-1}N_{j,\tau-1}\hat{\Gamma}_{i,j}^2
\end{aligned}
$$

$$+ (5+\rho)\sqrt{2Km_{\tau-1}N_{i,\tau-1}\log(6K\tau^2)}$$
$$+ 4\sqrt{2}K\log(6K\tau^2)$$

Let $L = 2, \mathrm{MV}_* \le \rho, \Delta_{\max}^{\mathrm{MV}} \le \frac{1}{4}+\rho$ and $\mathrm{GAP}_{\max} \le \frac{5}{4}+\rho$. We set the confidence parameter $\delta_t = 1/(12Kt^3)$ in the MV-UCB (Sani, Lazaric, and Munos 2012) algorithm and use Jensen's inequality, and then we have

$$\begin{aligned}
&\mathbb{E}[m_{\tau-1}\tilde{\mathcal{R}}_{m_{\tau-1}}(\text{MV-UCB})] \\
\le&\mathbb{E}\Bigg[12(5+\rho)^2(H_1^{\mathrm{MV}} + 4H_2^{\mathrm{MV}})\ln(6Km_{\tau-1}) \\
&+ 288(5+\rho)^4 H_3^{\mathrm{MV}}\frac{\ln^2(6Km_{\tau-1})}{m_{\tau-1}} + 9K + K\Delta_{\max}^{\mathrm{MV}}\Bigg] \\
\le&12(5+\rho)^2(H_1^{\mathrm{MV}} + 4H_2^{\mathrm{MV}})\ln(6K\mathbb{E}[m_{\tau-1}]) \\
&+ 288(5+\rho)^4 H_3^{\mathrm{MV}}\frac{\ln^2(6K\mathbb{E}[m_{\tau-1}])}{\mathbb{E}[m_{\tau-1}]} + 9K + K\Delta_{\max}^{\mathrm{MV}}.
\end{aligned}$$

At timestep τ, we have

$$(\tau-1)\widehat{\mathrm{MV}}_{\tau-1}(\mathcal{A}) - L < (1-\alpha)\mathrm{MV}_0\tau.$$

Thus,

$$\begin{aligned}
&\sum_{i=0}^{K} N_{i,\tau-1}\widehat{\mathrm{MV}}_i - \frac{1}{\tau-1}\sum_{i=0}^{K}\sum_{j\neq i} N_{i,\tau-1}N_{j,\tau-1}\hat{\Gamma}_{i,j}^2 - L \\
&< (1-\alpha)\mathrm{MV}_0(m_{\tau-1} + N_0(\tau-1) + 1)
\end{aligned}$$

Rearranging the terms, we have

$$\begin{aligned}
&\alpha\mathrm{MV}_0 N_0(\tau-1) \\
\le&(1-\alpha)\mathrm{MV}_0(m_{\tau-1}+1) \\
&- \left(\sum_{i=1}^{K} N_{i,\tau-1}\widehat{\mathrm{MV}}_i - \frac{1}{\tau-1}\sum_{i=1}^{K}\sum_{\substack{j\neq i\\ j\neq 0}} N_{i,\tau-1}N_{j,\tau-1}\hat{\Gamma}_{i,j}^2\right) \\
&+ \frac{2}{\tau-1}N_0(\tau-1)\sum_{i=1}^{K} N_{i,\tau-1}\hat{\Gamma}_{0,i}^2 + L \\
\le&(1-\alpha)\mathrm{MV}_0(m_{\tau-1}+1) \\
&- \left(\sum_{i=1}^{K} N_{i,\tau-1}\widehat{\mathrm{MV}}_i - \frac{1}{\tau-1}\sum_{i=1}^{K}\sum_{\substack{j\neq i\\ j\neq 0}} N_{i,\tau-1}N_{j,\tau-1}\hat{\Gamma}_{i,j}^2\right) \\
&- m_{\tau-1}\tilde{\mathcal{R}}_{m_{\tau-1}}(\text{MV-UCB}) + m_{\tau-1}\tilde{\mathcal{R}}_{m_{\tau-1}}(\text{MV-UCB}) \\
&+ 2N_0(\tau-1) + L \\
\le&(1-\alpha)\mathrm{MV}_0(m_{\tau-1}+1) \\
&- \left(\sum_{i=1}^{K} N_{i,\tau-1}\widehat{\mathrm{MV}}_i - \frac{1}{\tau-1}\sum_{i=1}^{K}\sum_{\substack{j\neq i\\ j\neq 0}} N_{i,\tau-1}N_{j,\tau-1}\hat{\Gamma}_{i,j}^2\right) \\
&- \mathrm{MV}_* m_{\tau-1} + \sum_{i=1}^{K} N_{i,\tau-1}\mathrm{MV}_i \\
&- \frac{2}{m_{\tau-1}}\sum_{i=1}^{K}\sum_{\substack{j\neq i\\ j\neq 0}} N_{i,\tau-1}N_{j,\tau-1}\Gamma_{i,j}^2 \\
&+ m_{\tau-1}\tilde{\mathcal{R}}_{m_{\tau-1}}(\text{MV-UCB}) + 2N_0(\tau-1) + L \\
\le&-(\mathrm{MV}_* - (1-\alpha)\mathrm{MV}_0)(m_{\tau-1}+1) \\
&- \left(\sum_{i=1}^{K} N_{i,\tau-1}\widehat{\mathrm{MV}}_i - \frac{1}{\tau-1}\sum_{i=1}^{K}\sum_{\substack{j\neq i\\ j\neq 0}} N_{i,\tau-1}N_{j,\tau-1}\hat{\Gamma}_{i,j}^2\right) \\
&+ \left(\sum_{i=1}^{K} N_{i,\tau-1}\mathrm{MV}_i - \frac{2}{\tau-1}\sum_{i=1}^{K}\sum_{\substack{j\neq i\\ j\neq 0}} N_{i,\tau-1}N_{j,\tau-1}\Gamma_{i,j}^2\right) \\
&+ m_{\tau-1}\tilde{\mathcal{R}}_{m_{\tau-1}}(\text{MV-UCB}) + 2N_0(\tau-1) + L + \mathrm{MV}_* \\
\le&-(\Delta_0^{\mathrm{MV}} + \alpha\mathrm{MV}_0)(m_{\tau-1}+1) \\
&+ \left(\sum_{i=1}^{K} N_{i,\tau-1}\mathrm{MV}_i - \frac{2}{\tau-1}\sum_{i=1}^{K}\sum_{\substack{j\neq i\\ j\neq 0}} N_{i,\tau-1}N_{j,\tau-1}\Gamma_{i,j}^2\right) \\
&- \left(\sum_{i=1}^{K} N_{i,\tau-1}\widehat{\mathrm{MV}}_i - \frac{1}{\tau-1}\sum_{i=1}^{K}\sum_{\substack{j\neq i\\ j\neq 0}} N_{i,\tau-1}N_{j,\tau-1}\hat{\Gamma}_{i,j}^2\right) \\
&+ m_{\tau-1}\tilde{\mathcal{R}}_{m_{\tau-1}}(\text{MV-UCB}) + 2N_0(\tau-1) + L + \mathrm{MV}_*
\end{aligned}$$

Taking expectation of both sides, we have

$$\begin{aligned}
&(\alpha\mathrm{MV}_0 - 2)\mathbb{E}[N_0(\tau-1)] \\
\le&-(\Delta_0^{\mathrm{MV}} + \alpha\mathrm{MV}_0)\mathbb{E}[m_{\tau-1}+1] \\
&+ \mathbb{E}\Bigg[\Bigg(\sum_{i=1}^{K} N_{i,\tau-1}\mathrm{MV}_i \\
&- \frac{2}{\tau-1}\sum_{i=1}^{K}\sum_{\substack{j\neq i\\ j\neq 0}} N_{i,\tau-1}N_{j,\tau-1}\Gamma_{i,j}^2\Bigg) \\
&-\Bigg(\sum_{i=1}^{K} N_{i,\tau-1}\widehat{\mathrm{MV}}_i \\
&- \frac{1}{\tau-1}\sum_{i=1}^{K}\sum_{\substack{j\neq i\\ j\neq 0}} N_{i,\tau-1}N_{j,\tau-1}\hat{\Gamma}_{i,j}^2\Bigg)|\mathcal{F}\Bigg]\Pr[\mathcal{F}] \\
&+ \mathrm{GAP}_{\max}\tau\Pr[\bar{\mathcal{F}}] + \mathbb{E}[m_{\tau-1}\tilde{\mathcal{R}}_{m_{\tau-1}}(\text{MV-UCB})] \\
&+ L + \mathrm{MV}_* \\
\le&-(\Delta_0^{\mathrm{MV}} + \alpha\mathrm{MV}_0)\mathbb{E}[m_{\tau-1}+1] \\
&+ \mathbb{E}[(5+\rho)\sqrt{2Km_{\tau-1}\ln(6K\tau^2)} \\
&+ 4\sqrt{2}K\ln(6K\tau^2)|\mathcal{F}] \\
&+ \mathrm{GAP}_{\max} + \mathbb{E}[m_{\tau-1}\tilde{\mathcal{R}}_{m_{\tau-1}}(\text{MV-UCB})] \\
&+ L + \mathrm{MV}_*
\end{aligned}$$

$$
\begin{aligned}
&< -(\Delta_0^{\mathrm{MV}} + \alpha \mathrm{MV}_0)\mathbb{E}[m_{\tau-1}+1] \\
&\quad + 2(5+\rho)\cdot \\
&\quad \sqrt{2K\mathbb{E}[m_{\tau-1}+1]\ln(6K(\mathbb{E}[N_0(\tau-1)]\mathbb{E}[m_{\tau-1}+1]))} \\
&\quad + 8\sqrt{2}K\ln(6K(\mathbb{E}[N_0(\tau-1)]\mathbb{E}[m_{\tau-1}+1])) \\
&\quad + 12(5+\rho)^2(H_1^{\mathrm{MV}} + 4H_2^{\mathrm{MV}})\ln(6K\mathbb{E}[m_{\tau-1}]) \\
&\quad + 288(5+\rho)^4 H_3^{\mathrm{MV}} \frac{\ln^2(6K\mathbb{E}[m_{\tau-1}])}{\mathbb{E}[m_{\tau-1}]} \\
&\quad + 9K + K\Delta_{\max}^{\mathrm{MV}} + \mathrm{GAP}_{\max} + L + \mathrm{MV}_* \\
&< -(\Delta_0^{\mathrm{MV}} + \alpha \mathrm{MV}_0)\mathbb{E}[m_{\tau-1}+1] \\
&\quad + 2(5+\rho)\cdot \\
&\quad \sqrt{2K}\sqrt{\mathbb{E}[m_{\tau-1}+1]}\ln(6K\mathbb{E}[N_0(\tau-1)]\mathbb{E}[m_{\tau-1}+1]) \\
&\quad +(8\sqrt{2}K+12(5+\rho)^2(H_1^{\mathrm{MV}}+4H_2^{\mathrm{MV}}))\cdot \\
&\quad \ln(6K\mathbb{E}[N_0(\tau-1)]\mathbb{E}[m_{\tau-1}])+864(5+\rho)^4 K H_3^{\mathrm{MV}} \\
&\quad + (13+3\rho)K
\end{aligned}
$$

Let $m = \mathbb{E}[m_{\tau-1}+1] \geq 2$, $c_1 = 2(5+\rho)\sqrt{2K}$, $c_2 = 6K\mathbb{E}[N_0(\tau-1)]$, $c_3 = \Delta_0^{\mathrm{MV}} + \alpha\mathrm{MV}_0 \in (2,\rho)$, $c_4 = 8\sqrt{2}K+12(5+\rho)^2(H_1^{\mathrm{MV}}+4H_2^{\mathrm{MV}}) > 8\sqrt{2}K+12(5+\rho)(K-1)+48(K-1)$ where $\rho > \frac{2}{\alpha\mu_0} > 2$, $c_4 > 3c_1$, $c_4 > 12c_3$. The RHS of the above inequality can be written as a constant term plus

$$g_2(m) = -c_3 m + c_1\sqrt{m}\ln(c_2 m) + c_4\ln(m).$$

According to Lemma 2, we have

$$g_2(m) \leq \frac{48c_1c_4}{c_3}\left[\ln(\frac{3\sqrt{c_2}c_4}{c_3})\right]^2.$$

Then, we have

$$
\begin{aligned}
&(\alpha\mathrm{MV}_0 - 2)\mathbb{E}[N_0(\tau-1)] \\
\leq &\frac{48\cdot 2(5+\rho)\sqrt{2K}(8\sqrt{2}K+12(5+\rho)^2(H_1^{\mathrm{MV}}+4H_2^{\mathrm{MV}}))}{\Delta_0^{\mathrm{MV}}+\alpha\mathrm{MV}_0}\cdot \\
&\Bigg[\ln\Bigg(3\sqrt{6K\mathbb{E}[N_0(\tau-1)]}\cdot \\
&\left(8\sqrt{2}K + 12(5+\rho)^2(H_1^{\mathrm{MV}}+4H_2^{\mathrm{MV}})\right)\Bigg) \\
&- \ln(\Delta_0^{\mathrm{MV}}+\alpha\mathrm{MV}_0)\Bigg]^2 + 864(5+\rho)^4 K H_3^{\mathrm{MV}} \\
&+ (13+3\rho)K \\
\leq &\Bigg(\frac{1536(5+\rho)K\sqrt{K} + 1630(5+\rho)^3\sqrt{K}(H_1^{\mathrm{MV}}+4H_2^{\mathrm{MV}})}{\Delta_0^{\mathrm{MV}}+\alpha\mathrm{MV}_0} \\
&+ 864(5+\rho)^4 K H_3^{\mathrm{MV}} + (13+3\rho)K\Bigg)\cdot\Bigg[\ln\Bigg(3\sqrt{6K}\cdot \\
&\left(8\sqrt{2}K+12(5+\rho)^2(H_1^{\mathrm{MV}}+4H_2^{\mathrm{MV}})\right)\sqrt{\mathbb{E}[N_0(\tau-1)]}\Bigg) \\
&- \ln(\Delta_0^{\mathrm{MV}}+\alpha\mathrm{MV}_0)\Bigg]^2.
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
&\mathbb{E}[N_0(\tau-1)] \\
\leq &\Bigg(\frac{1536(5+\rho)K\sqrt{K}+1630(5+\rho)^3\sqrt{K}(H_1^{\mathrm{MV}}+4H_2^{\mathrm{MV}})}{(\alpha\mathrm{MV}_0-2)(\Delta_0^{\mathrm{MV}}+\alpha\mathrm{MV}_0)} \\
&+ \frac{864(5+\rho)^4 K H_3^{\mathrm{MV}} + (13+3\rho)K}{\alpha\mathrm{MV}_0 - 2}\Bigg)\cdot \\
&\Bigg[\ln\Bigg(3\sqrt{6K}\left(8\sqrt{2}K+12(5+\rho)^2(H_1^{\mathrm{MV}}+4H_2^{\mathrm{MV}})\right)\cdot \\
&\sqrt{\mathbb{E}[N_0(\tau-1)]}\Bigg) - \ln(\Delta_0^{\mathrm{MV}}+\alpha\mathrm{MV}_0)\Bigg]^2, \\
&\sqrt{\mathbb{E}[N_0(\tau-1)]} \\
\leq &\Bigg(\frac{1536(5+\rho)K\sqrt{K}+1630(5+\rho)^3\sqrt{K}(H_1^{\mathrm{MV}}+4H_2^{\mathrm{MV}})}{(\alpha\mathrm{MV}_0-2)(\Delta_0^{\mathrm{MV}}+\alpha\mathrm{MV}_0)} \\
&+ \frac{864(5+\rho)^4 K H_3^{\mathrm{MV}} + (13+3\rho)K}{\alpha\mathrm{MV}_0 - 2}\Bigg)^{\frac{1}{2}}\cdot \\
&\Bigg[\ln\Bigg(3\sqrt{6K}\left(8\sqrt{2}K+12(5+\rho)^2(H_1^{\mathrm{MV}}+4H_2^{\mathrm{MV}})\right)\cdot \\
&\sqrt{\mathbb{E}[N_0(\tau-1)]}\Bigg) - \ln(\Delta_0^{\mathrm{MV}}+\alpha\mathrm{MV}_0)\Bigg].
\end{aligned}
$$

According to Fact 2 with
$z = \sqrt{\mathbb{E}[N_0(\tau-1)]}$,
$c_1 = (\frac{1536(5+\rho)K\sqrt{K}+1630(5+\rho)^3\sqrt{K}(H_1^{\mathrm{MV}}+4H_2^{\mathrm{MV}})}{(\alpha\mathrm{MV}_0-2)(\Delta_0^{\mathrm{MV}}+\alpha\mathrm{MV}_0)} + \frac{864(5+\rho)^4 K H_3^{\mathrm{MV}}+(13+3\rho)K}{\alpha\mathrm{MV}_0-2})^{\frac{1}{2}}$ and
$c_2 = \frac{3\sqrt{6K}(8\sqrt{2}K+12(5+\rho)^2(H_1^{\mathrm{MV}}+4H_2^{\mathrm{MV}}))}{\Delta_0^{\mathrm{MV}}+\alpha\mathrm{MV}_0}$,

$$
\begin{aligned}
&\sqrt{\mathbb{E}[N_0(\tau-1)]} \\
\leq &2\Bigg(\frac{1536(5+\rho)K\sqrt{K}+1630(5+\rho)^3\sqrt{K}(H_1^{\mathrm{MV}}+4H_2^{\mathrm{MV}})}{(\alpha\mathrm{MV}_0-2)(\Delta_0^{\mathrm{MV}}+\alpha\mathrm{MV}_0)} \\
&+ \frac{864(5+\rho)^4 K H_3^{\mathrm{MV}} + (13+3\rho)K}{\alpha\mathrm{MV}_0 - 2}\Bigg)^{\frac{1}{2}}\cdot \\
&\ln\Bigg(\Bigg(\frac{1536(5+\rho)K\sqrt{K}+1630(5+\rho)^3\sqrt{K}(H_1^{\mathrm{MV}}+4H_2^{\mathrm{MV}})}{(\alpha\mathrm{MV}_0-2)(\Delta_0^{\mathrm{MV}}+\alpha\mathrm{MV}_0)} \\
&+ \frac{864(5+\rho)^4 K H_3^{\mathrm{MV}} + (13+3\rho)K}{\alpha\mathrm{MV}_0 - 2}\Bigg)^{\frac{1}{2}}\cdot \\
&\frac{3\sqrt{6K}(8\sqrt{2}K+12(5+\rho)^2(H_1^{\mathrm{MV}}+4H_2^{\mathrm{MV}}))}{\Delta_0^{\mathrm{MV}}+\alpha\mathrm{MV}_0}\Bigg), \\
&\mathbb{E}[N_0(\tau-1)] \\
\leq &\Bigg(\frac{6144(5+\rho)K\sqrt{K}+6520(5+\rho)^3\sqrt{K}(H_1^{\mathrm{MV}}+4H_2^{\mathrm{MV}})}{(\alpha\mathrm{MV}_0-2)(\Delta_0^{\mathrm{MV}}+\alpha\mathrm{MV}_0)}
\end{aligned}
$$

$$+\frac{3456(5+\rho)^4KH_3^{\mathrm{MV}}+4(13+3\rho)K}{\alpha\mathrm{MV}_0-2}\Bigg).$$

$$\Bigg[\ln\Bigg(\Bigg(\frac{1536(5+\rho)K\sqrt{K}+1630(5+\rho)^3\sqrt{K}(H_1^{\mathrm{MV}}+4H_2^{\mathrm{MV}})}{(\alpha\mathrm{MV}_0-2)(\Delta_0^{\mathrm{MV}}+\alpha\mathrm{MV}_0)}$$

$$+\frac{864(5+\rho)^4KH_3^{\mathrm{MV}}+(13+3\rho)K}{\alpha\mathrm{MV}_0-2}\Bigg)^{\frac{1}{2}}.$$

$$\frac{3\sqrt{6K}(8\sqrt{2}K+12(5+\rho)^2(H_1^{\mathrm{MV}}+4H_2^{\mathrm{MV}}))}{\Delta_0^{\mathrm{MV}}+\alpha\mathrm{MV}_0}\Bigg)\Bigg]^2.$$

Thus,

$$\begin{aligned}
&\mathbb{E}[N_0(\tau)]\\
=&\mathbb{E}[N_0(\tau-1)]+1\\
=&O\Bigg(\frac{\rho^3\sqrt{K}(H_1^{\mathrm{MV}}+4H_2^{\mathrm{MV}})+(\rho^4KH_3^{\mathrm{MV}}+\rho K)\tilde{\Delta}_0^{\mathrm{MV}}}{(\alpha\mathrm{MV}_0-2)\tilde{\Delta}_0^{\mathrm{MV}}}\cdot\\
&\Bigg[\ln\Bigg(\frac{\rho^3\sqrt{K}(H_1^{\mathrm{MV}}+4H_2^{\mathrm{MV}})+(\rho^4KH_3^{\mathrm{MV}}+\rho K)\tilde{\Delta}_0^{\mathrm{MV}}}{(\alpha\mathrm{MV}_0-2)\tilde{\Delta}_0^{\mathrm{MV}}}\cdot\\
&\frac{K\sqrt{K}+\rho^2\sqrt{K}(H_1^{\mathrm{MV}}+4H_2^{\mathrm{MV}})}{\tilde{\Delta}_0^{\mathrm{MV}}}\Bigg)\Bigg]^2\Bigg),
\end{aligned}$$

where $\tilde{\Delta}_0^{\mathrm{MV}}=\Delta_0^{\mathrm{MV}}+\alpha\mathrm{MV}_0$.

Theorem 5 follows from $\mathbb{E}[\mathcal{R}_T(\textsf{MV-CUCB})]\leq\mathbb{E}[\mathcal{R}_T(\textsf{MV-UCB})]+\frac{\mathbb{E}[N_0(T)]}{T}\Delta_0^{\mathrm{MV}}$. □