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

Yihan Du,1 Siwei Wang,1 Longbo Huang1
1Tsinghua University
duyh18@mails.tsinghua.edu.cn, {wangsw2020, longbohuang}@tsinghua.edu.cn

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 $\delta$ in previous algorithms to provide high-probability guarantees, the $\ln(1/\delta)$-dependent regrets will boost accordingly. Instead, we focus on a certainty...
### 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.

<table>
<thead>
<tr>
<th>Problem</th>
<th>Algorithm</th>
<th>Regret bound</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMAB</td>
<td>GenCB-CMAB (ours)</td>
<td>( O\left(H \ln T + \frac{K}{\alpha} \ln \left(\frac{H}{\alpha}\right)^2\right))</td>
<td>E</td>
</tr>
<tr>
<td>CMAB</td>
<td>ConUCB (Wu et al. 2016)</td>
<td>( O\left(H \ln \left(\frac{K}{\alpha}\right) + \sum_{i=1}^{K} \max\left{\frac{\Delta_i}{\alpha}, \Delta_0 - \Delta_i\right} \ln(T/\delta)\right))</td>
<td>H</td>
</tr>
<tr>
<td>CMAB</td>
<td>Lower Bound (Wu et al. 2016)</td>
<td>( O\left(\max\left{\frac{K}{\alpha}, \sqrt{KT}\right}\right))</td>
<td>E</td>
</tr>
<tr>
<td>CLB</td>
<td>GenCB-CLB (ours)</td>
<td>( O\left(d \ln(T)\sqrt{d} + \frac{d^2}{\alpha} \ln \left(\frac{d}{\alpha}\right)^2\right))</td>
<td>E</td>
</tr>
<tr>
<td>CLB</td>
<td>CLUCB (Kazerouni et al. 2017)</td>
<td>( O\left(d \ln \left(\frac{K}{\alpha}\right)\sqrt{d} + \frac{d^2}{\alpha} \ln \left(\frac{d}{\alpha}\right)^2\right))</td>
<td>H</td>
</tr>
<tr>
<td>CLB</td>
<td>CLUCB2 (Garcelon et al. 2020)</td>
<td>( O\left(d \ln \left(\frac{K}{\alpha}\right)\sqrt{d} + \frac{d^2}{\alpha} \ln \left(\frac{d}{\alpha}\right)^2\right))</td>
<td>H</td>
</tr>
<tr>
<td>CCCB</td>
<td>GenCB-CCCB (ours)</td>
<td>( O\left(d \ln(KT)\sqrt{d} + \frac{(K+d)^2}{\alpha} \ln \left(\frac{K+d}{\alpha}\right)^2\right))</td>
<td>E</td>
</tr>
<tr>
<td>CCCB</td>
<td>C3UCB (Zhang, Li, and Liu 2019)</td>
<td>( O\left(d \ln(KT)\sqrt{d} + \frac{d}{\alpha} \sqrt{\frac{d}{K}} \ln \left(\frac{K}{\alpha} T\right)\right))</td>
<td>H</td>
</tr>
</tbody>
</table>
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; Car- doso 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 $\mu_x$. 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 $x^*_t = \argmax_{x \in \mathcal{X}} \mu_x$, denote the optimal arm. The learning performance over a time horizon $T$ is measured by expected cumulative regret

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

where $\Delta_x = \mu_{x_T} - \mu_{x^*_T}$ 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 $0 < \mu_0 < \mu_{x^*_T}$ as previous works (Wu et al. 2016; Kazeroni et al. 2017; Zhang, Li, and Liu 2019) do.

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 at least $1-\alpha$ fraction of the reward of always playing $x_0$, i.e.,

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

Here $\alpha$ controls the strictness of the constraint, i.e., how conservative we want the learner 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; Kazeroni 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 $\mu_x$ is an arbitrary value.\footnote{This assumption can be relaxed to that $x_0$ has a random reward within a known interval $[r_0, r_0]$ ($r_0 > 0$) by slightly changing the right-hand-side of the $\alpha$ statements in our algorithms, and our analysis procedure still works. While previous works can remove this assumption by estimating $\mu_0$, this is due to that their constraints are imposed on the expected reward.} 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 $\mu_x = x^T \theta^*$, 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 $x \in \mathcal{X}$, $\mu_x$ 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 $A_S$ into a conservative bandit algorithm. In the algorithm, $m$ denotes the time horizon of $A_S$, 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

$$[K] \overset{def}{=} \{1, \ldots, K\}.$$
Next, we prove the regret bound of \( \text{GenCB} \). 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; Kazerooni et al. 2017; 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 \( \text{GenCB} \) as follows.

**Theorem 1.** Given a standard bandit problem and algorithm \( A_S \), with regret \( \mathbb{E}[R_T(A_S)] \leq B(T) \), \( \text{GenCB} \) (Algorithm 1) guarantees the sample-path reward constraint Eq. (2) and achieves a regret bound

\[
\mathbb{E}[R_T(\text{GenCB})] \leq B(T) + C\Delta_0,
\]

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 \( \text{GenCB} \). We use \( S_t \) 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 \( \tau \) denote the last timestep we play \( x_0 \), i.e., \( \tau \) is the last timestep such that \( r_s(\tau - 1) + N_0(\tau - 1)\mu_0 < (1 - \alpha)\mu_0\tau \). Rearranging the terms, and subtracting \((1 - \alpha)\mu_0N_0(\tau - 1)\) from both sides (note that \( \tau = N_0(\tau - 1) + m_{\tau - 1} + 1 \)), we have

\[
\alpha\mu_0N_0(\tau - 1) < (1 - \alpha)\mu_0(m_{\tau - 1} + 1) - r_s(\tau - 1)
\]

\[
= (1 - \alpha)\mu_0(m_{\tau - 1} + 1) + \sum_{t \in S_{\tau - 1}} (\mu_{x_t} - r_{t,x_t}) - \sum_{t \in S_{\tau - 1}} \mu_{x_t}.
\]

\[
\sum_{t \in S_{\tau - 1}} (\mu_{x_t} - r_{t,x_t})
\]

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 \( F\sqrt{m_{\tau - 1}\ln(1/\delta)} \). Taking expectation on both sides of Eq. (3), setting \( m = \mathbb{E}[m_{\tau - 1} + 1] \) and replacing \( \mathbb{E}[\sum_{t \in S_{\tau - 1}} \mu_{x_t}] \) with \( \mu_x \mathbb{E}[\sum_{t \in S_{\tau - 1}} (A_s)] \), we have

\[
\alpha\mu_0\mathbb{E}[N_0(\tau - 1)] < - (\Delta_0 + \alpha\mu_0)m + \mathbb{E}[B(m_{\tau - 1})]
\]

\[
+ (F\sqrt{m_{\tau - 1}\ln(1/\delta)} + 1)
\]

\[
\leq (\Delta_0 + \alpha\mu_0)m + B(m) + 2
\]

\[
+ F\sqrt{m\ln(\mathbb{E}[N_0(\tau - 1)] + m)}.
\]

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 \( G\ln(\sqrt{\mathbb{E}[N_0(\tau - 1)]} + m) \) where \( G \) is a constant factor that only depends on problem parameters. Then, we obtain \( \mathbb{E}[N_0(\tau - 1)] \leq \frac{\Delta_0}{\alpha\mu_0} [\ln(\frac{\Delta_0}{\alpha\mu_0})]^2 \). Thus, \( \mathbb{E}[N_0(T)] = \mathbb{E}[N_0(\tau - 1)] = \mathbb{E}[N_0(\tau - 1)] + 1 \leq C \), where \( C \triangleq \frac{\Delta_0}{\alpha\mu_0} [\ln(\frac{\Delta_0}{\alpha\mu_0})]^2 + 1 \) is independent of \( T \).

Combining the regrets for \( A_S \) and \( x_0 \), we obtain that \( \mathbb{E}[R_T(\text{GenCB})] \leq B(T) + C\Delta_0 \).

\[ \square \]

**Remark 1.** Theorem 1 shows that \( \text{GenCB} \) 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; Kazerooni et al. 2017; Garcelon et al. 2020; Zhang, Li, and Liu 2019)) with \( T \)-independent conservative regret.

Below, we apply \( \text{GenCB} \) to three widely studied CBPs, i.e., CMAB, CLB and CCCB. Here we only present the main theorems, and defer the algorithm pseudo-codes and proofs to the supplementary material (Du, Wang, and Huang 2020).

### 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 conservative exploration (Wu et al. 2016), which has extensive

---

**Algorithm 1: General Solution to Conservative Bandits (GenCB)

**Input:** Standard bandit problem and algorithm \( A_S \), regular arms \( X \), default arm \( x_0 \) with reward \( \mu_0 \), parameter \( \alpha \).

1. For each timestep \( t \), do
2.   For \( t = 1, 2, \ldots \) do
3.     if \( r_s(t - 1) + N_0(t - 1)\mu_0 \geq (1 - \alpha)\mu_0t \) then
4.       \( m \leftarrow m + 1; \)
5.       Play an arm \( x_t \) according to \( A_S \), observe \( r_{t,x_t} \) and update the statistical information;
6.       \( N_0(t) \leftarrow N_0(t - 1); \)
7.       \( r_s(t) \leftarrow r_s(t - 1) + r_{t,x_t}; \)
8.     else
9.       Play \( x_0 \) and receive reward \( \mu_0; \)
10. \( N_0(t) \leftarrow N_0(t - 1) + 1; \)
11. \( r_s(t) \leftarrow r_s(t - 1); \)

---
applications including clinical trials, online advertising and wireless network. In CMAB, $\mathcal{X} = [K]$ and $\mu_i (1 \leq i \leq K)$ can be an arbitrary value. Without lose of generality, we assume $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_K$ and denote $\mu_* = \mu_1$.

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 $x_t \leftarrow \arg\max_{x \in [K]} \left( \hat{\mu}_i + \sqrt{2 \ln m / N_i(t-1)} \right)$, 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

$$O\left( H \ln T + \frac{H \Delta_0}{\alpha_0 (\Delta_0 + \alpha_0)} \left[ \ln \left( \frac{H}{\alpha_0 (\Delta_0 + \alpha_0)} \right) \right]^2 \right),$$

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 $H$ (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 contextual combinatorial bandit problem under the safe exploration requirement. In CLB, $\mathcal{X}$ 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) $A_t \subset \mathcal{X}$ or the default arm $x_0$ at each timestep. The expected reward of base arm $x_e$ is $w^*_e = x^T_0 \theta^*$ and that of super arm $A$ is $f(A, w^*)$, where $\theta^*$ is an unknown parameter and $f$ satisfies two mild assumptions, i.e., monotonicity and Lipschitz continuous with parameter $\beta$ (Qin, Chen, and Zhu 2019). Similar to CLB, we assume $\|x\|_2 \leq L, \forall x \in \mathcal{X}$ and $\|\theta^*\|_2 \leq \beta$. At timestep $t$, the random reward of a base arm $x_e$ and a super arm $A$ are $w_{x,e} = w^* + \eta_{x,e} \in [0,1]$ and $r_{A,e} = f(A, w^*) + \eta_{A,e} \in [0,1]$, respectively. After pulling super arm $A_t$, we receive the random reward $r_{t,A_t}$ and observe a semi-bandit feedback, i.e., $u_{t,c}$ for each $c \in A_t$.

For CLB, we apply GenCB with the LinUCB algorithm (Abbasi-yadkori, Pál, and Szepesvári 2011) by replacing Line 5 in Algorithm 1 with $A_t \leftarrow \arg\max_{A \in \mathcal{X}} f(A, \tilde{w}_t)$. Here $\tilde{w}_{t,c} = x^T_0 \tilde{\theta}_{t-1} + (\sqrt{d \ln (2m^2 (1 + m KL^2 / \lambda))} + \sqrt{\lambda}) s^T x_t$ is a confidence ellipsoid that contains $\theta^*$ with high probability, and we define $\tilde{\theta}_{t-1} = V_{t-1}^{-1} b_t, V_t = \lambda I + \sum_{s=1}^{t} x_s x_s^T, b_t = \sum_{s=1}^{t} x_s x_s$, and $\lambda = \max\{1, L^2\}$. We name this version of the algorithm GenCB-CLB, whose key idea is to play a regular arm according to the principle 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

$$O \left( d \ln \left( \frac{LT}{\lambda} \right) \sqrt{T} + \frac{d^2 S^2 \lambda \Delta_0}{\alpha_0 (\Delta_0 + \alpha_0)} \left[ \ln \left( \frac{d \sqrt{\lambda}}{\alpha_0 (\Delta_0 + \alpha_0)} \right) \right]^2 \right),$$

where $\Delta_0 = \Delta_0 + \alpha_0$.

**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 CLUB2 (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 compact subset of $[0,1]^d$ and generated from certain combinatorial structure. The learner plays $\theta^*$ which satisfies two mild assumptions, i.e., monotonicity and Lipschitz continuous with parameter $\beta$ (Qin, Chen, and Zhu 2019). Similar to CLB, we assume $\|x\|_2 \leq L, \forall x \in \mathcal{X}$ and $\|\theta^*\|_2 \leq \beta$. At timestep $t$, the random reward of a base arm $x_e$ and a super arm $A$ are $w_{x,e} = w^* + \eta_{x,e} \in [0,1]$ and $r_{A,e} = f(A, w^*) + \eta_{A,e} \in [0,1]$, respectively. After pulling super arm $A_t$, we receive the random reward $r_{t,A_t}$ and observe a semi-bandit feedback, i.e., $u_{t,c}$ for each $c \in A_t$.

For CCCB, we apply GenCB with the C2UCB algorithm (Qin, Chen, and Zhu 2014), by replacing Line 5 in Algorithm 1 with $A_t \leftarrow \arg\max_{A \in \mathcal{X}} f(A, \tilde{w}_t)$. Here $\tilde{w}_{t,c} = x^T_0 \tilde{\theta}_{t-1} + (\sqrt{d \ln (2m^2 (1 + m KL^2 / \lambda))} + \sqrt{\lambda}) s^T x_t$ is the upper confidence bound of $w^*_{t,c}$, and we define $\tilde{\theta}_{t-1} = V_{t-1}^{-1} b_t, V_t = \lambda I + \sum_{s=1}^{t} x_s x_s^T, b_t = \sum_{s=1}^{t} x_s x_s$, and $\lambda = \max\{1, L^2\}$. 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-
We now extend CBPs to the mean-variance (Sani, Lazaric, and Munos 2012). Each arm \( x \) is associated with a measure mean-variance, which is formally defined as \( \text{MV}_x = \rho \mu_x - \sigma^2_x \), where \( \sigma^2_x \) is the reward variance and \( \rho \) is a weight parameter. Let \( x^\text{MV} \in \arg \max_{x \in [K]} \text{MV}_x \) 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 \( \text{MV}_{x,t} = \hat{\rho} \hat{\mu}_{x,t} - \hat{\sigma}^2_{x,t} \), where \( \hat{\rho}_{x,t} = \frac{1}{t} \sum_{s=1}^t Z_{x,s} \) and \( \hat{\sigma}^2_{x,t} = \frac{1}{t} \sum_{s=1}^t (Z_{x,s} - \hat{\mu}_{x,t})^2 \).

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

Due to the difficulty of the \( R_T^\text{MV}(A) \) metric, we follow the mean-variance bandit literature and use a more tractable measure mean-variance pseudo-regret (Sani, Lazaric, and Munos 2012) defined as:

\[
\tilde{R}_T^\text{MV}(A) = \frac{1}{T} \sum_{x \neq x^\text{MV}} N_{x,T} \Delta_x^\text{MV} + \frac{2}{T^2} \sum_{x \neq y \neq x} N_{x,T} N_{y,T} \Gamma_{x,y}^2,
\]

where \( N_{x,T} \) is a shorthand for \( N_x(T) \), \( \Delta_x^\text{MV} = \bar{\text{MV}}_{x^\text{MV}} - \bar{\text{MV}}_x \), and \( \Gamma_{x,y}^2 = \left( \mu_x - \mu_y \right)^2 \). It has been shown that any bound on \( R_T^\text{MV}(A) \) immediately translates into a bound on \( \tilde{R}_T^\text{MV}(A) \) (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 \( \tilde{R}_T^\text{MV}(A) \). Note that, in MV-CBP the measures \( \bar{\text{MV}}_T(A) \) and \( R_T^\text{MV}(A) \) are both normalized quantities over \( T \).

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

\[
\text{MV}_x(A) \geq (1 - \alpha) \text{MV}_0, \quad \forall t \in \{1, \ldots, T\}.
\]

Here \( \text{MV}_0 \) denotes the mean-variance of our default arm \( x_0 \) with known constant reward \( \mu_0 \) 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 (see the supplementary material (Du, Wang, and Huang 2020) for its proof).

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

\[
\bar{O}\left( \frac{\rho^2 \ln(KT)}{T} \left( H_1 + H_2 + \frac{\rho^2 \ln(KT)}{T} H_3 \right) \right)
\]

\[
+ \rho^3 \sqrt{K} (H_1^\text{MV} + A_2^\text{MV}) + (\rho^4 K H_3^\text{MV} + \rho) A_0^\text{MV} \Delta_0^\text{MV} (\alpha \text{MV}_0 - 2) \Delta_0^\text{MV},
\]

where

\[
H_1^\text{MV} = \sum_{i,j} \Delta_{ij}^\text{MV}^{-1} - 1, \quad H_2^\text{MV} = \sum_{i,j} \Delta_{ij}^\text{MV}^{-2}, \quad H_3^\text{MV} = \sum_{i,j} j > 1, i \neq j \Delta_{ij}^\text{MV}^{-2} - 2 \Delta_0^\text{MV} \Delta_0^\text{MV} + \alpha \text{MV}_0.
\]

---

4. Omit 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 $\alpha$MV$_0 > 2$ to guarantee that a pull of $x_0$ will not violate the constraint. 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, only partial results are presented here (see the supplementary material (Du, Wang, and Huang 2020) for full results).

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 $f(A, w^*) = \sum_{a \in A} w_{a}^*$. 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 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 $(1-\alpha)$MV$_0$, 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.

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 $T \cdot \bar{R}_{MV}^T$ 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 $(1-\alpha)$MV$_0T$. 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 ones both theoretically (incurs $T$-independent conservative regrets rather than $T$-dependent) and empirically. Moreover, we study a novel extension of CBPs 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.
Acknowledgements
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.

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.

References


