Constrained regret minimization for multi-criterion multi-armed bandits

Anmol Kagrecha[1] · Jayakrishnan Nair[2] · Krishna Jagannathan[3]

Abstract

We consider a stochastic multi-armed bandit setting and study the problem of constrained regret minimization over a given time horizon. Each arm is associated with an unknown, possibly multi-dimensional distribution, and the merit of an arm is determined by several, possibly conflicting attributes. The aim is to optimize a 'primary' attribute subject to user-provided constraints on other 'secondary' attributes. We assume that the attributes can be estimated using samples from the arms' distributions, and that the estimators enjoy suitable concentration properties. We propose an algorithm called Con-LCB that guarantees a logarithmic regret, i.e., the average number of plays of all non-optimal arms is at most logarithmic in the horizon. The algorithm also outputs a boolean flag that correctly identifies, with high probability, whether the given instance is feasible/infeasible with respect to the constraints. We also show that Con-LCB is optimal within a universal constant, i.e., that more sophisticated algorithms cannot do much better universally. Finally, we establish a fundamental trade-off between regret minimization and feasibility identification. Our framework finds natural applications, for instance, in financial portfolio optimization, where *risk constrained* maximization of expected return is meaningful.

✉ Anmol Kagrecha
akagrecha@gmail.com

Jayakrishnan Nair
jayakrishnan.nair@ee.iitb.ac.in

Krishna Jagannathan
krishnaj@ee.iitm.ac.in

[1] Stanford University, Stanford, USA

[2] IIT Bombay, Mumbai, India

[3] IIT Madras, Chennai, India

1 Introduction

The multi-armed bandit (MAB) problem is a fundamental construct in online learning, where a learner has to quickly identify the best option (a.k.a., arm) among a given set of options. In the stochastic MAB problem, each arm is associated with an (a priori unknown) reward distribution, and a sample from this distribution is revealed each time an arm is chosen (a.k.a., pulled). The classical goal is to use these samples to quickly identify the arm with the highest mean reward. The most popular metric to evaluate the performance of a learning algorithm is *regret*, which defined as a weighted sum of the expected number of plays of suboptimal arms, with the weights being equal to the sub-optimality gaps.

While the classical stochastic MAB formulation has been applied in various application scenarios, including clinical trials, portfolio optimization, anomaly detection, and telecommunication (Bouneffouf & Rish, 2019), it ignores a key aspect of most real-world decision-making problems—namely, that *they have multiple criteria of interest*. For example, when comparing testing kits in a clinical trial, one would want to keep track of the false-positive rate as well as the false-negative rate of each kit. Similarly, choosing the best financial portfolio involves balancing risk and reward. A wireless node deciding which channel to transmit on, or what transmission rate to use has to balance several criteria, including throughput, delay, and energy consumption. This multi-criterion nature of decision making is not always adequately captured by the classical MAB approach of optimizing a one-dimensional reward signal.

The most common approach for incorporating multiple arm attributes into MAB formulations is to define the reward as a suitable function (say a linear combination) of the attributes of interest. For example, risk-aware portfolio optimization can be cast as an MAB problem where the best arm is one that optimizes a certain linear combination of mean value and a suitable risk measure (such as standard deviation or Conditional Value at Risk) (see, for example, Sani et al. (2012), Vakili and Zhao (2016), Kagrecha et al. (2019)]. However, this approach assumes that the different attributes of interest can be expressed and compared on a common scale, which is not always reasonable. For example, how does one 'equate' the impact a certain increment in risk to that of an another increment in the mean return of a portfolio?

A more natural approach for multi-criterion decision making is to instead pose the optimal choice as the solution of a *constrained optimization*. In other words, optimize one attribute subject to constraints on the others. However, despite the modest literature on multi-criterion MABs (surveyed below), little attention has been paid to formulating, and designing algorithms for *constrained* multi-criterion MABs. This paper seeks to fill this gap. Formally, we assume that each arm is associated with a D-dimensional probability distribution, and the best arm is the one that minimizes a certain arm attribute subject to constraints on m other attributes. For this setting, we pursue regret minimization, i.e., we seek to minimize the average number of pulls of non-optimal arms.

To simplify the presentation, and to highlight the key aspects of this problem, we first consider a single constraint (i.e., $m = 1$). Each arm is associated with a probability distribution, and we consider the problem of optimizing an *objective attribute*, subject to a single *constraint attribute* satisfying a user-specified constraint. The algorithm design and performance evaluation for this special case (with $m = 1$) generalize readily to the multi-criterion problem; see Sect. 5. An example that fits this template that is of significant interest in the finance community: the optimization of the expected return,

subject to a risk constraint. Here, the objective attribute is the mean of an arm distribution, while the constraint attribute measuring risk could be Conditional Value at Risk (CVaR).

For this problem, we propose an algorithm, called Constrained Lower Confidence Bound (Con-LCB), that guarantees logarithmic regret, i.e., the average number of plays of all non-optimal arms (including those that violate the constraint) is at most logarithmic in the horizon. If Con-LCB is presented with an *infeasbile* instance, i.e., an instance where all arms violate the specified risk constraint, the algorithm in effect relaxes this constraint just enough to make at least one arm compliant. Another feature of Con-LCB is that at the end of the horizon, it outputs a boolean flag that correctly identifies with high probability, whether or not the given instance was feasible.

Finally, we establish fundamental lower bounds on the performance of any algorithm on this constrained regret minimization problem. Our results demonstrate a fundamental tradeoff between regret minimization and feasibility identification, similar to the well-known tradeoff between regret minimization and best arm identification in the classical (unconstrained) MAB problem (Bubeck et al., 2009).

The remainder of this paper is organized as follows. A brief survey of related literature is provided below. In Sect. 2, we introduce some preliminaries and formulate the constrained mean minimization problem. We present our algorithm (Con-LCB) and its performance guarantees in Sect. 3. Information-theoretic lower bounds on performance are discussed in Sect. 4. Finally, the general formulation for multi-criterion MABs is introduced in Sect. 5.

Related literature The literature related to multi-armed bandit problems is quite large. We refer the reader to Bubeck and Cesa-Bianchi (2012), Lattimore and Szepesvári (2020) for a comprehensive review. Here, we restrict ourselves to papers that consider (i) multi-objective MAB problems with vector rewards, and (ii) risk-aware arm selection.

For multi-objective MAB problems with vector rewards, different notions of optimality are considered. For example, Drugan and Nowe (2013), Yahyaa and Manderick (2015) consider the notion of Pareto optimality. In these papers, all dimensions are considered equally important and the aim is to play all Pareto-optimal arms an equal number of times. Another important notion of optimality is lexicographic optimality (see Ehrgott 2005). Here there is an order of importance among different dimensions. Tekin and Turğay (2018), Tekin (2019) consider the notion of lexicographic optimality for contextual bandit problems. In this line of work, the goal is to obtain higher reward in an important dimension and for tie-breaking, use rewards obtained in dimensions of lower importance.

Turning now to the literature on risk-aware arm selection, Sani et al. (2012), Galichet et al. (2013), Vakili and Zhao (2016), David and Shimkin (2016), Bhat and Prashanth (2019), Prashanth et al. (2020), Kagrecha et al. (2019) consider the problem of optimizing a risk metric alone or consider a linear combination of mean and a risk metric. Zimin et al. (2014) looks at the learnability of general functions of mean and variance, and Maillard (2013) proposes an optimization of the logarithm of moment generating function as a risk measure in a regret minimization framework. Cassel et al. (2018) look at path dependent regret and provides a general approach to study many risk metrics.

None of the above papers considers the *constrained* MAB problem, which frames the optimal arm as the solution of a constrained optimization problem. There has been some recent work for the constrained MAB problem. A constrained linear bandit setting is considered in Pacchiano et al. (2021) under the assumption that there is at least one arm which satisfies the constraints. The papers (Amani et al., 2019; Moradipari et al., 2019) consider the problem of maximizing the reward subject to satisfying a linear constraint with high probability. Constrained setting was also considered in David et al. (2018) and Chang

(2020). Both these papers consider a single constraint on arm selection; David et al. (2018) considers a constraint on VaR, and Chang (2020) considers an average cost constraint (each arm has a cost distribution that is independent of its reward distribution). All the papers above implicitly assume that the instance presented is feasible, whereas we address the issue of encountering an infeasible instance.

Finally, we note that constraints appear in an entirely different context in the family of bandit problems called "bandits with knapsacks." Here, the goal is to maximize rewards under (vector) budget constraints, where the learning agent ceases to pull arms (and accumulate reward) once any one of the budget components is exhausted. The problem was introduced in Badanidiyuru et al. (2018) and some recent work in this space includes (Agrawal & Devanur, 2016; Sankararaman & Slivkins, 2021).

2 Problem formulation

In this section, we describe the formulation of our constrained stochastic MAB problem. To keep the exposition simple, we consider a single constraint (i.e., $m = 1$) for now; the general formulation with multiple constraints (i.e., $m \geq 1$) is discussed in Sect. 5. Informally, under the regret minimization objective considered here, the goal is to play, as often as possible, the arm that optimizes the objective, subject to a constraint on another attribute.

Formally, consider a multi-armed bandit problem with K arms, labeled $1, 2, \ldots, K$. Each arm is associated with a (possibly multi-dimensional) probability distribution, with $\nu(k)$ denoting the (joint) distribution corresponding to arm $k \in [K]$[1]. Suppose that $\nu(k) \in \mathcal{C}$, the space of possible arm distributions. The objective and the constraint attributes are defined via functions g_0 and g_1 respectively, mapping $\mathcal{C}$ to $\mathbb{R}$. Additionally, the user provides a threshold $\tau \in \mathbb{R}$ which specifies an upper bound on the attribute g_1. An *instance* of the constrainted MAB problem is defined by (ν, τ), where $\nu = (\nu(k), 1 \leq k \leq K) \in \mathcal{C}^K$ (here, $\mathcal{C}^K$ denotes the K-fold cartesian product of $\mathcal{C}$). The arms that satisfy the constraint $g_1(\nu(\cdot)) \leq \tau$ are called *feasible arms*; the set of feasible arms is denoted by $\mathcal{K}(\nu)$, and the set of arms that do not satisfy the constraint is denoted by $\mathcal{K}(\nu)^c$. The instance (ν, τ) is said to be feasible if $\mathcal{K}(\nu) \neq \emptyset$, and is said to be infeasible if $\mathcal{K}(\nu) = \emptyset$.

Consider first a feasible instance. An optimal arm in this case is defined as an arm that minimizes $g_0(\nu(\cdot))$, subject to the constraint $g_1(\nu(\cdot)) \leq \tau$. Denote the optimal value as $g_0^* = \min_{k \in \mathcal{K}(\nu)} g_0(\nu(k))$. Arms which have $g_0(\nu(\cdot))$ larger than g_0^* (whether or not feasible) are referred to as *suboptimal* arms. Note that there can also exist infeasible arms with a smaller objective g_0 than g_0^*. We refer to such arms as *deceiver arms*; the set of deceiver arms is denoted by $\mathcal{K}^d(\nu)$, where $\mathcal{K}^d(\nu) = \{k \in [K] : g_1(\nu(k)) > \tau, g_0(\nu(k)) \leq g_0^*\}$. For a suboptimal arm k, the suboptimality gap is defined by $\Delta(k) := g_0(\nu(k)) - g_0^* > 0$. (Note that the suboptimality gap is also defined for infeasible, non-deceiver arms). Finally, for an infeasible arm k, the infeasibility gap is defined as $\Delta_\tau(k) = g_1(k) - \tau > 0$. Figure 1a provides a visual representation of a feasbile instance.

Next, consider an infeasible instance. The optimal arm in this case is defined in as the one with the smallest value of $g_1(\nu(\cdot))$. Let the optimal value be denoted by $g_1^* = \min_{k \in [K]} g_1(\nu(k)) > \tau$. This is equivalent to requiring that if the algorithm is faced with an infeasible instance, it must 'relax' the constraint just enough, until at least

---

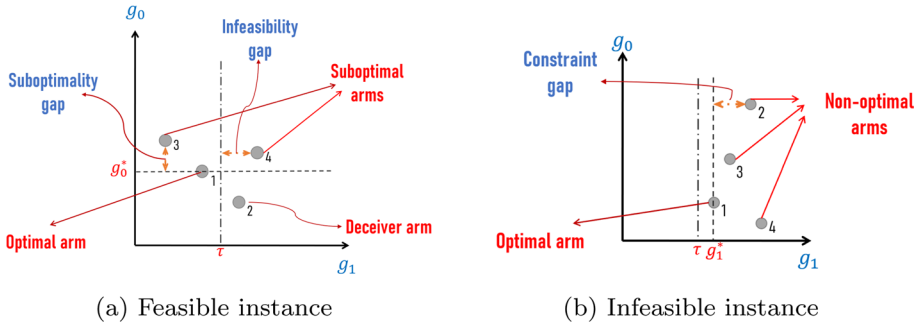[1] For a positive integer n, we denote $[n] := \{1, 2, \ldots, n\}$.

Fig. 1 The two types of instances

one arm satisfies the constraint. The *constraint gap* for an arm k that is not optimal is defined as $\Delta_{con}(k) = g_1(v(k)) - g_1^* > 0$. Figure 1b provides a visual representation of an infeasbile instance.

For any (feasible or infeasible) instance (v, τ), let the set of optimal arms be denoted by $\mathcal{K}^*(v)$. The total number of pulls (or horizon) is denoted by T. For an algorithm (a.k.a., policy) π, the number of pulls of an arm k over the first t pulls, for $t \in [T]$, is denoted by $N_k^\pi(t)$, though we often suppress the dependence on π for simplicity.

Consistency We now define the notion of *consistency* of an algorithm in this setting. A policy π is said to be *consistent* over a class of distributions $\mathcal{C}$, given a pre-specified constraint threshold τ, if for all instances (v, τ) such that $v \in \mathcal{C}^K$, it holds that $\mathbb{E}[N_k(T)] = o(T^a)$ for all $a > 0$ and for all (non-optimal) arms in $[K] \setminus \mathcal{K}^*(v)$. This definition is in line with the definition of consistency in the classical unconstrained regret minimization setting (see Lai and Robbins 1985).

Regret The formal definition of regret in the present setting is as follows. For a feasible instance, there are two types of regret: *suboptimality regret*

$$R_T^{sub} := \sum_{k \in \mathcal{K}(v) \setminus \mathcal{K}^*(v)} \Delta(k) \mathbb{E}[N_k(T)],$$

which is the regret caused due to the sampling of feasible, suboptimal arms, and *infeasibility regret*

$$R_T^{inf} := \sum_{k \in \mathcal{K}(v)^c} \Delta_\tau(i) \mathbb{E}[N_k(T)],$$

which is the regret caused due to the sampling of infeasible arms. For an infeasible instance, regret is caused by playing arms that are farther from the constraint boundary than the optimal arm. In this case, we define the *constraint regret* as

$$R_T^{con} := \sum_{k \in [K] \setminus \mathcal{K}^*(v)} \Delta_{con}(k) \mathbb{E}[N_k(T)].$$

(Note that our analysis essentially provides upper and lower bounds on $\mathbb{E}[N_k(T)]$ for all arms in $[K] \setminus \mathcal{K}^*(v)$. So alternative definitions of regret, involving linear combinations of the expected number of pulls of non-optimal arms, are also supported by our analysis.)

Infeasibility identification Next, we introduce an optional boolean flag called `feasibility_flag`, that the policy may set at the end of T plays, to indicate post facto whether it considered the instance as feasible (by setting the flag as `true`) or infeasible (by setting the flag as `false`). For the algorithms proposed in this paper, we provide bounds on the probability that this flag erroneously flags a feasible instance as infeasible and vice-versa. We also provide fundamental lower bounds on the probability of error under any consistent policy.

Concentration inequalities on attribute estimators Natually, MAB algorithms must estimate the attributes g_0 and g_1 corresponding to each arm using the data samples obtained by pulling that arm. We make the following assumption on concentration properties of these estimators. Suppose that for $i \in \{0, 1\}$, and distribution $F \in \mathcal{C}$, there exists an estimator $\hat{g}_{i,n}(F)$ of $g_i(F)$ using n i.i.d. samples from F, satisfying the following concentration inequality: There exists $a_i > 0$ such that for all $\Delta > 0$,

$$\Pr\left(\left|\hat{g}_{i,n}(F) - g_i(F)\right| \geq \Delta\right) \leq 2\exp(-a_i n \Delta^2). \tag{1}$$

Concentration inequalities of this form are available in a broad variety of settings. For example, if $g_i(F) = \mathbb{E}\left[h_i(X)\right]$, where X is a random vector distributed as F, then a concentration inequality of the form (1) is readily obtained if $h_i(X)$ is bounded (using the Hoeffding inequality), or σ-subGaussian. Similarly, if h_i is Lipschitz and X is a subGaussian random vector, concentration bounds of the form (1) can be obtained by invoking the results in Kontorovich (2014). Several examples where risk measures can be concentrated in this manner are provided in Cassel et al. (2018). Also, note that the specific form (1) of the concentration inequality is only assumed to simplify the description of our algorithms. Alternative forms of concentration guarantees (such as those known for the means of subexponential or heavy-tailed distributions) can also be supported by our algorithmic framework via trivial modifications to the confidence bounds.

3 Constrained regret minimization

In this section, we present an algorithm for constrained regret minimization, and provide performance guarantees for the same, assuming that we have estimators for each attribute that satisfy the concentration inequality (1).

The algorithm, which we refer to as *constrained lower confidence bound* (Con-LCB) algorithm (formal description presented as Algorithm 1) is based on the well-known principle of *optimism under uncertainty*. Con-LCB uses lower confidence bounds (LCBs) on attribute g_1 of each arm to maintain a set of *plausibly feasible* arms, and uses LCBs for attribute g_0 of the arms in this set to select the arm to be played. Note that LCBs are used for attribute g_0 because we are dealing with *minimization* of g_0. If, at some instant, the set of plausibly feasible arms maintained by Con-LCB becomes empty, the algorithm turns conservative and plays the arm which violates the constraint least, i.e., the one with the smallest LCB on g_1. Finally, at the end of T rounds, Con-LCB sets the feasiblity flag as `true` if the set of plausibly feasible arms is found to be non-empty, and `false` otherwise.

Algorithm 1 CON-LCB

---

procedure CON-LCB(T, K, τ)
    Play each arm once
    for $t = K + 1, \cdots, T$ **do**
        Set $\hat{\mathcal{K}}_t = \left\{ k : \hat{g}_{1, N_k(t-1)}(k) - \sqrt{\frac{\log(2T^2)}{a_1 N_k(t-1)}} \leq \tau \right\}$
        if $\hat{\mathcal{K}}_t \neq \varnothing$ **then**
            Set $\text{L}_0(k) = \hat{g}_{0, N_k(t-1)}(k) - \sqrt{\frac{\log(2T^2)}{a_0 N_k(t-1)}}$
            Play arm $k_t^\dagger \in \arg\min_{k \in \hat{\mathcal{K}}_t} \text{L}_0(k)$
        else
            Set $\text{L}_1(k) = \hat{g}_{1, N_k(t-1)}(k) - \sqrt{\frac{\log(2T^2)}{a_1 N_k(t-1)}}$
            Play arm $k_t^\dagger \in \arg\min_{k \in [K]} \text{L}_1(k)$
        end if
    end for
    if $\hat{\mathcal{K}}_T \neq \varnothing$ **then**
        Set `feasibility_flag = true`
    else
        Set `feasibility_flag = false`
    end if
end procedure

The remainder of this section is devoted to performance guarantees for CON-LCB. We consider feasible and infeasible instances separately.

Theorem 1 *Consider a feasible instance. Under CON-LCB, the expected number of pulls of a feasible but suboptimal arm k (i.e., satisfying $g_0(v(k)) > g_0^*$ and $g_1(v(k)) \leq \tau$), is bounded by*

$$\mathbb{E}[N_k(T)] \leq \frac{4 \log(2T^2)}{a_0 \Delta^2(k)} + 5.$$

The expected number of pulls of a deceiver arm k (i.e., satisfying $g_0(v(k)) \leq g_0^*$ and $g_1(v(k)) > \tau$ is bounded by

$$\mathbb{E}[N_k(T)] \leq \left(\frac{4 \log(2T^2)}{a_1 [\Delta_\tau(k)]^2} \right) + 2.$$

The expected number of pulls of an arm k which is an infeasible non-deceiver (i.e., satisfying $g_0(v(k)) > g_0^*$ and $g_1(v(k)) > \tau$) is bounded by

$$\mathbb{E}[N_k(T)] \leq \min \left(\frac{4 \log(2T^2)}{a_0 \Delta^2(k)}, \frac{4 \log(2T^2)}{a_1 [\Delta_\tau(k)]^2} \right) + 5.$$

The probability of incorrectly setting the `feasibility_flag` *is upper bounded by*

$$\Pr(\texttt{feasibility\_flag} = \texttt{false}) \leq \frac{1}{T}.$$

The main takeaways from Theorem 1 are as follows.

• The upper bound on the expected number of pulls for feasible, suboptimal arms is logarithmic in the horizon T, and inversely proportional to the square of the suboptimality gap. This is similar to bounds known for classical (unconstrained) MABs.

• For deceiver arms, the upper bound on the expected number of pulls, also logarithmic in T, is inversely proportional to the square of the feasiblity gap. This is similar to the bound one would obtain in a pure g_1 minimization problem for a hypothetical instance consisting of all the originally infeasible arms, and a single hypothetical (optimal) arm having g_1 equal to τ.

• The bound on the expected number of pulls of non-deceiver, infeasible arms involves a minimum of the dominant terms in the above two cases. Intuitively, this is because these arms can disambiguated in two ways: via the suboptimality gap, and the feasibility gap.

• The probability that the feasiblity flag incorrectly identifies the instance as infeasible is upper bounded by a power law in the horizon T. Note that the specific form of the probability of mis-identification bound is not fundamental; a small modification in the algorithm would make this bound a faster decaying power law at the expense of a multiplicative bump in regret. However, that this probability is not *much* smaller (for example, exponentially decaying in the horizon T) is a consequence of an inherent tension between regret minimization and feasiblity identification.

A similar tension is known to exist between regret minimization and best arm identification in the unconstrained setting; see Bubeck et al. (2009)). We provide lower bounds on the probability that any consistent algorithm makes a mistake in feasibility identification in Sect. 4.

Finally, we note that the suboptimality regret, as well as the infeasibility regret are logarithmic in the horizon T. Indeed, the suboptimality regret for a feasible instance is bounded as

$$R_T^{\text{sub}} \leq \sum_{k \in \mathcal{K}(v) \setminus \mathcal{K}^*(v)} \left(\frac{4 \log(2T^2)}{a_0 \Delta(k)} + 5\Delta(k) \right),$$

which is similar to regret bounds in the classical unconstrained setting. To express the infeasibility regret compactly, let us interpret $\Delta(k)$ as 0 (and $1/\Delta(k)$ as ∞) for deceiver arms. With this notation, the infeasibility regret of a feasible instance is bounded as

$$R_T^{\text{inf}} \leq \sum_{k \in \mathcal{K}(v)^c} 5\Delta_\tau(k) + \min\left(\frac{4\Delta_\tau(k) \log(2T^2)}{a_0 \Delta^2(k)}, \frac{4 \log(2T^2)}{a_1 \Delta_\tau(k)} \right).$$

Next, we move on to characterizing the performance of Con-LCB over infeasible instances.

Theorem 2 *Consider an infeasible instance. Under Con-LCB, the expected number of pulls of a non-optimal arm k is bounded by*

$$\mathbb{E}[N_k(T)] \leq \left(\frac{4 \log(2T^2)}{a_1 [\Delta_{\text{con}}(k)]^2} \right) + K + 2.$$

Moreover, the probability that the algorithm incorrectly flags the instance as feasible is bounded as $\Pr(\texttt{feasibility\_flag} = \texttt{true}) \leq \frac{K}{T}$ *for* $T > T^*(v)$*, where* $T^*(v)$ *is an instance-dependent constant.*

For an infeasible instance, the upper bound on the expected number $\mathcal{G}$ pulls of a non-optimal arm, logarithmic in the horizon, and inversely proportional to the square of the constraint gaps $\Delta_{\mathrm{con}}(k)$, is structurally similar to the bound one would obtain in pure g_1 minimization problem on the same instance. However, note that when faced with an infeasible instance, CON-LCB would only start playing the optimal arm regularly after *all K* arms appear to be infeasible; this explains the appearance of *K* in our bounds.

Here, the constraint regret is bounded as

$$R_T^{\mathrm{con}} := \sum_{k \in [K] \setminus \mathcal{K}^*(v)} (K+2)\Delta_{\mathrm{con}}(k) + \frac{4\log(2T^2)}{a_1 \Delta_{\mathrm{con}}(k)}$$

Finally, as before, the probability that the feasibility flag wrongly identifies the infeasible instance as feasible decays as a power law in the horizon for $T > T^*(v)$; the threshold T^* accounts for the time it takes for the algorithm to 'detect' that the instance is infeasibile with high probability.

4 Information theoretic lower bounds

In this section, we establish fundamental limits on the performance of algorithms for constrained regret minimization. First, we show that the regret bounds obtained for CON-LCB are asymptotically tight upto universal multiplicative constants (on a class of Gaussian bandit instances). We then prove a lower bound on the probability that any consistent algorithm misidentifies a feasible instance as infeasible or vice-versa. This result illustrates an inherent tension between regret minimization and feasibility identification—consistent algorithms (recall that consistency means regret is $o(T^a)$ for all $a > 0$) cannot have a misidentification probability that decays exponentially in the horizon, and algorithms that enjoy a mid-identification probability that decays exponentially in the horizon cannot be consistent.

To state our information theoretic lower bound on regret, suppose that the class of arm distributions $\mathcal{G}$ is the class of 2-dimensional Gaussian distributions with covariance matrix $\Sigma = \mathrm{diag}(1/2a_0, 1/2a_1)$. Let attribute g_0 be the mean of the first dimension, and g_1 be the mean of the second dimension. For the assumed covariance matrix structure, the standard empirical mean estimators for g_0 and g_1 satisfy the concentration properties stated in (1).

Theorem 3 *Let π be any consistent policy over the class of distributions $\mathcal{G}$ given the threshold $\tau \in \mathbb{R}$. Consider a feasible instance (v^f, τ), where $v^f \in \mathcal{G}^K$. For any feasible but suboptimal arm k,*

$$\liminf_{T \to \infty} \frac{\mathbb{E}\left[N_k^{\pi}(T)\right]}{\log(T)} \geq \frac{1}{a_0 \Delta^2(k)}.$$

For any deceiver arm k,

$$\liminf_{T \to \infty} \frac{\mathbb{E}\left[N_k^{\pi}(T)\right]}{\log(T)} \geq \frac{1}{a_1 [\Delta_{\tau}(k)]^2}.$$

Finally, for any infeasible non-deceiver arm k,

$$\liminf_{T \to \infty} \frac{\mathbb{E}\left[N_k^{\pi}(T)\right]}{\log(T)} \geq \frac{1}{2} \min \left(\frac{1}{a_0 \Delta^2(k)}, \frac{1}{a_1 [\Delta_{\tau}(k)]^2} \right).$$

Similarly, for an infeasible instance (v^i, τ), such that $v^i \in \mathcal{G}^K$, for any non-optimal arm k,

$$\liminf_{T \to \infty} \frac{\mathbb{E}\left[N_k^{\pi}(T)\right]}{\log(T)} \geq \frac{1}{a_1 [\Delta_{\mathrm{con}}(k)]^2}.$$

The proof of Theorem 3 can be found in "Appendix B". Comparing the lower bounds in Theorem 3 with the upper bounds for CON-LCB in Theorems 1 and 2, we conclude that CON-LCB is asymptotically optimal on regret, up to universal multiplicative constants.[2]

Next, we address the fundamental tradeoff between regret minimization and feasibility identification.

Theorem 4 *Consider a space of arm distributions $\mathcal{C}$, and a threshold τ such that $\mathcal{C}$ contains both feasible as well as infeasible arm distributions. There exists a feasible instance (v, τ) and an infeasible instance (v', τ) such that for any policy π that is consistent over $\mathcal{C}$,*

$$\limsup_{T \to \infty} -\frac{1}{T} \log \Big(\mathbb{P}_v(\texttt{feasibility\_flag = false})$$
$$+ \mathbb{P}_{v'}(\texttt{feasibility\_flag = true}) \Big) \leq 0.$$

Theorem 4 states that for any consistent algorithm, the probability that (v, τ) get misidentified as infeasible, and the probability that (v', τ) get misidentified as feasible, cannot both decay exponentially in the horizon. This is of course consistent with the power-law probability of misidentification under CON-LCB. In other words, slower-than-exponential decay of the probability of feasibility misidentification with respect to the horizon is an unavoidable consequence of the exploration-exploitation interplay in regret minimization. A similar tension between regret minimization and best arm identification was noted for the unconstrained MABs by Bubeck et al. (2009).

5 General framework for constrained MABs

In this section, we provide a general formulation for constrained stochastic MABs with multiple criteria. We allow each arm to be associated with a D-dimensional probability distibution, the goal being to optimize one (dominant) attribute associated with this

---

distribution, subject to constraints on m others. The algorithm design and performance evaluation performed in Sects. 2–4 for the special case of $m = 1$ extend naturally to this general formulation, which can in turn be applied in a variety of application scenarios. We illustrate a few here.

- For clinical trials, with the arms corresponding to various treatment protocols, the dominant attribute might, for example, correspond to the success/recovery probability, whereas the constraints might capture recovery time, severity of side effects, etc.

- For product/service rating, where the arms correspond to various service providers, the dominant attribute might correspond to product quality, with constraints capturing reliability, pricing, customer service, etc.

- In wireless networks, the arms might correspond to various access networks or channels, with, for example, the dominant attribute corresponding to throughput, and constraints capturing delay, energy efficiency, etc.

Formulation Consider a set of K arms, each associated with a D-dimensional probability distribution, with $v(k)$ denoting the joint distribution corresponding to arm $k \in [K]$. Suppose that $v(k) \in C$, the space of possible arm distributions. The objective and the constraints are defined by functions $g_0, g_1, \dots, g_m$. Specifically, the optimal arm is defined as that arm k that minimizes $g_0(v(k))$, subject to the constraints $\{g_i(v(k)) \leq \tau_i\}_{i=1}^m$ when the instance is feasible (i.e., at least one arm exists that satisfies the above constraints).

If the instance is infeasible, the optimal arm is defined via a ranked list of the constraints, that orders them by 'importance'. Without loss of generality assume that the order of importance increases from (g_1, τ_1) to (g_m, τ_m). The idea is to relax the constraints one-by-one, starting with the least important, until a compliant arm is found. Formally, for a given infeasible instance v, for $2 \leq i \leq m$, let $\mathcal{K}_i(v)$ denote the set of arms that satisfy the constraints $\{g_j(v(k)) \leq \tau_j\}_{j=i}^m$. Let us also define $\mathcal{K}_{m+1}(v) := [K]$. Now, let $i^*(v) = \min\{k \in [m] : \mathcal{K}_{k+1}(v) \neq \phi\}$. Here, $i^*(v)$ is the fewest number of constraints one must relax, in order of increasing importance, in order to have at least one compliant arm. An optimal arm is then defined to be $\arg\min\{g_{i^*(v)}(v(k)) : k \in \mathcal{K}_{i^*(v)+1}(v)\}$.

Similar to the case where $m = 1$, we make the following assumption to simplify algorithm design. Suppose that for $0 \leq i \leq m$, and $\overline{v} \in C$, there exist an estimator $\hat{g}_{i,n}(\overline{v})$ for $g_i(\overline{v})$ using n i.i.d. samples from $\overline{v}$, satisfying the following concentration inequality: There exists $a_i > 0$ such that for all $\Delta > 0$,

$$\Pr\left(\left|\hat{g}_{i,n}(\overline{v}) - g_i(\overline{v})\right| \geq \Delta\right) \leq 2\exp(-a_i n \Delta^2). \tag{2}$$

Note that this is simply an extension of our assumption (1).

Algorithm and performance guarantees For the above problem formulation, we propose Multi-Con-LCB, a simple extension of the Con-LCB algorithm that we presented before for the case of $m = 1$. Multi-Con-LCB uses upper confidence bounds on constrained attributes $\{g_i(v(k))\}_{i=1}^m$ for each arm k to maintain a set of plausibly feasible arms. The set of plausibly feasible arms for the constraint on function g_i at time $t \in [T]$ will be denoted by $\hat{\mathcal{K}}_{i,t}^{\dagger}$ for all values of $i \in [m]$. Using the sets of plausibly feasible arms for each constraint, we construct the set of arms that plausibly lie in the set $\mathcal{K}(v)$ and this set is denoted by $\hat{\mathcal{K}}_t$ for time instant $t \in [T]$. Formally, $\hat{\mathcal{K}}_t = \cap_{i=1}^m \hat{\mathcal{K}}_{i,t}^{\dagger}$. As this set might be empty, for $i \in \{2, \dots, m\}$, let the estimate for $\mathcal{K}_i(v)$ be $\hat{\mathcal{K}}_{i,t} = \cap_{j=i}^m \hat{\mathcal{K}}_{j,t}^{\dagger}$. It is also possible that the most important constraint is not satisfied, therefore let $\hat{\mathcal{K}}_{m+1,t} = [K]$. If the set $\hat{\mathcal{K}}_t$ is not empty, then the algorithm uses lower confidence bounds (LCBs) on g_0 to select the arm to be played. If the set $\hat{\mathcal{K}}_t$ turns out to be empty, then the algorithm finds the smallest index $\hat{i}^*$ such that the set $\hat{\mathcal{K}}_{\hat{i}^*+1,t}$ is not empty. The algorithm then plays the

arm with lowest LCB on $g_{\hat{i}^*}$. Finally, at the end of T rounds, MULTI-CON-LCB sets the feasibility flag as `true` if the set $\hat{\mathcal{K}}_T$ is not empty and `false` otherwise. The details are presented as Algorithm 2.

Algorithm 2 Multiple Constrained LCB

---

procedure MULTI-CON-LCB$(T, K, \{\tau_i\}_{i=1}^m)$
    Play each arm once
    for $t = K + 1, \cdots, T$ **do**
        for $i = 1, \cdots, m$ **do**
            Set $\hat{\mathcal{K}}_{i,t}^{\dagger} = \left\{ k : \hat{g}_{i,N_k(t-1)}(k) \leq \tau_i + \sqrt{\frac{\log(2T^2)}{a_i N_k(t-1)}} \right\}$
        end for
        Set $\hat{\mathcal{K}}_t = \cap_{i=1}^m \hat{\mathcal{K}}_{i,t}^{\dagger}$
        if $\hat{\mathcal{K}}_t \neq \varnothing$ **then**
            $k_{t+1}^{\dagger} \in \arg\min_{k \in \hat{\mathcal{K}}_t} \hat{g}_{0,N_k(t-1)}(k) - \sqrt{\frac{\log(2T^2)}{a_0 N_k(t-1)}}$
            Play arm $k_{t+1}^{\dagger}$
        else
            $\hat{i}^* = \arg\min_{i \in \{1, \cdots, m\}} \hat{\mathcal{K}}_{i+1,t} \neq \varnothing$
            $k_{t+1}^{\dagger} \in \arg\min_{k \in \hat{\mathcal{K}}_{\hat{i}^*+1,t}} \hat{g}_{\hat{i}^*,N_k(t-1)}(k) - \sqrt{\frac{\log(2T^2)}{a_{\hat{i}^*} N_k(t-1)}}$
            Play arm $k_{t+1}^{\dagger}$
        end if
    end for
    if $\hat{\mathcal{K}}_{T+1} \neq \varnothing$ **then**
        Set `feasibility_flag = true`
    else
        Set `feasibility_flag = false`
    end if
end procedure

The remainder of this section is devoted to performance guarantees for MULTI-CON-LCB. The suboptimality gap for an arm k is given by $\Delta(k) = \max(g_0(v_k) - g_0^*, 0)$. The infeasibility gap of an arm k for constraint i is given by $\Delta_{i,\tau_i}(k) = \max(g_i(v(k)) - \tau_i, 0)$. We restrict our attention to feasible instances here; infeasible instances can be handled on similar lines as Theorem 2 for CON-LCB.

Theorem 5 *Consider a feasible instance. Under MULTI-CON-LCB, the expected number of pulls of a feasible but suboptimal arm k (i.e., satisfying $g_0(v(k)) > g_0^*$ and $g_i(v(k)) \leq \tau_i$ for all $i \in [m]$), is bounded by*

$$\mathbb{E}\big[N_k(T)\big] \leq \frac{4\log(2T^2)}{a_0 \Delta^2(k)} + 2m + 3.$$

The expected number of pulls of a deceiver arm k (i.e., satisfying $g_0(v(k)) \leq g_0(v(1))$ and there exists a constraint indexed by $j \in [m]$ such that $g_j(v(k)) > \tau_j$) is bounded by

$$\mathbb{E}\big[N_k(T)\big] \le \min_{i \in [m]} \left(\frac{4 \log(2T^2)}{a_i [\Delta_{i,\tau_i}(k)]^2} \right) + m.$$

The expected number of pulls of a an arm k which is infeasible, but not a deceiver (i.e., satisfying $g_0(\nu(k)) > g_0(\nu(1))$ and there exists a constraint indexed by $j \in [m]$ such that $g_j(\nu(k)) > \tau_j$) is bounded by

$$\mathbb{E}\big[N_k(T)\big] \le \min \left(\frac{4 \log(2T^2)}{a_0 \Delta^2(k)}, \min_{i \in [m]} \left(\frac{4 \log(2T^2)}{a_i [\Delta_{i,\tau_i}(k)]^2} \right) \right) + 2m + 3.$$

The probability of incorrectly setting the `feasibility_flag` *is upper bounded by*

$$\Pr(\texttt{feasibility\_flag} = \texttt{false}) \le \frac{m}{T}.$$

We omit the proof of Theorem 5 in the appendix because it is very similar to the proof of Theorem 1. However, we state the key takeaways from Theorem 5 below.

• Similar to Theorem 1, the upper bound on the expected number of pulls of suboptimal arms is logarithmic in T and is inversely proportional to the suboptimality gap squared. Moreover, the upper bound for a deceiver arm is also logarithmic in T but there is a minimum over m terms because the deceiver arm can be identified by any of the constraints that the arm does not satisfy. Similarly, the upper bound for a non-deceiver, suboptimal arm is also logarithmic in T and there is a minimum over $m + 1$ terms because the arm can be identified as suboptimal, or infeasible for not satisfying one of the m constraints.

• With an increase in the number of constraints to m, the upper bounds on the expected number of pulls of non-optimal arms increase linearly in m and the probability of misidentification also gets scaled by a factor of m. The slight degradation in the guarantees is expected because with more constraints, the optimal arm has to satisfy more conditions, and it becomes harder to compare different arms.

6 Numerical experiments

In this section, we present a simple experiment to show the numerical performance of our algorithm Con-LCB.

We consider an instance with four arms and two criteria. Each arm is associated with a (two-dimensional) multivariate Gaussian distribution, different arms having the same covariance matrix Σ, but different mean vectors. The means corresponding to the first dimension are 0.3, 0.4, 0.2, and 0.5 and the means corresponding to the second dimension are 0.4, 0.4, 1.0, and 1.0. The covariance matrix is given by $\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$.
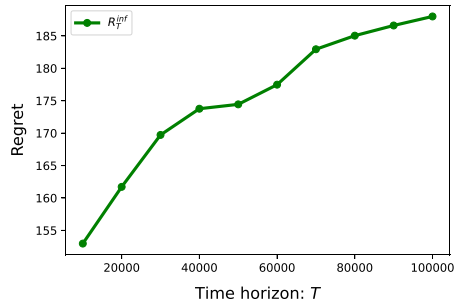
The goal is to maximize the pulls of the arm with the minimum mean of the first dimension, subject to a constraint on the mean of the second. Specifically, we require that the mean of the second dimension should be less than or equal to $\tau := 0.5$. The instance is summarized in Table 1. We use empirical averages as the estimators for the means and standard confidence bounds based on sub-Gaussianity assumption. We run the algorithm 1000 times for values of T in [10000, 100000]. The suboptimality regret R_T^{sub} and the infeasibility regret R_T^{inf} are plotted in Figure 2. Clearly, the regrets grow sub-linearly with

Table 1 Gaussian instance

Arm	Mean 1	Mean 2	Remarks
1	0.3	0.4	Optimal arm
2	0.4	0.4	Feasible, suboptimal
3	0.2	1.0	Deceiver
4	0.5	1.0	Infeasible, suboptimal
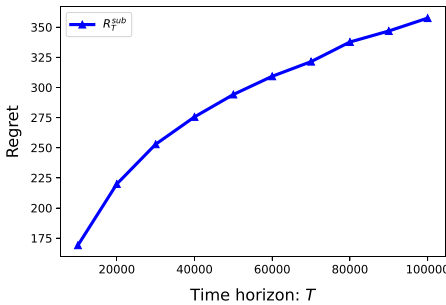
(a) Suboptimality regret (b) Infeasibility regret

Fig. 2 Regret of Gaussian instance

respect to the horizon. Also, in our experiments, the algorithm correctly detected the feasibility of the instance in all of the 1000 runs.

We consider another instance where arms are Beta distributed. The goal here is to maximize the pulls of the arm with the minimum mean, subject to a constraint on the variance. Specifically, we require that the variance of the arms should be less than or equal to $\tau := 0.1$. The instance is summarized in Table 2. We use the empirical average to estimate the mean of the arms and the standard unbiased estimator of variance to estimate the variance of the arms. The concentration bounds we use are based on fact that underlying arms are bounded between 0 and 1. We run the algorithm 100 times for values of T in [10000, 100000]. The suboptimality regret R_T^{sub} and the infeasibility regret R_T^{inf} are plotted in Figure 3. Similar to the previous case, the algorithm correctly detected feasibility in all of the 100 runs.

We also compare CON-LCB with an algorithm designed to optimize a single 'Lagrange relaxed' objective of the form $g_0 + \beta g_1$. Since there is no systematic way of setting β such that the original (constrained) optimal arm also minimizes this metric, this approach is very fragile, as we demonstrate in Figure 4. The instance we use is the Beta distributed instance in Table 2, and we are plotting the sum of the infeasible regret and suboptimal regret. When β is small ($\beta < 10/7$ here), a deceiver arm is optimal for the relaxed objective, and when β is large ($\beta > 5$ here), a suboptimal arm is optimal for the relaxed objective; the 'correct' range of β being instance dependent (and a priori unknown).

Table 2 Beta instance

Arm	Mean	Variance	Remarks
1	0.3	0.08	Optimal arm
2	0.4	0.06	Feasible, suboptimal
3	0.2	0.15	Deceiver
4	0.5	0.15	Infeasible, suboptimal
5	0.45	0.08	Feasible, suboptimal
6	0.3	0.15	Deceiver
7	0.6	0.15	Infeasible, suboptimal

7 Concluding remarks

In this paper, we have introduced a general formulation of multi-criterion constrained MABs, which is applicable when there are multiple attributes of interest associated with each arm. We propose algorithms that incur logarithmic regret, and also provide information theoretic lower bounds on the performance of any algorithm. An interesting departure from the classical MAB formulation is the aspect of instance feasibility; our algorithms predict, post facto, with a probability of error that decays as a power law in the horizon, whether the instance presented was feasible. Interestingly, this illustrates a fundamental tradeoff between regret minimization and feasibility detection. Finally, our algorithms 'auto-tune' to an infeasible instance, relaxing the constraints on arm attributes (in order of increasing importance), until a compliant arm is found.

The proposed framework and our algorithms can be applied in a wide range of application scenarios [see Bouneffouf and Rish (2019) for a survey]. Further, this work motivates the study of multi-criterion variants of other bandit formulations, including contextual bandits, combinatorial bandits, and correlated bandits.

Appendix A: Upper bounds

In this section, we prove Theorems 1 and 2. The bounds in Sections "Feasible instance: upper bounding the expected pulls of deceiver arms"–"Feasible instance: upper bounding the probability of misidentification" imply the statement of Theorem 1, and the bounds
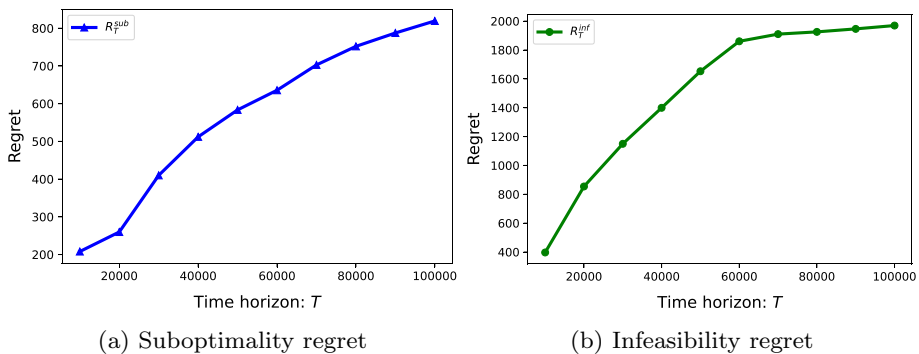

(a) Suboptimality regret   (b) Infeasibility regret

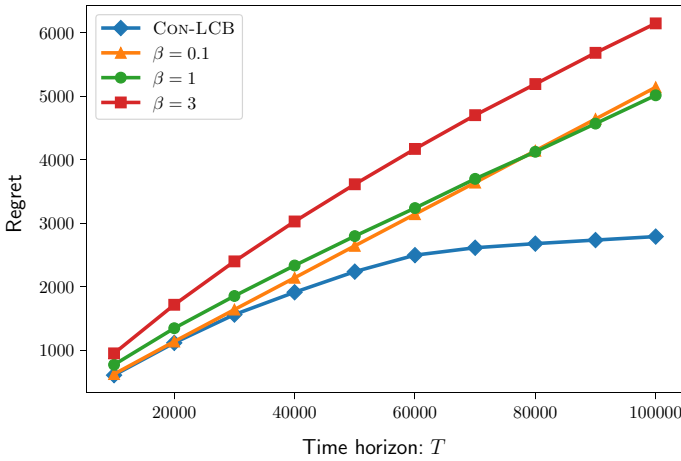**Fig. 3** Regret of Beta distributed instance

Fig. 4 Comparison with the Lagrangian relaxation

in "Sections Infeasible instance: upper bounding the expected pulls of non-optimal arms"–"Infeasible instance: upper bounding the probability of misidentification" imply the statement of Theorem 2.

A.1 Feasible instance: upper bounding the expected pulls of deceiver arms

For a deceiver arm k, we will define a good event $G_{1,k}$ where the estimator for g_1 is concentrated enough and derive an upper bound on the number of pulls of the deceiver arm.

$$G_{1,k} = \left\{ \forall n \in [T] \; |\hat{g}_{1,n}(k) - g_1(k)| < \sqrt{\frac{\log\left(2T^2\right)}{a_1 n}} \right\} \tag{3}$$

On $G_{1,k}$, we can lower bound the estimator for g_1 for the arm k as follows

$$\hat{g}_{1,n}(k) > g_1(k) - \sqrt{\frac{\log\left(2T^2\right)}{a_1 n}}$$

If the lower bound is greater than $\tau + \sqrt{\frac{\log\left(2T^2\right)}{a_1 n}}$, then arm k can't be in $\hat{\mathcal{K}}_t$. Hence, we can upper bound the number of pulls of an infeasible arm as follows

$$g_1(k) - \sqrt{\frac{\log\left(2T^2\right)}{a_1 n}} \leq \tau + \sqrt{\frac{\log\left(2T^2\right)}{a_1 n}}$$

$$\Rightarrow n \leq v_k := \lceil * \rceil \frac{4\log(2T^2)}{a_1 \Delta_\tau^2(k)} \tag{4}$$

Event $G_{1,k}^c$ is given by

$$G^c_{1,k} = \left\{ \exists n \in [T] \, |\hat{g}_{1,n}(k) - g_1(k)| \geq \sqrt{\frac{\log\left(2T^2\right)}{a_1 n}} \right\}$$

Using our assumption in Eq. 1 and a union bound, we can show

$$\mathsf{Pr}\left(G^c_{1,k}\right) \leq T \times \frac{1}{T^2} = \frac{1}{T}.$$

Now, let us upper bound the expected number of pulls of a deceiver arm k.

$$\mathbb{E}\left[N_k(T)\right] = \mathbb{E}\left[\mathbb{E}\left[N_k(T)|G_{1,k}\right]\right] + \mathbb{E}\left[\mathbb{E}\left[N_k(T)|G^c_{1,k}\right]\right]$$

$$\leq \lceil * \rceil \frac{4\log(2T^2)}{a_1 \Delta^2_\tau(k)}\left(1 - \frac{1}{T}\right) + T \times \frac{1}{T}$$

$$\leq \frac{4\log(2T^2)}{a_1 \Delta^2_\tau(k)} + 2$$

A.2 Feasible instance: upper bounding the expected pulls of feasible suboptimal arms

We will begin by showing that a feasible arm k remains in the set $\hat{\mathcal{K}}_t$ for $t \in [T]$ when estimator of g_1 is concentrated enough. We define an event $G_{1,k}$ for a feasible arm as done in Eq. (3). When $G_{1,k}$ holds, the estimator for g_1 is upper bounded by

$$\hat{g}_{1,n}(k) \leq g_1(k) + \sqrt{\frac{\log\left(2T^2\right)}{a_1 n}}$$

$$\leq \tau + \sqrt{\frac{\log\left(2T^2\right)}{a_1 n}} \quad (\because g_1(k) \leq \tau)$$

Hence, arm k is in $\hat{\mathcal{K}}_t$ for $t \in [T]$ when $G_{1,k}$ holds.

We are considering the case where an arm k is feasible but suboptimal. We will define a good event for arm k and bound the number of pulls on this good event. Without loss of generality, assume arm 1 is optimal.

$$G_{0,k} = \left\{ g_0(1) > \max_{n \in [T]} \hat{g}_{0,n}(1) - \sqrt{\frac{\log\left(2T^2\right)}{a_0 n}} \right\}$$

$$\cap \left\{ \hat{g}_{0,u_k}(k) - \sqrt{\frac{\log\left(2T^2\right)}{a_0 u_k}} > g_0(1) \right\} \cap G_{1,1} \cap G_{1,k} \qquad (5)$$

$$\text{where } u_k = \lceil * \rceil \frac{4\log(2T^2)}{a_0 \Delta^2(k)}$$

We will show that if $G_{0,k}$ holds, then $N_k(T) \leq u_k$. We will also show that $G_{0,k}^c$ holds with a small probability.

The proof is by contradiction. Suppose $G_{0,k}$ holds and $N_k(T) > u_k$, then there exists a $t \in [T]$ such that $N_k(t-1) = u_k$ and $A_t = k$. Using the definition of $G_{0,k}$,

$$\text{LCB}_k(t-1) = \hat{g}_{0,u_k}(k) - \sqrt{\frac{\log\left(2T^2\right)}{a_0 u_k}}$$

$$> g_0(1) \quad \text{(Definition of } G_{0,k})$$

$$> \text{LCB}_1(t-1) \quad \text{(Definition of } G_{0,k})$$

Hence, $A_t = \arg\min_j \text{LCB}_j(t) \neq k$, which is a contradiction. Therefore, if $G_{0,k}$ occurs, $N_k(T) \leq u_k$.

Now, consider the event $G_{0,k}^c$.

$$G_{0,k}^c = \left\{g_0(1) \leq \max_{n \in [T]} \hat{g}_{0,n}(1) - \sqrt{\frac{\log\left(2T^2\right)}{a_0 n}}\right\} \cup$$

$$\left\{\hat{g}_{0,u_k}(k) - \sqrt{\frac{\log\left(2T^2\right)}{a_0 u_k}} \leq g_0(1)\right\} \cup G_{1,1}^c \cup G_{1,k}^c.$$

Let us bound the probability of the first term above.

$$\Pr\left(g_0(1) \leq \max_{n \in [T]} \hat{g}_{0,n}(1) - \sqrt{\frac{\log\left(2T^2\right)}{a_0 n}}\right)$$

$$\leq \Pr\left(\bigcup_{n=1}^{T} \left\{g_0(1) \leq \hat{g}_{0,n}(1) - \sqrt{\frac{\log\left(2T^2\right)}{a_0 n}}\right\}\right)$$

$$\leq T \times \frac{1}{T^2} = \frac{1}{T} \quad \text{(Using Equation 1 and union bound)}$$

Now, let us bound the second event above. By the choice of u_k we have the following

$$\Delta(k) - \sqrt{\frac{\log\left(2T^2\right)}{a_0 u_k}} \geq \sqrt{\frac{\log\left(2T^2\right)}{a_0 u_k}}$$

Now,

$$\Pr\left(\hat{g}_{0,u_k}(k) - \sqrt{\frac{\log\left(2T^2\right)}{a_0 u_k}} \le g_0(1)\right)$$

$$= \Pr\left(g_0(k) - \hat{g}_{0,u_k}(k) \ge \Delta(k) - \sqrt{\frac{\log\left(2T^2\right)}{a_0 u_k}}\right) \quad (\text{Use } \Delta(k) = g_0(k) - g_0(1))$$

$$\le \Pr\left(g_0(k) - \hat{g}_{0,u_k}(k) \ge \sqrt{\frac{\log\left(2T^2\right)}{a_0 u_k}}\right) \quad (\text{By the choice of } u_k)$$

$$\le \frac{1}{T^2} \quad (\text{Using Equation 1})$$

We can show that $\Pr\left(G_{1,1}^c\right) \le 1/T$ and $\Pr\left(G_{1,k}^c\right) \le 1/T$ like in the previous subsection. Hence,

$$\Pr\left(G_{0,k}^c\right) \le \frac{3}{T} + \frac{1}{T^2}.$$

Hence, we can upper bound the number of pulls of feasible but suboptimal arms as follows

$$\mathbb{E}\left[N_k(T)\right] = \mathbb{E}\left[\mathbb{E}\left[N_k(T)|G_{0,k}\right]\right] + \mathbb{E}\left[\mathbb{E}\left[N_k(T)|G_{0,k}^c\right]\right]$$

$$\le \lceil * \rceil \frac{4\log(2T^2)}{a_0\Delta^2(k)}\left(1 - \frac{3}{T} - \frac{1}{T^2}\right) + T \times \left(\frac{3}{T} + \frac{1}{T^2}\right)$$

$$\le \frac{4\log(2T^2)}{a_0\Delta^2(k)} + 5.$$

A.3 Feasible instance: upper bounding the expected pulls of infeasible suboptimal arms

Consider arm k which is both suboptimal and infeasible. Define an event $G_{0,k}$ as done in Eq. (5). Recall that the upper bound on the pulls of the deceiver arms on event $G_{1,k}$ is denoted by v_k and the upper bound on the pulls of the feasible but suboptimal arms on event $G_{0,k}$ is denoted by u_k.

On the event $G_{0,k}$, if $u_k \ge v_k$, then due to concentration of estimator of g_1, this arm can't be played more than v_k times. If $u_k < v_k$, then due to suboptimality, this arm can't be played more than u_k times. We can show that the probability of $G_{0,k}^c$ is less than or equal to $3/T + 1/T^2$ as we did before. Hence, we can upper bound the pulls of infeasible and suboptimal arms as follows

$$\mathbb{E}\left[N_k(T)\right] = \mathbb{E}\left[\mathbb{E}\left[N_k(T)|G_{0,k}\right]\right] + \mathbb{E}\left[\mathbb{E}\left[N_k(T)|G_{0,k}^c\right]\right]$$

$$\le \min(u_k, v_k)\left(1 - \frac{3}{T} - \frac{1}{T^2}\right) + T \times \left(\frac{3}{T} + \frac{1}{T^2}\right)$$

$$\le \min(u_k, v_k) + 4$$

A.4 Feasible instance: upper bounding the probability of misidentification

Feasibility is correctly detected if for at least one of the feasible arms, event $G_{1,k}$ as defined in Eq. (3) holds. We had seen in Appendix A.2 that if estimator of g_1 is concentrated enough, a feasible arm always remains in the set $\hat{\mathcal{K}}_t$ of plausibly feasible arms.

Without loss of generality, assume that arm 1 is optimal. Then, we can lower bound the probability of correctly setting the flag as follows

$$\Pr(\texttt{feasibility\_flag} = \texttt{true}) \geq \Pr\big(G_{1,1}\big)$$

$$\geq 1 - \frac{1}{T}.$$

This upper bounds the probability of incorrectly setting the flag

$$\Pr(\texttt{feasibility\_flag} = \texttt{false}) \leq \frac{1}{T}.$$

A.5 Infeasible instance: upper bounding the expected pulls of non-optimal arms

In this section, we discuss the case when the given instance is infeasible. As defined before, the optimal choice for the algorithm is to play the arm with minimum g_1. We will upper bound the number of pulls of arms that have g_1 greater than the minimum.

For all the arms, we define good events $G_{1,k}$ as in Eq. (3) where the estimator for g_1 is concentrated enough. Let $\mathcal{E}_r = \cap_{k=1}^{K} G_{1,k}$. When event $\mathcal{E}_r$ occurs, the set $\hat{\mathcal{K}}_t$ becomes empty after at most $\sum_{k=1}^{K} v_k$ pulls where v_k is defined in Eq. (4). The analysis is similar to given in Appendix A.1.

Once the set $\hat{\mathcal{K}}_t$ becomes empty, the algorithm starts pulling arms with minimum lower confidence bound on estimator of g_1. We will upper bound the number of pulls for the non-optimal arms. Without loss of generality, assume that arm 1 has the lowest g_1. As we are dealing with an infeasible instance, $g_1(1) > \tau$. For a non-optimal arm k, we define the following good event

$$G_{r,k} = \left\{ \hat{g}_{1,w_k} - \sqrt{\frac{\log\left(2T^2\right)}{a_1 w_k}} > g_1(1) \right\} \cap \mathcal{E}_r$$

$$\text{where } w_k = \lceil * \rceil \frac{4\log(2T^2)}{a_1 \Delta_{\text{con}}^2(k)}.$$

One can check that w_k is greater than v_k because the gap $\Delta_{\text{con}}(k) = g_1(k) - g_1(1)$ is smaller than $\Delta_\tau(k) = g_1(k) - \tau$. It is easy to argue using LCB based arguments that if $G_{r,k}$ occurs, then arm k can't be pulled more than w_k times. This is similar to the proof given in Appendix A.2.

Let us upper bound the probability of $G_{r,k}^c$. Event $G_{r,k}^c$ is given by

$$G_{r,k}^c = \left\{ \hat{g}_{1,w_k} - \sqrt{\frac{\log\left(2T^2\right)}{a_1 w_k}} \le g_1(1) \right\} \cup \mathcal{E}_r^c.$$

Using analysis in Appendix A.1, we can show $\Pr\left(\mathcal{E}_r^c\right) \le K/T$. Let us bound the probability of the first term above. By the choice of w_k, we have

$$\Delta_{\mathrm{con}}(k) - \sqrt{\frac{\log\left(2T^2\right)}{a_1 w_k}} \ge \sqrt{\frac{\log\left(2T^2\right)}{a_1 w_k}}$$

Now,

$$\Pr\left(\hat{g}_{1,w_k} - \sqrt{\frac{\log\left(2T^2\right)}{a_1 w_k}} \le g_1(1) \right)$$

$$= \Pr\left(g_1(k) - \hat{g}_{1,w_k} \ge \Delta_{\mathrm{con}}(k) - \sqrt{\frac{\log\left(2T^2\right)}{a_1 w_k}} \right) \quad (\text{Use } \Delta_{\mathrm{con}}(k) = g_1(k) - g_1(1))$$

$$\le \Pr\left(g_1(k) - \hat{g}_{1,w_k} \ge \sqrt{\frac{\log\left(2T^2\right)}{a_1 w_k}} \right) \quad (\text{By the choice of } w_k)$$

$$\le \frac{1}{T^2} \quad (\text{Using Equation 1})$$

Hence, we can upper bound $\Pr\left(G_{r,k}^c\right)$ using a union bound

$$\Pr\left(G_{r,k}^c\right) \le \frac{K}{T} + \frac{1}{T^2} \le \frac{K+1}{T}.$$

When the instance is infeasible, the expected number of pulls of non-optimal arms are upper bounded by

$$\mathbb{E}\left[N_k(T)\right] = \mathbb{E}\left[\mathbb{E}\left[N_k(T)|G_{r,k}\right]\right] + \mathbb{E}\left[\mathbb{E}\left[N_k(T)|G_{r,k}^c\right]\right]$$

$$\le \lceil * \rceil \frac{4\log(2T^2)}{a_1 \Delta_{\mathrm{con}}^2(k)}\left(1 - \frac{K+1}{T}\right) + T \times \frac{K+1}{T}$$

$$\le \frac{4\log(2T^2)}{a_1 \Delta_{\mathrm{con}}^2(k)} + K + 2$$

A.6 Infeasible instance: upper bounding the probability of misidentification

In this subsection, we will upper bound the probability of incorrectly setting the `feasibility_flag`.

Firstly, we define T^* to be the minimum value of T for which the following holds:

$$T > \sum_{k=1}^{K} v_k \geq \sum_{k=1}^{K} \lceil * \rceil \frac{4 \log(2T^2)}{a_1 \Delta_\tau^2(k)}.$$

T^* exists because $C \log(T) = o(T)$ for a fixed $C \in (0, \infty)$.

For all the arms, we define good events $G_{1,k}$ as in Eq. (3) where the estimator for g_1 is concentrated enough. Let $\mathcal{E}_r = \cap_{k=1}^{K} G_{1,k}$. On event $\mathcal{E}_r$, for $t > T^*$, set of plausible feasible arms $\hat{\mathcal{K}}_t$ will remain empty.

We showed in the previous subsection that $\Pr(\mathcal{E}_r) \geq 1 - \frac{K}{T}$. Hence, we can bound the probability of incorrectly setting the `feasibility_flag` as

$$\Pr(\texttt{feasibility\_flag} = \texttt{true}) \leq \frac{K}{T} \text{ for } T > T^*.$$

Appendix B: Lower bounds

In this section, we will prove Theorems 3 and 4. To prove Theorem 3, we will first prove a result which lower bounds the pulls of non-optimal arms when the arm distributions belong to a general distribution class and the attributes are not necessarily means. Theorem 3 is proved in "Appendix B.1", Theorem 4 is proved in "Appendix B.2", and subsequent subsections provide the proof for the intermediate results.

The proof technique to derive lower bounds for the constrained bandit setting is very similar to the technique used for standard stochastic bandit setting. We begin by stating some important results that will be used later in the proofs.

We first state the divergence decomposition lemma for the constrained bandit setting. The proof of the lemma is similar to the proof of divergence decomposition for the standard setting and we leave it to the reader to verify the result [see Lemma 15.1, Lattimore and Szepesvári (2020)].

Lemma 1 *Consider two instances (v, τ) and (v', τ), where v and v' belong to a space of distributions $\mathcal{C}^K$. Fix some policy π and let $\mathbb{P}_v = \mathbb{P}_{(v,\tau),\pi}$ and $\mathbb{P}_{v'} = \mathbb{P}_{(v',\tau)\pi}$ be the probability measures on the constrained bandit model induced by the T-round interconnection between π and (v, τ) (respectively, π and (v', τ)). Then*

$$\text{KL}(\mathbb{P}_v, \mathbb{P}_{v'}) = \sum_{k=1}^{K} \mathbb{E}_v \big[N_k(T) \big] \text{KL}(P_i, P_i') \tag{6}$$

We also state the high probability Pinsker inequality [see Theorem 14.2, Lattimore and Szepesvári (2020)].

Lemma 2 *Let P and Q be probability measures on the same measurable space $(\Omega, \mathcal{F})$ and let $A \in \mathcal{F}$ be an arbitrary event. Then,*

$$P(A) + Q(A^c) \geq \frac{1}{2} \exp(-\text{KL}(P, Q)),$$

where $A^c = \Omega/A$ is the complement of A.

B.1 Lower bounding the pulls of non-optimal arms for the Gaussian case

We first state an instance-dependent lower bound for the expected pulls of non-optimal arms under consistent algorithms for any class of distributions $\mathcal{C}$ and a threshold $\tau \in \mathbb{R}$. For a feasible instance (v^f, τ), where $v^f \in \mathcal{C}^K$, define, for each non-optimal arm k,

$$\eta^f(v^f(k), g_0^*, \tau, \mathcal{C}) = \inf_{v'(k) \in \mathcal{C}} \{ \mathrm{KL}(v^f(k), v'(k)) : g_0(v'(k)) < g_0^*, \ g_1(v'(k)) \leq \tau \}.$$

Similarly, for an infeasible instance (v^i, τ), where $v' \in \mathcal{C}^K$, define, for each non-optimal arm k,

$$\eta^i(v^i(k), g_1^*, \mathcal{C}) = \inf_{v'(k) \in \mathcal{C}} \{ \mathrm{KL}(v^i(k), v'(k)) : g_1(v'(k)) < g_1^* \}.$$

Theorem 6 *Let π be a consistent policy over the class of distributions $\mathcal{C}$ given the threshold $\tau \in \mathbb{R}$. Then for a feasible instance (v^f, τ), where $v^f \in \mathcal{C}^K$, for any non-optimal arm k,*

$$\liminf_{T \to \infty} \frac{\mathbb{E}_v[N_k(T)]}{\log(T)} \geq \frac{1}{\eta^f(v^f(k), g_0^*, \tau, \mathcal{C})},$$

For an infeasible instance (v^i, τ), where $v' \in \mathcal{C}^K$, for any non-optimal arm k,

$$\liminf_{T \to \infty} \frac{\mathbb{E}_v[N_k(T)]}{\log(T)} \geq \frac{1}{\eta^i(v^i(k), g_1^*, \mathcal{C})}.$$

The bounds in "Appendix B.3" and "B.4" imply Theorem 6. The main message here is that the 'hardness' of a non-optimal arm k, which dictates the minimum number of pulls required on average to distinguish it from the optimal arm, is characterized by the reciprocal of the smallest perturbation in terms of KL divergence $\mathrm{KL}(v(k), \cdot)$ needed to 'make' the arm optimal.

Now, we will use Theorem 6 to prove Theorem 3.

Consider two d-dimensional Gaussian distribution v_1 and v_2 with means μ_1 and μ_2, and covariances Σ_1 and Σ_2. The KL divergence between these distributions is given by

$$\mathrm{KL}(v_1, v_2) = \log\left(\frac{|\Sigma_2|^{0.5}}{|\Sigma_1|^{0.5}} \right) + \frac{\mathrm{tr}(\Sigma_2^{-1}\Sigma_1) - \mathrm{tr}(\Sigma_1^{-1}\Sigma_1) + (\mu_1 - \mu_2)^T \Sigma_2^{-1}(\mu_1 - \mu_2)}{2}. \quad (7)$$

This is easy to derive but one can check Section 9 of these notes.

Let $\mathcal{G}$ be the class of 2-dimensional gaussian distributions with the covariance matrix $\Sigma = \mathrm{diagonal}(1/2a_0, 1/2a_1)$. Let attribute g_0 be the mean of the first dimension, and g_1 be the mean of the second dimension. Let $v_1, v_2 \in \mathcal{G}$, then using Eq. 7, the KL divergence between v_1 and v_2 is given by

$$\mathrm{KL}(v_1, v_2) = a_0(g_0(v_1) - g_0(v_2))^2 + a_1(g_1(v_1) - g_1(v_2))^2. \quad (8)$$

With $\mathcal{C} = \mathcal{G}$ and attributes as defined in Theorem 3, we can use Eq. 8 to calculate $\eta^f(v^f(k), g_0^*, \tau, \mathcal{C})$ and $\eta^i(v^i(k), g_1^*, \mathcal{C})$. In particular, for a feasible but suboptimal arm k, we have

$$\eta^f(\nu^f(k), g_0^*, \tau, \mathcal{C}) = a_0 \Delta^2(k),$$

for a deceiver arm k, we have

$$\eta^f(\nu^f(k), g_0^*, \tau, \mathcal{C}) = a_1 [\Delta_\tau(k)]^2,$$

and for an infeasible arm k which is not a deceiver, we have

$$\eta^f(\nu^f(k), g_0^*, \tau, \mathcal{C}) = a_0 \Delta^2(k) + a_1 [\Delta_\tau(k)]^2 \le 2 \max(a_0 \Delta^2(k), a_1 [\Delta_\tau(k)]^2).$$

Finally, for a non-optimal arm k for an infeasible instance, we have

$$\eta^i(\nu^i(k), g_1^*, \mathcal{C}) = a_1 [\Delta_{\text{con}}(k)]^2.$$

Using the results above and combining them with Theorem 6, we get Theorem 3.

B.2 Disambiguating between feasible and infeasible instances

Consider a feasible instance (ν, τ) where Arm 1 is the only feasible arm and therefore, the only optimal arm. Let $g_1^\dagger = \min_{k \in \{2,\dots,K\}} g_1(k)$ be the minimum value of g_1 for the set of infeasible arms. For Arm 1 we define

$$\eta(\nu(1), g_1^\dagger, \mathcal{C}) = \inf_{\nu'(1) \in \mathcal{C}} \{ \mathrm{KL}(\nu'(1), \nu(1)) : g_1(\nu'(1)) > g_1^\dagger \}.$$

Consider another instance (ν', τ) where $\nu'(j) = \nu(j)$ for $j \ne 1$ and $\nu'(1) \in \mathcal{C}$ such that $\mathrm{KL}(\nu(1), \nu'(1)) \le d_1 + \epsilon$ and $g_1(\nu'(1)) > g_1^\dagger$, where $d_1 = \eta(\nu(1), g_1^\dagger, \mathcal{C})$. Using divergence decomposition lemma, we have $\mathrm{KL}(\mathbb{P}_{\nu'}, \mathbb{P}_\nu) \le \mathbb{E}_\nu [N_1(T)](d_1 + \epsilon)$ and by using Lemma 2 we have

$$\mathbb{P}_\nu(A) + \mathbb{P}_{\nu'}(A^c) \ge \frac{1}{2} \exp(-\mathrm{KL}(\mathbb{P}_{\nu'}, \mathbb{P}_\nu)) \ge \frac{1}{2} \exp(-\mathbb{E}_{\nu'} [N_1(T)](d_1 + \epsilon)).$$

Let event $A = \{\texttt{feasibility\_flag} = \text{false}\}$. Taking logarithm on both sides and rearranging gives

$$-\frac{\log(\mathbb{P}_\nu(A) + \mathbb{P}_{\nu'}(A^c))}{T} \le \frac{\log(2) + (d_1 + \epsilon)\mathbb{E}_{\nu'}[N_1(T)]}{T}$$

RHS goes to zero as T goes to infinity. This follows from the definition of consistency and the fact that for instance (ν', τ), arm 1 is suboptimal. Hence, we have

$$\limsup_{T \to \infty} -\frac{\log(\mathbb{P}_\nu(A) + \mathbb{P}_{\nu'}(A^c))}{T} \le 0$$

This shows that at least for one of the instances, the probability of incorrect detection decays slower than exponential in T.

B.3 Feasible instances

Consider a class of distributions $\mathcal{C}$ and a threshold $\tau \in \mathbb{R}$. For a feasible instance (ν^f, τ), where $\nu^f \in \mathcal{C}^K$, define, for each non-optimal arm k,

$$\eta^f(v^f(k), g_0^*, \tau, \mathcal{C}) = \inf_{v'(k) \in \mathcal{C}} \{ \mathrm{KL}(v^f(k), v'(k)) : g_0(v'(k)) < g_0^*, \ g_1(v'(k)) \leq \tau \}.$$

We will show that

$$\liminf_{T \to \infty} \frac{\mathbb{E}_v[N_k(T)]}{\log(T)} \geq \frac{1}{\eta^f(v^f(k), g_0^*, \tau, \mathcal{C})}.$$

Proof Let $d_k = \eta^f(v^f(k), g_0^*, \tau, \mathcal{C})$ and fix any $\varepsilon > 0$. Let (v', τ) be a bandit instance with $v' \in \mathcal{C}^K$, and $v'(j) = v^f(j)$ for $j \neq k$ be such that $\mathrm{KL}(v^f(k), v'(k)) \leq d_k + \varepsilon$, $g_0(v'(k)) < g_0^*$, and $g_1(v'(k)) \leq \tau$. A distribution like $v'(k)$ exists because of the definition of d_k. Note that arm k is the unique optimal arm for bandit instance v'. Using divergence decomposition lemma, we have $\mathrm{KL}(\mathbb{P}_{v^f}, \mathbb{P}_{v'}) \leq \mathbb{E}_v[N_k(T)](d_k + \varepsilon)$ and by using Lemma 2 we have

$$\mathbb{P}_v(A) + \mathbb{P}_{v'}(A^c) \geq \frac{1}{2} \exp(-\mathrm{KL}(\mathbb{P}_{v^f}, \mathbb{P}_{v'})) \geq \frac{1}{2} \exp(-\mathbb{E}_{v^f}[N_k(T)](d_k + \varepsilon)).$$

Let event $A = \{N_k(T) > T/2\}$.

$$\mathbb{E}_{v^f}[N_k(T)] + \sum_{j \neq k} \mathbb{E}_{v'}[N_j(T)] \geq \frac{T}{2} (\mathbb{P}_{v^f}(A) + \mathbb{P}_{v'}(A^c)),$$

$$\geq \frac{T}{4} \exp(-\mathbb{E}_{v^f}[N_k(T)](d_k + \varepsilon)).$$

Rearranging and taking the limit inferior we get

$$\liminf_{T \to \infty} \frac{\mathbb{E}_{v^f}[N_k(T)]}{\log T} \geq \frac{1}{d_k + \varepsilon} \liminf_{T \to \infty} \frac{\log\left(\frac{T}{4(\mathbb{E}_{v^f}[N_k(T)] + \sum_{j \neq k} \mathbb{E}_{v'}[N_j(T)])} \right)}{\log T}$$

$$= \frac{1}{d_k + \varepsilon} \left(1 - \limsup_{T \to \infty} \frac{\log(\mathbb{E}_{v^f}[N_k(T)] + \sum_{j \neq k} \mathbb{E}_{v'}[N_j(T)])}{\log T} \right)$$

$$= \frac{1}{d_k + \varepsilon}.$$

The last equality follows from the definition of consistency. As an arm which is suboptimal or infeasible or both is played only $o(T^a)$ times in expectation for all $a > 0$, there exists a constant C_a for large enough T such that $\mathbb{E}_{v^f}[N_k(T)] + \sum_{j \neq k} \mathbb{E}_{v'}[N_j(T)]) \leq C_a T^a$. This gives

$$\limsup_{T \to \infty} \frac{\log(\mathbb{E}_{v^f}[N_k(T)] + \sum_{j \neq k} \mathbb{E}_{v'}[N_j(T)])}{\log T} \leq \limsup_{T \to \infty} \frac{C_a + a \log(T)}{\log(T)} = a.$$

As this holds for all $a > 0$ and ε was an arbitrary constant greater than zero, we have the result. $\qquad \square$

B.4 Infeasible instance

For an infeasible instance (v^i, τ), where $v' \in \mathcal{C}^K$, define, for each non-optimal arm k,

$$\eta^i(v_k^i, g_1^*, \mathcal{C}) = \inf_{v'(k) \in \mathcal{C}} \{ \mathrm{KL}(v_k^i, v'(k)) : g_1(v'(k)) < g_1^* \}.$$

We will show that for arm k

$$\liminf_{T \to \infty} \frac{\mathbb{E}_v \left[N_k(T) \right]}{\log(T)} \geq \frac{1}{\eta^i(v_k^i, g_1^*, \mathcal{C})}$$

Proof Let $d_k = \eta^i(v_k^i, g_1^*, \mathcal{C})$ and fix any $\varepsilon > 0$. Let (v', τ) be a bandit instance with $v'(j) = v^i(j)$ for $j \neq k$ and $v'(k) \in \mathcal{C}$ such that $\mathrm{KL}(v^i(k), v'(k)) \leq d_k + \varepsilon$ and $g_1(v'(k)) \leq g_1^*$. A distribution like $v'(k)$ exists because of the definition of d_k. Observe that the instance (v', τ) could be a feasible instance. Nonetheless, arm k is the unique optimal arm irrespective of the feasibility of instance (v', τ). Using divergence decomposition lemma, we have $\mathrm{KL}(\mathbb{P}_{v^i}, \mathbb{P}_{v'}) \leq \mathbb{E}_{v^i} \left[N_k(T) \right] (d_k + \varepsilon)$ and by using Lemma 2 we have

$$\mathbb{P}_{v^i}(A) + \mathbb{P}_{v'}(A^c) \geq \frac{1}{2} \exp(-\mathrm{KL}(\mathbb{P}_{v^i}, \mathbb{P}_{v'})) \geq \frac{1}{2} \exp(-\mathbb{E}_{v^i} \left[N_k(T) \right] (d_k + \varepsilon)).$$

Let event $A = \{ N_k(T) > T/2 \}$.

$$\mathbb{E}_{v^i} \left[N_k(T) \right] + \sum_{j \neq k} \mathbb{E}_{v'} \left[N_j(T) \right] \geq \frac{T}{2} (\mathbb{P}_{v^i}(A) + \mathbb{P}_{v'}(A^c)),$$

$$\geq \frac{T}{4} \exp(-\mathbb{E}_{v^i} \left[N_k(T) \right] (d_k + \varepsilon)).$$

Rearranging and taking the limit inferior we get

$$\liminf_{T \to \infty} \frac{\mathbb{E}_{v^i} \left[N_k(T) \right]}{\log T} \geq \frac{1}{d_k + \varepsilon} \liminf_{T \to \infty} \frac{\log \left(\frac{T}{4(\mathbb{E}_{v^i}[N_k(T)] + \sum_{j \neq k} \mathbb{E}_{v'}[N_j(T)])} \right)}{\log T}$$

$$= \frac{1}{d_k + \varepsilon} \left(1 - \limsup_{T \to \infty} \frac{\log(\mathbb{E}_{v^i} \left[N_k(T) \right] + \sum_{j \neq k} \mathbb{E}_{v'} \left[N_j(T) \right])}{\log T} \right)$$

$$= \frac{1}{d_k + \varepsilon}.$$

The last equality follows from the definition of consistency. As v^i is an infeasible instance, arm k is suboptimal and is played only $o(T^a)$ times in expectation for all $a > 0$. For instance v', arm k is the unique optimal arm. Therefore, all the other arms are played only $o(T^a)$ times in expectation for all $a > 0$. Hence, there exists a constant C_a for large enough T such that $\mathbb{E}_{v^i} \left[N_k(T) \right] + \sum_{j \neq k} \mathbb{E}_{v'} \left[N_j(T) \right]) \leq C_a T^a$. This gives

$$\limsup_{T \to \infty} \frac{\log(\mathbb{E}_{v^i} \left[N_k(T) \right] + \sum_{j \neq k} \mathbb{E}_{v'} \left[N_j(T) \right])}{\log T} \leq \limsup_{T \to \infty} \frac{C_a + a \log(T)}{\log(T)} = a.$$

As this holds for all $a > 0$ and ε was an arbitrary constant greater than zero, we have the result. □

Declarations

Conflict of interest The authors frequently collaborate with researchers from IIT Bombay (@iitb.ac.in), researchers from IIT Madras (@iitm.ac.in), and researchers from Stanford University (@stanford.edu). Jayakrishnan Nair received his PhD from Caltech (@caltech.edu) in 2012 and Krishna Jagannathan received his PhD from MIT (@mit.edu) in 2010.

References

Agrawal, S., & Devanur, N. (2016). Linear contextual bandits with knapsacks. *Advances in Neural Information Processing Systems, 29*.

Amani, S., Alizadeh, M., & Thrampoulidis, C. (2019). Linear stochastic bandits under safety constraints. *Advances in Neural Information Processing Systems, 32*.

Badanidiyuru, A., Kleinberg, R., & Slivkins, A. (2018). Bandits with knapsacks. *Journal of the ACM (JACM), 65*(3), 1–55.

Bhat, S. P., & Prashanth, L. A. (2019). Concentration of risk measures: A wasserstein distance approach. In *Advances in neural information processing systems* (pp. 11739–11748).

Bouneffouf, D., & Rish, I. (2019). A survey on practical applications of multi-armed and contextual bandits. CoRR arXiv:1904.10040.

Bubeck, S., Munos, R., & Stoltz, G. (2009). Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory* (pp. 23–37). Springer.

Bubeck, S., & Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning, 5*(1), 1–122.

Cassel, A., Mannor, S., & Zeevi, A. (2018). A general approach to multi-armed bandits under risk criteria. In *Conference on learning theory* (pp. 1295–1306). PMLR.

Chang, H. S. (2020). An asymptotically optimal strategy for constrained multi-armed bandit problems. *Mathematical Methods of Operations Research* 1–13.

David, Y., & Shimkin, N. (2016). Pure exploration for max-quantile bandits. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 556–571). Springer.

David, Y., Szörényi, B., Ghavamzadeh, M., Mannor, S., & Shimkin, N. (2018). PAC bandits with risk constraints. In *ISAIM*.

Drugan, M. M., & Nowe, A. (2013). Designing multi-objective multi-armed bandits algorithms: A study. In *The 2013 international joint conference on neural networks (IJCNN)* (pp. 1–8). IEEE.

Ehrgott, M. (2005). *Multicriteria optimization* (Vol. 491). Springer.

Galichet, N., Sebag, M., & Teytaud, O. (2013). Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. In *Asian conference on machine learning* (pp. 245–260).

Kagrecha, A., Nair, J., & Jagannathan, K. (2019). Distribution oblivious, risk-aware algorithms for multi-armed bandits with unbounded rewards. In *Advances in neural information processing systems* (pp. 11272–11281).

Kontorovich, A. (2014). Concentration in unbounded metric spaces and algorithmic stability. In *International conference on machine learning* (pp. 28–36).

Lai, T. L., & Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics, 6*(1), 4–22.

Lattimore, T., & Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Maillard, O.-A. (2013). Robust risk-averse stochastic multi-armed bandits. In *International conference on algorithmic learning theory* (pp. 218–233). Springer.

Moradipari, A., Amani, S., Alizadeh, M., & Thrampoulidis, C. (2019). Safe linear Thompson sampling. arXiv preprint arXiv:1911.02156.

Pacchiano, A., Ghavamzadeh, M., Bartlett, P., & Jiang, H. (2021). Stochastic bandits with linear constraints. In *International conference on artificial intelligence and statistics* (pp. 2827–2835). PMLR.

Prashanth, L. A., Jagannathan, K., & Kolla, R. K. (2020). Concentration bounds for CVaR estimation: The cases of light-tailed and heavy-tailed distributions. In *International conference on machine learning* (to appear).

Sani, A., Lazaric, A., & Munos, R. (2012). Risk-aversion in multi-armed bandits. In *Advances in neural information processing systems* (pp. 3275–3283).

Sankararaman, K. A., & Slivkins, A. (2021). Bandits with knapsacks beyond the worst case. *Advances in Neural Information Processing Systems, 34,* 23191–23204.

Tekin, C. (2019). The biobjective multiarmed bandit: learning approximate lexicographic optimal allocations. *Turkish Journal of Electrical Engineering & Computer Sciences, 27*(2), 1065–1080.

Tekin, C., & Turğay, E. (2018). Multi-objective contextual multi-armed bandit with a dominant objective. *IEEE Transactions on Signal Processing, 66*(14), 3799–3813.

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

Yahyaa, S., & Manderick, B. (2015). Thompson sampling for multi-objective multi-armed bandits problem. In *European symposium on artificial neural networks, computational intelligence and machine learning (ESANN)* (p. 47). Presses universitaires de Louvain.

Zimin, A., Ibsen-Jensen, R., & Chatterjee, K. (2014). Generalized risk-aversion in stochastic multi-armed bandits. arXiv preprint arXiv:1405.0833.