Mathematical Methods of Operations Research (2020) 91:545–557
https://doi.org/10.1007/s00186-019-00697-3

ORIGINAL ARTICLE

An asymptotically optimal strategy for constrained multi-armed bandit problems

Hyeong Soo Chang[1]

Received: 3 September 2018 / Revised: 29 November 2019 / Published online: 2 January 2020

Abstract
This note considers the model of "constrained multi-armed bandit" (CMAB) that generalizes that of the classical stochastic MAB by adding a feasibility constraint for each action. The feasibility is in fact another (conflicting) objective that should be kept in order for a playing-strategy to achieve the optimality of the main objective. While the stochastic MAB model is a special case of the Markov decision process (MDP) model, the CMAB model is a special case of the constrained MDP model. For the asymptotic optimality measured by the probability of choosing an optimal feasible arm over infinite horizon, we show that the optimality is achievable by a simple strategy extended from the ϵ_t-greedy strategy used for unconstrained MAB problems. We provide a finite-time lower bound on the probability of correct selection of an optimal near-feasible arm that holds for all time steps. Under some conditions, the bound approaches one as time t goes to infinity. A particular example sequence of $\{\epsilon_t\}$ having the asymptotic convergence rate in the order of $(1 - \frac{1}{t})^4$ that holds from a sufficiently large t is also discussed.

Keywords Multi-armed bandit · Constrained stochastic optimization · Simulation optimization · Constrained Markov decision process

1 Introduction

Many practical problems, e.g., in games (Browne et al. 2012), in prediction (Cesa-Bianchi and Lugosi 2006), in networking (Mahajan and Teneketzis 2007), and problems such as clinical trials, Ad placement in the Internet, etc. [see also, e.g., Tekin and Liu (2013), Santner and Tamhane (1984) and the references therein] have been studied with a model of stochastic multi-armed bandit (MAB) [see the books, e.g., Berry and Fristedt (1985), Gittins et al. (2011) and Cesa-Bianchi and Lugosi (2006)

✉ Hyeong Soo Chang
hschang@sogang.ac.kr

[1] Department of Computer Science and Engineering, Sogang University, Seoul 04107, Korea

for in depth cover of the topic and the related ones]. The usual setup of those problems have only one main objective function to be optimized. We can add some complexity to those problems by considering another objective function whose performance metric *conflicts* that of the original objective function. For example, in (wireless) communication networks, a trade-off exists between achieving a "small" delay (or "high" throughput) and "low" power consumption. To minimize the power consumption, we need to transmit with the lowest power level available. On the other hand, to maximize the throughput (or to minimize the delay) we need to transmit with the highest available power level because it will increase the probability of successful transmission. We can consider the problem of selecting an optimal feasible power level among all available powers that keeps the delay cost below some given bound. In Ad placement, we can consider choosing an optimal feasible Ad that maximizes some revenue that keeps the marketing cost below some bound.

Formally, we consider a stochastic MAB problem where there is a finite set A of arms and one arm in A needs to be sequentially played. Unlike the classical set up, in our case, when a in A is played at discrete time $t \geq 1$, the player not only obtains a sample bounded reward $X_{a,t} \in \Re$ drawn from an unknown reward-distribution associated with a, whose unknown expectation and variance are μ_a and $\sigma^2_{R,a}$, respectively, but also obtains a sample bounded cost $Y_{a,t} \in \Re$ drawn from an unknown cost-distribution associated with a, whose unknown expectation and variance are C_a and $\sigma^2_{C,a}$, respectively. Sample rewards and costs across arms are all independent for all time steps. That is, $X_{a,t}$, $X_{b,s}$, $Y_{p,t'}$, and $Y_{q,s'}$ are independent for all $a, b, p, q \in A$ and all $t, s, t', s' \geq 1$. For any fixed a in A, $X_{a,t}$'s and $Y_{a,t}$'s for $t \geq 1$ are identically distributed, respectively. We define the *feasible set* A_f of arms such that $A_f := \{a \in A | C_a \leq C\}$ for some real constant C (C is a problem parameter and we assume that $A_f \neq \emptyset$). Unlike the classical problem, our goal is to find an optimal *feasible* arm in $\arg\max_{a \in A_f} \mu_a$. We call this model "constrained MAB" (CMAB). (Note that for the sake of simplicity, we consider one constraint case. It is straightforward to extend our results into multiple-constraints case.)

In fact, the model of CMAB is a special case of the *constrained* Markov decision process (CMDP) model (Altman 1998; Denardo et al. 2013) (naming the model as *constrained* MAB based on this) while the model of the stochastic MAB is a special case of the unconstrained MDP model (see, e.g., Mahajan and Teneketzis 2007). It is important to note that in our problem setup, we add the assumption that all of the distributions of rewards and costs associated with all arms are *unknown* to the player.

When the CMAB problem parameters are all known in the model of CMDP, linear programming can be used to obtain an optimal policy that achieves the reward-optimality while keeping the constraint of the cost-feasibility (Altman 1998; Denardo et al. 2013). On the other hand, due to the assumption that the distributions of rewards and costs associated with all arms are unknown, we need to somehow blend a process of estimating the feasibility of each arm into an exploration-exploitation process for estimating the optimality of each arm. The methodology is a (simulation) process of iteratively updating estimates of the unknown parameter values, e.g., expectations, from samples of reward and cost and playing the bandit with an arm selected based on those information and obtaining new samples for further estimation eventually finding an optimal feasible arm.

We define a *strategy* (or algorithm) $\pi := \{\pi_t, t = 1, 2, \ldots\}$ as a sequence of mappings such that π_t maps from the set of past plays and rewards and costs and $m \geq 0$ random numbers, $H_{t-1} := (A \times \Re \times \Re \times [0, 1]^m)^{t-1}$ if $t \geq 2$ and $\emptyset$ if $t = 1$, to the set of all possible distributions over A. The tuple of m random numbers in $[0, 1]^m$ represents some exogenous randomness that controls the selection process in strategy. We denote the set of all possible strategies as Π. We let a random variable I_t^π denote the arm selected by π at time t.

The notion of the *asymptotic optimality* of a strategy was introduced by Robbins (1952) for the classical MAB problem, i.e., when $A_f = A$. We re-define it for the CMAB case: Let $\mu^* = \max_{a \in A_f} \mu_a$ and $A_f^* := \{a \in A_f | \mu_a = \mu^*\}$. For a given $\pi \in \Pi$, π is an *asymptotically optimal* strategy if $\sum_{a \in A_f^*} \Pr\{I_t^\pi = a\} \to 1$ as $t \to \infty$. Robbins studied a two-arm problem with Bernoulli reward distributions and Bather (1980) extended the problem into the general case where $|A| \geq 2$ and provided an asymptotically optimal "index-based" strategy. At each time each arm's certain performance index is obtained and an arm is selected based on the indices. The key idea of Bather was to ensure that each arm is played infinitely often by introducing some randomness into the index computation and to make the effect for an arm vanish as the number of times the arm has been played increases. The ϵ_t-greedy strategy (Auer et al. 2002) basically follows Bather's idea for general MAB problems: Set $\{\epsilon_t\}$ such that $\sum_{t=1}^{\infty} \epsilon_t = \infty$ and $\lim_{t \to \infty} \epsilon_t = 0$. The sequence ensures that each arm is played infinitely often and that the selection by the strategy becomes completely greedy in the limit. In addition, the value of ϵ_t plays the role of switching probability between greedy selection of the arm estimated as the current best and uniform selection over A. As ϵ_t goes to zero, the effect from uniform selection vanishes and the strategy achieves the asymptotic optimality. By analyzing its finite-time upper bound on the probability of wrong selection, Auer et al. (2002, Theorem 3) showed the convergence rate to zero is in the order of t^{-1}. Arguably, the ϵ_t-greedy policy is regarded as de facto standard algorithm of the stochastic MAB model in general when the distributions of the rewards associated with all arms are unknown [see, e.g., Auer et al. (2002), Kuleshov and Precup (2014) and Vermorel and Mohri (2005) that show the empirical competitiveness of the ϵ-greedy policy]. This naturally motivates studying how the ϵ-greedy strategy (adapted properly) works in the model of CMAB.

The goal of this brief note is to show in theory that under some conditions, a simple extension of the ϵ_t-greedy strategy, called "constrained ϵ_t-greedy" strategy achieves the asymptotic (near) optimality for CMAB problems. Our approach is to establish a finite-time lower bound on the probability of selecting an optimal near-feasible arm that holds for all time t, where the near-feasibility is measured by some deviation parameter, and then to show that the lower bound approaches one as t increases for any positive value of the deviation. In doing so, we observe that a lower-bound probability on the probability of selecting an optimal near-feasible arm can be obtained by a product of conditional probabilities. Conditioning is based on the event that each arm is pulled (sampled) "sufficiently" often and under the sufficiency, the set of the arms identified as feasible arms with the samples "approximates" the true set of feasible arms by an error. With the error, getting away from the constraints, we can deal with solving certain unconstrained MAB—obtaining a lower bound on each conditional

probability term in the product can be viewed as a problem or a property studied in the literature in the context of solving unconstrained MAB. To prove our main result in this note, we borrow (some parts of) the proof technique used in the literature that studied those problems. That is, the proof of the convergence result build upon previously existing ones and this must be obvious in some sense because the constrained ϵ_t-greedy" strategy "subsumes" the ϵ_t-greedy strategy. However, note that it is not trivial to combine all of the relevant results into one framework after proper adaptation for the novel problem. Along with theoretical study on convergence, we provide a particular example sequence of $\{\epsilon_t\}$ which makes the asymptotic convergence rate in the order of $(1 - \frac{1}{t})^4$ from a sufficiently large t.

2 Related works

Related with our model, much attention has been paid to the so called "Budgeted MAB (BMAB)" problem as a variant of the MAB problem that adds a certain constraint for optimality (see, e.g., Ding et al. 2013; Watanabe et al. 2017; Zhou and Tomlin 2018). In our terms, given a policy π, consider the sum of the random costs obtained by following π (i.e., playing the MAB machine according to π) $T > 0$ times, i.e., $\sum_{t=1}^{T} Y_{I_t^\pi,t}$. [Zhou and Tomlin (2018) considered an extended model of Ding et al. (2013) such that an arm can be played multiple times. Our argument applies similarly to their case.] It is then a random variable that takes the value of the sum of the random costs obtained over the sample path induced by following π. Let the stopping time $Q^\pi(B) = \min\left\{T \mid \sum_{t=1}^{T} Y_{I_t^\pi,t} > B\right\}$ where $B > 0$ is a problem parameter called "budget". The player stops playing the MAB machine at time $Q^\pi(B)$ once it consumes up all of the budget given by B. We now take the expected value of the sum of the random rewards obtained by following π over the sample path of length $Q^\pi(B) - 1$ and want to maximize the expected value over all possible π. That is, the goal of the BMAB problem is to obtain $\max_{\pi\in\Pi} E\left[\sum_{t=1}^{Q^\pi(B)-1} X_{I_t^\pi,t}\right]$ or a policy that achieves it.

As we can see, the (budget) constraint on the *played arm sequence* in BMAB is fundamentally different from the *feasibility constraint on each arm* in CMAB. Due to this, the optimality is different in the two models. Note that some arms in CMAB are infeasible but we do not know the infeasibility of the arms before we play the bandit machine. We need to find out, *as we play*, which arms are infeasible. Moreover, some arms in CMAB are non-optimal and we do not know the reward-optimality of the feasible arms before we play CMAB. We need to find out an optimal feasible arm as we play CMAB. That is, the feasibility (in terms of the cost-objective function) needs to be considered as another objective like the reward-optimality (in terms of the reward-objective function). At the same time, they are related such that the reward-optimality is constrained by the feasibility. Unless an arm is feasible, it is not optimal in our problem setting. Another important point is that the model of CMAB is a special case of the CMDP model (Altman 1998; Denardo et al. 2013) as we mentioned before. It seems that BMAB is not directly related with CMDP.

Our problem setting can be viewed as constrained (combinatorial) optimization problem when the objective function and the constraint function value can be obtained (but cannot be evaluated explicitly) by *sequential sampling/simulation of a solution at a time*. Note that we do not draw multiple samples of reward and cost at a single time step. We do not impose any assumption on the reward and the cost distributions (e.g., normality). Moreover, "(approximately) optimal sampling plan" or "optimal simulation-budget allocation" is not computed in advance as these or subset of these are common assumption and approaches in constrained simulation-optimization literature under to the topic of constrained "ranking and selection" [see, e.g., Pasupathy et al. (2014), Hunter and Pasupathy (2013), Park and Kim (2015) and the references therein]. Consider an optimization problem Ψ given in a general form of $\max_{i \in F}(\mu_i := E_w[r(i, w)])$, subject to $F = \{i \in S | \sigma_i := E_w[c(i, w)] \leq C\}$, where $S = \{1, 2, \ldots, n\}$ is a finite set of solutions, F is a finite set of feasible solutions, w is a random vector supported on a set $\Omega \subset \mathbb{R}^d$, $r : A \times \Omega \to \mathbb{R}$ is an objective function, and $c : A \times \Omega \to \mathbb{R}$ is a constraint function. The expectations are taken with respect to a fixed *but unknown* distribution P of w and all finite. Assume that samples w^j, $j = 1, 2, 3, \ldots$, of independent realizations of w can be generated by sampling from P and the values of $r(i, w^j)$ and $c(i, w^j)$ can be obtained for any $i \in S$ and $w^j \in \Omega$. The goal of Ψ is to find an optimal feasible solution in $\arg\max_{i \in F} \mu_i$. In this view, our approach of the ϵ-greedy strategy in CMAB can be used as a "stochastic search" (Spall 2003) for *expected-value constrained* combinatorial optimization problems [see, e.g., Wang and Ahmed (2008) and Lan and Zhou (2016) for the related works and the practical example problems].

The model considered in "profitable bandits" (Achab et al. 2018) also obtains a random reward and a random cost for playing an arm. The goal is to find an optimal policy maximizing the expected cumulative profit where the profit is the difference between the values of the reward and the cost. In other words, the reward and the cost are linearly related. There is no constraint on the feasibility of each arm.

The work by Locatelli et al. (2016) studies a "pure exploration" problem in the stochastic MAB model where the goal is to find a set of arms whose cost means are larger than a threshold, i.e., that are feasible in our terms. Therefore this model considers only one "dimension" of our model. In our model, we not only consider examining each arm for feasibility *but also finding a best arm among such feasible arms* (solving a "contest" problem) *at the same time*.

Our focus is on studying the behavior of the ϵ-greedy strategy with respect to the instantaneous regret over infinite horizon. This subject is also important in the sense of solving the "best arm identification" problem as in Bubeck et al. (2011) and Audibert et al. (2010). Our work can be viewed as a parallel work to those works in the literature in the direction of "pure exploration" for the stochastic MAB model. Even if other performance metric, called "the expected regret," has been studied well in the (recent) literature [see, e.g., Cesa-Bianchi and Lugosi (2006) and the references therein] since Lai and Robbins (1985) work and in particular Auer et al. (2002) work, the expected regret is defined in the *expected* sense for the average behaviour. On the other hand, the instantaneous regret captures the "*transient*" convergence behaviour. More precisely, the expected regret is re-interpreted as the expected loss relative to the cumulative expected reward of taking an optimal feasible arm due to the fact that

the algorithm does not always play an optimal feasible arm. In our terms, a possible definition would be $\mu^* T - \sum_{a \in A} \mu_a (\sum_{t=1}^{T} \Pr\{I_t^\pi = a\})$ if $T > 0$ is the horizon size. (In this case the loss is not always nonnegative because we consider the relative performance that includes the performances of the infeasible arms.) The regret is thus related with a finite-time behavior of the algorithm and in particular measures a degree of effectiveness in its exploration and exploitation process. If a policy achieves $O(1/t)$ bound on the instantaneous regret, then the policy achieves a logarithmic bound on the expected regret (simply from the definition). However, the other direction is not necessarily true. A policy that achieves a logarithmic bound on the expected regret does not necessarily achieve $O(1/t)$ bound on the instantaneous regret. That is, the two performance measures are not equivalent. In fact, when Auer et al. (2002) presented an instantaneous regret bound of the ϵ-greedy strategy, they noted that the instantaneous regret is a *stronger* notion than the expected regret. This must be obvious from the definitions.

To the author's best knowledge, there has been no known work that analyzes the instantaneous behaviour of UCB or its variants. There has been also no known work that analyzes the expected behaviour of UCB or its variants, i.e., $\sum_{a \in A} \Pr\{I_t^{\mathrm{UCB}} = a\}$ over finite time t. (Of course, the expected finite-time behaviour of UCB *relative to the best policy*, i.e., the expected regret, has been extensively studied in the literature.) This is most probably because UCB (and its variants) are deterministic (the decision to take a particular arm at a specific time depends only on the known history up to the time and the decision distribution is concentrated only on an arm). It appears to be non-trivial to obtain a bound on $\Pr\{I_t^{\mathrm{UCB}} = a\}$, $a \in A$, making difficult to compare the convergence behaviors, e.g., the convergence rate, with respect to the instantaneous regret between UCB (and its variants) and the ϵ-greedy strategy. Studying the advantage and the disadvantage of the ϵ-greedy strategy over UCB (and its variants) including other (heuristic) policies, is important. The emphasis here is on *algorithmic development and establishment of the asymptotic optimality* of the algorithm.

3 Algorithm

Once I_t^π in A is realized by the constrained ϵ_t-greedy strategy (referred to as π in what follows) at time t, the bandit is played with the arm and a sample reward of $X_{I_t^\pi, t}$ and a sample cost of $Y_{I_t^\pi, t}$ are obtained independently. We let $T_a(t) := \sum_{n=1}^{t} [I_n^\pi = a]$ denote the number of times a has been selected by π during the first t time steps, where $[\cdot]$ denotes the indicator function, i.e., $[I_t^\pi = a] = 1$ if $I_t^\pi = a$ and 0 otherwise. The sample average-reward $\bar{X}_{T_a(t)}$ for a in A is then given such that $\bar{X}_{T_a(t)} = \frac{1}{T_a(t)} \sum_{n=1}^{t} X_{a,n} [I_n^\pi = a]$ if $T_a(t) \geq 1$ and 0 otherwise, where $X_{a,n}$ is the sample reward observed at time n by playing a as mentioned before. Similarly, the sample average-cost $\bar{Y}_{T_a(t)}$ for a in A is given such that $\bar{Y}_{T_a(t)} = \frac{1}{T_a(t)} \sum_{n=1}^{t} Y_{a,n} [I_n^\pi = a]$ if $T_a(t) \geq 1$ and 0 otherwise, where $Y_{a,n}$ is the sample cost observed at time n by playing a. Note that $E[X_{a,t}] = \mu_a$ and $E[Y_{a,t}] = C_a$ for all t.

We refer to the process of selecting an arbitrary arm a in A with the same probabilities of $1/|A|$ for the arms in A as *uniform selection U over A* and the selected

arm by the uniform selection over A is denoted as $U(A)$. We formally describe the constrained ϵ_t-greedy strategy, π, below.

The constrained ϵ_t-greedy strategy

1. **Initialization:** Select $\epsilon_t \in (0, 1]$ for $t = 1, 2, \ldots$ Set $t = 1$ and $T_a(0) = 0$ for all $a \in A$ and $\bar{X}_0 = \bar{Y}_0 = 0$.
2. **Loop:**

   2.1 Obtain $A_t = \{a \in A | T_a(t) \neq 0 \wedge \bar{Y}_{T_a(t)} \leq C\}$.

   2.2 With probability $1 - \epsilon_t$,

   Greedy Selection: $I_t^\pi \in \arg\max_{a \in A_t} \bar{X}_{T_a(t)}$ if $A_t \neq \emptyset$ (ties broken arbitrarily).

   Otherwise, $I_t^\pi = U(A)$.

   And with probability ϵ_t,

   Random Selection: $I_t^\pi = U(A)$.

   2.3 Play the bandit with I_t^π and obtain $X_{I_t^\pi,t}$ and $Y_{I_t^\pi,t}$ independently.

   2.4 $T_{I_t^\pi}(t) \leftarrow T_{I_t^\pi}(t-1) + 1$ and $t \leftarrow t + 1$.

4 Convergence

To analyze the behavior of the constrained ϵ_t-greedy strategy, we define a set of approximately feasible arms: For a given $\kappa \in \Re$, $A_f^\kappa := \{a \in A | C_a \leq C + \kappa\}$. Given $\delta \geq 0$, any set $A_f^{\pm\delta}$ in $\mathcal{P}(A)$ is referred to as a δ-*feasible set* of arms *if* $A_f^{-\delta} \subseteq A_f^{\pm\delta} \subseteq A_f^{\delta}$ where $\mathcal{P}(A)$ is the power set of A. We say that an arm a in A is δ-*feasible* for a given $\delta \geq 0$ if a δ-feasible set exists and a is in the set. In the sequel, we further assume that the reward and the cost distributions all have the support in $[0, 1]$ for simplicity. That is, $X_{a,t}$ and $Y_{a,t}$ are in $[0, 1]$ for any a and t.

The following theorem provides a lower bound on the probability that the arm selected by π at t is equal to a best arm in some δ-feasible set $A_f^{\pm\delta}$ in terms of the parameters, $\{\epsilon_t\}$, $|A|$, δ, and $\rho := \min_{a,b \in A} |\mu_a - \mu_b|$.

Theorem 4.1 *Assume that the reward and the cost distributions associated with all arms in A have the support in* $[0, 1]$. *Let* $x_t := \frac{1}{2|A|} \sum_{n=1}^{t} \epsilon_n$ *for all* $t \geq 1$. *Then for all* $\delta \geq 0$ *and* $t \geq 1$, *we have that*

$$\Pr\left\{ I_t^\pi \in \arg\max_{a \in A_f^{\pm\delta}} \mu_a \text{ for some } \delta\text{-feasible } A_f^{\pm\delta} \in \mathcal{P}(A) \right\}$$
$$\geq \left(1 - \frac{\epsilon_t}{|A|}\right)\left(1 - |A| e^{-\frac{x_t}{5}}\right)\left(1 - 2|A| e^{-2\delta^2 x_t}\right)\left(1 - 2|A| e^{-\frac{\rho^2}{2} x_t}\right).$$

In the proof below, some parts are based on the proof idea of the results in Auer et al. (2002) and Wang and Ahmed (2008). Before the proof, note that for any t when for all $a \in A$, $T_a(t) \geq x_t$, it is not possible that $\sum_{a \in A} x_t > t$. This is because $|A| x_t \leq t/2$ from $0 < x_t \leq \frac{t}{2|A|}$, where this comes from the condition that $\epsilon_t \in (0, 1]$ for all $t \geq 1$.

We can see from the lower bound that the conditions of $\sum_{t=1}^{\infty} \epsilon_t = \infty$ and $\epsilon_t \to 0$ as $t \to \infty$ are necessary for the convergence to one as $t \to \infty$ because this makes $x_t \to \infty$. Furthermore, the lower bound shows that the convergence speed depends on the values of δ and ρ. If $\rho \neq 0$ but close to zero, the strategy will need a sufficiently large number of samples (depending on the value of δ) to distinguish the arms with the almost same (by ρ) values of the reward expectations. For the case where $\rho = 0$, we discuss below.

Let $\eta := \min_{a \in A} |C_a - C|$. The value of η represents another degree of the problem difficulty. Suppose that $\eta \neq 0$. Then the convergence to selection of an optimal 0-feasible arm at $t \to \infty$ is guaranteed with any δ in $(0, \eta)$ under some conditions (cf., Corollary 4.2). If δ is close to zero, because η is close to zero, x_t needs to be sufficiently large to compensate the small δ. Because $\Pr\{\forall a \in A, T_a(t) \geq x_t\}$ approaches one as x_t increases (cf., the proof below), this means that a large number of samples for each arm is necessary in order for π to figure out the feasibility with a high confidence. The convergence would be slow in general. At the extreme case, if $\eta = 0$ or if there exists an arm that satisfies the constraint *by equality*, then δ should be zero for the convergence because the 0-feasible set is uniquely equal to A_f. In this case or the case where $\rho = 0$, the lower bound in the theorem statement does not provide any useful result. (We provide a related remark in the conclusion.) The asymptotic optimality needs to be approximated by asymptotic near-optimality by fixing δ and/or ρ (arbitrarily) close to zero.

Finally, if the value x_t of the (normalized) cumulative sum of the switching probabilities up to time t is small, e.g., if the strategy spends rather more on greedy selection (exploitation) than random selection (exploration), the speed would be slow. That is, the convergence speed depends on the degree of switching between exploration and exploitation. We now provide the proof of Theorem 4.1.

Proof We first observe that the probability that a δ-feasible current-best arm is selected at time t by π from some δ-feasible set for a given $\delta \geq 0$ is lower bounded as follows:

$$
\begin{aligned}
&\Pr\left\{ I_t^{\pi} \in \arg\max_{a \in A_f^{\pm\delta}} \mu_a \text{ for some } \delta\text{-feasible set } A_f^{\pm\delta} \in \mathcal{P}(A) \right\} \\
&\geq \left(1 - \frac{\epsilon_t}{|A|}\right) \Pr\{\forall a \in A, T_a(t) \geq x_t\} && (1) \\
&\quad \times \Pr\left\{ A_f^{-\delta} \subseteq A_t \subseteq A_f^{\delta} | \forall a \in A, T_a(t) \geq x_t \right\} && (2) \\
&\quad \times \Pr\left\{ I_t^{\pi} \in \arg\max_{a \in A_t} \mu_a | A_f^{-\delta} \subseteq A_t \subseteq A_f^{\delta} \wedge \forall a \in A, T_a(t) \geq x_t \right\} && (3)
\end{aligned}
$$

We now provide a lower bound for each probability term except $(1 - \epsilon_t/|A|)$ in the product as given above.

Let $T_a^R(t)$ be a random variable whose value is the number of plays in which arm a was chosen at random by uniform selection (denoted as U^R) in **Random Selection** of the step 2.2 up to time t. That is, $T_a^R(t) = \sum_{n=1}^{t} [I_n^{\pi} = U^R(A)]$. Then for the first

Pr term in (1), we have that

$$\begin{aligned}\Pr\{\forall a \in A, T_a(t) \geq x_t\} &\geq \Pr\{\forall a \in A, T_a^R(t) \geq x_t\} \\ &= 1 - \Pr\{\exists a \in A T_a^R(t) < x_t\} \\ &\geq 1 - \sum_{a \in A} \Pr\{T_a^R(t) \leq x_t\} \\ &\quad \text{by Boole's inequality (Union bound)}\end{aligned}$$

We then apply Bernstein's inequality (Uspensky 1937) (stated for the completeness): Let $X_1, \ldots, X_j$ be random variables with range $[0, 1]$ and $\sum_{i=1}^{j} \mathrm{Var}[X_i | X_{i-1}, \ldots, X_1] = \sigma^2$. Let $S_j = X_1 + \cdots + X_j$. Then for all $h \geq 0$,

$$\Pr\{S_j \leq E[S_j] - h\} \leq e^{-\frac{h^2/2}{\sigma^2 + h/2}}.$$

Because $E[T_a^R(t)] = \frac{1}{|A|} \sum_{n=1}^{t} \epsilon_n = 2x_t$ and $\mathrm{Var}[T_a^R(t)] = \sum_{n=1}^{t} \frac{\epsilon_n}{|A|}(1 - \frac{\epsilon_n}{|A|}) \leq \frac{1}{|A|} \sum_{n=1}^{t} \epsilon_n = 2x_t$ by observing that $T_a^R(t)$ is the sum of t independent Bernoulli random variables, we have that by substituting $T_a^R(t)$ into S_t,

$$\Pr\{T_a^R(t) \leq 2x_t - x_t\} \leq e^{-\frac{x_t^2/2}{\sigma^2 + x_t/2}} \leq e^{-\frac{x_t^2}{2x_t + x_t/2}} = e^{-\frac{x_t}{5}}.$$

It follows that

$$\Pr\{\forall a \in A, T_a^R(t) \geq x_t\} \geq 1 - \sum_{a \in A} e^{-\frac{x_t}{5}} = 1 - |A| e^{-\frac{x_t}{5}}.$$

For the second probability term in (2), letting the event $\{\forall a \in A, T_a(t) \geq x_t\}$ be E

$$\begin{aligned}&\Pr\{A_f^{-\delta} \subseteq A_t \subseteq A_f^{\delta} | \forall a \in A, T_a(t) \geq x_t\} \\ &= 1 - \Pr\{\exists a \in A \bar{Y}_{T_a(t)} - C_a > \delta | E\} - \Pr\{\exists a \in A \bar{Y}_{T_a(t)} - C_a < -\delta | E\} \\ &= 1 - \sum_{a \in A} \Pr\{\bar{Y}_{T_a(t)} - C_a > \delta | E\} - \sum_{a \in A} \Pr\{\bar{Y}_{T_a(t)} - C_a < -\delta | E\} \\ &\geq 1 - \sum_{a \in A} e^{-2\delta^2 T_a(t)} - \sum_{a \in A} e^{-2\delta^2 T_a(t)} \\ &\geq 1 - 2|A| e^{-2\delta^2 x_t},\end{aligned}$$

where the lower bound on the last equality is achieved by Hoeffding's inequality (Hoeffding 1963): For random variables $X_1, \ldots, X_j$ with range $[0, 1]$ such that $E[X_i | X_1, \ldots, X_{i-1}] = \gamma$ for all i, $\Pr\{X_1 + \cdots + X_j\} \leq j\gamma - h\} \leq e^{-2h^2/j}$ for all $h \geq 0$.

For the third probability term in (3), let i_t^* denote any fixed arm in the set $\arg\max_{a \in A_t} \mu_a$. Let $\Delta_a = \mu_{i_t^*} - \mu_a$ for $a \in A_t \backslash \{i_t^*\}$. Then letting the event

$\{A_f^{-\delta} \subseteq A_t \subseteq A_f^{\delta} \wedge \forall a \in A, T_a(t) \geq x_t\}$ be E'

$$
\begin{aligned}
&\Pr\{I_t^{\pi} \notin \arg\max_{a \in A_t} \mu_a | A_f^{-\delta} \subseteq A_t \subseteq A_f^{\delta} \wedge \forall a \in A, T_a(t) \geq x_t\} \\
&\leq \sum_{a \in A_t \setminus \arg\max_{b \in A_t} \mu_b} \left(\prod_{c \in \arg\max_{b \in A_t} \mu_b} \Pr\{\bar{X}_{T_a(t)} > \bar{X}_{T_c(t)} | E'\} \right) \\
&\leq \sum_{a \in A_t \setminus \{i_t^*\}} \Pr\{\bar{X}_{T_a(t)} > \bar{X}_{T_{i_t^*}(t)} | E'\} \\
&\leq \sum_{a \in A_t \setminus \{i_t^*\}} \Pr\left\{\bar{X}_{T_a(t)} > \mu_a + \frac{\Delta_a}{2} | E'\right\} + \Pr\left\{\bar{X}_{T_{i_t^*}(t)} < \mu_{i_t^*} - \frac{\Delta_a}{2} | E'\right\} \\
&\leq \sum_{a \in A_t \setminus \{i_t^*\}} \left(e^{-2\left(\frac{\Delta_a}{2}\right)^2 T_a(t)} + e^{-2\left(\frac{\Delta_a}{2}\right)^2 T_{i_t^*}(t)} \right) \text{ by Hoeffding's inequality} \\
&\leq \sum_{a \in A_t \setminus \{i_t^*\}} 2e^{-2\left(\frac{\Delta_a}{2}\right)^2 x_t} \leq 2|A| e^{-2\left(\frac{\min_{a,b \in A} |\mu_a - \mu_b|}{2}\right)^2 x_t}.
\end{aligned}
$$

It follows that the third term is lower bounded by $1 - 2|A|e^{-2(\frac{\rho}{2})^2 x_t}$.

Putting the lower bounds of the three probability terms in (1), (2), and (3) together, we have the stated result that

$$
\begin{aligned}
&\Pr\left\{ I_t^{\pi} \in \arg\max_{a \in A_f^{\pm\delta}} \mu_a \text{ for some } \delta\text{-feasible } A_f^{\pm\delta} \in \mathcal{P}(A) \right\} \\
&\geq \left(1 - \frac{\epsilon_t}{|A|}\right) \left(1 - |A| e^{-\frac{x_t}{5}}\right) \left(1 - 2|A| e^{-2\delta^2 x_t}\right) \left(1 - 2|A| e^{-\frac{\rho^2}{2} x_t}\right).
\end{aligned}
$$

□

The following corollary is immediate. It states that the asymptotic optimality is achievable by π when $\eta \neq 0$ and $\rho \neq 0$ under the conditions on $\{\epsilon_t\}$.

Corollary 4.1 *Suppose that* $\sum_{t=1}^{\infty} \epsilon_t = \infty$ *and* $\lim_{t \to \infty} \epsilon_t = 0$ *and that* $\eta \neq 0$ *and* $\rho \neq 0$. *Then* $\lim_{t \to \infty} \Pr\{I_t^{\pi} \in \arg\max_{a \in A_f} \mu_a\} = 1$.

Proof From $\sum_{t=1}^{\infty} \epsilon_t = \infty$, $x_t \to \infty$ as $t \to \infty$. And ϵ_t goes to zero and $\rho \neq 0$. Therefore from Theorem 4.1, $\lim_{t \to \infty} \Pr\left\{ I_t^{\pi} \in \arg\max_{a \in A_f^{\pm\delta}} \mu_a \text{ for some } \delta\text{-feasible } A_f^{\pm\delta} \in \mathcal{P}(A) \right\} = 1$ if δ is fixed in $(0, \infty)$. Because $\eta \neq 0$, we observe that $A_f^{-\delta} = A_f = A_f^{\delta}$ for any $\delta \in (0, \eta)$ implying the event $\{I_t^{\pi} \in \arg\max_{a \in A_f^{\pm\delta}} \mu_a \text{ for some } \delta\text{-feasible } A_f^{\pm\delta} \in \mathcal{P}(A)\}$ is equal to $\{I_t^{\pi} \in \arg\max_{a \in A_f} \mu_a\}$ for such δ. □

We provide a particular example of the sequence $\{\epsilon_t\}$ such that the convergence rate can be obtained.

Corollary 4.2 *Assume that for* $t \geq 1$, $\epsilon_t = \min\{1, \frac{k}{t}\}$ *where* $k > 1$. *Then for* $t \geq k$ *we have that for any* $\delta \geq 0$,

$$\Pr\left\{ I_t^\pi \in \arg\max_{a \in A_f^{\pm\delta}} \mu_a \text{ for some } \delta\text{-feasible } A_f^{\pm\delta} \in \mathcal{P}(A) \right\} \geq \left(1 - \frac{k}{|A|t}\right)\left(1 - \frac{\beta(k, |A|, \delta, \rho)}{t^{\alpha(k,|A|,\delta,\rho)}}\right)^3,$$

where $\alpha(k, |A|, \delta, \rho) = \min\left\{ \frac{k}{10|A|}, \frac{\delta^2 k}{|A|}, \frac{k\rho}{4|A|} \right\}$ *and* $\beta(k, |A|, \delta, \rho) = \max\left\{ |A| k^{\frac{k}{10|A|}}, 2|A| k^{\frac{\delta^2 k}{|A|}}, 2|A| k^{\frac{k\rho}{4|A|}} \right\}$.

Proof From the assumption on $\{\epsilon_t\}$, $x_t = \frac{1}{2|A|}\sum_{n=1}^{k-1} \epsilon_n + \frac{1}{2|A|}\sum_{n=k}^{t} \epsilon_n = \frac{k-1}{2|A|} + \frac{k}{2|A|}\sum_{n=k}^{t} \frac{1}{n} \geq \frac{k-1}{2|A|} + \frac{k}{2|A|}\ln(\frac{t+1}{k}) \geq \frac{k}{2|A|}\ln(\frac{t}{k})$. Then by using $x_t \geq \frac{k}{2|A|}\ln(\frac{t}{k})$ in the lower bound given in Theorem 4.1, for $t \geq k$ and $\delta \geq 0$,

$$\begin{aligned}
&\Pr\left\{ I_t^\pi \in \arg\max_{a \in A_f^{\pm\delta}} \mu_a \text{ for some } \delta\text{-feasible } A_f^{\pm\delta} \in \mathcal{P}(A) \right\} \\
&\geq \left(1 - \frac{k}{|A|t}\right)\left(1 - \frac{|A| k^{\frac{k}{10|A|}}}{t^{\frac{k}{10|A|}}}\right)\left(1 - \frac{2|A| k^{\frac{\delta^2 k}{|A|}}}{t^{\frac{\delta^2 k}{|A|}}}\right)\left(1 - \frac{2|A| k^{\frac{k\rho}{4|A|}}}{t^{\frac{k\rho}{4|A|}}}\right) \\
&\geq \left(1 - \frac{k}{|A|t}\right)\left(1 - \frac{\beta(k, |A|, \delta, \rho)}{t^{\alpha(k,|A|,\delta,\rho)}}\right)^3.
\end{aligned}$$

□

For example that if $\alpha(k, |A|, \delta, \rho) \geq 1$, i.e., $k \geq \max\{\frac{4|A|}{\rho}, \frac{|A|}{\delta^2}, 10|A|\}$, $\Pr\left\{ I_t^\pi \in \arg\max_{a \in A_f^{\pm\delta}} \mu_a \text{ for some } \delta\text{-feasible } A_f^{\pm\delta} \right\} = \Theta((1 - 1/t)^4)$, i.e., the probability is in the order of $(1 - 1/t)^4$ for $t \geq k$. In general, if δ and/or ρ is small, in order to make $\alpha(\cdot) \geq 1$, k needs to be sufficiently large. The convergence rate is achieved asymptotically.

5 Concluding remark

As we mentioned before, if there exists an arm that achieves the equality constraint or if $\eta = 0$, then the finite-time bound in Theorem 4.1 does not provide any useful result because δ needs to be set zero. When $\rho = 0$, we have the same issue. It seems that describing a finite-time behavior of the strategy including both cases (e.g., by obtaining a useful finite-time bound) is difficult. We leave this as a future study. However, we remark that these cases do not break the convergence or the asymptotic optimality of the constrained ϵ_t-greedy strategy. This is because as long as the condition that $\sum_{t=1}^{\infty} \epsilon_t = \infty$ and $\epsilon_t \to 0$ holds, in fact, we still preserve the property that each action in A is played infinitely often in the constrained ϵ_t-greedy strategy. This can be seen by the fact that $T_a(t)$ goes to infinity for each $a \in A$ with probability one as $t \to \infty$. The sample average of $\bar{Y}_{T_a(t)}$ and $\bar{X}_{T_a(t)}$ will then eventually converge to the true average of C_a and μ_a, respectively, in the limit (simply by the law of large

numbers). The probability that the constraint ϵ_t-greedy strategy selects an optimal feasible arm will approach one in the limit.

In the ϵ-greedy algorithm, the set of feasible arms is estimated such that it consists of all arms whose empirical cost mean is at most C. We can add another degree of toleration into the ϵ-greedy algorithm itself, e.g., in the step 2.1 when it estimates the feasibility set. This will result in another degree of approximation in the result of Theorem 3.1.

A possible future work is to incorporate the approach by Locatelli et al. (2016) to develop a policy in CMAB. However, it seems to be non-trivial to analyze the convergence behavior of the resulting policy. This would be because the algorithm of Locatelli et al. (2016) is a deterministic index-based algorithm. Note that in our setting when a policy selects an arm for estimating an optimal feasible arm, at the same time the selection needs to be used for estimating the feasibility of each arm. Relating their feasibility-estimation part into the optimal-arm estimation part to obtain a bound on $\Pr\{I_t^\pi = a\}, a \in A$, seems difficult.

The necessary steps to perform the expected regret analysis in our setting would need to first define "constrained" expected regret. A possible definition of the expected regret of $\pi \in \Pi$ after the first T plays would be $T\mu^* - \sum_{a\in V} \mu_a E[T_a(T)] = T\mu^* - \sum_{a\in V} \mu_a(\sum_{t=1}^T \Pr\{I_t^\pi = a\})$, where if $V = A$, then the regret is the expected loss that occurs by not always playing an optimal feasible arm. If $V = A_f$ instead, then the regret is defined as the relative performance that includes the performances of only the feasible arms. The loss in this case is always nonnegative. If a policy that achieves a tight or even "reasonable" bound on this regret definition needs to be developed, the policy would need to obviously consider interdependency between measuring or estimating feasibility and ranking the arms. It would be the interdependency that makes the actual analysis about bounding the expected regret challenging. For example, one can consider extending UCB into a policy that selects the current best arm according to the indexes based on the sample reward-average among the arms whose sample cost-averages are below a given constraint value. But it seems that the technique used to prove the upper bound on the usual expected regret with respect to UCB for Theorem 1 in Auer et al. (2002) cannot simply be extended to bound the expected regret given as above because the events associated with not only ranking but also feasibility need to be defined and somehow manipulated in a "combined" way.

This note focused on one particular objective function, the asymptotic optimality, for CMAB problems. It is an important issue to consider another performance metric like the expected regret, and analyze the performance of a policy. Studying about other performance metrics is beyond of the scope of this note and is a good future work. Finally, investigating the theoretical results of the ϵ-greedy strategy and showing the advantage or the disadvantage over other policies, e.g., UCB (and its variants), by some experimental studies is an important future work.

References

Achab M, Clemencon S, Garivier A (2018) Profitable bandits. In: Proceedings of the 10th Asian conference on machine learning, vol 95, pp 694–709

Altman E (1998) Constrained Markov decision processes. Chapman & Hall, London
Audibert J-Y, Bubeck S, Munos R (2010) Best arm identification in multi-armed bandits. In Proceedings of the 23rd international conference on learning theory (COLT)
Auer P, Cesa-Bianchi N, Fisher P (2002) Finite-time analysis of the multiarmed bandit problem. Mach Learn 47:235–256
Bather J (1980) Randomized allocation of treatments in sequential trials. Adv Appl Probab 12(1):174–182
Berry D, Fristedt B (1985) Bandit problems: sequential allocation of experiments. Chapman & Hall, London
Browne CB, Powley E, Whitehouse D, Lucas SM, Cowling PI, Rohlfshagen P, Tavener S, Perez D, Samothrakis S, Colton S (2012) A survey of Monte Carlo tree search methods. IEEE Trans Comput Intell AI Games 4(1):1–43
Bubeck S, Munos R, Stoltz G (2011) Pure exploration in finitely armed and continuous armed bandits. Theor Comput Sci 412:1832–1852
Cesa-Bianchi N, Lugosi G (2006) Prediction, learning, and games. Cambridge University Press, Cambridge
Denardo EV, Feinberg EA, Rothblum UG (2013) The multi-armed bandit, with constraints. Ann Oper Res 208(1):37–62
Ding W, Qin T, Zhang XD, Liu TY (2013) Multi-armed bandit with budget constraint and variable costs. In: Proceedings of the 27th AAAI conference on artificial intelligence, pp 232–238
Gittins J, Glazebrook K, Weber R (2011) Multi-armed bandit allocation indices. Wiley, Hoboken
Hoeffding W (1963) Probability inequalities for sums of bounded random variables. J Am Stat Assoc 58:13–30
Hunter SR, Pasupathy R (2013) Optimal sampling laws for stochastically constrained simulation optimization on finite sets. INFORMS J Comput 25(3):527–542
Kuleshov V, Precup D (2014) Algorithms for the multi-armed bandit problem. arXiv:1402.6028
Lai T, Robbins H (1985) Asymptotically efficient adaptive allocation rules. Adv Appl Math 6:4–22
Lan G, Zhou Z (2016) Algorithms for stochastic optimization with expectation constraints. arXiv:1604.03887
Locatelli A, Gutzeit M, Carpentier A (2016) An optimal algorithm for the thresholding bandit problem. In: Proceedings of the 33rd international conference on machine learning, pp 1690–1698
Mahajan A, Teneketzis D (2007) Multi-armed bandit problems. In: Hero AO, Castanon DA, Cochran D, Kastella K (eds) Foundations and applications of sensor management. Springer, Boston
Park C, Kim S (2015) Penalty function with memory for discrete optimization via simulation with stochastic constraints. Oper Res 63(5):1195–1212
Pasupathy R, Hunter SR, Pujowidianto NA, Lee LH, Chen C (2014) Stochastically constrained ranking and selection via SCORE. ACM Trans Modeling Comput Simul 25, Article 1
Robbins H (1952) Some aspects of the sequential design of experiments. Bull Am Math Soc 58:527–535
Santner T, Tamhane A (1984) Design of experiments: ranking and selection. CRC Press, Boca Raton
Spall JC (2003) Introduction to stochastic search and optimization: estimation, simulation, and control. Wiley, Hoboken
Tekin C, Liu M (2013) Online learning methods for networking. Found Trends Netw 8(4):281–409
Uspensky JV (1937) Introduction to mathematical probability. McGraw-Hill, London
Vermorel J, Mohri M (2005) Multi-armed bandit algorithms and empirical evaluation. In: Gama J, Camacho R, Brazdil PB, Jorge AM, Torgo L (eds) Machine learning: ECML 2005, vol 3720. Lecture notes in computer science. Springer, Berlin, pp 437–448
Wang W, Ahmed S (2008) Sample average approximation of expected value constrained stochastic systems. Oper Res Lett 36:515–519
Watanabe R, Komiyama J, Nakamura A, Kudo M (2017) KL-UCB-based policy for budgeted multi-armed bandits with stochastic action costs. IEICE Trans Fundam Electron Commun Comput Sci E100–A(11):2470–2486
Zhou DP, Tomlin CJ (2018) Budget-constrained multi-armed bandits with multiple plays. In: Proceedings of the 32nd AAAI conference on artificial intelligence, pp 4572–4579