PAC Lower Bounds and Efficient Algorithms for The Max K-Armed Bandit Problem

Yahel David YAHEL83@GMAIL.COM
Nahum Shimkin SHIMKIN@EE.TECHNION.AC.IL
Department of Electrical Engineering, Technion–Israel Institute of Technology, 32000 Haifa, Israel

Abstract

We consider the Max K-Armed Bandit problem, where a learning agent is faced with several stochastic arms, each a source of i.i.d. rewards of unknown distribution. At each time step the agent chooses an arm, and observes the reward of the obtained sample. Each sample is considered here as a separate item with the reward designating its value, and the goal is to find an item with the highest possible value. Our basic assumption is a known lower bound on the *tail function* of the reward distributions. Under the PAC framework, we provide a lower bound on the sample complexity of any (ϵ, δ)-correct algorithm, and propose an algorithm that attains this bound up to logarithmic factors. We provide an analysis of the robustness of the proposed algorithm to the model assumptions, and further compare its performance to the simple non-adaptive variant, in which the arms are chosen randomly at each stage.

1. Introduction

In the classic stochastic multi-armed bandit (MAB) problem, the learning agent faces a set K of stochastic arms, and wishes to maximize the cumulative reward (in the regret formulation), or find the arm with the highest expected reward (the pure exploration problem). This model has been studied extensively in the statistical and learning literature, see for example (Bubeck & Cesa-Bianchi, 2012) for a comprehensive survey.

We consider a variant of the MAB problem called the Max K-Armed Bandit problem (Max-Bandit for short). In this variant, the objective is to obtain a sample with the highest possible reward (namely, the highest value in the support of the probability distribution of any arm). More precisely, considering the PAC setting, the objective is to return an (ϵ, δ)-correct sample, namely a sample whose value is ϵ-close to the overall best with a probability larger than $1-\delta$. In addition, we wish to minimize the sample complexity, namely the expected number of samples observed by the learning algorithm before it terminates. To minimize the sample complexity, a learning agent should ideally be able to focus on sampling the best arm, at the expense of the other, sub-optimal ones. To facilitate that, some assumptions on the reward distributions, or at least their tails, are required. These assumptions may be parametric, such as those used in (Cicirello & Smith, 2005; Streeter & Smith, 2006a; Carpentier & Valko, 2014). In this paper we use a more relaxed approach, by assuming a known lower bound on the tail of each distribution near its maximal value (Assumption 1). This assumption is further relaxed in Section 5.

The scenario considered in the Max-Bandit problem is most relevant when a single best item needs to be selected from a large collections of items, which are divided among several distinct sets. Here each set represented as a single arm. These sets may represent, for example, parts that come from different sources or produced by different processes, or job candidates that are referred by different employment agencies. Relevant problems include finding the best single match to certain genetic characteristics in several distinct populations, choosing the best channel among different frequency bands in a cognitive radio wireless network, or searching in multiple data bases for a match to a given image or document. Another application of interest is the search of a chemical compound which is suited for a given task (e.g., (Amis, 2004; Apostolidis et al., 2004)). Here, one empirically searches for a material (or a composition of materials) that provides the best result for the given task, with the number of tested compositions often reaching hundreds of thousands variations, coming from different families of materials.

For the classical MAB problem, algorithms that find the best arm (in terms of its expected reward) in the PAC

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48. Copyright 2016 by the author(s).

sense were presented in (Even-Dar et al., 2002; Audibert & Bubeck, 2010; Kalyanakrishnan et al., 2012; Gabillon et al., 2012; Karnin et al., 2013), and lower bounds on the sample complexity were presented in (Mannor & Tsitsiklis, 2004; Audibert & Bubeck, 2010; Kaufmann et al., 2016). The essential difference with respect to this work is in the objective, which is to find an (ϵ, δ)-correct sample in our case.

The Max-Bandit problem was apparently first proposed in (Cicirello & Smith, 2005). For reward distribution functions in a specific family, an algorithm with an upper bound on the sample complexity that increases as $\frac{-\ln(\delta)}{\epsilon^2}$ was provided in (Streeter & Smith, 2006a). For the case of discrete rewards, another algorithm was presented in (Streeter & Smith, 2006b), without performance analysis. Later, a similar model in which the objective is to maximize the expected value of the largest sampled reward for a given number of samples (n) was studied in (Carpentier & Valko, 2014). In that work the attained best reward is compared with the expected reward obtained by an oracle that samples the best arm n time. An algorithm is suggested and shown to secure an upper bound of order $n^{-b/((b+1)\alpha)}$ on that difference, where $\alpha > 0$ and $b > 0$ are determined by the properties of the distribution functions and b decreases as they are further away from a specific functions family. Recently, a similar model in which the goal is to find the arm for which the value of a given quantile (τ) is the largest was studied in (Szörényi et al., 2015). Their model can be compared to ours by allowing an error ϵ of the same size as the given quantile. In this case, the bound on the sample complexity provided in (Szörényi et al., 2015) increases as $\frac{-\ln(\tau)-\ln(\delta)}{\tau^2}$.

Our basic assumption in the present paper is that a known lower bound ($G_*(\epsilon)$, formally defined in Section 2) is available on the tail distributions, namely on the probability that the reward of each given arm will be close to its maximum. A special case is a lower bound on the probability densities near the maximum. Under that assumption, we provide an algorithm for which the sample complexity increases at most as $\frac{-\ln(G_*(\epsilon)\delta)}{G_*(\epsilon)}$. In the context of (Streeter & Smith, 2006a), $G_*(\epsilon) \simeq \epsilon$ and in the context of (Szörényi et al., 2015) $G_*(\epsilon) = \tau$. Therefore, the proposed algorithm provides an improvement by a factor of ϵ^{-1} over the result of (Streeter & Smith, 2006a), which was obtained for a more specific model, and an improvement by the same factor over the result of (Szörényi et al., 2015) which was derived for a similar, but different objective. To compare with the result in (Carpentier & Valko, 2014), we note that by considering the expected maximal value as the maximal possible value, it follows that $G_*(\epsilon) \simeq \epsilon^{\alpha}$. With a choice of $\delta = \frac{1}{n^2}$ in our algorithm, we obtain that the expected deficit of the largest sample with respect to the maximal reward possible is at most of order $O(\frac{\ln(n)}{n^{1/\alpha}})$ (as compared to $O(n^{-b/((b+1)\alpha)})$ with $b > 0$). Furthermore, we provide a lower bound on the sample complexity of every (ϵ, δ)-correct algorithm, which is shown to coincide, up to a logarithmic term, with the upper bound derived for the proposed algorithm. To the best of our knowledge, this is the first lower bound for the Max-Bandit problem. In addition, we analyze the robustness of the algorithm to our choice of the tail function bound $G_*(\epsilon)$, both for the case where this choice is too optimistic (i.e., the actual distributions do not obey the assumed bound) and for the case where our choice it overly conservative.

A basic feature of the Max-Bandit problem (and the associated algorithms) is the goal of quickly focusing on the best arm (in term of maximal reward), and sampling from that arm as much as possible. It is natural to compare the obtained results with an alternative approach, which ignores the distinction between arms, and simply draws a sample from a random arm at each round. This can be interpreted as mixing the items associated with each arm before sampling; we accordingly refer to this variant as the unified-arm problem. This problem actually coincides with the so-called infinitely-many armed bandit model studied in (Berry et al., 1997; Teytaud et al., 2007; Wang et al., 2008; Chakrabarti et al., 2009; Bonald & Proutiere, 2013), for the specific case of deterministic arms studied in (David & Shimkin, 2014) and (David & Shimkin, 2015). As may be expected, the unified-arm approach provides the best results when the reward distribution of all arms are identical. However, when many arms are suboptimal, the multi-armed approach provides superior performance.

The paper proceeds as follows. In the next section we present our model. In Section 3 we provide a lower bound on the sample complexity of every (ϵ, δ)-correct algorithm. In Section 4 we present an (ϵ, δ)-correct algorithm, and provide an upper bound on its sample complexity. The algorithm is simple and its bound has the same order as the lower bound up to a logarithmic term in $\frac{|K|}{\epsilon}$ (where $|K|$ stands for the number of arms). Then, in Section 5, we provide an analysis of the algorithm's performance for the case in which our assumption does not hold. In Section 6, we consider for comparison the unified-arm approach. In Section 7 we close the paper by some concluding remarks. Certain proofs are differed to the Appendix due to space limitations.

2. Model Definition

We consider a finite set of arms, denoted by K. At each stage $t = 1, 2, \ldots$ the learning agent chooses an arm $k \in K$, and a real valued reward is obtained from that arm. The rewards obtained from each arm k are independent and identically distributed, with a distribution function

(CDF) $F_k(\mu)$, $\mu \in \mathbb{R}$. We denote the maximal possible reward of each arm by $\mu_k^* = \inf_{\mu\in\mathbb{R}}\{\mu|F_k(\mu) = 1\}$, assumed finite, and the maximal reward among all arms by $\mu^* = \max_{k\in K} \mu_k^*$. The tail function $G_k(\epsilon)$ of each arm is defined as follows.

Definition 1. *For every arm $k \in K$, the tail function $G_k(\epsilon)$ is defined by*

$$G_k(\epsilon) \triangleq 1 - F_k(\mu_k^* - \epsilon), \quad \epsilon \geq 0\,.$$

For example, when $\boldsymbol{\mu}$ is uniform on $[a, b]$, then $G(\epsilon) = \frac{\epsilon}{b-a}$. In addition, we note that CDFs are nondecreasing functions and therefore the tail functions are nondecreasing. It should be observed that $G_k(\epsilon)$ does not reveal the maximal value μ_k^*, which remains unknown.

Throughout the paper, we shall use the following assumption.

Assumption 1. *There exists a known function $G_*(\epsilon)$ and a known constant $\epsilon_0 > 0$ such that, for every $k \in K$ and $0 \leq \epsilon \leq \epsilon_0$, it holds that*

$$G_k(\epsilon) \geq G_*(\epsilon)\,, \tag{1}$$

We note that for every $k \in K$, $P(\boldsymbol{\mu}_k > \mu_k^* - \epsilon) \geq G_*(\epsilon)$ where $\boldsymbol{\mu}_k$ stands for a random variable with distribution F_k. Furthermore, noting that the tail functions are non-negative and non-increasing, we assume the same for their lower bound $G_*(\epsilon)$. Moreover, for convenience we shall assume that $G_*(\epsilon)$ is strictly decreasing in ϵ, and denote its inverse function by $G_*^{-1}(\epsilon)$.

An important special-case is when one assumes that the probability density function (pdf) of each arm is lower bounded by a certain constant $A > 0$, so that $G_*(\epsilon) = A\epsilon$. We shall often use the more general bound of the form $G_*(\epsilon) = A\epsilon^\beta$ to illustrate our results.

An algorithm for the Max-Bandit model samples an arm at each time step, based on the observed history so far (i.e., the previously selected arms and observed rewards). We require the algorithm to terminate after a random number T of samples, which is finite with probability 1, and return a reward V which is the maximal reward observed over the entire period. An algorithm is said to be (ϵ, δ)-*correct* if

$$P(V > \mu^* - \epsilon) > 1 - \delta\,.$$

The expected number of samples $E[T]$ taken by the algorithm is the *sample complexity*, which we wish to minimize.

3. A Lower Bound

Before turning to our proposed algorithm, we provide a lower bound on the sample complexity of any (ϵ, δ)-correct algorithm. The bound is established under Assumption 1, and the additional provision that $G_*(\epsilon)$ is concave. The case of non-concave $G_*(\epsilon)$ turns out to be more complicated for analysis, and it is currently unclear whether our lower bound holds in that case.

For example, when $G_*(\epsilon) = A\epsilon^\beta$ for some known constants $A > 0$ and $\beta > 0$,

$$P(\boldsymbol{\mu}_k > \mu_k^* - \epsilon) \geq A\epsilon^\beta\,, \tag{2}$$

the required concavity holds for $\beta \leq 1$. The bound in Equation 2 is usually referred as β-regularity and is similar to those assumed in (Berry et al., 1997), (Wang et al., 2008), (David & Shimkin, 2014) and (Carpentier & Valko, 2015).

The following result specifies our lower bound.

Theorem 1. *Let k^* denote some optimal arm, such that $\mu_{k^*}^* = \mu^*$. Let Assumption 1 holds with a concave function $G_*(\epsilon)$ and let $\epsilon \leq \epsilon_0$ and $\delta \leq \frac{3}{20}e^{-3}$. Then, for every (ϵ, δ)-correct algorithm,*

$$E[T] \geq \sum_{k\in K\setminus\{k^*\}} \frac{1}{32G_*(\Theta_k)} \ln\left(\frac{3}{20\delta}\right) \tag{3}$$

where $\Theta_k = \min\{\max(\epsilon, \mu^* - \mu_k^*), \epsilon_0\}$.

We note that the specific requirement on δ is not fundamental, and can be released at the cost of a smaller constant in the bound.

This lower bound can be interpreted as summing over the minimal number of times that each arm, other than the optimal arm k^*, needs to be sampled. It is important to observe that if there are several optimal arms, only one of them is excluded from the summation. Indeed, the bound is large when there are several optimal (or near-optimal) arms, as the denominator of the summand is small for such arms. This follows since the algorithm needs to obtain more samples to verify that a given arm is ϵ-optimal.

The proof of Theorem 1 proceeds by considering any given set of reward distributions that obeys the Assumption, and showing that if an algorithm samples some suboptimal arm less than a certain number of times, it cannot be (ϵ, δ)-correct for some related set of reward distributions for which this arm is optimal.

Proof of Theorem 1. We begin by defining the following set of hypotheses $\{H_0, H_1, \ldots, H_{|K|}\}$, where $F_l^{H_k}(\mu)$ stands for the CDF of arm l under hypothesis k and $\mathbf{1}_\Theta$ stands for the indicator function of the set Θ. Hypothesis H_0 is the true hypothesis, namely,

$$F_k^{H_0}(\mu) = F_k(\mu) \quad \forall k \in K.$$

For $k = 1, \ldots, |K|$, we define H_k as follows. For each arm $l \neq k$, its CDF coincides with the true one, namely,

$$F_l^{H_k}(\mu) = F_l(\mu), \quad l \neq k.$$

For arm k, we construct a CDF $F_k^{H_k}$ such that its maximal value is $\mu_k^{*,H_k} = \mu^* + \epsilon$, while it still satisfies Assumption 1. To define $F_k^{H_k}$, we use the notation

$$F_*(\mu) = \begin{cases} 1 - G_*(\mu^* + \epsilon - \mu) & \mu < \mu^* + \epsilon \\ 1 & \mu \geq \mu^* + \epsilon \end{cases}$$

where ϵ is provided to the algorithm. We consider two cases.

Case 1: $\mu_k^* < \mu^* + \epsilon - \epsilon_0$. Let

$$\begin{aligned} F_k^{H_k}(\mu) =& \gamma_{k,1} F_k(\mu) \mathbf{1}_{(-\infty, \mu_k^*)}(\mu) \\ &+ \gamma_{k,1} F_k(\mu_k^*) \mathbf{1}_{[\mu_k^*, \mu^* + \epsilon - \epsilon_0)}(\mu) \\ &+ F_*(\mu) \mathbf{1}_{[\mu^* + \epsilon - \epsilon_0, \infty)}(\mu) \,, \end{aligned}$$

where $\gamma_{k,1} = 1 - G_*(\epsilon_0)$.

Case 2: $\mu_k^* \geq \mu^* + \epsilon - \epsilon_0$. Define $P_k^\epsilon \triangleq 1 - G_*(\epsilon_0) + G_*(\mu^* + \epsilon - \mu_k^*) \leq 1$, and let

$$\overline{\mu}_k = \sup_{\mu \leq \mu_k^*} \{\mu | F_k(\mu) \leq P_k^\epsilon\}$$

denote the value for which F_k reaches probability P_k^ϵ. Set

$$\begin{aligned} F_k^{H_k}(\mu) =& \gamma_{k,2} F_k(\mu) \mathbf{1}_{(-\infty, \overline{\mu}_k)}(\mu) \\ &+ \left(F_k(\mu) + (\gamma_{k,2} - 1) F_k(\overline{\mu}_k)\right) \mathbf{1}_{[\overline{\mu}_k, \mu_k^*)}(\mu) \\ &+ F_*(\mu) \mathbf{1}_{[\mu_k^*, \infty)}(\mu) \end{aligned}$$

where $\gamma_{k,2} = 1 - \frac{G_*(\mu^* + \epsilon - \mu_k^*)}{F_k(\overline{\mu}_k)}$.

By Lemma 1, which is provided in Section 8.1 in the Appendix, it follows that assumption 1 holds under all of the hypotheses $\{H_0, H_1, \ldots, H_{|K|}\}$.

If hypothesis H_k ($k \neq 0$) were true, then $\mu_k^* \geq \mu_l^* + \epsilon$ for all $l \neq k$, hence the algorithm should provide a reward from arm k with probability larger than $1 - \delta$. We use E_k^H and P_k^H to denote the expectation and probability, respectively, under the algorithm being considered and hypothesis H_k. For every $k \in K$ let

$$t_k = \frac{1}{16 \gamma_k} \ln\left(\frac{3}{20\delta}\right),$$

where

$$\gamma_k = \begin{cases} G_*(\epsilon_0) & \mu_k^* < \mu^* + \epsilon - \epsilon_0 \\ G_*(\mu^* + \epsilon - \mu_k^*) & \mu_k^* \geq \mu^* + \epsilon - \epsilon_0 \end{cases},$$

and let T_k stand for the number of samples from arm k.

Suppose now that our algorithm is (ϵ, δ)-correct under H_0, and that $E_0^H[T_k] \leq t_k$ for some $k \in K$. We will show that this algorithm cannot be (ϵ, δ)-correct under hypothesis H_k. Therefore, an (ϵ, δ)-correct algorithm must have $E_0^H[T_k] > t_k$ for all $k \in K$.

Define the following events , for $k \in K$:

- $A_k = \{T_k \leq 4t_k\}$. It easily follows from $4t_k \left(1 - P_0^H(A_k)\right) \leq E_0^H[T_k]$ that if $E_0^H[T_k] \leq t_k$, then $P_0^H(A_k) \geq \frac{3}{4}$.
- Let B_k stand for the event under which the chosen arm at termination is k, and B_k^C for its complement. Since $P_0^H(B_{k'}) > \frac{1}{2}$ can hold for one arm at most, it follows that

$$\exists k' : P_0^H\left(B_k^C\right) > \frac{1}{2}, \; \forall k \neq k'$$

- Let C_k to be the event under which all the samples obtained from arm k are on the interval $(-\infty, \mu_k^*]$. Clearly, $P_0^H(C_k) = 1$.
- For $k \in K$ for which $\mu_k^* < \mu^* + \epsilon - \epsilon_0$, $\overline{\mu}_k$ is still defined as before, so $\overline{\mu}_k = \mu_k^*$ (and $F_k(\overline{\mu}_k) = 1$). Now, for every $k \in K$, we let D_k denote the event under which for any number of samples $t \leq 4t_k$ from arm k, the number of samples which are on the interval $(-\infty, \overline{\mu}_k]$ is bounded as follows:

$$D_k \triangleq \left\{ \max_{1 \leq t \leq 4t_k} \sum_{i=1}^{t} \left(x_i^k - F_k\left(\overline{\mu}_k\right)\right) < 15 t_k F_k\left(\overline{\mu}_k\right) \right\}$$

where x_i^k is a RV which equals to 1 if the i-th sample from arm k is on that interval and 0 otherwise. Below we upper bound $P_0^H(D_k)$ using Kolmogorov's inequality.

Kolmogorov's inequality states that the sum $S_t = \sum_{i=1}^t z_i$ of zero-mean iid random variables (z_i) satisfies $P(\max_{1 \leq t \leq n} |S_t| \geq a) \leq \frac{\text{Var}[S_n]}{a^2}$ (Theorem 22.4, in p. 287 of (Patrick, 1995)). By applying it to the RVs $y_i^k = x_i^k - F_k(\overline{\mu}_k)$, we obtain

$$P_0^H\left(D_k^C\right) \leq \frac{Var(\sum_{i=1}^{4t_k} y_i^k)}{\left(15 t_k F_k\left(\overline{\mu}_k\right)\right)^2} = \frac{4 t_k F_k\left(\overline{\mu}_k\right)\left(1 - F_k\left(\overline{\mu}_k\right)\right)}{\left(15 t_k F_k\left(\overline{\mu}_k\right)\right)^2},$$

where D_k^C is the complementary of D_k.

So, for the case of $\mu_k^* < \mu^* + \epsilon - \epsilon_0$, by the fact that $F_k(\overline{\mu}_k) = 1$, it follows that $P_0^H(D_k) = 1$.

For the case of $\mu_k^* \geq \mu^* + \epsilon - \epsilon_0$, it follows that $G_*(\cdot) \leq 1$ by its definition, so, again by definition we obtain that $F_k(\overline{\mu}_k) \geq G_*(\mu^* + \epsilon - \mu) = \gamma_k$ and therefore $t_k F_k(\overline{\mu}_k) \geq \frac{1}{16} \ln\left(\frac{3}{20\delta}\right)$. So it follows that since $\delta \leq \frac{3}{20} e^{-3}$ by assumption $P_0^H(D_k) \geq 1 - \frac{64}{225 \ln\left(\frac{3}{20\delta}\right)} \geq \frac{9}{10}$. For simplicity, we use the bound $P_0^H(D_k) \geq \frac{9}{10}$ for every $k \in K$.

Define now the intersection event $S_k = A_k \cap B_k^C \cap C_k \cap D_k$. We have just shown that for every $k \neq k'$ it holds that $P_0^H(A_k) \geq \frac{3}{4}$, $P_0^H(B_k^C) > \frac{1}{2}$, $P_0^H(C_k) = 1$ and $P_0^H(D_k) \geq \frac{9}{10}$, from which it follows that $P_0^H(S_k) > \frac{3}{20}$ for $k \neq k'$.

Now, we let h be the history of the process (the sequence of chosen arms and obtained rewards). For every $k \in K$, we denote the number of rewards under $\overline{\mu}_k$ by N_k. For a given history, at time t', for every $k \in K$, the probability of choosing the next arm is the same under H_0 and under H_k. Also, by the hypotheses definition, the reward probability is the same, unless the chosen arm is k. Therefore, as under the event C_k, P_k^H is absolutely continuous w.r.t. P_0^H, by the definition of the hypotheses,

$$\frac{dP_k^H}{dP_0^H}(h) = \left(1 - \frac{\gamma_k}{F_k(\overline{\mu}_k)}\right)^{N_k}, \quad h \in C_k$$

where γ_k is defined before. Note that for $\mu_k^* < \mu^* + \epsilon - \epsilon_0$ it holds that $F_k(\overline{\mu}_k) = 1$ and it therefore follows that $\gamma_{k,1} = 1 - \frac{\gamma_k}{F_k(\overline{\mu}_k)}$. Also, note that for $\mu_k^* \geq \mu^* + \epsilon - \epsilon_0$ it follows that $\gamma_{k,2} = 1 - \frac{\gamma_k}{F_k(\overline{\mu}_k)}$.

Now we assume that the intersection event S_k occurs. Then, $\{A_k \cap D_k\}$ occurs, so $N_k \leq 16 t_k F_k(\overline{\mu}_k)$. Therefore, for $\alpha_k = \frac{\gamma_k}{F_k(\overline{\mu}_k)} \leq 1$,

$$\frac{dP_k^H}{dP_0^H}(h) \geq (1-\alpha_k)^{\frac{1}{\alpha_k}\ln\left(\frac{3}{20\delta}\right)}, \quad h \in S_k\,.$$

Now, by the fact that $(1-\epsilon)^{\frac{1}{\epsilon}} \geq e^{-1}$, we obtain the following inequalities,

$$\begin{aligned} P_k^H\left(B_k^C\right) \geq P_k^H\left(S_k\right) &= E_0^H\left[\frac{dP_k^H}{dP_0^H}(h) I\left(h \in S_k\right)\right] \\ &\geq E_0^H\left[(1-\alpha_k)^{\frac{1}{\alpha_k}\ln\left(\frac{3}{20\delta}\right)} I\left(h \in S_k\right)\right] \\ &\geq (1-\alpha_k)^{\frac{1}{\alpha_k}\ln\left(\frac{3}{20\delta}\right)} P_0^H\left(S_k\right) \\ &> \frac{3}{20} e^{-\ln\frac{3}{20\delta}} \geq \delta, \quad \forall k \neq k'\,. \end{aligned}$$

We found that if an algorithm is (ϵ, δ)-correct under hypothesis H_0 and $E_0[T_k] \leq t_k$ for some $k \neq k'$, then, under hypothesis H_k this algorithm returns a sample that is smaller by at least ϵ than the maximal possible reward with probability of δ or more, hence the algorithm is not (ϵ, δ)-correct. Therefore, any (ϵ, δ)-correct algorithm must satisfy $E_0[T_k] > t_k$ for all of arms except possibly for one (namely, for the one k' for which $P_0\left(B_{k'}^C\right) \leq \frac{1}{2}$, if such k' exists). In addition $t_{k^*} \geq t_{k'}$, where k^* is the optimal arm (namely, $\mu_{k^*}^* = \mu^*$). Hence,

$$E_0^H[T] \geq \sum_{k \in K \setminus \{k^*\}} \frac{1}{16 G_*(\min\left(\epsilon_0, \epsilon + \mu^* - \mu_k^*\right)} \ln\left(\frac{3}{20\delta}\right).$$

Now, by the fact that G_* is concave, it follows that $tG_*(y) + (1-t)G_*(0) \leq G_*(ty)$ where $y = \mu^* + \epsilon - \mu_k^*$. So, for the case of $\epsilon \leq \mu^* - \mu_k^*$, for $t = \frac{\epsilon}{y}$, by the fact that G_* is non-negative, it follows that $G_*(y) \geq 2G_*(\epsilon)$ and for the case of $\epsilon < \mu^* - \mu_k^*$, for $t = \frac{\mu^* - \mu_k^*}{y}$, it follows that $G_*(y) \geq 2G_*(\mu^* - \mu_k^*)$. Then since G_* is a non-decreasing function, the lower bound is obtained.

□

Algorithm 1 Maximal Confidence Bound (Max-CB) Algorithm

1: **Input:** The tail function bound $G_* = \{G_*(\epsilon'), 0 \leq \epsilon' \leq \epsilon_0\}$ and its inverse function G_*^{-1}, constants $\delta > 0$ and $\epsilon > 0$.
Define $L = 6\ln\left(|K|\left(1 + \frac{-\ln(\delta)}{G_*(\epsilon)}\right)\right) - \ln(\delta)$.
2: **Initialization:** Counters $C(k) = N_0$, $k \in K$, where $N_0 = \lfloor \frac{L}{G_*(\epsilon_0)} \rfloor + 1$.
3: Sample N_0 times from each arm.
4: Compute $Y_{C(k)}^k = V_{C(k)}^k + U(C(k))$ and set $k^* \in \arg\max_{k \in K} Y_{C(k)}^k$ (with ties broken arbitrary), where $V_{C(k)}^k$ is the largest reward observed so far from arm k and
$$U(C(k)) = G_*^{-1}\left(\frac{L}{C(k)}\right).$$
5: If $U(C(k^*)) < \epsilon$, stop and return the largest sampled reward.
Else, sample once from arm k^*, set $C(k^*) = C(k^*)+1$ and return to step 4.

4. Algorithm

Here we provide an (ϵ, δ)-correct algorithm. The algorithm is based on sampling the arm which has the highest upper confidence bound on its *maximal* reward.

The algorithm starts by sampling a fixed number of times from each arm. Then, it repeatedly calculates an index for each arm which can be interpreted as an upper bound on the maximal reward of this arm, and samples once from the arm with the largest index. The algorithm terminates when the number of samples from the arm with the largest index is above a certain threshold. This idea is similar to that in the UCB1 Algorithm of (Auer et al., 2002).

Theorem 2. *Under Assumption 1, for any $\epsilon \leq \epsilon_0$ and δ such that $L + \ln(\delta) \geq 10$, Algorithm 1 is (ϵ, δ)-correct with a sample complexity of*

$$E[T] \leq \sum_{k \in K} \frac{L}{G_*\left(\Theta_k\right)} + |K|, \tag{4}$$

where $L = 6\ln\left(|K|\left(1 + \frac{-\ln(\delta)}{G_*(\epsilon)}\right)\right) - \ln(\delta)$ as defined in the algorithm, and $\Theta_k = \min\left\{\max\left(\epsilon, \mu^* - \mu_k^*\right), \epsilon_0\right\}$.

As observed by comparing the bounds in Equations (3) and (4), the upper bound in Theorem 2 has the same depen-

dence of ϵ and $\ln(\delta^{-1})$, up to a logarithmic term. It should be noted though that while the lower bound is currently restricted to concave tail function bounds, the algorithm and its bound are not restricted to this case.

To establish Theorem 2, we first bound the probability of the event under which the upper bound of the best arm is below the maximal reward, using an extreme value bound. Then, we bound the largest number of samples after which the algorithm terminates under the assumption that the upper bound of the best arm is above the maximal reward.

Proof of Theorem 2. We denote the time step of the algorithm by t, the value of the counter $C(k)$ at time step t by $C^t(k)$ and $L' \triangleq L + \ln(\delta)$. Recall that T stands for the random final time step. By the condition in step 5 of the algorithm, for every arm $k \in K$, it follows that,

$$C^T(k) \leq \lfloor \frac{L' - \ln(\delta)}{G_*(\epsilon)} \rfloor + 1. \tag{5}$$

Note that by the fact that for $x \geq 6$ it follows that $\frac{d6\ln(x)}{dx} \leq 1$, and by the fact that for $x_0 = \exp\left(1\frac{2}{3}\right)$ it follows that $x_0 > 6\ln(x_0) = 10$ it is obtained that

$$\begin{aligned} L'' &\triangleq |K| \left(\frac{-\ln(\delta)}{G_*(\epsilon)} + 1 \right) \\ &> 6\ln\left(|K| \left(\frac{-\ln(\delta)}{G_*(\epsilon)} + 1 \right) \right) = L', \end{aligned}$$

for $L' \geq 10$. So, by the fact that $T = \sum_{k \in K} C^T(i)$, for $L' \geq 10$ it follows that

$$\begin{aligned} T &\leq |K| \left(\frac{L' - \ln(\delta)}{G_*(\epsilon)} + 1 \right) < |K| \left(\frac{L'' - \ln(\delta)}{G_*(\epsilon)} + 1 \right) \\ &\leq L''^2 = e^{\frac{L'}{3}}. \end{aligned} \tag{6}$$

Now, we begin with proving the (ϵ, δ)-correctness property of the algorithm. Recall that for every arm $k \in K$ the rewards are distributed according to the CDF $F_k(\mu)$. Let assume w.l.o.g. that $\mu_1^* = \mu^*$. Then, for $N > 0$ and by the fact that $(1-\epsilon)^{\frac{1}{\epsilon}} \leq e^{-1}$ for every $\epsilon \in (0, 1]$, for $U(N) = G_*^{-1}\left(\frac{L}{N}\right)$ it follows that

$$\begin{aligned} P\left(V_N^1 \leq \mu^* - U(N)\right) &= \left(F_1\left(\mu^* - U(N)\right)\right)^N \\ &\leq \left(1 - \left(\frac{L' - \ln(\delta)}{N}\right)\right)^N \\ &\leq \delta e^{-L'}, \end{aligned} \tag{7}$$

where V_N^k is the largest reward observed from arm $k \in K$ after this arm has been sampled for N times. Hence, at every time step t, by the definition of $Y_{C^t(1)}^1$ and Equations (6) and (7), by applying the union bound, it follows that

$$\begin{aligned} P\left(Y_{C^t(1)}^1 \leq \mu^*\right) &= P\left(V_{C^t(1)}^1 \leq \mu^* - U(C^t(1))\right) \\ &\leq \sum_{N=1}^{\exp\left(\frac{L'}{3}\right)} P\left(V_N^1 \leq \mu^* - U(N)\right) \\ &\leq \delta e^{-\frac{2L'}{3}}. \end{aligned} \tag{8}$$

Since by the condition in step 5, it is obtained that when the algorithm stops

$$V_{C^t(k^*)}^{k^*} > Y_{C^t(k^*)}^{k^*} - \epsilon,$$

and by the fact that for every time step

$$Y_{C^t(k^*)}^{k^*} \geq Y_{C^t(1)}^1,$$

it follows by Equation (8) that

$$P\left(V_{C^t(k^*)}^{k^*} \leq \mu^* - \epsilon\right) \leq P\left(Y_{C^t(1)}^1 \leq \mu^*\right) \leq \delta e^{-\frac{2L'}{3}}.$$

Therefore, it follows that the algorithm returns a reward greater than $\mu^* - \epsilon$ with a probability larger than $1 - \delta$. So, it is (ϵ, δ)-correct.

For proving the bound on the expected sample complexity of the algorithm we define the following sets:

$$M(\epsilon) = \{l \in K | \mu^* - \mu_l^* < \epsilon\}$$

and

$$N(\epsilon) = \{l \in K | \mu^* - \mu_l^* \geq \epsilon\}.$$

As before, we assume w.l.o.g. that $\mu_1^* = \mu^*$. For the case in which

$$E_1 \triangleq \bigcap_{1 \leq t < T} \left\{ Y_{C^t(1)}^1 \geq \mu^* \right\},$$

occurs, since $V_{C^t(k)}^k \leq \mu_k^*$ for every $k \in K$, and every time step, it follows that the necessary condition for sampling from arm k,

$$Y_{C^k(1)}^k \geq Y_{C^t(1)}^1,$$

occurs only when the event

$$E_2(t) \triangleq \left\{ \mu_k^* + U\left(C^t(k)\right) \geq \mu^* \right\},$$

occurs. But

$$E_2(t) \subseteq \left\{ C^t(k) \leq \frac{L' - \ln(\delta)}{G_*\left(\mu^* - \mu_k^*\right)} \right\}.$$

Therefore, it is obtained that

$$C^T(k) \leq \max\left(\lfloor \frac{L' - \ln(\delta)}{G_*\left(\mu^* - \mu_k^*\right)} \rfloor + 1, N_0 \right). \tag{9}$$

By using the bound in Equation (5) for the arms in the set $M(\epsilon)$, the bound in Equation (9) for the arms in the set $N(\epsilon)$ and the bound in Equation (6), it is obtained that

$$E[T] \leq (1 - P(E_1))\, e^{\frac{L'}{3}} + P(E_1)\, \Phi(\epsilon), \tag{10}$$

where

$$\Phi(\epsilon) \triangleq \sum_{k \in N(\epsilon)} \left(\lfloor \frac{L' - \ln(\delta)}{G_*(\min(\epsilon_0, \mu^* - \mu_k^*))} \rfloor + 1 \right) + \sum_{k \in M(\epsilon)} \left(\lfloor \frac{L' - \ln(\delta)}{G_*(\epsilon)} \rfloor + 1 \right).$$

In addition, by Equation (8), the bound in Equation (6) and by applying the union bound, it follows that

$$P(E_1) \geq 1 - \sum_{t=1}^{T} P\left(Y^1_{C^t(1)} < \mu^*\right) \geq 1 - \delta e^{-\frac{2L'}{3}} e^{\frac{L'}{3}} = 1 - \delta e^{-\frac{L'}{3}}.$$

So,

$$1 - P(E_1) \leq \delta e^{-\frac{L'}{3}}. \tag{11}$$

Furthermore, by the definitions of L', the sets $N(\epsilon)$ and $M(\epsilon)$ and since $\epsilon \leq \epsilon_0$, it can be obtained that

$$\Phi(\epsilon) \leq \sum_{k \in K} \lfloor \frac{L}{G_*(\Theta_k)} \rfloor + 1. \tag{12}$$

where $\Theta_k = \min\{\max(\epsilon, \mu^* - \mu_k^*), \epsilon_0\}$. Therefore, by Equation (10), (11) and (12) the bound on the sample complexity is obtained.

□

5. Robustness

The performance bounds presented for our algorithm depend directly on the choice of the lower bound G_* on the tail functions. A natural question is what happens if our choice of G_* is too optimistic, so that Assumption 1 is violated. In the opposite direction, how tight is our bound when our choice of G_* is to conservative? We address these two questions in turn.

5.1. Optimistic Tails Estimate

Here Equation (1) does not hold for $G_*(\epsilon)$, but holds for $G'_*(\epsilon) = \alpha G_*(\epsilon)$ for some $\alpha < 1$. The fact that Equation (1) does not hold for $G_*(\epsilon)$ leads to the situation in which the probability $P\left(Y^k_{C(k)} < \mu_k^*\right)$ is larger (where $Y^k_{C(k)}$ is the index calculated in step 4 of the algorithm) than the value on which the proof of Theorem 2 relies. In the following proposition we provide the (ϵ, δ)-correctness and sample complexity of Algorithm 1.

Proposition 1. *Suppose that Assumption 1 does not hold for $G_*(\epsilon)$, but holds for $G'_*(\epsilon) = \alpha G_*(\epsilon)$ for some $\alpha < 1$. Then Algorithm 1 is (ϵ', δ')-correct with*

$$\epsilon' = G_*^{-1}\left((|K|L)^{1-\alpha} (G_*(\epsilon))^{\alpha}\right) \text{ and } \delta' = \delta^{\alpha},$$

and sample complexity bound

$$E[T] \leq \sum_{k \in K} \frac{L}{G_*(\overline{\Theta}_k)} + \delta^{\alpha} \left(|K| \frac{L}{G_*(\epsilon)}\right)^{1-\alpha} + |K|,$$

where L is as in Theorem 2 and $\overline{\Theta}_k = \min\{\max(\epsilon, \mu^* - \epsilon' - \mu_k^*), \epsilon_0\}$.

The proof of the above proposition bases on the proof of Theorem 2 and is provided in Section 8.2 in the Appendix.

5.2. Conservative Tails Estimate

Here, Assumption 1 holds for the provided function $G_*(\epsilon)$ and also holds for $G'_*(\epsilon) = \alpha G_*(\epsilon)$ for some $\alpha > 1$. Therefore, in this case the probability $P\left(Y^k_{C(k)} < \mu_k^*\right)$ is smaller than the value on which the proof of Theorem 2 relies. So, Algorithm 1 returns an ϵ-optimal value with a larger probability. The probability of returning a false value is given in the following proposition.

Proposition 2. *When Assumption 1 holds for $G_*(\epsilon)$, and also for $G'_*(\epsilon) = \alpha G_*(\epsilon)$ for some $\alpha > 1$, and $L + \ln(\delta) \geq 10$, Algorithm 1 is (ϵ, δ')-correct where $\delta' = \delta e^{-(\alpha-1)L}$ (ϵ and δ are provided to the algorithm) with the sample complexity provided in Theorem 2*

For proving the above proposition we base on a minor variation of the proof of Theorem 2. The proof is provided in Section 8.3 in the Appendix.

6. Comparison with the Unified-Arm Model

In this section, we analyze the improvement in the sample complexity obtained by utilizing the multi arm framework (the ability to choose from which arm to sample at each time step) compared to a model in which all the arms are unified into a single arm, so that the sample is effectively obtained from a random arm. In the unified-arm model, when the agent samples from this unified arm, one of the original arms is chosen uniformly at random, and a reward is sampled from this arm. The CDF of the unified arm is therefore $F(\mu) = \frac{1}{|K|} \sum_{k \in K} F_k(\mu)$, and the corresponding maximal reward is $\mu^* = \max_k \mu_k^*$. Assumption 1, implies that $1 - F(\mu) \geq \frac{G_*(\mu^* - \mu)}{|K|}$.

In the remainder of this section, we provide a lower bound on the sample complexity and an (ϵ, δ)-correct algorithm that attains the same order of this bound for the unified-arm model. (Note that the lower bound in Theorem 1 is

Algorithm 2 Unified-Arm Algorithm

1: **Input:** The tail function bound $G_* = \{G_*(\epsilon'), 0 \leq \epsilon' \leq \epsilon_0\}$ and its inverse function G_*^{-1}, constants $\delta > 0$ and $\epsilon > 0$.
2: Sample $\lceil \frac{-\ln(\delta)|K|}{G_*(\epsilon)} \rceil + 1$ times from the unified-arm.
3: Return the best sample.

meaningless for $|K| = 1$.) Then, we discuss which approach (multi-armed or unified-arm) is better for different model parameters, and provide examples that illustrate these cases.

6.1. Lower Bound

The following Theorem provides a lower bound on the sample complexity for the unified-arm model.

Theorem 3. *For every (ϵ, δ)-correct algorithm, under Assumption 1, when $G_*(\epsilon)$ is concave, $\epsilon \leq \epsilon_0$ and $\delta \leq \frac{3}{20}e^{-3}$, it holds that*

$$E[T] \geq \frac{|K|}{16G_*(\epsilon)} \ln\left(\frac{3}{20\delta}\right). \tag{13}$$

The proof is provided in Section 8.4 in the Appendix and is based on a similar idea to that of Theorem 1.

6.2. Algorithm

In Algorithm 2, a fixed number of instances is sampled, and the algorithm chooses the best one among them. In the following Theorem we provide a bound on the sample complexity achieved by Algorithm 2.

Theorem 4. *Under Assumption 1, Algorithm 2 is (ϵ, δ)-correct, with a sample complexity bound of*

$$E[T] \leq \frac{|K|\ln(\delta^{-1})}{G_*(\epsilon)} + 2.$$

The proof is provided in Section 8.5 in the Appendix. Note that the upper bound on the sample complexity is of the same order as the lower bound in Theorem 3.

6.3. Comparison and Examples

To find when the multi-armed algorithm is useful, we may compare the upper bound on the sample complexity provided in Theorem 2 for Algorithm 1 (multi-armed case) with the lower bound for the unified-arm model in Theorem 3. We consider two extreme cases.

Case 1: Suppose that arm 1 is best: $\mu_1^* = \mu^*$, while all the other arms fall short significantly compared to the required accuracy ϵ: $\mu_k^* \ll \mu^* - \epsilon$, for $k \neq 1$.

Here $\frac{1}{\epsilon} \gg \frac{1}{\left(\max\left(\epsilon, \mu^* - \mu_k^*\right)\right)}$, for $k \neq 1$. Hence the upper bound on sample complexity of Algorithm 1 (multi-armed case) will be smaller than the lower bound for the unified-arm model in Theorem 3. We now provide an example which illustrates case 1 numerically.

Example 1 (Case 1)**.** *Let $|K| = 10^4$, $\mu_1^* = 0.9$, $\mu_k^* = 0.1\ \forall k \neq 1$, $G_*(\epsilon) = A\epsilon$ and $A = 0.01$. For $\epsilon = 10^{-4}$ and $\delta = 10^{-3}$ the sample complexity attained by Algorithm 1 is 3.52×10^8. The lower bound for the unified-arm model is 3.13×10^9. The sample complexity attained by Algorithm 2 (for the unified-arm model model) is 6.9×10^{10}.*

Case 2: Consider next the opposite case, where there are many optimal arms and few that are worse: say $\mu_1^* \ll \mu^* - \epsilon$, while $\mu_k^* = \mu^*$ for all $k \neq 1$.
Here $\frac{1}{\epsilon} = \frac{1}{\left(\max\left(\epsilon, \mu^* - \mu_k^*\right)\right)}$, for $k \neq 1$. Hence, since there is a logarithmic-in-$\frac{|K|}{\epsilon}$ multiplicative factor in the upper bound on the sample complexity of Algorithm 1, this bound will be larger than the lower bound for the unified-arm model in Theorem 3. The following example illustrates case 2 numerically.

Example 2 (Case 2)**.** *Let $|K|$, $G_*(\epsilon)$, δ and ϵ remain the same as in Example 1, and let $\mu_1^* = 0.1$ and $\mu_k^* = 0.9$ for $k \neq 1$. The sample complexity of Algorithm 1 is 1.56×10^{12}, which is larger than the sample complexity of Algorithm 2 which is 6.9×10^{10}.*

As shown in Example 2, in some cases the bound on the sample complexity of the multi-armed Algorithm 1 is larger than that of the unified-arm Algorithm 2. We shall further comment on these finding in our concluding remarks.

7. Conclusion

We have considered in this paper the Max K-armed Bandit problem in the PAC setting, under the assumption of a known lower bound on the tail function of reward distributions. We provided a lower bound on the sample complexity of any algorithm, and a UCB-type sampling algorithm whose sample complexity is essentially of the same order up to logarithmic terms.

We have further analyzed the robustness of our algorithm to the violation of Assumption 1 on the tail functions, and bounded the resulting deterioration in performance.

The performance of the multi-armed Algorithm 1 was compared to a simple unified-arm approach. The benefits of Algorithm 1, which aims to focus sampling on the best arms, are clear when there are few optimal arms (in term of their maximal reward), but might diminish when many arms are close to optimal. Combining these two approaches into a single algorithm that excels in either case remains a challenge for future works.

References

Amis, Eric J. Combinatorial materials science: Reaching beyond discovery. *Nature Materials*, 3(2):83–85, 2004.

Apostolidis, Athanasios, Klimant, Ingo, Andrzejewski, Damian, and Wolfbeis, Otto S. A combinatorial approach for development of materials for optical sensing of gases. *Journal of Combinatorial Chemistry*, 6(3): 325–331, 2004.

Audibert, Jean-Yves and Bubeck, Sébastien. Best arm identification in multi-armed bandits. In *COLT-23th Conference on Learning Theory*, pp. 41–53, 2010.

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

Berry, Donald A., Chen, Robert W., Zame, Alan, Heath, David C., and Shepp, Larry A. Bandit problems with infinitely many arms. *The Annals of Statistics*, 25(5): 2103–2116, 1997.

Bonald, Thomas and Proutiere, Alexandre. Two-target algorithms for infinite-armed bandits with Bernoulli rewards. In *Advances in Neural Information Processing Systems 26*, pp. 2184–2192. Curran Associates, Inc., 2013.

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

Carpentier, Alexandra and Valko, Michal. Extreme bandits. In *Advances in Neural Information Processing Systems 27*, pp. 1089–1097. Curran Associates, Inc., 2014.

Carpentier, Alexandra and Valko, Michal. Simple regret for infinitely many armed bandits. In *Proceedings of the 32nd International Conference on Machine Learning, ICML*, pp. 1133–1141, 2015.

Chakrabarti, Deepayan, Kumar, Ravi, Radlinski, Filip, and Upfal, Eli. Mortal multi-armed bandits. In *Advances in Neural Information Processing Systems 21*, pp. 273–280. Curran Associates, Inc., 2009.

Cicirello, Vincent A. and Smith, Stephen .F. The max k-armed bandit: A new model of exploration applied to search heuristic selection. In *Proceedings of the National Conference on Artificial Intelligence*, volume 20, pp. 1355–1361, 2005.

David, Yahel and Shimkin, Nahum. Infinitely many-armed bandits with unknown value distribution. In *Machine Learning and Knowledge Discovery in Databases*, pp. 307–322. Springer, 2014.

David, Yahel and Shimkin, Nahum. Refined algorithms for infinitely many-armed bandits with deterministic rewards. In *Machine Learning and Knowledge Discovery in Databases*, pp. 464–479. Springer, 2015.

Even-Dar, Eyal, Mannor, Shie, and Mansour, Yishay. PAC bounds for multi-armed bandit and markov decision processes. In *COLT-15th Conference on Learning Theory*, pp. 255–270. 2002.

Gabillon, Victor, Ghavamzadeh, Mohammad, and Lazaric, Alessandro. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems 25*, pp. 3212–3220. Curran Associates, Inc., 2012.

Kalyanakrishnan, Shivaram, Tewari, Ambuj, Auer, Peter, and Stone, Peter. PAC subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Conference on Machine Learning, ICML*, pp. 655–662, 2012.

Karnin, Zohar Shay, Koren, Tomer, and Somekh, Oren. Almost optimal exploration in multi-armed bandits. In *Proceedings of the 30th International Conference on Machine Learning, ICML*, pp. 1238–1246, 2013.

Kaufmann, Emilie, Cappé, Olivier, and Garivier, Aurélien. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

Mannor, Shie and Tsitsiklis, John N. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5:623–648, 2004.

Patrick, Billingsley. Probability and measure. *A Wiley-Interscience Publication, John Wiley & Sons, New York*, 1995.

Streeter, Matthew J. and Smith, Stephen F. An asymptotically optimal algorithm for the max k-armed bandit problem. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, pp. 135–142, 2006a.

Streeter, Matthew J. and Smith, Stephen F. A simple distribution-free approach to the max k-armed bandit problem. In *Proceedings of the 12th international conference on Principles and Practice of Constraint Programming*, pp. 560–574. Springer, 2006b.

Szörényi, Balázs, Busa-Fekete, Róbert, Weng, Paul, and Hüllermeier, Eyke. Qualitative multi-armed bandits: A quantile-based approach. In *Proceedings of the 32nd International Conference on Machine Learning, ICML*, pp. 1660–1668, 2015.

Teytaud, Olivier, Gelly, Sylvain, and Sebag, Michèle. Anytime many-armed bandits. In *CAP*, Grenoble, France, 2007.

Wang, Yizao, Audibert, Jean-Yves, and Munos, Rémi. Algorithms for infinitely many-armed bandits. In *Advances in Neural Information Processing Systems 21*, pp. 1729–1736. Curran Associates, Inc., 2008.