Bandits with Budgets: Regret Lower Bounds and Optimal Algorithms

Richard Combes, Chong Jiang, Srikant Rayadurgam

Bandits with Budgets: Regret Lower Bounds and Optimal Algorithms

Richard Combes
Centrale-Supelec, L2S
Gif-sur-Yvette, France
richard.combes@supelec.fr

Chong Jiang
Electrical and Computer Engineering
University of Illinois at Urbana-Champaign, USA
chongjiang@gmail.com

R. Srikant
Electrical and Computer Engineering
University of Illinois at Urbana-Champaign, USA
rsrikant@illinois.edu

ABSTRACT

We investigate multi-armed bandits with budgets, a natural model for ad-display optimization encountered in search engines. We provide asymptotic regret lower bounds satisfied by any algorithm, and propose algorithms which match those lower bounds. We consider different types of budgets: scenarios where the advertiser has a fixed budget over a time horizon, and scenarios where the amount of money that is available to spend is incremented in each time slot. Further, we consider two different pricing models, one in which an advertiser is charged for each time her ad is shown (i.e., for each impression) and one in which the advertiser is charged only if a user clicks on the ad. For all of these cases, we show that it is possible to achieve $O(\log(T))$ regret. For both the cost-per-impression and cost-per-click models, with a fixed budget, we provide regret lower bounds that apply to any uniformly good algorithm. Further, we show that B-KL-UCB, a natural variant of KL-UCB, is asymptotically optimal for these cases. Numerical experiments (based on a real-world data set) further suggest that B-KL-UCB also has the same or better finite-time performance when compared to various previously proposed (UCB-like) algorithms, which is important when applying such algorithms to a real-world problem.

Categories and Subject Descriptors

I.2.6 [**Computing Methodologies**]: Learning; G.3 [**Mathematics of Computing**]: Probability and Statistics

Keywords

ad-display optimization; search engines; multi-armed bandits; learning; budgets; UCB; KL-UCB

1. INTRODUCTION

The multi-armed bandit (MAB) involves an decision maker who samples from several statistical populations with unknown distributions (also called "arms"), with the goal of maximizing the cumulative sum of drawn samples (called the "rewards"). The objective is to minimize the *regret*, which is the difference between the sum of rewards obtained by a given sampling strategy, and that of the best sampling strategy if the distribution of each arm were known. The number of arms can be finite (discrete bandits), countably infinite (infinite-armed bandits) or uncountably infinite (continuous bandits). MABs are stylized models for sequential decision problems with uncertainty, featuring in particular the so-called "exploration-exploitation" trade-off. MABs have been an active subject of research since the 30's, [24], [21].

For discrete bandits with uncorrelated arms, a notable result is [20], showing that in the asymptotic regime $T \to \infty$ (with T denoting the time horizon), there exists a regret lower bound for any algorithm that achieves $O(\log(T))$ regret for any input distribution, and provides an algorithm whose regret matches this lower bound. Further research has provided computationally simple, asymptotically optimal algorithms [19], [12], [16], with good finite-time behaviour.

More recent research has focused on so-called structured MABs, where the unknown parameters of the problem (say the expected values of the arms) have a certain structure and lie in some set known to the decision maker. The goal is to quantify the performance gain due to a given type of structure, and both regret lower bounds and asymptotically optimal algorithms have been proposed for certain structures. Structured MABs are interesting because they naturally arise in the design of computer systems (at large), for instance: wireless networks [9], shortest-path routing [13], search engines [23] and ad-display optimization [22].

For discrete bandits several structures have been studied: unimodal [7], combinatorial [5], arms with lower bounded differences [4] to name but a few. Continuous bandits are by definition bandits with correlated arms, since the expected reward (as a function of the arm) is assumed to be continuous. Many natural structures have been considered, including: Lipshitz continuous [17], unimodal [28], strongly convex [10].

In this paper we study the problem of discrete MABs with budgets, where the number of times a given arm may be selected is upper bounded by a number called the budget. The budget of an arm need not be deterministic: it may be a random variable, and may depend on the sample path (the successive rewards of arms). MABs with budgets are a natural model for ad-display optimization (e.g., Google Ad-Words). Given a search query, several advertisers would like to display an ad and the search engine must choose which ad to display. The chosen ad is displayed to a user (this is termed an impression) who may or may not click on it. The corresponding advertiser is charged either when her ad is shown (cost-per-impression), or clicked (cost-per-click). Each advertiser has a maximal amount of money she can spend, so that any ad cannot be displayed infinitely many times. Uncertainty is due to the fact

SIGMETRICS'15 June 15 - 19, 2015, Portland, OR, USA

ACM 978-1-4503-3486-0/15/06 ...$15.00.
http://dx.doi.org/10.1145/2745844.2745847.

that the probability for a given ad to be clicked (also known as the click-through-rate or CTR) is unknown and must be learnt.

Our model is a generalization of the models considered in [22],[14]. We will consider three cases:

- Cost-Per-Impression (CPI): an arm may be played a deterministic number of times
- Cost-Per-Click (CPC): an arm may be played until its accumulated reward is above a deterministic number
- General budgets: the maximal number of plays of an arm is an arbitrary non-decreasing function of time, and may depend on the sample path

It is noted that the general budgets assumption allows for feedback. This is well-suited for ad-display optimization because in practice advertisers may change their future budget allocations based on the historical (sample path) click-through-rates.

Our contribution

(a) For general budgets, we demonstrate that B-KL-UCB, a natural variant of KL-UCB, achieves $O(\log(T))$ regret, improving the results of [2], [22] which give an upper bound of $O(\sqrt{T})$. The proof uses a coupling argument, showing that, when we consider an arbitrary algorithm π and the optimal algorithm $\pi^\star$ run on the same sample path, at any given time, the expected value of the best arm available to π is higher than that available to $\pi^\star$. This induces a type of majorization order that allows us to prove the result.

(b) Next we consider the CPC and CPI case where the budget of each arm is a linear function of the time horizon, and we prove asymptotic (when the time horizon goes to infinity) regret lower bounds satisfied by any algorithm achieving $O(\log(T))$ regret regardless of problem parameters. The technique for proving the lower bound is different than the one introduced by Lai and Robbins in the seminal paper [20], and uses an inequality of [27] by reducing the problem to a single classical hypothesis test at the end of the time horizon. This technique might be useful beyond the scope of this article, as it renders the proofs significantly shorter than the original one proposed by Lai and Robbins.

(c) We provide finite-time regret upper bounds for B-KL-UCB. As a consequence, we prove that B-KL-UCB is asymptotically optimal in the CPI case, as well as in the CPC case if a simple separation assumption on the budgets is satisfied (which would most likely be the case in practice). For instance the set of budget vectors that do not satisfy this condition has Lebesgue measure zero.

(d) We assess the finite-time performance of B-KL-UCB using numerical experiments. The simulation parameters (number of advertisers and CTRs) are extracted from a publicly available data set [1]. We confirm the intuition provided by our theoretical results that B-KL-UCB works significantly better than UCB-type algorithms based on Hoeffding's inequality (such as the ones proposed in [14, 22]) which do not take into account the variance of the rewards, and lower bound the Kullback-Leibler (KL) divergence by twice the square distance (Pinsker's inequality). Indeed, in practice the values of the arms are small (most popular ads have a CTR of 2% or less), hence have low variance when they are modeled as Bernoulli random variables. An algorithm which is a heuristic modification of the PD-BwK algorithm proposed in [2] performs similarly to B-KL-UCB in simulations, although it lacks a corresponding problem-dependent regret bound and additionally requires knowledge of the time horizon.

Related Work

First, for arbitrarily large budgets, the problem reduces to the classical multi-armed bandit problem [20], and B-KL-UCB reduces to KL-UCB, which is known to be asymptotically optimal for that problem. The regret lower bounds also reduce to the classical one from [20]. Also, one can notice that bandits with budgets are an instance of sleeping bandits [18], which are bandit problems where not all arms may be selected at a given time. However, in [18], the available arms are chosen by an oblivious adversary, so that arms available at a given time are arbitrary but *may not depend on the arms selected previously*. Hence there is no straightforward extension of [18] to our setting. A different but related setting is that of bandits with a single knapsack constraint, considered in [25, 26]. Namely, all arms may be played until a weighted sum (with known weights) of the number of draws of each arms exceeds a known constant. The crucial difference is that in this model the optimal policy draws *a single arm* (maximizing the ratio between its expected reward and its weight), while in our setting the optimal policy (in general) plays several arms.

Another related problem is the knapsack bandit studied in [2]. There are several constraints on the weighted sum of rewards obtained on the different arms. Any arm might be selected, until one of the constraints is violated, and then the problem stops. There is a similarity between knapsack bandits and bandits with budgets (explored in the simulations section). However the results of [2] are quite different from ours: [2] considers minimax regret (for a given T, the regret on the worst problem instance, which in general depends on T), while we study problem-dependent regret, where a fixed instance is considered and we study the regret as T goes to infinity (as in [20]). Specifically, the authors obtain a minimax regret of $O(\sqrt{T})$ (up to multiplicative logarithmic terms). It is also noted that the algorithms in [2] rely on knowledge of T, whereas our algorithm does not.

The rest of the paper is organized as follows: in section 2 we define the model considered for bandits with budgets. In section 3 we prove that the optimal policy for each of the models considered here is the greedy policy (i.e. the one which plays the available arm with the highest expected value). In section 4 we provide lower bounds on the regret of any uniformly good algorithm in the CPI and CPC case. In section 5 we provide regret upper bounds for algorithm B-KL-UCB and demonstrate its asymptotic optimality in the CPI and CPC cases. In section 6 we assess the finite time performance of B-KL-UCB and its competitors by numerical experiments. Section 7 concludes the paper. For ease of reading, proofs and intermediate results are found in section 10. Some additional intermediate results are found in the appendix.

2. THE MODEL

We consider a bandit problem with a finite number of arms $K \geq 1$ and time horizon $T \geq 0$. Time is discrete, at time $n \in \{1, ..., T\}$ a decision maker is provided with a set of allowed arms $A(n) \subset \{1, ..., K\}$, and selects an arm $k(n) \in A(n)$. Then she receives a reward $X_{k(n)}(t_{k(n)}(n))$, where $t_k(n)$ is the number of times arm k has been selected between time 1 and n. We assume that the rewards $(X_k(i))_{1\leq k\leq K, i\geq 0}$ are independent, and that $X_k(i)$ is a Bernoulli random variable with parameter μ_k. We define $S_k(t) = \sum_{i=1}^{t} X_k(i)$ to be the accumulated rewards obtained from arm k after selecting it t times. We denote by $\mu = (\mu_1, \ldots, \mu_K) \in [0,1]^K$ the parameters of the problem. We assume that there exists functions $n \mapsto c_k(n)$ called budgets, so that the allowed set of arms can be written as: $A(n) = \{1 \leq k \leq K : t_k(n) \leq c_k(n)\}$. It is noted that $c_k(n)$ is not assumed to be deterministic and is possibly sample path dependent. We call sample path dependent any quantity that depends on the rewards $(X_k(i))_{1\leq k\leq K, i\geq 0}$.

We will consider three possible models for the availability of arms:

- Cost-per-impression (CPI): $c_k(n) = Tc_k$ for all $n \leq T$ with $c_k \geq 0$ a constant.
- Cost-per-click (CPC): $c_k(n) = \tau_k$ for all $n \leq T$, where $\tau_k = \min\{t : S_k(t) \geq Tc_k\}$ and $c_k \geq 0$ a constant.
- General budgets: $n \mapsto c_k(n)$ an increasing, possibly sample path dependent function.

We denote by $\mathcal{F}_n$ the σ-algebra generated by

$$\{A(1), \ldots, A(n+1), X_{k(1)}(t_{k(1)}(1)), \ldots, X_{k(n)}(t_{k(n)}(n))\}.$$

We consider adaptive policies, so that $k(n)$ is $\mathcal{F}_{n-1}$ measurable for all n. We denote by Π the set of adaptive policies. When the decision rule considered is not clear from a context we denote it with a superscript, for instance $k^\pi(n)$ is the arm selected at time n by policy $\pi \in \Pi$. We define $\pi^\star$ to be the oracle policy (which knows μ) and maximizes the expected accumulated sum of rewards: $\sum_{k=1}^K \mu_k \mathbb{E}[t_k^{\pi^\star}(T)]$. We further define the *regret* of decision rule π by:

$$R^\pi(T) = \sum_{k=1}^K \mu_k \mathbb{E}[t_k^{\pi^\star}(T)] - \sum_{k=1}^K \mu_k \mathbb{E}[t_k^\pi(T)].$$

The regret of policy π is the loss in accumulated reward due to the the fact that parameters μ are unknown to π. We say that policy π is uniformly good if, for all problem instances, $R^\pi(T) = O(\log(T))$ when $T \to \infty$.

In this article we present our results when rewards are Bernoulli distributed, mainly for simplicity and due to the fact that the model originates from ad-display optimization where rewards (click / no click) are indeed Bernoulli distributed. However, it should be clear that the regret upper bounds apply without modification to any bounded reward distribution in $[0, 1]$. Furthermore, both upper and lower bounds hold for rewards in a one-dimensional exponential family, (provided that they are sub-Gaussian), by replacing the Bernoulli KL divergence with the appropriate divergence measure. For instance, for Gaussian rewards with known variance, our results hold where the KL divergence is taken equal to the square distance divided by twice the variance. See the discussion in [12] for additional clarification.

3. PRELIMINARY RESULTS

3.1 Some notations

We assume that the arms are indexed such that $\mu_1 > \ldots > \mu_K$. For both the CPI and CPC cases we define $c = (c_1, \ldots, c_K)$ to be the budget vector. We define $I(p, q) = p\log(\frac{p}{q}) + (1-p)\log(\frac{1-p}{1-q})$ to be the KL divergence between Bernoulli distributions of parameters p and q. We use the convention that the value of an empty sum is zero, so that $\sum_{k'=1}^0 ... = 0$.

3.1.1 CPI case

In the CPI case we define $k^\star = \min\{k : \sum_{k'=1}^k c_{k'} \geq 1\}$ to be the last arm played by a greedy policy with knowledge of the μ_k's, which would play the arms in increasing order (until their respective budgets are exhausted). We define the fraction of time that such a policy would play arm $k^\star$: $\overline{c} = 1 - \sum_{k'=1}^{k^\star-1} c_{k'}$. It is noted that $\overline{c} > 0$, and that $\overline{c} = 1$ if $k^\star = 1$.

3.1.2 CPC case

Consider the CPC case. We recall the definition of τ_k, $\tau_k = \min\{t : S_k(t) \geq Tc_k\}$ which is the number of plays of arm k until Tc_k successes are realized. We define the random variable $\tilde{k}$ to be the last arm played by a policy which would play the arms in increasing order (until their respective budgets are exhausted):

$$\tilde{k} = \begin{cases} \min\{k : \sum_{k'=1}^k \tau_{k'} \geq T\}, & \text{if } \sum_{k=1}^K \tau_k \geq T \\ K, & \text{otherwise} \end{cases}.$$

We define the random variable $\overline{\tau}$ to be the number of plays of arm $\tilde{k}$: $\overline{\tau} = T - \sum_{k=1}^{\tilde{k}-1} \tau_k$. It is noted that $\overline{\tau} = T$ if $\tilde{k} = 1$.

We will relate these random quantities to the deterministic quantities obtained by taking expectations over sample paths. That is, we define $d_k = c_k/\mu_k$, the expected fraction of time that arm k could possibly be played, so that a CPI model with budgets of Td_k emulates this CPC model with budgets of Tc_k. We then define $k^\star = \min\{k : \sum_{k'=1}^k d_{k'} \geq 1\}$, the last arm played by the greedy policy with knowledge of the μ_k's, modulo the randomness in the budgets. It is noted that the definition of $k^\star$ is not the same for CPI and CPC. Finally, we define $\overline{d} = 1 - \sum_{k=1}^{k^\star-1} d_k$, the fraction of time that such a policy would play arm $k^\star$.

3.1.3 High probability events

We use the following convention throughout the remainder of the article: For a given event $\mathcal{A}$, we say that $\mathcal{A}$ occurs with high probability (w.h.p.) iff there exists a function $p_\mathcal{A}(\mu, c)$ such that for all T: $1 - \mathbb{P}[\mathcal{A}] \leq p_\mathcal{A}(\mu, c)T^{-1}$. Also we say that $\mathcal{A}$ occurs with small probability if its complement occurs w.h.p. It is noted that any event that occurs with small probability incurs only a *constant regret*. Denote by $r(T)$ the regret of a sample path, and consider $\mathcal{A}$ an event that occurs w.h.p., then, since $r(T) \leq T$:

$$\begin{aligned} R^\pi(T) = \mathbb{E}[r(T)] &= \mathbb{E}[r(T)\mathbf{1}\{\mathcal{A}\}] + \mathbb{E}[r(T)\mathbf{1}\{\mathcal{A}^c\})] \\ &\leq \mathbb{E}[r(T)\mathbf{1}\{\mathcal{A}\}] + p_\mathcal{A}(\mu, c). \end{aligned}$$

Hence given an event $\mathcal{A}$ which occurs with small probability, when analysing the regret of algorithms, one may simply ignore any sample path on which $\mathcal{A}$ occurs, at the expense of a constant regret term.

3.2 Optimal policy

In the case of general budgets, calculating the expected reward of the optimal policy is not completely straightforward. This is due to the fact that the set of available arms $A(n)$ is a random variable, and depends on the arms selected at instants $\{1, \ldots, n-1\}$ as well as the rewards $(X_k(i))_{1 \leq k \leq K, i \geq 0}$.

Define $\hat{\pi}$ to be the (greedy) policy which plays the arms in increasing order until their budgets are exhausted, i.e., $k^{\hat{\pi}}(n) = \min A^{\hat{\pi}}(n)$. It turns out that in the general budgets case (so in the CPI and CPC cases as well), we have that $\pi^\star = \hat{\pi}$ from which we can characterize the value of $\pi^\star$.

PROPOSITION 1. *For general budgets, we have that $\pi^\star = \hat{\pi}$, i.e. the greedy policy is optimal.*

In the CPI case, the reward of $\pi^\star$ is $\overline{R}T$ with:

$$\overline{R} = \sum_{k=1}^{k^\star-1} c_k\mu_k + \overline{c}\mu_{k^\star}.$$

In the CPC case the expected accumulated reward of $\pi^\star$ is:

$$\mathbb{E}[\mu_{\tilde{k}}\overline{\tau} + \sum_{k=1}^{\tilde{k}-1} \mu_k\tau_k].$$

4. REGRET LOWER BOUNDS

To simplify the regret lower and upper bounds we define $\Delta = \min_{k \neq k'} |\mu_k - \mu_{k'}|$. For $0 < \epsilon < \Delta$ we define:

$$\delta_k^\epsilon = \sum_{k'>k} \frac{\mu_k - \mu_{k'}}{I(\mu_{k'} + \epsilon, \mu_k)}.$$

with the convention that $\delta_k = \delta_k^0$.

Theorems 4.1 and 4.2 give lower bounds on the regret of *any* uniformly good algorithm.

THEOREM 4.1. *Consider the CPI case. For any uniformly good policy $\pi \in \Pi$, we have that for all $k > k^\star$:*

$$\liminf_{T\to\infty} \frac{\mathbb{E}[t_k^\pi(T)]}{\log(T)} \geq \frac{1}{I(\mu_k, \mu_{k^\star})}.$$

By corollary the regret satisfies the lower bound:

$$\liminf_{T\to\infty} \frac{R^\pi(T)}{\log(T)} \geq \delta_{k^\star}$$

THEOREM 4.2. *Consider the CPC case. For any uniformly good policy $\pi \in \Pi$, we have that for all $k > k^\star$:*

$$\liminf_{T\to\infty} \frac{\mathbb{E}[t_k^\pi(T)]}{\log(T)} \geq \frac{1}{I(\mu_k, \mu_{k^\star})}.$$

By corollary the regret satisfies the lower bound:

$$\liminf_{T\to\infty} \frac{R^\pi(T)}{\log(T)} \geq \delta_{k^\star}$$

For both the CPI and CPC cases, it is noted that arms $k \leq k^\star$ do not contribute to the regret lower bound, and that the minimal number of times an arm $k > k^\star$ may be played depends only on its expected value and the value of $\mu_{k^\star}$. In fact it is as if arms below $k^\star$ do not matter at all for our analysis. This will be made clear in light of the matching upper bounds derived in section 5. Furthermore, note that when the budgets are large enough, (for instance by setting $c_1 = 1$ in the CPI case, and letting $c_1 \to \infty$ in the CPC case), we have that $k^\star = 1$, so that Theorems 4.1 and 4.2 reduce to the well known result of Lai and Robbins [20].

The proof technique is similar to that of [4, 3, 6], and uses a reduction to a hypothesis test between two point hypotheses (a Neyman-Pearson test). However, the way in which we choose our hypothesis test is able to precisely recover the Lai and Robbins lower bound in [20], whereas the results in [4] do not do so. In particular, consider a given uniformly good algorithm π and two parameters μ and λ such that π must have a different behaviour under μ and λ. Say π plays a certain arm $O(T)$ times under λ, but only $O(\log(T))$ times under μ. Then we argue that the algorithm must be a hypothesis test with risk $O(T^{-1})$ between hypotheses $H_0 = \{\mu\}$ and $H_1 = \{\lambda\}$. Of course the original proof [20] used such an argument, but involved some manipulations of likelihood ratios, whereas we use an inequality of [27] which reduces these calculations to essentially a single line. Also note that, contrary to [4], we do not treat the arms played at times $n \in \{1, ..., T\}$ as a series of tests, but simply argue that the number of times each arm has been sampled by the end of the time horizon $(t_1(T), ..., t_K(T))$ can be used as a test statistic.

Finally, it should be noted that both Theorems 4.1 and 4.2 are still valid when the rewards are not Bernoulli, and instead belong to a parametric family of distributions for which one can define the KL divergence. In that case one may simply replace the Bernoulli KL divergence $I(\cdot, \cdot)$ by the relevant divergence measure, e.g., for Gaussian rewards with fixed variance one may replace $I(\cdot, \cdot)$ by the square distance divided by twice the variance.

5. REGRET UPPER BOUNDS

In this section we analyse the regret of B-KL-UCB, an algorithm which is asymptotically optimal (in most cases of interest), i.e., its regret matches the lower bounds given in section 4. It is a natural extension of KL-UCB [12] proposed for bandits with independent arms, which reaches the Lai-Robbins bound [20].

We define the empirical reward of arm k at time n: $\hat{\mu}_k(n) = S_k(t_k(n))/t_k(n)$ if $t_k(n) > 0$ and $\hat{\mu}_k(n) = 0$ otherwise. We introduce the (KL-UCB) index of arm k at time n:

$$b_k(n) = \sup\{q \in [\hat{\mu}_k(n), 1] : t_k(n) I(\hat{\mu}_k(n), q) \leq f(n)\},$$

with $f(n) = \log(n) + 3\log(\log(n))$. The B-KL-UCB algorithm is the rule that picks the *available arm* with largest index:

Algorithm 1 B-KL-UCB

for $n = 1, 2, \ldots, T$ **do**
  pull arm $k(n) = \arg\max_{k \in A(n)} b_k(n)$
end for

5.1 General budgets

Theorem 5.1 proves that B-KL-UCB achieves $O(\log(T))$ regret in the general budgets case. This in particular proves that in the gradual budget case considered in [14] (where $c_k(n)$ is deterministic and proportional to n), we also have $O(\log(T))$ regret, which is an improvement on the $O(\sqrt{T})$ upper bound derived in [14].

The proof is based on the following coupling argument: we show that if π = B-KL-UCB and the optimal policy $\pi^\star$ are run on the same sample path, then we have that, at all time instants n, $\min A^\pi(n) \leq \min A^{\pi^\star}(n)$. Hence either $k^\pi(n) \leq k^{\pi^\star}(n)$, which incurs no regret, or we have that $k^\pi(n) > k^{\pi^\star}(n) \geq \min A^{\pi^\star}(n) \geq \min A^\pi(n)$, which happens only $O(\log(T))$ times. We derive and use Lemma 10.3, an intermediate result shown in appendix. Lemma 10.3 enables us to deal with bandit problems where the available set of arms is a stochastic process and might depend on the past decisions, hence we believe it could be useful beyond the scope of this article, to analyse problems such as sleeping bandits [18] and knapsack bandits [2].

THEOREM 5.1. *Consider general budgets. Under policy π = B-KL-UCB, for all $0 < \epsilon < \Delta$ the regret admits the upper bound:*

$$R^\pi(T) \leq f(T) \sum_{k=2}^{K} \frac{\mu_1 - \mu_k}{I(\mu_k + \epsilon, \mu_{k-1})} + CK(\log(\log(T)) + \epsilon^{-2}).$$

with $C > 0$ a constant independent of μ, c and ϵ.

5.2 CPI case

Theorem 5.2, gives a finite-time regret upper bound for B-KL-UCB in the CPI case, from which we can deduce that B-KL-UCB is asymptotically optimal.

THEOREM 5.2. *(i) Under policy π = B-KL-UCB, for all $0 < \epsilon < \Delta$ the regret admits the upper bound:*

$$R^\pi(T) \leq f(T)\delta_{k^\star}^\epsilon + CK(\log(\log(T)) + \epsilon^{-2}) + C_0(c, \mu).$$

with $C > 0$ a constant independent of μ, c and ϵ, and $C_0(c, \mu) > 0$ a function independent of T and ϵ.

(ii) By corollary:

$$\limsup_{T\to\infty} \frac{R^\pi(T)}{\log(T)} \leq \delta_{k^\star},$$

i.e., B-KL-UCB is asymptotically optimal.

REMARK 1. *Note that Theorem 5.2 is not simply a specialization of Theorem 5.1 to the the CPI case, as the coefficients of $f(T)$ are different in the two cases. In particular, there is no proxy for $k^\star$ in the general budget case, whereas we exploit the existence of $k^\star$ to tighten the upper bound in Theorem 5.2.*

5.3 CPC case

Theorem 5.3, gives a finite-time regret upper bound for B-KL-UCB in the CPC case, from which we can deduce that B-KL-UCB is asymptotically optimal. In the derived regret upper bound, the dominant term (the multiplicative term in front of the $\log(T)$) is a convex combination of $\delta_{k^\star}$ and $\delta_{k^\star+1}$. By Theorem 5.2, those quantities represent the asymptotic regret in the CPI case where the last played arm by $\hat{\pi}$ is $k^\star$ and $k^\star+1$ respectively. Furthermore if we add the separation assumption $\sum_{k=1}^{k^\star} d_k > 1$, then the asymptotic regret is that of the CPI case. Since the regret lower bound of theorem 4.2 is met by the upper bound, B-KL-UCB is asymptotically optimal.

The proof of Theorem 5.3 involves upper bounding the number of times a sub-optimal arm might be played, and we do so by decomposing this number based on the expected value of the best arm available (i.e $\min A(n)$). As in the general budgets case, Lemma 10.3 is instrumental here. The proof is completed by studying the concentration of τ_k and $\tilde{k}$ and $\overline{\tau}$, based on classical concentration inequalities.

THEOREM 5.3. *(i) Under policy π = B-KL-UCB, there exists $\alpha(T) \in [0,1]$ such that, for all $0 < \epsilon < \Delta$ the regret admits the upper bound:*

$$R^\pi(T) \leq f(T)\left[\alpha(T)\delta^\epsilon_{k^\star} + (1-\alpha(T))\delta^\epsilon_{k^\star+1}\right] + CK(\log(\log(T)) + \epsilon^{-2}) + C_1(c,\mu),$$

with $C > 0$ a constant independent of μ, c and ϵ, and $C_1(c,\mu) > 0$ a function independent of T and ϵ.

(ii) By corollary:

$$\lim\sup_{T\to\infty} \frac{R^\pi(T)}{\log(T)} \leq \max(\delta_{k^\star}, \delta_{k^\star+1}).$$

(iii) If $\sum_{k=1}^{k^\star} d_k > 1$ we have $\alpha(T) \to_{T\to\infty} 1$ so that

$$\lim\sup_{T\to\infty} \frac{R^\pi(T)}{\log(T)} \leq \delta_{k^\star},$$

i.e., B-KL-UCB is asymptotically optimal.

6. NUMERICAL EXPERIMENTS

6.1 Data set and simulation parameters

We now compare the finite-time performance of B-KL-UCB with that of previously proposed algorithms. The simulation parameters, namely the values of K (the number of arms) and μ (the vector of reward probabilities), are extracted from a publicly available data set [1]. The data set describes user queries and displayed ads for a popular search engine, over the course of one day.

For our purposes, this dataset is a set of keywords, each containing a set of ads. Each ad has been subject to some number of impressions, a fraction of which have resulted in clicks. These simulations will use the empirical CTRs based on a keyword from this dataset. Since the number of ad impressions in the dataset is heavily skewed, using the click-through rate of an ad with only a few impressions would be prone to quantization effects (e.g., many arms with CTRs of exactly $\frac{1}{2}, \frac{1}{3}, \ldots$), so we first prune away any ad with fewer than 100 impressions. The histogram of click-through rates is shown in Figure 1. Indeed, the CTRs tend to be small. We filter the keywords present in the data set, and select those which contain at least 3 ads, 10^5 total impressions across those ads, and an overall click-through rate (total number of clicks divided by total number of impressions) of at least 1%. We chose keyword id #158 in the dataset, which we will refer to as keyword β. We then set K to be the number of different ads that have been displayed when β was requested, and for $1 \leq k \leq K$, we estimate μ_k by the empirical click probability for k, that is the number of clicks on k divided by the number of impressions for k. We obtain $K = 28$, and the values of $\mu_1, ..., \mu_K$ are shown in Figure 2 and Table 1.

Please note that the data is anonymized, so that each keyword and each ad is represented as a number, from which it is not possible to retrieve the actual query or the identity of the advertiser. The values of the budgets c are not available, so in the simulations to follow, we extract from the data only the K and μ of keyword β, and assign an equal budget to every arm. The budget is used as a parameter in our simulations, since it is unknown.

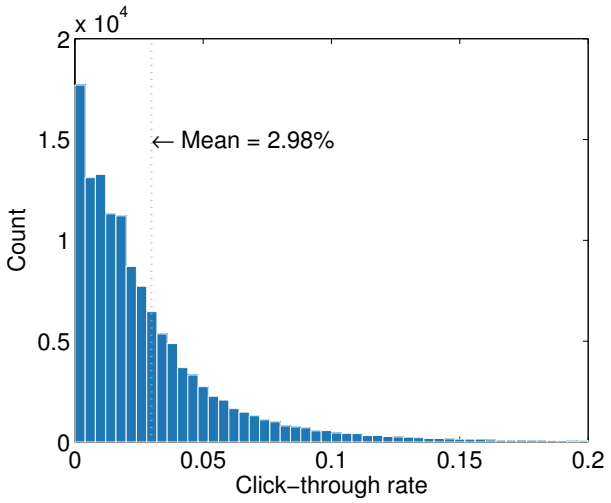

Figure 1: Histogram of CTRs for all ads with $\geq$ 100 impressions, from the KDD Cup dataset.

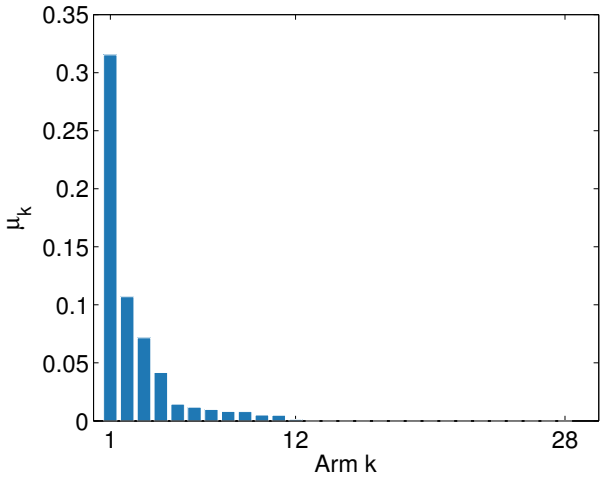

Figure 2: Plot of μ_k vs. k for the 28 ads with keyword β.

0.3153	0.1070	0.0716	0.0417	0.0144	0.0118
0.0099	0.0082	0.0081	0.0050	0.0049	0.0013

Table 1: List of the 12 non-zero entries in μ for keyword β.

6.2 Competing algorithms

We assess the performance of several algorithms identified as follows:

- B-KL-UCB: The algorithm proposed in this article.
- B-UCB1: The algorithms proposed in [14], [22]. It is noted that the two algorithms are not identical, but are nearly so. Roughly, those algorithms behave the same as B-KL-UCB except that the KL-UCB index $b_k(n)$ is replaced by the UCB index $\hat{\mu}_k(n) + \sqrt{2\log(n)/t_k(n)}$. Since they give the same performance, we only show the performance of one of them in the interest of readability.
- Balance-BwK (Balance Bandits with Knapsacks): an adaptation of the first algorithm proposed in [2] to bandits with budgets.
- PD-BwK (Primal Dual Bandits with Knapsacks): an adaptation of the second algorithm proposed in [2] to bandits with budgets.

In the knapsack bandit problem studied in [2], there are multiple resources and each arm consumes some combination thereof. The problem terminates when any one of the resources is exhausted. This is somewhat similar to our problem, where each arm's budget can be thought of as a resource. However, in our problem, even if the budget for one of the arms is exhausted, we can continue to play the other arms. Thus, while the algorithms in [2] do not directly apply to our model, nevertheless we attempt to modify those algorithms to fit our model and study how well they perform compared to our algorithm. In particular, the Balance-BwK and PD-BwK algorithms we consider here are tuned versions of the original algorithms proposed in [2], which take into account the additional structure. Namely, there are fewer unknown parameters in a problem instance of bandits with budgets than in bandits with knapsacks, e.g. resource k is known a priori to be consumed only when arm k is played. For completeness we provide a full description, including pseudo-code, of the tuned versions of Balance-BwK and PD-BwK in subsection 6.4.

6.3 Numerical results

The regret of each studied algorithm is calculated by averaging its sample path regret over 4000 independent runs.

First, we investigate the regret $R^\pi(T)$ as a function of the arm budgets (which determine $k^\star$). We fix a time horizon of $T = 1000K = 28000$. We consider uniform budgets so that for all k and n, $c_k(n) = cT$ where c is a parameter. We calculate the regret as a function of c.

Recall that for large budgets, the problem reduces to the classical bandit problem (and $k^\star = 1$). As budgets decrease, $k^\star$ transitions to $2, 3, \ldots, K$. We plot the regret of the various algorithms as we change the budget, in Figure 3 for the CPI model and in Figure 4 for the CPC model. These results show B-KL-UCB out-performs the other three algorithms across the entire range of $k^\star$, although our variant of PD-BwK stays a close second.

Next, we investigate the regret $R^\pi(T)$ as a function of the time horizon T. In order to fix $k^\star$ while letting time progress, the budgets must grow linearly with time. Instead of restarting the simulation with different budgets and time horizons, for simplicity of

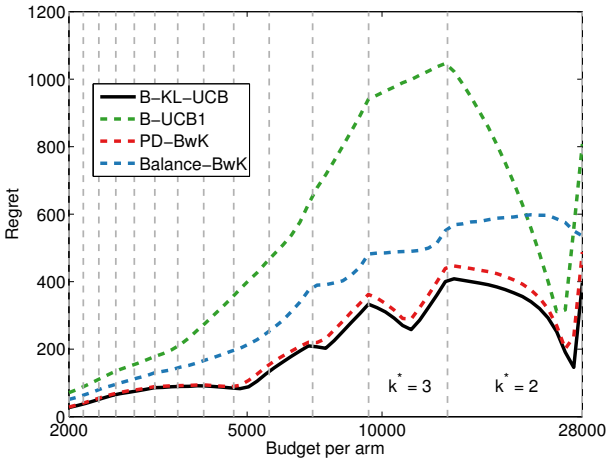

Figure 3: Plot of regret at time $T = 28000$ vs. the budget Tc given to each arm, under the CPI model. The dotted vertical lines demarcate $k^\star$ transitions.

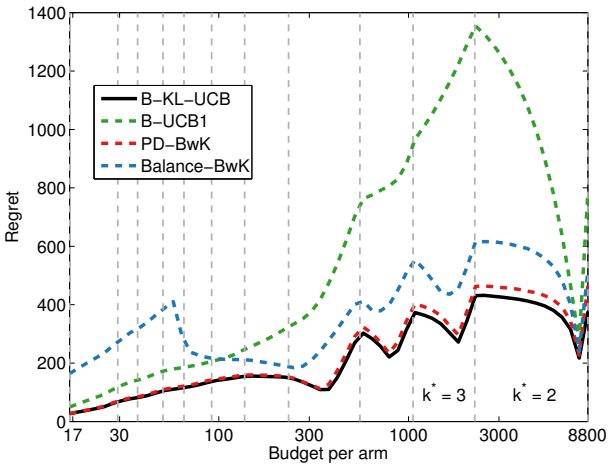

Figure 4: Plot of regret at time $T = 28000$ vs. the budget Tc given to each arm, under the CPC model. The dotted vertical lines demarcate $k^\star$ transitions.

simulation we use incremental budgets (by replacing Tc_k with nc_k in the RHS of the CPI and CPC definitions of $c_k(n)$, which removes all dependence on T) and a fixed $T = 10^6$. For the CPI model, we present two plots where $k^\star = 6$; in Figure 5, $\sum_{k=1}^{k^\star} c_k = 1$, and in Figure 6, $\sum_{k=1}^{k^\star} c_k > 1$. Similarly, for the CPC model, we again set $k^\star = 6$ and show two plots; in Figure 7, $\sum_{k=1}^{k^\star} d_k = 1$, and in Figure 8, $\sum_{k=1}^{k^\star} d_k > 1$. The results confirm that B-KL-UCB and PD-BwK again out-perform the other two algorithms, with very similar regrets. Furthermore, despite our upper bound for the regret not being tight in the $\sum_{k=1}^{k^\star} d_k = 1$ case, empirically we do not see any degradation in performance, suggesting that perhaps B-KL-UCB is optimal even when the separation assumption is violated. It should be noted that B-KL-UCB performs at least as well as both of the modified BwK algorithms, even though the BwK algorithms require knowledge of the time horizon T and B-KL-UCB does not.

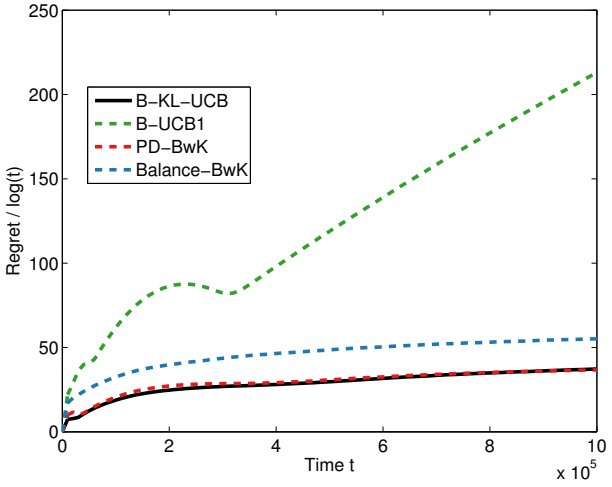

Figure 5: Plot of regret vs. time, with $k^\star = 6$ and $\sum_{k=1}^{k^\star} c_k = 1$, under the CPI model. Each arm is given the same incremental budget per timestep of $1/6$.

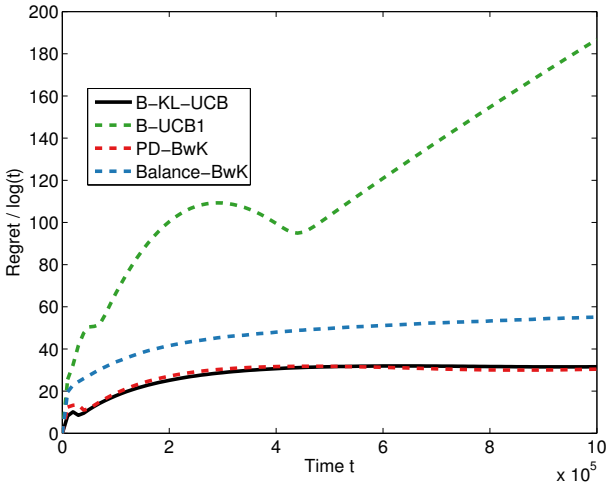

Figure 6: Plot of regret vs. time, with $k^\star = 6$ and $\sum_{k=1}^{k^\star} c_k > 1$, under the CPI model. Each arm is given the same incremental budget per timestep of $1/5.5$.

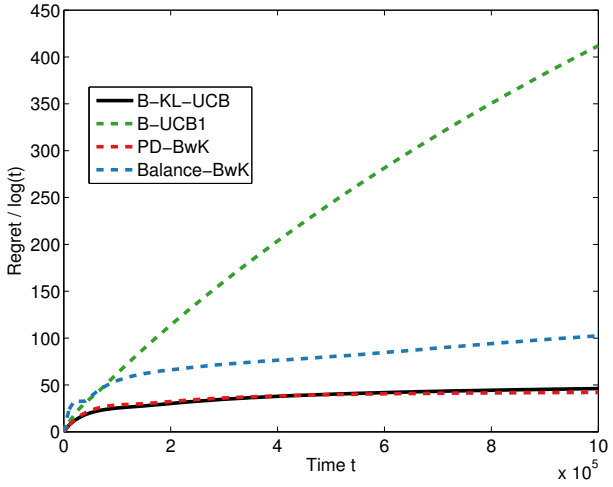

Figure 7: Plot of regret vs. time, with $k^\star = 6$ and $\sum_{k=1}^{k^\star} d_k = 1$, under the CPC model. Each arm is given the same incremental budget per timestep of 0.00489.

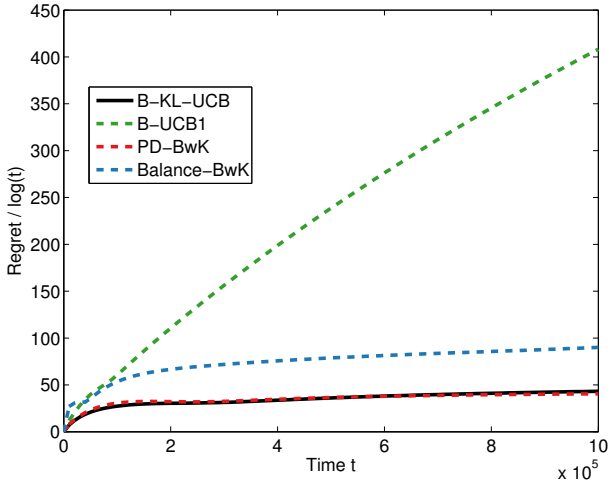

Figure 8: Plot of regret vs. time, with $k^\star = 6$ and $\sum_{k=1}^{k^\star} d_k > 1$, under the CPC model. Each arm is given the same incremental budget per timestep of 0.006.

6.4 BwK algorithms

For both BwK algorithms, we use the so-called confidence radius of an arm

$$\mathrm{rad}\left(\nu, N\right) = \sqrt{\frac{C_{rad}\nu}{N}} + \frac{C_{rad}}{N},$$

where $C_{rad} = \log\left(TK\left(K+1\right)\right)$, ν stands for the current estimate of the expected reward from the arm, and N stands for the number of times that the arm has been played so far. We will also assume the budgets are fixed at the start, so that $n \mapsto c_k(n)$ is a constant.

The idea behind Balance BwK is to ensure that the budgets of the best arms are simultaneously exhausted at T. However, this is not possible since the μ_k's are unknown; therefore, we attempt to exhaust the budgets simultaneously using the current confidence-bound adjusted estimates of the μ_k's. Specifically, we divide time into phases of K time slots each, and we do the following at each the beginning of each phase:

(i) Based on the current estimates of the rewards of the arms, we identify the set of best arms, which collectively have enough budget to be the only arms played. During this process, we also compute an estimate of the number of times each of these arms can be played over the time horizon.

(ii) The probability of playing an arm is simply this estimated number of times it can be played, divided by T.

Now we provide more details about the above computation. To compute $\mathcal{D}$, first we sort the arms (by decreasing order) based on their index $u_k(n)$, settling ties arbitrarily. Next, we iterate through this list, assigning probability mass $\frac{c_k}{L_{n,k}}$ to $\mathcal{D}_k$. We do so until we have accumulated probability 1. If the budget of all arms has been exhausted, assign any remaining probability to the virtual arm with 0 reward and 0 consumption.

The idea behind PD-BwK is to think of each arm's total budget as a resource, each with a fictitious "price", internal to the algorithm. Initially, all of the prices are equal, but as arms are played,

Algorithm 2 Balance-BwK

for each phase $p = 0, 1, 2, \ldots$ **do**
  for each arm $k = 1, \ldots, K$ **do**
    compute UCB estimate for the reward vector,
    $u_{n,k} = \min\left\{\hat{\mu}_k(n) + \mathrm{rad}\left(\hat{\mu}_k(n), t_k\left(n\right)\right), 1\right\}$
    if model is CPI **then**
      resource consumption vector is known a priori, $L_{n,k} = 1$
    else if model is CPC **then**
      compute LCB estimate for the resource consumption vector,
      $L_{n,k} = \max\left\{\hat{\mu}_k(n) - \mathrm{rad}\left(\hat{\mu}_k(n), t_k\left(n\right)\right), 0\right\}$
    end if
  end for
  compute a distribution $\mathcal{D}$ over arms, described in detail below

  for $t = 1, \ldots, K$ **do**
    choose an arm k as an independent sample from $\mathcal{D}$
    if k has enough budget remaining **then**
      pull k
    else
      pull the virtual arm with 0 reward
    end if
    halt if time horizon is met
  end for
end for

Algorithm 3 PD-BwK

set $\epsilon = \sqrt{\log\left(K+1\right)/B}$, where $B = \min\left\{T, \min_k Tc_k\right\}$
in the first K rounds, pull each arm once
initialize the price vector, $v_1 = \mathbf{1}_{K+1}$
for $n = K+1, \ldots, T$ **do**
  for each arm $k = 1, \ldots, K$ **do**
    compute UCB estimate for the reward vector,
    $u_{n,k} = \min\left\{\hat{\mu}_k(n) + \mathrm{rad}\left(\hat{\mu}_k(n), t_k\left(n\right)\right), 1\right\}$
    if model is CPI **then**
      resource consumption vector is known a priori, $L_{n,k} = 1$
    else if model is CPC **then**
      compute LCB estimate for the resource consumption vector,
      $L_{n,k} = \max\left\{\hat{\mu}_k(n) - \mathrm{rad}\left(\hat{\mu}_k(n), t_k\left(n\right)\right), 0\right\}$
    end if
  end for
  $y_n = v_n / \left(\mathbf{1}^T v_n\right)$
  pull arm $j \in \arg\min_{k\in\{1,\ldots,K\}} \left\{\dfrac{y_{K+1} + y_k L_{n,k}}{u_{n,k}}\right\}$
  $v_{n+1,j} = v_{n,j} \cdot (1+\epsilon)^{L_{n,j}}$
  $v_{n+1,K+1} = v_{n,K+1} \cdot (1+\epsilon)$
end for

their remaining budgets (resources) decrease. As each resource becomes more scarce, we respond by multiplicatively increase its price. Additionally, since there is a finite time horizon, the remaining number of time steps is also a resource, with its own price that increases every time step. We then define the "cost" of playing arm k to be the expected total price of all resources consumed: the price of resource k multiplied by the expected consumption of resource k, plus the price of time (multiplied by one, the number of time steps that will be consumed). If we knew the μ_k's, a greedy policy approach would be to always play the arm that maximized the expected reward divided by the expected cost. However, since the μ_k's are unknown, we replace these deterministic quantities (expected consumption of resource k, expected reward from playing arm k) by their confidence-bound adjusted estimates. For the CPI model, we can simplify this and replace the expected consumption of resource k by 1, since it is known a priori that each play of an arm reduces the remaining budget by exactly 1. We note that the way in which prices are increased has to be carefully chosen, and is a function of the time horizon T. As an implementation detail, we actually track the logarithm of the prices and use the corresponding additive update rule, in order to improve numerical stability.

7. CONCLUSION

In this work we have investigated bandits with budgets, which are a natural model for ad-display optimization encountered in search engines. We use the same approach as in the study of the classical bandit: we provide asymptotic regret lower bounds satisfied by any algorithm, and propose algorithms which match those lower bounds. For general budgets we have shown that it is possible to achieve $O(\log(T))$ regret. For CPI and CPC budgets we have provided regret lower bounds that apply to any uniformly good algorithm. Further, we have shown that B-KL-UCB, a natural variant of KL-UCB, is asymptotically optimal. Numerical experiments (based on a real-world data set) further suggest that B-KL-UCB outperforms previously proposed UCB-like algorithms (by a significant margin), so that designing asymptotically optimal algorithms is not purely a theoretical pursuit and yields schemes with good finite-time performance. This is of interest when applying those algorithms to practical problems such as ad-display optimization.

8. ACKNOWLEDGEMENTS

This research was partially supported by AFOSR Grant FA 9550-10-1-0573.

9. REFERENCES

[1] Kdd cup challenge. https://www.kddcup2012.org/c/kddcup2012-track2.

[2] A. Badanidiyuru, R. Kleinberg, and A. Slivkins. Bandits with knapsacks. In *Proc. of FOCS*, 2013.

[3] J. Broder and P. Rusmevichientong. Dynamic pricing under a general parametric choice model. *Operations Research*, 60(4):965–980, 2012.

[4] S. Bubeck, V. Perchet, and P. Rigollet. Bounded regret in stochastic multi-armed bandits. In *Proc. of COLT*, 2013.

[5] N. Cesa-Bianchi and G. Lugosi. Combinatorial bandits. *J. Comput. Syst. Sci.*, 78(5):1404–1422, 2012.

[6] S. S. Chandramouli. Multi armed bandit problem: some insights. Accessed: 2013-09-26.

[7] R. Combes and A. Proutiere. Unimodal bandits: Regret lower bounds and optimal algorithms. In *Proc. of ICML*, 2014.

[8] R. Combes and A. Proutiere. Unimodal bandits: Regret lower bounds and optimal algorithms. Technical Report, http://arxiv.org/abs/1405.5096, 2014.

[9] R. Combes, A. Proutiere, D. Yun, J. Ok, and Y. Yi. Optimal rate sampling in 802.11 systems. In *Proc. of IEEE INFOCOM*, 2014.

[10] E. W. Cope. Regret and convergence bounds for a class of continuum-armed bandit problems. *IEEE Trans. Automat. Contr.*, 54(6):1243–1253, 2009.

[11] A. Garivier. Informational confidence bounds for self-normalized averages and applications. In *Proc. of ITW*, 2013.

[12] A. Garivier and O. Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proc. of COLT*, 2011.

[13] A. Gyorgy, T. Linder, G. Lugosi, and G. Ottucsak. The on-line shortest path problem under partial monitoring. *Journal of Machine Learning Research*, 2007.

[14] C. Jiang and R. Srikant. Bandits with budgets. In *Proc. of CDC*, 2013.

[15] T. Kailath. The divergence and bhattacharyya distance measures in signal selection. *IEEE Trans. Communications*, 15(1):52 – 60, February 1967.

[16] E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Proc. of ALT*, 2012.

[17] R. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Proc. of NIPS*, 2004.

[18] R. Kleinberg, A. Niculescu-Mizil, and Y. Sharma. Regret bounds for sleeping experts and bandits. In *Proc. of COLT*, 2008.

[19] T. Lai. Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, 15(3):1091–1114, 09 1987.

[20] T. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–2, 1985.

[21] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

[22] A. Slivkins. Dynamic ad allocation: Bandits with budgets. http://arxiv.org/abs/1306.0155, 2013.

[23] A. Slivkins, F. Radlinski, and S. Gollapudi. Ranked bandits in metric spaces: learning diverse rankings over large document collections. *Journal of Machine Learning Research*, 2013.

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

[25] L. Tran-Thanh, A. Chapman, E. M. de Cote, A. Rogers, and N. R. Jennings. Epsilon-first policies for budget-limited multi-armed bandits. In *Proc. of AAAI*, 2010.

[26] L. Tran-Thanh, A. Chapman, A. Rogers, and N. R. Jennings. Knapsack based optimal policies for budget-limited multi-armed bandits. In *Proc. of AAAI*, 2012.

[27] A. B. Tsybakov. *Introduction to non-parametric estimation*. Springer, 2008.

[28] J. Yu and S. Mannor. Unimodal bandits. In *Proc. of ICML*, 2011.

10. PROOFS

10.1 Ordering lemma

We define the ordered majorization property: given x and y in $\mathbb{R}^K$, we write $x \lesssim y$ iff $\sum_{k=1}^K x_k = \sum_{k=1}^K y_k$ and for all k: $\sum_{k'=1}^k x_{k'} \leq \sum_{k'=1}^k y_{k'}$. In fact, if x and y are taken as elements of the simplex of $\mathbb{R}^K$ (so that they represent probability distributions on $\{1,\dots,K\}$), the ordered majorization property is equivalent to the strong stochastic order (ordering of c.d.f's). Also, consider $a \in \mathbb{R}^K$ with $k \mapsto a_k$ non-increasing, then we have that $x \lesssim y$ implies: $\sum_{k=1}^K a_n x_n \leq \sum_{k=1}^K a_n y_n$.

The result of Lemma 10.1 states that if the greedy policy $\hat{\pi}$ and an arbitrary policy π are run on the same sample path, then vectors $t^\pi(n) = (t_1^\pi(n),\dots,t_1^\pi(n))$ and $t^{\pi^\star}(n) = (t_1^{\hat{\pi}}(n),\dots,t_1^{\hat{\pi}}(n))$ satisfy $t^\pi(n) \lesssim t^{\hat{\pi}}(n)$ at all time instants n. This ordering property has two non-trivial consequences: (i) it allows us to show that the greedy policy is in fact the optimal policy for general budgets (including the CPI and CPC case), and (ii) it constitutes the crux of our regret upper bound in the case of general budgets. Once again we believe that this is a general property in bandit problems (such as sleeping bandits) where the set of available arms is time-varying and might depend on the sample paths, so that Lemma 10.1 could be useful in analyzing those problems as well, although we have not explored that possibility here.

LEMMA 10.1. *Consider an arbitrary policy π and the greedy policy $\hat{\pi}$, then one has $t^\pi(n) \lesssim t^{\hat{\pi}}(n)$ a.s. for all $n \geq 1$.*

Proof. We proceed by induction. Clearly $t^\pi(0) = (0,0,\dots,0) \lesssim t^{\hat{\pi}}(0)$. Define $n' = \max\{n : t^\pi(n) \lesssim t^{\hat{\pi}}(n)\}$, and assume that $n' < \infty$. Since $t^\pi(n'+1) \lesssim t^{\hat{\pi}}(n'+1)$ is false, we must have $\min A^\pi(n'+1) > \min A^{\hat{\pi}}(n'+1)$. Define $k = \min A^\pi(n'+1)$, so we must have:

$$\sum_{k'=1}^k t_{k'}^\pi(n'+1) > \sum_{k'=1}^k t_{k'}^{\hat{\pi}}(n'+1),$$

which implies that there exists $k' \leq k$ such that $t_{k'}^\pi(n'+1) > t_{k'}^{\hat{\pi}}(n'+1)$, so that $k' \in A^{\hat{\pi}}(n'+1)$. By definition $\hat{\pi}$ selects the arm $\min A^{\hat{\pi}}(n'+1) \leq k' \leq k$, which is a contradiction. Hence such an $n' < \infty$ does not exist, which proves the result. □

10.2 Proof of Proposition 1

Proof. Consider any policy π such that $k^\pi(n)$ is $\mathcal{F}_{n-1}$ measurable. Define $Y_n^\pi = X_{k^\pi(n)}(t_{k^\pi(n)}(n))$ the reward observed at time n and define $M_n^\pi = \sum_{t=1}^n Y_n^\pi - \sum_{k=1}^K \mu_k t_k^\pi(n)$. Then $(M_n^\pi)_n$ is a martingale:

$$M_{n+1}^\pi = M_n^\pi + \sum_{k=1}^K \mathbf{1}\{k^\pi(n) = k\}(Y_n^\pi - \mu_k)$$

$$\mathbb{E}[M_{n+1}^\pi | \mathcal{F}_n] = M_n^\pi + \sum_{k=1}^K \mathbf{1}\{k^\pi(n) = k\}(\mu_k - \mu_k) = M_n^\pi.$$

so that $\mathbb{E}[M_T^\pi] = \mathbb{E}[M_0^\pi] = 0$. Hence the expected reward of π can be written as:

$$\mathbb{E}[r^\pi(T)] = \sum_{k=1}^K \mu_k \mathbb{E}[t_k^\pi(T)].$$

Using Lemma 10.1 , one has $t^\pi(T) \lesssim t^{\hat{\pi}}(T)$ a.s. Since $k \to \mu_k$ is decreasing we have:

$$\sum_{k=1}^K \mu_k t_k^\pi(T) \leq \sum_{k=1}^K \mu_k t_k^{\hat{\pi}}(T) \text{ a.s.}$$

Taking expectations we obtain that $\mathbb{E}[r^\pi(T)] \leq \mathbb{E}[r^{\hat{\pi}}(T)]$. Since the above reasoning is true for all policies we have proven that $\hat{\pi}$ is the optimal policy which concludes the proof. □

10.3 Lower bounds: intermediate results

The following results are instrumental for establishing our regret lower bounds. Lemma 10.2 is an inequality derived in [27] (first noted in [15]), which relates the risk of a hypothesis test between two point hypotheses to the KL divergence between them. Here P and Q represent the two probability distributions corresponding to the two point hypotheses, and the test is taken to be $\mathbf{1}\{\mathcal{A}\}$ with $\mathcal{A}$ an arbitrary event.

LEMMA 10.2 ([27]). *Consider two probability measures P and Q, both absolutely continuous with respect to a given measure. Denote by $KL(P||Q)$ the Kullback Leiber divergence between P and Q. Then for any event $\mathcal{A}$ we have:*

$$P(\mathcal{A}) + Q(\mathcal{A}^c) \geq (1/2)\exp\{-\min(KL(P||Q), KL(Q||P))\}$$

We will be considering change of measure arguments, and in order to avoid confusion, for a given parameter μ we denote by $\mathbb{P}_\mu$ and $\mathbb{E}_\mu$ the probability and expectation under μ. Given an algorithm π running on some parametric bandit, Proposition 2 allows us to calculate the KL divergence of the rewards observed by π, if π were run on the same bandit problem with parameters μ and λ.

PROPOSITION 2. *Consider a bandit problem where the reward of each arm lies in some parametric family, and denote $I(\cdot,\cdot)$ the corresponding KL divergence. Consider a given algorithm π and a given time horizon T. Denote by $Y^T = (Y(1), ..., Y(T))$ with $Y(n) = X_{k^\pi(n)}(t^\pi_k(n))$ the reward from the arm drawn at time n. Consider two parameters μ and λ, and define P and Q to be the distributions of Y^T under parameters μ and λ respectively. Then one has:*

$$KL(P||Q) = \sum_{k=1}^{K} \mathbb{E}_\mu[t^\pi_k(T)] I(\mu_k, \lambda_k).$$

Proof. The proof follows from a straightforward conditioning argument. □

Proposition 3 enables us to lower bound the regret of a given sample path based on the difference on the number of times arm k is selected by the optimal policy and a given policy π:

PROPOSITION 3. *For all $1 \leq k \leq K$ and all policies π we have the following inequality:*

$$\sum_{k'=1}^{K} \mu_k(t^{\pi^\star}_{k'}(T) - t^\pi_{k'}(T)) \geq |t^{\pi^\star}_k(T) - t^\pi_k(T)|\Delta.$$

with $\Delta = \min_{1\leq k'\leq K-1}(\mu_{k'} - \mu_{k'+1}) > 0$.

Proof. This holds as a straightforward consequence of majorization. □

10.4 Proof of Theorem 4.1

Consider a fixed uniformly good policy π. Consider $k > k^\star$ fixed, $\epsilon > 0$ fixed and define parameter λ, with $\lambda_k = \mu_{k^\star} + \epsilon$, and $\lambda_{k'} = \mu_{k'}$, $k' \neq k$. Define $\tilde{c} = \min(c_k, \overline{c})$, and the event $\mathcal{A} = \{t^\pi_k(T) \geq T\tilde{c}/2\}$. Denote by $Y^T = (Y(1), ..., Y(T))$ with $Y(n) = X_{k^\pi(n)}(t^\pi_k(n))$ the reward from the arm drawn at time n. Define P and Q the distributions of Y^T under parameters μ and λ respectively. From Proposition 2 we have:

$$KL(P||Q) = \sum_{k=1}^{K} \mathbb{E}_\mu[t^\pi_k(T)] I(\mu_k, \lambda_k) = \mathbb{E}_\mu[t^\pi_k(T)] I(\mu_k, \mu_{k^\star} + \epsilon).$$

Notice that $\mathbf{1}\{\mathcal{A}\}$ is a function of Y^T, and apply Lemma 10.2:

$$\begin{aligned} P(\mathcal{A}) + Q(\mathcal{A}^c) &\geq (1/2)\exp[-\min(KL(P||Q), KL(Q||P))] \\ &\geq (1/2)\exp[-KL(P||Q)], \end{aligned}$$

so by taking logarithms:

$$\mathbb{E}_\mu[t^\pi_k(T)] I(\mu_k, \mu_{k^\star} + \epsilon) \geq -\log(2) - \log(P(\mathcal{A}) + Q(\mathcal{A}^c)). \quad (1)$$

Let us now upper bound $P(\mathcal{A})$ and $Q(\mathcal{A}^c)$. Under parameter μ we have $t^{\pi^\star}_k(T) = 0$, and under λ we have $t^{\pi^\star}_k(T) = T\tilde{c}$. Applying Proposition 3 we lower bound the sample path regret as follows:

$$\begin{aligned} r(T) &\geq \Delta t^\pi_k(T)\mathbf{1}\{\mathcal{A}\} && \mathbb{P}_\mu\text{- a.s.} \\ r(T) &\geq \epsilon|\tilde{c} - t^\pi_k(T)|\mathbf{1}\{\mathcal{A}^c\} && \mathbb{P}_\lambda\text{- a.s.} \end{aligned}$$

We apply Proposition 3 twice, once under parameter μ, and another time under parameter λ (in this case Δ equals ϵ). When $\mathcal{A}$ occurs, $t^\pi_k(T) \geq T\tilde{c}/2$, and when $\mathcal{A}^c$ occurs, $|\tilde{c} - t^\pi_k(T)| \geq T\tilde{c}/2$, so taking expectations:

$$\begin{aligned} \mathbb{E}_\mu[r(T)] &\geq \Delta T(\tilde{c}/2) P(\mathcal{A}) \\ \mathbb{E}_\lambda[r(T)] &\geq \epsilon T(\tilde{c}/2) Q(\mathcal{A}^c) \end{aligned}$$

Since π is uniformly good, $\mathbb{E}_\mu[r(T)]$ and $\mathbb{E}_\lambda[r(T)]$ must be $O(\log(T))$ so $P(\mathcal{A})$ and $Q(\mathcal{A}^c)$ are $O(T^{-1}\log(T))$. In turn $-\log(P(\mathcal{A}) + Q(\mathcal{A}^c)) \sim_{T\to\infty} \log(T)$ and replacing in (1) we have that:

$$\liminf_{T\to\infty} \frac{\mathbb{E}_\mu[t^\pi_k(T)]}{\log(T)} \geq \frac{1}{I(\mu_k, \mu_{k^\star} + \epsilon)}.$$

The reasoning above is valid for any $\epsilon > 0$, letting $\epsilon \to 0$ in the above equation gives the announced result.

□

10.5 Proof of Theorem 4.2

We proceed as in the proof of Theorem 4.1. Consider a fixed uniformly good policy π. Consider $k > k^\star$ fixed, $\epsilon > 0$ fixed and define parameter λ, with $\lambda_k = \mu_{k^\star} + \epsilon$, and $\lambda_{k'} = \mu_{k'}$, $k' \neq k$.

Define $\tilde{d} = \min(d_k, \overline{d})$, and the event:

$$\mathcal{B} = \cup_{k=1}^{K}\{|Td_k - \tau_k| \leq T\tilde{d}/(4K)\}.$$

From Lemma B.1 (to be proved in the next section), $\mathcal{B}$ occurs w.h.p. (under both $\mathbb{P}_\mu$ and $\mathbb{P}_\lambda$). On all sample paths where $\mathcal{B}$ occurs we have the following inequalities:

$$\begin{aligned} T - \sum_{k=1}^{k^\star - 1} \tau_k &\geq T(\overline{d} - \tilde{d}/4) \geq 3T\tilde{d}/4, \\ T - \sum_{k=1}^{k^\star} \tau_k &\leq T(1 - \sum_{k=1}^{k^\star} d_k + \tilde{d}/4) \leq T\tilde{d}/4, \\ \tau_k &\geq T(d_k - \tilde{d}/(4K)) \geq 3T\tilde{d}/4. \end{aligned}$$

Define the event $\mathcal{A} = \{t^\pi_k(T) \geq T\tilde{d}/2\}$. Denote by $Y^T = (Y(1), ..., Y(T))$ with $Y(n) = X_{k^\pi(n)}(t^\pi_k(n))$ the reward from the arm drawn at time n. Define P and Q the distributions of Y^T under parameters μ and λ respectively. From Proposition 2 we have:

$$KL(P||Q) = \sum_{k=1}^{K} \mathbb{E}_\mu[t^\pi_k(T)] I(\mu_k, \lambda_k) = \mathbb{E}_\mu[t^\pi_k(T)] I(\mu_k, \mu_{k^\star} + \epsilon).$$

Notice that $\mathbf{1}\{\mathcal{A}\}$ is a function of Y^T, and apply Lemma 10.2:

$$\begin{aligned} P(\mathcal{A}) + Q(\mathcal{A}^c) &\geq (1/2)\exp[-\min(KL(P||Q), KL(Q||P))] \\ &\geq (1/2)\exp[-KL(P||Q)], \end{aligned}$$

so by taking logarithms:

$$\mathbb{E}_\mu[t^\pi_k(T)] I(\mu_k, \mu_{k^\star} + \epsilon) \geq -\log(2) - \log(P(\mathcal{A}) + Q(\mathcal{A}^c)). \quad (2)$$

Let us now upper bound $P(\mathcal{A})$ and $Q(\mathcal{A}^c)$. First it is noted that

$$\begin{aligned} P(\mathcal{A}) &= P(\mathcal{A} \cap \mathcal{B}) + P(\mathcal{A} \cap \mathcal{B}^c) \leq P(\mathcal{A} \cap \mathcal{B}) + P(\mathcal{B}^c) \\ &= P(\mathcal{A} \cap \mathcal{B}) + O(T^{-1}). \end{aligned}$$

since $\mathcal{B}$ occurs w.h.p. By the same reasoning, $Q(\mathcal{A}^c) \leq Q(\mathcal{A}^c \cap \mathcal{B}) + O(T^{-1})$, so we can restrict our attention to the events $\mathcal{A} \cap \mathcal{B}$ and $\mathcal{A}^c \cap \mathcal{B}$.

Applying Proposition 3 we lower bound the sample path regret as follows:

$$r(T) \geq \Delta |t_k^{\pi^\star}(T) - t_k^{\pi}(T)| \mathbf{1}\{\mathcal{A} \cap \mathcal{B}\} \qquad \mathbb{P}_\mu\text{- a.s.}$$
$$r(T) \geq \epsilon |t_k^{\pi^\star}(T) - t_k^{\pi}(T)| \mathbf{1}\{\mathcal{A}^c \cap \mathcal{B}\} \qquad \mathbb{P}_\lambda\text{- a.s.}$$

Under parameter μ, when event $\mathcal{A} \cap \mathcal{B}$ occurs, we have $t_k^{\pi^\star}(T) \leq T - \sum_{k=1}^{k^\star} \tau_k \leq T\tilde{d}/4$ and $t_k^{\pi}(T) \geq T\tilde{d}/2$. Similarly, under λ, when event $\mathcal{A}^c \cap \mathcal{B}$ occurs, we have $t_k^{\pi^\star}(T) \geq \min(\tau_k, T - \sum_{k=1}^{k^\star - 1} \tau_k) \geq 3T\tilde{d}/4$, and $t_k^{\pi}(T) \leq T\tilde{d}/2$. Replacing in the above inequalities and taking expectations we get:

$$\mathbb{E}_\mu[r(T)] \geq \Delta T(\tilde{d}/4) P(\mathcal{A} \cap \mathcal{B}),$$
$$\mathbb{E}_\lambda[r(T)] \geq \epsilon T(\tilde{d}/4) Q(\mathcal{A}^c \cap \mathcal{B}).$$

Since π is uniformly good, both $\mathbb{E}_\mu[r(T)]$ and $\mathbb{E}_\lambda[r(T)]$ are $O(\log(T))$, so that $P(\mathcal{A} \cap \mathcal{B})$ and $Q(\mathcal{A}^c \cap \mathcal{B})$ are $O(T^{-1}\log(T))$. In turn $-\log(P(\mathcal{A}) + Q(\mathcal{A}^c)) \sim_{T\to\infty} \log(T)$ and replacing in (1) we have that:

$$\lim \inf_{T\to\infty} \frac{\mathbb{E}_\mu[t_k^\pi(T)]}{\log(T)} \geq \frac{1}{I(\mu_k, \mu_{k^\star} + \epsilon)}.$$

Since the reasoning above is valid for any $\epsilon > 0$, letting $\epsilon \to 0$ in the above equation gives the announced result.

□

10.6 Proof of Theorem 5.1

Proof. Consider $0 < \epsilon < \Delta$ fixed. Define $d(n) = \mu_{k^{\pi^\star}(n)} - \mu_{k^\pi(n)}$, and write the sample path regret as: $r(T) = \sum_{n=1}^T d(n)$. Consider a time instant n such that $d(n) > 0$. Then $1 \leq k^{\pi^\star}(n) < k^\pi(n)$ and $d(n) \leq \mu_1 - \mu_{k^\pi(n)}$, so that:

$$r(T) \leq \sum_{k\geq 2}^K (\mu_1 - \mu_k)|B_k|, \tag{3}$$

with $B_k = \{n \leq T : k^\pi(n) = k, k^{\pi^\star}(n) \leq k-1\}$. Consider $n \in B_k$. From Lemma 10.1, we have that: $t^\pi(n) \lesssim t^{\pi^\star}(n)$, which implies that $\min A^\pi(n) \leq \min A^{\pi^\star}(n) = k^{\pi^\star}(n) \leq k-1$. Therefore we have $\min A^\pi(n) \leq k-1$, so that applying Lemma 10.3 we obtain:

$$\mathbb{E}[|B_k|] \leq \frac{f(T)}{I(\mu_k + \epsilon, \mu_{k-1})} + \epsilon^{-2} + C\log(\log(T)).$$

Taking expectations and replacing in (3) we get the announced result:

$$R^\pi(T) \leq f(T) \sum_{k\geq 2} \frac{\mu_1 - \mu_k}{I(\mu_k + \epsilon, \mu_{k-1})} + K(\epsilon^{-2} + C\log(\log(T)))$$

which concludes the proof. □

10.7 Proof of Theorem 5.2

Proof. Recall that for the optimal policy we have: $t_k(T) = c_k T$ for $k < k^\star$, $t_{k^\star}(T) = \overline{c}T$, and $t_k(T) = 0$ for $k > k^\star$. Therefore the regret of a sample path is:

$$r(T) = \overline{c}T\mu_{k^\star} + \sum_{k=1}^{k^\star - 1} c_k T\mu_k - \sum_{k=1}^K \mu_k t_k(T).$$

Using statement (i) of Lemma B.3, we have that $t_k(T) = c_k T$ for $k < k^\star$ w.h.p., therefore $t_{k^\star}(T) = \overline{c}T - \sum_{k>k^\star} t_k(T)$ and:

$$r(T) = \overline{c}T\mu_{k^\star} - \sum_{k\geq k^\star} \mu_k t_k(T) = \sum_{k>k^\star} (\mu_{k^\star} - \mu_k) t_k(T) \text{ w.h.p.}$$

Taking expectations:

$$R^\pi(T) \leq \sum_{k>k^\star} (\mu_{k^\star} - \mu_k)\mathbb{E}[t_k(T)] + O(1). \tag{4}$$

Since $\sum_{k=1}^{k^\star - 1} c_k T + \overline{c}T = T$, we have that $\max_{1\leq n\leq T}(\min A(n)) = k^\star$. Hence applying Lemma 10.3, for all $k > k^\star$ we have:

$$\mathbb{E}[t_k(T)] \leq \frac{f(T)}{I(\mu_k + \epsilon, \mu_{k^\star})} + C(\log(\log(T)) + \epsilon^{-2}).$$

with C a constant. Replacing in (4) we obtain the announced result:

$$R^\pi(T) \leq f(T) \sum_{k>k^\star} \frac{\mu_{k^\star} - \mu_k}{I(\mu_k + \epsilon, \mu_{k'})} + KC(\log(\log(T)) + \epsilon^{-2}),$$
$$= f(T)\delta_{k^\star}^\epsilon + KC(\log(\log(T)) + \epsilon^{-2}).$$

which concludes the proof. □

10.8 Proof of Theorem 5.3

Proof.

First statement

From Lemma B.2 we have $\tilde{k} \in \{k^\star, k^\star + 1\}$ w.h.p. Define the following events:

$$\mathcal{A} = \{\tilde{k} = k^\star\},$$
$$\mathcal{B} = \{\tilde{k} = k^\star + 1, t_{k^\star}(T) = \tau_{k^\star}\},$$
$$\mathcal{C} = \{\tilde{k} = k^\star + 1, t_{k^\star}(T) < \tau_{k^\star}\}.$$

We decompose the regret according to the occurence of $\mathcal{A}$, $\mathcal{B}$ or $\mathcal{C}$.

Regret of sample paths in $\mathcal{A}$

First, consider a sample path in $\mathcal{A}$. Define $\tilde{\tau} = T - \sum_{k=1}^{k^\star - 1} \tau_k$. It is noted that $\tilde{\tau} \geq 0$. The regret of such a sample path is:

$$r(T) = \tilde{\tau}\mu_{k^\star} + \sum_{k<k^\star} \tau_k\mu_k - \sum_{k=1}^K t_k(T)\mu_k.$$

Using statement (ii) of Lemma B.3 we have $t_k(T) = \tau_k$ for all $k < k^\star$, therefore $t_{k^\star}(T) = \tilde{\tau} - \sum_{k>k^\star} t_k(T)$ so that the regret is:

$$r(T) = \sum_{k>k^\star} t_k(T)(\mu_{k^\star} - \mu_k).$$

Since $\tilde{k} = k^\star$ we have that $\max_{1\leq n\leq T}(\min A(n)) = k^\star$. Taking expectations and applying Lemma 10.3:

$$\mathbb{E}[r(T)\mathbf{1}\{\mathcal{A}\}]$$
$$\leq \mathbb{P}[\mathcal{A}] f(T)\delta_{k^\star}^\epsilon + CK(\log(\log(T)) + \epsilon^{-2})$$

with C a constant.

Regret of sample paths in $\mathcal{B}$

Consider a sample path in $\mathcal{B}$. Define $\tilde{\tau} = T - \sum_{k=1}^{k^\star} \tau_k$. The regret is:

$$r(T) = \tilde{\tau}\mu_{k^\star+1} + \sum_{k\leq k^\star} \tau_k\mu_k - \sum_{k=1}^K t_k(T)\mu_k.$$

By the definition of $\mathcal{B}$ we have $t_{k^\star}(T) = \tau_{k^\star}$ and using statement (ii) of Lemma B.3 we have $t_k(T) = \tau_k$ for all $k < k^\star$. Therefore $t_{k^\star+1}(T) = \tilde{\tau} - \sum_{k>k^\star+1} t_k(T)$ and the regret is:

$$r(T) = \sum_{k>k^\star+1} t_k(T)(\mu_{k^\star+1} - \mu_k).$$

Since $\tilde{k} = k^\star + 1$ we have that $\max_{1\leq n\leq T}(\min A(n)) = k^\star + 1$. Taking expectations and applying Lemma 10.3:

$$\begin{aligned}
&\mathbb{E}[r(T)\mathbf{1}\{\mathcal{B}\}] \\
&\leq \sum_{k>k^\star+1} \mathbb{E}[t_k(T)\mathbf{1}\{\mathcal{B}\}](\mu_{k^\star+1} - \mu_k), \\
&\leq \sum_{k>k^\star+1} \frac{\mathbb{P}[\mathcal{B}](\mu_{k^\star+1} - \mu_k)f(T)}{I(\mu_k + \epsilon, \mu_{k^\star+1})} + \epsilon^{-2} + C\log(\log(T)) \\
&= \mathbb{P}[\mathcal{B}]f(T)\delta^\epsilon_{k^\star+1} + CK(\log(\log(T)) + \epsilon^{-2})
\end{aligned}$$

Regret of sample paths in $\mathcal{C}$

Finally consider sample paths in $\mathcal{C}$. Define $\tilde{\tau} = T - \sum_{k=1}^{k^\star} \tau_k$. The regret is:

$$r(T) = \tilde{\tau}\mu_{k^\star+1} + \sum_{k\leq k^\star} \tau_k\mu_k - \sum_{k=1}^{K} t_k(T)\mu_k.$$

Using statement (ii) of Lemma B.3 we have $t_k(T) = \tau_k$ for all $k < k^\star$. Therefore $t_{k^\star}(T) = \tilde{\tau} + \tau_{k^\star} - \sum_{k>k^\star} t_k(T)$. The regret is:

$$\begin{aligned}
r(T) &= \tilde{\tau}\mu_{k^\star+1} + \tau_{k^\star}\mu_{k^\star} - \left(\mu_{k^\star}t_{k^\star}(T) + \sum_{k>k^\star} t_k(T)\mu_k\right) \\
&= \tilde{\tau}\mu_{k^\star+1} + \tau_{k^\star}\mu_{k^\star} - \\
&\quad \left(\mu_{k^\star}(\tilde{\tau} + \tau_{k^\star} - \sum_{k>k^\star} t_k(T)) + \sum_{k>k^\star} t_k(T)\mu_k\right) \\
&= (\mu_{k^\star+1} - \mu_{k^\star})\tilde{\tau} + \sum_{k>k^\star} t_k(T)(\mu_{k^\star} - \mu_k).
\end{aligned}$$

Using the fact that $\mu_{k^\star+1} - \mu_{k^\star} < 0$ we have the upper bound:

$$r(T) \leq \sum_{k>k^\star} t_k(T)(\mu_{k^\star} - \mu_k).$$

Since $t_{k^\star}(T) < \tau_{k^\star}$ we have that $\max_{1\leq n\leq T}(\min A(n)) = k^\star$. Taking expectations and applying Lemma 10.3:

$$\begin{aligned}
\mathbb{E}[r(T)\mathbf{1}\{\mathcal{C}\}] &\leq \sum_{k>k^\star} \mathbb{E}[t_k(T)\mathbf{1}\{\mathcal{C}\}](\mu_{k^\star} - \mu_k), \\
&\leq \sum_{k>k^\star} \frac{\mathbb{P}[\mathcal{C}](\mu_{k^\star} - \mu_k)f(T)}{I(\mu_k + \epsilon, \mu_{k^\star})} + \epsilon^{-2} + C\log(\log(T)) \\
&= \mathbb{P}[\mathcal{C}]f(T)\delta^\epsilon_{k^\star} + CK(\log(\log(T)) + \epsilon^{-2})
\end{aligned}$$

Therefore, defining $\alpha(T) = \mathbb{P}[\mathcal{A}] + \mathbb{P}[\mathcal{C}]$ and noting that $\mathbb{P}[\mathcal{A}] + \mathbb{P}[\mathcal{B}] + \mathbb{P}[\mathcal{C}] \leq 1$ so that $\mathbb{P}[\mathcal{B}] \leq 1 - \alpha(T)$, we obtain the announced result:

$$\begin{aligned}
R^\pi(T) \leq &f(T)(\alpha(T)\delta^\epsilon_{k^\star} + (1 - \alpha(T))\delta^\epsilon_{k^\star+1}) \\
&+ 3CK(\log(\log(T)) + \epsilon^{-2}).
\end{aligned}$$

which proves the first statement of the theorem.

Second statement

Using Lemma B.1, we have that if $\sum_{k=1}^{k^\star} d_k > 1$, then $\sum_{k=1}^{k^\star} \tau_k > 1$ w.h.p, so that $\tilde{k} = k^\star$ w.h.p. Hence $\mathbb{P}[\mathcal{B}] \to_{T\to\infty} 0$ and $\mathbb{P}[\mathcal{C}] \to_{T\to\infty} 0$, and letting $T \to \infty$ in the first statement of the theorem yields the second statement.

□

10.9 Upper bounds: an intermediate result

LEMMA 10.3. *Consider arbitrary budgets. For any policy π, and $k > k'$, define the set of instants:*

$$B^\pi_{k,k'} = \{n : k(n) = k, \max_{1\leq n\leq T}(\min A^\pi(n)) \leq k'\}.$$

and consider any event $\mathcal{A}$. Then under policy B-KL-UCB, for all $0 < \epsilon < \mu_{k'} - \mu_k$ we have:

$$\mathbb{E}\left[|B_{k,k'}|\mathbf{1}\{\mathcal{A}\}\right] \leq \mathbb{P}[\mathcal{A}]\frac{f(T)}{I(\mu_k + \epsilon, \mu_{k'})} + \epsilon^{-2} + CK\log(\log(T)),$$

with $C > 0$ a constant.

Proof.

Consider k, k', ϵ and $\mathcal{A}$ fixed. Define $t_0 = \frac{f(T)}{I(\mu_k+\epsilon,\mu_{k'})}$. Decompose $B_{k,k'}$ as:

$$B_{k,k',1} = \{n \in B_{k,k'}, t_k(n) \leq t_0\} \quad \text{(i)}$$

$$B_2 = \cup_{k''}\tilde{B}_{k'',2},\ \tilde{B}_{k'',2} = \{n \leq T : b_{k''}(n) < \mu_{k''}\} \quad \text{(ii)}$$

$$B_{k,k',3} = \{n \in B_{k,k'} \setminus B_2, t_k(n) > t_0\}. \quad \text{(iii)}$$

and $B_{k,k'} \subset B_{k,k',1} \cup B_2 \cup B_{k,k',3}$.

(i) At each $n \in B_{k,k',1}$, $t_k(n)$ is incremented and $t_k(n) \leq t_0$, so $|B_{k,k',1}| \leq t_0$ surely.

(ii) From Lemma A.1, we have $\mathbb{E}[|\tilde{B}_{k'',2}|] \leq O(\log(\log(T)))$ for all k'', so $\mathbb{E}[|B_2|] \leq O(K\log(\log(T)))$ by union bound.

(iii) Consider $n \in B_{k,k',3}$. We are going to prove that we have $|\hat{\mu}_k(n) - \mu_k| \geq \epsilon$. First if $\hat{\mu}_k(n) \geq \mu_{k'}$ we have $|\hat{\mu}_k(n) - \mu_k| \geq \epsilon$ trivially since $\epsilon < \mu_{k'} - \mu_k$. Now assume that $\hat{\mu}_k(n) < \mu_{k'}$. We have that $k(n) = k$ and there exists $k'' \leq k'$ such that $k'' \in A(n)$ since

$$\min A(n) \leq \max_{1\leq n\leq T}(\min A(n)) \leq k'.$$

Hence $b_k(n) \geq b_{k''}(n) \geq \mu_{k''} \geq \mu_{k'}$ since $n \notin B_2$. Furthermore, $t_k(n) \geq t_0$. By the definition of $b_k(n)$, this implies:

$$\begin{aligned}
t_k(n)I(\hat{\mu}_k(n), \mu_{k'}) &\leq f(T) \\
t_0 I(\hat{\mu}_k(n), \mu_{k'}) &\leq f(T) \\
I(\hat{\mu}_k(n), \mu_{k'}) &\leq I(\mu_k + \epsilon, \mu_{k'})
\end{aligned}$$

By monotonicity of the KL-divergence, this implies $|\hat{\mu}_k(n) - \mu_k| \geq \epsilon$ in this case as well. We have proven that:

$$B_{k,k',3} \subset \{n : k(n) = k, |\hat{\mu}_k(n) - \mu_k| \geq \epsilon\}.$$

so that $\mathbb{E}[|B_{k,k',3}|] \leq \epsilon^{-2}$ using [8][Lemma B.2].

Putting it all together:

$$\begin{aligned}
&\mathbb{E}[|B_{k,k'}|\mathbf{1}\{\mathcal{A}\}] \\
&\leq \mathbb{E}[|B_{k,k',1}|\mathbf{1}\{\mathcal{A}\}] + \mathbb{E}[|B_2|\mathbf{1}\{\mathcal{A}\}] + \mathbb{E}[|B_{k,k',3}|\mathbf{1}\{\mathcal{A}\}] \\
&\leq \mathbb{E}[t_0\mathbf{1}\{\mathcal{A}\}] + \mathbb{E}[|B_2|] + \mathbb{E}[|B_{k,k',3}|] \\
&\leq t_0\mathbb{P}[\mathcal{A}] + O(K\log(\log(T))) + \epsilon^{-2},
\end{aligned}$$

which proves the announced result.

□

APPENDIX

A. CONCENTRATION INEQUALITY

The following concentration inequality derived in [11] is instrumental here.

LEMMA A.1 ([11]). *Consider $\{X(n)\}_{n\geq 1}$ i.i.d. Bernoulli with parameter μ. Define $S_t = (1/t)\sum_{n=1}^{t} X(n)$, then for all $\delta > 0$ we have that:*

$$\mathbb{P}[\sup_{1\leq t\leq T} tI(S_t,\mu) \geq \delta] \leq 2e\lceil \delta \log(T)\rceil e^{-\delta}.$$

By Pinsker's inequality, $I(p,q) \geq 2(p-q)^2$ so that for all $\delta \geq 0$:

$$\mathbb{P}[\sup_{1\leq t\leq T} \sqrt{t}|S_t-\mu| \geq \delta] \leq 4e\lceil \delta^2 \log(T)\rceil e^{-2\delta^2}.$$

B. UPPER BOUNDS: TECHNICAL RESULTS

We present some lemmas which are instrumental for the regret analysis of B-KL-UCB in the three cases of interest.

LEMMA B.1. *For all k and $\epsilon > 0$, we have:*

$$\tau_k/T \in [d_k-\epsilon, d_k+\epsilon] \text{ w.h.p.}$$

Proof. We have to prove that $\mathbb{P}[\tau_k/T \notin [d_k-\epsilon, d_k+\epsilon]] = O(T^{-1})$. Using a union bound:

$$\mathbb{P}[\tau_k/T \notin [d_k-\epsilon,d_k+\epsilon]] \leq \mathbb{P}[\tau_k \leq T(d_k-\epsilon)]+\mathbb{P}[\tau_k \geq T(d_k+\epsilon)]. \quad (5)$$

Consider the first term in the r.h.s. of (5). The event $\tau_k \leq T(d_k-\epsilon)$ implies:

$$\sum_{i=1}^{T(d_k-\epsilon)} X_k(i) \geq Tc_k,$$

$$\sum_{i=1}^{T(d_k-\epsilon)} (X_k(i)-\mu_k) \geq Tc_k - T(d_k-\epsilon)\mu_k = T\epsilon\mu_k.$$

Applying Hoeffding's inequality we obtain:

$$\mathbb{P}\left[\sum_{i=1}^{T(d_k-\epsilon)} (X_k(i)-\mu_k) \geq T\epsilon\mu_k\right] \leq \exp\left(-\frac{2T\epsilon^2\mu_k^2}{d_k-\epsilon}\right).$$

Consider the second term in the r.h.s. of (5). The event $\tau_k \geq T(d_k+\epsilon)$ implies:

$$\sum_{i=1}^{T(d_k+\epsilon)} X_k(i) \leq Tc_k,$$

$$\sum_{i=1}^{T(d_k+\epsilon)} (X_k(i)-\mu_k) \leq Tc_k - T(d_k+\epsilon)\mu_k = -T\epsilon\mu_k.$$

Applying Hoeffding's inequality again we obtain:

$$\mathbb{P}\left[\sum_{i=1}^{T(d_k+\epsilon)} (X_k(i)-\mu_k) \leq -T\epsilon\mu_k\right] \leq \exp\left(-\frac{2T\epsilon^2\mu_k^2}{d_k+\epsilon}\right).$$

Therefore:

$$\mathbb{P}[\tau_k/T \notin [d_k-\epsilon,d_k+\epsilon]] \leq \exp\left(-\frac{2T\epsilon^2\mu_k^2}{d_k-\epsilon}\right)+\exp\left(-\frac{2T\epsilon^2\mu_k^2}{d_k+\epsilon}\right) = O(T^{-1})$$

so $\tau_k/T \in [d_k-\epsilon, d_k+\epsilon]$ w.h.p., which is the announced result. □

LEMMA B.2. *In the CPC case, we have $\tilde{k} \in \{k^\star, k^\star+1\}$ w.h.p.*

Proof. Define $\overline{d} = 1-\sum_{k=1}^{k^\star-1} d_k$. By the definition of $k^\star$, $\overline{d} > 0$. Applying Lemma B.1 with $\epsilon = \overline{d}/(K+1)$, we have:

$$\sum_{k=1}^{k^\star-1} \tau_k \leq T\sum_{k=1}^{k^\star-1}(d_k+\overline{d}/(K+1)) = T(1-\overline{d}+\overline{d}(k^\star-1)/(K+1)) < T \text{ w.h.p.}$$

so $\tilde{k} \geq k^\star$ w.h.p.

If $k^\star = K$, then $\tilde{k} \leq k^\star+1$ trivially. Otherwise, by the same reasoning, define $\underline{d} = (\sum_{k=1}^{k^\star+1} d_k) - 1$. By the definition of $k^\star$, $\underline{d} > 0$. Applying Lemma B.1 with $\epsilon = \underline{d}/(K+1)$, we have:

$$\sum_{k=1}^{k^\star+1} \tau_k \geq T\sum_{k=1}^{k^\star+1}(d_k-\underline{d}/(K+1)) = T(1+\underline{d}-\underline{d}(k^\star+1)/(K+1)) > T \text{ w.h.p.}$$

so $\tilde{k} \leq k^\star+1$ w.h.p.

Therefore $\tilde{k} \in \{k^\star, k^\star+1\}$ w.h.p. which concludes the proof. □

LEMMA B.3. *Consider algorithm B-KL-UCB.*

(i) In the CPI case, for all $k < k^\star$ we have $t_k(T) = c_k T$ w.h.p.

(ii) In the CPC case, for all $k < k^\star$ we have $t_k(T) = \tau_k$ w.h.p.

Proof.

First consider the CPC case. Consider $k < k^\star$ fixed, and consider the event $\mathcal{A} = \{t_k(T) < \tau_k\}$. Consider $k' > k$. Using Lemma A.1 (first statement) with $\delta = f(T)$ we have that for all $1 \leq n \leq T$: $b_k(n) \geq \mu_k$ w.h.p. Using Lemma A.1 (second statement) with $\delta = 2\log(T)$ we have that:

$$\hat{\mu}_{k'}(n) \leq \mu_{k'} + \sqrt{2\log(T)/t_k(n)} \text{ w.h.p.}$$

Using Pinsker's inequality:

$$b_{k'}(n) \leq \hat{\mu}_{k'}(n) + \sqrt{\frac{2\log(T)}{t_{k'}(n)}},$$

so that:

$$b_{k'}(n) \leq \mu_{k'} + \sqrt{\frac{8\log(T)}{t_{k'}(n)}} \text{ w.h.p.}$$

Since k' is only selected at instants n such that $b_{k'}(n) \geq b_k(n)$, this implies that:

$$t_{k'}(T) \leq \frac{8\log(T)}{(\mu_k-\mu_{k'})^2} \text{ w.h.p.} \quad (6)$$

Define $\overline{d} = \sum_{k'=1}^{k^\star-1} d_{k'}$. We have $\overline{d} < 1$ by the definition of $k^\star$. If $t_k(T) < \tau_k$, from Lemma B.1, we have that $\sum_{k'=1}^{k^\star-1} \tau_{k'} \leq T(1+\overline{d})/2$ w.h.p. Using (6) we obtain that:

$$T = \sum_{k'=1}^{K} t_{k'}(T) \leq \sum_{k'\leq k} \tau_{k'} + \sum_{k'>k} t_{k'}(T)$$

$$\leq \sum_{k'=1}^{k^\star-1} \tau_{k'} + \sum_{k'>k} t_{k'}(T) \leq \frac{T(1+\overline{d})}{2} + O(\log(T)) < T \text{ w.h.p.}$$

for large T, a contradiction (recall that $\overline{d} < 1$ so that $(1+\overline{d})/2 < 1$). Therefore $\mathcal{A}$ occurs with small probability, and for all $k < k^\star$, $t_k(T) = \tau_k$ w.h.p., which concludes the proof.

The proof in the CPI case follows from the same argument.