ONLINE ADVERTISEMENTS AND MULTI-ARMED BANDITS

BY

CHONG JIANG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Doctoral Committee:

Professor R. Srikant, Chair
Associate Professor Carolyn Beck
Associate Professor Angelia Nedich
Professor Venugopal V. Veeravalli

ABSTRACT

We investigate a number of multi-armed bandit problems that model different aspects of online advertising, beginning with a survey of the key techniques that are commonly used to demonstrate the theoretical limitations and achievable results for the performance of multi-armed bandit algorithms. We then formulate variations of the basic stochastic multi-armed bandit problem, aimed at modeling how budget-limited advertisers should bid and how ad exchanges should choose whose ad to display, and study them using these techniques.

We first consider online ad auctions from the point of view of a single advertiser who has an average budget constraint. By modeling the rest of the bidders through a probability distribution (often referred to as the mean-field approximation), we develop a simple bidding strategy which can be implemented without any statistical knowledge of bids, valuations, and query arrival processes. The key idea is to use stochastic approximation techniques to automatically track long-term averages.

Next, we consider multi-armed bandits with budgets, modeling how ad exchanges select which ad to display. 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 each time their ad is shown, 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.

Finally, we consider the problem of multi-armed bandits with a large, possibly infinite number of correlated arms, modeling a retailer advertising a large number of related items. We assume that the arms have Bernoulli distributed rewards, where the probabilities of success are parametrized by known attribute vectors for each arm and an unknown vector which describes the preferences of the target audience. For this model, we seek an algorithm with a total regret that is sub-linear in time and independent of the number of arms. We present such an algorithm and analyze its performance, showing upper bounds on the total regret which apply uniformly in time, for both the finite and infinite arm cases.

ACKNOWLEDGMENTS

I would like to express my gratitude to Professor Srikant for being my advisor; I could not have done this without his guidance, support, and empirically infinite patience.

I would also like to thank the collaborators I've worked with over the years; Mindi Yuan, Professor Carolyn Beck, Joohwan Kim, and Richard Combes. It was a pleasure working with each of you.

Finally, I would like to thank my parents for instilling in me the lifelong interest in math and science that has shaped who I am, and for their constant love and support throughout my academic journey.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

With the rise in internet usage, there has been a corresponding rise in online retail and online advertising, evidenced by the growth of companies such as Amazon and Google. There are many optimization problems in both of these domains. For example, as a retailer, which items should we recommend to a given customer? Or, as an advertiser, how much should we bid in order to place one of our ads? Or, as an ad server, whose ads should we select to display? In general, different users will also have different preferences, which are not known in advance, and hence may need to be learned over time. Furthermore, these problems often have many players, each with their own objectives and possible actions, which suggest game-theoretic formulations. Finally, for online retailers and ad servers, mechanisms can be designed to optimize for a metric of their choosing (e.g., user satisfaction, advertiser profit, or social welfare). We will attempt to address some of these problems using the theory of multi-armed bandits.

A common problem in learning theory is to identify the best option from a set of options, without prior knowledge of how good each option is, while also minimizing the cost incurred in doing so. A classical model for this problem is the stochastic multi-armed bandit, where each option is represented by a slot machine (we will refer to each machine as an arm, since playing a machine is done through pulling said arm). At each time-step, we play one of the m arms and subsequently observe a stochastic reward from that arm, where each pull of an arm k generates i.i.d. samples from some unknown distribution. Our objective will be to maximize the sum of the rewards obtained during all time-steps, up to some time-horizon T. Any algorithm that specifies which arm we should play, given the past history of rewards, is known as a *policy*. The usual metric for measuring performance of a policy is that of *regret*, a measure of how much less total reward we obtain in expectation by following our chosen policy, rather than by following an optimal policy that knows a priori the means of the reward distributions. Since these distributions are unknown to our policy, we must learn them through repeated plays, but

we also wish to maximize our reward by choosing to play arms that are already estimated to be good. These two conflicting goals, exploration of the unknown and exploitation of the known, exemplify a fundamental trade-off present in a wide class of online machine learning problems.

In Chapter 3, we consider the problem of sponsored search, a type of online advertisement where paid links are shown next to the search results of relevant queries. This form of advertisement yields substantial revenue for search engine providers like Microsoft and Google, and draws both views and customers to websites. It is also a problem of theoretical interest, since bidders are competing for overlapping keywords and targeted demographics, and where the number of bidders, ads, and ad slots on webpages are all large. We consider this problem from the viewpoint of an advertiser bidding in such an ad auction, with the ability to bid strategically. However, because the number of other bidders is often very large, attempting to reason about their strategies is in general computationally infeasible, and we will assume the other bidders have stationary bid distributions. The goal then is to construct a simple policy to maximize our own profit, while meeting the budget constraints.

In Chapter 4, we again consider sponsored search, but this time from the perspective of the ad server. Again, even though advertisers can bid strategically, often they either choose not to for the sake of simplicity, or have attempted strategic play and subsequently converged on a stationary bid distribution (e.g., if they follow the policy from Chapter 3). The exact problem we will consider, then, is how the ad server should allocate ads in order to maximize ad relevance and profit (both related to the click-through rate), given the fact that each advertiser's budget is limited. This will be modeled as a multi-armed bandit problem where each arm is associated with not only a reward distribution, but a budget as well. We will consider two budget models that commonly appear in online advertising: cost-per-impression, which limits the number of times an ad can be shown, and cost-per-click, which limits the number of times an ad can be clicked. Additionally, a more general budget model, one that allows advertisers to change their budget allocation based on the past performance of their ad campaign, is also considered.

In Chapter 5, we consider the problem of item recommendation, which occurs in a variety of online contexts in order to deliver a more personalized set of results, rather than a static list to all viewers. This can be useful in displaying related items for sale, for example by Amazon, or in the ordering of local restaurants, such as on Yelp. In both of these examples, the number of items is large, relative to

the number of features. As a concrete example, imagine an online camera store, with hundreds of different camera models in stock. However, there are perhaps closer to ten features which people will compare when deciding which, if any, to purchase. There are permanent features of the camera itself, such as megapixel count, brand name, and year of introduction, as well as extrinsic features, such as price, review scores, and item popularity, all of which might be considered by the customer in order to decide whether or not to buy the camera. If purchased, the store gains a profit corresponding to the item. One can similarly interpret a listing of restaurants, with features such as type of cuisine, price range, and review scores.

Finally, we give an overview of the prior work in this area, for those who may be interested in a particular variation of the multi-armed bandit problem. For a general survey of multi-armed bandit problems and their variations, see [1, 2]. One of the earliest breakthroughs on the classical multi-armed bandit problem came from [3], who showed that under geometric discounting, the optimal policy assigns an index to each arm, now known as the Gittins index, and pulls the arm with the largest Gittins index. Other proofs of this optimality have been given later by [4, 5]. Reference [6] proved that a similar index-based result is nearly optimal in the "restless bandit" variation of this model, where the arms which are not pulled also evolve in time. While these policies greatly simplify a single m-dimensional problem into m 1-dimensional problems, it is still, in general, too computationally complex for online learning.

Reference [7] proved an achievable $O(m \cdot \log T)$ lower bound for the expected total regret of the stochastic multi-armed bandit problem in the case of independent arms. Related work in [8, 9, 10, 11, 12, 13] considered similar models with i.i.d. and Markov time dependencies for each arm, constructed index policies which are computationally much simpler, and extended the results to include "multiple plays" and "switching costs."

References [14, 15, 16] considered models with finite numbers of arms, with reward distributions that are correlated through a multivariate parameter z of dimension n, and obtained upper bounds on the regret of order $O(\sqrt{mT})$, $O(\sqrt{nT} \cdot \log T)$, and $O(\sqrt{nT})$, respectively. Reference [17] considered a model in which the expected rewards are affine functions of a scalar parameter z, but allowed the set of arms to be a bounded, convex region in $\mathbb{R}^n$, in which case m is uncountably infinite. They then derived a policy whose expected total regret is $\Theta(\sqrt{T})$. Reference [18] expanded this model to allow for a multivariate parameter z of

dimension n, and showed that the expected total regret (ignoring $\log T$ factors) is $\Theta(n\sqrt{T})$. Reference [19] independently considered a nearly identical model, and obtained similar results. Reference [20] considered a model in which the deterministic rewards are a Lipschitz-continuous function of the n-dimensional vector corresponding to each arm, and obtained an expected total regret (ignoring $\log T$ factors) of $\Theta(T^{\frac{n+1}{n+2}})$. This was generalized in [21], where the Lipschitz property is only required in the neighborhood of the reward function's maxima.

Reference [22] considered a non-stochastic version of the multi-armed bandit problem, in which the rewards are no longer drawn from an unknown distribution, but can instead be adversarially generated. The resultant total weak regret, calculated by comparison with the single arm which is best over the entire time horizon, is shown to be $O(\sqrt{mT})$. The change from logarithmic to polynomial regret in this model is due to having rewards which are time-dependent and potentially adversarially generated, instead of being drawn from a time-independent distribution.

Reference [23] investigated the finite-time regret of the multi-armed bandit problem, assuming bounded but otherwise arbitrary reward distributions. Using upper confidence bound (UCB) algorithms, where the confidence interval of an arm shrinks as the arm is subjected to more plays, they achieve a logarithmic upper bound on the regret, uniform over time, that scales with the "gaps" between the expected rewards for the arms. One algorithm they propose, UCB2, selects the arm with largest empirical mean plus confidence interval, plays it for a number of time-steps dependent on how often that particular arm has been selected in the past, and repeats this process until the time-horizon is reached. This achieves asymptotically optimal expected total regret, but with a suboptimal constant. A variant of UCB proposed by [24], KL-UCB, does attain the best constant possible. Finally, we note that a common idea used in crafting policies to solve the multi-armed bandit problem is that of the doubling trick [25, 26]. This technique is used to convert an algorithm which works on a time horizon T, along with its corresponding bound, into an anytime algorithm, with an upper bound that holds uniformly over time.

CHAPTER 2

COMMON TECHNIQUES FOR MAB PROBLEMS

2.1 Introduction

In this chapter, we will motivate and describe techniques used in proving both lower- and upper-bounds for the classical stochastic multi-armed bandit problem. These will serve as the simplest version of key proof ideas we will use in later chapters, when examining more complicated variations of this basic model.

We consider a bandit problem with two arms and a time horizon $T \geq 0$. Time is discrete; at time $n \in \{1, ..., T\}$ we can select an arm $k(n) \in \{1, 2\}$ to play. We then receive 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 $(X_k(i))_{k\in\{1,2\},i\geq 0}$ are independent, and that $X_k(i)$ is a Bernoulli random variable with parameter μ_k, where $\mu_1 > \mu_2$ (i.e., arm 1 is the best arm). We denote by $\mathcal{F}_n$ the σ-algebra generated by $\{X_{k(1)}(t_{k(1)}(1)), \ldots, X_{k(n)}(t_{k(n)}(n))\}$. We consider adaptive policies, so that $k(n)$ is $\mathcal{F}_{n-1}$ measurable for all n. We define $\pi^\star$ the oracle policy (which knows μ), and can therefore maximize the expected accumulated sum of rewards by always playing arm 1. We further define the *regret* of decision rule π by:

$$\begin{aligned} R^\pi(T) &= \sum_{k=1}^{2} \mu_k \mathbb{E}_\mu \left[t_k^{\pi^\star}(T)\right] - \sum_{k=1}^{2} \mu_k \mathbb{E}_\mu \left[t_k^{\pi}(T)\right] \\ &= (\mu_1 - \mu_2) \cdot \mathbb{E}_\mu \left[t_2^{\pi}(T)\right]. \end{aligned}$$

The regret of policy π is the loss in accumulated reward due to the fact that the parameters (μ_1, μ_2) are unknown to π. We say that policy π is uniformly good if, for all problem instances, $R^\pi(T) = O(\log(T))$ when $T \to \infty$. Finally, we will use $D_{KL}(\mathcal{P}||\mathcal{Q})$ to denote the KL divergence between two distributions $\mathcal{P}$ and $\mathcal{Q}$, and $D_{KL}(p||q)$ to denote the KL divergence between two Bernoulli distributions

with parameters p and q. Namely, $D_{KL}(p||q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$.

2.2 Lower Bounds

We will first investigate an asymptotic bound, originally by Lai and Robbins [7], which states:

Theorem 2.2.1 *For any uniformly good policy π,*

$$\liminf_{T\to\infty} \frac{R^{\pi}(T)}{\log(T)} \geq \sum_{k:\mu_k<\mu_1} \frac{\mu_1 - \mu_k}{D_{KL}(\mu_k||\mu_1)}.$$

Proof: We shall prove this for two arms for simplicity of exposition, but the proof is easily extended to the K-arm case. The main idea is to consider an alternative set of parameters, say (λ_1, λ_2) with $\lambda_1 < \lambda_2$ (i.e., arm 2 is the best arm). Let $\mathcal{P}$ and $\mathcal{Q}$ denote the probability distributions of $\left(X_{k(1)}\left(t_{k(1)}(1)\right), \ldots, X_{k(T)}\left(t_{k(T)}(T)\right)\right)$ when the parameters of X_k are μ and λ, respectively. Consider the probabilities of seeing any particular sequence of rewards; if μ and λ are "close," then the distributions $\mathcal{P}$ and $\mathcal{Q}$ are also close, and it becomes difficult to determine which of the two parameters gave rise to the observed reward sequence. This is formalized through a result of [27], which states that for any event $\mathcal{A}$, we have:

$$\mathcal{P}(\mathcal{A}) + \mathcal{Q}(\mathcal{A}^c) \geq \frac{1}{2}\exp\left\{-\min\left(D_{KL}(\mathcal{P}||\mathcal{Q}), D_{KL}(\mathcal{Q}||\mathcal{P})\right)\right\}. \tag{2.1}$$

We provide a short proof of this lemma at the end of this section. Using this, we now choose an event $\mathcal{A}$ which is likely under $\mathcal{Q}$ and unlikely under $\mathcal{P}$, in order to minimize the LHS of the above equation. In particular, we choose $\mathcal{A} = \left\{t_2(T) \geq \frac{T}{2}\right\}$, since arm 2 is sub-optimal under the parameter μ and can only be played $O(\log T)$ times on average as π is uniformly good, and similarly for arm 1 under λ. We can then use Markov's inequality to compute

$$\frac{T}{2}\cdot\mathcal{P}\left(t_2(T) \geq \frac{T}{2}\right) \leq \mathbb{E}\left[t_2(T)\right] \in O(\log T),$$
$$\frac{T}{2}\cdot\mathcal{Q}\left(t_2(T) < \frac{T}{2}\right) = \frac{T}{2}\cdot\mathcal{Q}\left(t_1(T) \geq \frac{T}{2}\right) \leq \mathbb{E}_{\lambda}\left[t_1(T)\right] \in O(\log T).$$

Combining this with (2.1) yields

$$\begin{aligned} D_{KL}\left(\mathcal{P}||\mathcal{Q}\right) &\geq \min\left(D_{KL}\left(\mathcal{P}||\mathcal{Q}\right), D_{KL}\left(\mathcal{Q}||\mathcal{P}\right)\right) \\ &\geq \log T - \log 4 - \log\left(\frac{T}{2}\cdot\left[\mathcal{P}\left(\mathcal{A}\right) + \mathcal{Q}\left(\mathcal{A}^{c}\right)\right]\right), \\ \frac{D_{KL}\left(\mathcal{P}||\mathcal{Q}\right)}{\log T} &\geq 1 - O\left(\frac{\log\log T}{\log T}\right). \end{aligned} \tag{2.2}$$

Now all that remains is to relate the regret $R\left(T\right)$ to $D_{KL}\left(\mathcal{P}||\mathcal{Q}\right)$. We have that

$$\mathcal{P}\left(x\right) = \prod_{t=1}^{T}\prod_{k=1}^{2}\left[\mu_{k}^{x_{t}}\cdot\left(1-\mu_{k}\right)^{\left(1-x_{t}\right)}\right]^{1_{\{k(t)=k\}}}$$

and similarly for $\mathcal{Q}\left(x\right)$, with λ in place of μ. Thus,

$$\begin{aligned} D_{KL}\left(\mathcal{P}||\mathcal{Q}\right) &= \int \log\frac{d\mathcal{P}\left(x\right)}{d\mathcal{Q}\left(x\right)} d\mathcal{P}\left(x\right) \\ &= \int \log\left(\prod_{t=1}^{T}\prod_{k=1}^{2}\left[\left(\frac{\mu_{k}}{\lambda_{k}}\right)^{x_{t}}\cdot\left(\frac{1-\mu_{k}}{1-\lambda_{k}}\right)^{\left(1-x_{t}\right)}\right]^{1_{\{k(t)=k\}}}\right) d\mathcal{P}\left(x\right) \\ &= \sum_{t=1}^{T}\sum_{k=1}^{2}\int 1_{\{k(t)=k\}}\left[x_{t}\log\left(\frac{\mu_{k}}{\lambda_{k}}\right) + \left(1-x_{t}\right)\log\left(\frac{1-\mu_{k}}{1-\lambda_{k}}\right)\right] d\mathcal{P}\left(x\right) \\ &= \sum_{t=1}^{T}\sum_{k=1}^{2}\mathbb{P}_{\mu}\left[k\left(t\right)=k\right]\cdot D_{KL}\left(\mu_{k}||\lambda_{k}\right) \\ &= \mathbb{E}_{\mu}\left[t_{1}\left(T\right)\right]\cdot D_{KL}\left(\mu_{1}||\lambda_{1}\right) + \mathbb{E}_{\mu}\left[t_{2}\left(T\right)\right]\cdot D_{KL}\left(\mu_{2}||\lambda_{2}\right). \end{aligned}$$

Recall the regret:

$$\begin{aligned} R\left(T\right) &= \left(\mu_{1}-\mu_{2}\right)\cdot\mathbb{E}_{\mu}\left[t_{2}^{\pi}\left(T\right)\right] \\ &= \frac{\mu_{1}-\mu_{2}}{D_{KL}\left(\mu_{2}||\lambda_{2}\right)}\cdot\left[D_{KL}\left(\mathcal{P}||\mathcal{Q}\right) - \mathbb{E}_{\mu}\left[t_{1}\left(T\right)\right]\cdot D_{KL}\left(\mu_{1}||\lambda_{1}\right)\right]. \end{aligned} \tag{2.3}$$

We can now select the "worst-case" λ in order to obtain the best regret bound. Since $\mathbb{E}_{\mu}\left[t_{1}\left(T\right)\right] \geq T - O\left(\log T\right)$ and $D_{KL}\left(\mathcal{P}||\mathcal{Q}\right) \sim \log T$, we must choose $\lambda_{1} = \mu_{1}$ in order to obtain any positive bound on the regret. To further make our lower-bound on $R\left(T\right)$ the largest possible, given a fixed bound on $D_{KL}\left(\mathcal{P}||\mathcal{Q}\right)$, we should choose λ so that $D_{KL}\left(\mu_{2}||\lambda_{2}\right)$ is the smallest possible. Since $\lambda_{2} >$

$\lambda_1 = \mu_1 > \mu_2$, we have that $D_{KL}(\mu_2||\lambda_2) \geq D_{KL}(\mu_2||\mu_1)$, and thus

$$\frac{R(T)}{\log T} \geq \frac{\mu_1 - \mu_2}{D_{KL}(\mu_2||\mu_1)} \cdot \frac{D_{KL}(\mathcal{P}||\mathcal{Q})}{\log T}.$$

Applying (2.2) and letting $T \to \infty$ yields the desired result,

$$\liminf_{T\to\infty} \frac{R^{\pi}(T)}{\log(T)} \geq \frac{\mu_1 - \mu_2}{D_{KL}(\mu_2||\mu_1)}.$$

□

It is easy to see how this proof can be adapted for $K > 2$ arms; assuming $\mu_1 > \mu_2 \geq \ldots \geq \mu_K$, we can bound each $\mathbb{E}_{\mu}[t_k(T)]$ by an associated choice of λ, which is identical to μ at all indices except k, since we must have $\lambda_k > \mu_1$ for the best arm (and hence the correct decision) to differ between μ and λ. The regret can then be directly computed from $R(T) = \sum_{k>1}(\mu_1 - \mu_k) \cdot \mathbb{E}_{\mu}[t_k(T)]$. As for extensions to other models, the key steps under this procedure are:

1. Construct the "closest possible" λ. In this case, we ended up choosing $\lambda_1 = \mu_1$ and $\lambda_2 \searrow \mu_1$; had we arbitrarily done so at the start, the proof would be simpler (avoiding the need for $D_{KL}(\mu_1||\lambda_1)$), but the derivation we present motivates this choice of λ. In general, any arm k that is played $\Omega(T)$ times must have $\lambda_k = \mu_k$ in order to obtain a meaningful bound, for the same reason that $\lambda_1 = \mu_1$ was necessary here.

2. Choose an event $\mathcal{A}$ which is likely under $\mathcal{Q}$ and unlikely under $\mathcal{P}$, and use the uniformly good property of π to bound $\mathcal{P}(\mathcal{A}) + \mathcal{Q}(\mathcal{A}^c)$. In this case, we chose $\mathcal{A} = \left\{t_2(T) \geq \frac{T}{2}\right\}$, and in general, any event involving a $\Theta(T)$ difference in the number of plays of arms can suffice.

3. Express the regret $R(T)$ in terms of

$$D_{KL}(\mathcal{P}||\mathcal{Q}) = \sum_k \mathbb{E}_{\mu}[t_k(T)] \cdot D_{KL}(\mu_k||\lambda_k),$$

   which can be related to $\mathcal{P}(\mathcal{A}) + \mathcal{Q}(\mathcal{A}^c)$ through (2.1), the Kailath bound.

We remark that this technique is similar to the decision-theoretic approach used in [28, 29, 30], except each of those apply the Kailath bound at every time-step, instead of once with a single decision rule at the end of the time-horizon T. For

example, in the case of [30], applying their lower-bound analysis to Bernoulli-distributed rewards leads to an incorrect (KL-divergence based) constant in the lower-bound. Specifically, they lower-bound the maximum of the two regrets under μ and λ respectively, but the proof exploits a symmetry present for Gaussian distributions with fixed variances and mean parameters of $\mu = (\mu_1, \mu_1 - \delta)$ and $\lambda = (\mu_1, \mu_1 + \delta)$, namely that a multiplicative factor of δ can be factored from the regrets under either parameter μ or λ. When applied to the Bernoulli case, it could be that $\mu_1 + \delta \geq 1$, forcing λ to be chosen differently and causing the analysis to produce an incorrect multiplicative constant for the $\Omega(\log T)$ lower-bound.

In contrast, our technique of a single decision at time T depends only on $\mathbb{E}_\mu[t_2(T)]$, which can be directly related to the total regret under μ, without needing to consider the regret under λ. Additionally, having a single decision provides intuition of why the obtained regret is unavoidable - an alternative hypothesized reward distribution λ can be similar enough that we mistake it for the true reward distribution μ with non-trivial probability, while the index of the best arm differs between μ and λ. If we explore insufficiently many times, then it may be the case that we are actually drawing samples according to λ, and this non-negligible probability of making a mistake that incurs order T regret would prevent this policy from being uniformly good.

For completeness, we provide a simple derivation of the Kailath bound, which is the key lemma needed to introduce the correct constant into the lower-bound proofs.

Lemma 2.2.2 *Suppose we have two hypotheses about the probability distribution of a random variable X, either $H_0 : X \sim \mathcal{P}$, or $H_1 : X \sim \mathcal{Q}$. For any decision rule $\psi : \mathcal{X} \to \{0, 1\}$,*

$$\frac{1}{2}\left(\mathbb{P}_1(\psi(X) = 0) + \mathbb{P}_0(\psi(X) = 1)\right) \geq \frac{1}{4}\exp\left(-D_{KL}(\mathcal{P}||\mathcal{Q})\right).$$

Proof: The LHS can be interpreted as the probability of error under the rule ψ, when we have an equal prior on H_0 and H_1. In this case, we know that the

likelihood ratio test is optimal, and thus

$$\begin{aligned}\frac{1}{2}\left(\mathbb{P}_1\left(\psi\left(X\right)=0\right)+\mathbb{P}_0\left(\psi\left(X\right)=1\right)\right) &\geq \frac{1}{2}\left[\mathbb{P}_1\left(\frac{\mathcal{P}\left(X\right)}{\mathcal{Q}\left(X\right)}>1\right)+\mathbb{P}_1\left(\frac{\mathcal{Q}\left(X\right)}{\mathcal{P}\left(X\right)}\geq 1\right)\right] \\ &= \frac{1}{2}\left[\int 1_{\left\{\frac{\mathcal{P}}{\mathcal{Q}}>1\right\}}\mathcal{Q}dx+\int 1_{\left\{\frac{\mathcal{Q}}{\mathcal{P}}\geq 1\right\}}\mathcal{P}dx\right] \\ &= \frac{1}{2}\int \min\left\{\mathcal{P},\mathcal{Q}\right\}dx,\end{aligned}$$

where in the interest of notational simplicity, we have omitted the x from $\mathcal{P}\left(x\right)$ and $\mathcal{Q}\left(x\right)$. Next, we relate $\min\left\{\mathcal{P},\mathcal{Q}\right\}$ to $\sqrt{\mathcal{P}\cdot\mathcal{Q}}$. Since $p+q=\min\left\{p,q\right\}+\max\left\{p,q\right\}$, we have that $\int\left[\min\left\{\mathcal{P},\mathcal{Q}\right\}+\max\left\{\mathcal{P},\mathcal{Q}\right\}\right]=2$. Next,

$$\begin{aligned}\left(\int\sqrt{\mathcal{P}\cdot\mathcal{Q}}dx\right)^2 &= \left(\int\sqrt{\min\left\{\mathcal{P},\mathcal{Q}\right\}\cdot\max\left\{\mathcal{P},\mathcal{Q}\right\}}dx\right)^2 \\ &\leq \left(\int\min\left\{\mathcal{P},\mathcal{Q}\right\}dx\right)\cdot\left(\int\max\left\{\mathcal{P},\mathcal{Q}\right\}dx\right) \\ &= \left(\int\min\left\{\mathcal{P},\mathcal{Q}\right\}dx\right)\cdot\left(2-\int\min\left\{\mathcal{P},\mathcal{Q}\right\}dx\right) \\ &\leq 2\int\min\left\{\mathcal{P},\mathcal{Q}\right\}dx,\end{aligned}$$

where we have used the Cauchy-Schwarz inequality. Finally, we relate $\sqrt{\mathcal{P}\cdot\mathcal{Q}}$ to $D_{KL}\left(\mathcal{P}||\mathcal{Q}\right)$.

$$\begin{aligned}\left(\int\sqrt{\mathcal{P}\cdot\mathcal{Q}}dx\right)^2 &= \exp\left(2\log\int\sqrt{\mathcal{P}\cdot\mathcal{Q}}dx\right)=\exp\left(2\log\int\mathcal{P}\sqrt{\frac{\mathcal{Q}}{\mathcal{P}}}dx\right) \\ &= \exp\left(2\log\mathbb{E}_0\left[\sqrt{\frac{\mathcal{Q}}{\mathcal{P}}}\right]\right)\geq\exp\left(\mathbb{E}_0\left[2\log\sqrt{\frac{\mathcal{Q}}{\mathcal{P}}}\right]\right) \\ &= \exp\left(\mathbb{E}_0\left[\log\frac{\mathcal{Q}}{\mathcal{P}}\right]\right)=\exp\left(-D_{KL}\left(\mathcal{P}||\mathcal{Q}\right)\right),\end{aligned}$$

where we have used Jensen's inequality. Putting it all together, we have that

$$\begin{aligned}&\frac{1}{2}\left(\mathbb{P}_1\left(\psi\left(X\right)=0\right)+\mathbb{P}_0\left(\psi\left(X\right)=1\right)\right) \\ \geq&\frac{1}{2}\int\min\left\{\mathcal{P},\mathcal{Q}\right\}dx\geq\frac{1}{4}\left(\int\sqrt{\mathcal{P}\cdot\mathcal{Q}}dx\right)^2\geq\frac{1}{4}\exp\left(-D_{KL}\left(\mathcal{P}||\mathcal{Q}\right)\right).\end{aligned}$$

As a corollary, if we take any event $\mathcal{A}$ and apply this lemma to the decision rule $\psi(X) = 1_{\{X \in \mathcal{A}\}}$, we obtain (2.1). $\square$

2.3 Upper Bounds

In this section we will consider a sequence of policies, in an attempt to asymptotically meet the lower bound of the previous section, i.e., find a policy π such that

$$\limsup_{T \to \infty} \frac{R(T)}{\log(T)} \leq \sum_{k:\mu_k < \mu_1} \frac{\mu_1 - \mu_k}{D(\mu_k || \mu_1)}.$$

We begin with perhaps the simplest way of making the tradeoff between exploration and exploitation: time division. Namely, choose $1 > \epsilon > 0$ and spend ϵ fraction of the time exploring the arms, and the remaining $1 - \epsilon$ fraction of the time exploiting the best arm found thus far. This is termed an ϵ-greedy algorithm, and serves as the simplest of adaptive allocation rules, albeit one where the use of feedback information is limited to the exploitation time-steps. This assignment of exploration and exploitation time-steps can be done either deterministically or randomly; here we examine the deterministic case where we have spent $\frac{\epsilon t}{K}$ time-steps exploring each of the K arms up to time-step t. This is also known as an epoch-greedy policy, since we can imagine partitioning time into a sequence of epochs, where each epoch is a single exploration pull (or a round of exploration pulls for each arm), followed by some number of exploitation pulls of the empirically best arm. The larger ϵ is, the more quickly the empirical best arm converges to the actual best arm, but this comes with a cost of purposefully playing sub-optimal arms long after the best arm has been identified. In other words, by increasing ϵ, we see a tradeoff where we improve short-term performance and worsen asymptotic performance. Of course, with any fixed ϵ, we will have spent $\frac{\epsilon T}{K}$ time-steps in expectation exploring each of the sub-optimal arms by the time-horizon T. Since each play of a sub-optimal arm incurs a constant regret, the exploration regret is

$$R_{explore}(T) = \frac{\epsilon T}{K} \cdot \sum_{k:\mu_k < \mu_1} (\mu_1 - \mu_k).$$

To bound the exploitation regret, both for ϵ-greedy and more sophisticated algo-

rithms, we will rely on the Hoeffding and Chernoff bounds, applied to a sequence of i.i.d. Bernoulli random variables $\{X_i\}$ with parameter p. In this case, the Hoeffding bound states that:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq p+\epsilon\right) \leq \exp\left(-2n\epsilon^2\right),$$
$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \leq p-\epsilon\right) \leq \exp\left(-2n\epsilon^2\right),$$

and the Chernoff bound states that:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq p+\epsilon\right) \leq \exp\left(-n\cdot D_{KL}\left(p+\epsilon||p\right)\right),$$
$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \leq p-\epsilon\right) \leq \exp\left(-n\cdot D_{KL}\left(p-\epsilon||p\right)\right),$$

where $D_{KL}\left(p||q\right) = p\cdot\log\left(\frac{p}{q}\right) + (1-p)\cdot\log\left(\frac{1-p}{1-q}\right)$ is the Kullback-Leibler divergence between two Bernoulli random variables. Note that the Hoeffding bound in this case is implied by the Chernoff bound and Pinsker's inequality, which states that $D_{KL}\left(p||q\right) \geq 2\left(p-q\right)^2$. Although this means that the Hoeffding bound is the weaker of the two bounds, we will nonetheless use it here, as well as in the next section to analyze the popular UCB1 algorithm.

Returning to the exploitation regret, recall that $\{X_k\left(i\right)\}_{i\geq 1}$ are i.i.d. Bernoulli random variables with parameter μ_k. We define $\tilde{\mu}_k\left(l\right) = \frac{1}{l}\sum_{i=1}^{l} X_k\left(i\right)$, the empirical mean of the rewards from pulling arm k, *during the exploration time-steps* up to epoch l. Then, consider the probability of choosing to exploit a sub-optimal arm k in epoch l,

$$\begin{aligned}
&\mathbb{P}\left(\tilde{\mu}_k\left(l\right) \geq \max_{k'}\tilde{\mu}_{k'}\left(l\right)\right) \leq \mathbb{P}\left(\tilde{\mu}_k\left(l\right) \geq \tilde{\mu}_1\left(l\right)\right)\\
\leq&\mathbb{P}\left(\tilde{\mu}_k\left(l\right) \geq \frac{\mu_1+\mu_k}{2}\right) + \mathbb{P}\left(\tilde{\mu}_1\left(l\right) \leq \frac{\mu_1+\mu_k}{2}\right)\\
=&\mathbb{P}\left(\frac{1}{l}\sum_{i=1}^{l} X_k\left(i\right) \geq \mu_k + \frac{\mu_1-\mu_k}{2}\right) + \mathbb{P}\left(\frac{1}{l}\sum_{i=1}^{l} X_1\left(i\right) \leq \mu_1 - \frac{\mu_1-\mu_k}{2}\right)\\
\leq&2\exp\left(-l\cdot\frac{\left(\mu_1-\mu_k\right)^2}{2}\right).
\end{aligned}$$

The number of exploitation pulls in epoch l is $\frac{1-\epsilon}{\epsilon}$ and the total number of epochs is ϵT, so we can compute the total regret as

$$\begin{aligned} R(T) &\leq R_{explore}(T) + \sum_{l=1}^{\epsilon T} \frac{1-\epsilon}{\epsilon} \cdot \sum_{k:\mu_k<\mu_1} (\mu_1 - \mu_k) \cdot \mathbb{P}\left(\tilde{\mu}_k(l) \geq \max_{k'} \tilde{\mu}_{k'}(l)\right) \\ &\leq R_{explore}(T) + \sum_{k:\mu_k<\mu_1} (\mu_1 - \mu_k) \cdot \sum_{l=1}^{\epsilon T} \frac{1-\epsilon}{\epsilon} \cdot 2\exp\left(-l \cdot \frac{(\mu_1-\mu_k)^2}{2}\right) \\ &\leq \sum_{k:\mu_k<\mu_1} (\mu_1 - \mu_k) \cdot \left\{ \frac{\epsilon T}{K} + \frac{2}{\epsilon} \cdot \frac{1}{\exp\left(\frac{(\mu_1-\mu_k)^2}{2}\right) - 1} \right\}, \end{aligned}$$

which is linear in T regardless of the choice of ϵ. To summarize, we have shown the following:

Theorem 2.3.1 *For the epoch-greedy policy, for any constant $\epsilon > 0$, we have that*

$$R(T) \leq \sum_{k:\mu_k<\mu_1} (\mu_1 - \mu_k) \cdot \left\{ \frac{\epsilon T}{K} + \frac{2}{\epsilon} \cdot \frac{1}{\exp\left(\frac{(\mu_1-\mu_k)^2}{2}\right) - 1} \right\}.$$

Starting with this ϵ-greedy algorithm with constant ϵ, it is a simple modification to make ϵ depend upon the current time-step n. In particular, if $\epsilon(n) \searrow 0$, the fraction of time spent on exploration (the dominant term in the regret analysis) will become $o(T)$, i.e., we will have sub-linear regret. For simplicity of analysis, let us again consider an epoch-greedy policy, except with increasing lengths of epochs. The intuition behind this is that for larger n, we become increasingly confident that the arm we choose to exploit - the empirical best arm - is in fact the true best arm, and thus the expected loss due to having chosen a sub-optimal arm to exploit diminishes. However, each time we choose to explore incurs a fixed cost, in exchange for gaining additional information in order to mitigate the expected loss per exploitation pull. As the probability of error and hence expected loss decreases, the fraction of time that we should spend on exploration should also decrease. Specifically, we would like to choose the number of exploitation time-steps in epoch l, denoted $g(l)$, so that the exploitation regret can be bounded

by a constant, matching the exploration regret. Recall the total regret

$$R(T) \leq \sum_{l=1}^{L} \left\{ \sum_{k:\mu_k<\mu_1} (\mu_1 - \mu_k) \cdot \left[\frac{1}{K} + g(l) \cdot \mathbb{P}\left(\tilde{\mu}_k(l) \geq \max_{k'} \tilde{\mu}_{k'}(l) \right) \right] \right\}$$
$$\leq \sum_{k:\mu_k<\mu_1} (\mu_1 - \mu_k) \cdot \sum_{l=1}^{L} \left[\frac{1}{K} + 2g(l) \cdot \exp\left(-l \cdot \frac{(\mu_1 - \mu_k)^2}{2} \right) \right].$$

If we now choose $g(l) = \frac{1}{K} \cdot \exp(c \cdot l)$ where $c \leq \frac{\min_{k:\mu_k<\mu_1} (\mu_1 - \mu_k)^2}{2}$, then the regret can be bounded by

$$R(T) \leq \frac{3L}{K} \cdot \sum_{k:\mu_k<\mu_1} (\mu_1 - \mu_k).$$

Note that L epochs corresponds to L explorations and $\sim \frac{1}{K} \cdot \exp(c \cdot L)$ exploitations. Since this sums to T time-steps, then the number of explorations is $L \sim \frac{\log KT}{c}$. Thus, we have the following result:

Theorem 2.3.2 *For the epoch-greedy policy, if the length of the epochs is chosen as* $g(l) = \frac{1}{K} \cdot \exp(c \cdot l)$ *where* $c \leq \frac{\min_{k:\mu_k<\mu_1} (\mu_1 - \mu_k)^2}{2}$*, then we have that*

$$\limsup_{T\to\infty} \frac{R(T)}{\log T} \leq \frac{3}{cK} \cdot \sum_{k:\mu_k<\mu_1} (\mu_1 - \mu_k).$$

Note that this regret bound only applies for c being sufficiently small, which means our policy would require knowledge of $\min_{k:\mu_k<\mu_1} (\mu_1 - \mu_k)$. We can also reformulate this epoch-greedy policy back into an ϵ-greedy policy, by noting that if we choose to explore with probability $\epsilon(t) = \frac{1}{c \cdot t}$, the result is also $\sim \frac{\log T}{c}$ exploration time-steps. By the Hoeffding bound, we can show that for any $\delta > 0$, each arm will have been explored at least $(1 - \delta) \cdot \frac{\log T}{cK}$ times with high probability, when T is large enough. By the same analysis as before (now conditioned on this high probability event), we can obtain the same regret bound, even when the exploration time-steps are chosen randomly and not deterministically:

Theorem 2.3.3 *For the* ϵ*-greedy policy, if we choose to explore with probability*

$\epsilon(t) = \dfrac{1}{c \cdot t}$ *where* $c \leq \dfrac{\min_{k:\mu_k<\mu_1} (\mu_1 - \mu_k)^2}{2}$*, then we have that*

$$\limsup_{T\to\infty} \frac{R(T)}{\log T} \leq \frac{3}{cK} \cdot \sum_{k:\mu_k<\mu_1} (\mu_1 - \mu_k).$$

It is important to note that for ϵ-greedy as well as for epoch-greedy, the schedule of explorations is effectively decided in advance and is not adaptive; only the choices of which arm to exploit are adaptive. As an example, if there are three arms with expected rewards of $\{0.5, 0.4, 0.1\}$ respectively, and we spend equal time exploring each arm, the second arm is more likely than the third to be confused for being the best arm. This asymmetry suggests that an optimal algorithm would explore different arms at different rates, and it is precisely the failure of these algorithms to do so that accounts for their sub-optimality.

2.3.1 UCB algorithms

So if there is an asymmetry in how often arms should be explored, how can we determine this adaptively? Instead of fixing the number of explorations, we could instead fix the probabilities of error (i.e., deciding a particular sub-optimal arm is the best arm) and adapt the number of explorations of each arm in order to meet those probabilities. Namely, if we wish to obtain a regret bound of $O(\log T)$, then one option is to give ourselves a $O(n^{-1})$ probability of choosing each sub-optimal arm $k > 1$ at each time-step n, as this integrates into a total of $O(\log T)$ regret.

To accomplish this, we will turn to a common theme of algorithms designed to solve multi-armed bandit problems, "index policies" that at each time-step n assign an index $b_k(n)$ to each arm k and play the arm with the largest index, breaking ties arbitrarily. In particular, these indices will be "optimistic," meaning they should over-estimate the expected reward. An intuitive justification for this is the following: Suppose we did not use optimistic indices. If we are unlucky with the first few plays of the best arm, it could be that its empirical average reward dips below the expected reward of a sub-optimal arm. The index of the best arm may be so low that this arm is never played again. Since this event occurs within a deterministic number of time-steps with positive probability, the result is linear regret. Thus, we restrict our consideration to optimistic policies where we over-estimate the expected reward. We can decompose each index (also

called an upper confidence bound, for reasons that will become clear soon) into a sum of two terms, the empirical average reward and an "exploration bonus" that captures the uncertainty in the mean we have due to having only finite samples. We will choose the exploration bonus to be large enough so that the probabilities of error satisfy $\mathbb{P}\left(b_k\left(n\right) > b_1\left(n\right)\right) \in O\left(n^{-1}\right)$, which is necessary if arm k is to be played $O\left(\log T\right)$ times. However, we should not choose an exploration bonus that is too large, as that would cause the sub-optimal arms to be played too often, exceeding the $\dfrac{\log T}{D\left(\mu_i||\mu_1\right)}$ lower-bound. We define $t_k\left(n\right)$ to be the number of times arm k has been played up to time-step n, and $\hat{\mu}_k\left(n\right) = \dfrac{1}{t_k\left(n\right)}\sum_{i=1}^{t_k(n)} X_k\left(i\right)$ to be the empirical mean of the rewards seen when pulling arm k up to time-step n. The index of arm k at time-step n is then defined to be

$$b_k\left(n\right) = \hat{\mu}_k\left(n\right) + \sqrt{\frac{K\log n}{t_k\left(n\right)}},$$

where K is a constant to be chosen later. Note that $b_k\left(n\right)$ is large when $T_k\left(n\right) < \log n$.

To understand the performance of UCB, let us first understand the conditions under which we might pull arm k instead of arm 1. This will happen only if $b_k\left(n\right) \geq b_1\left(n\right)$. If $t_k\left(n\right)$ is sufficiently large compared to $\log n$, then $b_k\left(n\right)$ will concentrate around μ_k and similarly, $b_1\left(n\right)$ will concentrate around μ_1. Therefore, the probability of $b_k\left(n\right) \geq b_1\left(n\right)$ will be small. To make this intuition precise, we first convert the condition $b_k\left(n\right) \geq b_1\left(n\right)$ into a condition involving the deviation of $b_k\left(n\right)$ from μ_k and $b_1\left(n\right)$ from μ_1. Namely, note that $b_k\left(n\right) \geq b_1\left(n\right)$ is false if $b_k\left(n\right) < \mu_1$ and $\mu_1 < b_1\left(n\right)$, and thus,

$$\left\{b_k\left(n\right) \geq b_1\left(n\right)\right\} \implies \left\{b_k\left(n\right) \geq \mu_1\right\} \text{ or } \left\{b_1\left(n\right) \leq \mu_1\right\}. \tag{2.4}$$

We now write the expected regret up to time T in terms of the above events as follows:

$$R\left(T\right) = \sum_{k:\mu_k<\mu_1} \left(\mu_1 - \mu_k\right)\mathbb{E}\left[t_k\left(T\right)\right],$$

where $\left(\mu_1 - \mu_k\right)$ is the reduction in expected reward each time arm k is played, and $\mathbb{E}\left[t_k\left(T\right)\right]$ is the expected number of times that arm k is played up to time T.

Next, by the definition of the UCB algorithm and Equation (2.4),

$$\begin{aligned}\mathbb{E}\left[t_k\left(T\right)\right] &\leq \mathbb{E}\left[\sum_{n=1}^{T-1} 1\left\{b_k\left(n\right) \geq b_1\left(n\right)\right\}\right] \\ &\leq \sum_{n=1}^{T-1} \mathbb{E}\left[1\left\{b_k\left(n\right) \geq \mu_1\right\}\right] + \sum_{n=1}^{T-1} \mathbb{E}\left[1\left\{b_1\left(n\right) \leq \mu_1\right\}\right] \\ &= \sum_{n=1}^{T-1} \mathbb{P}\left(b_k\left(n\right) \geq \mu_1\right) + \sum_{n=1}^{T-1} \mathbb{P}\left(b_1\left(n\right) \leq \mu_1\right).\end{aligned}$$

Consider the first term on the RHS,

$$\begin{aligned}\mathbb{P}\left(b_k\left(n\right) \geq \mu_1\right) &= \mathbb{P}\left(\frac{1}{t_k\left(n\right)} \sum_{j=1}^{t_k(n)} X_k\left(j\right) + \sqrt{\frac{K \log n}{t_k\left(n\right)}} \geq \mu_1\right) \\ &= \mathbb{P}\left(\frac{1}{t_k\left(n\right)} \sum_{k=j}^{t_k(n)} X_k\left(j\right) - \mu_k \geq \mu_1 - \mu_k - \sqrt{\frac{K \log n}{t_k\left(n\right)}}\right).\end{aligned} \tag{2.5}$$

To study the probability of such deviations from the mean, we recall the Hoeffding bound. However, we cannot directly apply the Hoeffding bound to Equation (2.5), since

1. $t_k\left(n\right)$ is a random variable, not a constant,
2. $\mu_1 - \mu_k - \sqrt{\frac{K \log n}{t_k\left(n\right)}}$ may not be positive, as required in the Hoeffding bound.

To handle the second issue above, we need to ensure that $t_k\left(n\right)$ is sufficiently large. In particular, suppose $t_k\left(n\right)$ is such that $\mu_1 - \mu_k - \sqrt{\frac{K \log n}{t_k\left(n\right)}} \geq \sqrt{\frac{K \log n}{t_k\left(n\right)}}$, or equivalently, $t_k\left(n\right) \geq \frac{4K \log n}{\left(\mu_1 - \mu_k\right)^2}$. Let $N = \frac{4K \log n}{\left(\mu_1 - \mu_k\right)^2}$, and let $Y\left(n\right)$ be a random variable denoting the arm that is played at time n. Then,

$$\begin{aligned}t_k\left(n\right) &= \sum_{j=1}^{n-1} 1\left\{Y\left(j\right) = k\right\} \\ &= \sum_{j=1}^{n-1} 1\left\{Y\left(j\right) = k\right\} \cdot 1\left\{t_k\left(j\right) < N\right\} + \sum_{j=1}^{n-1} 1\left\{Y\left(j\right) = k\right\} \cdot 1\left\{t_k\left(j\right) \geq N\right\}.\end{aligned}$$

Since $\sum_{j=1}^{n-1} 1\{Y(j)=k\}\cdot 1\{t_k(j)<N\} \leq N$ and $\{Y(j)=k\} \implies \{c_k(j) \geq c_1(j)\}$, we have

$$\begin{aligned} t_k(n) &\leq N + \sum_{j=1}^{n-1} 1\{Y(j)=k\} \cdot 1\{t_k(j) \geq N\} \\ &\leq N + \sum_{j=1}^{n-1} 1\{c_k(j) \geq c_1(j)\} \cdot 1\{t_k(j) \geq N\}, \end{aligned}$$

and using Equation (2.4),

$$\begin{aligned} \mathbb{E}[t_k(n)] &\leq N + \mathbb{E}\left[\sum_{j=1}^{n-1} 1\{c_k(j) \geq c_1(j), t_k(j) \geq N\}\right] \\ &\leq N + \sum_{j=1}^{n-1} \mathbb{P}(c_k(j) \geq \mu_1, t_k(j) \geq N) + \mathbb{P}(c_1(j) \leq \mu_1, t_k(j) \geq N). \end{aligned} \tag{2.6}$$

Note that

$$\begin{aligned} \{c_k(j) \geq \mu_1, t_k(j) \geq N\} &= \left\{ \frac{1}{t_k(j)} \sum_{l=1}^{t_k(j)} X_k(l) + \sqrt{\frac{K \log j}{t_k(j)}} \geq \mu_1, t_k(j) \geq N \right\} \\ &\subseteq \bigcup_{L=N}^{j-1} \left\{ \frac{1}{L} \sum_{l=1}^{L} X_k(l) + \sqrt{\frac{K \log j}{L}} \geq \mu_1 \right\}, \end{aligned}$$

since $t_k(j) \leq j-1$. Thus,

$$\begin{aligned} \mathbb{P}(c_k(j) \geq \mu_1, t_k(j) \geq N) &\leq \sum_{L=N}^{j-1} \mathbb{P}\left(\frac{1}{L} \sum_{l=1}^{L} X_k(l) + \sqrt{\frac{K \log j}{L}} \geq \mu_1 \right) \\ &= \sum_{L=N}^{j-1} \mathbb{P}\left(\frac{1}{L} \sum_{l=1}^{L} X_k(l) - \mu_k \geq \mu_1 - \mu_k - \sqrt{\frac{K \log j}{L}} \right) \\ &\leq \sum_{L=N}^{j-1} \mathbb{P}\left(\frac{1}{L} \sum_{l=1}^{L} X_k(l) - \mu_k \geq \sqrt{\frac{K \log j}{L}} \right) \\ &\leq \sum_{L=N}^{j-1} \exp\left(-2L \cdot \frac{K \log j}{L} \right) \\ &= \sum_{L=N}^{j-1} \frac{1}{j^{2K}} \leq \frac{1}{j^{2K-1}}. \end{aligned}$$

Combining, we have that

$$\sum_{j=1}^{n-1} \mathbb{P}\left(c_k\left(j\right) \geq \mu_1, t_k\left(j\right) \geq N\right) \leq \sum_{j=1}^{n-1} \frac{1}{j^{2K-1}} \leq \frac{\pi^2}{6},$$

with the choice $K = 1.5$. Similarly, we have that

$$\begin{aligned}
\mathbb{P}\left(c_1\left(j\right) \leq \mu_1, t_k\left(j\right) \geq N\right) &= \mathbb{P}\left(\frac{1}{t_1\left(j\right)} \sum_{l=1}^{t_1(j)} X_1\left(l\right) + \sqrt{\frac{K \log j}{t_1\left(j\right)}} \leq \mu_1, t_k\left(j\right) \geq N\right) \\
&\leq \sum_{L=N}^{j-1} \mathbb{P}\left(\frac{1}{L} \sum_{l=1}^{L} X_1\left(l\right) \leq \mu_1 - \sqrt{\frac{K \log j}{L}}\right) \\
&\leq \sum_{L=1}^{j-1} \exp\left(-2L \cdot \frac{K \log j}{L}\right) \leq \frac{1}{j^{2K-1}}, \\
\sum_{j=1}^{n-1} \mathbb{P}\left(c_1\left(j\right) \leq \mu_1, t_k\left(j\right) \geq N\right) &\leq \frac{\pi^2}{6}.
\end{aligned}$$

Putting everything together, we have that

$$\begin{aligned}
\mathbb{E}\left[t_k\left(T\right)\right] &\leq \frac{6 \log T}{\left(\mu_1 - \mu_k\right)^2} + \frac{\pi^2}{3}, \\
R\left(T\right) &\leq \sum_{k:\mu_k<\mu_1} \left[\frac{6 \log T}{\left(\mu_1 - \mu_k\right)} + \frac{\pi^2}{3} \cdot \left(\mu_1 - \mu_k\right)\right].
\end{aligned}$$

To summarize, we have shown the following theorem:

Theorem 2.3.4 *(UCB1) If we play the arm with the largest index*

$$b_k\left(n\right) = \hat{\mu}_k\left(n\right) + \sqrt{\frac{\log n}{t_k\left(n\right)}}$$

at each time-step, then

$$R\left(T\right) \leq \sum_{k:\mu_k<\mu_1} \left[\frac{6 \log T}{\left(\mu_1 - \mu_k\right)} + \frac{\pi^2}{3} \cdot \left(\mu_1 - \mu_k\right)\right].$$

If we replace the Hoeffding bound with the Chernoff bound, we can motivate the tighter result and algorithm of KL-UCB, found in [24]. Although this algorithm obtains asymptotically optimal regret, the exploration bonus no longer has a sim-

ple closed-form solution, and is slower in practice to compute indices because of this. Suppose we choose the exploration bonus $\epsilon_{k,n}$ so that $\mathbb{P}\left(b_1\left(n\right) \leq \mu_1\right) = \frac{1}{n\left(\log n\right)^c}$, for some $c \geq 2$. This looser bound (c.f. $\frac{1}{n^2}$ in the previous analysis) allows for more slack for making such an error, but still maintains the property that the integrated error over all n will be bounded. Then, we can compute

$$\begin{aligned}\mathbb{P}\left(b_1\left(n\right) \leq \mu_1\right) &= \mathbb{P}\left(\frac{1}{t_1\left(n\right)}\sum_{i=1}^{t_1(n)} X_{1,i} + \epsilon_{1,n} \leq \mu_1\right) \\ &= \exp\left(-t_1\left(n\right) \cdot D\left(\mu_1 - \epsilon_{1,n}||\mu_1\right)\right) = \frac{1}{n\left(\log n\right)^c},\end{aligned}$$

and thus $t_1\left(n\right) \cdot D\left(\mu_1 - \epsilon_{1,n}||\mu_1\right) = \log n + c\log\log n$. However, the exploration bonus ϵ should only depend on $\hat{\mu}_k$, t_k, and n, and not on μ_1 as that is unknown to the algorithm. Since $\hat{\mu}_1 \to \mu_1$ and we want $\epsilon_{1,n} \to 0$, we replace $D\left(\mu_1 - \epsilon_{1,n}||\mu_1\right)$ by $D\left(\hat{\mu}_1||\hat{\mu}_1 + \epsilon_{1,n}\right) = D\left(\hat{\mu}_1||b_1\right)$. That is, we set the index of arm 1 (and by symmetry, for any arm k) to be:

$$b_k = \max\left\{q \in \left[\hat{\mu}_k, 1\right] : t_k\left(n\right) \cdot D\left(\hat{\mu}_k||q\right) \leq \log n + c\log\log n\right\},$$

which is the index computed by KL-UCB. Using this policy, [24] shows that

$$\limsup_{T\to\infty} \frac{\mathbb{E}\left[t_k\left(T\right)\right]}{\log\left(T\right)} \leq \frac{1}{D\left(\mu_k||\mu_1\right)},$$

and thus proves the following theorem, which asymptotically matches the Lai and Robbins lower-bound:

Theorem 2.3.5 *(KL-UCB) If we play the arm with the largest index*

$$b_k = \max\left\{q \in \left[\hat{\mu}_k, 1\right] : t_k\left(n\right) \cdot D\left(\hat{\mu}_k||q\right) \leq \log n + 3\log\log n\right\}$$

at each time-step, then

$$\limsup_{T\to\infty} \frac{R\left(T\right)}{\log\left(T\right)} \leq \sum_{k:\mu_k<\mu_1} \frac{\mu_1 - \mu_k}{D\left(\mu_k||\mu_1\right)}. \tag{2.7}$$

The full proof of this requires a different concentration inequality [31], the sim-

plest version of which states that for all $\epsilon > 1$ and all $n \geq 1$,

$$\mathbb{P}\left\{\hat{\mu}_1(n) < \mu_1, D\left(\hat{\mu}_1(n)\,||\mu_1\right) \geq \frac{\epsilon}{t_1(n)}\right\} \leq e\left\lceil \epsilon \log n \right\rceil \exp\left(-\epsilon\right).$$

From this, it immediately follows that

$$\begin{aligned}\mathbb{P}\left(b_1 \leq \mu_1\right) &= \mathbb{P}\left(\hat{\mu}_1(n) < \mu_1, D\left(\hat{\mu}_1(n)\,||\mu_1\right) > \frac{\log n + 3\log\log n}{t_1(n)}\right) \\ &\leq e\left\lceil (\log n + 3\log\log n)\log n \right\rceil \frac{1}{n\left(\log n\right)^3},\end{aligned}$$

for which the leading term, of order $\dfrac{1}{n\log n}$, integrates to $\log\log n$. Thus the regret is written as

$$\begin{aligned}\mathbb{E}\left[t_k(T)\right] &\leq \sum_{n=1}^{T}\mathbb{P}\left(b_k(n) \geq \max_{k'\neq k} b_{k'}(n)\right) \leq \sum_{n=1}^{T}\mathbb{P}\left(b_k(n) \geq b_1(n)\right) \\ &\leq \sum_{n=1}^{T}\mathbb{P}\left(b_k(n) \geq \mu_1\right) + \sum_{n=1}^{T}\mathbb{P}\left(b_1(n) \leq \mu_1\right),\end{aligned}$$

where the second term is $\Theta\left(\log\log T\right)$. Next, to compute the first term,

$$\begin{aligned}\mathbb{P}\left(b_k(n) \geq \mu_1\right) &\leq \mathbb{P}\left(t_k(n) \leq \frac{\log n + 3\log\log n}{D\left(\hat{\mu}_k(n)\,||\mu_1\right)}\right), \\ \sum_{n=1}^{T}\mathbb{P}\left(b_k(n) \geq \mu_1\right) &\leq \sum_{n=1}^{T}\mathbb{P}\left(D\left(\frac{\sum_{i=1}^{n} X_{k,i}}{n}||\mu_1\right) \leq \frac{\log T + 3\log\log T}{n}\right) \\ &\leq n_0 + \sum_{n=n_0}^{T}\mathbb{P}\left(D\left(\frac{\sum_{i=1}^{n} X_{k,i}}{n}||\mu_1\right) \leq \frac{\log T + 3\log\log T}{n}\right),\end{aligned}$$

where $n_0 = \dfrac{\log T + 3\log\log T}{D_{KL}\left(\mu_k||\mu_1\right)}$, and we have re-indexed n in the second line from denoting actual time to denoting a count of how many times arm k has been played. The final term here can be shown to be $O\left(\sqrt{\log T}\right)$, although the proof is quite involved and not reproduced here. Putting everything together, we have that $\mathbb{E}\left[t_k(T)\right] \leq \dfrac{\log T}{D_{KL}\left(\mu_k||\mu_1\right)} + o\left(\log T\right)$, as desired. To summarize, despite requiring a sharper concentration inequality, the intuition behind the choice of the exploration bonus in KL-UCB is the same as it is for UCB1, namely to pick the exploration bonus large enough so that $\sum_{n=1}^{T}\mathbb{P}\left(b_1(n) \leq \mu_1\right) \in o\left(\log T\right)$.

2.3.2 Problem-independent bounds

In both the Lai and Robbins lower-bound and the result achieved by KL-UCB, the results are *problem-dependent*, meaning the constants in the RHS are a function of μ. In particular, as $\mu_2 \nearrow \mu_1$, we have that $\frac{\mu_1 - \mu_2}{D_{KL}(\mu_2||\mu_1)} \to \infty$. That is, even though for each problem instance we can obtain logarithmic regret, if the parameters are generated in way such that μ_2 and μ_1 can be arbitrarily close, then at any time T, we could have $R(T) \geq c \log T$ for any $c > 0$, i.e., $R(T) \in \omega(\log T)$. We now prove a *problem-independent* upper-bound, one which is not a function of μ, using the results from above. In particular, suppose $\mu = (\mu_1, \mu_1 - \delta)$. The key idea here is to choose a $\delta(T)$ that gives the worst-case (i.e., largest) regret for each time-horizon T. By Theorem 2.3.5, we can write the regret bound in our two-armed case,

$$\begin{aligned} \frac{R(T)}{\log(T)} &\leq \frac{\delta}{D(\mu_1 - \delta||\mu_1)} + o(1) \\ &\leq \frac{\delta}{2\delta^2} + o(1), \end{aligned} \tag{2.8}$$

where we have used Pinsker's inequality. We also have that $R(T) \leq \delta T$, as the regret per time-step is at most δ. Thus a better regret lower-bound is the minimum of these two bounds, and so we set these two (approximately) equal to determine a worst-case $\delta(T)$. Ignoring the $o(1)$ term for now, suppose $\delta(T) \cdot T = \frac{1}{2\delta(T)} \log T$, and thus $\delta(T) = \sqrt{\frac{\log T}{2T}}$. Substituting back into (2.8) yields the following theorem, showing a problem-independent upper-bound:

Theorem 2.3.6 *Under KL-UCB, we have that*

$$R(T) \leq \sqrt{\frac{T \log T}{2}} + o(\log T).$$

CHAPTER 3

BIDDING WITH AVERAGE BUDGETS

3.1 Introduction

In this chapter, we consider the problem of sponsored search from the perspective of a budget-constrained advertiser. Sponsored search is a type of online advertisement where paid links are shown next to the search results of relevant queries, yielding substantial revenue for search engine providers like Microsoft and Google. It is also a problem of theoretical interest, since bidders are competing for overlapping keywords and targeted demographics, and where the number of bidders, ads, and ad slots on webpages are all large. Even though the ad displayed for each query can be modeled as a second price auction, the aforementioned other constraints complicate the problem.

We consider a model from a single bidder's point of view, and in our model the bidder has an average budget constraint, expressed as a constraint on the expected payment, but which can be interpreted as being a constraint on the time average of the payment. As is common in this literature, if we model all other bidders' strategies by a probability distribution over their largest bid, then we get a simple static optimization problem for each bidder. However, the solution to this static optimization problem requires knowledge of the distribution of the maximum of others' bids, as well as the distribution of one's own valuation, and statistics of the arrival process of relevant queries. In more complicated settings with auctions involving multiple ad slots, we would also require click-through rates, the distributions of the second highest bid of the opponents, etc. The main contribution of this work is to show that the average budget model considered here allows us to compute the optimal bid using stochastic approximation without the aforementioned statistical details. We provide upper bounds and simulation results on both the expected regret in the profit, as well as any budget overdraft or underdraft, despite the limited statistical knowledge. For ease of reading, all proofs and

intermediate results are deferred to subsection 3.6.

The distributional assumption that we make about the opponents' bids is often called the mean-field approximation, which has been studied in a Markov decision problem context in [32, 33]. In [32], the focus is on learning the distribution of the valuation, while the focus of [33] is on budget constraints. Our model is closer to that of [33, 34], but our use of an average budget constraint rather than the strict budget constraint they use allows us to obtain a solution that does not require statistical knowledge of the system parameters. We show that under an average budget constraint, an under-bidding factor also appears in the solution, which we then estimate through stochastic approximation (SA) [35, 36, 37]. In [33], a Markov Decision Problem (MDP) must be solved to obtain the factor by which one under-bids an item's true valuation. However, this MDP is nearly intractable; therefore, a fluid approximation is used to calculate this factor in a more tractable manner. However, even this approximation still relies on the knowledge of the probability distribution of the opponents' maximum bid. In [34], under an additional assumption of homogeneous bidders, this under-bidding strategy is shown to be the unique fluid mean-field equilibrium.

Compared to an MDP formulation, the use of an SA approach (when used with a small but fixed stepsize) allows one to track non-stationary behavior on the bids, valuations, relevant query arrivals, etc. Additionally, our formulation allows us to change our bid at whatever time scale we choose. For example, suppose that bids are placed at multiple geographical locations, then computational considerations involved in keeping track of the massive amount of data generated from multiple bids may prevent a query-by-query update of the under-bidding factor. Then it is natural to fix a time period over which the under-bidding factor is held constant, and changed only at the beginning of each period based on the expenditure, aggregated over all locations, during the previous period. Our formulation allows us to incorporate such practical constraints.

We note that the average budget constraint has been used previously in [38] in a different context, from the point of view of a search engine provider. In that model, there is no bidding involved and the goal is to maximize either click-through rates or impressions subject to an average budget constraint.

3.2 Model and Algorithm

We define the sequence of queries to be indexed by $i \in \mathbb{N}$, with a valuation v_i, drawn i.i.d. from some distribution f_v. Suppose the highest bid from the other bidders is b'_i, drawn i.i.d. from $f_{b'}$. We also assume an average budget constraint of B for each time period, indexed by $t \in \mathbb{N}$. We allow the number of queries per time period to be random, and let N_t denote the total number of queries during the time periods $\{1, 2, \ldots, t\}$, where N is assumed to have independent and stationary increments. Thus the number of queries in time period t is $N_t - N_{t-1} \sim f_{N_1}$, where $N_0 = 0$. Two such examples of N are $N_t = r \cdot t$ and a Poisson process of rate r.

Since we observe only v_i, we seek to find a bidding strategy $b(v)$ that will maximize the expected reward while satisfying the average budget constraint, i.e.

$$\max_b \mathbb{E}_{v,b',N_1} \left[\sum_{i=1}^{N_1} (v_i - b'_i) \, I_{\left\{b_i > b'_i\right\}} \right]$$

subject to

$$\mathbb{E}_{v,b',N_1} \left[\sum_{i=1}^{N_1} b'_i I_{\left\{b_i > b'_i\right\}} \right] \leq B.$$

Using a Lagrange multiplier λ^*, we now find

$$\max_b \mathbb{E}_{v,b',N_1} \left[\sum_{i=1}^{N_1} (v_i - b'_i - \lambda^* b'_i) \, I_{\left\{b_i > b'_i\right\}} \right].$$

Note that the bid is a function of v and N_1, but we can treat these values as given when optimizing for b, so the above maximization can be equivalently written as

$$\begin{aligned} &\max_b \mathbb{E}_{b'} \left[\sum_{i=1}^{N_1} (v_i - b'_i (1 + \lambda^*)) \, I_{\left\{b_i > b'_i\right\}} \right] \\ =&\max_b \sum_{i=1}^{N_1} \int_0^{b_i} (v_i - b'_i (1 + \lambda)) \, f_{b'}(b') \, db'. \end{aligned}$$

Differentiating w.r.t. b and equating to 0 yields

$$\begin{aligned} &(v_i - b'_i (1 + \lambda^*)) = 0 \\ \implies &b_i = \frac{v_i}{1 + \lambda^*}. \end{aligned} \tag{3.1}$$

By complementary slackness, λ^* is either the positive solution to

$$\mathbb{E}_{v,b',N_1}\left[\sum_{i=1}^{N_1} b' I_{\left\{\frac{v}{1+\lambda^*}>b'\right\}}\right] - B = 0 \tag{3.2}$$

if it exists, or 0 otherwise. Note that this optimal bidding is similar in form to that found in [33], where they refer to $\dfrac{1}{1+\lambda^*}$ as the bid shading (a.k.a. under-bidding) factor.

Now we focus on computing λ^* when the distributions of b', v and N_1 are unknown. Later we will comment on how to use any available partial knowledge of these distributions. The form of (3.1) and (3.2) suggests the following stochastic approximation update rule, where ϵ_t is the stepsize:

$$\lambda_{t+1} = \left[\lambda_t + \epsilon_t\left(\sum_{i=N_t+1}^{N_{t+1}} b_i' I_{\left\{b_i>b_i'\right\}} - B\right)\right]^+ \tag{3.3}$$

and a bidding rule $b_i = \dfrac{v_i}{1+\lambda_{t(i)}}$, where $t\,(i)$ is the time period that includes query i, i.e.

$$N_{t(i-1)} < i \leq N_{t(i)}.$$

Applying convergence results from [37, 36], we can show that under mild conditions and a sufficiently slowly decreasing sequence ϵ_t, we have that $\lambda_t \to \lambda^*$ a.s., regardless of the initial condition λ_0. One can also use a fixed stepsize (i.e., ϵ_t is a constant) to track any possible non-stationarities in the random processes involved. In this case, instead of almost sure convergence, one can provide probabilistic guarantees on how close λ_t is to λ^* in steady-state.

While the convergence (or closeness) of λ_t to λ^* is important, it is also important to keep track of the total regret (the difference between the expected optimal payoff and the realized payoff under our algorithm) and the amount by which we overshoot or undershoot the budget (called overdraft and underdraft, respectively) as functions of time. More precisely, if the amount that the bidder has spent up to time t is $\tilde{B}_t$, then we define the underdraft to be $B \cdot t - \tilde{B}_t$. If the underdraft is negative, then one can also call it an overdraft.

Note that while convergence may hold for any λ_0, the time required to approach λ^* and the magnitude of the regret and budget underdrafts are dependent on the particular value of λ_0. This suggests that if we are given partial or approximate information about the system distributions, we should take into account $\hat{\lambda}^*$, the

solution to (3.2) using the available distributions in place of $f_{b'}$, f_v, and f_{N_1}, when deciding λ_0.

One extension to the basic model is to replace the second price auction with either a generalized second price (GSP) or Vickrey-Clarke-Groves (VCG) auction with m ad slots, which takes into account click-through rates. Another extension is to consider a fixed total budget over a finite time horizon. One could treat this using the average budget formulation proposed here, and stopping when the budget is exhausted or the time horizon is reached. Based on the simulation results that follow, any budget underdraft or regret at the end of the time horizon will be small.

3.3 Upper Bounds

To analyze the performance of the stochastic approximation policy, we make a slight change. First, let $\Lambda = [0, \lambda_{\max}]$ for some $\lambda_{\max} > 0$. Effectively, we are preventing our bids from going to zero, which is a reasonable assumption in practice. Then, the exact algorithm we use is the following:

$$\lambda_{t+1} = \min\left\{\left[\lambda_t + \epsilon_t\left(\sum_{i=N_t+1}^{N_{t+1}} b_i' I_{\left\{b_i > b_i'\right\}} - B\right)\right]^+, \lambda_{\max}\right\},$$
$$b_i = \frac{v_i}{1+\lambda_{t(i)}}.$$

Define the expected instantaneous budget exceedance

$$Y(\lambda_t) = \mathbb{E}_{v,b'}\left[b' \cdot 1_{\left\{\frac{v}{1+\lambda_t} > b'\right\}} - B\right],$$

the expected instantaneous reward

$$R(\lambda_t) = \mathbb{E}_{v,b'}\left[(v-b') \cdot 1_{\left\{\frac{v}{1+\lambda_t} > b'\right\}}\right],$$

the regret up to time T

$$Regret(T) = \sum_{t=1}^{T}\left\{R(\lambda^*) - \mathbb{E}\left[R(\lambda_t)\right]\right\},$$

and the overdraft up to time T

$$Overdraft\left(T\right)=\sum_{t=1}^{T}\mathbb{E}\left[Y\left(\lambda_{t}\right)\right].$$

Finally, we make the following assumptions:

1. $Y\left(\lambda^{*}\right)=0$.
2. $Y\left(\lambda\right)$ and $R\left(\lambda\right)$ are continuously differentiable over Λ.
3. Y, R are Lipschitz, and Y is strictly decreasing. In particular, this implies the existence of the following constants:

$$\begin{aligned}K_{R}&=\max_{\lambda\in\Lambda}\left|\frac{dR}{d\lambda}\left(\lambda\right)\right|<\infty,\\K_{Y,\min}&=\min_{\lambda\in\Lambda}\left|\frac{dY}{d\lambda}\left(\lambda\right)\right|>0,\\K_{Y,\max}&=\max_{\lambda\in\Lambda}\left|\frac{dY}{d\lambda}\left(\lambda\right)\right|<\infty.\end{aligned}$$

4. $\sigma^{2}=\sup_{\lambda\in\Lambda}E\left[\delta M^{2}\left(\lambda\right)\right]<\infty$.

Example 1 *Assumptions 1-4 are satisfied when v and b' are i.i.d. uniform on* $[0,1]$.

The following is a more realistic model of actual bids in practice.

Example 2 *Assumptions 1-4 are satisfied when v and b' are each i.i.d. truncated log-normal distributions, with (possibly different) parameters μ and σ^{2}, and support $[a,b]$. Namely, these are distributions with p.d.f.*

$$f_{\mu,\sigma^{2},a,b}\left(x\right)=\frac{1_{\{a\leq x\leq b\}}}{x\cdot Z}\cdot\exp\left(-\frac{\left(\log x-\mu\right)^{2}}{2\sigma^{2}}\right),$$

where Z is a normalizing constant.

The following theorems bound the regret and overdraft for the stochastic approximation bidding strategy to both be sublinear functions of the time horizon T, under an appropriate choice of ϵ_{t}. Note that bounding just the regret is insufficient, since one could play, for example, the strategy which always bids the valuation in order

to minimize the regret. This alternative strategy minimizes regret, but comes at the cost of over-bidding and thus violating the budget constraint by a significant amount, namely the expected total amount spent will exceed $B \cdot T$ by an amount linear in T.

Theorem 3.3.1 *Under assumptions 1-4, for any $\delta > 0$, with the choice of $\epsilon_t = \frac{1}{2K_{Y,\min} \cdot (1-\delta)^2} \cdot \frac{1}{t}$, we have that*

$$Regret\,(T) \lesssim \frac{K_R \cdot \sigma}{2K_{Y,\min} \cdot (1-\delta)} \cdot \sqrt{T \log T}.$$

The main idea behind this proof is to use a first order Taylor expansion of the instantaneous regret, as a function of the current estimate λ_t, about the true value λ. We then consider λ_t, which is a Martingale process due to the stochastic approximation being applied. The second moment of λ_t can then be bounded by $O\left(\frac{\log t}{t}\right)$ through solving a recurrence relation. The combination of these two and summing over t yields the desired result.

Theorem 3.3.2 *Under assumptions 1-4, for any $\delta > 0$, with the choice of $\epsilon_t = \frac{1}{2K_{Y,\min} \cdot (1-\delta)^2} \cdot \frac{1}{t}$, we have that*

$$|Overdraft\,(T)| \lesssim \frac{K_{Y,\max} \cdot \sigma}{2K_{Y,\min} \cdot (1-\delta)} \cdot \sqrt{T \log T}.$$

The proof of this theorem is identical to that of 3.3.1, except with a Taylor expansion of the absolute value of the overdraft instead of the regret.

3.4 Numerical Experiments

We verify that this algorithm works well using numerical simulations on synthetic data generated from fixed distributions (unknown to the algorithm). Following [39, 40], we use log-normal distributions to model bids and valuations. Namely, we set our budget to be $B = 0.3$ with time periods of 3 queries each, and generate $v \sim [\ln \mathcal{N}\,(0,1)]_0^{10}$ and $b' \sim [\ln \mathcal{N}\,(1,1)]_0^{10}$ independently, where $\ln \mathcal{N}\,(\mu, \sigma^2)$ is the log-normal distribution and $[X]_a^b$ indicates the truncated distribution of X on $[a, b]$. These assumptions model a budget-constrained problem where our valuations are less than the maximum bid from the rest of the bidders, in expectation.

If these parameters were known, solving (3.2) would yield the optimal Lagrange multiplier $\lambda^* \approx 1.18$. One goal of the simulation is then to verify how well our algorithm tracks this Lagrange multiplier without knowledge of the system parameters. We numerically simulate our algorithm starting with $\lambda_0 = 0$, $\lambda_0 = 1$, and $\lambda_0 = 1.5$, with a time-varying stepsize $\epsilon_t = t^{-0.5}$ and a time horizon of $3 \cdot 10^3$, and plot the empirical means based on 10^6 sample-paths in Figure 3.1. The same initial conditions and parameters, but with $\epsilon_t = t^{-0.8}$ and a time horizon of $3 \cdot 10^4$, are shown in Figure 3.2. Simulation results for constant ϵ_t and for $\epsilon_t = t^{-1}$ are omitted due to space considerations. In the figures, we plot the empirical behavior of λ_t, the regret $\sum_{i=1}^{N_{t+1}} (v_i - b'_i) \cdot \left(I_{\left\{ \frac{v_k}{1+\lambda^*} > b'_k \right\}} - I_{\left\{ \frac{v_k}{1+\lambda_k} > b'_k \right\}} \right)$, and the budget underdraft $\sum_{s=1}^{t} \left[B - \sum_{i=N_s+1}^{N_{s+1}} b'_i I_{\{b_i > b'_i\}} \right]$.

For $\epsilon_t = t^{-0.5}$, we see that the choice of λ_0 affects the initial transients of λ_t, and whether $\lambda_0 \gtrless \lambda^*$ determines the initial sign of the underdraft. Furthermore, after λ_t has become close to λ^*, the regret increases very slowly in t. For $\epsilon_t = t^{-0.8}$, we additionally see that the choice of λ_0 can affect the convergence rate of λ_t, with underestimates overshooting and then slowly descending towards λ^* due to the asymmetry in the update equation for λ_t.

In the simulations presented, the regret and budget underdraft are within a few percent of the total profit and budget, respectively, at the end of the time horizon. As would be expected, there is a dependence of the regret and budget underdraft on the initial condition λ_0, with smaller regret when λ_0 is close to λ^*. This suggests that whenever possible, the algorithm should be warm-started with λ_0 set to an approximation of λ^*. As mentioned before, if the exact distributions of v, b', and N_1 are unknown, but one has empirical estimates of these distributions (or perhaps full information for some), one could solve (3.2) for $\hat{\lambda}^*$ using those surrogate distributions. Additionally, one could also intentionally bias the initial choice λ_0 away from λ^*, possibly to avoid budget overdrafts or to speed up convergence.

3.5 Conclusion

We have considered the problem of auctions with a large number of bidders and an average budget constraint. Through the use of a mean-field approximation, we can formulate the problem as a static optimization problem for each player. One could solve this explicitly given knowledge of the opponents' bid distributions and

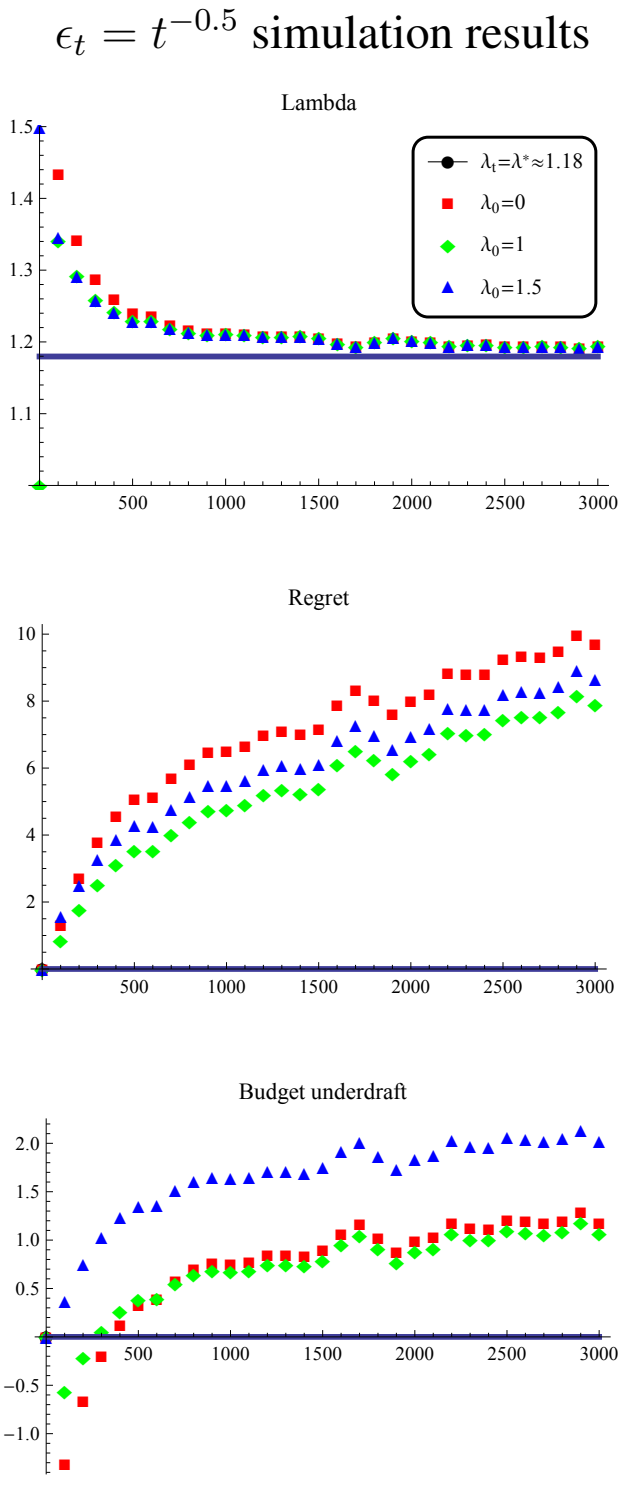

Figure 3.1: Empirical means of λ_t, regret, and budget underdraft vs. t in the $\epsilon_t = t^{-0.5}$ simulation. For comparison, at the end of the time-horizon, the optimal profit is 871 and the total budget is 300; the regret and budget underdraft are both approximately 1% of these values.

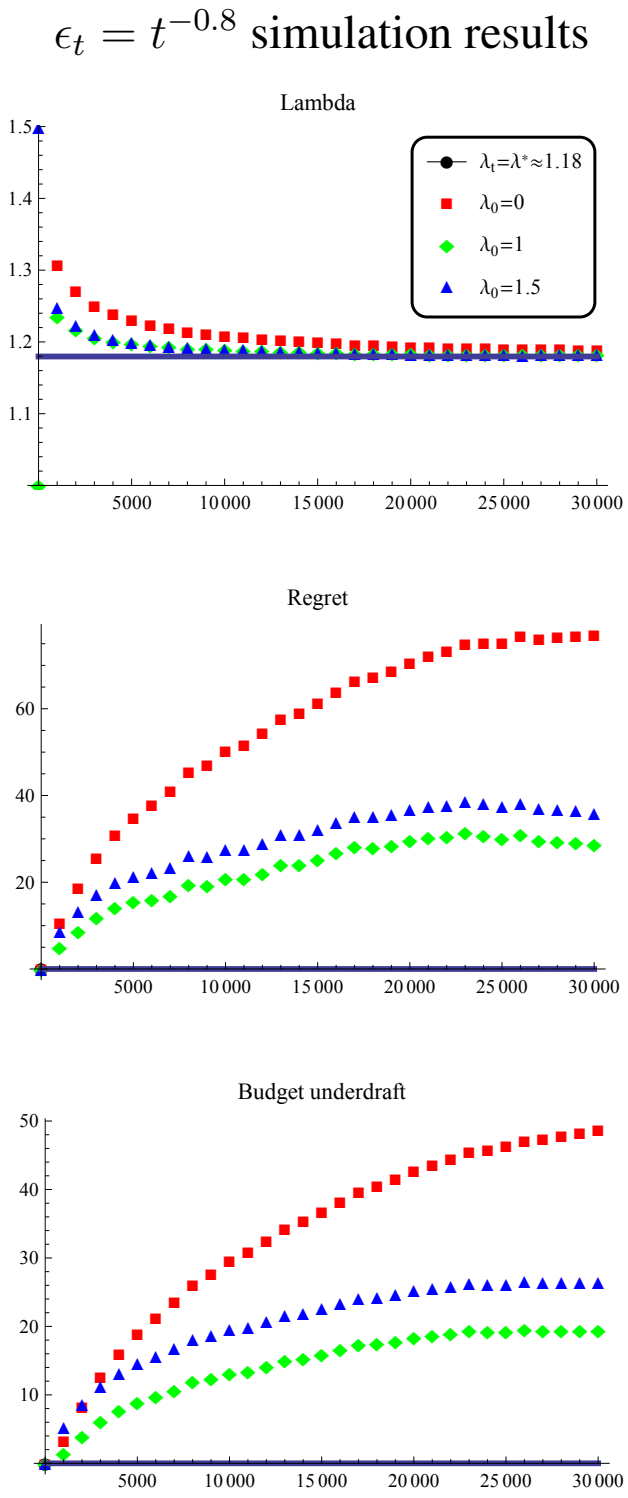

Figure 3.2: Empirical means of λ_t, regret, and budget underdraft vs. t in the $\epsilon_t = t^{-0.8}$ simulation. For comparison, at the end of the time-horizon, the optimal profit is $8.71 \cdot 10^3$ and the total budget is $3 \cdot 10^3$; the maximum absolute regret and budget underdraft are approximately 1% and 2% of these values, respectively.

the valuation distribution, but by applying stochastic approximation, we provide an algorithm that can converge to the optimal bid even without such information. Furthermore, this formulation is significantly simpler to analyze than the dynamic auctions considered previously under a strict budget constraint, and is capable of multiple extensions, including to VCG auctions that incorporate click-through rates. The computations involved in implementing this algorithm are rather minimal, especially if λ_t is updated per time period instead of per query. Our simulation results on synthetic data offer empirical support for bounds on the regret and any budget overdrafts or underdrafts, but of course depending on the assumed distributions, λ_0, choice of ϵ_t, and time horizon T. Warm-starting with an estimate for λ^* based on limited statistical knowledge is also considered.

3.6 Proofs

3.6.1 Proof of Theorem 3.3.1

Since $R(\lambda)$ is continuously differentiable over Λ by assumption 2, using a first order Taylor expansion and assumption 3,

$$\begin{aligned} R(\lambda) &= R(\lambda^*) + \frac{dR}{d\lambda}\left(\tilde{\lambda}\right) \cdot (\lambda - \lambda^*) \\ &\geq R(\lambda^*) - K_R \cdot (\lambda - \lambda^*), \end{aligned}$$

where $\tilde{\lambda}$ is between λ and λ^*. Thus,

$$\begin{aligned} Regret(T) &= \sum_{t=1}^{T} \{R(\lambda^*) - \mathbb{E}[R(\lambda_t)]\} \\ &\leq K_R \cdot \sum_{t=1}^{T} \mathbb{E}[(\lambda_t - \lambda^*)] \\ &\leq K_R \cdot \sum_{t=1}^{T} \sqrt{\mathbb{E}\left[(\lambda_t - \lambda^*)^2\right]} \\ &= K_R \cdot \sum_{t=1}^{T} \sqrt{W_t}. \end{aligned} \tag{3.4}$$

To obtain a bound on this, we will show that $W_t \in O\left(\frac{\log t}{t}\right)$.

Since $Y(\lambda)$ is also continuously differentiable over Λ, we again use a first order Taylor expansion,

$$\begin{aligned} Y(\lambda) &= Y(\lambda^*) + \frac{dY}{d\lambda}\left(\tilde{\lambda}\right) \cdot (\lambda - \lambda^*) \\ &\overset{(a)}{=} \frac{dY}{d\lambda}\left(\tilde{\lambda}\right) \cdot (\lambda - \lambda^*) \end{aligned} \tag{3.5}$$

where $\tilde{\lambda}$ is between λ and λ^*, (a) is by assumption 1 and $\frac{dY}{d\lambda}(\lambda) \in [-K_{Y,\max}, -K_{Y,\min}]$ is by assumption 3. We then have the following recurrence:

$$\begin{aligned} W_{t+1} =& \mathbb{E}\left[(\lambda_{t+1} - \lambda^*)^2\right] \\ =& \mathbb{E}\left[(\lambda_t - \lambda^* + \epsilon_t\left[Y(\lambda_t) + \delta M_t\right])^2\right] \\ =& W_t + 2\epsilon_t \mathbb{E}\left[(\lambda_t - \lambda^*) \cdot \left[Y(\lambda_t) + \delta M_t\right]\right] + \\ & \epsilon_t^2 \mathbb{E}\left[Y(\lambda_t)^2 + 2Y(\lambda_t) \cdot \delta M_t + \delta M_t^2\right] \\ \overset{(a)}{=}& W_t + 2\epsilon_t \mathbb{E}\left[\frac{dY}{d\lambda}\left(\tilde{\lambda}_t\right) \cdot (\lambda_t - \lambda^*)^2\right] + \\ & \epsilon_t^2 \left\{\mathbb{E}\left[\left(\frac{dY}{d\lambda}\left(\tilde{\lambda}_t\right)\right)^2 \cdot (\lambda_t - \lambda^*)^2\right] + \mathbb{E}\left[\delta M_t^2\right]\right\} \\ \overset{(b)}{\leq}& \left(1 - 2\epsilon_t K_{Y,\min} + \epsilon_t^2 K_{Y,\max}^2\right) W_t + \epsilon_t^2 \sigma^2, \end{aligned}$$

where (a) follows from δM_t being a Martingale difference sequence w.r.t. λ_t, i.e. $E\left[\delta M_t | \lambda_t\right] = 0$, and (b) follows from assumptions 3 and 4. Note that as long as $\epsilon_t \to 0$, then $\forall \delta > 0$, $\exists t_{\min}(\delta) \geq 1$ such that $\forall t > t_{\min}(\delta)$, $\left(1 - 2\epsilon_t K_{Y,\min} + \epsilon_t^2 K_{Y,\max}^2\right) \leq \left(1 - 2(1-\delta)\epsilon_t K_{Y,\min}\right)$. Choosing $\epsilon_t = \frac{K_\epsilon}{t+1} = \frac{1}{2K_{Y,\min} \cdot (1-\delta)} \cdot \frac{1}{t+1}$, which satisfies $\epsilon_t \to 0$, we then have

$$W_{t+1} \leq \left(1 - \frac{1}{t+1}\right) W_t + \frac{K_\epsilon^2 \sigma^2}{(t+1)^2}, \ \forall t \geq t_{\min}.$$

This can be solved by recursive substitution to yield

$$\begin{aligned} W_t &\leq W_{t_{\min}} \cdot \frac{t_{\min}}{t} + K_\epsilon^2 \sigma^2 \cdot \frac{\sum_{s=t_{\min}+1}^{t} \frac{1}{s}}{t} \\ &\leq W_{t_{\min}} \cdot \frac{t_{\min}}{t} + K_\epsilon^2 \sigma^2 \cdot \frac{\log t}{t}, \ \forall t \geq t_{\min}, \end{aligned}$$

where we have used the fact $t_{\min} \geq 1$. Since for any $\delta > 0$, $t_{\min}$ is finite, we have the following asymptotic inequality,

$$W_t \lesssim K_\epsilon^2 \sigma^2 \cdot \frac{\log t}{t}. \tag{3.6}$$

Thus, combining (3.4) and (3.6),

$$\begin{aligned} Regret\,(T) &\leq K_R \cdot \sum_{t=1}^{T} \sqrt{W_t} \\ &\lesssim K_R \cdot K_\epsilon \sigma \cdot \sum_{t=1}^{T} \sqrt{\frac{\log t}{t}} \\ &\sim \frac{K_R \cdot \sigma}{2K_{Y,\min} \cdot (1-\delta)} \cdot \sqrt{T \log T}. \end{aligned}$$

□

3.6.2 Proof of Theorem 3.3.2

The proof of this theorem follows that of Theorem 3.3.1 very closely.

$$\begin{aligned} |Overdraft\,(T)| &= \left| \sum_{t=1}^{T} \mathbb{E}\left[Y\left(\lambda_t\right)\right] \right| \\ &\overset{(a)}{=} \left| \sum_{t=1}^{T} \mathbb{E}\left[\frac{dY}{d\lambda}\left(\tilde{\lambda}_t\right) \cdot \left(\lambda_t - \lambda^*\right) \right] \right| \\ &\overset{(b)}{\leq} K_{Y,\max} \cdot \sum_{t=1}^{T} \sqrt{W_t} \\ &\overset{(c)}{\lesssim} \frac{K_{Y,\max} \cdot \sigma}{2K_{Y,\min} \cdot (1-\delta)} \cdot \sqrt{T \log T}, \end{aligned}$$

where (a) follows from (3.5), (b) from assumption 3 and the same technique as (3.4), and (c) from (3.6). □

CHAPTER 4

BANDITS WITH BUDGETS

4.1 Introduction

The multi-armed bandit (MAB) involves an agent 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 1930s, [41], [42].

For discrete bandits with uncorrelated arms, a notable result is [7], 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 [43], [24], [44], with good finite-time behavior.

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 [45], shortest-path routing [46], search engines [47] and ad-display optimization [48].

For discrete bandits several structures have been studied: unimodal [49], combinatorial [50], arms with lower bounded differences [30] 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 [51], unimodal [52], strongly convex [53].

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

Our model is a generalization of the models considered in [48],[54]. 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 below a deterministic number
- General budgets: the total number of plays of an arm is an arbitrary increasing 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 click-through-rates and amounts spent.

4.1.1 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 [55], [48] 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 [7], and uses an inequality of [56] 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 [57]. 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 [48, 54]) which do not take into account the variance of the rewards, and lower bound the 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), and 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 [55] 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.

4.1.2 Related Work

First, for arbitrarily large budgets, the problem reduces to the classical multi-armed bandit problem [7], 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 [7]. Also, one can notice that bandits with budgets are an instance of sleeping bandits [58], which are bandit problems where not all arms may be selected at a given time. However, in [58], 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 [58] to our setting. A different but related setting is that of bandits with a single knapsack constraint, considered in [59, 60]. 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 [55]. 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 [55] are quite different from ours: [55] 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 [7]). Specifically, the authors obtain a minimax regret of $O(\sqrt{T})$ (up to multiplicative logarithmic terms). It is also noted that the algorithms in [55] rely on knowledge of T, whereas our algorithm does not.

The rest of the chapter is organized as follows: In section 4.2 we define the model considered for bandits with budgets. In section 4.3 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 4.3.3 we provide lower bounds on the regret of any uniformly good algorithm in the CPI and CPC case. In section 4.3.4 we provide regret upper bounds for algorithm B-KL-UCB and demonstrate its asymptotic optimality in the CPI and CPC cases. In section 4.4 we assess the finite time performance of B-KL-UCB and its com-

petitors by numerical experiments. Section 4.5 concludes the chapter. For ease of reading, all proofs and intermediate results are found in section 4.6.

4.2 Model

We consider a bandit problem with a finite number of arms $1 \leq K < \infty$ with time horizon $T \geq 0$. Time is discrete; at time $n \in \{1, ..., T\}$ an agent 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 $S_k(t) = \sum_{i=1}^{t} X_k(i)$ 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 exist 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 $\pi^\star$ 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:

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

Simply said, the regret of policy π is the loss in accumulated reward due to 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 chapter we present our results when rewards are Bernoulli, mainly for simplicity and due to the fact that the model originates from ad-display optimization where rewards (click / no click) are indeed Bernoulli. 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 [24] for additional clarification.

4.3 Results

4.3.1 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 $I(p, q) = p \log(\frac{p}{q}) + (1-p) \log(\frac{1-p}{1-q})$ the KL divergence between Bernoulli distributions of parameters p and q.

4.3.1.1 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^*:

$$\overline{c} = \begin{cases} (1 - \sum_{k'=1}^{k^\star - 1} c_{k'}), & \text{if } k^\star \geq 2 \\ 1, & \text{otherwise} \end{cases}.$$

It is noted that $\overline{c} > 0$.

4.3.1.2 CPC case

In the CPC case 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 further define the random variable $\tau_k = \min\{t : S_k(t) \geq Tc_k\}$ to be the number of plays of arm k until Tc_k successes are realized, and the random variable $\overline{\tau}$ to be the number of plays of arm $\tilde{k}$:

$$\overline{\tau} = \begin{cases} T - \sum_{k=1}^{\tilde{k}-1} \tau_k, & \text{if } \tilde{k} \geq 2 \\ T, & \text{otherwise} \end{cases}.$$

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. Finally, we define $\overline{d} = 1 - \sum_{k=1}^{k^\star - 1} d_k > 0$, the fraction of time that such a policy would play arm k^*.

4.3.1.3 High probability events

We use the following convention throughout the remainder of the chapter: 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:

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

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

4.3.2 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:

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

4.3.3 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 4.3.1 and 4.3.2 give lower bounds on the regret of *any* uniformly good algorithm.

Theorem 4.3.1 *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 4.3.2 *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 our calculations of a lower bound on the regret, 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 4.3.4. 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 4.3.1 and 4.3.2 reduce to the well known

result of Lai and Robbins [7].

The proof technique is similar to that of [30, 29, 28], 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 [7], whereas the results in [30] 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 behavior 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 [7] used such an argument, but involved some manipulations of likelihood ratios, whereas we use an inequality of [56] which reduces these calculations to essentially a single line. Also note that, contrary to [30], 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, i.e., $(t_1(T), ..., t_K(T))$, can be used as a test statistic.

Finally, it should be noted that both Theorems 4.3.1 and 4.3.2 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.

4.3.4 Regret upper bounds

In this section we analyze 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 4.3.3. It is a natural extension of KL-UCB [24] proposed for bandits with independent arms, which reaches the Lai-Robbins bound [7].

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 all $1 \leq n \leq T$, select arm $k(n)$ such that $k(n) = \arg\max_{k \in A(n)} b_k(n)$.

4.3.4.1 General budgets

Theorem 4.3.3 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 [54] (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 [54].

The proof is based on the following coupling argument: we show that if $\pi =$ 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 we either have $k^\pi(n) \leq k^{\pi^\star}(n)$, 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 4.6.6, an intermediate result that enables us to deal with bandit problems where the available set of arms is a stochastic process and might depend on the past decisions, something we believe could be useful beyond the scope of this chapter, to analyze problems such as sleeping bandits [58] and knapsack bandits [55].

Theorem 4.3.3 *Consider general budgets. Under policy $\pi =$ 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.

4.3.4.2 CPI case

Theorem 4.3.4, 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 4.3.4 *1. Under policy π = B-KL-UCB, for all $0 < \epsilon < \Delta$ the regret admits the upper bound:*

$$R^{\pi}(T) \leq f(T)\delta^{\epsilon}_{k^\star} + CK(\log(\log(T)) + \epsilon^{-2})$$

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

2. 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 4.3.4 is not simply a specialization of Theorem 4.3.3 to the the CPI case, as the coefficients of $f(T)$ are different in the two cases. In particular, there is no notion of $k^\star$ in the general budget case, whereas we exploit the existence of a $k^\star$ to tighten the upper bound in Theorem 4.3.4.*

4.3.4.3 CPC case

Theorem 4.3.5, gives a finite-time regret upper bound of the regret of B-KL-UCB in the CPI 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 4.3.4, 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 4.3.2 is met by the upper bound, B-KL-UCB is asymptotically optimal.

The proof of Theorem 4.3.5 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 4.6.6 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 4.3.5 *1. 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 by:*

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

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

2. *By corollary:*

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

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

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

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

4.4 Numerical Experiments

4.4.1 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 [57]. 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 $\frac{1}{2}$, $\frac{1}{3}$, ...), so we first prune away any ad with fewer than 100 impressions. The histogram of click-through rates is shown in Figure 4.1. 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 α. We then set K to be the number of different ads that have been displayed when α 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 4.2 and Table 4.1.

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 α, and assign an equal budget to every arm. The budget is used as a parameter in our simulations, since it is unknown.

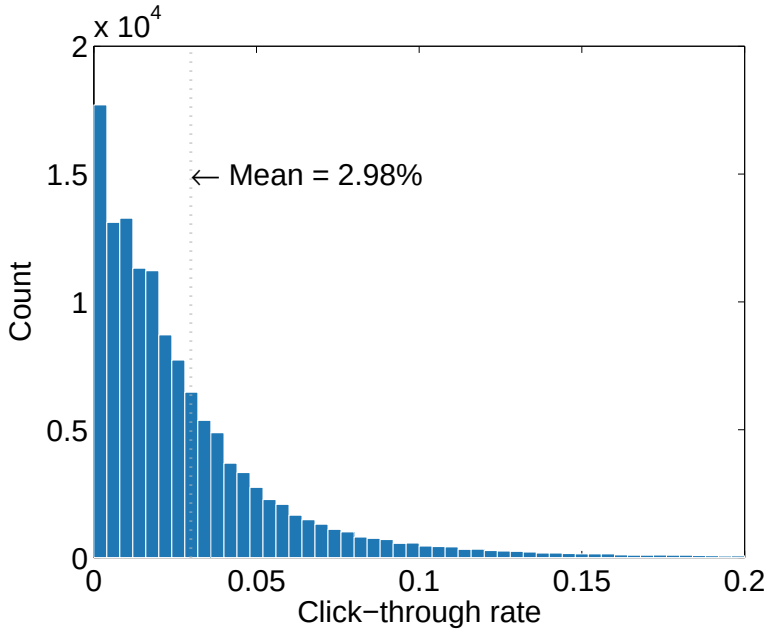

Figure 4.1: Histogram of CTRs for all ads with ≥ 100 impressions, from the KDD Cup dataset.

Table 4.1: List of the 12 non-zero entries in μ for keyword α.

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

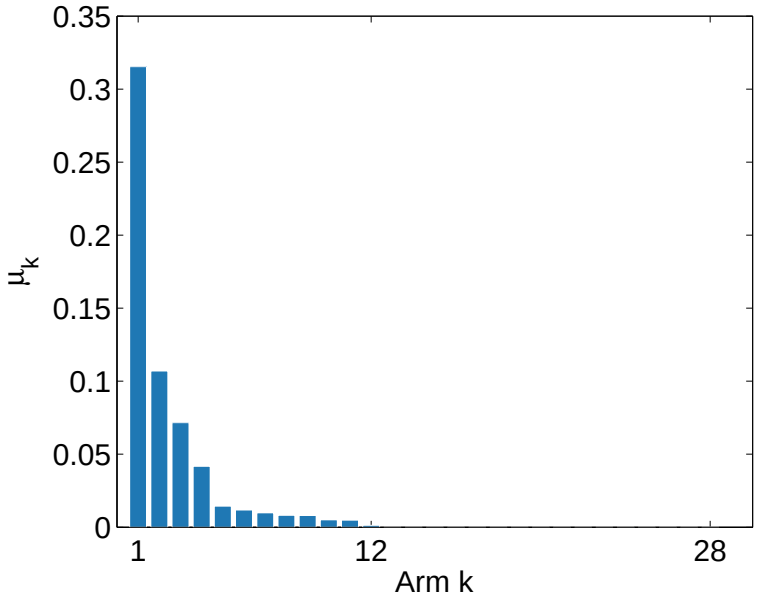

Figure 4.2: Plot of μ_k vs. k for the 28 ads with keyword α.

4.4.2 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 [54, 48]. It is noted that the two algorithms are not identical, but are nearly so. 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 [55] to bandits with budgets.
- PD-BwK (Primal Dual Bandits with Knapsacks): an adaptation of the second algorithm proposed in [55] to bandits with budgets.

In the knapsack bandit problem studied in [55], 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 [55] do not directly apply to our model, nevertheless we attempt to modify their 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 [55], which take into account the additional structure present. 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 descriptions (in pseudo-code) of the tuned versions of Balance-BwK and PD-BwK.

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

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 $c_k(n)$ is a

Algorithm 3 PD-BwK

set $\epsilon = \sqrt{\log(K+1)/B}$, where $B = \min\{T, \min_k Tc_k\}$
in the first K rounds, pull each arm once
initialize the prices for each arm and the price for time, $v_1 = \mathbf{1}_{K+1}$
for $n = K+1, \ldots, T$ **do**
  for each arm $k = 1, \ldots, K$ **do**
    compute UCB estimate for the reward vector,
    $u_{n,k} = \min\{\hat{\mu}_k(n) + \mathrm{rad}(\hat{\mu}_k(n), t_k(n)), 1\}$
    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,
      $L_{n,k} = \max\{\hat{\mu}_k(n) - \mathrm{rad}(\hat{\mu}_k(n), t_k(n)), 0\}$
    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

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 the beginning of each phase:

1. 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.

2. 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 $\dfrac{c_k}{L_{n,k}}$ 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, 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.

4.4.3 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 $T = 1000K = 28000$. 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 for the CPI model (Figure 4.3) and for the CPC model (Figure 4.4). These results show that B-KL-UCB outperforms 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 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 4.5, $\sum_{k=1}^{k^\star} c_k = 1$, and in Figure 4.6, $\sum_{k=1}^{k^\star} c_k > 1$. Similarly, for the CPC model, we again set $k^\star = 6$ and show two plots; in Figure 4.7, $\sum_{k=1}^{k^\star} d_k = 1$, and in Figure 4.8, $\sum_{k=1}^{k^\star} d_k > 1$. The results confirm that B-KL-UCB and PD-BwK again outperform 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.

4.5 Conclusion

In this chapter 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 the regret of 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 when the time horizon is unknown), 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.

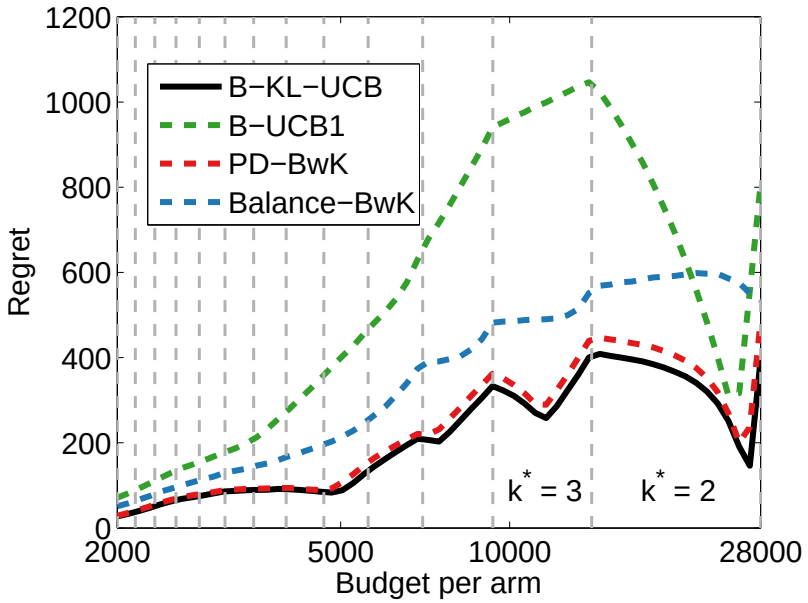

Figure 4.3: Plot of regret at time $T = 28000$ vs. the budget Tc given to each arm, under the CPI model. The dotted vertical lines demarcate $k^\star$ transitions.

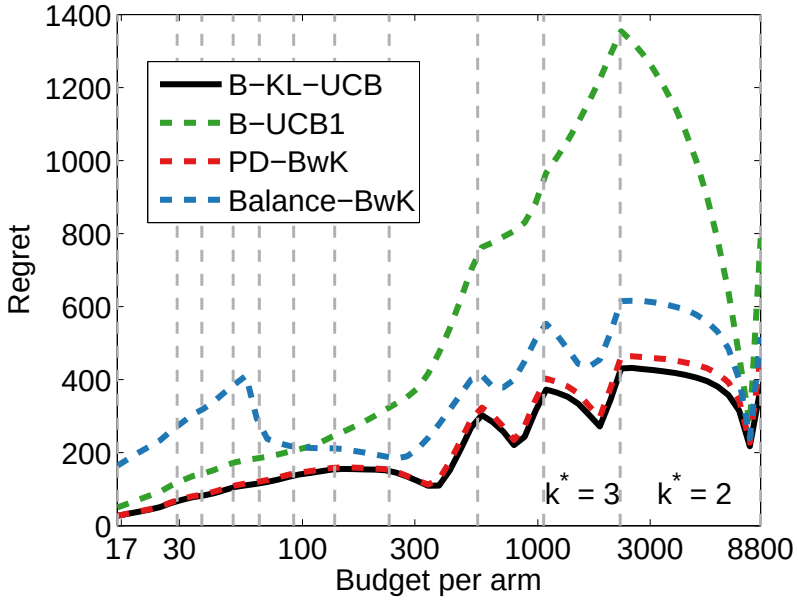

Figure 4.4: Plot of regret at time $T = 28000$ vs. the budget Tc given to each arm, under the CPC model. The dotted vertical lines demarcate $k^\star$ transitions.

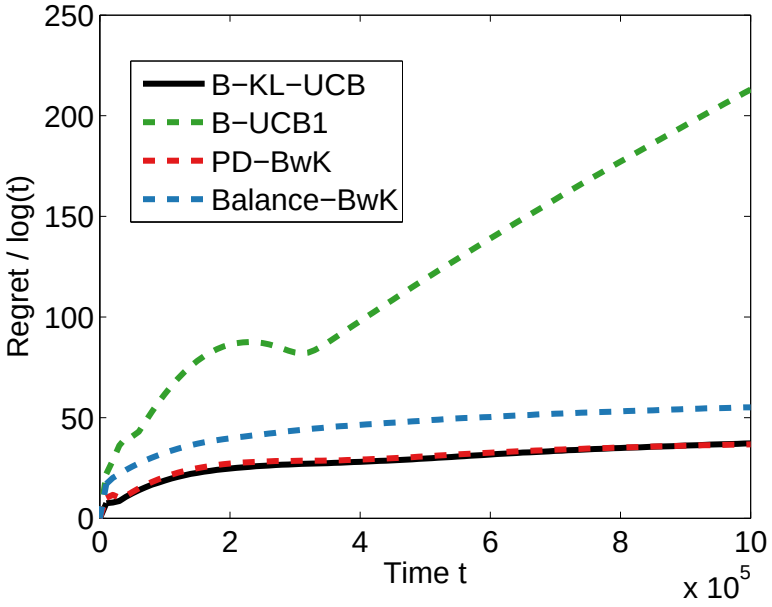

Figure 4.5: Plot of regret vs. time, with $k^\star = 6$ and $\sum_{k=1}^{k^\star} c_k = 1$, under the CPI model. Each arm is given the same incremental budget per timestep of $1/6$.

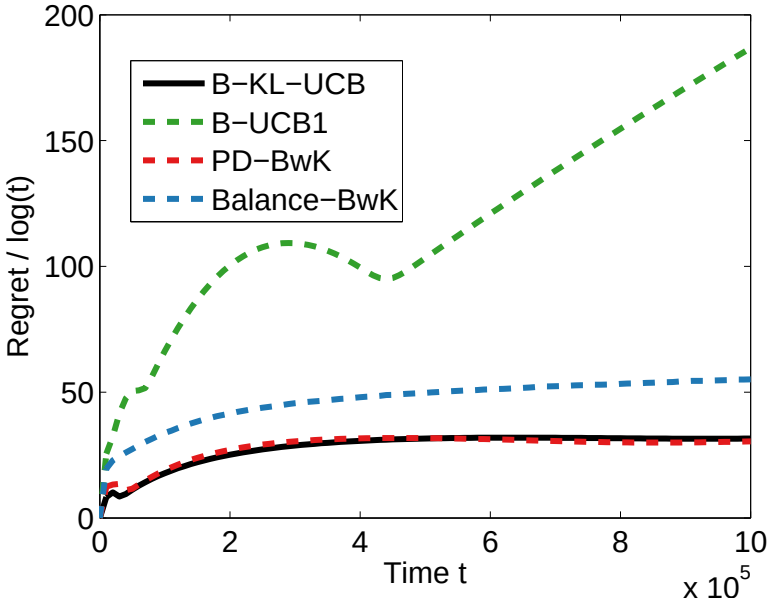

Figure 4.6: Plot of regret vs. time, with $k^\star = 6$ and $\sum_{k=1}^{k^\star} c_k > 1$, under the CPI model. Each arm is given the same incremental budget per timestep of $1/5.5$.

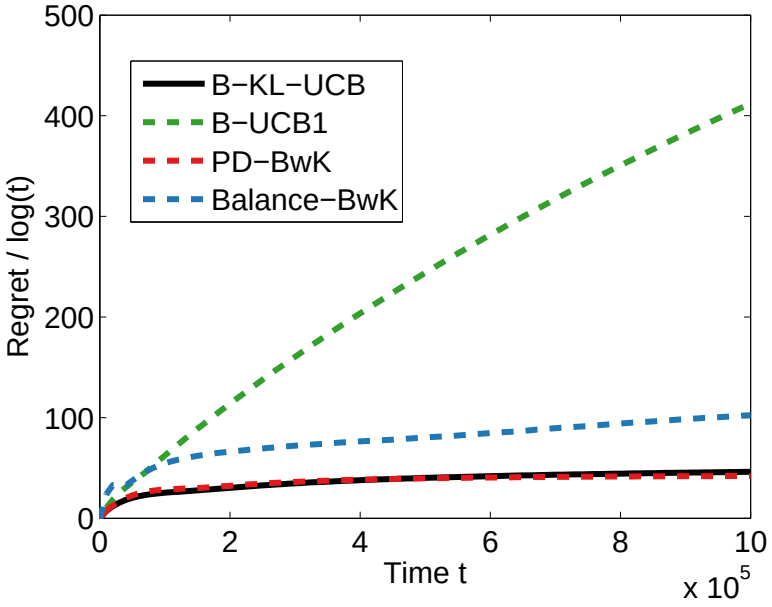

Figure 4.7: Plot of regret vs. time, with $k^\star = 6$ and $\sum_{k=1}^{k^\star} d_k = 1$, under the CPC model. Each arm is given the same incremental budget per timestep of 0.00489.

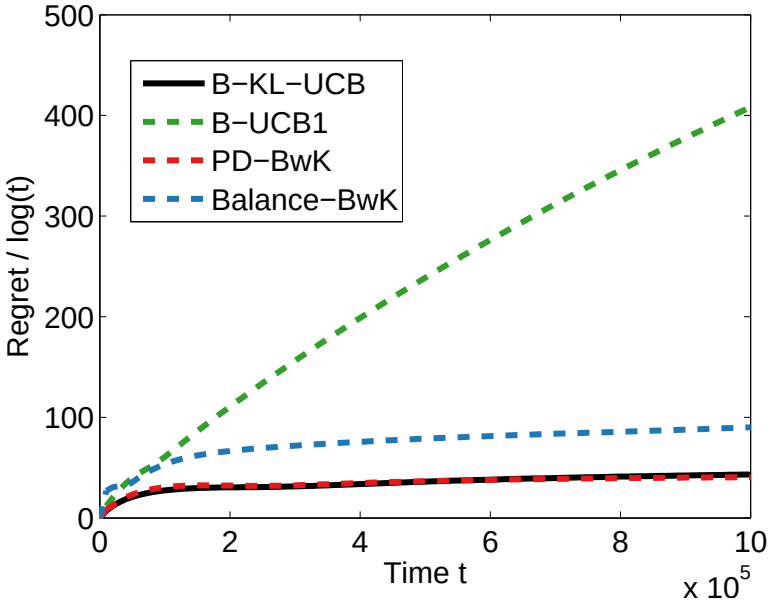

Figure 4.8: Plot of regret vs. time, with $k^\star = 6$ and $\sum_{k=1}^{k^\star} d_k > 1$, under the CPC model. Each arm is given the same incremental budget per timestep of 0.006.

4.6 Proofs

4.6.1 Concentration inequality

The following concentration inequality derived in [31] is instrumental here.

Lemma 4.6.1 *[31] Consider an i.i.d. Bernoulli sequence $\{X(n)\}_{n\geq 1}$ 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 $\delta \geq 0$:

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

4.6.2 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 $\{1,\ldots,K\}$), 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 4.6.2 states that if the greedy policy $\hat{\pi}$ and an arbitrary policy π are run on the same sample path, then vectors $t^{\pi}(n) = (t_1^{\pi}(n),\ldots,t_1^{\pi}(n))$ and $t^{\pi^\star}(n) = (t_1^{\hat{\pi}}(n),\ldots,t_1^{\hat{\pi}}(n))$ 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 4.6.2 could be useful in analyzing those problems as well, although we have not explored that possibility here.

Lemma 4.6.2 *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 $t^{\pi}(0) = (0, 0, \ldots, 0) \lesssim t^{\hat{\pi}}(0)$. 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:

$$\sum_{k'=1}^{k} t^{\pi}_{k'}(n'+1) > \sum_{k'=1}^{k} t^{\hat{\pi}}_{k'}(n'+1),$$

which implies that there exists $k' \leq k$ such that $t^{\pi}_{k'}(n'+1) > t^{\hat{\pi}}_{k'}(n'+1)$, 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. □

4.6.3 Proof of Proposition 1

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

$$M^{\pi}_{n+1} = M^{\pi}_n + \sum_{k=1}^{K} \mathbf{1}\{k^{\pi}(n) = k\}(Y^{\pi}_n - \mu_k)$$

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

so that $\mathbb{E}[M^{\pi}_T] = \mathbb{E}[M^{\pi}_0] = 0$. Hence the expected reward of π can be written as:

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

Using Lemma 4.6.2 , one has $t^{\pi}(T) \lesssim t^{\hat{\pi}}(T)$ a.s. Since $k \to \mu_k$ is decreasing we have:

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

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. □

4.6.4 Lower bounds: intermediate results

The following results are instrumental for establishing our regret lower bounds. Lemma 4.6.3 is an inequality derived in [56] (first noted in [27]), 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 $\mathbf{1}\{A\}$ with A an arbitrary event.

Lemma 4.6.3 *[[56]] 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:*

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

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:

$$\sum_{k'=1}^{K} \mu_k(t_{k'}^{\pi^\star}(T) - t_{k'}^{\pi}(T)) \geq |t_k^{\pi^\star}(T) - t_k^{\pi}(T)|\Delta$$

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

Proof. Straightforward consequence of majorization. □

4.6.5 Proof of Theorem 4.3.1

Proof: 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 $\mathcal{A} = \{t_k^{\pi}(T) \geq T\tilde{c}/2\}$. Denote by $Y^T = (Y(1), ..., Y(T))$ with $Y(n) = X_{k^{\pi(n)}}(t_k^{\pi}(n))$ 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:

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

Notice that $\mathbf{1}\{\mathcal{A}\}$ is a function of Y^T, and apply Lemma 4.6.3:

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

so by taking logarithms:

$$\mathbb{E}_\mu[t_k^{\pi}(T)]I(\mu_k, \mu_{k^\star} + \epsilon) \geq -\log(2) - \log(P(\mathcal{A}) + Q(\mathcal{A}^c)). \tag{4.1}$$

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

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

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, $t_k^{\pi}(T) \geq T\tilde{c}/2$, and when

$\mathcal{A}^c$ occurs, $|\tilde{c} - t_k^\pi(T)| \geq T\tilde{c}/2$, so taking expectations:

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

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 (4.1) we have that:

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

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

□

4.6.6 Proof of Theorem 4.3.2

Proof: We proceed as in the proof of Theorem 4.3.1. 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 4.6.4 (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:

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

Define the event $\mathcal{A} = \{t_k^\pi(T) \geq T\tilde{d}/2\}$. Denote by $Y^T = (Y(1), ..., Y(T))$ with

$Y(n) = X_{k^\pi(n)}(t_k^\pi(n))$ 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:

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

Notice that $\mathbf{1}\{\mathcal{A}\}$ is a function of Y^T, and apply Lemma 4.6.3:

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

so by taking logarithms:

$$\mathbb{E}_\mu[t_k^\pi(T)] I(\mu_k, \mu_{k^\star} + \epsilon) \geq -\log(2) - \log(P(\mathcal{A}) + Q(\mathcal{A}^c)). \tag{4.2}$$

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

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

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 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:

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

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:

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

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 (4.1) we have that:

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

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

□

4.6.7 Upper bounds: technical results

We present some lemmas which are instrumental for the regret analysis of B-KL-UCB in the 3 cases of interest.

Lemma 4.6.4 *For all k and $\epsilon > 0$, we have $\tau_k/T \in [d_k - \epsilon, d_k + \epsilon]$ 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:

$$\mathbb{P}[\tau_k/T \notin [d_k - \epsilon, d_k + \epsilon]] \leq \mathbb{P}[\tau_k \leq T(d_k - \epsilon)] + \mathbb{P}[\tau_k \geq T(d_k + \epsilon)]. \quad (4.3)$$

Consider the first term in the r.h.s. of (4.3). 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 (4.3). 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 4.6.5 *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 $k^\star$, $\overline{d} > 0$. Apply Lemma 4.6.4 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 $k^\star$, $\underline{d} > 0$. Apply Lemma 4.6.4 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 4.6.6 *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:

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

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

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

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

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.

From Lemma 4.6.1, for all k'' we have $\mathbb{E}[|\tilde{B}_{k'',2}|] \leq O(\log(\log(T)))$ so that by union bound $\mathbb{E}[|B_2|] \leq O(K\log(\log(T)))$.

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 [49][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. □

Lemma 4.6.7 *Consider algorithm B-KL-UCB.*

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

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 4.6.1 (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 4.6.1 (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:

$$t_{k'}(T) \leq \frac{8\log(T)}{(\mu_k - \mu_{k'})^2} \text{ w.h.p.} \tag{4.4}$$

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 4.6.4, we have that $\sum_{k'=1}^{k^\star-1} \tau_{k'} \leq T(1+\overline{d})/2$ w.h.p. Using (4.4) we obtain that:

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

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. □

4.6.8 Proof of Theorem 4.3.3

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:

$$r(T) \leq \sum_{k \geq 2}^{K} (\mu_1 - \mu_k)|B_k|, \tag{4.5}$$

with $B_k = \{n \leq T : k^\pi(n) = k, k^{\pi^\star}(n) \leq k-1\}$. Consider $n \in B_k$. From Lemma 4.6.2, 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 4.6.6 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 (4.5) 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. □

4.6.9 Proof of Theorem 4.3.4

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 4.6.7, 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:

$$R^\pi(T) \leq \sum_{k>k^\star} (\mu_{k^\star} - \mu_k)\mathbb{E}[t_k(T)] + O(1). \tag{4.6}$$

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 4.6.6, 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 (4.6) we obtain the announced result:

$$\begin{aligned} R^\pi(T) &\leq f(T) \sum_{k>k^\star} \frac{\mu_{k^\star} - \mu_k}{I(\mu_k + \epsilon, \mu_{k'})} + KC(\log(\log(T)) + \epsilon^{-2}), \\ &= f(T)\delta^\epsilon_{k^\star} + KC(\log(\log(T)) + \epsilon^{-2}) \end{aligned}$$

which concludes the proof. □

4.6.10 Proof of Theorem 4.3.5

Proof: <u>First statement</u>

From Lemma 4.6.5 we have $\tilde{k} \in \{k^\star, k^\star + 1\}$ w.h.p. Define the following

events:

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

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

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

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 4.6.7 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 4.6.6:

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

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 4.6.7 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 4.6.6:

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

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 4.6.7 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:

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

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 4.6.6:

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

Total regret

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:

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

which proves the first statement of the theorem.

Second statement

Using Lemma 4.6.4, 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. □

CHAPTER 5

LINEARLY PARAMETRIZED BANDITS

5.1 Introduction

In this chapter, we are motivated by problems in e-commerce related to the recommendation of one or several items from a large number of related items, where the items are distinguished by relatively few features. If the preferences of an individual customer (or cluster of similar customers) for the individual features are known, then we can predict the user's affinity for each item, and subsequently display the list of items in descending order of relevancy or expected profit. However, we may not know a priori what this preference vector is, so we wish to learn it in an online manner by sequentially presenting the user with an item, observing whether the item is purchased, and then updating an internal estimate of the preference vector. Since we assume that the number of actual features is much smaller than the total number of items, it may be beneficial to learn user preferences for the few individual features rather than for the many individual items. In this way, it is no longer necessary for each item to be investigated in order to estimate the expected reward from that item, reducing the amount of learning required. A key distinction of our model, when compared to previous work on parametrized bandits, is the incorporation of this inherently binary choice customers are faced with: to buy or not to buy.

5.2 Model

Our model consists of a multi-armed bandit with a set u of arms (items) and n underlying parameters (attributes). We will use u_k to indicate the vector of feature weights for item k, and thus we will interchangeably treat u as a matrix. Furthermore, we will assume that $\text{rank}(u) = n$. There is also a fixed but unknown

preference vector $z^* \in \mathbb{R}^n$, with $z^* \neq 0$. The quality $u_k^T z^*$ of arm k is a scalar indicating how desirable the item is to a user. We will use the logistic function f to define the expected reward of an arm k,

$$\mu_k = f\left(u_k^T z^*\right) = \frac{1}{1 + \exp\left(-u_k^T z^*\right)}.$$

Thus, the expected rewards of all of the arms are coupled through z^*. We will assume that the arms are ordered such that $\mu_1 > \sup_{k \neq 1} \mu_k$. At each time-step t up to a finite time horizon T, a policy will choose to pull exactly one arm, call this arm C_t, and a reward X_t will be obtained, where $X_t \sim \text{Ber}\left(\mu_{C_t}\right)$. We wish to find policies π which maximize the total expected reward, $\sum_{t=1}^T X_t$, or equivalently, minimize the expected total regret, $\mathbb{E}^\pi\left[\sum_{t=1}^T \left(\mu_1 - X_t\right)\right] = T\mu_1 - \mathbb{E}^\pi\left[\sum_{t=1}^T \mu_{C_t}\right]$.

5.3 Lower Bounds

In the subsequent section, we will describe a simple algorithm for this problem, which can be used whether the set U is finite or a continuum of arms, and show a regret upper-bound of $O\left(K \cdot \log\left(T\right) \cdot \log^*\left(T\right)\right)$, where $\log^*$ is the iterated logarithm function, which is unbounded but grows far slower than $\log\left(T\right)$. This two-phase algorithm is conceptually based on ϵ-greedy algorithms, and the key idea is that we can explore by sampling some rank-n set of arms and performing matrix inversion in order to estimate the preference vector z^*. Each time-step in an exploitation phase then greedily plays the best arm, given our current estimate of z^*. The choice of scheduling function corresponds to choosing how $\epsilon\left(t\right)$ should vary with time, in order to balance the regret incurred from exploration and exploitation. Although UCB-type algorithms have been designed for similar models [61], the technique of KL-UCB does not appear to fit readily into this context, and to the best of our knowledge no work has generated matching upper- and lower-bounds with the same constant, as has been done for the basic multi-armed bandit. In this section, we will provide a lower-bound which we conjecture is tight, using the decision-theoretic technique presented in Chapter 2. Because the upper-bound result follows the technique of ϵ-greedy policies, it does not explore the arms optimally and cannot provide a tight bound, even if we disregard the $\log^*\left(T\right)$ factor not present in the lower-bound. However, the simplicity of the

two-phase algorithm allows it to be computationally tractable in real-world applications, and captures the dominant $\log T$ factor in the regret bound, which is already a tremendous improvement over the $O(T)$ regret obtained by A/B testing techniques commonly used in industry.

The following theorem is based on the lower bound shown in Chapter 2, but the correlations between arms introduce some additional complications when calculating the multiplicative constant in front of the $\log T$. Instead of constructing a hypothesis test for each individual arm, we construct hypothesis tests for every direction, i.e., each v on the $(n-1)$-dimensional unit sphere. That is, instead of defining a new parameter vector $\lambda(v)$ that is "close" to μ in KL distance, the correlations between arms do not allow us that amount of generality, so we cannot have $\lambda(v)$ be identical to μ in all but one component. Instead, in this correlated arm model, $\lambda(v)$ is fully determined by the n-dimensional parameter $z(v)$ which is itself a scaled version of v, i.e., $\lambda(v)_k = f\left(u_k^T z(v)\right)$. Note that when the system is parametrized by $z(v)$ instead of z^*, the rewards of all arms (except the best arm u_1, for reasons discussed below) will change in general. This leads to terms in the KL divergence between the distribution of rewards under μ and $\lambda(v)$ to include terms from all arms except the best arm. Other than handling these extra terms, the proof here is no different than that for uncorrelated arms.

Theorem 5.3.1 *For any problem instance with finite arms, $u_1^T z^* \neq 0$, and $\mu_1 > \max_{k\neq 1} \mu_k$, for any uniformly good policy π, we have that:*

$$\liminf_{T\to\infty} \frac{R^\pi(T)}{\log T} \geq \max_{v\in V} \left\{ \min_k \frac{\mu_1 - \mu_k}{D_{KL}\left(\mu_k || \lambda(v)_k\right)} \right\},$$

where D_{KL} is the Kullback-Leibler divergence for Bernoulli random variables:

$$D_{KL}(p,q) = p\log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right)$$

and V, $\lambda(v)_k$ are defined below.

Proof: By assumption, $z^* \neq 0$ and $u_1^T z^* \neq 0$. We assert there exists a non-empty set of directions:

$$V = \left\{ v \subseteq \mathbb{R}^n : |v| = 1, \text{sgn}\left(u_1^T v\right) = \text{sgn}\left(u_1^T z^*\right), \max_{k\neq 1} u_k^T v > u_1^T v \right\} :$$

where the signum condition is satisfied on a half-sphere, and the value of $u_1^T v \to 0$

along the boundary of this surface. Note that for any other arm k (as long as u_1 and u_k are not co-linear), $u_k^T v$ will also be positive on an overlapping half-sphere, and hence there will be a point in the neighborhood of the first boundary such that $u_k^T v > u_1^T v$. Since the set of arms is full rank, such an arm k must exist, and thus V is non-empty.

We now fix a direction $v \in V$ and construct a scaled version, $z(v) = \left(\frac{u_1^T z^*}{u_1^T v}\right) \cdot v$. This choice of scaling is so that $\lambda(v)_1 = f\left(u_1^T z(v)\right) = f\left(u_1^T z^*\right) = \mu_1$. If this condition were not satisfied, then the $\sim T$ plays of arm 1 would allow us to easily discern between these two distributions using the empirical reward sequence. Recall that in order to craft the best lower bound, we must choose $\lambda(v)$ to make it as difficult as possible to discern whether rewards are being generated according to $\lambda(v)$ or μ, under the constraint that the index of the best arm differs between $\lambda(v)$ and μ; hence we have this $\lambda(v)_1 = \mu_1$ constraint.

We now have a well-defined $\lambda(v)$, where $\lambda(v)_k = f\left(u_k^T z(v)\right)$ for all k. Next, we write the KL divergence between the distribution of rewards under μ and $\lambda(v)$ (denoted $\mathcal{P}$ and $\mathcal{Q}$, respectively), as:

$$D_{KL}(\mathcal{P}||\mathcal{Q}) = \sum_{k \neq 1} \mathbb{E}_\mu[t_k(T)] \cdot D_{KL}(\mu_k||\lambda(v)_k). \tag{5.1}$$

Our choice of scaling for $z(v)$ has removed the dependence on $t_1(T)$, since $\mu_1 = \lambda(v)_1$ and thus $D_{KL}(\mu_1||\lambda(v)_1) = 0$. We now construct a hypothesis test to decide between H_μ and H_λ, whether μ or $\lambda(v)$ is the underlying parameter, respectively. We base the test on the observed number of times arm 1 is played, deciding in favor of H_μ if $t_1^\pi(T) \geq \frac{T}{2}$, and deciding H_λ otherwise. The probability of errors can then be computed as:

$$\mathbb{P}(e|H_\mu) = \mathbb{P}_\mu\left(t_1 < \frac{T}{2}\right) \leq \frac{2}{T} \cdot \mathbb{E}_\mu\left[\sum_{k \neq 1} t_k\right],$$
$$\mathbb{P}(e|H_\lambda) = \mathbb{P}_\lambda\left(t_1 \geq \frac{T}{2}\right) \leq \frac{2}{T} \cdot \mathbb{E}_\lambda[t_1],$$

where we have used Markov's inequality and the fact that $\sum_k t_k(T) = T$. Then,

since π is uniformly good, it follows that

$$\mathbb{E}_\mu \left[\sum_{k \neq 1} t_k \right] \in O\left(\log T\right),$$
$$\mathbb{E}_\lambda \left[t_1 \right] \in O\left(\log T\right),$$

since playing any sub-optimal arm $k \neq 1$ under μ incurs at least $\mu_1 - \sup_{k \neq 1} \mu_k > 0$ regret, and similarly, playing arm 1 under $\lambda\left(v\right)$ also incurs positive regret, as $v \in V$ implies $\max_{k \neq 1} u_k^T v > u_1^T v$, and thus $\sup_{k \neq 1} \lambda\left(v\right)_k > \lambda\left(v\right)_1$. Then, combined with the above bounds on the probabilities of error, we have that

$$T \cdot \left[\mathbb{P}\left(e | H_\mu\right) + \mathbb{P}\left(e | H_\lambda\right) \right] \in O\left(\log T\right). \tag{5.2}$$

Next, we use the Kailath bound [27],

$$\mathbb{P}\left(e | H_\mu\right) + \mathbb{P}\left(e | H_\lambda\right) \geq \frac{1}{2} \exp\left(- \min\left\{ D_{KL}\left(\mathcal{P} || \mathcal{Q}\right), D_{KL}\left(\mathcal{Q} || \mathcal{P}\right) \right\} \right),$$

and by combining (5.1) and (5.2), we have that

$$\begin{aligned}
\sum_k \mathbb{E}_\mu \left[t_k \right] \cdot D_{KL}\left(\mu_k || \lambda_k\right) &= D_{KL}\left(\mathcal{P} || \mathcal{Q}\right) \\
&\geq \min\left\{ D_{KL}\left(\mathcal{P} || \mathcal{Q}\right), D_{KL}\left(\mathcal{Q} || \mathcal{P}\right) \right\} \\
&\geq -\log 2 + \log T - \log\left[T \cdot \left(\mathbb{P}\left(e | H_\mu\right) + \mathbb{P}\left(e | H_\lambda\right) \right) \right] \\
&= \log T - O\left(\log \log T\right).
\end{aligned}$$

Finally, we relate the total regret to $\mathbb{E}_\mu \left[t_k \right]$, namely,

$$\begin{aligned}
R^\pi\left(T\right) &= \sum_k \mathbb{E}_\mu \left[t_k \right] \cdot \left[f\left(u_1^T z^* \right) - f\left(u_k^T z^* \right) \right] \\
&\geq \min_k \frac{\left[f\left(u_1^T z^* \right) - f\left(u_k^T z^* \right) \right]}{D_{KL}\left(\mu_k || \lambda_k\right)} \cdot \sum_k \mathbb{E}_\mu \left[t_k \right] \cdot D_{KL}\left(\mu_k || \lambda_k\right).
\end{aligned}$$

Combining the above, we have

$$\frac{R^\pi\left(T\right)}{\log T} \geq \min_k \frac{\mu_1 - \mu_k}{D_{KL}\left(\mu_k || \lambda_k\right)} - O\left(\frac{\log \log T}{\log T} \right),$$

and then letting $T \to \infty$ and $\epsilon \to 0$, by continuity of the KL divergence,

$$\liminf_{T\to\infty} \frac{R^{\pi}(T)}{\log T} \geq \min_{k} \frac{\mu_1 - \mu_k}{D_{KL}(\mu_k||\lambda_k)}.$$

Finally, since this holds for each $v \in V$, we can combine these individual bounds, and thus

$$\liminf_{T\to\infty} \frac{R^{\pi}(T)}{\log T} \geq \max_{v\in V} \left\{ \min_{k} \frac{\mu_1 - \mu_k}{D_{KL}(\mu_k||\lambda(v)_k)} \right\}.$$

□

5.4 Upper Bounds

In this section, we shall consider a variant of the previous model, but where the set of arms $U \subseteq \mathbb{R}^n$ is no longer finite, but instead uncountably infinite. In particular, we consider the unit sphere in n dimensions, with every point on the sphere being an arm. While this lacks the full generality of an arbitrary set of arms, this will make the analysis tractable. Since the expected rewards of the arms are coupled through an unknown parameter of dimension n, it is no longer necessary or even possible for each arm to be investigated in order to estimate the expected reward from that arm. Instead, we can estimate the underlying parameter; in this way, each pull can yield information about many arms. We present a simple algorithm, as well as bounds on the expected total regret as a function of time horizon when using this algorithm.

5.4.1 Two-Phase Algorithm

We first present an algorithmic description of a policy for the multi-armed bandit problem. This algorithm, which we call the Two-Phase Algorithm, will depend on a scheduling function $g : \mathbb{N}_1 \to \mathbb{N}_0$, such that g is strictly increasing. Since g is not surjective in general, its inverse g^{-1} is not defined over all of $\mathbb{N}_0$; however, we can extend the inverse image in the natural way to preserve monotonicity, by defining $g^{-1} : \mathbb{N}_0 \to \mathbb{N}_1$,

$$g^{-1}(t) = \max\{1 \cup \{l \in \mathbb{N}_1 : g(l) \leq t\}\}.$$

Earlier work on the analysis of this algorithm in the case of a finite number of arms, which showed an upper bound on the expected total regret that was independent of the number of arms m, can be found in [62]. In Theorem 5.4.2, we present an upper-bound to the expected total regret of this policy for the special case of a unit sphere of arms.

Algorithm 4 Two-Phase Algorithm

Require: Set of all arms U
Require: Set of n chosen arms $\Sigma = \{\Sigma_1, \ldots, \Sigma_n\} \subseteq U$, s.t. Σ has rank n
Require: Scheduling function $g : \mathbb{N}_1 \to \mathbb{N}_0$, strictly increasing
  $t \leftarrow 1, l \leftarrow 1$
  $q_u \leftarrow 0, \forall u \in \Sigma$
  loop
    for $u \in \Sigma$ **do**
      Pull arm $C_t \leftarrow u$, obtain reward X_t
      {Phase 1}
      $q_{C_t} \leftarrow q_{C_t} + 1_{\{X_t\}}$
      $t \leftarrow t + 1$
    end for
    Form the estimates $\hat{\alpha}_{u,l} \leftarrow \dfrac{q_u}{l},\ \forall u \in \Sigma$
    if $\hat{\alpha}_{u,l} \in (0,1),\ \forall u \in \Sigma$ **then**
      $$\hat{z}_l \leftarrow \left(\Sigma^T\right)^{-1} \begin{bmatrix} f^{-1}\left(\hat{\alpha}_{\Sigma_1,l}\right) \\ \vdots \\ f^{-1}\left(\hat{\alpha}_{\Sigma_n,l}\right) \end{bmatrix}$$
    else
      $\hat{z}_l \leftarrow \mathbf{0}_n$
    end if
    $C_{(l)} \leftarrow$ arbitrary choice from $\arg\max_{u \in U} \alpha_u(\hat{z}_l)$
    for $s \leftarrow 1$ **to** $g(l)$ **do**
      Pull arm $C_t \leftarrow C_{(l)}$, obtain reward X_t
      {Phase 2}
      $t \leftarrow t + 1$
    end for
    $l \leftarrow l + 1$
  end loop

The algorithm requires a selection of n arms,

$$\Sigma = \{\Sigma_1, \ldots, \Sigma_n\} \subseteq U, \text{ s.t. } \Sigma \text{ has rank } n.$$

Such a choice exists since we assume U has rank n. The algorithm proceeds in epochs; epoch l consists of n exploration pulls (called Phase 1), one for each arm

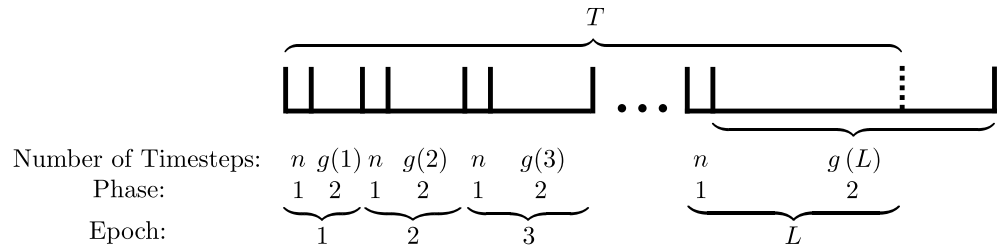

Figure 5.1: Given a time horizon T, we partition the T time-steps into Phase 1 and Phase 2 time-steps, grouped into a total of L epochs.

in Σ, and $g(l)$ exploitation pulls (called Phase 2). In other words, Phase 1 refines our estimate of z^*, and Phase 2 repeatedly pulls the best arm given our current estimate $\hat{z}_l$. If we impose a time horizon of T, epochs $1, 2, \ldots, L$ are appended until the time horizon T has been reached. The two phases are illustrated in Figure 5.1.

For each timestep t in Phase 1, an arm $k \in \Sigma$ is chosen, and the empirical count of successes q_k is incremented if $X_t = 1$. Prior to each Phase 2 timestep during epoch l, there have already been l Phase 1 pulls. We can then form empirical estimates for μ_k based on the Phase 1 time-steps, namely $\hat{\mu}_{k,l} = \dfrac{q_k}{l},\ \forall k \in \Sigma$. We define $\mathcal{G}_l = 1\left\{\hat{\mu}_{k,l} \in (0,1),\ \forall uk \in \Sigma\right\}$, where we call an epoch l a "good" epoch if $\mathcal{G}_l = 1$, and we can then form the current best estimate for z^*,

$$\hat{z}_l = \left(\Sigma^T\right)^{-1} \begin{bmatrix} f^{-1}\left(\hat{\mu}_{\Sigma_1,l}\right) \\ \vdots \\ f^{-1}\left(\hat{\mu}_{\Sigma_n,l}\right) \end{bmatrix},$$

since f being strictly increasing and continuous implies f^{-1} exists on $(0,1)$, and since Σ being an $n \times n$ matrix with full rank implies $\left(\Sigma^T\right)^{-1}$ exists. Otherwise, we call epoch l a bad epoch, and let $\hat{z}_l = \mathbf{0}_n$. Note that $\mathcal{G}_l = 1 \implies \mathcal{G}_{l+i} = 1,\ \forall i \geq 0$.

Then, choose an arm $C_{(l)} \in \arg\max_k f\left(u_k^T \hat{z}_l\right)$, settling ties arbitrarily, and pull this arm $g(l)$ times to form the current epoch's Phase 2.

In practice, LU decomposition, instead of matrix inversion, can be used to solve for $\hat{z}_l$. Also, since f is strictly increasing, the estimated best arm in a good epoch l can be computed as

$$C_{(l)} \in \arg\max_k f\left(u_k^T \hat{z}_l\right) = \arg\max_k \left(u_k^T \hat{z}_l\right).$$

We shall point out some of the features of this algorithm. Firstly, we were motivated by the need for computationally tractable algorithms at large scale, for

example in online advertising where m can on the order of thousands and T on the order of millions or more. We note that many of the current state-of-the-art algorithms require the solution to an optimization problem at every timestep; even when these are convex, the total number of optimizations needed is T and potentially intractable. In contrast, the algorithm proposed here requires only $g^{-1}(T)$ optimizations, which is sub-linear in T. Secondly, the algorithm is defined to run indefinitely; to obtain the total regret for any finite time horizon T, we simply terminate the algorithm when timestep T has been reached. This achieves the same outcome as an application of the doubling trick, in that the algorithm is not dependent on a time horizon T. Thirdly, in our exploration phases, the choice of arm exploits the correlation model that we have assumed in our problem, allowing us to remove the dependence on n. Finally, as we will see later, the lengths of the exploitation phases are chosen to grow in the epoch number. Thus, as we gain more information and are able to estimate z^* more accurately, we can spend a greater fraction of time-steps exploiting the arm we think is best; this is achieved by choosing a suitable scheduling function g to control the ratio of the number of exploitation (Phase 2) pulls versus exploration (Phase 1) pulls, as a function of the epoch number l. Our goal is to find an upper bound on $R(T)$, the expected total regret. In particular, we are interested in the asymptotic behavior of the upper bound as $T \to \infty$. The first theorem below can be found in [62], and is presented here as a comparison for the second result.

Theorem 5.4.1 *For any problem instance with finite arms, the two-phase algorithm with a given choice of scheduling function g s.t. $g(l) \in \exp(o(l))$, has that*

$$R(T) \leq O\left(n \cdot g^{-1}(T)\right).$$

The key idea behind this proof is similar to that behind the epoch-greedy policy, in that the probability of choosing a sub-optimal arm to exploit decreases exponentially with the epoch l. Instead of requiring knowledge of the gap between the best and second-best arms, we simply explore slightly more than required, thus avoiding a dependence on this constant.

Theorem 5.4.2 *For the problem of a continuum of arms, specifically the $(n-1)$-dimensional unit sphere, the two-phase algorithm, with the choice of scheduling*

function $g(l) = l$*, has that*

$$R^{\pi}(T) \in O\left(T^{\left(\frac{1}{2-\epsilon}\right)}\right), \textit{ for any } \epsilon > 0.$$

The main idea behind this proof is to consider what the expected regret in a single Phase 2 timestep is, as a function of the epoch l. With the choice of epoch length $g(l) = l^{1-\epsilon}$, we can show that the probability of choosing a "bad" arm decreases to 0, even when the set of good arms shrinks with the epoch l and converges asymptotically to the set consisting of just the best arm. Then, the expected regret is a combination of two terms: either playing an arm that is bad, which happens with low probability, or playing an arm which is still sub-optimal but whose regret can be bounded by an $o(1)$ function of the epoch l.

5.5 Conclusion

We have proposed a class of parametrized multi-armed bandit problems, in which the reward distribution is Bernoulli and independent across arms and across time, with a parameter that is a non-linear function of the scalar quality of an arm. The real-valued qualities are inner products between the unknown preference and known attribute vectors. Under this model, we are able to capture the fundamentally binary choice inherent in certain online machine learning problems.

Our proposed algorithm achieves an asymptotic expected total regret of $\tilde{O}\left(n\sqrt{T}\right)$ in the infinite arm, unit circle case. This is in contrast to the $\Omega\left(T^{\frac{n+1}{n+2}}\right)$ lower bound of [20]. The assumption of additional structure to the rewards, namely a logistic function of an inner product, instead of the more general Lipschitz condition, can be used to out-perform optimal algorithms which do not account for this structure. We conjecture that the lower bound on our problem is $\Omega(\sqrt{T})$ for the infinite arm case in general; if true, then this simple algorithm's performance is nearly optimal.

Finally, our algorithm can be implemented very efficiently as there are only $O(n)$ quantities tracked, and this does not scale with either the number of arms or the time-horizon. Also, since the exploration and exploitation phases are decoupled, the only history-dependent part of the algorithm, the optimization to determine which arm to pull, is only performed during $O(\sqrt{T})$ time-steps. In the infinite arm, unit circle case, this optimization itself is simply the normalization of

the current estimate $\hat{z}_l$. The basic idea of increasing the length of epochs is similar to that of UCB1, but because our algorithm uses a global count of the epoch instead of local counts for each arm, we are able to apply it to infinite-armed problems. Finally, we note that several extensions to this work are possible; multiple plays and time-dependent U and z^* would be directly applicable for e-commerce applications.

5.6 Proofs

In order to bound the regret during the exploitation phases, we will use the following idea: if the current estimate of z^*, namely $\hat{z}_l$, is "close enough" to z^*, then the regret per time-step will be bounded by some function of this distance. If $\hat{z}_l$ lies outside of this set, then the regret will be bounded by unity. In each successive epoch l, we reduce the size of this "close enough" set, while ensuring that the probability of lying within this set is close to 1. We bound the total regret during exploitation phases by a summation over epochs of this per-timestep regret, multiplied by the number of time-steps.

$$\begin{aligned} R_{exploit}(T) &\leq \sum_{l=1}^{L} \left[\left(1 \cdot 1_{\{\hat{z}_l \notin B_{z^*}(\Delta(l))\}} + r_{l,\Delta(l)} \cdot 1_{\{\hat{z}_l \in B_{z^*}(\Delta(l))\}} \right) \cdot g(l) \right] \\ &\leq \sum_{l=1}^{L} \left[\left(1_{\{\hat{z}_l \notin B_{z^*}(\Delta(l))\}} + r_{l,\Delta(l)} \right) \cdot g(l) \right], \end{aligned} \tag{5.3}$$

where $r_{l,\Delta(l)}$ is the maximum pseudo-regret incurred when $\hat{z}_l \in B_{z^*}(\Delta(l))$, a set which will be defined shortly. Finally we will choose these sequences $g(l)$ and $\Delta(l)$ to obtain the desired bound.

Lemma 5.6.1 *If* $\hat{z}_l \in B_{z^*}(\Delta(l))$ *and* $\Delta(l) \leq \sqrt{\frac{\lambda_{\min}\left(\Sigma^T\Sigma\right)}{n}} \cdot \|z^*\|_2$, *then* $r_{l,\Delta(l)} \leq k_3 \cdot [\Delta(l)]^2$, *where* $k_3 = \frac{\pi^2 n}{36 \cdot \|z^*\|_2 \cdot \lambda_{\min}(\Sigma^T\Sigma)}$.

Proof: Let θ_l be the central angle between z^* and $\hat{z}_l$. We can immediately bound the regret per time-step when playing the best arm given this estimate:

$$\begin{aligned} r_{l,\Delta(l)} &= f\left(u^{*T} z^*\right) - f\left(\hat{u}_l^T z^*\right) = f\left(\|z^*\|_2\right) - f\left(\|z^*\|_2 \cdot \cos\theta_l\right) \\ &\leq \frac{\|z^*\|_2}{4} \cdot (1 - \cos\theta_l) \leq \frac{\|z^*\|_2}{8} \cdot \theta_l^2. \end{aligned}$$

Next, we define the region of arms that are "close enough" at epoch l, $B_{z^*}(\Delta) = \left\{z \in \mathbb{R}^n : \left\|\Sigma^T z - \Sigma^T z^*\right\|_2 < \sqrt{n} \cdot \Delta\right\}$, and thus

$$\begin{aligned}\hat{z}_l \in B_{z^*}(\Delta(l)) &\iff \left\|\Sigma^T \hat{z}_l - \Sigma^T z^*\right\|_2 < \sqrt{n} \cdot \Delta(l) \\ &\implies \|\hat{z}_l - z^*\|_2 \leq \sqrt{\frac{n}{\lambda_{\min}\left(\Sigma^T \Sigma\right)}} \cdot \Delta(l).\end{aligned}$$

We want to relate this back to the central angle θ_l, as we know how to compute the regret from there. Now, if we suppose $\Delta(l) \leq \sqrt{\frac{\lambda_{\min}\left(\Sigma^T\Sigma\right)}{n}} \cdot \|z^*\|_2$, then by the previous implication it follows that $\|\hat{z}_l - z^*\|_2 \leq \|z^*\|_2$. Now consider θ_l, and by geometry,

$$\begin{aligned}\theta_l &\leq \sin^{-1}\left(\frac{\|\hat{z}_l - z^*\|_2}{\|z^*\|_2}\right) \leq \frac{\pi}{2} \cdot \frac{\|\hat{z}_l - z^*\|_2}{\|z^*\|_2} \\ &\leq \frac{\pi}{2\|z^*\|_2} \cdot \sqrt{\frac{n}{\lambda_{\min}\left(\Sigma^T\Sigma\right)}} \cdot \Delta(l).\end{aligned}$$

Thus, $r_{l,\Delta(l)} \leq k_3 \cdot [\Delta(l)]^2$, where $k_3 = \dfrac{\pi^2 n}{36 \cdot \|z^*\|_2 \cdot \lambda_{\min}\left(\Sigma^T\Sigma\right)}$. □

Proof of Theorem 5.4.2

Proof: Let $0 < \epsilon < 1$, and choose the sequences $\Delta(l) = k_4 \cdot l^{-\frac{1}{2}(1-\epsilon)}$ and $g(l) = l^{(1-\epsilon)}$, where $k_4 = \sqrt{\frac{\lambda_{\min}\left(\Sigma^T\Sigma\right)}{n}} \cdot \|z^*\|_2$. That is, the ratio of exploitation time-steps to exploration time-steps in epoch l grows roughly linearly, and the set of "good" estimates $\hat{z}_l$ shrinks roughly as $\dfrac{1}{\sqrt{l}}$. Define

$$\gamma_\Delta = \min_{k \in \Sigma} \min\left\{D\left(f\left(u_k^T z^* - \Delta\right)||\mu_k\right), D\left(f\left(u_k^T z^* + \Delta\right)||\mu_k\right)\right\},$$

which through the Chernoff bound, characterizes the probability of choosing an arm outside a "good" set. Then, by an intermediate result of [62], we have that

$$\begin{aligned}\gamma_\Delta &\geq 2\Delta^2 \cdot \exp\left(-\frac{\left(\|z^*\|_2 + \Delta\right)^2}{2}\right) \\ &\geq 2\Delta^2 \cdot \exp\left(-\frac{\left(\|z^*\|_2 + k_4\right)^2}{2}\right).\end{aligned}$$

Thus, we have that $\mathbb{P}\left(\hat{z}_l \notin B_{z^*}\left(\Delta\left(l\right)\right)\right) \cdot g\left(l\right) k_5\left(l\right)$ and $r_{l,\Delta(l)} \cdot g\left(l\right) \leq k_3 \cdot k_4^2 = \frac{\pi^2 \cdot \left\|z^*\right\|_2}{36}$, where

$$k_5\left(l\right) = 2n \cdot l^{(1-\epsilon)} \left[f\left(\left\|z^*\right\|_2\right)^l + \exp\left(-2k_4^2 l^{\epsilon} \cdot \exp\left(-\frac{\left(\left\|z^*\right\|_2 + k_4\right)^2}{2} \right) \right) \right]$$

since $k_1 = -\log\left[\max_{k\in\Sigma} \max\left\{\mu_k, 1-\mu_k\right\}\right] \geq -\log\left[f\left(\max_{k\in\Sigma}\left\|u_k\right\|_2 \cdot \left\|z^*\right\|_2\right)\right]$. Using (5.3),

$$\begin{aligned} R_{exploit}\left(T\right) \leq & \sum_{l=1}^{L} \left[\left(\mathbb{P}\left(\hat{z}_l \notin B_{z^*}\left(\Delta\left(l\right)\right)\right) + r_{l,\Delta(l)} \right) \cdot g\left(l\right) \right] \\ \leq & k_6 + \left(1 + \frac{\pi^2 \cdot \left\|z^*\right\|_2}{36} \right) \cdot L \end{aligned}$$

where $k_6 = \sum_{l=1}^{L'} k_5\left(l\right)$ and $L' = \max\left\{l : k_5\left(l\right) > 1\right\}$. In short, there is some finite number of epochs L' where we have not explored enough for the incremental regret bound to be small enough. For all $l > L'$, we can bound the total exploitation regret by a constant. Finally, since, $T \geq \sum_{l=1}^{L-1}\left\{n + g(l)\right\} \geq \frac{L^{2-\epsilon}}{2}$, we have that

$$\begin{aligned} R\left(T\right) \leq & k_6 + 2 \cdot \left(n + 1 + \frac{\pi^2 \cdot \left\|z^*\right\|_2}{36} \right) \cdot T^{\left(\frac{1}{2-\epsilon}\right)} \\ \in & O\left(nT^{\left(\frac{1}{2-\epsilon}\right)} \right). \end{aligned}$$

□

The choice of scheduling function g can be made βl, where β changes the trade-off between the constant term $O\left(\beta \cdot \sum_{l=1}^{L^*-1} g\left(l\right)\right)$ and the time-dependent term $O\left(\sqrt{\frac{T}{\beta}}\right)$. That is, the asymptotics can be improved at the expense of finite-time performance. Furthermore, if the time-horizon is known in advance, then the scheduling function can be chosen to minimize the sum of these two terms.

CHAPTER 6

FUTURE WORK

In this chapter we mention some possible directions of future work.

Lower Bounds

Our work in Chapter 2 is still in need of a corresponding lower-bound, although there are difficulties due to the interaction between regret and overdraft not present in other multi-armed bandit problems. We have some preliminary results for this problem, but more work is needed. Nonetheless, it may be possible to show that with a strict constraint on the overdraft, a lower-bound on the regret exists. More generally, different policies can lead to different tradeoffs between regret and overdraft, and characterizing what tradeoffs are achievable may also be worthwhile.

Mean Field Equilibrium

In Chapter 2, we have considered the problem of bidding in a repeated game against a field of bidders, and considered the highest bid amongst them to be i.i.d. samples from some fixed distribution. This mean field assumption is useful, but assumes we are the only active player in the system. One extension, then, could be to start with some fixed number of players, although each with differing budgets and valuations, and have each apply stochastic approximation to try and learn their optimal bid. The starting point could be a proof for the existence and uniqueness of a Nash equilibrium in the single stage game, and then perhaps under some conditions (such as valuations not being perfectly correlated between any two players), the under-bidding factors of each player can be shown to converge to the equilibrium values. This would show the mean field assumption is validated and leads to a mean field equilibrium, if all players follow the stochastic

approximation policy, even though individual bids are no longer independent over time.

A further extension would be to allow players to arrive and depart the system, for example with arrivals marked by a Poisson process, and i.i.d. exponential lifetimes, which would generalize the above extension with no arrivals or departures. Despite the lack of a steady-state in the sense that no player stays in the system, it may still be possible to characterize the convergence rate of the under-bidding factor in this dynamic environment, again leading to a regret bound.

Furthermore, if under certain conditions, such an equilibrium is unique, then we have a mechanism that can suggest to players a stochastic approximation policy, and due to the limited information and computational abilities of each player, they are best off following said policy; when all players do this, then all players converge to their optimal under-bidding factor.

Hybrid Learning Systems

For many optimization problems on big data, both online and offline learning algorithms exist, but have tradeoffs in complexity and convergence properties. Online algorithms often yield approximate solutions with larger error, but can allow one to take advantage of incremental data. For example, to run a SVM with quantile loss and incorporate billions of historical datapoints, one might use an offline algorithm that assumes data is generated i.i.d. [63, 64]. When the dataset increases by another million datapoints over one day, the entire regression is re-run. This assumption of stationarity and the slowness to update may cause unmodeled non-stationarities, and in particular, fast time-scale transients, to affect the performance of the algorithm in practice.

Consider the following heuristic in which offline learning is still periodically re-run, but in between those slow updates to the optimization parameters, online learning takes over and fine-tunes the parameters based on the incoming data. This combination of offline and online learning, which we refer to as a hybrid learning system, might have provable performance guarantees that are superior to each individually. Namely, the hybrid system could handle a wider class of models that can include non-stationarities and a fast response to parameter changes, when compared to offline learning, and offer better guaranteed convergence behavior, when compared to online learning.

REFERENCES

[1] A. Mahajan and D. Teneketzis, "Multi-armed bandit problems," in *Foundations and Applications of Sensor Management*, A. O. Hero, D. A. Castañón, D. Cochran, and K. Kastella, Eds. Springer-Verlag, 2007, ch. 6, pp. 121–151.

[2] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *CoRR*, vol. abs/1204.5721, 2012.

[3] J. C. Gittins and D. M. Jones, "A dynamic allocation index for the sequential design of experiments," *Progress in Statistics*, vol. 1, pp. 241–266, 1974.

[4] R. Weber, "On the Gittins index for multiarmed bandits," *The Annals of Applied Probability*, vol. 2, no. 4, pp. 1024–1033, Nov. 1992.

[5] J. N. Tsitsiklis, "A short proof of the Gittins index theorem," *The Annals of Applied Probability*, vol. 4, no. 1, pp. 194–199, Feb. 1994.

[6] P. Whittle, "Restless bandits: Activity allocation in a changing world," *Journal of Applied Probability*, vol. 25, pp. 287–298, 1988.

[7] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, pp. 4–22, 1985.

[8] R. Agrawal, D. Teneketzis, and V. Anantharam, "Asymptotically efficient adaptive allocation schemes for controlled i.i.d. processes: Finite parameter space," *IEEE Transactions on Automatic Control*, vol. 34, no. 3, pp. 258–267, Mar. 1989.

[9] R. Agrawal, D. Teneketzis, and V. Anantharam, "Asymptotically efficient adaptive allocation schemes for controlled Markov chains: Finite parameter space," *IEEE Transactions on Automatic Control*, vol. 34, no. 12, pp. 1249–1259, Dec. 1989.

[10] R. Agrawal, M. Hegde, and D. Teneketzis, "The multi-armed bandit problem with switching cost," in *26th IEEE Conference on Decision and Control, 1987*, vol. 26, Dec. 1987, pp. 1106–1108.

[11] R. Agrawal, "Sample mean based index policies with $o(\log n)$ regret for the multi-armed bandit problem," *Advances in Applied Probability*, vol. 27, no. 4, pp. 1054–1078, Dec. 1995.

[12] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-Part I: I.I.D. rewards," *IEEE Transactions on Automatic Control*, vol. 32, no. 11, pp. 968–976, Nov. 1987.

[13] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-Part II: Markovian rewards," *IEEE Transactions on Automatic Control*, vol. 32, no. 11, pp. 977–982, Nov. 1987.

[14] N. Abe, A. W. Biermann, and P. M. Long, "Reinforcement learning with immediate rewards and linear hypotheses," *Algorithmica*, vol. 37, no. 4, pp. 263–293, 2003.

[15] P. Auer, "Using confidence bounds for exploitation-exploration trade-offs," *Journal of Machine Learning Research*, vol. 3, pp. 397–422, Mar. 2003.

[16] J.-Y. Audibert and S. Bubeck, "Minimax policies for adversarial and stochastic bandits," 2009.

[17] A. J. Mersereau, P. Rusmevichientong, and J. N. Tsitsiklis, "A structured multiarmed bandit problem and the greedy policy," *IEEE Transactions on Automatic Control*, vol. 54, no. 12, pp. 2787–2802, Dec. 2009.

[18] P. Rusmevichientong and J. N. Tsitsiklis, "Linearly parameterized bandits," *Mathematics of Operations Research*, vol. 35, no. 2, pp. 395–411, May 2010.

[19] V. Dani, T. P. Hayes, and S. M. Kakade, "Stochastic linear optimization under bandit feedback," in *Proc. of the 21st Annual Conference on Learning Theory*, Helsinki, Finland, July 2008, pp. 363–374.

[20] R. Kleinberg, A. Slivkins, and E. Upfal, "Multi-armed bandits in metric spaces," in *Proc. of the 40th annual ACM symposium on Theory of computing*, Victoria, British Columbia, Canada, 2008, pp. 681–690.

[21] S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvári, "Online optimization in x-armed bandits," vol. 21, pp. 201–208, 2009.

[22] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM Journal on Computing*, vol. 32, no. 1, pp. 48–77, 2002.

[23] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2, pp. 235–256, 2002.

[24] A. Garivier and O. Cappé, "The kl-ucb algorithm for bounded stochastic bandits and beyond," *Journal of Machine Learning Research - Proceedings Track*, vol. 19, pp. 359–376, 2011.

[25] G. Stoltz, "Incomplete information and internal regret in prediction of individual sequences," Ph.D. dissertation, University of Paris-Sud, Nov. 2005. [Online]. Available: http://eprints.pascal-network.org/archive/00001692/

[26] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. New York, NY: Cambridge University Press, 2006.

[27] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE Trans. Communications*, vol. 15, no. 1, pp. 52 – 60, February 1967.

[28] S. S. Chandramouli, "Multi armed bandit problem: some insights," accessed: 2013-09-26. [Online]. Available: http://www.columbia.edu/sc3102/bandit.pdf

[29] J. Broder and P. Rusmevichientong, "Dynamic pricing under a general parametric choice model," *Operations Research*, vol. 60, no. 4, p. 965?980, 2012.

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

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

[32] K. Iyer, R. Johari, and M. Sundararajan, "Mean field equilibria of dynamic auctions with learning," in *Proceedings of the 12th ACM Conference on Electronic Commerce*, ser. EC '11, 2011, longer version available at http://ssrn.com/abstract=1799085. [Online]. Available: http://doi.acm.org/10.1145/1993574.1993631 pp. 339–340.

[33] R. Gummadi, P. B. Key, and A. Proutiere, "Optimal bidding strategies in dynamic auctions with budget constraints," in *Proceedings of the 49th Annual Allerton Conference on Communication, Control, and Computing*, 2011, longer version available at http://ssrn.com/abstract=2066175. pp. 588–588.

[34] S. Balseiro, O. Besbes, and G. Weintraub, "Auctions for online display advertising exchanges: Approximations and design," *Columbia Business School Research Paper*, 2012.

[35] H. Robbins and S. Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics*, pp. 400–407, 1951.

[36] H. Kushner, "Stochastic approximation: a survey," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 1, pp. 87–96, 2010. [Online]. Available: http://dx.doi.org/10.1002/wics.57

[37] V. Borkar, *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008. [Online]. Available: http://books.google.com/books?id=_wIKYYieLbUC

[38] B. Tan and R. Srikant, "Online advertisement, optimization and stochastic networks," *IEEE Transactions on Automatic Control*, vol. 57, no. 11, pp. 2854–2868, 2012.

[39] S. Lahaie and D. M. Pennock, "Revenue analysis of a family of ranking rules for keyword auctions," in *Proceedings of the 8th ACM Conference on Electronic Commerce*. ACM, 2007, pp. 50–56.

[40] D. R. M. Thompson and K. Leyton-Brown, "Revenue optimization in the generalized second-price auction," in *Proceedings of the 14th ACM Conference on Electronic Commerce*, June 2013, to appear.

[41] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3-4, pp. 285–294, 1933.

[42] H. Robbins, "Some aspects of the sequential design of experiments," *Bulletin of the American Mathematical Society*, vol. 58, no. 5, pp. 527–535, 1952.

[43] T. Lai, "Adaptive treatment allocation and the multi-armed bandit problem," *The Annals of Statistics*, vol. 15, no. 3, pp. 1091–1114, 09 1987.

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

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

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

[47] 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.

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

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

[50] N. Cesa-Bianchi and G. Lugosi, "Combinatorial bandits," *J. Comput. Syst. Sci.*, vol. 78, no. 5, pp. 1404–1422, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.jcss.2012.01.001

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

[52] J. Yu and S. Mannor, "Unimodal bandits," in *Proc. of ICML*, 2011.

[53] E. W. Cope, "Regret and convergence bounds for a class of continuum-armed bandit problems," *Automatic Control, IEEE Transactions on*, vol. 54, no. 6, pp. 1243–1253, 2009.

[54] C. Jiang and R. Srikant, "Bandits with budgets," in *Proc. of the 52nd IEEE Conference on Decision and Control*, 2013.

[55] A. Badanidiyuru, R. Kleinberg, and A. Slivkins, "Bandits with knapsacks," in *Proc. of FOCS*, 2013.

[56] A. B. Tsybakov, *Introduction to Non-Parametric Estimation*. Springer, 2008.

[57] "Kdd cup challenge," https://www.kddcup2012.org/c/kddcup2012-track2.

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

[59] 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.

[60] 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.

[61] S. Filippi, O. Cappé, A. Garivier, and C. Szepesvári, "Parametric bandits: The generalized linear case," *Advances in Neural Information Processing Systems*, vol. 23, pp. 1–9, 2010.

[62] C. Jiang and R. Srikant, "Parametrized stochastic multi-armed bandits with binary rewards," in *Proc. of the American Control Conference*, 2011, pp. 119–124.

[63] S. Boyd, C. Cortes, M. Mohri, and A. Radovanovic, "Accuracy at the top," in *Advances in Neural Information Processing Systems*, 2012, pp. 962–970.

[64] S. Boyd, C. Cortes, C. Jiang, M. Mohri, A. Radovanovic, and J. Skaf, "Large-scale distributed optimization for improving accuracy at the top," 2012.