Unimodal Bandits: Regret Lower Bounds and Optimal Algorithms

Richard Combes RCOMBES@KTH.SE
KTH, Royal Institute of technology, Stockholm, Sweden

Alexandre Proutiere ALEPRO@KTH.SE
KTH, Royal Institute of technology, Stockholm, Sweden

Abstract

We consider stochastic multi-armed bandits where the expected reward is a unimodal function over partially ordered arms. This important class of problems has been recently investigated in (Cope, 2009; Yu & Mannor, 2011). The set of arms is either discrete, in which case arms correspond to the vertices of a finite graph whose structure represents similarity in rewards, or continuous, in which case arms belong to a bounded interval. For discrete unimodal bandits, we derive asymptotic lower bounds for the regret achieved under any algorithm, and propose OSUB, an algorithm whose regret matches this lower bound. Our algorithm optimally exploits the unimodal structure of the problem, and surprisingly, its asymptotic regret does not depend on the number of arms. We also provide a regret upper bound for OSUB in non-stationary environments where the expected rewards smoothly evolve over time. The analytical results are supported by numerical experiments showing that OSUB performs significantly better than the state-of-the-art algorithms. For continuous sets of arms, we provide a brief discussion. We show that combining an appropriate discretization of the set of arms with the UCB algorithm yields an order-optimal regret, and in practice, outperforms recently proposed algorithms designed to exploit the unimodal structure.

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

1. Introduction

Stochastic Multi-Armed Bandits (MAB) (Robbins, 1952; Gittins, 1989) constitute the most fundamental sequential decision problems with an exploration vs. exploitation trade-off. In such problems, the decision maker selects an arm in each round, and observes a realization of the corresponding unknown reward distribution. Each decision is based on past decisions and observed rewards. The objective is to maximize the expected cumulative reward over some time horizon by balancing exploitation (arms with higher observed rewards should be selected often) and exploration (all arms should be explored to learn their average rewards). Equivalently, the performance of a decision rule or algorithm can be measured through its expected *regret*, defined as the gap between the expected reward achieved by the algorithm and that achieved by an oracle algorithm always selecting the best arm. MAB problems have found many fields of application, including sequential clinical trials, communication systems, economics, see e.g. (Cesa-Bianchi & Lugosi, 2006; Bubeck & Cesa-Bianchi, 2012).

In their seminal paper (Lai & Robbins, 1985), Lai and Robbins solve MAB problems where the successive rewards of a given arm are i.i.d., and where the expected rewards of the various arms are not related. They derive an asymptotic (when the time horizon grows large) lower bound of the regret satisfied by any algorithm, and present an algorithm whose regret matches this lower bound. This initial algorithm was quite involved, and many researchers have tried to devise simpler and yet efficient algorithms. The most popular of these algorithms are UCB (Auer et al., 2002) and its extensions, e.g. KL-UCB (Garivier & Cappé, 2011; Cappé et al., 2013) (note that KL-UCB algorithm was initially proposed in (Lai, 1987), see (2.6)). When the expected rewards of the various arms are not related (Lai & Robbins, 1985), the regret of the best algorithm is essentially of the order $O(K\log(T))$ where K denotes the number of arms, and T is the time horizon. When

K is very large or even infinite, MAB problems become more challenging. Fortunately, in such scenarios, the expected rewards often exhibit some structural properties that the decision maker can exploit to design efficient algorithms. Various structures have been investigated in the literature, e.g., Lipschitz (Agrawal, 1995; Kleinberg et al., 2008; Bubeck et al., 2008), linear (Dani et al., 2008), convex (Flaxman et al., 2005).

We consider bandit problems where the expected reward is a unimodal function over partially ordered arms as in (Yu & Mannor, 2011). The set of arms is either discrete, in which case arms correspond to the vertices of a finite graph whose structure represents similarity in rewards, or continuous, in which case arms belong to a bounded interval. This unimodal structure occurs naturally in many practical decision problems, such as sequential pricing (Yu & Mannor, 2011) and bidding in online sponsored search auctions (B., 2005).

Our contributions. We mainly investigate unimodal bandits with finite sets of arms, and are primarily interested in cases where the time horizon T is much larger than the number of arms K.

(a) For these problems, we derive an asymptotic regret lower bound satisfied by any algorithm. This lower bound does not depend on the structure of the graph, nor on its size: it actually corresponds to the regret lower bound in a classical bandit problem (Lai & Robbins, 1985), where the set of arms is just a neighborhood of the best arm in the graph.

(b) We propose OSUB (Optimal Sampling for Unimodal Bandits), a simple algorithm whose regret matches our lower bound, i.e., it optimally exploits the unimodal structure. The asymptotic regret of OSUB does not depend on the number of arms. This contrasts with LSE (Line Search Elimination), the algorithm proposed in (Yu & Mannor, 2011) whose regret scales as $O(\gamma D \log(T))$ where γ is the maximum degree of vertices in the graph and D is its diameter. We present a finite-time analysis of OSUB, and derive a regret upper bound that scales as $O(\gamma \log(T)+K)$. Hence OSUB offers better performance guarantees than LSE as soon as the time horizon satisfies $T \geq \exp(K/\gamma D)$. Although this is not explicitly mentioned in (Yu & Mannor, 2011), we believe that LSE was meant to address bandits where the number of arms is not negligible compared to the time horizon.

(c) We further investigate OSUB performance in non-stationary environments where the expected rewards smoothly evolve over time but keep their unimodal structure.

(d) We conduct numerical experiments and show that OSUB significantly outperforms LSE and other classical bandit algorithms when the number of arms is much smaller than the time horizon.

(e) Finally, we briefly discuss systems with a continuous set of arms. We show that using a simple discretization of the set of arms, UCB-like algorithms are order-optimal, and actually outperform more advanced algorithms such as those proposed in (Yu & Mannor, 2011). This result suggests that in discrete unimodal bandits with a very large number of arms, it is wise to first prune the set of arms, so as to reduce its size to a number of the order of $\sqrt{T}/\log(T)$.

2. Related work

Unimodal bandits have received relatively little attention in the literature. They are specific instances of bandits in metric spaces (Kleinberg, 2004; Kleinberg et al., 2008; Bubeck et al., 2008). In this paper, we add unimodality and show how this structure can be optimally exploited. Unimodal bandits have been specifically addressed in (Cope, 2009; Yu & Mannor, 2011). In (Cope, 2009), bandits with a continuous set of arms are studied, and the author shows that the Kiefer-Wolfowitz stochastic approximation algorithm achieves a regret of the order of $O(\sqrt{T})$ under some strong regularity assumptions on the reward function. In (Yu & Mannor, 2011), for the same problem, the authors present LSE, an algorithm whose regret scales as $O(\sqrt{T}\log(T))$ without the need for a strong regularity assumption. The LSE algorithm is based on Kiefer's golden section search algorithm. It iteratively eliminates subsets of arms based on PAC-bounds derived after appropriate sampling. By design, under LSE, the sequence of parameters used for the PAC bounds is pre-defined, and in particular does not depend of the observed rewards. As a consequence, LSE may explore too much sub-optimal parts of the set of arms. For bandits with a continuum set of arms, we actually show that combining an appropriate discretization of the decision space (i.e., reducing the number of arms to $\sqrt{T}/\log(T)$ arms) and the UCB algorithm can outperform LSE in practice (this is due to the adaptive nature of UCB). Note that the parameters used in LSE to get a regret of the order $O(\sqrt{T}\log(T))$ depend on the time horizon T.

In (Yu & Mannor, 2011), the authors also present an extension of the LSE algorithm to problems with discrete sets of arms, and provide regret upper bounds of this algorithm. These bounds depends on the structure of the graph defining unimodal structure, and on the number of arms as mentioned previously. LSE performs better than classical bandit algorithms only when the number of arms is very large, and actually becomes comparable to the time horizon. Here we are interested in bandits with relatively small number of arms.

Non-stationary bandits have been studied in

(Hartland et al., 2007; Garivier & Moulines, 2008; Slivkins & Upfal, 2008; Yu & Mannor, 2011). Except for (Slivkins & Upfal, 2008), these papers deal with environments where the expected rewards and the best arm change abruptly. This ensures that arms are always well separated, and in turn, simplifies the analysis. In (Slivkins & Upfal, 2008), the expected rewards evolve according to independent brownian motions. We consider a different, but more general class of dynamic environments: here the rewards smoothly evolve over time. The challenge for such environments stems from the fact that, at some time instants, arms can have expected rewards arbitrarily close to each other.

Finally, we should mention that bandit problems with structural properties such as those we address here can often be seen as specific instances of problems in the control of Markov chains, see (Graves & Lai, 1997). We leverage this observation to derive regret lower bounds. However, algorithms developed for the control of generic Markov chains are often too complex to implement in practice. Our algorithm, OSUB, is optimal and straightforward to implement.

3. Model and Objectives

We consider a stochastic multi-armed bandit problem with $K \geq 2$ arms. We discuss problems where the set of arms is continuous in Section 6. Time proceeds in rounds indexed by $n = 1, 2, \ldots$. Let $X_k(n)$ be the reward obtained at time n if arm k is selected. For any k, the sequence of rewards $(X_k(n))_{n\geq 1}$ is i.i.d. with distribution and expectation denoted by ν_k and μ_k respectively. Rewards are independent across arms. Let $\mu = (\mu_1, \ldots, \mu_K)$ represent the expected rewards of the various arms. At each round, a decision rule or algorithm selects an arm depending on the arms chosen in earlier rounds and their observed rewards. We denote by $k^\pi(n)$ the arm selected under π in round n. The set Π of all possible decision rules consists of policies π satisfying: for any $n \geq 1$, if $\mathcal{F}_n^\pi$ is the σ-algebra generated by $(k^\pi(t), X_{k^\pi(t)}(t))_{1\leq t\leq n}$, then $k^\pi(n+1)$ is $\mathcal{F}_n^\pi$-measurable.

3.1. Unimodal Structure

The expected rewards exhibit a *unimodal* structure, similar to that considered in (Yu & Mannor, 2011). More precisely, there exists an undirected graph $G = (V, E)$ whose vertices correspond to arms, i.e., $V = \{1, \ldots, K\}$, and whose edges characterize a partial order (initially unknown to the decision maker) among expected rewards. We assume that there exists a unique arm $k^\star$ with maximum expected reward $\mu^\star$, and that from any sub-optimal arm $k \neq k^\star$, there exists a path $p = (k_1 = k, \ldots, k_m = k^\star)$ of length m (depending on k) such that for all $i = 1, \ldots, m-1$, $(k_i, k_{i+1}) \in E$ and $\mu_{k_i} < \mu_{k_{i+1}}$. We denote by $\mathcal{U}_G$ the set of vectors μ satisfying this unimodal structure.

This notion of unimodality is quite general, and includes, as a special case, classical unimodality (where G is just a line). Note that we assume that the decision maker knows the graph G, but ignores the best arm, and hence the partial order induced by the edges of G.

3.2. Stationary and non-stationary environments

The model presented above concerns stationary environments, where the expected rewards for the various arms do not evolve over time. In this paper, we also consider non-stationary environments where these expected rewards could evolve over time according to some deterministic dynamics. In such scenarios, we denote by $\mu_k(n)$ the expected reward of arm k at time n, i.e., $\mathbb{E}[X_k(n)] = \mu_k(n)$, and $(X_k(n))_{n\geq 1}$ constitutes a sequence of independent random variables with evolving mean. In non-stationary environments, the sequences of rewards are still assumed to be independent across arms. Moreover, at any time n, $\mu(n) = (\mu_1(n), \ldots \mu_K(n))$ is unimodal with respect to some fixed graph G, i.e., $\mu(n) \in \mathcal{U}_G$ (note however that the partial order satisfied by the expected rewards may evolve over time).

3.3. Regrets

The performance of an algorithm $\pi \in \Pi$ is characterized by its *regret* up to time T (where T is typically large). The way regret is defined differs depending on the type of environment.

Stationary Environments. In such environments, the regret $R^\pi(T)$ of algorithm $\pi \in \Pi$ is simply defined through the number of times $t_k^\pi(T) = \sum_{1\leq n\leq T} \mathbf{1}\{k^\pi(n) = k\}$ that arm k has been selected up to time T: $R^\pi(T) = \sum_{k=1}^K (\mu^\star - \mu_k)\mathbb{E}[t_k^\pi(T)]$. Our objectives are (1) to identify an asymptotic (when $T \to \infty$) regret lower bound satisfied by *any* algorithm in Π, and (2) to devise an algorithm that achieves this lower bound.

Non-stationary Environments. In such environments, the regret of an algorithm $\pi \in \Pi$ quantifies how well π tracks the best arm over time. Let $k^\star(n)$ denote the optimal arm with expected reward $\mu^\star(n)$ at time n. The regret of π up to time T is hence defined as: $R^\pi(T) = \sum_{n=1}^T \left(\mu^\star(n) - \mathbb{E}[\mu_{k^\pi(n)}(n)]\right)$.

4. Stationary environments

In this section, we consider unimodal bandit problems in stationary environments. We derive an asymptotic lower bound of regret when the reward distributions belong to a parametrized family of distributions, and propose OSUB, an algorithm whose regret matches this lower bound.

4.1. Lower bound on regret

To simplify the presentation, we assume here that the reward distributions belong to a parametrized family of distributions. More precisely, we define a set of distributions $\mathcal{V} = \{\nu(\theta)\}_{\theta\in[0,1]}$ parametrized by $\theta \in [0,1]$. The expectation of $\nu(\theta)$ is denoted by $\mu(\theta)$ for any $\theta \in [0,1]$. $\nu(\theta)$ is absolutely continuous with respect to some positive measure m on $\mathbb{R}$, and we denote by $p(x,\theta)$ its density. The Kullback-Leibler (KL) divergence number between $\nu(\theta)$ and $\nu(\theta')$ is: $KL(\theta,\theta') = \int_{\mathbb{R}} \log(p(x,\theta)/p(x,\theta'))p(x,\theta)m(dx)$. We denote by $\theta^\star$ a parameter (it might not be unique) such that $\mu(\theta^\star) = \mu^\star$, and we define the minimal divergence number between $\nu(\theta)$ and $\nu(\theta^\star)$ as: $I_{\min}(\theta,\theta^\star) = \inf_{\theta\in[0,1]:\mu(\theta')\geq\mu^\star} KL(\theta,\theta')$.

Finally, we say that arm k has parameter θ_k if $\nu_k = \nu(\theta_k)$, and we denote by Θ_G the set of all parameters $\theta = (\theta_1,\ldots,\theta_K) \in [0,1]^K$ such that the corresponding expected rewards are unimodal with respect to graph G: $\mu = (\mu_1,\ldots,\mu_K) \in \mathcal{U}_G$. Of particular interest is the family of Bernoulli distributions: the support of m is $\{0,1\}$, $\mu(\theta) = \theta$, and $I_{\min}(\theta,\theta^\star) = I(\theta,\theta^\star)$ where $I(\theta,\theta^\star) = \theta\log(\frac{\theta}{\theta^\star}) + (1-\theta)\log(\frac{1-\theta}{1-\theta^\star})$ is KL divergence number between Bernoulli distributions of respective means θ and $\theta^\star$.

We are now ready to derive an asymptotic regret lower in parametrized unimodal bandit problems as defined above. Without loss of generality, we restrict our attention to so-called uniformly good algorithms, as defined in (Lai & Robbins, 1985) (uniformly good algorithms exist as shown later on). We say that $\pi \in \Pi$ is uniformly good if for all $\theta \in \Theta_G$, we have that $R^\pi(T) = o(T^a)$ for all $a > 0$.

Theorem 4.1 *Let $\pi \in \Pi$ be a uniformly good algorithm, and assume that $\nu_k = \nu(\theta_k) \in \mathcal{V}$ for all k. Then for any $\theta \in \Theta_G$,*

$$\liminf_{T\to+\infty} \frac{R^\pi(T)}{\log(T)} \geq c(\theta) = \sum_{(k,k^*)\in E} \frac{\mu^\star - \mu_k}{I_{\min}(\theta_k,\theta^\star)}. \quad (1)$$

The above theorem is a consequence of results in optimal control of Markov chains (Graves & Lai, 1997). All proofs are presented in (Combes & Proutiere, 2014). As in classical discrete bandit problems, the regret scales at least logarithmically with time (the regret lower bound derived in (Lai & Robbins, 1985) is obtained from Theorem 4.1 assuming that G is the complete graph). We also observe that the unimodal structure, if optimally exploited, can bring significant performance improvements: the regret lower bound does not depend on the size K of the decision space. Indeed $c(\theta)$ includes only terms corresponding to arms that are neighbors in G of the optimal arm (as if one could learn without regret that all other arms are sub-optimal).

In the case of Bernoulli rewards, the lower regret bound becomes $\log(T)\sum_{(k,k^*)\in E} \frac{\mu^\star-\mu_k}{I(\theta_k,\theta^\star)}$. Note that LSE and GLSE, the algorithms proposed in (Yu & Mannor, 2011), have performance guarantees that do not match our lower bound: when G is a line, LSE achieves a regret bounded by $41/\Delta^2\log(T)$, whereas in the general case, GLSE incurs a regret of the order of $O(\gamma D\log(T))$ where γ is the maximal degree of vertices in G, and D is its diameter. The performance of LSE critically depends on the graph structure, and the number of arms. Hence there is an important gap between the performance of existing algorithms and the lower bound derived in Theorem 4.1. In the next section, we close this gap and propose an asymptotically optimal algorithm.

4.2. The OSUB Algorithm

We now describe OSUB, a simple algorithm whose regret matches the lower bound derived in Theorem of 4.1 for Bernoulli rewards, i.e., OSUB is asymptotically optimal. The algorithm is based on KL-UCB proposed in (Lai, 1987; Cappé et al., 2013), and uses KL-divergence upper confidence bounds to define an *index* for each arm. OSUB can be readily extended to systems where reward distributions are within one-parameter exponential families by simply modifying the definition of arm indices as done in (Cappé et al., 2013). In OSUB, each arm is attached an index that resembles the KL-UCB index, but the arm selected at a given time is the arm with maximal index within the neighborhood in G of the arm that yielded the highest empirical reward. Note that since the sequential choices of arms are restricted to some neighborhoods in the graph, OSUB is not an index policy. To formally describe OSUB, we need the following notation. For $p \in [0,1]$, $s \in \mathbb{N}$, and $n \in \mathbb{N}$, we define:

$$F(p,s,n) = \sup\{q \geq p : sI(p,q) \leq \log(n) + c\log(\log(n))\}, \quad (2)$$

with the convention that $F(p,0,n) = 1$, and $F(1,s,n) = 1$, and where $c > 0$ is a constant. Let $k(n)$ be the arm selected under OSUB at time n, and let $t_k(n)$ denote the number of times arm k has been selected up to time n. The empirical reward of arm k at time n is $\hat{\mu}_k(n) = \frac{1}{t_k(n)}\sum_{t=1}^n \mathbf{1}\{k(t)=k\}X_k(t)$, if $t_k(n) > 0$ and $\hat{\mu}_k(n) = 0$ otherwise. We denote by $L(n) = \arg\max_{1\leq k\leq K} \hat{\mu}_k(n)$ the index of the arm with the highest empirical reward (ties are broken arbitrarily). Arm $L(n)$ is referred to as the *leader* at time n. Further define $l_k(n) = \sum_{t=1}^n \mathbf{1}\{L(t) = k\}$ the number of times arm k has been the leader up to time n. Now the index of arm k at time n is defined as:

$$b_k(n) = F(\hat{\mu}_k(n), t_k(n), l_k(L(n))).$$

Finally for any k, let $N(k) = \{k' : (k',k) \in E\} \cup \{k\}$ be the neighborhood of k in G. The pseudo-code of OSUB is

presented below.

Algorithm OSUB

Input: graph $G = (V, E)$
For $n \geq 1$, select the arm $k(n)$ where:

$$k(n) = \begin{cases} L(n) & \text{if } \frac{l_{L(n)}(n)-1}{\gamma+1} \in \mathbb{N}, \\ \arg\max_{k \in N(L(n))} b_k(n) & \text{otherwise,} \end{cases}$$

where γ is the maximal degree of nodes in G and ties are broken arbitrarily.

Note that OSUB forces us to select the current leader often: $L(n)$ is chosen when $l_{L(n)}(n) - 1$ is a multiple of $\gamma + 1$. This ensures that the number of times an arm has been selected is at least proportional to the number of times this arm has been the leader. This property significantly simplifies the regret analysis, but it could be removed.

4.3. Finite-time analysis of OSUB

Next we provide a finite time analysis of the regret achieved under OSUB. Let Δ denote the minimal separation between an arm and its best adjacent arm: $\Delta = \min_{1 \leq k \leq K} \max_{k':(k,k') \in E} \mu_{k'} - \mu_k$. Note that Δ is not known a priori.

Theorem 4.2 *Assume that the rewards lie in [0,1] (i.e., the support of ν_k is included in $[0, 1]$, for all k), and that $(\mu_1, \ldots, \mu_K) \in \mathcal{U}_G$. The number of times suboptimal arm k is selected under OSUB satisfies: for all $\epsilon > 0$ and all $T \geq 3$,*

$$\mathbb{E}[t_k(T)] \leq \begin{cases} (1+\epsilon)\frac{\log(T)+c\log(\log(T))}{I(\mu_k,\mu^\star)} & \text{if } (k,k^\star) \in E, \\ \quad +C_1 \log\log(T) + \frac{C_2}{T^{\beta(\epsilon)}} & \\ \frac{C_3}{\Delta^2} & \text{otherwise,} \end{cases}$$

where $\beta(\epsilon) > 0$, and $0 < C_1 < 7$, $C_2 > 0$, $C_3 > 0$ are constants.

To prove this upper bound, we analyze the regret accumulated (i) when the best arm $k^\star$ is the leader, and (ii) when the leader is arm $k \neq k^\star$. (i) When $k^\star$ is the leader, the algorithm behaves like KL-UCB restricted to the arms around $k^\star$, and the regret at these rounds can be analyzed as in (Cappé et al., 2013). (ii) Bounding the number of rounds where $k \neq k^\star$ is not the leader is more involved. To do this, we decompose this set of rounds into further subsets (such as the time instants where k is the leader and its mean is not well estimated), and control their expected cardinalities using concentration inequalities. Along the way, we establish Lemma 4.3, a new concentration inequality of independent interest.

Lemma 4.3 *Let $\{Z_t\}_{t \in \mathbb{Z}}$ be a sequence of independent random variables with values in $[0, B]$. Define $\mathcal{F}_n$ the σ-algebra generated by $\{Z_t\}_{t \leq n}$ and the filtration $\mathcal{F} = (\mathcal{F}_n)_{n \in \mathbb{Z}}$. Consider $s \in \mathbb{N}$, $n_0 \in \mathbb{Z}$ and $T \geq n_0$. We define $S_n = \sum_{t=n_0}^{n} B_t(Z_t - \mathbb{E}[Z_t])$, where $B_t \in \{0, 1\}$ is a $\mathcal{F}_{t-1}$-measurable random variable. Further define $t_n = \sum_{t=n_0}^{n} B_t$. Define $\phi \in \{n_0, \ldots, T+1\}$ a $\mathcal{F}$-stopping time such that either $t_\phi \geq s$ or $\phi = T + 1$.*

Then we have that: $\mathbb{P}[S_\phi \geq t_\phi \delta \ , \ \phi \leq T] \leq \exp(-2s\delta^2 B^{-2})$. As a consequence: $\mathbb{P}[|S_\phi| \geq t_\phi \delta \ , \ \phi \leq T] \leq 2\exp(-2s\delta^2 B^{-2})$.

Lemma 4.3 concerns the sum of products of i.i.d. random variables and of a previsible sequence, evaluated at a stopping time (for the natural filtration). We believe that concentration results for such sums can be instrumental in bandit problems, where typically, we need information about the empirical rewards at some specific random time epochs (that often are stopping times). Refer to (Combes & Proutiere, 2014) for a proof. A direct consequence of Theorem 4.2 is the asymptotic optimality of OSUB in the case of Bernoulli rewards:

Corollary 4.4 *Assume that rewards distributions are Bernoulli (i.e for any k, $\nu_k \sim$ Bernoulli(θ_k)), and that $\theta \in \Theta_G$. Then the regret achieved under π=OSUB satisfies: $\limsup_{T \to +\infty} R^\pi(T)/\log(T) \leq c(\theta)$.*

5. Non-stationary environments

We now consider time-varying environments. We assume that the expected reward of each arm varies smoothly over time, i.e., it is Lipschitz continuous: for all $n, n' \geq 1$ and $1 \leq k \leq K$: $|\mu_k(n) - \mu_k(n')| \leq \sigma|n - n'|$.

We further assume that the unimodal structure is preserved (with respect to the same graph G): for all $n \geq 1$, $\mu(n) \in \mathcal{U}_G$. Considering smoothly varying rewards is more challenging than scenarios where the environment is abruptly changing. The difficulty stems from the fact that the rewards of two or more arms may become arbitrarily close to each other (this happens each time the optimal arm changes), and in such situations, regret is difficult to control. To get a chance to design an algorithm that efficiently tracks the best arm, we need to make some assumption to limit the proportion of time when the separation of arms becomes too small. Define for $T \in \mathbb{N}$, and $\Delta > 0$:

$$H(\Delta, T) = \sum_{n=1}^{T} \sum_{(k,k') \in E} \mathbf{1}\{|\mu_k(n) - \mu_{k'}(n)| < \Delta\}.$$

Assumption 1 *There exists a function Φ and Δ_0 such that for all $\Delta < \Delta_0$: $\limsup_{T \to +\infty} H(\Delta, T)/T \leq \Phi(K)\Delta$.*

5.1. OSUB with a Sliding Window

To cope with the changing environment, we modify the OSUB algorithm, so that decisions are based on past choices and observations over a time-window of fixed duration equal to $\tau+1$ rounds. The idea of adding a sliding window to algorithms initially designed for stationary environments is not novel (Garivier & Moulines, 2008); but here, the unimodal structure and the smooth evolution of rewards make the regret analysis more challenging.

Define: $t_k^\tau(n) = \sum_{t=n-\tau}^n \mathbf{1}\{k(t)=k\}$; $\hat{\mu}_k^\tau(n) = (1/t_k^\tau(n))\sum_{t=n-\tau}^n \mathbf{1}\{k(t)=k\}X_k(t)$ if $t_k^\tau(n)>0$ and $\hat{\mu}_k^\tau(n)=0$ otherwise; $L^\tau(n) = \arg\max_{1\le k\le K}\hat{\mu}_k^\tau(n)$; $l_k^\tau(n) = \sum_{t=n-\tau}^n \mathbf{1}\{L^\tau(t)=k\}$. The index of arm k at time n then becomes: $b_k^\tau(n) = F(\hat{\mu}_k^\tau(n), t_k^\tau(n), l_k^\tau(L^\tau(n)))$. The pseudo-code of SW-OSUB is presented below.

Algorithm SW-OSUB

Input: graph $G=(V,E)$, window size $\tau+1$
For $n\ge 1$, select the arm $k(n)$ where:

$$k(n) = \begin{cases} L^\tau(n) & \text{if } \frac{l^\tau_{L^\tau(n)}(n)-1}{\gamma+1}\in\mathbb{N}, \\ \arg\max_{k\in N(L^\tau(n))} b_k^\tau(n) & \text{otherwise.}\end{cases}$$

5.2. Regret Analysis

In non-stationary environments, achieving sublinear regrets is often not possible. In (Garivier & Moulines, 2008), the environment is subject to abrupt changes or breakpoints. It is shown that if the density of breakpoints is strictly positive, which typically holds in practice, then the regret of any algorithm has to scale linearly with time. We are interested in similar scenarios, and consider smoothly varying environments where the number of times the optimal arm changes has a positive density. The next theorem provides an upper bound of the regret per unit of time achieved under SW-OSUB. This bound holds for any non-stationary environment with σ-Lipschitz rewards.

Theorem 5.1 *Let Δ: $2\tau\sigma<\Delta<\Delta_0$. Assume that for any $n\ge 1$, $\mu(n)\in\mathcal{U}_G$ and $\mu^\star(n)\in[a,1-a]$ for some $a>0$. Further suppose that $\mu_k(\cdot)$ is σ-Lipschitz for any k. The regret per unit time under π =SW-OSUB with a sliding window of size $\tau+1$ satisfies: if $a>\sigma\tau$, then for any $T\ge 1$,*

$$\frac{R^\pi(T)}{T} \le \frac{H(\Delta,T)}{T}(1+\Delta) + \frac{C_1 K\log(\tau)}{\tau(\Delta-4\tau\sigma)^2} + \gamma\left(1+g_0^{-1/2}\right)\frac{\log(\tau)+c\log(\log(\tau))+C_2}{2\tau(\Delta-2\tau\sigma)^2},$$

where C_1, C_2 are positive constants and $g_0=(a-\sigma\tau)(1-a+\sigma\tau)/2$.

Corollary 5.2 *Assume that for any $n\ge 1$, $\mu(n)\in\mathcal{U}_G$ and $\mu^\star(n)\in[a,1-a]$ for some $a>0$, and that $\mu_k(\cdot)$ is σ-Lipschitz for any k. Set $\tau=\sigma^{-3/4}\log(1/\sigma)/8$. The regret per unit of time of π =SW-OSUB with window size $\tau+1$ satisfies:*

$$\limsup_{T\to\infty}\frac{R^\pi(T)}{T} \le C\Phi(K)\sigma^{\frac{1}{4}}\log\left(\frac{1}{\sigma}\right)(1+Kj(\sigma)),$$

for some constant $C>0$, and some function j such that $\lim_{\sigma\to 0^+} j(\sigma)=0$.

These results state that the regret per unit of time achieved under SW-OSUB decreases and actually vanishes when the speed at which expected rewards evolve decreases to 0. Also observe that the dependence of this regret bound in the number of arms is typically mild (in many practical scenarios, $\Phi(K)$ may actually not depend on K).

The proof of Theorem 5.1 relies on the same types of arguments as those used in stationary environments. To establish the regret upper bound, we need to evaluate the performance of the KL-UCB algorithm in non-stationary environments (the result and the corresponding analysis are presented in (Combes & Proutiere, 2014)).

6. Continuous Set of Arms

In this section, we briefly discuss the case where the decision space is continuous. The set of arms is $[0,1]$, and the expected reward function $\mu:[0,1]\to\mathbb{R}$ is assumed to be Lipschitz continuous, and unimodal: there exists $x^\star\in[0,1]$ such that $\mu(x')\ge\mu(x)$ if $x'\in[x,x^\star]$ or $x'\in[x^\star,x]$. Let $\mu^\star=\mu(x^\star)$ denote the highest expected reward. A decision rule selects at any round $n\ge 1$ an arm x and observes the corresponding reward $X(x,n)$. For any $x\in[0,1]$, $(X(x,n))_{n\ge1}$ is an i.i.d. sequence. We make the following additional assumption on function μ.

Assumption 2 *There exists $\delta_0>0$ such that (i) for all x,y in $[x^\star,x^\star+\delta_0]$ (or in $[x^\star-\delta_0,x^\star]$), $C_1|x-y|^\alpha\le|\mu(x)-\mu(y)|$; (ii) for $\delta\le\delta_0$, if $|x-x^*|\le\delta$, then $|\mu(x^*)-\mu(x)|\le C_2\delta^\alpha$.*

This assumption is more general than that used in (Yu & Mannor, 2011). In particular it holds for functions with a *plateau* and a *peak*: $\mu(x)=\max(1-|x-x^\star|/\epsilon,0)$. Now as for the case of a discrete set of arms, we denote by Π the set of possible decision rules, and the regret achieved under rule $\pi\in\Pi$ up to time T is: $R^\pi(T)=T\mu^\star-\sum_{n=1}^T\mathbb{E}[\mu(x^\pi(n))]$, where $x^\pi(n)$ is the arm selected under π at time n.

There is no known precise asymptotic lower bound for continuous bandits. However, we know that for our problem, the regret must be at least of the order of $O(\sqrt{T})$ up to

logarithmic factor. In (Yu & Mannor, 2011), the authors show that the LSE algorithm achieves a regret scaling as $O(\sqrt{T}\log(T))$, under more restrictive assumptions. We show that combining discretization and the UCB algorithm as initially proposed in (Kleinberg, 2004) yields lower regrets than LSE in practice (see Section 7), and is order-optimal, i.e., the regret grows as $O(\sqrt{T}\log(T))$.

For $\delta > 0$, we define a discrete bandit problem with $K = \lceil 1/\delta \rceil$ arms, and where the rewards of k-th arm are distributed as $X((k-1)/\delta, n)$. The expected reward of the k-th arm is $\mu_k = \mu((k-1)/\delta)$. Let π be an algorithm running on this discrete bandit problem. The regret of π for the initial continuous bandit problem is at time T: $R^\pi(T) = T\mu^\star - \sum_{k=1}^{\lceil 1/\delta \rceil} \mu_k \mathbb{E}[t_k^\pi(T)]$. We denote by UCB($\delta$) the UCB algorithm (Auer et al., 2002) applied to the discretized bandit. In the following proposition, we show that when $\delta = (\log(T)/\sqrt{T})^{1/\alpha}$, UCB($\delta$) is order-optimal. In practice, one may not know the time horizon T in advance. In this case, using the "doubling trick" (see e.g. (Cesa-Bianchi & Lugosi, 2006)) would incur an additional logarithmic multiplicative factor in the regret.

Proposition 1 *Consider a unimodal bandit on $[0,1]$ with rewards in $[0,1]$ and satisfying Assumption 2. Set $\delta = (\log(T)/\sqrt{T})^{1/\alpha}$. The regret under UCB($\delta$) satisfies:*

$$\limsup_{T\to\infty} \frac{R^\pi(T)}{\sqrt{T}\log(T)} \leq C_2 3^\alpha + 16/C_1.$$

7. Numerical experiments

7.1. Discrete bandits

We compare the performance of our algorithm to that of KL-UCB (Cappé et al., 2013), LSE (Yu & Mannor, 2011), UCB (Auer et al., 2002), and UCB-U. The latter algorithm is obtained by applying UCB restricted to the arms which are adjacent to the current leader as in OSUB. We add the prefix "SW" to refer to Sliding Window versions of these algorithms.

Stationary environments. In our first experiment, we consider $K = 17$ arms with Bernoulli rewards of respective averages $\mu = (0.1, 0.2, ..., 0.9, 0.8, \ldots, 0.1)$. The rewards are unimodal (the graph G is simply a line). The regret achieved under the various algorithms is presented in Figure 1 and Table 1. The parameters in LSE algorithm are chosen as suggested in Proposition 4.5 (Yu & Mannor, 2011). Regrets are calculated averaging over 50 independent runs. OSUB significantly outperforms all other algorithms. The regret achieved under LSE is not presented in Figure 1, because it is typically much larger than that of other algorithms. This poor performance can be explained by the non-adaptive nature of LSE, as already discussed earlier. LSE can beat UCB when the number of arms is not negligible compared to the time horizon (e.g. in Figure 4 in (Yu & Mannor, 2011), $K = 250.000$ and the time horizon is less than $3K$): in such scenarios, UCB-like algorithms perform poorly because of their initialization phase (all arms have to be tested once).

T	1000	10000	100000
UCB	30.1	35.1	39
KL-UCB	18.8	21.4	23
UCB-U	8.5	11.7	13.9
OSUB	5.8	5.9	6
LSE	36.3	271.5	999.1

Table 1. $R^\pi(T)/\log(T)$ for different algorithms – 17 arms.

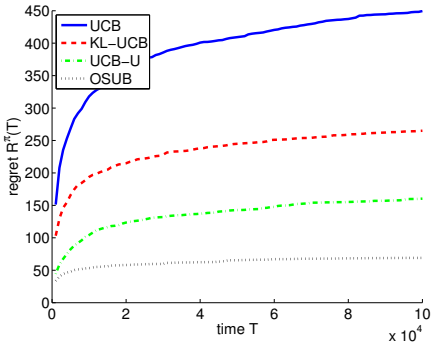

Figure 1. Regret vs. time in stationary environments – $K = 17$ arms.

In Figure 2, the number of arms is 129, and the expected rewards form a triangular shape as in the previous example, with minimum and maximum equal to 0.1 and 0.9, respectively. Similar observations as in the case of 17 arms can be made. We deliberately restrict the plot to small time horizons: this corresponds to scenarios where LSE can perform well.

Non-stationary environments. We now investigate the per-

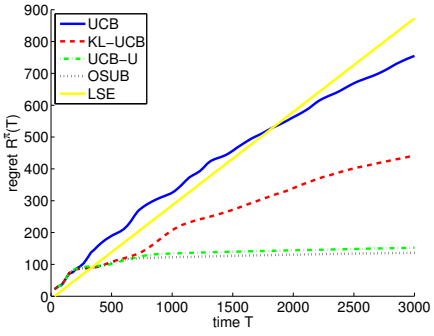

Figure 2. Regret vs. time in stationary environments – $K = 129$ arms.

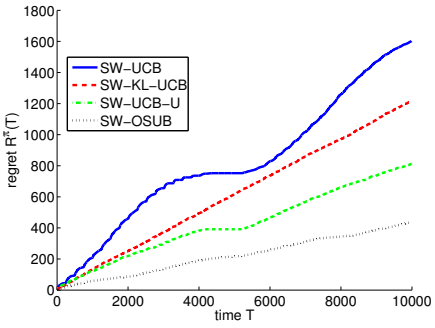

Figure 3. Regret vs. time in a slowly varying environment – $K = 10$ arms, $\sigma = 10^{-3}$.

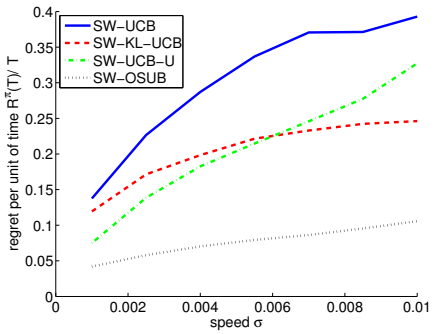

Figure 4. Regret per unit of time $R^\pi(T)/T$ vs. speed σ – $K = 10$ arms.

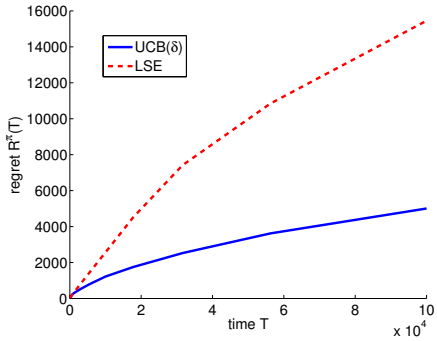

Figure 5. Regret vs. time for a continuous set of arms.

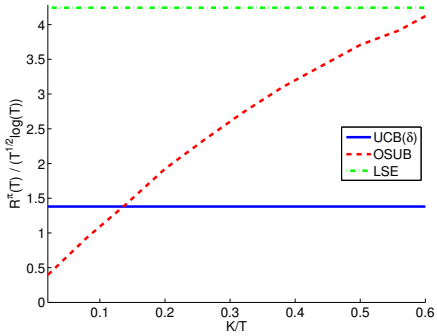

Figure 6. Normalized regret vs. K/T, $T = 5.10^4$ for a continuous set of arms.

formance of SW-OSUB in a slowly varying environment. There are $K = 10$ arms whose expected rewards form a moving triangle: for $k = 1, \dots, K$, $\mu_k(n) = (K-1)/K - |w(n)-k|/K$, where $w(n) = 1+(K-1)(1+\sin(n\sigma))/2$. Figure 3 presents the regret as a function of time under various algorithms when the speed at which the environment evolves is $\sigma = 10^{-3}$. The window size are set as follows for the various algorithms: $\tau = \sigma^{-4/5}$ for SW-UCB and SW-KL-UCB (the rationale for this choice is explained in (Combes & Proutiere, 2014)), $\tau = \sigma^{-3/4}\log(1/\sigma)/8$ for SW-UCB-U and OSUB. In Figure 4, we show how the speed σ impacts the regret per time unit. SW-OSUB provides the most efficient way of tracking the optimal arm.

7.2. Continuous bandits

In Figure 5, we compare the performance of the LSE and UCB(δ) algorithms when the set of arms is continuous. The expected rewards form a triangle: $\mu(x) = 1/2 - |x - 1/2|$ so that $\mu^\star = 1/2$ and $x^\star = 1/2$. The parameters used in LSE are those given in (Yu & Mannor, 2011), whereas the discretization parameter δ in UCB(δ) is set to $\delta = \log(T)/\sqrt{T}$. UCB($\delta$) significantly outperforms LSE at any time: an appropriate discretization of continuous bandit problems might actually be more efficient than other methods based on ideas taken from classical optimization theory.

Figure 6 compares the regret of the discrete version of LSE (with optimized parameters), and of OSUB as the number of arms K grows large, $T = 50,000$. The average rewards of arms are extracted from the triangle used in the continuous bandit, and we also provide the regret achieved under UCB(δ). OSUB outperforms UCB(δ) even if the number of arms gets as large as 7500! OSUB also beats LSE unless the number of arms gets bigger than $0.6 \times T$.

8. Conclusion

In this paper, we address stochastic bandit problems with a unimodal structure, and a finite set of arms. We provide asymptotic regret lower bounds for these problems and design an algorithm that asymptotically achieves the lowest regret possible. Hence our algorithm optimally exploits the unimodal structure of the problem. Our preliminary analysis of the continuous version of this bandit problem suggests that when the number of arms become very large and comparable to the time horizon, it might be wiser to prune the set of arms before actually running any algorithm.

References

Agrawal, R. The continuum-armed bandit problem. *SIAM J. Control and Optimization*, 33(6):1926–1951, November 1995.

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

B., Edelman. Strategic bidder behavior in sponsored search auctions. In *Proc. of Workshop on Sponsored Search Auctions, ACM Electronic Commerce*, pp. 192–198, 2005.

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

Bubeck, S., Munos, R., Stoltz, G., and Szepesvári, C. Online optimization in x-armed bandits. In *Advances in Neural Information Processing Systems 22*, 2008.

Cappé, O., Garivier, A., Maillard, O., Munos, R., and Stoltz, G. Kullback-leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41(3): 516–541, June 2013.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Combes, R. and Proutiere, A. Unimodal bandits: Regret lower bounds and optimal algorithms. Technical Report, people.kth.se/˜alepro/pdf/tr-icml2014.pdf, 2014.

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

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. In *Proc. of Conference On Learning Theory (COLT)*, pp. 355–366, 2008.

Flaxman, A., Kalai, A. T., and McMahan, H. B. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proc. of ACM/SIAM symposium on Discrete Algorithms (SODA)*, pp. 385–394, 2005.

Garivier, A. and Cappé, O. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proc. of Conference On Learning Theory (COLT)*, 2011.

Garivier, A. and Moulines, E. On upper-confidence bound policies for non-stationary bandit problems. In *Proc. of Algorithmic Learning Theory (ALT)*, 2008.

Gittins, J.C. *Bandit Processes and Dynamic Allocation Indices*. John Wiley, 1989.

Graves, T. L. and Lai, T. L. Asymptotically efficient adaptive choice of control laws in controlled markov chains. *SIAM J. Control and Optimization*, 35(3):715–743, 1997.

Hartland, C., Baskiotis, N., Gelly, S., Teytaud, O., and Sebag, M. Change point detection and meta-bandits for online learning in dynamic environments. In *Proc. of conférence francophone sur l'apprentissage automatique (CAp07)*, 2007.

Kleinberg, R., Slivkins, A., and Upfal, E. Multi-armed bandits in metric spaces. In *Proc. of the 40th annual ACM Symposium on Theory of Computing (STOC)*, pp. 681–690, 2008.

Kleinberg, R. D. Nearly tight bounds for the continuum-armed bandit problem. In *Proc. of the conference on Neural Information Processing Systems (NIPS)*, 2004.

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

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

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

Slivkins, A. and Upfal, E. Adapting to a changing environment: the brownian restless bandits. In *Proc. of Conference On Learning Theory (COLT)*, pp. 343–354, 2008.

Yu, J. and Mannor, S. Unimodal bandits. In *Proc. of International Conference on Machine Learning (ICML)*, pp. 41–48, 2011.