Unimodal Bandits

Jia Yuan Yu JIAYUAN.YU@ENS.FR
Ecole Normale Supérieure, HEC Paris, CNRS, France.

Shie Mannor SHIE@EE.TECHNION.AC.IL
Department of Electrical Engineering, Technion, Haifa, Israel.

Abstract

We consider multiarmed bandit problems where the expected reward is unimodal over partially ordered arms. In particular, the arms may belong to a continuous interval or correspond to vertices in a graph, where the graph structure represents similarity in rewards. The unimodality assumption has an important advantage: we can determine if a given arm is optimal by sampling the possible directions around it. This property allows us to quickly and efficiently find the optimal arm and detect abrupt changes in the reward distributions. For the case of bandits on graphs, we incur a regret proportional to the maximal degree and the diameter of the graph, instead of the total number of vertices.

1. Introduction

Unimodal reward functions occur naturally in various decision problems, *e.g.*, single-peak preferences economics and voting theory (Mas-Collel et al., 1995). We consider the unimodality property in an uncertain setting: that of the stochastic multiarmed bandit. This setting is composed of stochastic sources—or arms—with unknown reward distributions, but unimodal expected value with respect to some partial order. Our goal is to use this property to quickly find an arm with the highest expected value. We do this in both the one-dimensional setting—with a continuum of arms—and the graphical setting—where each arm corresponds to a vertex in a graph. Such graphical decision problems are central to the study of social and communication networks.

Our first contribution is an algorithm whose regret is proportional to the maximum degree of the graph and its diameter, whereas the number of arms can be exponential in the diameter. In the one-dimensional setting, although existing methods already achieve the optimal expected regret, we present an algorithm that does so more efficiently thanks to the unimodality assumption, and also prove a new probably approximately correct (PAC) regret bound.

An additional consequence of unimodality is the ability to efficiently detect abrupt changes in the arms' reward distributions. Our second contribution is a method that simultaneously detects change-points and minimizes the regret. This method occasionally samples one arm in each direction around the optimal arm so as to balance the regret due to choosing suboptimal arms and the regret due to delays in detecting changes.

This paper is organized as follows. In Section 2, we situate our work with respect to related works. We present our model with motivating examples and assumptions in Section 3. In Section 4, we present an intermediate result in the one-dimensional setting, *i.e.*, a sampling scheme that finds an approximately optimal arm with high probability. Sections 5 and 6 contain our main results on graphical bandits and change-points. Section 5 considers a bandit problem where the arms have a graphical structure and the expected reward is an unimodal function on the graph. We define in Section 6 the notion of change-points in the reward distributions and present an algorithm that simultaneously minimizes regret and detects changes. We conclude by discussing open problems in Section 7.

2. Related Works

Multiarmed bandit problems are central to machine learning and adaptive control due to numerous applications. With a finite set of arms, efficient index-based solutions exist, whether the rewards are stochastic or adversarial (Auer et al., 2002a;b). The case of a continuum of arms is also of great interest due to

Appearing in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

applications, such as the design of auction mechanisms (Blum et al., 2003) and routing (Bansal et al., 2003). The one-dimensional case was first studied in (Agrawal, 1995) with a Hölder condition. Under similar conditions, Kleinberg (2004) presents an algorithm based on discretization that achieves a regret of $O(T^{2/3})$, which is optimal up to a logarithmic factor. A regret of the order of $O(\sqrt{T}\log T)$ is shown under additional assumptions in (Auer et al., 2007). The common assumption in bandit problems with an infinite number of arms is a dependence between the rewards of nearby arms. This notion of dependence also helps in the case of bandits with finitely many arms, *e.g.*, (Pandey et al., 2007).

Our model in Section 4 is a special case of the model of (Kleinberg, 2004), with the additional assumption that the expected reward is unimodal over an interval $[0,1]$ of arms. This is similar to the assumptions of (Cope, 2009), with the notable difference that we do not require the expected reward function to be three-times differentiable. Cope (2009) shows that for a multidimensional unimodal expected reward function, the Kiefer-Wolfowitz stochastic approximation algorithm achieves a regret of the order of $O(\sqrt{T})$. It is however well-known that Kiefer-Wolfowitz type algorithms require suitable differentiability assumptions and that finite-time convergence results are generally unavailable for these algorithms (Agrawal, 1995). Correspondingly, the regret guarantees of (Cope, 2009) are asymptotic and without explicit constants. In Section 4, we give a new method that achieves a regret of the order of $O(\sqrt{T}\log T)$ in finite time under the unimodality assumption—this regret bound is also tight up to a logarithmic factor. In contrast to (Cope, 2009), this method does not require convexity or differentiability. This method is based on one-dimensional line search combined with appropriate sampling and is therefore particularly efficient. Our algorithm iteratively eliminates subsets of arms; this approach is reminiscent of algorithms such as the successive elimination algorithm for the classical bandit problem (Even-Dar et al., 2002).

Bandit problems on tree graphs are special cases of more general bandit problems in topological spaces (Kleinberg et al., 2008; Bubeck et al., 2008). In this paper, we motivate and specifically study bandit problems with a graphical structure. Our algorithm gives some insight on the dependence of the regret on the characteristics of a graph. A related problem is online learning on graphs where the rewards form a nonstochastic (adversarial) and fully observed individual sequence (Cesa-Bianchi et al., 2009a;b). In our work, the rewards are stochastic and only partially observed. Our unimodality assumption for bandits on graphs is similar to labeled graphs with bitonic paths in the graph theory literature (cf. (Müller-Hannemann & Weihe, 2001; Spinrad, 2003)). However, our expected rewards—corresponding to the labels—are unknown and observed through stochastic samples.

Bandit problems with abrupt changes—at unknown time instants—in the reward distributions is a generalization of two classical models. The stochastic bandit corresponds to case without changes, whereas the adversarial bandit corresponds to the case of changes at every time instant. This generalization gives the nonstationary bandit problem of (Hartland et al., 2006; Garivier & Moulines, 2008; Yu & Mannor, 2009). We adopt a similar notion of change-points in Section 6, but do not require additional assumptions such as side observations (Yu & Mannor, 2009) or knowledge of the frequency of changes (Garivier & Moulines, 2008). Our solution approach is also completely different, relying on sparse and efficient sampling.

3. Unimodality: Examples and Assumptions

First, let us motivate the assumption of unimodal expected rewards with a bandit problem where the set of arms is the interval $[0,1]$ of the real line. Consider a sequential pricing problem with the goal of maximizing the total revenue from the sale of a sequence of identical items. We may think of the arms as the possible prices for the item. At each time instant t, the agent chooses a price x_t from an interval $[0,1]$. The reward of the arm $x \in [0,1]$ at time t is a random variable $r_t(x) = x\ w_{t,x}$. For a fixed x, the sequence $w_{1,x}, w_{2,x}, \ldots$ is a sequence of i.i.d. Bernoulli random variables. Each Bernoulli random variable $w_{t,x}$ corresponds to whether an item is sold at time t. Hence, for a fixed x, the sequence $r_1(x), r_2(x), \ldots$ is also an i.i.d. random sequence. The expected reward of a fixed arm x at every time t is

$$\bar{r}(x) \triangleq \mathbb{E}[x\ w_{t,x}] = x \Pr(w_{t,x} = 1), \quad x \in [0,1],$$

which is independent of t. The function $\bar{r}$ is the *expected reward function* over the set of arms $[0,1]$.

For a fixed time t, we do not know the probability $\Pr(w_{t,x} = 1)$ of sale of the item (*i.e.*, success) for each price x. However, we assume that this success probability is a monotone decreasing function of the price. Hence, if we have two arms x and y such that $x \leq y$, then $\Pr(w_{t,y} = 1) \leq \Pr(w_{t,x} = 1)$. This leads to the following unimodality property on the expected reward function $\bar{r}$, which is the main assumption in Section 4.

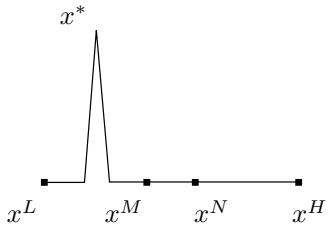


Figure 1. Cases to avoid: sharp peak and flat plateau.

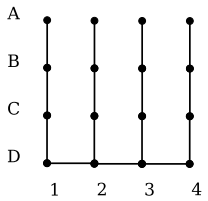


Figure 2. Graphical bandit with two features: quality (D, C, B, A) and price (1, 2, 3, 4).

Assumption 3.1 (Unimodality)**.** The expected reward function $\bar{r}$ is unimodal, *i.e.*, there exists an *optimal arm* $x^* \in [0,1]$ such that $\bar{r}$ is monotonically increasing in the interval $[0, x^*]$, and monotonically decreasing in the interval $[x^*, 1]$.

The function $\bar{r}$ in the preceding example also satisfies the following sufficient condition for unimodality.

Remark 1 (Sufficient condition). Suppose that the function f is differentiable and monotone decreasing over $[0,1]$, $f(0) > 0$, and that $xf'(x)$ is strictly decreasing. Then the function $g(x) = xf(x)$ is unimodal.

We also need the following assumption to avoid expected reward functions such as the one depicted in Figure 1. First, we need an upper bound on the rate of increase of $\bar{r}$ to avoid sharp peaks that induce high regret even when we choose arms very close to the best arm. Second, we need a lower bound on the rate of increase of $\bar{r}$ to avoid intervals where there is too little separation between expected rewards.

Assumption 3.2 (Strong max)**.** Let $\varphi \approx 1.618$ denote the golden ratio. Assume that the function $\bar{r}$ is unimodal with the maximum at x^*. There exist Lipschitz constants $C_H > C_L > 0$ such that $|\bar{r}(x) - \bar{r}(y)| \leq C_H |x - y|$ for all pairs $x, y \in [0,1]$, and such that $|\bar{r}(x) - \bar{r}(y)| \leq C_L |x - y|$ for $x, y \in [x^* - C_L, x^* + C_L]$ and $|\bar{r}(x) - \bar{r}(y)| \geq C_L/\varphi^3 |x - y|$ for $x, y \in [0, x^* - C_L]$ or $x, y \in [x^* + C_L, 1]$.

3.1. Unimodality in Graphical Bandits

In addition to having an unimodal structure, the expected reward of the arms may have other interesting structures, such as those arising in the context of networks, *e.g.*, social networks and communication networks. One natural structure is obtained by associating the arms of the bandit to vertices in a graph and capturing the partial order relation with edges in the graph.

Figure 2 illustrates a concrete but simplified example, where each vertex of a tree graph corresponds a vector (x, y) of two features: a price x and a quality parameter y for a service offered by a firm. For every fixed quality, the offerings can be ordered by price; moreover, for a given base price, the offerings can be ordered by quality. The expected profit function is unimodal along every path. At each time instant, the firm's task is to make an offer (x, y) to a customer, with the ultimate goal of maximizing its long-term average profit.

Another example arises in online ad auctions (cf. (Varian, 2009)). From an advertiser's perspective, a partial order can be induced by nesting keyphrases, *e.g.*, "laptop" $\subset$ "buy laptop" $\subset$ "buy laptop in the US." Each arm is a vector with two components: a search keyphrase and a corresponding bid. Due to the large number of possible bid vectors, suppose that the advertiser may only choose from a small subset of arms. Every time a user searches one of the keyphrases, the advertiser's ad is displayed along with ads from other advertisers. To the advertiser, the expected revenue of each arm is unknown due to the presence of competitors and unknown user behavior.

In these examples, it is natural to assume that there exists an optimal vertex v^* with maximal expected reward $\bar{r}(v^*)$. Moreover, the farther a vertex v_i is from v^*, the lower is its expected reward. This gives rise to an unimodal structure in the expected reward, which is made precise in the following assumption.

Assumption 3.3 (Unimodality on graphs)**.** Let $\mathcal{G}$ be an undirected tree over the set of vertices $\mathcal{V}$. The expected reward function $\bar{r} : \mathcal{V} \to [0,1]$ is unimodal along every path $(v_1, \ldots, v_j)$ of $\mathcal{G}$, *i.e.*, there exists a vertex v_M in every path $(v_1, \ldots, v_j)$ such that

$$\bar{r}(v_1) < \bar{r}(v_2) < \ldots < \bar{r}(v_M)$$
$$\text{and} \quad \bar{r}(v_M) > \ldots > \bar{r}(v_{j-1}) > \bar{r}(v_j).$$

Remark 2. Observe that the expected reward $\bar{r}(v_i)$ corresponds to a label for vertex v_i. A graph with a unimodal reward function is therefore partially ordered according to the labels, but this partial order relation is a priori unknown.

Since there are finitely many arms in the graphical bandit, we replace the strong maximum assumption by the following assumption on the separation in expect reward of neighboring vertices.

Assumption 3.4 (Separation of expected rewards)**.** There exists a positive constant $D_L \in (0, 1]$ such that $D_L \leq |\bar{r}(v_i) - \bar{r}(v_{i+1})|$, for every pair of neighboring vertices v_i and v_{i+1}.

3.2. The Notion of Regret

Suppose that an algorithm generates a (random) sequence of actions $x_1, x_2, \ldots$, the corresponding total expected reward is $\sum_{t=1}^{T} \bar{r}(x_t)$. The *regret* of this algorithm is $L_T \triangleq \sum_{t=1}^{T} \left(\bar{r}(x^*) - \mathbb{E}[\bar{r}(x_t)]\right)$. This notion of expected regret is similar to that of (Lai & Robbins, 1985; Auer et al., 2002a), and we use it in Sections 4 and 5.

When the sequence of reward distributions contains change-points, we do not have a fixed expected reward function $\bar{r}$ for all time instants. In that case, we use the following notion of regret:

$$L_T^{\mathrm{C}} \triangleq \left(\sum_{t=1}^{T} \max_{x \in [0,1]} \bar{r}_t(x)\right) - \left(\sum_{t=1}^{T} \mathbb{E}[\bar{r}_t(x_t)]\right),$$

where $\bar{r}_t$ denotes the expected reward function at time t. The baseline of comparison for the regret L_T^{C} is the sum of *optimal* expected rewards, which differs entirely from the baseline used in the adversarial bandit problem (cf. (Auer et al., 2002b)). We call this the *non-stationary regret* and use it in Section 6.

4. Unimodal Bandit in One Dimension

In this section, we present some intermediate results in the simple one-dimensional unimodal bandit problem without abrupt changes in reward distribution. Various solutions with optimal performance guarantees already exist (Agrawal, 1995; Kleinberg, 2004; Auer et al., 2007; Kleinberg et al., 2008; Bubeck et al., 2008; Cope, 2009). However, we present a new efficient algorithm that hinges on the unimodality of the expected reward function; as a result, it only keeps four indices at any given time.

First, we present a simple algorithm for the stochastic multiarmed bandit problem with a probably approximately correct (PAC) guarantee. The Sampling Algorithm works on a finite set of arms $\{1, \ldots, m\}$. It takes two parameters ϵ and δ as input and samples the arms sequentially—in the order arm 1, arm 2, …, arm m, arm 1, …, and stops after $(4m/\epsilon^2)\log(2m/\delta)$ samples.

Theorem 4.1 (Theorem 1 of (Even-Dar et al., 2002))**.** *With probability* $1-\delta$*, the Sampling Algorithm outputs an arm* i^* *that is* ϵ*-optimal,* i.e.*, which has average reward* $\bar{r}(i^*) \geq \max_{i=1,\ldots,m} \bar{r}(i) - \epsilon$*. We say that this algorithm is* (ϵ, δ)*-PAC.*

Algorithm 1 Sampling Algorithm

1: **Input:** A set of m arms, $\epsilon > 0$, and $\delta > 0$.
2: Sample all arms $1, \ldots, m$ sequentially, until each arm has been sampled $(4/\epsilon^2)\log(2m/\delta)$ times.
3: Let the sample-average reward of arm i be denoted by $\widehat{r}(i)$. Output the arm $\arg\max_{i=1,\ldots,m} \widehat{r}(i)$.

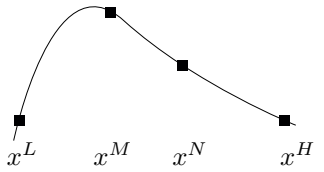


Figure 3. The four sampled arms and their expected rewards during one iteration. The interval $[x^N, x^H]$ is eliminated with high probability at the end of this iteration. If $|x^H - x^L| = 1$, then the distances between the points are: $|x^M - x^L| = |x^H - x^N| = \varphi^{-2}$ and $|x^N - x^M| = \varphi^{-3}$.

4.1. The LSE Algorithm

Our proposed LSE Algorithm (Algorithm 2) works as follows. At every iteration of the main loop, it narrows down the *sampling interval* $[x^L, x^H]$ in which the true optimal arm x^* lies with high probability. During each iteration, the LSE Algorithm runs the Sampling Algorithm over four arms x^L, x^M, x^N, x^H. These four arms are chosen as in Kiefer's golden section search algorithm (Kiefer, 1953). Three of these arms are kept from one iteration to the next; the fourth arm is new. At the end of each iteration, the algorithm narrows down the sampling interval by a constant factor $1/\varphi$, where φ is the golden ratio. Figure 3 illustrates the four arms and their true expected rewards during one iteration. Over time, the goal is to eliminate intervals that do not contain the optimal arm with high probability, such as $[x^N, x^H]$ in Figure 3, at the end of each iteration, and hence narrow down to smaller and smaller sampling intervals that contain the optimal arm x^* with high probability.

Remark 3 (Notation). Although the arms x^L, x^M, x^N, x^H change from one iteration to another, for simplicity of notation, we shall sometimes omit the subscript n. Hence, we write x^L instead of x_n^L when the n-th iteration is implicitly understood.

Remark 4. Observe that, from one iteration of the LSE Algorithm to the next, three of the four arms do not change. Hence, we can reuse samples from previous iterations to improve efficiency.

We first show that the expected regret of the LSE Algorithm is of the order of $O(\sqrt{T}\log T)$ and then give a PAC bound on the regret.

Algorithm 2 Line Search Elimination Algorithm

Input: Sequences ϵ_n and δ_n for $n = 1, 2, \ldots$.
(**Initialization**.) Set $[x^L, x^H] = [0, 1]$. Set x^M such that $(x^H - x^M)/(x^M - x^L) = \varphi$ (cf. Figure 3), where φ is the golden ratio. Set x^N in $[x^M, x^H]$ such that $(x^N - x^L)/(x^H - x^N) = \varphi$.
for iterations $n = 1, 2, \ldots$ **do**
  (Find (ϵ_n, δ_n)-PAC arm.) Run the Sampling Algorithm on the arms $\{x^L, x^M, x^N, x^H\}$ with parameters ϵ_n and δ_n. Let x_n^* denote the output.
  (Interval elimination.)
  if $x_n^* = x^N$ or $x_n^* = x^H$ **then**
    Eliminate the interval $[x^L, x^M]$. Update the points $x^L := x^M$ and $x^M := x^N$, and $x^N = (x^L + \varphi x^H)/(1 + \varphi)$.
  else
    Eliminate $[x^N, x^H]$. Update the points $x^H := x^N$, $x^N := x^M$, and $x^M = (\varphi x^L + x^H)/(1 + \varphi)$.
  end if
end for

Lemma 4.2. *Suppose that Assumptions 3.1 and 3.2 hold, and that $1/\varphi^n > C_L$. Then, for the n-th iteration of the LSE Algorithm, we have:* $\Pr\left(x^* \notin [x_n^L, x_n^H]\right) \leq \sum_{i=1}^n \delta_i$.

The proofs of all the results appear in (Yu & Mannor, 2011).

Theorem 4.3 (Expected regret of LSE)**.** *Let T be known and let $\varphi = (1 + \sqrt{5})/2$ denote the golden ratio. Suppose that we employ the LSE Algorithm with $\delta_n = 8/T$ and $\epsilon_n = C_L/\varphi^{n+3}$ for a total of N intervals $n = 1, \ldots, N$. Suppose that the assumptions of Lemma 4.2 hold. Then the regret of the LSE Algorithm is at most*

$$L_T \leq Z(C_H/C_L^2)\sqrt{1 + C_L^2 T}\log T + 2\log_\varphi^2(1 + C_L^2 T),$$

where $Z = 32\varphi^7/(\varphi - 1) \leq 1504$.

Remark 5. This upper bound is tight up to a logarithmic factor; it is independent of the number of arms and the difference between the best and second-best distributions, but depends on the Lipschitz constants.

Remark 6. For this and the subsequent results, we assume that the time horizon T is known. This assumption can be easily removed by employing the doubling trick (Cesa-Bianchi & Lugosi, 2006). This trick consists of starting with an arbitrary horizon, then, at the end of that horizon, we reset and restart our algorithm with a new horizon twice as long.

Theorem 4.4 (PAC-arm)**.** *Suppose that the assumptions of Theorem 4.3 hold. Let T be fixed. The LSE Algorithm with parameters*

$$\epsilon_n = \frac{C_L \epsilon}{C_H \varphi^3}, \quad \delta_n = \frac{16\varphi^6 C_H^2}{C_L^2 \epsilon^2 T}\delta\log(8/\delta), \quad \text{for all } n,$$

outputs an ϵ-optimal arm with probability $1 - \delta$ after T steps.

4.2. The LSE Algorithm for m Arms

When there is a finite number m of arms in a (totally-ordered) chain graph, a regret of $O(\log m \log T)$ can be achieved with the following version of the LSE Algorithm, as opposed to $O(m \log T)$ for a bandit algorithm that ignores unimodality.

Definition 4.1 (Finite LSE Algorithm)**.** The LSE Algorithm can be applied to a Finite number m of arms as follows. To each arm v_i of $(v_1, \ldots, v_m)$, we assign the interval $[(i-1)/m, i/m) \subset [0, 1]$, for $i = 1, \ldots, m$. Then, we replace every reference in the LSE Algorithm to an arm x in the interval $[(i-1)/m, i/m)$ by the arm v_i. The algorithm terminates and outputs x^L when $x^H - x^L < 1/m$.

Proposition 4.5 (Finite unimodal bandit)**.** *Suppose that there are m arms in a chain graph and that Assumptions 3.3 and 3.4 hold. Let T be fixed and known. Suppose that we follow the Finite LSE Algorithm with parameters $\epsilon_n = D_L$ and $\delta_n = 8/T$ for all n. Then, the expected regret is at most*

$$\widehat{L}_T(m) \leq \frac{16}{D_L^2}\left(\frac{1}{1 - 1/\varphi} + \frac{4\log_\varphi^2 m}{T}\right)\log T + 8\log_\varphi m.$$

5. Unimodal Bandits on Graphs

In this section, we consider bandits on graphs with unimodal rewards, as set out in Section 3.1. The objective is to exploit both the graphical structure and the unimodality assumption in order to find the best arm. By combining the Finite LSE Algorithm with a greedy search method, we obtain the GLSE Algorithm (Algorithm 3). Although the performance of the LSE Algorithm has explicit guarantees, the performance of the graphical algorithm, however, depends critically on characteristics of the graph that affect the number of iterations of the main loop.

Let d denote the maximal degree among vertices of the graph $\mathcal{G}$, and ℓ the diameter of the graph (*i.e.*, the longest shortest path between two vertices). Our main result in this section asserts that the GLSE Algorithm incurs a regret of the order of $O(d\ell\log(dT) + \ell\log\ell)$. Observe that a naive application of a standard bandit algorithm would yield a regret of $O(m\log T)$ which is potentially much larger since $m = O(d^\ell)$ in the worst case.

Algorithm 3 Graphical LSE (GLSE) Algorithm

Input: a tree $\mathcal{G}$, positive integers τ, and sequences ϵ_n and δ_n for $n = 1, 2, \ldots$.
(**Initialization**.) Pick a vertex x^* as the root.
repeat
  Run the Sampling Algorithm on the root x^* and all its neighbors until each vertex has been sampled τ times. Let y denote the output.
  if $y \neq x^*$ **then**
    Remove the root x^* and take the sub-tree rooted in y as the new tree $\mathcal{G}$.
    Find the longest path $\mathcal{P}$ through y in $\mathcal{G}$.
    Follow the Finite LSE Algorithm on the path $\mathcal{P}$ with parameters ϵ_n and δ_n. Let z denote the output.
    Remove from $\mathcal{G}$ the vertices of $\mathcal{P}$ except z.
    Set the new root: $x^* := z$.
  end if
until $y = x^*$
Return x^*.

Theorem 5.1 (Unimodal graphical bandit)**.** *Suppose that Assumptions 3.3 and 3.4 hold. The GLSE Algorithm with parameters $\epsilon_n = D_L$, $\delta_n = 8/T$, and $\tau = (4/D_L^2)\log(2(d+1)T)$ incurs a regret*

$$L_T \leq \frac{2d\ell}{D_L^2}\log\big(2(d+1)T\big) + \frac{8\ell}{(1-1/\varphi)D_L^2}\log T + \frac{32\ell\log T}{D_L^2 T}\log_\varphi^2 m + 4\ell\log_\varphi m.$$

5.1. Experiment

We compared the performance of the GLSE Algorithm on a a star-shaped graph—as in Figure 2—with 50 branches, each of which contains 500 vertices. The reward distributions for all vertices are normal, with variance 1 and expected value in the interval $[0, 1]$. A vertex v^* is chosen uniformly at random and assigned the highest expected reward $\bar{r}(v^*)$. The expected reward of other vertices decreases in constant steps as we move away from v^*.

Figure 4 shows the empirical average regret of the GLSE and UCB1 Algorithms, which ignores the unimodal structure of rewards. The two sets of four plots correspond to four typical runs of the experiment, which illustrates the variance in actual performance. Observe that in contrast to the GLSE Algorithm, the UCB1 algorithm incurs high regret initially because every arm is sampled separately. The regret of the GLSE Algorithm exhibits a small jump when it begins line search on a new maximal path of a sub-tree.

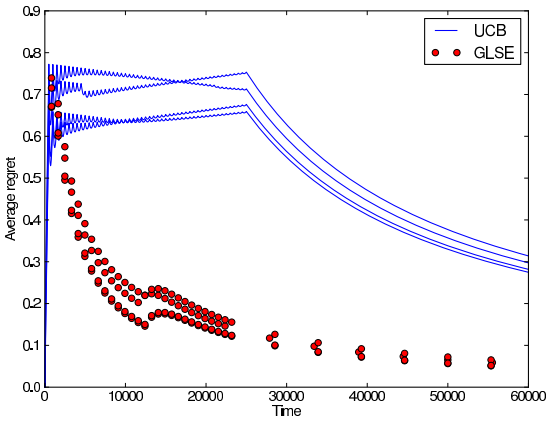


Figure 4. Convergence of the expected reward of the GLSE and UCB1 Algorithms.

6. Unimodality and Change-points

Bandit problems with unknown change-points are notoriously hard. Even when the number of changes is fixed and known a priori, the worst-case regret L_T^C is at least of the order of $\Omega(\sqrt{T})$ (Garivier & Moulines, 2008). Using the unimodality assumption, we show an efficient method of detecting change-points that does not require prior knowledge of the frequency of changes, and that incurs a regret that matches this lower bound up to a logarithmic factor. Although this method can be employed in the bandits on graphs setting of Section 6, we present the method in the one-dimensional setting of Section 4 for simplicity. The extension to graphs is technical and therefore omitted.

First, we define the notion of abrupt changes in the reward distributions. We assume that there is a sequence of change-points $\nu_1, \nu_2, \ldots$ such that between consecutive change-points ν_i and ν_{i+1}, the rewards of the arms are i.i.d. random variables. Similarly to adversarial learning problems (cf. (Cesa-Bianchi & Lugosi, 2006)), both the change-points $\nu_1, \nu_2, \ldots$ and the reward distributions are unknown. We are interested in the case where the change-point occur with low frequency. Our objective is to detect changes while excluding infinitesimal changes.

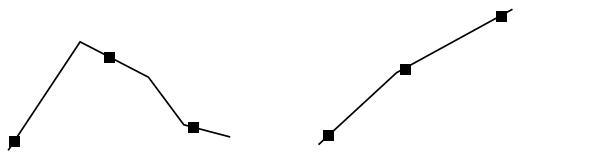
Assumption 6.1. Let ν denote a change-point. Let $\bar{r}_{\nu-1}$ and $\bar{r}_{\nu+1}$ denote the expected reward functions before and after ν. Let $x^*(\nu-1)$ and $x^*(\nu+1)$ denote optimal arms for $\bar{r}_{\nu-1}$ and $\bar{r}_{\nu+1}$. We assume that there exists a constant $\beta > 0$ such that $|x^*(\nu-1) - x^*(\nu+1)| > \beta$ for every change-point ν.

Remark 7. By Assumptions 3.2 and 6.1, if we do not detect a change-point ν where $|x^*(\nu-1) - x^*(\nu+1)| \leq \beta$, then we incur a regret of at most $C_H\beta$ per time step, where β can be chosen appropriately small.

Algorithm 4 Adaptive LSE Algorithm

Input: θ, τ, ϕ, sequences ϵ_n and δ_n for $n = 1, 2, \ldots$.
(**Initialization**.) Same as the LSE Algorithm.
for iterations $n = 1, 2, \ldots$ **do**
  Insert lines 4–10 of the LSE Algorithm.
  (Change Detection.) Every θ time steps, run the Sampling Algorithm on the four arms $\{x_L - \phi, x_L, x_H, x_H + \phi\}$ until each has been sampled τ times. Let y denote the output.
  if $y = x_L - \phi$ or $y = x_H + \phi$ **then**
    Reset algorithm: $[x^L, x^H] := [0, 1]$ and $n := 1$.
  end if
end for

A general approach for detecting change-points is to compare the empirical average rewards over successive windows. As in the previous section, we consider a unimodal setting where the expected reward function between successive change-points is unimodal. In this setting, a simpler approach is to detect when the current best arm is superseded by an arm to its left or its right. We take the second approach in this paper. Figure 5 suggests how a change-point may be detected by sampling arms to the left and right of the current optimal arm.



(a) Expected reward function before ν.

(b) Expected reward function after ν.

Figure 5. Typical expected reward function before and after change-point ν.

The following Adaptive LSE Algorithm is obtained by inserting a change-detection phase to the LSE Algorithm.

Theorem 6.1. *Suppose that Assumptions 3.1, 3.2, and 6.1 hold. Further, suppose that there are k change-points up to time T. For $T \geq 2/\beta$, $\phi = \beta/2$, $\theta = \sqrt{T}$, and $\tau = 16\log(8T)/(C_L^2\beta^2)$, the non-stationary regret of the Adaptive LSE Algorithm is at most*

$$L_T^{\mathrm{C}} \leq \frac{64}{C_L^2\beta^2}(\sqrt{T}+1)\log(8T) + 1 + 1504\frac{C_H}{C_L^2}\, k\sqrt{1+C_L^2 T}\log T + 2k\log_\varphi^2(1+C_L^2 T).$$

This upper bound is of the order of $O(k\sqrt{T}\log T)$. It matches the lower bound for the non-stationary bandit up to a logarithmic factor (cf. (Garivier & Moulines, 2008)). In contrast to (Garivier & Moulines, 2008), our solution does not require prior knowledge of the number of change-points k, but instead requires the unimodality assumption.

6.1. Experiment

We conduct an experiment on an interval of arms $[0, 1]$. Between consecutive change-points, the reward distribution of each arm is a normal distribution with unit variance and expected value shown in Figure 6(a). The superposition of Figures 6(b) and 6(a) shows that from one iteration to the next, the sampling interval of the Adaptive LSE scales down toward the arm with optimal expected reward and resets after a change-point. Observe that a reset occurs at time 2×10^3 because the optimal arm has been erroneously eliminated.

7. Discussion

The unimodal expected reward property allows efficient solutions to bandit problems with a large number of partially ordered arms—*e.g.*, continuous intervals and graphs, which are central to the study of voting theory, repeated auctions, and social networks. An open problem remains to extend our techniques to high-dimensional spaces.

Acknowledgments

We would like to thank the reviewers for their comments. J. Y. Yu was supported in part by French National Research Agency (ANR, project EXPLO-RA, ANR-08-COSI-004), the PASCAL2 Network of Excellence under EC grant no. 216886, and a fellowship from FQRNT. S. Mannor was partially supported by the Israel Science Foundation under contract 890015.

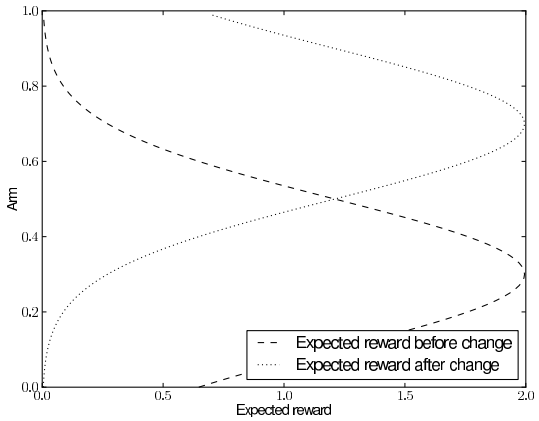
References

Agrawal, R. The continuum-armed bandit problem. *SIAM J. Control Optim.*, 33:1926–1951, 1995.

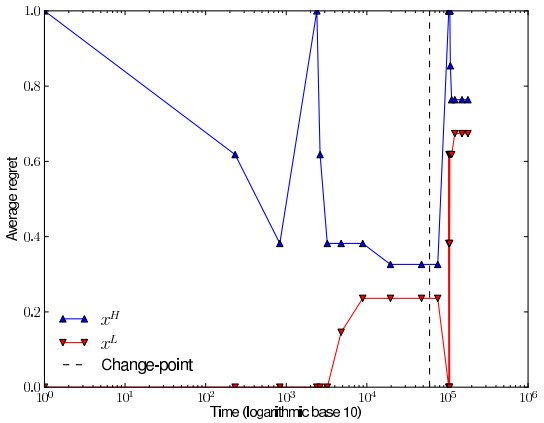
Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002a.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002b.

Auer, P., Ortner, R., and Szepesvári, C. Improved rates for the stochastic continuum-armed bandit problem. In *Proc. COLT*, 2007.



(a) Expected reward functions before and after the change-point, with optimal arms 0.3 and 0.7 respectively.



(b) Sampling intervals narrowing down toward the optimal arms of Figure 6(a).

Figure 6. Expected reward and the sampling interval during one run of the Adaptive LSE.

Bansal, N., Blum, A., Chawla, S., and Meyerson, A. Online oblivious routing. In *Proc. of SPAA*, 2003.

Blum, A., Kumar, V., Rudra, A., and Wu, F. Online learning in online auctions. In *Symp. on Discrete Alg.*, pp. 202–204, 2003.

Bubeck, S., Munos, R., Stoltz, G., and Szepesvári, C. Online optimization in X-armed bandits. In *Proc. NIPS*, 2008.

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

Cesa-Bianchi, N., Gentile, C., and Vitale, F. Fast and optimal prediction on a labeled tree. In *Proc. COLT*, 2009a.

Cesa-Bianchi, N., Gentile, C., and Vitale, F. Learning unknown graphs. In *Proceedings of Algorithmic Learning Theory*, 2009b.

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

Even-Dar, E., Mannor, S., and Mansour, Y. PAC bounds for multi-armed bandit and Markov decision processes. In *Proc. COLT*, 2002.

Garivier, A. and Moulines, E. On upper-confidence bound policies for non-stationary bandit problems. In *Proc. EWRL*, 2008.

Hartland, C., Gelly, S., Baskiotis, N., Teytaud, O., and Sebag, M. Multi-armed bandit, dynamic environments and meta-bandits. In *Workshops of NIPS*, 2006.

Kiefer, J. Sequential minimax search for a maximum. *Proc. Amer. Math. Soc.*, 4(3):502–506, 1953.

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

Kleinberg, R., Slivkins, A., and Upfal, E. Multi-armed bandits in metric spaces. In *Proc. STOC*, 2008.

Lai, T. L. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

Mas-Collel, A., Whinston, M. D., and Green, J. R. *Microeconomic Theory*. Oxford University Press, 1995.

Müller-Hannemann, M. and Weihe, K. *Algorithm Engineering*, Pareto Shortest Paths is Often Feasible in Practice. Springer, 2001.

Pandey, S., Chakrabarti, D., and Agarwal, D. Multi-armed bandit problems with dependent arms. In *Proc. ICML*, 2007.

Spinrad, J. P. *Efficient graph representations*. American Mathematical Society, 2003.

Varian, H. R. Online ad auctions. *American Economic Review*, 99(2):430âĂŞ–434, 2009.

Yu, J. Y. and Mannor, S. Unimodal bandits. `http://www.math.ens.fr/~jiayu/unimodal.pdf`.

Yu, J. Y. and Mannor, S. Piecewise-stationary bandit problems with side observations. In *Proc. ICML*, 2009.