Piecewise-stationary Bandit Problems with Side Observations

Jia Yuan Yu JIA.YU@MCGILL.CA
Department Electrical and Computer Engineering, McGill University, Montréal, Québec, Canada.

Shie Mannor SHIE.MANNOR@MCGILL.CA; SHIE@EE.TECHNION.AC.IL
Department Electrical and Computer Engineering, McGill University, Montréal, Québec, Canada.
Department of Electrical Engineering, Technion, Haifa, Israel.

Abstract

We consider a sequential decision problem where the rewards are generated by a piecewise-stationary distribution. However, the different reward distributions are unknown and may change at unknown instants. Our approach uses a limited number of side observations on past rewards, but does not require prior knowledge of the frequency of changes. In spite of the adversarial nature of the reward process, we provide an algorithm whose regret, with respect to the baseline with perfect knowledge of the distributions and the changes, is $O(k \log(T))$, where k is the number of changes up to time T. This is in contrast to the case where side observations are not available, and where the regret is at least $\Omega(\sqrt{T})$.

1. Introduction

In some learning scenarios, the agent is confronted with an adversarial opponent that can be very general and difficult to model, and is therefore modelled as an arbitrary non-stochastic process. In other scenarios, the opponent is stochastic, which may be characterized and adapted to. What about opponents that fall between these two extremes? An instance of the adversarial scenario is the expert problem (Littlestone & Warmuth, 1994), where the agent observes sequentially the performance of a number of experts, and (choosing one expert at each time step) tries to match the performance of the best expert in retrospect. An instance of the stochastic scenario is the multi-armed bandit problem (Lai & Robbins, 1985), where each of n arms has a fixed reward distribution, and where the agent tries to obtain the performance of the best arm by picking and observing one arm each time step—without observing the reward of any other arm.

We consider a model that combines the bandit and expert models, and shall refer to the arms of the bandit and the experts interchangeably. The reward process of the arms is non-stationary on the whole, but stationary on intervals. This piecewise-stationary reward process is similar to that of the non-stationary bandit problem of (Hartland et al., 2006; Garivier & Moulines, 2008), or that of the multiple change-point detection problem of (Akakpo, 2008).

In our variant of the non-stationary bandit problem, the agent has the benefit of querying and observing some of the past outcomes of arms that have not been picked. This is the same benefit available to the agent in the expert problem (cf. Herbster & Warmuth, 1998). The following examples motivate our model.

Example 1.1 (Investment options)**.** Consider the problem of choosing every day one of n investment options, say mutual funds. Our model assumes that the outcomes of these investments undergo changes reflecting changes in market conditions. Otherwise, the outcomes remains stationary over the periods between two changes, *e.g.*, they follow bearish or bullish trends. Suppose that the outcomes of the previous day's investment options are revealed today, *e.g.*, in the newspaper. Suppose that observing the outcome of each option requires a query (looking up a price history), which incur a querying cost. By limiting the number of queries allowed at each step, we can model the trade-off between the cost of observations and the regret due to insufficient observations.

Example 1.2 (Dynamic pricing with feedback)**.** As a second example, we consider a vendor whose task is to sell commodity X. Potential customers arrive sequentially, one after the other, and the demand for

Appearing in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009. Copyright 2009 by the author(s)/owner(s).

commodity X (for various prices) is modelled as a stationary process that may nonetheless change abruptly at unknown instants. To each customer, the vendor offers one of n possible prices. If the customer accepts, a corresponding profit is made. Bargaining is not an option, but after each transaction, the vendor has the leisure to ask the customer is the outcome would have been different had a different price been offered (*e.g.*, through a short survey). A partial goal is to achieve as much profit as if the distribution of the demand were known at all times (even though unknown changes occur at unknown instants). A second goal is to minimize the cost associated with conducting surveys for feedback. A similar problem of dynamic pricing with partial-monitoring is also described in (Cesa-Bianchi & Lugosi, 2006).

We present the setting in Section 2, followed by a survey of related works in Section 3. We present a solution and its guarantee in Section 4. In Section 5, we compare our solution with other solutions via simulation. In Section 6, we conclude with a discussion.

2. Setting

We consider the following sequential decision problem. Let $\{A_1, \ldots, A_n\}$ denote the n arms of a multi-armed bandit—or n experts of an online learning problem. Let $b_1, b_2, \ldots$ be a sequence of reward vectors in $\mathbb{R}^n$. The element $b_t(i)$ of b_t, for $i = 1, \ldots, n$ and $t = 1, 2, \ldots$, represents the reward associated with the i-th arm A_i at time t. With an abuse of notation, we shall write $b_t(A_i)$ interchangeably with $b_t(i)$. We assume that the rewards take values in the unit interval $[0, 1]$, *i.e.*, $b_t(i) \in [0, 1]$ for all i and t.

2.1. Reward Process

In our model, the source of rewards is piecewise-stationary: *i.e.*, it changes its distribution arbitrarily and at arbitrary time instants, but otherwise remains stationary. The reward process $b_1, b_2, \ldots$ is an independent sequence of random variables that undergoes abrupt changes in distribution at unknown time instants $\nu_1, \nu_2, \ldots$, which are called *change-points*. By convention, we let $\nu_1 = 1$. Let f_t denote the distribution (probability density function) of b_t. Hence, $b_{\nu_1}, \ldots, b_{\nu_2 - 1}$ are i.i.d. with common distribution f_{ν_1}, as is the case for stochastic learning problems (cf. Lai & Robbins, 1985). Likewise, $b_{\nu_j}, b_{\nu_j + 1}, \ldots, b_{\nu_{j+1} - 1}$ are i.i.d. with distribution f_{ν_j}, for $j = 1, 2, \ldots$. The intervals are illustrated as follows:

$$\underbrace{b_1, b_2, \ldots, b_{\nu_2 - 1}}_{\text{distribution } f_{\nu_1}}, \underbrace{b_{\nu_2}, \ldots, b_{\nu_3 - 1}}_{\text{distribution } f_{\nu_2}}, \ldots, \underbrace{b_{\nu_j}, \ldots, b_{\nu_{j+1} - 1}}_{\text{distribution } f_{\nu_j}} \ldots.$$

Similarly to adversarial learning problems (cf. (Cesa-Bianchi & Lugosi, 2006)), both the change-points $\nu_1, \nu_2, \ldots$ and the distributions $f_{\nu_1}, f_{\nu_2}, \ldots$ are unknown. We can think of an opponent deciding the time instants (and frequency) of the changes, as well as the distribution after each change.

Remark 1. It is important that the changes occur at arbitrary instants. Otherwise, we only need to reset an algorithm for the multi-armed bandit problem at the appropriate instants.

The model of piecewise-stationary rewards combines two important models. If there are no changes, then we recover the stochastic source of the multi-armed bandit problem. If there is no constraint on the number of changes, we obtain the source of rewards adopted by the adversarial model of prediction with expert advice. We consider the interesting case where the frequency of changes is between these two extremes, *i.e.*, where the number of change-points

$$k(T) \triangleq \sum_{t=1}^{T-1} \mathbf{1}_{[f_t \neq f_{t+1}]}$$

up to time T increases with T. To simplify notation, we shall simply write k in place of $k(T)$.

2.2. Decision-maker

At each time step $t > 1$, the agent picks an arm $a_t \in \{A_1, \ldots, A_n\}$ and makes ℓ (where $1 \leq \ell \leq n$) observations on the individual arm-rewards $b_{t-1}(1), \ldots, b_{t-1}(n)$. This is captured in the following assumption.

Assumption 2.1 (Partial observation)**.** At time 1, the agent chooses an action a_1 and an ℓ-subset S_1 of the arms $\{A_1, \ldots, A_n\}$. At every time step $t > 1$, the agent chooses (deterministically) an ℓ-subset S_t and takes an action a_t that is a function of the reward observations

$$\{b_j(i) \mid j = 1, \ldots, t-1, \quad A_i \in S_j\}.$$

Partial observation allows us to capture querying costs associated with observations, and to quantify the total query budget.

2.3. Notion of Regret

At each time instant t, the agent chooses and activates an arm $a_t \in \{A_1, \ldots, A_n\}$ and receives the corresponding reward $b_t(a_t)$. Let β_t denote the mean of the reward vector b_t. The agent's baseline—or objective—is the reward accumulated by picking at each instant t an arm with the maximal expected reward. Letting k

denote the number of changes in reward distribution up to time T, the baseline is

$$\sum_{t=1}^{T} \max_{i=1,\ldots,n} \beta_t(i) = \max_{\sigma_1,\ldots,\sigma_T \,:\, k \text{ changes}} \sum_{t=1}^{T} \mathbb{E}[b_t(\sigma_t)],$$

where the maximum is taken over sequences of arms with only as many changes as change-points in the reward sequence $b_1, \ldots, b_T$, *i.e.*, over the set

$$\{\sigma_1, \ldots, \sigma_T \mid \sigma_{\nu_j} = \ldots = \sigma_{\nu_{j+1}-1} \quad \text{for } j = 1, \ldots, k\}.$$

Despite the appearance, this objective is reasonable when the number of changes k is small; it is also the same objective as in the classical stochastic multi-armed bandit problems. Hence, for a given reward process $b_1, b_2, \ldots$, we define the expected regret of the agent at time T as

$$R_T \triangleq \sum_{t=1}^{T} \max_{i=1,\ldots,n} \beta_t(i) - \sum_{t=1}^{T} \mathbb{E}[b_t(a_t)], \qquad (1)$$

where the expectation $\mathbb{E}$ is taken with respect to the sequence $b_1, b_2, \ldots$.

3. Related Works

In this section, we survey results concerning related models. The different models are distinguished by the source of the reward process, the observability of the rewards, and the baseline for the notion of regret.

3.1. Stochastic Multi-armed Bandit

In stochastic multi-armed bandit problems (Lai & Robbins, 1985; Auer et al., 2002a), the reward sequence $b_1, b_2, \ldots$ is a sequence of i.i.d. random vectors from a common unknown distribution $\beta_1 = \beta_2 = \ldots$. The reward observations are limited to rewards $b_1(a_1), b_2(a_2), \ldots$ corresponding to the arms chosen by the agent. This invites the agent to trade-off exploring the different arms to estimate their distributions and exploiting the arms with the highest empirical reward. The notion of regret is the same as ours (1). However, the optimal reward of the baseline can be obtained by a single fixed arm. In such problems, the optimal expected regret is of the order of $O(n \log(T))$, which may be obtained by a number of algorithms, *e.g.*, (Lai & Robbins, 1985; Auer et al., 2002a; Kocsis & Szepesvári, 2006).

3.2. Adversarial Expert Problem

Many learning problems take the adversarial setting, *e.g.*, prediction with expert advice, etc.—see (Cesa-Bianchi & Lugosi, 2006) for a comprehensive review. The sequence of rewards achieved by the experts is arbitrary; *i.e.*, no assumption is made regarding the joint distribution of $b_1, b_2, \ldots$. This approach essentially makes provisions for the worst-case sequence of reward. At time t, the past reward vectors $b_1, \ldots, b_{t-1}$ are observable by the agent. In this case, the notion of *adversarial regret* is adopted, whose baseline is the reward accumulated by the best fixed expert, *i.e.*, $\max_{i=1,\ldots,n} \sum_{t=1}^{T} b_t(i)$. For every sequence $b_1, b_2, \ldots$, the (expected) adversarial regret

$$\max_{i=1,\ldots,n} \sum_{t=1}^{T} b_t(i) - \sum_{t=1}^{T} \mathbb{E}[b_t(a_t)]$$

is of the order of $O(\sqrt{T \log(n)})$—see (Cesa-Bianchi & Lugosi, 2006) for a detailed account. A similar bound holds when the observations are limited to the chosen arms: $b_1(a_1), b_2(a_2), \ldots$ (Auer et al., 2002b).

The baseline in the adversarial case is limited to a single fixed expert, whereas our baseline in (1) is the optimal expected reward. Our baseline, which contains as many switches as changes in distribution, is similar to the baseline defined by appropriately chosen shifting policies in (Herbster & Warmuth, 1998). The fixed-share algorithm or one of its variants (Herbster & Warmuth, 1998; Auer et al., 2002b) can be applied to our setting, if the number of changes $k(T)$ is given in advance, yielding a regret of $O(\sqrt{nkT \log(nT)})$. We present an algorithm with a regret of $O(nk \log(T))$ without prior knowledge of $k(T)$. It should be noted that when $k(T)$ is of the same order as T, it is hopeless to minimize the regret of (1): consider an adversary that picks the new distribution after each change-point.

3.3. Non-stationary Bandits

Our problem is reminiscent of the non-stationary bandit problem of (Hartland et al., 2006; Garivier & Moulines, 2008). The reward process and the notion of regret are similarly defined, as in Section 2. However, in those works, observation of the past rewards is limited to the chosen arms; hence, at time t, the agent's choice a_t is a function of $b_1(a_1), b_2(a_2), \ldots$. Using a statistical change detection test, Hartland *et al.* present a partial solution for instances where the best arm is not superseded by another arm following a change. In the event that an oracle reveals a-priori the number of changes $k(T)$ up to time T, Garivier and Moulines provide solutions that achieve a regret of $O(n\sqrt{kT \log(T)})$; a lower-bound of $\Omega(\sqrt{T})$ is also shown.

With respect to the above non-stationary bandit model, the distinguishing feature of our model is that, in addition to activating an arm at each time instant, the agent may query the past reward of one or more arms. We show that with T queries in total, the regret is bounded by $O(nk\log(T))$. Hence, queries reduce significantly the regret with respect to the results of (Garivier & Moulines, 2008).

3.4. Change Detection

Another problem related to ours is that of fault detection-isolation (Lai, 2001). In that problem, the goal is to detect the change, and classify the post-change distribution within a finite set of possibilities. In our problem, the distribution after the change can be arbitrarily different. Moreover, we do not classify; instead, we apply a minimum-regret learning algorithm. In contrast to the change-detection literature, we consider a more complex setting, where the sequence $b_1, b_2, \ldots$ may have multiple (arbitrarily many) changes. The problem of joint-detection of multiple change-points is addressed in (Akakpo, 2008) and references therein.

4. Multi-armed Bandits with Queries

In this section, we present an algorithm for our setting and provide its performance guarantee. We begin with two assumptions. We shall use as a component of our solution a typical multi-armed bandit algorithm described in the first assumption. The second assumption describes a limitation of our algorithm.

Assumption 4.1 (MAB algorithm for $k = 1$)**.** Consider a multi-armed bandit where there are no distribution changes (except at time 1). Let the i.i.d. reward sequence $b_1, b_2, \ldots$ have distribution β. Let $A_{i^{(1)}}$ and $A_{i^{(2)}}$ denote, respectively, the arm with the highest and second-highest mean. Let Δ denote their mean difference:

$$\Delta = \beta(i^{(1)}) - \beta(i^{(2)}).$$

Let $\mathcal{A}$ be an algorithm that guarantees a regret of at most $Cn\log(T)/\Delta$, for some constant C. At each step $t > 1$, algorithm $\mathcal{A}$ receives as input the reward $b_{t-1}(a_{t-1})$ obtained in the previous step, and outputs a new arm choice a_t. Examples of candidate algorithms include those of (Lai & Robbins, 1985; Auer et al., 2002a).

In this paper, we are concerned with detecting abrupt changes bounded from below by some threshold; we exclude infinitesimal changes in the following assumption.

Assumption 4.2. Recall that $\beta_{\nu_j}(i)$ and $\beta_{\nu_{j+1}}(i)$ denote the pre-change and post-change distributions of the arm A_i at the change-point ν_{j+1}. There exists a known value $\epsilon > 0$ such that, for each $j = 1, 2, \ldots$, there exists an arm A_i such that

$$\left|\beta_{\nu_j}(i) - \beta_{\nu_{j+1}}(i)\right| > 2\epsilon.$$

4.1. The WMD Algorithm

Our algorithm (Figure 1) detects changes in the mean of a process, in the spirit of statistical methods for detecting an abrupt change of distribution in an otherwise i.i.d. sequence of random variables (see (Lai, 2001) for a survey). The algorithm partitions the time horizon into intervals of equal length τ. Hence, for $m = 1, 2, \ldots$, the m-th interval is comprised of the time instants $(m-1)\tau+1, (m-1)\tau+2, \ldots, m\tau$. The algorithm computes iteratively empirical mean vectors $\widehat{b}_1, \widehat{b}_2, \ldots$ over intervals (windows) of length τ, in the following fashion:

$$\underbrace{b_1, b_2, \ldots, b_\tau}_{\widehat{b}_1}, \underbrace{b_{\tau+1}, \ldots, b_{2\tau}}_{\widehat{b}_2}, \ldots, \underbrace{b_{(m-1)\tau+1}, \ldots, b_{m\tau}}_{\widehat{b}_m} \ldots.$$

The algorithm follows a multi-armed bandit algorithm $\mathcal{A}$ with a regret guarantee in the absence of changes (Assumption 4.1). When it detects a mean shift with respect to a threshold given by Assumption 4.2, it reset the sub-algorithm $\mathcal{A}$.

4.2. WMD Regret

The following theorem bounds the expected regret of the WMD algorithm.

Theorem 4.1 (WMD regret)**.** *Suppose that Assumption 2.1 holds. Suppose that the agent employs the WMD algorithm with a sub-algorithm satisfying Assumption 4.1, a threshold ϵ satisfying Assumption 4.2, and intervals of length $\tau = \lfloor \frac{n}{\ell} \rfloor \cdot \lfloor \frac{\log(T)}{2\epsilon^2} \rfloor$. Then, for every sequence of change-points $\nu_1, \nu_2, \ldots$ and every choice of post-change distributions $f_{\nu_1}, f_{\nu_2}, \ldots$, the expected regret is bounded as follows:*

$$R_T \leq \frac{7}{\epsilon^2}\frac{kn}{\ell}\log(T) + \frac{C}{\Delta}kn\log(T) + \frac{6C}{\Delta}n^2, \qquad (2)$$

where C is the constant of Assumption 4.1.

Remark 2. The WMD algorithm does not require prior knowledge of the number of distribution changes $k(T)$.

Remark 3 (Query-regret trade-off). The bound of Theorem 4.1 indicates a way to trade-off the number of queries ℓ per step and the expected regret per step. Suppose that an increasing function C_Q assigns a cost, in the same unit as the rewards and the regret, to the rate of queries ℓ. The corresponding new objective thus becomes the sum of two components:

Input: interval length $\tau > 0$, threshold $\epsilon > 0$, and ℓ queries per step. Initialize $r := 1$.
At each step t:

1. (Follow $\mathcal{A}$.) Follow the action of an algorithm $\mathcal{A}$ satisfying Assumption 4.1.

2. (Querying policy.) If t belongs to the m-th interval except its first step, *i.e.*, if $t \in [(m-1)\tau+2, \ldots, m\tau]$, let $\Sigma_{t-1}(i)$ denote the number of queries arm A_i has received since the start of the m-th interval until step $t-1$. Order the arms $\{A_1, \ldots, A_n\}$ according to $\Sigma_{t-1}(1), \ldots, \Sigma_{t-1}(n)$. Query the set S_t of arms that received the fewest queries. Update the following elements of the empirical mean $\widehat{b}_m$:
$$\widehat{b}_m(i) := \frac{\Sigma_{t-1}(i)\, \widehat{b}_m(i) + b_{t-1}(i)}{\Sigma_{t-1}(i) + 1}, \quad \text{for every } i \in S_t.$$

3. (Detect change.) At the start of the m-th interval, *i.e.*, if $t = (m-1)\tau + 1$ for some $m = 3, 4, \ldots$. If $\left\|\widehat{b}_m - \widehat{b}_r\right\|_\infty > \epsilon$, reset (*i.e.*, re-instantiate) algorithm $\mathcal{A}$ and set $r := m$. The index r denotes the last interval at which the algorithm $\mathcal{A}$ was reset.

Figure 1. Windowed mean-shift detection (WMD) algorithm

query cost and regret. This overall expected cost-per-step at time T is $C_Q(\ell) + R_T/T$. With the implicit assumption that the bound (2) is tight in the duration T, the number of changes k, and the query rate ℓ, this cost can be optimized with respect to ℓ. If each query is assigned a constant cost c_q, *i.e.*, $C_Q(\ell) = c_q \cdot \ell$, then the (non-discrete) optimal query rate is $\ell^* = \sqrt{(7kn/C_Q)\log(T)/T}$. This is the type of optimization problem that has to be resolved in Example 1.2.

Proof of Theorem 4.1. The proof is composed of five steps. In the first step, we identify the components of the regret. In the second step, we analyze the empirical means computed by the WMD algorithm. In the successive steps, we bound the components of the regret. The components are combined in the final step.

Step 1. Let $L(T)$ denote the expected number of intervals after a change-point ν_j occurs before it is detected by the WMD algorithm (*i.e.*, algorithm $\mathcal{A}$ is reset). Let $N(T)$ be the expected number of false detections up to T, *i.e.*, instances when the algorithm $\mathcal{A}$ resets when no change-point has occurred since the last time $\mathcal{A}$ was reset. Observe that the total number of times the algorithm resets is bounded from above by $k + N(T)$. Hence, over a T-step horizon, there are at most $k + N(T)$ interval-periods during which the algorithm resets once and the source distribution does not change. By Assumption 4.1, during each such period, the expected regret is of the order of $Cn\log(\Gamma)/\Delta$, where Γ is the length of the period. Since log is a concave function and $\Gamma \leq T$, the expected regret over all such periods is at most $(Cn/\Delta)(k + N(T))\log(T)$. The algorithm may also incur regret during the delay between distribution change and its detection. Since there are k distribution changes, each ocurring at most $\lceil L(T) \rceil \tau$ time steps before the algorithm $\mathcal{A}$ resets. Hence, the total regret of this algorithm is at most
$$k(L(T)+1)\tau + \frac{Cn}{\Delta}(k + N(T))\log(T). \qquad (3)$$

Next, we bound $N(T)$ and $L(T)$, starting with $N(T)$. Observe that the term $k\tau$ accounts for the regret within intervals during which a change occurs. Hence, for the remainder of the proof, we consider only the intervals that do not contain a distribution change.

Step 2 (Empirical means). Consider the empirical mean over each interval that does not contain a change. Let $\gamma = \tau/\lfloor \frac{n}{\ell} \rfloor = \lfloor \frac{\log(T)}{2\epsilon^2} \rfloor$. Observe that by the construction of the WMD algorithm, after the end of the m-th interval, spanning time steps $(m-1)\tau + 1, \ldots, m\tau$, every arm is queried either γ or $\gamma + 1$ times. In the former case, the empirical mean for an arm A_i is the mean of γ i.i.d. random variables
$$\widehat{b}_m(i) = \frac{1}{\gamma} \sum_{t=(m-1)\tau+1}^{m\tau} b_t(i) \cdot \mathbf{1}_{[i \in S_t]}.$$

The expression for the latter case (of $\gamma + 1$ queries) is similar, and is omitted in this proof.

Step 3 (Number of false detections). Suppose that in the opponent action sequences during the m-th and r-th intervals are generated from the same distribution with expected value denoted, with an abuse of notation, by β_m. Observe that

$$\begin{aligned} N(T) &\le \lceil T/\tau \rceil P\left(\left\|\widehat{b}_m - \widehat{b}_r\right\|_\infty > \epsilon\right) \\ &\le n\lceil T/\tau\rceil \min_{i=1,\ldots,n} P\left(\left|\widehat{b}_m(i) - \widehat{b}_r(i)\right| > \epsilon\right), \end{aligned}$$

since there are at most $\lceil T/\tau \rceil$ intervals. Observe that, for every $i = 1, \ldots, n$,

$$\begin{aligned} &P\left(\left|\widehat{b}_m(i) - \widehat{b}_r(i)\right| > \epsilon\right) \\ &= P\left(\left|\widehat{b}_m(i) - \widehat{b}_r(i)\right| > \epsilon, \left|\widehat{b}_m(i) - \beta_m(i)\right| \le \epsilon\right) \\ &\quad + P\left(\left|\widehat{b}_m(i) - \widehat{b}_r(i)\right| > \epsilon, \left|\widehat{b}_m(i) - \beta_m(i)\right| > \epsilon\right) \\ &\le P\left(\left|\widehat{b}_m(i) - \widehat{b}_r(i)\right| > \epsilon \mid \left|\widehat{b}_m(i) - \beta_m(i)\right| \le \epsilon\right) \\ &\quad + P\left(\left|\widehat{b}_m(i) - \beta_m(i)\right| > \epsilon\right) \\ &\le P\left(\left|\widehat{b}_r(i) - \beta_m(i)\right| > 2\epsilon\right) \\ &\quad + P\left(\left|\widehat{b}_m(i) - \beta_m(i)\right| > \epsilon\right) \qquad (4) \\ &\le \exp\left(-8\gamma\epsilon^2\right) + \exp\left(-2\gamma\epsilon^2\right), \end{aligned}$$

where the last inequality follows by Step 2 and Hoeffding's Inequality (recall that $\widehat{b}_r(i)$ and $\widehat{b}_m(i)$ have the same distribution). Hence, we have

$$N(T) \le 2n \exp\left(-2\gamma\epsilon^2\right)(T/\tau + 1). \qquad (5)$$

Step 4 (Delay in change detection). Next, we bound $L(T)$. Suppose that there is a reset at the s-th interval. The WMD algorithm compares successively the empirical means $\widehat{b}_{s+1}, \widehat{b}_{s+2}, \ldots$ to $\widehat{b}_s$. Suppose that the following change occurs during the $(m-1)$-th interval. Let β_m and β_s denote the expected reward vectors during m-th and s-th intervals, respectively. By the same argument as the first occurrence of an event in an i.i.d. random sequence, we have

$$L(T) \le 1 \Big/ P\left(\left\|\widehat{b}_m - \widehat{b}_s\right\|_\infty > \epsilon\right). \qquad (6)$$

Observe that, for every $i = 1, \ldots, n$,

$$\begin{aligned} &P\left(\left\|\widehat{b}_m - \widehat{b}_s\right\|_\infty > \epsilon\right) \\ &\ge P\left(\left|\widehat{b}_m(i) - \widehat{b}_s(i)\right| > \epsilon\right) \\ &\ge P\left(\left|\widehat{b}_s(i) - \beta_m(i)\right| > 3\epsilon/2, \left|\widehat{b}_m(i) - \beta_m(i)\right| \le \epsilon/2\right) \\ &= P\left(\left|\widehat{b}_s(i) - \beta_m(i)\right| > \frac{3\epsilon}{2}\right) P\left(\left|\widehat{b}_m(i) - \beta_m(i)\right| \le \frac{\epsilon}{2}\right) \\ &\ge P\left(\left|\widehat{b}_s(i) - \beta_m(i)\right| > \frac{3\epsilon}{2}\right)\left(1 - e^{-\gamma\epsilon^2/2}\right), \qquad (7) \end{aligned}$$

where the equality is due to independence, and the final inequality follows by Hoeffding's Inequality. Next, we bound the first term of (7). Suppose, without loss of generality, that $\beta_s(i) > \beta_m(i)$; let δ denote $\beta_s(i) - \beta_m(i)$; we obtain, for some i:

$$\begin{aligned} &P\left(\left|\widehat{b}_s(i) - \beta_m(i)\right| > 3\epsilon/2\right) \\ &= P\left(\left|\widehat{b}_s(i) - \beta_s(i) + \delta\right| > 3\epsilon/2\right) \\ &= 1 - P\left(\delta - 3\epsilon/2 \le \widehat{b}_s(i) - \beta_s(i) \le \delta + 3\epsilon/2\right) \\ &\ge 1 - P\left(\delta - 3\epsilon/2 \le \widehat{b}_s(i) - \beta_s(i)\right) \\ &\ge 1 - \exp(-2\gamma(\delta - 3\epsilon/2)^2) \ge 1 - \exp(-\gamma\epsilon^2/2), \quad (8) \end{aligned}$$

where the last two inequalities follows from the fact that $\delta = \beta_s(i) - \beta_m(i) > 2\epsilon$ for some i by Assumption 4.2. Hence, (7), (8) and (6) give

$$L(T) \le 2 \big/ \left(1 - \exp\left(-\gamma\epsilon^2/2\right)\right). \qquad (9)$$

Step 5 (Tying up). By combining (3) with (5) and (9), we find that expected regret is at most:

$$\begin{aligned} &\frac{2k\tau}{(1 - \exp(-\gamma\epsilon^2/2))} + k\tau \\ &\quad + \frac{Cn}{\Delta}\Big(k + 2n\exp\left(-2\gamma\epsilon^2\right)(T/\tau + 1)\Big)\log(T), \end{aligned}$$

from which (2) follows by substituting the values $\tau = \lfloor \frac{n}{\ell} \rfloor \cdot \lfloor \frac{\log(T)}{2\epsilon^2} \rfloor$ and $\gamma = \lfloor \frac{\log(T)}{2\epsilon^2} \rfloor$. □

5. Simulations

In this section, we present an empirical comparison of the WMD algorithm with other algorithms for multi-armed bandit problems. As reference, we consider two algorithms based on upper confidence bounds: the UCB1 algorithm of (Auer et al., 2002a), and the Discounted-UCB algorithm of (Kocsis & Szepesvári, 2006; Garivier & Moulines, 2008). For comparison purpose, we employ the UCB1 algorithm as the component $\mathcal{A}$ of the WMD algorithm. The resulting combination is referred to as the WMD-UCB algorithm.

For our setting, we take a bandit with 4 arms, whose rewards are piecewise-stationary with Bernoulli distributions. The sequence of expected rewards $\beta_t(i)$ for each arm A_i is illustrated in Figure 2. The Discounted-UCB algorithm is provided with prior knowledge of the number k of changes to come in the reward sequence, and its parameters are accordingly set to optimal values (Garivier & Moulines, 2008). Neither the UCB1, nor the WMD-UCB algorithm require this prior information. However, the WMD-UCB algorithm has the privilege to query the previous rewards of some of the arms.

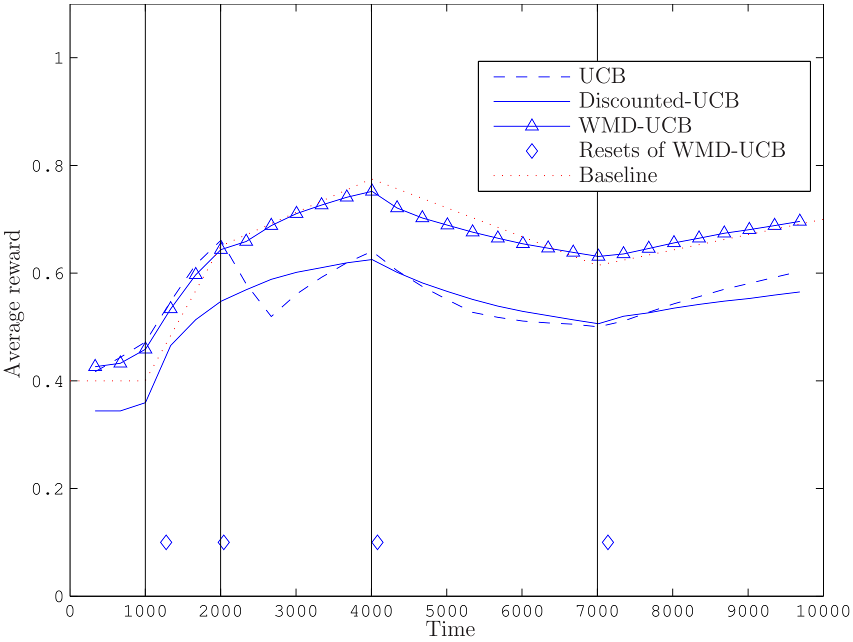

Figure 3. Average expected regret of UCB, Discounted-UCB, and WMD-UCB against the bandit of Figure 2. The WMD-UCB uses the threshold $\epsilon = 0.3$ and makes 1 query per step. Changes in the reward sequence distribution are indicated by vertical lines, whereas instants at which the WMD-UCB algorithm resets are indicated by diamonds. The baseline of our notion of regret (1) is also plotted.

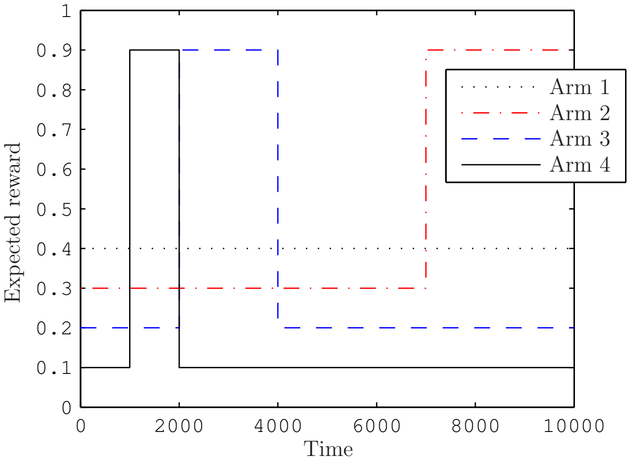

Figure 2. Expected reward of the arms of a 4-armed bandit.

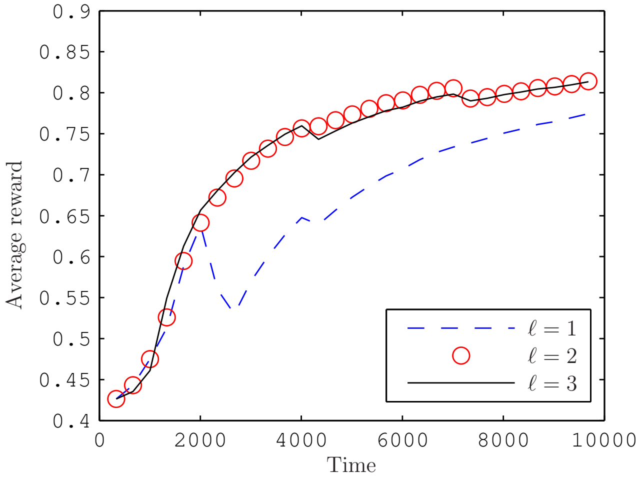

Figure 4. Average expected regret of the WMD-UCB algorithm with 1, 2, and 3 queries per step against a 4-armed bandit. The threshold parameter ϵ is 0.2.

Figure 3 shows the evolution of the average reward of the three algorithms. In this experiment, the WMD-UCB algorithm queries only one arm per step; its average reward is close to optimal with respect to the baseline of (1). Figure 4 illustrates the benefit of increasing the number of queries per step of the WMD-UCB algorithm (with the interval length τ held fixed).

6. Discussions

The WMD algorithm uses a very simple scheme to detect changes in the mean. In its place, we may employ more sophisticated change-detection schemes, *e.g.*, CUSUM (Page, 1954) and the Shiryayev-Roberts rule (Shiryayev, 1963). Modifications are nonetheless required to make them applicable to our problem: the reward distributions must be parametrized; and the pre-change distribution is unknown and must be estimated (cf. Mei, 2006). There also exist schemes that detect changes when the reward process follows one of many Markovian processes (Fuh, 2004), as is the case for restless bandit problems. Despite the drawback of complexity, these schemes detect changes with optimal delay, and do not require prior knowledge of the parameter ϵ of Assumption 4.2. Yet, they also incur a regret of the order of $\log(T)$ due to an inevitable logarithmic delay to detection (Lorden, 1971; Pollak, 1987). This provides, in our model, a lower-bound on the regret of $\Omega(k \log(T))$ for every algorithm that detect the unknown changes and react thereafter.

The side information obtained through queries can be applied to two purposes: detecting changes and improving the performance of the multi-armed bandit algorithm of Assumption 4.1. In this paper, the queries serve only the purpose of change detection. Because of the aforementioned lower regret-bound intrinsic to change detection schemes, we have neglected the question of accelerating the exploration of the sub-algorithm of Assumption 4.1. The action elimination method of (Even-Dar et al., 2006) presents another possible improvement to the sub-algorithm of Assumption 4.1. As a further improvement to the detection component of the WMD algorithm, it is sufficient, when the distribution changes are not adversarial, to limit detections to changes where the current best arm is no longer the best. Finally, it would be interesting to consider different models of querying for side information. For instance, the case when queries may succeed or fail according to an i.i.d. random sequence, or the case where the agent queries two arms and then picks the best of the two.

Acknowledgments

This research was funded by the NSERC Postgraduate Graduate Scholarship, the FQRNT Doctoral Research Scholarship, the ISF under grant 890015, and the Horev Fellowship. We thank the anonymous reviewers for their comments.

References

Akakpo, N. (2008). Detecting change-points in a discrete distribution via model selection. Preprint. http://arxiv.org/abs/0801.0970.

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

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

Cesa-Bianchi, N., & Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge University Press.

Even-Dar, E., Mannor, S., & Mansour, Y. (2006). Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *J. Mach. Learn. Res.*, *7*, 1079–1105.

Fuh, C. D. (2004). Asymptotic operating characteristics of an optimal change point detection in hidden Markov models. *Ann. Statist.*, 2305–2339.

Garivier, A., & Moulines, E. (2008). On upper-confidence bound policies for non-stationary bandit problems. Preprint. http://arxiv.org/abs/0805.3415.

Hartland, C., Gelly, S., Baskiotis, N., Teytaud, O., & Sebag, M. (2006). Multi-armed bandit, dynamic environments and meta-bandits. Preprint. http://hal.archives-ouvertes.fr/hal-00113668/en/.

Herbster, M., & Warmuth, M. K. (1998). Tracking the best expert. *Machine Learning*, *32*, 151–178.

Kocsis, L., & Szepesvári, C. (2006). Discounted-UCB. 2nd PASCAL Challenges Workshop.

Lai, T. L. (2001). Sequential analysis: some classical problems and new challenges. *Statistica Sinica*, *11*, 303–408.

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

Littlestone, N., & Warmuth, M. (1994). The weighted majority algorithm. *Information and Computation*, *108*, 212–261.

Lorden, G. (1971). Procedures for reacting to a change in distribution. *Ann. Math. Statist.*, *42*, 1897–1908.

Mei, Y. J. (2006). Sequential change-point detection when unknown parameters are present in the pre-change distribution. *Ann. Statist.*, *34*, 92–122.

Page, E. S. (1954). Continuous inspection scheme. *Biometrika*, *41*, 100–115.

Pollak, M. (1987). Average run lengths of an optimal method of detecting a change in distribution. *Ann. Statist.*, *15*, 749–779.

Shiryayev, A. N. (1963). On optimum methods in quickest detection problems. *Theory Probab. Appl.*, *8*, 22–46.