

---

# Piecewise-stationary Bandit Problems with Side Observations

---

Jia Yuan Yu and Shie Mannor\*

Department Electrical and Computer Engineering, McGill University, Montréal, Québec, Canada.  
jia.yu@mcgill.ca; shie.mannor@mcgill.ca

## 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  $\Omega(\sqrt{T})$ . We also show that our bound is tight for a natural class of algorithms. An earlier version of this work appears in [YM09].

## 1 Introduction

In stochastic multi-armed bandit problem [LR85], rewards have fixed distributions, but the agent obtains feedback only for the chosen arms. In adversarial expert problems [LW94], rewards are arbitrary individual sequences, but the agent obtains feedback on every expert's performance. In these problems, feedback have distinct roles. For the bandit problem, feedback is used to provide confidence bounds on the arm rewards. For the expert problem, feedback is used to compute a Hannan-consistent policy.

We consider a bandit model that combines aspects from both the stochastic bandit and adversarial expert problem. 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 [HGB<sup>+</sup>06, GM08], or that of the multiple change-point detection problem of [Aka08]. In our variant of the bandit problem, the feedback is not as limited as the stochastic bandit problem, nor as extensive as the adversarial expert problem: we give the agent the benefit of querying and observing some of the past outcomes of arms that have not been picked. In our setting, the feedback serves both the purpose

of detecting changes in reward distribution and of computing confidence bounds. By assigning negative rewards as query costs, it is also possible to quantify the optimize the desired amount of feedback. 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 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 if 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 [CBL06].*

## 2 Setting

We consider the following sequential decision problem. Let  $\{A_1, \dots, A_n\}$  denote the  $n$  arms of a multi-armed bandit—or  $n$  experts of an online learning problem. Let  $b_1, b_2, \dots$  be a sequence of reward vectors in  $\mathbb{R}^n$ . The element  $b_t(i)$  of  $b_t$ , for  $i = 1, \dots, n$  and  $t = 1, 2, \dots$ , represents the reward

---

\*S. Mannor is also with the Department of Electrical Engineering, Technion, Haifa, Israel. shie@ee.technion.ac.il

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  $b_1, b_2, \dots$  is an independent sequence of random variables that undergoes abrupt changes in distribution at unknown time instants  $\nu_1, \nu_2, \dots$ , 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,  $b_{\nu_1}, \dots, b_{\nu_2-1}$  are i.i.d. with common distribution  $f_{\nu_1}$ , as is the case for stochastic learning problems (cf. [LR85]). Likewise,  $b_{\nu_j}, b_{\nu_j+1}, \dots, b_{\nu_{j+1}-1}$  are i.i.d. with distribution  $f_{\nu_j}$ , for  $j = 1, 2, \dots$ . The intervals are illustrated as follows:

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

Similarly to adversarial learning problems (cf. [CBL06]), both the change-points  $\nu_1, \nu_2, \dots$  and the distributions  $f_{\nu_1}, f_{\nu_2}, \dots$  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 oblivious 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

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

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  $a_t \in \{A_1, \dots, A_n\}$  and makes  $\ell$  (where  $1 \leq \ell \leq n$ ) observations on the individual arm-rewards  $b_{t-1}(1), \dots, b_{t-1}(n)$  of the previous step. This is captured in the following assumption.

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

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

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  $a_t \in \{A_1, \dots, A_n\}$  and receives the corresponding reward  $b_t(a_t)$ . Let  $\beta_t$  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

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

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

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

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  $b_1, b_2, \dots$ , we define the expected regret of the agent at time  $T$  as

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

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

## 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 [LR85, ACBF02], the reward sequence  $b_1, b_2, \dots$  is a sequence of i.i.d. random vectors from a common unknown distribution  $\beta_1 = \beta_2 = \dots$ . The reward observations are limited to rewards  $b_1(a_1), b_2(a_2), \dots$  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, a number of algorithms, e.g., [LR85, ACBF02, KS06], achieve the optimal expected regret of the order of  $O(n \log(T)/\Delta^2)$ , where  $\Delta$  denotes the difference in mean between the best and second-best arms.

### 3.2 Adversarial Expert Problem

Many learning problems take the adversarial setting, e.g., prediction with expert advice, etc.—see [CBL06] 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  $b_1, b_2, \dots$ . This approach essentially makes provisions for the worst-case sequence of reward. At time  $t$ , the past reward vectors  $b_1, \dots, b_{t-1}$  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.*,  $\max_{i=1,\dots,n} \sum_{t=1}^T b_t(i)$ . For every sequence  $b_1, b_2, \dots$ , the (expected) adversarial regret

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

is of the order of  $O(\sqrt{T \log(n)})$ —see [CBL06] for a detailed account. A bound of  $O(\sqrt{T n \log(n)})$  holds when the observations are limited to the chosen arms:  $b_1(a_1), b_2(a_2), \dots$  [ACBFS02].

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 [HW98]. The fixed-share algorithm or one of its variants [HW98, ACBFS02] can be applied to our setting, if the number of changes  $k$  is given in advance, yielding a regret of  $O(\sqrt{nkT \log(T)})$  [Aue02]. We present an algorithm with a regret of  $O(nk \log(T))$  without prior knowledge of  $k$ . It should be noted that when  $k$  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 [HGB<sup>+</sup>06, GM08]. 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  $b_1(a_1), b_2(a_2), \dots$ . 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  $k$  up to time  $T$ , Garivier and Moulines provide upper-confidence bound algorithms that achieve a regret of  $O(n\sqrt{kT} \log(T)/\Delta^2)$ , and show a lower-bound of  $\Omega(\sqrt{T})$  for the regret.

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 current reward of one or more arms. We show that with  $T$  queries in total, the regret is bounded by  $O(nk \log(T)/\Delta^2)$ . Hence, queries reduce significantly the regret with respect to the results of [GM08].

## 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  $b_1, b_2, \dots$  have

mean  $\beta$ . Let  $A_{i^{(1)}}$  and  $A_{i^{(2)}}$  denote, respectively, the arm with the highest and second-highest mean. Let  $\Delta$  denote their mean difference:  $\Delta = \beta(i^{(1)}) - \beta(i^{(2)})$ . Let  $\mathcal{A}$  be an algorithm that guarantees a regret of at most  $Cn \log(T)/\Delta^2$ , 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 [LR85, ACBF02].

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 means of the arm  $A_i$  at the change-point  $\nu_{j+1}$ . There exists a known value  $\epsilon > 0$  such that, for each  $j = 1, 2, \dots$ , there exists an arm  $A_i$  such that  $|\beta_{\nu_j}(i) - \beta_{\nu_{j+1}}(i)| > 2\epsilon$ .

### 4.1 The WMD Algorithm

Our algorithm (Table 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 [Lai01] for a survey). The algorithm partitions the time horizon into intervals of equal length  $\tau$ . Hence, for  $m = 1, 2, \dots$ , the  $m$ -th interval is comprised of the time instants  $(m-1)\tau + 1, (m-1)\tau + 2, \dots, m\tau$ . The algorithm computes iteratively empirical mean vectors  $\hat{b}_1, \hat{b}_2, \dots$  over intervals (windows) of length  $\tau$ , in the following fashion:

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

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. The proof appears in [YM09].

**Theorem 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  $\epsilon$  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  $\nu_1, \nu_2, \dots$  and every choice of post-change distributions  $f_{\nu_1}, f_{\nu_2}, \dots$ , the expected regret is bounded as follows:

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

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  $k$ .

**Remark 3 (Query-regret trade-off)** The bound of Theorem 1 indicates a way to trade-off the number of queries  $\ell$  per step and the expected regret per step. Suppose that an increasing

Input: interval length  $\tau > 0$ , threshold  $\epsilon > 0$ , and  $\ell$  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  $t \in [(m-1)\tau + 2, \dots, m\tau]$ , 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  $\{A_1, \dots, A_n\}$  according to  $\Sigma_{t-1}(1), \dots, \Sigma_{t-1}(n)$ . Query the set  $S_t$  of arms that received the fewest queries. Update the following elements of the empirical mean  $\hat{b}_m$ :

$$\hat{b}_m(i) := \frac{\Sigma_{t-1}(i) \hat{b}_m(i) + b_{t-1}(i)}{\Sigma_{t-1}(i) + 1}, \quad \text{for every } i \in S_t.$$

3. (Detect change.) At the start of the  $m$ -th interval, *i.e.*, if  $t = (m-1)\tau + 1$  for some  $m = 3, 4, \dots$ . If  $\|\hat{b}_m - \hat{b}_r\|_\infty > \epsilon$ , 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.

Algorithm 1: Windowed mean-shift detection (WMD) algorithm

function  $C_Q$  assigns a cost, in the same unit as the rewards and the regret, to the rate of queries  $\ell$ . The corresponding new objective thus becomes the sum of two components: 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  $\ell$ , this cost can be optimized with respect to  $\ell$ . 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.

**Remark 4** 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 [Pag54] and the Shiriyayev-Roberts rule [Shi63]. 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.* [Mei06]). There also exist schemes that detect changes when the reward process follows one of many Markovian processes [Fuh04], 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  $\epsilon$  of Assumption 4.2. Yet, they also incur a regret of the order of  $\log(T)$  due to an inevitable logarithmic delay to detection [Lor71]. 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.

## References

- [ACBF02] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.
- [ACBFS02] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Computing*, 32(1):48–77, 2002.
- [Aka08] N. Akakpo. Detecting change-points in a discrete distribution via model selection. Preprint, 2008.
- [Aue02] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *JMLR*, 3:397–422, 2002.
- [CBL06] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.
- [Fuh04] C. D. Fuh. Asymptotic operating characteristics of an optimal change point detection in hidden Markov models. *Ann. Statist.*, pages 2305–2339, 2004.
- [GM08] A. Garivier and E. Moulines. On upper-confidence bound policies for non-stationary bandit problems. Preprint. <http://arxiv.org/abs/0805.3415>, 2008.
- [HGB<sup>+</sup>06] C. Hartland, S. Gelly, N. Baskiotis, O. Teytaud, and M. Sebag. Multi-armed bandit, dynamic environments and meta-bandits. Preprint. <http://hal.archives-ouvertes.fr/hal-00113668/en/>, 2006.
- [HW98] M. Herbster and M. K. Warmuth. Tracking the best expert. *Machine Learning*, 32(2):151–178, 1998.
- [KS06] L. Kocsis and C. Szepesvári. Discounted-UCB. 2nd PASCAL Challenges Workshop, 2006.
- [Lai01] T. L. Lai. Sequential analysis: some classical problems and new challenges. *Statistica Sinica*, 11:303–408, 2001.
- [Lor71] G. Lorden. Procedures for reacting to a change in distribution. *Ann. Math. Statist.*, 42:1897–1908, 1971.
- [LR85] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.
- [LW94] N. Littlestone and M. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.
- [Mei06] Y. J. Mei. Sequential change-point detection when unknown parameters are present in the pre-change distribution. *Ann. Statist.*, 34(1):92–122, 2006.
- [Pag54] E. S. Page. Continuous inspection scheme. *Biometrika*, 41:100–115, 1954.
- [Shi63] A. N. Shiriyayev. On optimum methods in quickest detection problems. *Theory Probab. Appl.*, 8:22–46, 1963.
- [YM09] J. Y. Yu and S. Mannor. Piecewise-stationary bandit problems with side observations. In *Proceedings of ICML*, 2009.