Adaptively Tracking the Best Arm with an Unknown Number of Distribution Changes

Peter Auer[†] AUER@UNILEOBEN.AC.AT

Pratik Gajane[†] PRATIK.GAJANE@UNILEOBEN.AC.AT

Ronald Ortner[†] RONALD.ORTNER@UNILEOBEN.AC.AT

[†] *Montanuniversität Leoben*
Erzherzog Johann-Straße 3
8700 Leoben, Austria

Abstract

We consider the variant of the stochastic multi-armed bandit problem where the stochastic reward distributions may change abruptly several times. In contrast to previous work, we are able to achieve (nearly) optimal mini-max regret bounds *without knowing the number of changes.*

Keywords: multi-armed bandits, non-stationary rewards, switching bandits

1. Introduction

The classical multi-armed bandit (MAB) is the simplest problem setting that gives rise to the exploration-exploitation dilemma inherent to all reinforcement learning problems (see Bubeck and Cesa-Bianchi, 2012, for a survey). In this setup, a learner has access to a number of available actions, also called "arms" in reference to the arm of a slot machine or a one-armed bandit. The learner has to repeatedly select one of these arms, which yields a reward generated from the unknown reward process of the selected arm. The learner's aim is to maximize the sum of the gathered rewards. In the usual stochastic MAB problem, the reward process for an arm is assumed to be a distribution which remains stationary. In this article, though, we consider the stochastic MAB problem with non-stationary reward distributions. Following Garivier and Moulines (2011), we call this the *switching bandits* problem. As a motivation, consider the problem of real-time content optimization of websites which aims to serve targeted and relevant content to individuals. In order to do so, the website needs to learn which content (represented by an arm of the MAB) the users are most likely to be interested in. The user interest in the content of a website (for example, news) is likely to vary over time. For additional motivating examples and practical applications of this problem setting, see Garivier and Moulines (2011), Hartland et al. (2006), Koulouriotis and Xanthopoulos (2008), and the references therein.

1.1 Related work

Non-stationary MAB problems where reward distributions vary over time have been previously studied in the literature, see e.g. Kocsis and Szepesvári (2006), Hartland et al. (2006), Slivkins and Upfal (2008), Yu and Mannor (2009). Besbes et al. (2014) consider the same problem and provide an algorithm with an upper bound on the regret depending on the variation in the reward distributions.

Garivier and Moulines (2011) consider the same setting as this article and provide two algorithms: *discounted* UCB and *sliding-window* UCB. They achieve a regret bound of $\tilde{O}\left(\sqrt{LT}\right)$ where L is the number of changes and T is the total number of steps. This is mini-max optimal up to $\log T$ factors, which are hidden in the $\tilde{O}$- notation. Allesiardo et al. (2017) provide an algorithm called SER4, whose regret is also bounded by $O(\sqrt{LT})$. To achieve the given regret bounds, all these algorithms need to be tuned with the knowledge of the number of changes L. Without knowing L, the best previous bounds on the regret are either of order $L\sqrt{T}$ or T^b with $b > 1/2$ (Besbes et al., 2014).

1.2 Outline

The article is structured as follows. In Section 2, we formally define the problem setting at hand and introduce some relevant notation. Section 3 presents our algorithm and the respective regret bound, followed by a sketch of the regret analysis in Section 4. The final Section 5 discusses future directions.

2. Problem setting

For this article, we consider only bandit problems with two arms. We expect that our approach can be extended to any number of arms without much difficulty. We also assume that the horizon T is known. An unknown horizon can be handled by the doubling trick (Besson and Kaufmann, 2018).

At each time $t = 1, 2, \ldots, T$, a bandit algorithm $\mathfrak{A}$ selects an arm $a_t \in \{1, 2\}$ and, consequently, receives a reward $r_t \sim D_t(a_t)$ distributed according to some unknown distributions $D_t(a)$ depending on time t and arm a. We assume that the support of all these distributions is $[0, 1]$. All reward distributions are selected by an oblivious adversary in advance before the first step of the algorithm. The reward distributions may change for an unknown number of times L,

$$\begin{aligned} &\mathcal{L} = \{1 < t \leq T \mid \exists a : D_{t-1}(a) \neq D_t(a)\}, \\ &L = |\mathcal{L}|. \end{aligned}$$

We call the steps in $\mathcal{L}$ when a reward distribution changes, change points. These are abrupt changes: the reward distributions remain constant during certain periods and change only at change points.

Let $\mu_t(a)$ be the mean of distribution $D_t(a)$. Then the expected regret of an algorithm $\mathfrak{A}$ is given by

$$R_{\mathfrak{A}} = \sum_{t=1}^{T} \max_a \mu_t(a) - \mathbb{E}\left[\sum_{t=1}^{T} \mu_t(a_t)\right].$$

The expectation is over the arms selected by the algorithm, depending on the observed rewards. The goal of the algorithm is to minimize the regret. Note that our regret formulation considers regret against an optimal non-stationary policy. This notion of regret is similar to the "regret against arbitrary strategies" introduced by Auer et al. (2003) for the non-stochastic bandit problem.

3. The adaptive switching algorithm AdSwitch

Our algorithm AdSwitch (shown as Algorithm 1) proceeds in episodes, where a new episode starts when the algorithm detects a change in one of the arms. Each episode starts with an estimation phase (Step 2), in which both arms are chosen alternatingly until their means can be distinguished (Step 3). After the estimation phase, the algorithm exploits the empirical best arm (Step 5). However, in order to detect changes, at each step with some probability (Step 4) an exploration phase is started that checks whether a change has occurred (comparing the resulting empirical means to that obtained from the estimation phase). During such exploration phases the algorithm checks for changes of certain magnitudes (denoted by d_i). Note that our algorithm does not require knowledge of the number of changes L. The following theorem provides an upper bound on the expected regret of our algorithm.

Theorem 1 *For a switching bandit problem with two arms and L changes, the expected regret of* AdSwitch *with sufficiently large C_1 and C_2 is upper-bounded by*

$$C(\log T)\sqrt{(L+1)T}$$

for a suitable constant C.

4. Sketch of the analysis

First, we observe that by choosing the parameter C_2 sufficiently large, the probability that at least one false detection of a change occurs in Step 4 is bounded by $1/T$ (using Hoeffing's inequality and a union bound over the involved time steps). Thus we assume that there are no false change detections and that during the execution of the algorithm we have

$$k \leq L.$$

We decompose the expected regret $R_{\mathfrak{A}}$ as

$$R_{\mathfrak{A}} \leq R_1 + R_2 + R_3,$$

where R_1 is the regret during the estimation phase, R_2 is the regret due to sampling in Step 4 (under the assumption that the estimated means are close to their true means, cf. Section 4.2), and R_3 is the regret incurred in the exploitation phase due to changes.

4.1 Regret in the estimation phase

Let $R_{k,1}$ be the regret during the estimation phase of episode k. Let $\sigma_1 < \sigma_2$ be any two consecutive change points in the estimation phase with the understanding that σ_1 might be

Algorithm 1 AdSwitch

```
   Input: Time horizon $T$
   Parameters: $C_1, C_2 > 0$
1: Initialize $k = 0$.
   For each episode $k$, let $\tau_k^0$ be the time when episode $k$ starts.
   Estimation of $\hat{\Delta}_k$:
2: Sample both arms alternatingly until the condition of Step 3 is met. Let $\hat{\mu}_a[t_1, t_2]$ be the empirical mean for arm $a$ for samples obtained from times $t \in [t_1, t_2)$.
3: If at time $t$ there is a $\sigma$, $\tau_k^0 \leq \sigma < t$, with
       $\big|\hat{\mu}_1[\sigma, t] - \hat{\mu}_2[\sigma, t]\big| > \sqrt{\frac{C_1 \log T}{t - \sigma}}$,
   then set $\hat{\mu}_{k,a} = \hat{\mu}_a[\sigma, t]$, $\hat{\Delta}_k = |\hat{\mu}_{k,1} - \hat{\mu}_{k,2}|$, $\bar{a}_k = \arg\max_a \hat{\mu}_{k,a}$, $\underline{a}_k = \arg\min_a \hat{\mu}_{k,a}$, $\tau_k = t$, and proceed with Step 4.
   Exploitation and checking:
4: (Checking) At the current time step $t$, check for changes of the arms with some probability. That is, let $d_i = 2^{-i}$ and $I_k = \max\{i : d_i \geq \hat{\Delta}_k\}$. Then for any $i$ from $\{1, 2, \ldots, I_k\}$ with probability $p_{k,i} = d_i \sqrt{\frac{k+1}{T}}$, sample both arms alternatingly for $s_i = 2\left\lceil \frac{C_2 \log T}{d_i^2} \right\rceil$ steps to check for changes of size $d_i$. If for any arm $a$,
       $\hat{\mu}_{\bar{a}_k}[t, t+s] - \hat{\mu}_{\underline{a}_k}[t, t+s] \notin \left[\hat{\Delta}_k - \frac{d_i}{4}, \hat{\Delta}_k + \frac{d_i}{4}\right]$,
   then set $k \leftarrow k + 1$, and start a new episode at Step 2.
5: (Exploitation) If no checking is performed at current time step $t$, then select arm $\bar{a}_k$ and repeat Step 4.
```

equal to the start of the phase (and thus not necessarily a change point), τ_k^0, and σ_2 might be equal to the end of the phase τ_k (and hence also not necessarily a change point).

We consider the regret during the steps $t \in [\sigma_1, \sigma_2)$. Let $\Delta = |\mu_{\sigma_1}(1) - \mu_{\sigma_1}(2)|$ be the true gap during these steps. By Hoeffding's inequality we have with probability $1 - 1/T^3$ that

$$\big|\hat{\mu}_1[\sigma_1, \sigma_2] - \hat{\mu}_2[\sigma_1, \sigma_2]\big| > \Delta - \sqrt{\frac{3 \log T}{\sigma_2 - \sigma_1}}.$$

Thus Step 3 implies that

$$\sigma_2 - \sigma_1 \leq \frac{\max\{12, 4C_1\} \log T}{\Delta^2},$$

and the regret during these steps is at most

$$\min\left\{\frac{\max\{12, 4C_1\} \log T}{\Delta}, \Delta(\sigma_2 - \sigma_1)\right\} \leq \sqrt{\max\{12, 4C_1\}(\sigma_2 - \sigma_1) \log T}.$$

Summing over all such change point intervals in the estimation phases we get

$$\sum_k R_{k,1} \leq \sqrt{\max\{12, 4C_1\} T (2L + 1) \log T},$$

since the sum of the length of those intervals is at most T, and there are at most $2L+1$ such intervals.

4.2 Regret due to sampling in exploitation phase

Next, we bound the regret due to sampling in Step 4 of the algorithm when checking for changes. However, we only consider those steps in which our estimates are close to the true means. That is, let

$$S_k = \{\tau_k \leq t < \tau_{k+1}^0 : \max_a |\hat{\mu}_{k,a} - \mu_t(a)| \leq \hat{\Delta}_k/4\}$$

be the time steps in episode k for which the estimated means $\hat{\mu}_{k,a}$ are close to the true means. Then we define $R_{k,2}$ to be the regret in episode k due to sampling in Step 4 of the algorithm for time steps in S_k. For $t \in S_k$, we have

$$\mu_t(\bar{a}_k) - \mu_t(\underline{a}_k) \leq (\hat{\mu}_{k,\bar{a}_k} + \hat{\Delta}_k/4) - (\hat{\mu}_{k,\underline{a}_k} - \hat{\Delta}_k/4) = 3\hat{\Delta}_k/2.$$

Therefore, using the definitions of s_i, $p_{k,i}$, d_i, and I_k as given in Step 4 of the algorithm, the expected regret — conditioned on the past before step τ_k — caused by sampling in these time steps can be bounded by

$$\begin{aligned}
R_{k,2} &\leq \frac{3}{2}\hat{\Delta}_k |S_k| \sum_{i=1}^{I_k} p_{k,i} s_i \\
&\leq 6\hat{\Delta}_k |S_k| \sum_{i=1}^{I_k} p_{k,i} \frac{C_2 \log T}{d_i^2} \\
&\leq 6C_2(\log T)\hat{\Delta}_k |S_k| \sqrt{\frac{k+1}{T}} \sum_{i=1}^{I_k} \frac{1}{d_i} \\
&\leq 6C_2(\log T)\hat{\Delta}_k |S_k| \sqrt{\frac{k+1}{T}} \cdot 2^{I_k+1} \\
&\leq 6C_2(\log T)\hat{\Delta}_k |S_k| \sqrt{\frac{k+1}{T}} \cdot \frac{2}{\hat{\Delta}_k} \\
&\leq 12C_2(\log T)|S_k| \sqrt{\frac{k+1}{T}}.
\end{aligned}$$

Summing over k gives

$$R_2 = \sum_k R_{k,2} \;\leq\; 12C_2(\log T)\sqrt{\frac{L+1}{T}} \sum_k |S_k| \;\leq\; 12C_2(\log T)\sqrt{(L+1)T}.$$

4.3 Regret in the exploitation phase due to changes

Using the notation of the previous section, it remains to bound the regret $R_{k,3}$ in the time steps of each episode k that are not in S_k. We partition these time steps into intervals

$[\alpha_1, \beta_1), [\alpha_2, \beta_2), \dots$ of consecutive time steps with maximal size and no changes:

$$\begin{aligned}
\alpha_1 &= \min\{\alpha \geq \tau_k | \max_a |\mu_\alpha(a) - \hat{\mu}_{k,a}| > \hat{\Delta}_k/4\}, \\
\beta_j &= \min\{\beta > \alpha_j | \exists a : \mu_\beta(a) \neq \mu_{\alpha_j}(a)\}, \\
\alpha_{j+1} &= \min\{\alpha \geq \beta_j | \max_a |\mu_\alpha(a) - \hat{\mu}_{k,a}| > \hat{\Delta}_k/4\}.
\end{aligned}$$

Let $\epsilon_j = \max_a |\mu_{\alpha_j}(a) - \hat{\mu}_{k,a}|$ be the deviation of the true means from the estimated means. Let i_j be such that

$$d_{i_j} < \epsilon_j \leq 2d_{i_j}.$$

Since we are dealing with time steps not in S_k, the deviation of estimated means from their true means is necessarily greater than $\hat{\Delta}_k/4$. Combining this with the previous inequality, we get

$$d_{i_j} \geq \epsilon_j/2 > \hat{\Delta}_k/8.$$

The regret for a time step t in such an interval is at most

$$|\mu_t(1) - \mu_t(2)| \leq \hat{\Delta}_k + 2\epsilon_j \leq \hat{\Delta}_k + 4d_{i_j} \leq 12d_{i_j}.$$

If $\beta_j - \alpha_j \leq 2s_{i_j}$, then $d_{i_j} \leq \sqrt{\frac{8C_2 \log T}{\beta_j - \alpha_j}}$ and the regret in interval $[\alpha_j, \beta_j)$ is at most

$$12d_{i_j}(\beta_j - \alpha_j) \leq 36\sqrt{C_2(\log T)(\beta_j - \alpha_j)}.$$

Summing over such intervals contributes at most $36\sqrt{C_2(\log T)LT}$ to the overall regret. We now consider the contribution to the regret if $\beta_j - \alpha_j > 2s_{i_j}$. To simplify notation, we assume in the following that $\beta_j - \alpha_j > 2s_{i_j}$ for all $j = 1, 2, \dots$. The expected regret over all these intervals is given by the sum of the regret terms within the intervals multiplied by the probability, that no change is detected in any of the previous intervals. Let R^k_j be the expected regret over the intervals $[\alpha_j, \beta_j), [\alpha_{j+1}, \beta_{j+1}), \dots$, given that no change is detected before α_j. In the following, we prove by induction that

$$R^k_j \leq 12\sqrt{\frac{T}{k+1}} + \sum_{j' \geq j} 48\sqrt{C_2(\log T)(\beta_{j'} - \alpha_{j'})}.$$

By Step 4 of the algorithm, the change of means in interval $[\alpha_j, \beta_j)$ will be detected with high probability, if a check for changes of size d_{i_j} is activated and the remaining size of the interval is at least s_{i_j}. Since the probability that the algorithm does not check for a change of size d_{i_j} during $\beta_j - \alpha_j - s_{i_j}$ steps is given by $\sum_{h=1}^{\beta_j - \alpha_j - s_{i_j}} (1 - p_{k,i_j})^h$, we have

$$\begin{aligned}
R^k_j &\leq 12d_{i_j} \left[\sum_{h=1}^{\beta_j - \alpha_j - s_{i_j}} (1 - p_{k,i_j})^h + s_{i_j} \right] + (1 - p_{k,i_j})^{\beta_j - \alpha_j - s_{i_j}} R^k_{j+1} \\
&\leq 12d_{i_j} \left[\frac{1 - (1 - p_{k,i_j})^{\beta_j - \alpha_j - s_{i_j}}}{p_{k,i_j}} + s_{i_j} \right] + (1 - p_{k,i_j})^{\beta_j - \alpha_j - s_{i_j}} R^k_{j+1}
\end{aligned}$$

$$
\begin{aligned}
&\leq 12\frac{d_{i_j}}{p_{k,i_j}}\left[1-(1-p_{k,i_j})^{\beta_j-\alpha_j-s_{i_j}}\right]+12d_{i_j}s_{i_j}+(1-p_{k,i_j})^{\beta_j-\alpha_j-s_{i_j}}R^k_{j+1}\\
&\leq 12\sqrt{\frac{T}{k+1}}\left[1-(1-p_{k,i_j})^{\beta_j-\alpha_j-s_{i_j}}\right]+48\frac{C_2\log T}{d_{i_j}}\\
&\quad+12(1-p_{k,i_j})^{\beta_j-\alpha_j-s_{i_j}}\sqrt{\frac{T}{k+1}}+\sum_{j'\geq j+1}48\sqrt{C_2(\log T)(\beta_{j'}-\alpha_{j'})}\\
&\leq 12\sqrt{\frac{T}{k+1}}+\sum_{j'\geq j}48\sqrt{C_2(\log T)(\beta_{j'}-\alpha_{j'})}.
\end{aligned}
$$

Summing over k and using $\sum_{k=1}^{L}\frac{1}{\sqrt{k+1}}=\sum_{k=2}^{L+1}\frac{1}{\sqrt{k}}\leq 2\sqrt{L+1}$ we get

$$
\sum_k R_{k,3}\leq 24\sqrt{(L+1)T}+48\sqrt{C_2(\log T)LT},
$$

which concludes the proof sketch for Theorem 1.

5. Discussion and further directions

We have provided an algorithm which tracks the best of two arms in a multi-armed bandit problem with changing reward distributions. The salient feature of our work is that the algorithm does not need to know the number of changes in advance and still obtains a regret bound of $\tilde{O}\left(\sqrt{LT}\right)$, which is optimal up to logarithmic terms.

Concerning future work, the first obvious step concerns the extension to an arbitrary number of arms, which we expect to pose no major problems. Another direction would be to consider adversarial instead of stochastic bandits. Extending our approach to reinforcement learning in changing Markov decision processes (MDPs) is a more challenging direction to pursue.

Acknowledgments

This work has been supported by the Austrian Science Fund (FWF): I 3437-N33 in the framework of the CHIST-ERA ERA-NET (DELTA project).

References

Robin Allesiardo, Raphaël Féraud, and Odalric-Ambrym Maillard. The non-stationary stochastic multi-armed bandit problem. *International Journal of Data Science and Analytics*, 3(4):267–283, Jun 2017.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2003.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in Neural Information Processing Systems 27, NIPS 2014*, pages 199–207. 2014.

Lilian Besson and Emilie Kaufmann. What doubling tricks can and can't do for multi-armed bandits. *CoRR*, abs/1803.06971, 2018. URL `http://arxiv.org/abs/1803.06971`.

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

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *Proceedings of the 22nd International Conference on Algorithmic Learning Theory*, ALT 2011, pages 174–188. Springer, 2011.

Cédric Hartland, Sylvain Gelly, Nicolas Baskiotis, Olivier Teytaud, and Michèle Sebag. Multi-armed bandit, dynamic environments and meta-bandits. *NIPS-2006 workshop, Online trading between exploration and exploitation*, 2006.

Levente Kocsis and Csaba Szepesvári. Discounted UCB. *2nd PASCAL Challenges Workshop*, 2006.

Dimitris E. Koulouriotis and A.S. Xanthopoulos. Reinforcement learning and evolutionary algorithms for non-stationary multi-armed bandit problems. *Applied Mathematics and Computation*, 196(2):913 – 922, 2008.

Alex Slivkins and Eli Upfal. Adapting to a changing environment: The Brownian restless bandits. In *21st Conference on Learning Theory (COLT)*, 2008.

Jia Yuan Yu and Shie Mannor. Piecewise-stationary bandit problems with side observations. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML 2009, pages 1177–1184, 2009.