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

FWF Der Wissenschaftsfonds.

Peter Auer, Pratik Gajane, Ronald Ortner

{auer,pratik.gajane,ronald.ortner}@unileoben.ac.at
Montanuniversität Leoben

MONTAN UNIVERSITÄT www.unileoben.ac.at CiT

Introduction

- Multi-Armed Bandits (MAB): Simple setting for exploration-exploitation trade-off.
- Classical stochastic MAB: Stationary reward distributions.
- ➜ **Switching Bandits** [1]: **Stochastic MAB with non-stationary reward distributions** .
- Application : Real-time content optimization of websites.

Problem setting

- At each time $t = 1, 2, \ldots, T$, algorithm $\mathfrak{A}$ selects an arm $a_t \in \{1, 2\}$.
- It receives $r_t \sim D_t(a_t)$ with mean $\mu_t(a_t)$.
- Reward distributions D_t may change abruptly at certain times.
- Algorithm has no knowledge about the number of changes L.
- Optimize regret $\mathbf{R}_{\mathfrak{A}} = \sum_{t=1}^{T} \max_{a} \mu_t(a) - \mathbb{E}\left[\sum_{t=1}^{T} \mu_t(a_t)\right]$.

Algorithm AdSwitch (Sketch)

- **Episodic algorithm** with each episode having two phases.
- **Estimation phase**: Both arms are selected alternatingly, until better arm has been identified.
- **Exploitation and checking phase**:
  - Mostly exploit the empirical best arm.
  - Sometimes sample both arms to check for change. If a change is detected then a new episode is started.

Regret Bound

Theorem 1. *When* AdSwitch *is run with with sufficiently large C_1 and C_2, then its regret for a switching bandit problem with two arms and L changes is at most*

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

Discussion and Further Directions

- The best known lower bound is $\Omega\left(\sqrt{LT}\right)$, which holds even when L is given to the algorithm.
- Previously, upper bounds $\tilde{O}\left(\sqrt{LT}\right)$ were known only for algorithms which receive L as input [1, 2].
- Our algorithm can be extended for K arms and achieves $O\left(K(\log T)\sqrt{(L+1)T}\right)$ regret.
- A version of our algorithm achieves variational regret $\tilde{O}\left(V^{1/3}T^{2/3}\right)$ (which is order-optimal [3]) without knowing the variation $V = \sum_t \max_a |\mu_t(a) - \mu_{t-1}(a)|$.
- Future work: Switching adversarial bandits and Markov Decision Processes.

Key references

[1] Aurélien Garivier and Eric Moulines: On upper-confidence bound policies for switching bandit problems, ALT 2011.

[2] Robin Allesiardo, Raphael Féraud and Odalric-Ambrym Maillard: The non-stationary stochastic multi-armed bandit problem, IJDSA 2017.

[3] Omar Besbes, Yonatan Gur, and Assaf Zeevi: Stochastic Multi-Armed-Bandit Problem with Non-stationary Rewards, NIPS 2014.

Algorithm AdSwitch

1: **Input:** Time horizon T

2: **Parameters:** $C_1, C_2 > 0$

3: Initialize $k = 0$

For each episode k, let τ_k^0 be the time when episode k starts.

Estimation of $\hat{\Delta}_k$:

4: Sample both arms alternatingly until the condition of Step 5 is met.

Let $\hat{\mu}_a[t_1, t_2]$ be the empirical mean for arm a for samples obtained from times $t \in [t_1, t_2)$.

5: If at time t there is a σ, $\tau_k^0 \leqslant \sigma < t$, with

$$|\hat{\mu}_1[\sigma, t] - \hat{\mu}_2[\sigma, t]| > \sqrt{\frac{C_1 \log T}{t - \sigma}},$$

then set

$$\hat{\mu}_{k,a} = \hat{\mu}_a[\sigma, t] \qquad \text{and} \qquad \hat{\Delta}_k = |\hat{\mu}_{k,1} - \hat{\mu}_{k,2}|,$$

$$\bar{a}_k = \arg\max_a \hat{\mu}_{k,a} \qquad \text{and} \qquad \underline{a}_k = \arg\min_a \hat{\mu}_{k,a}$$

and proceed with Step 6.

Exploitation and checking:

6: Let $d_i = 2^{-i}$ and $I_k = \max\{i : d_i \geqslant \hat{\Delta}_k\}$.

Randomly choose i from $\{1, 2, \ldots, I_k\}$ with probabilities $\mathbf{p_{k,i} = d_i\sqrt{\frac{k+1}{T}}}$.

With the remaining probability select arm $\bar{a}_k$ and repeat step 6.

If an i is chosen, sample both arms alternatingly for $\mathbf{s_i = 2\left\lceil\frac{C_2 \log T}{d_i^2}\right\rceil}$ steps to check for changes of size d_i: if for any arm a,

$$\hat{\mu}_{\bar{a}_k}[t, t+s_i] - \hat{\mu}_{\underline{a}_k}[t, t+s_i] \notin \left[\hat{\Delta}_k - \frac{d_i}{4}, \hat{\Delta}_k + \frac{d_i}{4}\right],$$

then set $k \leftarrow k + 1$, and start a new episode at Step 4.