Achieving Optimal Dynamic Regret for Non-stationary Bandits without Prior Information

Peter Auer[1] AUER@UNILEOBEN.AC.AT
Yifang Chen[2] YIFANG@USC.EDU
Pratik Gajane[1] PRATIK.GAJANE@UNILEOBEN.AC.AT
Chung-Wei Lee[2] LEECHUNG@USC.EDU
Haipeng Luo[2] HAIPENGL@USC.EDU
Ronald Ortner[1] RONALD.ORTNER@UNILEOBEN.AC.AT
Chen-Yu Wei[2] CHENYU.WEI@USC.EDU
[1]*Montanuniversität Leoben*
[2]*University of Southern California*

Editors: Alina Beygelzimer and Daniel Hsu

[1]Abstract

This joint extended abstract introduces and compares the results of (Auer et al., 2019) and (Chen et al., 2019), both of which resolve the problem of achieving optimal dynamic regret for non-stationary bandits without prior information on the non-stationarity. Specifically, Auer et al. (2019) resolve the problem for the traditional multi-armed bandits setting, while Chen et al. (2019) give a solution for the more general contextual bandits setting. Both works extend the key idea of (Auer et al., 2018) developed for a simpler two-armed setting.

1. Introduction

We consider the classical multi-armed bandit problem (Auer et al., 2002) and its generalization, the contextual bandit problem (Auer et al., 2002; Langford and Zhang, 2008), in non-stationary environments. The learning protocol of the general contextual bandits problem is as follows. Let $\mathcal{X}$ be a context space and $[K] \triangleq \{1, \dots, K\}$ be the set of arms. Ahead of time, the environment decides on T distributions $\mathcal{D}_1, \dots, \mathcal{D}_T$ supported on context-reward pairs $\mathcal{X} \times [0,1]^K$. In round $t = 1, \dots, T$, the environment samples $(x_t, r_t) \sim \mathcal{D}_t$ and reveals x_t to the learner, then the learner selects an arm $a_t \in [K]$ and observes $r_t(a_t)$. For a fixed set of policies Π consisting of mappings from $\mathcal{X}$ to $[K]$, the dynamic regret of the learner is defined as the difference between the total expected reward of *the best sequence of policies* and that of the learner:

$$\text{Reg} = \sum_{t=1}^{T} \max_{\pi \in \Pi} \mathbb{E}_{(x,r)\sim\mathcal{D}_t}[r(\pi(x))] - \sum_{t=1}^{T} r_t(a_t).$$

The classical multi-armed bandit problem is a special case of the contextual bandit problem, where the context is unavailable or ignored by the learner. The set of policies Π consists of the K policies $\pi_1, \dots, \pi_K$ for the arms $a \in [K]$ with $\pi_a(x) = a$ for any context x. In this case we denote

1. Joint extended abstract of (Auer et al., 2019) and (Chen et al., 2019).

by $\mu_t(a) = \mathbb{E}_{(x,r)\sim\mathcal{D}_t}[r(\pi_a(x))]$ the expected reward of arm a at time t, and the dynamic regret simplifies as

$$\text{Reg} = \sum_{t=1}^{T} \max_a \mu_t(a) - \sum_{t=1}^{T} r_t(a_t).$$

We measure the non-stationarity of the environment by the total number of changes S or by the total variation V:

$$S \triangleq \begin{cases} 1 + \sum_{t=2}^{T} \mathbf{1}\left\{\exists a : \mu_t(a) \neq \mu_{t-1}(a)\right\} & \text{for multi-armed bandits (Auer et al., 2019),} \\ 1 + \sum_{t=2}^{T} \mathbf{1}\{\mathcal{D}_t \neq \mathcal{D}_{t-1}\} & \text{for contextual bandits (Chen et al., 2019),} \end{cases}$$

$$V \triangleq \begin{cases} \sum_{t=2}^{T} \max_a |\mu_t(a) - \mu_{t-1}(a)| & \text{for multi-armed bandits (Auer et al., 2019),} \\ \sum_{t=2}^{T} \|\mathcal{D}_t - \mathcal{D}_{t-1}\|_{\text{TV}} & \text{for contextual bandits (Chen et al., 2019).} \end{cases}$$

Our goal is to develop algorithms with dynamic regret that is optimal in terms of the relevant parameters T, K, and S or V, without any prior knowledge about the amount of non-stationarity S or V. Specifically our main results are:[2]

Theorem 1 (Auer et al., 2019) *For multi-armed bandits, there exists an algorithm (*ADSWITCH*) that achieves* $\mathbb{E}[\text{Reg}] = \tilde{\mathcal{O}}\left(\sqrt{KST}\right)$ *without knowing* S.

Theorem 2 (Chen et al., 2019) *For contextual bandits, there exists an algorithm (*ADA-ILTCB$^+$*) that achieves* $\text{Reg} = \tilde{\mathcal{O}}\left(\min\left\{\sqrt{K(\log|\Pi|)ST}, (K(\log|\Pi|)V)^{\frac{1}{3}}T^{\frac{2}{3}} + \sqrt{K(\log|\Pi|)T}\right\}\right)$ *with high probability and without knowing* S *or* V.

A few remarks are in order. First, the analysis of (Chen et al., 2019) reveals that the algorithm of (Auer et al., 2019) in fact also achieves a better regret bound $\tilde{\mathcal{O}}(\min\{\sqrt{KST}, (KV)^{\frac{1}{3}}T^{\frac{2}{3}} + \sqrt{KT}\})$ without any modification. Second, it is well-known that these bounds are optimal up to logarithmic factors by a simple extension of (Auer et al., 2002) and by (Besbes et al., 2014). Third, since multi-armed bandits are a special case of contextual bandits, the result of (Chen et al., 2019) essentially subsumes that of (Auer et al., 2019), except for logarithmic terms and at the cost of much larger constants. Finally, for contextual bandits the algorithm of (Chen et al., 2019) is efficient assuming access to some ERM oracle, a common assumption made in most prior works on efficient contextual bandits; see (Chen et al., 2019) for the formal definition.

Related work. It is well-known how to achieve the same regret bounds having knowledge of S and V. As far as we know, the only prior results with unknown S or V in the bandit setting and with regret bounds $\tilde{\mathcal{O}}(S^\alpha T^{1-\alpha})$ or $\tilde{\mathcal{O}}(V^\alpha T^{1-\alpha})$ for some $\alpha \in (0,1)$ are the following:[3] Karnin

2. $\tilde{\mathcal{O}}(\cdot)$ hides logarithmic dependence on T, K, and also $1/\delta$ in the case of high probability statements where δ is the confidence level.
3. Note that importantly the exponents of S (or V) and T sum up to 1 in these results. Achieving $\tilde{\mathcal{O}}(S^\alpha T^\beta)$ or $\tilde{\mathcal{O}}(V^\alpha T^\beta)$ without knowing S or V but with $\alpha + \beta > 1$ is trivial and is also much weaker since for some values of S and V (that are sublinear in T), the bounds become vacuous. Also note that in this discussion we ignore the dependence on other parameters such as K.

and Anava (2016) developed an algorithm with regret $\tilde{\mathcal{O}}\left(V^{0.82}T^{0.18} + T^{0.77}\right)$ for two-armed non-contextual bandits, the first result of this kind; Luo et al. (2018) developed an oracle-efficient algorithm with regret $\tilde{\mathcal{O}}(\min\{S^{\frac{1}{4}}T^{\frac{3}{4}}, V^{\frac{1}{5}}T^{\frac{4}{5}} + T^{\frac{3}{4}}\})$ for general contextual bandits; the work of Auer et al. (2018), a preliminary version of (Auer et al., 2019), studied a two-armed non-contextual setting and is the first to obtain the optimal bound $\tilde{\mathcal{O}}(\sqrt{ST})$; finally, Cheung et al. (2019) achieved $\tilde{\mathcal{O}}(V^{\frac{1}{3}}T^{\frac{2}{3}} + T^{\frac{3}{4}})$ regret for linear bandits, although one can verify that their Bandit-over-Bandits technique in fact applies to a harder "agnostic" setting and achieves $\tilde{\mathcal{O}}(\sqrt{ST} + T^{\frac{3}{4}})$ for a stronger definition of regret.[4] For a more extensive discussion on other related work, we refer the reader to the related work sections of (Auer et al., 2019) and (Chen et al., 2019).

2. Techniques

In this section we briefly discuss the main techniques to achieve our results and the connections between (Auer et al., 2019) and (Chen et al., 2019). At a high level, both works built on the same key observations from Auer et al. (2018), but the structure of the algorithms is quite different.

2.1. Key Observations

In the traditional stochastic multi-armed bandit problem, the reward distribution does not change over time. Thus, after identifying the sub-optimality of an arm, the learner does not need to draw this arm anymore. However, in non-stationary environments, the learner has to occasionally draw a previously sub-optimal arm, in order to detect whether this arm has changed and has become the best arm. This poses an additional challenge in the exploration-exploitation trade-off: in order to detect changes more accurately, the learner has to draw sub-optimal arms more often, which potentially incurs more regret.

The observation used by Auer et al. (2018) is that for a change of $2^{-\frac{k}{2}}$ in the mean of some arm, the learner only needs $\mathcal{O}\left(2^k\right)$ samples of this arm to confirm the change. Furthermore, for a sub-optimal arm a with an identified optimality gap $\Delta_a = \max_{a'} \mu(a') - \mu(a)$, the learner only needs to make sure not to miss a change of amount larger than Δ_a. This observation provides a solution to the exploration-exploitation trade-off: for a sub-optimal arm with optimality gap Δ_a, the learner will detect whether its mean has changed by more than $2^{-i/2}$, $i = 1, \ldots, \lceil 2\log_2(1/\Delta_a)\rceil$. Detecting such changes requires different amounts of samples: larger changes require fewer samples, while smaller changes require more samples, but at the same time smaller changes incur less regret. Exploiting these observations is the key to achieve the optimal regret.

Another important aspect is that the exploration steps for change detection need to be *consecutive*, as opposed to the common strategy of exploring with some probability independently at each time step. This idea has already been used by Karnin and Anava (2016) and Luo et al. (2018).

In the following two subsections, we compare the exploration schemes of (Auer et al., 2019) and (Chen et al., 2019), both of which utilize the ideas above to ensure a sufficient amount of samples for estimating distribution changes. These exploration schemes are paired with specific *non-stationarity tests* based on standard concentration inequalities to decide whether reward distributions have significantly changed. Details about the tests can be found in the two papers.

4. Specifically, this agnostic setting refers to the one originally studied in (Auer et al., 2002) for the algorithm Exp3.S.

2.2. Sampling Obligation

The algorithm of Auer et al. (2019) maintains for each sub-optimal arm a a set $\mathcal{S}(a)$ of *sampling obligations*, each represented as a pair (k, t).[5] At time t, for a sub-optimal arm a with estimated optimality gap Δ_a, sampling obligations are added at random according to the following rule (for some constants C_1 and C_2):

$$\text{For all } k \geq 1 \text{ with } 2^{-\frac{k}{2}} \geq C_1 \Delta_a, \text{ with probability } C_2 2^{-\frac{k}{2}}, \text{ add } (k, t) \text{ to } \mathcal{S}(a).$$

A sampling obligation (k, t) is removed from $\mathcal{S}(a)$ once arm a has been drawn for 2^k times since time t, or when any non-stationarity is detected and the algorithm restarts. In the light of the discussion in Section 2.1, the role of a sampling obligation (k, t) is to detect a mean change of order $2^{-\frac{k}{2}}$. To serve the sampling obligations, at each time the algorithm selects the arm that has been selected least recently among the arms which are either possibly optimal or have at least one sampling obligation.

2.3. Replay Phase

A naive way to extend the idea of Auer et al. (2018) to the contextual bandits setting is to treat each policy as an arm. However, this leads to a regret bound and computational complexity both polynomial in $|\Pi|$, which is prohibitive. Thanks to prior works such as (Dudík et al., 2011) and (Agarwal et al., 2014), there are statistically and computationally efficient ways to estimate the expected reward of all policies simultaneously. The algorithm of Chen et al. (2019) is built upon the ILOVETOCONBANDITS algorithm of Agarwal et al. (2014).

The original ILOVETOCONBANDITS is designed for stationary environments. It finds a sparse distribution Q_j over Π at time $t = 2^j$, using the data collected from time interval $[1, 2^j - 1]$. For $t = 2^j, \ldots, 2^{j+1} - 1$, the learner samples arms using the following distribution:

$$Q_j^{\nu_j}(a|x_t) \triangleq (1 - K\nu_j) \sum_{\pi \in \Pi : \pi(x_t) = a} Q_j(\pi) + \nu_j, \qquad \text{where } \nu_j = \tilde{\Theta}\left(\sqrt{\frac{1}{K 2^j}}\right).$$

In non-stationary environments, for detection purpose the algorithm of Chen et al. (2019) maintains a set $\mathcal{S}$ of *replay phases*, each of which can again be represented as a pair (k, t). At time $t \in [2^j, 2^{j+1} - 1]$, several replay phases are generated according to the following rule (for some constant C_3):

$$\forall k = 1, \ldots, j-1, \quad \text{with probability } C_3 2^{-\frac{k}{2}}, \text{ add } (k, t) \text{ to } \mathcal{S},$$

which is very similar to the rule of (Auer et al., 2019). A replay phase (k, t) is removed at time $t + 2^k$, or when any non-stationarity is detected and the algorithm restarts. The selection of arms is based on the following rule: if $\mathcal{S}$ is empty, then draw $a_t \sim Q_j^{\nu_j}(\cdot|x_t)$; otherwise, uniformly sample an index k from the set $\{m : \exists(m, \tau) \in \mathcal{S}\}$, and sample $a_t \sim Q_k^{\nu_k}(\cdot|x_t)$. Note that in the latter case the algorithm reuses previous distributions (hence the name replay), and the role of replays is in the same vein as that of sampling obligations used by Auer et al. (2019).

5. The notation/representation here is slightly different from that of (Auer et al., 2019) for conciseness and easier comparison.

Acknowledgments. We thank Dylan Foster, Akshay Krishnamurthy, and Ruihao Zhu for in-depth discussions on the Bandit-over-Bandits approach of (Cheung et al., 2019). The work of (Auer et al., 2019) has been supported by the Austrian Science Fund (FWF): I 3437-N33 in the framework of the CHIST-ERA ERA-NET (DELTA project). The work of (Chen et al., 2019) is supported by NSF Grant #1755781.

References

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert E Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best arm with an unknown number of distribution changes. In *14th European Workshop on Reinforcement Learning*, 2018.

Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *32nd Annual Conference on Learning Theory (COLT)*, 2019.

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

Yifang Chen, Chung-Wei Lee, Haipeng Luo, and Chen-Yu Wei. A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In *32nd Annual Conference on Learning Theory (COLT)*, 2019.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Learning to optimize under non-stationarity. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2019. Full version available at https://ssrn.com/abstract=3261050.

M. Dudík, D. Hsu, S. Kale, N. Karampatziakis, J. Langford, L. Reyzin, and T. Zhang. Efficient optimal learning for contextual bandits. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2011.

Zohar S Karnin and Oren Anava. Multi-armed bandits: Competing with optimal sequences. In *Advances in Neural Information Processing Systems 29*, 2016.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems 21*, 2008.

Haipeng Luo, Chen-Yu Wei, Alekh Agarwal, and John Langford. Efficient contextual bandits in non-stationary worlds. In *31st Annual Conference on Learning Theory (COLT)*, 2018.