Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)
Doctoral Consortium Track

Dynamic Bandits with Temporal Structure

Qinyi Chen*
Massachusetts Institute of Technology
qinyic@mit.edu

Abstract

In this work, we study a dynamic multi-armed bandit (MAB) problem, where the expected reward of each arm evolves over time following an auto-regressive model. We present an algorithm whose per-round regret upper bound almost matches the regret lower bound, and numerically demonstrate its efficacy in adapting to the changing environment.

1 Introduction

The multi-armed bandit (MAB) [Thompson, 1933; Robbins, 1952] is a framework for studying online decision-making under uncertainty. In the basic stochastic formulation, there are k arms, each having a fixed, unknown and independent reward distribution, and the goal is to identify the arm with the highest expected reward while experimenting with the arms over the course of T rounds [Auer *et al.*, 2002b]. Another line of work that is widely studied in the MAB literature is the adversarial setting, which allows the reward distributions to be determined by an *adversary* to change arbitrarily in time [Auer *et al.*, 2002a]. Due to the unpredictability of the adversary, their goal is to approach the performance of the best static benchmark, i.e., the single best arm one can pull in hindsight.

The assumptions in either stochastic or adversarial setting, however, are often deemed too stringent and no longer valid in most real-world applications, such as online advertising, online auctions and dynamic pricing, where the reward distributions in fact evolve with time following some intrinsic temporal structure. Consider, for example, a web search engine in which the goal is to maximize its revenue by showing users ads with the highest click through rate (CTR). The changes of CTR usually scale linearly with T and can be modeled using some temporal structure.

Motivated by this, in our work, we relax the previously restrictive assumptions on the reward distributions by studying a *dynamic bandit problem*, in which the reward of each arm has a temporal structure and the amount of changes is linear in T. To impose temporal structure, we consider a simple yet popular auto-regressive (AR) model, which is one of the most commonly studied time series model and is widely adopted in the prediction of future behavior based on past information. We assume the expected reward of each arm $i \in [k]$, $r_i(t)$, undergoes an independent AR-1 process, defined as follows:

$$r_i(t+1) = \alpha(r_i(t) + \epsilon_i(t)) = \alpha R_i(t)\,,$$

where $r_i(t)$ ($R_i(t)$) is the expected (observed) reward of arm i at round t, $\alpha \in (0,1)$ is the AR parameter that measures temporal correlation over time, and $\epsilon_i(t) \sim N(0,\sigma)$ is an independent noise across arms and rounds.

The setup in our dynamic bandit problem bears some resemblance to the settings in the *restless bandits*, where the reward distribution of each arm changes at each round according to some known stochastic process. Some works in this class, e.g., [Ortner *et al.*, 2012], study the restless MAB with Markovian rewards, where the expected reward of each arm transitions according to a finite-state Markov chain. Nevertheless, they critically assume that the state space is discrete, which no longer holds true when we work with infinite state spaces in the AR setting. There are also works that study restless bandits in non-stationary environments with sublinear amount of changes, e.g., [Besbes *et al.*, 2014]. This assumption again becomes invalid in the AR setting, where the rewards change much more rapidly and the amount of changes scale with T. In comparison to the existing works, the AR temporal structure poses new challenges due to its fast-changing nature. It is thus imperative to design new learning algorithms with strong performance guarantees in this new dynamic setting.

To measure the performance of algorithms in our setting, we use our notion of *per-round steady state dynamic regret*, where the dynamic regret at a given round is the difference between the collected reward and the highest expected reward at that round. The notion of dynamic regret is more demanding than the notion of static regret used in stochastic and adversarial MAB problems. Unlike the static one, it requires the algorithm to adapt well to the changes, since it compares the reward of algorithms against the policy that, at each round, pulls the arm with the highest expected reward.

2 Main Contributions

In our work, we introduce a novel setting of dynamic bandits where the rewards evolve according to an AR-based temporal structure. In addition, we contribute in the following aspects.

Regret lower bound. We characterize the per-round steady state regret lower bound as a function of the stochastic rate of change σ and the temporal correlation of the rewards α. In particular, when α is close to one, we show the per-round regret of any algorithm is at least $O(k\alpha^2\sigma^2)$. Our result shows that in the AR setting, the per-round regret measured against

*This work is a joint work with Negin Golrezaei (Massachusetts Institute of Technology) and Djallel Bouneffouf (IBM Research).

a strong dynamic benchmark does not converge to zero as T increases, implying that any good algorithm should keep exploring over time. This stands in contrast to stationary or adversarial bandits with the time-invaraiant benchmark, for which there are algorithms [Auer *et al.*, 2002b; Auer *et al.*, 2002a] that asymptotically achieves zero per-round regret as T goes to infinity. In the dynamic setting, each arm undergoes $O(T)$ amount of changes. The presence of missing values along with the rapid changes of each arm's expected reward makes it difficult for any algorithm to adapt to the changing environment in time. Our result also agrees with the lower bound provided in [Besbes *et al.*, 2014], which states that when the total variation of expected rewards scale linearly with T, the regret also grows linearly.

An alternating-and-restarting algorithm. We present and analyze an algorithm for the dynamic AR bandits. A special case of an AR model is a Brownian motion (random walk) process, which is used to model temporal structure in dynamic bandits problems [Slivkins and Upfal, 2008]. We note that the algorithm designed for Brownian dynamic bandits cannot be used directly for AR dynamic bandits. When the rewards evolve as AR processes, an extra challenge is that the correlation between past and future information decays exponentially fast. Consequently, the algorithm needs to not only adapt to the rapidly changing environment, but also discard past information that is less relevant for prediction of the future. We achieve these goals by presenting an algorithm called `AR2`, which stands for "Alternating and Restarting" algorithm for dynamic "AR" bandits. `AR2` relies on two mechanisms: (i) it alternates between exploration and exploitation throughout the horizon. During exploitation, `AR2` goes with a *superior arm* that is expected to have high reward based on recently collected information; during exploration, it pulls one of the *triggered arms*, which have the potential to outperform the superior arm. (ii) `AR2` also adopts a restarting mechanism. Restarting allows `AR2` to get rid of potentially misleading information and examine the arms that are not necessarily triggered but may have the potential to surpass triggered arms. We establish an upper bound on the per-round steady state regret of `AR2`, which is at most $\tilde{O}(\alpha^2\sigma^2k^3)$, and show that it is close to the regret lower bound. From a technical point of view, the analysis of the upper bound is rather intricate, as it involves distributing the regret incurred at each round over subsequent rounds, and identifying a high-probability event on which the rewards do not experience drastic changes.

Numerical studies. We numerically verify the efficacy of `AR2` by comparing it against several benchmark algorithms designed for both stationary and non-stationary settings. We show that our algorithm outperforms traditional algorithms (e.g., "explore-then-commit" [Garivier *et al.*, 2016] and UCB [Auer *et al.*, 2002b]), existing algorithms designed for non-stationary bandits (e.g., REXP3 [Besbes *et al.*, 2014]), and a modified UCB algorithm that takes into account the AR temporal structure. We show that `AR2` continues to perform well in a more general setting where the expected reward of each arm $i \in [k]$ follows an independent AR-1 process with heterogeneous AR parameters α_i and stochastic rate of change σ_i. Finally, we demonstrate via numerical studies that in the presence of certain levels of noise in our knowledge of AR parameters, the performance of `AR2` remains robust.

3 Conclusion and Future Directions

In this work, we study a dynamic MAB problem where the expected reward of each arm follows an AR-based temporal structure. We show, both theoretically and numerically, the efficacy of our learning algorithm `AR2` in a dynamically changing environment. The alternating and restarting mechanism of `AR2` enables us to dynamically adapt to changes, as well as discard irrelevant past information at a suitable rate.

There are a number of interesting future directions of this work that we intend to pursue. One such direction is to model the evolution of expected rewards using more sophisticated temporal structures, such as those with seasonality and trends. Many real-world dynamics, such as the shift of product demands and the variation of patient arrival rates at hospitals, display seasonality and periodicity and are more accurately captured using these temporal structures. This would allow us to apply our algorithms to more realistic settings. Another research area of interest is to add a mechanism that learns the AR parameter. In this work, we have assumed full knowledge of the AR parameter α; however, in many real-world applications, the extent of correlation between future and past rewards is unknown and needs to be learned over time. We intend to improve `AR2` to learn α when we lack knowledge of the AR parameter, and provide theoretical guarantees for this more challenging setting. With the addition of the new mechanism, we would like to develop methods that not only effectively adapt to the dynamic environment, but also provide us with insights into the temporal structure of the environment.

References

[Auer *et al.*, 2002a] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The non-stochastic multi-armed bandit problem. *SIAM journal of computing*, 32:48–77, 2002.

[Auer *et al.*, 2002b] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

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

[Garivier *et al.*, 2016] Aurelien Garivier, Tor Lattimore, and Emilie Kaufmann. On explore-then-commit strategies. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[Ortner *et al.*, 2012] Ronald Ortner, Daniil Ryabko, Peter Auer, and Rémi Munos. Regret bounds for restless markov bandits. In *Proceedings of the 23rd International Conference on Algorithmic Learning Theory*, ALT'12, page 214–228, Berlin, Heidelberg, 2012. Springer-Verlag.

[Robbins, 1952] Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527 – 535, 1952.

[Slivkins and Upfal, 2008] Alex Slivkins and Eli Upfal. Adapting to a changing environment: the brownian restless bandits. In *21st Conference on Learning Theory (COLT)*, pages 343–354, 2008.

[Thompson, 1933] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285 – 294, 1933.