JMLR: Workshop and Conference Proceedings vol 49:1–5, 2016

An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits

Peter Auer * AUER@UNILEOBEN.AC.AT
Chair for Information Technology
Montanuniversitaet Leoben
Franz-Josef-Strasse 18, A-8700 Leoben, Austria

Chao-Kai Chiang CHAOKAI@GMAIL.COM
University of California, Los Angeles
Computer Science Department
4732 Boelter Hall
Los Angeles, CA 90095

Abstract

We present an algorithm that achieves almost optimal pseudo-regret bounds against adversarial and stochastic bandits. Against adversarial bandits the pseudo-regret is $O\left(K\sqrt{n\log n}\right)$ and against stochastic bandits the pseudo-regret is $O\left(\sum_i(\log n)/\Delta_i\right)$. We also show that no algorithm with $O\left(\log n\right)$ pseudo-regret against stochastic bandits can achieve $\tilde{O}\left(\sqrt{n}\right)$ expected regret against adaptive adversarial bandits. This complements previous results of Bubeck and Slivkins (2012) that show $\tilde{O}\left(\sqrt{n}\right)$ expected adversarial regret with $O\left((\log n)^2\right)$ stochastic pseudo-regret.

1. Introduction

We consider the multi-armed bandit problem, which is the most basic example of a sequential decision problem with an exploration-exploitation trade-off. In each time step $t = 1, 2, \ldots, n$, the player has to play an arm $I_t \in \{1, \ldots, K\}$ from this fixed finite set and receives reward $x_{I_t}(t) \in [0, 1]$ depending on its choice[1]. The player observes only the reward of the chosen arm, but not the rewards of the other arms $x_i(t)$, $i \neq I_t$. The player's goal is to maximize its total reward $\sum_{t=1}^n x_{I_t}(t)$, and this total reward is compared to the best total reward of a single arm, $\sum_{t=1}^n x_i(t)$. To identify the best arm the player needs to explore all arms by playing them, but it also needs to limit this exploration to often play the best arm. The optimal amount of exploration constitutes the exploration-exploitation trade-off.

Different assumptions on how the rewards $x_i(t)$ are generated have led to different approaches and algorithms for the multi-armed bandit problem. In the original formulation (Robbins, 1952) it is assumed that the rewards are generated independently at random, governed by fixed but unknown probability distributions with means μ_i for each arm $i = 1, \ldots, K$. This type of bandit problem is called *stochastic*. The other type of bandit problem that we consider in this paper is called non-stochastic or *adversarial* (Auer et al., 2002b). Here the rewards may be selected arbitrarily by

* Extended abstract. Full version appears as [arXiv:1605.08722v1].

1. We assume that the player knows the total number of time steps n.

an adversary and the player should still perform well for any selection of rewards. An extensive overview of multi-armed bandit problems is given in (Bubeck and Cesa-Bianchi, 2012).

A central notion for the analysis of stochastic and adversarial bandit problems is the regret $R(n)$, the difference between the total reward of the best arm and the total reward of the player:

$$R(n) = \max_{1 \le i \le K} \sum_{t=1}^{n} x_i(t) - \sum_{t=1}^{n} x_{I_t}(t).$$

Since the player does not know the best arm beforehand and needs to do exploration, we expect that the total reward of the player is less than the total reward of the best arm. Thus the regret is a measure for the cost of not knowing the best arm. In the analysis of bandit problems we are interested in high probability bounds on the regret or in bounds on the expected regret. Often it is more convenient, though, to analyze the pseudo-regret

$$\overline{R}(n) = \max_{1 \le i \le K} \mathbb{E}\left[\sum_{t=1}^{n} x_i(t) - \sum_{t=1}^{n} x_{I_t}(t)\right]$$

instead of the expected regret

$$\mathbb{E}\left[R(n)\right] = \mathbb{E}\left[\max_{1 \le i \le K} \sum_{t=1}^{n} x_i(t) - \sum_{t=1}^{n} x_{I_t}(t)\right].$$

While the notion of pseudo-regret is weaker than the expected regret with $\overline{R}(n) \le \mathbb{E}\left[R(n)\right]$, bounds on the pseudo-regret imply bounds on the expected regret for adversarial bandit problems with *oblivious* rewards $x_i(t)$ selected independently from the player's choices. The pseudo-regret also allows for refined bounds in stochastic bandit problems.

1.1. Previous results

For adversarial bandit problems, algorithms with high probability bounds on the regret are known (Bubeck and Cesa-Bianchi, 2012, Theorem 3.3): with probability $1 - \delta$,

$$R_{\mathrm{adv}}(n) = O\left(\sqrt{n \log(1/\delta)}\right).$$

For stochastic bandit problems, several algorithms achieve logarithmic bounds on the pseudo-regret, e.g. Auer et al. (2002a):

$$\overline{R}_{\mathrm{sto}}(n) = O\left(\log n\right).$$

Both of these bounds are known to be best possible.

While the result for adversarial bandits is a worst-case — and thus possibly pessimistic — bound that holds for any sequence of rewards, the strong assumptions for stochastic bandits may sometimes be unjustified. Therefore an algorithm that can adapt to the actual difficulty of the problem is of great interest. The first such result was obtained by Bubeck and Slivkins (2012), who developed the SAO algorithm that with probability $1 - \delta$ achieves

$$R_{\mathrm{adv}}(n) \le O\left((\log n)\sqrt{n \log(n/\delta)}\right)$$

regret for adversarial bandits and

$$\overline{R}_{\mathrm{sto}}(n) = O\left((\log n)^2\right)$$

pseudo-regret for stochastic bandits.

It has remained as an open question if a stochastic pseudo-regret of order $O\left((\log n)^2\right)$ is necessary or if the optimal $O(\log n)$ pseudo-regret can be achieved while maintaining an adversarial regret of order $\sqrt{n}$.

1.2. Summary of new results

We give a twofold answer to this open question. We show that stochastic pseudo-regret of order $O\left((\log n)^2\right)$ is necessary for a player to achieve high probability adversarial regret of order $\sqrt{n}$ against an oblivious adversary, and to even achieve expected regret of order $\sqrt{n}$ against an adaptive adversary. But we also show that a player can achieve $O(\log n)$ stochastic pseudo-regret and $\tilde{O}(\sqrt{n})$ adversarial *pseudo-regret* at the same time. This gives, together with the results of (Bubeck and Slivkins, 2012), a quite complete characterization of algorithms that perform well both for stochastic and adversarial bandit problems.

More precisely, for any player with stochastic pseudo-regret bound of order $O\left((\log n)^\beta\right)$, $\beta < 2$, and any $\epsilon > 0$, $\alpha < 1$, there is an adversarial bandit problem for which the player suffers $\Omega(n^\alpha)$ regret with probability $\Omega(n^{-\epsilon})$. Furthermore, there is an adaptive adversary against which the player suffers $\Omega(n^\alpha)$ expected regret. Secondly, we construct an algorithm with

$$\overline{R}_{\mathrm{sto}}(n) = O(\log n)$$

and

$$\overline{R}_{\mathrm{adv}}(n) = O\left(\sqrt{n \log n}\right).$$

At first glance these two results may appear contradictory for $\alpha - \epsilon > 1/2$, as the lower bound seems to suggest a pseudo-regret of $\Omega(n^{\alpha-\epsilon})$. This is not the case, though, since the regret may also be negative. Indeed, consider an adversarial multi-armed bandit that initially gives higher rewards for one arm, and from some time step on gives higher rewards for a second arm. A player that detects this change and initially plays the first arm and later the second arm, may outperform both arms and achieve negative regret. But if the player misses the change and keeps playing the first arm, it may suffer large regret against the second arm.

In our analysis we use both mechanisms. For the lower bound on the pseudo-regret we show that a player with little exploration (which is necessary for small stochastic pseudo-regret) will miss such a change with significant probability and then will suffer large regret. For the upper bound we explicitly compensate possible large regret that occurs with small probability by negative regret that occurs with sufficiently large probability. For the lower bound on the expected regret we construct an adaptive adversary that prevents such negative regret. Consequently, our results exhibit one of the rare cases where there is a significant gap between the achievable pseudo-regret and the achievable expected regret.

2. Statement of results

We consider multi-armed bandit problems with rewards $x_i(t) \in [0,1]$ with arms $i = 1, \dots, K$ and time steps $t = 1, \dots, n$. We assume that the number of time steps n is known to the player.

In stochastic multi-armed bandit problems the rewards are generated independently at random with a fixed average reward $\mu_i = \mathbb{E}[x_i(t)]$ for each arm i. An important quantity is the gap $\Delta_i = \mu^* - \mu_i$ which is the distance to the optimal average reward $\mu^* = \max_i \mu_i$. The goal of the player is to achieve low pseudo-regret which for a stochastic bandit problem can be written as $\overline{R}_{\text{sto}}(n) = \sum_{i=1}^{K} \Delta_i \mathbb{E}[T_i(n)]$, where $T_i(n)$ is the total number of plays of arm i.

In adversarial bandit problems the rewards are selected by an adversary. If this is done before the player interacts with the environment, then the adversary is called *oblivious*. If the selection of rewards $x_i(t)$, $1 \le i \le K$, depends on the player's previous choices, then the adversary is called *adaptive*.

Theorem 1 *Let $\alpha < 1$, $\epsilon > 0$, $\beta < 2$, and $C > 0$. Consider a player that achieves pseudo-regret*

$$\overline{R}_{\text{sto}}(n) \le C(\log n)^{\beta}$$

for any stochastic bandit problem with two arms and gap $\Delta = 1/8$. Then for large enough n there is an adversarial bandit problem with two arms and an oblivious adversary such that the player suffers regret

$$R_{\text{obl}}(n) \ge n^{\alpha}/8 - 4\sqrt{n \log n}$$

with probability at least $1/(16n^{\epsilon}) - 2/n^2$. Furthermore, there is an adversarial bandit problem with two arms and an adaptive adversary such that the player suffers expected regret

$$\mathbb{E}[R_{\text{ada}}(n)] \ge \frac{n^{\alpha-\epsilon}}{128} - 3\sqrt{n \log n}.$$

Theorem 2 *There are constants C_{sto} and C_{adv}, such that for large enough n and any $\delta > 0$, our SAPO algorithm (Stochastic and Adversarial Pseudo-Optimal) achieves the following bounds on the pseudo-regret:*

- *For stochastic bandit problems with gaps Δ_i such that $\sum_{i:\Delta_i>0} \frac{\log(n/\delta)}{\Delta_i} \le \sqrt{nK \log(n/\delta)}$,*

$$T_i(n) \le C_{\text{sto}} \frac{\log(n/\delta)}{\Delta_i^2}$$

  with probability $1 - \delta$ for any arm i with $\Delta_i > 0$, and thus

$$\overline{R}_{\text{sto}}(n) \le C_{\text{sto}} \sum_{i:\Delta_i>0} \frac{\log(n/\delta)}{\Delta_i} + \delta n.$$

- *For adversarial bandit problems*

$$\overline{R}_{\text{ada}}(n) \le C_{\text{adv}} K \sqrt{n \log(n/\delta)} + \delta n.$$

Our bound for stochastic bandits is optimal up to a constant. The linear dependency on K of our bound for adversarial bandits is an artifact of our current analysis and can be improved to $O\left(\sqrt{nK \log(n/\delta)}\right)$. This bound is optimal up to a factor $\sqrt{\log n}$.

Our SAPO algorithm follows the general strategy of the SAO algorithm (Bubeck and Slivkins, 2012) by essentially employing an algorithm for stochastic bandit problems that is equipped with

additional tests to detect non-stochastic arms. A different approach is taken in (Seldin and Slivkins, 2014): here the starting point is an algorithm for adversarial bandit problems that is modified by adding an additional exploration parameter to achieve also low pseudo-regret in stochastic bandit problems. While this approach has not yet allowed for the tight $O(\log n)$ regret bound in stochastic bandit problems (they achieve a $O\left(\log^3 n\right)$ bound), the approach is quite flexible and more generally applicable than the SAO and SAPO algorithms.

Acknowledgments

We thank the anonymous reviewers for their very valuable comments. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 231495 (CompLACS) and from the Austrian Science Fund (FWF) under contract P 26219-N15.

References

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002a.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multi-armed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002b.

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.

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *COLT - The 25th Annual Conference on Learning Theory*, pages 42.1–42.23, 2012.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.*, 58 (5):527–535, 1952.

Yevgeny Seldin and Aleksandrs Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1287–1295, 2014.