Proceedings of Machine Learning Research vol 65:1–4, 2017

Open Problem: First-Order Regret Bounds for Contextual Bandits

Alekh Agarwal[⋆] ALEKHA@MICROSOFT.COM

Akshay Krishnamurthy[†] AKSHAY@CS.UMASS.EDU

John Langford[⋆] JCL@MICROSOFT.COM

Haipeng Luo[⋆] HAIPENG@MICROSOFT.COM

Robert E. Schapire[⋆] SCHAPIRE@MICROSOFT.COM

[⋆] *Microsoft Research, NYC*

[†] *University of Massachusetts, Amherst*

Abstract

We describe two open problems related to first order regret bounds for contextual bandits. The first asks for an algorithm with a regret bound of $\widetilde{\mathcal{O}}(\sqrt{L_\star K \ln N})$ where there are K actions, N policies, and $L_\star$ is the cumulative loss of the best policy. The second asks for an optimization-oracle-efficient algorithm with regret $\widetilde{\mathcal{O}}(L_\star^{2/3}\text{poly}(K, \ln(N/\delta)))$. We describe some positive results, such as an inefficient algorithm for the second problem, and some partial negative results.

Keywords: contextual bandits, first-order regret bounds, oracle-efficient algorithms

1. Introduction and Preliminaries

The contextual bandits problem (Langford and Zhang, 2008) is a generalization of classic multi-armed bandits, where a learner sequentially interacts with the environment to find a good policy, despite limited feedback. Contextual bandit learning arises in recommendation, advertising, and elsewhere, and has been successfully deployed in real-world systems (Agarwal et al., 2016).

The theory for contextual bandits has focused on developing computationally efficient algorithms with optimal worst-case statistical performance. However, there is little progress in designing more adaptive algorithms that can perform much better when the environment is relatively benign. Given that adaptivity is possible in many other online learning settings (e.g. the expert setting (Freund and Schapire, 1997)), here we ask whether a specific yet standard first-order data-dependent bound is achievable for contextual bandits. We believe that this is not only a challenging puzzle, but also one that can further improve the practicality of contextual bandits.

The contextual bandits problem is defined as follows. Let $\mathcal{X}$ be an arbitrary context space and K be the number of actions. A mapping $\pi : \mathcal{X} \to [K]$ is called a policy and the learner is given a fixed set of policies Π. For simplicity, we assume Π is a finite set but with a huge cardinality $N = |\Pi|$. The whole learning process proceeds in rounds. At each round $t = 1, \ldots, T$, the environment first decides a pair of context $x_t \in \mathcal{X}$ and loss vector $\ell_t \in [0,1]^K$, and then reveals x_t to the learner. The learner then picks an action $a_t \in [K]$ and observes its loss $\ell_t(a_t)$. The regret of the learner against a fixed policy $\pi \in \Pi$ after T rounds is defined as: $\mathcal{R}(\pi) = \sum_{t=1}^T \ell_t(a_t) - \ell_t(\pi(x_t))$.

We consider two settings. In the *stochastic setting*, the context-loss pairs are i.i.d. from a distribution $\mathcal{D}$, and we denote by $\pi_\star = \text{argmin}_{\pi\in\Pi} \mathbb{E}_{(x,\ell)\sim\mathcal{D}}\, \ell(\pi(x))$ the optimal policy and $L_\star =$

$T\mathbb{E}_{(x,\ell)\sim\mathcal{D}}\,\ell(\pi_\star(x))$ the optimal cumulative loss. In the *adversarial setting*, the context-loss pairs are decided in an arbitrary way before the game starts,[1] and, overloading notation, we define $\pi_\star = \operatorname{argmin}_{\pi\in\Pi}\sum_{t=1}^{T}\ell_t(\pi(x_t))$ and $L_\star = \sum_{t=1}^{T}\ell_t(\pi_\star(x_t))$ to be the optimal policy and loss respectively. We measure performance by the regret against the optimal policy, that is, $\mathcal{R}(\pi_\star)$.

The classic algorithm for contextual bandits is EXP4 (Auer et al., 2002), which enjoys a worst-case optimal regret bound $\mathbb{E}[\mathcal{R}(\pi_\star)] = \mathcal{O}(\sqrt{TK\ln N})$ in the adversarial setting, but requires maintaining a weight for each policy and is thus inefficient. To circumvent this computational obstacle, many existing works assume access to an optimization oracle:

Definition 1 *An optimization oracle takes any set of context and loss pairs $S = \{(x_i,\ell_i)\}_{i\in\mathcal{I}}$ and outputs* $\operatorname{argmin}_{\pi\in\Pi}\sum_{(x,\ell)\in S}\ell(\pi(x))$. *An algorithm is* oracle-efficient *if it runs in poly$(T,K,\ln N)$ time given access to an optimization oracle (excluding the running time of the oracle itself).*

Agarwal et al. (2014) derive a statistically-optimal oracle-efficient algorithm for the stochastic setting with other recent algorithmic advances (Rakhlin and Sridharan, 2016; Syrgkanis et al., 2016a,b) and hardness results (Hazan and Koren, 2016).

2. The Problems

Our goal is to obtain more adaptive and data-dependent regret bounds that can be much smaller than $\widetilde{\mathcal{O}}(\sqrt{T})$ in some cases. Specifically, we ask the following two questions:[2]

(Q1) Is there an algorithm which guarantees $\mathcal{R}(\pi_\star) = \widetilde{\mathcal{O}}(\sqrt{L_\star K\ln(N/\delta)} + \text{poly}(K,\ln(N/\delta)))$ with probability $1-\delta$, in either the adversarial or the stochastic setting?[3]

(Q2) Is there an oracle-efficient algorithm (Def. 1) for the stochastic setting that guarantees $\mathcal{R}(\pi_\star) = \widetilde{\mathcal{O}}((L_\star^{2/3}+1)\text{poly}(K,\ln(N/\delta)))$ with probability $1-\delta$?

The bound in Problem (Q1) is a very natural and standard first-order bound. Indeed, it is well-known that with full information, which is the classical experts problem (Freund and Schapire, 1997), one can get $\widetilde{\mathcal{O}}(\sqrt{L_\star\ln(N/\delta)})$ regret using the exponential weights algorithm.[4] Even for multi-armed bandits where Π is just K fixed-action policies, bounds of the form $\widetilde{\mathcal{O}}(\sqrt{L_\star K\ln(K/\delta)})$ are known (Allenberg et al., 2006; Foster et al., 2016). Similar bounds also exist in more challenging bandit variants (Neu, 2015). Therefore, replacing the dependence on T by $L_\star$ in the regret bound is natural; however, no such bounds are known for contextual bandits, and obtaining one appears to be challenging, even with an inefficient algorithm in the stochastic setting.

We note that by treating each policy as an arm, the aforementioned bandit algorithms produce a $\widetilde{\mathcal{O}}(\sqrt{L_\star N\ln(N/\delta)})$ bound. However, the polynomial dependence on N makes this unacceptable.

However, in Section 3 we will show that $\widetilde{\mathcal{O}}((L_\star^{2/3}+1)\text{poly}(K,\ln(N/\delta)))$ is achievable with an inefficient algorithm even in the adversarial setting. Therefore, in Problem (Q2), we ask whether the same bound can be achieved by an oracle-efficient algorithm in the simpler stochastic setting. In Section 4 we mention why some existing methods and their variants fail to achieve this. In fact, we will consider resolving the following relaxed version of (Q2) as important progress:

(Q2') Is there an oracle-efficient algorithm for the stochastic setting that guarantees $o(\sqrt{T})$ regret as long as $L_\star = O(T^\alpha)$ for some constant $0\le\alpha<1$?

1. For simplicity we only consider an oblivious adversary here.
2. For simplicity we even assume the value $L_\star$ is given to the algorithm ahead of time.
3. We use $\widetilde{\mathcal{O}}$ to suppress dependence on $\ln T$, $\ln K$ and $\ln\ln(N/\delta)$.
4. For conciseness we often omit the lower order term that is independent of $L_\star$, such as $K\ln(N/\delta)$.

3. Positive Results

Turning first to Problem (Q1), we show that using the GREEN clipping trick (Allenberg et al., 2006) on the action distribution produced by EXP4 achieves $\widetilde{\mathcal{O}}(L_\star^{2/3}\text{poly}(K, \ln(N/\delta)))$ regret with high probability. Specifically, the algorithm uses the EXP4 update rule $p_{t+1}(\pi) \propto p_t(\pi)\exp(-\eta\tilde{\ell}_t(\pi(x_t)))$ where $\tilde{\ell}_t(a)$ is a loss estimate to be described shortly and η is the learning rate. The projected distribution over actions given context x_t is defined as: $p_t(a|x_t) = \sum_{\pi:\pi(x_t)=a} p_t(\pi)$ for any a. The algorithm picks action a with probability $\tilde{p}_t(a|x_t) \propto \mathbf{1}\{p_t(a|x_t) \geq \gamma\}p_t(a|x_t)$ for some fixed threshold γ, and the loss estimate is then defined as $\tilde{\ell}_t(a) = \ell_t(a)\mathbf{1}\{a_t = a\}/\tilde{p}_t(a|x_t)$ for all a.

Let $\hat{L} = \sum_{t=1}^T \ell_t(a_t)$ be the cumulative loss of the learner. Following the analysis of Allenberg et al. (2006) gives the bound:

$$\hat{L} \leq \sum_{t=1}^T \tilde{\ell}_t(\pi_\star(x_t)) + K\gamma\hat{L} + \frac{\ln N}{\eta} + \eta\sum_{t=1}^T \tilde{\ell}_t(a_t)$$

where the last term is at most $\eta\hat{L}/\gamma$ since $\tilde{p}_t(a_t|x_t)$ is at least γ.[5] Using Freedman's inequality one can bound $\sum_{t=1}^T \tilde{\ell}_t(\pi_\star(x_t))$ by $L_\star + 2\sqrt{\frac{L_\star \ln(1/\delta)}{\gamma}} + \frac{\ln(1/\delta)}{\gamma}$. Finally setting $\gamma = \sqrt{\eta/K}$, $\eta = \min\{L_\star^{-2/3}K^{-1/3}(\ln N)^{2/3}, 1/(2K)\}$ and re-arranging show that the regret is $\mathcal{O}(L_\star^{2/3}(K\ln N)^{1/3} + K\ln N)$ with probability at least $1-\delta$.

Turning to (Q2), we mention two attempts. First for the stochastic setting, a standard analysis of an explore-then-exploit approach using Bernstein's inequality produces an $\mathcal{O}((TL_\star K\ln(N/\delta))^{1/3} + \sqrt{TK\ln(N/\delta)})$ regret bound. While this improves upon the original $T^{2/3}$ regret of Langford and Zhang (2008), the $\sqrt{T}$ term still makes the bound dominated by results of Agarwal et al. (2014).

We also mention without giving details that using the GREEN trick with the CONTEXT-FTPL algorithm of Syrgkanis et al. (2016a) in their small-separator setting gives an expected regret bound of $\mathcal{O}(L_\star^{3/4}\text{poly}(d, K, \ln N))$ where d is the size of the separating set. Interestingly, this bound has no dependence on T and holds in the adversarial setting, and the algorithm is oracle-efficient. However, it crucially requires a small separator set, which is not available in general.

4. Negative Results

In this section we apply the clipping trick to some other algorithms for the stochastic setting. For simplicity, we consider the (inefficient) POLICY_ELIMINATION algorithm of Dudík et al. (2011), upon which oracle-efficient algorithms were built (Dudík et al., 2011; Agarwal et al., 2014), in a very simplified setup and show its failure in getting the desired bound.

Specifically, assume there are only two actions a_g and a_b, $\sqrt{T}$ contexts $x^1, \ldots, x^{\sqrt{T}}$, and $\sqrt{T}+1$ policies $\pi_0, \pi_1, \ldots, \pi_{\sqrt{T}}$ such that $\pi_0(x) = a_g$ for all x and $\pi_j(x)$ is a_g if $x \neq x^j$ and a_b otherwise. The marginal distribution over the contexts is uniform, and the loss for a_g and a_b is always 0 and 1 respectively. Note that in this case $\pi_\star = \pi_0$ and $L_\star = 0$.

Let $\Pi_0 = \Pi$, γ be a clipping threshold, $\mu \in [0, 1/2K]$ be some minimum probability and $q^\mu(a) = (1-K\mu)q(a) + \mu$ for $q \in \Delta^K$. Consider the following variant of POLICY_ELIMINATION: at each time t, find $p_t \in \Delta^{\Pi_{t-1}}$ subject to $\mathbb{E}_x\left[\frac{1}{p_t^\mu(\pi(x)|x)}\right] \leq 4K$, $\forall \pi \in \Pi_{t-1}$; then play a with probability

5. Note that here we must depart from the analysis of Allenberg et al. (2006), which is why we are getting $L_\star^{2/3}$.

proportional to $\mathbf{1}\{p_t^\mu(a|x_t) \geq \gamma\} p_t^\mu(a|x_t)$; finally update $\Pi_t = \{\pi \in \Pi_{t-1} : \pi(x_t) \neq a_t \vee \ell_t(a_t) = 0\}$ so that policies that are observed to make a mistake are eliminated. Now one can verify that

$$p_t = \frac{1}{2}\mathbf{1}_{\pi_i} + \frac{1}{2(|\Pi_{t-1}| - 2)} \sum_{\substack{j \neq 0, i \\ \pi_j \in \Pi_{t-1}}} \mathbf{1}_{\pi_j},$$

for any $i \neq 0$ s.t. $\pi_i \in \Pi_{t-1}$ satisfies the constraint ($\mathbf{1}_\pi$ denotes the point mass at π). However, since $p_t^\mu(a_b|x_i) = 1/2$, no matter what γ is on each round the algorithm either makes a mistake with probability at least $1/(2\sqrt{T})$, or has made $\sqrt{T}$ mistakes to eliminate all bad policies. Therefore the expected regret is $\Omega(\sqrt{T})$. Several variants of this algorithm were also considered and all of them were shown to have $\Omega(\sqrt{T})$ regret, indicating that some very different techniques are needed.

Prize We are offering a prize of $250 for positive or negative resolutions to (Q1) or (Q2).

Acknowledgements We thank Gergely Neu for formative discussions about these problems.

References

A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. E. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *ICML*, 2014.

A. Agarwal, S. Bird, M. Cozowicz, L. Hoang, J. Langford, S. Lee, J. Li, D. Melamed, G. Oshri, O. Ribas, S. Sen, and A. Slivkins. A multiworld testing decision service. *arXiv:1606.03966*, 2016.

C. Allenberg, P. Auer, L. Györfi, and G. Ottucsák. Hannan consistency in on-line learning in case of unbounded losses under partial monitoring. In *ALT*, 2006.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 2002.

M. Dudík, D. Hsu, S. Kale, N. Karampatziakis, J. Langford, L. Reyzin, and T. Zhang. Efficient optimal learning for contextual bandits. In *UAI*, 2011.

D. Foster, Z. Li, T. Lykouris, K. Sridharan, and E. Tardos. Learning in games: Robustness of fast convergence. In *NIPS*, 2016.

Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 1997.

Elad Hazan and Tomer Koren. The computational power of optimization in online learning. In *STOC*, 2016.

J. Langford and T. Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *NIPS*, 2008.

G. Neu. First-order regret bounds for combinatorial semi-bandits. In *COLT*, 2015.

A. Rakhlin and K. Sridharan. Bistro: An efficient relaxation-based method for contextual bandits. In *ICML*, 2016.

V. Syrgkanis, A. Krishnamurthy, and R. E. Schapire. Efficient algorithms for adversarial contextual learning. In *ICML*, 2016a.

V. Syrgkanis, H. Luo, A. Krishnamurthy, and R. E. Schapire. Improved regret bounds for oracle-based adversarial contextual bandits. In *NIPS*, 2016b.