Contextual bandits with continuous actions: Smoothing, zooming, and adapting

Akshay Krishnamurthy AKSHAY@CS.UMASS.EDU
John Langford JCL@MICROSOFT.COM
Aleksandrs Slivkins SLIVKINS@MICROSOFT.COM
Chicheng Zhang CHICHZHAN@MICROSOFT.COM
Microsoft Research, New York City

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

We study contextual bandit learning for any competitor policy class and continuous action space. We obtain two qualitatively different regret bounds: one competes with a smoothed version of the policy class under no continuity assumptions, while the other requires standard Lipschitz assumptions. Both bounds exhibit data-dependent "zooming" behavior and, with no tuning, yield improved guarantees for benign problems. We also study adapting to unknown smoothness parameters, establishing a price-of-adaptivity and deriving optimal adaptive algorithms that require no additional information.
Keywords: Contextual bandits, Lipschitz bandits, Nonparametric learning

We consider contextual bandits, a setting in which a learner repeatedly makes an action on the basis of contextual information and observes a loss for the action, with the goal of minimizing cumulative loss over a series of rounds. Contextual bandit learning has received much attention, and has seen substantial success in practice (e.g., Auer et al., 2002; Langford and Zhang, 2007; Agarwal et al., 2014, 2017). This line of work mostly considers small, finite action sets, yet in many real-world problems actions are chosen from an interval, so the set is continuous and infinite.

How can we learn to make actions from continuous spaces based on loss-only feedback?

We could assume that nearby actions have similar losses, for example that the losses are Lipschitz continuous as a function of the action (following Agrawal, 1995, and a long line of subsequent work). Then we could discretize the action set and apply generic contextual bandit techniques (Kleinberg, 2004) or more refined "zooming" approaches (Kleinberg et al., 2019; Bubeck et al., 2011; Slivkins, 2014) that are specialized to the Lipschitz structure.

However, this approach has several drawbacks. A global Lipschitz assumption is crude and limiting; actual problems exhibit more complex loss structures where smoothness varies with location, often with discontinuities. Second, prior works incorporating context — including the zooming approaches — employ a nonparametric benchmark set of policies, which yields a poor dependence on the context dimension and prevents application beyond low-dimensional context spaces. Finally, existing algorithms require knowledge of the Lipschitz constant, which is typically unknown.

Here we show that it is possible to avoid all of these drawbacks with a conceptually new approach, resulting in a more robust solution for managing continuous action sets. The key idea is to *smooth*

Extended abstract. Full version appears as arXiv:1902.01520, v2.

Type	Setting	Params	Regret Bound	Status
Smooth	Worst-case	$h \in (0,1]$	$\Theta\left(\sqrt{T/h}\right)$	New
Smooth	Data-dependent	$h \in (0,1]$	$O\left(\min_\epsilon T\epsilon + \theta_h(\epsilon)\right)$	New
Smooth	Adaptive $h \in (0,1]$	None	$\Theta\left(\sqrt{T}/h\right)$	New
Lipschitz	Worst-case	$L \geq 1$	$\Theta\left(T^{2/3}L^{1/3}\right)$	Generalized
Lipschitz	Data-dependent	$L \geq 1$	$O\left(\min_\epsilon TL\epsilon + \psi_L(\epsilon)/L\right)$	Generalized
Lipschitz	Adaptive $L \geq 1$	None	$\Theta(T^{2/3}\sqrt{L})$	New

Table 1: A summary of results for contextual bandits on the interval $[0,1]$ action space. T is the number of rounds, h is the smoothing bandwidth, $\theta_h(\epsilon) \leq 1/(h\epsilon)$ is the *smoothing coefficient*, L is the Lipschitz constant, and $\psi_L(\epsilon) \leq 1/\epsilon^2$ is the *policy zooming coefficient*. All algorithms take T and Π as additional inputs. Logarithmic dependence on $|\Pi|$ and T is suppressed in all upper bounds.

the actions: each action a is mapped to a well-behaved distribution over actions, denoted $\texttt{Smooth}(a)$, such as a uniform distribution over a small interval around a (when the action set is the interval $[0,1]$). This approach leads to provable guarantees with no assumptions on the loss function, since the loss for a smoothed action is always well behaved. Essentially, we may focus on estimation considerations while ignoring approximation issues. We recover prior results that assume a small Lipschitz constant, but the guarantees are meaningful in much broader scenarios.

Our algorithms work with any competitor policy set Π of mappings from context to actions, which we smooth as above. We measure performance by comparing the learner's loss to the loss of the best smoothed policy, and our guarantees scale with $\log|\Pi|$, regardless of the dimensionality of the context space. This recovers results for nonparametric policy sets, but more importantly accommodates *parametric* policies that scale to high-dimensional context spaces. Further, in some cases we are able to exploit benign structure in the policy set and the instance to obtain faster rates.

We design algorithms that require no knowledge of problem parameters and are *optimally adaptive*, matching lower bounds that we prove here. For the class of problems we consider, we show how this can be done with a unified algorithmic approach.

Our contributions, specialized to the interval $[0,1]$ action set for clarity, are:

1. We define a new notion of *smoothed regret* where policies map contexts to distributions over actions. These distributions are parametrized by a bandwidth h governing the spread. We show that the optimal worst-case regret bound with bandwidth h is $\Theta(\sqrt{T/h\log|\Pi|})$, which requires no smoothness assumptions on the losses (first row of Table 1).
2. We obtain data-dependent guarantees in terms of a *smoothing coefficient*, which can yield much faster rates in favorable instances (second row of Table 1).
3. We obtain an adaptive algorithm with $\sqrt{T}/h$ regret bound for all bandwidths, simultaneously. Further we show this to be optimal, demonstrating a price of adaptivity (third row of Table 1).
4. We obtain analogous results when the losses are L-Lipschitz (rows 3-6 of Table 1). Notably, our data-dependent result here is in terms of a *policy zooming coefficient*, generalizing and improving zooming results from prior work. We also demonstrate a price of adaptivity in the Lipschitz case.

Our results hold in much more general settings, and also apply to the non-contextual case, where we obtain several new guarantees.

References

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

Alekh Agarwal, Sarah Bird, Markus Cozowicz, Luong Hoang, John Langford, Stephen Lee, Jiaji Li, Dan Melamed, Gal Oshri, Oswaldo Ribas, Siddhartha Sen, and Alex Slivkins. Making contextual decisions with low technical debt. *arxiv:1606.03966*, 2017.

Rajeev Agrawal. The continuum-armed bandit problem. *SIAM Journal on Control and Optimization*, 1995.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 2002.

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 2011.

Robert Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems*, 2004.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Bandits and experts in metric spaces. *Journal of the ACM*, 2019. To appear. Merged and revised version of conference papers in *ACM STOC 2008* and *ACM-SIAM SODA 2010*. Also available at `http://arxiv.org/abs/1312.1277`.

John Langford and Tong Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. In *Advances in Neural Information Processing Systems*, 2007.

Aleksandrs Slivkins. Contextual bandits with similarity information. *The Journal of Machine Learning Research*, 2014.