Smooth Contextual Bandits: Bridging the Parametric and Non-differentiable Regret Regimes (Extended Abstract)

Yichun Hu YH767@CORNELL.EDU
Nathan Kallus KALLUS@CORNELL.EDU
Xiaojie Mao XM77@CORNELL.EDU
Cornell University

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

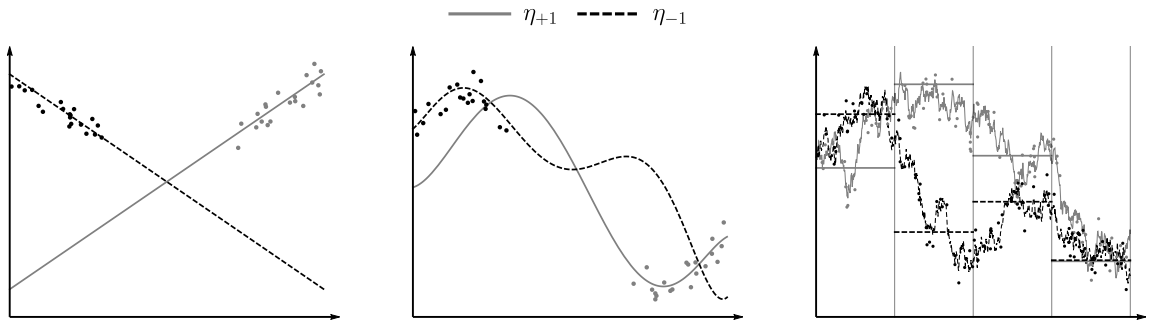
We study a nonparametric contextual bandit problem where the expected reward functions belong to a Hölder class with smoothness parameter β. We show how this interpolates between two extremes that were previously studied in isolation: non-differentiable bandits ($\beta \leq 1$), where rate-optimal regret is achieved by running separate non-contextual bandits in different context regions, and parametric-response bandits (satisfying $\beta = \infty$), where rate-optimal regret can be achieved with minimal or no exploration due to infinite extrapolatability. We develop a novel algorithm that carefully adjusts to all smoothness settings and we prove its regret is rate-optimal by establishing matching upper and lower bounds, recovering the existing results at the two extremes. In this sense, our work bridges the gap between the existing literature on parametric and non-differentiable contextual bandit problems and between bandit algorithms that exclusively use global or local information, shedding light on the crucial interplay of complexity and regret in contextual bandits. [1]

Keywords: Contextual bandits, local polynomial regression, minimax regret, margin condition

1. Introduction

In many domains, including healthcare and e-commerce, we frequently encounter the following decision-making problem: we sequentially and repeatedly receive context information X (*e.g.*, features of patients or users), need to choose an action $A \in \{-1, +1\}$ (*e.g.*, whether to treat a patient with invasive therapy or whether expose a user to our ad), and receive a reward $Y(A)$ (*e.g.*, patient's health outcome or user's click) corresponding to the chosen action. Our goal is to collect the most reward over time. When contexts X and potential rewards $Y(-1), Y(+1)$ are drawn from a stationary, but unknown, distribution, this setting is modeled by the stochastic bandit problem (Bubeck and Cesa-Bianchi 2012). A special case is the multi-armed bandit (MAB) problem where there is no contextual information (Lai and Robbins 1985, Auer et al. 2002). In these problems, we quantify the quality of an algorithm in terms of its *regret* for every horizon T: the expected additional cumulative reward up to time T that we would obtain if we had full knowledge of the stationary context-reward distribution. Since we only observe the reward of the chosen action, $Y(A)$, and never that of the unchosen action, $Y(-A)$, we face the oft-noted trade-off between exploration and exploitation: we are motivated to greedily exploit the arm we currently think is best for the context so to collect the

1. Extended abstract. Full version appears as [arXiv 1909.02553, v2].



(a) A linear response bandit: samples in one context region are fully informative about expected rewards in any other context region.

(b) A nonparametric-response bandit: samples offer only limited extrapolation to learn expected rewards at nearby context values.

(c) A non-differentiable-response bandit: rate-optimal regret obtainable by reducing the contextual bandit into multiple, separate MAB problems.

Figure 1: The fundamental nature of contextual bandit problems depends crucially on the assumed structure of expected reward functions, η_a.

highest reward right now, but we also need to explore the other arm for fear of missing better options in the future due to lack of information. This trade-off crucially depends on how we model the relationship between the context and the reward, *i.e.*, $\eta_a(x) = \mathbb{E}\left[Y(a) \mid X = x\right]$, for $a = \pm 1$. In the stochastic setting, previous literature has considered two extreme cases in isolation: a parametric reward model, usually linear (Goldenshluger and Zeevi 2013, Bastani and Bayati 2015, Bastani et al. 2017); and a nonparametric, non-differentiable reward model (Rigollet and Zeevi 2010, Perchet and Rigollet 2013). We review these below before describing our contribution.

Linear-response bandit. One extreme is the linear-response bandit where the expected reward function is assumed to be linear in context, $\eta_a(x) = \theta_a^\top x$ (Goldenshluger and Zeevi 2013, Bastani and Bayati 2015). This parametric assumption imposes a global structure on the expected reward function and permits extrapolation, since *all* samples from arm a are informative about the finite-dimensional parameters θ_a regardless of the context (see Fig. 1a). This global structure almost entirely obviates the need for forced exploration. In particular, Bastani et al. (2017) proved that, under mild conditions, the greedy algorithm is rate optimal for linear reward models, achieving logarithmic regret. Consequently, the result shows that the classic trade-off that characterizes contextual bandit problems is often not present in linear-response bandits. At the same time, while theoretically regret is very low, linear-response bandit algorithms may have linear regret in practice since the parametric assumption usually fails to hold exactly.

Non-differentiable nonparametric-response bandit. Another line of literature considers nonparametric reward models that satisfy a Hölder continuity condition (Rigollet and Zeevi 2010, Perchet and Rigollet 2013), which is a potentially weaker form of Lipschitz continuity. In stark contrast to the linear case, such functions need not even be differentiable. In nonparametric-response bandit, extrapolation is limited, since only nearby samples are informative about the reward functions at each context value (Fig. 1b). Thus, we need to take a more localized learning strategy: we have to actively explore in *every* context region and learn the expected reward functions using nearby samples. In the non-differentiable extreme, Rigollet and Zeevi (2010) showed that one can

		Smoothness		
		$\beta \leq 1$	$1 \leq \beta < \infty$	$\beta = \infty$
Margin Sharpness	$0 \leq \alpha < 1$	*Rigollet and Zeevi (2010)*	*— This paper —*	Bastani et al. (2017)
	$\alpha = 1$			Goldenshluger and Zeevi (2013)
	$\alpha > 1$	Perchet and Rigollet (2013)		Bastani et al. (2017)

Table 1: The lay of the literature on stochastic contextual bandits in terms of our smoothness perspective. Our work shows that (up to polylogs) the minimax regret rate $\tilde{\Theta}(T^{\frac{\beta+d-\alpha\beta}{2\beta+d}})$ reigns across *all* regimes.

achieve rate-optimal regret by partitioning the context space into small hypercubes and running completely separate MAB algorithms (*e.g.*, UCB) within each hypercube in isolation (Fig. 1c). In other words, we can almost ignore the contextual structure because we obtain so little information across contexts. At best, this achieves regret strictly worse than $\sqrt{T}$ in rate whenever the dimension of contexts is more than 2 and the bandit problem has nontrivial optimal decision rule. This rate *cannot* be imporved without further restrictions on reward models.

Our contribution: smooth contextual bandits. In this paper, we consider a nonparametric-response bandit problem with *smooth* expected reward functions. This bridges the gap between the infinitely-smooth linear-response bandit and the unsmooth non-differentiable-response bandit. We characterize the smoothness of the expected reward functions in terms of a Hölder smoothness parameter β, roughly indicating the highest order of continuous derivatives. Table 1 summarizes the landscape of the current literature and where our paper lies in terms of function smoothness and in terms of the sharpness α of the margin.

We propose a novel algorithm for every level of smoothness $1 \leq \beta < \infty$ that interpolates between the fully-global learning of the linear-response bandit ($\beta = \infty$) and the fully-local learning of the non-differentiable bandit ($0 < \beta \leq 1$). In particular, when $\beta > 1$, we must leverage information across farther-apart contexts and running separate MAB algorithms will be suboptimal. And, because $\beta < \infty$, we must ensure sufficient exploration everywhere. The smoother the expected reward functions, the more global reward information we incorporate. Moreover, our algorithm judiciously balances exploration and exploitation: it exploits only when we have certainty about which arm is optimal, and it explores economically in a shrinking margin region with fast diminishing error costs. As a result, our algorithm achieves regret bounded by $\tilde{O}(T^{\frac{\beta+d-\alpha\beta}{2\beta+d}})$. We show that, for any algorithm, there exists an instance on which it must have regret lower bounded by the same rate, showing that our algorithm is rate optimal and establishing the the minimax regret rate.

While this rate has the same *form* as the regret in the non-differentiable case, our results extend to the smooth ($\beta > 1$) regime where our algorithm can attain much lower regret, arbitrarily ap-

proaching polylogarithmic rates as smoothness increases. Our algorithm is fundamentally different, leveraging contextual information from farther away as smoothness increases without deteriorating estimation resolution, and our analysis is necessarily much finer. Our work connects seemingly disparate contextual bandit problems, and reveals the whole spectrum of minimax regret over varying levels of function complexity.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1846210.

References

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2-3):235–256, May 2002. ISSN 0885-6125.

Hamsa Bastani and Mohsen Bayati. Online decision-making with high-dimensional covariates. *Available at SSRN 2661896*, 2015.

Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits, 2017.

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. ISSN 1935-8237.

Alexander Goldenshluger and Assaf Zeevi. A linear response bandit problem. *Stochastic Systems*, 3(1): 230–261, 2013.

T.L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.*, 6(1):4–22, March 1985. ISSN 0196-8858.

Vianney Perchet and Philippe Rigollet. The multi-armed bandit problem with covariates. *Ann. Statist.*, 41(2): 693–721, 04 2013.

Philippe Rigollet and Assaf Zeevi. Nonparametric bandits with covariates, 2010.