Nearly Minimax-Optimal Regret for Linearly Parameterized Bandits

[*] **Yingkai Li** YINGKAI.LI@U.NORTHWESTERN.EDU
Department of Computer Science, Northwestern University

Yining Wang YNWANG.YINING@GMAIL.COM
Machine Learning Department, School of Computer Science, Carnegie Mellon University

Yuan Zhou YZHOUCS@INDIANA.EDU
Computer Science Department, Indiana University at Bloomington
Department of Industrial and Enterprise Systems Engineering, University of Illinois at Urbana-Champaign

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

We study the linear contextual bandit problem with finite action sets. When the problem dimension is d, the time horizon is T, and there are $n \leq 2^{d/2}$ candidate actions per time period, we (1) show that the minimax expected regret is $\Omega(\sqrt{dT \log T \log n})$ for every algorithm, and (2) introduce a Variable-Confidence-Level (VCL) SupLinUCB algorithm whose regret matches the lower bound up to iterated logarithmic factors. Our algorithmic result saves two $\sqrt{\log T}$ factors from previous analysis, and our information-theoretical lower bound also improves previous results by one $\sqrt{\log T}$ factor, revealing a regret scaling quite different from classical multi-armed bandits in which no logarithmic T term is present in minimax regret (Audibert and Bubeck, 2009). Our proof techniques include variable confidence levels and a careful analysis of layer sizes of SupLinUCB (Chu et al., 2011) on the upper bound side, and delicately constructed adversarial sequences showing the tightness of elliptical potential lemmas on the lower bound side.

1. Introduction

Linearly parameterized contextual bandit is an important class of sequential decision making models that incorporate contextual information with a linear function (Auer, 2002; Chu et al., 2011; Abbasi-Yadkori et al., 2011; Dani et al., 2008). There are $T \geq 1$ time periods, conveniently denoted as $\{1, 2, \cdots, T\}$, and a fixed but unknown d-dimensional regression model θ. Throughout this paper we will assume the model is normalized, meaning that $\|\theta\|_2 \leq 1$. At each time period t, a *policy* π is presented with an *action set* $\mathcal{A}_t = \{x_{it}\} \subseteq \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$. The policy then chooses, based on the feedback from previous time periods $\{1, 2, \cdots, t-1\}$, either deterministically or randomly an action $x_{it} \in \mathcal{A}_t$ and receives a reward $r_t = x_{it}^\top \theta + \varepsilon_t$, where $\{\varepsilon_t\}$ are independent centered sub-Gaussian random variables with variance proxy 1, representing noise during the reward collection procedure. The objective is to design a good policy π that tries to maximize its expected cumulative reward $\mathbb{E}[R^T] = \mathbb{E}\sum_{t=1}^T r_t$.

To provide a unified evaluation criterion, we adopt the concept of *worst-case regret* and aim to find a policy that has the smallest possible worst-case regret. More specifically, we are interested in

[*] Extended abstract. Full version appears as arXiv:1904.00242. Author names listed in alphabetical order.

Table 1: Previous results and our results on upper and lower bounds of $\mathfrak{R}(T; n, d)$.

		Upper bound	Lower bound
$n < \infty$	Previous result	$O(\sqrt{dT}\log^{3/2}(nT))$ (Auer, 2002; Chu et al., 2011)	$\Omega(\sqrt{dT})$ (Chu et al., 2011)
	Our result	$O(\sqrt{dT\log T\log n})\cdot \mathrm{poly}(\log\log(nT))$	$\Omega(\sqrt{dT\log n\log(T/d)})$
$n = \infty$	Previous result	$O(d\sqrt{T}\log T)$ (Abbasi-Yadkori et al., 2011)	$\Omega(d\sqrt{T})$ (Dani et al., 2008)
	Our result	N/A	$\Omega(d\sqrt{T\log T})$

the following defined *minimax regret*

$$\mathfrak{R}(T; n, d) := \inf_{\pi} \sup_{\theta\in\mathbb{R}^d, |\mathcal{A}_t|\le n} \mathbb{E}[R^T]. \tag{1}$$

Note that for $n = \infty$, the supremum is taken over all closed $\mathcal{A}_t \subseteq \{x \in \mathbb{R}^d : \|x\|_2 \le 1\}$ for all t.

1.1. Our results

The main results of this paper are the following two theorems that upper and lower bound the minimax regret $\mathfrak{R}(T; n, d)$ for various problem parameter values.

Theorem 1 (Upper bound) *For any $n < \infty$, the minimax regret $\mathfrak{R}(T; n, d)$ can be asymptotically upper bounded by $\mathrm{poly}(\log\log(nT))\cdot O(\sqrt{dT\log T\log n})$.*

Theorem 2 (Lower bound) *For any small constant $\epsilon > 0$, and any n, d, such that $n \le 2^{d/2}$ and $T \ge d(\log_2 n)^{1+\epsilon}$, the minimax regret $\mathfrak{R}(T; n, d)$ can be asymptotically lower bounded by $\Omega(1)\cdot\sqrt{dT\log n\log(T/d)}$.*

Comparing Theorems 1 and 2, we see that the upper and lower bounds nearly match each other up to iterated logarithmic terms when n (the number of actions per time period) is not too large.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2011.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the Conference on Learning Theory (COLT)*, 2009.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of the Conference on Learning Theory (COLT)*, 2008.