High-Dimensional Sparse Linear Bandits

Botao Hao, Tor Lattimore, Mengdi Wang

DeepMind

Princeton University

Contribution

- A fresh view on sparse linear bandits in the **high-dimensional regime**.
- A $\Theta(n^{2/3})$ minimax optimal regret bound when the feature vectors admit a well-conditioned exploration distribution.
- Provide an example where carefully balancing the trade-off between **information** and **regret** is necessary.

Problem Setting

- **Model:**
  At each round t, the agent chooses an action $A_t \in \mathcal{A} \subseteq \mathbb{R}^d$ (finite, fixed action set) and receives a reward:
  $$Y_t = \langle A_t, \theta^* \rangle + \eta_t, \ t \in [n],$$
  where $\|\theta^*\|_0 = s \ll d$ and η_t is sub-Gaussian noise.
  Interested in **high-dimensional regime:** $d > n$.
- **Hardness result:**
  Unfortunately, there exists a $\Omega(\sqrt{dsn})$ **minimax lower bound** in general.
  Minimax bounds do not tell the whole story!
- **Why?**
  A crude maximisation over **all environments** hides much of the rich structure of linear bandits with sparsity.
- **What we want to tell:**
  Derive a **sharp** $\Omega(\text{poly}(s)n^{2/3})$ lower bound in high-dimensional regime under the condition that *"the feature vectors admit a well-conditioned exploration distribution"*.

Related Work

Comparisons with existing results on regret upper bounds and lower bounds for sparse linear/contextual bandits. Here, K is the number of arms and τ is a problem-dependent parameter that may have a complicated form and vary across different literature.

	Regret upper bound	Assumptions
Abbasi-Yadkori et al. [2012]	$O(\sqrt{sdn})$	none
Bastani and Bayati [2020] Wang et al. [2018]	$O(\tau K s^2(\log(n))^2)$ $O(\tau K s^3 \log(n))$	linear contextual bandit with i.i.d context, margin condition, compatibility condition over an optimal action set
Kim and Paik [2019]	$O(\tau s\sqrt{n})$	linear contextual bandit with i.i.d context, compatibility condition, non-standard noise assumption
Lattimore et al. [2015]	$O(s\sqrt{n})$	action set is hypercube
This paper	$O(C_{\min}^{-2/3} s^{2/3} n^{2/3})$	action set spans $\mathbb{R}^d$
	Regret lower bound	
Multi-task bandits	$\Omega(\sqrt{sdn})$	N.A.
This paper	$\Omega(C_{\min}^{-1/3} s^{1/3} n^{2/3})$	N.A.

Minimax Lower Bound

Definition. Let $\mathcal{P}(\mathcal{A})$ be the space of probability measures over $\mathcal{A}$ and we define
$$C_{\min}(\mathcal{A}) = \sup_{\mu \in \mathcal{P}(\mathcal{A})} \sigma_{\min}\Big(\mathbb{E}_{A\sim\mu}[AA^\top]\Big),$$
where $\sigma_{\min}(\cdot)$ is the minimum eigenvalue of a square matrix.

Remark.

- When $C_{\min}(\mathcal{A})$ is independent of d, n, we say *"feature vectors admit a well-conditioned exploration distribution"*. Sampling uniformly from the corners of each set shows that $C_{\min}(\mathcal{A}) \geq 1$ for the former and $C_{\min}(\mathcal{A}) \geq 1/d$ for the latter.

Theorem (Minimax Lower Bound). For any policy π, there exists s-sparse parameter $\theta \in \mathbb{R}^d$ and an action set $\mathcal{A}$ where $C_{\min}(\mathcal{A})$ is independent of d, n such that
$$R_\theta(n) \gtrsim \min\Big(C_{\min}^{-\frac{1}{3}}(\mathcal{A}) s^{\frac{1}{3}} n^{\frac{2}{3}}, \sqrt{dn}\Big),$$
where $\gtrsim$ hides universal constants only.

Remark.

- When $d > n^{1/3}s^{2/3}$ the bound is $\Omega(n^{2/3})$, which is **independent of the dimension**.
- When $d \leq n^{1/3}s^{2/3}$, we recover the standard $\Omega(\sqrt{sdn})$ dimension-dependent lower bound up to a $\sqrt{s}$-factor, even though feature vectors admit a well-conditioned exploration distribution.

Matching Upper Bound

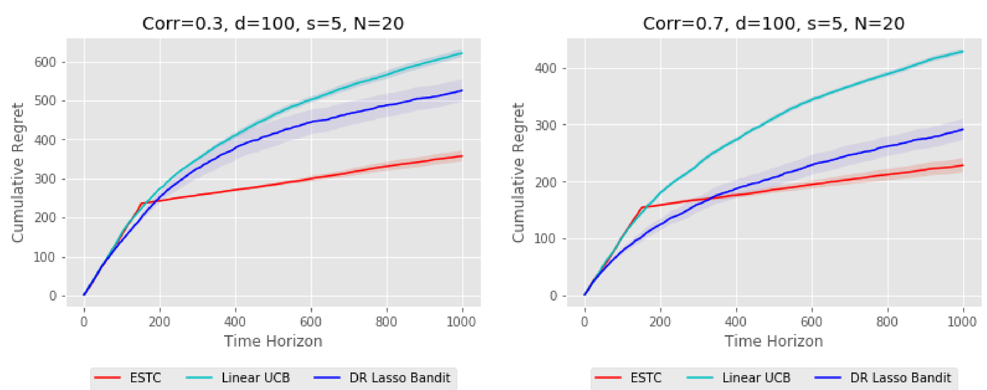
Theorem. Assume the action set $\mathcal{A}$ spans $\mathbb{R}^d$. The regret upper bound of explore-the-sparsity-then-commit (ESTC) algorithm satisfies
$$R_\theta(n) \lesssim C_{\min}^{-\frac{2}{3}}(\mathcal{A}) s^{\frac{2}{3}} n^{\frac{2}{3}}.$$

Algorithm.

1. Given an action set $\mathcal{A}$, the algorithm first solves the following optimization problem to find the most informative design:
$$\pi_e = \max_{\mu \in \mathcal{P}(\mathcal{A})} \sigma_{\min}\Big(\int_{x\in\mathcal{A}} xx^\top d\mu(x)\Big).$$
2. Independently pulls arms following π_e by n_1 rounds and collects samples as $\{(A_1, y_1), \ldots, (A_{n_1}, y_{n_1})\}$. Then the algorithm calculates the lasso estimator $\hat{\theta}_{n_1}$.
3. Executes the greedy action $A_t = \text{argmax}_{x\in\mathcal{A}} \langle x, \hat{\theta}_{n_1}\rangle$ for the remaining $n - n_1$ rounds.

Experiments

We compare ESTC (our algorithm) with LinUCB and doubly-robust (DR) lasso bandits on a linear contextual bandits: action set from $N(0_N, V)$, where N is the number of arms, $V_{ii} = 1$ and $V_{ik} = \rho^2$ for every $i \neq k$. Larger ρ favorable to DR-lasso.



Proof Hints

- **Hard problem instance.**

1. Construct a **low regret** action set $\mathcal{S}$ (*sparse*) and an **informative** action set $\mathcal{H}$ (*half of the hypercube*) as follows:
$$\mathcal{S} = \Big\{x \in \mathbb{R}^d \Big| x_j \in \{-1, 0, 1\} \text{ for } j \in [d-1], \|x\|_1 = s-1, x_d = 0\Big\},$$
$$\mathcal{H} = \Big\{x \in \mathbb{R}^d \Big| x_j \in \{-1, 1\} \text{ for } j \in [d-1], x_d = 1\Big\}.$$
2. Original bandit $\theta = (\underbrace{\varepsilon, \ldots, \varepsilon}_{s-1}, 0, \ldots, 0, -1)$, for some small $\varepsilon > 0$.
   Remark: sampling from the corner of $\mathcal{H}$ provides **more information** to infer θ than from $\mathcal{S}$ but leads to **high regret** due to the last coordinate -1.
3. Alternative bandit $\tilde{\theta}$. We denote a set $\mathcal{S}'$ as
$$\mathcal{S}' = \Big\{x \in \mathbb{R}^d \Big| x_j \in \{-1, 0, 1\} \text{ for } j \in \{s, s+1, \ldots, d-1\},$$
$$x_j = 0 \text{ for } j = \{1, \ldots, s-1, d\}, \|x\|_1 = s-1\Big\}.$$
   and $\tilde{x} = \text{argmin}_{x\in\mathcal{S}'} \mathbb{E}_\theta[\sum_{t=1}^n \langle A_t, x\rangle^2]$. Construct the alternative bandit $\tilde{\theta}$ as $\tilde{\theta} = \theta + 2\varepsilon\tilde{x}$.

- **Key steps: calculating the KL divergence.**
  Define $T_n(\mathcal{H}) = \sum_{t=1}^n \mathbb{I}(A_t \in \mathcal{H})$. The KL divergence between $\mathbb{P}_\theta$ and $\mathbb{P}_{\tilde{\theta}}$ is upper bounded by
$$\text{KL}\left(\mathbb{P}_\theta, \mathbb{P}_{\tilde{\theta}}\right) \leq 2\varepsilon^2\Big(\underbrace{n(s-1)^2/d}_{I_1} + \underbrace{\kappa^2(s-1)\mathbb{E}_\theta[T_n(\mathcal{H})]}_{I_2}\Big).$$

  Remark. I_1 is the contribution from actions in the low-regret action set $\mathcal{S}$, while I_2 is due to actions in $\mathcal{H}$. The fact that actions in $\mathcal{S}$ are not very informative is captured by the presence of the dimension in the denominator of the first term.

Conclusion

- **Summary.** In this paper, we show that $\Theta(n^{2/3})$ is the optimal rate in the high-dimensional regime when a suitable exploratory distribution exists.
- **Future direction.** It is unclear how the regret lower bound depends on $C_{\min}(\mathcal{A})$ in the data-rich regime and if $C_{\min}(\mathcal{A})$ is the best quantity to describe the shape of action set $\mathcal{A}$.

Reference

[1]. Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Online-to-confidence-set conversions and application to sparse stochastic bandits. [2]. Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates.
[3]. Gi-Soo Kim and Myunghee Cho Paik. Doubly-robust lasso bandit.
[4]. Tor Lattimore, Koby Crammer, and Csaba Szepesvari. Linear multi-resource allocation with semi-bandit feedback.