Information-Regret Trade-Off in Sparse Linear Bandits and Online RL

Botao Hao
Deepmind

Joint works with Tor Lattimore, Mengdi Wang, Csaba Szepesvári

High-Dimensional RL and Bandits

- RL and bandits achieve great success in recent years.

- Value function approximation requires high-dimensional features ⇒

  1. **high sample-complexity** 2. **poor interpretability**.

High-Dimensional RL and Bandits

- RL and bandits achieve great success in recent years.

- Value function approximation requires high-dimensional features ⇒

  1. **high sample complexity** 2. **poor interpretability**.

- A natural solution from supervised learning: **sparse representation.**

Q: Does sparsity still help in sequential decision making problems?

Story I: Stochastic Sparse Linear Bandits

Stochastic Sparse Linear Bandits

- At each round $t \in [n]$, the agent chooses an action $A_t \in \mathcal{A} \subseteq \mathbb{R}^d$ and receives a reward:

$$Y_t = \langle A_t, \theta^* \rangle + \eta_t.$$

where $\|\theta^*\|_0 = s \ll d$ and η_t is 1-sub-Gaussian noise. Assume for any $a \in \mathcal{A}$, $\|a\|_\infty \leq 1$ and $|\mathcal{A}| = K$.

- **Data-poor regime:** $d \gtrsim n$; **data-rich regime:** $d \lesssim n$.
- Cumulative regret:

$$\mathfrak{R}_{\theta^*}(n; \pi) = \mathbb{E}\left[\sum_{t=1}^{n} \langle x^*, \theta^* \rangle - \sum_{t=1}^{n} Y_t\right],$$

where x^* is the optimal action.

Does Sparsity Help?

- **Without sparsity**, the well-known minimax lower bound for linear bandits is $\Omega(d\sqrt{n})$.

Does Sparsity Help?

- **Without sparsity**, the well-known minimax lower bound for linear bandits is $\Omega(d\sqrt{n})$.
- **With sparsity**, there exists a $\Omega(\sqrt{dsn})$ minimax lower bound in general (no additional assumption on $\mathcal{A}$ and θ^*)[1].
- Unfortunately, sparsity **does not help much:(**

[1] Section 24.3 in Bandit Algorithm (2020).

Does Sparsity Help?

- **Without sparsity**, the well-known minimax lower bound for linear bandits is $\Omega(d\sqrt{n})$.
- **With sparsity**, there exists a $\Omega(\sqrt{dsn})$ minimax lower bound in general (no additional assumption on $\mathcal{A}$ and θ^*).
- Unfortunately, sparsity **does not help much:(**

Minimax bounds do not tell the whole story!

Why? A crude maximisation over **all environments** hides much of the rich structure of sparse linear bandits.

Recap on Sparse Linear Regression

- Consider a sparse linear regression:

$$y_i = \langle x_i, \theta^* \rangle + \eta_i, i = 1, \ldots, n,$$

where θ^* is s-sparse, η_i is 1-sub-Gaussian, x_i is i.i.d. random design.

Recap on Sparse Linear Regression

- Consider a sparse linear regression:

$$y_i = \langle x_i, \theta^* \rangle + \eta_i, i = 1, \ldots, n,$$

where θ^* is s-sparse, η_i is 1-sub-Gaussian, x_i is i.i.d. random design.

- **Sparsity does help!** If the design matrix is **well-conditioned**:

$$\sigma_{\min}(\mathbb{E}[x_i x_i^\top]) \geq C_{\min} \text{ (constant)},$$

where $\sigma_{\min}(\cdot)$ is the minimum eigenvalue, Lasso can reduce the parameter estimation error to

$$\left\| \widehat{\theta}_{\text{lasso}} - \theta^* \right\|_2 \lesssim \frac{1}{C_{\min}} \sqrt{\frac{s \log(d)}{n}}.$$

Recap on Sparse Linear Regression

- Consider a sparse linear regression:

$$y_i = \langle x_i, \theta^* \rangle + \eta_i, i = 1, \ldots, n,$$

  where θ^* is s-sparse, η_i is 1-sub-Gaussian, x_i is i.i.d. random design.

- **Sparsity does help!** If the design matrix is **well-conditioned**:

$$\sigma_{\min}(\mathbb{E}[x_i x_i^\top]) \geq C_{\min} \text{ (constant)},$$

  where $\sigma_{\min}(\cdot)$ is the minimum eigenvalue, Lasso can reduce the parameter estimation error to

$$\left\| \widehat{\theta}_{\text{lasso}} - \theta^* \right\|_2 \lesssim \frac{1}{C_{\min}} \sqrt{\frac{s \log(d)}{n}}.$$

Q: Will sparsity help in linear bandits under similar assumptions?

Our Contribution

- **When does sparsity help?** Derive a sharp $\Theta(\text{poly}(s)n^{2/3})$ minimax rate in the **data-poor regime** under the condition:

  "the action set admits a well-conditioned exploratory policy".

- **What should we learn?** Carefully balancing the trade-off between **information and regret** is **necessary** in sparse linear bandits.
- **How to achieve this?** Information-directed sampling can **adapt** to different information-regret structures.

Exploratory Policy

Definition. Let $\mathcal{P}(\mathcal{A})$ be the space of probability measures over $\mathcal{A}$. Then we define

$$C_{\min}(\mathcal{A}) = \sup_{\mu \in \mathcal{P}(\mathcal{A})} \sigma_{\min}\left(\mathbb{E}_{A \sim \mu}\left[AA^\top\right]\right).$$

Remarks.

- When $C_{\min}(\mathcal{A})$ is a constant, we say

  "action set $\mathcal{A}$ admits a well-conditioned exploratory policy".
- What is **information**? Pulling arms according to this exploratory policy, we collect information (well-conditioned data).

A Novel Minimax Lower Bound

Theorem (Minimax Lower Bound). For any policy π, there exists an action set $\mathcal{A}$ where $C_{\min}(\mathcal{A})$ is a constant and s-sparse parameter $\theta \in \mathbb{R}^d$ such that

$$\mathfrak{R}_\theta(n;\pi) \gtrsim \min\left(C_{\min}^{-\frac{1}{3}}(\mathcal{A}) s^{\frac{1}{3}} n^{\frac{2}{3}}, \sqrt{dn}\right),$$

where $\gtrsim$ just hides universal constants.

- When $d > n^{1/3}s^{2/3}$ the lower bound is $\Omega(n^{2/3})$, which is **independent of the dimension**.
- This lower bound is (nearly) sharp:
  - $O(s^{2/3}n^{2/3})$ achieved by explore-then-commit.
  - $O(\sqrt{sdn})$ achieved by optimism-based algorithm.

Why $n^{2/3}$? Some actions are **informative**, but also **high regret**!

Hard Bandit Problem Instance

- $\mathcal{A} = \mathcal{S} \cup \mathcal{H}$ with a **low regret** action set $\mathcal{S}$ (sparse) and an **informative** action set $\mathcal{H}$ (half of the hypercube):

$$\mathcal{S} = \left\{ x \in \mathbb{R}^d \middle| x_j \in \{-1, 0, 1\} \text{ for } j \in [d-1], \|x\|_1 = s-1, x_d = 0 \right\},$$
$$\mathcal{H} = \left\{ x \in \mathbb{R}^d \middle| x_j \in \{-1, 1\} \text{ for } j \in [d-1], x_d = 1 \right\}.$$

- True parameter θ^*: for some small $\varepsilon > 0$

$$\theta^* = (\underbrace{\varepsilon, \dots, \varepsilon}_{s-1}, 0, \dots, 0, -1).$$

Hard Bandit Problem Instance

- $\mathcal{A} = \mathcal{S} \cup \mathcal{H}$ with a **low regret** action set $\mathcal{S}$ (sparse) and an **informative** action set $\mathcal{H}$ (half of the hypercube):

$$\mathcal{S} = \left\{ x \in \mathbb{R}^d \middle| x_j \in \{-1, 0, 1\} \text{ for } j \in [d-1], \|x\|_1 = s-1, x_d = 0 \right\},$$
$$\mathcal{H} = \left\{ x \in \mathbb{R}^d \middle| x_j \in \{-1, 1\} \text{ for } j \in [d-1], x_d = 1 \right\}.$$

- True parameter θ^*: for some small $\varepsilon > 0$

$$\theta^* = \big(\underbrace{\varepsilon, \ldots, \varepsilon}_{s-1}, 0, \ldots, 0, -1\big).$$

- **Sampling uniformly from the corners** of $\mathcal{H}$ (exploratory policy) ensures the covariance matrix is well-conditioned so that Lasso can be used for learning θ^* faster than OLS (more information), but suffer high regret due to the **last coordinate -1**.

Explore-Then-Commit

Theorem. Assume $\mathcal{A}$ spans $\mathbb{R}^d$. The regret upper bound of explore-the-sparsity-then-commit (ESTC) algorithm satisfies

$$\mathfrak{R}_{\theta^*}(n; \pi^{\text{ESTC}}) \lesssim C_{\min}^{-\frac{2}{3}}(\mathcal{A}) s^{\frac{2}{3}} n^{\frac{2}{3}}.$$

- Optimal in **data-poor regime** but sub-optimal in **data-rich regime**.

Algorithm.

1. ESTC finds the most informative design:

$$\pi_e = \max_{\mu \in \mathcal{P}(\mathcal{A})} \sigma_{\min}\Big(\int_{x \in \mathcal{A}} x x^\top d\mu(x)\Big).$$

2. Pull arms following π_e by n_1 rounds and compute the Lasso estimator $\widehat{\theta}_{n_1}$.
3. Execute the greedy action $A_t = \operatorname{argmax}_{x \in \mathcal{A}} \langle x, \widehat{\theta}_{n_1} \rangle$ for the remaining $n - n_1$ rounds.

Optimism-Based Algorithms

In general, optimism-based algorithms π^{opt} choose

$$A_t = \operatorname*{argmax}_{a \in \mathcal{A}} \max_{\widetilde{\theta} \in \mathcal{C}_t} \langle a, \widetilde{\theta} \rangle,$$

where $\mathcal{C}_t$ is some sparsity-aware confidence set.

- Optimal in **data-rich regime**. Online-to-confidence-set conversion approach [2] has $O(\sqrt{dsn})$ regret bound.
- Sub-optimal in **data-poor regime**. There exists a sparse linear bandit instance characterized by θ such that for the data-poor regime, we have

$$\mathfrak{R}_\theta(n; \pi^{\text{opt}}) \gtrsim n.$$

Q: Can we have an algorithm that is optimal in both regimes?

[2] Online-to-Confidence-Set Conversions and Application to Sparse Stochastic Bandits (2012).

Information Directed Sampling

Define $\mathfrak{BR}(n;\pi) = \mathbb{E}\left[\sum_{t=1}^{n}\langle x^*, \theta^*\rangle - \sum_{t=1}^{n} Y_t\right]$. IDS (Russo and Van Roy (2014)) balances the information gain about the optimal action and single-round regret.

Theorem.[3] For an arbitrary action set, the following regret bound holds

$$\mathfrak{BR}(n;\pi^{\text{IDS}}) \lesssim \sqrt{nds}\,.$$

When $\mathcal{A}$ is exploratory and has sparse optimal actions, the following regret bound holds

$$\mathfrak{BR}(n;\pi^{\text{IDS}}) \lesssim \min\left\{\sqrt{nds}, \frac{sn^{2/3}}{(2C_{\min}(\mathcal{A}))^{1/3}}\right\}.$$

IDS is nearly optimal in both regimes!

[3] Information Directed Sampling for Sparse Linear Bandits (2021).

Information Directed Sampling

Theorem.[4] For an arbitrary action set, the following regret bound holds

$$\mathfrak{BR}(n; \pi^{\mathrm{IDS}}) \lesssim \sqrt{nds}\,.$$

When $\mathcal{A}$ is exploratory and has sparse optimal actions, the following regret bound holds

$$\mathfrak{BR}(n; \pi^{\mathrm{IDS}}) \lesssim \min\left\{\sqrt{nds}, \frac{sn^{2/3}}{(2C_{\min}(\mathcal{A}))^{1/3}}\right\}.$$

IDS is nearly optimal in both regimes!

BONUS: efficient implementation is available through an empirical Bayesian approach for sparse posterior sampling.

[4] Information Directed Sampling for Sparse Linear Bandits (2021).

Open Problem

Q: Is $C_{\min}(\mathcal{A})$ the **fundamental quantity** to characterize the problem?

A: Perhaps not. If $\mathcal{A}$ is a full binary hypercube such that $C_{\min}(\mathcal{A}) = 1$, there exists an algorithm to achieve $O(s\sqrt{n})$ regret[5].

More finite-time instance-dependent analysis is needed!

[5] Linear multi-resource allocation with semi-banditfeedback (2015).

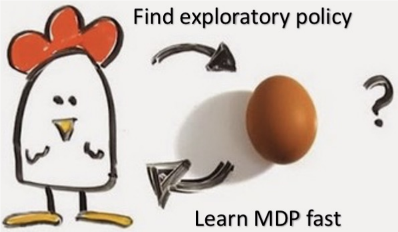
Story II: Online Sparse RL

HD Statistics v.s. Sparse Bandits v.s. Sparse RL

- **HD statistics.** "Best of both worlds": high representation power with many features while sparsity leads to efficient estimation.
- **Sparse bandits.** Existence of an exploratory policy $\Rightarrow$ dimension-free $\Theta(n^{2/3})$ rate. **Information and regret trade-off**.

HD Statistics v.s. Sparse Bandits v.s. Sparse RL

- **HD statistics.** "Best of both worlds": high representation power with many features while sparsity leads to efficient estimation.
- **Sparse bandits.** **Existence** of an exploratory policy $\Rightarrow$ dimension-free $\Theta(n^{2/3})$ rate. **Information and regret trade-off**.
- **Sparse RL.** Even though there **exists** an exploratory policy, **finding** the exploratory policy is also hard!



Episodic MDP

- States $\mathcal{X}$, actions $\mathcal{A}$, episode length H, transition kernel P, reward function r, policy π.
- Value function:

$$V_1^{\pi}(x) := \mathbb{E}^{\pi}\left[\sum_{h'=1}^{H} r(x_{h'}, a_{h'}) \middle| x_1 = x\right],$$

- Cumulative regret:

$$\mathfrak{R}(N; \pi) = \sum_{n=1}^{N} \left(V_1^*(x_1^n) - V_1^{\pi_n}(x_1^n)\right),$$

  where $V_1^*(\cdot)$ is the optimal value function, x_1^n is from some initial state distribution, N is the number of episodes.
- Sparse linear function approximation: $Q^{\pi}(x, a) \approx \phi(x, a)^{\top} w_{\pi}$ where $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ be a feature map, w_{π} is s-sparse.
- Earlier works focus on on-policy policy-evaluation, e.g. Lasso-TD (GLMH 2011).

Hardness of Online Sparse RL

Exploratory policy. We call a policy π *exploratory* if $\sigma_{\min}(\Sigma^\pi)$ is a constant, where

$$\Sigma^\pi := \mathbb{E}^\pi \left[\frac{1}{H} \sum_{h=1}^{H} \phi(x_h, a_h) \phi(x_h, a_h)^\top \right].$$

Hardness of Online Sparse RL

Exploratory policy. We call a policy π *exploratory* if $\sigma_{\min}(\Sigma^\pi)$ is a constant, where

$$\Sigma^\pi := \mathbb{E}^\pi \left[\frac{1}{H} \sum_{h=1}^{H} \phi(x_h, a_h) \phi(x_h, a_h)^\top \right] .$$

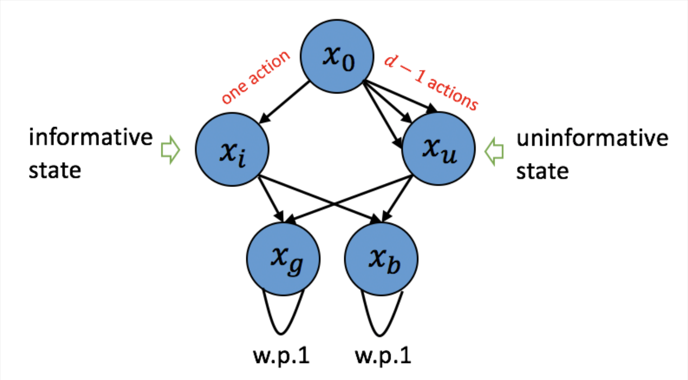
Theorem (Minimax Lower Bound). For any algorithm π, there **exists an exploratory policy** and a sparse linear MDP[6], such that for any $N \leq d$,

$$\mathfrak{R}(N; \pi) \geq \frac{1}{128} Hd \, .$$

This is in contrast to sparse linear bandits, where the existence of an exploratory policy is sufficient for dimension-free regret.

[6] The MDP kernel can be sparsely linear represented by the feature.

Hard MDP Problem Instance



- Only one of a large set of actions leading to the informative state **deterministically**.
- The exploratory policy has to visit that informative state to produce well-conditioned data

If We Have Oracle Access to an Exploratory Policy

Theorem (Regret Upper Bound) Consider a sparse linear MDP. Assume the learner has **oracle access** to an exploratory policy π_e. Online Lasso-fitted-Q-iteration can achieve a dimension-free sub-linear regret bound:

$$\mathfrak{R}(N;\pi) \lesssim H^{\frac{4}{3}} s^{\frac{2}{3}} N^{\frac{2}{3}}.$$

If We Have Oracle Access to an Exploratory Policy

Theorem (Regret Upper Bound) Consider a sparse linear MDP. Assume the learner has **oracle access** to an exploratory policy π_e. Online Lasso-fitted-Q-iteration can achieve a dimension-free sub-linear regret bound:

$$\mathfrak{R}(N; \pi) \lesssim H^{\frac{4}{3}} s^{\frac{2}{3}} N^{\frac{2}{3}}.$$

 # Online Lasso-FQI builds on the explore-then-commit template and uses Lasso to fit Q-function.

For exploratory policy:

- Only **existence** $\Rightarrow$ linear regret lower bound.
- **Existence** and **oracle access** $\Rightarrow$ sublinear regret upper bound.

Conclusion

- Exploiting sparsity in bandits and online RL is not as "easy" as in the high-dimensional statistics.
- Bandits: information and regret trade-off; RL: find exploratory policy without solving MDP.
- Future work: under what conditions, sparsity can help when minimizing regret in online RL?

Reference:

Botao Hao, Tor Lattimore, Mengdi Wang. High-Dimensional Sparse Linear Bandits (NeurIPS 2020).

Botao Hao, Tor Lattimore, Csaba Szepesvári, Mengdi Wang. Online Sparse Reinforcement Learning (AISTATS 2021).

Botao Hao, Tor Lattimore, Wei Deng. Information Directed Sampling for Sparse Linear Bandits (Under review).


thank you!