Stochastic Linear Bandits with Sparsity

Siyu Xie

Department of Statistics, Northwestern University

March 8, 2024

Outline

Outline

Sparse Linear Stochastic Bandits

At the beginning of round t, the learner receives a decision set $A_t \subset \mathbb{R}^d$. They then choose an action $A_t \in \mathcal{A}_t$ and receive a reward

$$X_t = \langle \theta_*, A_t \rangle + \eta_t \tag{1}$$

where η_t is zero-mean noise and $\theta_* \in \mathbb{R}^d$ is an unknown vector. But in this sparse setting, the parameter vector θ_* is assumed to have many zero entries.

The following hold:

1. (*Sparse parameter*) There exist known constants m_0 and m_2 such that $\|\theta_*\|_0 \leq m_0$ and $\|\theta_*\|_2 \leq m_2$.
2. (*Bounded mean rewards*) $\langle \theta_*, a \rangle \leq 1$ for all $a \in \mathcal{A}_t$ and all rounds t.
3. (*Subgaussian noise*) The noise is conditionally 1-subgaussian:

$$\text{for all } \lambda \in \mathbb{R}, \quad \mathbb{E}[\exp(\lambda \eta_t) | \mathcal{F}_{t-1}] \leq \exp(\lambda^2/2) a.s.$$

where $F_t = \sigma(A_1, X_1, \cdots, A_t, X_t, A_{t+1})$.

Outline

Elimination on the Hypercube

Warm-up

- Consider the case where the action set is the d-dimensional hypercube: $\mathcal{A}_t = \mathcal{A} = [-1, 1]^d$. Denote the true parameter vector by $\theta = \theta_*$.
  - The hypercube is notable as an action set because it enjoys perfect separability.
- For each dimension $i \in [d]$, the value of $A_{ti} \in [-1, 1]$ can be chosen independently of A_{tj} for $j \neq i$.
- The optimal action is $a^* = \text{sign}(\theta)$, where

$$\text{sign}(\theta)_i = \text{sign}(\theta_i) = \begin{cases} 1, & \text{if } \theta_i > 0; \\ 0, & \text{if } \theta_i = 0; \\ -1, & \text{if } \theta_i < 0. \end{cases}$$

Elimination on the Hypercube

- So learning the optimal action amounts to learning the sign of θ_i for each dimension.
  - A disadvantage of this structure is that in the worst case the sign of each θ_i must be learned independently, which in Chapter 24 we show leads to a worst-case regret of $R_n = \Omega(d\sqrt{n})$.
  - On the positive side, the separability means that θ_i can be estimated in each dimension independently while paying absolutely no price for this experimentation when $\theta_i = 0$.
- It turns out that this allows us to design a policy for which $R_n = O(\|\theta\|_0 \sqrt{n})$, even without knowing the value of $\|\theta\|_0$.

Elimination on the Hypercube

- Let $\mathcal{G}_t = \sigma(A_1, X_1, \ldots, A_t, X_t)$ be the σ-algebra containing information up to time $t-1$.
- Suppose $(A_{ti})_{i=1}^d$ are chosen to be conditionally independent given $\mathcal{G}_{t-1}$, and further assume for some specific $i \in [d]$ that A_{ti} is sampled from a Rademacher distribution so that $\mathbb{P}(A_{ti} = -1|\mathcal{G}_{t-1}) = \mathbb{P}(A_{ti} = 1|\mathcal{G}_{t-1}) = 1/2$. Then

$$\begin{aligned}
\mathbb{E}[A_{ti}X_t|\mathcal{G}_{t-1}] &= \mathbb{E}\left[A_{ti}\left(\sum_{j=1}^{d} A_{tj}\theta_j + \eta_t\right)\middle|\mathcal{G}_{t-1}\right] \\
&= \theta_i\mathbb{E}[A_{ti}^2|\mathcal{G}_{t-1}] + \sum_{j\neq i}\theta_j\mathbb{E}[A_{tj}A_{ti}|\mathcal{G}_{t-1}] + \mathbb{E}[A_{ti}\eta_t|\mathcal{G}_{t-1}] \\
&= \theta_i.
\end{aligned}$$

Idea of SETC

- This looks quite promising, but we should also check the variance.
- Using our assumptions that (η_t) is conditionally 1-subgaussian and that $\langle \theta, a \rangle \leq 1$ for all actions a, we have

$$\mathbb{V}[A_{ti} X_t | \mathcal{G}_{t-1}] = \mathbb{E}[A_{ti}^2 X_t^2 | \mathcal{G}_{t-1}] - \theta_i^2 = \mathbb{E}[(\langle \theta, A_t \rangle + \eta_t)^2 | \mathcal{G}_{t-1}] - \theta_i^2 \leq 2$$

- All the policy does is treat all dimensions independently.
- How long this takes depends on $|\theta_i|$, but note that if $|\theta_i|$ is small, then the price of exploring is also limited. The policy that results from this idea is called **selective explore-then-commit**.

Selective Explore-then-commit Algorithm (SETC)

Algorithm 13 Selective explore-then-commit

Input: n and d

Set $E_{1i} = 1$ and $C_{1i} = \mathbb{R}$ for all $i \in [d]$

for $t = 1, \ldots, n$ **do**

For each $i \in [d]$ sample $B_{ti} \sim \text{RADEMACHER}$

Choose action:

$$(\forall i) \qquad A_{ti} = \begin{cases} B_{ti} & \text{if } 0 \in C_{ti}, \\ 1 & \text{if } C_{ti} \subset (0, \infty], \\ -1 & \text{if } C_{ti} \subset [-\infty, 0). \end{cases}$$

Play A_t and observe X_t

Construct empirical estimators

$$(\forall i)\, T_i(t) = \sum_{s=1}^{t} E_{si} \qquad \hat{\theta}_{ti} = \frac{\sum_{s=1}^{t} E_{si} A_{si} X_s}{T_i(t)}$$

Elimination on the Hypercube

Algorithm 13 Selective explore-then-commit (Cont'd)

7: Construct confidence intervals:

$$(\forall i) \qquad W_{ti} = 2\sqrt{\left(\frac{1}{T_i(t)} + \frac{1}{T_i(t)^2}\right)\log\left(n\sqrt{2\,T_i(t)+1}\right)}$$

$$(\forall i) \qquad \mathcal{C}_{t+1,i} = [\hat{\theta}_{ti} - W_{ti}, \hat{\theta}_{ti} + W_{ti}]$$

8: Update exploration parameters:

$$(\forall i) \qquad E_{t+1,i} = \begin{cases} 0 & \text{if } 0 \notin \mathcal{C}_{t+1,i} \text{ or } E_{ti} = 0 \\ 1 & \text{otherwise.} \end{cases}$$

9: **end for**

Elimination on the Hypercube

SETC

Theorem 23.2

There exists a universal constants C, $C' > 0$ such that the regret of SETC satisfies

$$R_n \leq 3\|\theta\|_1 + C \sum_{i:\theta_i \neq 0} \frac{\log(n)}{|\theta_i|} \quad \textit{and} \quad R_n \leq 3\|\theta\|_1 + C'\|\theta\|_0 \sqrt{n \log(n)}$$

Lemma 23.3

Define $\tau_i = n \wedge \max\{t : E_{ti} = 1\}$, and let $F_i = \mathbb{I}\{\theta_i \notin \mathcal{C}_{\tau_i+1,i}\}$ be the event that θ_i is not in the confidence interval constructed at time τ_i. Then $\mathbb{P}(F_i) \leq 1/n$.

Elimination on the Hypercube

Proof of Theorem 23.2 Recall the definition of the regret. Using the fact that the optimal action is $a^* = \text{sign}(\theta)$, we have the following regret decomposition:

$$R_n = \sum_{t=1}^{n} \max_{a \in \mathcal{A}} \langle \theta, a \rangle - \mathbb{E}\left[\sum_{t=1}^{n} \langle \theta, A_t \rangle\right] = \sum_{i=1}^{d} \underbrace{\left(n|\theta_i| - \mathbb{E}\left[\sum_{i=1}^{n} A_{ti}\theta_i\right]\right)}_{R_{ni}}$$

If $\theta_i = 0$, then $R_{ni} = 0$. Hence, it suffices to bound R_{ni} for each i with $|\theta_i| > 0$.

Elimination on the Hypercube

Proof of Theorem 23.2 (Cont'd) Suppose that $|\theta_i| > 0$ for some i and the failure event F_i given in Lemma 23.3 does not occur. Then $\theta_i \in \mathcal{C}_{\tau_i+1,t}$, and by the definition of the algorithm, $A_{ti} = \text{sign}(\theta_i)$ for all $t \geq \tau_i$. Therefore,

$$R_n = n|\theta_i| - \mathbb{E}\left[\sum_{i=1}^{n} A_{ti}\theta_i\right] = \mathbb{E}\left[\sum_{i=1}^{n} |\theta_i|(1 - A_{ti}\text{sign}(\theta_i))\right]$$
$$\leq 2n|\theta_i|\mathbb{P}(F_i) + |\theta_i|\mathbb{E}[\mathbb{I}\{F_i^c\}\tau_i]$$

Proof of Theorem 23.2 (Cont'd)

$$R_n \leq 2n|\theta_i|\mathbb{P}(F_i) + |\theta_i|\mathbb{E}[\mathbb{I}\{F_i^c\}\tau_i]$$

Since τ_i is the first round t when $0 \notin \mathcal{C}_{t+1}$, it follows that if F_i does not occur, then $\theta_i \in \mathcal{C}_{\tau_i,i}$ and $0 \in \mathcal{C}_{\tau_i,i}$. Thus, the width of the confidence interval $\mathcal{C}_{\tau_i,i}$ must be at least $|\theta_i|$, and so

$$2W_{\tau_i-1,i} = 4\sqrt{\left(\frac{1}{\tau_i - 1} + \frac{1}{(\tau_i - 1)^2}\right)\log(n\sqrt{2\tau_i - 1})} \geq |\theta_i|$$

Elimination on the Hypercube

Proof of Theorem 23.2 (Cont'd) after rearranging, this shows for some universal constant $C > 0$ that

$$\mathbb{I}\{F_i^c\}(\tau_i - 1) \leq 1 + \frac{C \log(n)}{\theta_i^2}.$$

combining this result with

$$R_n \leq 2n|\theta_i|\mathbb{P}(F_i) + |\theta_i|\mathbb{E}[\mathbb{I}\{F_i^c\}\tau_i]$$

leads to

$$R_n \leq 2n|\theta_i|\mathbb{P}(F_i) + |\theta_i| + \frac{C \log(n)}{|\theta_i|},$$

where Lemma 23.3 bounds $\mathbb{P}(F_i)$.

Outline

Online to Confidence Set Conversion

- A new plan is needed to relax the assumption that the action set is a hypercube. The idea is to modify the ellipsoidal confidence set used in stochastic linear bandits (Th 20.5) to have a smaller radius
- We will see that modifying the algorithm in Chapter 19 to use the smaller confidence intervals improves the regret to $R_n = O(\sqrt{dpn}\log(n))$.

Online to Confidence Set Conversion

- The prediction problem considered is **online linear prediction** under the squared loss.
  - Also known as **online linear regression**
- The learner interacts with an environment in a sequential manner where in each round $t \in N^+$:
  1. The environment chooses $X_t \in \mathbb{R}$ and $A_t \in \mathbb{R}^d$ in an arbitrary fashion.
  2. The value of A_t is revealed to the learner (but not X_t).
  3. The learner produces a real-valued prediction $\hat{X}_t \in \mathbb{R}$ in some way.
  4. The environment reveals X_t to the learner and the loss is $\left(X_t - \hat{X}_t\right)^2$.

Online to Confidence Set Conversion

- The regret of the learner relative to a linear predictor that uses the weights $\theta \in \mathbb{R}^d$ is

$$\rho_n(\theta) = \sum_{t=1}^{n} \left(X_t - \hat{X}_t\right)^2 - \sum_{t=1}^{n} (X_t - \langle \theta, A_t \rangle)^2$$

- We say that the learner enjoys a regret guarantee B_n relative to $\Theta \subseteq \mathbb{R}^d$ if for any strategy of the environment,

$$\sup_{\theta \in \Theta} \rho_n(\theta) \leq B_n.$$

Online to Confidence Set Conversion

- First, we show how to use such a learning algorithm to construct a confidence set.
- Take any learner for online linear regression, and assume the environment generates X_t in a stochastic manner like in linear bandits:

$$X_t = \langle \theta_*, A_t \rangle + \eta_t.$$

With elementary algebra,

$$\begin{aligned} Q_t &= \sum_{t=1}^{n} \left(\hat{X}_t - \langle \theta_*, A_t \rangle \right)^2 = \rho_n(\theta_*) + 2 \sum_{t=1}^{n} \eta_t \left(\hat{X}_t - \langle \theta_*, A_t \rangle \right) \\ &\leq B_n + 2 \sum_{t=1}^{n} \eta_t \left(\hat{X}_t - \langle \theta_*, A_t \rangle \right), \end{aligned}$$

where the first equality serves as the definition of Q_t.

Online to Confidence Set Conversion

- The fluctuations can be controlled with high probability using a little concentration analysis. Let

$$Z_t = \sum_{s=1}^{t} \eta_s \left(\hat{X}_s - \langle \theta_*, A_s \rangle \right)$$

- Since $\hat{X}_t$ is chosen based on information available at the beginning of the round, $\hat{X}_t$ is $\mathcal{F}_{t-1}$-measurable, and so

$$\text{for all } \lambda \in \mathbb{R}, \quad \mathbb{E}\left[\exp\left(\lambda\left(Z_t - Z_{t-1}\right)\right) \mid \mathcal{F}_{t-1}\right] \leq \exp\left(\lambda^2 \sigma_t^2 / 2\right),$$

where $\sigma_t^2 = \left(\hat{X}_t - \langle \theta_*, A_t \rangle \right)^2$.

Online to Confidence Set Conversion

- The uniform self-normalised tail bound (Theorem 20.4) with $\lambda = 1$ implies that

$$\mathbb{P}\left(\text{exists } t \geq 0 \text{ such that } |Z_t| \geq \sqrt{(1+Q_t)\log\left(\frac{1+Q_t}{\delta^2}\right)} \right) \leq \delta.$$

- Provided this low-probability event does not occur, then we have

$$Q_t \leq B_t + 2\sqrt{(1+Q_t)\log\left(\frac{1+Q_t}{\delta^2}\right)}.$$

  - While both sides depend on Q_t, the left-hand side grows linearly, while the right-hand side grows sublinearly in Q_t. This means that the largest value of Q_t that satisfies the above inequality is finite.

Online to Confidence Set Conversion

- A tedious calculation then shows this value must be less than

$$\beta_t(\delta) = 1 + 2B_t + 32 \log\left(\frac{\sqrt{8} + \sqrt{1 + B_t}}{\delta}\right).$$

- By piecing together the parts, we conclude that with probability at least $1 - \delta$ the following holds for all t :

$$Q_t = \sum_{s=1}^{t} \left(\hat{X}_s - \langle \theta_*, A_s \rangle\right)^2 \leq \beta_t(\delta)$$

Online to Confidence Set Conversion

- We could define $\mathcal{C}_{t+1}$ to be the set of all θ such that the above holds with θ_* replaced by θ.
- But there is one additionally subtlety, which is that the resulting confidence interval may be unbounded (think about the case that $\sum_{s=1}^{t} A_s A_s^T$ is not invertible).
- In Chapter 19 we overcame this problem by regularising the least squares estimator. Since we have assumed that $\|\theta_*\|_2 \leq m_2$, the previous display implies that

$$\|\theta_*\|_2^2 + \sum_{s=1}^{t} \left(\hat{X}_s - \langle \theta_*, A_s \rangle\right)^2 \leq m_2^2 + \beta_t(\delta)$$

- All together, we have the following theorem:

Online to Confidence Set Conversion

Theorem 23.4

Let $\delta \in (0,1)$ and assume that $\theta_* \in \Theta$ and $\sup_{\theta \in \Theta} \rho_t(\theta) \leq B_t$. If

$$\mathcal{C}_{t+1} = \left\{ \theta \in \mathbb{R}^d : \|\theta\|_2^2 + \sum_{s=1}^{t} \left(\hat{X}_s - \langle \theta, A_s \rangle \right)^2 \leq m_2^2 + \beta_t(\delta) \right\},$$

then $\mathbb{P}$ (exists $t \in \mathbb{N}$ such that $\theta_* \notin \mathcal{C}_{t+1}$) $\leq \delta$.

Online to Confidence Set Conversion

- The confidence set in Theorem 23.4 is not in the most convenient form. By defining $V_t = I + \sum_{s=1}^{t} A_s A_s^T$ and $S_t = \sum_{s=1}^{t} A_s \hat{X}_s$ and $\hat{\theta}_t = V_t^{-1} S_t$ and performing some algebraic calculation, one can see that

$$\|\theta\|_2^2 + \sum_{s=1}^{t} \left(\hat{X}_s - \langle \theta, A_s \rangle\right)^2 = \left\|\theta - \hat{\theta}_t\right\|_{V_t}^2 + \sum_{s=1}^{t} \left(\hat{X}_s - \left\langle \hat{\theta}_t, A_s \right\rangle\right)^2 + \left\|\hat{\theta}_t\right\|_2^2.$$

- Using this, the confidence set can be rewritten in the familiar form of an ellipsoid:

$$\mathcal{C}_{t+1} = \left\{ \theta \in \mathbb{R}^d : \left\|\theta - \hat{\theta}_t\right\|_{V_t}^2 \leq m_2^2 + \beta_t(\delta) - \left\|\hat{\theta}_t\right\|_2^2 - \sum_{s=1}^{t} \left(\hat{X}_s^2 - \left\langle \hat{\theta}_t, A_s \right\rangle\right)^2 \right\}.$$

Algorithm 14 Online linear predictor UCB (OLR-UCB)

Input: Online linear predictor and regret bound B_t, confidence parameter $\delta \in (0, 1)$

1: **for** $t = 1, \ldots, n$ **do**

2: Receive action set $\mathcal{A}_t$

3: Computer confidence set:

$$\mathcal{C}_t = \left\{ \theta \in \mathbb{R}^d : \|\theta\|_2^2 + \sum_{s=1}^{t-1} \left(\hat{X}_s - \langle \theta, A_s \rangle \right)^2 \leq m_2^2 + \beta_t(\delta) \right\}$$

4: optimistic action

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

5: Feed A_t to the online linear predictor and obtain prediction $\hat{X}_t$

6: Play A_t and receive reward X_t

7: Feed X_t to online linear predictor as feedback

8: **end for**

Online to Confidence Set Conversion

- It is not obvious that $\mathcal{C}_{t+1}$ is not empty because the radius could be negative.
- Theorem 23.4 shows, however, that with high probability $\theta_* \in \mathcal{C}_{t+1}$.
- At last we have established all the conditions required for Theorem 19.2, which implies the following theorem bounding the regret of Algorithm 14:

Theorem 23.5

With probability at least $1 - \delta$ the pseudo-regret of OLR-UCB satisfies

$$\hat{R}_n \leq \sqrt{8dn\left(m_2^2 + \beta_{n-1}(\delta)\right)\log\left(1 + \frac{n}{d}\right)}$$

Outline

Sparse Online Linear Prediction

Theorem 23.6

There exists a strategy π for the learner such that for any $\theta \in \mathbb{R}^d$, the regret $\rho_n(\theta)$ of π against any strategic environment such that $\max_{t\in[n]} \|A_t\|_2 \leq L$ and $\max_{t\in[n]} |X_t| \leq X$ satisfies

$$\rho_n(\theta) \leq cX^2\|\theta\|_0 \left\{ \log\left(e + n^{1/2}L\right) + C_n \log\left(1 + \frac{\|\theta\|_1}{\|\theta\|_0}\right)\right\} + \left(1 + X^2\right) C_n$$

where $c > 0$ is some universal constant and $C_n = 2 + \log_2 \log\left(e + n^{1/2}L\right)$.

Sparse Online Linear Prediction

- Note that $C_n = O(\log\log(n))$, so by dropping the dependence on X and L, we have

$$\sup_{\theta:\|\theta\|_0\le m_0,\|\theta\|_2\le L} \rho_n(\theta) = O\left(m_0 \log(n)\right).$$

- As a final catch, the rewards (X_t) in sparse linear bandits with subgaussian noise are not necessarily bounded. However, the subgaussian property implies that with probability $1-\delta$, $|\eta_t| \le \log(2/\delta)$. By choosing $\delta = 1/n^2$ and Assumption 23.1, we have

$$\mathbb{P}\left(\max_{t\in[n]} |X_t| \ge 1 + \log\left(2n^2\right)\right) \le \frac{1}{n}$$

Sparse Online Linear Prediction

- Putting all the pieces together shows that the expected regret of OLR-UCB when using the predictor provided by Theorem 23.6 and when $\|\theta\|_0 \leq m_0$ satisfies

$$R_n = O\left(\sqrt{dnm_0}\log(n)^2\right)$$

Outline

Notes

1. The strategy achieving the bound in Theorem 23.6 is not computationally efficient. In fact we do not know of any polynomial time algorithm with logarithmic regret for this problem. The consequence is that **OLR-UCB does not yet have an efficient implementation**.
2. **While we focused on the sparse case, the results and techniques apply to other settings.** For example, we can also get alternative confidence sets from results in online learning even for the standard non-sparse case. Or one may consider additional or different structural assumptions on θ.

Notes

3. **When the online linear regression results are applied, it is important to use the tightest possible, data-dependent regret bounds** B_n. In online learning most regret bounds start as tight, data-dependent bounds, which are then loosened to get further insight into the structure of problems. The gains from using data-dependent bounds can be significant.

Notes

4. **The confidence set used by OLR-UCB depends on the sparsity parameter m_0, which must be known in advance.** No algorithm can enjoy a regret of $O\left(\sqrt{\|\theta_*\|_0 \, dn}\right)$ for all $\|\theta_*\|_0$ simultaneously (see Chapter 24).
5. **The bound in Theorem 23.5 still depends on the ambient dimension. In general this is unavoidable.** For this reason it recently became popular to study the contextual setting with changing actions and make assumptions on the distribution of the contexts so that techniques from high-dimensional statistics can be brought to bear. We can refer to papers by Kim and Paik [2019] and *Bastani and Bayati [2020]*.