Lecture 11
Page 1

CS 5 9000 – RL
Linear bandit
Agenda
- Linear regression
- Regret bound

---

* **Theorem**: For $\delta \in (0,1)$, with probability at least $1-\delta$, for any $t \in [n]$, we have:

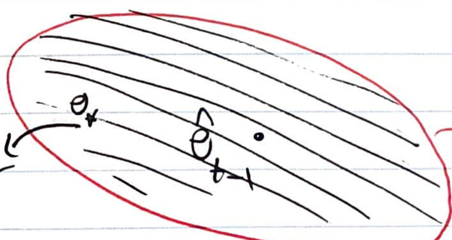
$$\|\hat{\theta}_t - \theta_*\|_{V_t(V)} \leq \sqrt{\beta_t(\delta)} ; \sqrt{\lambda}\,\|\theta_*\| + \sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det V_t(V)}{\det(V)}\right)}$$

for $V = \lambda I$.

Furthermore, if $\|\theta_*\| \leq S$, define confidence interval/set

$$C_t(\delta) = \left\{\theta \in \mathbb{R}^d : \|\hat{\theta}_{t-1} - \theta\|_{V_{t-1}(V)} \leq \sqrt{\lambda}\, S + \sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(V_{t-1}(V))}{\det V}\right)}\right.$$

Then, $\mathbb{P}(\text{exist } t \in [n] ;\ \theta_* \notin C_t(\delta)) \leq \delta$




$\to C_t(\delta)$

true parameter lives in it.

$$\|\theta - \hat{\theta}_{t-1}\|^2_{V_{t-1}(V)} \leq \beta_t(\delta)$$

Note: Setting $V = \lambda I \to \det(V) = \lambda^d$

Can we simplify $\det(V_t(\lambda I))$ ?

Remember that $V_t(\lambda) = \lambda I + \sum_{s=1}^{t} A_s A_s^T$ for $1 \leq s \leq t$

Assume that all $a \in D_s$ ; $\|a\| \leq L$ for all $1 \leq s \leq t$.
Also note that $V_t(\lambda)$ is positive definite matrix (why?)

Let $\varsigma_1, \ldots, \varsigma_d$ denote the eigenvalues of $V_t(\lambda)$.

Therefore, $\det(V_t(\lambda)) = \prod_{i=1}^{d} \varsigma_i$ and the $\operatorname{trace}(V_t(\lambda)) = \sum_{i=1}^{d} \varsigma_i$ . By inequality of arthmatic and geometric mean of positive numbers we have:

$$\sqrt[d]{\prod_{i=1}^{d} \varsigma_i} \leq \frac{\sum_{i=1}^{d} \varsigma_i}{d}$$

Therefore,

$$\Rightarrow \det(V_t(\lambda)) \leq \left(\frac{\operatorname{trace}(V_t(\lambda))}{d}\right)^d$$

Let's simplify the trace:

we Know that, $trace(V_t(\lambda)) = trace(\lambda I) + \sum_{s=1}^{t} trace(A_s A_s^T)$

$$= d\lambda + \sum_{s=1}^{t} \|A_s\|_2^2 \leq d\lambda + tL^2$$

⊗ $\log(\det(V_t(\lambda))) \leq d \log\left(\lambda + \frac{tL^2}{d}\right)$

Now using these simplifications:

$$\beta_t(\delta) \leq \sqrt{\lambda}\, S + \sqrt{2\log\left(\frac{1}{\delta}\right) + d\log\left(\lambda + \frac{tL^2}{d}\right) - d\log(\lambda)}$$

$$\leq \sqrt{\lambda}\, S + \sqrt{2\log\left(\frac{1}{\delta}\right) + d\log\left(\lambda + \frac{nL^2}{d}\right) - d\log(\lambda)}$$

---

Proof of theorem ✱:

$$\hat{\theta}_t = V_t(\lambda)^{-1} \sum_{s=1}^{t} A_s X_s$$

remember that $X_s = \langle A_s, \theta_* \rangle + \eta_s$, therefore

$$\hat{\theta}_t = V_t(\lambda)^{-1}\left(\sum_{s=1}^{t} A_s A_s^T \theta_* + \underbrace{\sum_{s=1}^{t} A_s \eta_s}_{S_t}\right)$$

$$= V_t(\lambda)^{-1} V_t \theta_* + V_t(\lambda)^{-1} S_t$$

Using this equality, we have:

$$\|\hat{\theta}_t - \theta^*\|_{V_t(\lambda)} = \| V_t(\lambda)^{-1} S_t + (\underbrace{V_t(\lambda)^{-1} V_t}_{V_t(\lambda)^{-1} V_t(\lambda)} - I)\theta_* \|_{V_t(\lambda)}$$

Using the fact that $\|\circ\|_{V_t(\lambda)}$ is a norm, and triangle inequality

$$\|\hat{\theta}_t - \theta_*\|_{V_t(\lambda)} \leqslant \|S_t\|_{V_t^{-1}(\lambda)} + \|(V_t(\lambda)^{-1} V_t - I)\theta_*\|_{V_t(\lambda)}$$

$$= \|S_t\|_{V_t(\lambda)^{-1}} + \|\lambda V_t(\lambda)^{-1} \theta_*\|_{V_t(\lambda)}$$

$$\leqslant \|S_t\|_{V_t(\lambda)^{-1}} + \lambda^{\frac{1}{2}} \|\theta_*\|.$$

From the past we know that

$$P\left(t \in [n]; \|S_t\|_{V_t(\lambda)^{-1}} \geqslant 2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(V_t(\lambda))}{\det(V)}\right)\right) \leqslant \delta$$

Ergo $\|\hat{\theta}_t - \theta_*\|_{V_t(\lambda)} \leqslant 2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(V_t(\lambda))}{\det(\lambda I)}\right) + \underbrace{\lambda^{\frac{1}{2}}\|\theta_*\|}_{\leqslant S}$

which the statement of the theorem.

We proved this theorem for almost any sequence $A_t$.

---

stochastic linear bandit:

- At each time step $t$, the agent is given a decision set $D_t$, from which it needs to choose an action. The reward of choosing $A_t \in D_t$ is as follows:

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

---

what is the oracle's expected reward at time $t$?

$$\max_{a \in D_t} \langle a, \theta_* \rangle .$$ : Note that this can be random!

Since $D_t$ can be random, or adversarially chosen, let $\hat{R}_n$ denote random regret defined as follows:

$$\hat{R}_n = \sum_{t=1}^{n} \max_{a \in D_t} \langle a, \theta_* \rangle - \sum_{t=1}^{n} X_t$$

---

Consider a setting where $|\langle a, \theta_* \rangle| < 1$ and $\|a\| < L$ for all $a \in \cup_t D_t$.

Lin UCB (LinRel, OFUL)

Psedu code of Lin UCB

- At time step $t$, comput $\hat{\theta}_{t-1} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \left( \sum_{s=1}^{t-1} \left(X_s - \langle A_s, \theta \rangle\right)^2 + \lambda \|\theta\|_2^2 \right)$

the estimate we have at the begining of time $t$

i.e. $\hat{\theta}_{t-1} = V_{t-1}(\lambda)^{-1} \sum_{s=1}^{t-1} A_s X_s$

- Construct the confidene st $C_t(\delta)$

$$C_t(\delta) = \{\theta \in \mathbb{R}^d ; \|\theta - \hat{\theta}_{t-1}\|^2_{V_{t-1}(\lambda)} \leq \beta_t(\delta)\}$$

- Optimism step: Choose an optimal arm of the most optimistic model

$$A_t = \underset{a \in D_t}{\operatorname{argmax}} \max_{\theta \in C_t} \langle a, \theta \rangle$$

and $\tilde{\theta}_t$ is the corresponding optimitc model

---

Theorem: The regret of Lin UCB satifies:

$$\hat{R}_A \leq \sqrt{8n\, \beta_n(\delta) \log\left(\frac{\det V_n(\lambda)}{\det V}\right)}$$

with probability at least $1-\delta$.