6.2 Linear Bandits - Handling Large Action Spaces

Let D be a compact set of decision. On each round, we must choose $x_t \in D$, each of which results in a reward r_t.

Pseudo code : The Linear UCB algorithm

Input : λ, β_t

1: for $t = 0, 1, \cdots$ do

2: Execute $x_t = \underset{x \in D}{\operatorname{argmax}} \max_{\mu \in BALL_t} \mu \cdot x$

and observe r_t

3: Update $BALL_{t+1}$

4: End for.

Note:

① We assume $E[r_t | x_t = x]$ is a fixed linear function,

i.e. $E[r_t | x_t = x] = \mu^* \cdot x \in [-1, 1]$ for all $x \in D$

② We assume $E[r_t] \in [-1, 1]$. Then the noise sequence

$\eta_t = r_t - \mu^* \cdot x_t$ is a martingale difference seq.

③ If $x_0, \cdots, x_{T-1}$ are the decisions made in the game,

the cumulative regret is defined by

$$R_T = T\mu^* \cdot x^* - \sum_{t=0}^{T-1} \mu^* \cdot x_t$$

where $x^* \in D$ is an optimal decision for μ^*,

i.e. $x^* \in \underset{x \in D}{\operatorname{argmax}} \, \mu^* \cdot x$

x^* exists since D is compact

④ Our goal is to keep R_T as small as possible.

1. The LinUCB algorithm.

At episode t, we define an uncertain region $BALL_t$.

The center of $BALL_t$ is $\hat{\mu}_t$, which is the solution of:

$$\hat{\mu}_t = \arg\max_{\mu} \sum_{\tau=0}^{t-1} \|\mu \cdot x_\tau - r_\tau\|_2^2 + \lambda\|\mu\|_2^2$$

$$= \Sigma_t^{-1} \sum_{\tau=0}^{t-1} r_\tau x_\tau$$

where λ is the parameter and Σ_t satisfies:

$$\Sigma_t = \lambda I + \sum_{\tau=0}^{t-1} x_\tau x_\tau', \quad \Sigma_0 = \lambda I$$

Then $BALL_t$ is defined as

$$BALL_t = \{\mu \mid (\hat{\mu}_t - \mu)' \Sigma_t (\hat{\mu}_t - \mu) \le \beta_t\}$$

2. Upper and lower bounds

Assume $x \in \mathbb{R}^d$

> Theorem 6.3. Suppose that the expeted costs are bounded by 1, i.e. $|\mu^* \cdot x| \le 1$ for all $x \in D$; $\|\mu^*\| \le W$, $\|x\| \le B$ for all $x \in D$, and that η_t is σ^2 sub-Guassian. Set
>
> $$\lambda = \sigma^2 / W^2, \quad \beta_t := \sigma^2\left(2 + 4d\log\left(1 + \frac{tB^2W^2}{d}\right) + 8\log(4/\delta)\right)$$
>
> we have that with prob. $\ge 1-\delta$, for all $T \ge 0$
>
> $$R_T \le c\sigma\sqrt{T}\left(d\log\left(1 + \frac{TB^2W}{d\sigma^2}\right) + \log(4/\delta)\right),$$
>
> where c is an absolute constant. i.e. $R_T \sim O(d\sqrt{T})$

Following shows R_T of Th. 6.3 is best.

> Theorem 6.4 [lower bound] There exist a distribution ove linear bandit problems (i.e. $\Delta(\mu)$) with rewards bounded by 1, in martingale, and $\sigma^2 \le 1$, s.t. for every algorithm, we have for $n \ge \max\{256, d^2/16\}$,
>
> $$E_\mu E[R_T] \ge \frac{1}{2500} d\sqrt{T}$$

w.r.t. randomness in the problem & algorithm.

We will eliminate the dependencies in Th 6.3.

Let Σ_D denote the D-optimal design matrix from :

Th 3.2 : Suppose $\mathcal{X} \subset \mathbb{R}^d$ is a compact set. There exists a distribution ρ on $\mathcal{X}$ s.t.

① ρ is supported on at most $d(d+1)/2$ points

② Define $\Sigma = E_{x\sim\rho}[xx^T]$, we have

$$\|x\|^2_{\Sigma^{-1}} \le d$$

and ρ is referred to as the D-optimal design.

Coordinate transformation:

$$\tilde{x} = \Sigma_D^{-1/2} x, \quad \tilde{\mu}^* = \Sigma_D^{1/2} \mu^*$$

s.t. $\tilde{x} \cdot \tilde{\mu}^* = x' \Sigma_D^{-1/2} \cdot \Sigma_D^{1/2} \mu^* = x \cdot \mu^*$

we hold the expected reward function.

We have $\underset{\Uparrow B}{\|\tilde{x}\|^2} = \|x\|^2_{\Sigma_D^{-1}} \le d$ (by Th. 3.2)

(assume $\|r\| \le 1$)

and $\underset{\Uparrow W}{\|\tilde{\mu}^*\|} = \|\mu^*\|_{\Sigma_D} = \sqrt{(\mu^*)' \Sigma_D \mu^*} = \sqrt{E_{x\sim\rho}[(\mu^* \cdot x)^2]} \le 1$

Following shows that we could remove the dependencies on B and W from previous theorem, due to $B \le \sqrt{d}$ and $W \le 1$.

Corollary 6.5 Suppose that the expected rewards are bounded in martingale by 1, $\|\tilde{\mu}^* \cdot \tilde{x}\| \le 1$ for all $x \in D$; and that η_t is σ^2 sub-Gaussian. Suppose linUCB is implemented in the $\tilde{x}$ coordinate system, with following settings:

$$\lambda = \sigma^2, \quad \beta_t := \sigma^2 \left(2 + 4d\log(1+t) + 8\log(4/\delta) \right)$$

with prob. $\geq 1-\delta$, for all $T \geq 0$,

$$R_T \leq c\sigma\sqrt{T}\left(d\log\left(1+\frac{T}{\sigma^2}\right) + \log(4/\delta) \right)$$