Linear Bandits

Least squares review

Suppose I measure $x_1, \dots, x_T \in \mathbb{R}^d$ then observe

$y_t = \langle x_t, \theta_* \rangle + \xi_t$ where ξ_t is mean-zero σ^2-sub-Gaussian

$$\hat{\theta} = \operatorname{argmin}_\theta \sum_{t=1}^{T} (y_t - \langle x_t, \theta \rangle)^2$$

$$= \operatorname{argmin}_\theta \|y - X\theta\|_2^2$$

$$= (X^T X)^{-1} X^T y$$

$$= (X^T X)^{-1} X^T (X\theta_* + \xi)$$

$$= \theta_* + (X^T X)^{-1} X^T \xi$$

$$X = \begin{bmatrix} - x_1^T - \\ \vdots \\ - x_T^T - \end{bmatrix} \in \mathbb{R}^{T \times d} \qquad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_T \end{bmatrix} \in \mathbb{R}^T$$

$W = X(X^T X)^{-1}$ then $\hat{\theta} = \theta_* + W^T \xi$

$$\mathbb{E}[(\hat{\theta} - \theta_*)(\hat{\theta} - \theta_*)^T] = \mathbb{E}[(X^T X)^{-1} X^T \xi \xi^T X (X^T X)^{-1}]$$

$$= (X^T X)^{-1} X^T \mathbb{E}[\xi \xi^T] X (X^T X)^{-1}$$

$$\preceq \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1}$$

$$= \sigma^2 (X^T X)^{-1}$$

Let $z \in \mathbb{R}^d$ be some new point.

$$\mathbb{E}[\langle z, \hat{\theta} \rangle] = \langle z, \theta_* \rangle \qquad \mathbb{V}(\langle z, \hat{\theta} \rangle) = \mathbb{E}[\langle z, \hat{\theta} - \theta_* \rangle^2]$$

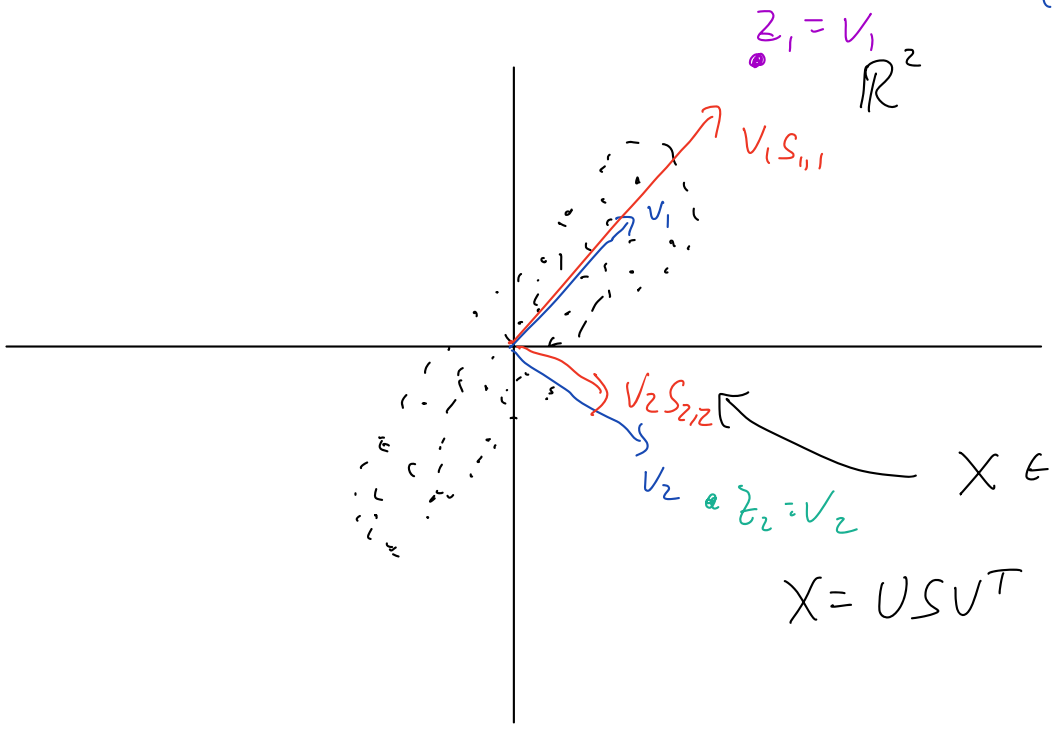
$$= \mathbb{E}[z^T (\hat{\theta} - \theta_*)(\hat{\theta} - \theta_*)^T z]$$

For any PSD $A \in \mathbb{R}^{d \times d}$

Define $\|x\|_A^2 = x^T A x$.

$= \sigma^2 z^T (X^T X)^{-1} z$.

$= \sigma^2 \|z\|^2_{(X^T X)^{-1}}$



$X = USV^T$

$U \in \mathbb{R}^{T \times 2}$

$S \in \mathbb{R}^{2 \times 2}$

$V = \mathbb{R}^{2 \times 2}$

$$U = \begin{bmatrix} - & u_1^T & - \\ & \vdots & \\ - & u_T & - \end{bmatrix}$$

$U^T U = I_2$

$V^T V = I_2$

$$V = \begin{bmatrix} 1 & 1 \\ v_1 & v_2 \\ 1 & 1 \end{bmatrix}$$

$x_i = u_i^T S V$

$= u_{i,1} s_{11} v_1 + u_{i,2} s_{22} v_2$

$(X^T X)^{-1} = (V S U^T U S V^T)^{-1}$

$= (V S^2 V^T)^{-1}$

$= V S^{-2} V^T$

V is orthonormal

$V^{-1} = V^T$

If $z = v_1$ then

$$z_1^T (X^T X)^{-1} z_1 = v_1^T V S^{-2} V^T v_1$$

$$= e_1^T S^{-2} e_1$$

$$= \frac{1}{s_{1,1}^2}$$

If $z_2 = v_2$

$$z_2^T (X^T X)^{-1} z_2 = \frac{1}{s_{2,2}^2}$$

How do I choose $x_1, \dots, x_T \in \mathcal{X}$ to minimize $\mathbb{E}[\langle z, \hat{\theta} - \theta_* \rangle^2]$?

Suppose I have a pool of potential measurements $\mathcal{X} \subset \mathbb{R}^d$.

For simplicity assume $|\mathcal{X}| = n$.

For any $X^T X = \sum_{i=1}^T x_i x_i^T$ $\exists \lambda \in \Delta_{\mathcal{X}} = \{w_i \geq 0 : \sum_{i=1}^n w_i = 1\}$ s.t. $X^T X = T \sum_{x \in \mathcal{X}} \lambda_x x x^T$.

$$=: T \cdot A(\lambda)$$

$$\boxed{A(\lambda) = \sum_{x \in \mathcal{X}} \lambda_x x x^T}$$

(Always exists by $\lambda_x := \frac{\sum_{i=1}^T \mathbb{1}\{x_i = x\}}{T}$)

$$\mathbb{E}[\langle z, \hat{\theta} - \theta_* \rangle^2] \leq \sigma^2 z^T (X^T X)^{-1} z$$

$$= \sigma^2 \cdot \frac{z^T A(\lambda)^{-1} z}{T}.$$

$$A(\lambda) = \mathbb{E}_{x \sim \lambda}[x x^T]$$

Idea: Choose $\lambda \in \Delta_{\mathcal{X}}$ to minimize RHS! Before seeing any data then choose $x_1, \dots, x_T$ "in proportion" to $\lambda \in \Delta_{\mathcal{X}}$.

A-optimality

$$\sum_{i=1}^{d} \mathbb{E}\left[\langle e_i, \hat{\theta} - \theta_* \rangle^2\right] = \sum_{i=1}^{d} \mathbb{E}\left[(\hat{\theta}-\theta_*)^T e_i e_i^T (\hat{\theta}-\theta_*)\right]$$

$$= \mathbb{E}\left[(\hat{\theta}-\theta_*)^T (\hat{\theta}-\theta_*)\right]$$

$$= \mathbb{E}\left[\|\hat{\theta}-\theta_*\|_2^2\right]$$

$$\mathbb{E}\left[\|\hat{\theta}-\theta_*\|_2^2\right] = \sum_{i=1}^{d} \mathbb{E}\left[\langle e_i, \hat{\theta}-\theta_*\rangle^2\right] \qquad [C_i]_j = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$$

$$\leq \sum_{i=1}^{d} \frac{\sigma^2 \; e_i^T A(\lambda)^{-1} e_i}{T}$$

$$= \frac{\sigma^2}{T} \cdot \text{Trace}\left(A(\lambda)^{-1}\right).$$

A-optimal design: $\underset{\lambda \in \Delta_K}{\text{argmin}} \; \text{Trace}\left(A(\lambda)^{-1}\right)$.

E-optimal

$$\max_{u:\|u\|_2 \leq 1} \mathbb{E}\left[\langle u, \hat{\theta}-\theta_*\rangle^2\right] \leq \frac{\sigma^2}{T} \max_{u:|u|\leq 1} u^T A(\lambda)^{-1} u$$

$$= \frac{\sigma^2}{T} \lambda_{max}\left(A(\lambda)^{-1}\right)$$

$$= \frac{\sigma^2}{T} \cdot \frac{1}{\lambda_{min}\left[A(\lambda)\right]}$$

(e, v_i) are eigen pairs

$Av_i = e_i v_i$

$$f_E(\lambda) = \lambda_{max}\left(A(\lambda)^{-1}\right)$$

D-optimal

$$g_D(\lambda) = \log|A(\lambda)|$$

λ^* is D-optimal if $\lambda^* = \underset{\lambda}{\text{argmax}} \; g_D(\lambda)$.

D-optimal maximizes product of eigs of $A(\lambda)$.

$\mathcal{N}(\mu, \Sigma)$ has density $\frac{1}{(2\pi|\Sigma|)^{d/2}} \exp\left(-(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$.

G-optimal

$$f_G(\lambda) = \max_{x \in \mathcal{X}} \|x\|^2_{A(\lambda)^{-1}} = \max_{x \in \mathcal{X}} x^T A(\lambda)^{-1} x$$

G-opt minimizes $\max_{x \in \mathcal{X}} \mathbb{E}[\langle x, \hat{\theta} - \theta_* \rangle^2]$.

Kiefer-Wolfowitz Equivalence Theorem

Let $\lambda_* = \operatorname{argmax}_\lambda g_D(\lambda)$. Then

$$g_D(\lambda_*) = \max_\lambda g_D(\lambda)$$

$$f_G(\lambda_*) = \min_\lambda f_G(\lambda) = d.$$

Intuition for $f_G(\lambda_*) = d$:

$$\max_x \|x\|^2_{A(\lambda)^{-1}} \geq \sum_x \lambda_x \|x\|^2_{A(\lambda)^{-1}}$$

$$= \sum_x \lambda_x \, x^T A(\lambda)^{-1} x$$

$$= \sum_x \lambda_x \operatorname{Tr}\left(x^T A(\lambda)^{-1} x\right)$$

$$= \sum_x \lambda_x \operatorname{Tr}\left(A(\lambda)^{-1} x x^T\right)$$

$$= \operatorname{Tr}\left(A(\lambda)^{-1} A(\lambda)\right)$$

$$= d$$

How do I turn $\lambda \in \Delta_X$ into $x_1, \ldots, x_T$?

If $T\lambda \in \mathbb{N}^n$ then take $T_x = \lambda_x T$.

1) One idea is to $x_i \overset{iid}{\sim} \lambda$. $i = 1, \ldots, T$.

$$\Rightarrow \sum_{i=1}^{n} x_i x_i^T = A(\lambda) \cdot T$$

Problem:

$$A(\lambda)^{-1} = \left(\mathbb{E}\left[\sum_{i=1}^{n} x_i x_i^T \right] \right)^{-1} \neq \mathbb{E}\left[\left(\sum_{i=1}^{n} x_i x_i^T \right)^{-1} \right]$$

Not going to work

2) Another idea is to take $T_x = \lceil \lambda_x T \rceil$

Problem: might go over budget: $\sum_x T_x >> T$.

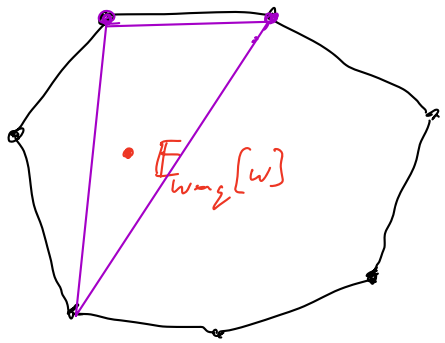
<u>Fact</u> For every $\lambda \in \Delta_X$, $\exists\, \lambda' \in \Delta_X$ s.t. $\|\lambda'\|_0 \leq \frac{d(d+1)}{2}$ that satisfies $A(\lambda) = A(\lambda')$.

Why? $A(\lambda) = \sum_x \lambda_x \cdot xx^T$.

<u>Caratheodory's Theorem</u> Let $S \subset \mathbb{R}^p$ and let $q \in \Delta_S$ Then $\exists q'$: $\mathbb{E}_{w \sim q}[w] = \mathbb{E}_{w \sim q'}[w]$. and $\|q'\|_0 \leq p+1$.

Proof by picture

$\mathcal{X}$



G-optimal

Procedure Input: $\mathcal{X}$, budget T

1.) Solve $\hat{\lambda} = \underset{\lambda\in\Delta_{\mathcal{X}}}{\operatorname{argmin}} \max_{x\in\mathcal{X}} \|x\|^2_{A(\lambda)^{-1}}$

2.) Find a $\frac{d(d+1)}{2}$ sparse solution $\lambda_\star$: $A(\lambda_\star) = A(\hat{\lambda})$.

3.) Set $T_x = \lceil \lambda_{\star,x} T \rceil$ times, observe $y(x) := \langle x, \theta_\star \rangle + z$ T_x times to obtain $\{(x_i, y_i)\}_{i=1}^{T_x}$

4) Compute $\hat{\theta} = (X^T X)^{-1} X^T y$

Lemma $\sum_x T_x = \sum_x \lceil \lambda_{\star,x} T \rceil \le \frac{d(d+1)}{2} + \sum_x \lambda_{\star,x} T = \frac{d(d+1)}{2} + T$

With probability at most $1-\delta$ we have for all $x \in \mathcal{X}$

$$|\langle x, \hat{\theta} - \theta_\star \rangle| \le \|x\|_{A(\hat{\lambda})^{-1}} \sqrt{\frac{2\sigma^2 \log(2|\mathcal{X}|/\delta)}{T}} = \sqrt{\frac{2d\sigma^2 \log(2|\mathcal{X}|/\delta)}{T}}.$$

For any z we have $z^T(\hat{\theta} - \theta_\star) = z^T (X^T X)^{-1} X^T \zeta =: w^T \zeta$

$\mathbb{E}[w^T \zeta] = 0$, $\mathbb{E}[(w^T \zeta)^2] = w^T \mathbb{E}[\zeta\zeta^T] w = \le \sigma^2 \|w\|_2^2$

$\mathbb{E}[\exp(\lambda w^T \zeta)] = \mathbb{E}[\exp(\lambda \sum_t w_t \zeta_t)]$ ζ_t is σ^2-sub-Gaussian

$= \mathbb{E}[\prod_t \exp(\lambda w_t \zeta_t)]$

$$= \prod_t \mathbb{E}\left[\exp(\lambda \omega_t \eta_t)\right]$$

$$\le \prod_t \exp\left(\lambda^2 \omega_t^2 \sigma^2/2\right)$$

$$= \exp\left(\frac{\lambda^2\sigma^2}{2} \cdot \sum_t \omega_t^2\right)$$

$$= \exp\left(\frac{\lambda\sigma^2}{2}\|\omega\|_2^2\right)$$

For any $\lambda' \in \mathbb{R}$ we have $\mathbb{E}[\exp(\lambda' \eta_t)] \le \exp(\lambda'^2\sigma^2/2)$

$$\|\omega\|_2^2 = \omega^T\omega = z^T(X^TX)^{-1}X^TX(X^TX)^{-1}z = z^T(X^TX)^{-1}z$$

$\Rightarrow$ $(\hat\theta - \theta_*)^T z$ is $\left(\sigma^2\|z\|^2_{(X^TX)^{-1}}\right)$-sub-Gaussian.

For any $z \in \mathcal{X}$

$$\mathbb{P}\left((\hat\theta-\theta_*)^T z > \varepsilon\right) \le \exp\left(-\frac{\varepsilon^2}{2\sigma^2\|z\|^2_{(X^TX)^{-1}}}\right) \quad \text{(Chernoff)}$$

$$\le \exp\left(-\frac{T\varepsilon^2}{2d\sigma^2}\right)$$

$$\|z\|^2_{(X^TX)^{-1}} = z^T(X^TX)^{-1}z = z^T\left(\sum_x \lceil \lambda_x T \rceil xx^T\right)^{-1}z$$

$$\le z^T\left(\sum_x \lambda_x T\, xx^T\right)^{-1}z$$

$$= \frac{z^T A(\lambda)^{-1} z}{T} = \frac{\|z\|^2_{A(\lambda)^{-1}}}{T} \le \frac{d}{T}$$

$z \in \mathcal{X}$

$$\mathbb{P}\left(\bigcup_{x\in\mathcal{X}}\{|x^T(\hat\theta-\theta_*)| > \varepsilon\}\right) \le 2\sum_{x\in\mathcal{X}} \mathbb{P}\left(x^T(\hat\theta-\theta_*) > \varepsilon\right)$$

$$\le 2|\mathcal{X}|\exp\left(-\frac{T\varepsilon^2}{2d\sigma^2}\right)$$ Set $= \delta$ solve for ε. ∎