Fix $x_1, \ldots, x_n \in \mathbb{R}^d$ and $\theta_\star \in \mathbb{R}^d$. (known / unknown)

$\forall i$ Observe

$$y_i = x_i^T \theta_\star + \varepsilon_i, \quad \{\varepsilon_i\}_i \text{ are IID mean-zero 1-sub-Gaussian R.V.}$$

Least squares estimate:

$$\hat{\theta} = \operatorname*{argmin}_\theta \sum_{i=1}^n (\langle x_i, \theta \rangle - y_i)^2$$

Observe $\{(x_i, y_i)\}_{i=1}^n$

$$= \left(\sum_{i=1}^n x_i x_i^T \right)^{-1} \sum_{i=1}^n x_i y_i$$

$$= \left(\sum_{i=1}^n x_i x_i^T \right)^{-1} \left(\sum_{i=1}^n x_i x_i^T \right) \theta_\star + \left(\sum_{i=1}^n x_i x_i^T \right)^{-1} \sum_{i=1}^n x_i \varepsilon_i$$

$$= \theta_\star + \left(\sum_{i=1}^n x_i x_i^T \right)^{-1} \sum_{i=1}^n x_i \varepsilon_i$$

$$\mathbb{E}[\hat{\theta}] = \theta_\star \qquad \mathrm{Cov}(\hat{\theta}) = \left(\sum_{i=1}^n x_i x_i^T \right)^{-1}$$

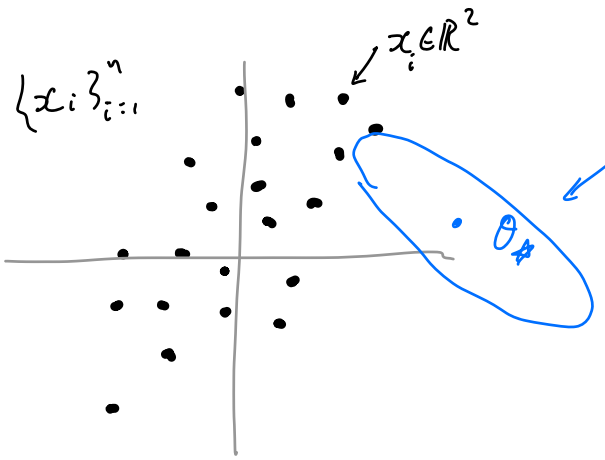
Fix any $z \in \mathbb{R}^d$. With probability at least $1-\delta$

$$\langle z, \hat{\theta} - \theta_\star \rangle \le \|z\|_{\left(\sum_{i=1}^n x_i x_i^T\right)^{-1}} \sqrt{2 \log(1/\delta)}.$$

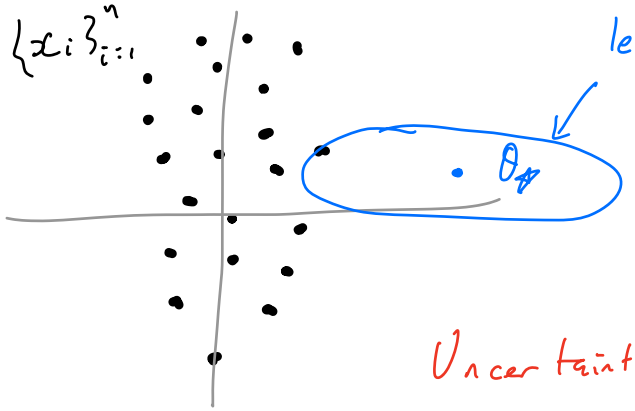
$\left(\|x\|_A^2 = x^T A x \right)$

Proof: Show $\langle z, \hat{\theta} - \theta_\star \rangle$ is $\|z\|^2_{\left(\sum_{i=1}^n x_i x_i^T\right)^{-1}}$-sub-Gaussian. Apply Chernoff bound. ∎

Intuition: If $\varepsilon_i \sim \mathcal{N}(0,1)$ then $\hat{\theta} \sim \mathcal{N}(\theta_\star, (\sum_{i=1}^n x_i x_i^T)^{-1})$



level set of distribution of $\hat{\theta}$



level set of distribution of $\hat{\theta}$

Uncertainty of $\hat{\theta} - \theta_\star$

depends only on $\{x_i\}_{i=1}^n$ and <u>not</u> $\theta_\star$ or $\{\varepsilon_i\}$

Experimental Design: Exploits this observation to choose $\{x_i\}$ in order to obtain a desired covariance shape.

Given choice of $\{x_i\}_{i=1}^n$ and then observe $y_i = \langle x_i, \theta_\star \rangle + \varepsilon_i$

results in $\hat{\theta}$ w/ $\mathbb{E}[\hat{\theta}] = \theta_\star$ $\quad Cov(\hat{\theta}) = \left(\sum_i x_i x_i^T\right)^{-1}$

Given some pool of points $\mathcal{X} \subset \mathbb{R}^d$

Choose $\{x_i\}_{i=1}^n$

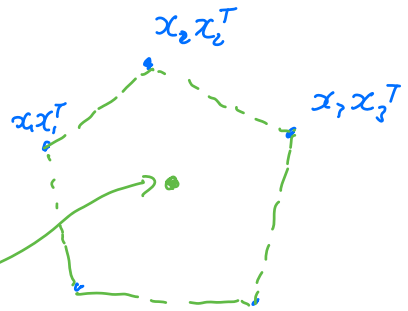
For every choice of points $\{x_i\}_{i=1}^n \subset \mathcal{X}$, $\exists \lambda \in \Delta_{\mathcal{X}}$

$$\Delta_{\mathcal{X}} = \left\{ p \in \mathbb{R}^{|\mathcal{X}|} : \sum_{x \in \mathcal{X}} p_x = 1, \; p_x \geq 0 \;\; \forall x \in \mathcal{X} \right\}$$

s.t. $$\sum_{i=1}^n x_i x_i^T = n \sum_{x \in \mathcal{X}} \lambda_x x x^T$$

why? $\lambda_x = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{x_i = x\}$.

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



Exp. Design Objectives

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

$$\mathbb{E}\left[\|\hat{\theta} - \theta_*\|_2^2 \right] = Tr\left(\left(\sum_{i=1}^n x_i x_i^T \right)^{-1} \right)$$

"minimize average error over all directions"

- E-optimality $f_E(\lambda) = \sup_{u:\|u\|_2 \le 1} u^T A(\lambda)^{-1} u$

$$\sup_{u:\|u\|_2\le 1} \mathbb{E}\left[\langle u, \hat{\theta} - \theta_* \rangle^2\right] = \sup_u u^T \left(\sum_{i=1}^n x_i x_i^T\right)^{-1} u.$$

"minimize worst-case error direction"

- G-optimality $f_G(\lambda) = \max_{x \in \mathcal{X}} x^T A(\lambda)^{-1} x$

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

$$\max_{x\in\mathcal{X}} \mathbb{E}\left[\langle x, \hat{\theta} - \theta_* \rangle^2\right] = \max_{x \in \mathcal{X}} x^T \left(\sum_{i=1}^n x_i x_i^T\right)^{-1} x$$

$$= \max_{x \in \mathcal{X}} \|x\|^2_{(\sum_i x_i x_i^T)^{-1}}$$

- D-optimality $f_D(\lambda) = -\log(|A(\lambda)|)$

$$= \log(|A(\lambda)^{-1}|)$$

Recall: if $Z \sim \mathcal{N}(0, \Sigma)$

then entropy of $Z = \frac{1}{2}\log(2\pi e |\Sigma|)$

"minimize entropy of $\hat{\theta}$"

Lemma (Kiefer-Wolfowitz 1960)

For any $\mathcal{X} \subset \mathbb{R}^d$: $\dim(\text{span}(\mathcal{X})) = d$, there exists a $\lambda^* \in \Delta_{\mathcal{X}}$ such that

- $\min_\lambda f_D(\lambda) = f_D(\lambda^*)$
- $\min_\lambda f_G(\lambda) = f_G(\lambda^*)$
- $f_G(\lambda^*) = d$
- $\text{support}(\lambda^*) \le \frac{(d+1)d}{2}$.

Caratheodory Theorem.

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

$$\ge \min_\lambda \sum_{x \in \mathcal{X}} \lambda_x \, x^T A(\lambda)^{-1} x$$

$$= \min_\lambda \text{Tr}\Big(\underbrace{\sum_{x \in \mathcal{X}} \lambda_x x x^T}_{= I} A(\lambda)^{-1} \Big)$$

$$= d$$

$x \in \mathbb{R}^d$ is s-sparse if $|\{ i \in [d] : x_i \neq 0 \}| = s$.

Proposition Fix $\mathcal{X} \subset \mathbb{R}^d$. Assume $Y = \langle X, \theta_* \rangle + \varepsilon$ w/ ε 1-sub-Gaussian. If $\lambda^* = \min_{\lambda \in \Delta_{\mathcal{X}}} f_G(\lambda)$ is a $\frac{(d+1)d}{2}$-sparse solution, and we pull arm $x \in \mathcal{X}$ exactly $\lceil 3\lambda^*_x \rceil$ times, then w.p. $\geq 1-\delta$

$$|\langle x, \hat{\theta} - \theta_* \rangle| \leq \underbrace{\sqrt{\frac{2d \log(2|\mathcal{X}|/\delta)}{3}}}_{=\varepsilon} \quad \forall x \in \mathcal{X}$$

$$\mathbb{P}\left(\bigcup_{x \in \mathcal{X}} \{ |\langle x, \hat{\theta} - \theta_* \rangle| > \varepsilon \} \right) \leq \exp\left(-\frac{3\varepsilon^2}{2d}\right) \cdot 2|\mathcal{X}|$$

and $\underbrace{\text{total \# pulls}}_{=n} \leq 3 + \frac{(d+1)d}{2}$.

From above, for any $x \in \mathcal{X}$ w.p. $\geq 1-\delta'$

$$\langle x, \hat{\theta} - \theta_* \rangle \leq \|x\|_{\left(\sum_{i=1}^n x_i x_i^T\right)^{-1}} \sqrt{2 \log(1/\delta')}$$

$$\sum_{i=1}^n x_i x_i^T = \sum_{x \in \mathcal{X}} \lceil 3\lambda^*_x \rceil x x^T \geq \sum_{x \in \mathcal{X}} 3\lambda^*_x x x^T$$

$$\leq \|x\|_{\left(3\sum_{x \in \mathcal{X}} \lambda^*_x x x^T\right)^{-1}} \sqrt{2 \log(1/\delta')}$$

$$\leq \max_{x' \in \mathcal{X}} \underbrace{\|x\|_{\left(\sum_{x \in \mathcal{X}} \lambda^*_x x x^T\right)^{-1}}}_{=\sqrt{d}} \sqrt{\frac{2 \log(1/\delta')}{3}}$$

$$= \sqrt{\frac{2d \log(1/\delta')}{3}}$$

$A \succeq B$ if $A - B \succeq 0$ (PSD)

$$\|x\|_{\left(3\sum_{x'\in X}\lambda^*_{x'}x'x'^T\right)^{-1}} = \sqrt{x^T\left(3\sum_{x'\in X}\lambda_{x'}x'x'^T\right)^{-1}x}$$

$$= \frac{1}{\sqrt{3}}\cdot\|x\|_{A(A^*)^{-1}}$$

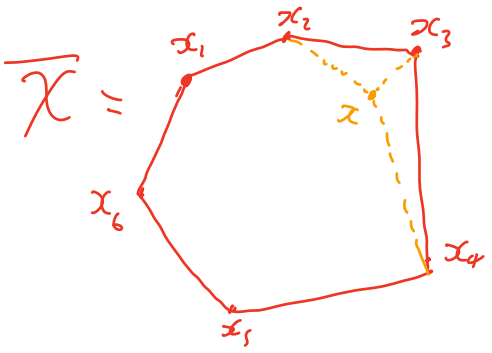
Caratheodory

Let $x_1,\dots,x_n \in \mathbb{R}^d$ and define

$$\overline{X} = \left\{\sum_{i=1}^n p_i x_i : \sum_{i=1}^n p_i = 1,\ p_i \geq 0\ \forall i\right\}$$

Then for any $x \in \overline{X}$, $\exists\, p \in \Delta_n$: $\sum_{i=1}^n p_i x_i = x$

$\|p\|_0 = d+1$



Input: Finite set $\mathcal{X} \subset \mathbb{R}^d$, confidence level $\delta \in (0,1)$.
Let $\widehat{\mathcal{X}}_1 \leftarrow \mathcal{X}, \ell \leftarrow 1$
while $|\mathcal{X}_\ell| > 1$ **do**
Let $\widehat{\lambda}_\ell \in \triangle_{\mathcal{X}_\ell}$ be a $\frac{d(d+1)}{2}$-sparse minimizer of $f(\lambda) = \max_{x \in \mathcal{X}_\ell} \|x\|^2_{(\sum_{x \in \mathcal{X}_\ell} \lambda_x x x^\top)^{-1}}$
$\epsilon_\ell = 2^{-\ell}$, $\tau_\ell = 2d\epsilon_\ell^{-2} \log(4\ell^2|\mathcal{X}|/\delta)$
Pull arm $x \in \mathcal{X}_\ell$ exactly $\lceil \widehat{\lambda}_{\ell,x} \tau_\ell \rceil$ times and construct the least squares estimator $\widehat{\theta}_\ell$ using only the observations of this round
$\mathcal{X}_{\ell+1} \leftarrow \mathcal{X}_\ell \setminus \{x \in \mathcal{X}_\ell : \max_{x' \in \mathcal{X}_\ell} \langle x' - x, \widehat{\theta}_\ell \rangle > 2\epsilon_\ell\}$
$\ell \leftarrow \ell + 1$
Output: $\mathcal{X}_\ell$

Regret $R_T = \max_{x \in \mathcal{X}} \mathbb{E}\left[\sum_{t=1}^{T} \langle x, \theta_* \rangle - \langle x_t, \theta_* \rangle \right]$

x_t ↑ arm algorithm played at time t

$= \sum_{x \in \mathcal{X}} \Delta_x \mathbb{E}[T_x]$

$\Delta_x = \max_{x' \in \mathcal{X}} \langle x' - x, \theta_* \rangle$

$T_x = \sum_{t=1}^{T} \mathbb{1}\{x = x_t\}$

Assume $\max_{x \in \mathcal{X}} |\langle x, \theta_* \rangle| \le 1$. Let $x^* = \operatorname{argmax}_{x \in \mathcal{X}} \langle x, \theta_* \rangle$

Lemma With prob at least $1-\delta$, we have $x^* \in \mathcal{X}_\ell$ and $\max_{x \in \mathcal{X}_\ell} \Delta_x \le 8\epsilon_\ell$ for all $\ell \in \mathbb{N}$.

What does linear structure get us?

- If we ignored structure our elim. alg of last time has $R_T \le \sqrt{|\mathcal{X}| T \log(T)}$.

- This algorithm that exploits linear structure,

we will show $R_T \le \sqrt{d\,T \log(|\mathcal{X}|\,T)}$.

I can "cover" $\mathcal{X}$ with just $O(2^d)$ landmarks

$$\Rightarrow \quad R_T \le \sqrt{d^2\,T \log(T)}$$

Proof. Consider $V \subset \mathcal{X}$ and event

$$\mathcal{E}^c_{x,\ell}(V) = \left\{ |\langle x, \hat{\theta}_\ell(V) - \theta^* \rangle| \le \varepsilon_\ell \right\}$$

$\uparrow$

The $\hat{\theta}_\ell$ that would arise if $\mathcal{X}_\ell = V$.

$$\mathbb{P}\left(\bigcup_{\ell=1}^{\infty} \bigcup_{x \in \mathcal{X}_\ell} \mathcal{E}^c_{x,\ell}(\mathcal{X}_\ell) \right) \le \sum_{\ell=1}^{\infty} \mathbb{P}\left(\bigcup_{x \in \mathcal{X}_\ell} \mathcal{E}^c_{x,\ell}(\mathcal{X}_\ell) \right)$$

$$= \sum_{\ell=1}^{\infty} \sum_{V \subset \mathcal{X}} \mathbb{P}\left(\left\{ \bigcup_{x \in V} \mathcal{E}^c_{x,\ell}(V) \right\} \cap \{ \mathcal{X}_\ell = V \} \right)$$

$$= \sum_{\ell=1}^{\infty} \sum_{V \subset \mathcal{X}} \mathbb{P}\left(\bigcup_{x \in V} \mathcal{E}^c_{x,\ell}(V) \right) \mathbb{P}(\mathcal{X}_\ell = V)$$

apply proposition

$$\le \sum_{\ell=1}^{\infty} \sum_{V \subset \mathcal{X}} \frac{|V|}{|\mathcal{X}|} \frac{\delta}{2\ell^2} \mathbb{P}(\mathcal{X}_\ell = V) \qquad (|V| \le |\mathcal{X}|)$$

$$\le \sum_{\ell=1}^{\infty} \frac{\delta}{2\ell^2} \le \delta$$

x_* is kicked out at round ℓ if

$$\max_{x \in X_\ell} \langle x - x_*, \hat{\theta}_\ell \rangle > 2\varepsilon_\ell$$

For any x, on the good event

$$\langle x - x_*, \hat{\theta}_\ell \rangle = \langle x - x_*, \hat{\theta}_\ell - \theta_* \rangle + \langle x - x_*, \theta_* \rangle$$

$$= \underbrace{\langle x, \hat{\theta}_\ell - \theta_* \rangle}_{\leq \varepsilon_\ell} \underbrace{- \langle x_*, \hat{\theta}_\ell - \theta_* \rangle}_{\leq \varepsilon_\ell} - \underbrace{\Delta_x}_{>0}$$

$$\leq 2\varepsilon_\ell.$$

Suppose some $x \in X_\ell$ has $\Delta_x > 4\varepsilon_\ell$,

then

$$\max_{x' \in X_\ell} \langle x' - x, \hat{\theta}_\ell \rangle \geq \langle x_* - x, \hat{\theta}_\ell \rangle$$

$$= \langle x_*, \hat{\theta}_\ell - \theta_* \rangle - \langle x, \hat{\theta}_\ell - \theta_* \rangle + \underbrace{\Delta_x}_{> 4\varepsilon_\ell}$$

$$> -2\varepsilon_\ell + 4\varepsilon_\ell$$

$$> 2\varepsilon_\ell \Rightarrow x \text{ is kicked out.}$$

$$\Delta_x > 4\varepsilon_\ell \Rightarrow x \notin \mathcal{X}_{\ell+1} \Rightarrow \Delta_x \le 4\varepsilon_\ell = 8\varepsilon_{\ell+1}$$

$$\Rightarrow \max_{x \in \mathcal{X}_\ell} \Delta_x \le 8\varepsilon_\ell \quad \forall \ell.$$

$$R_T = \sum_{x \in \mathcal{X}} \Delta_x \mathbb{E}[T_x].$$ Fix any $\nu \ge 0$

$$\sum_{x \in \mathcal{X}} \Delta_x T_x = \sum_{x \in \mathcal{X}: \Delta_x \le \nu} \Delta_x T_x + \sum_{x \in \mathcal{X}: \Delta_x > \nu} \Delta_x T_x$$

$$\le \nu T + \sum_{\ell=1}^{\infty} \sum_{x \in \mathcal{X}_\ell : \Delta_x > \nu} \underbrace{\Delta_x}_{\le 8\varepsilon_\ell} \underbrace{\lceil \tau_\ell \hat{\lambda}_{\ell,x} \rceil}_{\le \tau_\ell \hat{\lambda}_{\ell,x} + \mathbb{1}\{\hat{\lambda}_{\ell,x} > 0\}}$$

$a \vee b = \max\{a, b\}$

$$\le \nu T + \sum_{\ell=1}^{\infty} 8\varepsilon_\ell \sum_{x \in \mathcal{X}_\ell : \Delta_x > \nu} \left(\tau_\ell \hat{\lambda}_{\ell,x} + \mathbb{1}\{\hat{\lambda}_{\ell,x} > 0\} \right)$$

$\Delta = \min_{x \ne x_\star} \langle x_\star - x, \theta_\star \rangle$

$$\le \nu T + \sum_{\ell=1}^{\lceil \log_2(8/\nu \vee \Delta^{-1}) \rceil} 8\varepsilon_\ell \left(\tau_\ell + \frac{(d+1)d}{2} \right)$$

$$\tau_\ell = 8\varepsilon_\ell^{-2} d \log\left(\frac{4\ell^2|\mathcal{X}|}{\delta}\right)$$

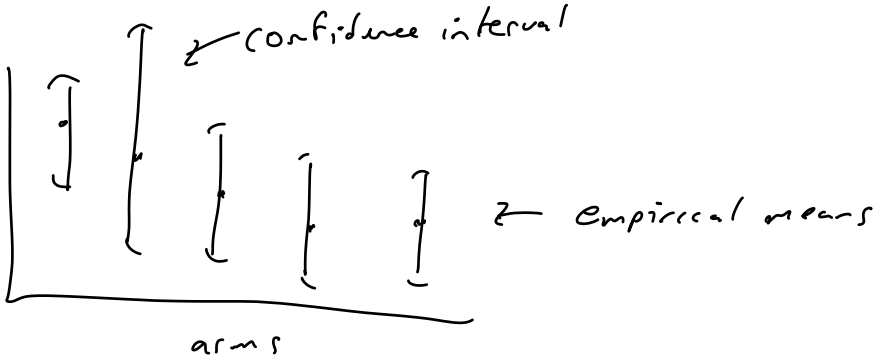
$$\le \nu T + cd \log\left(\frac{\log_2((8/\nu) \vee \Delta^{-1})|\mathcal{X}|}{\delta}\right) \sum_{\ell=1}^{\lceil \log_2(8/\nu \vee \Delta^{-1}) \rceil} \underbrace{\varepsilon_\ell^{-1}}_{2^\ell} + \sum_{\ell=1}^{\cdots} \frac{(d+1)d}{2} 2^\ell$$

$$\le \nu T + c' d \log\left(\frac{\log_2((8/\nu) \vee \Delta^{-1})|\mathcal{X}|}{\delta}\right) (\nu \vee \Delta)^{-1}$$

To get expected regret choose $\delta = \frac{1}{T}$

If $\nu = 0$ then $R_T \leq \frac{c' d \log\left(T \log(\delta^{-1}) |\mathcal{X}|\right)}{\Delta}$

$\min \nu$, $\max \Delta \rightarrow$ $R_T \leq c'' \sqrt{dT \log(|\mathcal{X}| T)}$



Suppose we've pulled some arms to get $\{(x_t, y_t)\}_{t \leq s}$ to construct $\hat{\theta}_s$. Consider constructing a set

$C_s = \{\theta \in \mathbb{R}^d : \theta$ is not ruled out as θ_* w.p. $\geq 1-\delta\}$

UCB Alg.

for $t = 1, 2, \ldots$

$x_t = \underset{x \in \mathcal{X}}{\arg\max} \max_{\theta \in C_{t-1}} \langle x, \theta \rangle$

Thompson Sampling

Input prior distribution P_0 defined over $\mathbb{R}^d$

for $t = 0, 1, \ldots$

$\tilde{\theta}_t \sim P_t$

$x_t = \underset{x \in \mathcal{X}}{\operatorname{argmax}} \langle x, \tilde{\theta}_t \rangle$

Observe $y_t = \langle x_t, \theta^* \rangle + \varepsilon_t$

Update posterior dist P_{t+1}

$\varepsilon_t \sim \sigma^2$-subGaussian

In practice: for some $\lambda > 0$

$$P_t = \mathcal{N}\left(\left(\sum_{s=1}^{t} x_s x_s^T + \lambda I \right)^{-1} \sum_{s=1}^{t} x_s y_s, \sigma^2 \left(\sum_{s=1}^{t} x_s x_s^T + \lambda I \right)^{-1} \right)$$