$$\text{Fix } \underset{\text{known}}{x_1, \dots, x_n} \in \mathbb{R}^d \text{ and } \underset{\text{unknown}}{\theta_*} \in \mathbb{R}^d.$$

$\forall i$ Observe

$$y_i = x_i^T \theta_* + \varepsilon_i \,, \quad \{\varepsilon_i\}_i \text{ are IID mean-zero}$$

1-sub-Gaussian R.V.

Least squares estimate:

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

$$\hat{\theta} = \arg\min_\theta \sum_{i=1}^n (\langle x_i, \theta \rangle - y_i)^2$$

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

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

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

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

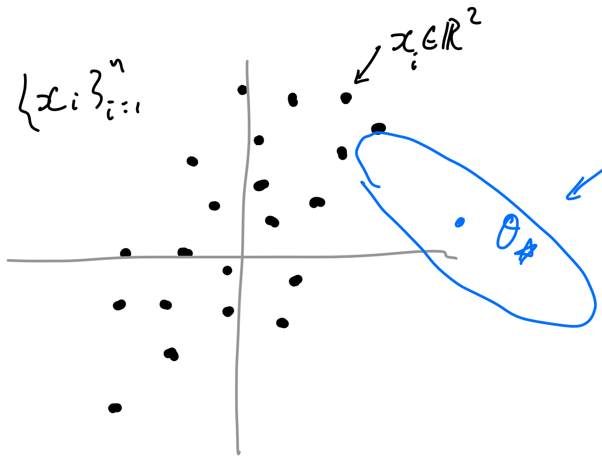
Fix any $z \in \mathbb{R}^d$. With probability at least $1-\delta$

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

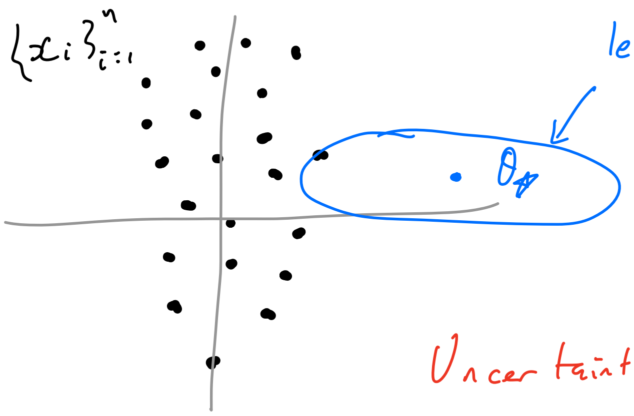
$\left(\|x\|_A^2 = x^T A x \right)$

Proof: Show $\langle z, \hat{\theta} - \theta_* \rangle$ is $\|z\|^2_{(\sum_i x_i x_i^T)^{-1}}$-sub-Gaussian. Apply Chernoff bound. $\blacksquare$

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



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



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

Uncertainty of $\hat{\theta} - \theta_*$ depends only on $\{x_i\}_{i=1}^n$ and <u>not</u> θ_* or $\{y_i\}$

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 $y_i = \langle x_i, \theta_* \rangle + \varepsilon_i$ results in $\hat{\theta}$ w/ $\mathbb{E}[\hat{\theta}] = \theta_*$ $\quad \mathrm{Cov}(\hat{\theta}) = \left(\sum_i x_i x_i^T\right)^{-1}$

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

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

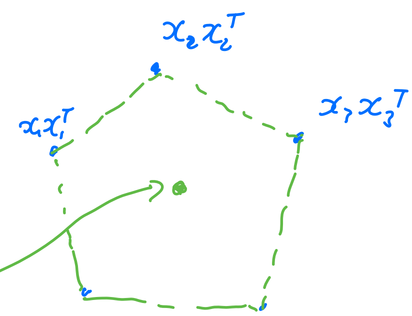
For every choice of points $\{x_i\}_{i=1}^n \subset \mathcal{X}$, $\exists \lambda \in \Delta_{\mathcal{X}}$

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

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

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

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



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

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

"minimize worst-case error direction"

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

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

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

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

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

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

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

Caratheodory Theorem.

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

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

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

$$= d$$

$x \in \mathbb{R}^d$ 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 $\lceil \tau \lambda_x^* \rceil$ times, then w.p. $\geq 1-\delta$

$$|\langle x, \hat{\theta} - \theta_* \rangle| \leq \sqrt{\frac{2d \log(2|\mathcal{X}|/\delta)}{\tau}} \qquad \forall x \in \mathcal{X}$$

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

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

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

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

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

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

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

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

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

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

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

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 $x^* = \arg\max_{x \in \mathcal{X}} \langle x, \theta_* \rangle$

Lemma: With prob at least $1-\delta$, we have $x^* \in \mathcal{X}_\ell$ and $\max_{x \in \mathcal{X}_\ell} \Delta_x \le 8\varepsilon_\ell$ 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 $R_T \leq \sqrt{d\,T \log(|\mathcal{X}|T)}$.

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

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