Consider a finite set of vectors $x_1, \dots, x_n \in \mathbb{R}^d$

For all $i \in [n]$ we observe $y_i = \langle x_i, \theta^* \rangle + \varepsilon_i$

for ε_i (Independent) mean-0, sub-Gaussian. $\mathbb{E}[\varepsilon_i] = 0$, $\mathbb{E}[\exp(\lambda \varepsilon_i)] \le \exp(\lambda^2/2)$

Ex. $y_i \in \{-1, 1\}$ denoting "click/no-click"

$\varepsilon_i \in [-\langle x_i, \theta^* \rangle, 1 - \langle x_i, \theta^* \rangle]$ (bounded)

$\Rightarrow \mathbb{E}[\exp(\lambda \varepsilon_i)] \le \exp(\lambda^2/8)$

Ex. $y_i \sim \mathcal{N}(\langle \theta_*, x_i \rangle, 1)$ denoting "linger time"

Given $\{(y_i, x_i)\}_{i=1}^n$ estimate θ_*.

$$y_i = \langle x_i, \theta^* \rangle + \varepsilon_i$$

$X \in \mathbb{R}^{n \times d}$ $Y = \mathbb{R}^n$

Most natural estimator is LS (also MLE for Gaussian noise)

$$X = \begin{pmatrix} — x_1^T — \\ \vdots \\ — x_n^T — \end{pmatrix} \qquad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

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

$$= \left(\sum_{i=1}^n x_i x_i^T \right)^{-1} \left(\sum_{i=1}^n x_i y_i \right) \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$= (X^T X)^{-1} X^T Y \qquad Y = X\theta^* + \varepsilon$$

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

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

$$\mathbb{E}\left[(\hat{\theta}-\theta^*)(\hat{\theta}-\theta^*)^T\right] = \mathbb{E}\left[(X^TX)^{-1}X^T\varepsilon\varepsilon^TX(X^TX)^{-1}\right]$$

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

$$\leq (X^TX)^{-1}X^TX(X^TX)^{-1} \quad \leftarrow \quad \mathbb{E}[\varepsilon_i^2] \leq 1$$

$$\mathbb{E}[\varepsilon_i\varepsilon_j] = 0$$

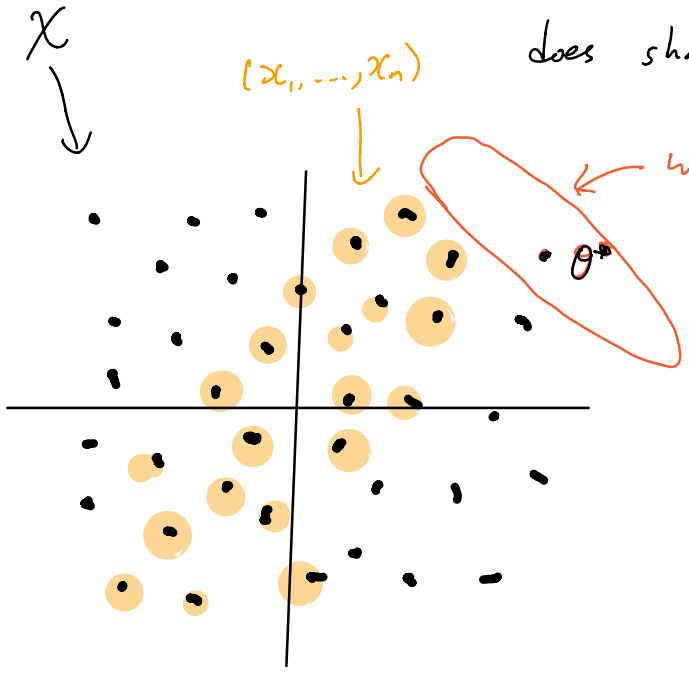
$$= (X^TX)^{-1}$$

Ex. If $\varepsilon \sim N(0,1)$ then

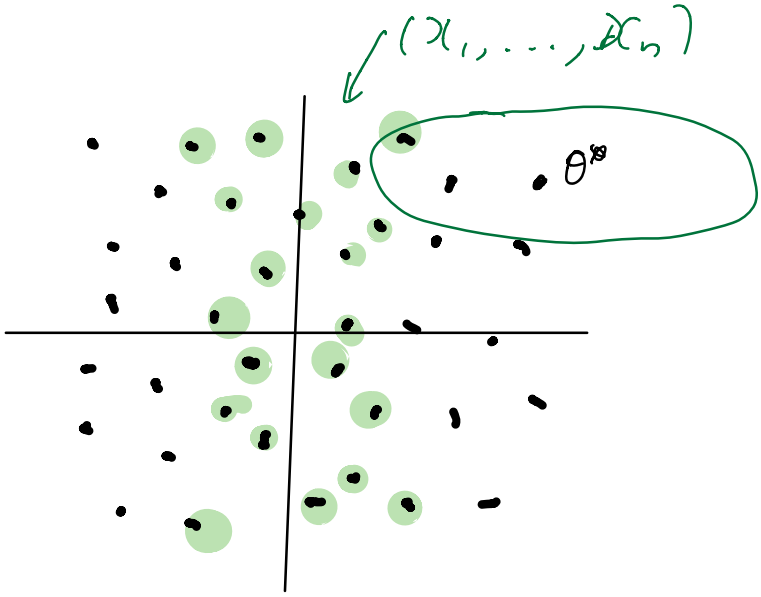
$\hat{\theta}$ has mean θ^*, covariance $(X^TX)^{-1}$

$$\hat{\theta} - \theta^* \sim \mathcal{N}(0, (X^TX)^{-1})$$

Now suppose we have a pool of points $\mathcal{X}$ and we choose $(x_1,\dots,x_n)$ from it. How does shape of uncertainty look?



← w.h.p. $\hat{\theta}$ will land in this ellipse



For any PSD matrix A define $\|x\|_A^2 = x^T A x$.

A is PSD if $x^T A x \geq 0$ for all $x \in \mathbb{R}^d$.

Proposition Fix $x_1, \ldots, x_n$. Let $y_i = \langle x_i, \theta^* \rangle + \varepsilon_i \ \forall i$ where $\mathbb{E}[\varepsilon_i] = 0$ and $\mathbb{E}(\exp\{\lambda \varepsilon_i\}) \leq e^{\lambda^2/2}$ and $\varepsilon_i \perp \varepsilon_j$. Then for any $z \in \mathbb{R}^d$ w.p. $\geq 1 - \delta$, $\hat{\theta} = \operatorname{argmin}_\theta \sum_i (y_i - x_i^T \theta^*)^2$ satisfies

$$\langle z, \hat{\theta} - \theta^* \rangle \leq \sqrt{2 \|z\|^2_{(X^T X)^{-1}} \log(1/\delta)}.$$

Which follows from $\mathbb{P}(\langle z, \hat{\theta} - \theta^* \rangle \geq t) \leq \exp\left(-t^2 / 2\|z\|^2_{(X^T X)^{-1}}\right)$.

Proof $\hat{\theta} = \theta^* + (X^T X)^{-1} X^T \varepsilon$ $\qquad w := X(X^T X)^{-1} z$.

$$\Rightarrow \quad z^T(\hat{\theta} - \theta^*) = z^T (X^T X)^{-1} X^T \varepsilon =: w^T \varepsilon = \sum_{i=1}^n w_i \varepsilon_i$$

$$\mathbb{E}\left[\exp\left(\lambda \sum_i w_i \varepsilon_i\right)\right] = \mathbb{E}\left[\prod_{i=1}^{n} \exp(\lambda w_i \varepsilon_i)\right]$$

$$= \prod_{i=1}^{n} \mathbb{E}\left[\exp(\lambda w_i \varepsilon_i)\right] \qquad \text{(Independence of } \varepsilon_i\text{)}$$

$$\leq \prod_{i=1}^{n} \exp\left(\lambda^2 w_i^2 / 2\right) \qquad \text{(sub-Gaussian)}$$

$$= \exp\left(\sum_{i=1}^{n} \lambda^2 w_i^2 / 2\right)$$

$$= \exp\left(\lambda^2 \|w\|_2^2 / 2\right)$$

Through Chernoff bound technique

$$\mathbb{P}\left(w^T \varepsilon \geq t\right) \leq \exp\left(-\frac{t^2}{2\|w\|_2^2}\right)$$

$$\|w\|_2^2 = \left\| X (X^T X)^{-1} z \right\|_2^2$$

$$= z^T (X^T X)^{-1} \cancel{X^T X (X^T X)^{-1}} z$$

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

$$\mathbb{P}\left(\langle \hat{\theta} - \theta^*, z \rangle \geq t\right) = \mathbb{P}\left(\langle w, \varepsilon \rangle \geq t\right)$$

$$\leq \exp\left(\frac{-t^2}{2\|w\|_2^2}\right)$$

$$= \exp\left(-\frac{t^2}{2\|z\|_{(X^T X)^{-1}}^2}\right)$$

$\blacksquare$

So given cov of $\hat{\theta} - \theta^*$ is $(X^TX)^{-1}$

and $\langle z, \hat{\theta} - \theta^* \rangle \le \sqrt{2 \|z\|^2_{(X^TX)^{-1}} \log(1/\delta)}$.

Clearly, $(X^TX)^{-1}$ is key quantity.

For any $x_1, \dots, x_n \in \mathcal{X}$ $\exists \lambda \in \Delta_{\mathcal{X}} := \{ \nu \in \mathbb{R}_+^{|\mathcal{X}|} : \sum_{x \in \mathcal{X}} \nu_x = 1 \}$

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

($\mathcal{X}$: pool of points)

$$A(\lambda) = \sum_{x \in \mathcal{X}} \lambda_x x x^T$$ ← Design matrix.

Idea: instead of choosing $(x_1, \dots, x_n)$, simply design $\lambda \in \Delta_{\mathcal{X}}$ ignoring integer effects.

Design Criteria

A-optimal: $f_A(\lambda) = \mathrm{Tr}(A(\lambda)^{-1})$ to minimize $\mathbb{E}[\|\hat{\theta} - \theta_*\|_2^2]$

$\mathbb{E}[\|\hat{\theta} - \theta^*\|_2^2] = \mathbb{E}[\|(X^TX)^{-1}X^T\varepsilon\|_2^2] \le \mathrm{Tr}((X^TX)^{-1})$ (equality if $\mathbb{E}(\varepsilon_i^2) = 1$)

$\asymp \frac{1}{n} \mathrm{Tr}(A(\lambda)^{-1})$

E-optimal: $f_E(\lambda) = \lambda_{max}(A(\lambda)^{-1})$ to min. $\sup_{\|u\|_2 \le 1} \mathbb{E}[\langle u, \hat{\theta} - \theta_* \rangle^2]$

D-optimal: $g_D(\lambda) = \log|A(\lambda)|$. maximizing this quantity is equivalent to minimizing to entropy of $\hat{\theta}$ if $\varepsilon_i \sim \mathcal{N}(0,1)$.

G-optimal: $f_G(\lambda) = \max_{x \in \mathcal{X}} \|x\|^2_{A(\lambda)^{-1}}$ to min. $\max_{x \in \mathcal{X}} \mathbb{E}[\langle x, \hat{\theta} - \theta_* \rangle^2]$

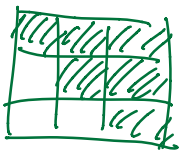
Theorem/ Kiefer-Wolfowitz (1960)

For finite set $\mathcal{X}$ w/ $\dim(\text{span}(\mathcal{X})) = d$ there exists $\lambda^* \in \Delta_{\mathcal{X}}$ such that:

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

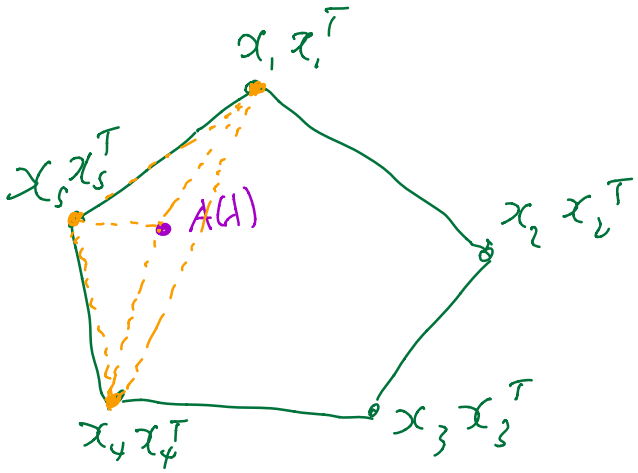
Fix any $\lambda \in \Delta_{\mathcal{X}}$ and consider

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



← degrees of freedom $\le \frac{(d+1)d}{2}$

Consider each xx^T as a point in $\mathbb{R}^{(d+1)d/2}$



In picture dim $p=2$

Caratheodory says any pt w/in convex hull of n points in p-dimensions can be represented by a convex combination of just $p+1$ points.

Proposition Fix $T \in \mathbb{N}$ and let λ^* be the G-optimal soln w/ $\text{support}(\lambda^*) \le \frac{d(d+1)}{2}$. Pull arm $x \in \mathcal{X}$ exactly $\lceil T \lambda_x^* \rceil$ times and compute $\hat{\theta} = (X^T X)^{-1} X^T y$ where X are stacked vectors and y are observations. Then w.p. $\ge 1-\delta$ forall $x \in \mathcal{X}$ we have

$$|\langle \hat{\theta} - \theta^*, x \rangle| \le \sqrt{2 \|x\|^2_{(X^T X)^{-1}} \log(2|\mathcal{X}|/\delta)}$$

$$= \sqrt{x^T \left(\sum_{x' \in \mathcal{X}} \lceil T \lambda^*_{x'} \rceil x' x'^T \right)^{-1} x \cdot 2 \log(2|\mathcal{X}|/\delta)}$$

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

$$= \|x\|_{A(\lambda^*)^{-1}} \cdot \sqrt{\frac{2 \log(2|\mathcal{X}|/\delta)}{T}}$$

$$\le \sqrt{\frac{2d \log(2|\mathcal{X}|/\delta)}{T}}$$

Note: total pulls $\le T + \text{support}(\lambda^*) \le T + \frac{(d+1)d}{2}$.