Lecture 3

Recap sparse regression

$X \in \mathbb{R}^{n\times d}$, $X_1, \dots, X_d \in \mathbb{R}^n$ with $\|X_i\|^2 = n$

$\underline{w} \sim N(0, I_n)$ $[\, k \ll n \ll d \,]$

$\underline{y} = X\beta^\circ + \underline{w}$ for unknown k-sparse $\beta^\circ \in \mathbb{R}^d$

Lasso $\hat{\underline{\beta}} = \arg\min_{\beta} \{\|X\beta - \underline{y}\|^2 \mid \|\beta\|_1 \le R\}$ $\quad * \; (R = \|\beta^\circ\|_1)$

Fast rate for Lasso

assume $X \in \mathbb{R}^{n\times d}$, $S \subseteq [d]$, $|S| \le k$, $\gamma > 0$ so that

(γ, S)-RE: $\forall u \in C(S)$, $\frac{1}{n}\|Xu\|^2 \ge \gamma \cdot \|u\|^2$ $\quad * \; \frac{1}{n}\|Xu\|^2 = \langle u, \overbrace{(\tfrac{1}{n}X^TX)}^{\text{rank}=n} u\rangle$

where $C(S) = \{u \in \mathbb{R}^d \mid \underbrace{\|u_S\|_1}_{\sum_{i\in S}|u_i|} \ge \underbrace{\|u_{\bar{S}}\|_1}_{\sum_{i\notin S}|u_i|}\}$ $\quad \to$ all eigenvalues $\ge \gamma$

Theorem (fast rate)

$\forall \beta^\circ$, support$(\beta^\circ) \subseteq S$, lasso $\hat{\underline{\beta}}$ satisfies for $R = \|\beta^\circ\|_1$, $\forall X$, (γ, S)-RE

$$\mathbb{E}\left[\frac{1}{n}\|X(\hat{\underline{\beta}} - \beta^\circ)\|^2\right] \le \frac{k}{\gamma \cdot n} \cdot O(\log d)$$

Proof (fast rate)

$\underline{u} = \hat{\underline{\beta}} - \beta^\circ$ need to show $\mathbb{E}[\|X\underline{u}\|^2] \le O(\frac{k}{\gamma}\log d)$

from slow rate proof

$\frac{1}{2}\|X\underline{u}\|^2 \le \|\underline{u}\|_1 \cdot \|X^T\underline{w}\|_\infty$ (A)

$\mathbb{E}[\|X\underline{u}\|^2] \le n \cdot O(\log d)$ (B)

from exercises

Claim 1: $\underline{u} \in C(S)$

claim 2: $\forall \underline{u} \in C(S)$, $\|\underline{u}\|_1 \le 2\sqrt{k} \cdot \|\underline{u}\|_2$

together

$$\frac{1}{2}\|X\underline{u}\|^{\cancel{2}} \le \|\underline{u}\|_1 \cdot \|X^T\underline{w}\|_\infty \quad *A$$

$$\le 2\sqrt{k} \cdot \|\underline{u}\|_2 \cdot \|X^T\underline{w}\|_\infty \quad *\text{claim 2}$$

$$\le \frac{2\sqrt{k}}{\sqrt{\gamma n}} \|\cancel{X\underline{u}}\|_2 \cdot \|X^T\underline{w}\|_\infty \quad *(\gamma, S)\text{-RE}, \; \|x\| = \|x\|_2 = \sqrt{x_1^2 + \cdots + x_n^2}$$

$$\to \frac{1}{2}\|X\underline{u}\| \le \frac{2\sqrt{k}}{\sqrt{\gamma n}} \cdot \|X^T\underline{w}\|_\infty$$

$$\frac{1}{4}\|X\underline{u}\|^2 \le \frac{4k}{\gamma n} \cdot \|X^T\underline{w}\|_\infty^2 \quad \Rightarrow \mathbb{E}[\|X\underline{u}\|^2] = O(\frac{k}{\gamma}\log d) \quad *B$$

$$\|X\underline{u}\|^2 \le \frac{16k}{\gamma n} \cdot \|X^T\underline{w}\|_\infty^2$$

Linear regression with outliers

Motivation: LS is sensitive to outliers. Single outlier can completely change outcome.

Want: estimator robust to (certain kinds of) outliers

Setup: $X \in \mathbb{R}^{n\times d}$, $\frac{1}{n}X^T X = I_d$

observation $\underline{y} = X\beta^\circ + \underline{\eta}$ for unknown $\beta^\circ \in \mathbb{R}^d$

$\{\underline{\eta}\} = \{-\underline{\eta}\}$ symmetry

$\{\underline{\eta}\} = \underbrace{\{\underline{\eta}_1\} \otimes \cdots \otimes \{\underline{\eta}_n\}}_{\text{product distribution}}$ independence

Example (Huber loss)

$\underline{w} \sim N(0, I_n)$, arbitrary $Q \in \mathbb{R}^n$ (unknown to algorithm)

$$\underline{\eta}_i = \begin{cases} \underline{w}_i, & \text{with prob. } \alpha \\ +Q_i, & \text{with prob. } \frac{1}{2}(1-\alpha) \\ -Q_i, & \text{with prob. } \frac{1}{2}(1-\alpha) \end{cases}$$

outlier noise

sign oblivious to choice of β°

$\mathbb{E}[\underline{\eta}_i] = 0$

$\mathbb{E}[\underline{\eta}_i^2] = \alpha \cdot 1 + (1-\alpha)\cdot Q_i^2$

can be unbounded

Idea: better loss function

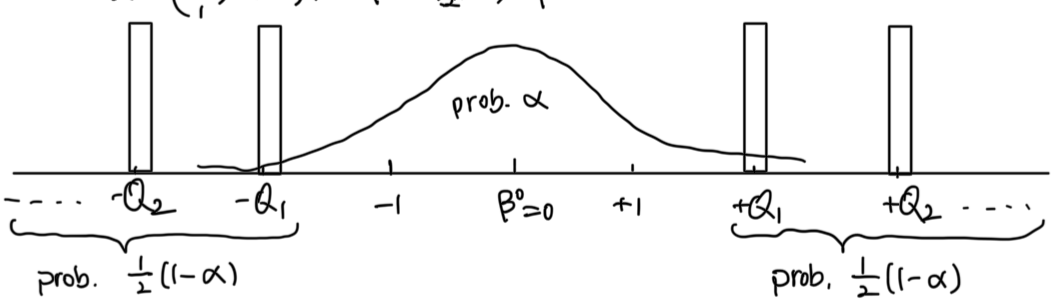
$$f(\beta) = \underbrace{\sum_i \Phi[(X\beta)_i - y_i]}_{} \quad \text{for some } \Phi: \mathbb{R} \to \mathbb{R}^+$$

$= \Phi[X\beta - \underline{y}]$ for simplicity

* In LS, $\Phi[t] = t^2$

example:

$X = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}^{d=1}$, $y_i = \beta^\circ + \underline{\eta}_i$, $\beta^\circ = 0$



median of $\underline{y}_1, \cdots, \underline{y}_n$ gives good estimate for β_i.

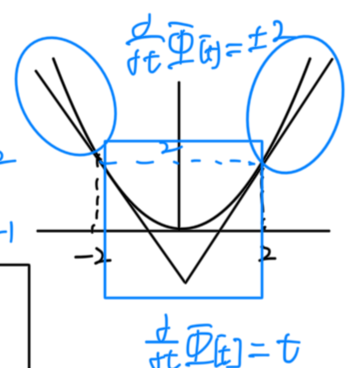
$$\text{median}(\underline{y}_1, \cdots, \underline{y}_n) = \arg\min_\beta \sum_i \underbrace{|\underline{y}_i - \beta|}_{\text{candidate loss } \Phi[t] = |t|}$$

Better: Combine with LS loss <u>Huber loss</u>

$$\Phi[t] = \begin{cases} \frac{1}{2}t^2, & \text{if } |t| \le 2 \\ 2|t| - 2, & \text{if } |t| > 2 \end{cases}$$

* similar to L2

* similar to L1



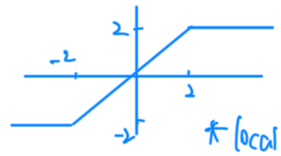
> **Theorem**
>
> Let $\alpha = \min_i \mathbb{P}(|\underline{\eta}_i| \le 1)$,
>
> Let $C = \max_i \frac{\|X_i\|}{\sqrt{d}}$ rows $X_1, \cdots, X_n \in \mathbb{R}^d$ of X,
>
> Suppose $n \ge O\left(\frac{d^2}{C^2\alpha^2}\right)$, then $\|\hat{\underline{\beta}} - \beta^\circ\|^2 \le O\left(\frac{d}{\alpha^2 n}\right)$.

Proof $\underline{u} = \hat{\beta} - \beta^0$

objective: $\underline{f}(\beta^0 + u) = \frac{1}{n}\sum_{i=1}^{n} \Phi\big[\overbrace{\langle x_i, u\rangle - \underline{y}_i}^{(Xu - y)_i}\big]$

gradient: $\nabla \underline{f}(\beta^0 + u) = \frac{1}{n}\sum_{i=1}^{n} \Phi'\big[\langle x_i, u\rangle - \underline{y}_i\big] \cdot x_i$

where $\Phi'[t] = \mathrm{sign}(t) \cdot \min\{|t|, 2\}$



optimality: $\underline{f}(\hat{\underline{\beta}}) \le \underline{f}(\beta^0)$

claim: $\forall u$ with $\|u\| \le R$, $\underline{f}(\beta^0 + u) \ge \underline{f}(\beta^0) - \|\nabla \underline{f}(\beta^0)\| \cdot \|u\| + \frac{\alpha}{10} \cdot \|u\|^2$

* local quadratic lower bound on $\underline{f}$ and β^0

theorem from claim

$$\cancel{\underline{f}(\beta^0)} \ge \underline{f}(\beta^0 + \underline{u}) = \cancel{\underline{f}(\beta^0)} + \langle \nabla \underline{f}(\beta^0), u\rangle + \frac{\alpha}{10}\|u\|^2$$

$$\longrightarrow \frac{\alpha}{10}\|u\|^2 \le -\langle \nabla \underline{f}(\beta^0), u\rangle$$

$$\frac{\alpha}{10}\|u\|^{\cancel{2}} \le \cancel{\|u\|} \cdot \|\nabla \underline{f}(\beta^0)\|$$

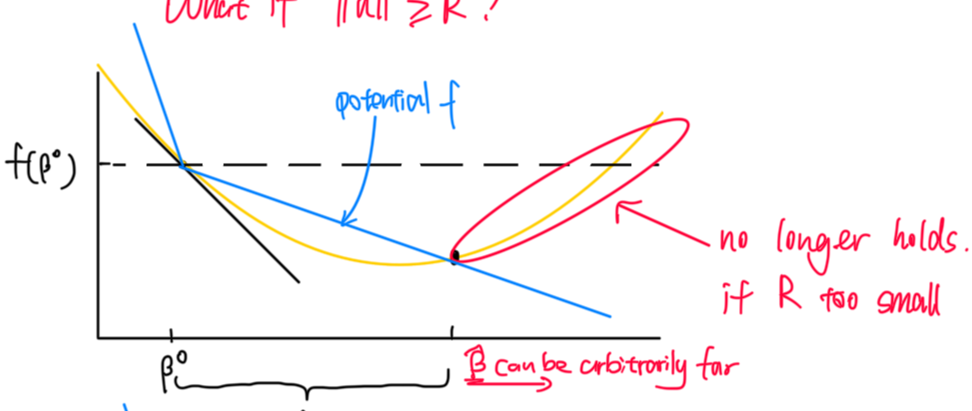
$$\frac{\alpha^2}{100}\|u\|^2 \le \|\nabla \underline{f}(\beta^0)\|^2$$

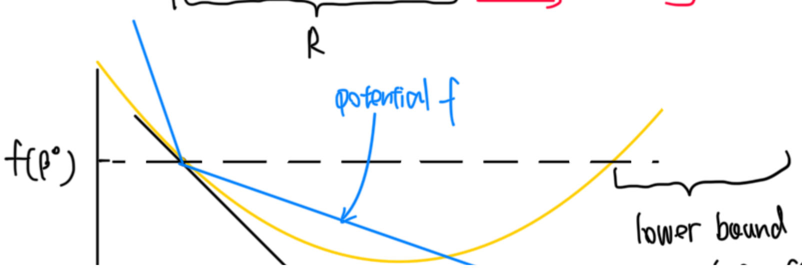
$$\|u\|^2 \ge \frac{100}{\alpha^2}\|\nabla \underline{f}(\beta^0)\|^2$$

$$\longrightarrow \mathbb{E}\big[\|\nabla \underline{f}(\beta^0)\|^2\big] = \underbrace{\frac{1}{n^2}\sum_{i=1}^{n}\|x_i\|^2 \cdot \mathbb{E}\big[\underbrace{\Phi'[y_i]^2}_{\le 4}\big]}_{|t| \le 2} + \underbrace{\cancel{\frac{1}{n^2}\sum_{i,j}\langle x_i, x_j\rangle \cdot \underbrace{\mathbb{E}\big[\Phi'[y_i] \cdot \Phi'[y_j]\big]}_{0 \text{ symmetric!!!}}}}_{|t| > 2}$$

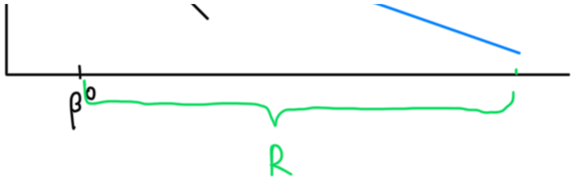
$$\le \frac{4}{n^2}\underbrace{\sum_{i=1}^{n}\|x_i\|^2}_{X^T X = dn \text{ orthogonal}} = \frac{4d}{n}$$

What if $\|u\| \ge R$?





lower bound for f beyond R



(uses convexity of f)

Claim:

w.h.p., $\forall u$ with $\|u\| \in R$, $\underline{f}(\beta^0+u) \geq \underline{f}(\beta^0) - \underbrace{\sqrt{\frac{400d}{n}}}_{C_1}\|u\| + \underbrace{\frac{\alpha}{10}}_{C_2}\|u\|^2$

if $C_1 \cdot R < C_2 \cdot R^2$, then $\|\underbrace{\underline{\beta} - \beta^0}_{\underline{u}}\|^2 \leq \frac{C_1^2}{C_2^2} = \frac{400d}{\alpha^2 n} \cdot 100 \precsim O\left(\frac{d}{\alpha^2 \cdot n}\right)$

$\mathbb{E}[\|\nabla \underline{f}\|^2] = \frac{4d}{n}$

$\rightarrow \|\nabla f\|^2 \leq \frac{400d}{n}$

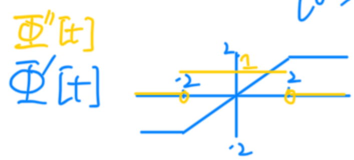
$\rightarrow$ need $R > \frac{C_1}{C_2} = O\left(\sqrt{\frac{d}{\alpha^2 \cdot n}}\right)$

first order Taylor expansion

$$\underline{f}(\beta^0+u) = \underline{f}(\beta^0) + \underbrace{\langle \nabla \underline{f}(\beta^0), u \rangle}_{\geq -\sqrt{\frac{400d}{n}}\|u\|} + \overbrace{\left\langle u, \left(\int_0^1 \underbrace{H[\underline{f}(\beta^0 + t \cdot u)]}_{\text{Hessian ???}} \cdot (1-t)\, dt\right) \cdot u \right\rangle}^{=\underline{M}(u)}$$

$$H[\underline{f}(\beta^0 + t \cdot u)] = \frac{1}{n}\sum_{i=1}^{n} \Phi''[\langle X_i, u \rangle - \underline{\eta}_i] \cdot X_i X_i^T$$

$*$ $\Phi''[t] = [\![\, |t| \leq 2 \,]\!] = \begin{cases} 1, & |t| \leq 2 \\ 0, & |t| > 2 \end{cases}$



need to show: $\underline{M}(u) \geq \frac{\alpha}{10} \cdot \|u\|^2$ ...

$$\underline{M}(u) = \frac{1}{n}\sum_{i=1}^{n}\left(\int_0^1 \underbrace{[\![\, |t \cdot \langle X_i, u \rangle - \underline{\eta}_i| \leq 2 \,]\!]}_{\geq [\![\, \cancel{t}\, |\langle X_i, u\rangle| \leq 1 \,]\!] \cdot [\![\, |\underline{\eta}_i| \leq 1 \,]\!]} \cdot (1-t)\, dt\right) \cdot \langle X_i, u \rangle^2$$

$*$ all terms ≥ 0

$|\langle X_i, u\rangle| \leq \|X_i\| \cdot \|u\| \leq C \cdot \sqrt{d} \cdot \|u\| \leq C\sqrt{d} \cdot R$

always satisfied if we choose $R = \frac{1}{C \cdot \sqrt{d}}$

if $R = \frac{1}{C \cdot \sqrt{d}}$, then

$$\underline{M}(u) \geq \underbrace{\frac{1}{n}\sum_{i=1}^{n} [\![\, |\underline{\eta}_i| \leq 1 \,]\!] \cdot \langle X_i, u \rangle^2 \cdot \overset{1/2}{\cancel{\int_0^1 (1-t)\, dt}}}_{= \underline{M}^{low}(u)}$$

$$\begin{aligned}
\mathbb{E}[\underline{M}^{low}(u)] &= \frac{1}{2n}\sum_{i=1}^{n} \mathbb{P}(|\underline{\eta}_i| \leq 1) \cdot \langle X_i, u \rangle^2 \\
&\geq \frac{\alpha}{2n}\sum_{i=1}^{n} \langle X_i, u \rangle^2 \\
&= \frac{\alpha}{2n} \cdot \langle u, X^T X u \rangle \\
&= \frac{\alpha}{2} \cdot \|u\|^2
\end{aligned}$$

$*$ $\alpha = \min_i \mathbb{P}(|\underline{\eta}_i| \leq 1)$

w.h.p. $\forall u$, $\underline{M}^{low}(u) \geq \frac{\alpha}{3}\|u\|^2$. ???