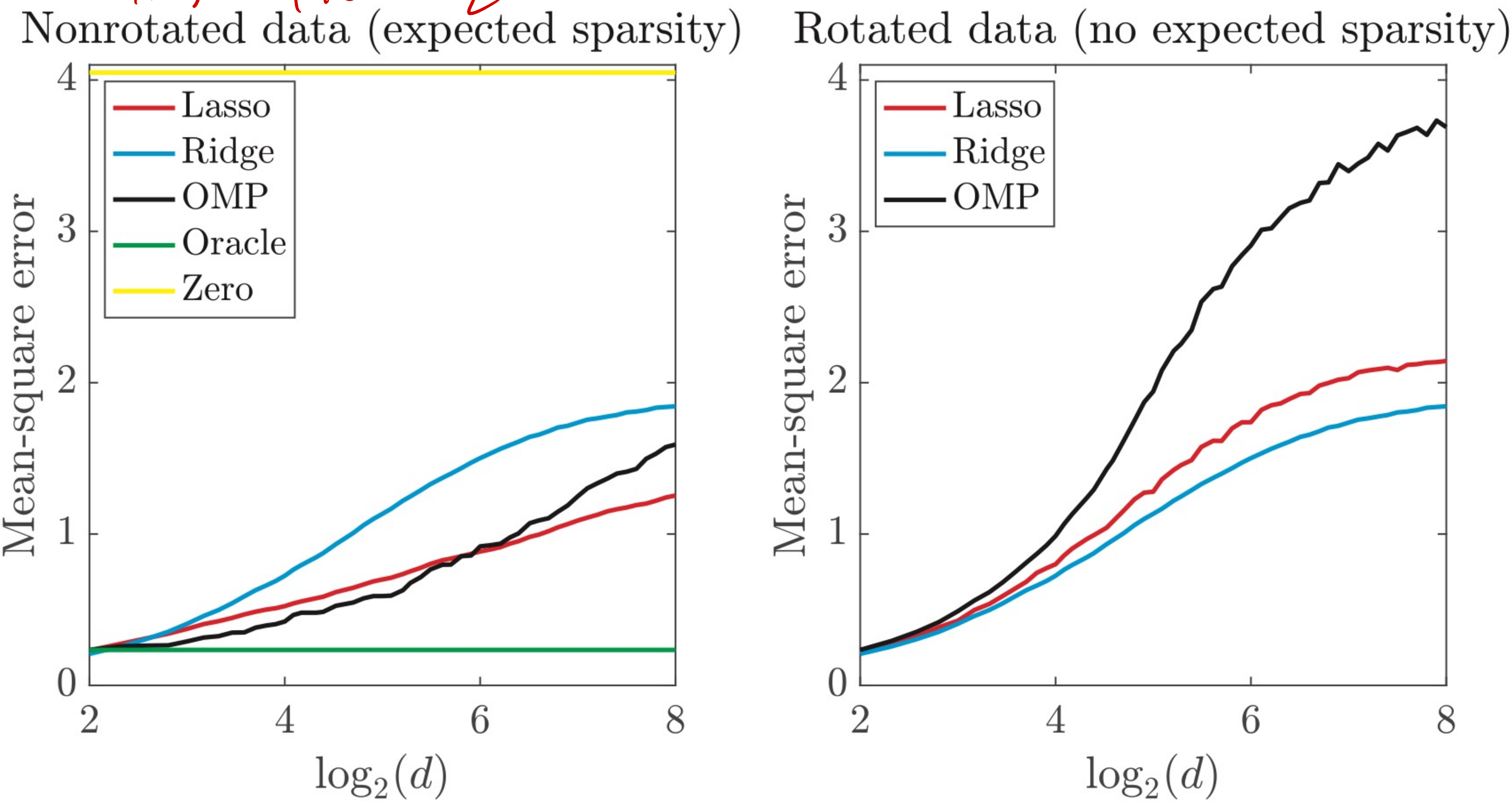
$\|y - \Phi\theta\|_2$ $\varphi(x_i) \longrightarrow R\varphi(x_i)$

$\|y - \Phi R^T\theta\|_2$ $\theta_\gamma \longrightarrow R\theta_\gamma$



Sparse methods – Model Selection

$$R(f) = \mathbb{E}\, \ell(y, f(x)) \quad , \quad x \in \mathcal{X} \in \mathbb{R}^d$$

$$\varphi : \mathcal{X} \to \mathbb{R}^d$$

Models: linear model $f(x) = \underline{\varphi(x)^T \theta}$

excess risk of OLS $\frac{\sigma^2 d}{n} \longrightarrow$ regularization $\boxed{\ell_2}$

dimension independent quantities $\sim \frac{1}{\sqrt{n}} \times \square$

$$\ell_2 \longrightarrow \boxed{\ell_1 \,|\, \ell_0}$$

Least-squares: $\hat{R}(\theta) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - \varphi(x_i)^T\theta)^2$

$$= \boxed{\frac{1}{2n} \| y - \Phi \theta \|_2^2}$$

$y \in \mathbb{R}^n$, Φ: design matrix / data $\in \mathbb{R}^{n \times d}$

Fixed design assumption: $\forall i,\ y_i = \varphi(x_i)^T \theta_* + \varepsilon_i$, with $\mathbb{E}\varepsilon_i = 0$, $\mathbb{E}\varepsilon_i^2 = \sigma^2$, independent

$$\boxed{y = \Phi\theta_* + \varepsilon} \Leftarrow$$

Criterion: "in sample" test error

$$\boxed{R(\theta) - R^* = \frac{1}{2n} \| \Phi (\theta - \theta_*) \|_2^2}$$

Φ: deterministic.

Lecture 2: $\hat{\theta} = \arg\min \hat{R}(\theta)$.

$$\mathbb{E} R(\hat{\theta}) = \frac{\sigma^2 \operatorname{rank}(\Phi)}{n}$$

A new analysis for OLS: $\hat{\theta} \in \underset{\theta \in \Theta}{\operatorname{argmin}} \hat{R}(\theta) = \frac{1}{2n}\|y - \Phi\theta\|_2^2$

characterize $\hat{\theta}$ by $\hat{R}(\hat{\theta}) \le \hat{R}(\theta_*)$

$$\Leftrightarrow \|y - \Phi\hat{\theta}\|_2^2 \le \|y - \Phi\theta_*\|_2^2 \qquad + \left(y = \Phi\theta_* + \varepsilon \right)$$

$$\Leftrightarrow \|\Phi(\theta_* - \hat{\theta}) + \varepsilon\|_2^2 \le \|\varepsilon\|_2^2$$

$$\Leftrightarrow \|\Phi(\theta_* - \hat{\theta})\|_2^2 - 2\varepsilon^\top \Phi(\hat{\theta} - \theta_*) + \|\varepsilon\|_2^2 \le \|\varepsilon\|_2^2$$

$$\Leftrightarrow \|\Phi(\hat{\theta} - \theta_*)\|_2^2 \le 2\varepsilon^\top \frac{\Phi(\hat{\theta} - \theta_*)}{\|\Phi(\hat{\theta} - \theta_*)\|_2} \|\Phi(\hat{\theta} - \theta_*)\|_2$$

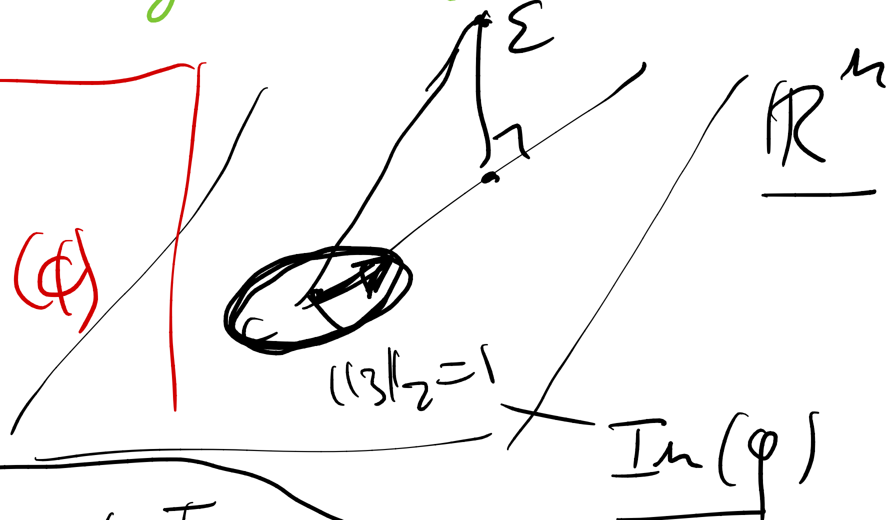
$$\Leftrightarrow \|\Phi(\hat{\theta} - \theta_*)\|_2 \le \frac{2\varepsilon^\top \Phi(\hat{\theta} - \theta_*)}{\|\Phi(\hat{\theta} - \theta_*)\|_2}$$

$$\Rightarrow \underbrace{\|\Phi(\hat{\theta} - \theta_*)\|_2^2}_{\text{"} = 2nR(\hat{\theta})} \le \sup_{\theta \in \Theta} \left(\frac{2\varepsilon^\top \Phi(\theta - \theta_*)}{\|\Phi(\theta - \theta_*)\|_2} \right)^2$$

$$\mathbb{E}_\varepsilon\left[\sup_{\theta\in\mathbb{R}^d}\left(2\varepsilon^T \boxed{\frac{\Phi(\theta-\theta_*)}{\|\Phi(\theta-\theta_*)\|_2}}\right)^2\right] = \mathbb{E}_\varepsilon\left[\sup_{\substack{z\in \mathrm{Im}(\Phi)\\ \|z\|_2=1}} (z^T\varepsilon)^2\right]$$

$$\mathbb{R}^n \ni z \Rightarrow \|z\|_2 = 1,\quad z\in\mathrm{Im}(\Phi)$$

$$\mathbb{E}[2nR(\hat\theta)] \le 4\sigma^2\,\mathrm{rank}(\Phi)$$



$$z^* = \frac{\Pi_\Phi\,\varepsilon}{\|\Pi_\Phi\varepsilon\|}$$

where $\Pi_\Phi \in \mathbb{R}^{n\times n}$ = orthogonal projection on $\mathrm{Im}(\Phi)$

$$\Rightarrow \Pi_\Phi^2 = \Pi_\Phi$$

$$\Pi_\Phi = \Phi(\Phi^T\Phi)^{-1}\Phi^T \qquad \Phi\in\mathbb{R}^{n\times d}$$

$$\Pi_\Phi(\Phi a) = \Phi(\Phi^T\Phi)^{-1}\Phi^T\Phi a = \Phi a$$

$$\mathbb{E}_\varepsilon\left(\sup_{\substack{\|z\|_2=1\\ z\in\mathrm{Im}(\Phi)}}(\varepsilon^T z)^2\right) = \mathbb{E}_\varepsilon\left(\varepsilon^T\overbrace{\frac{\Pi_\Phi\varepsilon}{\|\Pi_\Phi\varepsilon\|_2}}^{z}\right)^2 = \mathbb{E}_\varepsilon\left(\frac{(\varepsilon^T\Pi_\Phi\varepsilon)^2}{\varepsilon^T\Pi_\Phi^2\varepsilon}\right) = \mathbb{E}[\varepsilon^T\Pi_\Phi\varepsilon]$$

$$= \mathbb{E}\,\mathrm{tr}\,\Pi_\Phi\varepsilon\varepsilon^T = \sigma^2\,\mathrm{tr}\,\Pi_\Phi = \sigma^2\,\mathrm{rank}(\Phi)$$

Variable selection by ℓ_0-penalty

Model $y = \Phi\theta_* + \varepsilon$ with $\|\theta_*\|_0 = \#$ of nonzero components of θ_*

$\|\theta_*\|_0 \leq k$

⚠ if the support $\text{support}(\theta_*) = \{j, \theta_j \neq 0\}$.

$\Rightarrow$ excess risk is $\dfrac{\sigma^2 k}{n}$

Goal : achieve $\dfrac{\sigma^2 k}{n} \log d$ price of adaptivity

$\dfrac{d}{n} \longrightarrow \dfrac{k \log d}{n} \ll \dfrac{d}{n}$

(1) Assume k is known $\longrightarrow$ $\hat{\theta} = \underset{\|\theta\|_0 \leq k}{\arg\min}\ \hat{R}(\theta)$

(2) k unknown

$$\|\Phi(\hat{\theta}-\theta_*)\|_2^2 \le 4 \sup_{\|\theta\|_0 \le h} \left(\varepsilon^T \frac{\Phi(\theta-\theta_*)}{\|\Phi(\theta-\theta_*)\|_2} \right)^2 \qquad \underline{2h \le d}$$

$$\le 4 \sup_{\|\theta-\theta_*\|_0 \le 2h} \left(\qquad \right) \qquad \frac{\varepsilon^T \Phi(\theta-\theta_*)}{\|\Phi(\theta-\theta_*)\|_2}$$

$$\le 4 \sup_{\substack{B \subset \{1,\dots,d\} \\ |B| \le 2h}} \underbrace{\sup_{supp(\theta-\theta_*) \subset B}}_{\|\Pi_{\Phi_B}\varepsilon\|_2^2}$$

if $\Phi_B \in \mathbb{R}^{n\times|B|}$ submatrix of Φ with columns in B

$$\le 4 \sup_{|B| \le 2h} \|\Pi_{\Phi_B}\varepsilon\|_2^2 \simeq \sigma^2 h \log d$$

$\overset{?}{=} 2h$ (by monotonicity)

Assumption: ε is gaussian iid, $\varepsilon_i \sim \mathcal{N}(0,\sigma^2)$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$$

$$\Rightarrow A\varepsilon \sim \mathcal{N}(0, \sigma^2 AA^T) \quad \text{because } Cov(A\varepsilon) = A\, Cov(\varepsilon) A^T$$

$$\Phi_B \varepsilon \sim \mathcal{N}(0, \sigma^2 \Pi_{\Phi_B})$$

Π_{Φ_B} is an orthogonal projection on a space of dim $\le 2h$

WLOG $\Pi_{\Phi_B} = \begin{pmatrix} 1 \cdot\cdot 1 & 0 \\ 0 & 0 \end{pmatrix}$

$$\Pi_{\Phi_B}\varepsilon \sim \mathcal{N}\left(0,\ \sigma^2\, \Pi_{\Phi_B}\right)$$

$$= R\, \mathrm{Diag}\begin{pmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} R^T$$

basis of eigenvectors

$$R^T \Pi_{\Phi_B}\varepsilon \sim \mathcal{N}\big(0,\ \sigma^2 \underbrace{R^T R}_{I}\, \mathrm{Diag}(\ )\, \underbrace{R^T R}_{I}\big)$$

$$\sim \mathcal{N}\left(0,\ \sigma^2\, \mathrm{Diag}\begin{pmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}\right)$$

$\Pi_{\Phi_B}\varepsilon$ is gaussian with mean zero and cov mat $\sigma^2 I_{2h}$.

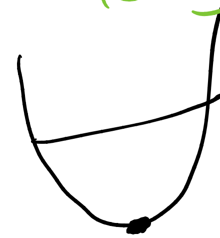
goal: $$\mathbb{E} \max_{|B|=2h} \|\Pi_{\Phi_B}\varepsilon\|_2^2$$

$$\hookrightarrow \mathbb{E} = \sigma^2\, 2h$$

Lemma: max of Gaussians

$\varepsilon_B = \Pi_{\phi_B}(\varepsilon)$.

$\mathbb{E} \max_{|B|=2k} \|\varepsilon_B\|_2^2$?



By Jensen's inequality:

$$\mathbb{E}(e^{Z}) \geq e^{\mathbb{E}Z}$$

$$\log(\mathbb{E}e^{sZ}) \geq s\,\mathbb{E}Z$$

$$\frac{1}{s}\log(\mathbb{E}e^{sZ}) \geq \mathbb{E}Z \qquad \text{"Chernoff"}$$

$$\mathbb{E} \max_{|B|=2k} \|\varepsilon_B\|_2^2 \leq \frac{1}{s}\log\left(\mathbb{E}e^{s\max_B \|\varepsilon_B\|_2^2}\right) = \frac{1}{s}\log\left(\mathbb{E}\max_B e^{s\|\varepsilon_B\|_2^2}\right)$$

$$\leq \frac{1}{s}\log\left(\sum_{|B|=2k}\mathbb{E}e^{s\|\varepsilon_B\|_2^2}\right) \qquad \text{max} \leq \text{sum}$$

$$= \frac{1}{s}\log\left(\sum_{|B|=2k}\prod_{i=1}^{2k}\underline{\mathbb{E}e^{s(\varepsilon_B)_i^2}}\right) \qquad \text{by independence}$$

$$= \frac{1}{s}\log\left(\sum_{|B|=2k}\left(\frac{1}{\underbrace{1-2s\sigma^2}_{1/2}}\right)^{k}\right)$$

$$= 4\sigma^2\log\left[2^{2k}\ \#\text{ of subsets with cardinality} \leq 2k\right] \qquad s=\frac{1}{4\sigma^2}$$

$$= 4\sigma^2 k + \sigma^2\log\binom{d}{2k}$$

Lemma: $\varepsilon \sim \mathcal{N}(0,\sigma^2)$

$$\mathbb{E}e^{s\varepsilon^2} = \frac{1}{\sqrt{1-2s\sigma^2}}$$

$\boxed{\text{Lemma 8.1}}$

Lemma 1: # of subsets $B \subset \{1, \dots, d\}$, $|B| = 2h$

$$= \binom{d}{2h} \Rightarrow \log \binom{d}{2h} \le 2h\left[1 + \log \frac{d}{2h}\right].$$

$$\frac{d(d-1)\cdots(d-2h+1)}{(2h)!}$$

$$\binom{d}{2h} \le \left(\frac{de}{2h}\right)^{2h}$$

Lemma 8.3 Sach

- Algorithm: $\min\limits_{\|\theta\|_0 \le h} \|y - \phi\theta\|_2^2$

search

- ℓ_0-penalty $\Rightarrow \min\limits_{2h} \frac{1}{n} \|y - \phi\theta\|_2^2 + \lambda \|\theta\|_0$

$\frac{\sigma^2 h \log d}{n}$, with no assumptions on ϕ, hard to compute

ℓ_1-norm penalty: estimator

$$\hat{\theta} = \arg\min_{\theta} \hat{R}(\theta) + \lambda\|\theta\|_1$$

$$\underbrace{\frac{1}{2n}\|y-\Phi\theta\|_2^2 + \lambda\|\theta\|_1}_{\text{LASSO}}$$

where $\|\theta\|_1 = \sum_{j=1}^{d} |\theta_j|$

Why is it leading to sparse solutions?

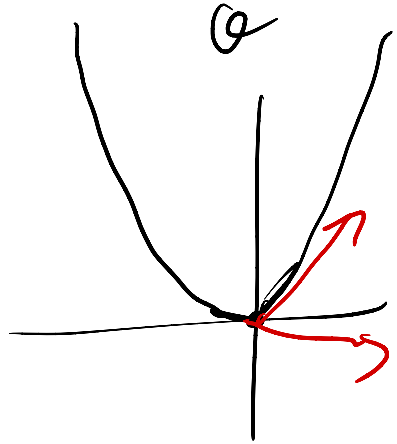
in 1D: $\min_{\theta} \frac{1}{2}|\theta-y|^2 \to$ argmin is $\theta = y$

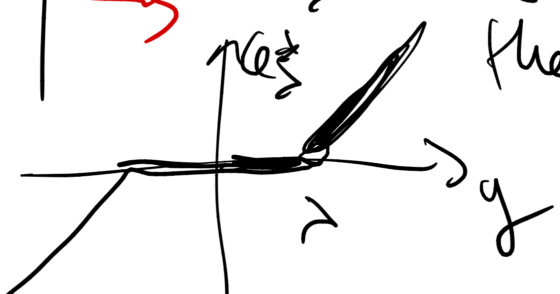
$\min_{\theta} \underbrace{\frac{1}{2}|\theta-y|^2 + \lambda|\theta|}_{=F(\theta)} \longrightarrow$ argmin θ_λ.

$$F'(0_+) = -y+\lambda$$
$$F'(0_-) = -y-\lambda$$

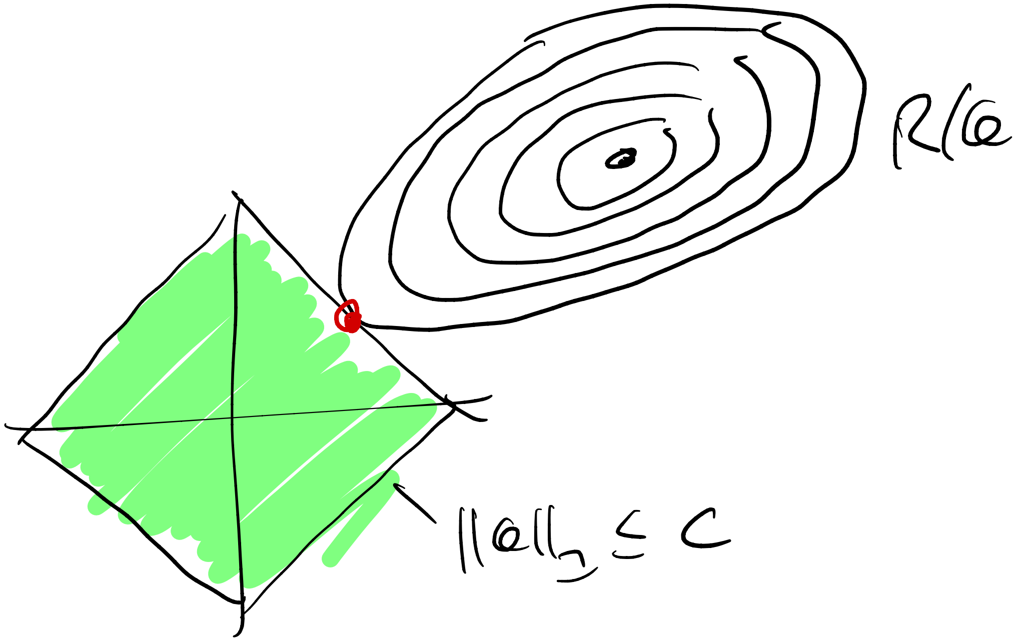
$\theta_\lambda > 0$ if $F'(0_+) < 0$ then $\theta = y-\lambda$

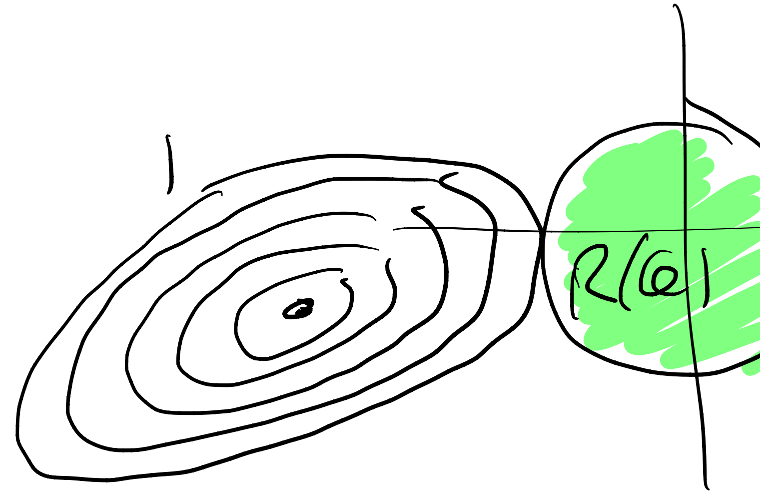
$$\theta_\lambda = 0 \iff \begin{cases} \lambda - y = F'(0_+) \geq 0 \\ -\lambda - y = F'(0_-) \leq 0 \end{cases} \iff |y| \leq \lambda$$

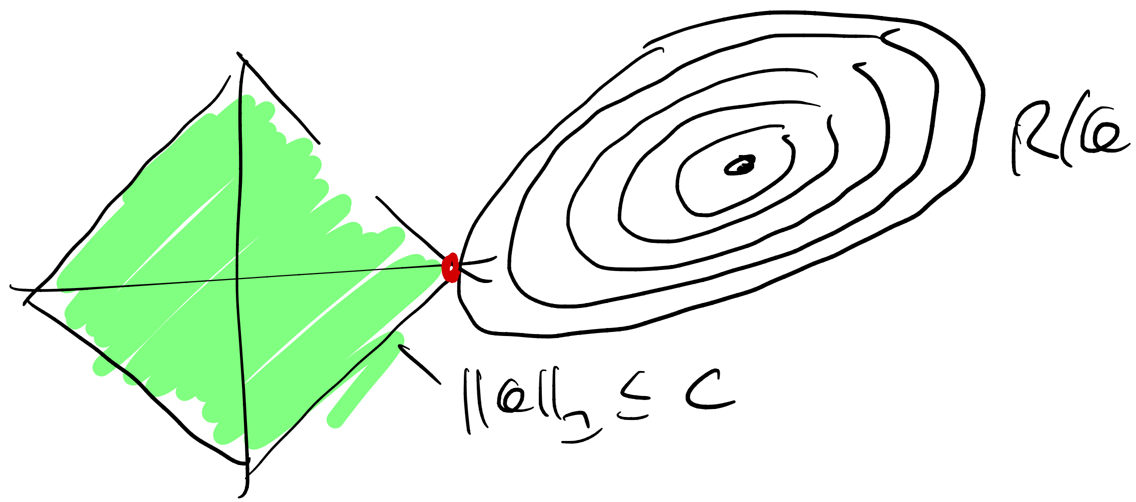




<u>in 2D</u> : $\min R(\theta) \quad s.t. \quad \|\theta\|_1 \leq C$





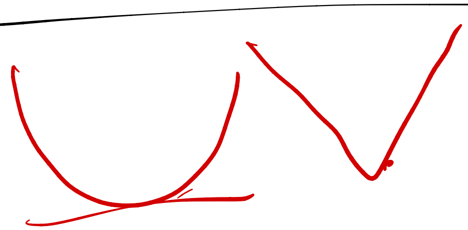


Algorithms to minimize with L_1 penalty

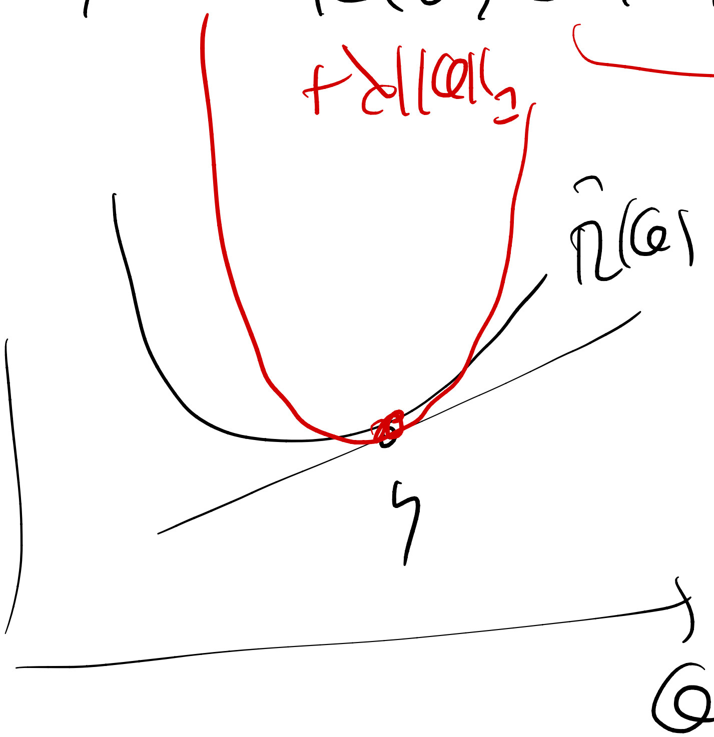
"proximal method" $\quad \min \hat{R}(\theta) + \lambda \|\theta\|_1$ } non smooth

convex smooth function

$\hat{R}''(\theta)$ have eigenvalues in $[0, L]$.



$$\forall \eta, \theta. \quad \hat{R}(\theta) \le \hat{R}(\eta) + \hat{R}'(\eta)^T(\theta - \eta) + \frac{L}{2}\|\theta - \eta\|_2^2 + \lambda\|\theta\|_1$$



$\hookrightarrow \arg\min_\theta \Rightarrow$ gradient $= \hat{R}'(\eta) + L(\theta - \eta) = 0$

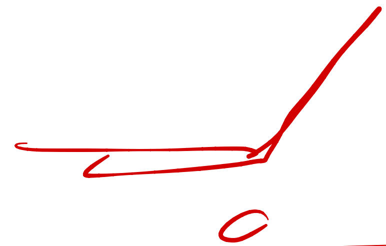
$\Leftrightarrow \theta = \eta - \frac{1}{L}\hat{R}'(\eta)$

GD

$\arg\min_\theta$ can be done in closed form

$$\min_{\theta} \; R(\theta) + \lambda \|\theta\|_2^2 \quad \checkmark \quad \rightarrow \text{ minimizers are at the kink}$$

$$\min \; R\Big(\sum a_j (w_j^T x + b_j)_+ \Big)$$



Analysis $\quad \inf \hat{R}(\theta) \quad \text{s.t.} \quad \|\theta\|_1 \le D \;\Rightarrow\; \hat{\theta}$

$$\hat{R}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_\theta(x_i)), \qquad f_\theta(x_i) = \varphi(x_i)^T \theta$$

$\boxed{\text{RANDOM DESIGN}}$

if Lipschitz cont. max loss (Lipschitz constant G) $\quad \|\varphi(x)\|_\infty \le R$

we have

$$R(\hat{\theta}) - \inf_{\|\theta\|_1 \le D} R(\theta) \le \frac{GRD\sqrt{\log d}}{\sqrt{n}}$$

$$D = \|\theta_*\|_1 \;\rightarrow\; \|\theta_*\|_1 \frac{\sqrt{\log d}}{\sqrt{n}}$$

$\boxed{\text{HIGH DIMENSIONAL PROPERTY}}$

Lasso $\sqrt{\frac{\log d}{n}}$ = slow rate
$\frac{\log d}{n}$ = fast rate) Lasso

Not covered = fast rates for Lasso if low-correlation in Φ.

$\frac{1}{n}\Phi^T\Phi \sim$ Diagonal

Fixed design slow rate for Lasso

Fixed design slow rate for Lasso:

$$\hat{\theta} = \arg\min \frac{1}{2n}\|y - \Phi\theta\|_2^2 + \lambda\|\theta\|_1$$

Model: $y = \Phi\theta_* + \varepsilon$

(Goal: $\frac{1}{n}\|\Phi(\hat{\theta} - \theta_*)\|_2^2 \leq \bullet$)

Lemma: if $\|\Phi^T\varepsilon\|_\infty \leq \frac{n\lambda}{2}$, then $\|\hat{\theta}\|_1 \leq 3\|\theta_*\|_1$

$$\frac{1}{n}\|\Phi(\hat{\theta} - \theta_*)\|_2^2 \leq 3\lambda\|\theta_*\|_1$$

Lemma: $a^T b \leq \|a\|_1 \cdot \|b\|_\infty$

$\sum a_i b_i \leq \|b\|_\infty \sum |a_i|$

Proof:

$$\frac{1}{2n}\|y - \Phi\hat{\theta}\|_2^2 + \lambda\|\hat{\theta}\|_1 \leq \frac{1}{2n}\|y - \Phi\theta_*\|_2^2 + \lambda\|\theta^*\|_1$$

$$\Leftrightarrow \|\varepsilon - \Phi(\hat{\theta} - \theta_*)\|_2^2 + 2n\lambda\|\hat{\theta}\|_1 \leq \|\varepsilon\|_2^2 + 2n\lambda\|\theta^*\|_1$$

$$\|\Phi(\hat{\theta} - \theta_*)\|_2^2 - 2\varepsilon^T\Phi(\hat{\theta} - \theta_*) + 2n\lambda\|\hat{\theta}\|_1 \leq 2n\lambda\|\theta^*\|_1$$

$$\Leftrightarrow \|\Phi(\hat{\theta} - \theta_*)\|_2^2 \leq 2\varepsilon^T\Phi(\hat{\theta} - \theta^*) + 2n\lambda\|\theta^*\|_1 - 2n\lambda\|\hat{\theta}\|_1$$

$$\leq 2\|\hat{\theta} - \theta_*\|_1 \underbrace{\|\Phi^T\varepsilon\|_\infty}_{\leq \frac{n\lambda}{2}} + 2n\lambda\|\theta^*\|_1 - 2n\lambda\|\hat{\theta}\|_1$$

$$\leq \underbrace{\|\hat{\theta} - \theta_*\|_1}_{\leq \|\hat{\theta}\|_1 + \|\theta_*\|_1} n\lambda + 2n\lambda\|\theta_*\|_1 - 2n\lambda\|\hat{\theta}\|_1$$

$$\leq 3n\lambda\|\theta_*\|_1 - n\lambda\|\hat{\theta}\|_1$$

What is $\|\frac{\Phi^T\varepsilon}{n}\|_\infty$?

$$= \left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\varphi(x_i)\right\|_\infty$$

$$= \max_{j\in\{1,\dots,d\},\, \eta\in\{-1,1\}} \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\varphi(x_i)_j\eta$$

Final result: assume ε_i is almost surely bounded by σ

$\mathbb{E}\varepsilon_i = 0$

$\mathbb{E}\varepsilon_i^2 \leq \sigma^2$

. assume $\|\varphi(x_i)\|_\infty \leq R$

$$\mathbb{P}\left(\|\Phi^T\varepsilon\|_\infty > \frac{n\lambda}{2} \right) \underset{\text{union bound}}{\leq} \sum_{j=1}^{d} \mathbb{P}\left(|\Phi^T\varepsilon|_j > \frac{n\lambda}{2} \right)$$

$$\leq \sum_{j=1}^{d} \mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} \underbrace{\varepsilon_i \varphi(x_i)_j}_{\text{zero mean},\ |\ | \leq \sigma R} \right| > \frac{n\lambda}{2} \right)$$

Hoeffding

$$\leq 2d\, e^{-n\lambda^2/\sigma^2 R^2}$$

$$\leq 2d \exp\left(\frac{-n\lambda^2}{\sigma^2 R^2} \right) = \delta$$

up to constants

$$\boxed{\lambda = \frac{\sigma R}{\sqrt{n}} \log\left(\frac{2d}{\delta} \right)}$$

$\Rightarrow$ with proba $\geq 1-\delta$, $\frac{1}{n}\|\Phi(\theta - \theta_*)\|_2^2 \leq \text{cste}\, \|\theta_*\|_1 \frac{\sigma R}{\sqrt{n}}\left(\sqrt{\log d} + \sqrt{\log \frac{1}{\delta}}\right)$