Variable selection: ℓ_0-penalty $\sigma^2 k \log d / n$

ℓ_1-penalty: slow rate as $\sqrt{\frac{\log d}{n}}$

Impact of dimension $\sigma^2 d / n$

$\frac{1}{n^{2/2+d}} \sim \frac{1}{n^{2/d}}$ — k-NN — curse of dim

variable selection $\Rightarrow$ d smaller $\Rightarrow$ k

① How to pick k "good" variables

② Can it work?

Need assumptions

Focus on least-squares with linear models

ℓ_2-regularization — class

class last week

ℓ_1 or ℓ_0

Constrained least-squares regression

$(x_i, y_i) \in \mathcal{X} \times \mathbb{R}$, $i = 1, \dots, n$, $f_\theta(x) = \theta^\top \underbrace{\varphi(x)}_{\text{feature vector}} \in \mathbb{R}^d$

Design matrix $\Phi \in \mathbb{R}^{n \times d} = \begin{pmatrix} \varphi(x_1)^\top \\ \vdots \\ \varphi(x_n)^\top \end{pmatrix}$, $y \in \mathbb{R}^n$ → vector of outputs responses labels

Empirical risk:

$$\frac{1}{n} \sum_{i=1}^{n} |y_i - \theta^\top \varphi(x_i)|^2 = \frac{1}{n} \|y - \Phi\theta\|_2^2$$

Criterion for performance

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

Fixed design assumption: $x_1, \dots, x_n$ are deterministic

$y_1, \dots, y_n$ are random

$\forall i$, $y_i = \varphi(x_i)^\top \theta_* + \varepsilon_i$ model is well-specified

ε Gaussian with mean 0 and covariance matrix $\sigma^2 I$

For OLS: $\hat{\theta} = \arg\min \frac{1}{n}\|y-\phi\theta\|_2^2$ $\quad \hat{\theta} = (\phi^T\phi)^{-1}\phi^T y$

Direct computations: $y = \phi\theta_* + \varepsilon$

$\mathbb{E}\,\|\phi(\hat{\theta}-\theta_*)\|_2^2 = \sigma^2 d$

Add constraints: $\theta \in \Theta$

$\hat{\theta} = \arg\min_{\theta\in\Theta} \frac{1}{n}\|y-\phi\theta\|_2^2 \Rightarrow \|y-\phi\hat{\theta}\|_2^2 \le \|y-\phi\theta_*\|_2^2$ $\quad \theta_* \in \Theta$

$y - \phi\hat{\theta} = \phi\theta_* - \phi\hat{\theta} + \varepsilon$

$= \phi(\theta_* - \hat{\theta}) + \varepsilon$

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

$\|\phi(\hat{\theta}-\theta_*)\|_2^2 \le 2\varepsilon^T\phi(\hat{\theta}-\theta_*)$

Option 1: $\|\phi(\hat{\theta}-\theta_*)\|_2^2 \le 2\|\varepsilon\|_2 \cdot \|\phi(\hat{\theta}-\theta_*)\|_2$ (candy-Schwarz)

$\|\phi(\hat{\theta}-\theta_*)\|_2 \le 2\|\varepsilon\|_2$

$\frac{1}{n}\|\phi(\hat{\theta}-\theta_*)\|_2^2 \le \frac{4\|\varepsilon\|_2^2}{n}$ with $\mathbb{E}\|\varepsilon\|_2^2 = \sigma^2 d$

Option 2: $\|\phi(\hat{\theta}-\theta_*)\|_2 \le 2\varepsilon^T\left(\frac{\phi(\hat{\theta}-\theta_*)}{\|\phi(\hat{\theta}-\theta_*)\|_2}\right)$

$\le 2\sup_{\theta\in\Theta} \varepsilon^T\left(\frac{\phi(\theta-\theta_*)}{\|\phi(\theta-\theta_*)\|_2}\right)$

ℓ_0-penalty / ℓ_0-constraint: $\|\theta\|_0 = \#$ of non-zero components of θ

$= \sum_{j=1}^{d} |\theta_j|^0$ / "$0^0 = 0$"

(*) $= \{\theta \in \mathbb{R}^d, \|\theta\|_0 \le k\}$

$\theta^T \varphi(x_i) = \sum_{j=1}^{d} \theta_j \varphi(x_i)_j$

$\hat{\theta} = \underset{\|\theta\|_0 \le k}{\operatorname{argmin}} \|y - \Phi\theta\|_2^2 \Rightarrow$ Algorithms: ① exhaustive search a sets $|A| \le k$ (can restrict to $|A| = k$)

$= \underset{|A| \le k}{\operatorname{argmin}} \left[\min_{\theta_{A^c} = 0} \|y - \Phi\theta\|_2^2 \right]$

function of A is non-increasing

support of θ

There are $\binom{d}{k} = \frac{d!}{k!(d-k)!} \le \frac{d^k}{k!}$ such sets

$\le \left(\frac{ed}{k}\right)^k \Rightarrow$ see back

② greedy: efficient but no optimal

$$\|\phi(\hat{\theta}-\theta_*)\|_2 \le 2\varepsilon^T\left(\frac{\phi(\hat{\theta}-\theta_*)}{\|\phi(\hat{\theta}-\theta_*)\|_2}\right)$$

$$\le 2\sup_{\|\theta\|_0\le k} \varepsilon^T\left(\frac{\phi(\theta-\theta_*)}{\|\phi(\theta-\theta_*)\|_2}\right)$$

$\|\theta_*\|_0 \le k$

$\hookrightarrow \|\theta-\theta_*\|_0 \le 2k$

$$\le 2\sup_{\|\theta-\theta_*\|_0\le 2k} \varepsilon^T\frac{\phi(\theta-\theta_*)}{\|\phi(\theta-\theta_*)\|_2}$$

$$\le 2\sup_{|B|\le 2k}\ \sup_{(\theta-\theta_*)_{B^C}=0} \varepsilon^T\frac{\phi(\theta-\theta_*)}{\|\phi(\theta-\theta_*)\|_2}$$

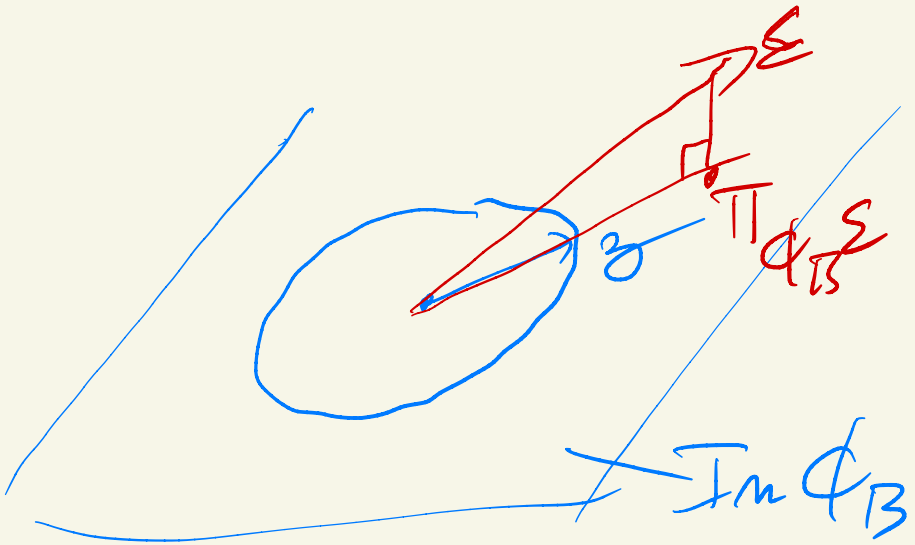
$$\le 2\sup_{|B|\le 2k}\left[\sup_{z \text{ s.t } \|z\|_2=1,\ z\in \mathrm{Im}\,\phi_B} z^T\varepsilon\right] M\begin{pmatrix} d \\ \vdots \\ B \end{pmatrix}$$

$$= \|\Pi_{\phi_B}\varepsilon\|_2$$

$\|z\|_2=1$, $z\in \mathrm{Im}\,\phi_B$

ϕ_B: submatrix of ϕ with columns indexed by B

$$2 \sup_{|B| \le 2h} \left\{ \sup_{z \text{ s.t. } \|z\|_2 = 1,\ z \in \mathrm{Im}\,\phi_B} z^T \varepsilon \right\} = \|\Pi_{\phi_B} \varepsilon\|_2$$



$$z^T \varepsilon = \left(\Pi_{\phi_B}\varepsilon + \varepsilon - \Pi_{\phi_B}\varepsilon\right)^T z$$

$$= z^T \Pi_{\phi_B}\varepsilon + 0 \le \|z\|_2 \|\Pi_{\phi_B}\varepsilon\|_2 \quad \text{by Cauchy Schwarz}$$

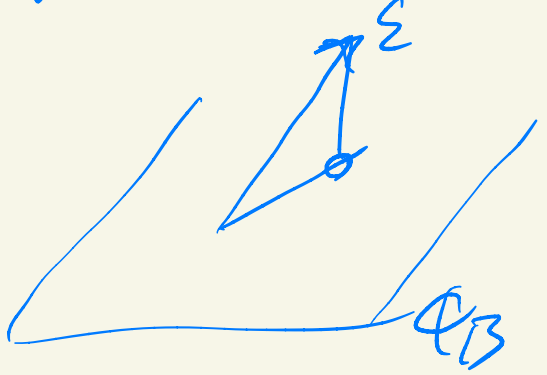
Summary

$$\|\phi(\hat{\theta} - \theta_*)\|_2 \le 2 \sup_{|B| \le 2h} \|\Pi_{\phi_B}\varepsilon\|_2$$

ε is gaussian with covariance matrix $\sigma^2 I$

$\Pi_{\phi_B}\varepsilon$ is gaussian

If ε is gaussian with mean zero and covariance matrix $\sigma^2 I$



$\Pi_{\Phi_B}\varepsilon$ is a gaussian in dimension $|B|$ and covariance matrix $\sigma^2 I$

$\in \mathbb{R}^d$

WLOG, $\Phi_B = \text{span}(e_1, \dots, e_{|B|})$ canonical basis of $\mathbb{R}^d$

$\mathbb{E}\|\Pi_{\Phi_B}\varepsilon\|^2 = \sigma^2|B|$

$\sigma^2\Pi_\Phi$

$$\mathbb{E}\sup_{|B|=2h} \|\Pi_{\Phi_B}\varepsilon\|_2^2 - \|\beta_B\|_2^2$$

$$\mathbb{E}\, e^{s\|\beta_B\|^2} \left(\frac{1}{1-2\sigma^2 s}\right)^{|B|/2}$$

$$m = \binom{d}{2h}$$

h

Lemma 1: $Z_1, \dots, Z_n$ random variables not necessarily independent

$$\mathbb{E}\max\{Z_1,\dots,Z_n\} \le \frac{1}{s}\log\left(\mathbb{E}e^{s\max\{Z_1,\dots,Z_n\}}\right)$$

$$\le \frac{1}{s}\log\left(\mathbb{E}\max\{e^{sZ_1},\dots,e^{sZ_n}\}\right) \quad \text{(exp is increasing)}$$

$$\le \frac{1}{s}\log\left(\mathbb{E}e^{sZ_1}+\dots+\mathbb{E}e^{sZ_n}\right)$$

$\forall s>0$

$\mathbb{E}Y \le \frac{1}{s}\log(\mathbb{E}e^{sY})$

$t \mapsto e^{st}$ convex

$\mathbb{E}(e^{sY}) \ge e^{s\mathbb{E}Y}$

Jensen's

Lemma 2: if y is gaussian with covariance matrix $\sigma^2 I$

$y \in \mathbb{R}^n$

then $\mathbb{E}e^{s\|y\|_2^2} = \mathbb{E}\prod_{i=1}^n e^{sy_i^2} \overset{\text{indep}}{=} \prod_{i=1}^n \mathbb{E}e^{sy_i^2} = \left(\mathbb{E}e^{sy_1^2}\right)^n$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\int e^{sy_1^2} e^{-\frac{1}{2}\frac{y_1^2}{\sigma^2}}\,dy_1\right)^n = \left(\frac{1}{\sqrt{e^{-\frac{1}{2}y_1^2\left[\frac{1}{\sigma^2}-2s\right]}}}\right)^n$$

$$= \left(\frac{1}{\sqrt{1-2s\sigma^2}}\right)^n = \left(\frac{1}{1-2s\sigma^2}\right)^{\frac{n}{2}}$$

$s < \frac{1}{2\sigma^2}$

$\mathbb{E} \sup_{|B|=2h} \|\Pi_{\mathcal{L}_B} \varepsilon\|_2^2 \quad \|\beta_B\|_2^2 \quad \mathbb{E} e^{s\|\beta_B\|^2} \le \left(\frac{1}{1-2\sigma^2 s}\right)^{|B|/2}$

$m = \binom{d}{2h}^h$

$\boxed{\begin{matrix} 0.32 \\ 0.50 \end{matrix}}$

$$\mathbb{E} \sup_{|B|=2h} \|\Pi_{\mathcal{L}_B} \varepsilon\|_2^2 \le \frac{1}{s} \log\left[m \frac{1}{(1-2\sigma^2 s)^h} \right] \qquad \left(s = 1/4\sigma^2\right)$$

$$\le 4\sigma^2 \log\left[m 2^h \right]$$

$$\le 4\sigma^2 \left[\log \binom{d}{2h} + h \log 2 \right]$$

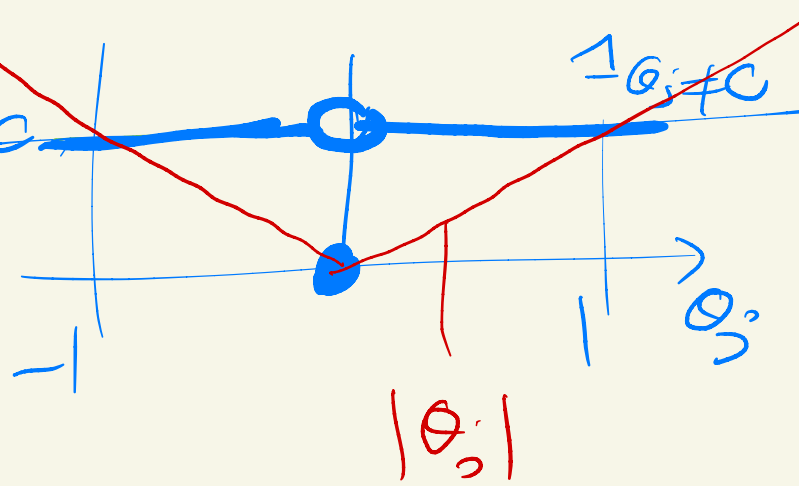
price of adaptivity

$$\le \log\left(\left(\frac{ed}{2h}\right)^{2h} \right) = 2h \log\left(\frac{ed}{h}\right)$$

$$= 2h + \boxed{2h \log \frac{d}{h}}$$

$$\le 8\sigma^2 h \boxed{\log \frac{d}{h}} + \sigma^2 \square h$$

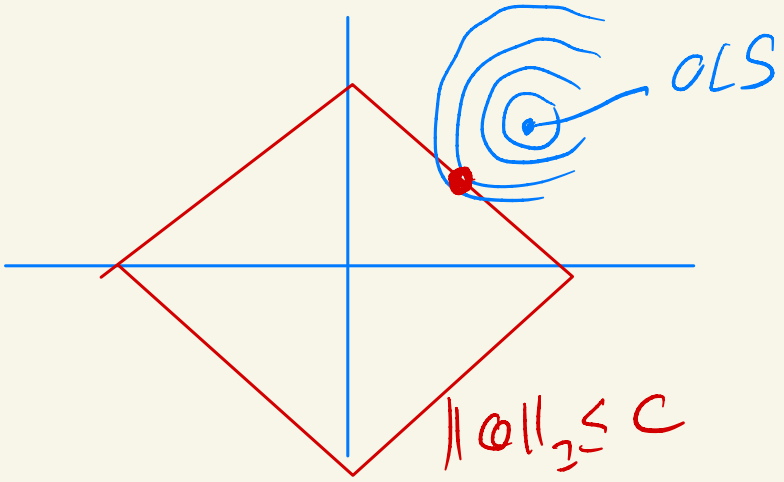
From ℓ_0 to ℓ_1 : $\|\theta\|_0 = \sum_{j=1}^{d} 1_{\theta_j \neq 0}$

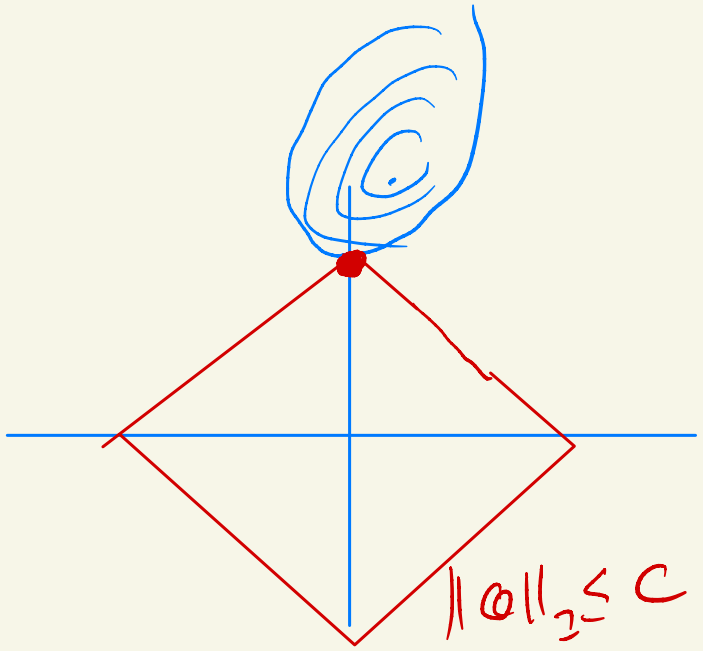


$\|\theta\|_1$ is the convex envelope of $\|\theta\|_0$ on $[-1,1]^d$

Constrained or penalized by ℓ_1-norm

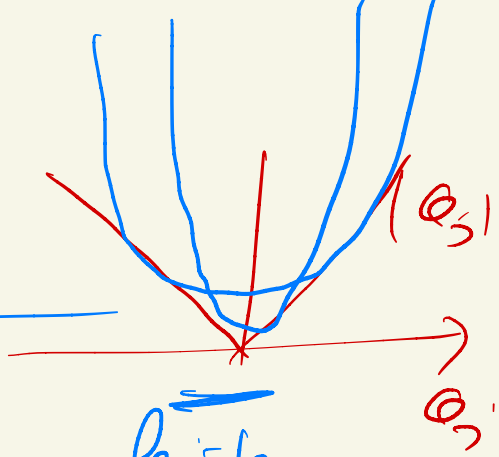
$\min \|\phi\theta - y\|^2$ s.t $\|\theta\|_1 \leq c$





Lasso : $\min_{\theta} \frac{1}{2n} \|y - \phi\theta\|_2^2 + \lambda \|\theta\|_1$

$$+\lambda \sum_j |\theta_j|$$



① Algorithms

"η-trick" : $|\theta_j| = \inf_{\eta_j > 0} \frac{1}{2} \frac{\theta_j^2}{\eta_j} + \frac{1}{2} \eta_j$

with $\eta_j^* = |\theta_j|$

$$\min_{\theta \in \mathbb{R}^d} \min_{\eta \in \mathbb{R}_+^d} \frac{1}{2n} \|y - \phi\theta\|_2^2 + \frac{\lambda}{2} \sum_j \frac{\theta_j^2}{\eta_j} + \frac{\lambda}{2} \sum_j \eta_j$$

with η fixed = quadratic problem in θ $\Rightarrow$ closed form

with θ fixed $\Rightarrow$ $\eta_j^* = |\theta_j|$

$\Rightarrow$ alternate optimization

For exact algos: proximal methods $\Rightarrow$ see blog post

(2) Analysis → slow rate with "no" assumptions $\sqrt{\frac{\log d}{n}}$

→ fast rate with strong assumptions $\frac{\sigma^2 k \log d}{n}$ ⟹ see Dach

---

Consider $\hat{\theta} = \operatorname{argmin} \frac{1}{2n} \|y - \phi\theta\|_2^2 + \lambda \Omega(\theta)$ — norm

$y = \phi\theta_* + \varepsilon$

$$\frac{1}{2n} \|y - \phi\hat{\theta}\|_2^2 + \lambda \Omega(\hat{\theta}) \leq \frac{1}{2n} \|y - \phi\theta_*\|_2^2 + \lambda \Omega(\theta_*)$$

(because $\hat{\theta}$ is a global minimizer)

$$\frac{1}{2n} \left[\|\phi(\theta_* - \hat{\theta})\|_2^2 + \|\varepsilon\|_2^2 + 2\varepsilon^\top \phi(\theta_* - \hat{\theta}) \right] + \lambda \Omega(\hat{\theta}) \leq \frac{1}{2n} \|\varepsilon\|_2^2 + \lambda \Omega(\theta_*)$$

$$\begin{aligned}
\|\Phi(\hat{\theta} - \theta_*)\|_2^2 &\leqslant 2\Omega^*(\Phi^\top \varepsilon)\Omega(\hat{\theta} - \theta_*) + 2n\lambda\Omega(\theta_*) - 2n\lambda\Omega(\hat{\theta}) \\
&\leqslant n\lambda\Omega(\hat{\theta} - \theta_*) + 2n\lambda\Omega(\theta_*) - 2n\lambda\Omega(\hat{\theta}) \\
&\leqslant n\lambda\Omega(\hat{\theta}) + n\lambda\Omega(\theta_*) + 2n\lambda\Omega(\theta_*) - 2n\lambda\Omega(\hat{\theta}) \leqslant 3n\lambda\Omega(\theta_*) - n\lambda\Omega(\hat{\theta}).
\end{aligned}$$

Lemma: $\Omega^*(u) = \sup_{\Omega(z) \leq 1} z^\top u$ ⟹ dual norm

Dual norm : $\Omega^*(u) = \sup_{\Omega(z) \le 1} z^T u$

$\Omega = \ell_2 = \Omega^* = \ell_2$

$\Omega = \ell_1 : \Omega^* = \ell_\infty$

$\Omega = \ell_\infty : \Omega^* = \ell_1$

$\Omega = \ell_p , \Omega^* = \ell_q$

$\frac{1}{p} + \frac{1}{q} = 1$

$\sup_{\|z\|_1 \le 1} z^T u = \sum_{j=1}^{d} z_j u_j$

$\le \sum_j |z_j| \max_j |u_j|$

$\le \max_j |u_j|$

$= \|u\|_\infty$

Property $= z^T u \le \Omega(z) \Omega^*(u)$

"Cauchy-Schwarz"

Holder-ln

Proof

$z^T u = \frac{z^T u}{\Omega(z)} \cdot \Omega(z) \le \Omega(z) \sup_{\Omega(z') \le 1} z'^T u = \Omega(z) \Omega^*(u)$

Proposition: if $\Omega^*(\Phi^\top \varepsilon) \leq \frac{n\lambda}{2}$ then $\frac{1}{n}\|\Phi(\hat\theta - \theta_*)\|_2^2 \leq 3\lambda\Omega(\theta_*)$

"performance"

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

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

(because $\hat\theta$ is a global minimizer)

$$\frac{1}{2n}\left[\|\Phi(\theta_* - \hat\theta)\|_2^2 + \|\varepsilon\|_2^2 + 2\varepsilon^\top\Phi(\theta_* - \hat\theta)\right] + \lambda\Omega(\hat\theta) \leq \frac{1}{2n}\|\varepsilon\|_2^2 + \lambda\Omega(\theta_*)$$

$$\begin{aligned}
\|\Phi(\hat\theta - \theta_*)\|_2^2 &\leqslant 2\Omega^*(\Phi^\top\varepsilon)\Omega(\hat\theta - \theta_*) + 2n\lambda\Omega(\theta_*) - 2n\lambda\Omega(\hat\theta) \\
&\leqslant n\lambda\Omega(\hat\theta - \theta_*) + 2n\lambda\Omega(\theta_*) - 2n\lambda\Omega(\hat\theta) \\
&\leqslant n\lambda\Omega(\hat\theta) + n\lambda\Omega(\theta_*) + 2n\lambda\Omega(\theta_*) - 2n\lambda\Omega(\hat\theta) \leqslant 3n\lambda\Omega(\theta_*) - n\lambda\Omega(\hat\theta).
\end{aligned}$$

Using $\varepsilon^\top\Phi(\hat\theta - \theta_*) \leq \Omega(\hat\theta - \theta_*)\Omega^*(\Phi^\top\varepsilon)$

Proposition: if $\Omega^*(\Phi^T\varepsilon) \le \frac{n\lambda}{2}$ then $\frac{1}{n}\|\Phi(\hat\theta - \theta_*)\|_2^2 \le 3\lambda\, s(\theta_*)$

"performance"

Consequences: ℓ_1-norm

Band with high probability

$$\mathbb{P}\left(\Omega^*(\Phi^T\varepsilon) \ge \frac{n\lambda}{2}\right) = \mathbb{P}\left(\|\Phi^T\varepsilon\|_\infty \ge \frac{n\lambda}{2}\right)$$

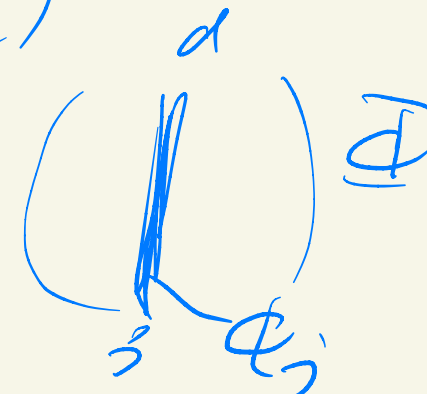
$$= \mathbb{P}\left(\sup_{j\in\{1,\dots,d\}} |(\Phi^T\varepsilon)_j| \ge \frac{n\lambda}{2}\right)$$

union bound →

$$\le \sum_{j=1}^{d} \mathbb{P}\left(|(\Phi^T\varepsilon)_j| \ge \frac{n\lambda}{2}\right)$$

If ε is Gaussian with mean zero and covariance matrix $\sigma^2 I$,

$(\Phi^T\varepsilon)_j = \Phi_j^T\varepsilon \in \mathbb{R}$ is Gaussian with mean 0 and variance $\Phi_j^T\Phi_j$



Assumption $= \frac{1}{n}\Phi_j^T\Phi_j = \frac{1}{n}\sum_{i=1}^{n} \varphi(x_i)_j^2 =$ bounded by 1 $\Rightarrow \alpha \le 1/n$

Lemma: If $Z \sim$ Gaussian with mean 0 and variance σ^2

$$P(|Z| \geq t) \leq 2e^{-t^2/2\sigma^2}$$

$$L = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{|Z| \geq t} e^{-z^2/2\sigma^2} dz = \frac{1}{\sqrt{2\pi}} \int_{|z'| \geq \frac{t}{\sigma}} e^{-z'^2/2} dz'$$

$$\leq 2e^{-t^2/2\sigma^2}$$

with $z' = \frac{z}{\sigma}$

$$P(Z \geq t) = P(e^{\lambda Z} \geq e^{\lambda t}) \leq \frac{E e^{\lambda Z}}{e^{\lambda t}}$$

— Hoeffding

$$E e^{\sigma Z} = e^{\frac{1}{2}\sigma^2 Z^2}$$

$$\mathbb{P}\left(|(\Phi^T\varepsilon)_j| \geq \frac{n\lambda}{2}\right) \leq 2\exp\left(-\frac{\left(\frac{n^2\lambda^2}{4}\right)}{2n}\right)$$

gaussian with variance n

$$= 2\exp\left(-\frac{n\lambda^2}{8}\right)$$

$$\mathbb{P}\left(\Omega^\infty(\Phi^T\varepsilon) \geq \frac{n\lambda}{2}\right) \leq 2d\exp\left(-\frac{n\lambda^2}{8}\right)$$

Consequence: for the lasso, with probability

$$\geq 1 - \frac{2d}{\delta}e^{-n\lambda^2/8}$$

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

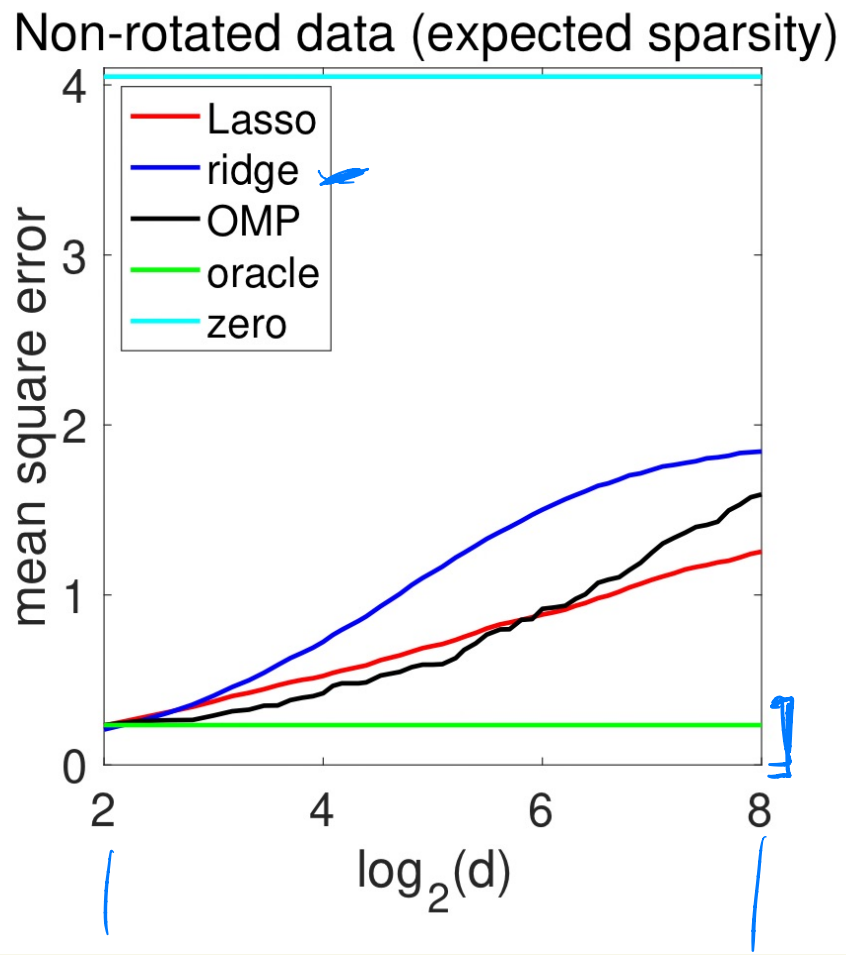
$$\leq 3\|\theta_*\|_1\sqrt{\frac{8}{n}\left(\log 2d + \log\frac{1}{\delta}\right)}$$

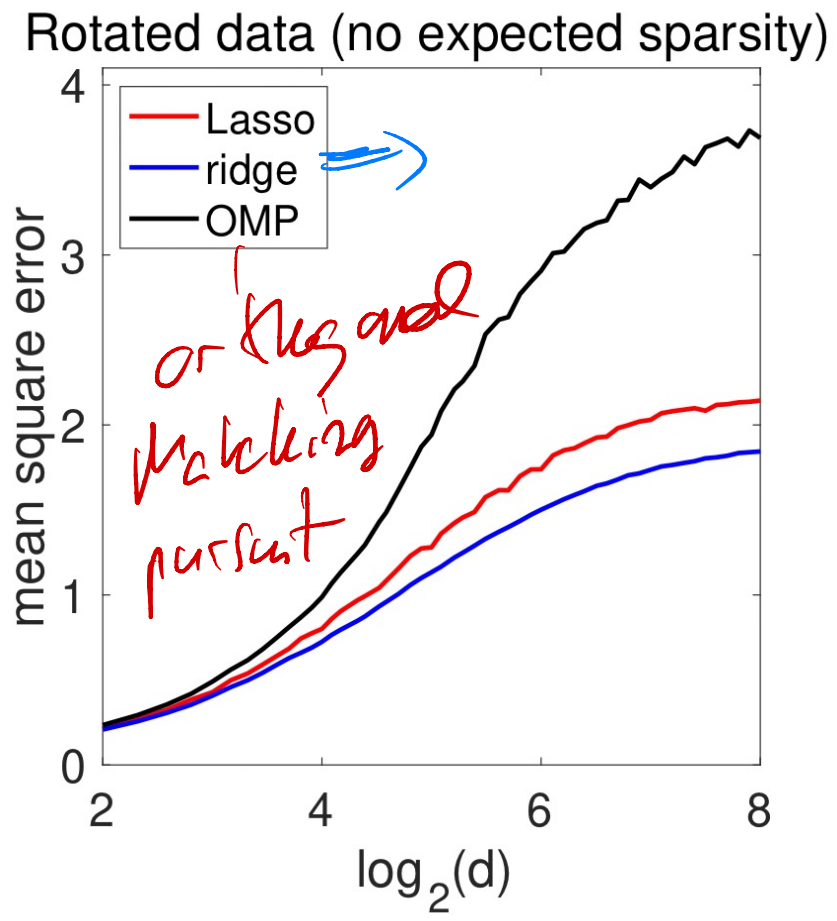
$$\sqrt{\log d/n}$$

$$\frac{n\lambda^2}{8} = \log\frac{2d}{\delta}$$

$$\lambda = \sqrt{\frac{8}{n}\log\frac{2d}{\delta}}$$

least-squares : $y = \Phi\theta_s + \varepsilon$ only non-zero components in the first d components





$d = 4$

$d = 128$

$\Phi \leftarrow \Phi R$ - rotation

$\Phi\theta = \Phi R R^T \theta$ $\quad \|\theta\|_2 = \|R^T\theta\|_2$