Last class: Wed 19/11 at 2pm

Sparse methods $\Rightarrow$ deal with large d

d = dimension of x

Two classes: • linear models

$\theta^\top \varphi(x)$ with $\theta \in \mathbb{R}^d$

with no regularization: $\frac{\sigma^2 d}{n}$ (OLS) $\longrightarrow$ Ridge regression $O(\frac{1}{\sqrt{n}})$

(squared) ℓ_2-norm regularization

• "non-linear" models (local averaging / kernel methods)

$\hookrightarrow O\left(\frac{1}{n^{2/d+2}}\right)$

• Today: variable selection for linear models

goal: if θ_* has only k non zero components we could achieve $\frac{\sigma^2 k}{n} \log d$.

Two types of penalty: ℓ_0 and ℓ_1

$\|\theta_*\|_0$ = # of non zero elements of $\theta_* \leq k$

⚠ only relevant when ...

Goal: $\min \mathbb{E}\, \ell(y, \theta^\top \varphi(x))$

$\hookrightarrow$ linear models + assume $y_i = \varphi(x_i)^\top \theta_* + \varepsilon_i$

+ fixed design.

Goal: estimate $\varphi(x_i)^\top \theta_*$

$\hookrightarrow$ square loss

Empirical risk

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

Excess risk of θ: $\frac{1}{n}\|\Phi(\theta - \theta_*)\|_2^2$

where $\Phi \in \mathbb{R}^{n \times d}$ design matrix

$\hat{\Sigma} = \frac{1}{n}\Phi^\top\Phi$ empirical covariance matrix

New proof technique for least-squares: $\hat{\Theta} = \arg\min_{\Theta \in \Theta} \hat{R}(\Theta) = \frac{1}{n}\|y - \Phi\Theta\|_2^2$

constrained

exact

set of allowed $\Theta \in \mathbb{R}^d$

assuming $\frac{\Phi^T\Phi}{n}$ has full rank

$\boxed{n \geq d}$

which model: $y = \Phi\Theta_* + \varepsilon$ $\boxed{\Theta_* \in \Theta}$

a criterion: $\frac{1}{n}\|\Phi(\Theta - \Theta_*)\|_2^2$

Start with $\|y - \Phi\hat{\Theta}\|_2^2 \leq \|y - \Phi\Theta_*\|_2^2$

$\|\varepsilon + \Phi(\Theta_* - \hat{\Theta})\|_2^2 \leq \|\varepsilon\|_2^2$ using model

$\cancel{\|\varepsilon\|_2^2} + \|\Phi(\hat{\Theta} - \Theta_*)\|_2^2 - 2\varepsilon^T\Phi(\hat{\Theta} - \Theta_*) \leq \cancel{\|\varepsilon\|_2^2}$ by expanding

$\Rightarrow \|\Phi(\hat{\Theta} - \Theta_*)\|_2^2 \leq 2\varepsilon^T\Phi(\hat{\Theta} - \Theta_*) = \|\Phi(\hat{\Theta} - \Theta_*)\|_2 \cdot 2\varepsilon^T \frac{\Phi(\hat{\Theta} - \Theta_*)}{\|\Phi(\hat{\Theta} - \Theta_*)\|_2}$

$\Rightarrow \|\Phi(\hat{\Theta} - \Theta_*)\|_2 \leq 2\varepsilon^T \frac{\Phi(\hat{\Theta} - \Theta_*)}{\|\Phi(\hat{\Theta} - \Theta_*)\|_2}$

when $\sum_{i=1}^{n} a_i \varepsilon_i \sim \sqrt{n}$

$a^T\varepsilon \sim \sqrt{n}$

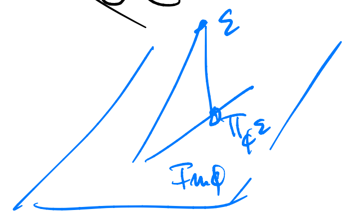
$$\boxed{\|\Phi(\hat{\Theta} - \Theta_*)\|_2 \leq \sup_{\Theta \in \Theta} 2\varepsilon^T \frac{\Phi(\Theta - \Theta_*)}{\|\Phi(\Theta - \Theta_*)\|_2}}$$

"Safety check": what if $\Theta = \mathbb{R}^d$?

$$\mathbb{E}\|\Phi(\hat{\Theta} - \Theta_*)\|_2^2 \leq \mathbb{E}\left(\sup_{\Theta \in \Theta} 2\varepsilon^T \boxed{\frac{\Phi(\Theta - \Theta_*)}{\|\Phi(\Theta - \Theta_*)\|_2}}\right)^2 \leq 4\sigma^2 d$$

$$\leq \mathbb{E} \sup_{\substack{z \in \operatorname{Im}\Phi \\ \|z\|_2 = 1}} (\varepsilon^T z)^2 = \mathbb{E}\|\Pi_\Phi \varepsilon\|^2 = \sigma^2 \operatorname{tr} \Pi_\Phi^2 = \sigma^2 d.$$

$z = \frac{\Pi_\Phi \varepsilon}{\|\Pi_\Phi \varepsilon\|}$



"$z \in \mathbb{R}^n$, $\|z\|_2 = 1$

$z \in \operatorname{Im}\Phi$ $\Phi \in \mathbb{R}^{n \times d}$

Variable selection by ℓ_0-penalty: $\Theta = \{\theta \in \mathbb{R}^d, \|\theta\|_0 \le k\}$

ℓ_0-penalty = # of non zeros $= |\text{Support}(\theta)|$

Assume $\theta_* \in \Theta$

$A = \text{support of } \theta_* = \{i, (\theta_*)_i \ne 0\}$

$|A| \le k$

$$\hat\theta = \operatorname*{argmin}_{\theta \in \Theta} \|y - \Phi\theta\|_2^2$$

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

$$\le 4 \sup_{\substack{B \subset \{1, \dots, d\} \\ |B| \le 2k}} \ \sup_{\text{supp}(\theta - \theta_*) \subset B} \left(\varepsilon^T \underbrace{\frac{\Phi(\theta - \theta_*)}{\|\Phi(\theta - \theta_*)\|_2}}_{z \in \text{Im}\,\Phi_B,\ \|z\|_2 = 1}\right)^2$$

if Φ_B is the submatrix of Φ with columns indexed by B

$$\le 4 \sup_{|B| \le 2k} \|\Pi_{\Phi_B} \varepsilon\|^2$$

$$\boxed{\|\Phi(\hat\theta - \theta_*)\|_2^2 \le 4 \sup_{|B| = 2k} \|\Pi_{\Phi_B} \varepsilon\|^2} \quad \le \boxed{\sigma^2 k \log \frac{d}{k}}$$

taking expectations: $\mathbb{E}\|\Pi_{\Phi_B}\varepsilon\|_2^2 = \mathbb{E}\,\text{tr}(\Pi_{\Phi_B} \varepsilon\varepsilon^T \Pi_{\Phi_B}) = \sigma^2 \text{tr}\, \Pi_{\Phi_B}^2 = \sigma^2 \text{tr}\, \Pi_{\Phi_B} \le \sigma^2 |B| \le 2\sigma^2 k$

Two "lemmas":

1. sup of norms of gaussian
2. counting # of B's

$$\le 16\sigma^2 \log\left[2^k \cdot \underbrace{\#B \text{ s.t. } (|B| = 2k)}_{\binom{d}{2k} \le}\right]$$

Lemma 8.1 If $z \in \mathbb{R}^n$ has gaussian distribution with mean 0 covariance matrix $\sigma^2 I$

if $s < \frac{1}{2\sigma^2}$ $\mathbb{E}[e^{s\|z\|_2^2}] = \frac{1}{(1-2\sigma^2 s)^{n/2}}$

Proof: $\mathbb{E}(e^{s\|z\|_2^2}) = \left(\mathbb{E}e^{s z_i^2}\right)^n$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\int e^{sz^2} e^{-z^2/2\sigma^2} dz\right)^n = \left(\sqrt{\frac{\tau^2}{\sigma^2}}\right)^n = \left(\frac{1}{1-2\sigma^2 s}\right)^{n/2}$$

$$e^{-\frac{z^2}{2}\left[\frac{1}{\sigma^2} - 2s\right]} = e^{-z^2/2\tau^2} \quad \text{with } \tau^2 = \frac{1}{\frac{1}{\sigma^2} - 2s}$$

Lemma 8.2: if $u_1, \dots, u_m$ are any random variables

$\forall s > 0$, $\mathbb{E}(\max(u_1, \dots, u_m)) \leq \frac{1}{s}\log\left(\sum_{i=1}^m \mathbb{E}[e^{su_i}]\right)$

proof:

$$\exp(s\,\mathbb{E}\max(u_1, \dots, u_m)) \underset{\uparrow\ \text{Jensen}}{\leq} \mathbb{E}\exp(s\max(u_1, \dots, u_m)) = \mathbb{E}\max[e^{su_1}, \dots, e^{su_m}]$$

$$\leq \mathbb{E}[e^{su_1} + \dots + e^{su_m}]$$

---

goal: $\mathbb{E}\sup_{|B|=2k} \underbrace{\|\Pi_{\Phi_B}\varepsilon\|_2^2}_{u_B} \underset{\uparrow\ \text{lemma 8.2}}{\leq} \frac{1}{s}\log\left(\sum_{|B|=2k} \mathbb{E}[e^{s\|\Pi_{\Phi_B}\varepsilon\|^2}]\right)$

$z = \Pi_{\Phi_B}\varepsilon$ is gaussian, has zero mean and covariance $\sigma^2 \|\Pi_{\Phi_B}\|^2 = \sigma^2 \Pi_{\Phi_B}$

$\|z\|^2$

lemma 8.1 $\leq \frac{1}{s}\log\left(\sum_{|B|=2k} \frac{1}{(1-2\sigma^2 s)^k}\right) = 4\sigma^2 \log(2^k \times \#\text{ of } B\text{'s})$ $\quad s = \frac{1}{4\sigma^2}$

goal: $\mathbb{E} \sup_{|B|=2k} \underbrace{\|\Pi_{\Phi_B}\varepsilon\|_2^2}_{u_B} \underset{\text{lemma 8.2}}{\leq} \frac{1}{s}\log\left(\sum_{|B|=2k} \mathbb{E}\left[e^{s\|\Pi_{\Phi_B}\varepsilon\|^2}\right]\right)$

$z = \Pi_{\Phi_B}\varepsilon$ is gaussian, los zero mean and covariance $\sigma^2 \Pi_{\Phi_B}^2 = \sigma^2 \Pi_{\Phi_B}$

$\|B\|^2$

lemma 8.1 $\leq \frac{1}{s}\log\left(\sum_{|B|\leq 2k} \frac{1}{(1-2\sigma^2 s)^k}\right) \overset{s=\frac{1}{4\sigma^2}}{=} 4\sigma^2 \log(2^k \times \#\text{ of } B\text{'s})$

$\mathbb{E} \sup_{|B|=2k} \|\Pi_{\Phi_B}\varepsilon\|_2^2 \leq 4\sigma^2\left[k\log 2 + \log\binom{d}{2k}\right]$

$\leq 2k\left(1 + \log\frac{d}{2k}\right)$

lemma 8.3

Excess risk $\leq 4\sigma^2\left(\cdot k + \cdot k\log\frac{d}{k}\right) + \frac{4}{n}$

$\sim \frac{\sigma^2 k}{n} \cdot + \frac{\sigma^2 k \log d/k}{n} \cdot$

price of adaptivity

10.33

10.53

From ℓ_0-constraint to ℓ_0-penalty : $\hat{\theta} = \arg\min \frac{1}{n} \|y - \Phi\theta\|_2^2 + \lambda \|\theta\|_0$

Prop 8.2 : $\lambda \propto \frac{\sigma^2}{n} \log d$ then

excess risk $\leq \frac{\sigma^2 k \log d}{n}$

Algorithms for $\min \frac{1}{n} \|y - \Phi\theta\|_2^2 + \lambda \|\theta\|_0$

or such that $\|\theta\|_0 \leq k$

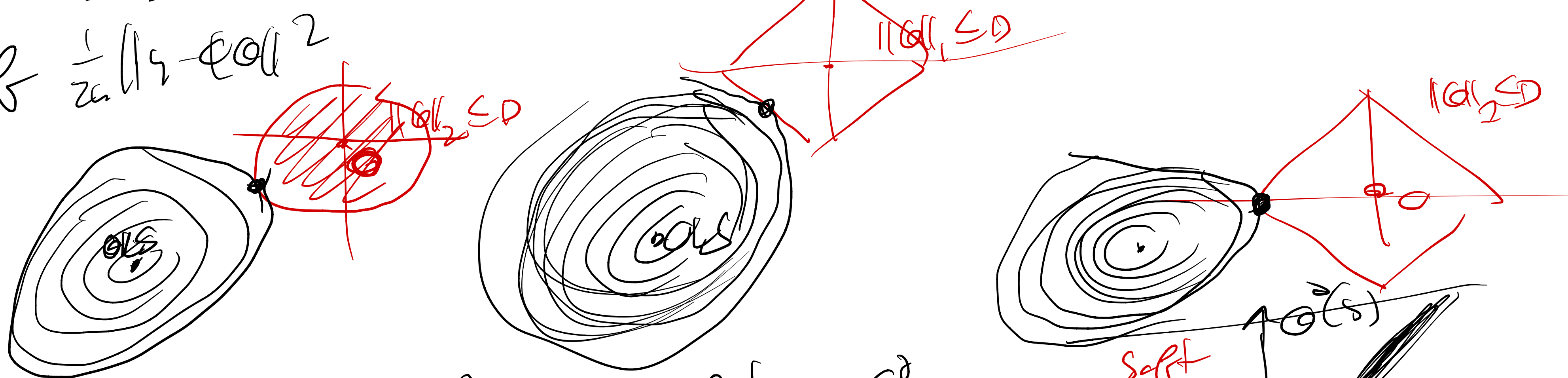
Classical : greedy $\Leftrightarrow$ orthogonal matching pursuit

ℓ_1-penalty : $\|\theta\|_1 = \sum_{j=1}^{d} |\theta_j|$

cost function : $\frac{1}{2n}\|y-\Phi\theta\|_2^2 + \lambda\|\theta\|_1 = F(\theta)$

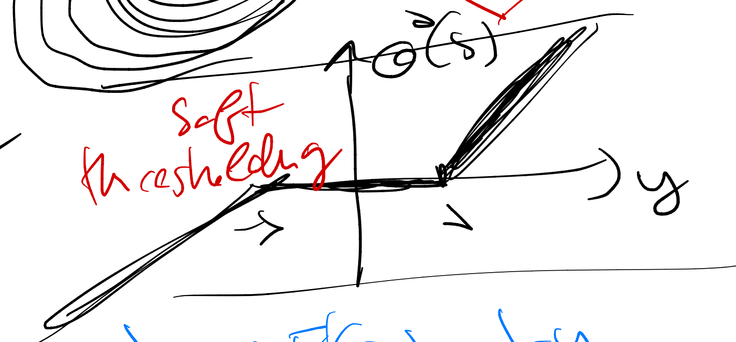
Why does ℓ_1-norm lead to zeros in θ

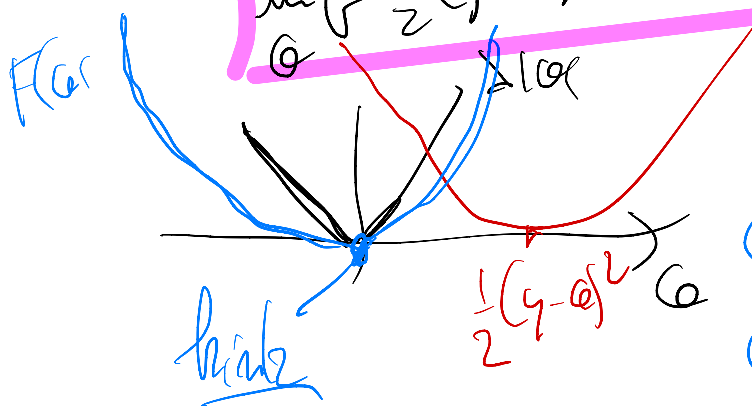
(1) $\min \frac{1}{2n}\|y-\Phi\theta\|^2$



(2) in 1D : $\min_\theta \frac{1}{2}(y-\theta)^2 \longrightarrow$ solution $\theta^* = y$

$$\boxed{\min_\theta \frac{1}{2}(y-\theta)^2 + \lambda|\theta|} = F(\theta)$$





if $\theta > 0$, $F'(\theta) = \theta - y + \lambda \Rightarrow F'(\theta_+) = \lambda - y$

if $\theta < 0$, $F'(\theta) = \theta - y - \lambda \rightarrow F'(\theta_-) = -\lambda - y$

(1) $\theta^* = 0 \Leftrightarrow F'(\theta_+) \geq 0$, $F'(\theta_-) \leq 0 \Leftrightarrow \lambda - y \geq 0$, $-\lambda - y \leq 0$ $\Leftrightarrow |y| \leq \lambda$

(2) $\theta^* > 0 \Leftrightarrow F'(\theta_+) < 0 \Leftrightarrow y > \lambda$ and then $\theta_y = y - \lambda$

Iterative soft thresholding $\in$ proximal method

$\min F(\theta) + \lambda \|\theta\|_1$ F is convex and L-smooth

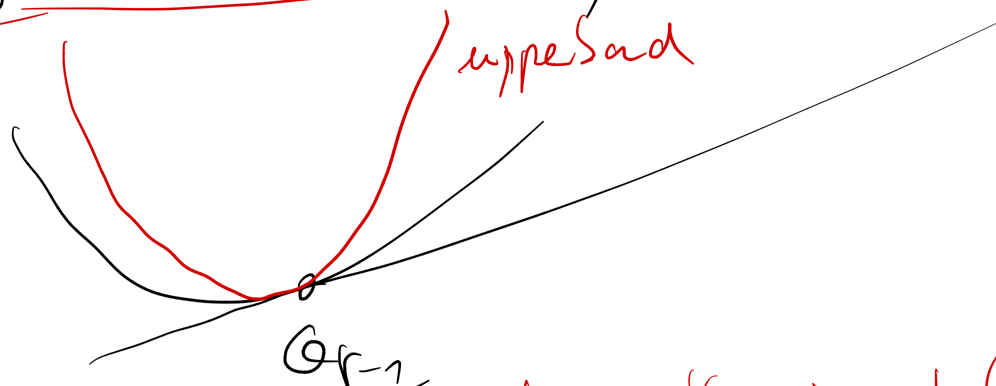
review: to minimize F, GD with $\gamma = \frac{1}{L}$

$$\theta_t = \theta_{t-1} - \frac{1}{L} F'(\theta_{t-1})$$ was converging at speed $\frac{1}{t}$ or $\exp\left(-\frac{t}{\kappa}\right)$

condition number of F

new interpretation

$$F(\theta) \le \underbrace{F(\theta_{t-1}) + F'(\theta_{t-1})^T(\theta - \theta_{t-1}) + \frac{L}{2}\|\theta - \theta_{t-1}\|^2}_{\text{upper bound}}$$



$\underset{\theta}{\arg\min}$ $\Rightarrow$ gradient: $F'(\theta_{t-1}) + L(\theta - \theta_{t-1}) = 0$

$\Leftrightarrow \theta = \theta_{t-1} - \frac{1}{L} F'(\theta_{t-1})$

$$F(\theta) + \lambda\|\theta\|_1 \le \underbrace{F(\theta_{t-1}) + F'(\theta_{t-1})^T(\theta - \theta_{t-1}) + \frac{L}{2}\|\theta - \theta_{t-1}\|^2 + \lambda\|\theta\|_1}$$

$\theta_t = \underset{\theta}{\arg\min}$ $\Rightarrow$ (1) same CV rate as $\lambda = 0$

$\Rightarrow$ (2) soft thresholding as fcn of GD

Analysis: ① Random design, constrained, Lipschitz loss
② Fixed design, penalized, square loss

---

① $\hat{R}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \varphi(x_i)^T\theta)$. $\quad \hat\theta \in \arg\min_{\|\theta\|_1 \le D} \hat{R}(\theta)$ with $\|\theta_*\|_1 \le D$

$\mathbb{E}(R(\hat\theta)) - R(\theta_*) \le 2$ uniform deviations

$\le 2\cdot 2\cdot$ Rademacher complexity of $\ell(y, \theta^T\varphi(x))$

$\le 2\cdot 2\cdot G\cdot$ Rademacher complexity of $\theta^T\varphi(x)$

↳ Lipschitz constant of loss ℓ

$$\le 4G\,\mathbb{E}\sup_{\|\theta\|_1\le D} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i\, \varphi(x_i)^T\theta$$

↳ Rademacher variable $\in \{-1,1\}$.

$$\le 4G\,\mathbb{E}\left(D \left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\varphi(x_i)\right\|_\infty\right) = 4GD\,\mathbb{E}\sup(\pm u_j)$$

$$u_j = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\varphi(x_i)_j$$

$$\boxed{\frac{4GD\|\varphi\|\sqrt{\log d}}{\sqrt{n}}}$$

slow rate / $D = \|\theta_*\|_2 \sim k$

<u>exercise for this class</u>

Finish the proof using Lemma 8.2 and 8.3.2

$\frac{2\sqrt{\log d}}{n}$ ($\frac{2\log d}{n}$

② Fixed design + square loss

Section 8.3.3 : $\hat{\theta} = \arg\min \underbrace{\frac{1}{2n}\|y - \Phi\theta\|_2^2 + \lambda\|\theta\|_1}_{\text{Lasso}}$

if $\lambda \propto \frac{\sigma}{\sqrt{n}}\sqrt{\log d}$

excess risk $\sim \|\theta_*\|_1 \frac{\sigma}{\sqrt{n}}\sqrt{\log d}$ — high-dim phenomena

$\sim k$

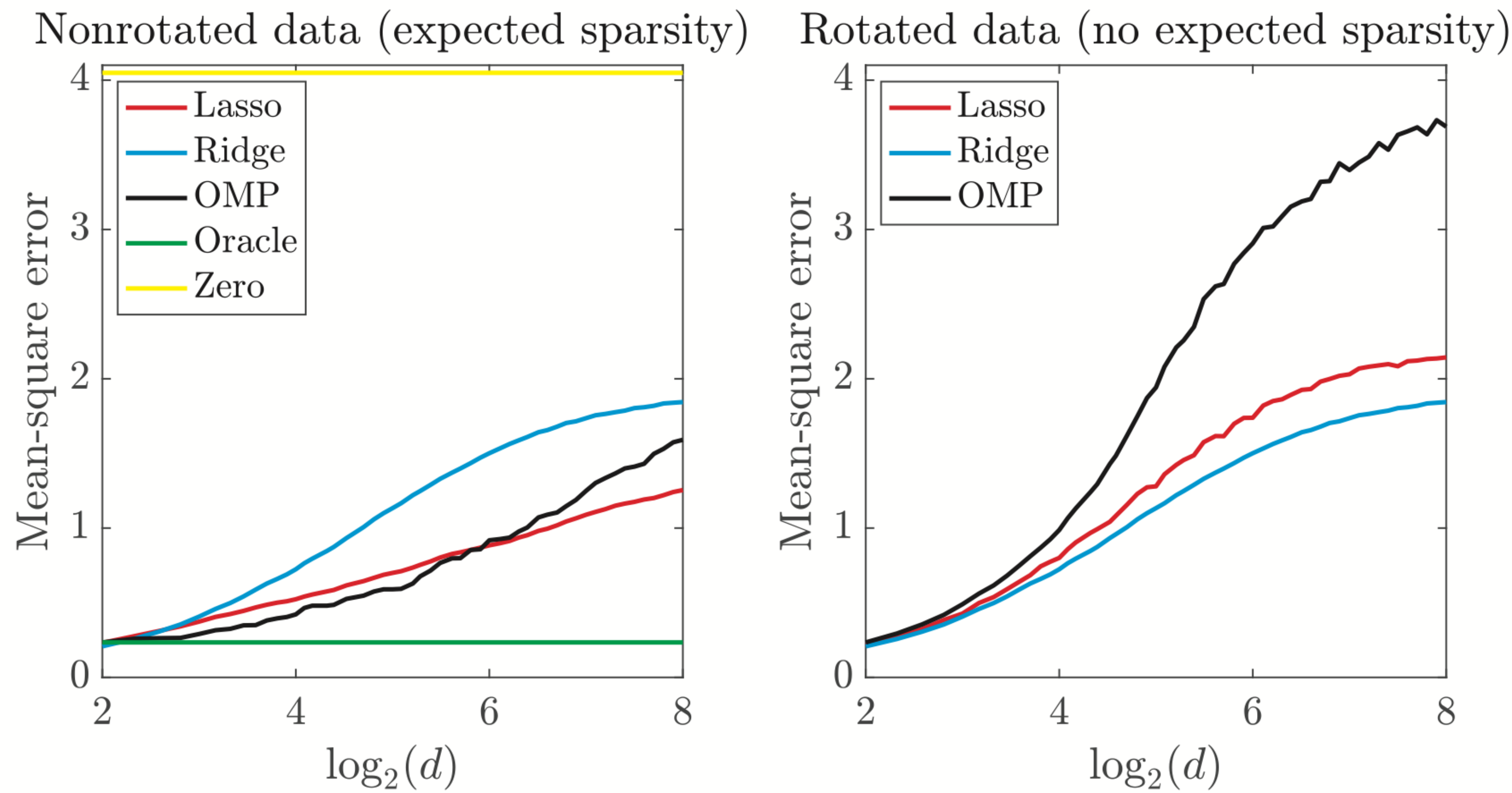
slow rate

Section 8.3.4 : fast rates of order $\left| \frac{\sigma^2 k}{n} \log d \times \text{cte} \right|$

if $\frac{\Phi^T\Phi}{n}$ close to being diagonal

$y_i = x_i^T \theta_* + \varepsilon$

θ_* is supported on 4 coordinates and x_i are all $\mathcal{N}(0,1)$



$x_i \to R x_i$

↳ rotation

$$\frac{\sigma^2 d}{n} \Rightarrow \sigma^2 \frac{\mathrm{rank}(\Phi)}{n}$$