

03/10

Least-squares regression

Training data: $(x_i, y_i) \in \mathcal{X} \times \mathbb{R}, i=1, \dots, n$

$$\ell(y, f(x)) = (y - f(x))^2$$

Expected risk: $R(f) = \underline{\mathbb{E}}[\ell(y, f(x))]$

$\hookrightarrow$ testing data
 $= \arg \min_z \mathbb{E}(\ell(y, z) | x)$

Bayes predictor:

$$f^*(x) = \mathbb{E}[y | x]$$

Empirical risk: $\hat{R}(f) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, f(x_i))$

$$\underline{\text{Model}} = f_{\theta}(x) = \varphi(x)^T \theta$$

linear model

! in the parameter θ

$\varphi: \mathcal{X} \rightarrow \mathbb{R}^d$
feature map

Ordinary least-squares (OLS)

$$\hat{R}(\theta) = \hat{R}(\theta) = \frac{1}{n} \sum_{i=1}^n (y_i - \underbrace{\theta^T \phi(x_i)}_{\in \mathbb{R}^d})^2$$

$\underbrace{\hspace{1.5cm}}_{\text{new notation}} \quad \underbrace{\hspace{1.5cm}}_{\in \mathbb{R}} \quad \underbrace{\hspace{1.5cm}}_{\in \mathbb{R}^d} \quad \underbrace{\hspace{1.5cm}}_{\in \mathbb{R}^d}$

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

$y \in \mathbb{R}^n$ label response output
 vector

$$= \frac{1}{n} \sum_i (y_i - \underbrace{(\Phi \theta)_i}_{\phi(x_i)^T \theta})^2$$

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

$$\begin{pmatrix} \phi(x_1)^T \\ \vdots \\ \phi(x_n)^T \end{pmatrix}$$

OLS estimator: $\hat{\theta} \in \arg \min \hat{R}(\theta)$

Computation : $\hat{\Theta} \in \arg \min_{\Theta \in \mathbb{R}^d} \frac{1}{n} \|y - \Phi \Theta\|_2^2 = \hat{R}(\Theta)$

$$\hat{R}'(\Theta) \in \mathbb{R}^d$$

$$\hat{R}(\Theta) = \frac{1}{n} [y^T y - 2\Theta^T \Phi^T y + \Theta^T \Phi \Theta]$$

$$\hat{R}'(\Theta) = \frac{-2\Phi^T y + 2\Phi^T \Phi \Theta}{n}$$

$$\frac{\Phi^T y}{n}$$

Assumption: $\frac{1}{n} \Phi^T \Phi \in \mathbb{R}^{d \times d}$ full rank $\Leftrightarrow \text{rank}(\Phi) = d$

empirical covariance
second moment matrix

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n \varphi(x_i) \varphi(x_i)^T \in \mathbb{R}^{d \times d}$$

$$\text{rank}(\hat{\Sigma}) \leq n$$

! $d \leq n$

$$\Phi = \begin{pmatrix} \varphi(x_1)^T \\ \vdots \\ \varphi(x_n)^T \end{pmatrix}$$

$$\hat{R}(\theta) = 0 \quad (=) \quad \frac{\Phi^T \Phi}{n} \theta = \frac{\Phi^T y}{n} \quad \text{normal equations}$$

$$\hat{\Sigma} = \frac{1}{n} \Phi^T \Phi$$

$$\hat{\theta} = \left(\frac{\Phi^T \Phi}{n} \right)^{-1} \frac{\Phi^T y}{n} \quad \begin{array}{l} \text{OLS} \\ \text{estimator} \end{array}$$

$$\hat{\theta} = (\Phi^T \Phi)^{-1} \Phi^T y$$

Algorithms :

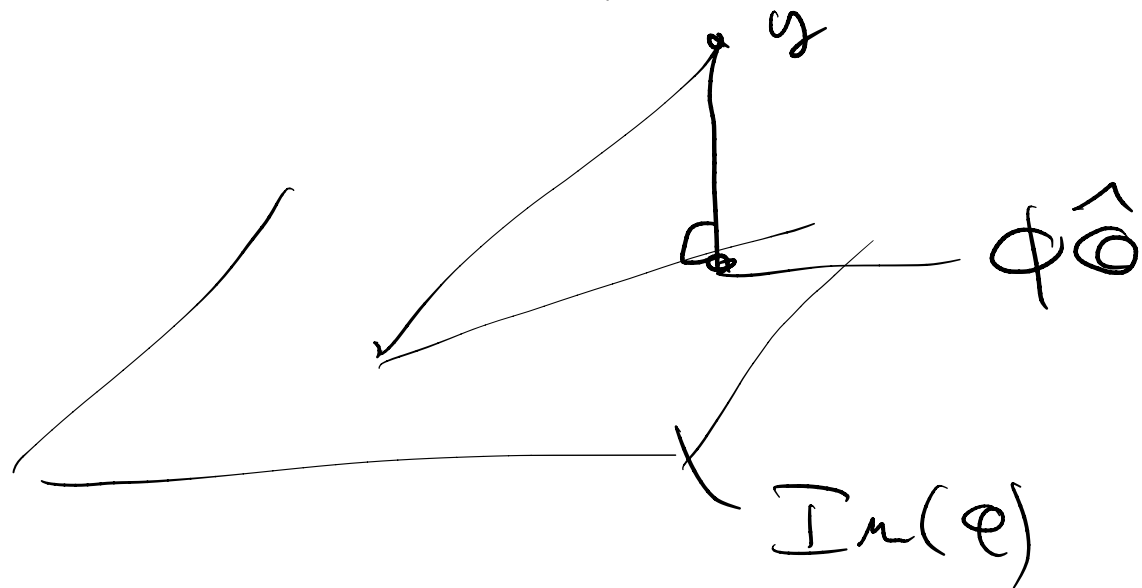
- ① solving linear systems $O(d^3) + O(d^2 n)$
QR factorization
- ② gradient descent (stochastic)
 $O(dn)$

geometric interpretation

$$\hat{\theta} = \arg \min_{\theta \in \mathbb{R}^d} \|y - \phi\theta\|_2^2$$

$$y \in \mathbb{R}^n, \quad \text{Im}(\phi) = \{ \phi\theta, \theta \in \mathbb{R}^d \}$$
$$= \left\{ \sum_i \phi(i; x) \theta_i, \theta \in \mathbb{R}^d \right\}$$

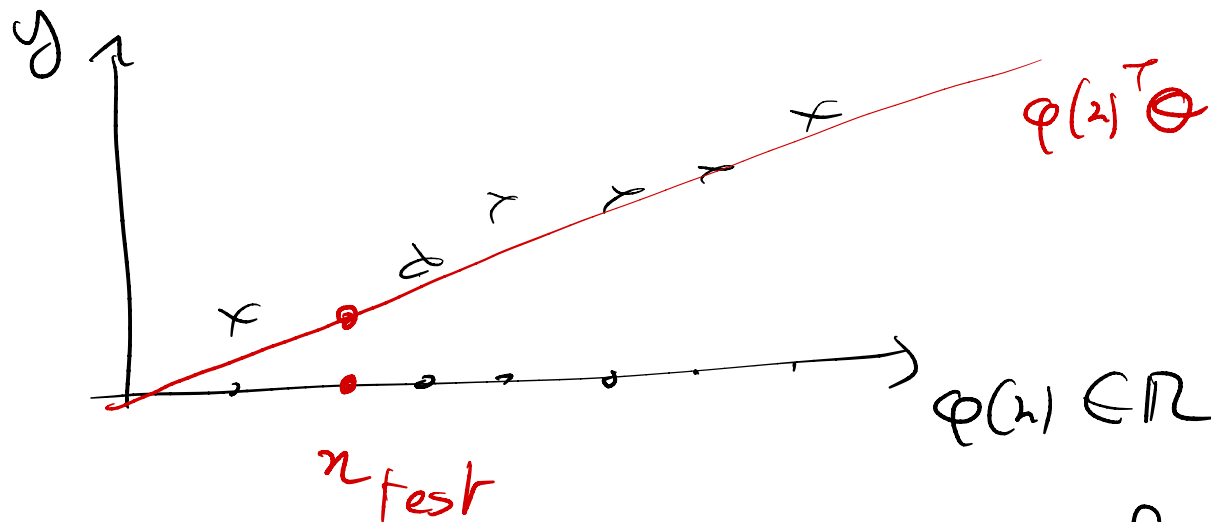
$$n \times \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}^d$$



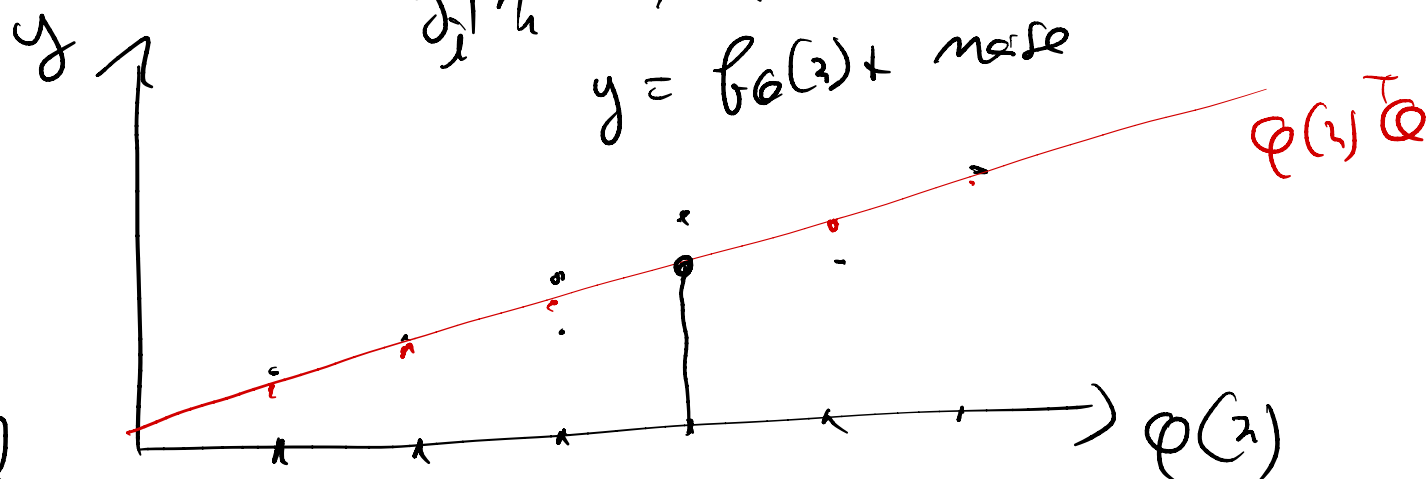
$$\hat{\theta} = (\phi^T \phi)^{-1} \phi^T y \in \mathbb{R}^d$$

Analysis = random design / fixed design

Random design = "classical ML" (x's) are both random



Fixed design = $x_1, \dots, x_n$ assumed fixed deterministic
 y_i / x_i random
 $y = f_0(x) + \text{noise}$



$\Rightarrow$ distribution
 y^n
 $=$ empirical

in-sample prediction error

Assumptions

$$y_i = \phi(x_i)^T \Theta_* + \varepsilon_i$$

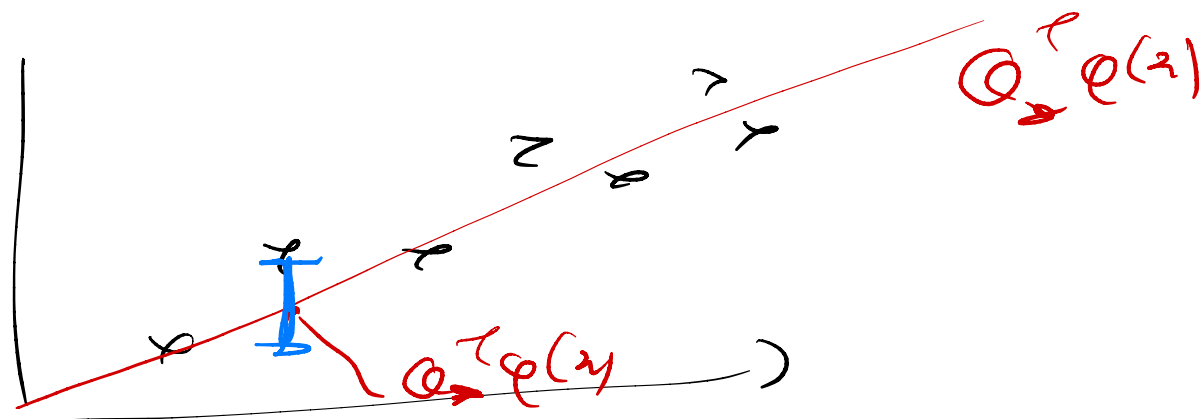
note (indep.)

$$\mathbb{E} \varepsilon_i = 0$$

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

well-specified linear model

$$y = \Phi \Theta_* + \varepsilon$$



$$\begin{aligned} R(\Theta) &= \frac{1}{n} \mathbb{E}_{\varepsilon} \|y - \Phi \Theta\|_2^2 = \frac{1}{n} \mathbb{E}_{\varepsilon} \|\varepsilon + \Phi(\Theta_* - \Theta)\|_2^2 \\ &= \underbrace{\frac{1}{n} \mathbb{E}_{\varepsilon} \|\varepsilon\|_2^2}_{\sigma^2} + \underbrace{\frac{1}{n} \mathbb{E}_{\varepsilon} (\|\Phi(\Theta_* - \Theta)\|_2^2)}_{\text{bias}} + \underbrace{\frac{2}{n} \mathbb{E}_{\varepsilon} \left(\varepsilon^T \Phi(\Theta_* - \Theta) \right)}_{0} \end{aligned}$$

predictor

test ε

$$\begin{aligned} \mathbb{E} \|\varepsilon\|^2 &= \mathbb{E} \text{tr} \varepsilon \varepsilon^T \\ &= \mathbb{E} \text{tr} \varepsilon \varepsilon^T \\ &= \text{tr} \mathbb{E} \varepsilon \varepsilon^T \\ &= \text{tr} \sigma^2 \mathbf{I} = n \sigma^2 \end{aligned}$$

$$\begin{aligned} &= \sigma^2 + \frac{1}{n} \|\Phi(\Theta_* - \Theta)\|_2^2 \\ &= \sigma^2 + (\Theta_* - \Theta)^T \frac{\Phi^T \Phi}{n} (\Theta_* - \Theta) \end{aligned}$$

$$\mathbb{E}_{\varepsilon} \|\varepsilon + z\|^2 = \underbrace{\|\mathbb{E}\varepsilon + z\|^2}_{\|\phi(\theta - \theta_*)\|^2} + \underbrace{\mathbb{E}\varepsilon^2}_{\sigma^2}$$

"Test" (in-sample) exercise

$$R(\theta) - R_* = (\theta - \theta_*)^T \hat{\Sigma} (\theta - \theta_*)$$

Goal = compute $\mathbb{E} R(\hat{\theta})$ for $\hat{\theta} = \text{OLS}$

$$\begin{aligned} \hat{\theta} &= (\phi^T \phi)^{-1} \phi^T y = (\phi^T \phi)^{-1} \phi^T (\underbrace{\phi \theta_* + \varepsilon}_{\text{training model}}) \\ &= \theta_* + (\phi^T \phi)^{-1} \phi^T \varepsilon \end{aligned}$$

$$\mathbb{E} \varepsilon \varepsilon^T = \sigma^2 I$$

Consequence =

Bias: $\mathbb{E} \hat{\theta} - \theta_* = 0$

unbiased estimator

Covariance: $\mathbb{E} (\hat{\theta} - \theta_*) (\hat{\theta} - \theta_*)^T = \mathbb{E} \left[(\phi^T \phi)^{-1} \phi^T \varepsilon \varepsilon^T \phi (\phi^T \phi)^{-1} \right]$

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

$$R(\hat{\theta}) - R_{\star} = (\hat{\theta} - \theta_{\star})^T \hat{\Sigma} (\hat{\theta} - \theta_{\star})$$

$$= \text{tr} \hat{\Sigma} (\hat{\theta} - \theta_{\star}) (\hat{\theta} - \theta_{\star})^T$$

$$E(\hat{\theta} - \theta_{\star}) = 0$$

$$E(\hat{\theta} - \theta_{\star}) (\hat{\theta} - \theta_{\star})^T = \frac{\sigma^2}{n} \left(\hat{\Sigma} \right)^{-1}$$

test excess risk
estimator

$$E[R(\hat{\theta}) - R_{\star}] = (E\hat{\theta} - \theta_{\star})^T \hat{\Sigma} (E\hat{\theta} - \theta_{\star})$$

$$+ \text{tr} \hat{\Sigma} E(\hat{\theta} - \theta_{\star}) (\hat{\theta} - \theta_{\star})^T$$

square of bias
variance

always true, for all $\hat{\theta}$

$$= 0 + \text{tr} \hat{\Sigma} \frac{\sigma^2}{n} \hat{\Sigma}^{-1}$$

$$= \text{tr} \left(\hat{\Sigma} \hat{\Sigma}^{-1} \right) \cdot \frac{\sigma^2}{n} = \text{tr} I \cdot \frac{\sigma^2}{n}$$

$$= \frac{\sigma^2 d}{n}$$

$\leq$

excess risk for OLS

10^25
10^45

$$E(a-z)^2 = (Ea-z)^2 + \text{var}(z)$$

$$E\|a-z\|_2^2 = \|Ea-z\|^2 + \text{tr Cov}(a)$$

$$E(a-z)^T M (a-z) = (Ea-z)^T M (Ea-z) + \text{tr } M \text{Cov}(a)$$

Model : $y_i = \phi(x_i)^T \theta + \varepsilon_i$

$$\mathbb{E} \varepsilon_i = 0$$

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

Estimator : OLS . $\hat{\theta} = (\Phi^T \Phi)^{-1} \Phi^T y$

normal
equations

Expected test risk: $\mathbb{E}_{\text{training data}} R(\hat{\theta}) - R_{\star} = \frac{\sigma^2 d}{n}$

with $\Phi^T \Phi$ invertible ($n \geq d$)

Regularization

$$\hat{R}(\theta) + \lambda \|\theta\|_2^2 = \frac{1}{n} \|y - \Phi \theta\|_2^2 + \lambda \|\theta\|_2^2$$

($+\lambda \|\theta\|_2^2$) chapter 8

RIDGE
REGRESSION

$$\hat{R}(\theta) + \frac{\lambda}{2} \|\theta\|_2^2 = \frac{1}{n} \|y - \Phi\theta\|_2^2 + \lambda \|\theta\|_2^2$$

($+ \lambda \|\theta\|_2^2$) chapter 8

RIDGE
REGRESSION

Computation of $\hat{\theta}_\lambda = \arg \min_{\theta} \hat{R}(\theta) + \frac{\lambda}{2} \|\theta\|_2^2$ $R_\lambda(\theta)$

$$\text{gradient } R'_\lambda(\theta) = \frac{2}{n} \Phi^T (\Phi\theta - y) + 2\lambda\theta = 0$$

$$\left(\frac{\Phi^T \Phi}{n} + \lambda I \right) \theta = \frac{\Phi^T y}{n}$$

$\frac{\Phi^T \Phi}{n}$ positive semi-definite (eigenvalues ≥ 0)
("PSD")

$\forall \lambda > 0, \quad \frac{\Phi^T \Phi}{n} + \lambda I$ invertible

$$\Rightarrow \text{Estimator: } \hat{\theta}_\lambda = \left(\hat{\Sigma} + \lambda I \right)^{-1} \frac{\Phi^T y}{n}$$

$$\hat{\Theta}_\lambda = (\hat{\Sigma} + \lambda I)^{-1} \frac{\Phi^T y}{n} = (\hat{\Sigma} + \lambda I)^{-1} \frac{\Phi^T}{n} [\underbrace{\Phi \Theta_0}_{\text{model}} + \varepsilon]$$

Test excess risk: $R(\Theta) - R_\star = (\Theta - \Theta_\star)^T \hat{\Sigma} (\Theta - \Theta_\star)$

For any estimator $\hat{\Theta} = \mathbb{E}[R(\hat{\Theta}) - R_\star] =$

(1) squared bias: $(\mathbb{E}(\hat{\Theta}) - \Theta_\star)^T \hat{\Sigma} (\mathbb{E}(\hat{\Theta}) - \Theta_\star)$

(2) variance: $\text{tr} \hat{\Sigma} \left[\mathbb{E}(\hat{\Theta} - \mathbb{E}(\hat{\Theta}))(\hat{\Theta} - \mathbb{E}(\hat{\Theta}))^T \right]$

(1) Bias: $\mathbb{E} \hat{\Theta}_\lambda = (\hat{\Sigma} + \lambda I)^{-1} \hat{\Sigma} \Theta_\star = (\hat{\Sigma} + \lambda I)^{-1} (\underbrace{\hat{\Sigma} + \lambda I}_{\text{red}} - \underbrace{\lambda I}_{\text{red}}) \Theta_\star$

$$= \Theta_\star - \underbrace{\lambda (\hat{\Sigma} + \lambda I)^{-1} \Theta_\star}_{\text{red}}$$

$\Rightarrow$ squared bias of $\hat{\Theta}_\lambda = \underbrace{(\lambda (\hat{\Sigma} + \lambda I)^{-1} \Theta_\star)^T}_{\text{red}} \hat{\Sigma} \underbrace{(\lambda (\hat{\Sigma} + \lambda I)^{-1} \Theta_\star)}_{\text{red}}$

$$= \lambda^2 \Theta_\star^T \hat{\Sigma} (\hat{\Sigma} + \lambda I)^{-2} \Theta_\star$$

(2) covariance matrix of $\hat{\Theta}_\lambda = \mathbb{E}(\hat{\Theta}_\lambda - \mathbb{E}(\hat{\Theta}_\lambda))(\hat{\Theta}_\lambda - \mathbb{E}(\hat{\Theta}_\lambda))^T$

(2) covariance matrix of $\hat{\theta}_\lambda = E(\hat{\theta}_\lambda - E(\hat{\theta}_\lambda) | (\hat{\theta}_\lambda - E(\hat{\theta}_\lambda))^T$
 with $\hat{\theta}_\lambda = (\hat{\Sigma} + \lambda I)^{-1} \frac{\Phi^T}{n} [\Phi \theta_0 + \varepsilon]$

Lemma: $y = M\varepsilon + c \Rightarrow$ covariance matrix of y is $M(E\varepsilon\varepsilon^T)M^T$
 with $E\varepsilon = c$

Proof: $Ey = M \cdot E\varepsilon + c = c$
 cov matrix = $E(y - c)(y - c)^T = E(M\varepsilon(M\varepsilon)^T)$
 $= E(M\varepsilon\varepsilon^T M^T)$

cov-mat of $\hat{\theta}_\lambda$
 with $M = (\hat{\Sigma} + \lambda I)^{-1} \frac{\Phi^T}{n}$
 $E\varepsilon\varepsilon^T = \sigma^2 I$

$$= (\hat{\Sigma} + \lambda I)^{-1} \frac{\Phi^T}{n} \sigma^2 I \frac{\Phi}{n} (\hat{\Sigma} + \lambda I)^{-1}$$

$$= \frac{\sigma^2}{n} (\hat{\Sigma} + \lambda I)^{-1} \hat{\Sigma} (\hat{\Sigma} + \lambda I)^{-1}$$

$$= \frac{\sigma^2}{n} \hat{\Sigma} (\hat{\Sigma} + \lambda I)^{-2}$$

OLS = $\frac{\sigma^2}{n} \hat{\Sigma}^{-1}$

Combinatorics

Bias

squared bias:

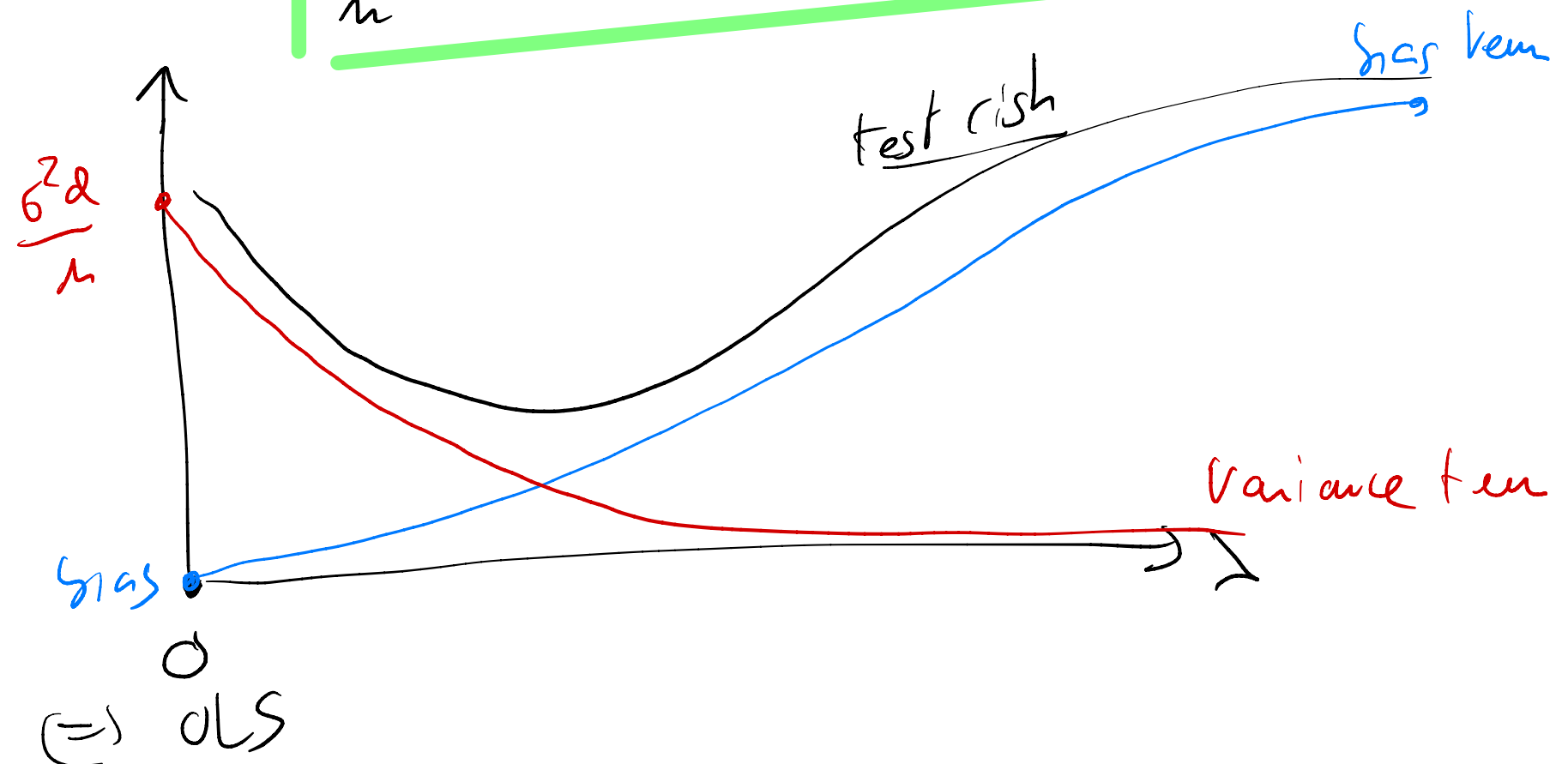
$$\lambda^2 Q_0^T \hat{\Sigma} (\hat{\Sigma} + \lambda I)^{-2} Q_0$$

variance = $\text{tr} \hat{\Sigma} \cdot (\text{cov matrix of } \hat{Q}_1)$

$$= \text{tr} \hat{\Sigma} \frac{\sigma^2}{n} \hat{\Sigma} (\hat{\Sigma} + \lambda I)^{-2}$$

variance

$$= \frac{\sigma^2}{n} \cdot \text{tr} \hat{\Sigma}^2 (\hat{\Sigma} + \lambda I)^{-2}$$



variance term 1)

$$\frac{\sigma^2}{n} \text{tr } \hat{\Sigma}^2 (\bar{E} + \lambda I)^{-2}$$

$$= \frac{\sigma^2}{n} \sum_{j=1}^d \lambda_j^2 (\lambda_j + \lambda)^{-2} = \frac{\sigma^2}{n} \sum_{j=1}^d \left(\frac{\lambda_j}{\lambda_j + \lambda} \right)^2$$

when $\lambda \rightarrow \infty = \frac{\sigma^2}{n}$. # of non zero eigenvalues

$\hat{\Sigma}$ has eigenvalues
 $\lambda_1, \dots, \lambda_d$

Simple upper bound:

① squared bias: $\lambda^2 \mathbf{Q}_\sigma^T \hat{\Sigma} (\hat{\Sigma} + \lambda \mathbf{I})^{-2} \mathbf{Q}_\sigma$

$$= \lambda \mathbf{Q}_\sigma^T \left(\underbrace{\hat{\Sigma} (\hat{\Sigma} + \lambda \mathbf{I})^{-1}}_{\text{eigenvalues} \in (c, 1)} \cdot \underbrace{\lambda (\hat{\Sigma} + \lambda \mathbf{I})^{-1}}_{\text{eigenvalues} \in (c, 1)} \right) \mathbf{Q}_\sigma$$

squared $\leq \underbrace{\lambda \|\mathbf{Q}_\sigma\|_2^2}_{\text{eigenvalues} \in [c, 1]}$

② variance

$$\frac{\sigma^2}{n} \text{tr} \hat{\Sigma}^2 (\hat{\Sigma} + \lambda \mathbf{I})^{-2}$$

$$\frac{\sigma^2}{n\lambda} \text{tr} \left(\hat{\Sigma} \underbrace{\hat{\Sigma} (\hat{\Sigma} + \lambda \mathbf{I})^{-1} \lambda (\hat{\Sigma} + \lambda \mathbf{I})^{-1}}_{\text{eigenvalues} \in (c, 1)} \right)$$

$$\leq \underbrace{\frac{\sigma^2}{n\lambda} \text{tr} \hat{\Sigma}}_{\text{eigenvalues} \in (c, 1)}$$

overall we showed:

$$E(R(\hat{\theta}_2) - R_0) \leq \underbrace{\lambda \|\theta_0\|^2}_{\text{squared bias}} + \underbrace{\frac{\sigma^2}{n\lambda} \text{tr} \hat{\Sigma}}_{\text{variance}}$$

squared
bias

variance

minimized with respect to λ at

$$\lambda^* = \frac{\sigma}{\sqrt{n}} \frac{\sqrt{\text{tr} \hat{\Sigma}}}{\|\theta_0\|_2}$$

$$\frac{\|\theta_0\| \sigma \sqrt{\text{tr} \hat{\Sigma}}}{\sqrt{n}} \text{ with value}$$

λ^* : dimension independent

λ^* : $1/\sqrt{n}$ instead of $1/n$
 $\hookrightarrow \frac{\sigma^2}{n}$

$$\lambda = \frac{\sigma}{\sqrt{n}} \frac{\sqrt{\text{tr} \Sigma}}{\|a\|_2} \sim \frac{hg \sqrt{hg^2 m^{-2}}}{m} \sim hg^2 m^{-2}$$

$$\text{risk} = \frac{2 \|a\| \sigma \sqrt{\text{tr} \Sigma}}{\sqrt{n}} \sim m hg hg m^{-1} \sim hg^2$$

dimensionality analysis

$$y = \underset{hg}{\underbrace{\varphi(x)^T}} \underset{hg m^{-1}}{\underbrace{a}} \underset{m}{\underbrace{1}} + \underset{hg \sim \sigma}{\underbrace{\varepsilon}}$$

$$\text{risk} = \|y - \hat{y}\|^2 \sim hg^2$$

$$\text{regularizer: } \lambda \|a\|^2 \sim hg^2 \\ \Rightarrow \lambda \sim hg^2 m^{-2}$$

Random design: test risk $y = \varphi(x)^T \theta_* + \varepsilon$

$$R(\theta) = \mathbb{E} (y - \varphi(x)^T \theta)^2$$

$$= \underbrace{\sigma_y^2}_{\sigma^2} + (\theta - \theta_*)^T \Sigma (\theta - \theta_*)$$

where $\Sigma = \mathbb{E} [\varphi(x) \varphi(x)^T]$

expected covariance matrix

$$\mathbb{E} R(\hat{\theta}_{OLS}) = \sigma^2$$
$$= \frac{\sigma^2}{n} \text{tr}(\hat{\Sigma}^{-1} \Sigma)$$

if input data are Gaussian, $\frac{\sigma^2}{n} \frac{d}{1 - \frac{d}{n}}$

true rates

$$\frac{\sigma^2 d}{n}$$

no regularization

homework

$$\frac{\sigma^2}{\sqrt{n}}$$

dimension
independent

regularization

penalty
 $\geq \| \theta \|_2^2$
 $\geq \| \theta \|_1$

SGD