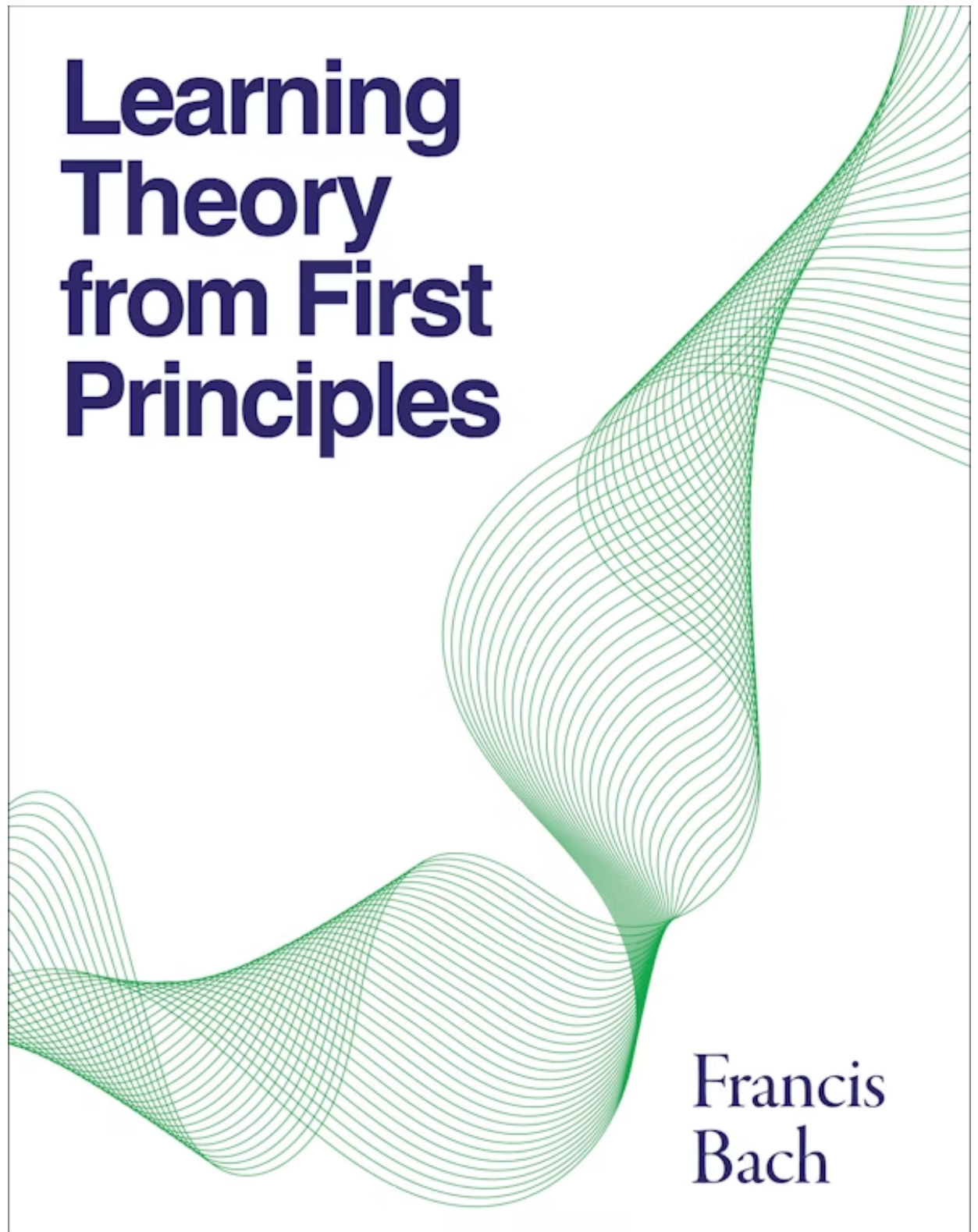


fbachorsay2017@gmail.com

Learning Theory from First Principles

Francis
Bach



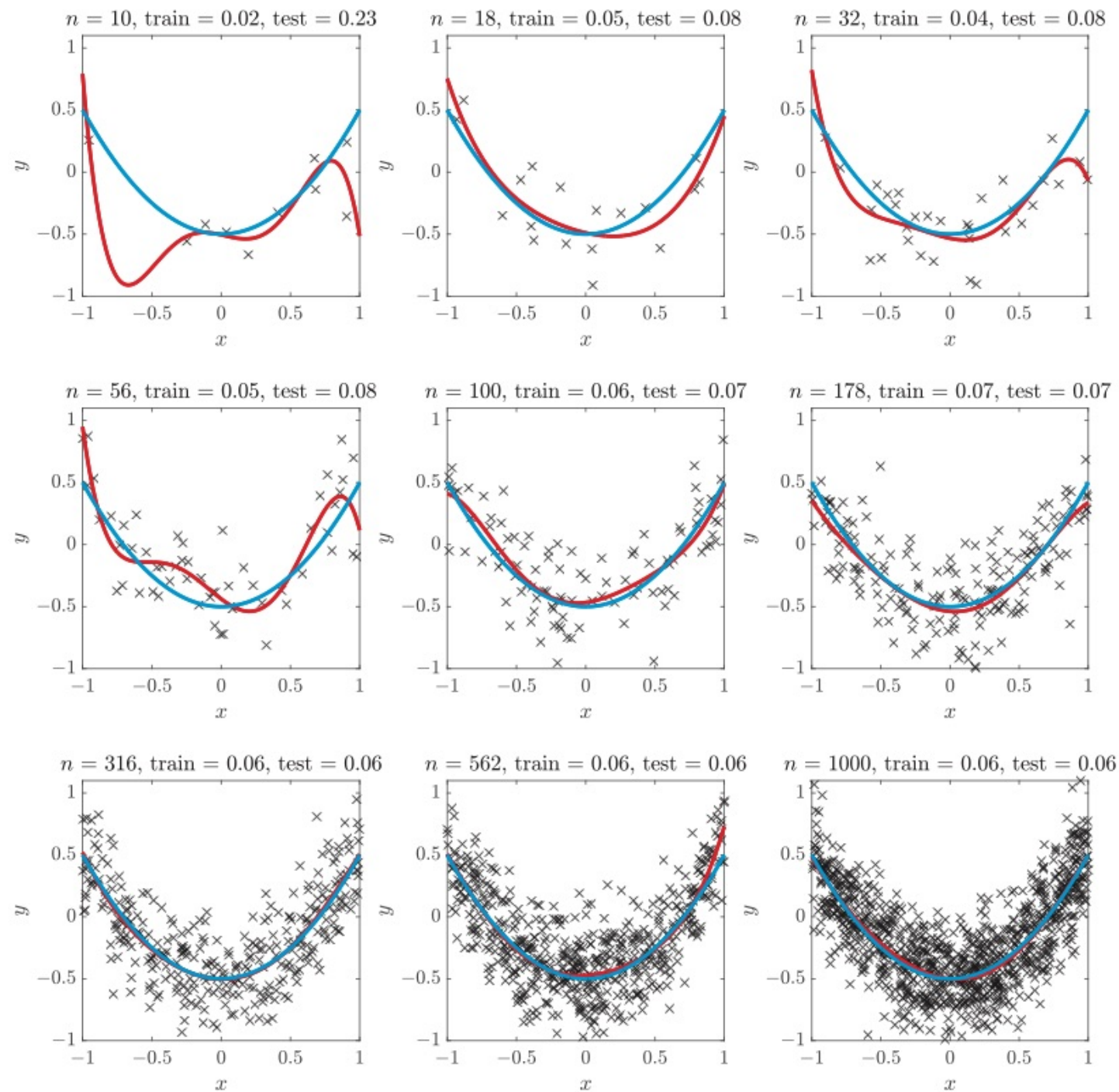
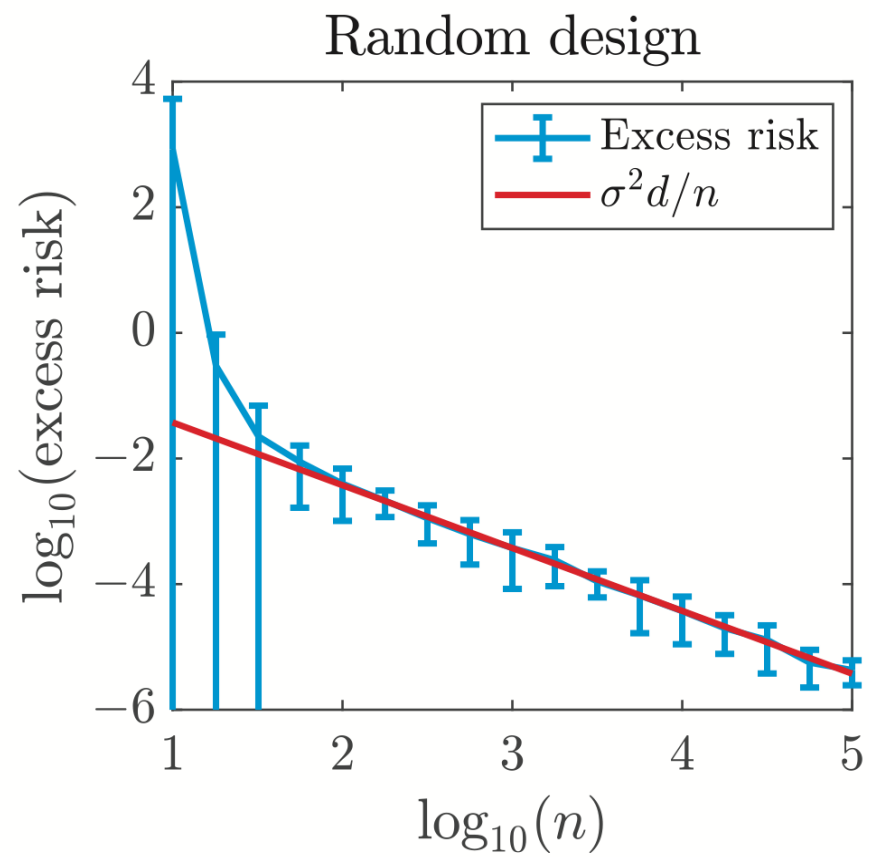
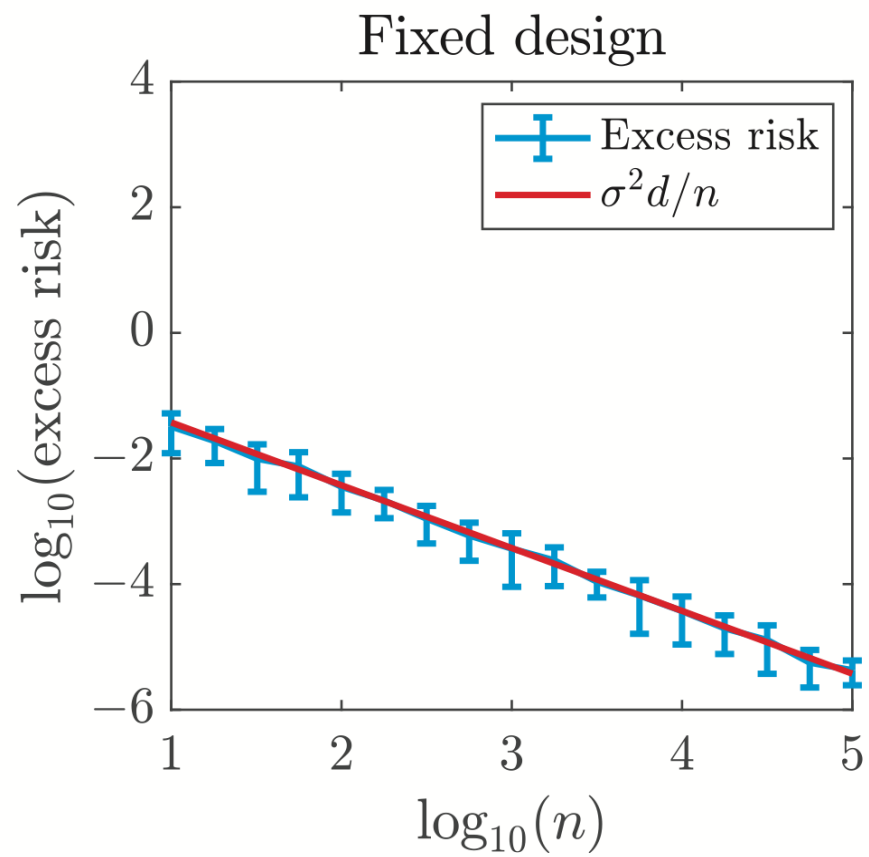
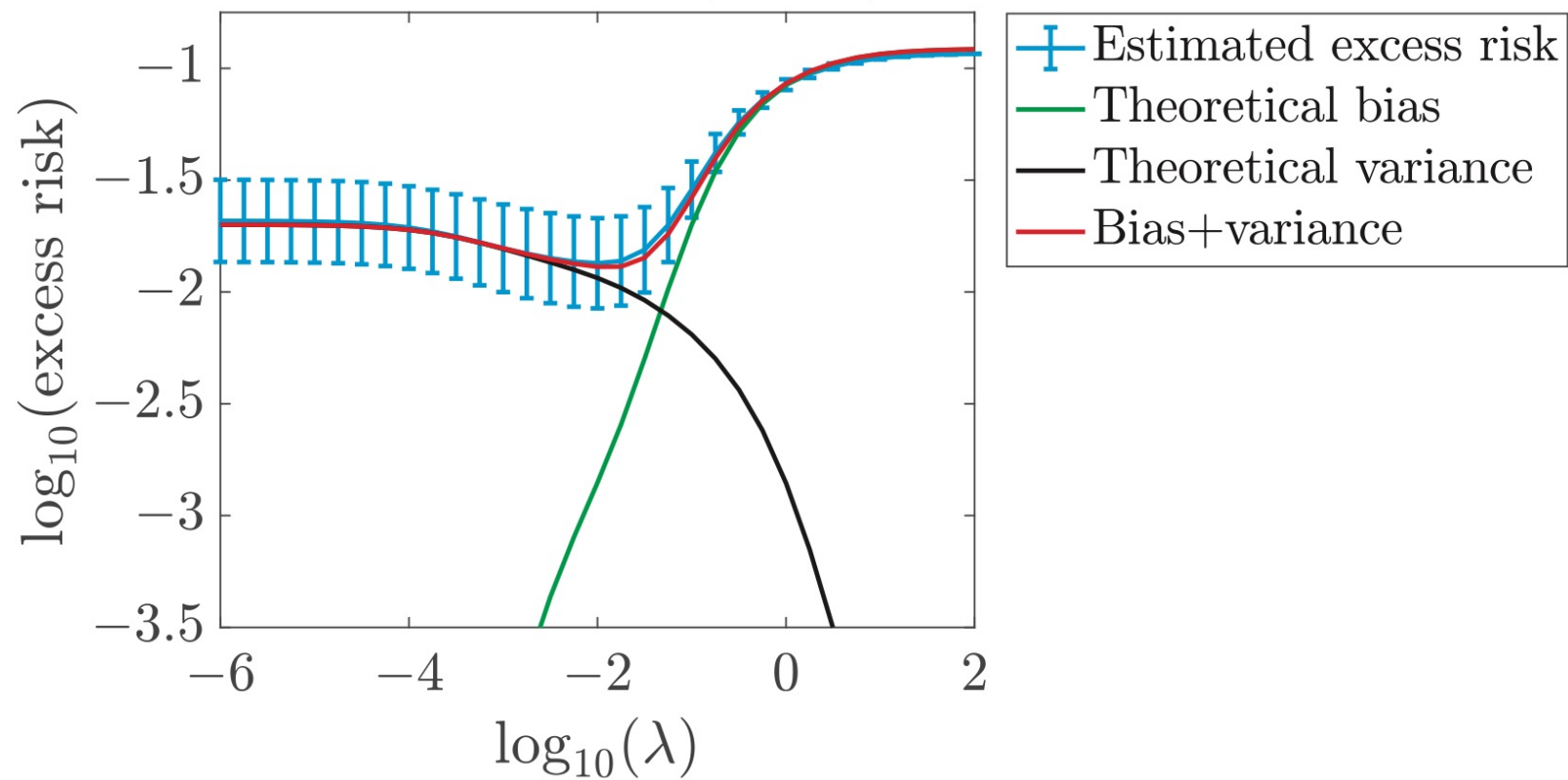


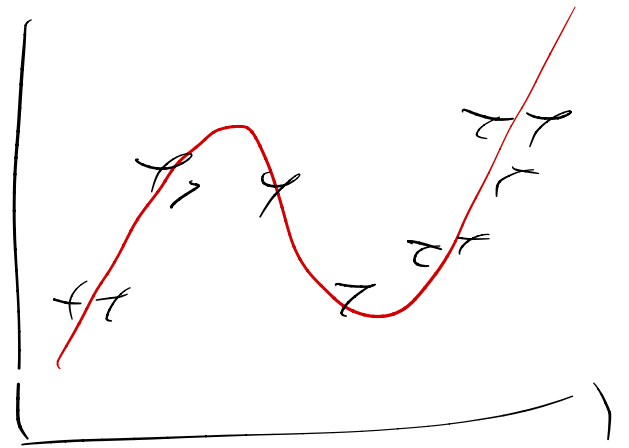
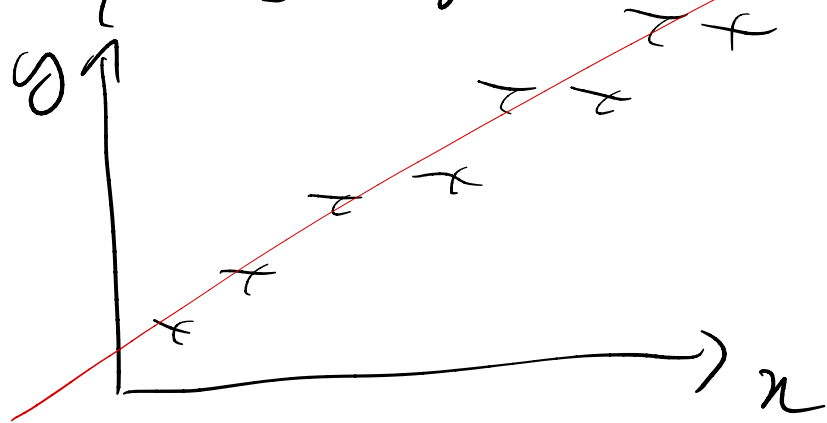
Figure 3.1. Polynomial regression in one dimension with a varying number of observations (from $n = 10$ to $n = 1,000$ with training and testing errors. Blue: Optimal prediction, red: estimated prediction by OLS with degree-5 polynomials.



Polynomial regression ($d = 5$) - Ridge



Least-squares regression



Distribution $p(x, y)$, $x, y \in \mathcal{X} \times \mathcal{Y}$

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

$f: \mathcal{X} \rightarrow \mathcal{Y}$

Training data (x_i, y_i) $i = 1 \dots n$

IID

Bayes predictor $f^*(x) = \arg \min_{f \in \mathcal{F}} \mathbb{E}[\ell(y, f(x))]$

$\mathbb{E}[\ell(y, f(x))]$

$\ell(y, z) = (y - z)^2$ (square loss) $\Rightarrow f^*(x) = \mathbb{E}(y|x)$
 $y \in \mathbb{R}$

Linear model

$$f_{\Theta}: \mathcal{X} \rightarrow \mathbb{R}$$

$$f_{\Theta}(x) = \Theta^T \varphi(x) \text{ where } \varphi: \mathcal{X} \rightarrow \mathbb{R}^d$$

feature vector

$$= \sum_{j=1}^d \Theta_j \varphi(x)_j$$

Empirical risk minimizer

$$\hat{\mathcal{R}}(f_{\Theta}) = \hat{\mathcal{R}}(\Theta) = \frac{1}{n} \sum_{i=1}^n (y_i - \varphi(x_i)^T \Theta)^2$$

↑
empirical risk

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

$$(\Phi \Theta)_i = \varphi(x_i)^T \Theta$$

Matrix notations = $y \in \mathbb{R}^n$ response vector

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

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

$$\hat{R}(\theta) = \frac{1}{n} \sum_{i=1}^n (y_i - \phi(x_i)^T \theta)^2 \quad \left| \begin{array}{l} \text{Ordinary least-squares (OLS)} \\ \text{estimator} \end{array} \right.$$

$$= \frac{1}{n} \|y - \Phi \theta\|_2^2 \quad \left| \begin{array}{l} \theta \in \arg \min_{\theta} \hat{R}(\theta) \end{array} \right.$$

gradient $\frac{\partial \hat{R}(\theta)}{\partial \theta} = \frac{1}{n} \sum_{i=1}^n 2(y_i - \phi(x_i)^T \theta)(-\phi(x_i))$

$$= 2 \left[\frac{1}{n} \sum_{i=1}^n \phi(x_i) \phi(x_i)^T \theta - \frac{2}{n} \sum_{i=1}^n y_i \phi(x_i) \right]$$

design matrix = $\sum \phi(x_i) \phi(x_i)^T \in \mathbb{R}^{d \times d}$
 empirical (non centered)
 covariance matrix

$$\hat{R}'(\theta) = \frac{2}{n} \Phi^T (\Phi \theta - y) = \left[2 - \frac{1}{n} \Phi^T \Phi \theta - \frac{1}{n} \Phi^T y \right] \in \mathbb{R}^d$$

$$f(u) = u^T u \Rightarrow f'(u) = 2u \quad \left| \quad f(u) = g(\phi u) \Rightarrow f'(u) = \Phi^T g'(\phi u) \right|$$

Lemma: $\phi(x)^T \theta \cdot \phi(z) = \phi(x) \phi(z)^T \theta$ $\left| \quad \theta^T M \theta = \frac{1}{n} \theta^T \Phi \Phi^T \theta = \text{tr } M \theta \theta^T \right.$
 $= \sum_{i,j} M_{ij} \theta_i \theta_j$

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

Ordinary least-squares (OLS) estimator

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

$\hat{\theta} \in \arg \min_{\theta} \hat{R}(\theta)$

Gradient

$$\hat{R}'(\theta) = \frac{1}{n} \sum_{i=1}^n 2(y_i - \phi(x_i)^T \theta)(-\phi(x_i))$$

$$= 2 \left[\frac{1}{n} \sum_{i=1}^n \phi(x_i) \phi(x_i)^T \theta - \frac{2}{n} \sum_{i=1}^n y_i \phi(x_i) \right]$$

$d \times d$ matrix $= \sum$
empirical (non centered)
covariance matrix $\in \mathbb{R}^{d \times d}$

$$\hat{R}'(\theta) = \frac{2}{n} \Phi^T (\Phi \theta - y) = 2 \left[\frac{1}{n} \Phi^T \Phi \theta - \frac{1}{n} \Phi^T y \right] \in \mathbb{R}^d$$

$$\hat{R}'(\theta) = 0 \quad (\Rightarrow) \quad \frac{1}{n} \Phi^T \Phi \theta = \frac{1}{n} \Phi^T y$$

linear system

normal equations

$$\theta = \left(\frac{1}{n} \Phi^T \Phi \right)^{-1} \frac{1}{n} \Phi^T y \quad \text{if } \frac{1}{n} \Phi^T \Phi \text{ invertible}$$

OLS estimator

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

$$\hat{Q}'(\theta) = 2 \left(\frac{1}{n} \Phi^T \Phi \theta - \frac{\Phi^T y}{n} \right) = 0 \Leftrightarrow \frac{\Phi^T \Phi}{n} \theta = \frac{\Phi^T y}{n}$$

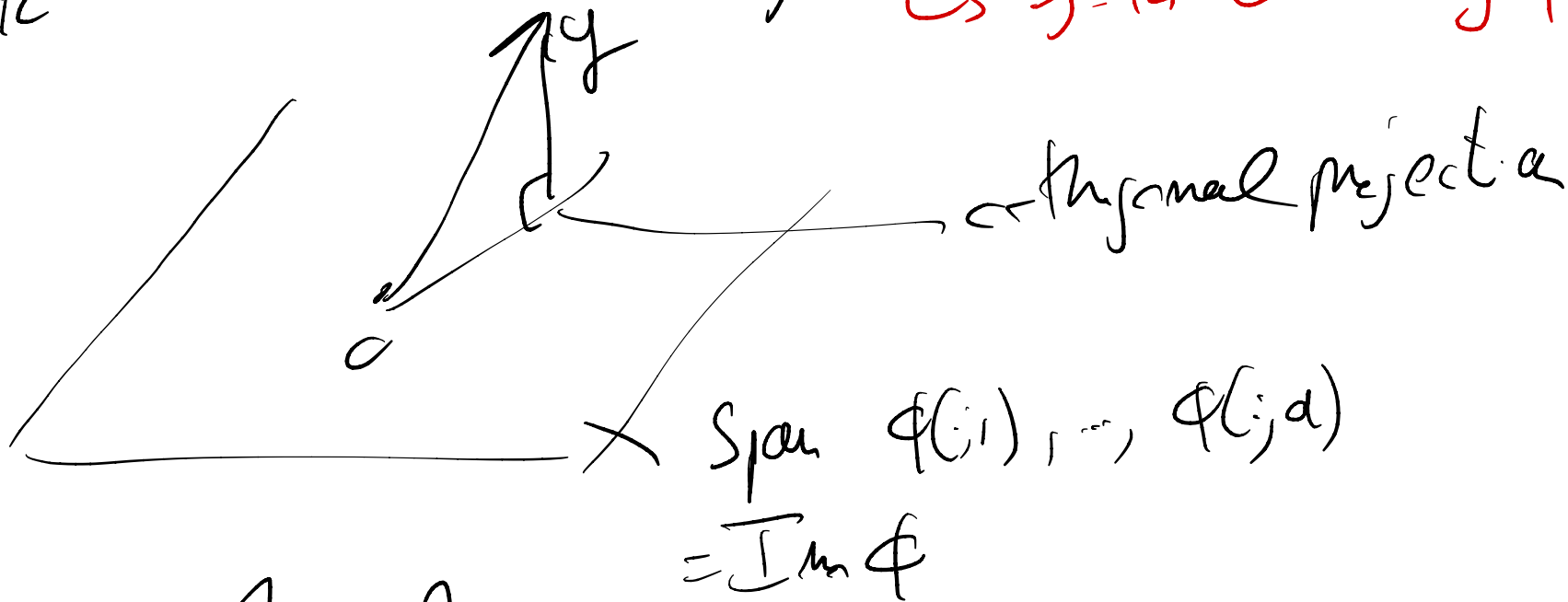
$\Phi \in \mathbb{R}^{n \times d}$

normal equations

Geometric interpretation in $\mathbb{R}^n$

$$\min_{\theta \in \mathbb{R}^d} \|y - \Phi\theta\|_2^2 = \left\| y - \sum_{j=1}^d \phi(:,j) \theta_j \right\|_2^2$$

$\Rightarrow$ $\phi(:,j)$ is j -th column of Φ



Computing $\hat{\theta}_{OLS}$?

$$\hat{\mathcal{L}}(\theta) = \frac{1}{n} \|y - \Phi\theta\|_2^2$$

$$\hat{\mathcal{L}}'(\theta) = 2 \left(\frac{1}{n} \Phi^T \Phi \theta - \frac{\Phi^T y}{n} \right) \Rightarrow \frac{\Phi^T \Phi}{n} \theta = \frac{\Phi^T y}{n}$$

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

normal eqns etc

Computing $\hat{\theta}_{OLS}$

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

SGD: $\mathcal{O}(dn)$

machine
precision

Bad way

Average way

$\underbrace{\text{inv}(\Phi^T \Phi)}_{d \times d} * (\Phi^T y)$

$(\Phi^T \Phi) \setminus (\Phi^T y)$

complexity

$\mathcal{O}(d^2 n)$

$\rightarrow$ get $\Phi^T \Phi$

$\mathcal{O}(d^3)$

$\rightarrow$ invert

Good way: QR factorization

solve

$R^T R \theta = R^T \Phi^T y$
 $R \theta = Q^T y$

$\Phi = QR$ — triangular

orthogonal

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

$$(x_i, y_i) \in X \times \mathbb{R} \quad \varphi: X \rightarrow \mathbb{R}^d$$

Model: (data) : $y_i = \varphi(x_i)^\top \theta_* + \varepsilon_i$

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

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

indep.

fixed design

$$\hat{R}(\theta) = \frac{1}{n} \|y - \Phi \theta\|_2^2 \quad \text{empirical risk}$$

$$R(\theta) = \mathbb{E}_y \hat{R}(\theta) \quad \text{with } y = \Phi \theta_* + \varepsilon$$

expanded risk

$$\hat{R}(\theta) = \frac{1}{n} \mathbb{E}_y \left[\|\Phi(\theta_* - \theta) + \varepsilon\|_2^2 \right] = \frac{1}{n} \|\Phi(\theta_* - \theta)\|_2^2 + \frac{1}{n} \sigma^2$$

in sample error

$$\mathbb{E} \varepsilon = 0 \quad \text{Cov}(\varepsilon) = \mathbb{E}(\varepsilon \varepsilon^\top) = \sigma^2 I$$

lemma: $\mathbb{E}_y \|\underline{y} - \underline{a}\|_2^2 = \|\mathbb{E} \underline{y} - \underline{a}\|_2^2 + \text{trace Cov}(\underline{y})$

Expected risk: $R(\alpha) = \frac{1}{n} \|\Phi(\theta - \theta_*)\|_2^2 + \sigma^2$

excess risk: $R(\hat{\theta}) - \min_{\theta} R(\theta) = \left[\frac{1}{n} \|\Phi(\hat{\theta} - \theta_*)\|_2^2 \right]$

$y = \Phi\theta_* + \varepsilon$ [estimator $\hat{\theta} = (\Phi^T \Phi)^{-1} \Phi^T y$

$\hat{\theta} = (\Phi^T \Phi)^{-1} \Phi^T (\Phi\theta_* + \varepsilon)$

Ordinary
least squares

$\hat{\theta} = \theta_* + (\Phi^T \Phi)^{-1} \Phi^T \varepsilon$

(OLS)

Goal: compute $\mathbb{E}[R(\hat{\theta})]$

$\mathbb{E}\left[\frac{1}{n} \|\Phi(\hat{\theta} - \theta_*)\|_2^2\right] = \frac{1}{n} \|\Phi(\mathbb{E}\hat{\theta} - \theta_*)\|_2^2 + \frac{1}{n} \text{trace cov}(\Phi\hat{\theta})$

$\rightarrow 0 + \frac{\sigma^2}{n} \text{tr} \Phi(\Phi^T \Phi)^{-1} \Phi^T \Phi(\Phi^T \Phi)^{-1} \Phi^T$
 $\frac{\sigma^2}{n} \text{tr} \Phi(\Phi^T \Phi)^{-1} \Phi^T \Phi(\Phi^T \Phi)^{-1} \Phi^T = \frac{\sigma^2}{n} \text{trace}(\mathbf{I}) = \frac{\sigma^2 d}{n}$

Ridge Regression a.k.a. ℓ_2 -regularized LS
LS with weight decay

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

regularized
empirical risk

$(\lambda \|\theta\|_2)$ in def 7

$\lambda \|\theta\|_2$?

if $\lambda > 0$, $\hat{R}_\lambda$ is strongly convex $\rightarrow$ unique minimum
even if $d > n$

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

gradient = 0 $\Rightarrow$ $\left(\frac{\Phi^T \Phi}{n} + \lambda I \right) \theta = \frac{1}{n} \Phi^T y$

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

Estimator $\hat{\theta}_\lambda = (\hat{Z} + \lambda I)^{-1} \frac{\Phi^T y}{n} = \underbrace{(\hat{Z} + \lambda I)^{-1} \hat{Z}}_{\text{Bias}} Q_* + \underbrace{(\hat{Z} + \lambda I)^{-1} \Phi^T \Sigma}_{\text{Variance}}$

Model $y = \Phi Q_* + \Sigma$

$\mathbb{E}(\hat{\theta}_\lambda) = Q_* - \lambda (\hat{Z} + \lambda I)^{-1} Q_*$

Excess risk $R(\hat{\theta}_\lambda) = \frac{1}{n} \|\Phi(\hat{\theta}_\lambda - Q_*)\|_2^2 =$
 $= (\hat{\theta}_\lambda - Q_*)^T \hat{Z} (\hat{\theta}_\lambda - Q_*)$

Goal: $\mathbb{E}[R(\hat{\theta}_\lambda) - \sigma^2]$

$= (\mathbb{E}\hat{\theta}_\lambda - Q_*)^T \hat{Z} (\mathbb{E}\hat{\theta}_\lambda - Q_*) + \text{trace} \hat{Z} \text{Cov}(\hat{\theta}_\lambda)$

$= (-\lambda (\hat{Z} + \lambda I)^{-1} Q_*)^T \hat{Z} (-\lambda (\hat{Z} + \lambda I)^{-1} Q_*) + \text{trace} \hat{Z} (\hat{Z} + \lambda I)^{-1} \frac{\sigma^2}{n}$

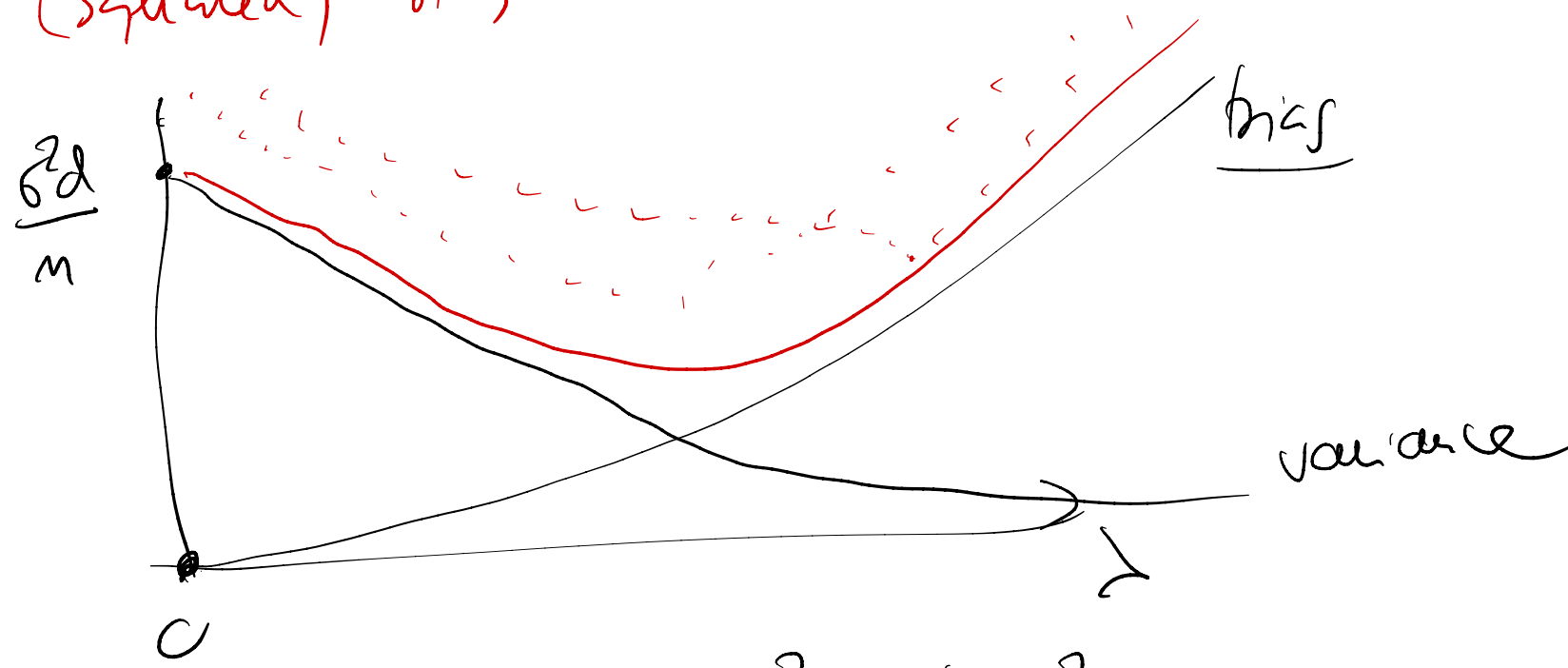
$= \underbrace{\lambda^2 Q_*^T \hat{Z} (\hat{Z} + \lambda I)^{-2} Q_*}_{\text{(squared) bias}} + \underbrace{\frac{\sigma^2}{n} \text{trace} \hat{Z}^2 (\hat{Z} + \lambda I)^{-2}}_{\text{variance}}$

$+ \sigma^2 d/n$

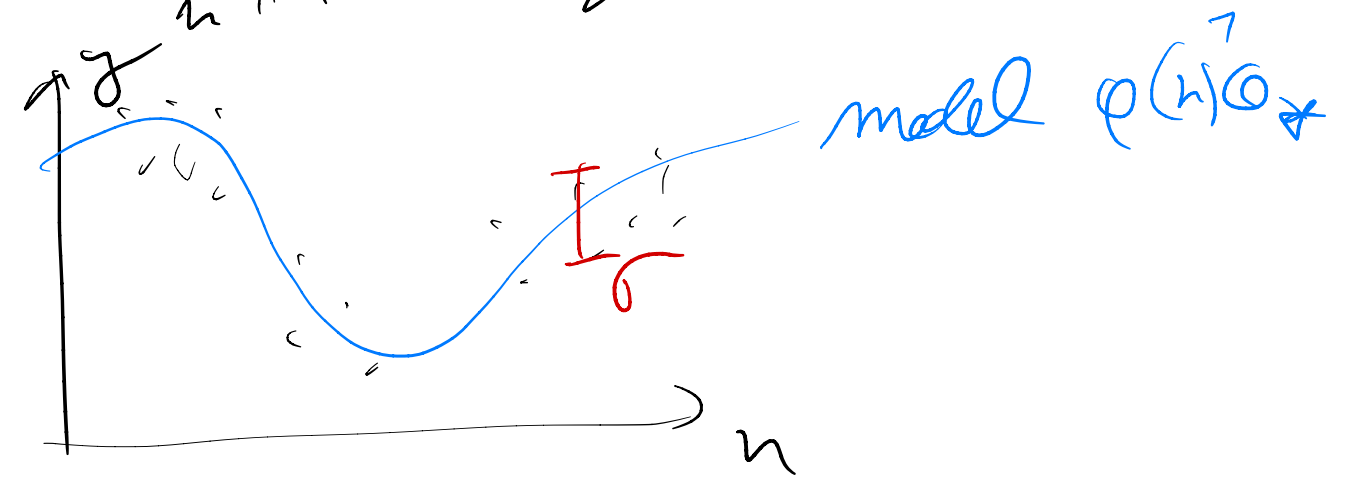
$wx+b$
 $\downarrow$
 weight bias
 $\lambda = 0$

Lemma: $\text{Cov}(\hat{\theta}_\lambda) = \text{Cov}((\hat{Z} + \lambda I)^{-1} \frac{\Phi^T \Sigma}{n}) = (\hat{Z} + \lambda I)^{-1} \frac{\Phi^T \Sigma \Phi}{n} (\hat{Z} + \lambda I)^{-1}$
 $= (\hat{Z} + \lambda I)^{-2} \frac{\hat{Z}}{n} \sigma^2$

testing error $\Rightarrow$ $\underbrace{\lambda^2 \phi_0^T \hat{\Sigma} (\hat{\Sigma} + \lambda I)^{-2} \phi_0}_{\text{(squared) bias}} + \underbrace{\frac{\sigma^2}{n} \text{trace } \hat{\Sigma}^2 (\hat{\Sigma} + \lambda I)^{-2}}_{\text{variance}}$



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



$$\underbrace{\frac{1}{n} \phi^T \hat{\Sigma} (\hat{\Sigma} + \lambda I)^{-2} \phi}_{\text{(squared) bias}} \leq \frac{\lambda}{2} \|\phi\|_2^2$$

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

$$\text{variance} \leq \frac{\sigma^2}{2n\lambda} \text{tr} \hat{\Sigma}$$

eigenvalue decomposition (SVD) "eig"
 $\hat{\Sigma} = \frac{1}{n} \Phi^T \Phi$ symmetric positive semidefinite
 $d \times d$ eigenvalues are ≥ 0

$$\hat{\Sigma} = \sum_{i=1}^d \lambda_i u_i u_i^T \quad \text{with } (u_1, \dots, u_d) \text{ orthonormal basis}$$

$$= U \text{diag}(\lambda) U^T$$

$$\text{variance} = \frac{\sigma^2}{n} \sum_{i=1}^n \frac{\lambda_i^2}{(\lambda_i + \lambda)^2} \leq \frac{\sigma^2}{2n\lambda} \left(\sum_i \lambda_i \right) = \frac{\sigma^2}{2n\lambda} \text{tr}(\hat{\Sigma})$$

$$\text{lemma: } \frac{\lambda_i \lambda}{(\lambda_i + \lambda)^2} \leq \frac{1}{2}$$

$$\text{bias} \leq \frac{\lambda}{2} \|\phi\|_2^2$$

$$\underbrace{\lambda^2 \mathbf{Q}_x^T \hat{\Sigma} (\hat{\Sigma} + \lambda \mathbf{I})^{-2} \mathbf{Q}_x}_{\text{(squared) bias}} \leq \frac{\lambda}{2} \|\mathbf{Q}_x\|^2$$

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

$$\leq \frac{\lambda}{2} \|\mathbf{Q}_x\|_2^2 + \frac{\sigma^2}{2n\lambda} \text{tr } \hat{\Sigma}$$

minimized when

$$\lambda = \lambda_* = \frac{\sigma}{\sqrt{n}} \frac{\sqrt{\text{tr } \hat{\Sigma}}}{\|\mathbf{Q}_x\|_2}$$

value

$$\frac{\|\mathbf{Q}_x\|_2}{\sqrt{n}} \sigma \sqrt{\text{tr } \hat{\Sigma}}$$

$$\text{tr } \frac{1}{n} \sum_i \mathbf{q}(x_i) \mathbf{q}(x_i)^T = \frac{1}{n} \sum_i \|\mathbf{q}_i\|_2^2$$

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

$$\frac{\frac{1}{2} \frac{a}{b} + \frac{b}{2a} \geq \sqrt{\frac{a}{b}}}{\frac{1}{2} \frac{a}{b} + \frac{b}{2a} - \sqrt{\frac{a}{b}} = \frac{1}{2} \left(\sqrt{\frac{a}{b}} - \sqrt{\frac{b}{a}} \right)^2 \geq 0}$$

$$\lambda^* = \sqrt{\frac{a}{b}}$$

Dimensional analysis

$$y = \underbrace{\varphi(n)}_{h_g}^T \underbrace{\Theta}_{h_g m^{-1}} + \underbrace{\varepsilon}_{h_g}$$

$$\mathbb{E}(\varepsilon^2) = \sigma^2 \quad h_g^2$$

$$\text{Risk} \quad \underbrace{\frac{1}{n} \|y - \varphi\Theta\|_2^2}_{h_g^2} + \underbrace{\lambda \|\Theta\|^2}_{\frac{h_g^2}{n^2} = m^2}$$

$h_g^2 m^{-2}$

$$\lambda = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{\text{tr}(\hat{\Sigma})}{\|\Theta\|_2}} = \frac{h_g \cdot m}{h_g m^{-1}} = m$$

Assume $\|\varphi(n)\|^2 \leq 1$

Imagine
more