Least-square regression

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

$x_i \in X$, $y_i \in \mathbb{R}$

Square loss: $R(f) = \mathbb{E}\left[(y - f(x))^2\right]$ test error

$\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} (y_i - f(x_i))^2$ train error

Model: $f_\theta(x) = \varphi(x)^\top \theta$ where $\varphi : X \to \mathbb{R}^d$ feature map

linear model wrt parameters $\theta \in \mathbb{R}^d$

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

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

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

(1) Ordinary least-squares regression

OLS: $\min \frac{1}{n} \| y - \Phi\theta \|_2^2$ $\longrightarrow$ $\frac{\sigma^2 d}{n}$

excess risk

(2) Ridge regression ——— $+ \lambda \|\theta\|_2^2$

(3) lower bounds

$$\hat{\theta} \in \arg\min_{\theta} \frac{1}{n} \|y - \Phi\theta\|_2^2 = F(\theta)$$

Assuption $\frac{\Phi^T\Phi}{n} \in \mathbb{R}^{d\times d}$ empirical (non centered) covariance matrix

$\hat{\Sigma}$

invertible $\Rightarrow n \geq d$.

① Alg: $F'(\theta) = \frac{1}{n}\Phi^T(\Phi\theta - y) = 0$

residual vector

$\Leftrightarrow \Phi^T\Phi\theta = \Phi^T y$ normal equations

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

Complexity $O(d^3 + d^2 n)$

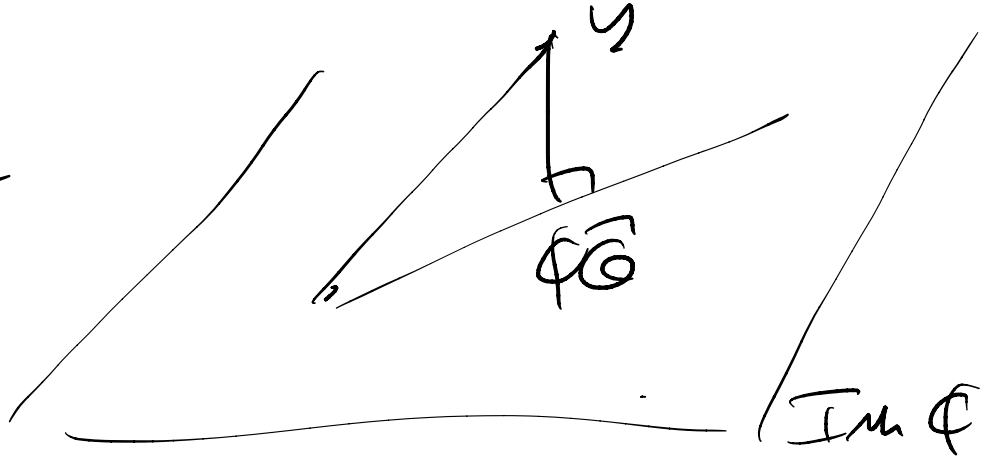
↳ $d^2 n$: $\Phi^T\Phi$

↳ d^3 using Cholesky

② geometric interpretation

$\min_{\theta \in \mathbb{R}^d} \|y - \Phi\theta\|_2^2 \Leftrightarrow$ projection of $y \in \mathbb{R}^n$ onto the column space of Φ

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

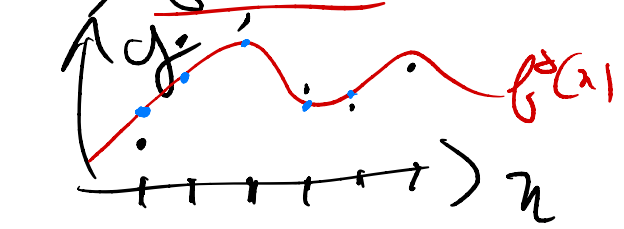


③ Analysis ⟶ random design

(x_i, y_i) iid from $p(x,y)$

goal: $R(f_{\hat{\theta}})$ small (rest risk)

⟶ fixed design: $x_1, \dots, x_n$ are deterministic



y_i: random

goal: predict well only at $x_1, \dots, x_n$

Assumptions ① fixed design ($x_1, \dots, x_n$ deterministic)

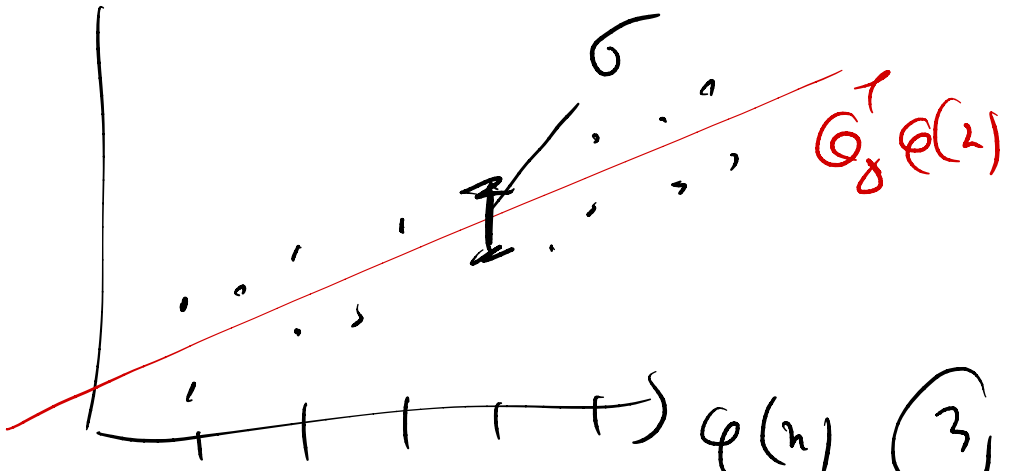
② assume model: $y_i = \varphi(x_i)^T \theta_* + \varepsilon_i$

where $\mathbb{E}\varepsilon_i = 0$, $\mathbb{E}\varepsilon_i^2 = \sigma^2$, $\varepsilon_1, \dots, \varepsilon_n$ independent

"well-specified" model

$\mathbb{E}(y|x) = \varphi(x)^T \theta_* = f^*(x)$ Bayes predictor "target" function

ε_i Gaussian



③ estimator: $\hat{\theta} = (\Phi^T \Phi)^{-1} \Phi^T y$

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

"In sample" expected risk. Assume same predictor θ

$$R(\theta) = \mathbb{E}\left(\frac{1}{n} \sum_i (y_i - \varphi(x_i)^T \theta)^2 \right)$$

deterministic

↳ Testing y_i has distribution $y_i = \varphi(x_i)^T \theta_* + \varepsilon_i$

$$= \frac{1}{n} \mathbb{E} \| y - \Phi\theta \|_2^2 = \frac{1}{n} \mathbb{E} \| \underbrace{\Phi\theta_* + \varepsilon}_{\text{model}} - \Phi\theta \|_2^2$$

$$= \frac{1}{n} \mathbb{E} \| \varepsilon + \Phi(\theta_* - \theta) \|_2^2$$

$$= \frac{1}{n} \mathbb{E} \|\varepsilon\|_2^2 + \frac{1}{n} \mathbb{E} \| \Phi(\theta - \theta_*) \|_2^2 + \frac{2}{n} \mathbb{E}(\varepsilon^T \Phi(\theta_* - \theta))$$

$\frac{1}{n} \sum_i \mathbb{E}\varepsilon_i^2 = \sigma^2$; $\frac{1}{n} \mathbb{E} \| \Phi(\theta - \theta_*) \|_2^2$; 0

$$R(\theta) = \sigma^2 + (\theta - \theta_*)^T \frac{\Phi^T \Phi}{n} (\theta - \theta_*) \neq \|\theta - \theta_*\|^2$$

test risk $R(\theta) = \sigma^2 + (\theta - \theta_*)^T \frac{\Phi^T\Phi}{n} (\theta - \theta_*)$

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

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

goal: $\mathbb{E}_\varepsilon R(\hat{\theta}) - \sigma^2$ since $\mathbb{E}\varepsilon = 0 \Rightarrow$ "unbiased"

$= \mathbb{E}\left[\left[(\Phi^T\Phi)^{-1}\Phi^T\varepsilon\right]^T \frac{\Phi^T\Phi}{n} (\Phi^T\Phi)^{-1}\Phi^T\varepsilon\right]$

$= \mathbb{E}\, \varepsilon^T \left[\Phi (\Phi^T\Phi)^{-1} \frac{\Phi^T\Phi}{n} (\Phi^T\Phi)^{-1}\Phi^T\right]\varepsilon$

$= \frac{\mathbb{E}}{n} \operatorname{tr} \varepsilon^T \Phi (\Phi^T\Phi)^{-1}\Phi^T\varepsilon = \frac{\mathbb{E}}{n} \operatorname{tr} \Phi (\Phi^T\Phi)^{-1}\Phi^T \varepsilon\varepsilon^T$

$= \frac{\sigma^2}{n} \mathbb{E} \operatorname{tr} \Phi(\Phi^T\Phi)^{-1}\Phi^T = \frac{\sigma^2}{n} \operatorname{tr} I_d = \frac{\sigma^2 d}{n}$

$\mathbb{E}(\varepsilon\varepsilon^T)_{ij} = \sigma^2 I_{n=}$

Ridge regression : $\frac{1}{n}\|y - \Phi\theta\|_2^2 + \lambda\|\theta\|_2^2$ MAP

Random design

Loss-based

Data-driven $\lambda\|\theta\|_2$

$\hat{\theta} = \arg\min \|y - \Phi\theta\|_2$

$\|y - \Phi\theta\|_1$

$\mathrm{tr}AB = \mathrm{tr}BA$

$\mathrm{tr}\,\delta = \lambda$ if $\delta \in \mathbb{R}$

$2^5 5^2$

$3^4 2^6$

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

empirical risk — regularizata

$\exp\left(\quad\right) \propto \exp\left(\quad\right)\exp\left(-\lambda\|\theta\|_2^2\right)$

likelihood — prior

$\Leftrightarrow$ Maximum A posteriori estimata

$\Leftrightarrow$ optimize

with $y|\theta,\varphi(x)$ gaussian

θ gaussian

Estimator: no assumption on $\Phi \in \mathbb{R}^{n\times d}$

$\frac{1}{2}F'(\theta) = \frac{1}{n}\Phi^T(\Phi\theta - y) + \lambda\theta = 0$

$\Leftrightarrow \boxed{\hat{\theta}_\lambda = (\Phi^T\Phi + n\lambda I)^{-1}\Phi^T y}$

Estimator: $\hat{\theta}_\lambda = (\Phi^T\Phi + n\lambda I)^{-1}\Phi^T y = (\Phi^T\Phi + n\lambda I)^{-1}\Phi^T\Phi\theta_* + (\Phi^T\Phi + n\lambda I)^{-1}\Phi^T\varepsilon$

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

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

bias: $\hat{\theta}_\lambda - \theta_* = (\Phi^T\Phi + n\lambda I)^{-1}\Phi^T\varepsilon - n\lambda(\Phi^T\Phi + n\lambda I)^{-1}\theta_*$

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

V $\qquad$ B

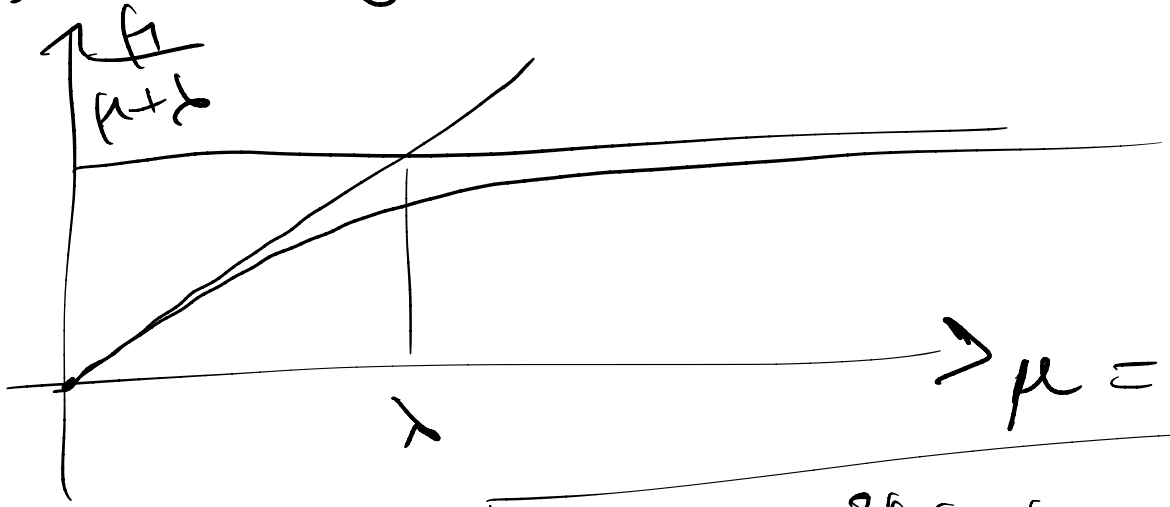
$V = \frac{1}{n}\mathbb{E}\,\mathrm{tr}\,\Phi(\Phi^T\Phi + n\lambda I)^{-1}\Phi^T\underbrace{\varepsilon\varepsilon^T}_{\sigma^2 I}\Phi(\Phi^T\Phi + n\lambda I)^{-1}\Phi^T$

$= \frac{\sigma^2}{n}\mathrm{tr}\,(\underbrace{\Phi^T\Phi}_{n\hat{\Sigma}} + n\lambda I)^{-2}(\Phi^T\Phi)^2 = \frac{\sigma^2}{n}\mathrm{tr}\,\hat{\Sigma}^2(\hat{\Sigma} + \lambda I)^{-2}$ (simple of $\sigma^2 d/n$)

$=$ degrees of freedom

Variance $\frac{\sigma^2}{n} \operatorname{tr}\, \hat{\Sigma}^2 (\hat{\Sigma}+\lambda I)^{-2} = \sum_{i=1}^{d} \frac{\lambda_i^2}{(\lambda_i+\lambda)^2}$ ≈ "counting" the eigenvalues above λ

λ_i are eigenvalues of $\hat{\Sigma}$



$\frac{\sigma^2}{n} df(\lambda)$

Bias: $n\lambda^2 \left\| \Phi (\Phi\Phi + n\lambda I)^{-1} \theta_0 \right\|_2^2$

$= \theta_*^T (\Phi^T\Phi + n\lambda I)^{-2} \Phi^T\Phi\, \theta_* \, n\lambda^2$

$= \lambda^2 \theta_*^T (\hat{\Sigma} + \lambda I)^{-2} \hat{\Sigma}\, \theta_*$

increasing in λ

Risk = variance + bias

$$\frac{\sigma^2}{n} \operatorname{tr}\hat{\Sigma}^2(\hat{\Sigma}+\lambda I)^{-2} + \lambda^2 \theta_*^T \hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-2}\theta_*$$

Lemma: $\forall \mu, \forall \lambda \geq c$ $(\mu+\lambda)^{-2}\mu \leq \frac{1}{2\lambda} \Leftrightarrow (\mu+\lambda)^2 \geq 2\lambda\mu$

applied to μ = eigenvalues of $\hat{\Sigma}$

$$\Rightarrow (\hat{\Sigma}+\lambda I)^{-2}\hat{\Sigma} \preceq \frac{1}{2\lambda} I$$

$$V \leq \frac{\sigma^2}{2n}\frac{\operatorname{tr}\hat{\Sigma}}{\lambda} \qquad B \leq \frac{\lambda}{2}\|\theta_*\|^2$$

variance — bias

optimal $\lambda^* = \frac{\sigma}{\sqrt{n}} \frac{\sqrt{\operatorname{tr}\hat{\Sigma}}}{\|\theta_*\|_2}$ and optimal performance $\frac{\sigma}{\sqrt{n}} \frac{\sqrt{\operatorname{tr}\hat{\Sigma}}}{R}\|\theta_*\|_2$

dimension independent bound $\neq \frac{\sigma^2 d}{n}$

Reg : adding regularizer

(1) explicitly $\hat{R}(\beta) + \frac{\lambda}{2}\|\beta\|_2^2$

$+ \lambda\|\beta\|_1$

$\hat{R}(\beta)$ such that $\|\beta\|_2 \leq D$

(2) Computational regularization.

single pass SGD

Random design : $(x_1, y_1), \ldots, (x_n, y_n)$ IID

$\neq$ fixed design

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

Model : $y_i = \varphi(x_i)^T\theta_* + \varepsilon_i$ $\Rightarrow$ expected risk

$$\mathbb{E}\|y - \varphi(x)^T\theta\|^2 \qquad (x, y) \text{ testing data}$$

$$y = \varphi(x)^T\theta_* + \varepsilon$$

$$= \sigma^2 + (\theta - \theta_*)^T \mathbb{E}\varphi(x)\varphi(x)^T(\theta - \theta_*)$$

$$= \sigma^2 + (\theta - \theta_*)^T \Sigma (\theta - \theta_*).$$

excess risk of OLS $= \frac{\sigma^2}{n}\mathbb{E}\left[\mathrm{tr}\,\hat{\Sigma}^{-1}\Sigma\right]$.

$\varphi(x)$ is Gaussian $\Rightarrow \frac{\sigma^2 d}{n}\left(1 - \frac{d+1}{n}\right.$

Morning : ERM initial square loss. | Statistics
Rademacher complexity

Afternoon : Convex optimization | optim

SGD ⟶ combination

⟶ linear models
in finite dim.

Morning : kernel methods

neural networks

eft.