Least-squares regression

last week: $\frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i))$ empirical risk

output input

training data

$R(f) = E[\ell(y, f(x))]$.

losses

square

function classes

linear

(1) What is the optimal prediction?

$$f^*(x) = \arg\min_z \mathbb{E}_{p(y|x)}\left[\ell(y, z)\right], \quad \ell(y,z) = (y-z)^2$$

$$= \mathbb{E}(y|x)$$

(2) Model

$$f(x) = f_\theta(x), \quad \theta \in \Theta$$

$$= \theta^T \varphi(x)$$

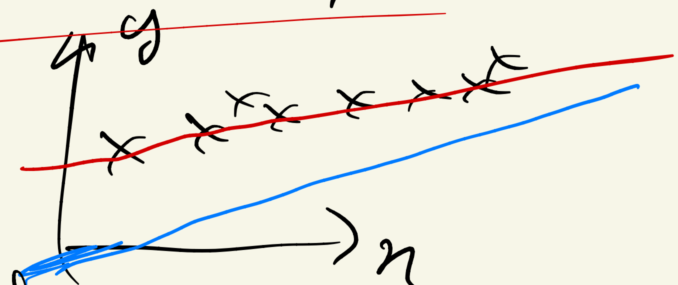
↑ parameters

Least-squares regression

Data (x_i, y_i), $i = 1, \dots, n$, with $x_i \in \mathcal{X}$, $y_i \in \mathbb{R}$

feature function $\varphi : \mathcal{X} \to \mathbb{R}^d$

Method: $\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \theta^T \varphi(x_i)\right)^2 = F(\theta)$



Notes: (1) often $\varphi(x) = \begin{pmatrix} x \\ 1 \end{pmatrix}$

(2) Vector/matrix notation

$y \in \mathbb{R}^n$, $\Phi \in \mathbb{R}^{n \times d}$, $\Phi = \begin{pmatrix} \varphi(x_1)^T \\ \vdots \\ \varphi(x_n)^T \end{pmatrix}$

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

with $\|z\|_2^2 = \sum_{j=1}^{d} z_j^2$ Euclidean norm

$$= \frac{1}{n} \sum_{i=1}^{n} \left(y_i - (\Phi\theta)_i\right)^2$$

Minimizing $F(\theta) = \frac{1}{n}\|y - \Phi\theta\|_2^2 \Rightarrow$ convex $F'(\theta) = 0$

$+ \lambda/2 \|\theta\|_2^2$

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

$$F(\theta) = \frac{1}{n}\sum_i (y_i - \varphi(x_i)^T\theta)^2 = \frac{1}{n}\sum_{i=1}^n \Big(y_i - \sum_{j=1}^d \varphi(x_i)_j \theta_j\Big)^2$$

$$F'(\theta) = -\frac{1}{n}\sum_i 2(y_i - \varphi(x_i)^T\theta)\varphi(x_i) = -\frac{2}{n}\sum_i y_i\varphi(x_i) + \frac{2}{n}\sum_i \varphi(x_i)\varphi(x_i)^T\theta$$

$$\frac{\partial F}{\partial \theta_j}(\theta) = -\frac{1}{n}\sum_i 2\Big(y_i - \sum_{j'=1}^d \varphi(x_i)_{j'}\theta_{j'}\Big)\varphi(x_i)_j$$

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

$$\hat{\Sigma}_{jj} = \frac{1}{n}\sum_{i=1}^n \varphi(x_i)_j^2$$

$$\Phi = \begin{pmatrix} \varphi(x_1)^T \\ \vdots \\ \varphi(x_n)^T \end{pmatrix} \in \mathbb{R}^{n\times d}$$

$$\hat{\Sigma} = \frac{1}{n}\Phi^T\Phi = \frac{1}{n}\sum_{i=1}^n \varphi(x_i)\varphi(x_i)^T$$

matrix of second order moment (non centered covariance matrix)

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

(normal equations) — linear system

$\frac{1}{n}\Phi^T\Phi \in \mathbb{R}^{d\times d}$, $\theta \in \mathbb{R}^d$, $\frac{1}{n}\Phi^T y \in \mathbb{R}^d$

Assumption: $\frac{1}{n}\Phi^T\Phi$ invertible

$\iff \operatorname{rank}\Phi = d$ (remember that $\Phi \in \mathbb{R}^{n\times d}$)

This imposes that $n \geq d$

Solution: $\theta = \left(\frac{1}{n}\Phi^T\Phi\right)^{-1}\frac{\Phi^T y}{n}$ (+ λI) 1 line of code

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

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

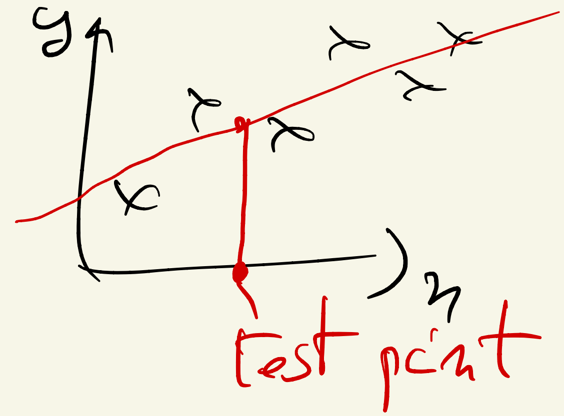
complexity? $O(d^3)$ only for small d + $O(d^2 n)$ dominant

gradient descent / stochastic $\to O(dn)$

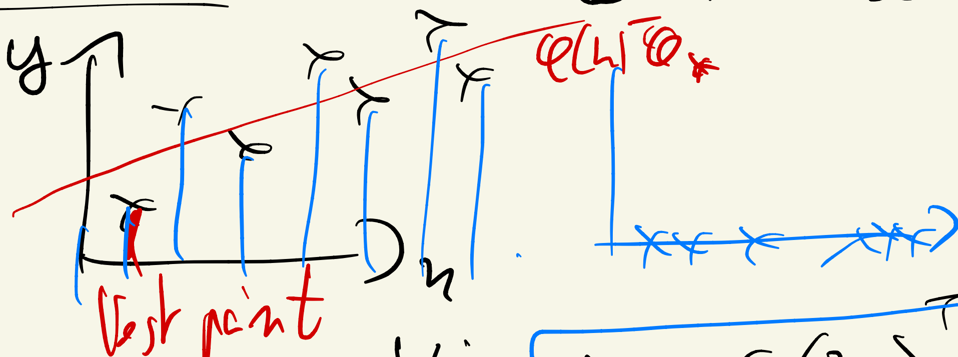
Summary: $F(\theta) = \frac{1}{n}\|y - \Phi\theta\|_2^2$ and $\hat{\theta} = \left(\frac{1}{n}\Phi^T\Phi\right)^{-1}\frac{1}{n}\Phi^T y$

(x_i, y_i) sampled iid from $dp(x,y)$ $\Rightarrow$ generalization performance

Random design: $\rightarrow$ goal is to minimize $R(\theta) = \mathbb{E}(y - \varphi(\theta)^T x)^2$



Fixed design: (x_i)'s are deterministic



Assumptions: $\forall i$, $y_i = \varphi(x_i)^T\theta_* + \varepsilon_i$ — only source of randomness

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

$\mathbb{E}\,\varepsilon_i^2 = \sigma^2$ noise variance

independent

goal: find θ such that

$\mathbb{E}_y(F(\theta))$ is as small as possible

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

Mo

$$F(\theta) = \hat{R}(\theta) = \frac{1}{n}\|y - \Phi\theta\|_2^2$$ with model $y = \Phi\theta^* + \varepsilon \in \mathbb{R}^n$

with $\mathbb{E}\,\varepsilon = 0$ $\quad (\mathbb{E}\,\varepsilon_i^2 = \sigma^2)$

$\mathbb{E}(\varepsilon\varepsilon^T) \in \mathbb{R}^{n\times n}$
$\quad \parallel$
$\quad \sigma^2 I$

$$(\mathbb{E}(\varepsilon\varepsilon^T))_{ij} = \mathbb{E}\,\varepsilon_i\varepsilon_j = \sigma^2 \text{ if } i=j$$
$$= 0 \text{ if } i \neq j$$

goal: $\min R(\theta) = \mathbb{E}\hat{R}(\theta)$

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

$$\|z\|_2^2 = z^T z$$

deterministic

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

$$= \frac{1}{n}\sum_{i=1}^n \underbrace{\mathbb{E}\,\varepsilon_i^2}_{\sigma^2} + \frac{1}{n}(\theta^* - \theta)^T\Phi^T\Phi(\theta^* - \theta) + 0$$

$$\boxed{R(\theta) = \sigma^2 + (\theta^* - \theta)^T\hat{\Sigma}(\theta^* - \theta)}$$

σ^2: R^*

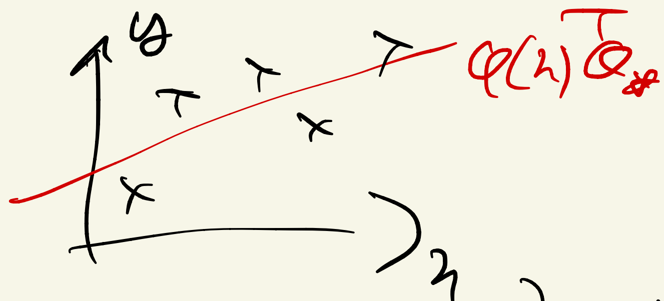
minimized for $\theta = \theta^*$

$$\Rightarrow \quad R(\theta) - R^* = (\theta^* - \theta)^T\hat{\Sigma}(\theta^* - \theta) \quad \text{excess risk}$$

different cost: $\|\theta^* - \theta\|_2^2$ not used

$R(\theta) - R^* = (\theta-\theta_*)^T \Sigma (\theta-\theta_*)$ with model $y = \Phi\theta_* + \varepsilon$

$\hat\theta = \hat\Sigma^{-1} \frac{\Phi^T y}{n}$ (normal equations)



$\mathbb{E}\varepsilon = 0$

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

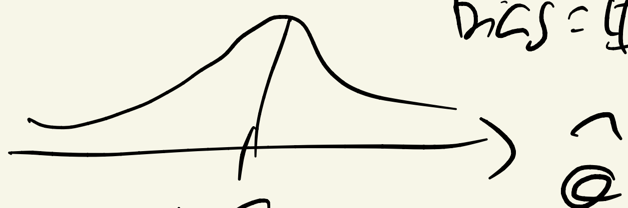
$\mathbb{E}[R(\hat\theta) - R_*] = \mathbb{E}[(\hat\theta-\theta_*)^T \hat\Sigma (\hat\theta-\theta_*)]$ (tr)

with $\hat\theta = \hat\Sigma^{-1}\frac{\Phi^T y}{n}$
$= \hat\Sigma^{-1}\frac{\Phi^T(\Phi\theta_*+\varepsilon)}{n}$

$\hat\theta = \theta_* + \hat\Sigma^{-1}\frac{\Phi^T\varepsilon}{n}$

<u>Consequences</u>: $\mathbb{E}\hat\theta = \theta_*$ unbiased

bias $= \mathbb{E}\hat\theta - \theta_*$



θ_*

Covariance matrix

$\mathbb{E}(\hat\theta-\theta_*)(\hat\theta-\theta_*)^T = \mathbb{E}\left[\frac{\hat\Sigma^{-1}\Phi^T\varepsilon\varepsilon^T\Phi\hat\Sigma^{-1}}{n^2}\right]$ ($\varepsilon\varepsilon^T \to \sigma^2 I$)

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

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

Consequences

$\mathbb{E}R(\hat\theta) - R^*$
$= \mathrm{tr}\left(\hat\Sigma\,\mathbb{E}(\hat\theta-\theta_*)(\hat\theta-\theta_*)^T\right)$
$= \mathrm{tr}\,\hat\Sigma\cdot\frac{\sigma^2}{n}\hat\Sigma^{-1}$
$= \mathrm{tr}\,\frac{\sigma^2}{n}I = \frac{\sigma^2 d}{n}$

Excess risk = $\frac{\sigma^2 d}{n}$ + dimension

$R(\hat\theta) - R^* \leq$ pessimistic

- it's an equality for any n
- it's only for fixed design → see book for analysis for random design
- it is "optimal" ⇒ if $y = \varphi(x)^T \theta_* + \varepsilon$ → see precise statements in book
- it is disappointing → need for regularization
- why random design harder ?

Regularization: replace $F(\theta) = \frac{1}{n}\|y - \Phi\theta\|_2^2$ $\frac{\lambda}{2}\theta^T\theta$

by $F(\theta) + \frac{\lambda}{2}\|\theta\|_2^2$ $\Rightarrow$ ridge regression

$F_\lambda(\theta)$ $+\lambda\|\theta\|_1$ $\Rightarrow$ Lasso

gradient: $F_\lambda'(\theta) = F'(\theta) + \lambda\theta$

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

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

normal equations: $(\hat{\Sigma} + \lambda I)\theta = \frac{1}{n}\Phi^T y$. Always unique solution when $\lambda > 0$

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

goal: compute bias and variance + excess risk.

$\hat{\theta}_\lambda = (\hat{\Sigma} + \lambda I)^{-1}\frac{\Phi^T y}{n} = (\hat{\Sigma} + \lambda I)^{-1}\frac{\Phi^T}{n}(\underbrace{\Phi\theta_* + \varepsilon}_{\text{model}}) = (\hat{\Sigma} + \lambda I)^{-1}\hat{\Sigma}\theta_* + (\hat{\Sigma} + \lambda I)^{-1}\frac{\Phi^T\varepsilon}{n}$

Bias: $\mathbb{E}\hat{\theta}_\lambda = (\hat{\Sigma} + \lambda I)^{-1}\hat{\Sigma}\theta_* = (\hat{\Sigma} + \lambda I)^{-1}(\hat{\Sigma} + \lambda I - \lambda I)\theta_*$

$= \theta_* - \lambda(\hat{\Sigma} + \lambda I)^{-1}\theta_*$

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

$$\hat{\theta}_\lambda = \underbrace{(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}\theta_*}_{\mathbb{E}\hat{\theta}_\lambda} + (\hat{\Sigma}+\lambda I)^{-1}\frac{1}{n}\Phi^T\varepsilon$$

$$\operatorname{var}(\hat{\theta}_\lambda) = \mathbb{E}\left[(\hat{\theta}_\lambda - \mathbb{E}\hat{\theta}_\lambda)(\hat{\theta}_\lambda - \mathbb{E}\hat{\theta}_\lambda)^T\right] = \mathbb{E}\left[\frac{(\hat{\Sigma}+\lambda I)^{-1}\Phi^T\varepsilon\varepsilon^T\Phi(\hat{\Sigma}+\lambda I)^{-1}}{n^2}\right]$$

($\varepsilon\varepsilon^T \to \sigma^2 I$)

covariance matrix

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

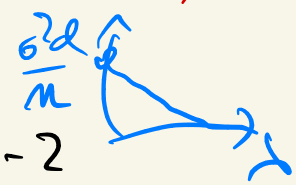
$\to$ when $\lambda = 0$: $\dfrac{\hat{\Sigma}^{-1}}{n}$

Main result:

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

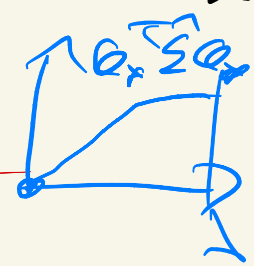
Lemma:

$$\mathbb{E}_Z (Z-a)^T M (Z-a) = (\mathbb{E}Z - a)^T M (\mathbb{E}Z - a) + \operatorname{tr} M \operatorname{var}(Z)$$



$$= \underbrace{\operatorname{tr}\frac{\hat{\Sigma}^2(\hat{\Sigma}+\lambda I)^{-2}\sigma^2}{n}}_{\text{variance term}} + \underbrace{\left(-\lambda(\hat{\Sigma}+\lambda I)^{-1}\theta_*\right)^T\hat{\Sigma}\left(-\lambda(\hat{\Sigma}+\lambda I)^{-1}\theta_*\right)}_{\text{bias term}}$$

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



goal: getting upper bounds and optimizing over λ

variance: $\frac{\sigma^2}{n} \operatorname{tr} \hat\Sigma^2 (\hat\Sigma + \lambda I)^{-2}$ — $\operatorname{tr} \hat\Sigma \times$ matrix

bias: $\lambda^2 \theta_*^\top \hat\Sigma (\hat\Sigma + \lambda I)^{-2} \theta_*$

requirement: $\hat\Sigma$ may not be invertible

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

matrix with eigenvalues $\mu(\mu+\lambda)^{-2}$ where μ is an eigenvalue of $\hat\Sigma$.

Lemma: $\mu(\mu+\lambda)^{-2} \leq \frac{1}{2\lambda} \quad \forall \lambda, \mu \geq 0$

$\Leftrightarrow (\mu+\lambda)^{-2} \leq \frac{1}{2\lambda\mu} \Leftrightarrow (\mu+\lambda)^2 \geq 2\lambda\mu$ true

bias $\leq \frac{1}{2} \lambda \|\theta_*\|^2$

variance $\leq \frac{\sigma^2}{2n} \lambda^{-1} (\operatorname{tr} \hat\Sigma)$

"optimal" tradeoff: $\lambda^2 = \frac{\sigma^2}{n} \frac{\operatorname{tr} \hat\Sigma}{\|\theta_*\|^2} \Rightarrow$ Risk $= \frac{\sigma}{\sqrt{n}} \|\theta_*\| \sqrt{\operatorname{tr} \hat\Sigma}$

- not optimal
- useless in practice

compare to $\sigma^2 d/n$

Homogeneity

$$y = \varphi(x)^\top \theta + \varepsilon$$

(units: y ~ kg; $\varphi(x)$ ~ m; θ ~ kg·m^{-1})

$$\mathbb{E}\,\varepsilon^2 = \sigma^2 \qquad \sigma \sim \text{kg}$$

Risk + $\lambda \|\theta\|^2$

kg^2 kg^2·m^{-2}

$$\|\varphi(x)\| \leq R, \quad \operatorname{tr}\hat\Sigma = \operatorname{tr}\frac{1}{n}\sum_i \varphi(x_i)\varphi(x_i)^\top \leq R^2 \sim m^2$$

$$\text{Expected risk} \sim \frac{\sigma}{\sqrt{n}}\sqrt{\operatorname{tr}\hat\Sigma}\,\|\theta_*\| \sim \text{kg}^2$$

(σ ~ kg; $\sqrt{\operatorname{tr}\hat\Sigma}$ ~ m; $\|\theta_*\|$ ~ kg·m^{-1})

kg^2

$$\lambda^2 = \frac{\sigma^2 \operatorname{tr}\hat\Sigma}{n\,\|\theta_*\|^2} \sim \frac{\text{kg}^2 m^2}{\text{kg}^2 m^{-2}} = m^4$$

$$\lambda \sim m^2$$

× Least squares
- ordinary: $\frac{\sigma^2 d}{n}$
- regularization: $\frac{1}{n\lambda} + \lambda$ —
  - $\frac{1}{n\lambda}$: variance
  - λ: bias

× Extensions
- square loss ← stats, optim
- linear parameters
- beyond L^2 — L_1