SDS 387
Linear Models

Fall 2024

Lecture 17 - Thu, Oct 29, 2024

Instructor: Prof. Ale Rinaldo

- Last time: convergence of gradient descent when $\text{rank}(\Phi) = d < n$

  We saw that convergence to the minimum is slower when

  $$\hat{\Sigma} = \frac{\Phi^\top \Phi}{n}$$ is rank-deficient (has rank $< d$).

- Classically, it is alway assumed that $\text{rank}(\hat{\Sigma}) = d$. But what if $d_{\min}(\hat{\Sigma}) = 0$?

- Suppose that $\underset{n \times d}{\Phi}$ has more columns than rows $(d > n)$. What happens?

  $\hat{\beta}$ is still obtained as solution to the normal equations

  $$\Phi^\top \Phi \hat{\beta} = \Phi^\top y \qquad \bigstar$$

  But now there are infinitely many solutions! That is if say $\tilde{\beta}$ solves $\bigstar$, then $\tilde{\beta} + u$ is also a solution for every $u \in \text{kernel}(\Phi)$

- Furthermore, for any solution $\tilde{\beta}$ to ☆, we have that

$$\Phi \tilde{\beta} = Y$$

↳ we interpolate the data (overfitting)
HW

- Among the infinitely many solutions, one is somewhat "canonical": it is the one with smallest Euclidean norm!

It can be calculated using Moore-Penrose pseudo-inverse:

→ pseudo-inverse

$$\underbrace{\hat{\beta}_{MN}}_{\text{min-norm}} = (\Phi^\top \Phi)^+ \Phi^\top Y$$

where for a matrix $\underset{m \times n}{A}$ its Moore-Penrose pseudo-inverse is $\underset{n \times m}{A^+}$ a unique matrix satisfying the conditions:

i) $AA^+A = A$ (AA^+ maps columns of A to themselves, it is an identity on $C(A)$)

ii) $A^+AA^+ = A^+$

iii) $\underset{m \times m}{AA^+}$ and $\underset{n \times n}{A^+A}$ are symmetric

Notice that AA^+ and A^+A are idempotent (see properties i) and ii)) and symmetric. So, AA^+ is the orthogonal projector onto $C(A)$ and A^+A is the orthogonal projector onto $C(A^\top)$, the row space of A.

Useful identities $\quad A^{+} = (A^{T}A)^{+}A = A^{T}(AA^{T})^{+}$

$$\hookrightarrow \quad \hat{\beta}_{MN} = \Phi^{+} Y$$

if $\quad A^{T}A$ is invertible $\quad A^{+} = (A^{T}A)^{-1}A$

$\quad AA^{T}$ is invertible $\quad A^{+} = A^{T}(AA^{T})^{-1}$

- If $\underset{m\times n}{A} = \underset{m\times r}{U}\,\underset{r\times r}{\Sigma}\,\underset{r\times n}{V^{T}}$ $\quad r = \text{rank}(A)$ $\quad$ then

$$\underset{n\times m}{A^{+}} = V\,\Sigma^{-1}\,U^{T}$$

- So for regression:

$$\hat{\beta}_{MN} = \Phi^{+}Y = \text{argmin}\,\{\, \|\beta\| \text{ s.t. } \Phi\beta = Y \,\}$$

<u>Fact</u>: gradient descent initialized at 0 (or at any point in the row-span of Φ) converges to $\hat{\beta}_{MN}$

Properties of OLS in fixed design settings and assuming well-specifie model

From now on let's assume that the data are of the form

$$Y_i = \Phi_i^T \beta^* + \varepsilon_i \qquad \text{where } \varepsilon_1, \dots, \varepsilon_n \overset{iid}{\sim} (0, \sigma^2)$$

$i = 1, \dots, n$

$\Phi_1, \dots, \Phi_n$ are deterministic vectors in $\mathbb{R}^d$

$\downarrow$

Assume $\text{rank}(\Phi) = d$ $\longleftarrow$ $\underset{n \times d}{\Phi}$ has Φ_i^T as its ith row

Note : if we further assume that $\varepsilon_1, \dots, \varepsilon_n \overset{iid}{\sim} N(0, \sigma^2)$ then the likelihood of $Y_1, \dots, Y_n$ is

$$\left(\frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \right)^d \exp\left\{ - \frac{\| Y - \Phi \beta^* \|^2}{2\sigma^2} \right\}$$

and $\hat{\beta}$ is the MLE of β^*

OLS

Next, for any $\beta \in \mathbb{R}^d$, the risk of β is

$$R(\beta) = \mathbb{E}_Y \left[\frac{\| Y - \Phi \beta \|^2}{n} \right] = \mathbb{E}_\varepsilon \left[\frac{1}{n} \| \Phi (\beta^* - \beta) + \varepsilon \|^2 \right]$$

expectation wrt Y

$$= (\beta^* - \beta)^T \frac{\Phi^T \Phi}{n} (\beta^* - \beta) + \underbrace{\mathbb{E}_\varepsilon \left[\frac{\|\varepsilon\|^2}{n} \right]}_{= \sigma^2}$$

$$= \underbrace{\|\beta^* - \beta\|^2_{\hat{\Sigma}}}_{\text{estimation error}} + \underbrace{\sigma^2}_{= R(\beta^*)} = \|\beta^* - \beta\|^2_{\hat{\Sigma}} + R(\beta^*)$$

irreducible error
(smallest possible risk)

The quantity $R(\beta) - R(\beta^*) \geq 0$ is the <u>excess risk</u>

So, let's look at the expected excess risk of $\tilde{\beta} \rightarrow$ any estimator of β^*

$$\mathbb{E}\left[R(\tilde{\beta})\right] - R(\beta^*) = \mathbb{E}\left[(\beta^* - \tilde{\beta})^T \hat{\Sigma} (\beta^* - \tilde{\beta})\right]$$

$=$ add and subtract $\mathbb{E}[\tilde{\beta}]$

$$= (\beta^* - \mathbb{E}[\tilde{\beta}])^T \hat{\Sigma} (\beta^* - \mathbb{E}[\tilde{\beta}]) +$$

$$\mathbb{E}\left[(\tilde{\beta} - \mathbb{E}[\tilde{\beta}])^T \hat{\Sigma} (\tilde{\beta} - \mathbb{E}[\tilde{\beta}])\right]$$

$$+ 2\,\mathbb{E}\left[\cancel{(\tilde{\beta} - \mathbb{E}[\tilde{\beta}])^T \hat{\Sigma}} (\beta^* - \mathbb{E}[\tilde{\beta}])\right]$$

$= 0$ by linearity of $\mathbb{E}[\cdot]$

$$= \underbrace{\mathbb{E}\left[\|\tilde{\beta} - \mathbb{E}[\tilde{\beta}]\|^2_{\hat{\Sigma}}\right]}_{\text{variance term for } \tilde{\beta}} + \|\beta^* - \mathbb{E}[\tilde{\beta}]\|^2_{\hat{\Sigma}}$$

Bias term
($= 0$ if $\mathbb{E}[\tilde{\beta}] = \beta^*$)

$\downarrow$

bias - variance decomposition of excess risk

If we choose to use $\hat{\beta}$ the OLS estimator, then

i) $\mathbb{E}[\hat{\beta}] = \beta^*$

ii) $\text{Var}[\hat{\beta}] = \frac{\sigma^2}{n} \hat{\Sigma}^{-1}$

Pf/ $\mathbb{E}[\hat{\beta}] = (\Phi^T\Phi)^{-1}\Phi^T \underbrace{\mathbb{E}[Y]}_{\Phi\beta^*} = \beta^*$ because $\Phi^T\Phi$ is invertible

$$\text{Var}[\hat{\beta}] = \text{Var}[(\Phi^T\Phi)^{-1}\Phi^T Y] \qquad \left(\text{Var}[AY] = A\,\text{Var}[Y]A^T\right)$$

$$= (\Phi^T\Phi)^{-1}\Phi^T \underbrace{\text{Var}[Y]}_{\sigma^2 I} \Phi(\Phi^T\Phi)^{-1}$$

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

■

Using these facts, we can establish that

$$\mathbb{E}[R(\hat{\beta})] - R(\beta^*)] = \sigma^2 \frac{d}{n} \rightarrow 0 \quad \text{if } d = o(n)$$

Pf/ Because $\mathbb{E}[\hat{\beta}] = \beta^*$ we only need to analyze the variance term:

$$\mathbb{E}[\|\hat{\beta} - \beta^*\|^2_{\hat{\Sigma}}] = \mathbb{E}\left[\|(\Phi^T\Phi)^{-1}\Phi^T(\Phi\beta^* + \varepsilon) - \beta^*\|^2_{\hat{\Sigma}}\right]$$

$$= \mathbb{E}\left[\|\cancel{\beta^*} + (\Phi^T\Phi)^{-1}\Phi^T\varepsilon \cancel{- \beta^*}\|^2_{\hat{\Sigma}}\right]$$

$$= \mathbb{E}\left[\|\hat{\Sigma}^{-1}\frac{\Phi^T}{n}\varepsilon\|^2_{\hat{\Sigma}}\right]$$

$$= \mathbb{E}\left[\varepsilon^T \frac{\Phi}{n} \hat{\Sigma}^{-1} \cancel{\hat{\Sigma}} \cancel{\hat{\Sigma}^{-1}} \frac{\Phi^T}{n} \varepsilon \right]$$

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

$$= \mathbb{E}\left[\frac{1}{n} \varepsilon^T H \varepsilon \right]$$

$$= \frac{1}{n} \mathbb{E}\left[\mathrm{tr}\left(H \varepsilon \varepsilon^T \right) \right]$$

$$= \frac{1}{n} \mathrm{tr}\left(H \underbrace{\mathbb{E}[\varepsilon \varepsilon^T]}_{\sigma^2 I_n} \right)$$

$$= \frac{\sigma^2}{n} \mathrm{tr}(H)$$

Exercise!

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

Remarks

i) This rate is optimal (in d, n and σ^2)

ii) an analogous bound holds with high probability but it requires more advanced tools

iii) This bound implies that the risk of $\hat{\beta}$ is

$$\mathbb{E}\left[R(\hat{\beta})\right] = \mathbb{E}_{Y_{new} \in \mathbb{R}^n}\left[\frac{\| Y_{new} - \Phi \hat{\beta} \|^2}{n} \right]$$

fresh new set of samples

$$= \sigma^2 \left(1 + \frac{d}{n} \right)$$

This is called the out-of-sample risk or expected test error

If we insted we compute the in-sample expected risk:

$$\underbrace{\mathbb{E}\left[\hat{R}(\hat{\beta})\right]}_{\text{expected training error}} = \underbrace{\mathbb{E}_Y}_{\text{expectation wrt to some data used to compute } \hat{\beta}}\left[\frac{\|Y - \Phi\hat{\beta}\|^2}{n}\right]$$

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