Lecture 16/17

Chapter 4

- Linear Regression
  - Ordinary Least Squares
  - Ridg Regression

Linear Regression:

Consider the problem

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

We can think of this as approximating

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

Let's write compactly

$$\min_{\theta \in \mathbb{R}^d} R_n(\theta) := \frac{1}{n} \| y - X\theta \|_2^2$$

where $y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$, $X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$

An optimal solution θ_n^{OLS} is called an ordinary least squares estimator.

We will assume for now that X has full column rank

So $d < n$

$$X = \begin{bmatrix} \\ \\ \\ \\ \\ \end{bmatrix}$$

Lemma: θ_n^{OLS} exists and is unique and is given by

$$\theta_n^{OLS} = (X^T X)^{-1} X^T y$$

Proof: $\hat{R}$ is coercive and continuous and therefore has a minimizer.

Taking the gradient yields

$$0 = \nabla R_n(\theta) = \frac{2}{n} X^T (X\theta - y)$$

Solve for θ.

Lemma: The prediction

$$X\theta_n^{OLS} = X(X^TX)^{-1}X^Ty \text{ is}$$

the orthogonal projection of y onto $\text{Range}(X) \in \mathbb{R}^d$.

So θ_n^{OLS} is obtained by

① project y onto $\text{Range}(X)$ to get $\hat{y}$

② Solve $X\theta = \hat{y}$ for θ.

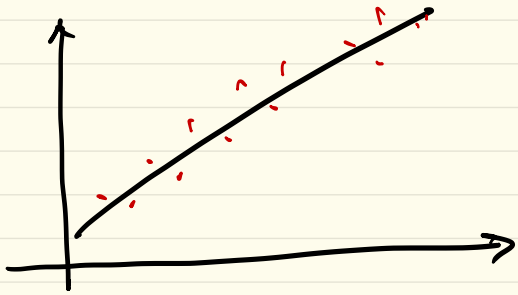
Next, let's estimate the generalization error:

$$R(\theta_n^{OLS}) - \underbrace{\min_{\theta} R(\theta)}_{R^{***}}$$

Two cases:

① Fixed design:

- $\exists \theta^*$ s.t. $y_i = x_i^T \theta^* + \varepsilon_i \quad \forall i$
- All ε_i are independent with $\mathbb{E}[\varepsilon_i] = 0$ and $\mathbb{E}[\varepsilon_i^2] \leq \sigma^2$.



② Random Design:

Both (x_i, y_i) are random.

More on this later.

Assume the fixed design setting. Set $\hat{\Sigma} = \frac{1}{n} X^T X$ and suppose $\hat{\Sigma}$ is invertible. Define the inner product

$$\langle \theta, \theta' \rangle_{\hat{\Sigma}} = \theta^T \hat{\Sigma} \theta'$$

and induced norm

$$\|\theta\|_{\hat{\Sigma}}^2 = \langle \theta, \theta \rangle_{\hat{\Sigma}} = \|\hat{\Sigma}^{1/2} \theta\|_2^2$$

Lemma:

$$R(\theta) - R^* = \|\theta - \theta^*\|_{\hat{\Sigma}}^2 \quad \forall \theta.$$

and $R^* = \sigma^2$?

pf:

$$R(\theta) = \mathop{\mathbb{E}}_{y} \frac{1}{n} \|y - X\theta\|_2^2$$

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

$$= \|\theta^* - \theta\|_{\hat{\Sigma}}^2 + \frac{1}{n} \mathop{\mathbb{E}}_{\varepsilon} \|\varepsilon\|_2^2$$

$$= \|\theta^* - \theta\|_{\hat{\Sigma}}^2 + \sigma^2$$

So $R^* = \sigma^2$. ☑

Recall for any estimator $\hat{\theta}_n$, we may write

$$\mathbb{E}\|\hat{\theta}_n - \theta^*\|_{\hat{\Sigma}}^2 = \underbrace{\|\mathbb{E}\hat{\theta}_n - \theta^*\|_{\hat{\Sigma}}^2}_{\text{Bias}(\hat{\theta}_n)} + \underbrace{\mathbb{E}\|\hat{\theta}_n - \mathbb{E}\hat{\theta}_n\|_{\hat{\Sigma}}^2}_{\text{Var}(\hat{\theta}_n)}$$

Lemma: $\text{Bias}(\hat{\theta}_n^{OLS}) = 0$

$$\mathbb{E}(\hat{\theta}_n^{OLS} - \theta^*)(\hat{\theta}_n^{OLS} - \theta^*)^T = \frac{\sigma^2}{n} \hat{\Sigma}^{-1}$$

pf: Recall $\theta_n^{OLS} = (X^TX)^{-1}X^Ty$

$= (X^TX)^{-1}X^T(X\theta^* + \varepsilon)$

$= \theta^* + (X^TX)^{-1}X^T\varepsilon$

So $\mathbb{E}\theta_n^{OLS} = \theta^*$ and

$\mathbb{E}(\theta_n^{OLS} - \theta^*)(\theta_n^{OLS} - \theta^*)^T$

$= (X^TX)^{-1}X^T(\underbrace{\mathbb{E}\varepsilon\varepsilon^T}_{\sigma^2 I})X(X^TX)^{-1}$

$= \sigma^2(X^TX)^{-1} = \frac{\sigma^2}{n}\hat{\Sigma}^{-1}$ ☐

Cor: $\mathbb{E}R(\theta_n^{OLS}) - R(\theta^*) = \frac{\sigma^2 d}{n}$

pf: $Var(\theta_n^{OLS}) = \mathbb{E}\|\theta_n^{OLS} - \theta^*\|_{\hat{\Sigma}}^2$

$= \mathbb{E}\,tr\left(\hat{\Sigma}(\theta_n^{OLS} - \theta^*)(\theta_n^{OLS} - \theta^*)^T\right)$

$= \frac{\sigma^2}{n}tr(I) = \frac{d\sigma^2}{n}$ ☐

Remark:

$$\mathbb{E}\, R_n(\theta_n^{OLS}) = \mathop{\mathbb{E}}_{y_1 \cdots y_n} \frac{1}{n} \| X\theta_n^{OLS} - y \|_2^2$$

$$= \mathop{\mathbb{E}}_{y_1 \cdots y_n} \frac{1}{n} \| (X(X^TX)^{-1}X^T - I)y \|_2^2$$

$$= \mathop{\mathbb{E}}_{\varepsilon} \frac{1}{n} \| (X(X^TX)^{-1}X^T - I)(X\theta^* + \varepsilon) \|_2^2$$

$$= \mathop{\mathbb{E}}_{\varepsilon} \frac{1}{n} \| \underbrace{(X(X^TX)^{-1}X^T}_{P} - I)\varepsilon \|_2^2$$

$$= \frac{\sigma^2}{n} \cdot \mathrm{tr}\left((P-I)(P-I)^T\right)$$

$$= \frac{\sigma^2}{n} \mathrm{tr}(P-I) = \frac{\sigma^2}{n}(n-d)$$

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

downward bias.

Better estimators than OLS trade bias for variance.

Defn: For $\lambda > 0$, define the ridge least-squares estimator

$$\hat{\theta}_n^\lambda = \arg\min_\theta \frac{1}{n} \|y - X\theta\|_2^2 + \lambda \|\theta\|_2^2$$

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

pf: Set $\nabla = 0$ and solve.

Lemma: $\text{Bias}(\hat{\theta}_n^\lambda) = \lambda^2 \theta^{*T} (\hat{\Sigma} + \lambda I)^{-2} \hat{\Sigma} \theta^*$

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

pf: $\mathbb{E}\hat{\theta}_n^\lambda = \frac{1}{n}(\hat{\Sigma}+\lambda I)^{-1} X^T X \theta^*$

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

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

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

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

So $\mathrm{Bias}(\hat{\theta}_n^\lambda) = \lambda^2 \left\|(\hat{\Sigma}+\lambda I)^{-1}\theta^*\right\|_{\hat{\Sigma}}^2$

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

$\mathrm{Var}\hat{\theta}_n^\lambda = \mathbb{E}\left\|\hat{\theta}_n^\lambda - \mathbb{E}\hat{\theta}_n^\lambda\right\|_{\hat{\Sigma}}^2$

$= \mathbb{E}\left\|\frac{1}{n}(\hat{\Sigma}+\lambda I)^{-1}X^T y - \frac{1}{n}(\hat{\Sigma}+\lambda I)^{-1}X^T X\theta^*\right\|_{\hat{\Sigma}}^2$

$= \mathbb{E}\left\|\frac{1}{n}(\hat{\Sigma}+\lambda I)^{-1}X^T\varepsilon\right\|_{\hat{\Sigma}}^2$

$= \frac{\sigma^2}{n}\mathrm{tr}\left(\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-1}\right)$ ☒

The quantity (evals of $\hat{\Sigma}$)

$$\mathrm{tr}\left(\hat{\Sigma}^2(\hat{\Sigma}+\lambda I)^{-2}\right) = \sum_{i=1}^{d} \frac{\lambda_i^2}{(\lambda_i+\lambda)^2} \leq d$$

is called the <u>degrees of freedom</u>

<u>Summary:</u>

$$\mathbb{E}\,\mathcal{R}(\hat{\theta}_\lambda) - \mathcal{R}^* = \lambda^2 \theta^{*T}(\hat{\Sigma}+\lambda I)^{-2}\hat{\Sigma}\theta^* + \frac{\sigma^2}{n}\mathrm{tr}\left(\hat{\Sigma}^2(\hat{\Sigma}+\lambda I)^{-2}\right)$$

<u>Remarks:</u>

- No explicit dependence on d
- Optimizing over λ can lead to a lower value than θ_n^{OLS}.

Note: $\lambda^2 \theta^{*T}(\hat{\Sigma}+\lambda I)^{-2}\hat{\Sigma}\theta^* \le \lambda \|(\hat{\Sigma}+\lambda I)^{-2}\lambda\hat{\Sigma}\|_{op}\|\theta^*\|_2^2$

$\le \lambda\|\theta^*\|_2^2$

$\frac{\sigma^2}{n}\mathrm{tr}\left(\hat{\Sigma}^2(\hat{\Sigma}+\lambda I)^{-2}\right) \le \frac{\sigma^2}{\lambda n}\mathrm{tr}\left(\hat{\Sigma}\,\lambda\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-2}\right)$

$\le \frac{\sigma^2}{2\lambda n}\mathrm{tr}(\hat{\Sigma})$

Optimizing over λ gives

$$\bar{\lambda} = \frac{\sigma}{\|\theta^*\|_2}\cdot\sqrt{\frac{\mathrm{tr}(\hat{\Sigma})}{n}}$$

Thm $\mathbb{E}\,R(\theta_n^{\bar{\lambda}}) - R^* \le \sigma\|\theta^*\|_2\sqrt{\frac{\mathrm{tr}(\hat{\Sigma})}{n}}$

Note $\mathrm{tr}(\hat{\Sigma}) = \frac{1}{n}\sum_{i=1}^n \|x_i\|^2$

typically independent of d.

- So worse rate in n but better in σ and no 'dependence on d'!

Random Design:

Suppose $\exists \theta^* \in \mathbb{R}^d$ s.t.

- $y = x^T\theta^* + \varepsilon$
- ε is independent from x, and $\mathbb{E}\varepsilon = 0$, $\mathbb{E}\varepsilon^2 = \sigma^2$

Define $\Sigma = \mathbb{E}xx^T$

Lemma: $R(\theta) - R^* = \|\theta - \theta^*\|_\Sigma^2$ and $R^* = \sigma^2$.

Pf:

$$R(\theta) = \mathbb{E}(y - x^T\theta)^2$$
$$= \mathbb{E}(x^T(\theta^* - \theta) + \varepsilon)^2$$
$$= (\theta^* - \theta)^T \mathbb{E}xx^T(\theta^* - \theta) + \sigma^2.$$

□

Lemma:

$$\mathbb{E}\, R(\theta_n^{OLS}) - R^* = \frac{\sigma^2}{n}\mathbb{E}\, tr(\Sigma\hat{\Sigma}^{-1})$$

Pf: Recall $\theta_n^{OLS} = \theta^* + \frac{1}{n}\hat{\Sigma}^{-1}X^T\varepsilon$

So

$$\mathbb{E}R(\theta_n^{OLS}) - R^* = \mathbb{E}\|\theta_n^{OLS} - \theta^*\|_\Sigma^2$$

$$= \frac{1}{n^2}\mathbb{E}\,\varepsilon^T X\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}X^T\varepsilon$$

$$= \frac{1}{n^2}tr\left(X\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}X^T\mathbb{E}(\varepsilon\varepsilon^T)\right)$$

$$= \frac{\sigma^2}{n^2}tr\left(\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}(X^TX)\right)$$

$$= \frac{\sigma^2}{n}tr\left(\hat{\Sigma}^{-1}\Sigma\right) \quad \blacksquare$$

So we must compute

$$\mathbb{E}\,\mathrm{tr}(\Sigma\hat{\Sigma}^{-1}) = \mathbb{E}\,\mathrm{tr}(\Sigma^{1/2}\hat{\Sigma}^{-1}\Sigma^{1/2})$$

$$\Sigma^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T\right)^{-1}\Sigma^{1/2}$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n}(\Sigma^{-1/2}x_i)(\Sigma^{-1/2}x_i)^T\right)^{-1}$$

$$= n\cdot\left(\sum_{i=1}^{n} z_i z_i^T\right)^{-1}$$

where $z_i = \Sigma^{-1/2}x_i$

If $x_i \sim N(0,\Sigma)$, then $z_i \sim N(0,I)$ and $(Z^TZ)^{-1}$ has a special distribution.

Can compute $\mathbb{E}(Z^TZ)^{-1} = \frac{1}{n-d-1}I$