# Lecture 18

## Last time

- Intro to estimation
- Maximum likelihood estimation

## Today

- Ordinary least squares
- Excess risk

## Least Squares

Suppose we wished to solve the problem

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

where

Random or fixed      iid $N(0, \sigma^2)$

$$y_i = \theta^{*T} x_i + \varepsilon_i$$

As we talked about last time this is known as the empirical risk and it approximates the population risk

$$R(\theta) = \mathbb{E}\, R_n(\theta).$$

We will use the following matrix notation

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

where

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad \text{and} \quad X = \begin{bmatrix} - x_1^T - \\ \vdots \\ - x_n^T - \end{bmatrix}.$$

We will assume $d \leq n$ and $X$ has rank $d$. An optimal solution of $(*)$ is called "ordinary least squares estimator" $\theta^{OLS}$.

**Lemma:** $\theta^{OLS}$ is exist and is unique. Further, it is given by

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

**Proof:** The first statement follows since $R_n(\cdot)$ is strongly convex. Then, using first order optimality yields

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

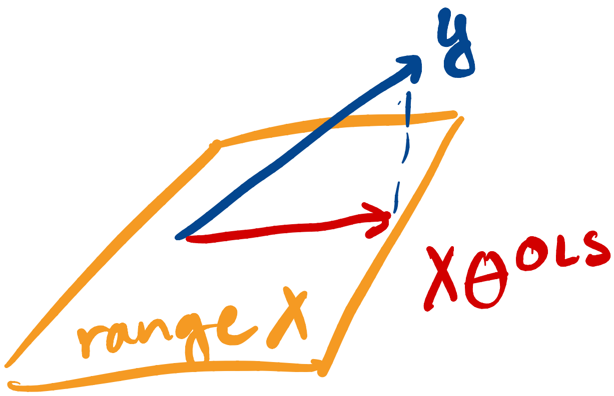
and so the formula for $\theta^{OLS}$ follows. $\square$

The OLS estimator has a nice geometric interpretation.

**Lemma:** The predictions

$$X\theta^{OLS} = X(X^TX)^{-1}X^Ty$$

is the orthogonal projection of $y$ onto $\text{range}(X) \subseteq \mathbb{R}^n$. $\dashv$




Thus we can see $\theta^{OLS}$ as solving

① $\hat{y} = \text{Proj}_{\text{range}X}\, y$

② Solve $X\theta = \hat{y}$.

## Excess risk

Natural question:

How close is $R(\theta^{OLS})$ to $\min_{\theta \in \mathbb{R}^d} R(\theta)$?

This is called the excess risk

We will study it in two situations

▹ **Fixed design:** Assume that $X$ is deterministic and we study

$$\mathbb{E}\, R(\theta^{OLS}) - \underbrace{\min_{\theta \in \mathbb{R}^d} R(\theta)}_{R^*}$$

Expectation w.r.t. $\varepsilon$ only.

▷ **Random design:** Assume that $X$ is random and we take the expectation w.r.t. $X$ as well.

We focus on the fixed design setting first. Define

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

which by assumption is invertible. Any positive definite matrix defines an inner product via

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

which induces a norm

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

Let's characterize the generalization error for any $\theta \in \mathbb{R}^d$.

**Lemma (8):** Suppose $X$ is fixed. Then,

$y = \theta^{*T}$

$$R(\theta) - R^* = \|\theta - \theta^*\|_{\hat{\Sigma}}^2 \qquad \forall\, \theta \in \mathbb{R}^d$$

and, moreover, $R^* = \sigma^2$. $\dashv$

**Proof:** Expanding

$$\begin{aligned} R(\theta) &= \frac{1}{n}\, \mathbb{E}_{\varepsilon} \|y - X\theta\|_2^2 \\ &= \frac{1}{n}\, \mathbb{E}_{\varepsilon} \|X(\theta^* - \theta) + \varepsilon\|_2^2 \\ &= \|\theta^* - \theta\|_{\hat{\Sigma}}^2 + \frac{1}{n}\, \mathbb{E}\|\varepsilon\|_2^2 \\ &= \|\theta^* - \theta\|_{\hat{\Sigma}}^2 + \sigma^2. \end{aligned}$$

Thus, $R^* = R(\theta^*) = \sigma^2$ and the result follows. $\square$

Recall from our computation last class that

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

Next we study these two for the OLS estimator.

**Lemma (✕):** Suppose that $X$ is fixed. Then,

$$\mathrm{Bias}_{\hat{\Sigma}}(\hat{\theta}) = 0 \quad \text{and} \quad \mathbb{E}(\theta^{OLS}-\theta^*)(\theta^{OLS}-\theta^*)^T = \frac{\sigma^2}{n}\hat{\Sigma}^{-1}.$$

**Proof:** Recall

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

So by linearity of expectation, $\mathbb{E}\theta^{OLS} = \theta^*$.

Further,

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

$$= \sigma^2(X^TX)^{-1}$$

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

$\square$

As a direct corollary of Lemmas (⌧) and (✕) we obtain a characterization of the excess risk.

**Corollary:** Suppose $X$ is fixed. Then

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

**Proof:** We know that since $\theta^{OLS}$ is unbiased, the excess risk is $\mathrm{Var}_{\hat{\Sigma}}(\hat{\theta})$. Thus,

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

Cyclic invariance of the trace $\rightarrow$

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

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

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

$\square$

Another natural question given that we don't have access to $R$, how close is $R_n(\theta^{OLS})$ to $R(\theta^{OLS})$?

**Lemma:** Suppose that $X$ is fixed. Then,

$$\mathbb{E}\,R_n(\theta^{OLS}) = \sigma^2 - \frac{d}{n}\sigma^2.$$

**Proof:** Expanding

$$\mathbb{E}\,R_n(\theta^{OLS}) = \frac{1}{n}\mathbb{E}_\varepsilon\|X\theta^{OLS} - y\|_2^2$$

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

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

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

$$= \frac{1}{n} \operatorname{tr}\left( (P - I)\, \mathbb{E}\varepsilon\varepsilon^T (P-I)^T \right)$$

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

Projection onto a $(n-d)$-dim subspace.

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

$\square$