

# Excess Risk Decomposition

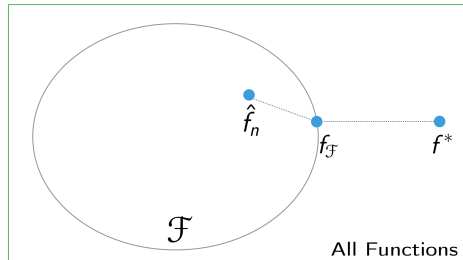
Julia Kempe & David S. Rosenberg

NYU CDS

January 29, 2019

## Excess Risk Decomposition

# Error Decomposition



$$f^* = \arg \min_f \mathbb{E} \ell(f(x), y)$$

$$f_{\mathcal{F}} = \arg \min_{f \in \mathcal{F}} \mathbb{E} \ell(f(x), y)$$

$$\hat{f}_n = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \ell(f(x_i), y_i)$$

- **Approximation Error** (of  $\mathcal{F}$ ) =  $R(f_{\mathcal{F}}) - R(f^*)$
- **Estimation error** (of  $\hat{f}_n$  in  $\mathcal{F}$ ) =  $R(\hat{f}_n) - R(f_{\mathcal{F}})$

## Definition

The **excess risk** compares the risk of  $f$  to the Bayes optimal  $f^*$ :

$$\text{Excess Risk}(f) = R(f) - R(f^*)$$

- Can excess risk ever be negative?

# Excess Risk Decomposition for ERM

- The excess risk of the ERM  $\hat{f}_n$  can be decomposed:

$$\begin{aligned}\text{Excess Risk}(\hat{f}_n) &= R(\hat{f}_n) - R(f^*) \\ &= \underbrace{R(\hat{f}_n) - R(f_{\mathcal{F}})}_{\text{estimation error}} + \underbrace{R(f_{\mathcal{F}}) - R(f^*)}_{\text{approximation error}}.\end{aligned}$$

# Approximation Error

Approximation error  $R(f_{\mathcal{F}}) - R(f^*)$  is

- a property of the class  $\mathcal{F}$
- the penalty for restricting to  $\mathcal{F}$  (rather than considering all possible functions)

*Bigger  $\mathcal{F}$  mean smaller approximation error.*

Concept check: Is approximation error a random or non-random variable?

# Estimation Error

Estimation error  $R(\hat{f}_n) - R(f_{\mathcal{F}})$

- is the performance hit for choosing  $f$  using finite training data
- is the performance hit for minimizing empirical risk rather than true risk

With *smaller*  $\mathcal{F}$  we expect *smaller* estimation error.

*Under typical conditions:* “With infinite training data, estimation error goes to zero.”

Concept check: Is estimation error a random or non-random variable?

- Given a loss function  $\ell : \mathcal{A} \times \mathcal{Y} \rightarrow \mathbf{R}$ .
- Choose hypothesis space  $\mathcal{F}$ .
- Use an optimization method to find ERM  $\hat{f}_n \in \mathcal{F}$ :

$$\hat{f}_n = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \ell(f(x_i), y_i).$$

- Data scientist's job:
  - choose  $\mathcal{F}$  to balance between approximation and estimation error.
  - as we get more training data, use a bigger  $\mathcal{F}$



- We've been cheating a bit by writing “argmin”.
- In practice, we need a method to find  $\hat{f}_n \in \mathcal{F}$ .
- For nice choices of loss functions and classes  $\mathcal{F}$ , we can get arbitrarily close to a minimizer
  - But takes time – is it worth it?
- For some hypothesis spaces (e.g. neural networks), we don't know how to find  $\hat{f}_n \in \mathcal{F}$ .

# Optimization Error

- In practice, we don't find the ERM  $\hat{f}_n \in \mathcal{F}$ .
- We find  $\tilde{f}_n \in \mathcal{F}$  that we hope is good enough.
- **Optimization error:** If  $\tilde{f}_n$  is the function our optimization method returns, and  $\hat{f}_n$  is the empirical risk minimizer, then

$$\text{Optimization Error} = R(\tilde{f}_n) - R(\hat{f}_n).$$

- Can optimization error be negative? Yes!
- But

$$\hat{R}(\tilde{f}_n) - \hat{R}(\hat{f}_n) \geq 0.$$

# Error Decomposition in Practice

- Excess risk decomposition for function  $\tilde{f}_n$  returned by algorithm:

$$\begin{aligned}\text{Excess Risk}(\tilde{f}_n) &= R(\tilde{f}_n) - R(f^*) \\ &= \underbrace{R(\tilde{f}_n) - R(\hat{f}_n)}_{\text{optimization error}} + \underbrace{R(\hat{f}_n) - R(f_{\mathcal{F}})}_{\text{estimation error}} + \underbrace{R(f_{\mathcal{F}}) - R(f^*)}_{\text{approximation error}}\end{aligned}$$

- Concept check: It would be nice to have a concrete example where we find an  $\tilde{f}_n$  and look at its error decomposition. Why is this usually impossible?
- But we could construct an artificial example, where we know  $P_{\mathcal{X} \times \mathcal{Y}}$  and  $f^*$  and  $f_{\mathcal{F}} \dots$