LECTURE 16

Cross-Validation and Regularization

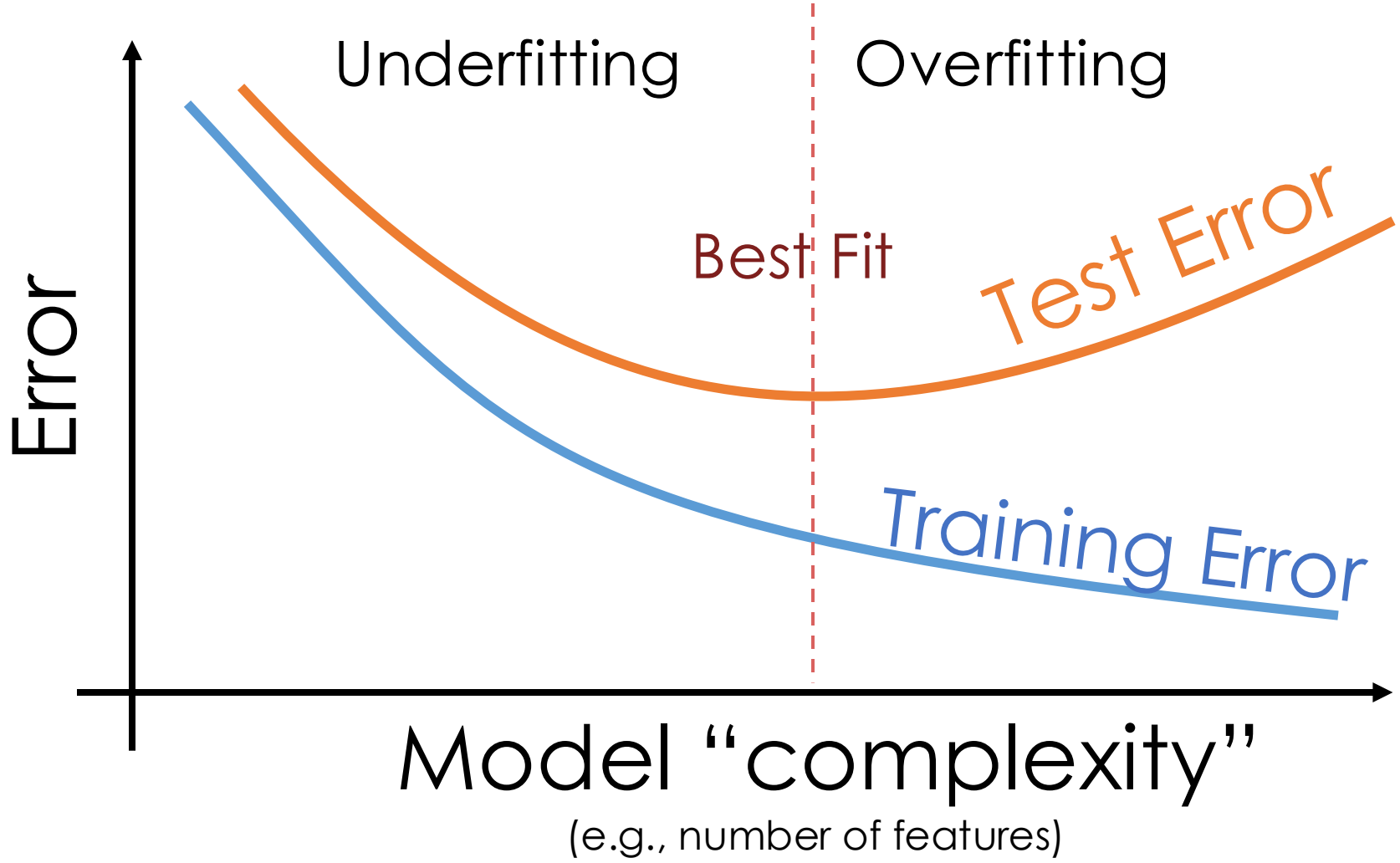
Different methods for ensuring the generalizability of our models to unseen data.

Sean Kang

Cross Validation

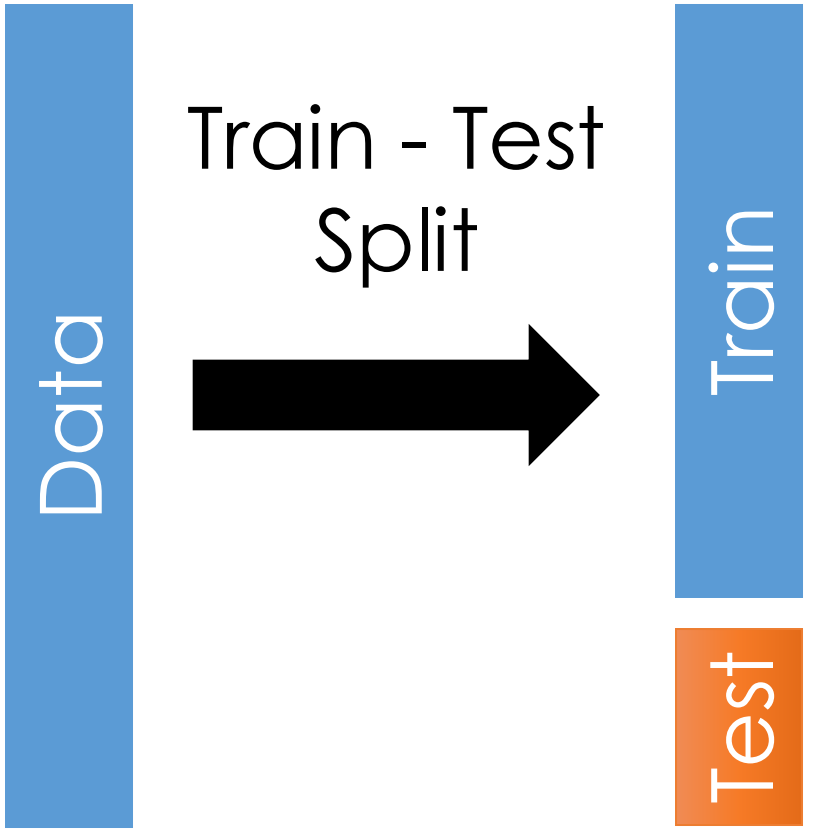
Training vs **Test** Error

Training error typically *under estimates* **test error**.



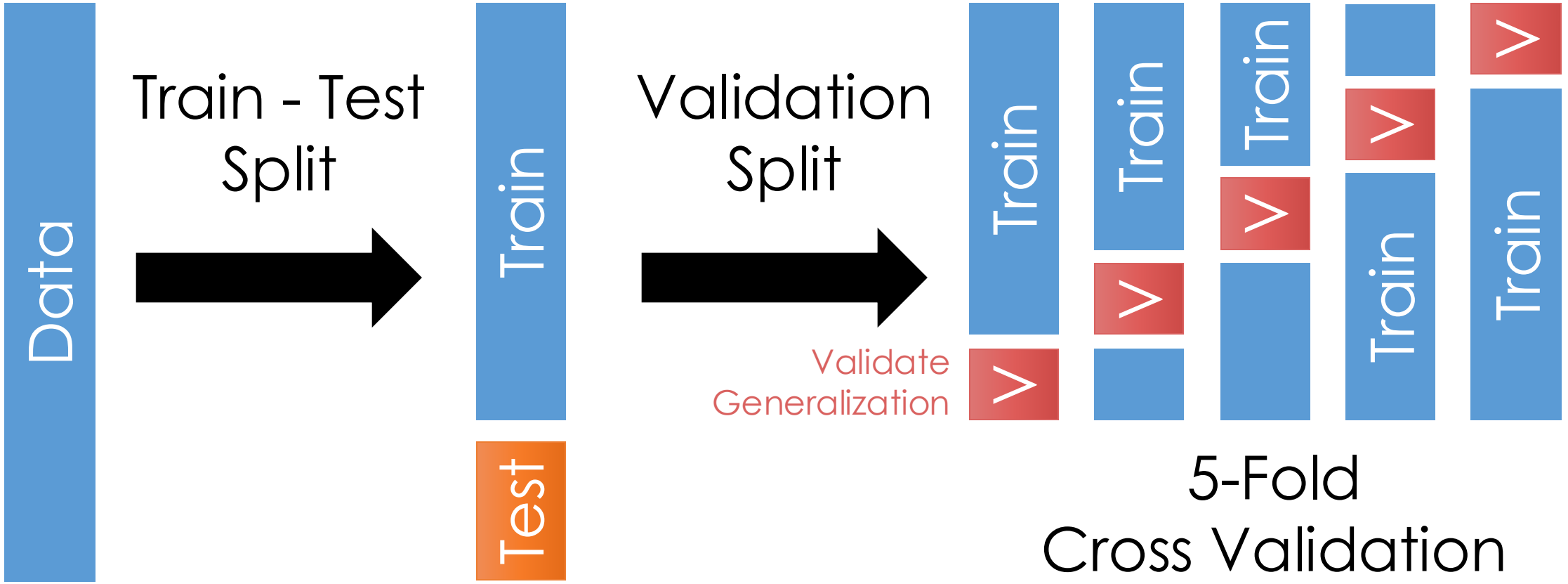
Generalization: *The Train-Test Split*

- **Training Data:** used to fit model
- **Test Data:** check generalization error
- How to split?
  - Randomly, Temporally, Geo…
  - Depends on application (usually randomly)
- What size? (90%-10%)
  - Larger training set – more complex models
  - Larger test set – better estimate of generalization error
  - Typically between 75%-25% and 90%-10%



You can only use the test dataset once after deciding on the model.

Generalization: *Validation Split*



*Cross validation **simulates multiple train test-splits** on the training data.*

Recipe for Successful Generalization

1. Split your data into **training** and **test** sets (90%, 10%)

2. Use **only the training data** when designing, training, and tuning the model
   - Use **cross validation** to test *generalization* during this phase
   - **Do not look at the test data**

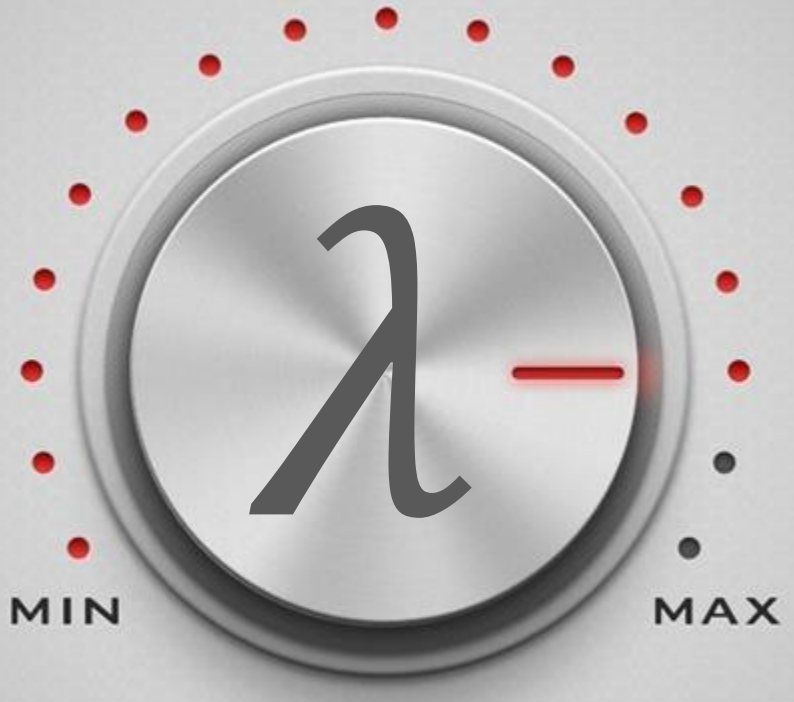
1. Commit to your final model and train once more using **only the training data.**

2. Test the final model using the **test data.** If accuracy is not acceptable return to (2). (*Get more test data if possible.*)

3. Train **on all available data** and ship it!

Regularization

Parametrically Controlling the *Model Complexity*



Basic Idea

- We add regularization to a model to reduce the overfitting of a model.
- This is a method of adding a penalty to the cost or loss function (RMSE, MSE remember?)
- This reduces the magnitude of the coefficient or weight to almost zero.
- The penalty can either be L1 or L2 norm.
- L1 norm is Lasso
- L2 norm is Ridge
- This will increase bias but reduce variance

Bias - Variance Trade Off

* Trade off between the error of a complex model versus the error due to the model's sensitivity to data

High Bias Model – Too simple and cannot capture the patterns of the data.

High Variance Model – Too complex and fits the training data too well thus leads to high variance error on test data

Ideal Model – Low variance and low bias that can handle generic new data.

More about Regularization

It reduces (to almost zero) the coefficients of the irrelevant variables

Side effect – it increases the bias

How to know the level of regularization to use?

- Use k-fold cross validation to find the optimal level

Ridge Regression

"Ridge Regression" is a term for the following specific combination of model, loss, and regularization:

- Model: $\hat{\mathbb{Y}} = \mathbb{X}\theta$
- Loss: Squared loss
- Regularization: L2 regularization

The **objective function** we minimize for Ridge Regression is average squared loss, plus an added penalty:

$$\hat{\theta}_{\text{ridge}} = \arg\min_{\theta} \frac{1}{n} ||\mathbb{Y} - \mathbb{X}\theta||_2^2 + \lambda \sum_{j=1}^{d} \theta_i^2$$

Ridge Regression

Also called L2 Regression

Adds the square of the coefficients

LASSO Regression

"LASSO Regression" is a term for the following specific combination of model, loss, and regularization:

- Model: $\hat{\mathbb{Y}} = \mathbb{X}\theta$
- Loss: Squared loss
- Regularization: L1 regularization

The **objective function** we minimize for LASSO Regression is average squared loss, plus an added penalty:

$$\hat{\theta}_{\text{LASSO}} = \arg\min_{\theta} \frac{1}{n} ||\mathbb{Y} - \mathbb{X}\theta||_2^2 + \lambda \sum_{j=1}^{d} |\theta_i|$$

LASSO Regression

LASSO – Least Absolute Shrinkage Selector Operator

Adds the Sum of the absolute value of the coefficients

Summary of Regression Methods

Name	Model	Loss	Reg.	Objective	Solution
Ridge Regression	$\hat{\mathbb{Y}} = \mathbb{X}\theta$	Squared loss	L2	$\frac{1}{n}\|\|\mathbb{Y} - \mathbb{X}\theta\|\|_2^2 + \lambda \sum_{j=1}^{d} \theta_i^2$	$\hat{\theta}_{\text{ridge}} = (\mathbb{X}^T\mathbb{X} + n\lambda I)^{-1}\mathbb{X}^T\mathbb{Y}$
LASSO	$\hat{\mathbb{Y}} = \mathbb{X}\theta$	Squared loss	L1	$\frac{1}{n}\|\|\mathbb{Y} - \mathbb{X}\theta\|\|_2^2 + \lambda \sum_{j=1}^{d} \|\theta_i\|$	No closed form

Fitting vs. Evaluating

While we may use a regularized objective function to determine our model's parameters, we still look at **(root) mean squared error** to evaluate our model's performance.

Which is better? LASSO or Ridge?

They are both similar but LASSO can produce coefficients that are exactly zero which provides better feature selection, in most situations. LASSO works well when there are smaller number of significant parameters, and others are not.