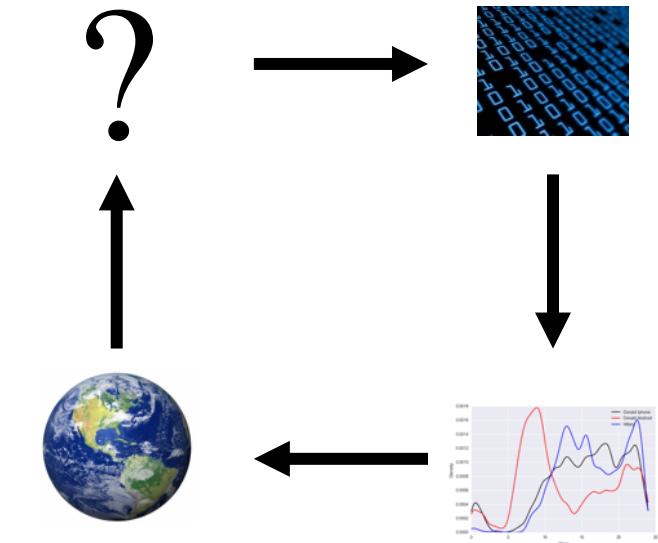


The Bias Variance Tradeoff and Regularization

Slides by:

Joseph E. Gonzalez

jegonzal@cs.berkeley.edu



Recap: Least Squares Regression

- Least squares regression:

$$\hat{\theta} = \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^n \underbrace{(y_i - f_{\theta}(x))^2}_{\text{Loss Function}}$$

- Generic framework
- Linear models:

$$f_{\theta}(x) = \sum_{j=1}^p \theta_j x_j$$

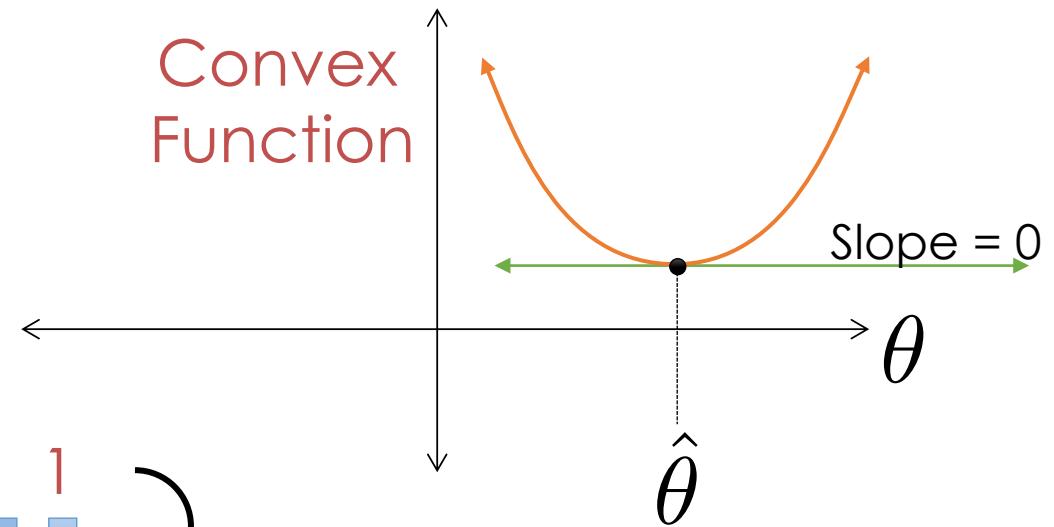
- Linear in the parameters

➤ Linear models:

$$f_{\theta}(x) = \sum_{j=1}^p \theta_j x_j$$

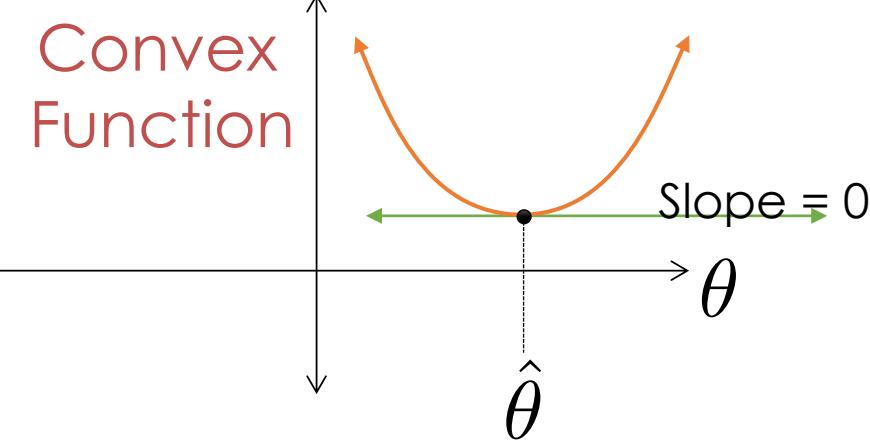
- Linear in the parameters
- Finding Optimal Parameters
 - Normal Equations:

$$\hat{\theta} = \left(\begin{matrix} n & p \\ X^T & X \end{matrix} \right)^{-1} \left(\begin{matrix} n & 1 \\ p & X^T \\ Y \end{matrix} \right)$$

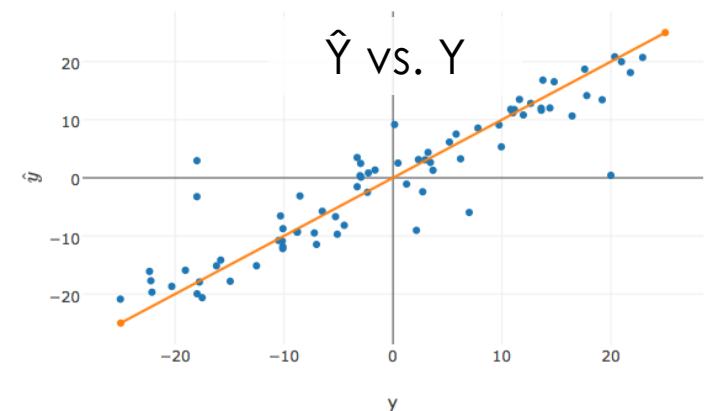
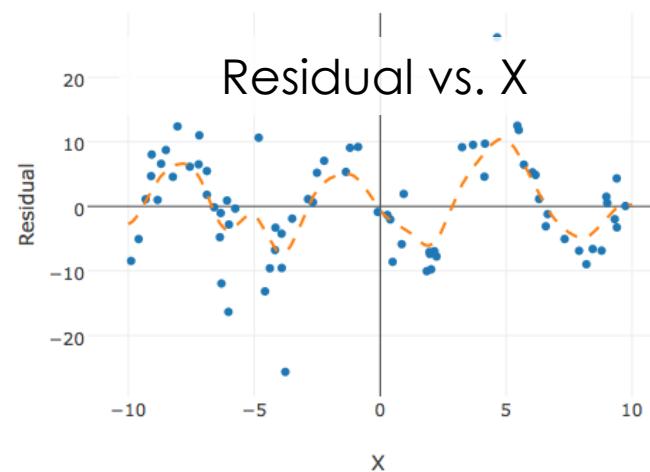
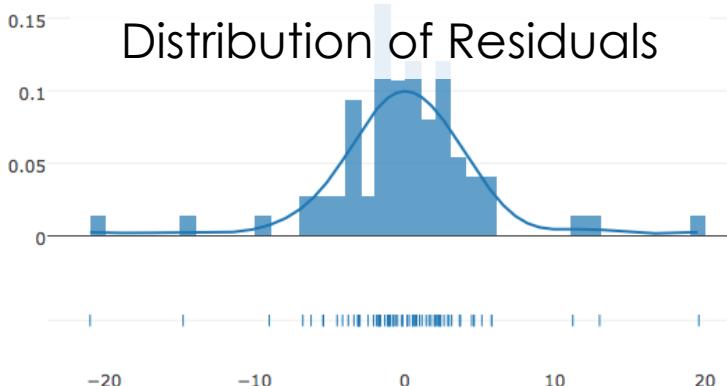


- Finding Optimal Parameters
- Normal Equations:

$$\hat{\theta} = p = \begin{pmatrix} n & p \\ X^T & X \end{pmatrix}^{-1} \begin{pmatrix} n & 1 \\ p & X^T & Y \end{pmatrix}$$



- Using software packages (e.g., linalg.solve, sklearn.linear_models)
- Diagnostics:

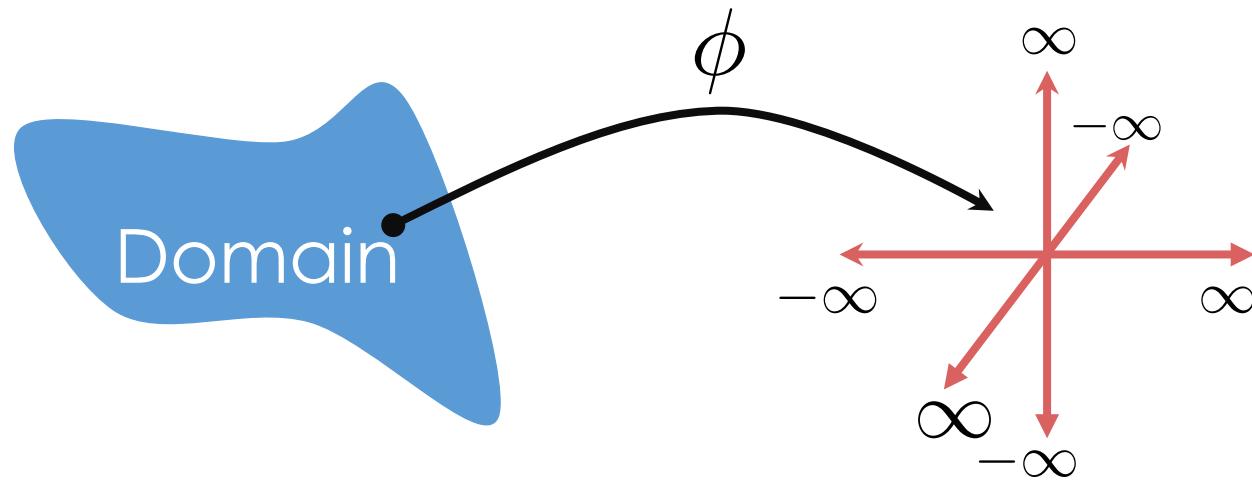


Recap: Feature Engineering

- Linear models with feature functions:

$$f_{\theta}(x) = \sum_{j=1}^p \theta_j \phi_j(x)$$

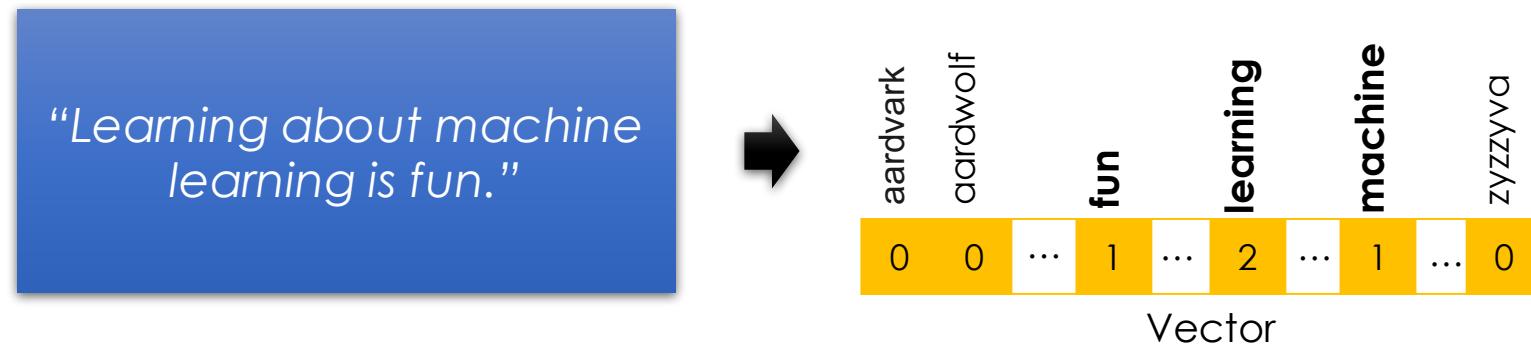
- Feature Functions: $\phi : \mathcal{X} \rightarrow \mathbb{R}^p$



➤ One-hot encoding: Categorical Data

| state | AL | ... | CA | ... | NY | ... | WA | ... | WY |
|-------|----|-----|----|-----|----|-----|----|-----|----|
| NY | 0 | ... | 0 | ... | 1 | ... | 0 | ... | 0 |
| WA | 0 | ... | 0 | ... | 0 | ... | 1 | ... | 0 |
| CA | 0 | ... | 1 | ... | 0 | ... | 0 | ... | 0 |

➤ Bag-of-words & N-gram: Text Data



➤ Custom Features: Domain Knowledge

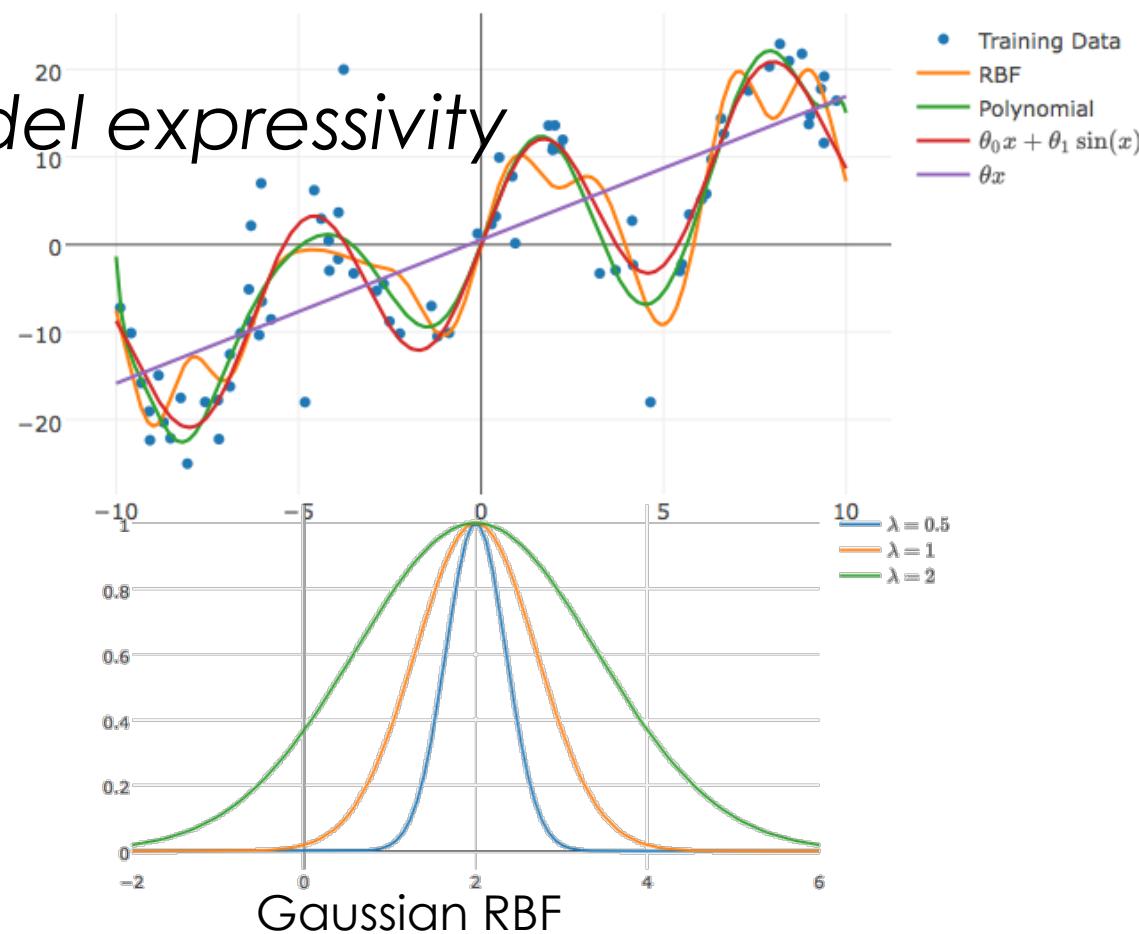
$$\phi(\text{lat}, \text{lon}, \text{amount}) = \frac{\text{amount}}{\text{Stores}[\text{ZipCode}[\text{lat}, \text{lon}]]}$$

- **Generic Features:** increase model expressivity
 - Polynomial and trigonometric:

$$\phi(x) = [x, x^2, \dots, x^p]$$

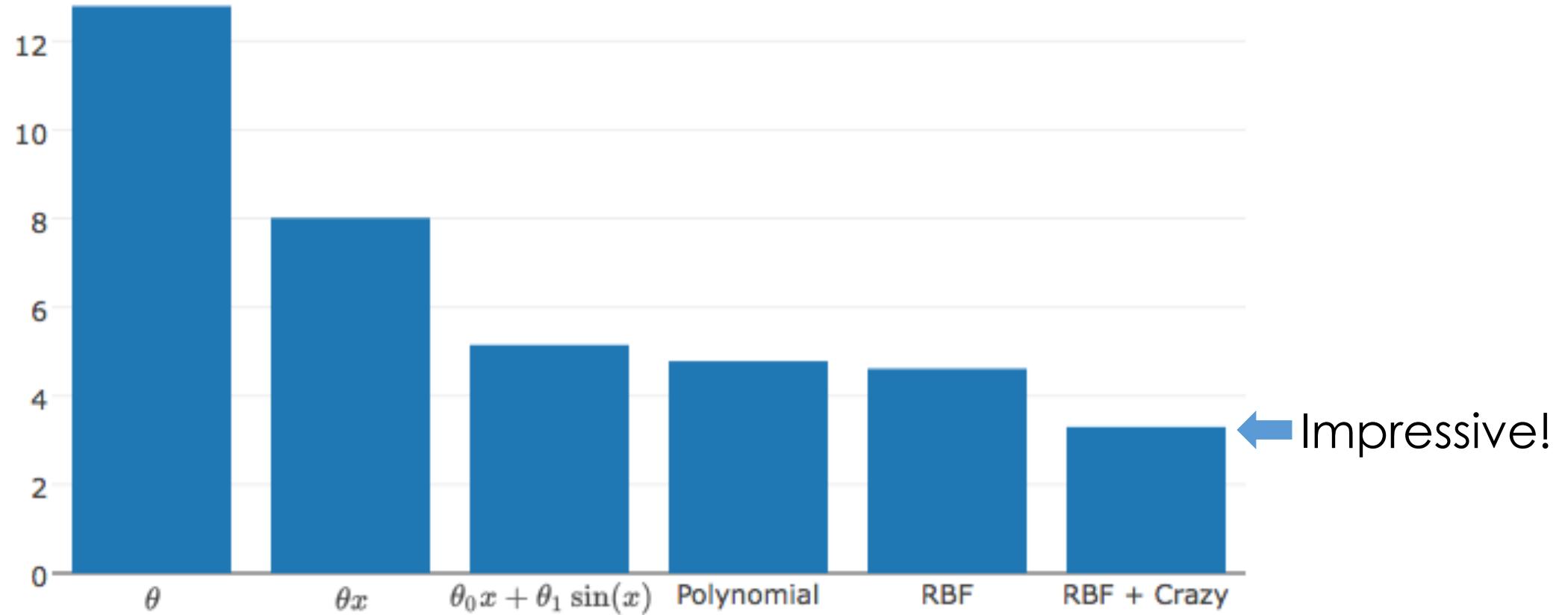
- **Gaussian Radial Basis Functions:**

$$\phi_{\lambda_i, \mu_i}(x) = \exp\left(-\frac{\|x - \mu_i\|_2^2}{\lambda_i}\right)$$

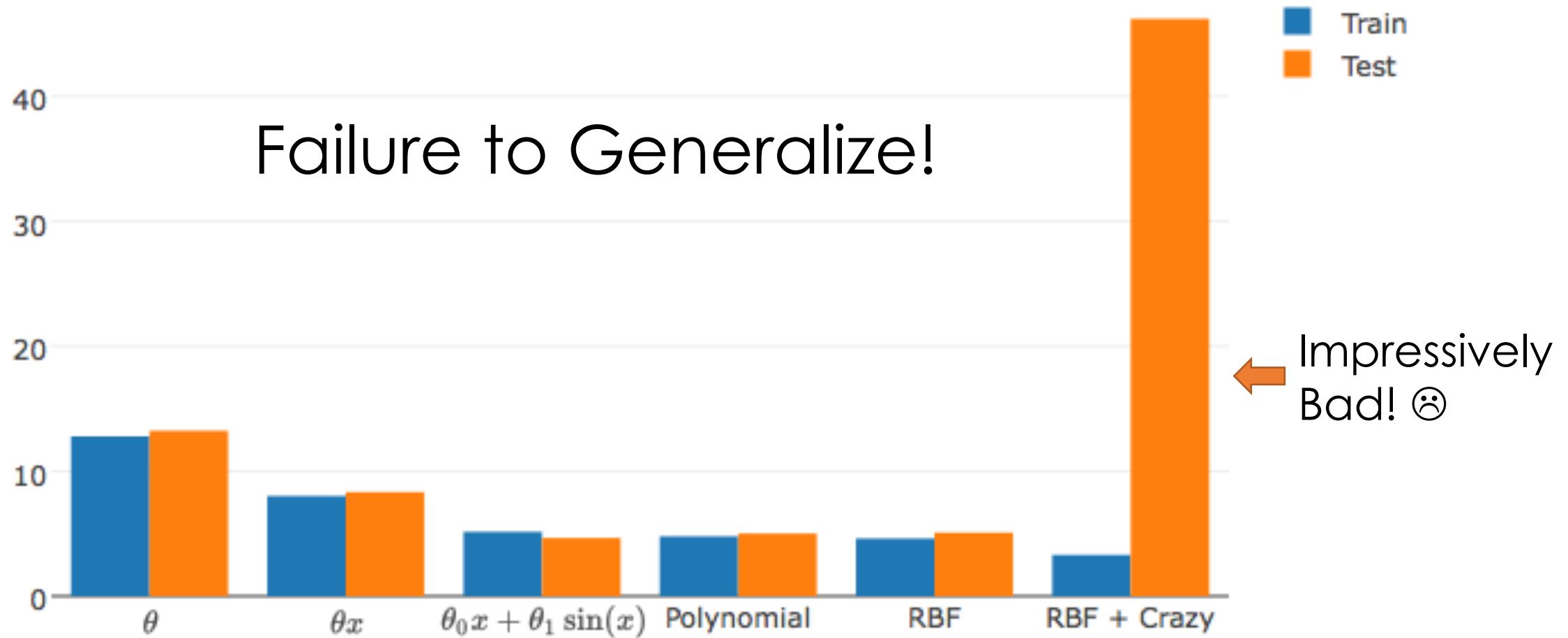


- **Hyper-parameters:** parameters that are not “learned” but instead selected using prior knowledge or through cross-validation.
 - Example: Poly. deg., μ_i , and σ_i

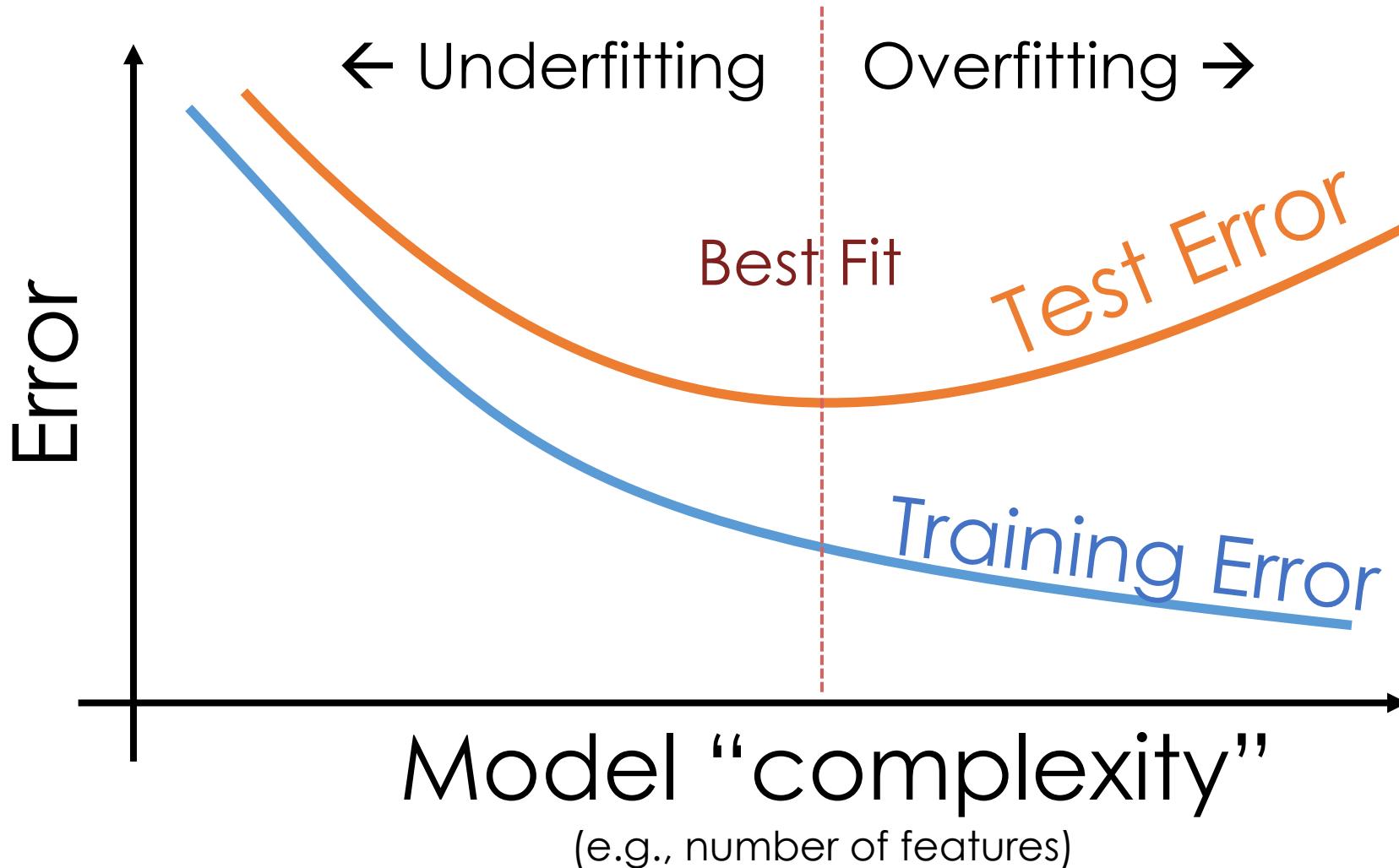
Training Error



Training vs Test Error

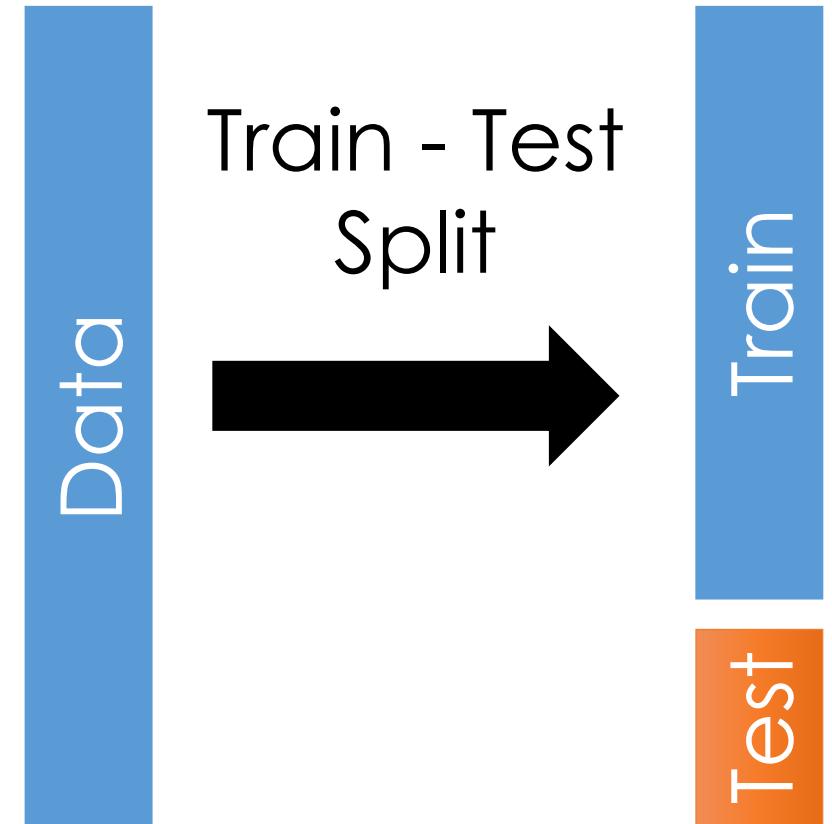


Training vs Test Error



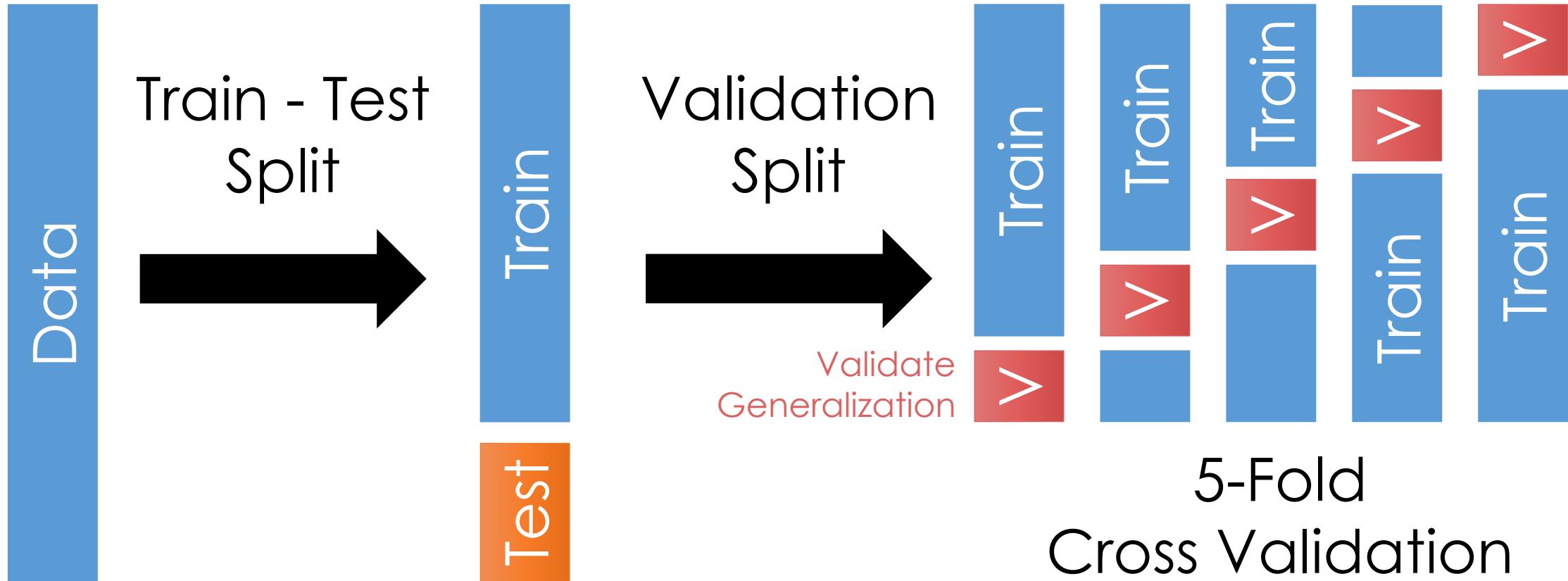
Generalization: The Train-Test Split

- **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.

Python Demo!

Regularization Notebook Part 1

Fundamental Challenges of Prediction

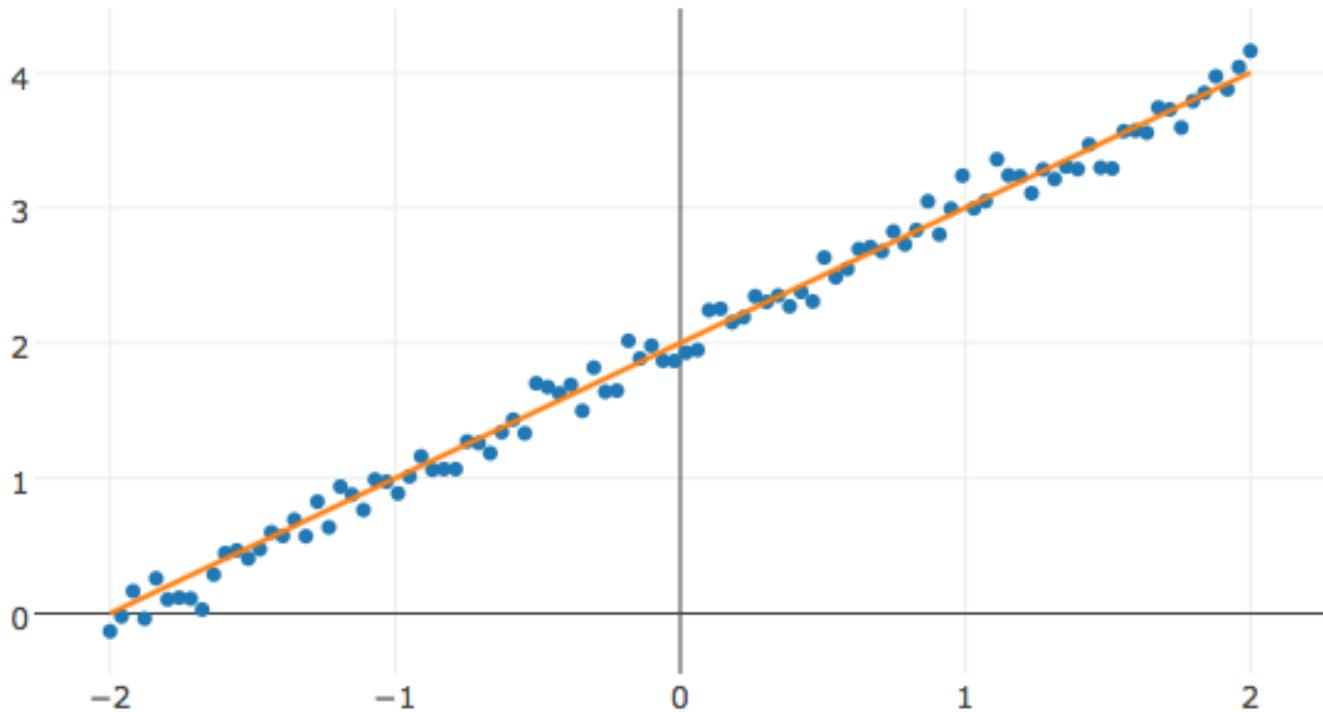
- **Noise:** the intrinsic variability in the process we are trying to model
- **Bias:** the expected deviation between the predicted value and the true value
- **Variance:** variability between the estimated value and the true value across different training datasets

Noise

The intrinsic variability in the process we are trying to predict

- measurement variability
- stochasticity
- missing information

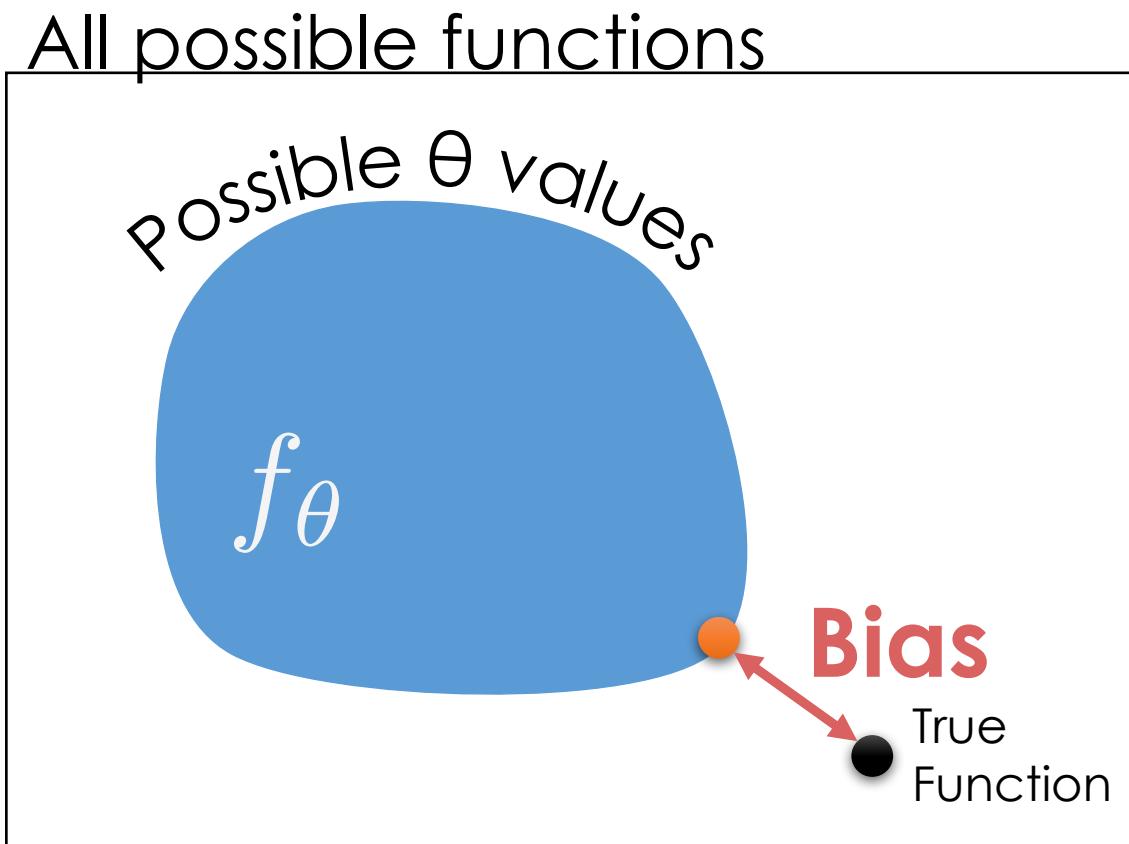
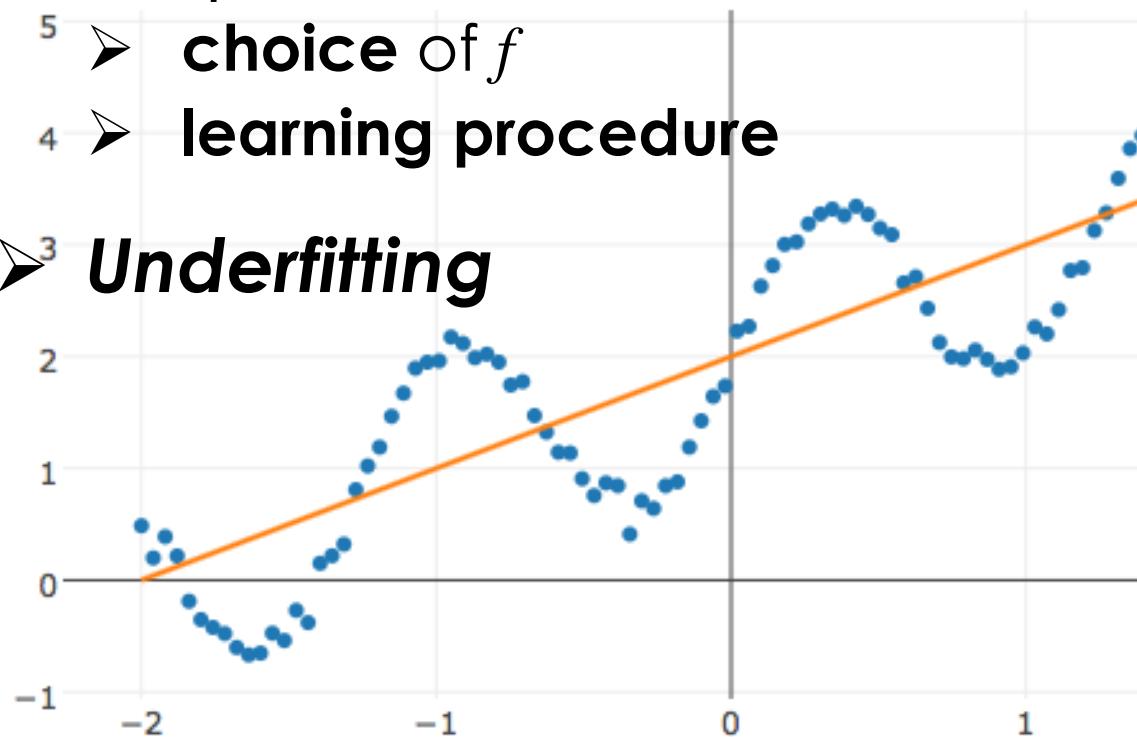
**Beyond our control
(usually)**



Bias

The expected deviation between the predicted value and the true value

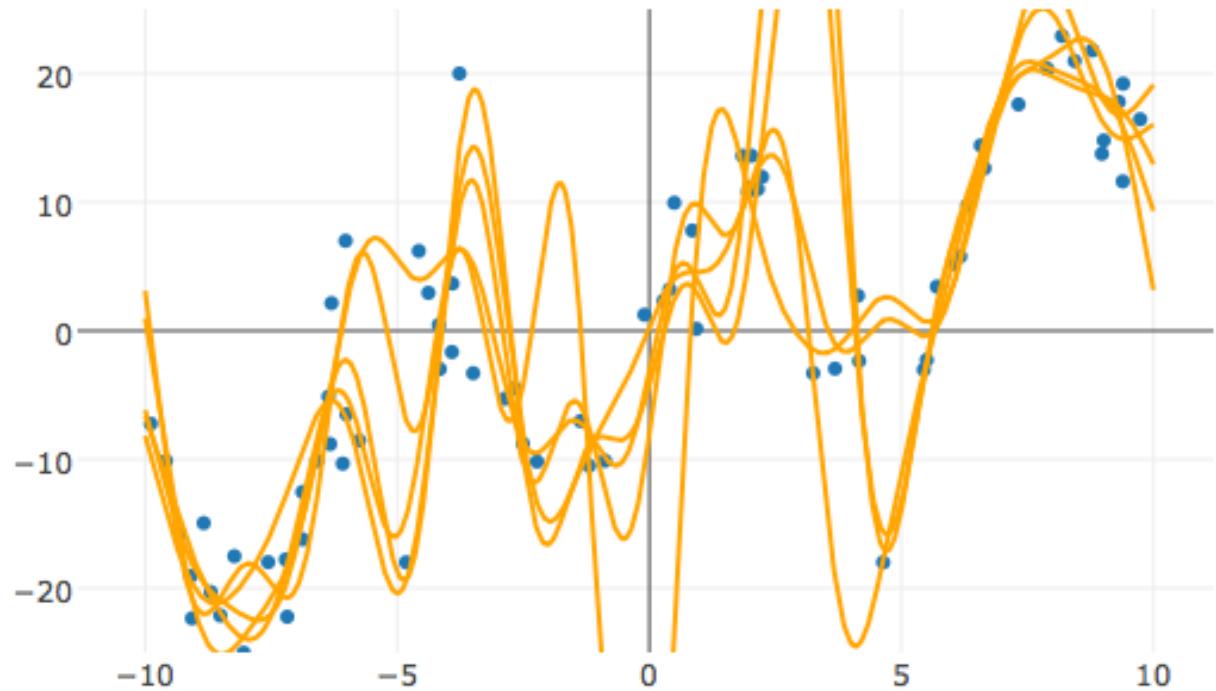
- Depends on both the:
 - choice of f
 - learning procedure
- **Underfitting**



Variance

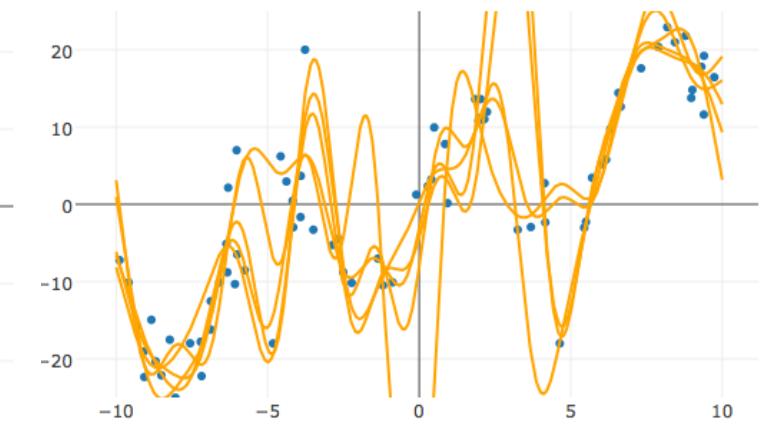
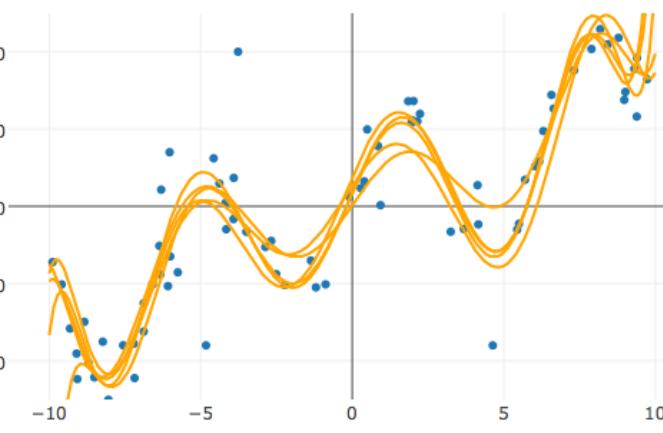
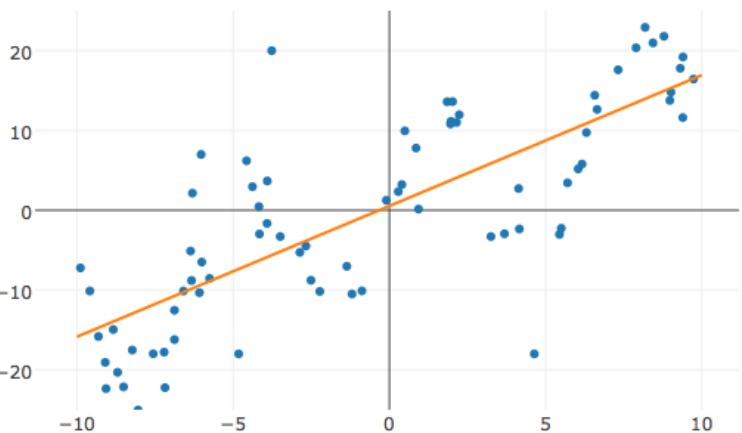
variability between the estimated value and the true value across different training datasets

- Sensitivity to variation in the training data
- Poor generalization
- **Overfitting**



The Bias-Variance Tradeoff

Variance



Bias

Analysis of Squared Error

- For the test point x the expected error:
 - Random variables are **red**

Assume noisy observations
→ y is a random variable

$$\mathbf{E} \left[(y - f_{\theta}(x))^2 \right]$$

True Function

Noise term:
 $\mathbf{E} [\epsilon] = 0$

Assume **training data** is random
→ θ is a random variable

Analysis of Squared Error

Goal:

$$\mathbf{E} \left[(y - f_{\theta}(x))^2 \right] =$$

Noise + (Bias)² + Variance

$$\mathbf{E} \left[(y - f_{\theta}(x))^2 \right] = \mathbf{E} \left[\underbrace{(y - h(\textcolor{red}{x}) + h(\textcolor{red}{x}) - f_{\theta}(\textcolor{red}{x}))^2}_{\text{Subtracting and adding } h(\textcolor{red}{x})} \right]$$

Useful Eqns:

$$y = h(\textcolor{red}{x}) + \epsilon$$
$$\mathbf{E} [\epsilon] = 0$$

$$\mathbf{E} \left[(y - f_{\theta}(x))^2 \right] = \mathbf{E} \left[\underbrace{(y - h(x) + h(x) - f_{\theta}(x))^2}_{a+b} \right]$$

Expanding in terms of a and b : $(a + b)^2 = a^2 + b^2 + 2ab$

$$\begin{aligned} &= \mathbf{E} \left[(y - h(x))^2 \right] + \mathbf{E} \left[(h(x) - f_{\theta}(x))^2 \right] \\ &\quad + 2\mathbf{E} [(y - h(x))(h(x) - f_{\theta}(x))] \end{aligned}$$

$2ab$

Useful Eqns:

$$y = h(x) + \epsilon$$

$$\mathbf{E} [\epsilon] = 0$$

Expanding Terms

$$+ 2\mathbf{E} [yh(x) - yf_{\theta}(x) - h(x)h(x) + h(x)f_{\theta}(x)]$$

$$\mathbf{E} \left[(y - f_{\theta}(x))^2 \right] = \mathbf{E} \left[(y - h(x) + h(x) - f_{\theta}(x))^2 \right]$$

$$= \mathbf{E} \left[(y - h(x))^2 \right] + \mathbf{E} \left[(h(x) - f_{\theta}(x))^2 \right]$$

Useful Eqns:

| |
|-----------------------------|
| $y = h(x) + \epsilon$ |
| $\mathbf{E} [\epsilon] = 0$ |

$$+ 2\mathbf{E} [yh(x) - yf_{\theta}(x) - h(x)h(x) + h(x)f_{\theta}(x)]$$

$$\underbrace{\hspace{10em}}_{y = h(x) + \epsilon}$$

$$(h(x) + \epsilon)h(x) - (h(x) + \epsilon)f_{\theta}(x)$$

$$= h(x)h(x) + \epsilon h(x) - h(x)f_{\theta}(x) - \epsilon f_{\theta}(x)$$

$$\mathbf{E} \left[(y - f_{\theta}(x))^2 \right] = \mathbf{E} \left[(y - h(x) + h(x) - f_{\theta}(x))^2 \right]$$

$$= \mathbf{E} \left[(y - h(x))^2 \right] + \mathbf{E} \left[(h(x) - f_{\theta}(x))^2 \right]$$

Useful Eqns:

$$y = h(x) + \epsilon$$

$$\mathbf{E} [\epsilon] = 0$$

$$+ 2\mathbf{E} [yh(x) - yf_{\theta}(x) - h(x)h(x) + h(x)f_{\theta}(x)]$$

$\underbrace{\hspace{10em}}_{y = h(x) + \epsilon}$

$$= h(x)h(x) + \boxed{\epsilon h(x)} - h(x)f_{\theta}(x) - \boxed{\epsilon f_{\theta}(x)}$$

$$+ 2\mathbf{E} [\epsilon h(x) - \epsilon f_{\theta}(x)]$$

$$\mathbf{E} \left[(y - f_{\theta}(x))^2 \right] = \mathbf{E} \left[(y - h(x) + h(x) - f_{\theta}(x))^2 \right]$$

$$= \mathbf{E} \left[(y - h(x))^2 \right] + \mathbf{E} \left[(h(x) - f_{\theta}(x))^2 \right]$$

Useful Eqns:

$$y = h(x) + \epsilon$$

$$\mathbf{E} [\epsilon] = 0$$

$$+ 2 \underbrace{\mathbf{E} [\epsilon h(x) - \epsilon f_{\theta}(x)]}$$

$$\mathbf{E} [\epsilon h(x) - \epsilon f_{\theta}(x)] = \mathbf{E} [\epsilon (h(x) - f_{\theta}(x))]$$

Independence

$$= \mathbf{E} [\epsilon] \mathbf{E} [(h(x) - f_{\theta}(x))]$$

Noise definition
(Useful eqns.)

$$= 0$$

$$\mathbf{E} \left[(y - f_{\theta}(x))^2 \right] =$$

$$\mathbf{E} \left[(y - h(x))^2 \right] +$$

Obs. Value

True Value

Noise Term

$$\mathbf{E} \left[(h(x) - f_{\theta}(x))^2 \right]$$

True Value

Pred. Value

Model
Estimation
Error

$$\mathbf{E} \left[(h(\textcolor{red}{x}) - f_{\theta}(\textcolor{red}{x}))^2 \right] = \text{ Next we will show....}$$

$$\mathbf{E} \left[(h(\textcolor{red}{x}) - \mathbf{E}_{\theta} [f_{\theta}(x)])^2 \right] + \mathbf{E} \left[(\mathbf{E}_{\theta} [f_{\theta}(\textcolor{red}{x})] - f_{\theta}(x))^2 \right]$$

- How?
- Adding and Subtracting what?

Useful Eqns:

$$\textcolor{red}{y} = h(\textcolor{red}{x}) + \epsilon$$

$$\mathbf{E} [\epsilon] = 0$$

$$\mathbf{E} \left[(h(\textcolor{red}{x}) - f_{\theta}(\textcolor{red}{x}))^2 \right] =$$

$$\mathbf{E} \left[(h(\textcolor{red}{x}) - \mathbf{E}_{\theta} [f_{\theta}(\textcolor{red}{x})] + \mathbf{E}_{\theta} [f_{\theta}(\textcolor{red}{x})] - f_{\theta}(\textcolor{red}{x}))^2 \right]$$

}

Adding and subtracting $\mathbf{E}_{\theta} [f_{\theta}(\textcolor{red}{x})]$

- Note expectation is over θ
- Random variable in x

Useful Eqns:

$$\textcolor{red}{y} = h(\textcolor{red}{x}) + \epsilon$$

$$\mathbf{E} [\epsilon] = 0$$

Useful Eqns:

$$y = h(x) + \epsilon$$

$$\mathbf{E}[\epsilon] = 0$$

$$\mathbf{E} \left[(h(\mathbf{x}) - f_{\theta}(\mathbf{x}))^2 \right] =$$

$$\mathbf{E} \left[\underbrace{(h(\mathbf{x}) - \mathbf{E}_{\theta}[f_{\theta}(\mathbf{x})])^2}_{a} + \underbrace{\mathbf{E}_{\theta}[f_{\theta}(\mathbf{x})] - f_{\theta}(\mathbf{x})}_{b}^2 \right]$$

Expanding in terms of a and b : $(a + b)^2 = a^2 + b^2 + 2ab$

$$\begin{aligned} & \mathbf{E} \left[(h(\mathbf{x}) - \mathbf{E}_{\theta}[f_{\theta}(\mathbf{x})])^2 \right] + \mathbf{E} \left[(\mathbf{E}_{\theta}[f_{\theta}(\mathbf{x})] - f_{\theta}(\mathbf{x}))^2 \right] \\ & + 2\mathbf{E} \left[(h(\mathbf{x}) - \mathbf{E}_{\theta}[f_{\theta}(\mathbf{x})]) (\mathbf{E}_{\theta}[f_{\theta}(\mathbf{x})] - f_{\theta}(\mathbf{x})) \right] \\ & \quad 2ab \end{aligned}$$

Useful Eqns:

$$y = h(x) + \epsilon$$

$$\mathbf{E}[\epsilon] = 0$$

$$\mathbf{E} [(h(x) - f_{\theta}(x))^2] =$$

$$\mathbf{E} [(h(x) - \mathbf{E}_{\theta}[f_{\theta}(x)])^2] + \mathbf{E} [(\mathbf{E}_{\theta}[f_{\theta}(x)] - f_{\theta}(x))^2]$$

$$+ 2\mathbf{E} [(h(x) - \mathbf{E}_{\theta}[f_{\theta}(x)]) (\mathbf{E}_{\theta}[f_{\theta}(x)] - f_{\theta}(x))]$$

Expanding Terms and applying linearity of expectation

$$+ 2\mathbf{E} [h(x)\mathbf{E}_{\theta}[f_{\theta}(x)]] - 2\mathbf{E} [h(x)f_{\theta}(x)]$$

$$- 2\mathbf{E} [\mathbf{E}_{\theta}[f_{\theta}(x)] \mathbf{E}_{\theta}[f_{\theta}(x)]] + 2\mathbf{E} [\mathbf{E}_{\theta}[f_{\theta}(x)] f_{\theta}(x)]$$

Useful Eqns:

$$y = h(x) + \epsilon$$

$$\mathbf{E}[\epsilon] = 0$$

$$\mathbf{E} \left[(h(x) - f_{\theta}(x))^2 \right] =$$

$$\mathbf{E} \left[(h(x) - \mathbf{E}_{\theta} [f_{\theta}(x)])^2 \right] + \mathbf{E} \left[(\mathbf{E}_{\theta} [f_{\theta}(x)] - f_{\theta}(x))^2 \right]$$

$$+ 2\mathbf{E} [h(x)\mathbf{E}_{\theta} [f_{\theta}(x)]] - 2\mathbf{E} [h(x)f_{\theta}(x)]$$

$$- 2\mathbf{E} [\mathbf{E}_{\theta} [f_{\theta}(x)] \mathbf{E}_{\theta} [f_{\theta}(x)]] + 2\mathbf{E} [\mathbf{E}_{\theta} [f_{\theta}(x)] f_{\theta}(x)]$$

Useful Eqns:

$$y = h(x) + \epsilon$$

$$\mathbf{E}[\epsilon] = 0$$

$$\mathbf{E} [(h(x) - f_\theta(x))^2] =$$

$$\mathbf{E} [(h(x) - \mathbf{E}_\theta [f_\theta(x)])^2] + \mathbf{E} [(\mathbf{E}_\theta [f_\theta(x)] - f_\theta(x))^2]$$

$$+ 2\mathbf{E} [h(x)\mathbf{E}_\theta [f_\theta(x)]] - 2\mathbf{E} [h(x)f_\theta(x)]$$

$$\int_x \int_\theta h(x)f_\theta(x)p(\theta, x)d\theta dx =$$
$$\int_x h(x) \int_\theta f_\theta(x)p(\theta, x)d\theta dx$$

$$- 2\mathbf{E} [h(x)\mathbf{E}_\theta [f_\theta(x)]]$$

$h(x)$ does not
depend on the θ

$$- 2\mathbf{E} [\mathbf{E}_\theta [f_\theta(x)] \mathbf{E}_\theta [f_\theta(x)]] + 2\mathbf{E} [\mathbf{E}_\theta [f_\theta(x)] f_\theta(x)]$$

Useful Eqns:

$$y = h(x) + \epsilon$$

$$\mathbf{E}[\epsilon] = 0$$

$$\mathbf{E} \left[(h(x) - f_{\theta}(x))^2 \right] =$$

$$\mathbf{E} \left[(h(x) - \mathbf{E}_{\theta} [f_{\theta}(x)])^2 \right] + \mathbf{E} \left[(\mathbf{E}_{\theta} [f_{\theta}(x)] - f_{\theta}(x))^2 \right]$$

$$-2\mathbf{E} [\mathbf{E}_{\theta} [f_{\theta}(x)] \mathbf{E}_{\theta} [f_{\theta}(x)]] + 2\mathbf{E} [\mathbf{E}_{\theta} [f_{\theta}(x)] f_{\theta}(x)]$$

$\mathbf{E}_{\theta} [f_{\theta}(x)]$ does not
depend on the θ

$$+ 2\mathbf{E} [\mathbf{E}_{\theta} [f_{\theta}(x)] \mathbf{E}_{\theta} [f_{\theta}(x)]]$$

$$\mathbf{E} \left[(h(\textcolor{red}{x}) - f_{\theta}(\textcolor{red}{x}))^2 \right] =$$

$$\mathbf{E} \left[(h(\textcolor{red}{x}) - \mathbf{E}_{\theta} [f_{\theta}(\textcolor{red}{x})])^2 \right] + \mathbf{(Bias)^2 Term}$$

$$\mathbf{E} \left[(\mathbf{E}_{\theta} [f_{\theta}(\textcolor{red}{x})] - f_{\theta}(\textcolor{red}{x}))^2 \right]$$

Variance Term

$$\mathbf{E} \left[(y - f_{\theta}(x))^2 \right] =$$

$$\mathbf{E} \left[(y - h(x))^2 \right] +$$

Noise Term

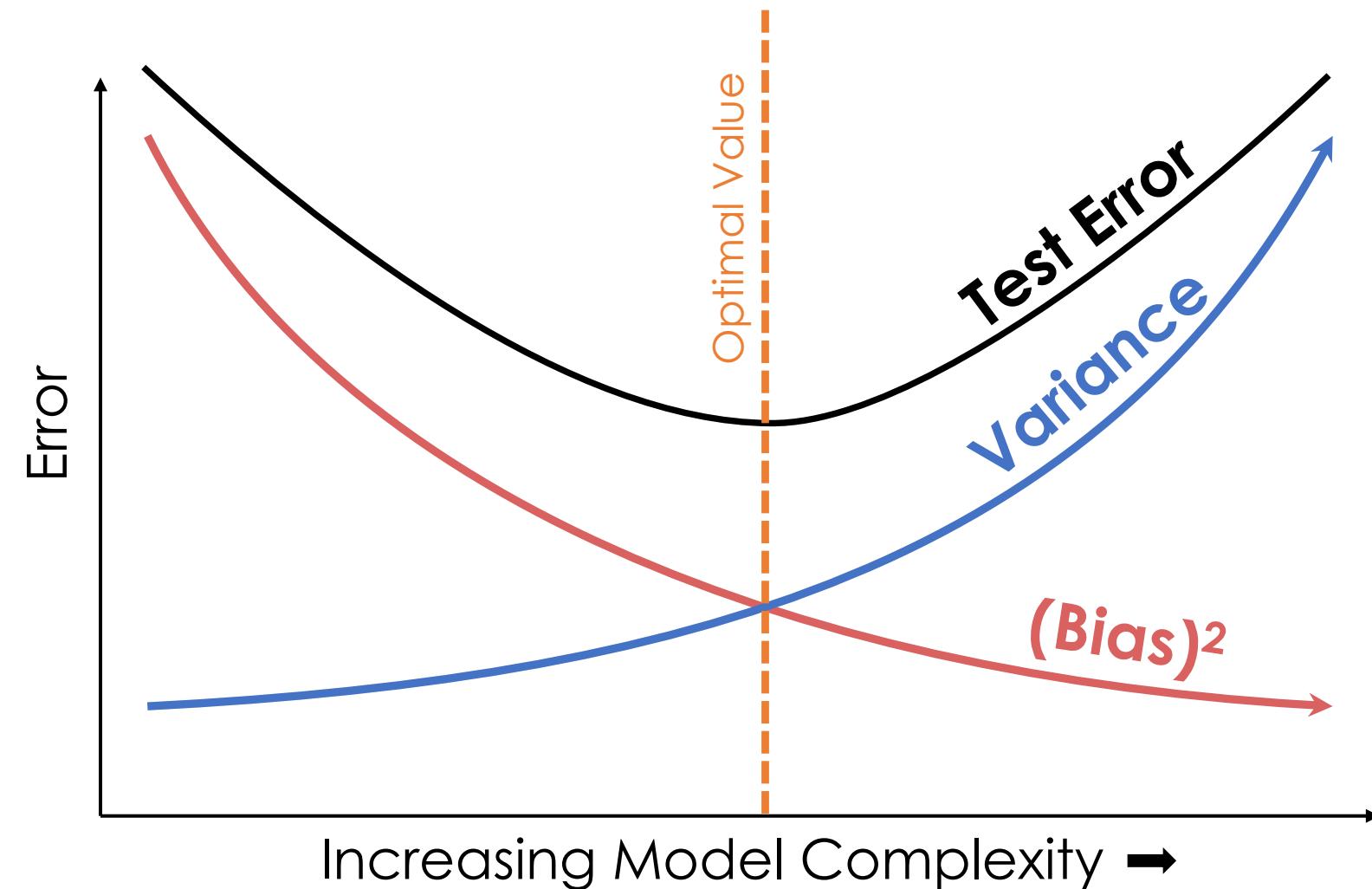
$$\mathbf{E} \left[(h(x) - \mathbf{E}_{\theta} [f_{\theta}(x)])^2 \right] +$$

(Bias)² Term

$$\mathbf{E} \left[(\mathbf{E}_{\theta} [f_{\theta}(x)] - f_{\theta}(x))^2 \right]$$

Variance Term

Bias Variance Plot

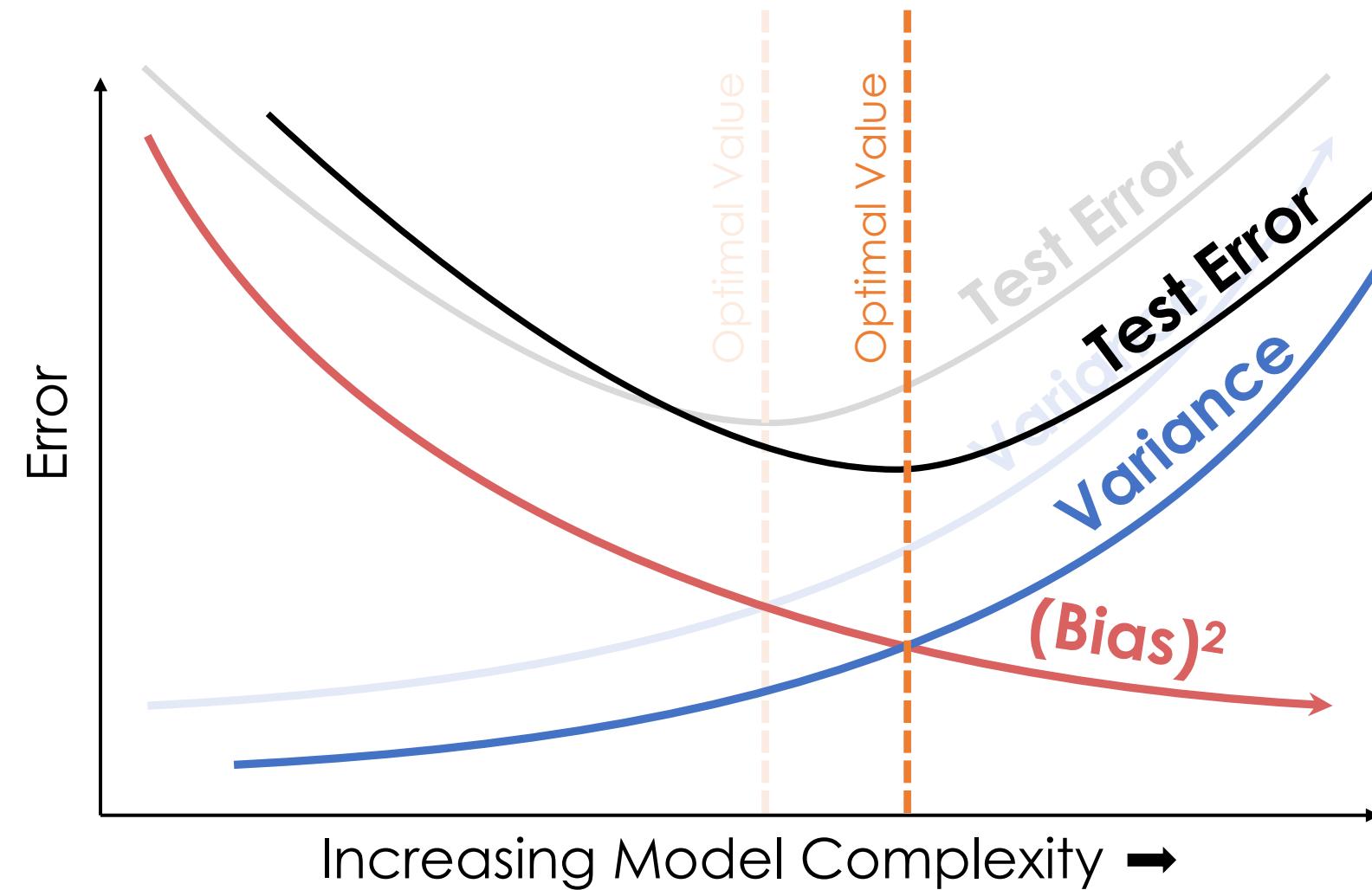


How do we compute test error?

- Use test set to pick best model?
 - No no no nooooo!
- Cross validation?
 - Yes!!!!
- What happens with more data?

$$\mathbf{E} [(y - h(x))^2] \quad \mathbf{Noise}$$
$$\mathbf{E} [(h(x) - \mathbf{E}[f_{\theta}(x)])^2] \quad (\mathbf{Bias})^2$$
$$\mathbf{E} [(\mathbf{E}[f_{\theta}(x)] - f_{\theta}(x))^2] \quad \mathbf{Variance}$$

Bias Variance Plot

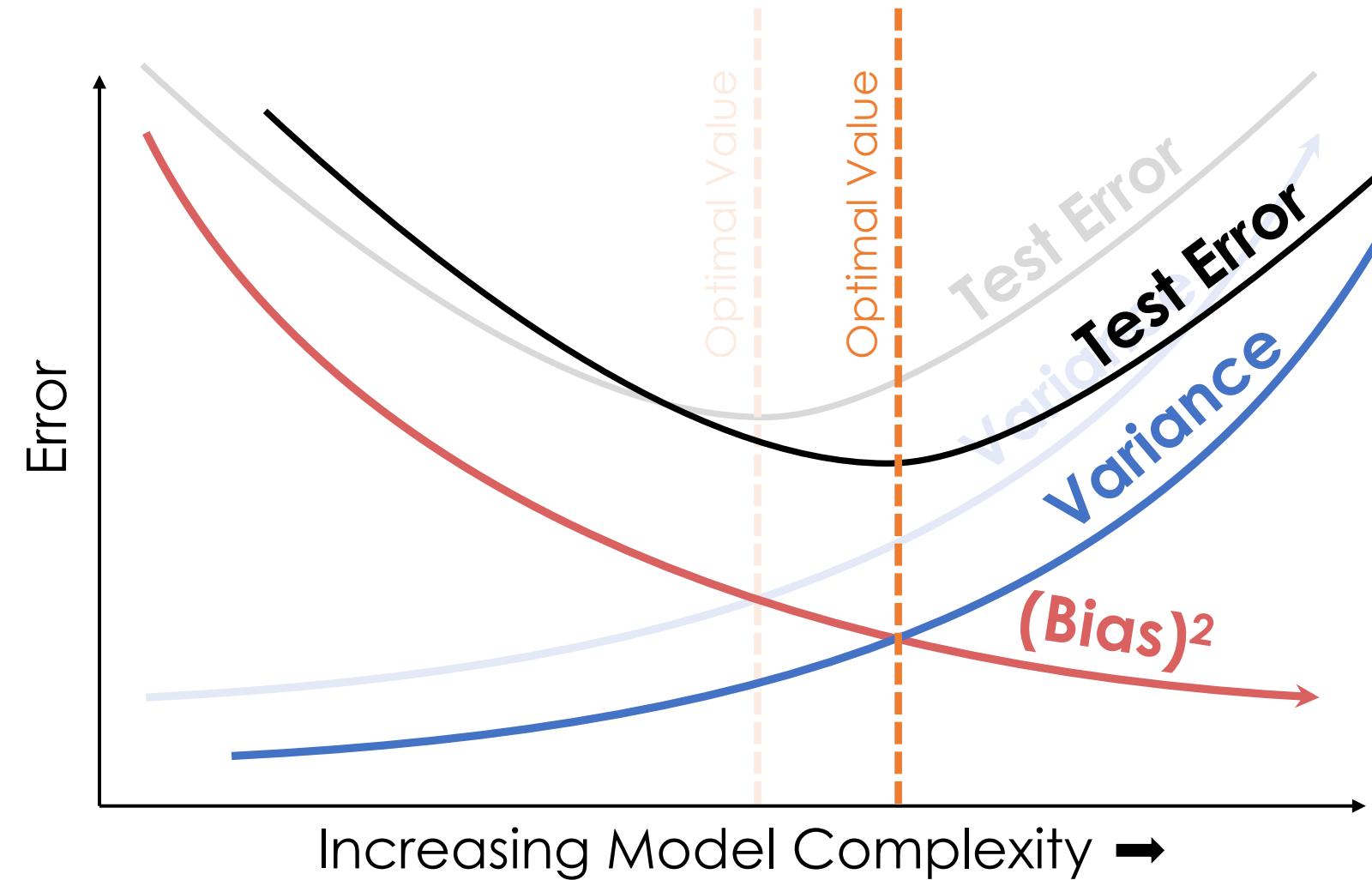


How do we compute test error?

- Use test set to pick best model?
 - No no no nooooo!
- Cross validation?
 - Yes!!!!
- What happens with more data?

$$\mathbb{E} [(y - h(x))^2] \quad \text{Noise}$$
$$\mathbb{E} [(h(x) - \mathbb{E}[f_\theta(x)])^2] \quad (\text{Bias})^2$$
$$\mathbb{E} [(\mathbb{E}[f_\theta(x)] - f_\theta(x))^2] \quad \text{Variance}$$

Bias Variance Plot



- Cross validation?
 - Yes!!!!
- What happens with more data?
- $\mathbf{E} [(\mathbf{y} - h(\mathbf{x}))^2]$ **Noise**
- $\mathbf{E} [(h(\mathbf{x}) - \mathbf{E} [f_{\theta}(\mathbf{x})])^2]$ **(Bias)²**
- $\mathbf{E} [(\mathbf{E} [f_{\theta}(\mathbf{x})] - f_{\theta}(\mathbf{x}))^2]$ **Variance**
- How do we tune the model complexity?
 - Number of features
 - Types of features
 - Different models
 - **Regularization**

Regularization

Parametrically Controlling the
Model Complexity

Basic Idea of Regularization

Fit the Data

Penalize
Complex Models

$$\hat{\theta} = \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^n \text{Loss}(y_i, f_{\theta}(x_i)) + \lambda \mathbf{R}(\theta)$$

Regularization
Parameter

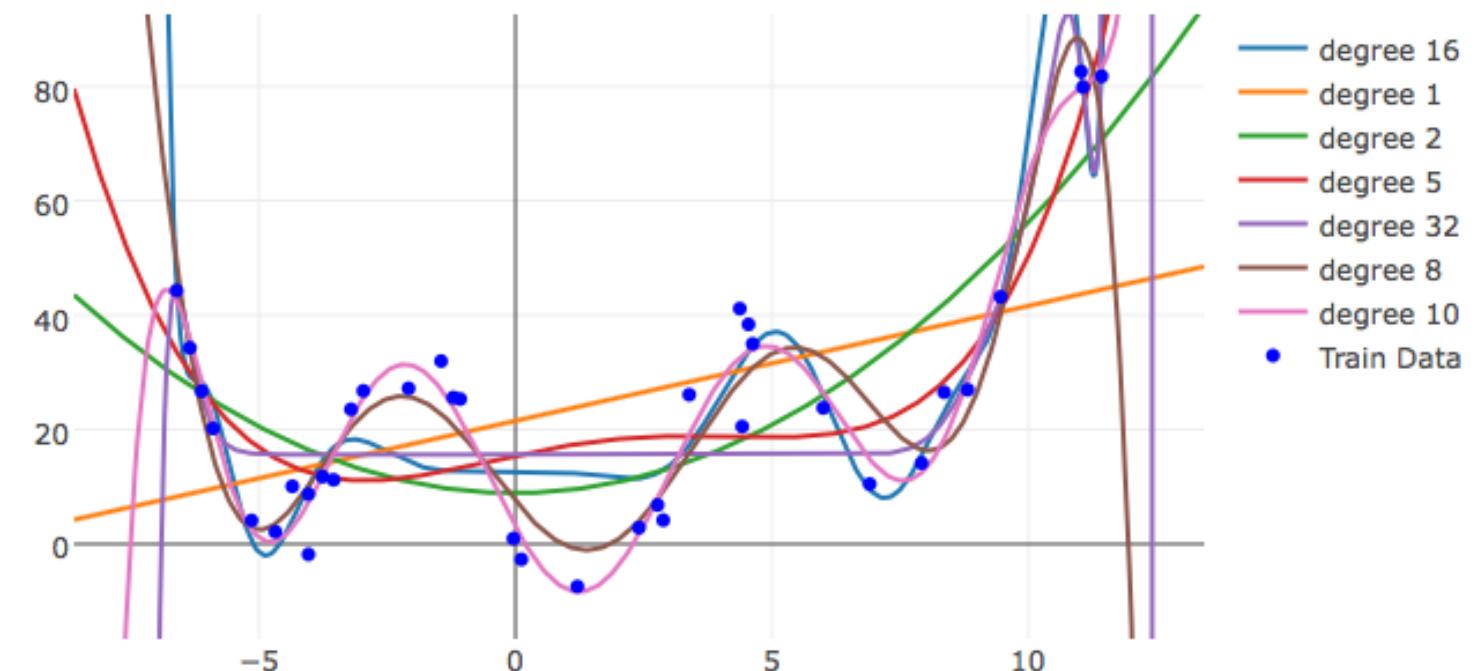
- How should we define $\mathbf{R}(\theta)$?
- How do we determine λ ?

The Regularization Function $R(\theta)$

Goal: Penalize model complexity

Recall earlier: $\phi(x) = [x, x^2, x^3, \dots, x^p]$

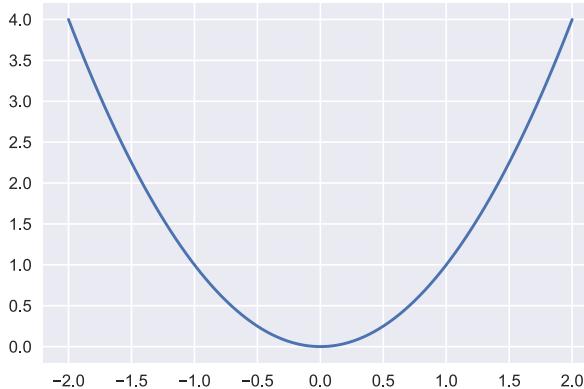
- More features → overfitting ...
- How can we control overfitting through θ
- **Proposal:** set weights = 0 to remove features



Common Regularization Functions

Ridge Regression
(L2-Reg)

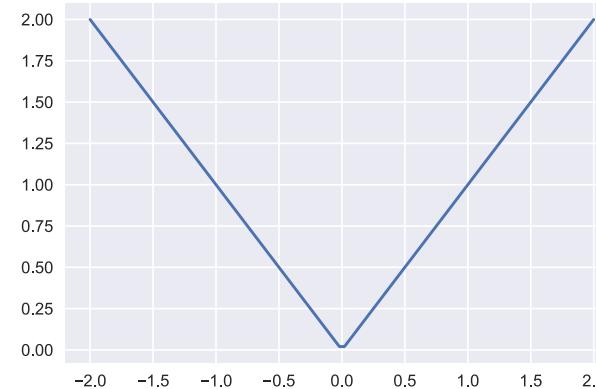
$$R_{\text{Ridge}}(\theta) = \sum_{i=1}^d \theta_i^2$$



- Distributes weight across related features (robust)
- Analytic solution (easy to compute)
- Does not encourage sparsity → small but non-zero weights.

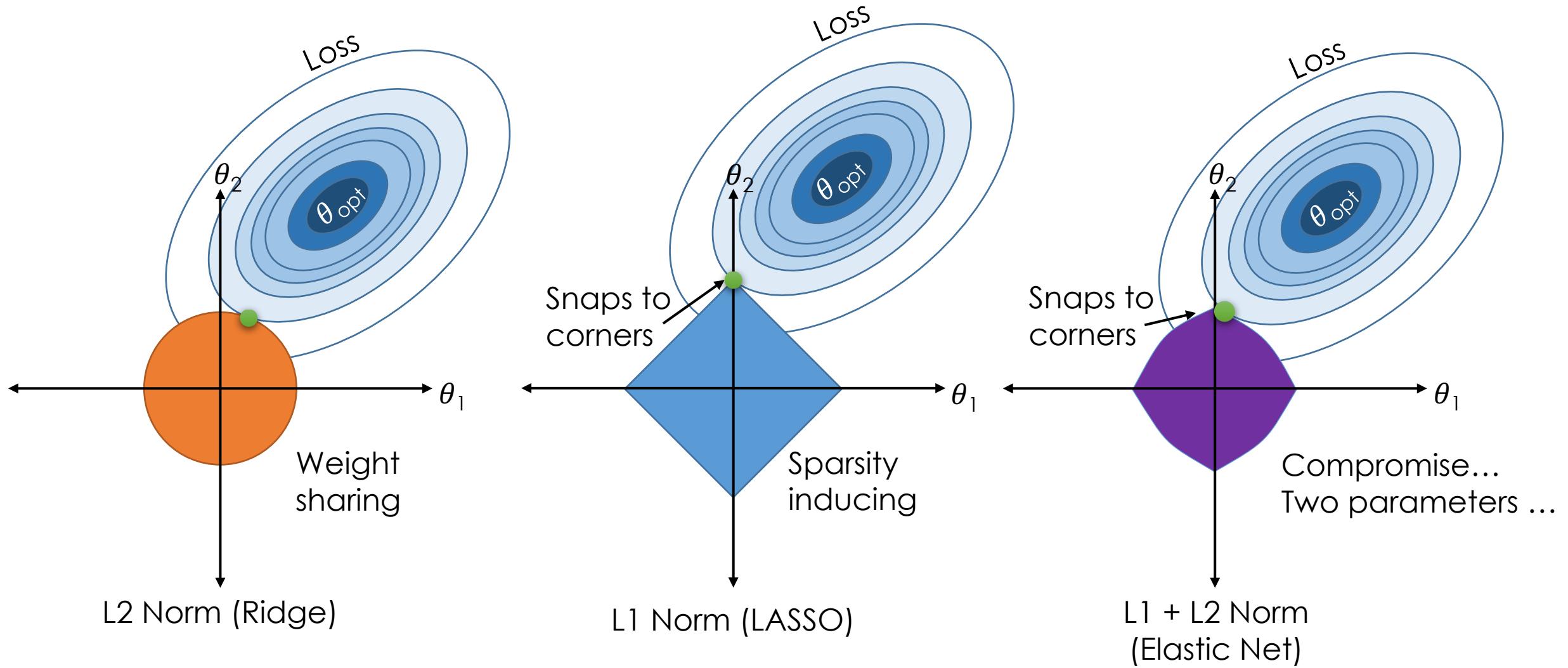
LASSO
(L1-Reg)

$$R_{\text{Lasso}}(\theta) = \sum_{i=1}^d |\theta_i|$$



- **Encourages sparsity** by setting weights = 0
 - Used to select informative features
- Does not have an analytic solution → numerical methods

Regularization and Norm Balls



Standardization and the Intercept Term

Height = θ_1 age_in_seconds + θ_2 weight_in_tons

Small

Large

➤ Regularization penalized dimensions equally

➤ **Standardization**

- Ensure that each dimensions has the same scale
- centered around zero

Standardization

For each dimension k :

$$z_k = \frac{x_k - \mu_k}{\sigma_k}$$

➤ **Intercept Terms**

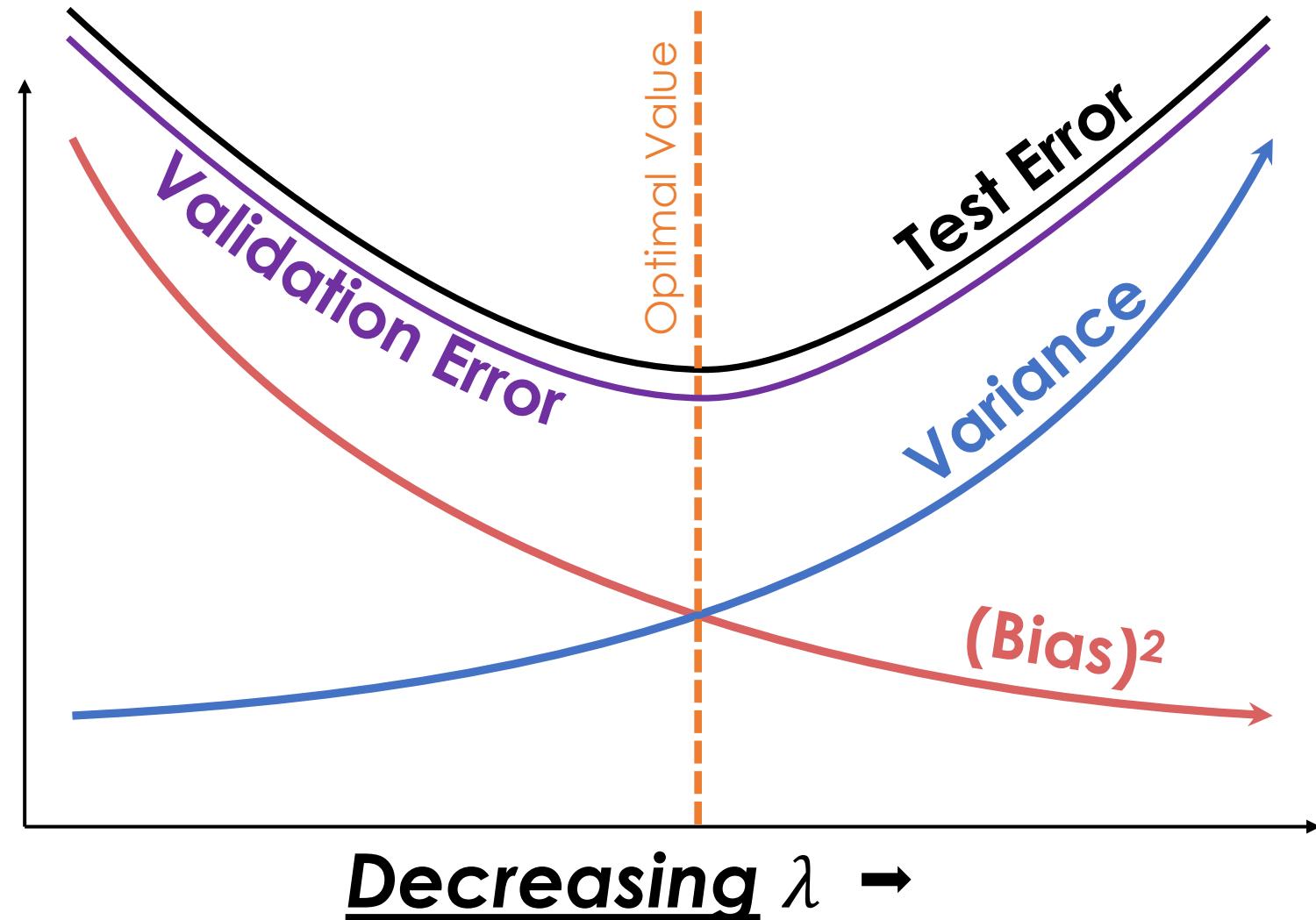
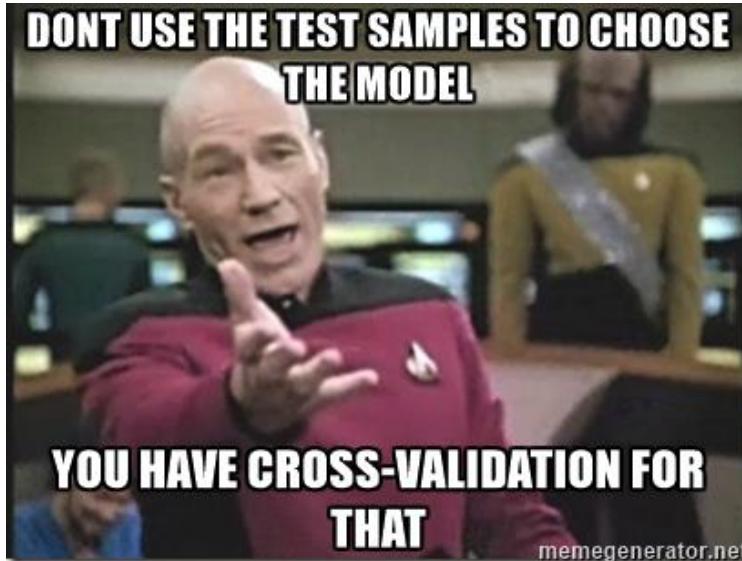
- Center y values (e.g., subtract mean)
- Don't regularize intercept term ← **Suggested**

Determining the Optimal λ

$$\hat{\theta} = \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^n \text{Loss}(y_i, f_{\theta}(x_i)) + \lambda \mathbf{R}(\theta)$$

- Value of λ determines bias-variance tradeoff
 - Larger values → more regularization → more bias → less variance

Determining the Optimal λ

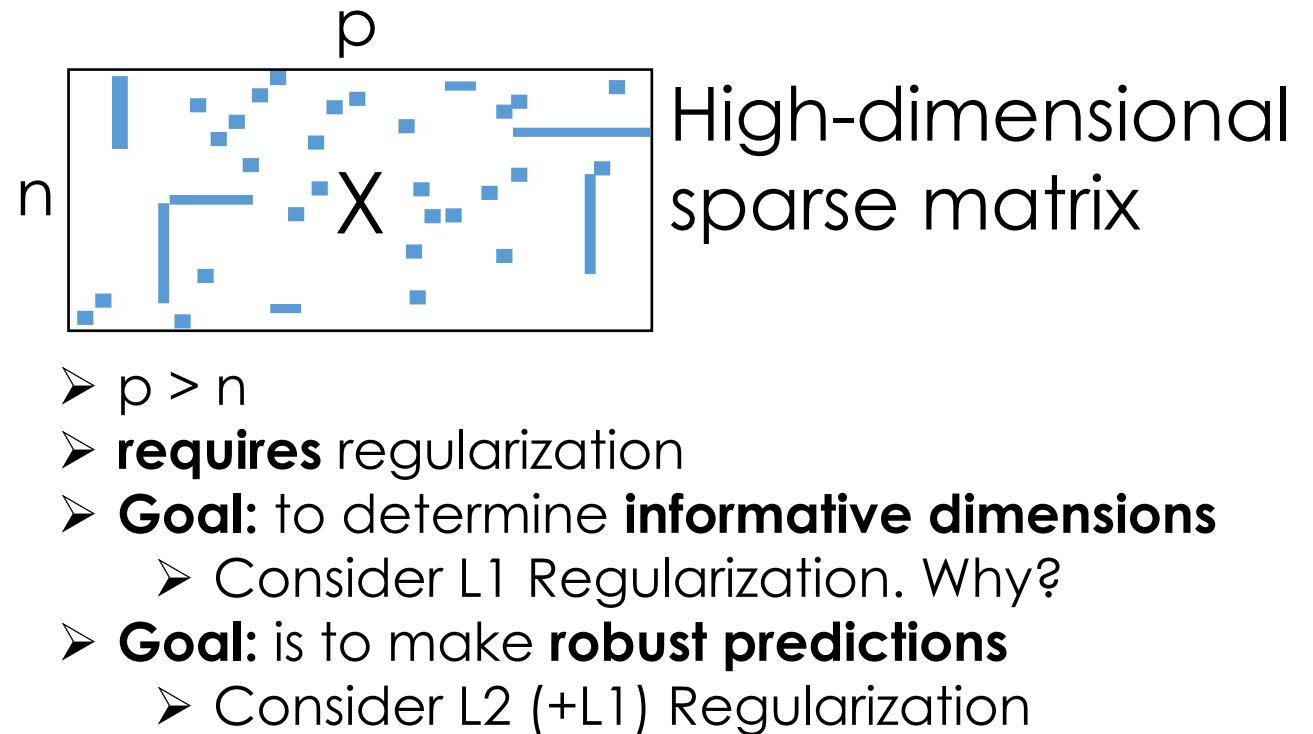
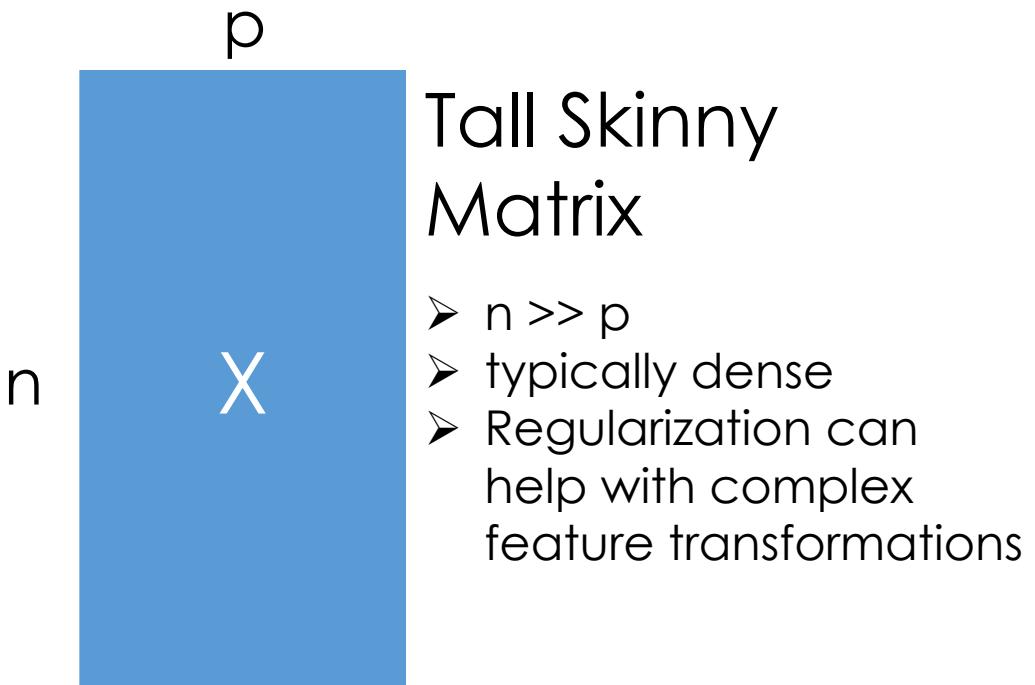


- Value of λ determines bias-variance tradeoff
 - Larger values \rightarrow more regularization \rightarrow more bias \rightarrow less variance
- Determined through cross validation

Python Demo!

Regularization and High-Dimensional Data

Regularization is often used with high-dimensional data



Connection to Bayesian Priors

- Ridge Regression:

Lik.: $y \sim \mathbf{N}(x^T \theta, \sigma_{\text{noise}}^2)$

Prior: $\theta \sim \mathbf{N}(0, 1/\lambda)$

Posterior
(Assume IID)
→
(proportional)

$$\prod_{I=1}^n \exp \left(-\frac{(y - x^T \theta)^2}{2\sigma_{\text{noise}}^2} - \lambda \frac{\theta^2}{2} \right)$$

- LASSO:

Lik.: $y \sim \mathbf{N}(x^T \theta, \sigma_{\text{noise}}^2)$

Prior: $\theta \sim \mathbf{Laplace}(0, p/\lambda)$

Posterior
(Assume IID)
→
(proportional)

$$\prod_{I=1}^n \exp \left(-\frac{(y - x^T \theta)^2}{2\sigma_{\text{noise}}^2} - \lambda \sum_{k=1}^p |\theta_k| \right)$$

- Regularization is often seen as applying a prior.

Summary

$$\mathbb{E} \left[(\mathbf{y} - f_{\theta}(\mathbf{x}))^2 \right] =$$

$$\mathbb{E} \left[(\mathbf{y} - h(\mathbf{x}))^2 \right] +$$

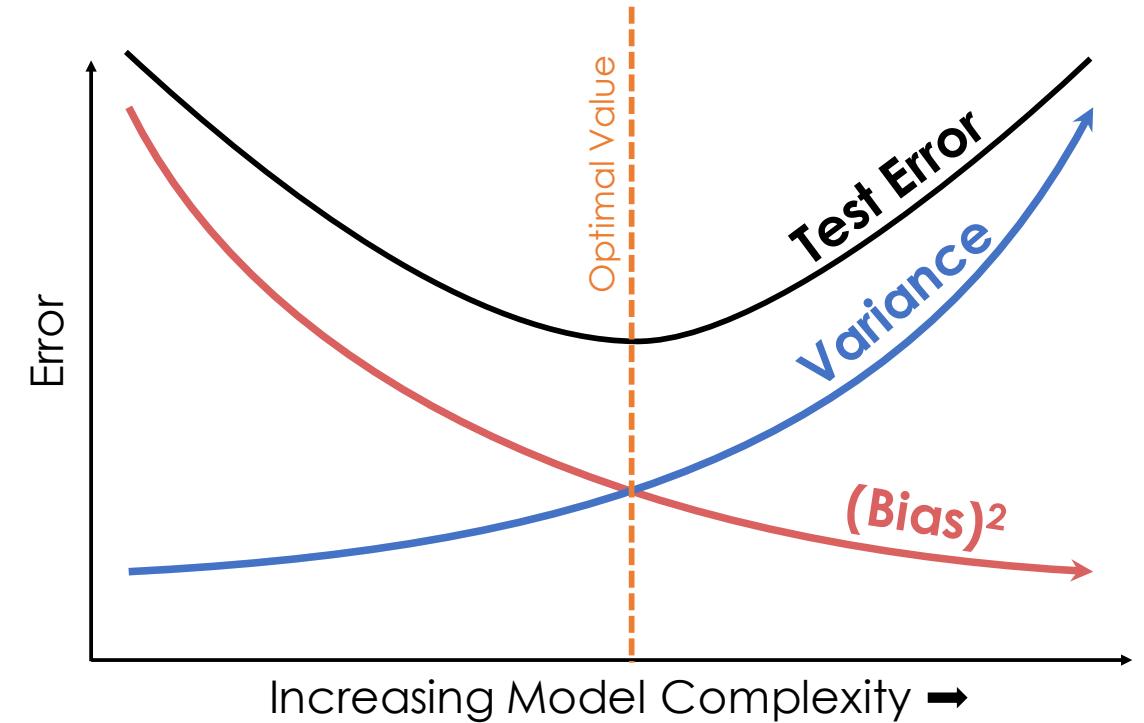
Noise Term

$$\mathbb{E} \left[(h(\mathbf{x}) - \mathbb{E}_{\theta} [f_{\theta}(\mathbf{x})])^2 \right] +$$

(Bias)² Term

$$\mathbb{E} \left[(\mathbb{E}_{\theta} [f_{\theta}(\mathbf{x})] - f_{\theta}(\mathbf{x}))^2 \right]$$

Variance Term



Regularization

$$\hat{\theta} = \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^n \text{Loss}(y_i, f_{\theta}(x_i)) + \lambda \mathbf{R}(\theta)$$

