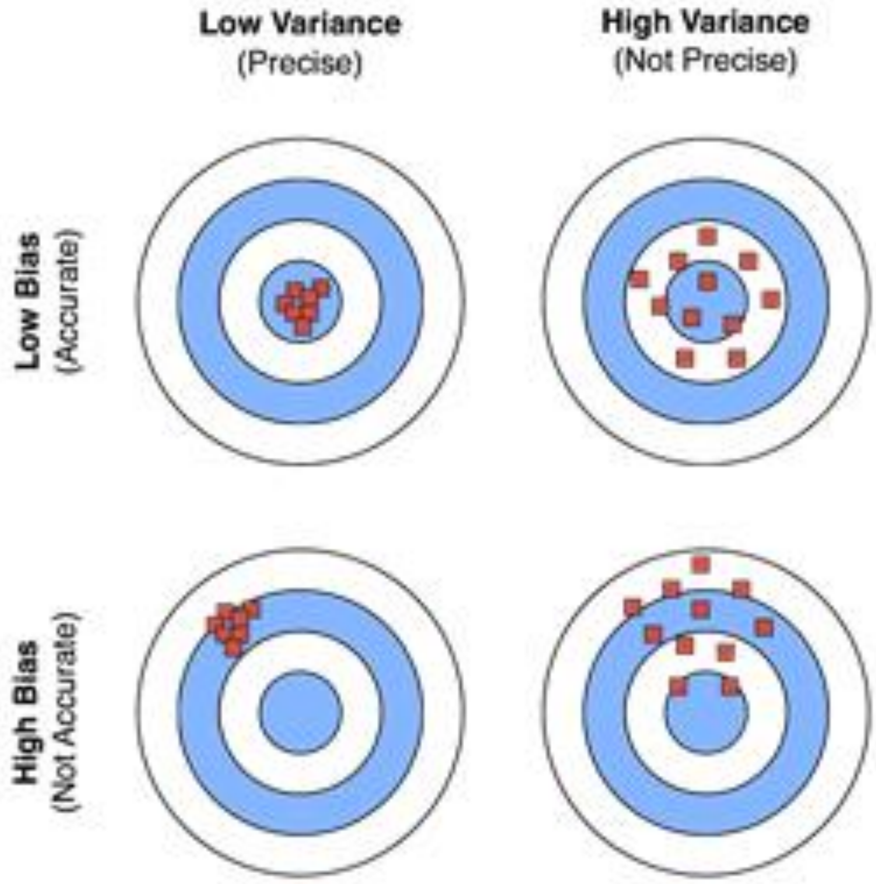
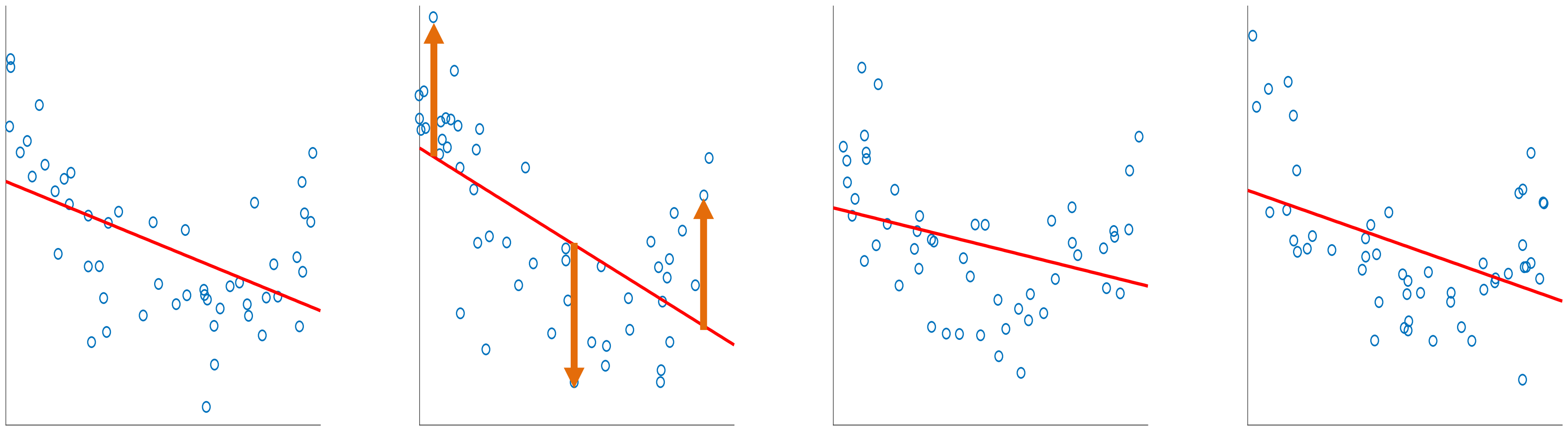
Regularization

A way to avoid overfitting


Low Variance
(Precise)
High Variance
(Not Precise)
Low Bias
(Accurate)
High Bias
(Not Accurate)

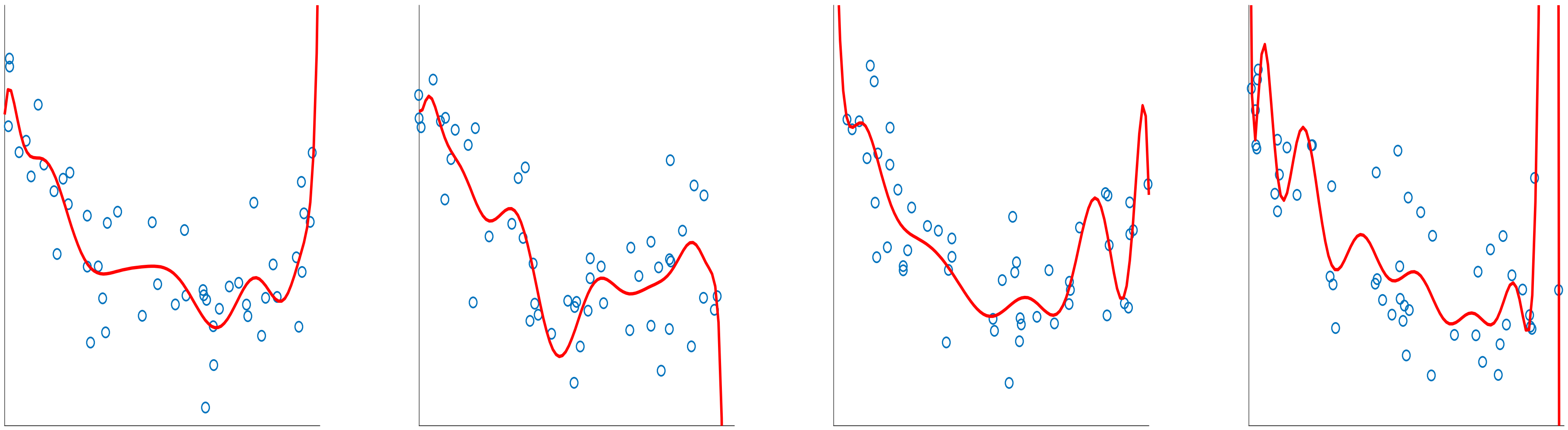
Bias



- Regardless of training sample, or size of training sample, model will produce consistent errors

Actual relationship may be quadratic

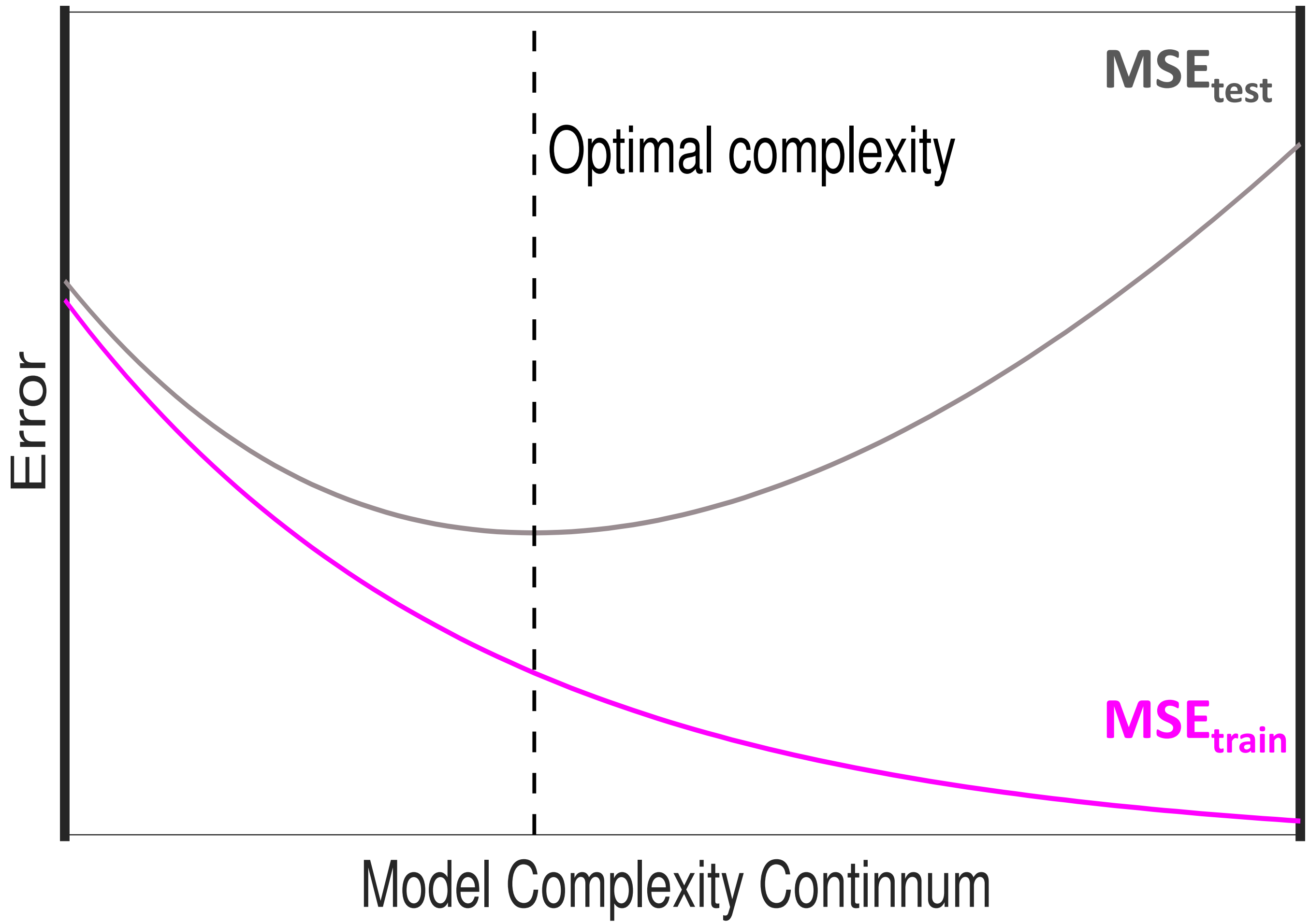
Variance



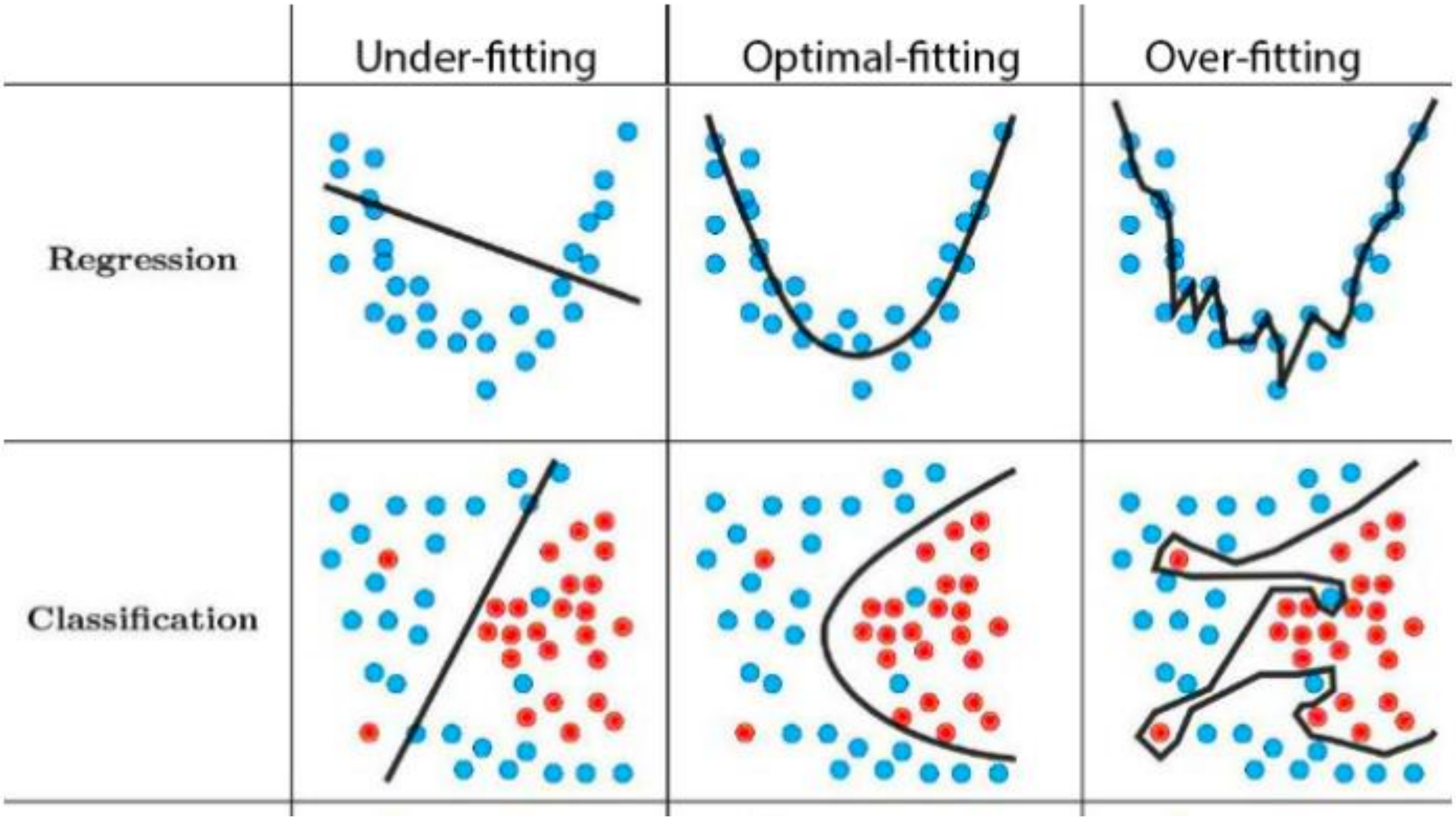
- Different samples of training data yield different model fits

We are trying to fit degree 8 polynomial

Bias-Variance Trade Off Is Revealed Via Test Set Not Training Set



Bias increases ⇦ ⇨ **Variance increases**

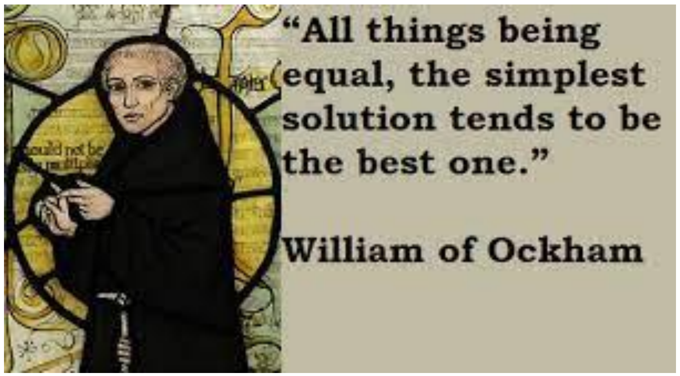


High bias ,
Low variance

Low bias ,
Low variance

Low bias ,
High variance

Regularization



- A controlled way of reducing the complexity of the fitted curve.
- We need to measure the complexity.
- We need to penalize the complex solutions.

Regularization: An Overview

- The idea of regularization revolves around modifying the criterion J; in particular, we add a regularization term that penalizes some specified properties of the model parameters
- $J_{reg}(\Theta) = J(\Theta) + \lambda R(\Theta)$
- where λ is a scalar that gives the weight (or importance) of the regularization term.
- Fitting the model using the modified loss function J_{reg} would result in model parameters with desirable properties (specified by R).

RIDGE and LASSO Regularizations

- In Ridge regularization,
- $J_{reg}(\Theta) = J(\Theta) + \lambda(\theta_1^2 + \cdots + \theta_d^2)$
- Note θ_0 is not used in the penalization

- In Lasso regularization,
- $J_{reg}(\Theta) = J(\Theta) + \lambda(|\theta_1| + \cdots + |\theta_d|)$
- Here also θ_0 is not penalized.

$J(\Theta)$: SSE, Ridge regression

- $J(\Theta) = \sum_{i=1}^{n} \left[\left(\sum_{j=1}^{d} x_{ij}\theta_j + \theta_0\right) - y_j\right]^2$ $J_{reg}(\Theta) = \sum_{i=1}^{n} \left[\left(\sum_{j=1}^{d} x_{ij}\theta_j + \theta_0\right) - y_j\right]^2 + \lambda\left(\sum_{j=1}^{d} \theta_j^2\right)$
- $\frac{\partial J_{reg}}{\partial \theta_j} = 2\left[\sum_{i=1}^{n}\left[\left(\sum_{j=1}^{d} x_{ij}\theta_j + \theta_0\right) - y_j\right]x_{ij} + \lambda\theta_j\right]$

This is for j = 1 to d. For θ_0 there is no regularization penalty (see the next slide)

- $\frac{\partial J_{reg}}{\partial \theta_0} = 2\left[\sum_{i=1}^{n}\left[\left(\sum_{j=1}^{d} x_{ij}\theta_j + \theta_0\right) - y_j\right]\right]$

$J(\Theta)$: SSE, Lasso regression

- $J(\Theta) = \sum_{i=1}^{n} \left[\left(\sum_{j=1}^{d} x_{ij}\theta_j + \theta_0\right) - y_j\right]^2$ $J_{reg}(\Theta) = \sum_{i=1}^{n} \left[\left(\sum_{j=1}^{d} x_{ij}\theta_j + \theta_0\right) - y_j\right]^2 + \lambda\left(\sum_{j=1}^{d} |\theta_j|\right)$
- This is not differentiable, hence we cannot do gradient descent,
  - Quadratic programming (using Lagrange multipliers) can be employed.

The gradient descent update rule is

$$\theta_j(k+1) = \theta_j(k) - \eta \frac{\partial J_{reg}}{\partial \theta_j}$$

Choosing λ

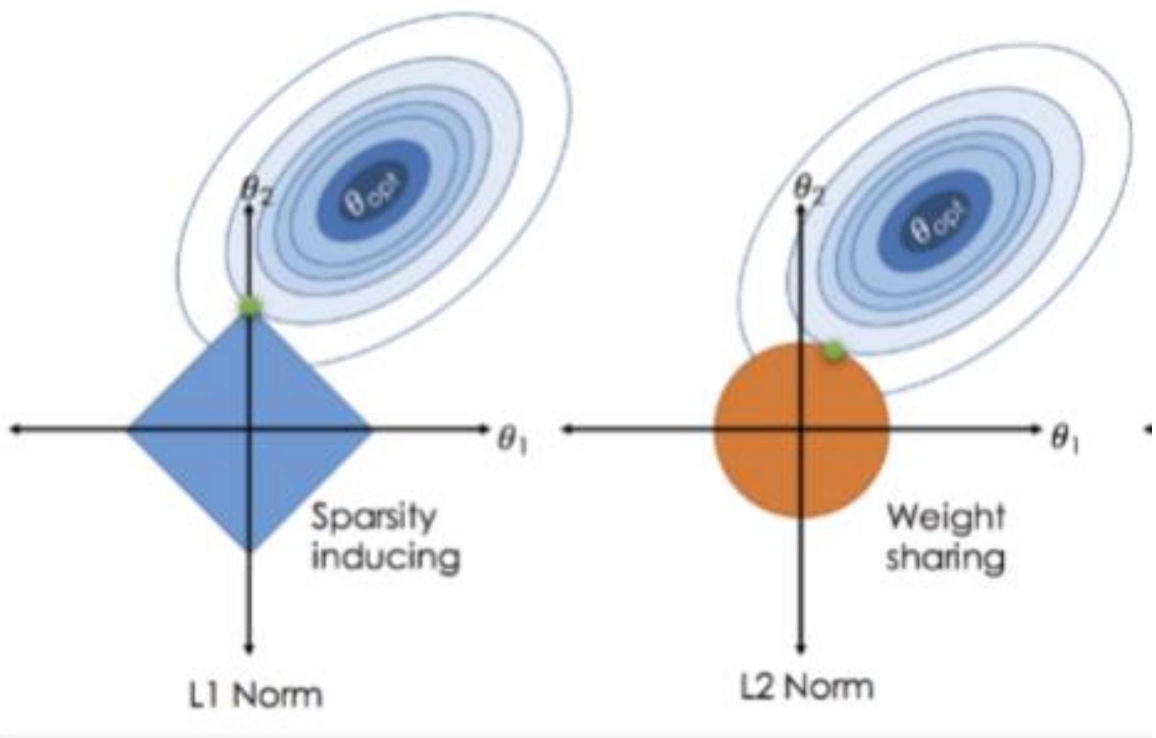
- In both ridge and LASSO regression, we see that the larger our choice of the **regularization parameter** λ, the more heavily we penalize large values in Θ**,**

- If λ is close to zero, we recover the SSE, i.e. ridge and LASSO regression is just ordinary regression.

- If λ is sufficiently large, the SSE term in the regularized loss function will be insignificant and the regularization term will force $\theta_1, \ldots, \theta_d$ to be close to zero. {note, θ_0 can escape this}

- To avoid ad-hoc choices, we should select λ using cross-validation.

Geometric Interpretation



Lasso regularization

Ridge regularization