

# Ridge Regression

---

Nipun Batra

IIT Gandhinagar

September 10, 2025

# Outline

1. Motivation: The Problem of Overfitting
2. Ridge Regression Formulation
3. Mathematical Derivation
4. Hyperparameter Selection
5. Examples and Applications
6. Implementation Details

# Motivation: The Problem of Overfitting

# The Problem: Overfitting in Linear Regression

## Important: Overfitting Challenge

As model complexity increases (higher polynomial degree), we often observe:

- Training error decreases
- Test error increases
- Model coefficients become very large

# The Problem: Overfitting in Linear Regression

## Important: Overfitting Challenge

As model complexity increases (higher polynomial degree), we often observe:

- Training error decreases
- Test error increases
- Model coefficients become very large

## Key Points: Key Insight

Large coefficient magnitudes often indicate overfitting!

# The Problem: Overfitting in Linear Regression

## Important: Overfitting Challenge

As model complexity increases (higher polynomial degree), we often observe:

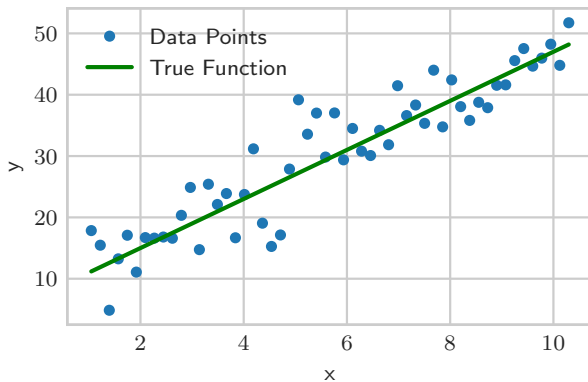
- Training error decreases
- Test error increases
- Model coefficients become very large

## Key Points: Key Insight

Large coefficient magnitudes often indicate overfitting!

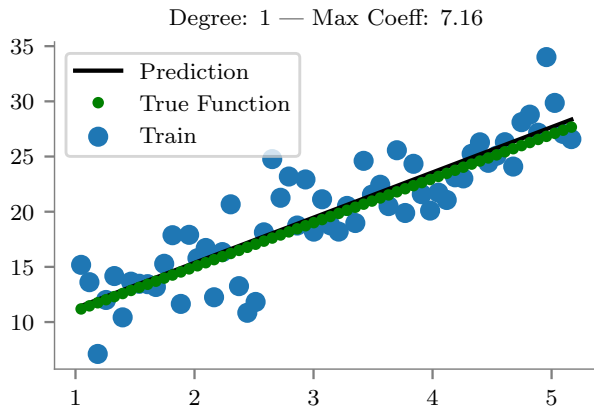
In polynomial  $f(x) = c_0 + c_1x + c_2x^2 + \dots + c_dx^d$ , watch  $\max |c_i|$

# Demonstration: Polynomial Degree vs Overfitting



Base Data Set

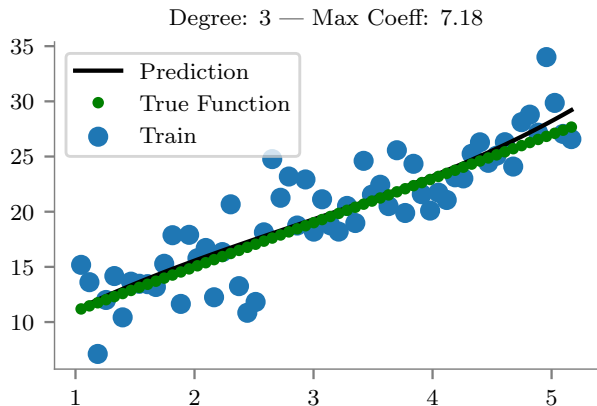
## Demonstration: Polynomial Degree vs Overfitting



Fit with Degree 1 - Underfitting

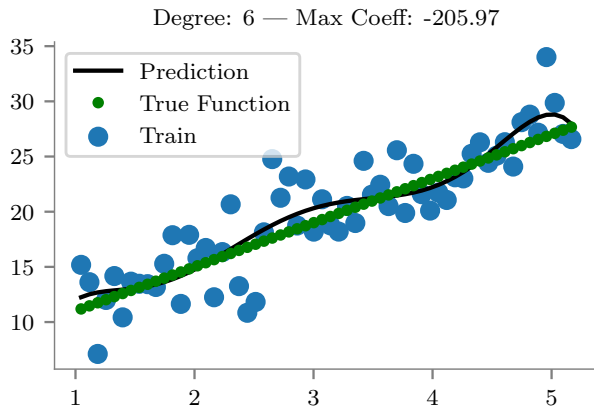


## Demonstration: Polynomial Degree vs Overfitting



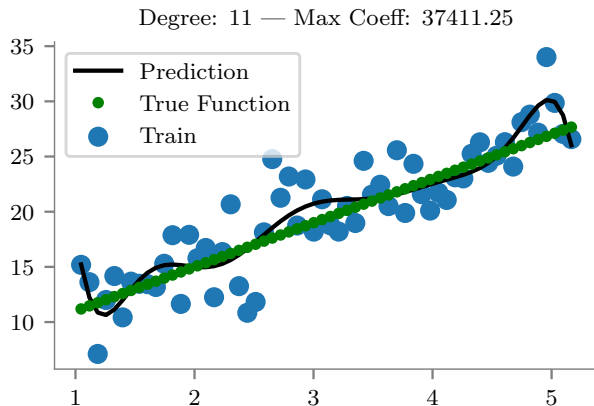
Fit with Degree 3 - Good Fit

## Demonstration: Polynomial Degree vs Overfitting



Fit with Degree 6 - Starting to Overfit

## Demonstration: Polynomial Degree vs Overfitting

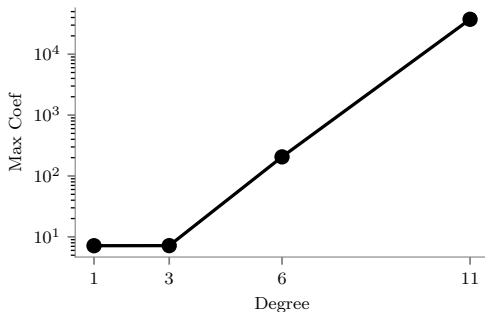


Fit with Degree 11 - Severe Overfitting

# Coefficient Explosion with Overfitting

## Key Points: Key Observation

As polynomial degree increases  $\rightarrow$  coefficients grow exponentially!



Coefficient Magnitudes vs Polynomial Degree

# The Central Question

## **Important: Critical Question**

How can we control coefficient magnitudes to prevent overfitting?

# The Central Question

## **Important: Critical Question**

How can we control coefficient magnitudes to prevent overfitting?

## **Key Points: Answer Preview**

Ridge regression adds a penalty term to shrink coefficients!

# Pop Quiz 1

## Answer this!

Which statement about overfitting is TRUE?

- A) Higher polynomial degree always improves generalization
- B) Large coefficients indicate good model fit
- C) Overfitting occurs when training error  $\gg$  test error
- D) Overfitting occurs when training error  $\ll$  test error

## Answer: Pop Quiz 1

### Answer this!

**D) Overfitting occurs when training error  $\ll$  test error**

Explanation:

- Training error becomes very small (model memorizes training data)
- Test error remains large (model fails to generalize)
- Large gap indicates overfitting



# Ridge Regression Formulation

## Solution: Regularization

### **Theorem: Ridge Regression Approach**

Add a penalty term to control coefficient magnitudes:

# Solution: Regularization

## Theorem: Ridge Regression Approach

Add a penalty term to control coefficient magnitudes:

## Definition: Constrained Formulation

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad & (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) \\ \text{subject to} \quad & \boldsymbol{\theta}^T \boldsymbol{\theta} \leq S \end{aligned}$$

where  $S > 0$  controls the size of the coefficient vector.

# Lagrangian Formulation

## Theorem: Equivalence Theorem

The constrained problem is equivalent to the unconstrained:

$$\min_{\boldsymbol{\theta}} \quad (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}^T \boldsymbol{\theta}$$

where  $\lambda \geq 0$  is the regularization parameter.

# Lagrangian Formulation

## Theorem: Equivalence Theorem

The constrained problem is equivalent to the unconstrained:

$$\min_{\boldsymbol{\theta}} \quad (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}^T \boldsymbol{\theta}$$

where  $\lambda \geq 0$  is the regularization parameter.

## Key Points: Key Insight

This transforms a constrained optimization into an unconstrained one with a penalty term.

# Understanding the Ridge Penalty

$$J(\boldsymbol{\theta}) = \underbrace{(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})}_{\text{Fit to data (MSE)}} + \underbrace{\lambda \boldsymbol{\theta}^T \boldsymbol{\theta}}_{\text{Penalty term}} \quad (1)$$

$$= \text{MSE}(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_2^2 \quad (2)$$

# Understanding the Ridge Penalty

$$J(\boldsymbol{\theta}) = \underbrace{(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})}_{\text{Fit to data (MSE)}} + \underbrace{\lambda \boldsymbol{\theta}^T \boldsymbol{\theta}}_{\text{Penalty term}} \quad (1)$$

$$= \text{MSE}(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_2^2 \quad (2)$$

## Key Points: Key Components

- **Data fitting term:** Ensures good fit to training data
- **Regularization term:**  $L_2$  penalty shrinks coefficients toward zero
- $\lambda$ : Controls trade-off between fitting vs. regularization

# Effect of Regularization Parameter $\lambda$

## Key Points: Parameter Effects

- $\lambda = 0$ : No regularization (standard linear regression)
- $\lambda$  small: Light regularization (slight shrinkage)
- $\lambda$  large: Heavy regularization (strong shrinkage)
- $\lambda \rightarrow \infty$ : Extreme regularization (coefficients  $\rightarrow 0$ )



# Effect of Regularization Parameter $\lambda$

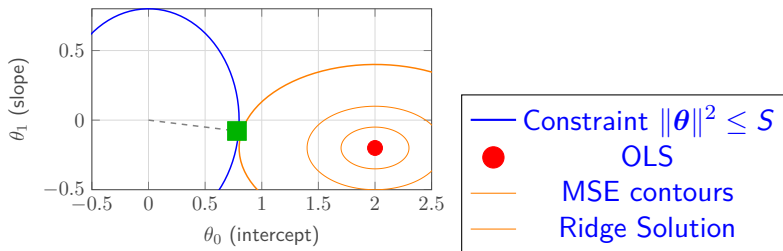
## Key Points: Parameter Effects

- $\lambda = 0$ : No regularization (standard linear regression)
- $\lambda$  small: Light regularization (slight shrinkage)
- $\lambda$  large: Heavy regularization (strong shrinkage)
- $\lambda \rightarrow \infty$ : Extreme regularization (coefficients  $\rightarrow 0$ )

## Important: Key Trade-off

Higher  $\lambda$  = more regularization = more bias, less variance

# Geometric Interpretation



Ridge solution where MSE contours touch constraint region

## Key Points: Key Insight

Ridge finds the minimum MSE point within the constraint  $\|\theta\|_2^2 \leq S$

# Mathematical Derivation

# Mathematical Derivation: Step 1

## Step 1: Set up the Lagrangian

For the constrained optimization problem:

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad & (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) \\ \text{s.t.} \quad & \boldsymbol{\theta}^T \boldsymbol{\theta} \leq S \end{aligned}$$

The Lagrangian is:

$$L(\boldsymbol{\theta}, \lambda) = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \lambda (\boldsymbol{\theta}^T \boldsymbol{\theta} - S)$$

where  $\lambda \geq 0$  is the Lagrange multiplier.

## Mathematical Derivation: Step 2

### Step 2: Apply KKT Conditions

For optimality, we need:

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = 0 \quad (\text{stationarity}) \quad (3)$$

$$\lambda \geq 0 \quad (\text{dual feasibility}) \quad (4)$$

$$\boldsymbol{\theta}^T \boldsymbol{\theta} - S \leq 0 \quad (\text{primal feasibility}) \quad (5)$$

$$\lambda(\boldsymbol{\theta}^T \boldsymbol{\theta} - S) = 0 \quad (\text{complementary slackness}) \quad (6)$$

# Mathematical Derivation: Step 2

## Step 2: Apply KKT Conditions

For optimality, we need:

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = 0 \quad (\text{stationarity}) \quad (3)$$

$$\lambda \geq 0 \quad (\text{dual feasibility}) \quad (4)$$

$$\boldsymbol{\theta}^T \boldsymbol{\theta} - S \leq 0 \quad (\text{primal feasibility}) \quad (5)$$

$$\lambda(\boldsymbol{\theta}^T \boldsymbol{\theta} - S) = 0 \quad (\text{complementary slackness}) \quad (6)$$

## Key Points: Two Cases

- **Case 1:**  $\lambda = 0 \Rightarrow$  No constraint active (standard OLS)
- **Case 2:**  $\lambda > 0 \Rightarrow \boldsymbol{\theta}^T \boldsymbol{\theta} = S$  (constraint is tight)

## Mathematical Derivation: Step 3

### Step 3: Compute the Gradient

Taking the derivative of the Lagrangian with respect to  $\theta$ :

$$\frac{\partial L}{\partial \theta} = \frac{\partial}{\partial \theta} \left[ (\mathbf{y} - \mathbf{X}\theta)^T (\mathbf{y} - \mathbf{X}\theta) + \lambda \theta^T \theta \right] \quad (7)$$

$$= \frac{\partial}{\partial \theta} \left[ \mathbf{y}^T \mathbf{y} - 2\mathbf{y}^T \mathbf{X}\theta + \theta^T \mathbf{X}^T \mathbf{X}\theta + \lambda \theta^T \theta \right] \quad (8)$$

$$= -2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X}\theta + 2\lambda \theta \quad (9)$$

## Mathematical Derivation: Step 4

### Step 4: Set Gradient to Zero

Setting  $\frac{\partial L}{\partial \theta} = 0$ :

$$-2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X} \theta + 2\lambda \theta = 0 \quad (10)$$

$$-\mathbf{X}^T \mathbf{y} + (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}) \theta = 0 \quad (11)$$

$$(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}) \theta = \mathbf{X}^T \mathbf{y} \quad (12)$$



## Mathematical Derivation: Step 4

### Step 4: Set Gradient to Zero

Setting  $\frac{\partial L}{\partial \boldsymbol{\theta}} = 0$ :

$$-2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X} \boldsymbol{\theta} + 2\lambda \boldsymbol{\theta} = 0 \quad (10)$$

$$-\mathbf{X}^T \mathbf{y} + (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}) \boldsymbol{\theta} = 0 \quad (11)$$

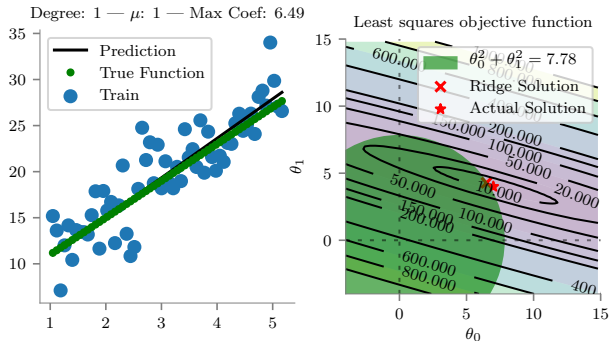
$$(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}) \boldsymbol{\theta} = \mathbf{X}^T \mathbf{y} \quad (12)$$

### Theorem: Ridge Regression Solution

$$\hat{\boldsymbol{\theta}}_{\text{ridge}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$$

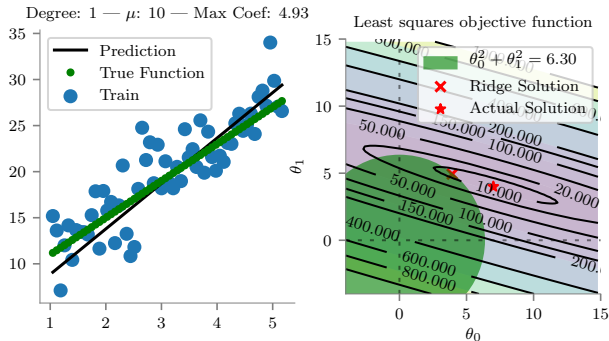
Compare with OLS:  $\hat{\boldsymbol{\theta}}_{\text{OLS}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$

# Effect of Regularization Parameter $\lambda$



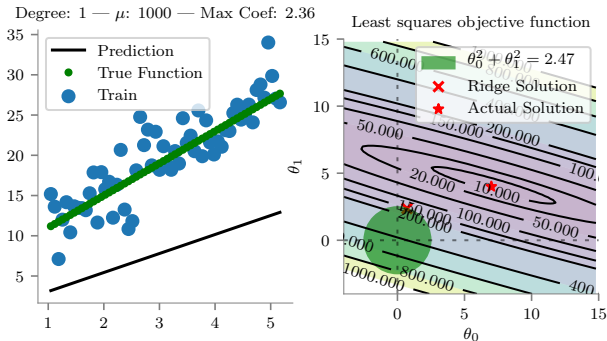
$\lambda = 1$  - Mild Regularization

# Effect of Regularization Parameter $\lambda$



$\lambda = 10$  - Moderate Regularization

# Effect of Regularization Parameter $\lambda$



$\lambda = 1000$  - Heavy Regularization

## Pop Quiz 2

### Answer this!

What happens to the Ridge regression solution as  $\lambda \rightarrow \infty$ ?

- A) Coefficients approach the OLS solution
- B) Coefficients approach zero
- C) Solution becomes undefined
- D) Training error becomes zero

Answer: Pop Quiz 2

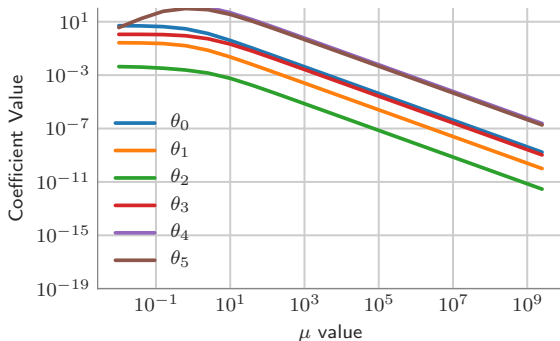
**Answer this!**

**B) Coefficients approach zero**

As  $\lambda \rightarrow \infty$ , the penalty term dominates:

$$\hat{\theta}_{\text{ridge}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y} \approx \lambda^{-1} \mathbf{I} \mathbf{X}^T \mathbf{y} \rightarrow \mathbf{0}$$

# Coefficient Shrinkage: Visual Evidence

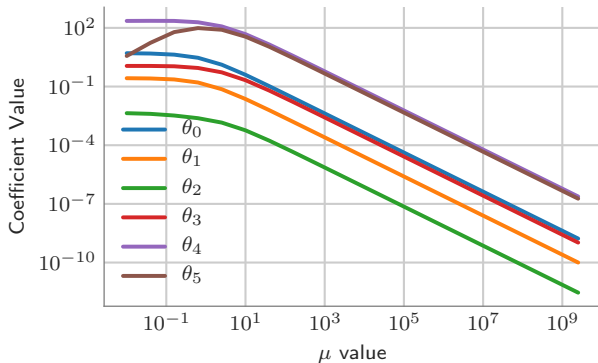


Coefficient Magnitudes vs  $\lambda$  (Real Estate Dataset)

## Important: Important Question

Do coefficients ever become exactly zero?

# Ridge Coefficient Behavior



Ridge Coefficients Shrink but Never Reach Zero



# Ridge vs. Lasso: Key Difference

## Key Points: Coefficient Behavior Comparison

- **Ridge ( $L_2$ ):** Coefficients shrink toward zero but remain non-zero
- **Lasso ( $L_1$ ):** Coefficients can become exactly zero (feature selection)

# Ridge vs. Lasso: Key Difference

## Key Points: Coefficient Behavior Comparison

- **Ridge ( $L_2$ ):** Coefficients shrink toward zero but remain non-zero
- **Lasso ( $L_1$ ):** Coefficients can become exactly zero (feature selection)

## Important: Important Insight

Ridge provides shrinkage, Lasso provides selection!

# Ridge Regression Solution

## Theorem: Ridge Solution Formula

$$\hat{\boldsymbol{\theta}}_{\text{ridge}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$$

## Key Property 1: Always Invertible

### Theorem: Invertibility Guarantee

$(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})$  is always positive definite for  $\lambda > 0$

# Key Property 1: Always Invertible

## Theorem: Invertibility Guarantee

$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})$  is always positive definite for  $\lambda > 0$

## Key Points: Why This Matters

- No singularity issues (unlike OLS)
- Always has unique solution
- Handles multi-collinearity gracefully

## Key Property 2: Coefficient Shrinkage

### Theorem: Shrinkage Effect

Ridge regression shrinks coefficients toward zero (but not exactly zero)

## Key Property 2: Coefficient Shrinkage

### Theorem: Shrinkage Effect

Ridge regression shrinks coefficients toward zero (but not exactly zero)

### Key Points: Shrinkage Benefits

- Reduces overfitting
- Stabilizes coefficient estimates
- Improves generalization

## Key Property 3: Bias-Variance Trade-off

### Theorem: Trade-off Effect

Ridge regression increases bias but reduces variance



## Key Property 3: Bias-Variance Trade-off

### Theorem: Trade-off Effect

Ridge regression increases bias but reduces variance

### Key Points: Net Effect

- Total error often decreases
- Better generalization to new data
- Controlled by  $\lambda$  parameter

# Hyperparameter Selection

# Choosing the Regularization Parameter $\lambda$

## **Important: Hyperparameter Selection**

How do we choose the optimal value of  $\lambda$ ?

# Choosing the Regularization Parameter $\lambda$

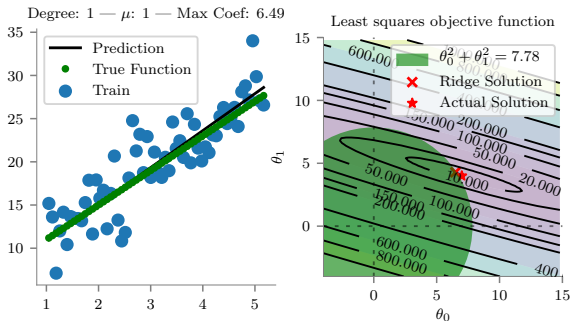
## Important: Hyperparameter Selection

How do we choose the optimal value of  $\lambda$ ?

## Theorem: Cross-Validation Approach

1. Split data into training and validation sets (k-fold CV)
2. For each candidate  $\lambda$  value:
  - Train ridge model on training data
  - Compute validation error
3. Select  $\lambda$  that minimizes validation error
4. Retrain on full dataset with chosen  $\lambda$

# Cross-Validation for Ridge Regression



Cross-validation curve showing optimal  $\lambda$

## Key Points: CV Pattern

Small  $\lambda$ : Overfitting    Large  $\lambda$ : Underfitting    Optimal  $\lambda$ : Best trade-off

# Bias-Variance Trade-off in Ridge Regression

## Theorem: Bias-Variance Decomposition

$$\text{Total Error} = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$

# Bias-Variance Trade-off in Ridge Regression

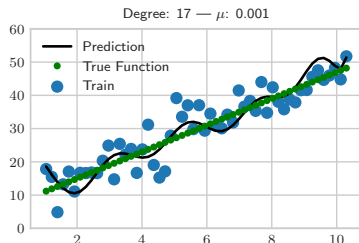
## Theorem: Bias-Variance Decomposition

$$\text{Total Error} = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$

## Key Points: Ridge Effect

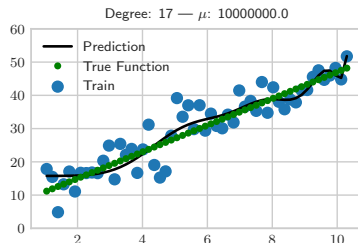
Regularization increases bias but reduces variance, often leading to lower total error.

# Small vs Large Regularization



**Small  $\lambda$  ( $\lambda \rightarrow 0$ ):**

- Low bias
- High variance
- Risk of overfitting



**Large  $\lambda$  ( $\lambda \rightarrow \infty$ ):**

- High bias
- Low variance
- Risk of underfitting



## Pop Quiz 3

### Answer this!

In ridge regression, as we increase  $\lambda$ , what happens to model bias and variance?

- A) Both bias and variance increase
- B) Both bias and variance decrease
- C) Bias increases, variance decreases
- D) Bias decreases, variance increases

## Answer: Pop Quiz 3

### Answer this!

#### **C) Bias increases, variance decreases**

Explanation:

- Increasing  $\lambda$  constrains coefficients more severely
- Model becomes simpler (higher bias)
- Less sensitive to training data variations (lower variance)
- This is the fundamental bias-variance trade-off!

# **Examples and Applications**

## Worked Example: Setup

### Example: Ridge Regression Example

Given the following simple dataset, compare OLS vs. Ridge regression with  $\lambda = 2$ :

Data:  $(x_1, y_1) = (1, 1)$ ,  $(x_2, y_2) = (2, 2)$ ,  $(x_3, y_3) = (3, 3)$ ,  
 $(x_4, y_4) = (4, 0)$

Model:  $y = \theta_0 + \theta_1 x$

## Worked Example: Setup

### Example: Ridge Regression Example

Given the following simple dataset, compare OLS vs. Ridge regression with  $\lambda = 2$ :

Data:  $(x_1, y_1) = (1, 1)$ ,  $(x_2, y_2) = (2, 2)$ ,  $(x_3, y_3) = (3, 3)$ ,  
 $(x_4, y_4) = (4, 0)$

Model:  $y = \theta_0 + \theta_1 x$

Step 1: Set up matrices

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 0 \end{bmatrix}, \quad \boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$$

## Worked Example: OLS Setup

Step 2: Ordinary Least Squares

$$\hat{\boldsymbol{\theta}}_{\text{OLS}} = (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{y})$$

## Worked Example: OLS Setup

### Step 2: Ordinary Least Squares

$$\hat{\theta}_{\text{OLS}} = (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{y})$$

### Step 3: Compute matrix products

$$\mathbf{X}^T \mathbf{X} = \begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix}$$

$$\mathbf{X}^T \mathbf{y} = \begin{bmatrix} 6 \\ 14 \end{bmatrix}$$

## Worked Example: Matrix Inverse

Step 4: Compute the inverse

For  $\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix}$ :

$$\det(\mathbf{X}^T\mathbf{X}) = 4 \cdot 30 - 10 \cdot 10 = 20$$

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{20} \begin{bmatrix} 30 & -10 \\ -10 & 4 \end{bmatrix}$$



## Worked Example: OLS Calculation

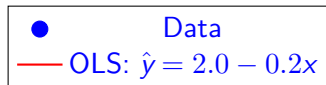
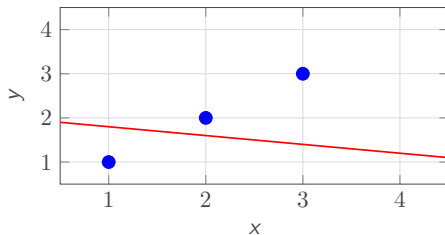
Step 5: Final matrix multiplication

$$\begin{aligned}\hat{\theta}_{\text{OLS}} &= (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{y}) \\ &= \frac{1}{20} \begin{bmatrix} 30 & -10 \\ -10 & 4 \end{bmatrix} \begin{bmatrix} 6 \\ 14 \end{bmatrix} \\ &= \frac{1}{20} \begin{bmatrix} 180 - 140 \\ -60 + 56 \end{bmatrix} = \frac{1}{20} \begin{bmatrix} 40 \\ -4 \end{bmatrix} = \begin{bmatrix} 2.0 \\ -0.2 \end{bmatrix}\end{aligned}$$

# OLS Final Result

## Theorem: OLS Result

$$\hat{y} = 2.0 - 0.2x \quad (\text{No regularization})$$



OLS fit to our example data

## Worked Example: Ridge Setup

Step 5: Ridge regression with  $\lambda = 2$

$$\hat{\boldsymbol{\theta}}_{\text{ridge}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} (\mathbf{X}^T \mathbf{y})$$

## Worked Example: Ridge Setup

Step 5: Ridge regression with  $\lambda = 2$

$$\hat{\boldsymbol{\theta}}_{\text{ridge}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} (\mathbf{X}^T \mathbf{y})$$

Step 6: Add regularization term

$$\begin{aligned} \mathbf{X}^T \mathbf{X} + \lambda \mathbf{I} &= \begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix} + 2\mathbf{I} \\ &= \begin{bmatrix} 6 & 10 \\ 10 & 32 \end{bmatrix} \end{aligned}$$

## Worked Example: Matrix Inverse

Step 7: Compute inverse

$$\det(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}) = 6 \cdot 32 - 10 \cdot 10 = 92$$

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1} = \frac{1}{92} \begin{bmatrix} 32 & -10 \\ -10 & 6 \end{bmatrix}$$

## Worked Example: Ridge Calculation

### Step 8: Matrix multiplication

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\text{ridge}} &= (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} (\mathbf{X}^T \mathbf{y}) \\ &= \frac{1}{92} \begin{bmatrix} 32 & -10 \\ -10 & 6 \end{bmatrix} \begin{bmatrix} 6 \\ 14 \end{bmatrix}\end{aligned}$$

## Worked Example: Ridge Calculation

### Step 8: Matrix multiplication

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\text{ridge}} &= (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} (\mathbf{X}^T \mathbf{y}) \\ &= \frac{1}{92} \begin{bmatrix} 32 & -10 \\ -10 & 6 \end{bmatrix} \begin{bmatrix} 6 \\ 14 \end{bmatrix}\end{aligned}$$

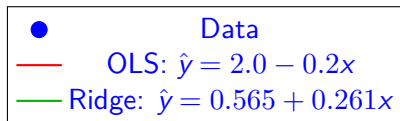
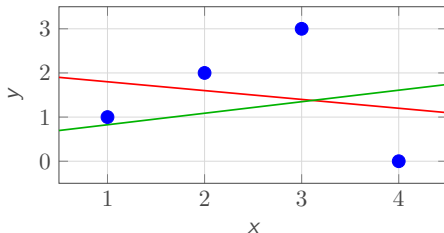
### Step 9: Compute products

$$\begin{aligned}&= \frac{1}{92} \begin{bmatrix} 32 \cdot 6 + (-10) \cdot 14 \\ (-10) \cdot 6 + 6 \cdot 14 \end{bmatrix} \\ &= \frac{1}{92} \begin{bmatrix} 192 - 140 \\ -60 + 84 \end{bmatrix} = \frac{1}{92} \begin{bmatrix} 52 \\ 24 \end{bmatrix} = \begin{bmatrix} 0.565 \\ 0.261 \end{bmatrix}\end{aligned}$$

## Ridge vs OLS: Final Comparison

### Theorem: Ridge Result

$$\hat{y} = 0.565 + 0.261x \quad (\text{With } \lambda = 2)$$



Ridge regression provides more stable coefficients



# Coefficient Magnitude Comparison

## Theorem: OLS vs Ridge Solutions

- **OLS:**  $\theta_{OLS} = \begin{bmatrix} 2.0 \\ -0.2 \end{bmatrix}$
- **Ridge:**  $\theta_{Ridge} = \begin{bmatrix} 0.565 \\ 0.261 \end{bmatrix}$

# Coefficient Magnitude Comparison

## Theorem: OLS vs Ridge Solutions

- **OLS:**  $\theta_{OLS} = \begin{bmatrix} 2.0 \\ -0.2 \end{bmatrix}$
- **Ridge:**  $\theta_{Ridge} = \begin{bmatrix} 0.565 \\ 0.261 \end{bmatrix}$

## L2 Norm Calculation

$$\|\theta_{OLS}\|_2^2 = (2.0)^2 + (-0.2)^2 = 4.04 \quad (13)$$

$$\|\theta_{Ridge}\|_2^2 = (0.565)^2 + (0.261)^2 = 0.387 \quad (14)$$

# Ridge Coefficient Shrinkage Result

## Important: Key Result

Ridge regression achieved a **90.4% reduction** in coefficient magnitude!

$$\frac{0.387}{4.04} = 0.096 \quad (\text{Ridge is 9.6\% of OLS magnitude})$$

## Key Points: Shrinkage Effect

Ridge systematically produces smaller coefficient magnitudes while maintaining prediction accuracy.

## Multi-collinearity

$(\mathbf{X}^T \mathbf{X})^{-1}$  is not computable when  $|\mathbf{X}^T \mathbf{X}| = 0$ .  
This was a drawback of using linear regression

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 4 \\ 1 & 3 & 6 \end{bmatrix}$$

The matrix  $\mathbf{X}$  is not full rank.

# Ridge Solution to Multi-collinearity

## Key Points: Ridge Advantage

With ridge regression, we invert  $\mathbf{X}^T\mathbf{X} + \mu\mathbf{I}$  instead of  $\mathbf{X}^T\mathbf{X}$

$$\mathbf{X}^T\mathbf{X} + \mu\mathbf{I} = \begin{bmatrix} 3 + \mu & 6 & 12 \\ 6 & 14 + \mu & 28 \\ 12 & 28 & 56 + \mu \end{bmatrix}$$

# Why Ridge Fixes Singularity

## Theorem: Key Result

The matrix  $\mathbf{X}^T\mathbf{X} + \mu\mathbf{I}$  is always full rank for  $\mu > 0$

# Why Ridge Fixes Singularity

## Theorem: Key Result

The matrix  $\mathbf{X}^T\mathbf{X} + \mu\mathbf{I}$  is always full rank for  $\mu > 0$

## Important: Another Interpretation

Ridge regression = regularization = fixing singularity issues!

# Why Ridge Fixes Singularity

## Theorem: Key Result

The matrix  $\mathbf{X}^T\mathbf{X} + \mu\mathbf{I}$  is always full rank for  $\mu > 0$

## Important: Another Interpretation

Ridge regression = regularization = fixing singularity issues!

## Key Points: Summary

Ridge regression elegantly handles multi-collinearity problems!



# The Intercept Penalty Problem

## Important: Critical Issue

Should we penalize the intercept  $\theta_0$  in ridge regression?

## Key Points: Two Approaches

- **Standard Ridge:**  $\hat{\theta} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$  (penalizes intercept)
- **No-intercept penalty:**  $\hat{\theta} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}^*)^{-1} \mathbf{X}^T \mathbf{y}$

## Modified Identity Matrix $\mathbf{I}^*$

Definition of  $\mathbf{I}^*$

$$\mathbf{I}^* = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

**Important: Key Point**

Zero in first position means NO penalty on intercept  $\theta_0$

# Demonstration: Two Simple Functions

## Example: Setup

Compare two functions with different intercepts:

- **Function 1:**  $f_1(x) = x$  (small intercept)
- **Function 2:**  $f_2(x) = x + 100$  (large intercept)

# Data Generation and Test Question

## Data Generation

For each function, generate data at  $x = 1, 2$ :

Function 1:  $(1, 1), (2, 2)$  (15)

Function 2:  $(1, 101), (2, 102)$  (16)

# Data Generation and Test Question

## Data Generation

For each function, generate data at  $x = 1, 2$ :

Function 1:  $(1, 1), (2, 2)$  (15)

Function 2:  $(1, 101), (2, 102)$  (16)

## Important: Test Question

How well can we predict  $y$  at  $x = 0$  using ridge regression with  $\lambda = 100$ ?

# Function 1: Setup and Data

**Theorem: Function 1:**  $y = x$

True value at  $x = 0$ :  $y = 0$

Data matrices

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

## Function 1: Matrix Computations

Matrix computations

$$\mathbf{X}^T \mathbf{X} = \begin{bmatrix} 2 & 3 \\ 3 & 5 \end{bmatrix} \quad (17)$$

$$\mathbf{X}^T \mathbf{y} = \begin{bmatrix} 3 \\ 5 \end{bmatrix} \quad (18)$$

## Function 1: Ridge with Standard $\mathbf{I}$

Standard Ridge:  $\mathbf{I}$  penalties both  $\theta_0$  and  $\theta_1$

$$\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I} = \begin{bmatrix} 2 & 3 \\ 3 & 5 \end{bmatrix} + 100 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 102 & 3 \\ 3 & 105 \end{bmatrix}$$



## Function 1: Standard Ridge Solution

Solution

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} 102 & 3 \\ 3 & 105 \end{bmatrix}^{-1} \begin{bmatrix} 3 \\ 5 \end{bmatrix} \quad (19)$$

$$\approx \begin{bmatrix} 0.029 \\ 0.047 \end{bmatrix} \quad (20)$$

## Function 1: Standard Ridge Solution

### Solution

$$\hat{\theta} = \begin{bmatrix} 102 & 3 \\ 3 & 105 \end{bmatrix}^{-1} \begin{bmatrix} 3 \\ 5 \end{bmatrix} \quad (19)$$

$$\approx \begin{bmatrix} 0.029 \\ 0.047 \end{bmatrix} \quad (20)$$

### Theorem: Prediction at $x = 0$

$$\hat{y}(0) = 0.029 + 0.047 \times 0 = 0.029$$

$$\text{Error: } |0.029 - 0| = 0.029$$

## Function 1: Ridge with Modified $\mathbf{I}^*$

Modified Ridge:  $\mathbf{I}^*$  does NOT penalize  $\theta_0$

$$\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}^* = \begin{bmatrix} 2 & 3 \\ 3 & 5 \end{bmatrix} + 100 \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 3 \\ 3 & 105 \end{bmatrix}$$

## Function 1: Modified Ridge Solution

Solution

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} 2 & 3 \\ 3 & 105 \end{bmatrix}^{-1} \begin{bmatrix} 3 \\ 5 \end{bmatrix} \quad (21)$$

$$\approx \begin{bmatrix} -0.001 \\ 0.048 \end{bmatrix} \quad (22)$$

## Function 1: Modified Ridge Solution

### Solution

$$\hat{\theta} = \begin{bmatrix} 2 & 3 \\ 3 & 105 \end{bmatrix}^{-1} \begin{bmatrix} 3 \\ 5 \end{bmatrix} \quad (21)$$

$$\approx \begin{bmatrix} -0.001 \\ 0.048 \end{bmatrix} \quad (22)$$

### Theorem: Prediction at $x = 0$

$$\hat{y}(0) = -0.001 + 0.048 \times 0 = -0.001$$

$$\text{Error: } |-0.001 - 0| = 0.001$$

## Function 2: Setup and Data

**Theorem: Function 2:**  $y = x + 100$

True value at  $x = 0$ :  $y = 100$

Data matrices

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 101 \\ 102 \end{bmatrix}$$

## Function 2: Matrix Computations

### Matrix computations

$$\mathbf{X}^T \mathbf{X} = \begin{bmatrix} 2 & 3 \\ 3 & 5 \end{bmatrix} \quad (\text{same as Function 1}) \quad (23)$$

$$\mathbf{X}^T \mathbf{y} = \begin{bmatrix} 203 \\ 305 \end{bmatrix} \quad (24)$$

## Function 2: Ridge with Standard I

Standard Ridge: penalizes large intercept heavily

$$\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I} = \begin{bmatrix} 102 & 3 \\ 3 & 105 \end{bmatrix} \quad (\text{same matrix})$$



## Function 2: Standard Ridge Solution

Solution

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} 102 & 3 \\ 3 & 105 \end{bmatrix}^{-1} \begin{bmatrix} 203 \\ 305 \end{bmatrix} \quad (25)$$

$$\approx \begin{bmatrix} 1.98 \\ 2.89 \end{bmatrix} \quad (26)$$

## Function 2: Standard Ridge Solution

### Solution

$$\hat{\theta} = \begin{bmatrix} 102 & 3 \\ 3 & 105 \end{bmatrix}^{-1} \begin{bmatrix} 203 \\ 305 \end{bmatrix} \quad (25)$$

$$\approx \begin{bmatrix} 1.98 \\ 2.89 \end{bmatrix} \quad (26)$$

### Theorem: Prediction at $x = 0$

$$\hat{y}(0) = 1.98 + 2.89 \times 0 = 1.98$$

**Error:**  $|1.98 - 100| = 98.02$  (TERRIBLE!)

## Function 2: Ridge with Modified $\mathbf{I}^*$

Modified Ridge: does NOT penalize intercept

$$\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}^* = \begin{bmatrix} 2 & 3 \\ 3 & 105 \end{bmatrix} \quad (\text{same as Function 1})$$

## Function 2: Modified Ridge Solution

Solution

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} 2 & 3 \\ 3 & 105 \end{bmatrix}^{-1} \begin{bmatrix} 203 \\ 305 \end{bmatrix} \quad (27)$$

$$\approx \begin{bmatrix} 99.91 \\ 1.05 \end{bmatrix} \quad (28)$$

## Function 2: Modified Ridge Solution

### Solution

$$\hat{\theta} = \begin{bmatrix} 2 & 3 \\ 3 & 105 \end{bmatrix}^{-1} \begin{bmatrix} 203 \\ 305 \end{bmatrix} \quad (27)$$

$$\approx \begin{bmatrix} 99.91 \\ 1.05 \end{bmatrix} \quad (28)$$

### Theorem: Prediction at $x = 0$

$$\hat{y}(0) = 99.91 + 1.05 \times 0 = 99.91$$

$$\text{Error: } |99.91 - 100| = 0.09 \text{ (EXCELLENT!)}$$

# Results Summary

Function	True $y(0)$	Standard I	Modified $I^*$
$f_1 : y = x$ Error	0	0.029 0.029	-0.001 0.001
$f_2 : y = x + 100$ Error	100	1.98 98.02	99.91 0.09

## Results Summary

Function	True $y(0)$	Standard I	Modified $I^*$
$f_1 : y = x$	0	0.029	-0.001
Error		0.029	0.001
$f_2 : y = x + 100$	100	1.98	99.91
Error		98.02	0.09

### Important: Key Insight

Penalizing the intercept creates **biased predictions** when data has non-zero mean!

## Results Summary

Function	True $y(0)$	Standard I	Modified $I^*$
$f_1 : y = x$	0	0.029	-0.001
Error		0.029	0.001
$f_2 : y = x + 100$	100	1.98	99.91
Error		98.02	0.09

### Important: Key Insight

Penalizing the intercept creates **biased predictions** when data has non-zero mean!

### Key Points: Solution

Use  $I^*$  to avoid penalizing the intercept, or normalize data first.



## Alternative: Data Normalization

### Theorem: Normalization Approach

Center the data to have zero mean, then use standard **I**

#### Function 2 with normalization

Original:  $(1, 101), (2, 102)$

Mean:  $\bar{x} = 1.5, \bar{y} = 101.5$

Centered:  $(-0.5, -0.5), (0.5, 0.5)$

# Benefits of Data Normalization

## Key Points: Why Normalize?

- Can use standard  $\mathbf{I}$  without bias
- Intercept becomes meaningful (deviation from mean)
- All features on similar scale
- More numerically stable

## Important: Best Practice

Always normalize data OR use  $\mathbf{I}^*$  for unbiased ridge regression!

# Implementation Details

# Ridge Regression via Gradient Descent

## Theorem: Gradient Descent Update Rule

Standard gradient descent step for ridge regression:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \alpha \nabla J(\boldsymbol{\theta}^{(t)})$$

# Ridge Regression via Gradient Descent

## Theorem: Gradient Descent Update Rule

Standard gradient descent step for ridge regression:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \alpha \nabla J(\boldsymbol{\theta}^{(t)})$$

## Ridge Gradient Computation

$$\nabla J(\boldsymbol{\theta}) = \nabla \left[ \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \frac{\lambda}{2} \|\boldsymbol{\theta}\|_2^2 \right] \quad (29)$$

$$= -\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \lambda\boldsymbol{\theta} \quad (30)$$

$$= -\mathbf{X}^T\mathbf{y} + \mathbf{X}^T\mathbf{X}\boldsymbol{\theta} + \lambda\boldsymbol{\theta} \quad (31)$$

## Ridge vs OLS: Gradient Descent Updates

### Theorem: Ridge Update (with shrinkage)

$$\begin{aligned}\boldsymbol{\theta}^{(t+1)} &= \boldsymbol{\theta}^{(t)} - \alpha(-\mathbf{X}^T \mathbf{y} + \mathbf{X}^T \mathbf{X} \boldsymbol{\theta}^{(t)} + \lambda \boldsymbol{\theta}^{(t)}) \\ &= (1 - \alpha\lambda) \boldsymbol{\theta}^{(t)} - \alpha(-\mathbf{X}^T \mathbf{y} + \mathbf{X}^T \mathbf{X} \boldsymbol{\theta}^{(t)})\end{aligned}$$

## Ridge vs OLS: Gradient Descent Updates

### Theorem: Ridge Update (with shrinkage)

$$\begin{aligned}\boldsymbol{\theta}^{(t+1)} &= \boldsymbol{\theta}^{(t)} - \alpha(-\mathbf{X}^T \mathbf{y} + \mathbf{X}^T \mathbf{X} \boldsymbol{\theta}^{(t)} + \lambda \boldsymbol{\theta}^{(t)}) \\ &= (1 - \alpha\lambda) \boldsymbol{\theta}^{(t)} - \alpha(-\mathbf{X}^T \mathbf{y} + \mathbf{X}^T \mathbf{X} \boldsymbol{\theta}^{(t)})\end{aligned}$$

### Theorem: OLS Update (no shrinkage)

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \alpha(-\mathbf{X}^T \mathbf{y} + \mathbf{X}^T \mathbf{X} \boldsymbol{\theta}^{(t)})$$

## Ridge vs OLS: Gradient Descent Updates

### Theorem: Ridge Update (with shrinkage)

$$\begin{aligned}\boldsymbol{\theta}^{(t+1)} &= \boldsymbol{\theta}^{(t)} - \alpha(-\mathbf{X}^T \mathbf{y} + \mathbf{X}^T \mathbf{X} \boldsymbol{\theta}^{(t)} + \lambda \boldsymbol{\theta}^{(t)}) \\ &= (1 - \alpha\lambda) \boldsymbol{\theta}^{(t)} - \alpha(-\mathbf{X}^T \mathbf{y} + \mathbf{X}^T \mathbf{X} \boldsymbol{\theta}^{(t)})\end{aligned}$$

### Theorem: OLS Update (no shrinkage)

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \alpha(-\mathbf{X}^T \mathbf{y} + \mathbf{X}^T \mathbf{X} \boldsymbol{\theta}^{(t)})$$

### Key Points: Key Insight

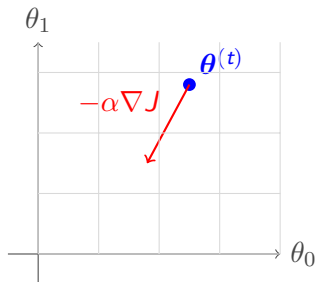
The  $(1 - \alpha\lambda)$  factor **shrinks** coefficients at each step!



## Visual: OLS Gradient Descent Step

### Theorem: OLS Update

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \alpha \nabla J(\boldsymbol{\theta}^{(t)})$$



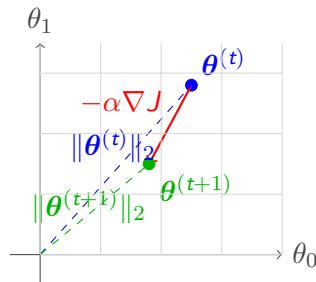
### Important: Step 1

Start at  $\boldsymbol{\theta}^{(t)}$  and compute negative gradient direction

# Visual: OLS Gradient Descent - Vector Sum

## Theorem: Vector Addition

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + (-\alpha \nabla J)$$



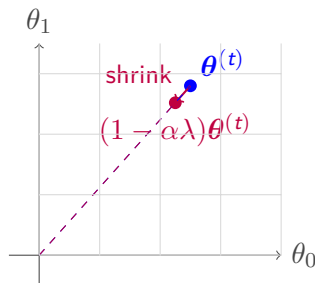
## Key Points: Result

OLS:  $\|\boldsymbol{\theta}^{(t+1)}\|_2$  depends only on gradient direction

## Visual: Ridge Gradient Descent - Shrinkage Step

**Theorem: Ridge Shrinkage**

First:  $\theta^{(t)} \rightarrow (1 - \alpha\lambda)\theta^{(t)}$



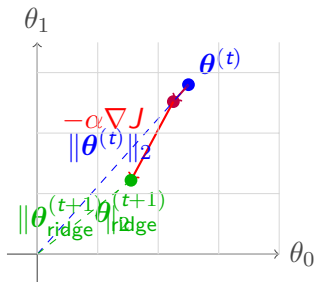
**Important: Ridge Step 1**

Shrink current parameters by factor  $(1 - \alpha\lambda) < 1$

## Visual: Ridge Gradient Descent - Complete Update

**Theorem: Ridge Complete Update**

$$\boldsymbol{\theta}^{(t+1)} = (1 - \alpha\lambda)\boldsymbol{\theta}^{(t)} - \alpha\nabla J$$

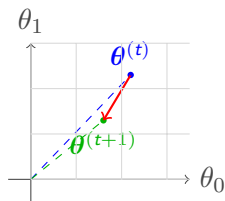


**Key Points: Key Insight**

Ridge:  $\|\boldsymbol{\theta}_{ridge}^{(t+1)}\|_2 < \|\boldsymbol{\theta}_{OLS}^{(t+1)}\|_2$  (smaller coefficients!)

# Side-by-Side Comparison: OLS vs Ridge Updates

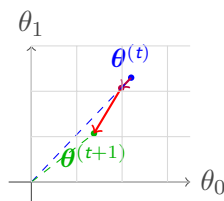
## OLS Gradient Descent



No shrinkage

$$\|\theta^{(t+1)}\|_2 = 1.98$$

## Ridge Gradient Descent



With shrinkage

$$\|\theta^{(t+1)}\|_2 = 1.72 < \text{OLS}$$

### Important: Ridge Effect

Ridge regression systematically produces **smaller coefficient magnitudes** at every gradient descent step!

# Summary: What We Learned

## Key Points: Ridge Regression Key Points

- **Problem:** Overfitting in linear regression with large coefficients
- **Solution:** Add  $L_2$  penalty  $\lambda \|\boldsymbol{\theta}\|_2^2$  to loss function
- **Effect:** Shrinks coefficients, improves generalization
- **Trade-off:** Higher bias, lower variance

## Key Formula & Next Steps

### Theorem: Ridge Regression Solution

$$\hat{\boldsymbol{\theta}}_{\text{ridge}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$$

# Key Formula & Next Steps

## Theorem: Ridge Regression Solution

$$\hat{\theta}_{\text{ridge}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$$

## Important: Next Steps

- Compare with Lasso regression ( $L_1$  penalty)
- Explore elastic net (combines  $L_1$  and  $L_2$ )
- Apply to real-world datasets