Ridge Regression

Nipun Batra

IIT Gandhinagar

September 9, 2025

Outline

- Motivation: The Problem of Overfitting
- Ridge Regression Formulation
- Mathematical Derivation
- Geometric Interpretation
- Hyperparameter Selection
- Examples and Applications
- Implementation Details

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:

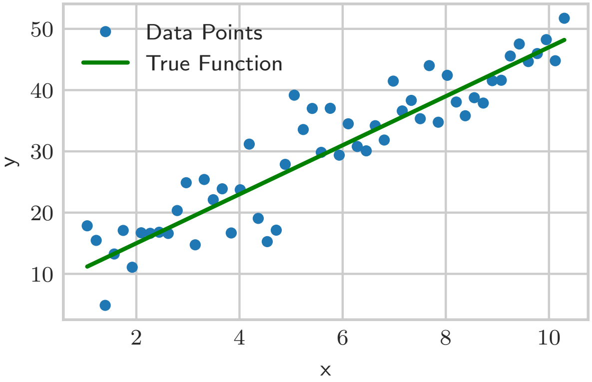
- Training error decreases
- Test error increases
- Model coefficients become very large

Key Points: Key Insight

Large coefficient magnitudes often indicate overfitting!

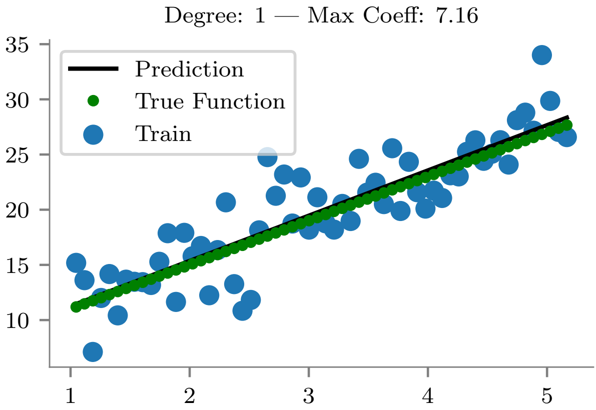
In polynomial $f(x) = c_0 + c_1 x + c_2 x^2 + \cdots + c_d x^d$, watch $\max |c_i|$

Demonstration: Polynomial Degree vs Overfitting



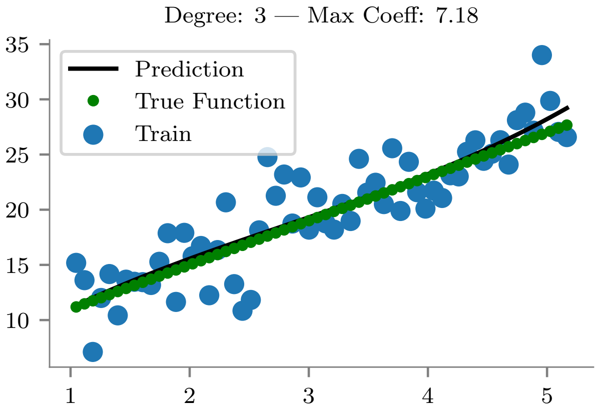
Base Data Set

Demonstration: Polynomial Degree vs Overfitting



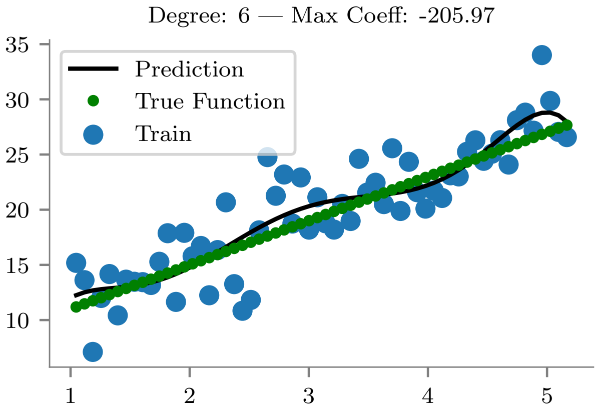
Fit with Degree 1 - Underfitting

Demonstration: Polynomial Degree vs Overfitting



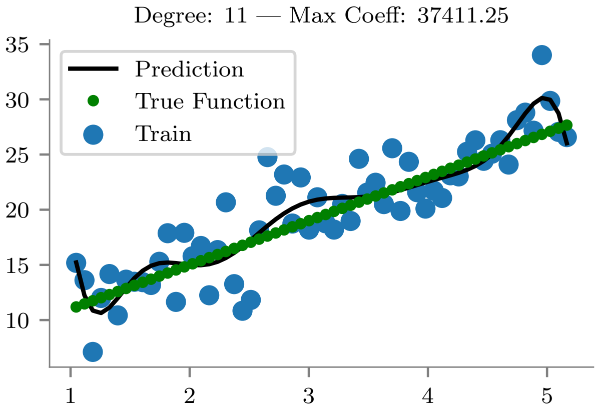
Fit with Degree 3 - Good Fit

Demonstration: Polynomial Degree vs Overfitting



Fit with Degree 6 - Starting to Overfit

Demonstration: Polynomial Degree vs Overfitting

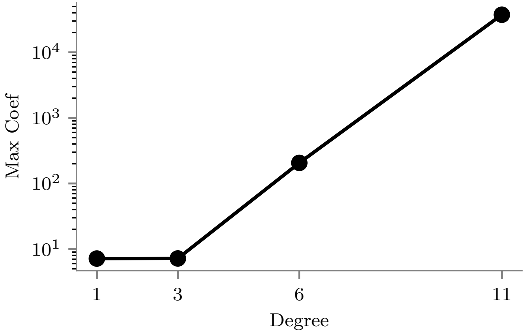


Fit with Degree 11 - Severe Overfitting

Coefficient Explosion with Overfitting

Key Points: Key Observation

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



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 $>>$ test error
D) Overfitting occurs when training error $<<$ test error

Answer: Pop Quiz 1

Answer this!

D) Overfitting occurs when training error $<<$ test error

Explanation:

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

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

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

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

Understanding the Ridge Penalty

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

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

Key Points: Key Components

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

Effect of Regularization Parameter λ

Key Points: Parameter Effects

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

Effect of Regularization Parameter λ

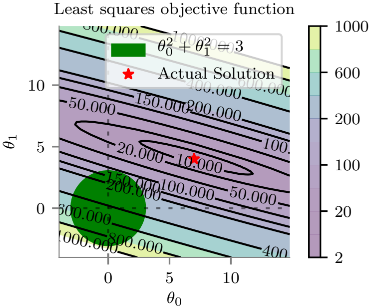
Key Points: Parameter Effects

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

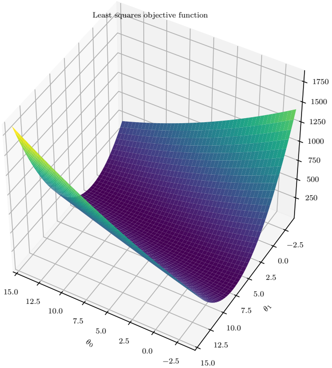
Important: Key Trade-off

Higher λ = more regularization = more bias, less variance

Geometric Interpretation



(a) Contour Plot



(b) Surface Plot

Ridge regression finds solution where error contours touch constraint circle

Key Points: Geometric Insight

Mathematical Derivation: Step 1

Step 1: Set up the Lagrangian

For the constrained optimization problem:

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

The Lagrangian is:

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

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

Mathematical Derivation: Step 2

Step 2: Apply KKT Conditions

For optimality, we need:

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

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

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

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

Mathematical Derivation: Step 2

Step 2: Apply KKT Conditions

For optimality, we need:

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

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

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

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

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 $\boldsymbol{\theta}$:

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

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

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

Mathematical Derivation: Step 4

Step 4: Set Gradient to Zero

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

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

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

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

Mathematical Derivation: Step 4

Step 4: Set Gradient to Zero

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

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

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

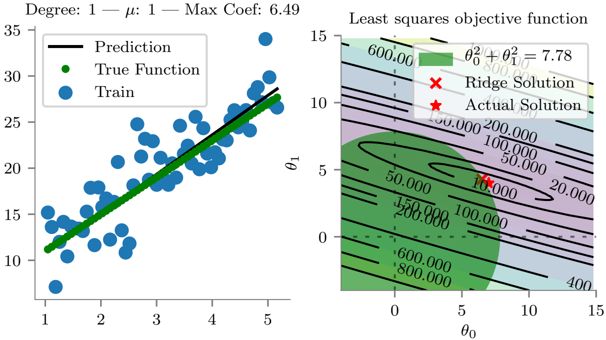
$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\boldsymbol{\theta} = \mathbf{X}^T\mathbf{y} \tag{12}$$

Theorem: Ridge Regression Solution

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

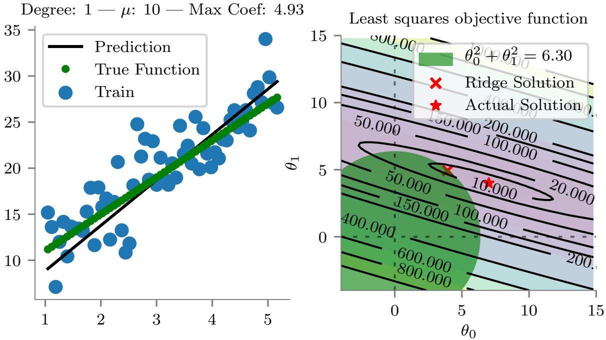
Compare with OLS: $\hat{\boldsymbol{\theta}}_{\text{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

Effect of Regularization Parameter λ



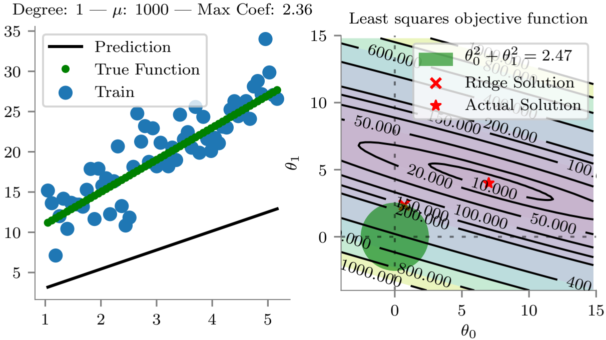
$\lambda = 1$ - Mild Regularization

Effect of Regularization Parameter λ



$\lambda = 10$ - Moderate Regularization

Effect of Regularization Parameter λ



$\lambda = 1000$ - Heavy Regularization

Pop Quiz 2

Answer this!

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

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

Answer: Pop Quiz 2

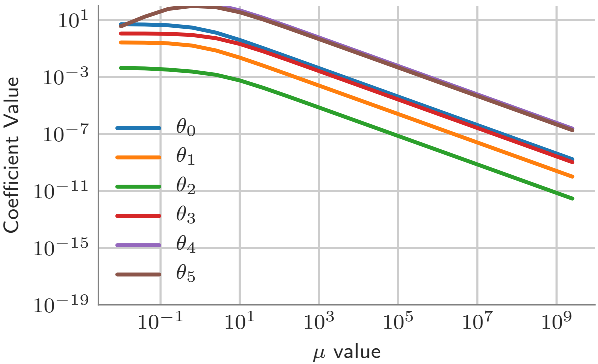
Answer this!

B) Coefficients approach zero

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

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

Coefficient Shrinkage: Visual Evidence

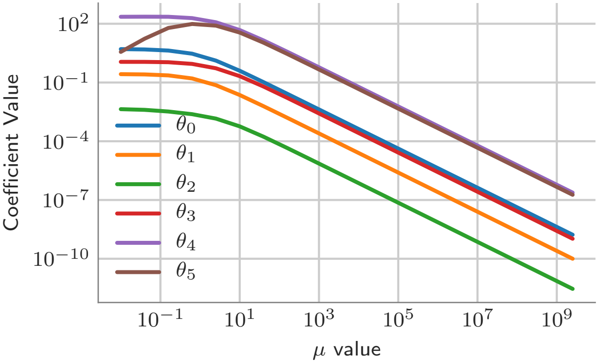


Coefficient Magnitudes vs λ (Real Estate Dataset)

Important: Important Question

Do coefficients ever become exactly zero?

Ridge vs. Lasso: Coefficient Behavior



Ridge Coefficients Shrink but Never Reach Zero

Key Points: Key Difference

- **Ridge (L_2):** Coefficients shrink toward zero but remain non-zero

Key Properties of Ridge Regression

Theorem: Ridge Solution Properties

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

Key Properties of Ridge Regression

Theorem: Ridge Solution Properties

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

Key Points: Important Properties

1. **Always invertible:** $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})$ is positive definite for $\lambda > 0$
2. **Shrinkage:** Coefficients are shrunk toward zero
3. **Bias-Variance trade-off:** Increases bias, reduces variance
4. **Computational efficiency:** Closed-form solution available

Choosing the Regularization Parameter λ

Important: Hyperparameter Selection

How do we choose the optimal value of λ?

Choosing the Regularization Parameter λ

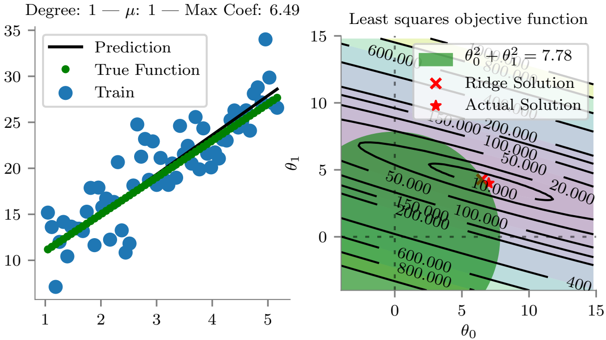
Important: Hyperparameter Selection

How do we choose the optimal value of λ?

Theorem: Cross-Validation Approach

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

Cross-Validation for Ridge Regression



Cross-validation curve for Ridge regression showing optimal λ

Key Points: CV Pattern

- Small λ: High variance (overfitting)
- Large λ: High bias (underfitting)

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

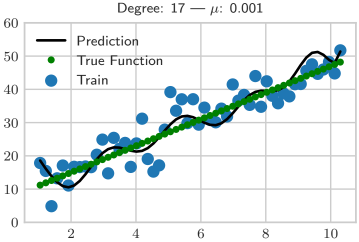
Theorem: Bias-Variance Decomposition

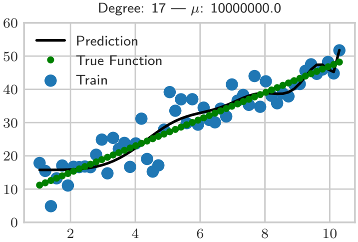
$$\text{Total Error} = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$

Key Points: Ridge Effect

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

Small vs Large Regularization





Small λ ($\lambda \to 0$):

- Low bias
- High variance
- Risk of overfitting

Large λ ($\lambda \to \infty$):

- High bias
- Low variance
- Risk of underfitting

Pop Quiz 3

Answer this!

In ridge regression, as we increase λ, 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 λ 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!

Worked Example: Setup

Example: Ridge Regression Example

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

Data: $(x_1, y_1) = (1, 1)$, $(x_2, y_2) = (2, 3)$, $(x_3, y_3) = (3, 2)$, $(x_4, y_4) = (4, 4)$
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: $(x_1, y_1) = (1, 1)$, $(x_2, y_2) = (2, 3)$, $(x_3, y_3) = (3, 2)$, $(x_4, y_4) = (4, 4)$
Model: $y = \theta_0 + \theta_1 x$

Step 1: Set up matrices

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

Worked Example: OLS Setup

Step 2: Ordinary Least Squares

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

Worked Example: OLS Setup

Step 2: Ordinary Least Squares

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

Step 3: Compute matrix products

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

$$\mathbf{X}^T\mathbf{y} = \begin{bmatrix} 10 \\ 28 \end{bmatrix}$$

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

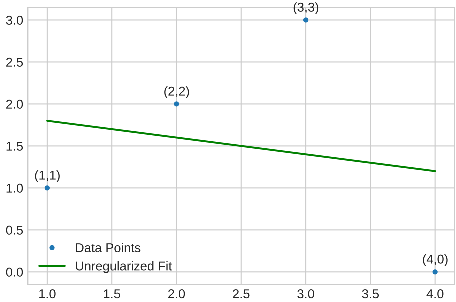
Step 5: Final matrix multiplication

$$
\begin{aligned}
\hat{\boldsymbol{\theta}}_{\mathsf{OLS}} &= (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{y}) \\
&= \frac{1}{20}\begin{bmatrix} 30 & -10 \\ -10 & 4 \end{bmatrix}\begin{bmatrix} 10 \\ 28 \end{bmatrix} \\
&= \frac{1}{20}\begin{bmatrix} 20 \\ 12 \end{bmatrix} = \begin{bmatrix} 1.0 \\ 0.6 \end{bmatrix}
\end{aligned}
$$

OLS Final Result

Theorem: OLS Result

$\hat{y} = 1.0 + 0.6x$ (No regularization)



Worked Example: Ridge Regression 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 Regression 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

$$\begin{aligned} \mathbf{X}^T\mathbf{X} + \lambda\mathbf{I} &= \begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix} + 2\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix} + \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \\ &= \begin{bmatrix} 6 & 10 \\ 10 & 32 \end{bmatrix} \end{aligned}$$

Worked Example: Ridge Regression Setup

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

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

Step 6: Add regularization term

$$\begin{aligned}
\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I} &= \begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix} + 2\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix} + \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \\
&= \begin{bmatrix} 6 & 10 \\ 10 & 32 \end{bmatrix}
\end{aligned}$$

Note: $\det(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}) = 6 \cdot 32 - 10 \cdot 10 = 192 - 100 = 92$

Worked Example: Ridge Result

Step 7: Final Ridge solution

$$
\begin{aligned}
\hat{\boldsymbol{\theta}}_{\text{ridge}} &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{y}) \\
&= \frac{1}{92}\begin{bmatrix} 32 & -10 \\ -10 & 6 \end{bmatrix}\begin{bmatrix} 10 \\ 28 \end{bmatrix} \\
&= \frac{1}{92}\begin{bmatrix} 32 \cdot 10 + (-10) \cdot 28 \\ (-10) \cdot 10 + 6 \cdot 28 \end{bmatrix} \\
&= \frac{1}{92}\begin{bmatrix} 320 - 280 \\ -100 + 168 \end{bmatrix} = \frac{1}{92}\begin{bmatrix} 40 \\ 68 \end{bmatrix} \\
&= \begin{bmatrix} 0.435 \\ 0.739 \end{bmatrix}
\end{aligned}
$$

Worked Example: Ridge Result

Step 7: Final Ridge solution

$$
\begin{aligned}
\hat{\boldsymbol{\theta}}_{\text{ridge}} &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{y}) \\
&= \frac{1}{92}\begin{bmatrix} 32 & -10 \\ -10 & 6 \end{bmatrix}\begin{bmatrix} 10 \\ 28 \end{bmatrix} \\
&= \frac{1}{92}\begin{bmatrix} 32 \cdot 10 + (-10) \cdot 28 \\ (-10) \cdot 10 + 6 \cdot 28 \end{bmatrix} \\
&= \frac{1}{92}\begin{bmatrix} 320 - 280 \\ -100 + 168 \end{bmatrix} = \frac{1}{92}\begin{bmatrix} 40 \\ 68 \end{bmatrix} \\
&= \begin{bmatrix} 0.435 \\ 0.739 \end{bmatrix}
\end{aligned}
$$

Theorem: Ridge Result

$\hat{y} = 0.435 + 0.739x$ (With $\lambda = 2$ regularization)

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.

Multi-collinearity

But with ridge regression, the matrix to be inverted is $\mathbf{X}^T\mathbf{X} + \mu\mathbf{I}$ and not $\mathbf{X}^T\mathbf{X}$.

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

The matrix $\mathbf{X}^T\mathbf{X}$ would be full rank for $\mu > 0$.

Multi-collinearity

But with ridge regression, the matrix to be inverted is $\mathbf{X}^T\mathbf{X} + \mu\mathbf{I}$ and not $\mathbf{X}^T\mathbf{X}$.

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

The matrix $\mathbf{X}^T\mathbf{X}$ would be full rank for $\mu > 0$.
Another interpretation of "regularisation"

Extension of the analytical model

For ridge with no penalty on θ_0

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{X}^T\mathbf{X} + \mu\mathbf{I}^*\right)^{-1}\mathbf{X}^T\mathbf{y}$$

where,

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

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

$$\begin{aligned}
\nabla J(\boldsymbol{\theta}) &= \nabla \left[\frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \frac{\lambda}{2} \|\boldsymbol{\theta}\|_2^2 \right] && (13) \\
&= -\mathbf{X}^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \lambda \boldsymbol{\theta} && (14) \\
&= -\mathbf{X}^T \mathbf{y} + \mathbf{X}^T \mathbf{X}\boldsymbol{\theta} + \lambda \boldsymbol{\theta} && (15)
\end{aligned}$$

Ridge vs OLS: Gradient Descent Updates

Theorem: Ridge Update (with shrinkage)

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

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!

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

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

Important: Next Steps

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