Ridge Regression – Geometric Intuition and Theory

1. Introduction

- **Linear Regression** works well but struggles in certain conditions:
    1. When features are **highly correlated (multicollinearity)**.
    2. When there are **many features** compared to the number of data points.
    3. When data contains **noise**, leading to unstable coefficients.
- **Ridge Regression** solves these issues by adding **regularization** to the model.

2. Linear Regression Recap

The ordinary least squares (OLS) regression formula:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

The cost function minimized is:

$$J(\beta) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

3. Problem with OLS

- If features are highly correlated:
    1. OLS estimates β\beta become **unstable**.
    2. Small changes in data → large swings in coefficients.
- Leads to **overfitting** and poor generalization.

4. What is Ridge Regression?

- Ridge Regression modifies the OLS cost function by adding a **penalty term** proportional to the **squared magnitude of coefficients**.

$$J(\beta) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Where:

- $\lambda \geq 0$ = regularization parameter.
- Larger λ → stronger penalty on large coefficients.

5. Effect of Regularization

- When we add $\lambda \sum \beta_j^2$: :
  1. The model is **forced to keep coefficients small**, reducing overfitting.
  2. Coefficients are **shrunk towards zero** but never become exactly zero (unlike Lasso regression).
- This **biases the model slightly**, but **reduces variance**, improving generalization.

---

6. Geometric Intuition

- **OLS Solution:**
  1. Finds β\beta that minimizes squared errors.
  2. Solution may be large coefficients if features are correlated.
- **Ridge Solution:**
  1. Adds a **constraint**:

$$\sum \beta_j^2 \leq c$$

  2. Geometrically:
     - OLS solution is the intersection of **elliptical contours** of the loss function with **constraint boundary** (a circle/ellipse).
     - Ridge regression **shrinks the solution toward the origin**, preventing extreme coefficient values.

---

7. Impact of λ\lambda

- $\lambda = 0$: Ridge = Linear Regression (no penalty).
- Small λ: Slight shrinkage, reduces variance.
- Large λ: Coefficients approach zero, model becomes too simple (possible underfitting).

---

8. Advantages of Ridge Regression

- Handles **multicollinearity** effectively.
- Reduces **model complexity**, preventing overfitting.
- Improves **prediction accuracy** on unseen data.

---

9. Limitations

- Does not perform **feature selection** (all coefficients remain non-zero).
- If many irrelevant features exist, **Lasso regression** may be better.

🧠 Key Takeaways

- Ridge Regression = **Linear Regression + L2 penalty** on coefficients.
- Cost function:

$$J(\beta) = \mathrm{MSE} + \lambda \sum \beta_j^2$$

- Helps prevent overfitting and stabilizes regression in high-dimensional or correlated feature spaces.
- Geometric intuition: Ridge **constrains the solution space**, shrinking coefficients toward zero but not eliminating them.
