Ridge Regression – Mathematical Formulation

1. Revision of Ridge Regression

- Ordinary Linear Regression minimizes:

$$J(\beta) = (y - X\beta)^T (y - X\beta)$$

- Ridge Regression modifies this by adding **L2 regularization**:

$$J(\beta) = (y - X\beta)^T (y - X\beta) + \lambda \beta^T \beta$$

Where:

- X = input feature matrix $(n \times p)$
- y = target vector $(n \times 1)$
- β = coefficient vector $(p \times 1)$
- λ = regularization parameter ($\lambda \geq 0$)

2. Goal

Q: Find β that minimizes J(β)

3. Step-by-Step Derivation

Step 1: Expand the cost function

$$J(\beta) = (y - X\beta)^T (y - X\beta) + \lambda \beta^T \beta$$

Expand the first term:

$$J(\beta) = y^T y - 2\beta^T X^T y + \beta^T X^T X \beta + \lambda \beta^T \beta$$

Step 2: Take derivative w.r.t. β

We compute:

$$\frac{\partial J}{\partial \beta} = -2X^T y + 2X^T X \beta + 2\lambda \beta$$

Step 3: Set derivative = 0

$$-2X^Ty + 2X^TX\beta + 2\lambda\beta = 0$$

Divide through by 2:

$$X^TX\beta + \lambda\beta = X^Ty$$

Step 4: Factorize terms

$$(X^TX + \lambda I)\beta = X^Ty$$

Where I is the identity matrix ($p \times p$).

Step 5: Solve for β

$$\boxed{\beta = (X^TX + \lambda I)^{-1}X^Ty}$$

---

4. Interpretation

- Ridge Regression solution is similar to Ordinary Least Squares (OLS):

$$\beta_{OLS} = (X^TX)^{-1}X^Ty$$

But with a **regularization term** λI added before inversion.

- This makes $X^TX + \lambda I$ **non-singular**, avoiding issues with multicollinearity (when X^TX is close to singular).

---

5. Ridge Regression for N-Dimensional Data

- Works the same way for multiple features (p > 1).
- The regularization term **shrinks all coefficients simultaneously**, reducing their magnitude but not making them exactly zero.

---

6. Effect of λ

- If $\lambda = 0$: Ridge reduces to Linear Regression.
- If $\lambda \to \infty$: All coefficients $\beta \approx 0$.
- Choosing λ is crucial:
  - Use **Cross-Validation** to find the optimal value.

🧠 Key Takeaways

- Ridge Regression solves:

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

- Adds **L2 penalty** to reduce overfitting and handle multicollinearity.
- Stabilizes regression solution when features are correlated or dataset is ill-conditioned.
