5 Key Points | Ridge Regression | Regularized Linear Models

1 How the coefficients get affected?

- In **Linear Regression**, coefficients (β) can become **very large**, especially when:
    1. There is **multicollinearity** (features are highly correlated).
    2. The dataset is **small or noisy**.
- **Ridge Regression** adds an **L2 penalty** ($\lambda \sum \beta^2$) to the cost function:

$$J(\beta) = \text{MSE} + \lambda \sum_{j=1}^{p} \beta_j^2$$

- This penalty **discourages large values of coefficients**:
    1. The model still tries to fit the data but **keeps weights small**, leading to a **simpler, more stable model**.

2 Higher values are impacted more

- Ridge penalizes **large coefficients more strongly** because the penalty term is squared (β_j^2):

$$\lambda \beta_j^2$$

- If:
    - $\beta_1 = 0.5$ → penalty = 0.25
    - $\beta_2 = 5$ → penalty = 25
- Meaning:
    - **Big coefficients are heavily reduced**, preventing the model from giving extreme importance to single feature.
    - **Small coefficients are only slightly affected**, preserving weaker but useful signals.

3 Impact on Bias-Variance Trade-off

- **Bias:** Error from incorrect assumptions (underfitting).
- **Variance:** Error from sensitivity to training data (overfitting).
- Ridge Regression:
    1. **Increases bias slightly** (coefficients are shrunk, model is simpler).
    2. **Reduces variance a lot** (model becomes more stable and generalizes better).
- Net effect:

1. **Lower overall error** on unseen data (better generalization).
2. **Trade-off:** We accept a small increase in bias to achieve a big decrease in variance.

---

4 Effect on the Loss Function

- Original Linear Regression minimizes:

$$J(\beta) = \sum (y_i - \hat{y}_i)^2$$

- Ridge modifies this to:

$$J(\beta) = \sum (y_i - \hat{y}_i)^2 + \lambda \sum \beta_j^2$$

Where:

- First term = **error** between predictions and actual values.
- Second term = **penalty** on large coefficients.

This:

- Forces model to **balance between accuracy and simplicity**.
- Even if a feature perfectly fits noise, the **penalty prevents its coefficient from blowing up**.

---

5 Why Ridge Regression is called so?

- The name "Ridge" comes from the **constraint region** of the solution:
  1. Ridge regression restricts coefficients inside a **circular (elliptical) boundary**:

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

- Geometrically:
  1. Solution is forced to stay near the **origin (small coefficients)**.
  2. This boundary forms a **ridge-shaped area** in the solution space.
  3. The optimizer finds the point where:
     - **Loss function contour** (ellipse) touches the **ridge constraint**.

---

6 A Practical Tip – Apply Ridge Regression

- Use Ridge Regression when:
  1. Dataset has **multicollinearity** (correlated features).
  2. Number of features **p > n** (more features than data points).

3. Model **overfits training data**.
4. You want to **reduce variance and improve generalization**.

- In practice:
  1. Use **cross-validation** to select the best λ\\lambda (regularization strength).
  2. Too high λ\\lambda → underfitting.
  3. Too low λ\\lambda → behaves like linear regression.

🎯 Summary of 5 Key Understandings

Key Point	Effect in Ridge Regression
1️⃣ **Coefficients get affected**	Shrunk toward zero, more stable
2️⃣ **Higher values impacted more**	Large weights penalized more heavily
3️⃣ **Bias-Variance Trade-off**	Increases bias slightly, greatly reduces variance
4️⃣ **Loss function effect**	Adds L2 penalty, balancing accuracy and simplicity
5️⃣ **Why "Ridge" name?**	Solution constrained in a “ridge” (circle/ellipse) region

---