Gradient Descent for Ridge-Penalized Quadratic Regression

Model Setup

We consider the quadratic regression model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \tag{1}$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. We introduce a ridge penalty on β_2. The objective function to minimize is:

$$L(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2) \right)^2 + \lambda \beta_2^2 \tag{2}$$

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

Gradient Descent Update Rule

At each iteration t, we update the parameters using:

$$\beta^{(t+1)} = \beta^{(t)} - \eta \nabla_\beta L(\beta^{(t)}) \tag{3}$$

where:

- η is the learning rate,
- $\nabla_\beta L(\beta)$ is the gradient of the loss function with respect to β.

Computing the Gradients

Define:

$$\hat{y}_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 \tag{4}$$

$$e_i = \hat{y}_i - y_i \quad \text{(prediction error)} \tag{5}$$

Then, the partial derivatives are:

$$\frac{\partial L}{\partial \beta_0} = 2 \cdot \frac{1}{n} \sum_{i=1}^{n} e_i \tag{6}$$

$$\frac{\partial L}{\partial \beta_1} = 2 \cdot \frac{1}{n} \sum_{i=1}^{n} e_i x_i \tag{7}$$

$$\frac{\partial L}{\partial \beta_2} = 2 \cdot \frac{1}{n} \sum_{i=1}^{n} e_i x_i^2 + 2\lambda\beta_2 \tag{8}$$

Thus, the gradient vector is:

$$\nabla_\beta L(\beta) = \begin{bmatrix} 2 \cdot \text{mean}(e) \\ 2 \cdot \text{mean}(e \cdot x) \\ 2 \cdot \text{mean}(e \cdot x^2) + 2\lambda\beta_2 \end{bmatrix} \tag{9}$$

Intuition

Each partial derivative measures how sensitive the loss function is to changes in a particular parameter:

- $\partial L/\partial \beta_0$: How the loss changes with a shift in intercept.
- $\partial L/\partial \beta_1$: How the loss changes with a change in the linear term.
- $\partial L/\partial \beta_2$: How the loss changes with a change in the quadratic term, including the effect of the ridge penalty.

The gradient descent update moves the parameters in the negative gradient direction to reduce the loss, and the penalty term shrinks β_2 to control model complexity.