Gradient Descent for ReLU Regression

Problem Setup

We are given:

- **Prediction function:** $\hat{y} = f(x; \theta) = \text{ReLU}(\theta^\top x)$
- **Loss function:** $L(\hat{y}, y) = (\hat{y} - y)^2$
- **Parameters:** $\theta = [w_0, w_1, w_2]$

Section (a): Symbolic Gradient Calculation

Our goal is to find the partial derivatives of the loss function $L(\hat{y}, y)$ with respect to each parameter in $\theta = [w_0, w_1, w_2]$.

Step 1: Compute $\frac{\partial L}{\partial \hat{y}}$

Since $L(\hat{y}, y) = (\hat{y} - y)^2$, we find:

$$\frac{\partial L}{\partial \hat{y}} = 2(\hat{y} - y)$$

Step 2: Compute $\frac{\partial \hat{y}}{\partial w_i}$

We have $\hat{y} = \text{ReLU}(\theta^\top x)$, where:

$$\theta^\top x = w_0 \cdot x_0 + w_1 \cdot x_1 + w_2 \cdot x_2$$

with $x = [x_0, x_1, x_2]$, where $x_0 = 1$ is the bias term. The derivative of the ReLU function is:

$$\text{ReLU}'(z) = \begin{cases} 1, & \text{if } z > 0 \\ 0, & \text{if } z \leq 0 \end{cases}$$

Thus,

$$\frac{\partial \hat{y}}{\partial w_i} = \text{ReLU}'(\theta^\top x) \cdot \frac{\partial(\theta^\top x)}{\partial w_i}$$

Since $\text{ReLU}'(\theta^\top x) = 1$ for $\theta^\top x > 0$ and 0 otherwise, we have:

$$\frac{\partial \hat{y}}{\partial w_i} = \begin{cases} x_i, & \text{if } \theta^\top x > 0 \\ 0, & \text{if } \theta^\top x \leq 0 \end{cases}$$

Step 3: Apply the Chain Rule

Using the chain rule:

$$\frac{\partial L}{\partial w_i} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_i} = 2(\hat{y} - y) \cdot x_i$$

Therefore, the partial derivatives for each component of θ are:

$$\frac{\partial L}{\partial w_0} = 2(\text{ReLU}(\theta^\top x) - y)x_0$$

$$\frac{\partial L}{\partial w_1} = 2(\text{ReLU}(\theta^\top x) - y)x_1$$

$$\frac{\partial L}{\partial w_2} = 2(\text{ReLU}(\theta^\top x) - y)x_2$$

These expressions represent the symbolic form of the gradient $\nabla_\theta L(f(x;\theta), y)$.

Section (b): Numerical Calculation with Given Values

Given:

- Initial parameter values: $w_0^0 = -1$, $w_1^0 = 1$, $w_2^0 = 0$
- Input $x = [1, 3, 2]$ and $y = 1$
- Learning rate $\eta = 0.1$

Step 1: Compute $\hat{y} = f(x;\theta)$

$\theta^\top x = w_0 \cdot x_0 + w_1 \cdot x_1 + w_2 \cdot x_2 = -1 \cdot 1 + 1 \cdot 3 + 0 \cdot 2 = -1 + 3 + 0 = 2$

Since $\text{ReLU}(2) = 2$, we have:

$$\hat{y} = f(x;\theta) = \text{ReLU}(2) = 2$$

Step 2: Compute the Loss $L(\hat{y}, y)$

Substitute $\hat{y} = 2$ and $y = 1$:

$$L(\hat{y}, y) = (2 - 1)^2 = 1$$

Step 3: Compute the Gradient $\nabla_\theta L(f(x;\theta), y)$

Since $\mathrm{ReLU}(\theta^\top x) = 2$ and $\mathrm{ReLU}'(\theta^\top x) = 1$ (because $\theta^\top x > 0$), we use $\frac{\partial L}{\partial w_i} = 2(\hat{y} - y) \cdot x_i$:

$$\frac{\partial L}{\partial w_0} = 2(2-1) \cdot 1 = 2$$

$$\frac{\partial L}{\partial w_1} = 2(2-1) \cdot 3 = 6$$

$$\frac{\partial L}{\partial w_2} = 2(2-1) \cdot 2 = 4$$

Step 4: Update θ with Gradient Descent

Now we can update each component of θ using the learning rate $\eta = 0.1$:

$$w_0^1 = w_0^0 - \eta \cdot \frac{\partial L}{\partial w_0} = -1 - 0.1 \cdot 2 = -1.2$$

$$w_1^1 = w_1^0 - \eta \cdot \frac{\partial L}{\partial w_1} = 1 - 0.1 \cdot 6 = 0.4$$

$$w_2^1 = w_2^0 - \eta \cdot \frac{\partial L}{\partial w_2} = 0 - 0.1 \cdot 4 = -0.4$$

Therefore, the updated parameter vector after this single step of gradient descent is:

$$\theta^1 = [w_0^1, w_1^1, w_2^1] = [-1.2, 0.4, -0.4]$$