Mean Squared Error (MSE) is a common loss function used in regression tasks to measure the average squared difference between the predicted values and the actual values.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where:

- n is the number of data points,
- y_i is the actual value of the dependent variable for data point i,
- $\hat{y}_i$ is the predicted value of the dependent variable for data point i.

for many iterations, any changes on w_t and b_t, the MSE shall be:

$$\text{MSE}_t = \left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right)_t$$

where:

$$\hat{y}_i = w_t x_i + b_t$$

and t is iteration at t (epoch)

In matrix form: Let's first expand the MSE:

$$\text{MSE} = \frac{1}{n}(\mathbf{y} - \mathbf{X}\mathbf{w_t} - \mathbf{b_t})^T(\mathbf{y} - \mathbf{X}\mathbf{w_t} - \mathbf{b_t})$$

if we define

$$\theta_t = \begin{bmatrix} b_t \\ w_t \end{bmatrix}$$

the MSE will be written as:

$$\text{MSE} = \frac{1}{n}\mathbf{e}^T\mathbf{e}$$

where

$$\mathbf{e} = \mathbf{X}\theta_t - \mathbf{y}$$

Notice we use

$$\mathbf{e}^T\mathbf{e} = \sum_{i=1}^{n} e_i^2$$

Explanation

1. **Vector** $\mathbf{e}$:
   - $\mathbf{e}$ is an n-dimensional column vector resulting from the difference between the predicted values ($\mathbf{X}\theta_t$) and the actual values ($\mathbf{y}$).
   - If $\mathbf{X}$ is an $n \times d$ matrix, θ_t is a d-dimensional column vector, and $\mathbf{y}$ is an n-dimensional column vector, then $\mathbf{e} = \mathbf{X}\theta_t - \mathbf{y}$ is also an n-dimensional column vector.
2. **Squared Error**:
   - The squared error $\mathbf{e}^T\mathbf{e}$ is a scalar value.
   - Here, $\mathbf{e}^T$ (the transpose of $\mathbf{e}$) is a $1 \times n$ row vector.
   - When multiplying $\mathbf{e}^T$ (a $1 \times n$ row vector) by $\mathbf{e}$ (an $n \times 1$ column vector), the result is a 1×1 scalar.

Partial Derivative

$$\mathrm{MSE} = \frac{1}{n}\mathbf{e}^T\mathbf{e}$$

Now, apply the chain rule:

$$\frac{\partial \mathrm{MSE}}{\partial \boldsymbol{\theta}_\mathbf{t}} = \frac{1}{n}\frac{\partial}{\partial \boldsymbol{\theta}_\mathbf{t}}(\mathbf{e}^T\mathbf{e})$$

Using the gradient of the squared error term, where $\mathbf{e}^T\mathbf{e} = \sum_{i=1}^{n} e_i^2$, the derivative with respect to $\boldsymbol{\theta}_\mathbf{t}$ is:

$$\frac{\partial}{\partial \boldsymbol{\theta}_\mathbf{t}}(\mathbf{e}^T\mathbf{e}) = 2\mathbf{X}^T\mathbf{e}$$

Thus:

$$\frac{\partial \mathrm{MSE}}{\partial \boldsymbol{\theta}_\mathbf{t}} = \frac{2}{n}\mathbf{X}^T\mathbf{e}$$

Substitute back the error term:

$$\mathbf{e} = \mathbf{X}\boldsymbol{\theta}_\mathbf{t} - \mathbf{y}$$

So the final expression is:

$$\frac{\partial \text{MSE}}{\partial \theta_{\mathbf{t}}} = \frac{2}{n}\mathbf{X}^T(\mathbf{X}\theta_{\mathbf{t}} - \mathbf{y})$$

Summary

The correct expression for the gradient of the MSE with respect to $\theta_{\mathbf{t}}$ is:

$$\frac{\partial \text{MSE}}{\partial \theta_{\mathbf{t}}} = \frac{2}{n}\mathbf{X}^T(\mathbf{X}\theta_{\mathbf{t}} - \mathbf{y})$$

This form does not include any unnecessary transpositions and directly applies the gradient correctly.

- Understand dMSE_dw, and dMSE_db