Mean Squared Error (MSE) Loss Function Derivation via MLE

1. Probabilistic Setup (Gaussian Noise Model)

Assume the true relationship between input x and output y is:

$$y = f(x;\theta) + \epsilon$$

where:

- $f(x;\theta)$ is the model (e.g., linear regression: $f(x;\theta) = \theta^T x$)
- $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is Gaussian noise with mean 0 and variance σ^2

The conditional distribution of y given x is:

$$p(y|x;\theta) = \mathcal{N}(f(x;\theta), \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - f(x;\theta))^2}{2\sigma^2}\right)$$

2. Likelihood Function

Given i.i.d. data $\{(x_i, y_i)\}_{i=1}^n$, the likelihood is:

$$L(\theta) = \prod_{i=1}^{n} p(y_i|x_i;\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - f(x_i;\theta))^2}{2\sigma^2}\right)$$

3. Log-Likelihood Function

Taking the natural logarithm:

$$\begin{aligned}
\ell(\theta) &= \log L(\theta) \\
&= \sum_{i=1}^{n} \left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_i - f(x_i;\theta))^2}{2\sigma^2}\right] \\
&= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - f(x_i;\theta))^2
\end{aligned}$$

4. Maximizing the Log-Likelihood

Maximizing $\ell(\theta)$ is equivalent to minimizing:

$$-\ell(\theta) = \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - f(x_i;\theta))^2$$

Dropping constants gives the Sum of Squared Errors (SSE):

$$\sum_{i=1}^{n}(y_i - f(x_i;\theta))^2$$

The Mean Squared Error (MSE) is:

$$\text{MSE}(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i;\theta))^2$$

5. Gradient of MSE

The gradient with respect to θ:

$$\nabla_\theta \text{MSE}(\theta) = \frac{2}{n}\sum_{i=1}^{n}(f(x_i;\theta) - y_i)\nabla_\theta f(x_i;\theta)$$

For linear regression ($f(x;\theta) = \theta^T x$):

$$\nabla_\theta \text{MSE}(\theta) = \frac{2}{n}\sum_{i=1}^{n}(\theta^T x_i - y_i)x_i$$

6. Estimating σ^2 via MLE

If σ^2 is unknown:

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(y_i - f(x_i;\theta))^2 = 0$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i;\theta))^2$$