1. Mean Squared Error (MSE)

Formula:

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

Gradient:

To optimize the model, we compute the gradient of the MSE with respect to the predictions ($\hat{y}_i$):

$$\frac{\partial MSE}{\partial \hat{y}_i} = -\frac{2}{n}(y_i - \hat{y}_i)$$

- The gradient is proportional to the error $(y_i - \hat{y}_i)$.
- **Large Errors**: Larger differences result in larger updates, which can cause instability, especially with outliers.
- **Small Errors**: Small differences lead to slower convergence near the minima.

Convergence:

- **Behavior**: Sensitive to large errors (outliers) because of the squared term.
- **Example**: Let's say the true value (y) is 5 and the prediction ($\hat{y}$) is 3:

  $$MSE = (5-3)^2 = 4$$

  Gradient:

  $$\frac{\partial MSE}{\partial \hat{y}} = -2(5-3) = -4$$

  A large gradient pushes the weights quickly, which is beneficial for large errors but can lead to instability if outliers dominate.

---

2. Mean Absolute Error (MAE)

Formula:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Gradient:

The derivative of MAE is not smooth due to the absolute value. The gradient is:

$$\frac{\partial MAE}{\partial \hat{y}_i} = \begin{cases} -\frac{1}{n} & \text{if } y_i > \hat{y}_i, \\ \frac{1}{n} & \text{if } y_i < \hat{y}_i. \end{cases}$$

- The gradient is constant for any error, meaning all updates are uniform regardless of the error magnitude.
- **Convergence**: MAE does not prioritize large errors, leading to slower convergence.

Example:

Let $y = 5$ and $\hat{y} = 3$:

$$MAE = |5 - 3| = 2$$

Gradient:

$$\frac{\partial MAE}{\partial \hat{y}} = -1$$

The gradient is constant, leading to a uniform update size, irrespective of the magnitude of the error.

---

3. Log Loss (Binary Cross-Entropy)

Formula:

For binary classification ($y \in \{0, 1\}$):

$$\text{Log Loss} = -\frac{1}{n} \sum_{i=1}^{n} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

- $\hat{y}_i$: The predicted probability for $y_i = 1$, i.e., $\hat{y}_i \in (0, 1)$.
- Penalizes incorrect predictions more as $\hat{y}_i$ approaches 0 or 1 incorrectly.

Gradient:

The gradient with respect to $\hat{y}_i$ is:

$$\frac{\partial \text{Log Loss}}{\partial \hat{y}_i} = \begin{cases} -\frac{y_i}{\hat{y}_i} & \text{if } y_i = 1, \\ \frac{1 - y_i}{1 - \hat{y}_i} & \text{if } y_i = 0. \end{cases}$$

- If the model is highly confident but wrong ($\hat{y}_i \to 0$ when $y_i = 1$), the gradient becomes large, encouraging significant updates.
- If the model is confident and correct ($\hat{y}_i \to 1$ when $y_i = 1$), the gradient becomes small, leading to finer updates.

Example:

Let $y = 1$ and $\hat{y} = 0.1$:

$$\text{Log Loss} = -[1 \cdot \log(0.1) + 0 \cdot \log(1 - 0.1)] = 2.302$$

Gradient:

$$\frac{\partial \text{Log Loss}}{\partial \hat{y}} = -\frac{1}{0.1} = -10$$

This large gradient encourages a significant update to $\hat{y}$, improving the prediction faster.

---

4. Comparing Convergence

Example Problem:

Suppose we are classifying whether a movie review is positive ($y = 1$) or negative ($y = 0$), and the model predicts probabilities ($\hat{y}$) for positivity.

Loss Function	$\hat{y} = 0.1, y = 1$	Gradient ($\frac{\partial Loss}{\partial \hat{y}}$)	Behavior
MSE	$(1 - 0.1)^2 = 0.81$	$-2(1 - 0.1) = -1.8$	Slower updates for probabilities near 0.
MAE	(	1 - 0.1	= 0.9)
Log Loss	$-\log(0.1) = 2.302$	-10	Fast updates for wrong predictions.

Key Insights

1. **MSE**:
   - Good for regression.
   - Sensitive to outliers and slower for classification.
2. **MAE**:

- Robust to outliers.
- Slower convergence due to uniform updates.

3. **Log Loss**:
   - Best for classification (e.g., NLP tasks like sentiment analysis).
   - Encourages faster convergence due to steep gradients for incorrect predictions.