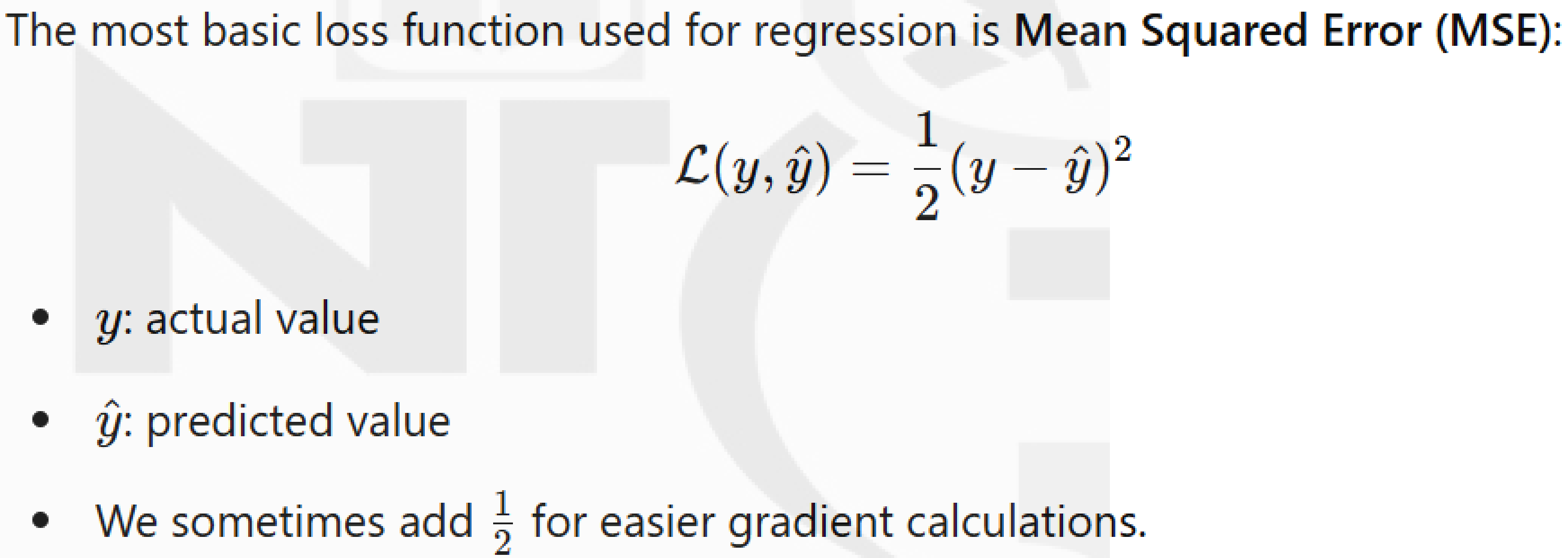
Understanding Gradient Boost Loss Function

MUKESH KUMAR

Loss Function for Regression

Mean Squared Error (MSE) Loss Function for Regression



The most basic loss function used for regression is **Mean Squared Error (MSE)**:

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

- y: actual value
- $\hat{y}$: predicted value
- We sometimes add $\frac{1}{2}$ for easier gradient calculations.

Gradient Boosting is inspired by **gradient descent** — we want to **minimize the loss** function.

To minimize any function, we move in the direction **opposite to the gradient**.

- So, we compute:

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

- Why w.r.t. y^?
  Because we are updating our **prediction** — so we need to know how a small change in prediction affects the loss.

What is a Residual?

- In regression, a **residual** is the difference between the actual and predicted value:

$$\text{residual} = y - \hat{y}$$

- This tells us **how far off** the prediction is.

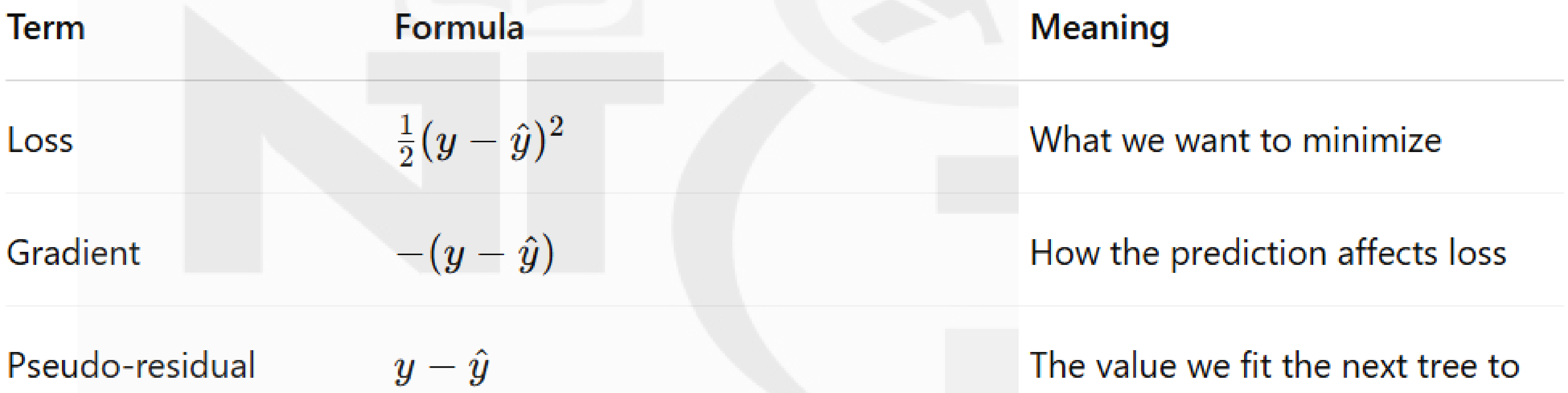
So What is a Pseudo-Residual?

- In gradient boosting, we generalize this idea:
  A **pseudo-residual** is:

$$r_i = -\frac{\partial \mathcal{L}(y_i, \hat{y}_i)}{\partial \hat{y}_i}$$

- This is the **negative gradient of the loss** w.r.t. prediction.

So, in MSE (regression):



Term	Formula	Meaning
Loss	$\frac{1}{2}(y - \hat{y})^2$	What we want to minimize
Gradient	$-(y - \hat{y})$	How the prediction affects loss
Pseudo-residual	$y - \hat{y}$	The value we fit the next tree to

Loss Function for Classification

Loss Function (Log-Loss)

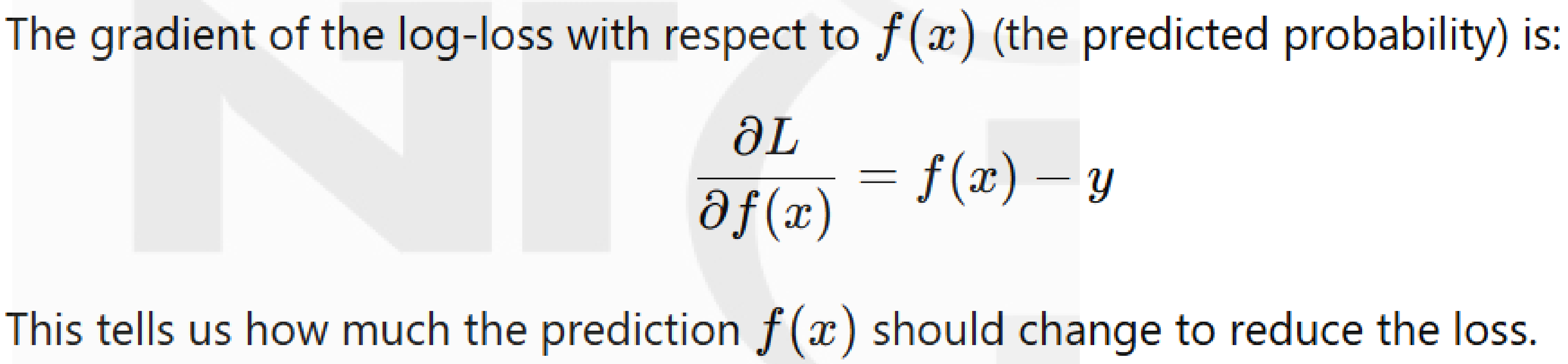
For binary classification, the log-loss is used as the loss function:

$$L(y, f(x)) = -\left[y \log(f(x)) + (1 - y) \log(1 - f(x))\right]$$

Where:

- y is the true label (0 or 1).
- $f(x)$ is the predicted probability for the positive class.

Gradient Calculation



The gradient of the log-loss with respect to $f(x)$ (the predicted probability) is:

$$\frac{\partial L}{\partial f(x)} = f(x) - y$$

This tells us how much the prediction $f(x)$ should change to reduce the loss.

Residuals

Residuals are the difference between the true label and the predicted value:

$$r_i = y_i - f(x_i)$$

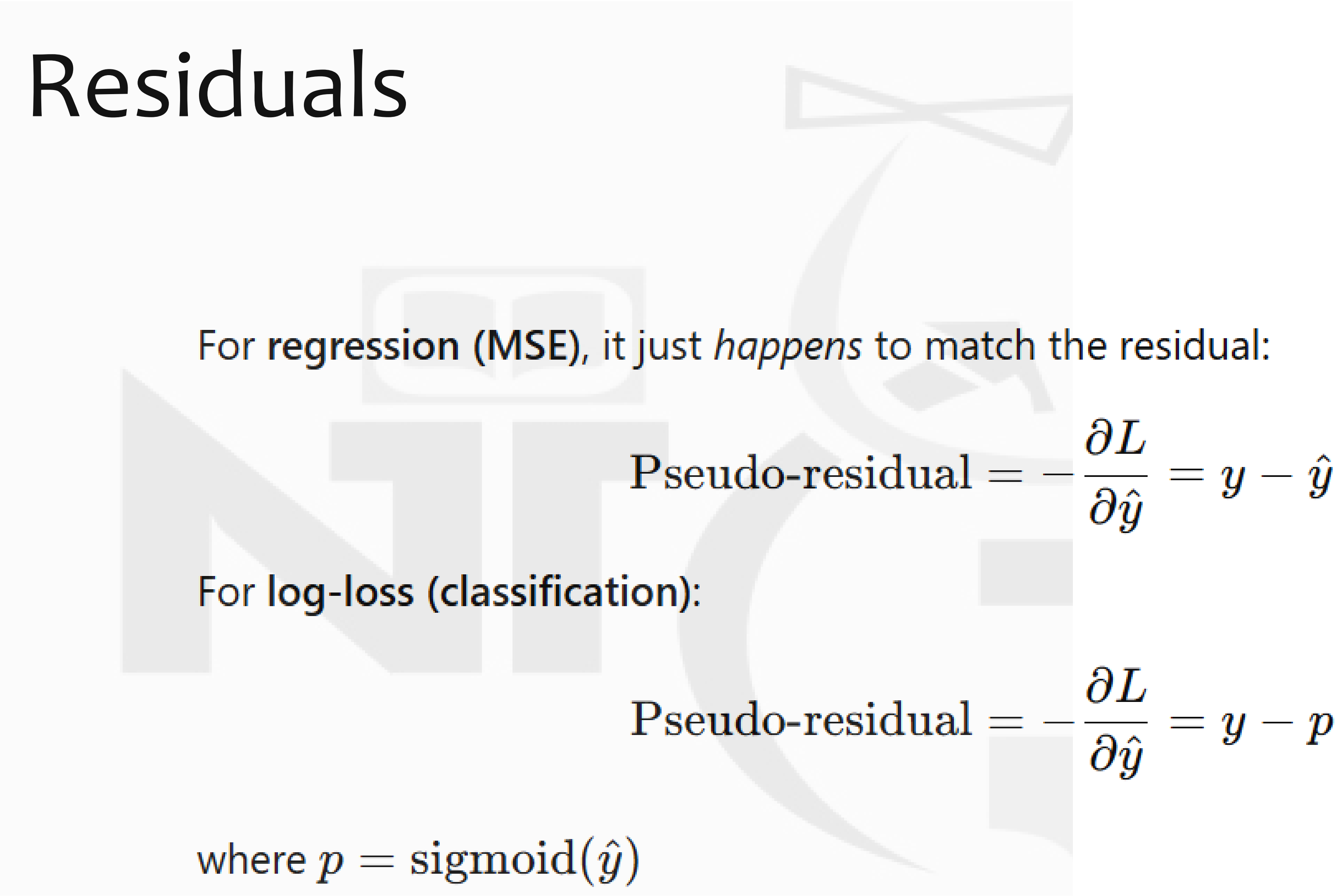
Pseudo-Residuals

The pseudo-residuals are the negative gradient of the loss:

$$r_i^{(t)} = y_i - f(x_i)$$

In each iteration, a new tree is fitted to these pseudo-residuals to adjust the model's predictions.

GB Residuals



For **regression (MSE)**, it just *happens* to match the residual:

$$\text{Pseudo-residual} = -\frac{\partial L}{\partial \hat{y}} = y - \hat{y}$$

For **log-loss (classification)**:

$$\text{Pseudo-residual} = -\frac{\partial L}{\partial \hat{y}} = y - p$$

where $p = \text{sigmoid}(\hat{y})$