WEIGHT INITIALIZATION IN NEURAL NETWORKS

01.

Setting all the weights to Zero or same constant

- setting all the weights to zero may not be a wiser way because it leads to a problem where weight multiplied by zero gives zeros to the forth coming layers as well. This may not let the neural network learn.
- Setting the weights to the same constant is also not recommended because it may lead to a problem of symmetry where the neural network will not learn new things.

02.

Initializing it to lower values

This is a problem as well because when we initialize low values to weights, the derivatives of weights becomes smaller, especially when using sigmoid functions (derivative is max 0.25) . So when using the chain rule, repeatedly multiplying the gradients with smaller values becomes even smaller and the updates to weights become very low or negligible. It is called vanishing gradient.

03.

Initializing higher values.

While initializing higher values to weights, it leads to higher derivatives and product of these derivatives becomes even higher. As a result, the updation of the weight becomes exponentially high. Where there is a lot of deviation from the old and new weights, this might lead to oscillations. Hence there is a possibility of skipping the global minima. It is called exploding gradient.

TWO TECHNIQUES OF WEIGHT INITIALIZATION

01.

Xavier (Glorot) Initialization

Best used for: Sigmoid or Tanh activation functions
Goal: Maintain consistent variance of activations and gradients across layers

- **Uniform distribution:**

$$W \sim \mathcal{U}\left[-\sqrt{\frac{6}{f_{in}+f_{out}}}, \sqrt{\frac{6}{f_{in}+f_{out}}}\right]$$

- **Normal distribution:**

$$W \sim \mathcal{N}\left(0, \frac{2}{f_{in}+f_{out}}\right)$$

He Initialization

- Best used for: ReLU and its variants (e.g., Leaky ReLU)
- Goal: Compensate for ReLU zeroing out half of the activations

02.

- **Uniform distribution:**

$$W \sim \mathcal{U}\left[-\sqrt{\frac{6}{f_{in}}}, \sqrt{\frac{6}{f_{in}}}\right]$$

- **Normal distribution:**

$$W \sim \mathcal{N}\left(0, \frac{2}{f_{in}}\right)$$