

Prompt : "Propose three  $d \times d$  initialization strategies for a 3-layer linear network  $y = W_3 W_2 W_1 x$  with  $d = 100$ : (i) orthogonal, (ii) scaled Gaussian, (iii) random uniform. For each, state a theoretical reason why it should preserve or control the output variance and gradient norms across layers."

Please be specific about the mathematical formulation for each initialization method.

**Answer :**

For a 3-layer linear network  $y = W_3 W_2 W_1 x$  with  $d = 100$ , here are three initialization strategies with their mathematical formulations and theoretical justifications:

## (i) Orthogonal Initialization

**Mathematical Formulation:**

- Sample each  $W_i \in \mathbb{R}^{(d \times d)}$  from the uniform distribution over the set of orthogonal matrices
- Equivalently: Generate random Gaussian matrix  $G$ , compute QR decomposition  $G = QR$ , set  $W_i = Q$
- Scale by gain factor  $g$ :  $W_i = g \cdot Q$  where typically  $g = 1$

**Theoretical Justification:**

Orthogonal matrices preserve norms exactly:  $\|Wx\|_2 = \|x\|_2$  when  $W$  is orthogonal.

For the forward pass, if  $x$  has variance  $\sigma_x^2$ :

- $\text{Var}(W_1 x) = \sigma_x^2$ ,  $\text{Var}(W_2 W_1 x) = \sigma_x^2$ ,  $\text{Var}(y) = \sigma_x^2$

For the backward pass, gradients propagate through  $W^T$ , which is also orthogonal:

- $\|\partial L / \partial x\|_2 = \|W_1^T W_2^T W_3^T \partial L / \partial y\|_2 = \|\partial L / \partial y\|_2$

This prevents vanishing/exploding gradients since singular values are all exactly 1, maintaining the condition number at 1 throughout training.

## (ii) Scaled Gaussian (Xavier/He Initialization)

**Mathematical Formulation:**

For **Xavier initialization** (suitable for linear/tanh):

- $W_i \sim N(0, \sigma^2)$  where  $\sigma^2 = 2/(d_{in} + d_{out}) = 2/(100 + 100) = 1/100$
- Each element:  $W_i[j,k] \sim N(0, 0.01)$

For **He initialization** (suitable for ReLU, but applicable here):

- $W_i \sim N(0, \sigma^2)$  where  $\sigma^2 = 2/d_i = 2/100 = 0.02$

#### Theoretical Justification:

Consider forward variance with independent inputs  $E[x] = 0$ ,  $\text{Var}(x_i) = \sigma_x^2$ :

- For  $y = Wx$ :  $\text{Var}(y_i) = \sum_j \text{Var}(W_{ij}x_j) = d \cdot \sigma^2_w \cdot \sigma_x^2$

To preserve variance ( $\text{Var}(y) = \sigma_x^2$ ), we need:

- $d \cdot \sigma^2_w = 1 \Rightarrow \sigma^2_w = 1/d = 1/100$

For backward pass, gradient  $\partial L / \partial x = W^T \partial L / \partial y$  follows similar analysis:

- $\text{Var}(\partial L / \partial x_i) = d \cdot \sigma^2_w \cdot \text{Var}(\partial L / \partial y)$

Xavier's  $2/(d_i + d_{ou})$  compromises between forward and backward variance preservation, preventing gradients from vanishing or exploding across the 3 layers.

### (iii) Random Uniform (LeCun Uniform)

#### Mathematical Formulation:

- $W_i \sim U(-a, a)$  where  $a = \sqrt{(3/d_i)} = \sqrt{(3/100)} \approx 0.173$
- Each element:  $W_{i[j,k]} \sim U(-0.173, 0.173)$

Alternatively, **Glorot Uniform**:

- $a = \sqrt{(6/(d_i + d_{ou}))} = \sqrt{(6/200)} \approx 0.173$

#### Theoretical Justification:

For uniform  $U(-a, a)$ , the variance is:

- $\text{Var}(W) = a^2/3$

To achieve  $\text{Var}(W) = 1/d$  for variance preservation:

- $a^2/3 = 1/d \Rightarrow a = \sqrt{(3/d)} = \sqrt{(3/100)}$

Forward pass analysis with  $y = Wx$ :

- $\text{Var}(y_i) = d \cdot (a^2/3) \cdot \sigma_x^2 = d \cdot (1/d) \cdot \sigma_x^2 = \sigma_x^2$

This maintains consistent activations across layers. The bounded support  $[-a, a]$  also provides additional numerical stability compared to Gaussian (no extreme outliers), which can be beneficial early in training when the network might be sensitive to large initial weights.

**Summary:** All three methods aim to keep the variance of activations and gradients stable across layers, preventing the vanishing/exploding gradient problem, but through different mechanisms: orthogonal via exact norm preservation, Gaussian via statistical variance matching, and uniform via bounded variance control.