Xavier/Glorot Initialization

Designed for Activations

sigmoid **tanh**

Var → Var → Var

Layer 1 Layer 2 Layer 3

Variance maintained ≈ 1.0

Uniform Distribution

$$W \sim U(-\sqrt{(6/(n_{in} + n_{out}))}, \sqrt{(6/(n_{in} + n_{out}))})$$

Gaussian Distribution

$$W \sim N(0, 2/(n_{in} + n_{out}))$$

Key Properties

- **Variance maintained** across layers during forward/backward pass
- Keeps activation variance approximately **1.0**
- **Prevents saturation** in early training stages
- Standard for **fully connected networks**
- Optimal for **tanh/sigmoid** activation functions

Calculation Examples

Example 1: Uniform Distribution

Input neurons (n_{in}):	784
Output neurons (n_{out}):	128
Weight matrix size:	784 × 128

1 Calculate sum

$$n_{in} + n_{out} = 784 + 128 = 912$$

2 Calculate fraction

$$6 / (n_{in} + n_{out}) = 6 / 912$$
$$= 0.006579$$

3 Calculate square root

$$\sqrt{(0.006579)} = 0.0811$$

Example 2: Gaussian Distribution

Input neurons (n_{in}):	256
Output neurons (n_{out}):	64
Weight matrix size:	256 × 64

1 Calculate sum

$$n_{in} + n_{out} = 256 + 64 = 320$$

2 Calculate variance

$$\sigma^2 = 2 / (n_{in} + n_{out}) = 2 / 320$$
$$= 0.00625$$

3 Standard deviation (σ)

$$\sigma = \sqrt{(0.00625)} = 0.0791$$

W ~ U(-0.0811, +0.0811)

Weight samples (partial)

```
[-0.0523  0.0701 -0.0198  0.0445]
[ 0.0312 -0.0789  0.0621 -0.0356]
[-0.0678  0.0234  0.0789 -0.0512]
[ 0.0456 -0.0123  0.0654  0.0289]
```

W ~ N(0, σ=0.0791)

Weight samples (partial)

```
[-0.0423  0.0612 -0.0891  0.0234]
[ 0.0756 -0.0189  0.0523 -0.0678]
[-0.0345  0.0801 -0.0267  0.0445]
[ 0.0589 -0.0734  0.0123  0.0690]
```

Example 3: Small Network

Input neurons (n_{in}):	10
Output neurons (n_{out}):	5
Weight matrix size:	10 × 5

1 Uniform Distribution calculation

$$n_{in} + n_{out} = 10 + 5 = 15$$
$$\sqrt{(6/15)} = \sqrt{(0.4)} = 0.6325$$

W ~ U(-0.6325, +0.6325)

Full weight matrix (10×5)

Example 4: Large Network Comparison

Input neurons (n_{in}):	2048
Output neurons (n_{out}):	1024
Weight matrix size:	2048 × 1024

1 Uniform Distribution

$$\sqrt{(6/(2048+1024))} = \sqrt{(6/3072)}$$
$$= \sqrt{(0.001953)} = 0.0442$$

2 Gaussian Distribution

```
[-0.3421  0.5234 -0.1823  0.4512 -0.2901]
[ 0.2134 -0.4823  0.3912  0.1567 -0.5123]
[-0.4567  0.2901  0.5421 -0.3234  0.1789]
[ 0.3789 -0.2156  0.4234 -0.5678  0.2912]
[-0.1234  0.4567 -0.3812  0.2345  0.5234]
[ 0.4912 -0.3567  0.1234 -0.4321  0.3678]
[-0.2678  0.5012 -0.4156  0.1923  0.3456]
[ 0.3234 -0.1789  0.4678 -0.2567  0.5123]
[-0.4123  0.2678  0.3567 -0.5234  0.1456]
[ 0.5678 -0.3012  0.2134  0.4789 -0.3901]
```

$\sigma = \sqrt{(2/3072)} = \sqrt{(0.000651)}$

$= 0.0255$

Uniform: W ~ U(-0.0442, +0.0442)
Gaussian: W ~ N(0, σ=0.0255)

💡 **Observation:** As the network grows larger, the initialization range becomes smaller. This prevents the sum of many inputs from becoming too large.