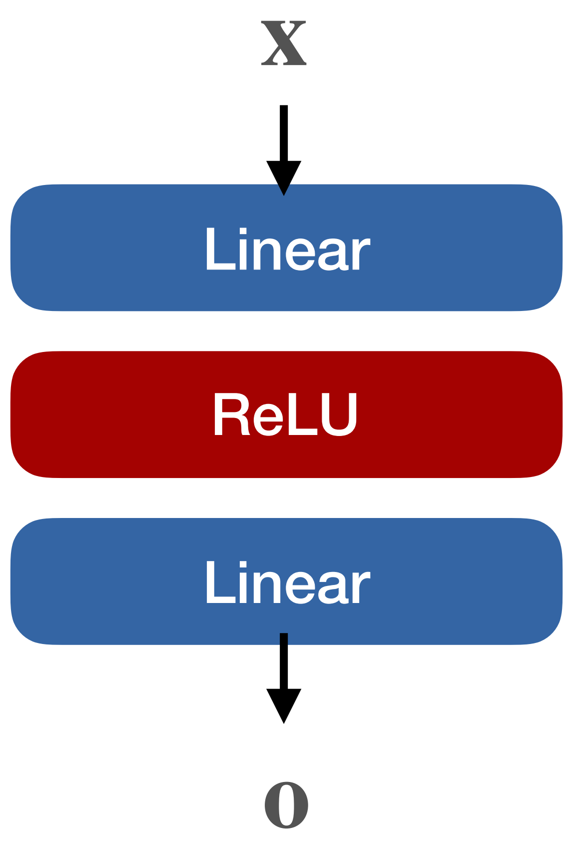
Xavier and Kaiming initialization

Xavier and Kaiming initialization

- Strategy to set variance σ^2 of Normal initialization
  - All activations are of similar scale

$$\mathbf{W}_1 \sim \mathcal{N}(\mu_1, \sigma_1^2 \mathbf{I})$$

$$\mathbf{W}_3 \sim \mathcal{N}(\mu_3, \sigma_3^2 \mathbf{I})$$



Random matrix multiplication

$$\mathbf{a}^\top \mathbf{x} \sim \mathcal{N}(\mu_a \sum_i \mathbf{x}_i, \|\mathbf{x}\|^2 \sigma_a^2) \qquad \text{for} \quad \mathbf{a} \sim \mathcal{N}(\mu_a, \sigma_a^2 \mathbf{I})\ {}_1$$

[1] for derivation see: https://en.m.wikipedia.org/wiki/Multivariate_normal_distribution

Random matrix multiplication

$$\mathbf{z}_i = \mathbf{W}_{i-1}\mathbf{z}_{i-1} \sim \mathcal{N}(0,\ \|\mathbf{z}_{i-1}\|^2\sigma^2_{W_{i-1}}\mathbf{I}) \quad \text{for} \quad \mathbf{W}_{i-1} \sim \mathcal{N}(0,\ \sigma^2_{W_{i-1}}\mathbf{I})$$

Random ReLU

$\mathbf{z}_{i+1} = \max(\mathbf{z}_i, 0)$ for $\mathbf{z}_i \sim \mathcal{N}(0, \sigma_i^2 \mathbf{I})$

$$\mathbb{E}[\|\mathbf{z}_{i+1}\|^2] = \frac{1}{2} n_{\mathbf{z}_i} \sigma_i^2$$

Putting things together

$$\mathbf{z}_i \sim \mathcal{N}(0,\, \sigma_i^2 \mathbf{I})$$

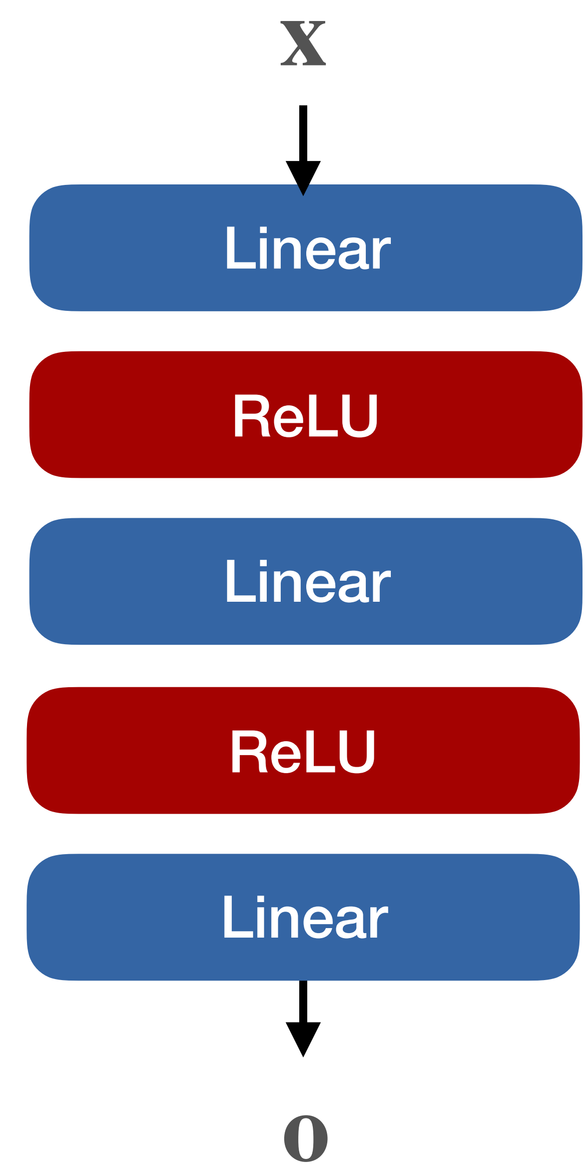
$$\|\mathbf{z}_{i+1}\|^2 \approx \mathbb{E}[\|\mathbf{z}_{i+1}\|^2] = \frac{1}{2} n_{\mathbf{z}_i} \sigma_i^2 \qquad \mathbf{z}_{i+2} \sim \mathcal{N}(0,\, \underbrace{\|\mathbf{z}_{\mathbf{i+1}}\|^2 \sigma_{W_{i+1}}^2}_{\sigma_{i+2}^2} \mathbf{I})$$

$$\sigma_{i+2} = \frac{1}{\sqrt{2}} \sigma_{W_{i+1}} \sigma_i \sqrt{n_{\mathbf{z}_i}}$$

$$\sigma_i = \prod_{k=0}^{(i-1)/2} \left(\frac{1}{\sqrt{2}} \sigma_{W_{2k+1}} \sqrt{n_{\mathbf{z}_{2k}}} \right) \sigma_x$$

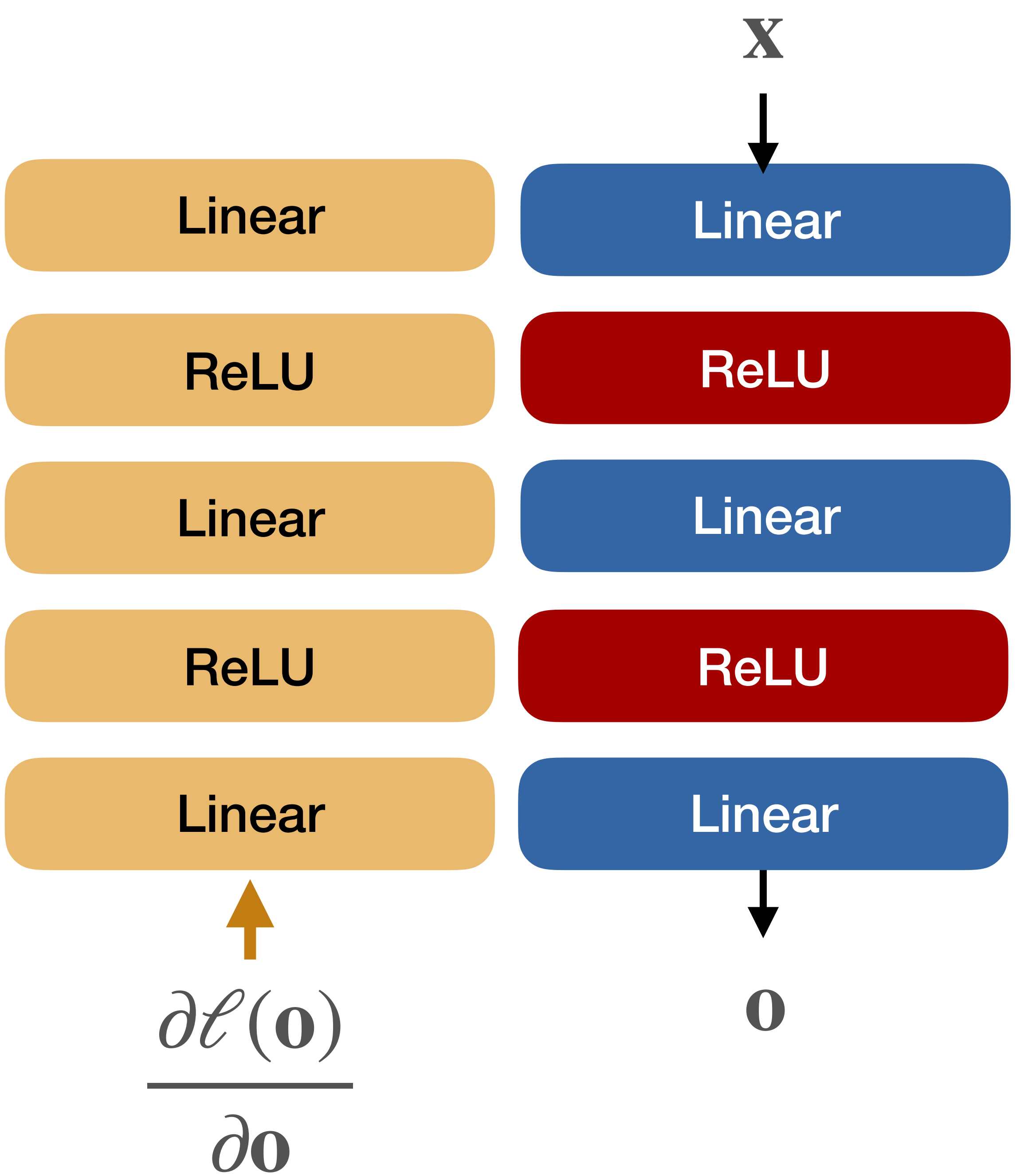
Randomly initialized network

$$\sigma_i = \prod_{k=0}^{(i-1)/2} \left(\frac{1}{\sqrt{2}} \sigma_{W_{2k+1}} \sqrt{n_{\mathbf{z}_{2k}}} \right) \sigma_x$$



Variance of back-propagation graph

$$\hat{\sigma}_i = \prod_{k=i/2}^{(N-1)/2} \left(\frac{1}{\sqrt{2}} \sigma_{W_{2k+1}} \sqrt{n_{\mathbf{z}_{2k+2}}} \right) \hat{\sigma}_N$$



Xavier initialization

- Try to keep both activations and gradient magnitude constant
- $\sigma_W = \sqrt{2}\sqrt{\dfrac{2}{n_{\mathbf{z}_i} + n_{\mathbf{z}_{i+1}}}}$

$$\sigma_i = \prod_{k=0}^{(i-1)/2} \left(\frac{1}{\sqrt{2}} \sigma_{W_{2k+1}} \sqrt{n_{\mathbf{z}_{2k}}} \right) \sigma_x$$

$$\hat{\sigma}_i = \prod_{k=i/2}^{(N-1)/2} \left(\frac{1}{\sqrt{2}} \sigma_{W_{2k+1}} \sqrt{n_{\mathbf{z}_{2k+2}}} \right) \hat{\sigma}_N$$

Kaiming initialization

- Try to keep either activation or gradient magnitude constant

$$\sigma_i = \prod_{k=0}^{(i-1)/2} \left(\frac{1}{\sqrt{2}} \sigma_{W_{2k+1}} \sqrt{n_{\mathbf{z}_{2k}}} \right) \sigma_x$$

$$\hat{\sigma}_i = \prod_{k=i/2}^{(N-1)/2} \left(\frac{1}{\sqrt{2}} \sigma_{W_{2k+1}} \sqrt{n_{\mathbf{z}_{2k+2}}} \right) \hat{\sigma}_N$$

- $\sigma_W = \sqrt{2} \frac{1}{\sqrt{n_{\mathbf{z}_i}}}$

- $\sigma_W = \sqrt{2} \frac{1}{\sqrt{n_{\mathbf{z}_{i+1}}}}$

Initialization in practice

- Xavier (default) is often good enough
- Initialize last layer to zero