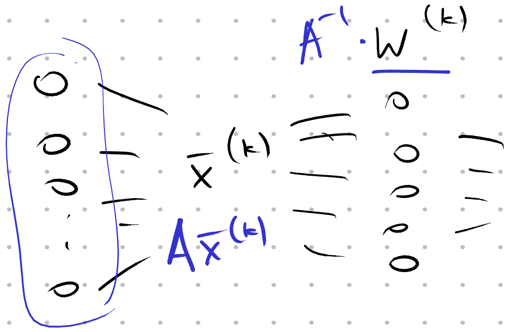
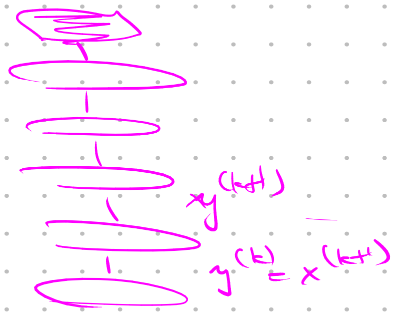


$$\bar{y}^{(k)} = h\left(A^{-1} W^{(k)} \cdot \bar{x}^{(k)} + \bar{w}_0^{(k)}\right)$$

vanishing gradients

exploding gradients



$$y = \bar{w}^T \bar{x} = \Sigma y_i = \Sigma w_i x_i$$

$$\mathrm{Var}[y_i] = \mathrm{Var}[w_i x_i] = \mathbb{E}[w_i^2 x_i^2] - \mathbb{E}[w_i x_i]^2 =$$

$$= \mathbb{E}[x_i]^2\,\mathrm{Var}[w_i] + \mathbb{E}[w_i]^2\,\mathrm{Var}[x_i] + \mathrm{Var}[w_i]\mathrm{Var}[x_i]$$

① $\mathbb{E} w_i = 0$, $\boxed{\mathbb{E} x_i = 0}$

$$\mathrm{Var}[y_i] = \mathrm{Var}[w_i]\,\mathrm{Var}[x_i] \quad \approx 1$$

$$\mathrm{Var}[y] = n \cdot \mathrm{Var}[y_i] = n \cdot \mathrm{Var}[w_i]\,\mathrm{Var}[x_i]$$

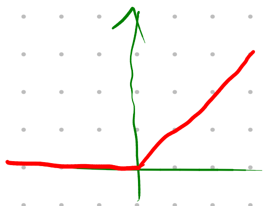
symmetric activ. functions

(Glorot, Bengio, 2010) $w_i \sim \mathrm{Unif}\left(\left[-\frac{\sqrt{3}}{\sqrt{n}}, \frac{\sqrt{3}}{\sqrt{n}}\right]\right)$ Xavier init

$$\mathrm{Var}[w_i] = \frac{(2\sqrt{3}/\sqrt{n})^2}{12} = \frac{1}{\cancel{3}n}\ ; \quad \mathrm{Var}[y] = \cancel{\tfrac{1}{3}}\,\mathrm{Var}[x_i]$$

② $\mathbb{E} w_i = 0$ $\quad h = \mathrm{ReLU}$ $\quad x^{(k+1)} = \mathrm{ReLU}(y^{(k)}) = \mathrm{ReLU}(\bar{w}_k^T \bar{x}^{(k)})$

$$\mathrm{Var}[y_i] = \mathrm{Var}[w_i]\left(\mathbb{E}[x_i]^2 + \mathrm{Var}[x_i]\right) =$$



Kaiming He

$$= \mathrm{Var}[w_i] \cdot \mathbb{E}[x_i^2]$$

$$\mathrm{Var}[y^{(k)}] = n^{(k)} \cdot \mathrm{Var}[w_i^{(k)}] \cdot \mathbb{E}[(x^{(k)})^2]$$

$\boxed{\text{He init}}$ $\quad x^{(k)} = \max(0, y^{(k-1)})$, $\quad \mathbb{E}\, y^{(k-1)} = 0 \implies \approx 2/n$

$$\mathbb{E}[(x^{(k)})^2] = \frac{1}{2}\mathrm{Var}[y^{(k-1)}], \quad \mathrm{Var}[y^{(k)}] = \frac{n^{(k)}}{2}\mathrm{Var}[w_i^{(k)}] \cdot \mathrm{Var}[y^{(k-1)}]$$

Batch Normalization

Covariate shift

Internal covariate shift

$\bar y^{(k)} = \bar x^{(k+1)}$

$\bar w^{(k)}$

$\bar x^{(k)}$

$\bar x$

$f(\bar x) = \frac{\bar x - \bar a}{\bar b}$

$\bar x^{(k+1)} = \mathrm{Norm}(\bar y^{(k)})$

Norm

$\bar y^{(k)} = F(\bar x^{(k)})$

F

$\bar x^{(k)}$

$\frac{\bar y^{(k)} - E[\bar y^{(k)}]}{\mathrm{Var}[\bar y^{(k)}]}$

$\bar W_0 + W \bar x^{(k)}$

$\bar W_0 + \Delta \bar W_0 + \Delta \bar W_0$

$\mathrm{Norm}(\bar y_n^{(k)}) = \frac{\bar y_n^{(k)} - E[\bar y_n^{(k)}]}{}$

$\frac{1}{N} \sum y_i^{(k)}$

Norm

$\bar y_1^{(k)}$ $\bar y_n^{(k)}$ $\bar y_N^{(k)}$

F F

$\bar x_1^{(k)}$ $\bar x_n^{(k)}$

D: $\bar X^{(k+1)}$

$\bar x_1$ $\bar x_n$ ... $\bar x_N$

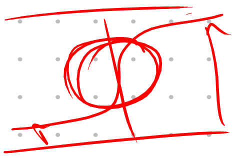
BN

BNorm

$\bar y_1^{(k)}$ $\bar y_n^{(k)}$ $\bar y_m^{(k)}$

$\bar x_1 \dots \bar x_k \dots \bar x_m$

$BN(\bar y_n^{(k)}) = \frac{\bar y_n^{(k)} - \mathrm{Avg}(\bar y_i^{(k)})}{\sqrt{\mathrm{Var}(\bar y_i^{(k)})}}$



BN — $\bar\gamma, \bar\beta$

$\bar x$

Norm

$BN(\bar x)_k = \gamma_k \cdot \frac{x_k - E[x_k]}{\sqrt{\mathrm{Var}[x_k]}} + \beta_k$