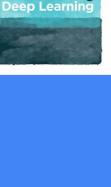
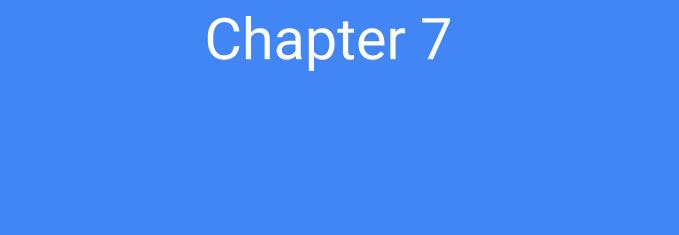


Gradients and initialization



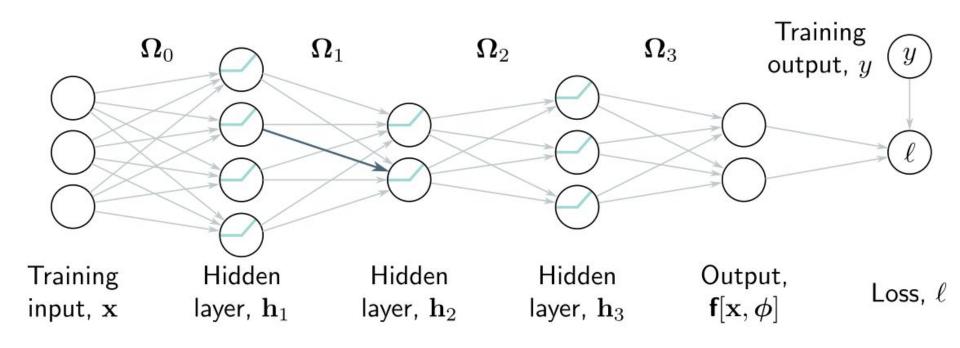
Understanding







Goal: to compute the derivatives of the loss ℓ with respect to each of the weights (arrows) and biases (not shown)



Backpropagation forward pass

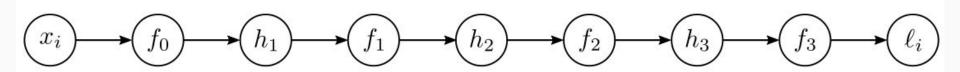


Figure 7.3 Backpropagation forward pass. We compute and store each of the intermediate variables in turn until we finally calculate the loss.

Backpropagation backward pass

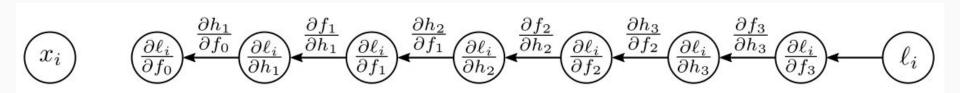


Figure 7.4 Backpropagation backward pass #1. We work backward from the end of the function computing the derivatives $\partial \ell_i/\partial f_{\bullet}$ and $\partial \ell_i/\partial h_{\bullet}$ of the loss with respect to the intermediate quantities. Each derivative is computed from the previous one by multiplying by terms of the form $\partial f_k/\partial h_k$ or $\partial h_k/\partial f_{k-1}$.

$$(x_i) \qquad (\frac{\partial \ell_i}{\partial f_0}) \stackrel{\partial h_1}{\longleftarrow} (\frac{\partial \ell_i}{\partial h_1}) \stackrel{\partial f_1}{\longleftarrow} (\frac{\partial \ell_i}{\partial f_1}) \stackrel{\partial h_2}{\longleftarrow} (\frac{\partial \ell_i}{\partial h_2}) \stackrel{\partial f_2}{\longleftarrow} (\frac{\partial \ell_i}{\partial f_2}) \stackrel{\partial h_3}{\longleftarrow} (\frac{\partial \ell_i}{\partial h_3}) \stackrel{\partial f_3}{\longleftarrow} (\frac{\partial \ell_i}{\partial f_3}) \stackrel{\partial \ell_i}{\longleftarrow} (\ell_i)$$

$$\frac{\partial \ell_{i}}{\partial f_{2}} = \frac{\partial h_{3}}{\partial f_{2}} \left(\frac{\partial f_{3}}{\partial h_{3}} \frac{\partial \ell_{i}}{\partial f_{3}} \right)
\frac{\partial \ell_{i}}{\partial h_{2}} = \frac{\partial f_{2}}{\partial h_{2}} \left(\frac{\partial h_{3}}{\partial f_{2}} \frac{\partial f_{3}}{\partial h_{3}} \frac{\partial \ell_{i}}{\partial f_{3}} \right)
\frac{\partial \ell_{i}}{\partial f_{1}} = \frac{\partial h_{2}}{\partial f_{1}} \left(\frac{\partial f_{2}}{\partial h_{2}} \frac{\partial h_{3}}{\partial f_{2}} \frac{\partial f_{3}}{\partial h_{3}} \frac{\partial \ell_{i}}{\partial f_{3}} \right)
\frac{\partial \ell_{i}}{\partial h_{1}} = \frac{\partial f_{1}}{\partial h_{1}} \left(\frac{\partial h_{2}}{\partial f_{1}} \frac{\partial f_{2}}{\partial h_{2}} \frac{\partial h_{3}}{\partial f_{2}} \frac{\partial f_{3}}{\partial h_{3}} \frac{\partial \ell_{i}}{\partial f_{3}} \right)
\frac{\partial \ell_{i}}{\partial f_{0}} = \frac{\partial h_{1}}{\partial f_{0}} \left(\frac{\partial f_{1}}{\partial h_{1}} \frac{\partial h_{2}}{\partial f_{1}} \frac{\partial f_{2}}{\partial h_{2}} \frac{\partial h_{3}}{\partial f_{2}} \frac{\partial f_{3}}{\partial h_{3}} \frac{\partial \ell_{i}}{\partial f_{3}} \right)$$

$$\underbrace{\begin{pmatrix} x_i \end{pmatrix}} \longrightarrow \underbrace{\begin{pmatrix} f_1 \end{pmatrix}} \longrightarrow \underbrace{\begin{pmatrix} f_2 \end{pmatrix}} \longrightarrow \underbrace{\begin{pmatrix} f_3 \end{pmatrix}} \longrightarrow \underbrace{\begin{pmatrix} \ell_i \end{pmatrix}}$$

$$f_k = \beta_k + \omega_k \cdot h_k$$

$$\frac{\partial f_k}{\partial \beta_k} = 1$$
 and $\frac{\partial f_k}{\partial \omega_k} = h$

$$egin{array}{lll} \mathbf{h}_1 &=& \mathbf{a}[\mathbf{f}_0] \ \mathbf{f}_1 &=& oldsymbol{eta}_1 + oldsymbol{\Omega}_1 \mathbf{h}_1 \ \mathbf{h}_2 &=& \mathbf{a}[\mathbf{f}_1] \ \mathbf{f}_2 &=& oldsymbol{eta}_2 + oldsymbol{\Omega}_2 \mathbf{h}_2 \ \mathbf{h}_3 &=& \mathbf{a}[\mathbf{f}_2] \ \mathbf{f}_3 &=& oldsymbol{eta}_3 + oldsymbol{\Omega}_3 \mathbf{h}_3 \ \ell_i &=& \mathbb{I}[\mathbf{f}_3, y_i], \end{array}$$

$$\partial \ell_i \quad \partial \mathbf{h}_3 \ \partial \mathbf{f}_3 \ \partial \ell_i$$
 $\mathbf{f}_0 = oldsymbol{eta}_0 + oldsymbol{\Omega}_0 \mathbf{x}_i$

 $\mathbf{h}_1 = \mathbf{a}[\mathbf{f}_0]$

 $\mathbf{h}_2 = \mathbf{a}[\mathbf{f}_1]$

 $\mathbf{h}_3 = \mathbf{a}[\mathbf{f}_2]$

 $\mathbf{f}_1 = \boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1 \mathbf{h}_1$

 $\mathbf{f}_2 = \boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2 \mathbf{h}_2$

 $\mathbf{f}_3 = \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3$

 $\ell_i = \mathbf{l}[\mathbf{f}_3, y_i],$

$$\frac{\partial \mathbf{f}}{\partial \mathbf{f}_2} = \frac{\partial}{\partial \mathbf{f}_2} \frac{\partial}{\partial \mathbf{h}_3} \frac{\partial}{\partial \mathbf{f}_3}$$

$$D_3 \times D_3, D_3 \times D_f, \text{ and } D_f \times 1$$

$$\partial \ell_i$$
 $\partial \mathbf{h}_3$ $\partial \mathbf{f}_3$ $\partial \ell_i$ $\partial \mathbf{h}_3$ $\partial \mathbf{h}_3$ $\partial \mathbf{h}_3$ $\partial \mathbf{h}_4$ $\partial \mathbf{h}_5$ $\partial \mathbf{h}_5$ $\partial \mathbf{h}_5$ $\partial \mathbf{h}_6$ $\partial \mathbf{h}_7$ $\partial \mathbf{h}_8$ $\partial \mathbf{h}_8$

$$\frac{\partial \ell_i}{\partial \mathbf{f}_2} = \frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2} \frac{\partial \mathbf{h}_3}{\partial \mathbf{h}_3} \frac{\partial \ell_i}{\partial \mathbf{f}_3}$$

 $\partial \ell_i$

$$D_3 \times D_3, D_3 \times D_f, \text{ and } D_f \times 1$$

$$= \frac{\partial \mathbf{h}_2}{\partial \mathbf{a}} \frac{\partial \mathbf{f}_2}{\partial \mathbf{r}} \left(\frac{\partial \mathbf{h}_3}{\partial \mathbf{a}} \frac{\partial \mathbf{f}_3}{\partial \mathbf{r}} \frac{\partial \ell_i}{\partial \mathbf{r}} \right)$$

$$\frac{\partial \mathbf{h}_2}{\partial \mathbf{f}_1} \frac{\partial \mathbf{f}_2}{\partial \mathbf{h}_2} \left(\frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2} \frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} \frac{\partial \ell_i}{\partial \mathbf{f}_3} \right)$$

$$egin{array}{lll} \mathbf{h}_1 &=& \mathbf{a}[\mathbf{f}_0] \ \mathbf{f}_1 &=& oldsymbol{eta}_1 + oldsymbol{\Omega}_1 \mathbf{h}_1 \ \mathbf{h}_2 &=& \mathbf{a}[\mathbf{f}_1] \end{array}$$

 $\mathbf{f}_2 = \boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2 \mathbf{h}_2$

 $\mathbf{h}_3 = \mathbf{a}[\mathbf{f}_2]$ $\mathbf{f}_3 = \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3$ $\ell_i = \mathbf{l}[\mathbf{f}_3, y_i],$

$$(x_i) \xrightarrow{} (f_0) \xrightarrow{} (h_1) \xrightarrow{} (f_1) \xrightarrow{} (h_2) \xrightarrow{} (f_2) \xrightarrow{} (h_3) \xrightarrow{} (f_3) \xrightarrow{} (\ell_i)$$

$$\underline{\partial \ell_i} \quad \underline{\partial \mathbf{h}_2} \ \underline{\partial \mathbf{h}_2} \ \underline{\partial \mathbf{h}_2} \ \underline{\partial \mathbf{h}_3} \ \underline{\partial \mathbf{h}_3} \ \underline{\partial \ell_i}) \qquad \mathbf{h}_1 \quad = \quad \mathbf{a}[\mathbf{f}_0]$$

$$\frac{\partial \ell_{i}}{\partial \mathbf{f}_{1}} = \frac{\partial \mathbf{h}_{2}}{\partial \mathbf{f}_{1}} \frac{\partial \mathbf{f}_{2}}{\partial \mathbf{h}_{2}} \left(\frac{\partial \mathbf{h}_{3}}{\partial \mathbf{f}_{2}} \frac{\partial \mathbf{f}_{3}}{\partial \mathbf{h}_{3}} \frac{\partial \ell_{i}}{\partial \mathbf{f}_{3}} \right)$$

$$\frac{\partial \ell_{i}}{\partial \mathbf{f}_{0}} = \frac{\partial \mathbf{h}_{1}}{\partial \mathbf{f}_{0}} \frac{\partial \mathbf{f}_{1}}{\partial \mathbf{h}_{1}} \left(\frac{\partial \mathbf{h}_{2}}{\partial \mathbf{f}_{1}} \frac{\partial \mathbf{f}_{2}}{\partial \mathbf{h}_{2}} \frac{\partial \mathbf{h}_{3}}{\partial \mathbf{f}_{2}} \frac{\partial \mathbf{f}_{3}}{\partial \mathbf{h}_{3}} \frac{\partial \ell_{i}}{\partial \mathbf{f}_{3}} \right)$$

 $egin{array}{lll} {f f}_1 & = & m{eta}_1 + m{\Omega}_1 {f h}_1 \ {f h}_2 & = & {f a}[{f f}_1] \ {f f}_2 & = & m{eta}_2 + m{\Omega}_2 {f h}_2 \ {f h}_3 & = & {f a}[{f f}_2] \ {f f}_3 & = & m{eta}_3 + m{\Omega}_3 {f h}_3 \end{array}$

 $\ell_i = \mathbf{l}[\mathbf{f}_3, y_i],$

The derivative $\partial \ell_i/\partial \mathbf{f}_3$ of the loss ℓ_i with respect to the network output \mathbf{f}_3 will depend on the loss function but usually has a simple form.

The derivative $\partial \mathbf{f}_3/\partial \mathbf{h}_3$ of the network output with respect to hidden layer \mathbf{h}_3 is:

$$\begin{split} \frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} &= \frac{\partial}{\partial \mathbf{h}_3} \left(\boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3 \right) = \boldsymbol{\Omega}_3^T \\ \frac{\partial \ell_i}{\partial \boldsymbol{\beta}_k} &= \frac{\partial \mathbf{f}_k}{\partial \boldsymbol{\beta}_k} \frac{\partial \ell_i}{\partial \mathbf{f}_k} \\ &= \frac{\partial}{\partial \boldsymbol{\beta}_k} \left(\boldsymbol{\beta}_k + \boldsymbol{\Omega}_k \mathbf{h}_k \right) \frac{\partial \ell_i}{\partial \mathbf{f}_k} \\ &= \frac{\partial \ell_i}{\partial \mathbf{f}_k}, \end{split}$$

 $\mathbf{f}_0 = oldsymbol{eta}_0 + oldsymbol{\Omega}_0 \mathbf{x}_i$ $\mathbf{h}_1 = \mathbf{a}[\mathbf{f}_0]$

 $\mathbf{f}_1 = \boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1 \mathbf{h}_1$

 $\mathbf{h}_2 = \mathbf{a}[\mathbf{f}_1]$ $\mathbf{f}_2 = \boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2 \mathbf{h}_2$ $\mathbf{h}_3 = \mathbf{a}[\mathbf{f}_2]$

 $\mathbf{f}_3 = \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3$ $\ell_i = \mathbf{l}[\mathbf{f}_3, y_i],$

$$egin{array}{lll} rac{\partial \ell_i}{\partial oldsymbol{eta}_k} &=& rac{\partial \mathbf{f}_k}{\partial oldsymbol{eta}_k} rac{\partial \ell_i}{\partial \mathbf{f}_k} \ &=& rac{\partial}{\partial oldsymbol{eta}_k} \left(oldsymbol{eta}_k + oldsymbol{\Omega}_k \mathbf{h}_k
ight) rac{\partial \ell_i}{\partial \mathbf{f}_k} \ &=& rac{\partial \ell_i}{\partial \mathbf{f}_k}, \ rac{\partial \ell_i}{\partial oldsymbol{eta}_k} & \partial \ell_i \end{array}$$

$$\frac{\partial \ell_i}{\partial \mathbf{\Omega}_k} = \frac{\partial \mathbf{f}_k}{\partial \mathbf{\Omega}_k} \frac{\partial \ell_i}{\partial \mathbf{f}_k}
= \frac{\partial}{\partial \mathbf{\Omega}_k} (\boldsymbol{\beta}_k + \mathbf{\Omega}_k \mathbf{h}_k) \frac{\partial \ell_i}{\partial \mathbf{f}_k}
= \frac{\partial \ell_i}{\partial \mathbf{f}_k} \mathbf{h}_k^T.$$

$$= \frac{\partial}{\partial \mathbf{\Omega}_{k}} (\boldsymbol{\beta}_{k} + \mathbf{\Omega}_{k} \mathbf{h}_{k}) \frac{\partial \ell_{i}}{\partial \mathbf{f}_{k}}$$

$$= \frac{\partial \ell_{i}}{\partial \mathbf{f}_{k}} \mathbf{h}_{k}^{T}.$$

$$\frac{\partial \ell_{i}}{\partial \boldsymbol{\beta}_{0}} = \frac{\partial \ell_{i}}{\partial \mathbf{f}_{0}}$$

$$\frac{\partial \ell_{i}}{\partial \mathbf{\Omega}_{0}} = \frac{\partial \ell_{i}}{\partial \mathbf{f}_{0}} \mathbf{x}_{i}^{T}$$

 $\mathbf{f}_0 = oldsymbol{eta}_0 + oldsymbol{\Omega}_0 \mathbf{x}_i$

 $\mathbf{f}_1 = \boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1 \mathbf{h}_1$

 $\mathbf{f}_2 = \boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2 \mathbf{h}_2$

 $\mathbf{f}_3 = \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3$

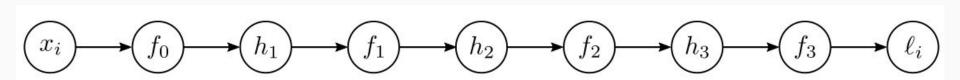
 $\ell_i = \mathbf{l}[\mathbf{f}_3, y_i],$

 $\mathbf{h}_1 = \mathbf{a}[\mathbf{f}_0]$

 $\mathbf{h}_2 = \mathbf{a}[\mathbf{f}_1]$

 $\mathbf{h}_3 = \mathbf{a}[\mathbf{f}_2]$

Forward pass



An example:

$$f_0 = \beta_0 + \omega_0 \cdot x_i$$

$$h_1 = \sin[f_0]$$

$$f_1 = \beta_1 + \omega_1 \cdot h_1$$

$$h_2 = \exp[f_1]$$

$$f_2 = \beta_2 + \omega_2 \cdot h_2$$

$$h_3 = \cos[f_2]$$

$$f_3 = \beta_3 + \omega_3 \cdot h_3$$

$$\ell_i = (f_3 - y_i)^2$$

Backward pass

$$\underbrace{\begin{pmatrix} \partial \ell_i \\ \partial f_0 \end{pmatrix}}^{\bullet} \underbrace{\begin{pmatrix} \partial h_1 \\ \partial h_1 \end{pmatrix}}^{\bullet} \underbrace{\begin{pmatrix} \partial f_1 \\ \partial h_1 \end{pmatrix}}^{\bullet} \underbrace{\begin{pmatrix} \partial \ell_i \\ \partial f_1 \end{pmatrix}}^{\bullet} \underbrace{\begin{pmatrix} \partial h_2 \\ \partial f_1 \end{pmatrix}}^{\bullet} \underbrace{\begin{pmatrix} \partial f_2 \\ \partial h_2 \end{pmatrix}}^{\bullet} \underbrace{\begin{pmatrix} \partial \ell_i \\ \partial f_2 \end{pmatrix}}^{\bullet} \underbrace{\begin{pmatrix} \partial h_3 \\ \partial f_2 \end{pmatrix}}^{\bullet} \underbrace{\begin{pmatrix} \partial h_3 \\ \partial h_3 \end{pmatrix}}^{\bullet} \underbrace{\begin{pmatrix} \partial f_3 \\ \partial h_3 \end{pmatrix}}^{\bullet} \underbrace{\begin{pmatrix} \partial \ell_i \\ \partial f_3 \end{pmatrix}}^{\bullet} \underbrace{\begin{pmatrix} \partial \ell_i \\ \partial f_3$$

$$\frac{\partial \ell_i}{\partial f_3}$$
, $\frac{\partial \ell_i}{\partial h_3}$, $\frac{\partial \ell_i}{\partial f_2}$, $\frac{\partial \ell_i}{\partial h_2}$, $\frac{\partial \ell_i}{\partial f_1}$, $\frac{\partial \ell_i}{\partial h_1}$, and $\frac{\partial \ell_i}{\partial f_0}$

$$f_k = eta_k + \omega_k.h_k$$
 $rac{\partial f_i}{\partial f_3} = 2(f_3 - y_i)$ $rac{\partial f_k}{\partial f_3} = 1$ and $rac{\partial f_k}{\partial f_k} = -h_i$

 h_2

 f_2

$$\frac{\partial \ell_{i}}{\partial h_{3}} = \frac{\partial f_{3}}{\partial h_{3}} \frac{\partial \ell_{i}}{\partial f_{3}}$$

$$\frac{\partial \ell_{i}}{\partial f_{2}} = \frac{\partial h_{3}}{\partial f_{2}} \left(\frac{\partial f_{3}}{\partial h_{3}} \frac{\partial \ell_{i}}{\partial f_{3}} \right)$$

$$\frac{\partial \ell_{i}}{\partial h_{2}} = \frac{\partial f_{2}}{\partial h_{2}} \left(\frac{\partial h_{3}}{\partial f_{2}} \frac{\partial f_{3}}{\partial h_{3}} \frac{\partial \ell_{i}}{\partial f_{3}} \right)$$

$$\frac{\partial \ell_{i}}{\partial f_{1}} = \frac{\partial h_{2}}{\partial f_{1}} \left(\frac{\partial f_{2}}{\partial h_{2}} \frac{\partial h_{3}}{\partial f_{2}} \frac{\partial f_{3}}{\partial h_{3}} \frac{\partial \ell_{i}}{\partial f_{3}} \right)$$

$$\frac{\partial \ell_{i}}{\partial h_{1}} = \frac{\partial f_{1}}{\partial h_{1}} \left(\frac{\partial h_{2}}{\partial f_{1}} \frac{\partial f_{2}}{\partial h_{2}} \frac{\partial h_{3}}{\partial f_{2}} \frac{\partial f_{3}}{\partial h_{3}} \frac{\partial \ell_{i}}{\partial f_{3}} \right)$$

 $\frac{\partial \ell_i}{\partial f_0} = \frac{\partial h_1}{\partial f_0} \left(\frac{\partial f_1}{\partial h_1} \frac{\partial h_2}{\partial f_1} \frac{\partial f_2}{\partial h_2} \frac{\partial h_3}{\partial f_2} \frac{\partial f_3}{\partial h_3} \frac{\partial \ell_i}{\partial f_3} \right)$

$$\frac{\partial f_k}{\partial \beta_k} = 1$$
 and $\frac{\partial f_k}{\partial \omega_k} = h_k$

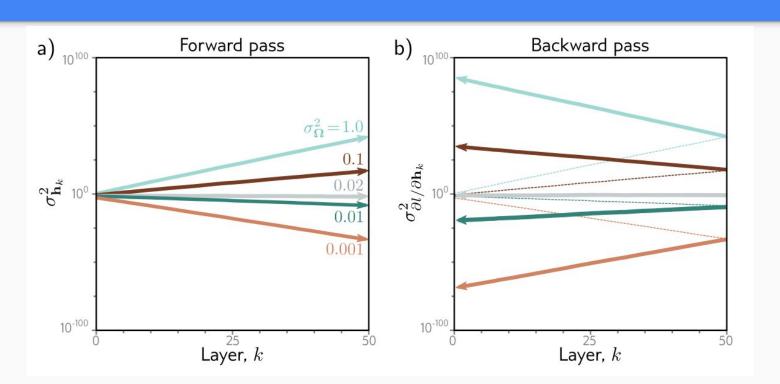
$$\frac{\partial f_0}{\partial \beta_0} = 1$$
 and $\frac{\partial f_0}{\partial \omega_0} = x_i$

Parameter initialization

$$\mathbf{f}_k = \boldsymbol{\beta}_k + \boldsymbol{\Omega}_k \mathbf{h}_k$$

= $\boldsymbol{\beta}_k + \boldsymbol{\Omega}_k \mathbf{a}[\mathbf{f}_{k-1}]$

vanishing gradient problem & exploding gradient problem



$$\mathbb{E}[f_i'] = \mathbb{E}\left[\beta_i + \sum_{j=1}^{D_h} \Omega_{ij} h_j\right]$$

$$= \mathbb{E}\left[\beta_i\right] + \sum_{j=1}^{D_h} \mathbb{E}\left[\Omega_{ij} h_j\right]$$

$$= \mathbb{E}\left[\beta_i\right] + \sum_{j=1}^{D_h} \mathbb{E}\left[\Omega_{ij}\right] \mathbb{E}\left[h_j\right]$$

 $= 0 + \sum_{i=1}^{D_h} 0 \cdot \mathbb{E}\left[h_j\right] = 0,$

$$\sigma_{f'}^{2} = \mathbb{E}[f_{i}'^{2}] - \mathbb{E}[f_{i}']^{2}$$

$$= \mathbb{E}\left[\left(\beta_{i} + \sum_{j=1}^{D_{h}} \Omega_{ij} h_{j}\right)^{2}\right] - 0$$

$$= \mathbb{E}\left[\left(\sum_{j=1}^{D_{h}} \Omega_{ij} h_{j}\right)^{2}\right]$$

$$= \sum_{j=1}^{D_{h}} \mathbb{E}\left[\Omega_{ij}^{2}\right] \mathbb{E}\left[h_{j}^{2}\right]$$

 $= \sum_{i=1}^{D_h} \sigma_{\Omega}^2 \mathbb{E}\left[h_j^2\right] = \sigma_{\Omega}^2 \sum_{i=1}^{D_h} \mathbb{E}\left[h_j^2\right],$

 $\sigma^2 = \mathbb{E}[(z - \mathbb{E}[z])^2] = \mathbb{E}[z^2] - \mathbb{E}[z]^2$

 $\sigma_{f'}^2 = \sigma_{\Omega}^2 \sum_{j=1}^{D_h} \frac{\sigma_f^2}{2} = \frac{1}{2} D_h \sigma_{\Omega}^2 \sigma_f^2$

He initialization (Kaiming Initialization)

$$\sigma_{\Omega}^2 = \frac{2}{D_h}$$



$$\sigma_{f'}^2 = \sigma_{\Omega}^2 \sum_{j=1}^{D_h} \frac{\sigma_f^2}{2} = \frac{1}{2} D_h \sigma_{\Omega}^2 \sigma_f^2$$

Initialization for both forward and backward pass

$$\sigma_{\Omega}^2 = \frac{2}{D_{h'}}$$

Initialization for both forward and backward pass

$$\sigma_{\Omega}^2 = \frac{4}{D_h + D_{h'}}$$

Understanding Deep Learning Chapter 8