1.
$$
\begin{aligned}
y \;=\;& \phi_0 + \phi_1 \mathrm{a}\Big[\psi_{01} + \psi_{11}\mathrm{a}[\theta_{01} + \theta_{11}x] + \psi_{21}\mathrm{a}[\theta_{02} + \theta_{12}x]\Big] \\
& + \phi_2 \mathrm{a}\Big[\psi_{02} + \psi_{12}\mathrm{a}[\theta_{01} + \theta_{11}x] + \psi_{22}\mathrm{a}[\theta_{02} + \theta_{12}x]\Big], \qquad (7.35)
\end{aligned}
$$

$$\frac{\partial y}{\partial \phi_0} = 1$$

$$\frac{\partial y}{\partial \phi_1} = a[\psi_{01} + \psi_{11} a[\theta_{01} + \theta_{11} x] + \psi_{21} a[\theta_{02} + \theta_{12} x]]$$

similar for ϕ_2

$$\frac{\partial y}{\partial \psi_{01}} = \phi_1 \mathbb{1}[\psi_{01} + \psi_{11} a[\theta_{01} + \theta_{11} x] + \psi_{21} a[\theta_{02} + \theta_{12} x] > 0]$$

similar for ψ_{02}

$$\frac{\partial y}{\partial \psi_{11}} = \phi_1 a[\theta_{01} + \theta_{11} x] \mathbb{1}[\psi_{01} + \psi_{11} a[\theta_{01} + \theta_{11} x] + \psi_{21} a[\theta_{02} + \theta_{12} x] > 0]$$

similar for $\psi_{12}, \psi_{21}, \psi_{22}$

$$
\begin{aligned}
\frac{\partial y}{\partial \theta_{01}} =\;& \phi_1 \psi_{11} \mathbb{1}[\theta_{01} + \theta_{11} x > 0] \mathbb{1}[\psi_{01} + \psi_{11} a[\theta_{01} + \theta_{11} x] + \psi_{21} a[\theta_{02} + \theta_{12} x] > 0] \\
& + \phi_2 \psi_{12} \mathbb{1}[\theta_{01} + \theta_{11} x > 0] \mathbb{1}[\psi_{02} + \psi_{12} a[\theta_{01} + \theta_{11} x] + \psi_{22} a[\theta_{02} + \theta_{12} x] > 0]
\end{aligned}
$$

similar for θ_{02}

$$
\begin{aligned}
\frac{\partial y}{\partial \theta_{11}} =\;& \phi_1 \psi_{11} x \mathbb{1}[\theta_{01} + \theta_{11} > 0] \mathbb{1}[\psi_{01} + \psi_{11} a[\theta_{01} + \theta_{11} x] + \psi_{21} a[\theta_{02} + \theta_{12} x] > 0] \\
& + \phi_2 \psi_{12} x \mathbb{1}[\theta_{01} + \theta_{11} x] \mathbb{1}[\psi_{02} + \psi_{12} a[\theta_{01} + \theta_{11} x] + \psi_{22} a[\theta_{02} + \theta_{12} x] > 0]
\end{aligned}
$$

similar for θ_{12}

3.
$$\frac{\partial \ell_i}{\partial \mathbf{f}_0} = \frac{\partial \mathbf{h}_1}{\partial \mathbf{f}_0}\frac{\partial \mathbf{f}_1}{\partial \mathbf{h}_1}\left(\frac{\partial \mathbf{h}_2}{\partial \mathbf{f}_1}\frac{\partial \mathbf{f}_2}{\partial \mathbf{h}_2}\frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2}\frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3}\frac{\partial \ell_i}{\partial \mathbf{f}_3}\right). \qquad (7.20)$$

$\frac{\partial l_i}{\partial f_3} : D_f \times 1$ $\qquad$ $\frac{\partial f_2}{\partial h_2} : D_2 \times D_3$ $\qquad$ $\frac{\partial h_1}{\partial f_0} : D_1 \times D_1$

$\frac{\partial f_3}{\partial h_3} : D_3 \times D_f$ $\qquad$ $\frac{\partial h_2}{\partial f_1} : D_2 \times D_2$

$\frac{\partial h_3}{\partial f_2} : D_3 \times D_3$ $\qquad$ $\frac{\partial f_1}{\partial h_1} : D_1 \times D_2$

5. $l_i = -(1-y_i)\log(1-\text{sig}[f(\underline{x}_i \phi)]) - y_i \log[\text{sig}[f(x_i,\phi)]]$

$$\frac{\partial \text{sig}(z)}{\partial z} = \frac{e^{-z}}{(1+e^{-z})^2} = \text{sig}(z)(1-\text{sig}(z))$$

$$\frac{\partial}{\partial f}\left(\log[1-\text{sig} f]\right) = \frac{-1}{\cancel{1-\text{sig}[f]}}\text{sig}(f)(\cancel{1-\text{sig}[f]})$$
$$= -\text{sig}[z]$$

$$\frac{\partial}{\partial f}(\log[\text{sig} f] = \frac{1}{\cancel{\text{sig}[f]}}\cancel{\text{sig}[f]}(1-\text{sig}[f])$$
$$= 1-\text{sig}[f]$$

$$\frac{\partial l_i}{\partial f} = (1-y_i)\text{sig}[f] - y_i(1-\text{sig}[f])$$
$$= \text{sig}[f] - y_i$$

6. $\underline{z} = \underline{\beta} + \underline{\Omega}\,\underline{h}$

$h \to (D_i \times 1)$
$z \to (D_o \times 1)$
$\Omega \to (D_o \times D_i)$

$$\frac{\partial \underline{z}}{\partial \underline{h}} = \begin{pmatrix} \frac{\partial z_1}{\partial h_1} & \frac{\partial z_2}{\partial h_1} & \cdots \frac{\partial z_j}{\partial h_1} \\ \frac{\partial z_1}{\partial h_2} & \ddots & \vdots \\ \vdots & \frac{\partial z_j}{\partial h_i} & \\ \frac{\partial z_1}{\partial h_i} & \cdots & \ddots\ \frac{\partial z_{D_o}}{\partial h_{D_i}} \end{pmatrix}$$

But, $z_j = \beta_j + \sum_{i}^{D_i} w_{ji} h_i$

so $\frac{\partial z_j}{\partial h_i} = w_{ji}$

Which makes up the transpose of $\underline{\Omega}$.

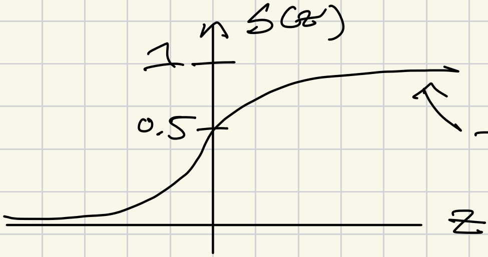
$$\frac{\partial z}{\partial \underline{h}} = \begin{pmatrix} w_{11} & w_{21} & \cdots & w_{j1} \cdots w_{D_o1} \\ w_{12} & \ddots & & \\ w_{i1}\vdots & & w_{ji} & \\ w_{1D_i} & & \ddots & w_{D_oD_i} \end{pmatrix} = \underline{\Omega}^T$$

7. $\mathrm{sig}[f] = \dfrac{1}{1+e^{-f}} = h$

$$\frac{\partial h}{\partial f} = \mathrm{sig}[f]\,[1-\mathrm{sig}[f]]$$

$$= \frac{1}{1+e^{-f}}\left(1-\frac{1}{1+e^{-f}}\right)$$

The gradients vanish in either case



the gradients get smaller and smaller.

9. $l[f]$

$$\underline{f} = \underline{\beta} + \underline{\Omega}\,\underline{h}$$

$$f_i = \beta_i + \textstyle\sum_j \Omega_{ij} h_j$$

$$\frac{\partial f_i}{\partial \Omega_{ij}} = h_j \rightarrow \frac{\partial \underline{f}}{\partial \underline{\Omega}} = \underline{h}^T$$

$$\frac{\partial l}{\partial \underline{\Omega}} = \frac{\partial l}{\partial \underline{f}}\,\frac{\partial \underline{f}}{\partial \underline{\Omega}} = \frac{\partial l}{\partial \underline{f}}\,\underline{h}^T$$

10. For regular ReLU, $\underline{h}' = a[\underline{f}]$, $\underline{f} = \underline{\beta} + \underline{\Omega}\,\underline{h}$

$$\frac{\partial}{\partial f} a[f] = \mathbb{1}[f>0] \leftarrow \text{indicator function.}$$

Now, instead of the derivative becoming 0, it is a small constant α.

so $\dfrac{\partial \underline{h}'}{\partial \underline{h}} = \mathbb{1}[\underline{f}>0] \odot \underline{\Omega}^T + \mathbb{1}[\underline{f}<0] \odot \alpha\,\underline{\Omega}^T$

14. $\mathrm{Var}[a] = \mathbb{E}[a^2] - \mathbb{E}[a]$

$\therefore \mathbb{E}[a^2] = \sigma^2.$

$\mathbb{E}[b^2] = \mathbb{E}[a^2 \cdot \mathbb{1}(a > 0)]$

Because a is centred around the mean, probability of a being positive is $1/2$. $\mathbb{E}[b] = 0$, because $\mathbb{E}[a] = 0$. Hence $\mathbb{E}[b^2] = \mathrm{Var}[b^2]$

$\mathbb{E}[b^2] = \frac{1}{2}\mathbb{E}[a^2] = \frac{1}{2}\sigma^2.$

15. The network won't train well because.

eg. $$\frac{\partial \ell_i}{\partial \mathbf{f}_0} = \frac{\partial \mathbf{h}_1}{\partial \mathbf{f}_0}\frac{\partial \mathbf{f}_1}{\partial \mathbf{h}_1}\left(\frac{\partial \mathbf{h}_2}{\partial \mathbf{f}_1}\frac{\partial \mathbf{f}_2}{\partial \mathbf{h}_2}\frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2}\frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3}\frac{\partial \ell_i}{\partial \mathbf{f}_3}\right).$$

Network with 3 hidden layers.

$$\frac{\partial \mathbf{f}_i}{\partial \mathbf{h}_j} = \frac{\partial}{\partial \mathbf{h}_j}[\boldsymbol{\beta}_i + \boldsymbol{\Omega}_i \mathbf{h}_i]$$
$$= \boldsymbol{\Omega}_i^T$$

If the weights are all initialised to zero then the gradients of the loss will be zero too.