Homework 1

Solution 1.2

Change in notation:

$$\text{Output of } Linear_1 : \boldsymbol{z^{(1)}} \to \boldsymbol{s}^{(1)}.$$
$$\text{Output of } f : \boldsymbol{z^{(2)}} \to \boldsymbol{z}^{(1)}.$$
$$\text{Output of } Linear_2 : \boldsymbol{z^{(3)}} \to \boldsymbol{s}^{(2)}.$$
$$\text{Output of } g : \text{Remains the same, } \boldsymbol{\hat{y}}.$$

Solution a)

1. `torch.nn.Linear`: $\text{Linear}(\boldsymbol{x}) = W\boldsymbol{x} + b$.
2. `torch.nn.ReLU`: $\text{ReLU}(\boldsymbol{x}) = \max(\boldsymbol{0}, \boldsymbol{x})$.
3. `torch.nn.Linear`: $\text{Linear}(\boldsymbol{x}) = W\boldsymbol{x} + b$.
4. `torch.nn.ReLU`: $\text{ReLU}(x) = \max(\boldsymbol{0}, \boldsymbol{x})$.
5. `torch.nn.MSELoss`: $l_{\text{MSE}}(\boldsymbol{\hat{y}}, \boldsymbol{y}) = ||\boldsymbol{\hat{y}} - \boldsymbol{y}||^2$.

Solution b)

Strictly using $\boldsymbol{x}, \boldsymbol{y}, W^{(1)}, W^{(2)}, b^{(1)}, b^{(2)}$:

Layer	Input	Output
Linear_1	$\boldsymbol{x}$	$W^{(1)}\boldsymbol{x} + b^{(1)}$
f	$W^{(1)}\boldsymbol{x} + b^{(1)}$	$\text{ReLU}(W^{(1)}\boldsymbol{x} + b^{(1)})$
Linear_2	$\text{ReLU}(W^{(1)}\boldsymbol{x} + b^{(1)})$	$W^{(2)}\text{ReLU}(W^{(1)}\boldsymbol{x} + b^{(1)}) + b^{(2)}$
g	$W^{(2)}\text{ReLU}(W^{(1)}\boldsymbol{x} + b^{(1)}) + b^{(2)}$	$\text{I}(W^{(2)}\text{ReLU}(W^{(1)}\boldsymbol{x} + b^{(1)}) + b^{(2)})$
Loss	$\boldsymbol{\hat{y}}, \text{I}(W^{(2)}\text{ReLU}(W^{(1)}\boldsymbol{x} + b^{(1)}) + b^{(2)})$	$(\boldsymbol{\hat{y}} - \text{I}(W^{(2)}\text{ReLU}(W^{(1)}\boldsymbol{x} + b^{(1)}) + b^{(2)}))(\boldsymbol{\hat{y}} - \text{I}(W^{(2)}\text{ReLU}(W^{(1)}\boldsymbol{x} + b^{(1)}) + b^{(2)}))^T$

Using intermediate variables:

Layer	Input	Output
Linear_1	$\boldsymbol{x}$	$\boldsymbol{s^{(1)}} = W^{(1)}\boldsymbol{x} + b^{(1)}$
f	$\boldsymbol{s^{(1)}}$	$\boldsymbol{z^{(1)}} = \text{ReLU}(\boldsymbol{s^{(1)}})$
Linear_2	$\boldsymbol{z^{(1)}}$	$\boldsymbol{s^{(2)}} = W^{(2)}\boldsymbol{z^{(1)}} + b^{(2)}$
g	$\boldsymbol{s^{(2)}}$	$\boldsymbol{\hat{y}} = \text{I}(\boldsymbol{s^{(2)}})$
Loss	$\boldsymbol{\hat{y}}, \boldsymbol{y}$	$\ell_{\text{MSE}} = (\boldsymbol{\hat{y}} - \boldsymbol{y})(\boldsymbol{\hat{y}} - \boldsymbol{y})^T$

Using components:

Layer	Input	Output
Linear_1	x_j	$s_i^{(1)} = \sum_j W_{ij}^{(1)} x_j + b_i^{(1)}$
f	$s_i^{(1)}$	$z_i^{(1)} = \text{ReLU}(s_i^{(1)})$
Linear_2	$z_i^{(1)}$	$s_k^{(2)} = \sum_i W_{ki}^{(2)} z_i^{(1)} + b_k^{(2)}$

Layer	Input	Output
g	$s_k^{(2)}$	$y_k = g(s_k^{(2)})$
Loss	$\hat{y}_k, y_k$	$\ell_{\text{MSE}} = \sum_k (\hat{y}_k - y_k)(\hat{y}_k - y_k)$

Solution c)

Dimensions

Following numerator layout:

$$\begin{aligned}
\boldsymbol{x} &: d_{\boldsymbol{x}} \times 1. \\
\boldsymbol{s}^{(1)} &: d_{\boldsymbol{s}^{(1)}} \times 1. \\
\boldsymbol{z}^{(1)} &: d_{\boldsymbol{z}^{(1)}} \times 1. \\
\boldsymbol{s}^{(2)} &: d_{\boldsymbol{s}^{(2)}} \times 1. \\
\hat{\boldsymbol{y}} &: d_{\hat{\boldsymbol{y}}} \times 1. \\
W^{(1)} &: d_{\boldsymbol{s}^{(1)}} \times d_{\boldsymbol{x}}. \\
W^{(2)} &: d_{\boldsymbol{s}^{(2)}} \times d_{\boldsymbol{z}^{(1)}}. \\
b^{(1)} &: d_{\boldsymbol{s}^{(1)}} \times 1. \\
b^{(2)} &: d_{\boldsymbol{s}^{(2)}} \times 1. \\
\frac{\partial \ell}{\partial W^{(2)}} &: d_{\boldsymbol{z}^{(1)}} \times d_{\boldsymbol{s}^{(2)}}. \\
\frac{\partial \ell}{\partial W^{(1)}} &: d_{\boldsymbol{x}} \times d_{\boldsymbol{s}^{(1)}}. \\
\frac{\partial \ell}{\partial b^{(2)}} &: 1 \times d_{\boldsymbol{s}^{(2)}}. \\
\frac{\partial \ell}{\partial b^{(1)}} &: 1 \times d_{\boldsymbol{s}^{(1)}}.
\end{aligned}$$

Where:

$$\begin{aligned}
d_{\boldsymbol{s}^{(1)}} &= d_{\boldsymbol{z}^{(1)}}. \\
d_{\boldsymbol{s}^{(2)}} &= d_{\hat{\boldsymbol{y}}}.
\end{aligned}$$

Gradient of $W^{(2)}$

Using chain rule and tensor notation:

$$\begin{aligned}
\frac{\partial \ell}{\partial W_{ij}^{(2)}} &= \sum_{k,l} \frac{\partial \ell}{\partial \hat{y}_k} \frac{\partial \hat{y}_k}{\partial s_l^{(2)}} \frac{\partial s_l^{(2)}}{\partial W_{ij}^{(2)}}. \\
&= \sum_{k,l} \frac{\partial \ell}{\partial \hat{y}_k} \frac{\partial \hat{y}_k}{\partial s_l^{(2)}} \frac{\partial}{\partial W_{ij}^{(2)}} \left(\sum_m W_{lm}^{(2)} z_m^{(1)} + b_l^{(2)} \right). \\
&= \sum_{k,l} \frac{\partial \ell}{\partial \hat{y}_k} \frac{\partial \hat{y}_k}{\partial s_l^{(2)}} z_m^{(1)} \delta_{il} \delta_{jm}. \\
&= \sum_{k} \frac{\partial \ell}{\partial \hat{y}_k} \frac{\partial \hat{y}_k}{\partial s_i^{(2)}} z_j^{(1)}. \\
&= \delta_i^{(2)} z_j^{(1)}.
\end{aligned}$$

In matrix form:

$$\frac{\partial \ell}{\partial W^{(2)}} = \begin{pmatrix} \frac{\partial \ell}{\partial W^{(2)}_{00}} & \frac{\partial \ell}{\partial W^{(2)}_{10}} & \cdots & \frac{\partial L}{\partial W^{(2)}_{d_{\boldsymbol{s}(2)}0}} \\ \frac{\partial \ell}{\partial W^{(2)}_{01}} & \frac{\partial \ell}{\partial W^{(2)}_{11}} & \cdots & \frac{\partial \ell}{\partial W^{(2)}_{d_{\boldsymbol{s}(2)}1}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \ell}{\partial W^{(2)}_{0d_{\boldsymbol{z}(1)}}} & \frac{\partial \ell}{\partial W^{(2)}_{1d_{\boldsymbol{z}(1)}}} & \cdots & \frac{\partial \ell}{\partial W^{(2)}_{d_{\boldsymbol{s}(2)}d_{\boldsymbol{z}(1)}}} \end{pmatrix}.$$

$$= \begin{pmatrix} z^{(1)}_0 \\ \vdots \\ z^{(1)}_{d_{\boldsymbol{z}(1)}} \end{pmatrix} \begin{pmatrix} \frac{\partial \ell}{\partial \hat{y}_0} & \cdots & \frac{\partial \ell}{\partial \hat{y}_{d_{\hat{\boldsymbol{y}}}}} \end{pmatrix} \begin{pmatrix} \frac{\partial \hat{y}_0}{\partial s^{(2)}_0} & \frac{\partial \hat{y}_0}{\partial s^{(2)}_1} & \cdots & \frac{\partial \hat{y}_0}{\partial s^{(2)}_{d_{\boldsymbol{s}(2)}}} \\ \frac{\partial \hat{y}_1}{\partial s^{(2)}_0} & \frac{\partial \hat{y}_1}{\partial s^{(2)}_1} & \cdots & \frac{\partial \hat{y}_1}{\partial s^{(2)}_{d_{\boldsymbol{s}(2)}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \hat{y}_{d_{\hat{\boldsymbol{y}}}}}{\partial s^{(2)}_0} & \frac{\partial \hat{y}_{d_{\hat{\boldsymbol{y}}}}}{\partial s^{(2)}_1} & \cdots & \frac{\partial \hat{y}_{d_{\hat{\boldsymbol{y}}}}}{\partial s^{(2)}_{d_{\boldsymbol{s}(2)}}} \end{pmatrix} = \boldsymbol{z^{(1)}} \frac{\partial \ell}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{s^{(2)}}}.$$

$$= \begin{pmatrix} z^{(1)}_0 \\ \vdots \\ z^{(1)}_{d_{\boldsymbol{z}(1)}} \end{pmatrix} \begin{pmatrix} \delta^{(2)}_0 & \cdots & \delta^{(2)}_{d_{\boldsymbol{s}(2)}} \end{pmatrix} = \boldsymbol{z^{(1)}} [\boldsymbol{\delta}^{(2)}]^T.$$

This results are for any activation function and any loss, in our case:

$$\frac{\partial \hat{y}_k}{\partial \boldsymbol{s}^{(2)}_i} = \frac{\partial}{\partial \boldsymbol{s}^{(2)}_i} \mathrm{g}\left(\boldsymbol{s}^{(2)}_k\right) = \delta_{ki}.$$

$$\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{s^{(2)}}} = I_{d_{\hat{\boldsymbol{y}}} \times d_{\boldsymbol{s}(2)}}.$$

And for the loss:

$$\begin{aligned} \frac{\partial \ell}{\partial \hat{y}_k} &= \frac{\partial}{\partial \hat{y}_k} \left[\sum_i (\hat{y}_i - y_i)^2 \right]. \\ &= \sum_i 2(\hat{y}_i - y_i)\delta_{ik}. \\ &= 2(\hat{y}_k - y_k). \end{aligned}$$

$$\frac{\partial L}{\partial \hat{\boldsymbol{y}}} = 2(\hat{\boldsymbol{y}} - \boldsymbol{y})^{\mathrm{T}}.$$

Inserting that in the formula for $\boldsymbol{\delta^{(2)}}$:

$$\delta^{(2)}_i = 2(\hat{y}_i - y_i).$$

$$\boldsymbol{\delta^{(2)}} = 2(\hat{\boldsymbol{y}} - \boldsymbol{y}) = 2 \begin{pmatrix} \hat{y}_0 - y_0 \\ \vdots \\ \hat{y}_{d_{\boldsymbol{s}(2)}} - y_{d_{\boldsymbol{s}(2)}} \end{pmatrix}.$$

And for $\frac{\partial \ell}{\partial W^{(2)}}$:

$$\frac{\partial \ell}{\partial W_{ij}^{(2)}} = 2(\hat{y}_i - y_i)z_j^{(1)}.$$

$$\frac{\partial \ell}{\partial W^{(2)}} = 2\begin{pmatrix} z_0^{(1)} \\ \vdots \\ z_{d_{z^{(1)}}}^{(1)} \end{pmatrix} \begin{pmatrix} \hat{y}_0 - y_0 & \dots & \hat{y}_{d_{s^{(2)}}} - y_{d_{s^{(2)}}} \end{pmatrix}.$$
$$= 2\boldsymbol{z}^{(1)}(\hat{\boldsymbol{y}} - \boldsymbol{y})^T$$

Gradient of $b^{(2)}$

Using chain rule and components and having into account the previous results for $W^{(2)}$:

$$\begin{aligned}
\frac{\partial \ell}{\partial b_i^{(2)}} &= \sum_{k,l} \frac{\partial \ell}{\partial \hat{y}_k} \frac{\partial \hat{y}_k}{\partial s_l^{(2)}} \frac{\partial s_l^{(2)}}{\partial b_i^{(2)}}. \\
&= \sum_{k,l} \frac{\partial l}{\partial \hat{y}_k} \frac{\partial \hat{y}_k}{\partial s_l^{(2)}} \frac{\partial}{\partial b_i^{(2)}} \left(\sum_m W_{lm}^{(2)} z_m^{(1)} + b_l^{(2)} \right). \\
&= \sum_{k,l} \frac{\partial \ell}{\partial \hat{y}_k} \frac{\partial \hat{y}_k}{\partial s_l^{(2)}} \delta_{il}. \\
&= \sum_{k} \frac{\partial \ell}{\partial \hat{y}_k} \frac{\partial \hat{y}_k}{\partial s_i^{(2)}}. \\
&= \delta_i^{(2)}. \\
&= 2(\hat{y}_i - y_i).
\end{aligned}$$

In matrix form:

$$\begin{aligned}
\frac{\partial L}{\partial b^{(2)}} &= \begin{pmatrix} \frac{\partial \ell}{\partial b_0^{(2)}} & \frac{\partial \ell}{\partial b_1^{(2)}} & \cdots \end{pmatrix} \\
&= \begin{pmatrix} \frac{\partial \ell}{\partial \hat{y}_0} & \cdots & \frac{\partial \ell}{\partial \hat{y}_{d_{\hat{y}}}} \end{pmatrix} \begin{pmatrix} \frac{\partial \hat{y}_0}{\partial s_0^{(2)}} & \frac{\partial \hat{y}_0}{\partial s_1^{(2)}} & \cdots & \frac{\partial \hat{y}_0}{\partial s_{d_{s^{(2)}}}^{(2)}} \\ \frac{\partial \hat{y}_1}{\partial s_0^{(2)}} & \frac{\partial \hat{y}_1}{\partial s_1^{(2)}} & \cdots & \frac{\partial \hat{y}_1}{\partial s_{d_{s^{(2)}}}^{(2)}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \hat{y}_{d_{\hat{y}}}}{\partial s_0^{(2)}} & \frac{\partial \hat{y}_{d_{\hat{y}}}}{\partial s_1^{(2)}} & \cdots & \frac{\partial \hat{y}_{d_{\hat{y}}}}{\partial s_{d_{s^{(2)}}}^{(2)}} \end{pmatrix}. \\
&= \frac{\partial \ell}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{s}^{(2)}}. \\
&= [\boldsymbol{\delta}^{(2)}]^T. \\
&= 2(\hat{\boldsymbol{y}} - \boldsymbol{y})^T. \\
&= 2\begin{pmatrix} \hat{y}_0 - y_0 & \dots & \hat{y}_{d_{s^{(2)}}} - y_{d_{s^{(2)}}} \end{pmatrix}.
\end{aligned}$$

Gradient of $W^{(1)}$

Using chain rule and tensor notation:

$$\begin{aligned}\frac{\partial \ell}{\partial W^{(1)}_{ij}} &= \sum_{k,l,m,n} \frac{\partial \ell}{\partial \hat{y}_k} \frac{\partial \hat{y}_k}{\partial s^{(2)}_l} \frac{\partial s^{(2)}_l}{\partial z^{(1)}_m} \frac{\partial z^{(1)}_m}{\partial s^{(1)}_n} \frac{\partial s^{(1)}_n}{\partial W^{(1)}_{ij}}. \\ &= \sum_{l,m,n} \delta^{(2)}_l \frac{\partial s^{(2)}_l}{\partial z^{(1)}_m} \frac{\partial z^{(1)}_m}{\partial s^{(1)}_n} \frac{\partial s^{(1)}_n}{\partial W^{(1)}_{ij}}. \\ &= \sum_{n} \delta^{(1)}_n \frac{\partial s^{(1)}_n}{\partial W^{(1)}_{ij}}. \\ &= \sum_{n} \delta^{(1)}_i x_j.\end{aligned}$$

$$\frac{\partial \ell}{\partial W^{(1)}} = \boldsymbol{x}\left[\boldsymbol{\delta^{(1)}}\right]^T.$$

Where $\boldsymbol{\delta^{(L=1)}}$ are the so called "errors" for the linear layer $L=1$. Then, we can compute $\frac{\partial \ell}{\partial W^{(1)}}$ in terms of the jacobians:

$$\begin{aligned}\frac{\partial \ell}{\partial W^{(1)}} &= \boldsymbol{x}\left[\boldsymbol{\delta^{(1)}}\right]^T. \\ \boldsymbol{\delta^{(1)}} &= \left[\frac{\partial \boldsymbol{s^{(2)}}}{\partial \boldsymbol{z^{(1)}}} \frac{\partial \boldsymbol{z^{(1)}}}{\partial \boldsymbol{s^{(1)}}}\right]^T \to \boldsymbol{x}\left[\left[\frac{\partial \boldsymbol{s^{(2)}}}{\partial \boldsymbol{z^{(1)}}} \frac{\partial \boldsymbol{z^{(1)}}}{\partial \boldsymbol{s^{(1)}}}\right]^T \boldsymbol{\delta^{(2)}}\right]^T. \\ \boldsymbol{\delta^{(2)}} &= \left[\frac{\partial \ell}{\partial \boldsymbol{\hat{y}}} \frac{\partial \boldsymbol{\hat{y}}}{\partial \boldsymbol{s^{(2)}}}\right]^T \to \boldsymbol{x} \frac{\partial \ell}{\partial \boldsymbol{\hat{y}}} \frac{\partial \boldsymbol{\hat{y}}}{\partial \boldsymbol{s^{(2)}}} \frac{\partial \boldsymbol{s^{(2)}}}{\partial \boldsymbol{z^{(1)}}} \frac{\partial \boldsymbol{z^{(1)}}}{\partial \boldsymbol{s^{(1)}}}.\end{aligned}$$

Gradient of $W^{(L)}$

The errors are easily generalizable:

$$\delta^{(L)}_i = \sum_{p,q} \delta^{(L+1)}_p \frac{\partial s^{(L+1)}_p}{\partial z^{(L)}_q} \frac{\partial z^{(L)}_q}{\partial s^{(L)}_i}.$$

In matrix form:

$$\begin{aligned}\boldsymbol{\delta}^{(L)} &= \left[\begin{pmatrix}\delta^{(L+1)}_0 & \cdots & \delta^{(L+1)}_{d_{\boldsymbol{s}(2)}}\end{pmatrix} \begin{pmatrix} \frac{\partial s^{(L+1)}_0}{\partial z^{(L)}_0} & \frac{\partial s^{(L+1)}_0}{\partial z^{(L)}_1} & \cdots & \frac{\partial s^{(L+1)}_0}{\partial z^{(L)}_{d_{\boldsymbol{z}(L)}}} \\ \frac{\partial s^{(L+1)}_1}{\partial z^{(L)}_0} & \frac{\partial s^{(L+1)}_1}{\partial z^{(L)}_1} & \cdots & \frac{\partial s^{(L+1)}_1}{\partial z^{(L)}_{d_{\boldsymbol{z}(L)}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial s^{(L+1)}_{d_{\boldsymbol{s}(L+1)}}}{\partial z^{(L)}_0} & \frac{\partial s^{(L+1)}_{d_{\boldsymbol{s}(L+1)}}}{\partial z^{(L)}_1} & \cdots & \frac{\partial s^{(L+1)}_{d_{\boldsymbol{s}(L+1)}}}{\partial z^{(L)}_{d_{\boldsymbol{z}(L)}}} \end{pmatrix} \begin{pmatrix} \frac{\partial z^{(L)}_0}{\partial s^{(L)}_0} & \frac{\partial z^{(L)}_0}{\partial s^{(L)}_1} & \cdots & \frac{\partial z^{(L)}_0}{\partial s^{(L)}_{d_{\boldsymbol{s}(L)}}} \\ \frac{\partial z^{(L)}_1}{\partial s^{(L)}_0} & \frac{\partial z^{(L)}_1}{\partial s^{(L)}_1} & \cdots & \frac{\partial z^{(L)}_1}{\partial s^{(L)}_{d_{\boldsymbol{s}(L)}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial z^{(L)}_{d_{\boldsymbol{z}(L)}}}{\partial s^{(L)}_0} & \frac{\partial z^{(L)}_{d_{\boldsymbol{z}(L)}}}{\partial s^{(L)}_1} & \cdots & \frac{\partial z^{(L)}_{d_{\boldsymbol{z}(L)}}}{\partial s^{(L)}_{d_{\boldsymbol{s}(L)}}} \end{pmatrix}\right]^T. \\ &= \left[\left[\boldsymbol{\delta^{(L+1)}}\right]^T \frac{\partial \boldsymbol{s^{(L+1)}}}{\partial \boldsymbol{z^{(L)}}} \frac{\partial \boldsymbol{z^{(L)}}}{\partial \boldsymbol{s^{(L)}}}\right]^T. \\ &= \left[\frac{\partial \boldsymbol{s^{(L+1)}}}{\partial \boldsymbol{z^{(L)}}} \frac{\partial \boldsymbol{z^{(L)}}}{\partial \boldsymbol{s^{(L)}}}\right]^T \boldsymbol{\delta^{(L+1)}}.\end{aligned}$$

Now, let's compute $\frac{\partial \boldsymbol{s^{(L+1)}}}{\partial \boldsymbol{z^{(L)}}}$ for a linear layer:

$$\frac{\partial s_i^{(L+1)}}{\partial z_j^{(L)}} = \frac{\partial}{\partial z_j^{(L)}} \left(\sum_k W_{ik}^{(L+1)} z_k^{(L)} + b_i^{(L+1)} \right).$$
$$\frac{\partial s_i^{(L+1)}}{\partial z_j^{(L)}} = W_{ij}^{(L+1)}.$$

$$\frac{\partial \boldsymbol{s}^{(\boldsymbol{L+1})}}{\partial \boldsymbol{z}^{(\boldsymbol{L})}} = W^{(L+1)}.$$

Taking into account the previous expressions, we can compute the gradient for any linear layer and any activation function:

$$\frac{\partial \ell}{\partial W^{(L)}} = \boldsymbol{z}^{(\boldsymbol{L-1})} \left[\boldsymbol{\delta}^{(\boldsymbol{L})} \right]^T.$$
$$\boldsymbol{\delta}^{(\boldsymbol{L})} = \left[W^{(L+1)} \frac{\partial \boldsymbol{z}^{(\boldsymbol{L})}}{\partial \boldsymbol{s}^{(\boldsymbol{L})}} \right]^T \boldsymbol{\delta}^{(\boldsymbol{L+1})}.$$
$$\boldsymbol{z}^{(\boldsymbol{0})} = \boldsymbol{x}.$$
$$\boldsymbol{\delta}^{(\boldsymbol{L}_{\max})} = \left[\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{s}^{(\boldsymbol{L}_{\max})}} \right]^T.$$

In regard to $\frac{\partial \boldsymbol{z}^{(\boldsymbol{L})}}{\partial \boldsymbol{s}^{(\boldsymbol{L})}}$, we can compute it for $g = I(\cdot)$ and $f = \mathrm{ReLU}(\cdot)$ (one of the most common cases):

$$f = \mathrm{ReLU}(\cdot) \rightarrow \frac{\partial z_i^{(L)}}{\partial s_j^{(L)}} = \max(0, \mathrm{sign}(s_j^L))\delta_{ij}.$$

$$\frac{\partial \boldsymbol{z}^{(\boldsymbol{L})}}{\partial \boldsymbol{s}^{(\boldsymbol{L})}} = I^{+\boldsymbol{s}^{(\boldsymbol{L})}}_{\boldsymbol{z}^{(\boldsymbol{L})} \times \boldsymbol{s}^{(\boldsymbol{L})}} = \begin{pmatrix} \max(0, \mathrm{sign}(s_0^{(L)})) & 0 & \dots & 0 \\ 0 & \max(0, \mathrm{sign}(s_1^{(L)})) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \max(0, \mathrm{sign}(s_{d_{\boldsymbol{s}^{(L)}}}^{(L)})) \end{pmatrix}$$

$$g = I(\cdot) \rightarrow \frac{\partial z_i^{(L)}}{\partial s_j^{(L)}} = \delta_{ij}.$$

$$\frac{\partial \boldsymbol{z}^{(\boldsymbol{L})}}{\partial \boldsymbol{s}^{(\boldsymbol{L})}} = I_{\boldsymbol{z}^{(\boldsymbol{L})} \times \boldsymbol{s}^{(\boldsymbol{L})}} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & 1 \end{pmatrix}.$$

Then, the "errors" for any linear layer are given by:

$$\begin{aligned} \delta_i^{(L)} &= \sum_{p,q} \delta_p^{(L+1)} \frac{\partial z_q^{(L)}}{\partial s_i^{(L)}} \frac{\partial}{\partial z_q^{(L)}} \left(\sum_l W_{pl}^{(L+1)} z_l^{(L)} + b_p^{(L+1)} \right). \\ &= \sum_{p,q} \delta_p^{(L+1)} W_{pq}^{(L+1)} \frac{\partial z_q^{(L)}}{\partial s_i^{(L)}}. \end{aligned}$$

$$\boldsymbol{\delta}^{(\boldsymbol{L})} = \left[W^{L+1} \frac{\partial \boldsymbol{z}^{(\boldsymbol{L})}}{\partial \boldsymbol{s}^{(\boldsymbol{L})}} \right]^T \boldsymbol{\delta}^{(\boldsymbol{L+1})}.$$

For a $\mathrm{ReLU}(\cdot)$:

$$\boldsymbol{\delta}^{(\boldsymbol{L})} = \left[W^{L+1} I^{+\boldsymbol{s}^{(\boldsymbol{L})}}_{\boldsymbol{z}^{(\boldsymbol{L})} \times \boldsymbol{s}^{(\boldsymbol{L})}} \right]^T \boldsymbol{\delta}^{(\boldsymbol{L+1})}.$$

As an example, let's particularize our computation of $\frac{\partial \ell}{\partial W^{(1)}}$:

$$\begin{aligned}\frac{\partial L}{\partial W^{(1)}} &= \boldsymbol{x}\left[\boldsymbol{\delta^{(1)}}\right]^T. \\ &= \boldsymbol{x}\left[\left[W^{(2)} I^{+\boldsymbol{s}^{(1)}}_{\boldsymbol{z}^{(1)}\times \boldsymbol{s}^{(1)}}\right]^T \boldsymbol{\delta^{(2)}}\right]^T. \\ &= 2\boldsymbol{x}(\hat{\boldsymbol{y}} - \boldsymbol{y})^T W^{(2)} I^{+\boldsymbol{s}^{(1)}}_{\boldsymbol{z}^{(1)}\times \boldsymbol{s}^{(1)}}.\end{aligned}$$

Gradient of $b^{(1)}$

Following the same idea:

$$\begin{aligned}\frac{\partial \ell}{\partial b_i^{(1)}} &= \sum_{k,l,m,n} \frac{\partial \ell}{\partial \hat{y}_k} \frac{\partial \hat{y}_k}{\partial s_l^{(2)}} \frac{\partial s_l^{(2)}}{\partial z_m^{(1)}} \frac{\partial z_m^{(1)}}{\partial s_n^{(1)}} \frac{\partial s_n^{(1)}}{\partial b_i^{(1)}}. \\ &= \sum_n \delta_n^{(1)} \frac{\partial s_n^{(1)}}{\partial b_i^{(1)}}. \\ &= \delta_i^{(1)}.\end{aligned}$$

$$\frac{\partial \ell}{\partial b^{(1)}} = \left[\boldsymbol{\delta^{(1)}}\right]^T.$$

In terms of the jacobians:

$$\frac{\partial \ell}{\partial b^{(1)}} = \frac{\partial \ell}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{s^{(2)}}} \frac{\partial \boldsymbol{s^{(2)}}}{\partial \boldsymbol{z^{(1)}}} \frac{\partial \boldsymbol{z^{(1)}}}{\partial \boldsymbol{s^{(1)}}}.$$

Gradient of $b^{(L)}$

For any linear layer, the gradient respect to the bias is:

$$\frac{\partial \ell}{\partial b_i^{(L)}} = \delta_i^{(L)}.$$

$$\frac{\partial \ell}{\partial b^{(L)}} = \left[\boldsymbol{\delta^{(L)}}\right]^T.$$

Where $\boldsymbol{\delta}^{(\boldsymbol{L}_{\max})}$ is given in the previous section. Let's particularize for our special case:

$$\frac{\partial L}{\partial b^{(1)}} = 2(\hat{\boldsymbol{y}} - \boldsymbol{y})^T W^{(2)} I^{+\boldsymbol{s}^{(1)}}_{\boldsymbol{z}^{(1)}\times \boldsymbol{s}^{(1)}}.$$

Summary

Shapes:

$$\begin{aligned}\boldsymbol{s^{(L)}} &= 1 \times d_{\boldsymbol{s}^{(L)}}. \\ \boldsymbol{z^{(L)}} &= 1 \times d_{\boldsymbol{z}^{(L)}}. \\ W^{(L)} &: d_{\boldsymbol{s}^{(L)}} \times d_{\boldsymbol{z}^{(L-1)}}. \\ b^{(L)} &: d_{\boldsymbol{s}^{(L)}} \times 1. \\ \frac{\partial \ell}{\partial W^{(L)}} &: d_{\boldsymbol{z}^{(L-1)}} \times d_{\boldsymbol{s}^{(L)}}. \\ \frac{\partial \ell}{\partial b^{(L)}} &: 1 \times d_{\boldsymbol{s}^{(L)}}.\end{aligned}$$

Where:

$$d_{\boldsymbol{z}^{(0)}} = d_{\boldsymbol{x}}.$$
$$d_{\boldsymbol{z}^{(L_{\max})}} = d_{\hat{\boldsymbol{y}}} = d_{\boldsymbol{y}}.$$

Backpropagation for a stack of linear layers in matrix form:

$$\frac{\partial \ell}{\partial W^{(L)}} = \boldsymbol{z}^{(\boldsymbol{L}-\mathbf{1})}\left[\boldsymbol{\delta}^{(\boldsymbol{L})}\right]^T.$$
$$\boldsymbol{\delta}^{(\boldsymbol{L})} = \left[W^{(L+1)}\frac{\partial \boldsymbol{z}^{(\boldsymbol{L})}}{\partial \boldsymbol{s}^{(\boldsymbol{L})}}\right]^T \boldsymbol{\delta}^{(\boldsymbol{L}+\mathbf{1})}.$$
$$\frac{\partial \ell}{\partial b^{(L)}} = \left[\boldsymbol{\delta}^{(\boldsymbol{L})}\right]^T.$$
$$\boldsymbol{z}^{(\mathbf{0})} = \boldsymbol{x}.$$
$$\boldsymbol{\delta}^{(\boldsymbol{L}_{\max})} = \left[\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{s}^{(\boldsymbol{L}_{\max})}}\right]^T.$$

$\frac{\partial \boldsymbol{z}^{(\boldsymbol{L})}}{\partial \boldsymbol{s}^{(\boldsymbol{L})}}$ for a $\mathrm{ReLU}(\cdot)$:

$$\frac{\partial \boldsymbol{z}^{(\boldsymbol{L})}}{\partial \boldsymbol{s}^{(\boldsymbol{L})}} = I^{+\boldsymbol{s}^{(L)}}_{\boldsymbol{z}^{(L)}\times \boldsymbol{s}^{(L)}} = \begin{pmatrix} \max(0,\mathrm{sign}(s^L_0)) & 0 & \cdots & 0 \\ 0 & \max(0,\mathrm{sign}(s^L_1)) & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \max(0,\mathrm{sign}(s^L_{d_{\boldsymbol{s}^{(L)}}})) \end{pmatrix}.$$

Parameter	Gradient	Gradient shape
$W^{(1)}$	$\boldsymbol{x}\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{s}^{(2)}}\frac{\partial \boldsymbol{s}^{(2)}}{\partial \boldsymbol{z}^{(1)}}\frac{\partial \boldsymbol{z}^{(1)}}{\partial \boldsymbol{s}^{(1)}} = 2\boldsymbol{x}(\hat{\boldsymbol{y}}-\boldsymbol{y})^T W^{(2)} I^{+\boldsymbol{s}^{(1)}}_{\boldsymbol{z}^{(1)}\times \boldsymbol{s}^{(1)}}.$	$d_{\boldsymbol{x}} \times d_{\boldsymbol{s}^{(1)}}.$
$b^{(1)}$	$\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{s}^{(2)}}\frac{\partial \boldsymbol{s}^{(2)}}{\partial \boldsymbol{z}^{(1)}}\frac{\partial \boldsymbol{z}^{(1)}}{\partial \boldsymbol{s}^{(1)}} = 2(\hat{\boldsymbol{y}}-\boldsymbol{y})^T W^{(2)} I^{+\boldsymbol{s}^{(1)}}_{\boldsymbol{z}^{(1)}\times \boldsymbol{s}^{(1)}}.$	$1 \times d_{\boldsymbol{s}^{(1)}}.$
$W^{(2)}$	$\boldsymbol{z}^{(\mathbf{1})}\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{s}^{(2)}} = 2\boldsymbol{z}^{(\mathbf{1})}(\hat{\boldsymbol{y}}-\boldsymbol{y})^T.$	$d_{\boldsymbol{z}^{(1)}} \times d_{\boldsymbol{s}^{(2)}}.$
$b^{(2)}$	$\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{s}^{(2)}} = 2(\hat{\boldsymbol{y}}-\boldsymbol{y})^T.$	$1 \times d_{\boldsymbol{s}^{(2)}}.$

Solution d)

With the change in notation:

$$\frac{\partial \boldsymbol{z}^{(\mathbf{2})}}{\partial \boldsymbol{z}^{(\mathbf{1})}} \rightarrow \frac{\partial \boldsymbol{z}^{(\mathbf{1})}}{\partial \boldsymbol{s}^{(\mathbf{1})}}.$$
$$\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}^{(\mathbf{3})}} \rightarrow \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{s}^{(\mathbf{2})}}.$$

$\frac{\partial \boldsymbol{z}^{(1)}}{\partial \boldsymbol{s}^{(1)}}$:

$$f = \mathrm{ReLU}(\cdot) \rightarrow \frac{\partial z^{(1)}_i}{\partial s^{(1)}_j} = \frac{\partial}{\partial s^{(1)}_j}\mathrm{ReLU}(s^{(1)}_i) = \max(0,\mathrm{sign}(s^{(1)}_j))\delta_{ij}.$$

$$\frac{\partial \boldsymbol{z}^{(\mathbf{1})}}{\partial \boldsymbol{s}^{(\mathbf{1})}} = I^{+\boldsymbol{s}^{(1)}}_{d_{\boldsymbol{z}^{(1)}}\times \boldsymbol{s}^{(1)}} = \begin{pmatrix} \max(0,\mathrm{sign}(s^{(1)}_0)) & 0 & \cdots & 0 \\ 0 & \max(0,\mathrm{sign}(s^{(1)}_1)) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \max(0,\mathrm{sign}(s^{(1)}_{d_{\boldsymbol{s}^{(1)}}})) \end{pmatrix}.$$

$\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{s}^{(2)}}$:

$$g = I(\cdot) \rightarrow \frac{\partial \hat{y}_i}{\partial s_j^{(2)}} = \frac{\partial}{\partial s_j^{(2)}} I(s_i^{(2)}) = \delta_{ij}, \;\; i = 0, \dots, d_{\hat{\boldsymbol{y}}}, \; j = 0, \dots, d_{\boldsymbol{s}^{(2)}}.$$

$$\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{s}^{(2)}} = I_{d_{\hat{y}} \times \boldsymbol{s}^{(2)}} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & 1 \end{pmatrix}.$$

$\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}$:

$$\begin{aligned} \frac{\partial \ell}{\partial \hat{y}_i} &= \frac{\partial}{\partial \hat{y}_i} \left[\sum_j (\hat{y}_j - y_j)^2 \right]. \\ &= \sum_j 2(\hat{y}_j - y_j)\delta_{ij}. \\ &= 2(\hat{y}_i - y_i). \end{aligned}$$

$$\begin{aligned} \frac{\partial \ell}{\partial \hat{\boldsymbol{y}}} &= 2(\hat{\boldsymbol{y}} - \boldsymbol{y})^{\mathrm{T}}. \\ &= 2 \begin{pmatrix} \hat{y}_0 - y_0 & \dots & \hat{y}_{d_{\hat{y}}} - y_{d_y} \end{pmatrix} \end{aligned}$$

Solution 1.3

Solution a)

In the case of **b)** the loss function (the replacement is done in the table with intermediate variables only) :

Layer	Input	Output
Linear_1	$\boldsymbol{x}$	$\boldsymbol{s}^{(1)} = W^{(1)}\boldsymbol{x} + b^{(1)}$
σ	$\boldsymbol{s}^{(1)}$	$\boldsymbol{z}^{(1)} = \sigma(\boldsymbol{s}^{(1)})$
Linear_2	$\boldsymbol{z}^{(1)}$	$\boldsymbol{s}^{(2)} = W^{(2)}\boldsymbol{z}^{(1)} + b^{(2)}$
σ	$\boldsymbol{s}^{(2)}$	$\hat{\boldsymbol{y}} = \sigma(\boldsymbol{s}^{(2)})$
Loss	$\hat{\boldsymbol{y}}, \boldsymbol{y}$	$\ell_{\mathrm{MSE}} = (\hat{\boldsymbol{y}} - \boldsymbol{y})(\hat{\boldsymbol{y}} - \boldsymbol{y})^T$

In the case of **c)** the jacobians $\frac{\partial \boldsymbol{z}^{(1)}}{\partial \boldsymbol{s}^{(1)}}$ and $\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{s}^{(2)}}$.

In the case of **d)** , we need to compute the derivatives so we can see the components explicitly, the derivative of σ is:

$$\sigma' = \sigma(1 - \sigma).$$

Then, $\frac{\partial \boldsymbol{z}^{(1)}}{\partial \boldsymbol{s}^{(1)}}$:

$$f = \sigma(\cdot) \rightarrow \frac{\partial z_i^{(1)}}{\partial s_j^{(1)}} = \frac{\partial}{\partial s_j^{(1)}} \sigma(s_i^{(1)}) = \sigma(s_i^{(1)})(1 - \sigma(s_i^{(1)}))\delta_{ij}.$$

$$\frac{\partial \boldsymbol{z}^{(1)}}{\partial \boldsymbol{s}^{(1)}} = \begin{pmatrix} \sigma(s_0^{(1)})(1 - \sigma(s_0^{(1)})) & 0 & \dots & 0 \\ 0 & \sigma(s_1^{(1)})(1 - \sigma(s_1^{(1)})) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma(s_{d_{s^{(1)}}}^{(1)})(1 - \sigma(s_{d_{s^{(1)}}}^{(1)})) \end{pmatrix}.$$

$\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{s}^{(2)}}$:

$$g = \sigma(\cdot) \rightarrow \frac{\partial \hat{y}_i}{\partial s_j^{(2)}} = \frac{\partial}{\partial s_j^{(2)}} \sigma(s_i^{(2)}) = \sigma(s_i^{(2)})(1 - \sigma(s_i^{(2)}))\delta_{ij}.$$

$$\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{s}^{(2)}} = \begin{pmatrix} \sigma(s_0^{(2)})(1-\sigma(s_0^{(2)})) & 0 & \cdots & 0 \\ 0 & \sigma(s_1^{(2)})(1-\sigma(s_1^{(2)})) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma(s_{d_{s^{(2)}}}^{(2)})(1-\sigma(s_{d_{s^{(2)}}}^{(2)})) \end{pmatrix}.$$

$\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}$ remains the same.

Solution b)

In the equations of **b)** only the loss function, $\ell_{\mathrm{MSE}} \rightarrow \ell_{\mathrm{BCE}}$

Layer	Input	Output
Linear_1	$\boldsymbol{x}$	$\boldsymbol{s}^{(1)} = W^{(1)}\boldsymbol{x} + b^{(1)}$
σ	$\boldsymbol{s}^{(1)}$	$\boldsymbol{z}^{(1)} = \sigma(\boldsymbol{s}^{(1)})$
Linear_2	$\boldsymbol{z}^{(1)}$	$\boldsymbol{s}^{(2)} = W^{(2)}\boldsymbol{z}^{(1)} + b^{(2)}$
σ	$\boldsymbol{s}^{(2)}$	$\hat{\boldsymbol{y}} = \sigma(\boldsymbol{s}^{(2)})$
Loss	$\hat{\boldsymbol{y}}, \boldsymbol{y}$	$\ell_{\mathrm{BCE}} = -\frac{1}{K}\left[\boldsymbol{y}^T \log(\hat{\boldsymbol{y}}) + (1-\boldsymbol{y})^T \log(1-\hat{\boldsymbol{y}})\right]$

In the equations of **c)** the derivative $\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}$.

In the equations of **d)**, since the derivative $\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}$ changes, so do its components, let's compute them and write the matrix representation:

$$\ell_{\mathrm{BCE}} = -\frac{1}{K}\sum_j \left[y_j \log(\hat{y}_j) + (1-y_j)\log(1-\hat{y}_j)\right].$$

$$\frac{\partial \ell_{\mathrm{BCE}}}{\partial \hat{y}_i} = \frac{1}{K}\frac{\hat{y}_i - y_i}{\hat{y}_i(1-\hat{y}_i)}.$$

$$\frac{\partial \ell_{\mathrm{BCE}}}{\partial \hat{\boldsymbol{y}}} = \frac{1}{K}\begin{pmatrix} \frac{\hat{y}_0 - y_0}{\hat{y}_0(1-\hat{y}_0)} & \frac{\hat{y}_1 - y_1}{\hat{y}_1(1-\hat{y}_1)} & \cdots & \frac{\hat{y}_{d_{\hat{y}}} - y_{d_{\hat{y}}}}{\hat{y}_{d_{\hat{y}}}(1-\hat{y}_{d_{\hat{y}}})} \end{pmatrix}.$$

Solution c)

Because the the calculation and the calculation of the gradient is faster and ReLU is good avoiding gradient vanishing.