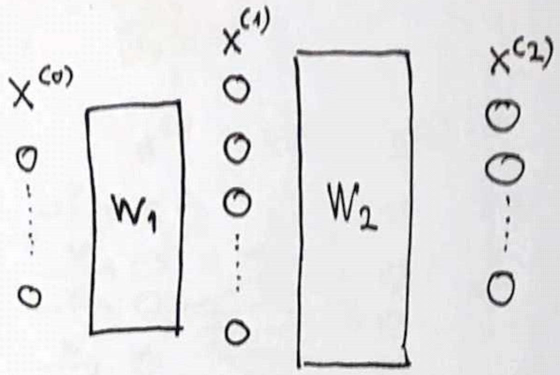
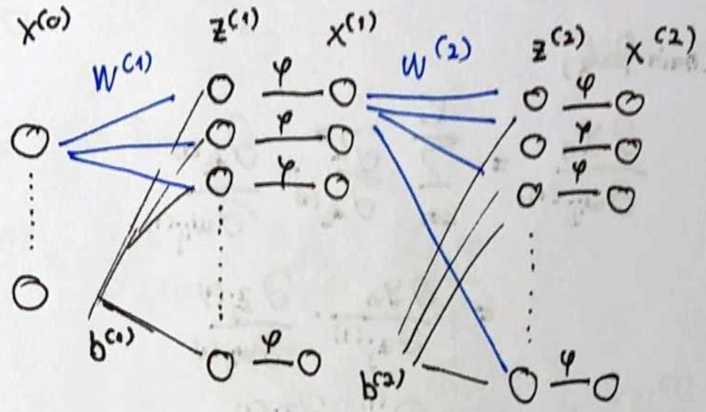
MNIST from Scratch

Model:



Detailed:



Params:

$W^{(1)} \in \mathbb{R}^{K \times 784}$

$W^{(2)} \in \mathbb{R}^{K \times 10}$

$b^{(1)} \in \mathbb{R}^{K}$

$b^{(2)} \in \mathbb{R}^{10}$

Loss:

$$\mathcal{L}_n(x^{(n)}) = \frac{1}{2}\left(\sum_{k=1}^{10}\left[f_k(x^{(n)}) - y_k^{(n)}\right]^2\right)$$

Goal:

$$\frac{\partial \mathcal{L}_n}{\partial w_{ij}^{(1)}} \qquad \frac{\partial \mathcal{L}_n}{\partial b_i^{(1)}}$$

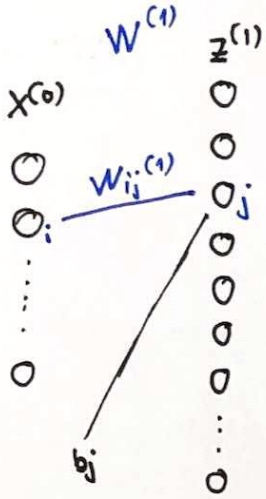
$$\frac{\partial \mathcal{L}_n}{\partial w_{ij}^{(2)}} \qquad \frac{\partial \mathcal{L}_n}{\partial b_i^{(2)}}$$

Notation: Let $\partial_j^{(1)} = \frac{\partial \mathcal{L}_n}{\partial z_j^{(1)}}$

$\partial_j^{(2)} = \frac{\partial \mathcal{L}_n}{\partial z_j^{(2)}}$

$\partial_j^{(3)} = \partial \mathcal{L}_n$

First Layer:



By the Chain Rule,

$$\frac{\partial \mathcal{L}_n}{\partial w_{ij}^{(1)}} = \sum_{k=1}^{K} \frac{\partial L^n}{\partial z_k^{(1)}} \cdot \frac{\partial z_k^{(1)}}{\partial w_{ij}^{(1)}}$$

$$= \frac{\partial \mathcal{L}_n}{\partial z_j^{(1)}} \cdot \frac{\partial z_j^{(1)}}{\partial w_{ij}^{(1)}}$$

$$= \delta_j^{(1)} \cdot \frac{\partial z_j^{(1)}}{\partial w_{ij}^{(1)}}$$

Since $z^{(1)} = \left[W^{(1)}\right]^T x^{(0)} + b^{(1)}$,

$$z_j^{(1)} = \sum_{k=1}^{D} \left[W^{(1)}\right]^T_{jk} x_k^{(0)} + b_j = \sum_{k=1}^{D} w_{kj}^{(1)} x_k^{(0)} + b_j$$

Now $$\frac{\partial z_j^{(1)}}{\partial w_{ij}^{(1)}} = \frac{\partial}{\partial w_{ij}^{(1)}} \left(\sum_{k=1}^{D} w_{kj}^{(1)} x_k^{(0)} + b_j \right)$$

$$= \sum_{k=1}^{D} \frac{\partial}{\partial w_{ij}^{(1)}} \left(w_{kj}^{(1)} x_k^{(0)} \right)$$

$$= x_i^{(0)}$$

Hence $$\frac{\partial \mathcal{L}_n}{\partial w_{ij}^{(1)}} = \delta_j^{(1)} \cdot x_i^{(0)}.$$

By the Chain Rule,

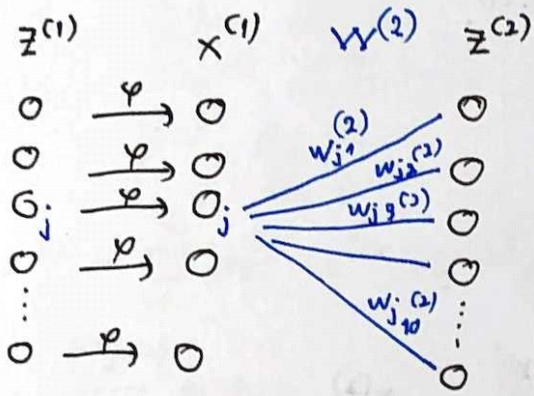
$$\frac{\partial \mathcal{L}_n}{\partial b_i^{(1)}} = \sum_{k=1}^{D} \frac{\partial \mathcal{L}_n}{\partial z_k^{(1)}} \cdot \frac{\partial z_k^{(1)}}{\partial b_i^{(1)}}$$

$$= \frac{\partial \mathcal{L}_n}{\partial z_i^{(1)}} \cdot \frac{\partial z_i^{(1)}}{\partial b_i^{(1)}}$$

Since $z_i^{(1)} = \sum_{k=1}^{D} w_{ki}^{(1)} x_k^{(0)} + b_i$, $\frac{\partial z_i^{(1)}}{\partial b_i^{(1)}} = 1$.

Hence $$\frac{\partial \mathcal{L}_n}{\partial b_i^{(1)}} = \frac{\partial \mathcal{L}_n}{\partial z_i^{(1)}} = \delta_i^{(1)}.$$

Now we focus on $\partial_j^{(1)} = \dfrac{\partial \mathcal{L}_n}{\partial z_j^{(1)}}$.



We have, by the Chain Rule,

$$\partial_j^{(1)} = \frac{\partial \mathcal{L}_n}{\partial z_j^{(1)}} = \frac{\partial \mathcal{L}_n}{\partial x_j^{(1)}} \cdot \frac{\partial x_j^{(1)}}{\partial z_j^{(1)}}$$

$$= \left(\sum_{k=1}^{10} \frac{\partial \mathcal{L}_n}{\partial z_k^{(2)}} \cdot \frac{\partial z_k^{(2)}}{\partial x_j^{(1)}} \right) \cdot \frac{\partial x_j^{(1)}}{\partial z_j^{(1)}}.$$

$$= \left[\sum_{k=1}^{10} \partial_k^{(2)} \cdot \frac{\partial z_k^{(2)}}{\partial x_j^{(1)}} \right] \cdot \frac{\partial x_j^{(1)}}{\partial z_j^{(1)}}.$$

We hae $x_j^{(1)} = \varphi(z_j^{(1)}) \Longrightarrow \dfrac{\partial x_j^{(1)}}{\partial z_j^{(1)}} = \varphi'(z_j^{(1)})$

We have $z_k^{(2)} = \left(W^{(2)T} x^{(1)} + b^{(2)} \right)_k$

$$= \sum_{i=1}^{K} (w^{2})^{T}_{ki} x_i^{(1)} + b_k^{(2)} = \sum_{i=1}^{K} w_{ik}^{(2)} x_i^{(1)} + b_k^{(2)}$$

Hence $\dfrac{\partial z_k^{(2)}}{\partial x_j^{(1)}} = \displaystyle\sum_{i=1}^{K} \frac{\partial}{\partial x_j^{(1)}} \left[w_{ik}^{(2)} x_i^{(1)} \right] + b_k^{(2)}$

$$= w_{jk}^{(2)}$$

Now $\partial_j^{(1)} = \left(\displaystyle\sum_{k=1}^{10} \partial_k^{(2)} \cdot w_{jk} \right) \varphi'(z_j^{(1)})$

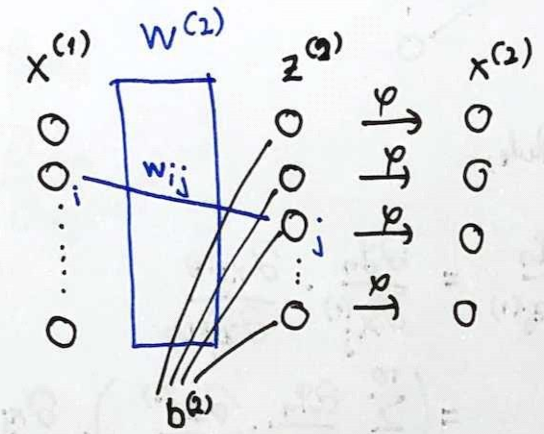
$$= \left(\sum_{k=1}^{10} w_{jk}^{(2)} \partial_k^{(2)} \right) \varphi'(z_j^{(1)})$$

Observe that $\partial^{(1)} = W^{(2)} \partial^{(2)} \odot \varphi'(z^{(1)})$.

Completing First Layer: $\begin{cases} \dfrac{\partial \mathcal{L}_n}{\partial w_{ij}^{(1)}} = \partial_j^{(1)} \cdot x_i^{(0)} \\ \dfrac{\partial \mathcal{L}_n}{\partial b_i^{(1)}} = \partial_i^{(1)}. \end{cases}$

Now consider the Second Layer:

Second Layer



We want $\dfrac{\partial \mathcal{L}_n}{\partial w_{ij}^{(2)}}, \dfrac{\partial \mathcal{L}_n}{\partial b_i^{(2)}} = ?$

$$\text{Now } \frac{\partial \mathcal{L}_n}{\partial w_{ij}^{(2)}} = \sum_{k=1}^{10} \frac{\partial z_k^{(2)}}{\partial w_{ij}^{(2)}} \cdot \frac{\partial \mathcal{L}^n}{\partial z_k^{(2)}}$$

$$= \sum_{k=1}^{10} \partial_k^{(2)} \cdot \frac{\partial z_k^{(2)}}{\partial w_{ij}^{(2)}}$$

$$= \partial_j^{(2)} \cdot \frac{\partial z_j^{(2)}}{\partial w_{ij}^{(2)}}$$

$$\text{We have } z_j^{(2)} = \left(\left(W^{(2)^T} x^{(1)} \right) + b^{(2)} \right)_j$$

$$= \sum_{k=1}^{K} w_{jk}^{(2)T} \cdot x_k^{(1)} + b_j^{(2)}$$

$$= \sum_{k=1}^{K} w_{kj}^{(2)} x_k^{(1)} + b_j^{(2)}.$$

$$\text{Now } \frac{\partial z_j^{(2)}}{\partial w_{ij}^{(2)}} = \sum_{k=1}^{K} \frac{\partial}{\partial w_{ij}^{(2)}} \left(w_{kj}^{(2)} x_k^{(1)} \right)$$

$$= x_i^{(1)}$$

$$\text{Hence } \frac{\partial \mathcal{L}_n}{\partial w_{ij}^{(2)}} = \partial_j^{(2)} x_i^{(1)}.$$

We have $\frac{\partial \mathcal{L}_n}{\partial b_i^{(2)}} = \sum_{k=1}^{K} \frac{\partial z_k^{(2)}}{\partial b_i^{(2)}} \cdot \frac{\partial \mathcal{L}_n}{\partial z_k^{(2)}}$

$= \frac{\partial \mathcal{L}_n}{\partial z_i^{(2)}} \cdot \frac{\partial z_i^{(2)}}{\partial b_i^{(2)}}$

$= \delta_i^{(2)} \cdot \frac{\partial z_i^{(2)}}{\partial b_i^{(2)}}$.

Since $z_i^{(2)} = \sum_{k=1}^{K} w_{ki}^{(2)} x_k^{(1)} + b_i^{(2)}$,

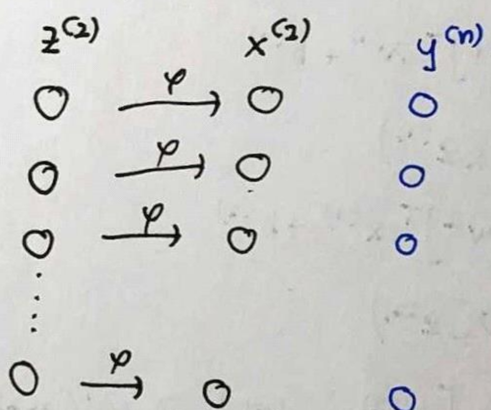
$\frac{\partial z_i^{(2)}}{\partial b_i^{(2)}} = 1$.

Therefore, $\frac{\partial \mathcal{L}_n}{\partial b_i^{(2)}} = \delta_i^{(2)} \cdot 1 = \delta_i^{(2)}$.

It remains to compute $\delta_j^{(2)}$.

By definition, $\delta_j^{(2)} = \frac{\partial \mathcal{L}_n}{\partial z_j^{(2)}}$.

We consider the relevant segment of the network:



By the Chain Rule,

$\delta_j^{(2)} = \frac{\partial \mathcal{L}_n}{\partial z_j^{(2)}} = \frac{\partial \mathcal{L}_n}{\partial x_j^{(2)}} \cdot \frac{\partial x_j^{(2)}}{\partial z_j^{(2)}}$

We have $x_j^{(2)} = \varphi(z_j^{(2)})$.

Hence $\frac{\partial x_j^{(2)}}{\partial z_j^{(2)}} = \varphi'(z_j^{(2)})$.

The interesting part is $\frac{\partial y^n}{\partial x_j^{(2)}}$.

We have $\mathcal{L}_n(x^{(n)}) = \frac{1}{2}\sum_{k=1}^{10}(x_k^{(2)} - y_k^{(n)})^2$

Hence $\frac{\partial \mathcal{L}_n}{\partial x_k^{(2)}} = \frac{1}{2}\cdot 2\,(x_k^{(2)} - y_k^{(n)})$

$= x_k^{(2)} - y_k^{(n)}$

$= [x^{(2)} - y^{(n)}]_k.$

Now $\partial_j^{(2)} = [x^{(2)} - y^{(n)}]_j\,\varphi'(z_j^{(2)})$

In vector form, $\partial^{(2)} = (x^{(2)} - y^{(n)}) \odot \varphi'(z^{(2)})$.

Hence we are ready to state forward and backward pass.

We have forward:

$x^{(0)} = x_n$

$z^{(1)} = [W^1]^T x^{(0)} + b^{(1)}$

$x^{(1)} = \phi(z^{(1)})$

$z^{(2)} = [W^2]^T x^{(1)} + b^{(2)}$

$x^{(2)} = \phi(z^{(2)})$

$K\times 1 \times 1 \times 10$

Backward: $\partial^{(2)} = [x^{(2)} - y^{(n)}] \odot \varphi'(z^{(2)})$

$\partial^{(1)} = W^{(2)}\partial^{(2)} \odot \varphi'(z^{(1)})$

$\begin{bmatrix} x_1^{(1)} \\ \vdots \\ x_K^{(1)} \end{bmatrix} [\partial_1^{(2)}\ \partial_2^{(2)} \dots \partial_{10}^{(2)}]$

$\frac{\partial \mathcal{L}_n}{\partial w_{ij}^{(2)}} = \partial_j^{(2)} x_i^{(1)} \Rightarrow \frac{\partial \mathcal{L}_n}{\partial W^{(2)}} = x^{(1)}(\partial^2)^T$ $\quad x_1\partial_1 \quad x_1\partial_2 \quad x_1\partial_3$

$\frac{\partial \mathcal{L}_n}{\partial b_j^{(2)}} = \partial_j^{(2)} \rightarrow \frac{\partial \mathcal{L}_n}{\partial b^{(2)}} = \partial^{(2)}$

$\frac{\partial \mathcal{L}_n}{\partial W^{(1)}} = x^{(0)}(\partial^1)^T$

$\frac{\partial \mathcal{L}_n}{\partial b^{(1)}} = \partial^{(1)}$