# Machine Learning: Neural Network Exercises

Fall Semester 2024, USI — TA: Vincent Herrmann (vincent.herrmann@usi.ch)

**Question 1**. Linear/affine transforms: Compute the derivative.

Scalar case: $u = wx + b$. The variables $w, x$ and $b$ are scalars.

$$\frac{\partial}{\partial b} wx + b =$$
$$\frac{\partial}{\partial w} wx + b =$$
$$\frac{\partial}{\partial x} wx + b =$$

Vectorized case: $\mathbf{u} = W\mathbf{x} + \mathbf{b}$, or $u_j = \left(\sum_k w_{jk} x_k\right) + b_j$, where $u_{\cdot}, x_{\cdot}$ and $b_{\cdot}$ are the scalar elements of the vectors $\mathbf{u} \in \mathbb{R}^d, \mathbf{x} \in \mathbb{R}^k$ and $\mathbf{b} \in \mathbb{R}^d$, and $w_{\cdot\cdot}$ are the scalar elements of the matrix $W \in \mathbb{R}^{d \times k}$.

$$\frac{\partial \mathbf{u}}{\partial \mathbf{b}} = \frac{\partial}{\partial \mathbf{b}} W\mathbf{x} + \mathbf{b} = \mathbb{I}$$
$$\frac{\partial u_j}{\partial b_i} = \frac{\partial}{\partial b_i} \left(\sum_k w_{jk} x_k\right) + b_j =$$

$$\frac{\partial \mathbf{u}}{\partial W} = \frac{\partial}{\partial W} W\mathbf{x} + \mathbf{b} =$$
$$\frac{\partial u_j}{\partial w_{il}} = \frac{\partial}{\partial w_{il}} \left(\sum_k w_{jk} x_k\right) + b_j =$$

$$\frac{\partial \mathbf{u}}{\partial \mathbf{x}} = \frac{\partial}{\partial \mathbf{x}} W\mathbf{x} + \mathbf{b} = W$$
$$\frac{\partial u_j}{\partial x_l} = \frac{\partial}{\partial x_l} \left(\sum_k w_{jk} x_k\right) + b_j =$$

Batched vectorized case: $U = XW^T + \mathbf{1} \otimes \mathbf{b}$, or $u_{nj} = \left(\sum_k w_{jk} x_{nk}\right) + b_j$. Now $U \in \mathbb{R}^{l \times d}$ and $X \in \mathbb{R}^{l \times k}$ are matrices with scalar elements $u_{..}$ and $x_{..}$, respectively.

$$\frac{\partial U}{\partial \mathbf{b}} = \frac{\partial}{\partial \mathbf{b}} XW^T + \mathbf{1} \otimes \mathbf{b} =$$

$$\frac{\partial u_{nj}}{\partial b_i} = \frac{\partial}{\partial b_i} \sum_k w_{jk} x_{nk} + b_j =$$

$$\frac{\partial U}{\partial W} = \frac{\partial}{\partial W} XW^T + \mathbf{1} \otimes \mathbf{b} =$$

$$\frac{\partial u_{nj}}{\partial w_{il}} = \frac{\partial}{\partial w_{il}} \sum_k w_{jk} x_{nk} + b_j =$$

$$\frac{\partial U}{\partial X} = \frac{\partial}{\partial X} XW^T + \mathbf{1} \otimes \mathbf{b} =$$

$$\frac{\partial u_{nj}}{\partial x_{ml}} = \frac{\partial}{\partial x_{ml}} \sum_k w_{jk} x_{nk} + b_j =$$

**Question 2**. Nonlinearity and Loss function: Compute the derivative.

Mean Squared Error: $E = \frac{1}{2} \sum_n \sum_k (z_{nk} - y_{nk})^2$

$$\frac{\partial E}{\partial z_{ml}} = \frac{\partial}{\partial z_{ml}} \frac{1}{2} \sum_n \sum_k (z_{nk} - y_{nk})^2 =$$

Sigmoid: $z = \sigma(u) = \frac{1}{1+e^{-u}}$

$$\frac{\partial z}{\partial u} = \frac{\partial}{\partial u} \sigma(u) = \frac{\partial}{\partial u} \frac{1}{1 + e^{-u}} =$$

Softmax: $s_j = \left(\sum_k e^{u_k}\right)^{-1}$

$$\frac{\partial s_j}{\partial u_i} = \frac{\partial}{\partial u_i} e^{u_j} \left(\sum_k e^{u_k}\right)^{-1} =$$

**Question 3**. Recall the multivariate chain rule: Let $r_i = f(x_1, x_2, ...)_i$ and $v_i = g(r_1, r_2, ...)_i$. Then, $\frac{\partial v_i}{\partial x_k} = \sum_j \frac{\partial v_i}{\partial r_j} \frac{\partial r_j}{\partial x_k}$. Compute the derivative of the following nested functions. As an additional exercise, write down the vectorized solutions.

$$\frac{\partial E}{\partial u_l} = \frac{\partial}{\partial u_l} \frac{1}{2} \sum_n (\sigma(u_n) - y_n)^2 =$$

$$\frac{\partial E}{\partial w_{il}} = \frac{\partial}{\partial w_{il}} \frac{1}{2} \sum_j (\sigma(\Big(\sum_k w_{jk} x_k\Big) + b_j) - y_j)^2 =$$

## ① SCALAR CASE

- $\frac{\partial}{\partial b}(Wx+b)=1$
- $\frac{\partial}{\partial W}(Wx+b)=x$
- $\frac{\partial}{\partial x}(Wx+b)=W$

## VECTORIAL CASE

- $\frac{\partial}{\partial b_i}\left(\left(\sum_k W_{jk}x_k\right)+b_j\right)=\begin{cases}1 & i=j\\ 0 & \text{OTHERWISE}\end{cases}=\delta_{ij}$
- $\frac{\partial}{\partial W}W\vec{x}+\vec{b}=I\otimes\vec{x}$
- $\frac{\partial}{\partial W_{il}}\left(\left(\sum_k W_{jk}x_k\right)+b_j\right)=\begin{cases}x_l & i=j\\ 0 & \text{OTHERWISE}\end{cases}=x_l\,\delta_{ij}$
- $\frac{\partial}{\partial x_l}\left(\left(\sum_k W_{jk}x_k\right)+b_j\right)=W_{jl}$

## BATCHED VECTORIZED CASE

- $\frac{\partial}{\partial \vec{b}}(XW^T+1\otimes b)=1\otimes I$
- $\frac{\partial}{\partial b_i}U_{nj}=\frac{\partial}{\partial b_i}\left(\sum_k W_{jk}X_{nk}+b_j\right)=\begin{cases}1 & i=j\\ 0 & \text{OTHERWISE}\end{cases}=\delta_{ij}$
- $\frac{\partial}{\partial W}(XW^T+1\otimes b)=\frac{\partial}{\partial W}WX^T=I\otimes W$
- $\frac{\partial}{\partial W_{il}}\left(\sum_k W_{jk}X_{nk}+b_j\right)=\begin{cases}X_{nl} & i=j\\ 0 & \text{OTHERWISE}\end{cases}=X_{nl}\,\delta_{ij}$

- $\frac{\partial}{\partial X}(XW^T + 1 \otimes b) = I \otimes W^T$

- $\frac{\partial}{\partial X_{me}}\left(\sum_k W_{jk} X_{nk} + b_j\right) = \begin{cases} X_{ne} & i=j \\ 0 & \text{OTHERWISE} \end{cases} = X_{ne}\delta_{ij}$

②

- $\frac{\partial}{\partial Z_{me}}\left(\frac{1}{2}\sum_n \sum_m (Z_{nk}-Y_{nk})^2\right) = \frac{1}{2}\frac{\partial}{\partial Z_{me}}(Z_{me}-Y_{me})^2 =$

$= 2 \cdot \frac{1}{2}(Z_{me}-Y_{me}) \cdot 1 = Z_{me}-Y_{me}$

- $\frac{\partial}{\partial v}\frac{1}{1+e^{-v}} = \frac{\partial}{\partial v}(1+e^{-v})^{-1} = -1(1+e^{-v})^{-2} \cdot (-1) \cdot e^{-v} = \frac{e^{-v}}{(1+e^{-v})^2} =$

$= \frac{1}{1+e^{-v}} \cdot \frac{e^{-v}}{1+e^{-v}} = \sigma(v) \cdot \left(\frac{1+e^{-v}-1}{1+e^{-v}}\right) = \sigma(v)(1-\sigma(v))$

- $\frac{\partial}{\partial v_j}\left(e^{v_i}\left(\sum_k e^{v_k}\right)^{-1}\right) = \frac{\partial}{\partial v_j} e^{v_i} \cdot \left(\sum_k e^{v_k}\right)^{-1} + e^{v_i} \cdot \frac{\partial}{\partial v_j}\left(\sum_k e^{v_k}\right)^{-1} =$

$= \delta_{ij} s_j + e^{v_i}\left(-e^{v_j}\left(\sum_k e^{v_k}\right)^{-2}\right) = \delta_{ij} s_j - s_i s_j$

③

- $\frac{\partial}{\partial v_e}\frac{1}{2}\sum_n (\sigma(v_n)-Y_n)^2 = (\sigma(v_e)-Y_e) \cdot \sigma'(v_e) =$

$= (\sigma(v_e)-Y_e) \cdot \frac{\partial}{\partial v_e}\left(\frac{1}{1+e^{-v}}\right) = (\sigma(v_e)-Y_e) \cdot (-(1+e^{-v})^{-2}(-e^{-v}))$

$= (\sigma(v_e)-Y_e)\frac{e^{-v_e}}{(1+e^{-v_e})^2} = (\sigma(v_e)-Y_e)\sigma(v_e)(1-\sigma(v_e))$

- $\frac{\partial}{\partial W_{ie}}\frac{1}{2}\sum_j (\sigma(\sum_k W_{jk}X_k)+b_j)-Y_j)^2 = \frac{\partial}{\partial W_{ie}}\frac{1}{2}(\sigma(W_{ie}X_e+b_i)-$

$-Y_i)^2 = (\sigma(W_{ie}X_e+b_i)-Y_i)\sigma'(W_{ie}X_e+b_i)X_e =$

$= (\sigma(W_{ie}X_e+b_i)-Y_i)\sigma(W_{ie}X_e+b_i)(1-\sigma(W_{ie}X_e+b_i))X_e$