# Introduction to Machine Learning (CSCI-UA.473): Homework 4

Instructor: Sumit Chopra

## Theory

### Question T1: Back propagation of a 2D Convolution Operation ( 15 points)

Let the input be an 2D gray scale image of size $m \times n$, denoted by the matrix $X \in \mathbb{R}^{m \times n}$. Let the parameters of the $p \times p$ convolution kernel be denoted by $[W, b]$, where $W \in \mathbb{R}^{p \times p}$ are the weights of the kernel and $b$ is the bias associated with the kernel. Let us denote by $L$ the loss function of your model and by $\delta$ the gradient of the loss with respect to the output of the convolution operation. Write the expression for the following:

**1. (5 points) Gradient of the loss function $L$ with respect to the inputs $X$ : $\frac{dL}{dX}$**

**Answer:**

$$y = W \cdot X + b$$
$$L = \frac{1}{N} \sum_{j=1}^{N} (y_j - t_j)^2$$

By Chain Rule, we have:

$$\frac{\partial y}{\partial X} = W$$
$$\frac{\partial L}{\partial y} = \frac{1}{N} \sum_{j=1}^{N} \frac{\partial (y_j - t_j)^2}{\partial y_j} = \frac{2}{N} \sum_{j=1}^{N} (y_j - t_j)$$
$$\frac{\partial L}{\partial X} = \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial X} = \frac{2}{N} \left[ \sum_{j=1}^{N} (y_j - t_j) \right] W$$
$$= \frac{2}{N} \left[ \sum_{j=1}^{N} (W_i \cdot x_i + b_i - t_j) \right] W$$

**2. (5 points) Gradient of the loss function $L$ with respect to the weights $W$ : $\frac{dL}{dW}$**

**Answer:**

$$y = W \cdot X + b$$
$$L = \frac{1}{N} \sum_{j=1}^{N} (y_j - t_j)^2$$

By Chain Rule, we have:

$$\begin{aligned}
\frac{\partial y}{\partial W} &= X^T \\
\frac{\partial L}{\partial y} &= \frac{1}{N} \sum_{j=1}^{N} \frac{\partial (y_j - t_j)^2}{\partial y_j} = \frac{2}{N} \sum_{j=1}^{N} (y_j - t_j) \\
\frac{\partial L}{\partial W} &= \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial W} = \frac{2}{N} \sum_{j=1}^{N} (y_j - t_j) x_j \\
&= \frac{2}{N} \sum_{j=1}^{N} (W_i \cdot x_i + b_i - t_j) x_j
\end{aligned}$$

**3. (5 points) Gradient of the loss function $L$ with respect to the bias $b$ : $\frac{dL}{db}$**

**Answer:**

$$\begin{aligned}
\frac{\partial y}{\partial b} &= 1 \\
\frac{\partial L}{\partial y} &= \frac{1}{N} \sum_{j=1}^{N} \frac{\partial (y_j - t_j)^2}{\partial y_j} = \frac{2}{N} \sum_{j=1}^{N} (y_j - t_j) \\
\frac{\partial L}{\partial b} &= \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial b} = \frac{2}{N} \sum_{j=1}^{N} (y_j - t_j) \\
&= \frac{2}{N} \sum_{j=1}^{N} (W_i \cdot x_i + b_i - t_j)
\end{aligned}$$

Please write all the steps that led you to the final expression. No points will be given if only the final expression is provided without the steps

## Question T2: Back propagation of other functions ( 15 points)

Compute the back propagation expression (the gradient of the loss function $L$ with respect to the input $x$, where $x \in \mathbb{R}^d$ is the 1D input vector of size $d$ ), for the following functions:

**1. (5 points) Tanh:** $f(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$

**Answer:**

$$f'(x) = \frac{\mathrm{d}f}{\mathrm{d}x} = \frac{\mathrm{d}\tanh(x)}{\mathrm{d}x} = 1 - \tanh^2(x)$$

General Case:

$$\begin{aligned}
a_k &= w_{kj} \cdot z_j + b_j \\
z_k &= f(a_k) = f(w_{kj} \cdot z_j + b_j) \\
L &\approx \frac{1}{2} \sum_{j=1}^{N} (a_j - y_j)^2
\end{aligned}$$

Derivation gives:

$$\begin{aligned}
\frac{\partial L_n}{\partial a_k} &= \delta_k \\
\frac{\partial a_k}{\partial w_{kj}} &= z_j \\
\frac{\partial a_k}{\partial z_j} &= w_{kj}
\end{aligned}$$

By Chain Rule, the gradient of the loss function $L$ is given by

$$\begin{aligned}
\delta_j &= \frac{\partial L}{\partial a_j} \\
&= \frac{\partial L}{\partial z_j} \cdot \frac{\partial z_j}{\partial a_j} \\
&= \left[ \sum_k \frac{\partial L}{\partial a_k} \frac{\partial a_k}{\partial z_j} \right] \cdot f'(a_j) \\
&= f'(a_j) \sum_k w_{kj} \delta_k \\
&= \left[1 - \tanh^2(a_j)\right] \sum_k w_{kj} \delta_k
\end{aligned}$$

$\forall j \in [0, d]$ such that

$$\nabla L_n = \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_d \end{bmatrix}$$

**2. (5 points) Max pooling:** $f(x) = \max_{i \in \{1,\ldots,d\}} x_i$

**Answer:** General Case:

$$\begin{aligned} a_k &= w_{kj} \cdot z_j + b_j \\ z_k &= f(a_k) = f(w_{kj} \cdot z_j + b_j) \\ L &\approx \frac{1}{2} \sum_{j=1}^{N} (a_j - y_j)^2 \end{aligned}$$

Derivation gives:

$$\begin{aligned} \frac{\partial L_n}{\partial a_k} &= \delta_k \\ \frac{\partial a_k}{\partial w_{kj}} &= z_j \\ \frac{\partial a_k}{\partial z_j} &= w_{kj} \end{aligned}$$

Assuming $z_k = a_k^* = \max_{i \in \{1,\ldots,d\}} a_i = f(a_k)$. By Chain Rule, the gradient of the loss function $L$ is given by

$$\delta_j = \frac{\partial L}{\partial a_j} = \begin{cases} 0 & \text{if } k \neq j \\ \sum_k w_{kj} \delta_k & \text{if } k = j \end{cases}$$

**3. (5 points) Average pooling:** $f(x) = \frac{1}{d} \sum_{i=1}^{d} x_i$

**Answer:** General Case:

$$\begin{aligned} a_k &= w_{kj} \cdot z_j + b_j \\ z_k &= f(a_k) = f(w_{kj} \cdot z_j + b_j) \\ L &\approx \frac{1}{2} \sum_{j=1}^{N} (a_j - y_j)^2 \end{aligned}$$

Derivation gives:

$$\frac{\partial L_n}{\partial a_k} = \delta_k$$
$$\frac{\partial a_k}{\partial w_{kj}} = z_j$$
$$\frac{\partial a_k}{\partial z_j} = w_{kj}$$

Since $f(x) = \frac{1}{d}\sum_{i=1}^{d} x_i$, then by Chain Rule, the gradient of the loss function $L$ is given by

$$\delta_j = \frac{\partial L}{\partial a_j} = \frac{1}{d}\sum_k w_{kj}\delta_k$$