Deep Learning 1 - Homework 1

Pedro M.P. Curvo
MSc Artificial Intelligence
University of Amsterdam
pedro.pombeiro.curvo@student.uva.nl

Question 1

a) b) c)

The linear model described previously can be written as:

$$Y = XW^T + B \tag{1}$$

Where Y is the output, X is the input, W is the weight matrix and B is the bias, with $Y \in \mathbb{R}^{S\times N}$, $X \in \mathbb{R}^{S\times M}$, $W \in \mathbb{R}^{N\times M}$ and $B \in \mathbb{R}^{1\times N}$.

In order to compute $\frac{\partial L}{\partial W}$, we can use the chain rule:

$$\begin{aligned}
\frac{\partial L}{\partial W} &= \frac{\partial L}{\partial Y}\frac{\partial Y}{\partial W} \\
&= \frac{\partial L}{\partial Y}\frac{\partial (XW^T + B)}{\partial W} \quad \text{(from equation 1)} \\
&= \frac{\partial L}{\partial Y}\frac{\partial (XW^T)}{\partial W} \quad \text{(since B does not depend on W)} \\
&= \frac{\partial L}{\partial Y}^T X
\end{aligned}$$

Similarly, to compute $\frac{\partial L}{\partial B}$, we can use the chain rule:

$$\begin{aligned}
\frac{\partial L}{\partial B} &= \frac{\partial L}{\partial Y}\frac{\partial Y}{\partial B} \\
&= \frac{\partial L}{\partial Y}\frac{\partial (XW^T + B)}{\partial B} \quad \text{(from equation 1)} \\
&= \frac{\partial L}{\partial Y}\frac{\partial B}{\partial B} \quad \text{(since B does not depend on W)} \\
&= \sum_{i=1}^{S} \frac{\partial L}{\partial Y_i}
\end{aligned}$$

Finally, to compute $\frac{\partial L}{\partial X}$, we can use the chain rule:

$$
\begin{aligned}
\frac{\partial L}{\partial X} &= \frac{\partial L}{\partial Y}\frac{\partial Y}{\partial X} \\
&= \frac{\partial L}{\partial Y}\frac{\partial (XW^T + B)}{\partial X} \quad \text{(from equation 1)} \\
&= \frac{\partial L}{\partial Y}\frac{\partial (XW^T)}{\partial X} \quad \text{(since B does not depend on W)} \\
&= \frac{\partial L}{\partial Y} W
\end{aligned}
$$

d)

Considering the element-wise activation h, given by:

$$Y = h(X) \Rightarrow Y_{ij} = h(X_{ij})$$

By applying the chain rule, we can compute $\frac{\partial L}{\partial X}$ as follows:

$$
\begin{aligned}
\frac{\partial L}{\partial X_{ij}} &= \frac{\partial L}{\partial Y_{ij}}\frac{\partial Y_{ij}}{\partial X_{ij}} \\
&= \frac{\partial L}{\partial Y_{ij}}\frac{\partial h(X_{ij})}{\partial X_{ij}}
\end{aligned}
$$

Which can have a simple notation of:

$$\frac{\partial L}{\partial X_{ij}} = \frac{\partial L}{\partial Y_{ij}} h'(X_{ij})$$

where $h'(X)$ is the derivative of the activation function h with respect to their input.

Now, this rule is applied to all elements of the matrix X, resulting in the following:

$$\frac{\partial L}{\partial X} = \frac{\partial L}{\partial Y} \odot h'(X)$$

where $\odot$ is the Hadamard product and h' is applied element-wise to the matrix X.

As we can see, since we are applying the Hadamard product on a derivative that is also applied element-wise, the result will have the same dimensions as the input matrix X. Hence, $\frac{\partial L}{\partial X} \in \mathbb{R}^{S \times M}$.

e)

As presented, the gradients can be given by:

$$
\begin{aligned}
\frac{\partial L}{\partial Z} &= Y \odot \left(\frac{\partial L}{\partial Y} - \left(\frac{\partial L}{\partial Y} \odot Y\right) 11^T\right) \\
\frac{\partial L}{\partial Y} &= -\frac{1}{S}\left(\frac{T}{Y}\right)
\end{aligned}
$$

First, we began by replacing the expression for $\frac{\partial L}{\partial Y}$ in the expression for $\frac{\partial L}{\partial Z}$:

$$\frac{\partial L}{\partial Z} = Y \odot \left(-\frac{1}{S} \left(\frac{T}{Y} \right) - \left(-\frac{1}{S} \left(\frac{T}{Y} \right) \odot Y \right) 11^T \right)$$

Now, since $-\frac{1}{S}$ is a scalar, we can take it out of the Hadamard product:

$$\frac{\partial L}{\partial Z} = -\frac{1}{S} Y \odot \left(\frac{T}{Y} - \left(\frac{T}{Y} \odot Y \right) 11^T \right)$$

Now, we can use the distributive property of the Hadamard product to simplify the expression:

$$\frac{\partial L}{\partial Z} = -\frac{1}{S} \left(Y \odot \frac{T}{Y} - Y \odot \left(\left(\frac{T}{Y} \odot Y \right) 11^T \right) \right)$$

Now, since the division is element-wise, we can cancel out the Hadamard product with the division:

$$\frac{\partial L}{\partial Z} = -\frac{1}{S} \left(T - Y \odot \left((T) \, 11^T \right) \right)$$

Now, we need to look at the matrix product of $T11^T$. 11^T gives us a matrix of ones, with dimensions $C \times C$. Naming the result as H, we have that:

$$H_{ij} = \sum_{k=1}^{C} T_{ik} 1 = \sum_{k=1}^{C} T_{ik}$$

But since T is a one-hot encoded matrix, we have that $\sum_j T_{ij} = 1$. Hence, $H_{ij} = 1$ for all i and j.

Now, we can simplify the expression for $\frac{\partial L}{\partial Z}$:

$$\begin{aligned} \frac{\partial L}{\partial Z} &= -\frac{1}{S} \left(T - Y \odot H \right) \\ &= -\frac{1}{S} \left(T - Y \right) \\ &= \frac{1}{S} \left(Y - T \right) \end{aligned}$$

With that, we can infer that:

$$\begin{aligned} \alpha &= \frac{1}{S} \\ M &= Y - T \end{aligned}$$

Since S is the number of samples, than $\alpha \in \mathbb{R}^+$ as we wanted to show.