Sapienza
Università di Roma

Advanced Machine Learning - Assignment #1

Group Members
Stefano D'Arrigo 1960500
Paola Antonicoli 1796554
Simone Fiorellino 1960415
Jeremy Sapienza 1960498

Oct 27 2021

1 Preliminary notions

Before proceeding and answering to the questions, we present some of the rules and derivations which will be exploited in the next sections. Let $W \in \mathbb{R}^{n \times m}$ and $A \in \mathbb{R}^{m \times l}$ be two real-value matrices. Recall that $Z = W \cdot A$ is defined as:

$$z_{i,j} = \mathbf{w}_{i,:} \cdot \mathbf{a}_{:,j} = \sum_{k=1}^{m} w_{i,k} \cdot a_{k,j} \tag{1}$$

The derivative of $z_{i,j}$ - which is a scalar - with respect to the vector $\mathbf{w}_{i,:}$ is defined as:

$$\begin{aligned}
&\frac{\partial z_{i,j}}{\partial \mathbf{w}_{i,:}} = \frac{\partial z_{i,j}}{\partial (w_{i,1}, w_{i,2}, ..., w_{i,m})} = \frac{\partial \mathbf{w}_{i,:} \cdot \mathbf{a}_{:,j}}{\partial (w_{i,1}, w_{i,2}, ..., w_{i,m})} \\
&= \frac{\partial (w_{i,1}a_{1,j} + w_{i,2}a_{2,j} + ... + w_{i,m}a_{m,j})}{\partial (w_{i,1}, w_{i,1}, ..., w_{i,m})} \\
&= \frac{\partial \left(\sum_{k=1}^{m} w_{i,k} \cdot a_{k,j}\right)}{\partial (w_{i,1}, w_{i,2}, ..., w_{i,m})} \\
&= \left[\frac{\partial \left(\sum_{k=1}^{m} w_{i,k} \cdot a_{k,j}\right)}{\partial w_{i,1}}, \frac{\partial \left(\sum_{k=1}^{m} w_{i,k} \cdot a_{k,j}\right)}{\partial w_{i,2}}, ..., \frac{\partial \left(\sum_{k=1}^{m} w_{i,k} \cdot a_{k,j}\right)}{\partial w_{i,m}}\right] \\
&= [a_{1,j}, a_{2,j}, ..., a_{m,j}] = \mathbf{a}_{:,j}^T
\end{aligned} \tag{2}$$

The same holds for $\frac{\partial z_{i,j}}{\partial \mathbf{a}_{:j}}$.
Moving forward, the derivative of $\mathbf{z}_{i,:}$ - which is a vector - with respect to the vector $\mathbf{w}_{i,:}$ is defined as:

$$\begin{aligned}
&\frac{\partial \mathbf{z}_{i,:}}{\partial \mathbf{w}_{i,:}} = \frac{\partial \mathbf{z}_{i,:}}{\partial (w_{i,1}, w_{i,2}, ..., w_{i,m})} \\
&= \left[\frac{\partial z_{i,1}}{\partial (w_{i,1}, w_{i,2}, ..., w_{i,m})}, \frac{\partial z_{i,2}}{\partial (w_{i,1}, w_{i,2}, ..., w_{i,m})}, ..., \frac{\partial z_{i,m}}{\partial (w_{i,1}, w_{i,2}, ..., w_{i,m})}\right] \\
&= [[a_{1,1}, a_{2,1}, ..., a_{m,1}], ..., [a_{1,l}, a_{2,l}, ..., a_{m,l}]] \\
&= \left[\mathbf{a}_{:,1}^T, ..., \mathbf{a}_{:,l}^T\right] = A^T
\end{aligned} \tag{3}$$

We can, then, derive a row $\mathbf{z}_{i,:} \in Z$ w.r.t the entire matrix W. Let's write down the product $W \cdot A$ explicitly:

$$Z = W \cdot A = \begin{pmatrix} w_{1,1} & w_{1,2} & ... & w_{1,m} \\ w_{2,1} & w_{2,2} & ... & w_{2,m} \\ ... & ... & ... & ... \\ w_{n,1} & w_{n,2} & ... & w_{n,m} \end{pmatrix} \cdot \begin{pmatrix} a_{1,1} & a_{1,2} & ... & a_{1,l} \\ a_{2,1} & a_{2,2} & ... & a_{2,l} \\ ... & ... & ... & ... \\ a_{m,1} & a_{m,2} & ... & a_{m,l} \end{pmatrix} \tag{4}$$

$$\mathbf{z}_{i,:} = \mathbf{w}_{i,:} \cdot A = \begin{pmatrix} w_{i,1} & w_{i,2} & ... & w_{i,m} \end{pmatrix} \cdot \begin{pmatrix} a_{1,1} & a_{1,2} & ... & a_{1,l} \\ a_{2,1} & a_{2,2} & ... & a_{2,l} \\ ... & ... & ... & ... \\ a_{m,1} & a_{m,2} & ... & a_{m,l} \end{pmatrix} \tag{5}$$

From Eq.(5), it's easy to conclude that $\frac{\partial \mathbf{z}_{i,:}}{\partial W}$ is:

$$\frac{\partial \mathbf{z}_{i,:}}{\partial W} = \frac{\partial \mathbf{z}_{i,:}}{\partial(\mathbf{w}_{1,:}, \mathbf{w}_{2,:}, ..., \mathbf{w}_{n,:})} = \left[O, O, ..., \frac{\partial \mathbf{z}_{i,:}}{\partial \mathbf{w}_{i,:}}, ..., O\right] = \left[O, O, ..., A^T, ..., O\right] \in \mathbb{R}^{(n\times m)\times l} \tag{6}$$

being $O \in \mathbb{R}^{l\times m}$ the zero matrix.

Finally, putting all together, the derivative of the matrix Z w.r.t. the matrix W, i.e. the jacobian, is:

$$\frac{\partial Z}{\partial W} = \left[\frac{\partial \mathbf{z}_{1,:}}{\partial W}, \frac{\partial \mathbf{z}_{2,:}}{\partial W}, ..., \frac{\partial \mathbf{z}_{n,:}}{\partial W}\right] = \left[\left[A^T, O, ..., O\right], \left[O, A^T, ..., O\right], ..., \left[O, ..., A^T\right]\right] \in \mathbb{R}^{(n\times m)\times(m\times l)} \tag{7}$$

It can be noticed that the tensor $\frac{\partial Z}{\partial W}$ is very sparse.
Since it will be useful in the following, we conclude this section by deriving a column $\mathbf{z}_{:,j} \in Z$ w.r.t. the entire matrix W. Again, $\mathbf{z}_{:,j}$ is:

$$\mathbf{z}_{:,j} = W \cdot \mathbf{a}_{:,j} = \begin{pmatrix} w_{1,1} & w_{1,2} & ... & w_{1,m} \\ w_{2,1} & w_{2,2} & ... & w_{2,m} \\ ... & ... & ... & ... \\ w_{n,1} & w_{n,2} & ... & w_{n,m} \end{pmatrix} \cdot \begin{pmatrix} a_{1,j} \\ a_{2,j} \\ ... \\ a_{m,j} \end{pmatrix} \tag{8}$$

The derivative is:

$$\frac{\partial \mathbf{z}_{:,j}}{\partial W} = \frac{\partial \mathbf{z}_{:,j}}{\partial(\mathbf{w}_{1,:}, \mathbf{w}_{2,:}, ..., \mathbf{w}_{n,:})} = \begin{bmatrix} \mathbf{a}_{:,j}^T \\ \mathbf{a}_{:,j}^T \\ ... \\ \mathbf{a}_{:,j}^T \end{bmatrix} \in \mathbb{R}^{n\times l} \tag{9}$$

2 Question 2

$$\underset{(1,4)}{\mathbf{a}_i^{(1)}} = \mathbf{x}_i \tag{10}$$

$$\underset{(10,1)}{\mathbf{z}_i^{(2)}} = \underset{(10,4)}{W^{(1)}} \underset{(4,1)}{\left(\mathbf{a}_i^{(1)}\right)^T} + \underset{(1,)}{b^{(1)}} \tag{11}$$

$$\underset{(10,1)}{\mathbf{a}_i^{(2)}} = \phi(\underset{(10,1)}{\mathbf{z}_i^{(2)}}) \tag{12}$$

$$\underset{(3,1)}{\mathbf{z}_i^{(3)}} = \underset{(3,10)}{W^{(2)}} \underset{(10,1)}{\mathbf{a}_i^{(2)}} + \underset{(1,)}{b^{(2)}} \tag{13}$$

$$\underset{(3,1)}{\mathbf{a}_i^{(3)}} = \psi(\underset{(3,1)}{\mathbf{z}_i^{(3)}}) \tag{14}$$

2.1 a)

$$note\ that\ (\Delta_i)_j = \begin{cases} 1, & if\ j = y_i \\ 0, & otherwise \end{cases} \tag{15}$$

$$
\begin{aligned}
\frac{\partial J}{\partial z_i^{(3)}}\left(\theta,\ \{x_i, y_i\}_{i=1}^{N}\right) &= \frac{\partial}{\partial z_i^{(3)}}\ \frac{1}{N}\sum_{i=1}^{N} - \log\left(\frac{\exp\left(z_i^{(3)}\right)_{y_i}}{\sum_{j=1}^{K}\exp\left(z_i^{(3)}\right)_j}\right) = \\
&= \frac{1}{N}\frac{\partial}{\partial z_i^{(3)}}\ \left[- \log\left(\frac{\exp\left(z_i^{(3)}\right)_{y_i}}{\sum_{j=1}^{K}\exp\left(z_i^{(3)}\right)_j}\right)\right] = \\
&= \frac{1}{N}\frac{\partial}{\partial z_i^{(3)}}\left[\log\left(\sum_{j=1}^{K}\exp\left(z_i^{(3)}\right)_j\right) - \log\left(\exp\left(z_i^{(3)}\right)_{y_i}\right)\right] = \\
&= \frac{1}{N}\left\{\frac{\partial}{\partial z_i^{(3)}}\left[\log\left(\sum_{j=1}^{K}\exp\left(z_i^{(3)}\right)_j\right)\right] - \frac{\partial}{\partial z_i^{(3)}}\left[\log\left(\exp\left(z_i^{(3)}\right)_{y_i}\right)\right]\right\} = \\
&= \frac{1}{N}\left\{\frac{\partial}{\partial z_i^{(3)}}\left[\log\left(\sum_{j=1}^{K}\exp\left(z_i^{(3)}\right)_j\right)\right] - \frac{\partial}{\partial z_i^{(3)}}\left(z_i^{(3)}\right)_{y_i}\right\} = \\
&= \frac{1}{N}\left(\frac{\exp\left(z_i^{(3)}\right)}{\sum_{j=1}^{K}\exp\left(z_i^{(3)}\right)_j} - \Delta_i\right) = \frac{1}{N}\left(\psi\left(z_i^{(3)}\right) - \Delta_i\right)
\end{aligned}
$$

2.2 b)

$$
\begin{aligned}
&\frac{\partial J}{\partial W^{(2)}}\left(\theta, \{x_i, y_i\}_{i=1}^{N}\right) = \frac{\partial}{\partial W^{(2)}}\left\{\frac{1}{N}\sum_{i=1}^{N} - \log\left[\frac{\exp\left(z_i^{(3)}\right)_{y_i}}{\sum_{j=1}^{K}\left(\exp\left(z_i^{(3)}\right)\right)_j}\right]\right\} \\
&\overset{\star}{=} \frac{\partial z^{(3)}}{\partial W^{(2)}} \cdot \frac{\partial}{\partial z^{(3)}}\left\{\frac{1}{N}\sum_{i=1}^{N} - \log\left[\frac{\exp\left(z_i^{(3)}\right)_{y_i}}{\sum_{j=1}^{K}\left(\exp\left(z_i^{(3)}\right)\right)_j}\right]\right\} \\
&= \sum_{i=1}^{N}\frac{\partial z_i^{(3)}}{\partial W^{(2)}} \cdot \frac{\partial}{\partial z_i^{(3)}}\left\{-\frac{1}{N}\log\left[\frac{\exp\left(z_i^{(3)}\right)_{y_i}}{\sum_{j=1}^{K}\left(\exp\left(z_i^{(3)}\right)\right)_j}\right]\right\} \\
&\overset{\star\star}{=} \sum_{i=1}^{N}\frac{\partial z_i^{(3)}}{\partial W^{(2)}} \cdot \frac{\partial J}{\partial z_i^{(3)}} = \sum_{i=1}^{N}\frac{\partial}{\partial W^{(2)}}\left(W^{(2)}a_i^{(2)} + b^{(2)}\right) \cdot \frac{\partial J}{\partial z_i^{(3)}} = \sum_{i=1}^{N}\frac{1}{N}\left(\psi\left(z_i^{(3)}\right) - \Delta_i\right)\left(a_i^{(2)}\right)^T
\end{aligned}
\tag{16}
$$

having ($\star$) by the chain rule and ($\star\star$) by the Equation of point *a)*.

The derivative $\frac{\partial \widetilde{J}}{\partial W^{(2)}}$ is trivial, since:

$$\begin{aligned}
\frac{\partial \widetilde{J}}{\partial W^{(2)}}\left(\theta, \{x_i, y_i\}_{i=1}^{N}\right) &= \frac{\partial J}{\partial W^{(2)}} + \frac{\partial R}{\partial W^{(2)}} = \frac{\partial J}{\partial W^{(2)}} + \frac{\partial}{\partial W^{(2)}}\left[\lambda\Big(\left\|W^{(1)}\right\|_2^2 + \left\|W^{(2)}\right\|_2^2\Big)\right] \\
&= \frac{\partial J}{\partial W^{(2)}} + \frac{\partial}{\partial W^{(2)}}\left[\lambda\left(\sum_{i=1}^{3}\sum_{j=1}^{10}\left(w_{ij}^{(2)}\right)^2\right)\right] = \sum_{i=1}^{N}\frac{1}{N}\left(\psi\left(z_i^{(3)}\right) - \Delta_i\right)\left(a_i^{(2)}\right)^T + 2\lambda W^{(2)}
\end{aligned} \tag{17}$$

2.3 c)

$$\begin{aligned}
\frac{\partial \widetilde{J}}{\partial W^{(1)}}\left(\theta, \{x_i, y_i\}_{i=1}^{N}\right) &= \frac{\partial J}{\partial W^{(1)}} + \frac{\partial R}{\partial W^{(1)}} = \frac{\partial J}{\partial z^{(3)}} \cdot \frac{\partial z^{(3)}}{\partial W^{(1)}} + \frac{\partial R}{\partial W^{(1)}} \\
&= \frac{\partial J}{\partial z^{(3)}} \cdot \frac{\partial z^{(3)}}{\partial a^{(2)}} \cdot \frac{\partial a^{(2)}}{\partial W^{(1)}} + \frac{\partial R}{\partial W^{(1)}} = \frac{\partial J}{\partial z^{(3)}} \cdot \frac{\partial z^{(3)}}{\partial a^{(2)}} \cdot \frac{\partial a^{(2)}}{\partial z^{(1)}} \cdot \frac{\partial z^{(2)}}{\partial W^{(1)}} + \frac{\partial R}{\partial W^{(1)}} \\
&= \sum_{i=1}^{N}\left[\frac{1}{N}\left(\psi\left(z_i^{(3)}\right) - \Delta_i\right)\right] \cdot \frac{\partial z^{(3)}}{\partial a^{(2)}} \cdot \frac{\partial a^{(2)}}{\partial z^{(1)}} \cdot \frac{\partial z^{(2)}}{\partial W^{(1)}} + \frac{\partial R}{\partial W^{(1)}} \\
&= \sum_{i=1}^{N}\left[\frac{1}{N}\left(\psi\left(z_i^{(3)}\right) - \Delta_i\right)\right] \cdot \left[W^{(2)}\right]^T \cdot \frac{\partial a^{(2)}}{\partial z^{(1)}} \cdot \frac{\partial z^{(2)}}{\partial W^{(1)}} + \frac{\partial R}{\partial W^{(1)}} \\
&= \sum_{i=1}^{N}\left[\left(\frac{1}{N}\left(\psi\left(z_i^{(3)}\right) - \Delta_i\right)\right) \cdot \left(W^{(2)}\right)^T\right] \odot \mathbb{I}\left\{z_i^{(3)} > 0\right\} \cdot \frac{\partial z^{(2)}}{\partial W^{(1)}} + \frac{\partial R}{\partial W^{(1)}} \\
&= \sum_{i=1}^{N}\left(a_i^{(1)}\right)^T \cdot \left\{\left[\left(\frac{1}{N}\left(\psi\left(z_i^{(3)}\right) - \Delta_i\right)\right) \cdot \left(W^{(2)}\right)^T\right] \odot \mathbb{I}\left\{z_i^{(3)} > 0\right\}\right\} + \frac{\partial R}{\partial W^{(1)}} \\
&= \sum_{i=1}^{N}\left(a_i^{(1)}\right)^T \cdot \left\{\left[\left(\frac{1}{N}\left(\psi\left(z_i^{(3)}\right) - \Delta_i\right)\right) \cdot \left(W^{(2)}\right)^T\right] \odot \mathbb{I}\left\{z_i^{(3)} > 0\right\}\right\} + 2\lambda W^{(1)}
\end{aligned} \tag{18}$$

where $\odot$ indicates the element-wise product (Hadamard product).

$$\begin{aligned}
\frac{\partial \widetilde{J}}{\partial b^{(1)}}\left(\theta, \{x_i, y_i\}_{i=1}^{N}\right) &= \frac{\partial J}{\partial b^{(1)}} + \frac{\partial R}{\partial b^{(1)}} = \frac{\partial J}{\partial z^{(3)}} \cdot \frac{\partial z^{(3)}}{\partial b^{(1)}} + \frac{\partial R}{\partial b^{(1)}} \\
&= \frac{\partial J}{\partial z^{(3)}} \cdot \frac{\partial z^{(3)}}{\partial a^{(2)}} \cdot \frac{\partial a^{(2)}}{\partial b^{(1)}} + \frac{\partial R}{\partial b^{(1)}} = \frac{\partial J}{\partial z^{(3)}} \cdot \frac{\partial z^{(3)}}{\partial a^{(2)}} \cdot \frac{\partial a^{(2)}}{\partial z^{(2)}} \cdot \frac{\partial z^{(2)}}{\partial b^{(1)}} + \frac{\partial R}{\partial b^{(1)}} \\
&= \sum_{i=1}^{N}\left[\left(\frac{1}{N}\left(\psi\left(z_i^{(3)}\right) - \Delta_i\right)\right) \cdot \left(W^{(2)}\right)^T\right] \odot \mathbb{I}\left\{z_i^{(3)} > 0\right\} + \frac{\partial R}{\partial b^{(1)}} \\
&= \sum_{i=1}^{N}\left[\left(\frac{1}{N}\left(\psi\left(z_i^{(3)}\right) - \Delta_i\right)\right) \cdot \left(W^{(2)}\right)^T\right] \odot \mathbb{I}\left\{z_i^{(3)} > 0\right\}
\end{aligned} \tag{19}$$

$$
\begin{aligned}
\frac{\partial \widetilde{J}}{\partial b^{(2)}}\left(\theta, \{x_i, y_i\}_{i=1}^{N}\right) &= \frac{\partial J}{\partial b^{(2)}} + \frac{\partial R}{\partial b^{(2)}} = \frac{\partial J}{\partial z^{(3)}} \cdot \frac{\partial z^{(3)}}{\partial b^{(2)}} + \frac{\partial R}{\partial b^{(2)}} \\
&= \sum_{i=1}^{N} \frac{1}{N}\left(\psi\left(z_i^{(3)}\right) - \Delta_i\right) + \frac{\partial R}{\partial b^{(2)}} = \sum_{i=1}^{N} \frac{1}{N}\left(\psi\left(z_i^{(3)}\right) - \Delta_i\right)
\end{aligned}
\tag{20}
$$