Deep Learning Assignment 1

Jochem Loedeman
12995282

1 MLP backprop and NumPy Implementation

1.1 Evaluating the Gradients

Question 1.1

(a) Given that

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}^T + \boldsymbol{B},$$

we deduce that

$$Y_{mn} = \sum_p X_{mp} W_{np} + B_{mn}.$$

We will now find the required derivatives.

(i)

$$\frac{\partial L}{\partial W_{ij}} = \sum_{m,n} \frac{\partial L}{\partial Y_{mn}} \frac{\partial Y_{mn}}{\partial W_{ij}}.$$

But

$$\frac{\partial Y_{mn}}{\partial W_{ij}} = \sum_p X_{mp} \frac{\partial W_{np}}{\partial W_{ij}} = \sum_p X_{mp} \delta_{ni} \delta_{pj} = X_{mj} \delta_{ni},$$

and hence

$$\frac{\partial L}{\partial W_{ij}} = \sum_{m,n} \frac{\partial L}{\partial Y_{mn}} X_{mj} \delta_{ni} = \sum_m \frac{\partial L}{\partial Y_{mi}} X_{mj}.$$

In matrix-vector notation, this is equivalent to

$$\frac{\partial L}{\partial \boldsymbol{W}} = \left(\frac{\partial L}{\partial \boldsymbol{Y}} \right)^T \boldsymbol{X}$$

(ii)

$$\frac{\partial L}{\partial b_i} = \sum_{m,n} \frac{\partial L}{\partial Y_{mn}} \frac{\partial Y_{mn}}{\partial b_i}$$

But

$$\frac{\partial Y_{mn}}{\partial b_i} = \frac{\partial B_{mn}}{\partial b_i} = \delta_{ni},$$

Since $B_{mn} = b_n$ for all $m = 1, \ldots, S$. Therefore,

$$\frac{\partial L}{\partial b_i} = \sum_{m,n} \frac{\partial L}{\partial Y_{mn}} \delta_{ni} = \sum_m \frac{\partial L}{\partial Y_{mi}}.$$

In matrix-vector notation, this is equivalent to

$$\frac{\partial L}{\partial \boldsymbol{b}} = \mathbf{1} \frac{\partial L}{\partial \boldsymbol{Y}}$$

where $\mathbf{1}$ is the $1 \times S$ ones-vector.

(iii)

$$\frac{\partial L}{\partial X_{ij}} = \sum_{m,n} \frac{\partial L}{\partial Y_{mn}} \frac{\partial Y_{mn}}{\partial X_{ij}}.$$

But

$$\frac{\partial Y_{mn}}{\partial X_{ij}} = \sum_p \frac{\partial X_{mp}}{\partial X_{ij}} W_{np} = \sum_p \delta_{mi}\delta_{pj} W_{np} = \delta_{mi} W_{nj},$$

so

$$\frac{\partial L}{\partial X_{ij}} = \sum_{m,n} \frac{\partial L}{\partial Y_{mn}} \delta_{mi} W_{nj} = \sum_n \frac{\partial L}{\partial Y_{in}} W_{nj}.$$

In matrix-vector notation, this is equivalent to

$$\frac{\partial L}{\partial \boldsymbol{X}} = \frac{\partial L}{\partial \boldsymbol{Y}} \boldsymbol{W}$$

(b) Like before, we have

$$\frac{\partial L}{\partial X_{ij}} = \sum_{m,n} \frac{\partial L}{\partial Y_{mn}} \frac{\partial Y_{mn}}{\partial X_{ij}}$$

But

$$\frac{\partial Y_{mn}}{\partial X_{ij}} = \frac{\partial h(X_{mn})}{\partial X_{ij}} = h'(X_{mn})\delta_{mi}\delta_{nj},$$

and therefore,

$$\frac{\partial L}{\partial X_{ij}} = \sum_{m,n} \frac{\partial L}{\partial Y_{mn}} h'(X_{mn})\delta_{mi}\delta_{nj} = \frac{\partial L}{\partial Y_{ij}} h'(X_{ij}).$$

Or, in matrix-vector notation:

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

(c) (i)

$$\frac{\partial L}{\partial X_{ij}} = \sum_{m,n} \frac{\partial L}{\partial Y_{mn}} \frac{\partial Y_{mn}}{\partial X_{ij}}.$$

The derivative of the softmax is

$$\begin{aligned}
\frac{\partial Y_{mn}}{\partial X_{ij}} &= \frac{\partial}{\partial X_{ij}} \frac{e^{X_{mn}}}{\sum_k e^{X_{mk}}} \\
&= \frac{\sum_k e^{X_{mk}} \cdot e^{X_{mn}} \delta_{mi}\delta_{nj} - e^{X_{mn}} \cdot \sum_k e^{X_{mk}} \delta_{mi}\delta_{kj}}{\left(\sum_k e^{X_{mk}}\right)^2} \\
&= \frac{\sum_k e^{X_{mk}} \cdot e^{X_{mn}} \delta_{mi}\delta_{nj} - e^{X_{mn}} \cdot e^{X_{mj}} \delta_{mi}}{\left(\sum_k e^{X_{mk}}\right)^2} \\
&= Y_{mn}\delta_{mi}\delta_{nj} - Y_{mn}Y_{mj}\delta_{mi} \\
&= Y_{mn}\delta_{mi}\left(\delta_{nj} - Y_{mj}\right)
\end{aligned}$$

Notice that when $m = i$ (i.e. when the same data point is considered), the derivative reduces to the usual softmax derivative for rank-1 tensors.
We get

$$\begin{aligned}
\frac{\partial L}{\partial X_{ij}} &= \sum_{m,n} \frac{\partial L}{\partial Y_{mn}} Y_{mn}\delta_{mi}\left(\delta_{nj} - Y_{mj}\right) \\
&= \sum_n \frac{\partial L}{\partial Y_{in}} Y_{in}\left(\delta_{nj} - Y_{ij}\right)
\end{aligned}$$

(ii)

$$\begin{aligned}
\frac{\partial L}{\partial X_{ij}} &= -\frac{1}{S}\sum_{m,n} T_{mn}\frac{\partial \log X_{mn}}{\partial X_{ij}} \\
&= -\frac{1}{S}\sum_{m,n} \frac{T_{mn}}{X_{mn}}\frac{\partial X_{mn}}{\partial X_{ij}} \\
&= -\frac{1}{S}\sum_{m,n} \frac{T_{mn}}{X_{mn}}\delta_{mi}\delta_{nj} \\
&= -\frac{1}{S}\frac{T_{ij}}{X_{ij}}
\end{aligned}$$

The equation can be vectorized by using the elementwise division operator, known as the Hadamard division.

$$\frac{\partial L}{\partial \boldsymbol{X}} = -\frac{1}{S}\,\boldsymbol{T} \oslash \boldsymbol{X}$$