First assignment in Deep learning 1 – 2023 – Paper 1

1 Question 1 - linear module (Adam Divak November)

(*a*) $\frac{\partial L}{\partial \mathbf{W}}$

Answer:

$$\begin{aligned} L &\in \mathbb{R} \\ \mathbf{X} &\in \mathbb{R}^{S\times M} \\ \mathbf{W} &\in \mathbb{R}^{N\times M} \\ \mathbf{B} &\in \mathbb{R}^{S\times N} \\ \mathbf{Y} &\in \mathbb{R}^{S\times N} \\ \mathbf{Y} &= \mathbf{X}\mathbf{W}^T + \mathbf{B} \end{aligned}$$

Apply chain rule to the loss to break the derivative into two parts. Use index notation to do the derivation for one element of the weight matrix:

$$\frac{\partial L(\mathbf{Y})}{\partial W_{kl}} = \sum_{i,j} \frac{\partial L(\mathbf{Y})}{\partial Y_{ij}} \frac{\partial Y_{ij}}{\partial W_{kl}} \tag{1}$$

First calculate the second part independently:

$$\begin{aligned} \frac{\partial Y_{ij}}{\partial W_{kl}} &= \frac{\partial}{\partial W_{kl}} (\mathbf{X}\mathbf{W}^T + \mathbf{B})_{ij} \\ &= \frac{\partial}{\partial W_{kl}} (\sum_p X_{ip} W_{jp} + B_{ij}) \\ &= \sum_p \frac{\partial}{\partial W_{kl}} X_{ip} W_{jp} \\ &= \sum_p X_{ip} \frac{\partial}{\partial W_{kl}} W_{jp} \\ &= \sum_p X_{ip} \delta_{kj} \delta_{lp} \\ &= X_{il} \delta_{kj} \end{aligned} \tag{2}$$

Substitute 2 into 1:

$$\begin{aligned} \frac{\partial L(\mathbf{Y})}{\partial W_{kl}} &= \sum_{i,j} \frac{\partial L(\mathbf{Y})}{\partial Y_{ij}} \frac{\partial Y_{ij}}{\partial W_{kl}} \\ &= \sum_{i,j} \frac{\partial L(\mathbf{Y})}{\partial Y_{ij}} X_{il} \delta_{kj} \\ &= \sum_{i} \frac{\partial L(\mathbf{Y})}{\partial Y_{ik}} X_{il} \end{aligned}$$

Turning into vector form:

$$\frac{\partial L(\mathbf{Y})}{\partial \mathbf{W}} = (\frac{\partial L(\mathbf{Y})}{\partial \mathbf{Y}})^T \mathbf{X} \tag{3}$$

Checking the shapes we get:

$$(\mathbb{R}^{S\times N})^T \times \mathbb{R}^{S\times M} \implies \mathbb{R}^{N\times S} \times \mathbb{R}^{S\times M} \implies \mathbb{R}^{N\times M} \tag{4}$$

which is what we expected for $\mathbf{W}$.

(*b*) $\frac{\partial L(\mathbf{Y})}{\partial \mathbf{b}}$

Answer: Apply chain rule to the loss to break the derivative into two parts. Use index notation to do the derivation for one element of the bias vector:

$$\frac{\partial L(\mathbf{Y})}{\partial b_k} = \sum_{i,j} \frac{\partial L(\mathbf{Y})}{\partial Y_{ij}} \frac{\partial Y_{ij}}{\partial b_k} \tag{5}$$

First calculate the second part independently:

$$\begin{aligned} \frac{\partial Y_{ij}}{\partial b_k} &= \frac{\partial}{\partial b_k}(\mathbf{X}\mathbf{W}^T + \mathbf{B})_{ij} \\ &= \frac{\partial}{\partial b_k}(\sum_p X_{ip}W_{jp} + B_{ij}) \\ &= \frac{\partial}{\partial b_k} B_{ij} \\ &\stackrel{1}{=} \frac{\partial}{\partial \mathbf{B}_{ik}} B_{ij} \\ &= \delta_{kj} \end{aligned} \tag{6}$$

Where in 1 we used the fact that the bias matrix is made up of identical rows of the bias vector, so $B_{ij} = b_j \forall i \in \{1 \dots S\}$

Substitute 6 into 5:

$$\begin{aligned} \frac{\partial L(\mathbf{Y})}{\partial b_k} &= \sum_{i,j} \frac{\partial L(\mathbf{Y})}{\partial Y_{ij}} \frac{\partial Y_{ij}}{\partial b_k} \\ &= \sum_{i,j} \frac{\partial L(\mathbf{Y})}{\partial Y_{ij}} \delta_{kj} \\ &= \sum_{i} \frac{\partial L(\mathbf{Y})}{\partial Y_{ik}} \end{aligned}$$

Turning into vector form:

$$\frac{\partial L(\mathbf{Y})}{\partial \mathbf{b}} = \frac{\partial L(\mathbf{Y})}{\partial \mathbf{Y}} \tag{7}$$

So the derivative of the bias is exactly the derivative of the error.

Checking the shapes we get:

$$\mathbb{R}^{S \times N} \tag{8}$$

which is what we expected for $\mathbf{B}$.

(*c*) $\frac{\partial L}{\partial \mathbf{X}}$

Answer:

Apply chain rule to the loss to break the derivative into two parts. Use index notation to do the derivation for one element of the input matrix:

$$\frac{\partial L(\mathbf{Y})}{\partial X_{kl}} = \sum_{i,j} \frac{\partial L(\mathbf{Y})}{\partial Y_{ij}} \frac{\partial Y_{ij}}{\partial X_{kl}} \tag{9}$$

First calculate the second part independently:

$$
\begin{aligned}
\frac{\partial Y_{ij}}{\partial X_{kl}} &= \frac{\partial}{\partial X_{kl}}(\mathbf{X}\mathbf{W}^T + \mathbf{B})_{ij} \\
&= \frac{\partial}{\partial X_{kl}}(\sum_p X_{ip}W_{jp} + B_{ij}) \\
&= \sum_p \frac{\partial}{\partial X_{kl}} X_{ip}W_{jp} \\
&= \sum_p W_{jp} \frac{\partial}{\partial X_{kl}} X_{ip} \\
&= \sum_p W_{jp}\delta_{ki}\delta_{lp} \\
&= W_{jl}\delta_{ki}
\end{aligned}
\tag{10}
$$

Substitute 10 into 9:

$$
\begin{aligned}
\frac{\partial L(\mathbf{Y})}{\partial W_{kl}} &= \sum_{i,j} \frac{\partial L(\mathbf{Y})}{\partial Y_{ij}} \frac{\partial Y_{ij}}{\partial X_{kl}} \\
&= \sum_{i,j} \frac{\partial L(\mathbf{Y})}{\partial Y_{ij}} W_{jl}\delta_{ki} \\
&= \sum_{j} \frac{\partial L(\mathbf{Y})}{\partial Y_{kj}} W_{jl}
\end{aligned}
$$

Turning into vector form:

$$
\frac{\partial L(\mathbf{Y})}{\partial \mathbf{X}} = \frac{\partial L(\mathbf{Y})}{\partial \mathbf{Y}}\mathbf{W} \tag{11}
$$

Checking the shapes we get:

$$
\mathbb{R}^{S\times N} \times \mathbb{R}^{N\times M} \implies \mathbb{R}^{S\times M} \tag{12}
$$

which is what we expected for $\mathbf{X}$.

(d) $\frac{\partial L}{\partial \mathbf{X}}$ for the activation function

---

Answer:

Apply chain rule to the loss to break the derivative into two parts. Use index notation to do the derivation for one element of the input matrix:

$$
\frac{\partial L(\mathbf{Y})}{\partial X_{kl}} = \sum_{i,j} \frac{\partial L(\mathbf{Y})}{\partial Y_{ij}} \frac{\partial Y_{ij}}{\partial X_{kl}} \tag{13}
$$

First calculate the second part independently. Given that the activation function is defined as being element-wise, we can simply use the univariate chain rule:

$$
\begin{aligned}
\frac{\partial Y_{ij}}{\partial X_{kl}} &= \frac{\partial h(X_{ij})}{\partial X_{kl}} \\
&= \frac{\partial h(X_{ij})}{\partial X_{ij}} \frac{\partial X_{ij}}{\partial X_{kl}} \\
&= \frac{\partial h(X_{ij})}{\partial X_{ij}} \delta_{ik}\delta_{jl}
\end{aligned}
\tag{14}
$$

Substitute 14 into 13:

$$
\begin{aligned}
\frac{\partial L(\mathbf{Y})}{\partial X_{kl}} &= \sum_{i,j} \frac{\partial L(\mathbf{Y})}{\partial Y_{ij}} \frac{\partial Y_{ij}}{\partial X_{kl}} \\
&= \sum_{i,j} \frac{\partial L(\mathbf{Y})}{\partial Y_{ij}} \frac{\partial h(X_{ij})}{\partial X_{ij}} \delta_{ik} \delta_{jl} \\
&= \frac{\partial L(\mathbf{Y})}{\partial Y_{kl}} \frac{\partial h(X_{kl})}{\partial X_{kl}}
\end{aligned}
$$

Turning into vector form, using the Hadamard (element-wise) product of two matrices:

$$
\frac{\partial L(\mathbf{Y})}{\partial \mathbf{X}} = \frac{\partial L(\mathbf{Y})}{\partial \mathbf{Y}} \circ \frac{\partial h(\mathbf{X})}{\partial \mathbf{X}} \tag{15}
$$

Checking the shapes we get:

$$
\mathbb{R}^{S \times N} \circ \mathbb{R}^{S \times M} \implies \mathbb{R}^{S \times N} \tag{16}
$$

which is true because $M = N$ in this case, and this is what we expected as a result.