Assignment 1. MLPs, CNNs and Backpropagation

Maurice Frank
11650656
maurice.frank@posteo.de

1 MLP backprop

1.1

1.1.a

$$\begin{aligned}
\frac{\partial \boldsymbol{L}}{\partial x_i^{(N)}} &= -\frac{\partial}{\partial x_i^{(N)}} \sum_i t_i \log x_i^{(N)} \\
&= -t_i \cdot \frac{1}{x_i^{(N)}} \\
&\Leftrightarrow \\
\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{x}^{(N)}} &= -[\dots \frac{t_i}{x_i^{(N)}} \dots] \\
&\in \mathbb{R}^{d_N}
\end{aligned} \tag{$\frac{\partial \boldsymbol{x}^{(N)}}{\partial \boldsymbol{L}}$}$$

$$\begin{aligned}
\frac{\partial x_i^{(N)}}{\partial \tilde{x}_j^{(N)}} &= \frac{\partial}{\partial \tilde{x}_j^{(N)}} \frac{\exp \tilde{x}_i^{(N)}}{\sum_k \exp \tilde{\boldsymbol{x}}_k^{(N)}} \\
&= \frac{(\frac{\partial}{\partial \tilde{x}_j^{(N)}} \exp \tilde{x}_i^{(N)}) \cdot \sum_k \exp \tilde{x}_k^{(N)} - \exp \tilde{x}_i^{(N)} \cdot \frac{\partial}{\partial \tilde{x}_j^{(N)}} \sum_k \exp \tilde{x}_k^{(N)}}{(\sum_k \exp \tilde{x}_k^{(N)})^2} \\
&= \frac{\delta_{ij} \exp \tilde{x}_j^{(N)}}{\sum_k \exp \tilde{x}_k^{(N)}} - \frac{\exp \tilde{x}_i^{(N)} \cdot \exp \tilde{x}_j^{(N)}}{(\sum_k \exp \tilde{x}_k^{(N)})^2} \\
&= \text{softmax}(\tilde{x}_j^{(N)}) \cdot (\delta_{ij} - \text{softmax}(\tilde{x}_i^{(N)})) \\
&\Rightarrow \\
\frac{\partial \boldsymbol{x}^{(N)}}{\partial \tilde{\boldsymbol{x}}^{(N)}} &= \begin{bmatrix} & \vdots & \\ \dots & \text{softmax}(\tilde{x}_j^{(N)}) \cdot (\delta_{ij} - \text{softmax}(\tilde{x}_i^{(N)})) & \dots \\ & \vdots & \end{bmatrix} \\
&\in \mathbb{R}^{d_N \times d_N}
\end{aligned} \tag{$\frac{\partial \boldsymbol{x}^{(N)}}{\partial \tilde{\boldsymbol{x}}^{(N)}}$}$$

$$\begin{aligned}
\frac{\partial \boldsymbol{x}^{(l<N)}}{\partial \tilde{\boldsymbol{x}}^{(l<N)}} &= \frac{\partial}{\partial \tilde{\boldsymbol{x}}^{(l<N)}} \max(0, \tilde{\boldsymbol{x}}^{(l<N)}) \\
&= \boldsymbol{x}^{(l<N)} \oslash \tilde{\boldsymbol{x}}^{(l<N)} \\
&\in \mathbb{R}^{d_l}
\end{aligned} \tag{$\frac{\partial \boldsymbol{x}^{(l<N)}}{\partial \tilde{\boldsymbol{x}}^{(l<N)}}$}$$

$$
\begin{aligned}
\frac{\partial \tilde{\boldsymbol{x}}^{(l)}}{\partial \boldsymbol{x}^{(l-1)}} &= \frac{\partial}{\partial \boldsymbol{x}^{(l-1)}} \boldsymbol{W}^{(l)} \boldsymbol{x}^{(l-1)} + \boldsymbol{b}^{(l)} \\
&= \boldsymbol{W}^{(l)} \\
&\in \mathbb{R}^{d_l \times d_{l-1}}
\end{aligned}
\tag{$\frac{\partial \tilde{\boldsymbol{x}}^{(l)}}{\partial \boldsymbol{x}^{(l-1)}}$}
$$

$$
\begin{aligned}
\frac{\partial \tilde{\boldsymbol{x}}^{(l)}}{\partial \boldsymbol{W}^{(l)}} &= \frac{\partial}{\partial \boldsymbol{W}^{(l)}} \boldsymbol{W}^{(l)} \boldsymbol{x}^{(l-1)} \\
&= \begin{bmatrix} \vdots \\ \frac{\partial \tilde{\boldsymbol{x}}_i^{(l)}}{\partial \boldsymbol{W}^{(l)}} \\ \vdots \end{bmatrix} \\
&\in \mathbb{R}^{d_l \times (d_l \times d_{l-1})} \\
&\text{with} \\
\frac{\partial \tilde{\boldsymbol{x}}_i^{(l)}}{\partial \boldsymbol{W}^{(l)}} &= \begin{bmatrix} \vdots \\ \boldsymbol{x}^{(l-1)^T} \\ \vdots \end{bmatrix} \\
&\in \mathbb{R}^{d_l \times d_{l-1}}
\end{aligned}
\tag{$\frac{\partial \tilde{\boldsymbol{x}}^{(l)}}{\partial \boldsymbol{W}^{(l)}}$}
$$

$$
\begin{aligned}
\frac{\partial \tilde{\boldsymbol{x}}^{(l)}}{\partial \boldsymbol{b}^{(l)}} &= \frac{\partial}{\partial \boldsymbol{b}^{(l)}} \boldsymbol{b}^{(l)} \\
&= \boldsymbol{b}^{(l)} \otimes \boldsymbol{b}^{(l)} \\
&\in \mathbb{R}^{?????}
\end{aligned}
\tag{$\frac{\partial \tilde{\boldsymbol{x}}^{(l)}}{\partial \boldsymbol{b}^{(l)}}$}
$$

Note the use of $\oslash$ for element-wise division and the use of δ for the Kronecker-Delta.

1.1.b

$$
\begin{aligned}
\frac{\partial \boldsymbol{L}}{\partial \tilde{\boldsymbol{x}}^{(N)}} &= \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{x}^{(N)}} \frac{\partial \boldsymbol{x}^{(N)}}{\partial \tilde{\boldsymbol{x}}^{(N)}} \\
&= \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{x}^{(N)}} \cdot \begin{bmatrix} & \vdots & \\ \dots & \text{softmax}(\tilde{x}_j^{(N)}) \cdot (\delta_{ij} - \text{softmax}(\tilde{x}_i^{(N)})) & \dots \\ & \vdots & \end{bmatrix}
\end{aligned}
\tag{$\frac{\partial \boldsymbol{L}}{\partial \tilde{\boldsymbol{x}}^{(N)}}$}
$$

$$
\begin{aligned}
\frac{\partial \boldsymbol{L}}{\partial \tilde{\boldsymbol{x}}^{(l<N)}} &= \frac{\partial \boldsymbol{L}}{\partial \tilde{\boldsymbol{x}}^{(l)}} \frac{\partial \boldsymbol{x}^{(l)}}{\partial \tilde{\boldsymbol{x}}^{(l)}} \\
&= \frac{\partial \boldsymbol{L}}{\partial \tilde{\boldsymbol{x}}^{(l)}} \cdot \boldsymbol{x}^{(l)} \oslash \tilde{\boldsymbol{x}}^{(l)}
\end{aligned}
\tag{$\frac{\partial \boldsymbol{L}}{\partial \tilde{\boldsymbol{x}}^{(l<N)}}$}
$$

$$
\begin{aligned}
\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{x}^{(l<N)}} &= \frac{\partial \boldsymbol{L}}{\partial \tilde{\boldsymbol{x}}^{(l+1)}} \frac{\partial \tilde{\boldsymbol{x}}^{(l+1)}}{\partial \boldsymbol{x}^{(l)}} \\
&= \frac{\partial \boldsymbol{L}}{\partial \tilde{\boldsymbol{x}}^{(l+1)}} \cdot \boldsymbol{W}^{(l+1)}
\end{aligned}
\tag{$\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{x}^{(l<N)}}$}
$$

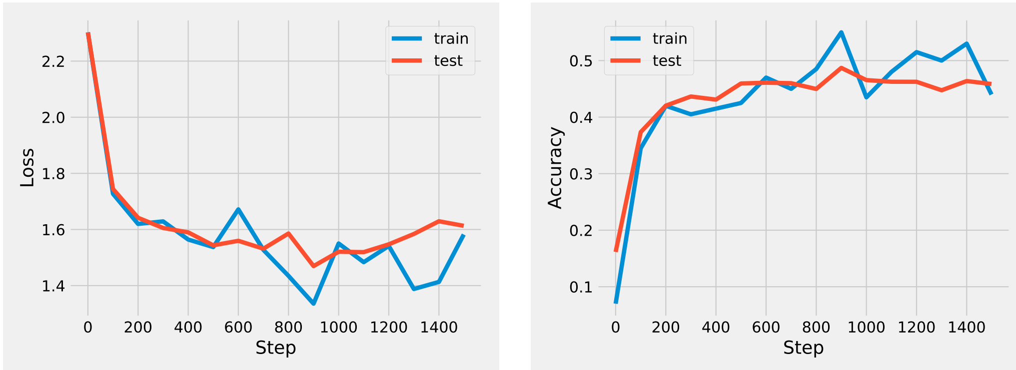

Figure 1: **Left** the loss and **right** the accuracy during training of the NumPy MLP implementation.

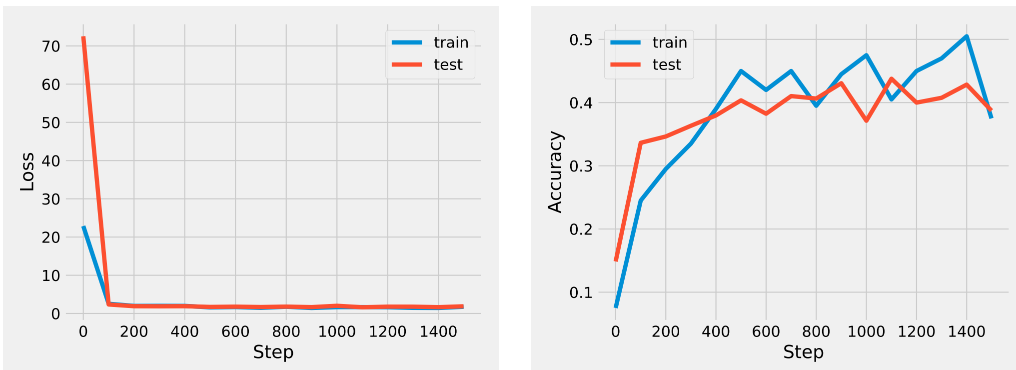

Figure 2: **Left** the loss and **right** the accuracy during training of the PyTorch MLP implementation.

$$\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{W}^{(l)}} = \frac{\partial \boldsymbol{L}}{\partial \tilde{\boldsymbol{x}}^{(l)}} \frac{\partial \tilde{\boldsymbol{x}}^{(l)}}{\partial \boldsymbol{W}^{(l)}} \tag{$\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{W}^{(l)}}$}$$

$$\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{b}^{(l)}} = \frac{\partial \boldsymbol{L}}{\partial \tilde{\boldsymbol{x}}^{(l)}} \frac{\partial \tilde{\boldsymbol{x}}^{(l)}}{\partial \boldsymbol{b}^{(l)}} \tag{$\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{b}^{(l)}}$}$$

1.2 NumPy MLP

2 PyTorch MLP