DDA3020 Assignment 3 Q2 Solution

May 15, 2024

1 Solution

Consider $p = \operatorname{softmax}(a), a \in \mathbb{R}^n$.

$$\begin{aligned}
&\frac{\partial p_i}{\partial a_j} = \frac{\partial}{\partial a_j}\frac{e^{a_i}}{\sum_k e^{a_k}} = e^{a_i} \cdot \frac{-e^{a_j}}{\left(\sum_k e^{a_k}\right)^2} = -\frac{e^{a_i}}{\sum_k e^{a_k}}\frac{e^{a_j}}{\sum_k e^{a_k}} = -p_i p_j \quad (i \neq j) \\
&\frac{\partial p_i}{\partial a_i} = \frac{\partial}{\partial a_i} \cdot \frac{e^{a_i}}{\sum_k e^{a_k}} = \frac{e^{a_i}\left(\left(\sum_k e^{a_k}\right) - e^{a_i}\right)}{\left(\sum_k e^{a_k}\right)^2} = \frac{e^{a_i}}{\sum_k e^{a_k}} \cdot \left(1 - \frac{e^{a_i}}{\sum_k e^{a_k}}\right) = p_i\left(1 - p_i\right) \\
&\therefore \frac{\partial p_i}{\partial a} = -p_i p + p_i \cdot e_i \quad \therefore \frac{\partial p}{\partial a} = -pp^\top + \operatorname{diag}(p)
\end{aligned}$$

Note that $\frac{\partial p}{\partial a}$ is an n by n matrix.
Consider $L = CE(y, x)$

$$\frac{\partial L}{\partial x} = \frac{\partial}{\partial x}\left(-\sum_i y_i \log x_i\right) = -\operatorname{diag}(x)^{-1} y$$

Let $p = \operatorname{softmax}(a)$.

$$\begin{aligned}
\frac{\partial L}{\partial a} &= \frac{\partial p}{\partial a}\frac{\partial L}{\partial p} \\
&= \left(-pp^\top + \operatorname{diag}(p)\right) \cdot \left(-\operatorname{diag}(p)^{-1} y\right) \\
&= p\left(1^\top y\right) - y \\
&= p - y
\end{aligned}$$

Hence, $\delta_1 = p - y$.

$$\begin{aligned}
\frac{\partial L}{\partial z} &= \frac{\partial h}{\partial z} \cdot \frac{\partial a}{\partial h} \cdot \frac{\partial L}{\partial a} \\
&= \operatorname{diag}(\mathbb{1}(z > 0)) \cdot V^\top \cdot \delta_1 \\
&= \left(V^\top \cdot \delta_1\right) \odot \mathbb{1}(z > 0)
\end{aligned}$$

Hence, $\delta_2 = \left(V^\top \delta_1\right) \odot \mathbb{1}(z > 0)$.

$$\frac{\partial L}{\partial v_i} = \frac{\partial a}{\partial v_i} \cdot \frac{\partial L}{\partial a} = [0 \cdots 0 \; h \; 0 \cdots 0] \cdot \delta_1 = \left(\delta_1\right)_i \cdot h \quad \therefore \frac{\partial L}{\partial V} = \delta_1 h^\top$$

$$\frac{\partial L}{\partial b_2} = \frac{\partial a}{\partial b_2} \cdot \frac{\partial L}{\partial a} = I \cdot \delta_1 = \delta_1$$

$$\frac{\partial L}{\partial w_i} = \frac{\partial z}{\partial w_i} \cdot \frac{\partial L}{\partial z} = [0 \cdots 0 \; x \; 0 \cdots 0] \cdot \delta_2 = \left(\delta_2\right)_i \cdot x \quad \therefore \frac{\partial L}{\partial W} = \delta_2 x^\top$$

$$\frac{\partial L}{\partial b_1} = \frac{\partial z}{\partial b_1} \cdot \frac{\partial L}{\partial z} = I \cdot \delta_2 = \delta_2$$

2 A crash course on matrix derivative

Let $\mathbf{x} \in \mathbb{R}^{n \times n}, \mathbf{X} \in \mathbb{R}^{m \times n}$ and other varibales be independent of $\mathbf{x}$ and $\mathbf{X}$.

$$\begin{aligned}
\frac{\partial \mathbf{x}^T \mathbf{a}}{\partial \mathbf{x}} &= \frac{\partial \mathbf{a}^T \mathbf{x}}{\partial \mathbf{x}} = \mathbf{a} \\
\frac{\partial A \mathbf{x}}{\partial \mathbf{x}} &= A^\top \\
\frac{\partial \mathbf{a}^T \mathbf{X} \mathbf{b}}{\partial \mathbf{X}} &= \mathbf{a} \mathbf{b}^T \\
\frac{\partial \mathbf{x}^T A \mathbf{x}}{\partial \mathbf{x}} &= \left(A + A^T\right) \mathbf{x} \\
\frac{\partial \|\mathbf{x}\|_2^2}{\partial \mathbf{x}} &= \frac{\partial \mathbf{x}^T \mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{x}
\end{aligned}$$

Let $f : \mathbb{R}^n \to \mathbb{R}, g : \mathbb{R}^m \to \mathbb{R}^n, f \circ g : \mathbb{R}^m \to \mathbb{R}$. Then

$$\frac{\partial f \circ g\left(\mathbf{x}\right)}{\partial \mathbf{x}} = \overbrace{\frac{\partial g\left(\mathbf{x}\right)}{\partial \mathbf{x}}}^{\in \mathbb{R}^{m \times n}} \overbrace{\frac{\partial f\left(g\left(\mathbf{x}\right)\right)}{\partial g\left(\mathbf{x}\right)}}^{\in \mathbb{R}^n} \in \mathbb{R}^m.$$