Question 1: The softmax function

Let $\mathbf{x} = [x_1, \ldots, x_N] \in \mathbb{R}^{1\times N}$ and $\sigma(\mathbf{x}) = \text{softmax}(\mathbf{x}) = [\sigma_1(\mathbf{x}), \ldots, \sigma_N(\mathbf{x})]$ with $\sigma_i(\mathbf{x}) = \frac{e^{x_i}}{\sum_{j=1}^N e^{x_j}}$.

(a) $D_{\mathbf{x}}\sigma(\mathbf{x}) \in \mathbb{R}^{N\times N} = \begin{bmatrix} \frac{\partial\sigma_1(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial\sigma_1(\mathbf{x})}{\partial x_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial\sigma_N(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial\sigma_N(\mathbf{x})}{\partial x_N} \end{bmatrix}$

Diagonal entries: $\boxed{\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}}$

$$\begin{aligned} \frac{\partial\sigma_i(\mathbf{x})}{\partial x_i} &= \frac{e^{x_i}\left(\sum_{j=1}^N e^{x_j}\right) - (e^{x_i})^2}{\left(\sum_{j=1}^N e^{x_j}\right)^2} && (1) \\ &= \sigma_i(\mathbf{x}) - \sigma_i^2(\mathbf{x}) && (2) \\ &= \sigma_i(\mathbf{x})\left(1 - \sigma_i(\mathbf{x})\right) && (3) \end{aligned}$$

Off-diagonal entries: $\boxed{\left(c\cdot f^{-1}\right)' = -c\cdot f^{-2}\cdot f' \text{ with } c = \text{const.}}$

$$\begin{aligned} \frac{\partial\sigma_i(\mathbf{x})}{\partial x_j} &= \frac{-e^{x_i}e^{x_j}}{\left(\sum_{j=1}^N e^{x_j}\right)^2} && (4) \\ &= -\sigma_i(\mathbf{x})\sigma_j(\mathbf{x}) && (5) \end{aligned}$$

Finally, $D_{\mathbf{x}}\sigma(\mathbf{x}) = \begin{bmatrix} \sigma_1(\mathbf{x})\left(1-\sigma_1(\mathbf{x})\right) & \ldots & -\sigma_N(\mathbf{x})\sigma_1(\mathbf{x}) \\ -\sigma_1(\mathbf{x})\sigma_2(\mathbf{x}) & \ldots & -\sigma_N(\mathbf{x})\sigma_2(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ -\sigma_1(\mathbf{x})\sigma_N(\mathbf{x}) & \ldots & \sigma_N(\mathbf{x})\left(1-\sigma_N(\mathbf{x})\right) \end{bmatrix}$

(b) Let $\mathbf{v} = [v_1, \ldots, v_N] \in \mathbb{R}^{1\times N}$ and $\mathbf{z} = [z_1, \ldots, z_N] = \mathbf{v}\cdot D_{\mathbf{x}}\sigma(\mathbf{x})$.
Specifically, for z_i with $i = 1$:

$$\begin{aligned} z_1 &= \mathbf{v}\cdot \text{col}_1\big(D_{\mathbf{x}}\sigma(\mathbf{x})\big) \\ &= v_1\sigma_1(\mathbf{x})(1-\sigma_1(\mathbf{x})) + v_2(-\sigma_1(\mathbf{x})\sigma_2(\mathbf{x})) + \ldots + v_N(-\sigma_1(\mathbf{x})\sigma_N(\mathbf{x})) \\ &= \sigma_1(\mathbf{x})\left(v_1 - \cancel{v_1\sigma_1(\mathbf{x})} - \sum_{j=1}^N v_j\sigma_j(\mathbf{x}) + \cancel{v_1\sigma_1(\mathbf{x})}\right) \\ &= \sigma_1(\mathbf{x})\left(v_1 - \mathbf{v}\cdot\sigma(\mathbf{x})^\top\right) \end{aligned}$$

Generally, $z_i = \sigma_i(\mathbf{x})\left(v_i - \mathbf{v}\cdot\sigma(\mathbf{x})^\top\right)$

Notation: $\text{col}_i(\cdot)$ returns the i-th column of the specified matrix.

(c) Let $\ell(\mathbf{z},\mathbf{t}) = -\sum_{i=1}^{N} t_i \ln(z_i)$ with $t \in [0,1]^{1\times N}$ e.g. $t = [0\,0\,1\,0]$ and $D_{\mathbf{z}}\ell(\mathbf{z},\mathbf{t}) = \left[\frac{\partial \ell(\mathbf{z},\mathbf{t})}{\partial z_1}, \ldots, \frac{\partial \ell(\mathbf{z},\mathbf{t})}{\partial z_N}\right] \in \mathbb{R}^{1\times N}$

$$\frac{\partial \ell(\mathbf{z},\mathbf{t})}{\partial z_i} = \frac{\partial}{\partial z_i}\Big(- t_i \cdot \ln(z_i)\Big) \tag{6}$$

$$= -\frac{t_i}{z_i} \text{ with } \boxed{\ln\,(x)' = \frac{1}{x}} \tag{7}$$

$D_{\mathbf{z}}\ell(\mathbf{z},\mathbf{t}) = \left[-\frac{t_1}{z_1}, \ldots, -\frac{t_N}{z_N}\right]$

(d) Division by zero when $z_i = 0$. This happens when any class gets 0 probability.

Question 3: A deeper network

(a)

$$\begin{aligned}
\frac{\partial \tanh}{\partial x}(x) &\overset{\text{def. of tanh}}{=} \frac{\partial}{\partial x}\frac{e^x - e^{-x}}{e^x + e^{-x}} \\
&\overset{\text{quotient rule}}{=} \frac{\left(\frac{\partial}{\partial x}(e^x - e^{-x})\right)(e^x + e^{-x}) - (e^x - e^{-x})\frac{\partial}{\partial x}(e^x + e^{-x})}{(e^x + e^{-x})^2} \\
&= \frac{(e^x + e^{-x})(e^x + e^{-x}) - (e^x - e^{-x})(e^x - e^{-x})}{(e^x + e^{-x})^2} \\
&= 1 - \frac{(e^x - e^{-x})^2}{(e^x + e^{-x})^2} = 1 - \tanh^2(x)
\end{aligned}$$

Question 4: A more stable softmax

(a)

$$\text{softmax}_i(x + c) = \frac{e^{x_i + c}}{\sum_{j=1}^{N} e^{x_j + c}} = \frac{e^c e^{x_i}}{e^c \sum_{j=1}^{N} e^{x_j}} = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}} = \text{softmax}_i(x)$$

(b)

$$\log \sigma_i(x) = \log \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}} = \log(e^{x_i}) - \log(\sum_{j=1}^{N} e^{x_j}) = x_i - \log(\sum_{j=1}^{N} e^{x_j})$$

diagonal:

$$\frac{\partial}{\partial x_i}\log(\sigma_i(x)) = 1 - \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}} = 1 - \sigma_i(x)$$

off-diagonal ($i \neq j$):

$$\frac{\partial}{\partial x_j}\log(\sigma_i(x)) = \underbrace{\frac{\partial}{\partial x_j} x_i}_{=0} - \frac{\partial}{\partial x_j}\log(\sum_{k=1}^{N} e^{x_k}) = -\frac{e^{x_j}}{\sum_{k=1}^{N} e^{x_k}} = -\sigma_j(x)$$

putting it together:

$$
\begin{aligned}
D_x \log(\sigma(x)) &\overset{\text{def. Jacobian}}{:=} \left(\frac{\partial}{\partial x_j} \log(\sigma_i(x))\right)_{ij} \\
&= \begin{pmatrix}
1-\sigma_1(x) & -\sigma_2(x) & \cdots & -\sigma_{N-1}(x) & -\sigma_N(x) \\
-\sigma_1(x) & 1-\sigma_2(x) & -\sigma_3(x) & \cdots & -\sigma_N(x) \\
\vdots & -\sigma_2(x) & \ddots & & \vdots \\
\vdots & \vdots & & \ddots & -\sigma_N(x) \\
-\sigma_1(x) & -\sigma_2(x) & \cdots & -\sigma_{N-1}(x) & 1-\sigma_N(x)
\end{pmatrix}
\end{aligned}
$$

(c)

$$z = v \cdot \underbrace{D_x \log(\sigma(x))}_{\text{matrix from b)}}$$

e.g. $i = 1$:

$$z_1 = v_1(1-\sigma_1(x)) - v_2\sigma_1(x) - \cdots - v_N\sigma_1(x)) = v_1 - \sigma_1(x)\sum_{j=1}^{N} v_j$$

general case:

$$z_i = v_i - \sigma_i(x)\sum_{j=1}^{N} v_j$$

(d)

$$\ell(z,t) = -\sum_{i=1}^{N} t_i z_i$$

$$\frac{\partial}{\partial z_i}\ell(z,t) = -\frac{\partial}{\partial z_i} t_i z_i = -t_i \Rightarrow D_z\ell(z,t) = \begin{pmatrix} -t_1 & \cdots & -t_N \end{pmatrix} = -t$$