ML W20 Exercise 4

Question 1 : The Softmax function

$\bar{x} \in \mathbb{R}^{1\times N}$, $\sigma(\bar{x}) = \text{softmax}(\bar{x}) = [\sigma_i(\bar{x})]_{i=1\dots N}$ where $\sigma_i(\bar{x}) = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}}$

(a) Derive the softmax's Jacobian $D_{\bar{x}}\sigma(\bar{x}) \in \mathbb{R}^{N\times N}$

Diagonal entries:

$$\frac{\partial \sigma_i(\bar{x})}{\partial x_i} = \frac{\partial}{\partial x_i}\left[\frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}}\right]$$

$$= \frac{e^{x_i}}{\sum e^{x_j}} - \left(\frac{e^{x_i}}{\sum e^{x_j}}\right)^2 \qquad \left(\frac{\partial}{\partial x}\frac{u}{v} = \frac{1}{v}\frac{\partial u}{\partial x} - \frac{u}{v^2}\frac{\partial v}{\partial x}\right)$$

$$= \sigma_i(\bar{x}) - \sigma_i^2(x)$$

Off diagonal entries:

$$\frac{\partial \sigma_i(\bar{x})}{\partial x_k}, k\neq i = \frac{\partial}{\partial x_k}\left[\frac{e^{x_i}}{\sum e^{x_j}}\right]$$

$$= \frac{0}{\sum e^{x_j}} - \frac{e^{x_i}}{(\sum e^{x_j})^2}\cdot e^{x_k}$$

$$= -\sigma_i(x)\,\sigma_k(x)$$

$$D_x\sigma(x) = \begin{bmatrix} \sigma_1(x)-\sigma_1^2(x) & -\sigma_1(x)\sigma_2(x) & \cdots \\ -\sigma_2(x)\sigma_1(x) & \sigma_2(x)-\sigma_2^2(x) & \\ \vdots & & \ddots \end{bmatrix}$$

(b) incoming vector $\bar{v} = (v_1 \dots v_N)$ will be multiplied by the Jacobian.

$$\bar{z} = \bar{v}\cdot D_x\sigma(\bar{x})$$

$$z_i = v_1 \cdot D_x\sigma(\bar{x})_{i1} + v_2 \cdot D_x\sigma(\bar{x})_{i2} + \dots + v_N \cdot D_x\sigma(\bar{x})_{iN}$$

$$= -\sum_{j\neq i} v_j\,\sigma_i(\bar{x})\,\sigma_j(\bar{x}) + v_i\sigma(\bar{x})_i - v_i\sigma_i^2(\bar{x})$$

$$= \sigma_i(\bar{x})\left[-\sum_{j\neq i} v_j\sigma_j(\bar{x}) + v_i - v_i\sigma_i(\bar{x})\right]$$

$$= \sigma_i(\bar{x})\left[v_i - \sum_{j=1}^{N} v_j\sigma_j(\bar{x})\right]$$

$$z_i = \sigma_i(\bar{x})\left[v_i - \bar{v}\cdot\sigma(\bar{x})^T\right]$$

(c) $\ell(\bar{z},\bar{t}) = -\sum_{i=1}^{N} t_i \ln(z_i)$, $\bar{t} \in [0,1]^{1\times N}$, $\sum_{i=1}^{N} t_i = 1$

$D_z\ell(\bar{z},\bar{t}) \in \mathbb{R}^{1\times N}$

$$D_z\ell(\bar{z},\bar{t})_i = \frac{\partial}{\partial z_i}\left[-\sum_{j=1}^{N} t_j \ln(z_j)\right]$$

$$= -t_i\cdot\frac{1}{z_i}$$

$$D_z\ell(\bar{z},\bar{t}) = \left[-\frac{t_i}{z_i}\right]_{i=1\dots N}$$

(d) The cross-entropy jacobian could have values that start to approach ∞ if z_i is sufficiently small. This will occur for $\sigma_i(x) \approx 0$ so when $e^{x_i} \ll \sum_{j=1}^{N} e^{x_j}$.

Question 3:

(a) $\partial/\partial x \tanh(x) = 1 - \tanh^2(x)$, $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$

$$\partial/\partial x \tanh(x) = \partial/\partial x \left[\frac{e^x - e^{-x}}{e^x + e^{-x}}\right]$$

$$= \frac{1}{e^x + e^{-x}}\left[e^x + e^{-x}\right] - \frac{1}{(e^x + e^{-x})^2}\left[e^x - e^{-x}\right]\left[e^x - e^{-x}\right] \qquad \left(\partial/\partial x \left[\frac{u}{v}\right] = \frac{1}{v}\frac{\partial u}{\partial x} - \frac{1}{v^2} \cdot u \cdot \frac{dv}{\partial x}\right)$$

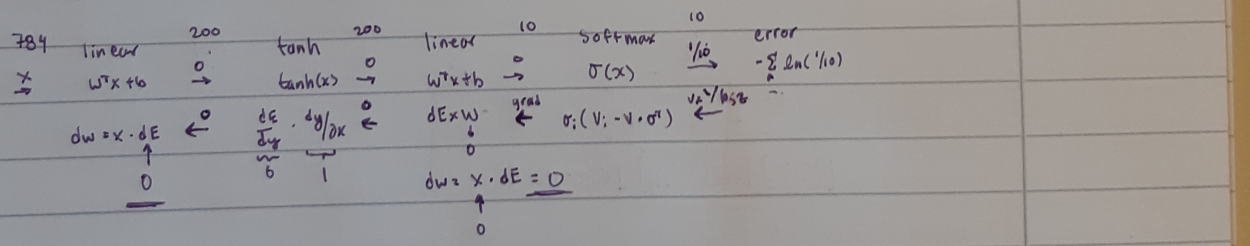
$$= 1 - \frac{(e^x - e^{-x})^2}{(e^x + e^{-x})^2}$$

$$= 1 - \tanh^2(x)$$

(b) fprop: $y = \tanh(x)$

bprop: $\partial E/\partial x = \underbrace{\frac{\partial E}{\partial y}}_{\text{input}} \cdot \underbrace{\partial y/\partial x}_{1-\tanh^2(x)}$

(c) If the values for W^T & b are initialized to zero the input to the upper layers will be zero, and then when propagating the gradient backwards it will lead to propagating zeroes and lead to no parameter updates! So the network doesn't learn.



Question 4:

(a) $\text{Softmax}(x) = \text{Softmax}(x+c)$ $\quad c \in \mathbb{R}$

$$\frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}} = \frac{e^{x_i + c}}{\sum_{j=1}^{N} e^{x_j + c}}$$

$$= \frac{e^c e^{x_i}}{\sum_{j=1}^{N} e^c e^{x_j}}$$

$$= \frac{e^c e^{x_i}}{e^c \sum_{j=1}^{N} e^{x_j}}$$

$$= \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}}$$

b) log-softmax Jacobian $D_x \log(\sigma(x)) \in \mathbb{R}^{N \times N}$

Diagonal entries:

$\partial / \partial x_i [\log(\sigma_i(x))]$: $\quad i, j \in 1..N$

$\frac{1}{\sigma_i(x)} \cdot \frac{\partial}{\partial x_i}[\sigma_i(x)]$

$\frac{\sum_j e^{x_j}}{e^{x_i}} \cdot \partial / \partial x_i \left[\frac{e^{x_i}}{\sum_j e^{x_j}} \right]$

$\frac{\sum_j e^{x_j}}{e^{x_i}} \cdot \left[\frac{e^{x_i}}{\sum_j e^{x_j}} - \frac{e^{x_i}}{(\sum_j e^{x_j})^2} \cdot \partial / \partial x_i \left(\sum_j e^{x_j} \right) \right]$

" " " $\cdot e^{x_i}$

$1 - \frac{e^{x_i}}{\sum_j e^{x_j}}$

$1 - \sigma_i(x)$

off diagonal:

$\partial / \partial x_K [\log(\sigma_i(x))]$ $K \neq i$, $\quad i, j, K \in 1..N$

$\frac{1}{\sigma_i(x)} \cdot \partial / \partial x_K [\sigma_i(x)]$

$\frac{\sum_j e^{x_j}}{e^{x_i}} \cdot \partial / \partial x_K \left[\frac{e^{x_i}}{\sum_j e^{x_j}} \right]$ $\qquad \partial / \partial x_K [\sum_j e^{x_j}]$

" $\cdot \left[\frac{0}{\sum_j e^{x_j}} - \frac{1}{(\sum_j e^{x_j})^2} \cdot [e^{x_i}] \cdot \overbrace{[e^{x_K}]} \right]$

$\frac{\sum_j e^{x_j}}{e^{x_i}} \left[- \frac{e^{x_i} e^{x_K}}{(\sum_j e^{x_i})^2} \right]$

$- \frac{e^{x_K}}{\sum_j e^{x_j}}$

$= - \sigma_K(x)$

$$D_x \log(\sigma(x)) = \begin{bmatrix} 1 - \sigma_1(x) & -\sigma_2(x) & -\sigma_3(x) \cdots \\ -\sigma_1(x) & 1 - \sigma_2(x) & -\sigma_3(x) \cdots \\ -\sigma_1(x) & -\sigma_2(x) & \ddots \\ \vdots & \vdots & \cdots \end{bmatrix}$$

(c) bprop: $\bar{z} = \bar{V} \cdot D_x \log(\sigma(\bar{x}))$ $\qquad \bar{V} = (v_1 \cdots v_N)$

$$\bar{z} = [v_1 \cdots v_N] \begin{bmatrix} 1 - \sigma_1(x) & -\sigma_2(x) & \cdots \\ -\sigma_1(x) & \ddots & \\ \vdots & & \end{bmatrix}$$

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

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

(d) Modified cross-entropy Criterion:

$l(\bar{z}, \bar{t}) = - \sum_{i=1}^{N} t_i z_i = -\bar{t} \cdot \bar{z}^T$

$D_z l(\bar{z}, \bar{t}) \in \mathbb{R}^{1 \times N} = \partial / \partial z \left[- \sum_{i=1}^{N} t_i z_i \right]$

$= [-t_1, -t_2, \ldots, -t_N]$