Solutions by **Andrew Lys** `andrewlys(at)u.e.`

1. **Back Propagation.**

(a) Let $\sigma(x) = \frac{1}{1+e^{-x}}$ be the sigmoid function. Let $\sigma_s(x)$ be the textbook softmax function, i.e.

$$\sigma_s(x) = \begin{bmatrix} \frac{e^{x_1}}{\sum_{i=1}^n e^{x_i}} \\ \vdots \\ \frac{e^{x_n}}{\sum_{i=1}^n e^{x_i}} \end{bmatrix}$$

The given softmax function in the homework is then $\sigma_s(z) \cdot z$.

Suppose $o[v]$ is computed with softmax. We define the activation energy to then be a vector:

$$a[v] = \begin{bmatrix} a[v][1] \\ \vdots \\ a[v][n] \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n w(1, u_i, v) \cdot o[u_i] \\ \vdots \\ \sum_{i=1}^n w(n, u_i, v) \cdot o[u_i] \end{bmatrix}$$

Then we have the following:

$$\begin{aligned} \frac{\partial \hat{y}}{\partial w(i,u,v)} &= \frac{\partial o[v_{out}]}{\partial w(i,u,v)} \\ &= \sum_j \frac{\partial o[v_{out}]}{\partial a[v][j]} \frac{\partial a[v][j]}{\partial w(i,u,v)} \\ &= \frac{\partial o[v_{out}]}{\partial a[v][i]} o[u] \end{aligned}$$

Let

$$\delta[v][i] = \frac{\partial o[v_{out}]}{\partial a[v][i]}$$

Then we have:

$$\frac{\partial \hat{y}}{\partial w(i,u,v)} = \delta[v][i] o[u]$$

If $o[v]$ is computed with sigmoid activation, we define the activation energy as usual, a scalar, and we have:

$$\frac{\partial \hat{y}}{\partial w(u,v)} = \frac{\partial o[v_{out}]}{a[v]} o[u]$$

We let

$$\gamma[v] = \frac{\partial o[v_{out}]}{a[v]}$$

Then we have:

$$\frac{\partial \hat{y}}{\partial w(u,v)} = \gamma[v] o[u]$$

Suppose v is the output node. We do the two cases separately.

i.

$$\begin{aligned} \delta[v][i] &= \frac{o[v]}{\partial a[v][i]} \\ &= \frac{\partial \sigma_s(a[v]) \cdot a[v]}{\partial a[v][i]} \\ &= \sum_j \frac{\partial}{\partial a[v][i]} \sigma_s(a[v])[j] \cdot a[v][j] \end{aligned}$$

$$
\begin{aligned}
&= \sum_j \sigma_s(a[v])[j]\delta_{ij} + a[v][j]\sigma_s(a[v])[i](\delta_{ij} - \sigma_s(a[v])[j]) \\
&= \sigma_s(a[v])[i] + a[v][i]\sigma_s(a[v])[i](1 - \sigma_s(a[v])[i]) \\
&- \sum_{j \neq i} a[v][j]\sigma_s(a[v])[i]\sigma_s(a[v])[j]
\end{aligned}
$$

ii.

$$
\begin{aligned}
\gamma[v] &= \frac{o[v]}{\partial a[v]} \\
&= \frac{\partial \sigma(a[v])}{\partial a[v]} \\
&= \sigma(a[v])(1 - \sigma(a[v]))
\end{aligned}
$$

If v is not the output node, suppose v is a parent node of v_{out}. We do the two cases separately.

i.

$$
\begin{aligned}
o[v_{out}] &= \sigma_s(a[v_{out}]) \cdot a[v_{out}] \\
\frac{\partial}{\partial a[v][i]}\sigma_s(a[v_{out}]) \cdot a[v_{out}] &= \sigma_s(a[v_{out}]) \cdot \frac{\partial a[v_{out}]}{\partial a[v][i]} + a[v_{out}] \cdot \frac{\partial \sigma_s(a[v_{out}])}{\partial a[v][i]} \\
&= \sigma_s(a[v_{out}]) \cdot w(v, v_{out}) + a[v_{out}] \cdot J\sigma_s(a[v_{out}]) \cdot w(v, v_{out})
\end{aligned}
$$

Where $w(v, v_out)$ is the vector of weights $(w(1, v, v_{out}), \ldots, w(n, v, v_{out}))$ and $J\sigma_s(a[v_{out}])$ is the Jacobian of the softmax function evaluated at $a[v_{out}]$.

ii.

$$
\frac{\partial o[v_{out}]}{\partial a[v]} = \sigma'(a[v_{out}])w(v, v_{out})
$$

If v is not a parent node of v_{out}, then we have a simple recursive formula for $\delta[v][i]$ and $\gamma[v]$.

i. We deal with the case where v is calculated with softmax activation. If v_{out} is calculated with softmax, We have:

$$
\begin{aligned}
\delta[v][i] &= \sum_{v_p \in \text{parents}(v)} \frac{\partial o[v_{out}]}{\partial o[v_p]} \frac{\partial o[v_p]}{\partial a[v][i]} \\
&= \sum_{v_p \in \text{parents}(v)} \left(\sigma_s(a[v_{out}]) \cdot w(v_p, v_{out}) + a[v_{out}]J\sigma_s(a[v_{out}])w(v_p, v_{out})\right) \delta^{(v_p)}[v][i]
\end{aligned}
$$

Where $\delta^{(v_p)}[v][i]$ is calculated as if v_p were the output node. In the case of sigmoid activation for $o[v_{out}]$, we have:

$$
\begin{aligned}
\delta[v][i] &= \sum_{v_p \in \text{parents}(v)} \frac{\partial o[v_{out}]}{\partial o[v_p]} \frac{\partial o[v_p]}{\partial a[v][i]} \\
&= \sum_{v_p \in \text{parents}(v)} \sigma'(a[v_{out}])w(v_p, v_{out})\delta^{(v_p)}[v][i]
\end{aligned}
$$

ii. We deal with the case where v is calculated with sigmoid activation. If v_{out} is calculated with sigmoid activation, we have:

$$
\begin{aligned}
\gamma[v] &= \sum_{v_p \in \text{parents}(v)} \frac{\partial o[v_{out}]}{\partial o[v_p]} \frac{\partial o[v_p]}{\partial a[v]} \\
&= \sum_{v_p \in \text{parents}(v)} \sigma'(a[v_{out}])w(v_p, v_{out})\gamma^{(v_p)}[v]
\end{aligned}
$$

Where $\gamma^{(v_p)}[v]$ is calculated as if v_p were the output node. We deal with the case where v_{out} is calculated with softmax activation. We have:

$$\begin{aligned}\gamma[v] &= \sum_{v_p \in \text{parents}(v)} \frac{\partial o[v_{out}]}{\partial o[v_p]} \frac{\partial o[v_p]}{\partial a[v]} \\ &= \sum_{v_p \in \text{parents}(v)} (\sigma_s(a[v_{out}]) \cdot w(v_p, v_{out}) + a[v_{out}] J\sigma_s(a[v_{out}]) w(v_p, v_{out}))\, \gamma^{(v_p)}[v]\end{aligned}$$