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:

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

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:

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

(b) Let $o[x] = x$. Let $u$ be the children of $x$, and let $a[u][i] = W^{(1)}x$. Then $\sigma(a[u]) = o[u]$. Let $v$ be an additional implied layer between $\hat{y}$ and $u$. Let $a[v] = W^{(2)}o[u]$ and $o[v] = a[v]$. Then $\hat{y} = \sigma_s(a[v]) \cdot a[v]$. With this notation we have:

$$\begin{aligned}
\nabla_{W^{(2)}} \ell^{sq}(\hat{y}, y) &= \nabla_{W^{(2)}} \frac{1}{2}(\hat{y} - y)^2 \\
&= (\hat{y} - y) \nabla_{W^{(2)}} \hat{y} \\
d\hat{y} &= d(\sigma_s(a[v])^\top a[v]) \\
&= \sigma_s(W^{(2)}o[u])^\top d(W^{(2)}o[u]) + d\sigma_s(W^{(2)}o[u])^\top (W^{(2)}o[u]) \\
&= \sigma_s(W^{(2)}o[u])^\top dW^{(2)}o[u] + (W^{(2)}o[u])^\top J\sigma_s(W^{(2)}o[u]) d(W^{(2)}o[u]) \\
&= \text{Tr}(\sigma_s(W^{(2)}o[u])^\top dW^{(2)}o[u]) + \text{Tr}(o[u]^\top W^{(2)\top} J\sigma_s(W^{(2)}o[u]) dW^{(2)}o[u]) \\
&= \text{Tr}(o[u]\sigma_s(W^{(2)}o[u])^\top dW^{(2)}) + \text{Tr}(o[u]o[u]^\top W^{(2)\top} J\sigma_s(W^{(2)}o[u]) dW^{(2)}) \\
\implies \frac{d\hat{y}}{dW^{(2)}} &= o[u]\sigma_s(W^{(2)}o[u])^\top + o[u]o[u]^\top W^{(2)\top} J\sigma_s(W^{(2)}o[u]) \\
\implies \nabla_{W^{(2)}} \ell^{sq}(\hat{y}, y) &= (\hat{y} - y) \left[o[u]\sigma_s(W^{(2)}o[u])^\top + o[u]o[u]^\top W^{(2)\top} J\sigma_s(W^{(2)}o[u])\right] \\
&= (\hat{y} - y) \left(o[u]\sigma_s(a[v])^\top + o[u]a[v]^\top J\sigma_s(a[v])\right)
\end{aligned}$$

Where $J\sigma_s(a[v])$ is the Jacobian of the softmax function evaluated at $a[v]$. For completeness, we have:

$$\begin{aligned}
J\sigma_s(z) &= \begin{bmatrix}
\sigma_s(z)[1](1 - \sigma_s(z)[1]) & -\sigma_s(z)[1]\sigma_s(z)[2] & \dots & -\sigma_s(z)[1]\sigma_s(z)[n] \\
-\sigma_s(z)[2]\sigma_s(z)[1] & \sigma_s(z)[2](1 - \sigma_s(z)[2]) & \dots & -\sigma_s(z)[2]\sigma_s(z)[n] \\
\vdots & \vdots & \ddots & \vdots \\
-\sigma_s(z)[n]\sigma_s(z)[1] & -\sigma_s(z)[n]\sigma_s(z)[2] & \dots & \sigma_s(z)[n](1 - \sigma_s(z)[n])
\end{bmatrix} \\
&= I\sigma_s(z) - \sigma_s(z)\sigma_s(z)^T
\end{aligned}$$

This gives us:

$$\begin{aligned}
\nabla_{W^{(2)}} \ell^{sq}(\hat{y}, y) &= (\hat{y} - y) \left[o[u]\sigma_s(a[v])^\top + o[u]a[v]^\top (I\sigma_s(a[v]) - \sigma_s(a[v])\sigma_s(a[v])^\top)\right] \\
&= (\hat{y} - y) \left[o[u]\sigma_s(a[v])^\top + o[u]a[v]^\top \sigma_s(a[v]) - o[u]a[v]^\top \sigma_s(a[v])\sigma_s(a[v])^\top\right]
\end{aligned}$$

Now we calculate $\nabla_{W^{(1)}} \ell^{sq}(\hat{y}, y)$.

$$\begin{aligned}
\nabla_{W^{(1)}} \ell^{sq}(\hat{y}, y) &= \nabla_{W^{(1)}} \frac{1}{2}(\hat{y} - y)^2 \\
&= (\hat{y} - y) \nabla_{W^{(1)}} \hat{y} \\
d\hat{y} &= d(\sigma_s(a[v])^\top a[v]) \\
&= \sigma_s(a[v])^\top d(W^{(2)}o[u]) + a[v]^\top d\sigma_s(W^{(2)}o[u]) \\
&= \sigma_s(a[v])^\top W^{(2)} do[u] + a[v]^\top J\sigma_s(W^{(2)}o[u]) d(W^{(2)}o[u]) \\
&= \sigma_s(a[v])^\top W^{(2)} d\sigma(W^{(1)}x) + a[v]^\top J\sigma_s(a[v]) W^{(2)} d\sigma(W^{(1)}x) \\
d\sigma(W^{(1)}x) &= \sigma'(W^{(1)}x) \odot dW^{(1)}x \\
&= (\sigma(W^{(1)}x)(1 - \sigma(W^{(1)}x))) \odot dW^{(1)}x \\
&= \text{Diag}(\sigma(W^{(1)}x)(1 - \sigma(W^{(1)}x))) dW^{(1)}x \\
\implies d\hat{y} &= \sigma_s(a[v])^\top W^{(2)} \,\text{Diag}(\sigma(W^{(1)}x)(1 - \sigma(W^{(1)}x))) dW^{(1)}x
\end{aligned}$$

$$
\begin{aligned}
&+ a[v]^\top J\sigma_s(a[v])W^{(2)}\,\mathrm{Diag}(\sigma(W^{(1)}x)(1-\sigma(W^{(1)}x)))dW^{(1)}x \\
&= \mathrm{Tr}\left[\left(\sigma_s(a[v])^\top + a[v]^\top J\sigma_s(a[v])\right)W^{(2)}\,\mathrm{Diag}(\sigma(W^{(1)}x)(1-\sigma(W^{(1)}x)))dW^{(1)}x\right] \\
&= \mathrm{Tr}\left[x\left(\sigma_s(a[v])^\top + a[v]^\top J\sigma_s(a[v])\right)W^{(2)}\,\mathrm{Diag}(\sigma(W^{(1)}x)(1-\sigma(W^{(1)}x)))dW^{(1)}\right] \\
\implies \frac{d\hat{y}}{dW^{(1)}} &= x\left(\sigma_s(a[v])^\top + a[v]^\top J\sigma_s(a[v])\right)W^{(2)}\,\mathrm{Diag}(\sigma(W^{(1)}x)(1-\sigma(W^{(1)}x)))
\end{aligned}
$$

Keep in mind that $\sigma$ is taken element wise in the above calculations, and we should have $\odot$ for element wise multiplication between $\sigma(W^{(1)}x)$ and $(1-\sigma(W^{(1)}x))$ inside the Diag operator, but the meaning is clear regardless. We then get the final result:

$$\nabla_{W^{(1)}}\ell^{sq}(\hat{y},y) = (\hat{y}-y)x\left(\sigma_s(a[v])^\top + a[v]^\top J\sigma_s(a[v])\right)W^{(2)}\,\mathrm{Diag}(\sigma(W^{(1)}x)(1-\sigma(W^{(1)}x)))$$

2. **Multiclass and Structured Prediction.**

(1) We have:

$$
\begin{aligned}
h_w(x) &= \arg\max_{y\in\{\pm1\}}\left\langle w, \frac{1}{2}y\phi(x)\right\rangle \\
\langle w,\phi(x)\rangle > 0 &\implies h_w(x) = 1 \\
\langle w,\phi(x)\rangle < 0 &\implies h_w(x) = -1 \\
\therefore \qquad h_w(x) &= \mathrm{sign}(\langle w,\phi(x)\rangle)
\end{aligned}
$$

Recall the binary hinge loss:

$$\ell^{hinge}(h(x),y) = [1-yh(x)]_+$$

In the binary case of multiclass prediction, we have:

$$
\begin{aligned}
\ell^{hinge}(w,(x,y)) &= \max_{y'\in\{\pm1\}}\left([[y'\neq y]] + \frac{1}{2}y'\langle w,\phi(x)\rangle - \frac{1}{2}y\langle w,\phi(x)\rangle\right) \\
y'\neq y &\implies [[y'\neq y]] + \frac{1}{2}y'\langle w,\phi(x)\rangle - \frac{1}{2}y\langle w,\phi(x)\rangle = 1 + \frac{1}{2}(-y)\langle w,\phi(x)\rangle - \frac{1}{2}y\langle w,\phi(x)\rangle \\
&= 1 - y\langle w,\phi(x)\rangle = 1 - yh_w(x) \\
y = y &\implies [[y'\neq y]] + \frac{1}{2}y'\langle w,\phi(x)\rangle - \frac{1}{2}y\langle w,\phi(x)\rangle = 0 \\
\therefore \qquad \ell^{hinge}(w,(x,y)) &= \max(0, 1-yh_w(x)) = [1-yh_w(x)]_+
\end{aligned}
$$