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:

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

(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:

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

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

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

$$
\begin{aligned}
&+ a[v]^\top J\sigma_s(a[v])W^{(2)}\operatorname{Diag}(\sigma(W^{(1)}x)(1-\sigma(W^{(1)}x)))dW^{(1)}x \\
&= \operatorname{Tr}\left[\left(\sigma_s(a[v])^\top + a[v]^\top J\sigma_s(a[v])\right)W^{(2)}\operatorname{Diag}(\sigma(W^{(1)}x)(1-\sigma(W^{(1)}x)))dW^{(1)}x\right] \\
&= \operatorname{Tr}\left[x\left(\sigma_s(a[v])^\top + a[v]^\top J\sigma_s(a[v])\right)W^{(2)}\operatorname{Diag}(\sigma(W^{(1)}x)(1-\sigma(W^{(1)}x)))dW^{(1)}\right]
\end{aligned}
$$

$$
\implies \frac{d\hat{y}}{dW^{(1)}} = x\left(\sigma_s(a[v])^\top + a[v]^\top J\sigma_s(a[v])\right)W^{(2)}\operatorname{Diag}(\sigma(W^{(1)}x)(1-\sigma(W^{(1)}x)))
$$

Keep in mind that σ 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:

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

2. **Multiclass and Structured Prediction.**

(a) We have:

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

Recall the binary hinge loss:

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

In the binary case of multiclass prediction, we have:

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

(b) (i) First we show that L_S^{hinge} is convex in w. We show that for every (x,y), $\ell^{hinge}(w,(x,y))$ is convex in w. Let u, $v\in\mathbb{R}^d$ and let $\lambda\in[0,1]$. Then we have:

$$
\begin{aligned}
\ell^{hinge}(\lambda u + (1-\lambda)v,(x,y)) &= \max_{y'\in\mathcal{Y}}(\Delta(y',y) + \langle \lambda u + (1-\lambda)v, \Psi(x,y')\rangle - \langle \lambda u + (1-\lambda)v, \Psi(x,y)\rangle) \\
&= \max_{y'\in\mathcal{Y}}[\lambda(\Delta(y',y) + \langle u,\Psi(x,y')\rangle - \langle u,\Psi(x,y)\rangle) + (1-\lambda)(\Delta(y',y) + \langle v,\Psi(x,y')\rangle - \langle v,\Psi(x,y)\rangle)] \\
&\leq \lambda\max_{y'\in\mathcal{Y}}(\Delta(y',y) + \langle u,\Psi(x,y')\rangle - \langle u,\Psi(x,y)\rangle) \\
&+ (1-\lambda)\max_{y'\in\mathcal{Y}}(\Delta(y',y) + \langle v,\Psi(x,y')\rangle - \langle v,\Psi(x,y)\rangle) \\
&= \lambda\ell^{hinge}(u,(x,y)) + (1-\lambda)\ell^{hinge}(v,(x,y))
\end{aligned}
$$

Thus,

$$
\begin{aligned}
L_S^{hinge}(\lambda u + (1-\lambda)v) &= \frac{1}{|S|}\sum_{(x,y)\in S}\ell^{hinge}(\lambda u + (1-\lambda)v,(x,y)) \\
&\leq \lambda\frac{1}{|S|}\sum_{(x,y)\in S}\ell^{hinge}(u) + (1-\lambda)\frac{1}{|S|}\sum_{(x,y)\in S}\ell^{hinge}(v) \\
&= \lambda L_S^{hinge}(u) + (1-\lambda)L_S^{hinge}(v)
\end{aligned}
$$

(ii) Notice that for all $y \in \mathcal{Y}$, by the definition of $h_w(x)$, we have:

$$\begin{aligned} \langle w, \Psi(x,y)\rangle &\le \langle w, \Psi(x, h_w(x))\rangle \\ \implies 0 &\le \langle w, \Psi(x, h_w(x))\rangle - \langle w, \Psi(x,y)\rangle \end{aligned}$$

Therefore, we have:

$$\begin{aligned} \ell^{\Delta}(h_w,(x,y)) &\le \Delta(h_w(x),y) + \langle w, \Psi(x,h_w(x))\rangle - \langle w, \Psi(x,y)\rangle \\ &\le \max_{y'\in\mathcal{Y}} (\Delta(y',y) + \langle w, \Psi(x,y')\rangle - \langle w, \Psi(x,y)\rangle) \end{aligned}$$

And so

$$L_S^{\Delta}(w) \le L_S^{hinge}(w)$$

(iii) Let w such that $L_S^{\Delta}(h_w) = 0$, then for all $(x,y) \in S$, we have $\Delta(h_w(x),y) = 0$. Therefore, taking the same w, we have, for $(x,y) \in S$ and $\Delta(y',y) > 0$:

$$\langle w, \Psi(x,h_w(x))\rangle > \langle w, \Psi(x,y')\rangle$$

We want an inequality as follows:

$$\langle \tilde{w}, \Psi(x,h_{\tilde{w}}(x))\rangle \ge \langle \tilde{w}, \Psi(x,y')\rangle + \Delta(y',y) \qquad \forall \Delta(y',y) > 0 \tag{1}$$

This would imply, for $\Delta(h_{\tilde{w}}(x),y) = 0$:

$$\begin{aligned} \ell^{hinge}(\tilde{w},(x,y)) = &\max_{\Delta(y',y)>0} [\Delta(y',y) + \langle \tilde{w}, \Psi(x,y')\rangle - \langle \tilde{w}, \Psi(x,y)\rangle] \quad (2)\\ &+ \max_{\Delta(y'',y)=0} [\langle \tilde{w}, \Psi(x,y'')\rangle - \langle \tilde{w}, \Psi(x,y)\rangle] \quad (3) \end{aligned}$$

We focus on the first term of the right hand side.

$$\begin{aligned} (2) &\le \max_{\Delta(y',y)>0} [\langle \tilde{w}, \Psi(x,h_{\tilde{w}}(x))\rangle - \langle \tilde{w}, \Psi(x,y)\rangle] \\ &= \langle \tilde{w}, \Psi(x,h_{\tilde{w}}(x))\rangle - \langle \tilde{w}, \Psi(x,y)\rangle \\ \Delta(h_{\tilde{w}}(x),y) = 0 &\implies h_{\tilde{w}}(x) = y \\ \therefore \quad (2) &\le 0 \end{aligned}$$

Now for the second term

$$\begin{aligned} (3) &\le \langle \tilde{w}, \Psi(x,h_{\tilde{w}}(x))\rangle - \langle \tilde{w}, \Psi(x,y)\rangle \\ &= 0 \end{aligned}$$

Therefore, if (1) holds, we have:

$$\ell^{hinge}(\tilde{w},(x,y)) = 0$$

To make (1) hold, we leverage the fact that $\mathcal{Y}$ is finite. Let

$$d = \min_{\Delta(y',y)>0} (\langle w, \Psi(x,h_w(x))\rangle - \langle w, \Psi(x,y')\rangle)$$

and

$$D = \max_{\Delta(y',y)>0} \Delta(y',y)$$

Thus, we just let

$$\tilde{w} = \frac{D+1}{d} w$$

Then we have:

$$\begin{aligned} \langle \tilde{w}, \Psi(x,h_{\tilde{w}}(x))\rangle - \langle \tilde{w}, \Psi(x,y')\rangle &= \frac{D+1}{d}\left(\langle w, \Psi(x,h_w(x))\rangle - \langle w, \Psi(x,y')\rangle\right) \\ &\ge D+1 > D = \Delta(y',y) \end{aligned}$$

And so (1) holds. Therefore, we have $L_S^{\Delta}(w) = 0 \implies L_S^{hinge}(w) = 0$.

(c) Fix a sample (x, y) and let

$$g_{y'}(w) = \Delta(y', y) + \langle w, \Psi(x, y')\rangle - \langle w, \Psi(x, y)\rangle$$
$$g(w) = \max_{y' \in \mathcal{Y}} g_{y'}(w)$$

Then

$$\ell^{hinge}(w, (x, y)) = g(w)$$

Let $\nabla_w g_{y'}(w)$ be a sub-gradient of $g_{y'}$ at w. Let

$$y_0 = \arg\max_{y' \in \mathcal{Y}} g_{y'}(w)$$

Then we have, for $v \in \mathbb{R}^d$:

$$\begin{aligned} g(v) - g(w) &= g(v) - g_{y_0}(w) \\ &\geq g_{y_0}(v) - g_{y_0}(w) \\ &\geq \nabla_w g_{y_0}(w)^\top (v - w) \end{aligned}$$

So $\nabla_w g_{y_0}(w)$ is a sub-gradient of $g(w)$ at w. Since $g_{y'}$ is differentiable, its gradient is a subgradient, so we have:

$$\begin{aligned} \nabla_w g_{y'}(w) &= \nabla_w \Delta(y', y) + \nabla_w \langle w, \Psi(x, y')\rangle - \nabla_w \langle w, \Psi(x, y)\rangle \\ &= \Psi(x, y') - \Psi(x, y) \end{aligned}$$

Therefore, we have:

$$\nabla_w \ell^{hinge}(w, (x, y)) = \nabla_w g_{y_0}(w) = \Psi(x, y_0) - \Psi(x, y)$$

In order to compute the subgradient, we need to determine what y_0 is. Without any assumptions on Δ or $\mathcal{Y}$, we just have to iterate over $\mathcal{Y}$. Assuming that computing $\Psi(x, y)$ takes constant time, we have that computing $g_{y'}(w)$ takes $O(d)$ operations. Therefore, computing the subgradient takes $O(d|\mathcal{Y}|)$ operations.