Assignment 1

Arvid Frydenlund

September 1, 2015

1 Q1

softmax($\mathbf{x}$) = softmax($\mathbf{x} + c$) (Note, remember softmax is defined as a function taking in a vector)

Considering element i

$$\frac{e^{x_i}}{\sum_j e^{x_j}} = \frac{e^{x_i+c}}{\sum_j e^{x_j+c}} \tag{1}$$

since,

$$\frac{e^{x_i} e^c}{\sum_j e^{x_j} e^c} = \frac{e^{x_i} \cancel{e^c}}{\sum_j e^{x_j} \cancel{e^c}} \tag{2}$$

QED

This is useful for the trick where $c = -\max_{\mathrm{i}}^{\mathrm{n}} \mathbf{x}_{\mathrm{i}}$, since the largest number in the exponential possible is now 0. That is it prevents overflow.

2 Q2 a)

The sigmoid function is

$$y = \frac{1}{1+e^{-x}} = \sigma(x) = (1+e^{-x})^{-1} = (z)^{-1} \tag{3}$$

Where $z = 1 + e^{-x} = 1 + e^r$, and $r = -x$

Then by simple chain rule

$$\frac{dy}{dx} = \frac{dy}{dz} * \frac{dz}{dr} * \frac{dr}{dx} \tag{4}$$

$$\frac{dy}{dx} = -(1+e^{-x})^{-2} * e^{-x} * (-1) = e^{-x} * \left(\frac{1}{1+e^{-x}}\right)^2 = e^{-x}\sigma(x)\sigma(x) \tag{5}$$

Since $e^{-x} = \frac{1-\sigma(x)}{\sigma(x)}$
because

$$\frac{1 - \frac{1}{1+e^{-x}}}{\frac{1}{1+e^{-x}}} = (1 + e^{-x}) - \frac{1 + e^{-x}}{1 + e^{-x}} = (1 + e^{-x}) - 1 = e^{-x} \tag{6}$$

Then

$$e^{-x}\sigma(x)\sigma(x) = \frac{1 - \sigma(x)}{\cancel{\sigma(x)}}\cancel{\sigma(x)}\sigma(x) = (1 - \sigma(x))(\sigma(x)) \tag{7}$$

QED

3 Q2 b)

Asume K classes, $\hat{y} = \text{softmax}(\theta) = [\hat{y}_1, ..., \hat{y}_K]$ =[softmax(θ_1), ...,softmax(θ_K)] $= g(\theta_i) = \frac{e^{\theta_i}}{\sum_c e^{\theta_c}}$

$$J = CE(y, \hat{y}) = -\sum_i^K y_i \log(\hat{y}_i) \tag{8}$$

Note that the sum is over all the classes and that y is 1-hot so that the sum only picks out one value

$$\nabla_\theta J = -\sum_i^K \frac{\partial}{\partial\theta} y_i \log(\hat{y}_i) = -\sum_i^K y_i \frac{\partial}{\partial\theta} \log(g(\theta_i)) = -\sum_i^K y_i \frac{1}{g(\theta_i)} \frac{\partial}{\partial\theta} g(\theta_i) \tag{9}$$

Considering $\frac{\partial}{\partial\theta} g(\theta_i)$, for an element j of θ, θ_j there are two cases.
Case 1) $j = i$ then note

$$\frac{\partial}{\partial\theta_j} g(\theta_i) = \frac{\partial}{\partial\theta_j} g(\theta_j) = \frac{\partial}{\partial\theta_i} g(\theta_i) \tag{10}$$

$$\frac{dg(\theta_i)}{d\theta_j} = \frac{d}{d\theta_j} \frac{e^{\theta_i}}{\sum_c e^{\theta_c}} = \frac{(\frac{d}{d\theta_j} e^{\theta_i})(\sum_c e^{\theta_c}) - (\frac{d}{d\theta_j} \sum_c e^{\theta_c})(e^{\theta_i})}{(\sum_c e^{\theta_c})^2} \tag{11}$$

Where $(\frac{d}{d\theta_j} \sum_c e^{\theta_c})(e^{\theta_i}) = \frac{d}{d\theta_j} e^{\theta_1} + ... + e^{\theta_j} + ... + e^{\theta_K} = e^{\theta_j} = e^{\theta_i}$
So

$$= \frac{(e^{\theta_i})(\sum_c e^{\theta_c}) - (e^{\theta_i})(e^{\theta_i})}{(\sum_c e^{\theta_c})(\sum_c e^{\theta_c})} = \frac{(e^{\theta_i})}{(\sum_c e^{\theta_c})} \frac{(\sum_c e^{\theta_c} - e^{\theta_i})}{(\sum_c e^{\theta_c})} = g(\theta_i)\left(\frac{\sum_c e^{\theta_c}}{\sum_c e^{\theta_c}} - \frac{e^{\theta_i}}{\sum_c e^{\theta_c}}\right) \tag{12}$$

$$= g(\theta_i)(1 - g(\theta_i)) = g(\theta_i)(1 - g(\theta_j)) \tag{13}$$

Now in the case there $j \neq i$

$$\frac{dg(\theta_i)}{d\theta_j} = \frac{d}{d\theta_j}\frac{e^{\theta_i}}{\sum_c e^{\theta_c}} = \frac{d}{d\theta_j}e^{\theta_i}\left(\sum_c e^{\theta_c}\right)^{-1} = -e^{\theta_i}\left(\sum_c e^{\theta_c}\right)^{-2} e^{\theta_j} \tag{14}$$

$$= \frac{-(e^{\theta_i})}{(\sum_c e^{\theta_c})}\frac{(e^{\theta_j})}{(\sum_c e^{\theta_c})} = -g(\theta_i)g(\theta_j) \tag{15}$$

Let $t_{ij} = 1$ if $i = j$ and $= 0$, if $i \neq j$, then we can combine both cases as

$$\frac{dg(\theta_i)}{d\theta_j} = g(\theta_i)(t_{ij} - g(\theta_i)) \tag{16}$$

Which can be confirmed with $i = j$, $g(\theta_i)(1 - g(\theta_i)) = g(\theta_i)(1 - g(\theta_j))$ and with $j \neq i$, $g(\theta_i)(0 - g(\theta_j)) = g(\theta_i)(-g(\theta_j)) = -g(\theta_i)(g(\theta_j))$

Now putting it all together we get

$$\nabla_\theta J = -\sum_i^K y_i \frac{1}{\cancel{g(\theta_i)}}\cancel{g(\theta_i)}(t_{ij} - g(\theta_i)) \tag{17}$$

Let $t = y = yt$, and since y is 1-hot, this gets vectorized as

$$= -y - g(\theta) = -y - \hat{y} = \hat{y} - y \tag{18}$$

4 Q2c)

Note I used 'Practical Guide to Matrix Calculus for Deep Learning' by Andrew Delong [HERE] for this question, which was very helpful in giving identities for matrix calculus. I am using the Hadamard (element-wise) product as $\odot$

Remember

$$\frac{\partial x^T a}{x} = \frac{\partial a^T x}{x} = a \tag{19}$$

Some definitions:

$\hat{y} = g(\theta) = g(hW_2 + b_2)$, which has dims $\mathbb{R}^{1*K}$ (same as y)

$\theta = hW_2 + b_2$, which has dims $\mathbb{R}^{1*K}$

$h = \sigma(z) = \sigma(xW_1 + b_1)$, which has dims $\mathbb{R}^{1*D_h}$ (same as the derivative $\sigma'(z)$)

$z = xW_1 + b_1$, which has dims $\mathbb{R}^{1*D_h}$

and $W_2 \in \mathbb{R}^{D_h*K}, W_1 \in \mathbb{R}^{D_x*D_h}, x \in \mathbb{R}^{1*D_x}$

Let $\triangle_2 = \hat{y} - y \in \mathbb{R}^{1*K}$ be the last error message

$$\nabla_x J = (\hat{y} - y)\bigg(\frac{d\theta}{dx}\bigg) = \bigg(\triangle_2\bigg(\frac{d\theta}{dh}\bigg)\bigg) \odot \bigg(\frac{dh}{dx}\bigg) \tag{20}$$

$$= \bigg(\triangle_2\bigg(\frac{d}{dh}hW_2 + b_2\bigg)\bigg) \odot \bigg(\frac{dh}{dx}\bigg) = (\triangle_2 W_2^T) \odot \bigg(\frac{dh}{dx}\bigg) = \bigg((\triangle_2 W_2^T) \odot \bigg(\frac{dh}{dz}\bigg)\bigg)\frac{dz}{dx} \tag{21}$$

$$= \bigg((\triangle_2 W_2^T) \odot (\sigma'(xW_1 + b_1))\bigg)W_1^T \tag{22}$$

Let $\triangle_1 = (\triangle_2 W_2^T) \odot (\sigma'(xW_1 + b_1)) = (\sigma'(xW_1 + b_1)) \odot (\triangle_2 W_2^T) \in \mathbb{R}^{1*D_h}$ be the hidden error message

Then $\triangle_1 W_1^T \in \mathbb{R}^{1*D_x}$ which is the size we were hoping for.

5 Q2d)

There are the same number of parameters as there are weights and biases which is the size of W_1, W_2, b_1, and b_2 which is $(D_x * D_h) + (D_h * K) + (1 * D_h) + (1 * K)$

6 Q3a)

From Slide 22 of Lecture 2, the global objective is

$$J(\theta) = \frac{1}{T}\sum_{t}^{T} \sum_{-c \le j \le c, j \ne 0} -\log(P(w_{t+j}|w_t)) \tag{23}$$

For a specific outer word w_i and an inner word $\hat{r}$

$$J_i = -\log(P(\text{word}_i|\hat{r}, W)) = -\log\left(\frac{e^{w_i^T\hat{r}}}{\sum_j^{|V|} e^{w_j^T\hat{r}}}\right) = -\bigg(w_i^T\hat{r} - log\bigg(\sum_j^{|V|} e^{w_j^T\hat{r}}\bigg)\bigg) \tag{24}$$

$$= -w_i^T\hat{r} + log\bigg(\sum_j^{|V|} e^{w_j^T\hat{r}}\bigg) \tag{25}$$

$$\nabla_{\hat{r}} J_i = \frac{\partial}{\partial\hat{r}}w_i^T\hat{r} + \frac{\partial}{\partial\hat{r}}log\bigg(\sum_j^{|V|} e^{w_j^T\hat{r}}\bigg) = -w_i + \frac{1}{\sum_j^{|V|} e^{w_j^T\hat{r}}}\bigg(\frac{\partial}{\partial\hat{r}}\sum_k^{|V|} e^{w_k^T\hat{r}}\bigg) \tag{26}$$

$$= -w_i + \frac{1}{\sum_j^{|V|} e^{w_j^T\hat{r}}}\bigg(\sum_k^{|V|} e^{w_k^T\hat{r}}\bigg)w_k^T = -w_i + \sum_k^{|V|}\frac{(e^{w_k^T\hat{r}})w_k^T}{\sum_j^{|V|} e^{w_j^T\hat{r}}} = -w_i + \sum_k^{|V|} P(\text{word}_k|\hat{r}, W))w_k^T \tag{27}$$

7 Q3b)

Just to write down the dimensions of all the variables:

$\hat{r}, w_i, \frac{\partial J_i}{\hat{r}}, \frac{\partial J_i}{w_j} \in \mathbb{R}^{1*D}$ and $w_k^T \hat{r} \in \mathbb{R}$, $\sigma(w_k^T \hat{r}) w_k, \sigma(w_k^T \hat{r}) \hat{r} \in \mathbb{R}^{1*D}$

There are two cases.

Case 1) $j = i$

$$\nabla_{w_j} J_i = -\hat{r} + \frac{1}{\sum_k^{|V|} e^{w_k^T \hat{r}}} \frac{\partial}{\partial w_j} \sum_l^{|V|} e^{w_l^T \hat{r}} \tag{28}$$

But the last partial is only not zero when $l = j$ so

$$\nabla_{w_j} J_i = -\hat{r} + \frac{e^{w_j^T \hat{r}} \hat{r}}{\sum_k^{|V|} e^{w_k^T \hat{r}}} \tag{29}$$

Case 2) $j \neq i$

$$\nabla_{w_j} J_i = -(0) + \frac{e^{w_j^T \hat{r}} \hat{r}}{\sum_k^{|V|} e^{w_k^T \hat{r}}} = \frac{e^{w_j^T \hat{r}} \hat{r}}{\sum_k^{|V|} e^{w_k^T \hat{r}}} \tag{30}$$

The two cases can be combined with t_{ij}, where t_{ij} is 1 when $i = j$ and 0 otherwise

$$\nabla_{w_j} J_i = -\hat{r}(t_{ij}) + \frac{e^{w_j^T \hat{r}} \hat{r}}{\sum_k^{|V|} e^{w_k^T \hat{r}}} \tag{31}$$

8 Q3c)a)

$$J_i(\hat{r}, w_i, w_1, ..., w_K) = -log(\sigma(w_i^T \hat{r})) - \sum_k^K \log(\sigma(-w_k^T \hat{r})) = z_1 + z_2 \tag{32}$$

$$\nabla_{\hat{r}} z_1 = \frac{-1}{\sigma(w_i^T \hat{r})} \sigma'(w_i^T \hat{r}) w_i = \frac{\cancel{-\sigma(w_i^T \hat{r})}(1 - \sigma(w_i^T \hat{r})) w_i}{\cancel{\sigma(w_i^T \hat{r})}} = (\sigma(w_i^T \hat{r}) - 1) w_i \tag{33}$$

$$\nabla_{\hat{r}} z_2 = -\sum_k^K \frac{1}{\sigma(-w_k^T \hat{r})} \sigma'(-w_k^T \hat{r})(-w_k) = \sum_k^K \frac{\cancel{\sigma(-w_k^T \hat{r})}(1 - \sigma(-w_k^T \hat{r})) w_k}{\cancel{\sigma(-w_k^T \hat{r})}} = \sum_k^K (1 - \sigma(-w_k^T \hat{r})) w_k \tag{34}$$

$$\nabla_{\hat{r}} J_i = (\sigma(w_i^T \hat{r}) - 1) w_i + \sum_k^K (1 - \sigma(-w_k^T \hat{r})) w_k \tag{35}$$

9 Q3c)b)

$$J_i(\hat{r}, w_i, w_1, ..., w_K) = -log(\sigma(w_i^T \hat{r})) - \sum_k^K \log(\sigma(w_k^T \hat{r})) = z_1 + z_2 \quad (36)$$

Case 1) $j = i$
Since $i \notin K$, then $\nabla_{\hat{r}} z_2 = 0$

$$\nabla_{\hat{r}} z_1 = \nabla_{\hat{r}} J_i = \frac{-1}{\sigma(w_i^T \hat{r})} \sigma'(w_i^T \hat{r})\hat{r} = \frac{-\cancel{\sigma(w_i^T \hat{r})}(1 - \sigma(w_i^T \hat{r}))\hat{r}}{\cancel{\sigma(w_i^T \hat{r})}} = (\sigma(w_i^T \hat{r}) - 1)\hat{r} \quad (37)$$

Case 2) $j \neq i$
$\nabla_{w_j} z_1 = 0$
Then for $\nabla_{w_j} z_1$ there are two cases, if $j \neq k$, then $\nabla_{w_j} z_2 = 0$.
If $j = k$,

$$\nabla_{w_j} z_2 = \frac{-1}{\sigma(-w_k^T \hat{r})} \sigma'(-w_k^T \hat{r})(-\hat{r}) = \frac{\cancel{\sigma(-w_k^T \hat{r})}(1 - \sigma(-w_k^T \hat{r}))\hat{r}}{\cancel{\sigma(-w_k^T \hat{r})}} = (1 - \sigma(-w_k^T \hat{r}))\hat{r} \quad (38)$$

A side note that $1 - \sigma(-x) = \sigma(x)$. Proof:

$$1 - \frac{1}{1 + e^{-(-x)}} = \frac{1 + e^x}{1 + e^x} - \frac{1}{1 + e^x} = \frac{1 - 1 + e^x}{1 + e^x} = \frac{e^x}{1 + e^x} \quad (39)$$

$$= \frac{1(e^x)}{(1/e^x + 1)(e^x)} = \frac{1}{(1/e^x + 1)} = \frac{1}{(e^{-x} + 1)} = \frac{1}{1 + e^{-x}} \quad (40)$$

The two cases can then be combined with t_{ij}, where t_{ij} is 1 when $i = j$ and 0 otherwise

$$\nabla_{w_j} J_i = \nabla_{w_j} z_1 + \nabla_{w_j} z_2 = (\sigma(w_k^T \hat{r}) - t_{ij})\hat{r} \quad (41)$$

Which works for if $j = k$ (i.e. only update the gradient for words which are in the negative samples and not all words cause that would be a lot unchanging updates)

10 Q3d)

$$J(\theta_{w_i}) = \sum_{-c \leq j \leq c, j \neq 0} - \log(P(v'_{w_{i+j}} | v_{w_i})) = \sum_{-c \leq j \leq c, j \neq 0} F(v'_{w_{i+j}} | v_{w_i}) \quad (42)$$

Where F is either softmax-CE or negative sampling

$$\nabla J = \sum_{-c \leq j \leq c, j \neq 0} F'(v'_{w_{i+j}} | v_{w_i}) \quad (43)$$

Which is easily found given the above parts for both $\nabla_{v'_{w_{i+j}}} J$ and $\nabla_{v_{w_i}} J$