Homework 3 - Theory Part

Reza Bayat and Arian Khorasani

November 2021

1 Question 1

(a)

$$g(x) = max\{0, x\} = \begin{cases} x & x \geq 0 \\ 0 & x < 0 \end{cases}$$

$$g'(x) = \mathbb{1}_{x>0}$$

(b)

$$\begin{aligned}
\sigma(x) &= \frac{1}{1+\exp(-x)} \\
\frac{d\sigma(x)}{dx} &= \frac{\frac{d1}{dx} \times (1+\exp(-x)) - \frac{d(1+\exp(-x))}{dx} \times 1}{(1+\exp(-x))^2} \\
&= \frac{0 \times (1+\exp(-x)) - (-\exp(-x)}{(1+\exp(-x))^2} \\
&= \left(\frac{1}{1+\exp(-x)}\right)\left(\frac{\exp(-x)}{1+\exp(-x)}\right) \\
&= \left(\frac{1}{1+\exp(-x)}\right)\left(1 - \frac{1}{1+\exp(-x)}\right) \\
&= \sigma(x) \cdot (1-\sigma(x))
\end{aligned}$$

(c)

$$\sigma(x) = \frac{1}{1+\exp(-x)} = \frac{e^x}{e^x+1} \quad , \quad \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

$$\begin{aligned}
\tanh\left(\frac{1}{2}x\right) + 1 &= \frac{e^{\frac{1}{2}x} - e^{-\frac{1}{2}x}}{e^{\frac{1}{2}x} + e^{-\frac{1}{2}x}} + 1 \\
&= \frac{e^x - 1}{e^x + 1} + 1 \\
&= \frac{e^x - 1}{e^x + 1} + \frac{e^x + 1}{e^x + 1} \\
&= \frac{2e^x}{e^x + 1} \\
\text{So } \sigma(x) &= \frac{1}{2}\left(\tanh\left(\frac{1}{2}x\right) + 1\right)
\end{aligned}$$

(d)

$$\begin{aligned}
\sigma(x) &= \frac{1}{1+\exp(-x)} \\
\ln \sigma(x) &= -\ln\left(1+e^{-x}\right) \\
\text{softplus }(x) &= \ln\left(1+e^{x}\right)
\end{aligned}$$

Then:

$$\ln \sigma(x) = -\operatorname{softplus}(-x)$$

(e)

$$\begin{aligned}
\operatorname{softplus}(x) - \operatorname{softplus}(-x) &= \ln\left(1+e^{x}\right) - \ln\left(1+e^{-x}\right) \\
&= \frac{\ln\left(1+e^{x}\right)}{\ln\left(1+e^{-x}\right)} \\
&= \frac{\ln e^{x}\left(1+e^{-x}\right)}{\ln\left(1+e^{-x}\right)} \\
&= \ln e^{x} \\
&= x
\end{aligned}$$

(f)

$$\operatorname{sign}(x) = \mathbf{1}_{x>0}(x) - \mathbf{1}_{x<0}(x)$$

(g)

$$\begin{aligned}
\frac{\partial \|\mathbf{x}\|_2^2}{\partial \mathbf{x}_i} &= \frac{\partial \sum_i \mathbf{x}_i^2}{\partial \mathbf{x}_i} = 2x_i \\
\frac{\partial \|\mathbf{x}\|_2^2}{\partial \mathbf{x}} &= \begin{bmatrix} 2x_1 \\ 2x_2 \\ \vdots \\ 2x_n \end{bmatrix} = 2\mathbf{x}
\end{aligned}$$

(h)

$$\begin{aligned}
\frac{\partial \|\mathbf{x}\|_1}{\partial \mathbf{x}_i} &= \operatorname{sign}\left(x_i\right) \\
\frac{\partial \|\mathbf{x}\|_1}{\partial \mathbf{x}} &= \begin{bmatrix} \operatorname{sign}\left(x_1\right) \\ \operatorname{sign}\left(x_2\right) \\ \vdots \\ \operatorname{sign}\left(x_n\right) \end{bmatrix}
\end{aligned}$$

(i)

$$\begin{aligned}
S(\mathbf{x})_i &= \frac{e^{x_i}}{\sum_j e^{\mathbf{x}_j}} \\
S(\mathbf{cx})_i &= \frac{e^{cx_i}}{\sum_j e^{\mathbf{cx}_j}} \\
&= \frac{(e^{x_i})^c}{\sum_j (e^{\mathbf{x}_j})^c}
\end{aligned}$$

$$
\begin{aligned}
\ln(S(\mathbf{cx})_i) &= \ln(\frac{(e^{x_i})^c}{\sum_j (e^{\mathbf{x}_j})^c}) \\
&= \frac{c\ln(e^{x_i})}{c\ln\sum_j (e^{\mathbf{x}_j})} \\
&= \frac{\ln(e^{x_i})}{\ln\sum_j (e^{\mathbf{x}_j})}
\end{aligned}
$$

$$
S(\mathbf{cx})_i = \frac{e^{x_i}}{\sum_j e^{\mathbf{x}_j}}
$$

(j)

$$
\begin{aligned}
S(\mathbf{x}+\mathbf{c})_i &= \frac{e^{(x_i+c)}}{\sum_j e^{(\mathbf{x}_j+c)}} \\
&= \frac{e^c e^{x_i}}{\sum_j e^c e^{\mathbf{x}_j}} \\
&= \frac{e^c e^{x_i}}{e^c \sum_j e^{\mathbf{x}_j}} \\
&= \frac{e^{x_i}}{\sum_j e^{\mathbf{x}_j}}
\end{aligned}
$$

(k)

$$
\frac{\partial S(\boldsymbol{x})}{\partial \boldsymbol{x}_j} = \begin{bmatrix} \frac{\partial S(\boldsymbol{x})_1}{\partial \boldsymbol{x}_j} & \frac{\partial S(\boldsymbol{x})_2}{\partial \boldsymbol{x}_j} & \dots & \frac{\partial S(\boldsymbol{x})_n}{\partial \boldsymbol{x}_j} \end{bmatrix}
$$

So when $i = j$:

$$
\frac{\partial S(\boldsymbol{x})_i}{\partial \boldsymbol{x}_j} = \frac{\partial}{\partial \boldsymbol{x}_j}\frac{e^{x_i}}{\sum e^{x_k}} = \frac{e^{x_i}\sum_k e^{x_k} - e^{x_i}e^{x_i}}{(\sum_k e^{x_k})^2} = S(\boldsymbol{x})_i - S^2(\boldsymbol{x})_i = S(\boldsymbol{x})_i - S(\boldsymbol{x})_i S(\boldsymbol{x})_j
$$

Otherwise:

$$
\frac{\partial S(\boldsymbol{x})_i}{\partial \boldsymbol{x}_j} = \frac{\partial}{\partial \boldsymbol{x}_j}\frac{e^{x_i}}{\sum_k e^{\boldsymbol{x}_k}} = \frac{0 - e^{\boldsymbol{x}_i}e^{\boldsymbol{x}_j}}{(\sum_k e^{\boldsymbol{x}_k})^2} = \frac{-e^{\boldsymbol{x}_i}e^{\boldsymbol{x}_j}}{(\sum_k e^{\boldsymbol{x}_k})^2} = -S(\boldsymbol{x})_i S(\boldsymbol{x})_j
$$

We can combine the above results using an indicator function:

$$
\frac{\partial S(\boldsymbol{x})_i}{\partial \boldsymbol{x}_j} = S(\boldsymbol{x})_i \mathbf{1}_{i=j} - S(\boldsymbol{x})_i S(\boldsymbol{x})_j
$$

(l) We know:

$$
\frac{\partial S(\boldsymbol{x})_i}{\partial \boldsymbol{x}_j} = S(\boldsymbol{x})_i \mathbf{1}_{i=j} - S(\boldsymbol{x})_i S(\boldsymbol{x})_j
$$

For $i = j$ the diagonal elements of Jacobian matrix:

$$
\frac{\partial S(\boldsymbol{x})_i}{\partial \boldsymbol{x}_j} = S(\boldsymbol{x})_i \mathbf{1}_{i=j} - S(\boldsymbol{x})_i S(\boldsymbol{x})_j
$$

And of other elements:

$$\left(\frac{\partial S(\boldsymbol{x})}{\partial \boldsymbol{x}}\right)_{i \neq j} = -S(\boldsymbol{x})_i S(\boldsymbol{x})_j$$

So using above equations:

$$\frac{\partial S(\boldsymbol{x})}{\partial \boldsymbol{x}} = \operatorname{diag}(S(\boldsymbol{x})) - S(\boldsymbol{x})S(\boldsymbol{x})^{\top}$$

$$\frac{\partial S(\boldsymbol{x})}{\partial \boldsymbol{x}} = \operatorname{diag}(S(\boldsymbol{x})) - \begin{bmatrix} S(\boldsymbol{x})_1 \\ S(\boldsymbol{x})_2 \\ \vdots \\ S(\boldsymbol{x})_n \end{bmatrix} \begin{bmatrix} S(\boldsymbol{x})_1 & S(\boldsymbol{x})_2 & \dots & S(\boldsymbol{x})_n \end{bmatrix}$$

(m)

$$\boldsymbol{y} = \sigma(\boldsymbol{x}) = \begin{bmatrix} \sigma(\boldsymbol{x}_1) \\ \sigma(\boldsymbol{x}_2) \\ \vdots \\ \sigma(\boldsymbol{x}_n) \end{bmatrix}$$

The Jacobian of $y = f(x) = \sigma(x)$ is defined below and it's diagonal.

$$\left(\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right) = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \sigma(\boldsymbol{x}_1)(1-\sigma(\boldsymbol{x}_1)) & & \\ & \ddots & \\ & & \sigma(\boldsymbol{x}_n)(1-\sigma(\boldsymbol{x}_n)) \end{bmatrix}$$

Then using following steps, we can show it has $O(n)$ time complexity.

$$\begin{aligned} \nabla_{\boldsymbol{x}} L = \left(\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right)^{\top} \nabla_{\boldsymbol{y}} L = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_n} \end{bmatrix} &= \begin{bmatrix} \sigma(\boldsymbol{x}_1)(1-\sigma(\boldsymbol{x}_1)) & & \\ & \ddots & \\ & & \sigma(\boldsymbol{x}_n)(1-\sigma(\boldsymbol{x}_n)) \end{bmatrix} \begin{bmatrix} (\nabla_{\boldsymbol{y}} L)_1 \\ (\nabla_{\boldsymbol{y}} L)_2 \\ \vdots \\ (\nabla_{\boldsymbol{y}} L)_n \end{bmatrix} \\ &= \begin{bmatrix} \sigma(\boldsymbol{x}_1)(1-\sigma(\boldsymbol{x}_1))(\nabla_{\boldsymbol{y}} L)_1 \\ \sigma(\boldsymbol{x}_2)(1-\sigma(\boldsymbol{x}_2))(\nabla_{\boldsymbol{y}} L)_2 \\ \vdots \\ \sigma(\boldsymbol{x}_n)(1-\sigma(\boldsymbol{x}_n))(\nabla_{\boldsymbol{y}} L)_n \end{bmatrix} \end{aligned}$$

And when $y = f(x) = S(x)$

$$\begin{aligned}
\nabla_{\boldsymbol{x}} L = \left(\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right)^{\top} \nabla_{\boldsymbol{y}} L &= \left(\operatorname{diag}(S(\boldsymbol{x})) - S(\boldsymbol{x})S(\boldsymbol{x})^{\top}\right)^{\top} \nabla_{\boldsymbol{y}} L \\
&= \left(\operatorname{diag}(S(\boldsymbol{x})) - S(\boldsymbol{x})S(\boldsymbol{x})^{\top}\right) \nabla_{\boldsymbol{y}} L \\
&= \operatorname{diag}(S(\boldsymbol{x}))\nabla_{\boldsymbol{y}} L - S(\boldsymbol{x})S(\boldsymbol{x})^{\top}\nabla_{\boldsymbol{y}} L \\
&= \operatorname{diag}(S(\boldsymbol{x}))\nabla_{\boldsymbol{y}} L - S(\boldsymbol{x})\left(S(\boldsymbol{x})^{\top}\nabla_{\boldsymbol{y}} L\right)
\end{aligned}$$

So, all the following parts have $O(n)$ time complexity, and in total, the time complexity is $O(n)$.

$$\operatorname{diag}(S(\boldsymbol{x}))\nabla_{\boldsymbol{y}} L$$

$$S(\boldsymbol{x})^{\top}\nabla_{y} L$$

$$S(\boldsymbol{x})\left(S(\boldsymbol{x})^{\top}\nabla_{\boldsymbol{y}} L\right)$$

2 Gradient computation for parameters optimization in a neural net for multiclass classification

2.a

The dimension of $\mathbf{b}^{(1)}$ is: $\mathbf{b}^{(\mathbf{1})} \in \mathbb{R}^{d_h}$.

$$\mathbf{h}^a = \mathbf{W}^{(1)T} \cdot \mathbf{x} + \mathbf{b}^{(1)} \tag{1}$$

$$\mathbf{h}^a_j = \mathbf{W}^{(1)T}_j \cdot \mathbf{x} + b^{(1)}_j \tag{2}$$

$$\mathbf{h}^s = g(\mathbf{h}^a) \tag{3}$$

Where $g(x)$ is the activation function (i.e., ReLU nonlinearity) applied element wise to the hidden layer. $g(x) = max(0, x)$.

2.b

The dimensions of $\mathbf{W}^{(2)}$ and $\mathbf{b}^{(2)}$ are: $\mathbf{W}^{(2)} \in \mathbb{R}^{m \times d_h}$, and $\mathbf{b}^{(2)} \in \mathbb{R}^m$.

$$\mathbf{o}^a = \mathbf{W}^{(2)T} \cdot \mathbf{h^s} + \mathbf{b}^{(2)} \tag{4}$$

$$\mathbf{o}^a_k = \mathbf{W}^{(2)T}_k \cdot \mathbf{h^s} + b^{(2)}_k \tag{5}$$

2.c

Let's define an m-class softmax: $softmax(x) = \frac{e^{x_i}}{\sum_{i=1}^m e^{x_i}}$. Therefore,

$$\mathbf{o}^s_k = \frac{e^{\mathbf{o}^a_k}}{\sum_{i=1}^m e^{\mathbf{o}^a_k}} \tag{6}$$

$$\begin{aligned}
\sum_{i=1}^m \mathbf{o}^s_k &= \sum_{i=1}^m \frac{e^{\mathbf{o}^a_k}}{\sum_{i=1}^m e^{\mathbf{o}^a_k}} && (7)\\
&= \frac{\sum_{i=1}^m e^{\mathbf{o}^a_k}}{\sum_{i=1}^m e^{\mathbf{o}^a_k}} && (8)\\
&= 1. && (9)
\end{aligned}$$

Since $e^x > 0$, the result will always be positive. This is crucial because we need the softmax to produce a probability distribution over our m output classes.

2.d

$$L(\mathbf{o}^a, y) = -\sum_{k=1}^{M} y_k \log\left(\frac{e^{\mathbf{o}_k^a}}{\sum_{i=1}^{m} e^{\mathbf{o}_k^a}}\right) \tag{10}$$

2.e

The set of parameters is:

$$\theta = \{\mathbf{W}^{(1)}, \mathbf{b}^{(1)}, \mathbf{W}^{(2)}, \mathbf{W}^{(2)}\} \tag{11}$$

Since $\mathbf{W}^{(1)} \in \mathbb{R}^{d_h \times d}$, $\mathbf{b}^{(1)} \in \mathbb{R}^{d_h}$, $\mathbf{W}^{(2)} \in \mathbb{R}^{m \times d_h}$, and $\mathbf{b}^{(2)} \in \mathbb{R}^{m}$, there is a total of $n_\theta = d_h \times (d+1) + m \times (d_h + 1)$ scalar parameters.

The empirical risk $\hat{R}$ associated with the lost function is:

$$\hat{R} = \frac{1}{n}\sum_{i=1}^{n} L(f_\theta(\mathbf{x}^{(i)}), y) \tag{12}$$

The optimization problem of training the network in order to find the optimal values of the parameters is:

$$\theta^* = argmin \hat{R}_\lambda(f_\theta, D_n) \tag{13}$$

Where we can add a regularization term, $\lambda\Omega(\theta)$, to the empirical risk.

2.f

$$\theta \leftarrow \theta - \eta \frac{\partial \hat{R}\lambda}{\partial \theta} \tag{14}$$

$$\leftarrow \theta - \eta \frac{\partial}{\partial \theta}\left(\frac{1}{n}\sum_{i=1}^{n} L(f_\theta(\mathbf{x}^{(i)}), y^{(1)})\right) \tag{15}$$

Again, we haven't, but we could add a regularization term at the end of this equation: $+\eta \frac{\partial}{\partial \theta}\Omega(\theta)$.

2.g

Cross Entropy Loss with Softmax function are used as the output layer extensively. Now we use the derivative of softmax [1] that we derived earlier to find the derivative of the cross entropy loss function.

First, we re-arrange our equation to allow for easy differentiation:

$$L(\mathbf{o}_k^a, y) = -\sum_{k=1}^{M} y_k \log\left(\frac{e^{\mathbf{o}_k^a}}{\sum_{i=1}^{m} e^{\mathbf{o}_i^a}}\right) \tag{16}$$

$$= \sum_{k=1}^{M} y_k - \log y_k * \left(\frac{e^{\mathbf{o}_k^a}}{\sum_{i=1}^{m} e^{\mathbf{o}_i^a}}\right) \tag{17}$$

$$= -\log y_k * \left(\frac{e^{\mathbf{o}_k^a}}{\sum_{i=1}^{m} e^{\mathbf{o}_i^a}}\right) - \sum_{j\neq y}^{M-1} y_j \log y_j * \left(\frac{e^{\mathbf{o}o_k^a}}{\sum_{i=1}^{m} e^{\mathbf{o}_i^a}}\right) \tag{18}$$

$$\tag{19}$$

Note that the right-hand term where $j \neq y$ is all zero because y_k is all 0, and our target y_k is 1, so

$$L(\mathbf{o}_k^a, y) = -\log\left(\frac{e^{\mathbf{o}_k^a}}{\sum_{i=1}^{m} e^{\mathbf{o}_i^a}}\right) \tag{20}$$

$$= -\log(e^{\mathbf{o}_k^a}) + \log(\sum_{i=1}^{m} e^{\mathbf{o}_i^a}) \tag{21}$$

$$= -\mathbf{o}_k^a + \log(\sum_{i=1}^{m} e^{\mathbf{o}_i^a}) \tag{22}$$

$$= -\mathbf{o}_k^a + \log(\sum_{i\neq k}^{m} e^{\mathbf{o}_i^a} + e^{\mathbf{o}_k^a}) \tag{23}$$

Now we are ready to take the derivative with respect to the output layer when we have the correct class k:

$$\frac{\partial L}{\partial \mathbf{o}_k^a} = -1 + \frac{e^{\mathbf{o}_k^a}}{\sum_{i\neq k}^{m} e^{\mathbf{o}_i^a} + e^{\mathbf{o}_k^a}} \tag{24}$$

$$= \frac{e^{\mathbf{o}_k^a}}{\sum_{k=1}^{m} e^{\mathbf{o}_k^a}} - 1 \tag{25}$$

$$= \mathbf{o}_k^s - 1 \tag{26}$$

Since $onehot_m(y) = 1$ when m is the target and is 0 otherwise, we see that the above is true. We can similarly take the derivative with respect to the output layer when we have the incorrect class i:

$$\frac{\partial L}{\partial \mathbf{o}_i^a} = \frac{e^{\mathbf{o}_i^a}}{\sum_{i \neq k}^m e^{\mathbf{o}_i^a} + e^{\mathbf{o}_k^a}} \tag{27}$$

$$= \mathbf{o}_i^s - 0 \tag{28}$$

2.h

$$\frac{\partial L}{\partial W_{kj}^{(2)}} = \sum_{i=1}^m \frac{\partial L}{\partial o_i^a} \frac{\partial o_i^a}{\partial W_{kj}^{(2)}} \tag{29}$$

$$\frac{\partial L}{\partial b_k^{(2)}} = \sum_{i=1}^m \frac{\partial L}{\partial o_i^a} \frac{\partial \mathbf{o}_i^a}{\partial b_k^{(2)}} \tag{30}$$

We have already defined $\frac{\partial L}{\partial o_k^a}$. $\frac{\partial o_k^a}{\partial W_{kj}^{(2)}}$ and $\frac{\partial \mathbf{o}_k^a}{\partial b_k^{(2)}}$ are given by:

$$\frac{\partial o_i^a}{\partial W_{kj}^{(2)}} = \frac{\partial}{\partial W_{kj}^{(2)}} (W_{kj}^{(2)} h_j^s + b_k^{(2)}) \tag{31}$$

$$= \frac{\partial}{\partial W_{kj}^{(2)}} (h_j^s W_{kj}^{(2)} + b_k^{(2)}) \tag{32}$$

$$= h_j^s \tag{33}$$

$$\frac{\partial \mathbf{o}_i^a}{\partial b_k^{(2)}} = \frac{\partial}{\partial b_k^{(2)}} (W_k j^{(2)} \mathbf{h}_j^s + b_k^{(2)}) \tag{34}$$

$$= 1 \tag{35}$$

2.i

$$\frac{\partial L}{\partial \mathbf{W}^{(2)}} = \frac{\partial L}{\partial \mathbf{o}^a} \frac{\partial \mathbf{o}^a}{\partial \mathbf{W}^{(2)}} \tag{36}$$

$$\frac{\partial \mathbf{o}^a}{\partial \mathbf{W}^{(2)}} = \frac{\partial}{\partial \mathbf{W}^{(2)}} (\mathbf{W}^{(2)T} \mathbf{h}^s + \mathbf{b}^{(2)}) \tag{37}$$

$$= \mathbf{h}^{sT} \tag{38}$$

and

$$\frac{\partial L}{\partial \mathbf{b}^{(2)}} = \frac{\partial L}{\partial \mathbf{o}^a} \frac{\partial \mathbf{o}^a}{\partial \mathbf{b}^{(2)}} \tag{39}$$

$$\frac{\partial \mathbf{o}^a}{\partial \mathbf{b}^{(2)}} = \frac{\partial}{\partial \mathbf{b}^{(2)}}(\mathbf{W}^{(2)T}\mathbf{h}^s + \mathbf{b}^{(2)}) \tag{40}$$

$$= \mathbf{1} \tag{41}$$

Where $\mathbf{W}^{(2)} \in \mathbb{R}^{m \times d_h}$, $\mathbf{b}^{(2)} \in \mathbb{R}^m$, $\mathbf{o}^a \in \mathbb{R}^m$, $\mathbf{h}^s \in \mathbb{R}^{d_h}$, $\mathbf{1} \in 1^m$ and $\frac{\partial L}{\partial o^a} \in \mathbb{R}^m$.

2.j

$$\frac{\partial L}{\partial h_j^s} = \sum_{k=1}^{m} \frac{\partial L}{\partial \mathbf{o}_k^a} \frac{\partial \mathbf{o}_k^a}{\partial h_j^s} \tag{42}$$

We have already defined $\frac{\partial L}{\partial \mathbf{o}_k^a}$ and we can calculate $\frac{\partial \mathbf{o}_k^a}{\partial h_j^s}$ by:

$$\frac{\partial \mathbf{o}_k^a}{\partial h_j^s} = \frac{\partial}{\partial h_j^s}(W_{kj}^{(2)} h_j^s + b^{(2)}) \tag{43}$$

$$= W_{kj}^{(2)} \tag{44}$$

2.k

$$\frac{\partial L}{\partial \mathbf{h}^s} = \frac{\partial L}{\partial \mathbf{o}^a} \frac{\partial \mathbf{o}^a}{\partial \mathbf{h}^s} \tag{45}$$

We have aleady defined $\frac{\partial L}{\partial \mathbf{o}^a}$. The gradient of $\frac{\partial \mathbf{o}^a}{\partial \mathbf{h}^s}$ is given by:

$$\frac{\partial \mathbf{o}^a}{\partial \mathbf{h}^s} = \frac{\partial}{\partial \mathbf{h}^s}(\mathbf{W}^{(2)T}\mathbf{h}^s + \mathbf{b}^{(2)}) \tag{46}$$

$$= \mathbf{W}^{(2)T} \tag{47}$$

Where $\mathbf{W}^{(2)} \in \mathbb{R}^{mxd_h}$, $\mathbf{h}^s \in \mathbb{R}^{d_h}$, $\mathbf{b}^{(2)} \in \mathbb{R}^m$, $\mathbf{o}^s \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^m$.

2.l

$$\frac{\partial L}{\partial h_j^a} = \frac{\partial L}{\partial h_j^s} \frac{\partial h_j^s}{\partial h_j^a} \tag{48}$$

Where:

$$\frac{\partial h_j^s}{\partial h_j^a} = \begin{cases} 0 & \text{if } h_j^a < 0 \\ 1 & \text{if } h_j^a > 0 \end{cases} \tag{49}$$

And is undefined if $h_j^a = 0$

2.m

$$\frac{\partial L}{\partial \mathbf{h}^a} = \frac{\partial L}{\partial \mathbf{h}^s} \frac{\partial \mathbf{h}^s}{\partial \mathbf{h}_j} \tag{50}$$

Where:

$$\frac{\partial \mathbf{h}^s}{\partial \mathbf{h}^a} = \mathbf{I}_{\{h_j^a > 0\}} \tag{51}$$

Where $\mathbf{I} \in \mathbb{R}^{d_h}$.

2.n

$$\frac{\partial L}{\partial W_{ji}^{(1)}} = \sum_{k=1}^{d_h} \frac{\partial L}{\partial h_k^a} \frac{\partial h_k^a}{W_{ji}^{(1)}} \tag{52}$$

We have already defined $\frac{\partial L}{\partial h_j^a}$. The gradient $\frac{\partial h_k^a}{W_{ji}^{(1)}}$ is given by:

$$\begin{aligned} \frac{\partial h_k^a}{W_{ji}^{(1)}} &= \frac{\partial}{W_{ji}^{(1)}} (W_{ji}^{(1)} x_i + b_j^{(1)}) && (53) \\ &= \frac{\partial}{\partial W_{ji}^{(1)}} (x_i W_{ji}^{(1)} + b_j^{(1)}) && (54) \\ &= x_i && (55) \end{aligned}$$

and

$$\frac{\partial L}{\partial b_j^{(1)}} = \sum_{k=1}^{d_h} \frac{\partial L}{\partial h_k^a} \frac{\partial h_k^a}{b_j^{(1)}} \tag{56}$$

Again, we have already defined $\frac{\partial L}{\partial h_k^a}$. The gradient $\frac{\partial h_k^a}{b_j^{(1)}}$ is given by:

$$\begin{aligned} \frac{\partial h_k^a}{b_j^{(1)}} &= \frac{\partial}{b_j^{(1)}} (W_{ji}^{(1)T} x_i + b_j^{(1)}) && (57) \\ &= 1 && (58) \end{aligned}$$

2.o

$$\frac{\partial L}{\partial \mathbf{W}^{(1)}} = \frac{\partial L}{\partial \mathbf{h}^a} \frac{\partial \mathbf{h}^a}{\mathbf{W}^{(1)}} \tag{59}$$

$$\begin{aligned} \frac{\partial \mathbf{h}^a}{\mathbf{W}^{(1)}} &= \frac{\partial}{\mathbf{W}^{(1)}} (\mathbf{W}^{(1)} \mathbf{x}^T + \mathbf{b}^{(1)}) && (60) \\ &= \mathbf{x}^T && (61) \end{aligned}$$

and

$$\frac{\partial L}{\partial \mathbf{b}^{(1)}} = \frac{\partial L}{\partial \mathbf{h}^a} \frac{\partial \mathbf{h}^a}{\mathbf{b}^{(1)}} \tag{62}$$

$$\begin{aligned} \frac{\partial \mathbf{h}^a}{\partial \mathbf{b}^{(1)}} &= \frac{\partial}{\mathbf{b}^{(1)}} (\mathbf{W}^{(1)} \mathbf{x}^T + \mathbf{b}^{(1)}) && (63) \\ &= \mathbf{1} && (64) \end{aligned}$$

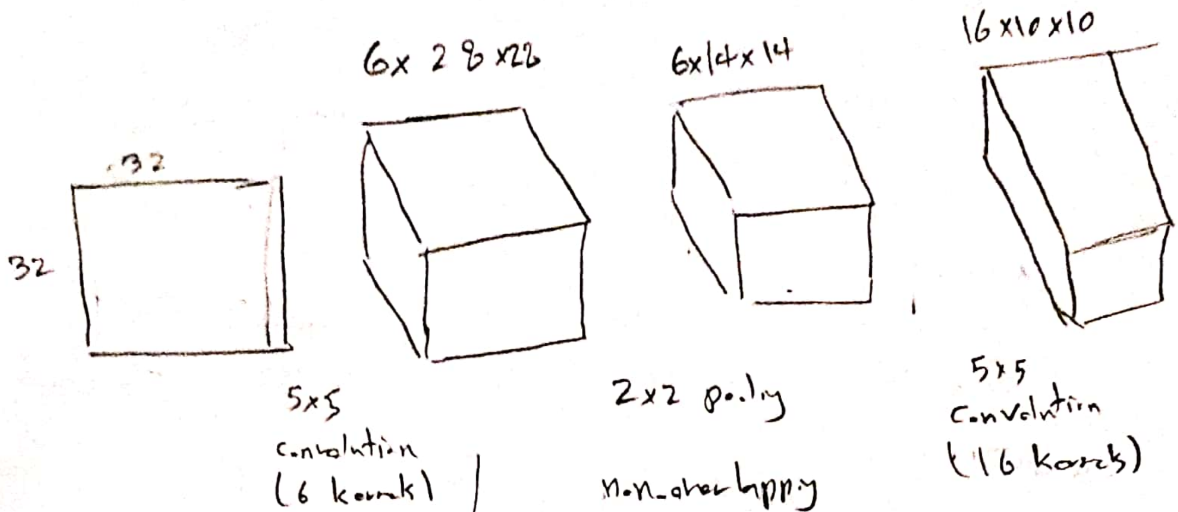
Where $\mathbf{h}^a \in \mathbb{R}^{d_h}$, $\mathbf{W}^{(1)} \in \mathbb{R}^{d_h \times d}$, $\mathbf{b}^{(1)} \in \mathbb{R}^{d_h}$, $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{1} \in \mathbb{R}^{d_h}$

2.p

$$\begin{aligned} \mathbf{h}_j^a &= \mathbf{b}_j^{(1)} + \sum_{i=1}^{d} \mathbf{W}_{ji}^{(1)} x_i \\ \frac{\partial L}{\partial x_k} &= \sum_j \frac{\partial L}{\partial \mathbf{h}_j^a} \frac{\partial \mathbf{h}_j^a}{\partial x_k} \\ \frac{\partial \mathbf{h}_j^a}{\partial x_k} &= \mathbf{W}_{jk}^{(1)} \\ \frac{\partial L}{\partial x_k} &= \sum_j \frac{\partial L}{\partial \mathbf{h}_j^a} \mathbf{W}_{jk}^{(1)} \end{aligned}$$

Q3:

a)



for the first layer $\Rightarrow 6\times(32-5+1)\times(32-5+1) = 6\times28\times28$

for the pooling layer $\Rightarrow 6\times\left(\frac{(28-2+2(0))}{2}+1\right)\times\left(\frac{28-2+2(0)}{2}+1\right) = 6\times14\times14$

for the last layer $\Rightarrow 16\times(14-5+1)\times(14-5+1) = 16\times10\times10$

b) params for input $= 0$

" " first layer $= 6(5\times5\times1) = 150$

" " last " $= 16(5\times5\times6) = 2400$

c) size of output $= \frac{\text{input size} - k + 2p}{s} + 1$

$6 = \frac{64-k+2(0)}{s} + 1 \Rightarrow \frac{64-k}{s} = 5$

we can chose many values for k and s; for example: $s=4$, $k=44$

s	1	2	3	---
k	59	54	49	

d) output size $= \frac{64 + 2p - d(k-1) - 1}{s} + 1$

$= \frac{64 + 2(1) - 2(k-1) - 1}{s} + 1 = 6$

$= \frac{67 - 2k}{s} = 5$

we also have many options for k and s; for example: $k = 26$, $s = 3$

it seems all values for s is prime number.

s	1	3	5	...
k	31	26	21	

---

e) output size $= \frac{64 + 2p - d(k-1) - 1}{s} + 1 = 6$

$= \frac{64 + 2(1) - 1(k-1) - 1}{s} + 1 = 6 \quad = \frac{66 - k}{s} = 5$

s	1	2	3	4	5
k	61	56	51	46	41