# IFT-6390 Fundamentals of Machine Learning

Professor: Ioannis Mitliagkas

Students: Abhay Puri (20209505), Saurabh Bodhe (20208545)

## Homework 3 - Theoretical Part

Solutions:

---

1) a) $(ReLU)\ g(x) = \max\{0, x\}$

this $\max\{0, x\}$ can be expanded as

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

Therefore, the first order derivative can be written as

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

and can be simplified as

$$\frac{d}{dx}(ReLU)g(x) = \begin{cases} \frac{d}{dx}0, & x < 0 \\ \frac{d}{dx}x, & x \geq 0 \end{cases}$$

Here invidual derivatives can be solved as follows:-

$\frac{d}{dx}x^n = nx^{n-1}$, and $\frac{d(0)}{dx} = rect'$

$$\frac{d}{dx}(ReLU)g(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases}$$

b) $\sigma(x) = \frac{1}{1+e^{-x}}$

$= (1+e^{-x})^{-1}$

also $(1+e^{-x}) = \frac{1}{\sigma(x)}$ ——— (1)

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

$e^{-x} = \frac{1-\sigma(x)}{\sigma(x)}$ — (2)

Therefore, the first order derivative

$\sigma'(x) = -1 \times (1+e^{-x})^{-2} \times e^{-x} \times (-1)$

$\sigma'(x) = e^{-x} \times \frac{1}{(1+e^{-x})^2}$

$\sigma'(x) = \frac{e^{-x}}{(1+e^{-x})^2}$ ——— (3)

Putting (1) & (2) in (3) We get

$\sigma'(x) = \frac{1-\sigma(x)}{\sigma(x)} \times \sigma^2(x)$

$$\boxed{\sigma'(x) = \sigma(x)(1-\sigma(x))}$$

---

c) $\sigma(x) = \frac{1}{2}\left(\tanh\left(\frac{1}{2}x\right) + 1\right)$

$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$

$\tanh\left(\frac{x}{2}\right) = \frac{e^{x/2} - e^{-\frac{x}{2}}}{e^{x/2} + e^{-\frac{x}{2}}}$

$$\tanh\left(\frac{1}{2}x\right)+1 = \frac{e^{\frac{1}{2}x}-e^{-\frac{1}{2}x}}{e^{\frac{1}{2}x}+e^{-\frac{1}{2}x}} + 1$$

$$= \frac{2e^{\frac{x}{2}}}{e^{\frac{x}{2}}+e^{-\frac{x}{2}}}$$

$$\frac{1}{2}\left(\tanh\left(\frac{x}{2}\right)+1\right) = \frac{e^{\frac{x}{2}}}{e^{\frac{x}{2}}+e^{-\frac{x}{2}}}$$

$$= \frac{e^{x/2}\ 1}{e^{x/2}\left(1+\frac{e^{-x/2}}{e^{x/2}}\right)}$$

$$\boxed{\frac{1}{2}\left(\tanh\left(\frac{x}{2}\right)+1\right) = \frac{1}{1+e^{-x}} = \sigma(x)}$$ Proved

---

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

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

Taking ln on ~~both~~ sides, We get

$$\ln\sigma(x) = \ln\left(\frac{1}{1+e^{-x}}\right)$$

$$\ln\sigma(x) = \ln(1)^{0} - \ln\left(1+e^{-x}\right)$$

$$\ln\sigma(x) = -\ln\left(1+e^{-x}\right) = -\text{softplus}(-x)$$

Hence Proved

---

e) $\text{softplus}(x) - \text{softplus}(-x) = x$

As We know that

$= \ln(1+e^x) - \ln(1+e^{-x})$

$= \ln\left(\frac{1+e^x}{1+e^{-x}}\right)$

$= \ln\left(\frac{1+e^x}{1+\frac{1}{e^x}}\right)$

$= \ln\left(\frac{e^x(1+e^x)}{(1+e^x)}\right)$

$= x$

---

f) $\text{Sign}(x) = 1_{x>0}(x) - 1_{x<0}(x)$

---

g) $\|x\|_2^2 = \sum_i x_i^2$

$\frac{\partial \|x\|_2^2}{\partial x_i} = 2x_i$

$\frac{\partial \|x\|_2^2}{\partial x} = \begin{bmatrix} 2x_1 \\ 2x_2 \\ \vdots \\ 2x_n \end{bmatrix} = 2x$

h) $\|x_1\| = \sum_i |x_i|$

$$\frac{\partial \|x\|_1}{\partial x} = \frac{\partial \sum_i |x_i|}{\partial x}$$

$$= \begin{bmatrix} \frac{\partial |x_1|}{\partial x_1} \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{\partial |x_2|}{\partial x_2} \\ \vdots \\ 0 \end{bmatrix} + \text{----}$$

$$= \begin{bmatrix} \frac{x_1}{|x_1|} \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{x_2}{|x_2|} \\ \vdots \\ 0 \end{bmatrix} + \text{----}$$

$$= \begin{bmatrix} \frac{x_1}{|x_1|} \\ \frac{x_2}{|x_2|} \\ \vdots \\ 0 \end{bmatrix} = \text{sign}(x)$$

---

i) $$S(x)_i = \frac{e^{x_i}}{\sum_j e^{x_j}}$$

$$S(c \cdot x_i) = \frac{e^{c \cdot x_i}}{\sum_j e^{c \cdot x_j}} \neq \frac{e^{x_i}}{\sum_j e^{x_j}}$$

Let us consider two cases :-

for $c = 0$,

$$S(c \cdot x_i) = \frac{e^{c x_i}}{\sum_j e^{c \cdot x_j}} = \frac{1}{N}$$

$N \Rightarrow$ No. of elements in $x$.

For $c \to \infty$

$$\lim_{c \to \infty} S(c \cdot x_i) = \frac{e^{c \cdot x_i}}{\sum_j e^{c \cdot x_j}} = 1$$ for maximum value of $x$

Hence, Softmax is not invariant under Scalar multiplication.

---

j) As We know that,

$$S(x)_i = \frac{e^{x_i}}{\sum_j e^{x_j}}$$

$$S(x+c)_i = \frac{e^{x_i + c}}{\sum_j e^{x_j + c}} = \frac{e^{x_i} \cdot e^{c}}{\sum_j e^{x_j} \cdot e^{c}}$$

$$= \frac{e^{c} \cdot e^{x_i}}{e^{c} \cdot \sum_j e^{x_j}} = \boxed{\frac{e^{x_i}}{\sum_j e^{x_j}}}$$

Hence, Softmax is translation invariant

---

k) $$\frac{\partial S(x)_i}{\partial x_j} = S(x)_i \, 1_{i=j} - S(x) \, S(x)_j$$

Consider 2 cases, $i = j$ and $i \neq j$

Case 1: $i = j$

$$\frac{\partial s(x)_i}{\partial x_i} = \frac{\partial}{\partial x_i} \frac{e^{x_i}}{\sum_j e^{x_j}}$$

Applying chain rule

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

$$= \frac{e^{x_i}}{\sum_j e^{x_j}} - \frac{e^{x_i} e^{x_i}}{(\sum_j e^{x_j})^2}$$

$$= s(x)_i - s(x)_i^2$$

$$= s(x)_i - s(x)_i s(x)_i \quad \text{——— (1)}$$

Case 2: $i \neq j$

$$\frac{\partial s(x)_i}{\partial x_i} = \frac{\partial \frac{e^{x_i}}{\sum_j e^{x_j}}}{\partial x_j}$$

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

$$= 0 - \frac{e^{x_i} e^{x_j}}{(\sum_j e^{x_j})^2}$$

$$= 0 - s(x)_i s(x)_j \quad \text{——— (2)}$$

Combining (1) & (2) we get

$$\boxed{\frac{\partial s(x)_i}{\partial x_j} = s(x)_i 1_{i=j} - s(x)_i s(x)_j}$$

l) As given in the question

$$\frac{\partial s(x)_i}{\partial x_j} = s(x)_i 1_{i=j} - s(x)_i s(x)_j$$

with condition $(i=j)$ diagonal elements of $\frac{\partial s(x)}{\partial x}$ are

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

Otherwise $i \neq j$, we have

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

We can write

$$\frac{\partial s(x)}{\partial x} = \text{diag}(s(x)) - \begin{bmatrix} s(x)_1 \\ s(x)_2 \\ \vdots \\ s(x)_n \end{bmatrix} \begin{bmatrix} s(x_1) & s(x_2) & \cdots & s(x)_n \end{bmatrix}$$

$$\boxed{\frac{\partial s(x)}{\partial x} = \text{diag}(s(x)) - s(x)\, s(x)^T}$$

---

m) $$y = \sigma(x) = \begin{bmatrix} \sigma(x_1) \\ \sigma(x_2) \\ \vdots \\ \sigma(x_n) \end{bmatrix}$$

Jacobian of $f(x) = \sigma(x)$ is diagonal and can be written as

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

$$\nabla_x L = \left(\frac{\partial y}{\partial x}\right)^T \nabla_y L$$

$$\nabla_x L = \left(\frac{\partial y}{\partial x}\right)^T \nabla_y L = \underbrace{\begin{bmatrix} \sigma(x_1)(1-\sigma(x_1)) & & \\ & \ddots & \\ & & \sigma(x_n)(1-\sigma(x_n)) \end{bmatrix}}_{d_{l2} \times d_{l2}} \underbrace{\begin{bmatrix} (\nabla_y L)_1 \\ (\nabla_y L)_2 \\ \vdots \\ (\nabla_y L)_n \end{bmatrix}}_{d_{l2} \times 1}$$

$$= \begin{bmatrix} \sigma(x_1)(1-\sigma(x_1))(\nabla_y L)_1 \\ \vdots \\ \sigma(x_n)(1-\sigma(x_n))(\nabla_y L)_n \end{bmatrix}_{d_{l2} \times 1}$$

For the case of $y = S(x)$, we have

$$\nabla_x L = \left(\frac{\partial y}{\partial x}\right)^T \nabla_y L = \left(\text{diag}(S(x)) - S(x)S(x)^T\right)^T \nabla_y L$$

$$= \left(\text{diag}(S(x)) - S(x)S(x)^T\right) \nabla_y L$$

$$= \text{diag}(S(x)) \nabla_y L - S(x)S(x)^T \nabla_y L$$

$$= \text{diag}(S(x)) \nabla_y L - S(x)\left(S(x)^T \nabla_y L\right)$$

$\underbrace{S(x)^T}_{1 \times n} \underbrace{\Delta_y L}_{n \times 1} = O(n)$

$\underbrace{S(x)}_{n \times 1} \left( \underbrace{S(x)^T \nabla_y L}_{1 \times 1} \right) = O(n)$

$$\text{diag}(S(x)) \nabla_y L = O(n)$$

The overall complexity is $O(n)$.

Q3. a) First Layer

Size = 32 × 32, Num of Kernels = 6, filter size = 5

Padding = 0, stride = 1

Output Dimension = $((32 - 5 + 2(0))/1) + 1$

$= 27 + 1 = 28 \times 28 \times 6$

Second Layer

Size = 28 × 28, pooling layer on this output : pooling layer Dimension : 2 × 2 with no overlapping., stride = 2

Output Dimension = $((28 - 2 + 2(0))/2) + 1$

$= 14 \times 14 \times 6$

Third Layer

Size = 14 × 14, Num of Kernels = 16, stride = 1, padding = 0, filter size = 5

Output Dimension = $((14 - 5 + 2(0))/1) + 1$

$= 10 \times 10 \times 16$

Dimensionality (scalar) output of last layer: 10 × 10 × 16 = 1600

b) parameters for last layer

$= 5 \times 5 \times 16 \times 6$

$= 2400$

---

c) input matrix $= 64 \times 64 \times 3$

Output shape of last layer $= (64, 6, 6)$

Output Dimension $= \left(\frac{i + 2p - k}{s}\right) + 1$

in case of dilations,

effective $k' = k + (k-1)(d-1)$

Assume we are not using padding or dilation.

output shape $= 6$

Stride $(s) = 1$

Padding $(p) = 0$

Dilation $(d) = 1$

$$6 = ((64 - k + 2(0))/1) + 1$$

$$6 = 64 - k + 0 + 1$$

$$k = 64 - 6 + 0 + 1$$

$$\boxed{k = 59}$$

---

d) Assume $d = 2$, $p = 1$

Output Shape $= 6$

Stride $(s) = 1$

$$6 = ((64 - (2k-1) + 2(1))/1) + 1$$

$$6 = ((64 - 2k + 1 + 2)) + 1$$

$$\boxed{k = 31}$$

---

e) Assume $d = 1$, $p = 1$

output Shape $= 6$

stride $(s) = 1$

$$6 = ((64 - k + 2(1))/1) + 1$$

$$6 = 64 - k + 2 + 1$$

$$k = 64 - 6 + 2 + 1$$

$$\boxed{k = 61}$$

---

Q2 a) Since $b^{(1)}$ contains one bias term for every neuron in the layer containing $d_h$ neurons, its dimension is $d_h \times 1$.

$$h^a = W^{(1)}x + b^{(1)}$$

Here each row of $W^{(1)}$ and $b^{(1)}$ represents the weight and bias for a neuron.

output of the $j^{th}$ neuron $= h_j^a = W_j^{(1)} x + b_j^{(1)}$

$$h^s = \phi(h^a) = \sin(h^a)$$

---

b) $W^{(2)}$ has dimensions $(m, d_h)$ and $b^{(2)}$ dimension $m$.

Activation function of the neurons of the output layer

$$O^a = W^{(2)} h^s + b^{(2)}$$

$$O_k^a = W_k^{(2)} h^s + b_k^{(2)}$$

---

c)

$$O^s = \text{softmax}(O^a)$$

$$O_k^s = \text{softmax}(O_k^a) = \frac{e^{O_k^a}}{\sum_{j=1}^{m} e^{O_j^a}}$$

Since $e^x > 0$, it is evident that $O_k^s > 0$

$$\text{So, } \sum_{k=1}^{m} O_k^s = \sum_{k=1}^{m} \text{softmax}(O_k^a) = \sum_{k=1}^{m} \frac{e^{O_k^a}}{\sum_{j=1}^{m} e^{O_j^a}}$$

$$= \frac{e^{O_1^a} + e^{O_2^a} + \cdots + e^{O_m^a}}{e^{O_1^a} + e^{O_2^a} + \cdots + e^{O_m^a}} = 1$$

d) $L(x, y) = -\log o_y^s(x)$

$$= -\log \text{softmax}(o_y^a)$$

$$= -\log \frac{e^{o_y^a(x)}}{\sum_{j=1}^{m} e^{o_j^a(x)}}$$

---

e) $\hat{R} \Rightarrow$ It is an estimate of the true risk and is defined as the average loss over the dataset.

$$\hat{R} = \frac{1}{n} \sum_{i=1}^{N} L\left(x^{(i)}, y^{(i)}\right)$$

$$= \frac{1}{n} \sum_{i=1}^{N} -\log \frac{e^{o^a_{y^{(i)}}(x^{(i)})}}{\sum_{j=1}^{m} e^{o_j^a(x^{(i)})}}$$

Parameters of $\theta$ are: $\{W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)}\}$

Scalar parameters: $n_\theta$

$$n_\theta = d \cdot d_h + d_h + m \cdot d_h + m$$

$$= d_h \cdot (d+1) + m \cdot (d_h + 1)$$

The Optimization problem is.

$$\hat{\theta} = \underset{\theta}{\text{argmin}}\ \hat{R}_\theta$$

or

$$\hat{\theta} = \underset{\theta}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{N} L_\theta\left(x^{(i)}, y^{(i)}\right)$$

f) $\theta := \theta - \eta \frac{d\hat{R}}{d\theta}$ $\quad \eta \Rightarrow$ Learning Rate $\quad \theta \Rightarrow \{W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)}\}$

for each iteration the update rule is shown below:

$$W^{(1)} := W^{(1)} - \eta \frac{d\hat{R}}{dW^{(1)}}$$

$$b^{(1)} := b^{(1)} - \eta \frac{d\hat{R}}{db^{(1)}}$$

$$W^{(2)} := W^{(2)} - \eta \frac{d\hat{R}}{dW^{(2)}}$$

$$b^{(2)} := b^{(2)} - \eta \frac{d\hat{R}}{db^{(2)}}$$

---

g) $\frac{\partial L}{\partial o^s} = -\frac{\partial \log o^s_y}{\partial o^s} = -o^{-s} \text{onehot}_m(y)$

$$= \begin{bmatrix} -\frac{1}{o^s_1} \\ \vdots \\ -\frac{1}{o^s_y} \\ \vdots \\ -\frac{1}{o^s_m} \end{bmatrix} - \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ -\frac{1}{o^s_y} \\ \vdots \\ 0 \end{bmatrix}$$

$$\frac{\partial o^s}{\partial o^a} = \frac{\partial \text{softmax}(o^a)}{\partial o^a}$$

$$\frac{\partial o^s}{\partial o^a} = \begin{bmatrix} o^s_1 - o^s_1 o^s_1 & \cdots & -o^s_1 o^s_m \\ \vdots & & \vdots \\ -o^s_m o^s_1 & \cdots & o^s_m - o^s_m o^s_m \end{bmatrix}$$

$$\frac{\partial L}{\partial o^a} = \frac{\partial L}{\partial o^s}\frac{\partial o^s}{\partial o^a}$$

$$= \begin{bmatrix} o_1^s - o_1^s o_1^s & \cdots & -o_1^s o_m^s \\ \vdots & & \vdots \\ -o_m^s o_1^s & \cdots & o_m^s - o_m^s o_m^s \end{bmatrix}_{m\times m} \begin{bmatrix} 0 \\ \vdots \\ -\frac{1}{o_y^s} \\ \vdots \\ 0 \end{bmatrix}_{m\times 1} = \begin{bmatrix} o_1^s \\ \vdots \\ o_y^s - 1 \\ \vdots \\ o_m^s \end{bmatrix}_{m\times 1}$$

$$= o^s - \text{onehot}_m(y)$$

---

h) $o_k^a = W_k^{(2)} h^s + b_k^{(2)}$

$$\frac{\partial o_k^a}{\partial W_{kj}^{(2)}} = \frac{\partial W_k^{(2)} h^s}{\partial W_{kj}^{(2)}} + \frac{\partial b_k^{(2)}}{\partial W_{kj}^{(2)}} = h_j^s \quad \text{—— (1)}$$

$$\frac{\partial L}{\partial W_{kj}^{(2)}} = \frac{\partial L}{\partial o_k^a}\frac{\partial o_k^a}{\partial W_{kj}^{(2)}} \quad \text{—— (2)}$$

Putting (1) in (2) we get

$$\boxed{\frac{\partial L}{\partial W_{kj}^{(2)}} = \frac{\partial L}{\partial o_k^a} h_j^s}$$

$$\frac{\partial o_k^a}{\partial b_k^{(2)}} = \frac{\partial W_k^{(2)} h^s}{\partial b_k^{(2)}} + \frac{\partial b_k^{(2)}}{\partial b_k^{(2)}} = 1 \quad \text{—— (3)}$$

$$\frac{\partial L}{\partial b_k^{(2)}} = \frac{\partial L}{\partial o_k^a}\frac{\partial o_k^a}{\partial b_k^{(2)}} \quad \text{—— (4)}$$

Putting (3) in (4) we get

$$\boxed{\frac{\partial L}{\partial b_k^{(2)}} = \frac{\partial L}{\partial o_k^a} \times (1)}$$

(i) $$\frac{\partial L}{\partial W^{(2)}} = \frac{\partial L}{\partial o^a} h^{s\,T}$$

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

Dimension of $h^s$: $d_h \times 1$

Dimension of $\frac{\partial L}{\partial W^{(2)}}$: $m \times d_h$

Dimension of $\frac{\partial L}{\partial b^{(2)}}$: $m \times 1$

---

j) $$\frac{\partial L}{\partial h_j^s} = \sum_{k=1}^{m} \frac{\partial L}{\partial o_k^a} \frac{\partial o_k^a}{\partial h_j^s}$$

$$o_k^a = W_k^{(2)} h^s + b_k^{(2)}$$

$$\frac{\partial o_k^a}{\partial h_j^s} = \frac{\partial W_k^{(2)} h^s}{\partial h_j^s} + \frac{\partial b_k^{(2)}}{\partial h_j^s} = W_{kj}^{(2)}$$

$$= \sum_{k=1}^{m} \frac{\partial L}{\partial o_k^a} \frac{\partial o_k^a}{\partial h_j^s} = \sum_{k=1}^{m} \frac{\partial L}{\partial o_k^a} W_{kj}^{(2)}$$

---

k) $$\frac{\partial L}{\partial h^s} = W^{(2)T} \frac{\partial L}{\partial o^a}$$

Dimension of $h^s$: $d_h \times 1$

Dimension of $\frac{\partial L}{\partial o^a}$: $m \times 1$

Dimension of $W^{(2)}$: $m \times d_h$

---

l) $$\frac{\partial L}{\partial h_j^a} = \frac{\partial L}{\partial h_j^s} \frac{\partial h_j^s}{\partial h_j^a}$$

$$h_j^s = \phi(h_j^a) = \sin(h_j^a)$$

$$\frac{\partial h_j^s}{\partial h_j^a} = \cos(h_j^a)$$

$$\frac{\partial L}{\partial h_j^a} = \frac{\partial L}{\partial h_j^s} \cos(h_j^a)$$

---

m) $$\frac{\partial L}{\partial h^a} = \frac{\partial L}{\partial h^s} \cos(h_j^a)$$

Dimension of $h^s = (d_h, 1)$

Dimension of $\frac{\partial L}{\partial h^s} = (d_h, 1)$

Dimension of $\frac{\partial L}{\partial h^a} = (d_h, 1)$

---

n) $$h_j^a = b_j^{(1)} + \sum_{i=1}^{d} W_{ji}^{(1)} x_i$$

chain rules gives:

$$\frac{\partial L}{\partial W_{kj}^{(1)}} = \frac{\partial L}{\partial h_k^a} \frac{\partial h_k^a}{\partial W_{kj}^{(1)}} = x_j$$

$$\frac{\partial L}{\partial b_j^{(1)}} = \frac{\partial L}{\partial h_j^a} \frac{\partial h_j^a}{\partial b_j^{(1)}} = 1$$

we have,

$$\frac{\partial L}{\partial W_{kj}^{(1)}} = \frac{\partial L}{\partial h_k^a} x_j$$

$$\frac{\partial L}{\partial b_j^{(1)}} = \frac{\partial L}{\partial h_j^a}$$

---

c) $$\frac{\partial L}{\partial W^{(1)}} = \frac{\partial L}{\partial h^a} x^T$$

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

$\frac{\partial L}{\partial h^a}$ has dimension $(d_h, 1)$.

$\frac{\partial L}{\partial W^{(1)}}$ has dimension $(d_h, d)$.

$x$ has dimension $(d, 1)$.

$$\frac{\partial L}{\partial b^{(1)}} = (d_h, 1)$$

---

d) We calculate $\frac{\partial L}{\partial h_j^a}$ in the previous question

and $h_j^a = b_j^{(1)} + \sum_{i=1}^{d} W_{j(i)}^{(1)} x_i$

chain rule gives:

$$\frac{\partial L}{\partial x_k} = \sum_j \frac{\partial L}{\partial h_j^a} \frac{\partial h_j^a}{\partial x_k}$$

$$\frac{\partial h^a}{\partial x_k} = W_{jk}^{(1)}$$

$$\frac{\partial L}{\partial x_k} = \sum_j \frac{\partial L}{\partial h_j^a} W_{jk}^{(1)}$$

---