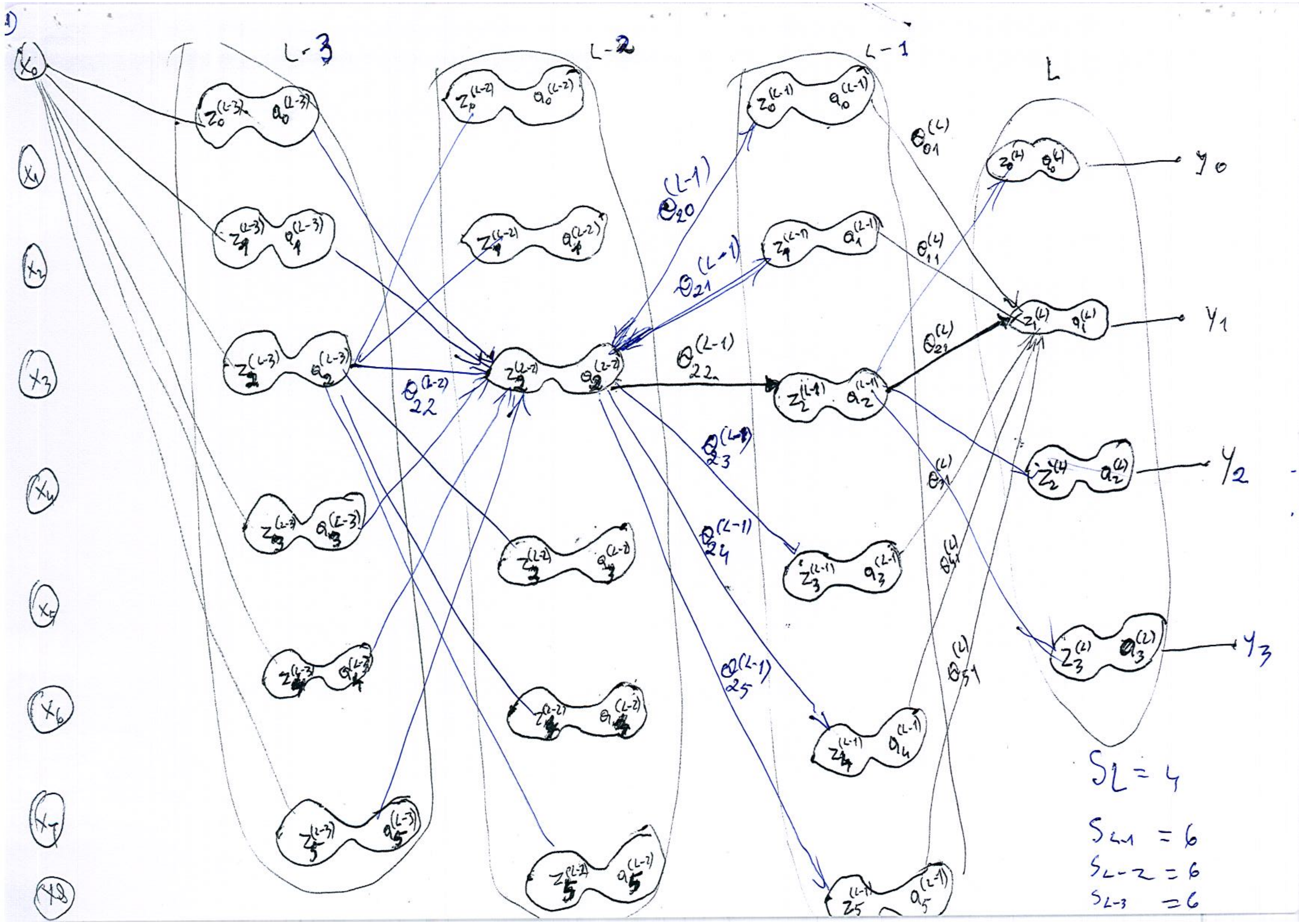
1)



$S_L = 4$

$S_{L-1} = 6$

$S_{L-2} = 6$

$S_{L-3} = 6$

2

$C_0(\Theta) = J_0(\Theta) = \sum_{j=0}^{S_L-1} (a_j^{(L)} - y_{0j})^2 = \sum_{j=0}^{S_L-1} C_{0j}$, where $C_{0j} = (\underbrace{a_j^{(L)}}_{\text{jth output of the output Layer (predicted)}} - \underbrace{y_{0j}^{(L)}}_{\text{j-th expected output}})^2$, 0 in C_0 and C_{0j} means just for one data training entry

$y_0 = [y_{00}, y_{01}, y_{02}, y_{03}]^T$.

$\boxed{C(\Theta) = \sum_{i=0}^{m-1} C_i(\Theta)}$ → Total cost for all training data

Let's work based on just one entry, entry 0.

Notes:

1) C_0 → is a function of $a^{(L)}$, of $a^{(L-1)}$, of $a^{(L-2)}$ ... . $C_0 = C_0(a^L(a^{L-1}(a^{L-2}(a^{L-3}(\ldots)))))$.

2) C_0 → is also a function of $\Theta^{(L)}$, of $\Theta^{(L-1)}$, of $\Theta^{(L-2)}$ ... . $C_0 = C_0(\Theta^{(L)})$; $C_0 = C_0(\Theta^{(L-1)})$ ..

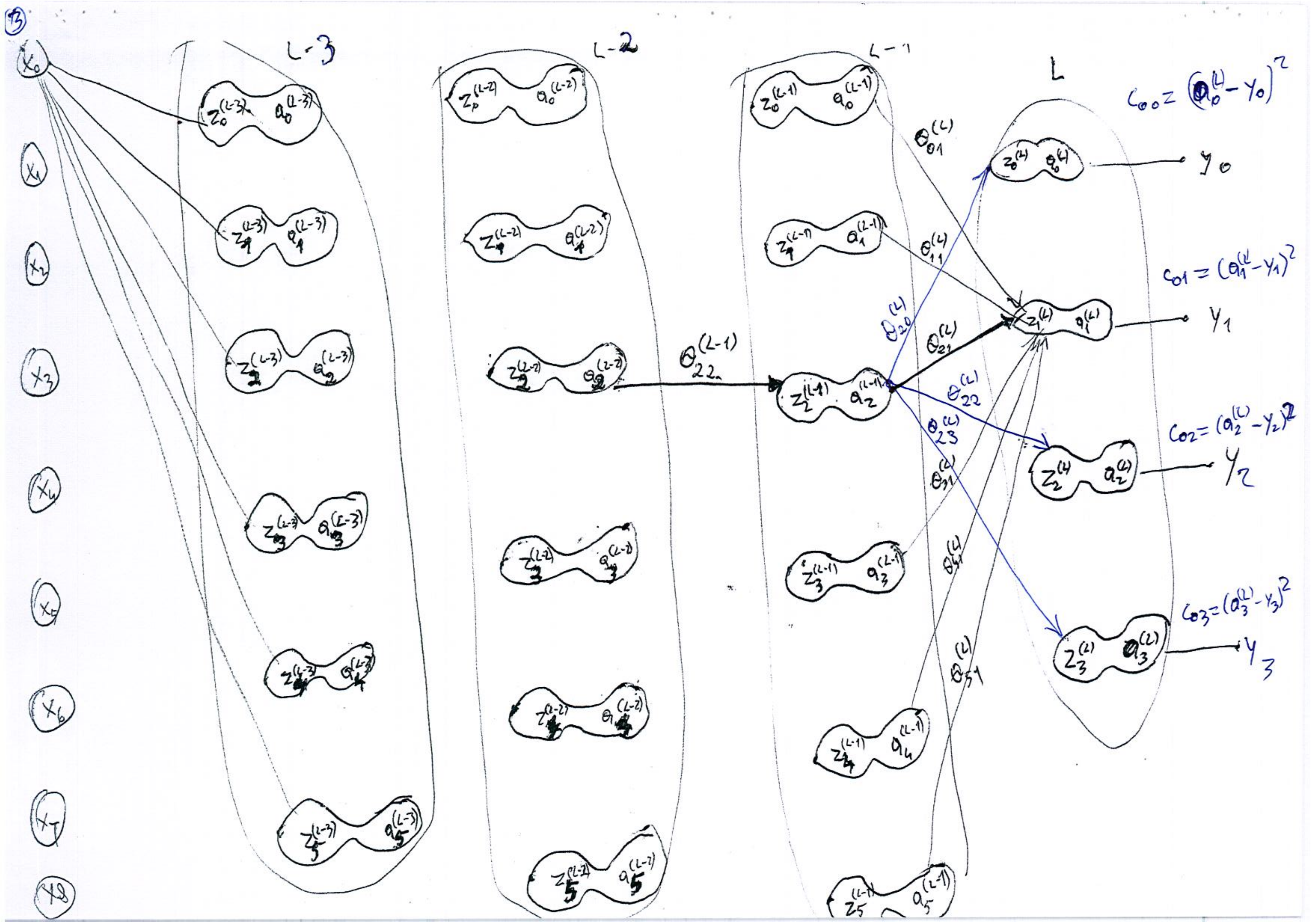
3) C_0 → is a function of any $\Theta_{jk}^{(l)}$ $\quad C_0 = C_0(\Theta_{jk}^{(l)})$, regardless in how coplicated way weights ($\Theta_{jk}^{(l)}$) and funcitions (activations functions) are connected with each other.

4) C_0 is a function of $\Theta_{jk}^{(l)}$ and a^l directly or through any composition of functions. (P.s. func $= f \circ g \circ h(x) = f(g(h(x)))$

5) $z_j^{(l)}$ is the input for $a_j^{(l)}$, α. $\boxed{z_1^{(L)} = \sum_{k=0}^{S_{L-1}-1} \Theta_{k1}^{(L)} \cdot a_k^{(L-1)}}$. $\boxed{z_j^{(L)} = \sum_{k=0}^{S_{L-1}-1} \Theta_{kj}^{(L)} a_k^{(L-1)}}$

6) $C_{01} = (a_1^{(L)} - y_{01})^2 \xleftarrow{\text{means is a function of}} C_0(a_1^{(L)}(z_1^{(L)}(\Theta^{(L)})))$ $\quad \boxed{a_j^{(l)} = g_j^{(l)}(z_j^{(l)})}$ Usually $\boxed{g_j^{(l)}(u) = \frac{1}{1+e^{-u}}}$

7) Reminder: For linear regression, logistic regression, during gradient Descent we had to calculate partial derivatives in order to update weights (which means updating or training the network here). $\Theta_j = \Theta_j - \alpha \cdot \frac{\partial C(\Theta)}{\partial \Theta_j}$ → How we calculate it?



13
L-3
L-2
L-1
L
x_0
x_1
x_2
x_3
x_4
x_5
x_6
x_7
x_8
$\theta_{22}^{(L-1)}$
$\theta_{01}^{(L)}$
$\theta_{11}^{(L)}$
$\theta_{20}^{(L)}$
$\theta_{21}^{(L)}$
$\theta_{22}^{(L)}$
$\theta_{23}^{(L)}$
$\theta_{31}^{(L)}$
$\theta_{41}^{(L)}$
$\theta_{51}^{(L)}$
$C_{00} = (a_0^{(L)} - y_0)^2$
$C_{01} = (a_1^{(L)} - y_1)^2$
$C_{02} = (a_2^{(L)} - y_2)^2$
$C_{03} = (a_3^{(L)} - y_3)^2$
y_0
y_1
y_2
y_3

7) In case of N. Networks we have Weights indexed $\theta_{jk}^{(L)}$, but essentially we need to do the same thing, calculate partial derivatives and update weights $\theta_{jk}^{(L)} = \theta_{jk}^{(L)} - \alpha \frac{\partial C(\theta)}{\partial \theta_{jk}^{(L)}}$ → How we calculate it in case of Neural Networks

8) $\frac{\partial C(\theta)}{\partial \theta_{kj}^{(L)}} = ?$ We do not pursue partial derivatives computation till the end. We will realize that there is a pattern in calculations, dynamic programming

9) $\frac{\partial C(\theta)}{\partial \theta_{kj}} = ?$

$$\boxed{\frac{\partial C_0(\theta)}{\partial \theta_{21}^{(L)}} = \frac{\partial C(\theta)}{\partial a_1^{(L)}} \cdot \frac{\partial a_1^{(L)}}{\partial z_1^{(L)}} \cdot \frac{\partial z_1^{(L)}}{\partial \theta_{21}^{(L)}}} \quad (1)$$

Note: θ_{21} is directly connected to output layer.

$$\boxed{\frac{\partial C_0(\theta)}{\partial \theta_{22}^{(L-1)}} = \frac{\partial C_0(\theta)}{\partial a_2^{(L-1)}} \cdot \frac{\partial a_2^{(L-1)}}{\partial z_2^{(L-1)}} \cdot \frac{\partial z_2^{(L-1)}}{\partial \theta_{22}^{(L-1)}}} \quad (2)$$

θ_{22} is an example of any weight not connected directly to output layer.

$$\frac{\partial C_0(\theta)}{\partial a_2^{(L-1)}} = \frac{\partial (C_{00}(\theta) + C_{01}(\theta) + C_{02}(\theta) + C_{03}(\theta))}{\partial a_2^{(L-1)}} =$$

$$= \frac{\partial C_{00}(\theta)}{\partial a_2^{(L-1)}} + \frac{\partial C_{01}(\theta)}{\partial a_2^{(L-1)}} + \frac{\partial C_{02}(\theta)}{\partial a_2^{(L-1)}} + \frac{\partial C_{03}(\theta)}{\partial a_2^{(L-1)}} =$$

$$\frac{\partial C_{00}}{\partial a_0^{(L)}} \cdot \frac{\partial a_0^{(L)}}{\partial z_0^{(L)}} \cdot \frac{\partial z_0^{(L)}}{\partial a_2^{(L-1)}} + \frac{\partial C_{01}}{\partial a_1^{(L)}} \cdot \frac{\partial a_1^{(L)}}{\partial z_1^{(L)}} \cdot \frac{\partial z_1^{(L)}}{\partial a_2^{(L-1)}} + \frac{\partial C_{02}}{\partial a_2^{(L)}} \cdot \frac{\partial a_2^{(L)}}{\partial z_2^{(L)}} \cdot \frac{\partial z_2^{(L)}}{\partial a_2^{(L-1)}} + \frac{\partial C_{03}}{\partial a_3^{(L)}} \cdot \frac{\partial a_3^{(L)}}{\partial z_3^{(L)}} \cdot \frac{\partial z_3^{(L)}}{\partial a_2^{(L-1)}}$$

⑮

$$\frac{\partial C_0(\theta)}{\partial a_2^{(L-1)}} = \sum_{j=0}^{S_L-1} \frac{\partial C_{0j}(\theta)}{\partial a_2^{(L-1)}} = \sum_{j=0}^{S_L-1} \frac{\partial C_{0j}}{\partial a_j^{(L)}} \cdot \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \cdot \frac{\partial z_j^{(L)}}{\partial a_2^{(L-1)}} \quad (3)$$

$$\frac{\partial C_{0j}}{\partial a_j^{(L)}} = \frac{\partial (a_j^{(L)} - y_j)^2}{\partial a_j^{(L)}} = \frac{\partial \sum_{j=0}^{S_L-1} (a_j^{(L)} - y_j)^2}{\partial a_j^{(L)}} = \frac{\partial (0 + 0 + (a_j^{(L)} - y_j)^2 + 0)}{\partial a_j^{(L)}} = 2(a_j^{(L)} - y_j)$$

Explanation

$$\frac{\partial C_{0j}}{\partial a_j^{(L)}} = 2(a_j^{(L)} - y_j) \quad (4) \qquad \frac{\partial C_{0j}}{\partial a_j^{(L)}} = \frac{\partial C_0(\theta)}{\partial a_j^{(L)}} \quad (5)$$

(4) ∧ (3) $$\boxed{\frac{\partial C_0(\theta)}{\partial a_2^{(L-1)}} = \sum_{j=0}^{S_L-1} 2(a_j^{(L)} - y_j) \cdot \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \cdot \frac{\partial z_j^{(L)}}{\partial a_2^{(L-1)}}} \quad (6)$$

Plugging in (6) in (2) =>

$$\boxed{\frac{\partial C_0(\theta)}{\partial \theta_{22}^{(L-1)}} = \underbrace{\left(\sum_{j=0}^{S_L-1} 2(a_j^{(L)} - y_j) \cdot \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \cdot \frac{\partial z_j^{(L)}}{\partial a_2^{(L-1)}}\right)}_{\frac{\partial C_0}{\partial a_2^{(L-1)}}} \cdot \frac{\partial a_2^{(L-1)}}{\partial z_2^{(L-1)}} \cdot \frac{\partial z^{(L-1)}}{\partial \theta_{22}^{(L-1)}}} \quad (7)$$

SUMMARY

$$\boxed{z_j^{(l)} = \sum_{k=0}^{S_{l}-1} a_{kj}^{(l)} \cdot a_k^{(l-1)}} \quad (1)$$

$$\boxed{a_j^{(l)}(z_j^{(l)}) = g_j^{(l)}(z_j^{(l)})} \quad (2)$$

$$\boxed{\text{ex. } g_j^{(l)}(u) = \sigma(u) = \frac{1}{1+e^{-u}}} \quad (3)$$

$$\boxed{C_{0j} = (a_j^{(L)} - y_{0j})^2; \quad \tilde{y}_{0j} = a_j^{(L)}; \quad C_{0j} = (\tilde{y}_j - y_{0j})^2} \quad (4)$$

$$\boxed{C_{01} = (a_1^{(L)} - y_{01})^2 = C_0(a_1^{(L)}(z_1^{(L)}(a^{(L)})), y)} \quad (5)$$

$C_0 \to$ cost considering just one trainning data entry, ex. (x_i, y_i) $i=0$

$$\boxed{C_0 = \sum_{j=0}^{S_L-1} C_{0j}} \quad (6) \equiv \boxed{C_0(\Theta) = \sum_{j=0}^{S_L-1} C_{0j}(\Theta)}$$

$$\boxed{C(\Theta) = \frac{1}{m}\sum_{i=0}^{m-1} C_i(\Theta)} \quad (7)$$

$$\boxed{\frac{\partial C_0(\Theta)}{\partial \Theta_{21}^{(L)}} = \frac{\partial C_0(\Theta)}{\partial a_1^{(L)}} \cdot \frac{\partial a_1^{(L)}}{\partial z_1^{(L)}} \cdot \frac{\partial z_1^{(L)}}{\partial \Theta_{21}^{(L)}}} \quad (8)$$

$$\boxed{\frac{\partial C_0(\Theta)}{\partial \Theta_{2j}^{(L)}} = \frac{\partial C_0(\Theta)}{\partial a_j^{(L)}} \cdot \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \cdot \frac{\partial z_j^{(L)}}{\partial \Theta_{2j}^{(L)}}} \quad (9)$$

$$\boxed{\frac{\partial C_0(\Theta)}{\partial \Theta_{32}^{(L-1)}} = \boxed{\frac{\partial C_0(\Theta)}{\partial a_2^{(L-1)}}} \cdot \frac{\partial a_2^{(L-1)}}{\partial z_2^{(L-1)}} \cdot \frac{\partial z_2^{(L-1)}}{\partial \Theta_{32}^{(L-1)}}} \quad (10)$$

$$\boxed{\boxed{\frac{\partial C_0(\Theta)}{\partial a_2^{(L-1)}}} = \sum_{j=0}^{S_L-1} \frac{\partial C_{0j}}{\partial a_j^{(L)}} \cdot \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \cdot \frac{\partial z_j^{(L)}}{\partial a_2^{(L-1)}}} \quad (11)$$

$$\boxed{\frac{\partial C_0(\Theta)}{\partial \Theta_{22}^{(L-1)}} = \boxed{\frac{\partial C_0(\Theta)}{\partial a_2^{(L-1)}}} \cdot \frac{\partial a_2^{(L-1)}}{\partial z_2^{(L-1)}} \cdot \frac{\partial z_2^{(L-1)}}{\partial \Theta_{22}^{(L-1)}}} \quad (12)$$

P.S $\dfrac{\partial C_{0j}}{\partial a_j^{(L)}} = \dfrac{\partial \left(\sum_{j=0}^{S_L-1} (a_j - y_{0j})^2\right)}{\partial a_j^{(L)}} = \boxed{(0 + .. + 2(a_j - y_{0j}) + 0)}$

$$\boxed{\frac{\partial C_0(\Theta)}{\partial \Theta_{k2}^{(L-1)}} = \boxed{\frac{\partial C_0(\Theta)}{\partial a_2^{(L-1)}}} \cdot \frac{\partial a_2^{(L-1)}}{\partial z_2^{(L-1)}} \cdot \frac{\partial z_2^{(L-1)}}{\partial \Theta_{k2}^{(L-1)}}} \quad (13)$$

$$\boxed{\frac{\partial C_{0j}}{\partial a_j^{(L)}} = 2(a_j - y_{0j}) = \frac{\partial C_0}{\partial a_j^{(L)}}} \quad (14)$$

Dynamic Programming !

How we Operate? The purpose is to (co)update $\theta_{jk}^{(l)}$ $\forall j, k, l$

- $l=L$: We start from layer (L). $l = L$. We start to calculate $\frac{\partial C_o(\theta)}{\partial \theta_{jk}^{(L)}}$. We accomplish this as in (9).
- $l = L-1$. What we accomplished in (9) i.e. $\frac{\partial C_o(\theta)}{\partial a_j^{(L)}}$, $\frac{\partial a_j^{(L)}}{\partial z_j^{(L)}}$, we use in (11) when we calculate $\frac{\partial C_o(\theta)}{\partial a_2^{(L-1)}}$ → which we use in (13) k-times.

Let's generalize (11).

$$\boxed{\frac{\partial C_o(\theta)}{\partial a_k^{(L-1)}} = \sum_{j=0}^{S_L-1} \frac{\partial C_{o_j}}{\partial a_j^{(L)}} \cdot \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \cdot \frac{\partial z_j^{(L)}}{\partial a_k^{(L-1)}}}, \quad \forall k : 0 .. S_{L-1} \tag{15}$$

Let's generalize (9)

$$\boxed{\frac{\partial C_o(\theta)}{\partial \theta_{kj}^{(L)}} = \frac{\partial C_o(\theta)}{\partial a_j^{(L)}} \cdot \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \cdot \frac{\partial z_j^{(L)}}{\partial \theta_{kj}^{(L)}}} \tag{16}$$

Let's generalize (13)

$$\boxed{\frac{\partial C_o(\theta)}{\partial \theta_{kj}^{(L-1)}} = \frac{\partial C_o(\theta)}{\partial a_j^{(L-1)}} \cdot \frac{\partial a_j^{(L-1)}}{\partial z_j^{(L-1)}} \cdot \frac{\partial z_j^{(L-1)}}{\partial \theta_{kj}^{(L-1)}}} \tag{17}$$

$$\boxed{\frac{\partial C_o}{\partial \theta_{kj}^{(l)}} = \frac{\partial C_o(\theta)}{\partial a_j^{(l)}} \cdot \frac{\partial a_j^{(l)}}{\partial z_j^{(l)}} \cdot \frac{\partial z_j^{(l)}}{\partial \theta_{kj}^{(l)}}} \tag{18}$$

Notes: The formula (8) could have been written initially as $\frac{\partial C_0}{\partial \Theta_{21}^{(L)}} = \frac{\partial C_{0j}}{\partial z_1^{(L)}} \cdot \frac{\partial z_1^{(L)}}{\partial \Theta_{21}^{(L)}}$ and then unfold as following

$$\frac{\partial C_0(\Theta)}{\partial \Theta_{21}^{(L)}} = \underbrace{\left(\frac{\partial C_{0j}}{\partial a_1^{(L)}} \cdot \frac{\partial a_1^{(L)}}{\partial z_1^{(L)}}\right)}_{\frac{\partial C_{0j}}{\partial z_1^{(L)}}} \cdot \frac{\partial z_1^{(L)}}{\partial \Theta_{21}^{(L)}}$$

. We could have done this for other analogical parts of other formulas.

i. e. :

$$(15') \boxed{\frac{\partial C_0}{\partial a_k^{(L-1)}} = \sum_{j=0}^{S_L-1} \frac{\partial C_{0j}}{\partial z_j^{(L)}} \cdot \frac{\partial z_j^{(L)}}{\partial a_k^{(L-1)}}} = \sum_{j=0}^{S_L-1} \underbrace{\left(\frac{\partial C_{0j}}{\partial a_j^{(L)}} \cdot \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}}\right)}_{\frac{\partial C_{0j}}{\partial z_j^{(L)}}} \frac{\partial z_j^{(L)}}{\partial a_{k-1}^{(L-1)}} \equiv (15)$$

Note: if $l == L$, then

$$\frac{\partial C_0(\Theta)}{\partial \Theta_{kj}^{(L)}} = \frac{\partial C_{0j}(\Theta)}{\partial \Theta_{kj}^{(L)}} \quad \left(\text{we can omit } \Theta\right)$$

proof

$$\left[\frac{\partial C_0}{\partial \Theta_{kj}^{(L)}} = \frac{\partial C_{0j}}{\partial \Theta_{kj}^{(L)}}\right] \text{ on earlier pages}$$

$$(18') \boxed{\frac{\partial C_0}{\partial \Theta_{kj}^{l}} = \underbrace{\frac{\partial C_0}{\partial a_j^{(l)}} \cdot \frac{\partial a_j^{(l)}}{\partial z_j^{(l)}}}_{= \frac{\partial C_0}{\partial z_j^{(l)}}} \cdot \frac{\partial z_j^{(l)}}{\partial \Theta_{kj}^{(l)}} = \frac{\partial C_0}{\partial a_j^{(l)}} \frac{\partial a_j^{(l)}}{\partial \Theta_{kj}^{(l)}}}$$

Note: Let's define auxiliary variable

$$\boxed{\delta_j^{l} := \frac{\partial C_0}{\partial z_j^{(l)}} = \frac{\partial C_0}{\partial a_j^{(l)}} \cdot \frac{\partial a_j^{(l)}}{\partial z_j^{(l)}}} \quad (19)$$

Then previous formula can be written as following:

$$(18'') \boxed{\frac{\partial C_0}{\partial \Theta_{kj}^{l}} = \delta_j^{(l)} \cdot \frac{\partial z_j^{(l)}}{\partial \Theta_{kj}^{(l)}}} \quad \left[\frac{\partial z_j^{(l)}}{\partial \Theta_{kj}^{(l)}} = \frac{\partial \left(\sum_{k=0}^{S_{l-1}-1} \Theta^{(l)} a_k^{(l-1)}\right)}{\partial \Theta_{kj}} = a_k^{(l-1)}\right] \Rightarrow (18''') \boxed{\frac{\partial C_0}{\partial \Theta_{kj}^{(l)}} = \delta_j^{l} \cdot a_k^{(l-1)}}$$

P.S

$$\left.\begin{aligned} dw_{kj} &:= \frac{\partial C_0}{\partial \Theta_{kj}^{(l)}} \\ dz &:= \frac{\partial C_0}{\partial z_j^{(l)}} = \delta_j^{l} \end{aligned}\right\} = \quad dw_{kj}^{(l)} = dz_j^{(l)} \cdot a_k^{(l-1)}$$

Notes

$$a_k^{(L-1)} \rightarrow z_j^{(L)} \rightarrow a_j^{(L)} \rightarrow C_0 \qquad \text{i.e} \quad C_0 = C_0\left(a_j^{(L)}\left(z_j^{(L)}\left(a_k^{(L-1)}\right)\right) \dots \right)$$

$$C_0 \rightarrow \sum \left(a_j^{(L)} \rightarrow z_j^{(L)}\right) \rightarrow a_k^{(L-1)}$$

$$C_0\left(\dots a_k^{(L-3)} \dots\right) = C\left(\dots a_j^{(L-1)}\left(z_j^{(L-1)}\left(a_k^{(L-2)}\right)\right) \dots \right)$$

$$a_k^{(L-2)} \rightarrow \boxed{\sum z_j^{(L-1)} \rightarrow a_j^{(L-1)}} \rightarrow$$  $$\rightarrow C_0 = C(+\Theta \cdot \Theta)$$

So far we've seen cases only for output layer L, or the hidden layer directly behind layer L.

Let's generalize more:

$$\frac{\partial C_0}{\partial a_{kj}^{(l)}} = \frac{\partial C_0}{\partial z_j^l} \cdot \frac{\partial z_j^l}{\partial a_{kj}^l}$$

When $l = L$, we have seen that

$$\frac{\partial C_0}{\partial z_j^l} = \frac{\partial C_0}{\partial a_j^l} \frac{\partial a_j^l}{\partial z_j^l} \qquad \frac{\partial C_0}{\partial a_j^l} = (a_j^l - y_{es});$$

Previously we attained:

$$\boxed{\frac{\partial C_0}{\partial a_k^{(L-1)}} = \sum_{j=0}^{S_L-1} \frac{\partial C_0}{\partial a_j^{(L)}} \cdot \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \cdot \frac{\partial z_j^{(L)}}{\partial a_k^{(L-1)}}}, \quad (15)'$$

we just replaced ∂C_{0j} with C_0

see the picture

$$\boxed{\frac{\partial C_0}{\partial a_k^{(L-2)}} = \sum_{j=0}^{S_{L-1}-1} \frac{\partial C_0}{\partial a_j^{(L-1)}} \cdot \frac{\partial a_j^{(L-1)}}{\partial z_j^{(L-1)}} \cdot \frac{\partial z_j^{(L-1)}}{\partial a_k^{(L-2)}}} \quad (20)$$

$$\boxed{\frac{\partial C_0}{\partial a_k^{(L-3)}} = \sum_{j=0}^{S_{L-2}-1} \frac{\partial C_0}{\partial a_j^{(L-2)}} \cdot \frac{\partial a_j^{(L-2)}}{\partial z_j^{(L-2)}} \cdot \frac{\partial z_j^{(L-2)}}{\partial a_k^{(L-3)}}} \quad (21)$$

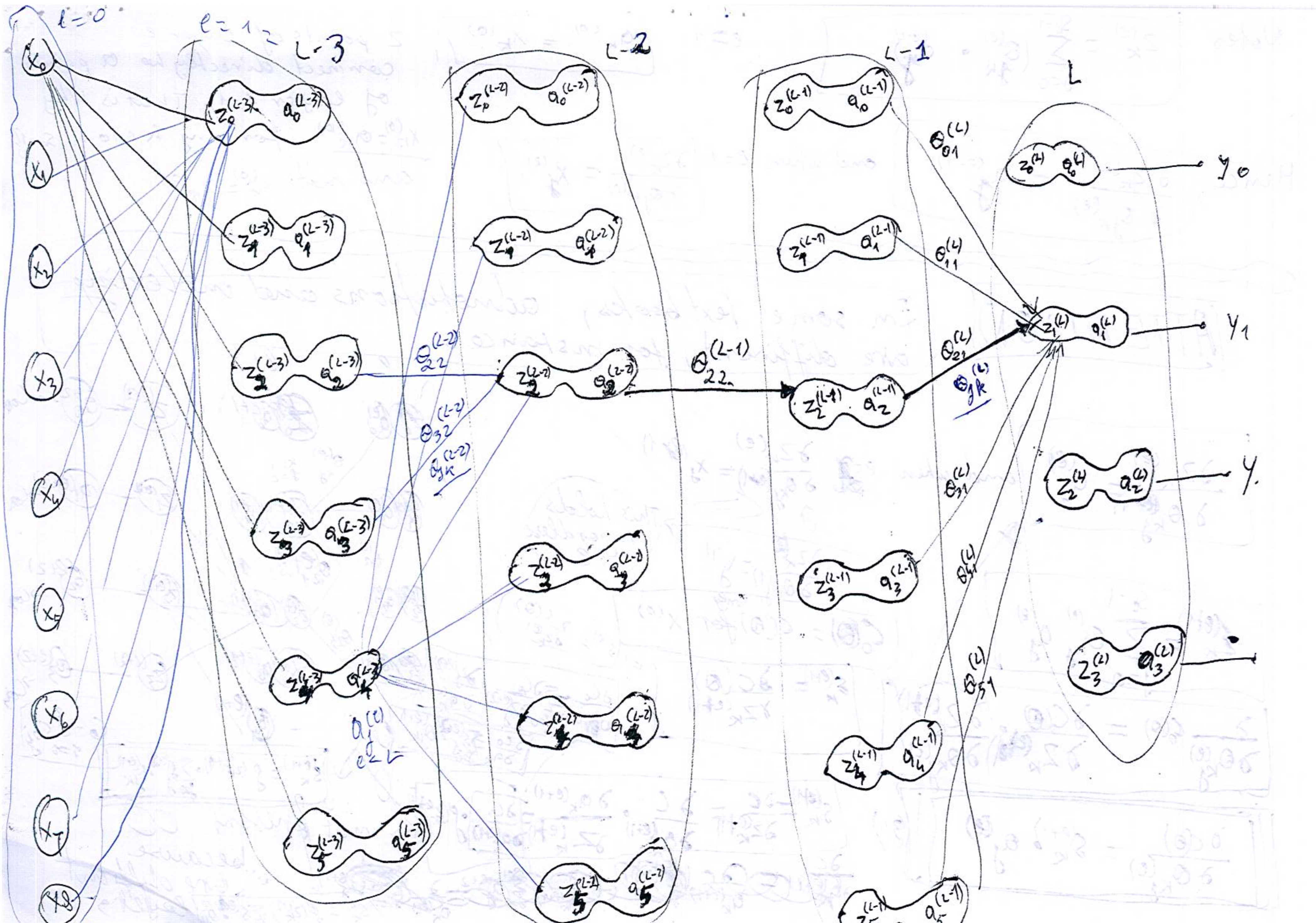
What about any $l \neq L$? ($l = L-1$, $L-2$, $L-3$, $L-4$...)

How we can provide relation between $\frac{\partial C_0}{a_k^{(l)}}$ and $\frac{\partial C_0}{a^{(l-1)}}$?

~~From~~ We can pick up whatever pair of consequent layers $(l-1, l)$

one have

$$\boxed{\frac{\partial C_0}{\partial a_k^{(l-1)}} = \sum_{j=0}^{S_l-1} \frac{\partial C_0}{\partial a_j^{(l)}} \cdot \frac{\partial a^{(l)}}{\partial z_j^{(l)}} \cdot \frac{\partial z_j^{(l)}}{\partial a_k^{(l-1)}}}$$



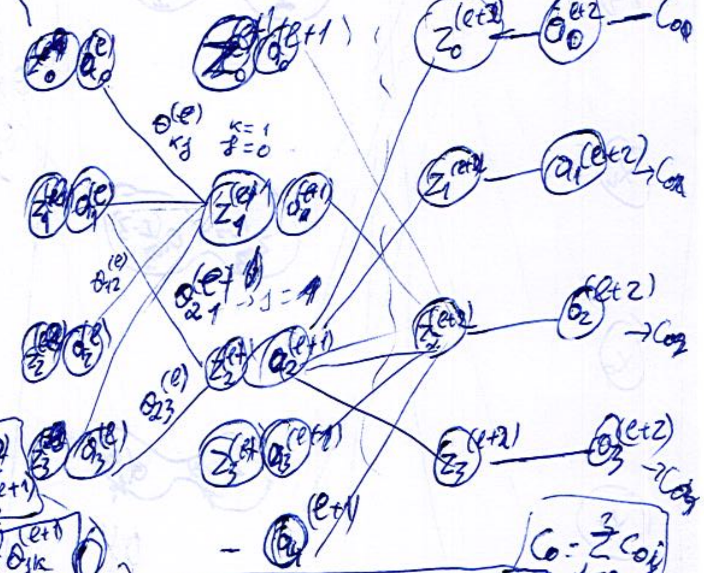
l=0
l=1=L-3
L-2
L-1
L
x0
x1
x2
x3
x4
x5
x6
x7
x8
y0
y1
y.
Θ22(L-2)
Θ32(L-2)
Θjk(L-2)
Θ22(L-1)
Θ01(L)
Θ11(L)
Θ21(L)
Θjk(L)
Θ31(L)
Θ41(L)
Θ51(L)
a(l)
l≥L

Notes $\boxed{z_k^{(e)} = \sum_{j=0}^{S_e-1} \left(\theta_{jk}^{(e)} \cdot a_j^{(e-1)}\right)}$ $e=1$ $\boxed{a_k^{(0)} = x_k^{(0)}}$; z points of layer e connect directly to a points of layer e-1. That is why $x_k^{(0)} = a_k^{(0)}$ for any $k \in \{0, \dots S_0-1\}$ and not $x_k^{(0)} = z_k^{(0)}$

Hence, $\boxed{\frac{\partial z_k^{(e)}}{\partial \theta_{jk}^{(e)}} = a_j^{(e-1)}}$ and when $e=1$ $\boxed{\frac{\partial z_k^{(e)}}{\partial \theta_{jk}^{(e)}} = x_j^{(e)}}$

ATTENTION In some textbooks, adnotations and indexing are diffirent, for instance:

e $e+1$



$\boxed{\frac{\partial z_k^{(e)}}{\partial \theta_{kj}^{(e+1)}} = a_j^{(e)}}$ and when $e=1$ $\frac{\partial z_k^{(e)}}{\partial \theta_{kj}^{(e+1)}} = x_j^{(e+1)}$

$\frac{\partial z_k^{(1)}}{\partial \theta_{kj}^{(1)}} = x_j^{(1)}$

This holds regardless e.

$\boxed{C_0(\theta) = C(\theta) \text{ for } x^{(0)} \text{ only}}$ $\boxed{\delta^{(e)} = \frac{\partial C(\theta)}{\partial z_k^{(e)}}}$

$\boxed{z_k^{(e+1)} = \sum_{j=0}^{S_e-1} \theta_{kj}^{(e)} a_j^{(e)}}$

$\boxed{\delta_k^{(e+1)} = \frac{\partial C(\theta)}{\partial z_k^{(e+1)}}}$

$\boxed{\frac{\partial C_0}{\partial a_k^{(e+1)}} = \sum_{j=0} \frac{\partial C_0}{\partial z_j^{(e+2)}} \cdot \frac{\partial z_j^{(e+2)}}{\partial a_k^{(e+1)}}}$

$\boxed{\frac{\partial C_0}{\partial a_k} = \sum_{j=0} \delta_j^{(e+2)} \cdot \theta_{jk}^{(e+1)}}$

$\boxed{\frac{\partial}{\partial \theta_{kj}^{(e)}} C_0(\theta) = \frac{\partial C_0(\theta)}{\partial z_k^{(e+1)}} \cdot \frac{\partial z_k^{(e+1)}}{\partial \theta_{kj}^{(e)}}}$

$\boxed{C_0 = \sum_{j=0}^{3} C_{0j}}$

$\boxed{\delta_k^{(e+1)} = g'(z_k^{(e+1)}) \cdot \sum_{j=0} \delta_j^{(e+2)} \cdot \theta_{jk}^{(e+1)}}$

$\boxed{\boxed{\frac{\partial C(\theta)}{\partial \theta_{kj}^{(e)}} = \delta_k^{(e+1)} \cdot a_j^{(e)}}}$ (3'')

$\boxed{\delta_k^{(e+1)} = \frac{\partial C}{\partial z_k^{(e+1)}} = \frac{\partial C}{\partial a_k^{(e+1)}} \cdot \frac{\partial a_k^{(e+1)}}{\partial z_k^{(e+1)}} = \frac{\partial C_0}{\partial a_k^{(e+1)}} \cdot g'(z_k^{(e+1)})}$

Do not present because are different layers

$\delta_k^{(e)} = g'(z_k^{(e)}) \cdot \sum \delta^{(e+1)} \cdot \theta_{jk}^{(e)}$

$$\boxed{\frac{\partial C_0}{\partial \theta_{jk}^{(l)}} = \frac{\partial C_0}{\partial z_k^{(l)}} \cdot \frac{\partial z_k^{(l)}}{\partial \theta_{jk}^{(l)}}} \quad (1)$$

This formula stands for any $l \in \{L, L-1, L-2 \dots\}$

sometimes we use this adnotation $\frac{\partial C}{\partial z_k^{(l)}} = \delta z_k^{(l)} = \delta_k^{(l)}$

$$\boxed{\delta_k^l \overset{\text{denote}}{:=} \frac{\partial C_0}{\partial z_k^{(l)}}} \quad (2)$$

$\forall l \in \{L, L-1, \dots 1\}$, Then formula (1) can be written as:

i.e. $\boxed{\delta_k^{l+1} = \frac{\partial C_0}{\partial z_k^{(l+1)}}} \quad (2')$

$$\boxed{\frac{\partial C_0}{\partial \theta_{jk}^{(l)}} = \delta_k^l \cdot \frac{\partial z_k^l}{\partial \theta_{jk}^{(l)}}} \quad (3) \rightarrow \forall l \in \{L, L-1, L-2, \dots\}$$

$$\boxed{\frac{\partial z_k^l}{\partial \theta_{jk}^{(l)}} = \frac{\partial \sum_{q=0}^{s_{l-1}} \theta_{qk}^{(l)} \cdot a_q^{(l-1)}}{\partial \theta_{jk}^{(l)}} = a_j^{(l-1)}}$$

Then $(3) \Rightarrow \boxed{\frac{\partial C_0}{\partial \theta_{jk}^{(l)}} = \delta_k^{(l)} \cdot a_j^{(l-1)}} \quad (3')$

$$\boxed{\delta_k^l = \frac{\partial C_0}{\partial z_k^l} = \underbrace{\boxed{\frac{\partial C_0}{\partial a_k^{(l)}}}}_{a} \cdot \underbrace{\boxed{\frac{\partial a_k^{(l)}}{\partial z_k^{(l)}}}}_{b}} \quad (4)$$

stands for any $l \in \{L, L-1, L-2 \dots\}$, because a_k and z_k are connected directly.

$$\boxed{\frac{\partial C_0}{\partial a_k^{(l)}}} \overset{a}{=} ?$$

a) $l = L$ $\quad \boxed{\frac{\partial C_0}{\partial a_k^{(L)}}} \overset{a}{=} \frac{\partial \left(\sum \frac{1}{2}(a_k^{(L)} - y_k)^2\right)}{\partial a_k^{(L)}} = \boxed{a_k^{(L)} - y_k}$; $\quad \boxed{\frac{\partial a_k^{(l)}}{\partial z_k^{(l)}}} \overset{b}{=} g'(z_k^{(l)})$

b) $l < L$ $\quad \boxed{\frac{\partial C_0}{\partial a_k^{(l)}}} \overset{a}{=} \sum_{j=0}^{s_{l+1}-1} \frac{\partial C_0}{\partial z_j^{(l+1)}} \frac{\partial z_j^{(l+1)}}{\partial a_k^{(l)}} = \boxed{\sum_{j=0}^{s_{l+1}-1} \delta_j^{l+1} \cdot \theta_{kj}^{(l+1)}}$; $\quad \boxed{\frac{\partial a_k^{(l)}}{\partial z_k^{(l)}}} \overset{b}{=} g'(z_k^{(l)})$

$\boxed{da^{(l)} = \theta^{(l+1)T} \cdot \delta^{(l+1)}}$

$\boxed{dA^{l-1} = \theta^{lT} \cdot dZ^l}$

$\frac{\partial C_0}{\partial a_k^{(l)}}$ ... ~~$= \sum_{j=0}^{s_{l+1}-1} \dots$~~

$dz^l = dA^{l+1} \cdot g'(\,)$

$dz^l = (\theta^{l+1})^T \cdot dZ^{l+1} \cdot g'(\,)$

a) $l = L$ $\quad \boxed{\delta_k^l = (a_k^{(l)} - y_k) \cdot g'(z_k^{(l)})} \quad (5)$

$l = L$: $\boxed{\delta_k^l = g'(z_k^{(l)}) * (a_k^{(l)} - y_k)} \quad (5)$ further $\delta^{(l)} = g'(z_k) .* \theta^{(l)} \cdot \delta^{(l+1)}$

b) $l < L$ $\quad \boxed{\delta_k^l = g'(z_k^{(l)}) \cdot \sum_{j=0}^{s_{l+1}-1} \delta_j^{(l+1)} \cdot \theta_{kj}^{(l)}} \quad (6)$ in different adnotation $\Longrightarrow \boxed{\delta_k^{(l)} = g'(z_k^{(l)}) .* \theta_k^{(l)} \cdot \delta^{(l+1)}} \rightarrow$ vectorized form

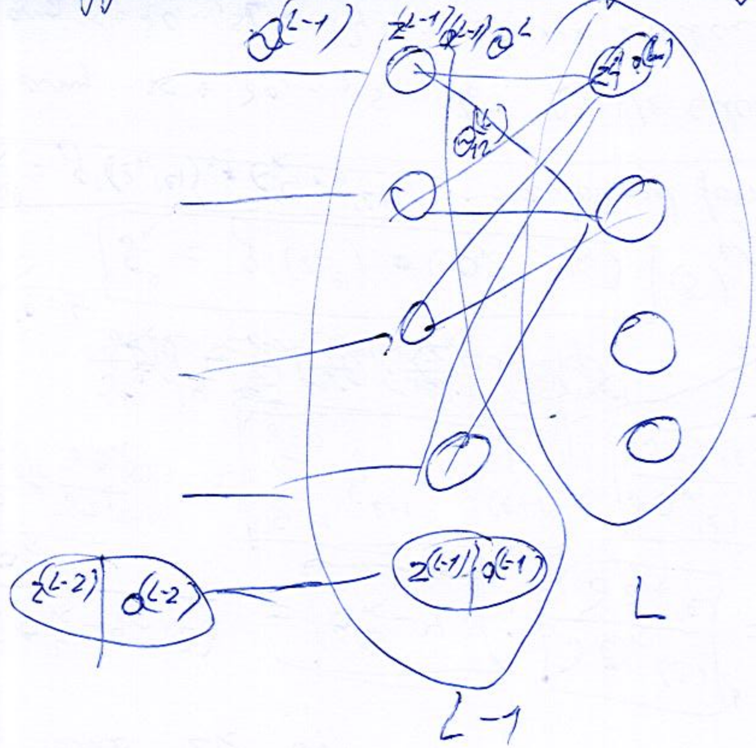
Algorithm:

1) We calculate $\boxed{\delta_k^l}$ as in (5) for $l = L$ and for any $k \in \{0, \dots, s_L - 1\}$; 2) We calculate $\frac{\partial C_0}{\partial \theta_{jk}^{(l)}}$ as in (3')

2) We calculate δ_k^l as in (6) for $l = L-1$ and any $k \in \{0, \dots, s_{L-1} - 1\}$; 2) we recalculate $\frac{\partial C_0}{\partial \theta_{jk}^{(l)}}$ as in (3')

3) We repeat (2) for $l = L-2, L-3 \dots$

if we use slightly different notations of layers for θ



$$z_j^{(L)} = \sum \theta_{jk}^{(L)} a_k^{(L-1)}$$

— — — —

$$\delta_k^{(l)} = g'(z_k^{(l)}) .* \theta^{(l+1)} \delta_k^{(l+1)}$$

$$\delta_{(k)}^{(l)} = \frac{\partial C}{\partial z_k^{(l)}}$$

$$dz_k^{(l)} = \delta_k^{(l)}$$

$$dz_k^{(l)} = g'(z_k^{(l)}) .* \underline{\theta^{(l+1)} \cdot dz_k^{(l+1)}}$$

$$\boxed{dz^{(l)} = (\theta^{l+1} \cdot dz^{(l+1)}) .* g(z^{(l)})}$$

standard matrix-multiplication (dot product)

elementwise

→ w.r.t. Andrews notation

recall (1) $\frac{\partial C}{\partial \theta_{jk}^{(l)}} = \frac{\partial C}{\partial z_k^{(l)}} \cdot \frac{\partial z_k^{(l)}}{\partial \theta_{jk}^{l}} \equiv \left[d\theta_{jk}^{(l)} = dz_k^{l} \cdot d\theta_{jk}^{(l)} \right] \geq \left[d\theta^{l} = dz^{l} \cdot a^{l} \right]$

$$\frac{\partial C}{\partial \theta_{jk}^{l}} = \left[\frac{\partial C}{\partial a_j^{(l)}} \cdot \frac{\partial a_j^{(l)}}{\partial z_j^{(l)}} \right] \frac{\partial z_j^{(l)}}{\partial \theta_{jk}}$$

$$dz_k^{(l)} = \frac{\partial C}{\partial a_j^{(l)}} \cdot \frac{\partial a_j^{l}}{\partial z_j^{(l)}} = da_j^{(l)} \cdot g'(z_j^{(l)})$$