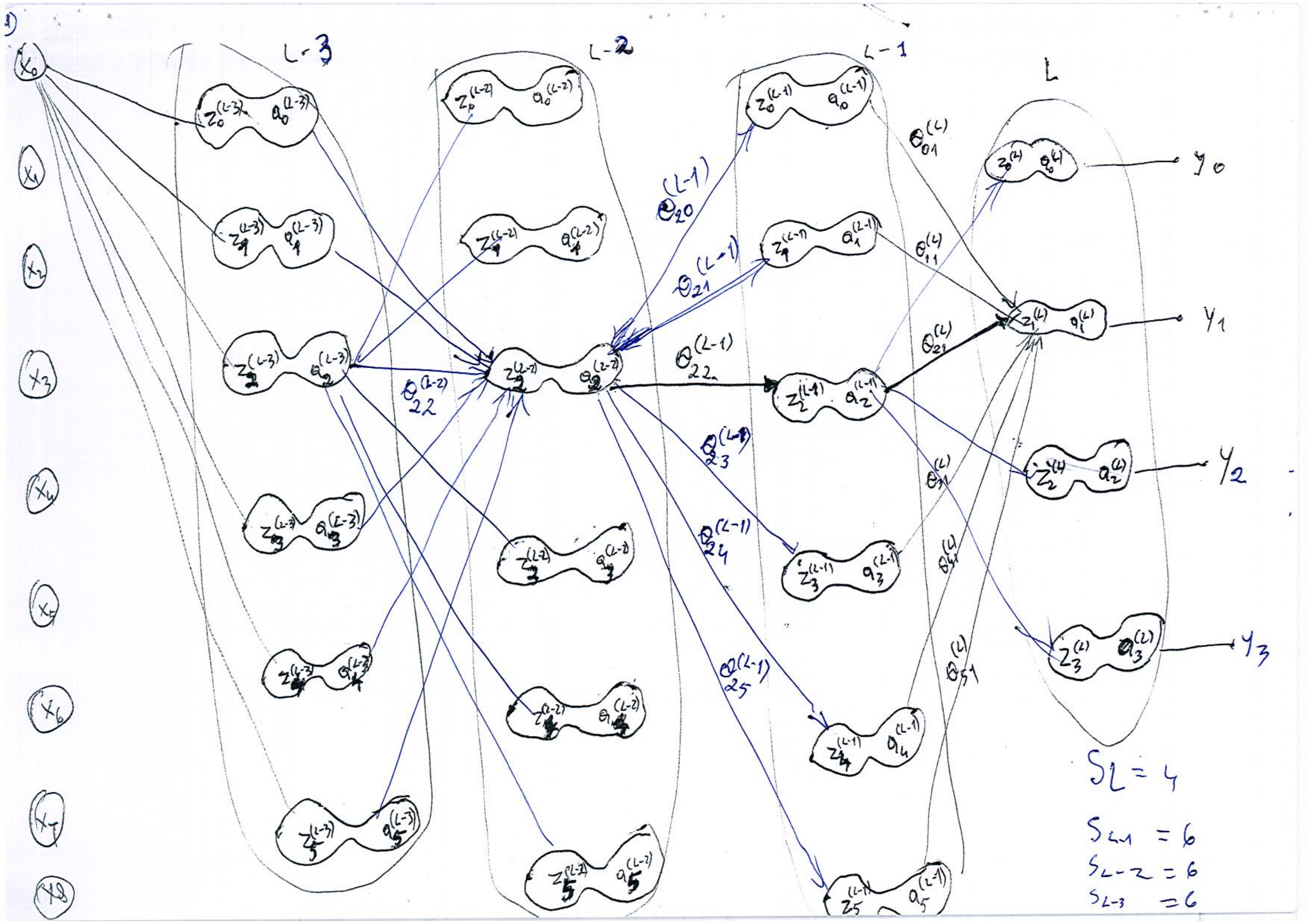


$S_L = 4$

$S_{L-1} = 6$

$S_{L-2} = 6$

$S_{L-3} = 6$

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 $\Theta^{(L-2)}$ .... $C_0 = C_0(\Theta^{(L)})$; $C_0 = C_0(\Theta^{(L-1)})$ ...

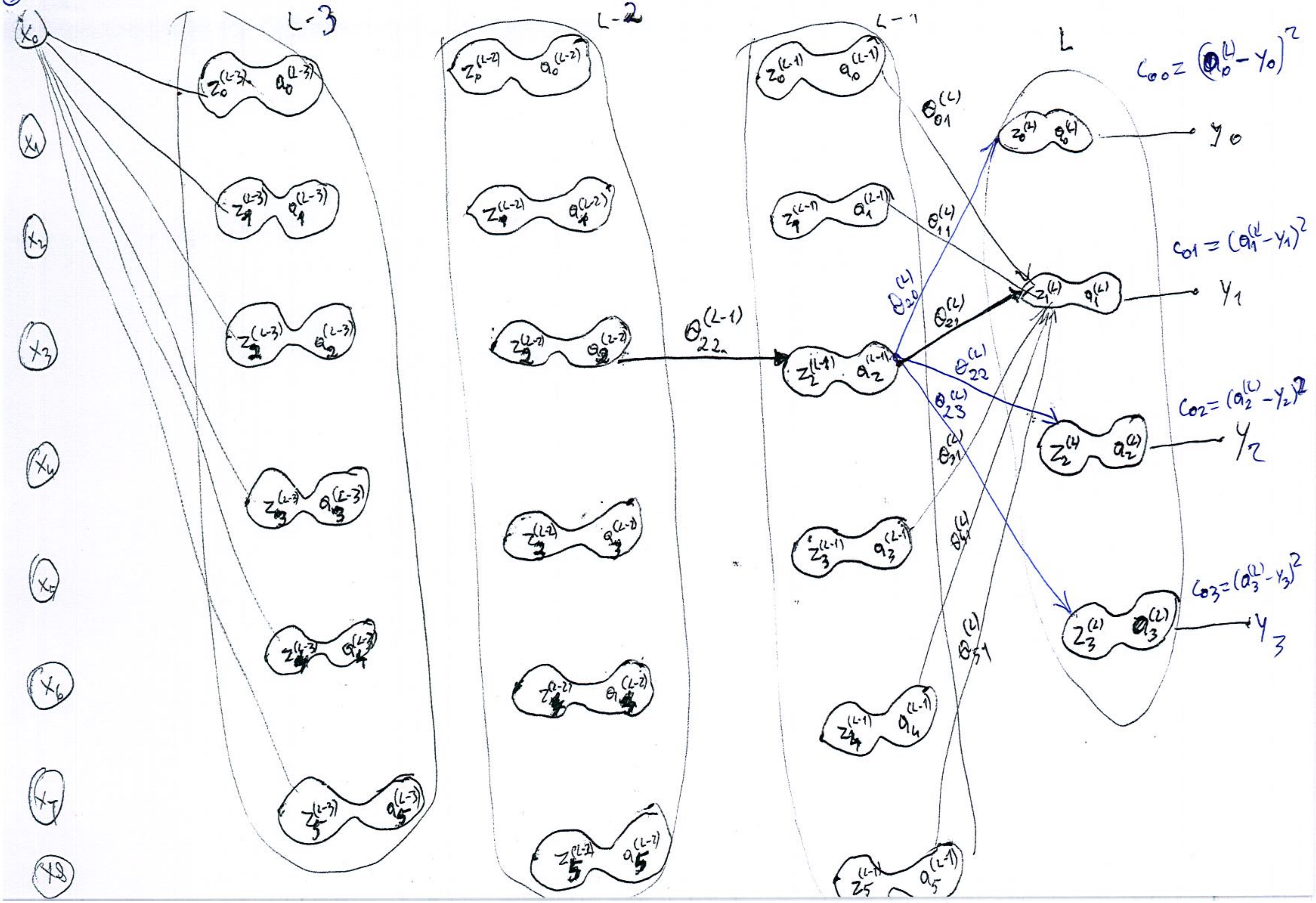
3) C_0 → is a function of any $\Theta_{jk}^{(\ell)}$ $\quad C_0 = C_0(\Theta_{jk}^{(\ell)})$, regardless in how coplicated way weights ($\Theta_{jk}^{(\ell)}$) and funcitions (activations functions) are connected with each other.

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

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

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

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 calculated it?



x_0 x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8
L-3
$z_0^{(L-3)}$ $a_0^{(L-3)}$
$z_1^{(L-3)}$ $a_1^{(L-3)}$
$z_2^{(L-3)}$ $a_2^{(L-3)}$
$z_3^{(L-3)}$ $a_3^{(L-3)}$
$z_4^{(L-3)}$ $a_4^{(L-3)}$
$z_5^{(L-3)}$ $a_5^{(L-3)}$
L-2
$z_0^{(L-2)}$ $a_0^{(L-2)}$
$z_1^{(L-2)}$ $a_1^{(L-2)}$
$z_2^{(L-2)}$ $a_2^{(L-2)}$
$z_3^{(L-2)}$ $a_3^{(L-2)}$
$z_4^{(L-2)}$ $a_4^{(L-2)}$
$z_5^{(L-2)}$ $a_5^{(L-2)}$
$\theta_{22}^{(L-1)}$
L-1
$z_0^{(L-1)}$ $a_0^{(L-1)}$
$z_1^{(L-1)}$ $a_1^{(L-1)}$
$z_2^{(L-1)}$ $a_2^{(L-1)}$
$z_3^{(L-1)}$ $a_3^{(L-1)}$
$z_4^{(L-1)}$ $a_4^{(L-1)}$
$z_5^{(L-1)}$ $a_5^{(L-1)}$
$\theta_{01}^{(L)}$ $\theta_{11}^{(L)}$ $\theta_{21}^{(L)}$ $\theta_{31}^{(L)}$ $\theta_{41}^{(L)}$ $\theta_{51}^{(L)}$
$\theta_{20}^{(L)}$ $\theta_{22}^{(L)}$ $\theta_{23}^{(L)}$
L
$z_0^{(L)}$ $a_0^{(L)}$
$z_1^{(L)}$ $a_1^{(L)}$
$z_2^{(L)}$ $a_2^{(L)}$
$z_3^{(L)}$ $a_3^{(L)}$
y_0 y_1 y_2 y_3
$C_{00} = (a_0^{(L)} - y_0)^2$
$C_{01} = (a_1^{(L)} - y_1)^2$
$C_{02} = (a_2^{(L)} - y_2)^2$
$C_{03} = (a_3^{(L)} - y_3)^2$

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 $\theta_{jk}^{(L)} = \theta_{jk}^{(L)} - \alpha \frac{\partial C(\theta)}{\partial \theta_{jk}^{(L)}}$. $\rightarrow$ 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Dynamic Programming!

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

- $\ell = L$ We start from layer (L). $\ell = L$. We start to calculate $\frac{\partial C_0(\Theta)}{\partial \Theta_{jk}^{(L)}}$. We accomplish this as in (9).
- $\ell = L-1$. What we accomplished in (9) i.e. $\frac{\partial C_0(\Theta)}{\partial a_j^{(L)}}$, $\frac{\partial a_j^{(L)}}{\partial z_j^{(L)}}$, we use in (11) when we calculate $\frac{\partial C_0(\Theta)}{\partial a_2^{(L-1)}}$ → which we use in (13) k-times.

Let's generalize (11). $$\boxed{\frac{\partial C_0(\Theta)}{\partial a_k^{(L-1)}} = \sum_{j=0}^{S_L-1} \frac{\partial C_{0j}}{\partial a_j^{(L)}} \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \cdot \frac{\partial z_j^{(L)}}{\partial a_k^{(L-1)}}}, \quad \forall k : 0 .. S_{L-1} \tag{15}$$

Let's generalize (9) $$\boxed{\frac{\partial C_0(\Theta)}{\partial \Theta_{kj}^{(L)}} = \frac{\partial C_0(\Theta)}{\partial a_j^{(L)}} \cdot \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \cdot \frac{\partial z_j^{(L)}}{\partial \Theta_{kj}^{(L)}}} \tag{16}$$

Let's generalize (13) $$\boxed{\frac{\partial C_0(\Theta)}{\partial \Theta_{kj}^{(L-1)}} = \frac{\partial C_0(\Theta)}{\partial a_j^{(L-1)}} \cdot \frac{\partial a_j^{(L-1)}}{\partial z_j^{(L-1)}} \cdot \frac{\partial z_j^{(L-1)}}{\partial \Theta_{kj}^{(L-1)}}} \tag{17}$$

$$\boxed{\frac{\partial C_0}{\partial \Theta_{kj}^{(\ell)}} = \frac{\partial C_0(\Theta)}{\partial a_j^{(\ell)}} \cdot \frac{\partial a_j^{(\ell)}}{\partial z_j^{(\ell)}} \cdot \frac{\partial z_j^{(\ell)}}{\partial \Theta_{kj}^{(\ell)}}} \tag{18}$$

Notes: The formulas (8) could have been written initially as $\frac{\partial C_0}{\partial \Theta_{21}^{(L)}} = \frac{\partial C_{0_1}}{\partial z_1^{(L)}} \cdot \frac{\partial z_1^{(L)}}{\partial \Theta_{21}^{(L)}}$ and then unfold as following

$$\frac{\partial C_0(\Theta)}{\partial \Theta_{21}^{(L)}} = \underbrace{\left(\frac{\partial C_{0_1}}{\partial a_1^{(L)}} \cdot \frac{\partial a_1^{(L)}}{\partial z_1^{(L)}}\right)}_{\frac{\partial C_{0_1}}{\partial z_1^{(L)}}} \cdot \frac{\partial z_1^{(L)}}{\partial \Theta_{21}^{(L)}}$$

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

i.e.:

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

Note: if $l == L$, then

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

$$\Big|\Big|\Big|$$

$$\left[\frac{\partial C_0}{\partial \Theta_{kj}^{(L)}} = \frac{\partial C_{0_j}}{\partial \Theta_{kj}^{(L)}}\right]$$

proof on earlier pages

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

Note: Let's define auxiliary variable

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

Then previous formula can be written as following:

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

P.S

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

Notes

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

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

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

$$a_k^{(L-2)} \rightarrow \boxed{\sum z_j^{(L-1)} \rightarrow a_j^{(L-1)}} \rightarrow C_0 = C(x|\theta : \theta)$$

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

Let's generalize more:

$$\frac{\partial C_0}{\partial \theta_{kj}^{(\ell)}} = \frac{\partial C_0}{\partial z_j^{\ell}} \cdot \frac{\partial z_j^{\ell}}{\partial \theta_{kj}^{\ell}}$$

When $\ell = L$, we have seen that

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

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

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

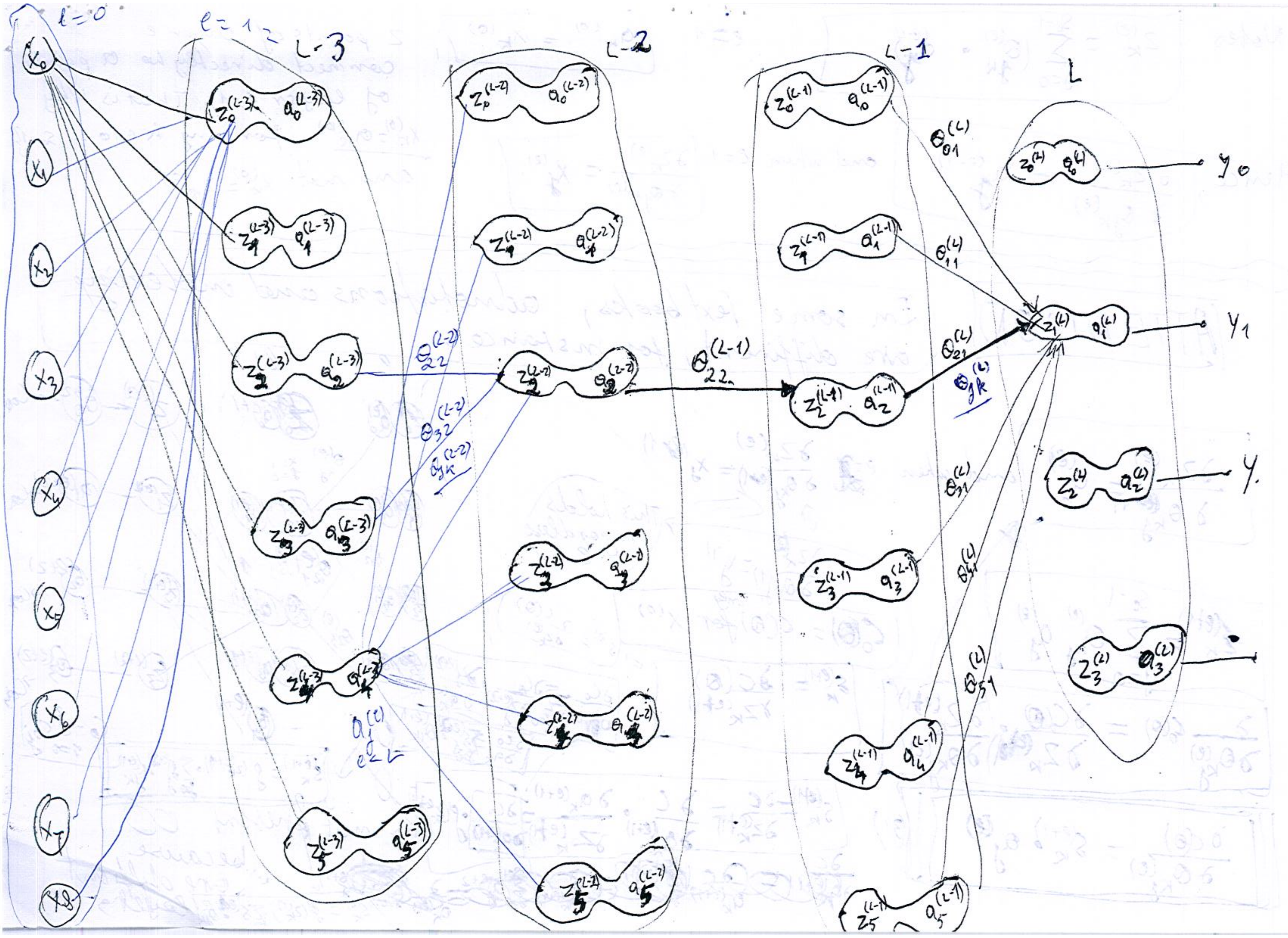
$$\boxed{\frac{\partial C_0}{\partial a_k^{(L-3)}} = \sum_{j=0}^{S_{L-2}-1} \frac{\partial C_0}{\partial a_j^{(L-2)}} \cdot \frac{\partial a_j^{(L-2)}}{\partial z_j^{(L-2)}} \cdot \frac{\partial z_j^{(L-2)}}{\partial a_k^{(L-3)}}} \quad (21)$$

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

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

~~From~~ We can pick up whatever pairs ~~of~~ consequent layrs $(\ell-1, \ell)$ and have

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



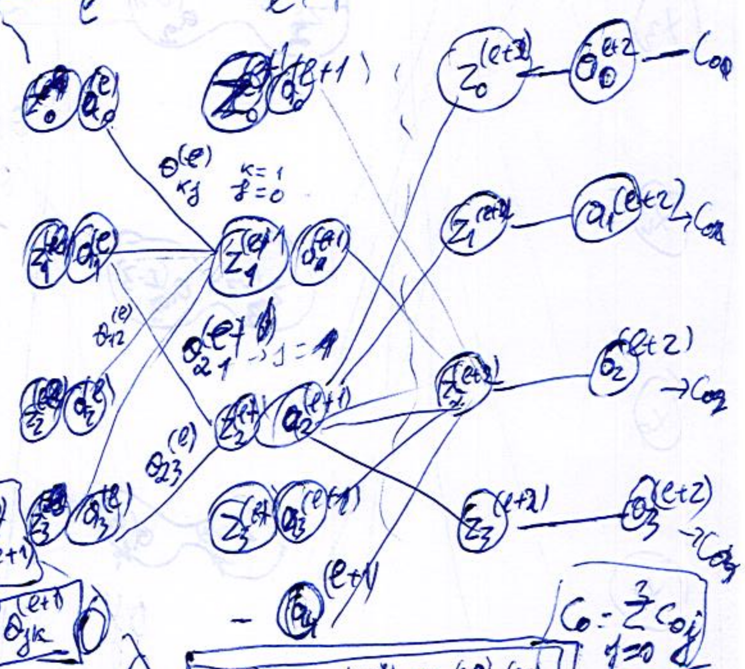
$l=0$
$l=1=L-3$
$L-2$
$L-1$
L
x_0
x_1
x_2
x_3
x_4
x_5
x_6
x_7
x_8
$z_0^{(L-3)}$ $a_0^{(L-3)}$
$z_1^{(L-3)}$ $a_1^{(L-3)}$
$z_2^{(L-3)}$ $a_2^{(L-3)}$
$z_3^{(L-3)}$ $a_3^{(L-3)}$
$z_4^{(L-3)}$ $a_4^{(L-3)}$
$z_5^{(L-3)}$ $a_5^{(L-3)}$
$a_j^{(l)}$
$l<L$
$\theta_{22}^{(L-2)}$
$\theta_{32}^{(L-2)}$
$\theta_{jk}^{(L-2)}$
$z_0^{(L-2)}$ $a_0^{(L-2)}$
$z_1^{(L-2)}$ $a_1^{(L-2)}$
$z_2^{(L-2)}$ $a_2^{(L-2)}$
$z_3^{(L-2)}$ $a_3^{(L-2)}$
$z_4^{(L-2)}$ $a_4^{(L-2)}$
$z_5^{(L-2)}$ $a_5^{(L-2)}$
$\theta_{22}^{(L-1)}$
$z_0^{(L-1)}$ $a_0^{(L-1)}$
$z_1^{(L-1)}$ $a_1^{(L-1)}$
$z_2^{(L-1)}$ $a_2^{(L-1)}$
$z_3^{(L-1)}$ $a_3^{(L-1)}$
$z_4^{(L-1)}$ $a_4^{(L-1)}$
$z_5^{(L-1)}$ $a_5^{(L-1)}$
$\theta_{01}^{(L)}$
$\theta_{11}^{(L)}$
$\theta_{21}^{(L)}$
$\theta_{jk}^{(L)}$
$\theta_{31}^{(L)}$
$\theta_{41}^{(L)}$
$\theta_{51}^{(L)}$
$z_0^{(L)}$ $a_0^{(L)}$
$z_1^{(L)}$ $a_1^{(L)}$
$z_2^{(L)}$ $a_2^{(L)}$
$z_3^{(L)}$ $a_3^{(L)}$
y_0
y_1
$y.$

Notes $\boxed{z_k^{(\ell)} = \sum_{j=0}^{S_\ell - 1} \left(\Theta_{jk}^{(\ell)} \cdot a_j^{(\ell-1)}\right)}$ $\ell = 1$ $\boxed{a_k^{(0)} = x_k^{(0)}}$; z points of layer ℓ connect directly to a points of layer $\ell - 1$. 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, $\boxed{\frac{\partial z_k^{(\ell)}}{\partial \Theta_{jk}^{(\ell)}} = a_j^{(\ell-1)}}$ and when $\ell = 1$ $\boxed{\frac{\partial z_k^{(\ell)}}{\partial \Theta_{jk}^{(\ell)}} = x_j^{(\ell)}}$

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

ℓ $\ell + 1$



$$\boxed{\frac{\partial z_k^{(\ell)}}{\partial \Theta_{kj}^{(\ell+1)}} = a_j^{(\ell)}}$$

and when $\ell = \cancel{1}$ $\frac{\partial z_k^{(\ell)}}{\partial \Theta_{kj}^{(\ell+1)}} = x_j^{(\ell+1)}$

$\Updownarrow$

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

This holds regardless ℓ.

$$\boxed{z_k^{(\ell+1)} = \sum_{j=0}^{S_\ell - 1} \Theta_{kj}^{(\ell)} a_j^{(\ell)}}$$

$\boxed{C_0(\Theta) = C(\Theta) \text{ only for } x^{(0)}}$

$\ell = 1$ $\boxed{\delta^{(\ell)} = \frac{\partial C(\Theta)}{\partial z_k^{(\ell)}}}$

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

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

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

$$\boxed{\frac{\partial}{\partial \Theta_{kj}^{(\ell)}} C_0(\Theta) = \frac{\partial C(\Theta)}{\partial z_k^{(\ell+1)}} \cdot \frac{\partial z_k^{(\ell+1)}}{\partial \Theta_{kj}^{(\ell)}}}$$

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

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

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

$$\boxed{\frac{\partial C(\Theta)}{\partial \Theta_{kj}^{(\ell)}} = \delta_k^{(\ell+1)} \cdot a_j^{(\ell)}} \quad (3'')$$

Do not present because are different layers

$\delta_k^{(\ell)} = g'(z_k^{(\ell)}) \cdot \sum \delta^{(\ell)} \cdot \Theta_{jk}^{(\ell)}$

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

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

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

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

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

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

$$\boxed{\frac{\partial C_0}{\partial \theta_{jk}^{(l)}} = \delta_k^l \cdot \frac{\partial z_k^l}{\partial \theta_{jk}^{(l)}}} \quad (3)$$

→ $\forall l \in \{L, L-1, L-2, \dots\}$

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

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

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

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

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

a) $l = L$

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

b) $l < L$

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

$\cancel{\frac{\partial C_0}{\partial a_k^{(l)}} = \sum \frac{\partial C_0}{\partial a_k^{(l+1)}} \cdot \frac{\partial a_k^{(l+1)}}{\partial z_k^{(l+1)}} \cdot \theta_{kj}^{(l)}}$

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

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

$dZ^l = dA^{l} \cdot g'(\;)$

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

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

$l = L$: $\boxed{\delta_k^l = g'(z_k^{(l)}) * (a_k^{(l)} - y_k)}$ (5')

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

b) $l < L$ $\boxed{\delta_k^l = g'(z_k^{(l)}) \cdot \sum_{j=0}^{S_{l+1}-1} \delta_j^{(l+1)} \cdot \theta_{kj}^{(l)}}$ (6) ⇒ $\boxed{\delta_k^{(l)} = g'(z_k^{(l)}) .* \theta_k^{(l)} \cdot \delta^{(l+1)}}$ → vectorized form

in different adnotations

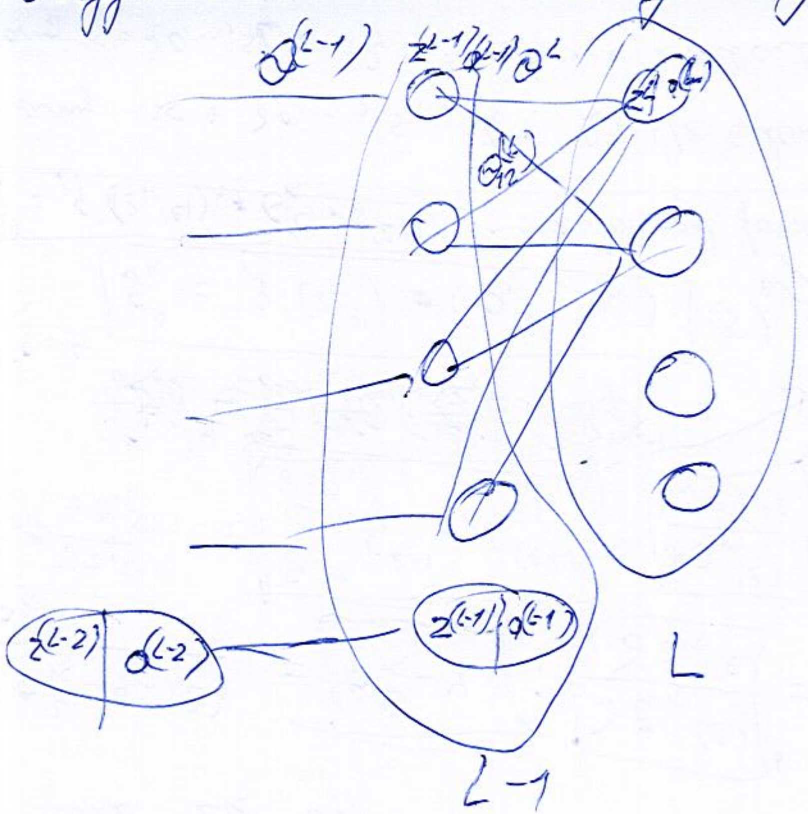
Algorithm:

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 calculate $\frac{\partial C_0}{\partial \theta_{jk}^{(l)}}$ as in (3')

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

it we use slightly different adnotations of layers for a



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

— — — —

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

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

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

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

$$\boxed{dz^{(l)} = (\underset{\text{standard matrix-multiplication (dot product)}}{\theta^{l+1} \cdot dz^{(l+1)}}) \underset{\text{elementwise}}{.*} g(z^{(l)})}$$

→ A.w.r.t. Andrews Adnotation

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

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

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