ANNEX 2: Gradient descent in multi-layer networks

- We define an error function to minimize:

$$E(y) = \frac{1}{2} . (r - y)^2$$

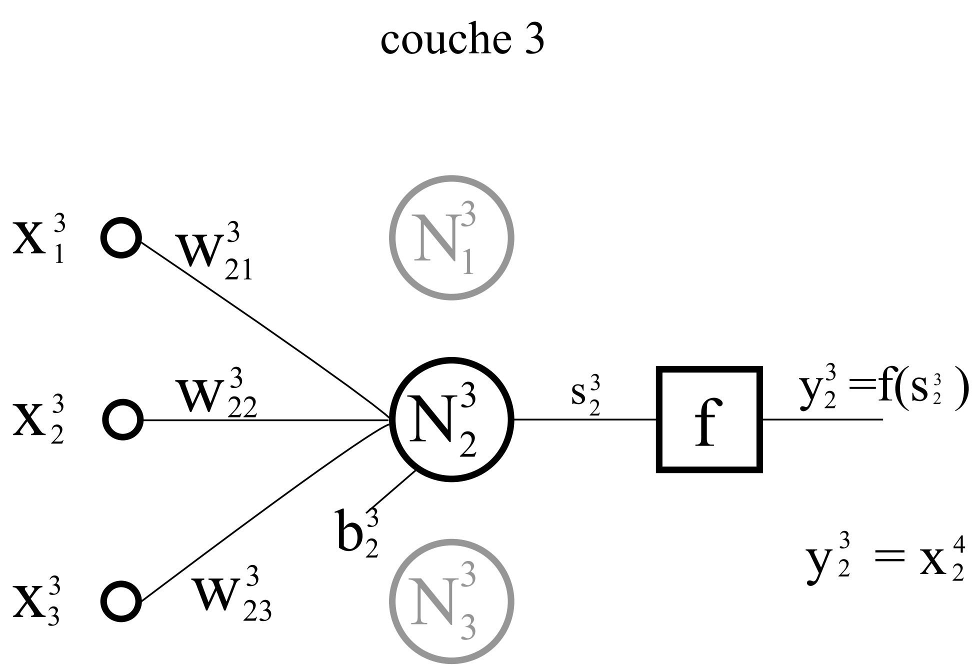
- For each weight w_{ij}^l , we characterize the impact of a change on E
  - Gradient descent on w_{ij}^l

$$w_{ij}^l \Leftarrow w_{ij}^l - \alpha . \frac{\partial E(y)}{\partial w_{ij}^l}$$

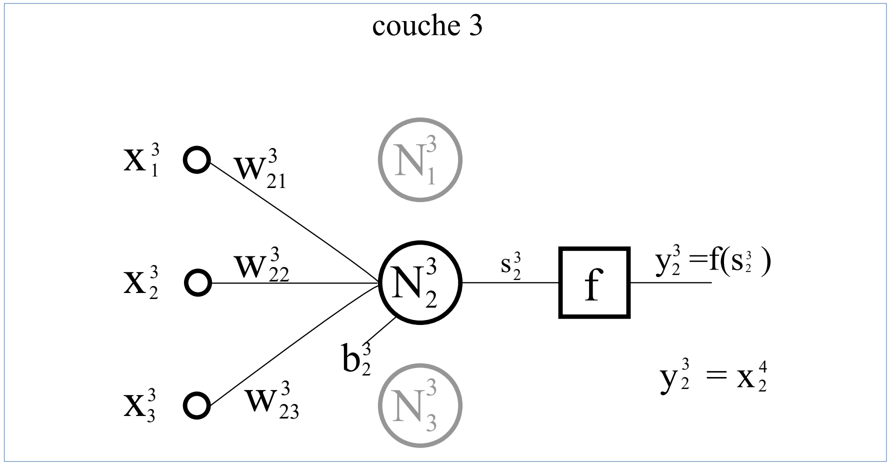
  - We have to find: $\frac{\partial E(y)}{\partial w_{ij}^l}$

ANNEX 2: Gradient descent in multi-layer networks

- **Gradient descent: notations**
  - We note l a layer, i the index of a neuron, L the last (output) layer.
  - A weight w_{ij}^l connects neuron N_i^l and neuron N_j^{l-1} ($N_j^{l-1} \rightarrow N_i^l$)
  - We note $s_i^l = s(W_i^l, X^l, b_i^l) = \Sigma\ w_{ij}^l \cdot x_i^l + b_i^l$ the weighted sum of neuron N_i^l of layer l.
  - We note $y_i^l = f(s_i^l)$ the output of neuron N_i^l after activation function f



ANNEX 2: Gradient descent in multi-layer networks



- **Gradient descent:**
  case of last layer L
  - We must find the term:

$$\frac{\partial E\left(y_i^L\right)}{\partial w_{ij}^L}$$

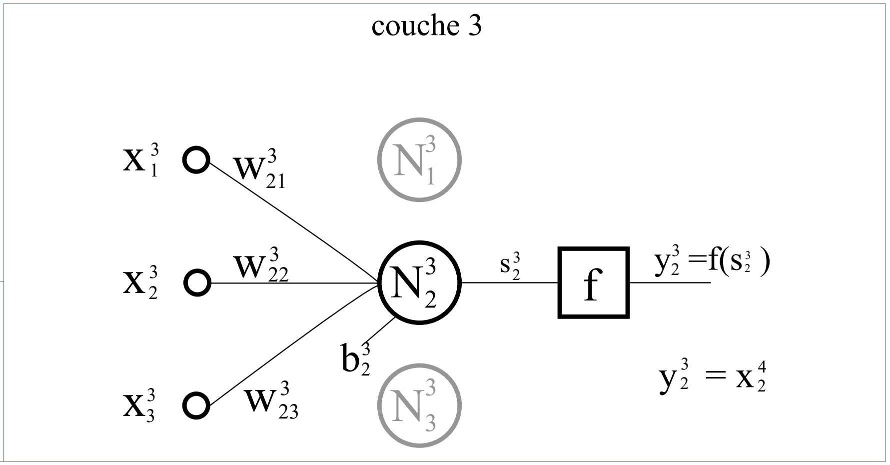
  - We can decompose with intermediate values w → s → y → E
    - Theorem of composed derivative functions

$$\frac{\partial E\left(y_i^L\right)}{\partial w_{ij}^L} = \frac{\partial E\left(y_i^L\right)}{\partial y_i^L} \cdot \frac{\partial y_i^L}{\partial s_i^L} \cdot \frac{\partial s_i^L}{\partial w_{ij}^L}$$

    - Three derivatives to find

ANNEX 2: Gradient descent in multi-layer networks

- **Gradient descent:**
  case of last layer L



$$\frac{\partial E(y_i^L)}{\partial w_{ij}^L} = \frac{\partial E(y_i^L)}{\partial y_i^L} \cdot \frac{\partial y_i^L}{\partial s_i^L} \cdot \frac{\partial s_i^L}{\partial w_{ij}^L}$$

- First term: $\dfrac{\partial E(y_i^L)}{\partial y_i^L}$

Derivation of composed functions:
(f o g)’ = (f’ o g).g’

- Error function: $E(y) = \dfrac{1}{2}.(r - y)^2$

$$\frac{\partial E(y)}{\partial y} = \frac{1}{2}.\frac{\partial (r-y)^2}{\partial y} = \cancel{\frac{2}{2}}.(r-y).\frac{\partial (r-y)}{\partial y}$$

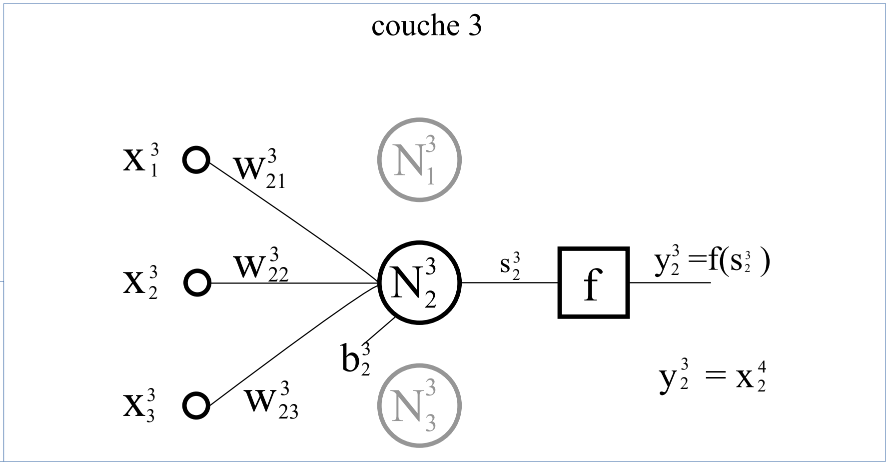
-1

$$\boxed{\frac{\partial E(y)}{\partial y} = -(r-y)}$$

ANNEX 2: Gradient descent in multi-layer networks

- **Gradient descent:**
  case of last layer L

$$\frac{\partial E(y_i^L)}{\partial w_{ij}^L} = \underline{\frac{\partial E(y_i^L)}{\partial y_i^L}} \cdot \frac{\partial y_i^L}{\partial s_i^L} \cdot \frac{\partial s_i^L}{\partial w_{ij}^L}$$



- second term: $\dfrac{\partial y_i^L}{\partial s_i^L} = \dfrac{\partial f(s_i^L)}{\partial s_i^L}$

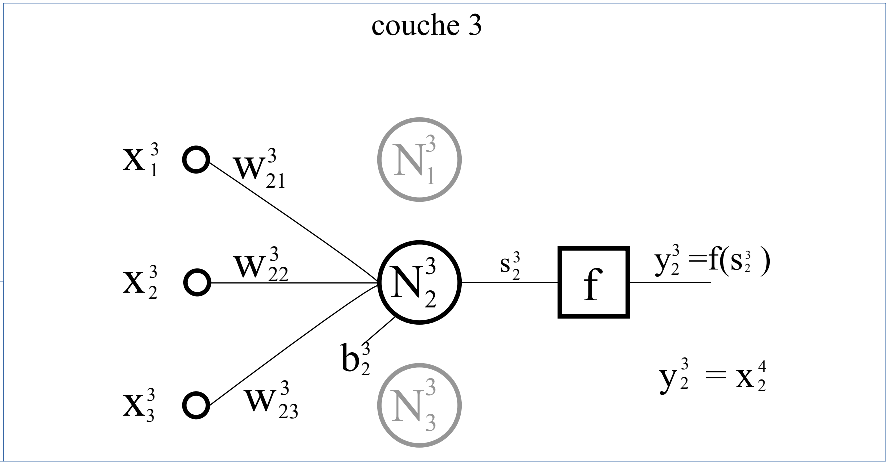
- It is just the derivative of activation function !

$$\boxed{\frac{\partial y_i^L}{\partial s_i^L} = f'_{(s_i^L)}}$$

ANNEX 2: Gradient descent in multi-layer networks

- **Gradient descent:**
  case of last layer L

$$\frac{\partial E(y_i^L)}{\partial w_{ij}^L} = \frac{\partial E(y_i^L)}{\partial y_i^L} \cdot \frac{\partial y_i^L}{\partial s_i^L} \cdot \frac{\partial s_i^L}{\partial w_{ij}^L}$$



- third term: $\dfrac{\partial s_i^L}{\partial w_{ij}^L}$

'constants'

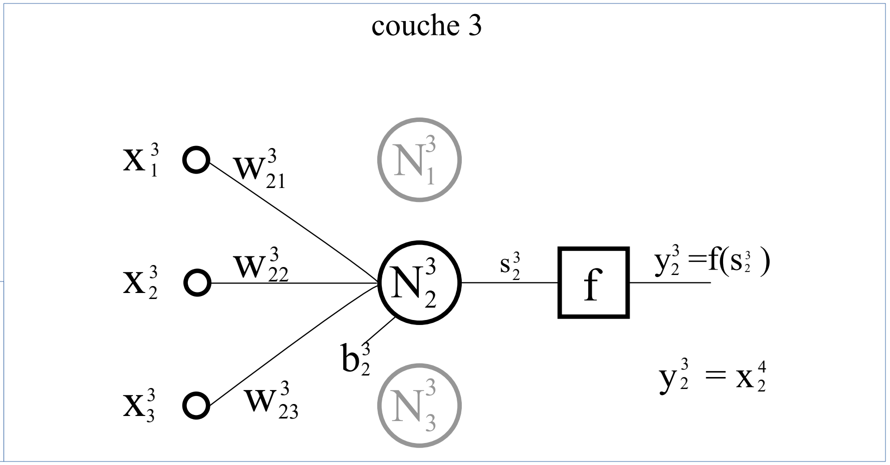
$$\frac{\partial s_i^L}{\partial w_{ij}^L} = \frac{\partial \left(w_{i1}^L . x_1^L + w_{i2}^L . x_2^L + \ldots + w_{ij}^L . x_i^L + \ldots + b_i^L\right)}{\partial w_{ij}^L} = \frac{\partial \left(w_{ij}^L . x_j^L\right)}{\partial w_{ij}^L}$$

$$\boxed{\frac{\partial s_i^L}{\partial w_{ij}^L} = x_j^L}$$

ANNEX 2: Gradient descent in multi-layer networks

- **Gradient descent:**
  case of last layer L

$$\frac{\partial E(y_i^L)}{\partial w_{ij}^L} = \frac{\partial E(y_i^L)}{\partial y_i^L} \cdot \frac{\partial y_i^L}{\partial s_i^L} \cdot \frac{\partial s_i^L}{\partial w_{ij}^L}$$



- We summarize:

$$\frac{\partial E(y_i^L)}{\partial w_{ij}^L} = -(r - y_i^L) \; . \; f'_{(s_i^L)} \; . \; x_j^L$$

- Gradient descent: $\quad w_{ij}^L \Leftarrow w_{ij}^L + \alpha \, . \, x_j^L \, . \, (r - y_i^L) \, . \, f'_{(s_i^L)}$

- We set $\quad \delta_i^L = (r - y_i^L) \; . \; f'_{(s(\ldots))}$

$$\rightarrow \boxed{w_{ij}^L \Leftarrow w_{ij}^L + \alpha \, . \, x_j^L \, . \, \delta_i^L}$$

Note :

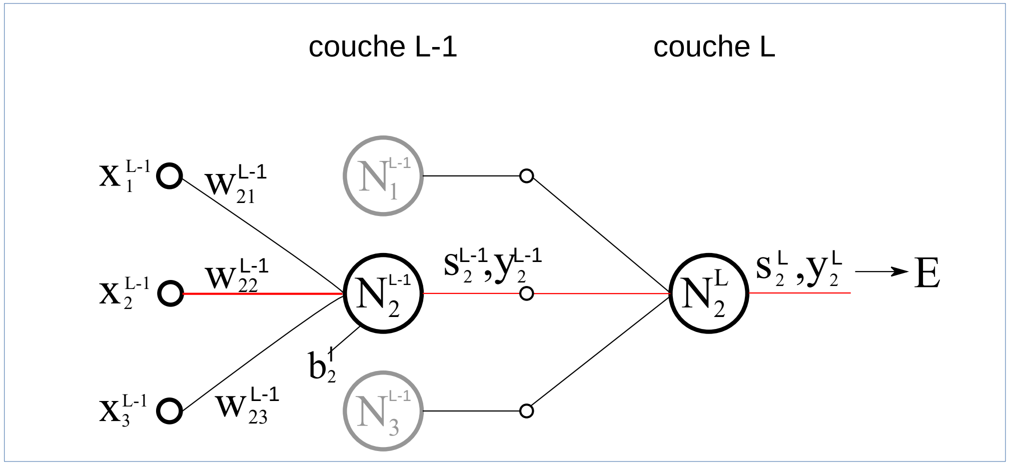
$$\delta_i^L = -\frac{\partial E(y_i^L)}{\partial s_i^L}$$

ANNEX 2: Gradient descent in multi-layer networks

- **Gradient descent:**
  case of hidden layer L-1
  - We must find

$$\frac{\partial E\left(y_m^L\right)}{\partial w_{ij}^{L-1}}$$

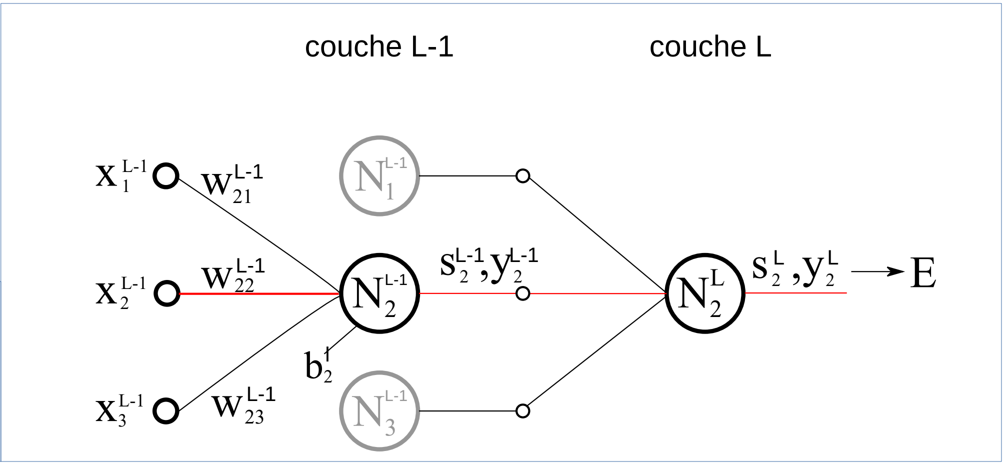
  - We decompose again:



$$\frac{\partial E\left(y_m^L\right)}{\partial w_{ij}^{L-1}} = \frac{\partial E\left(y_m^L\right)}{\partial y_i^{L-1}} \cdot \frac{\partial y_i^{L-1}}{\partial s_i^{L-1}} \cdot \frac{\partial s_i^{L-1}}{\partial w_{ij}^{L-1}}$$

ANNEX 2: Gradient descent in multi-layer networks

- **Gradient descent:**
  case of hidden layer L-1



$$\frac{\partial E(y_m^L)}{\partial w_{ij}^{L-1}} = \frac{\partial E(y_m^L)}{\partial y_i^{L-1}} \cdot \frac{\partial y_i^{L-1}}{\partial s_i^{L-1}} \cdot \frac{\partial s_i^{L-1}}{\partial w_{ij}^{L-1}}$$

- Second and third terms are the same:

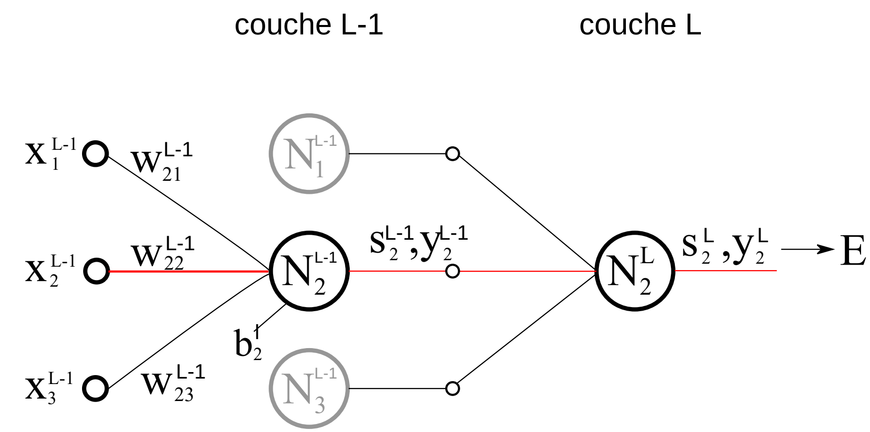
$$\boxed{\frac{\partial y_i^{L-1}}{\partial s_i^{L-1}} = f'_{(s_i^{L-1})}}$$

$$\boxed{\frac{\partial s_i^{L-1}}{\partial w_{ij}^{L-1}} = x_j^{L-1}}$$

ANNEX 2: Gradient descent in multi-layer networks

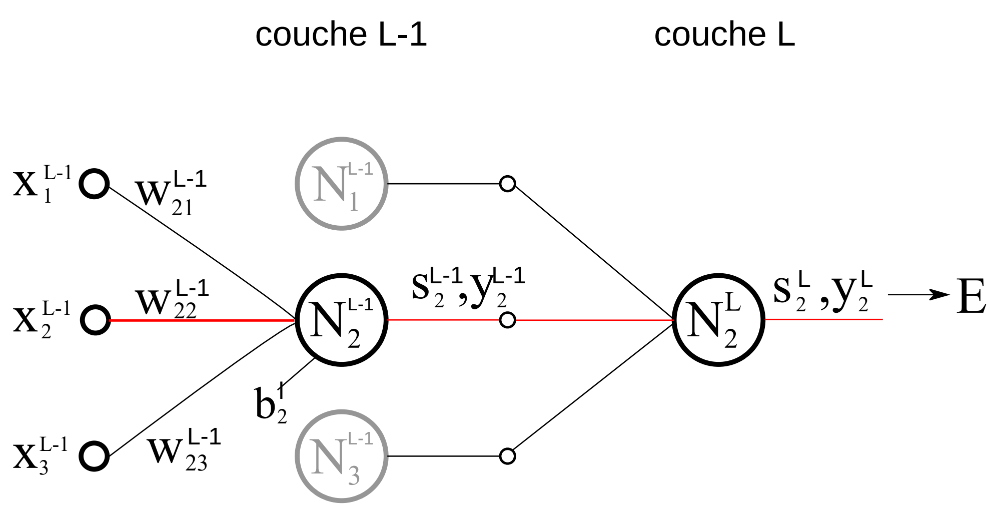
- **Gradient descent:**
  case of hidden layer L-1

$$\frac{\partial E\left(y_m^L\right)}{\partial w_{ij}^{L-1}} = \frac{\partial E\left(y_m^L\right)}{\partial y_i^{L-1}} \cdot \frac{\partial y_i^{L-1}}{\partial s_i^{L-1}} \cdot \frac{\partial s_i^{L-1}}{\partial w_{ij}^{L-1}}$$



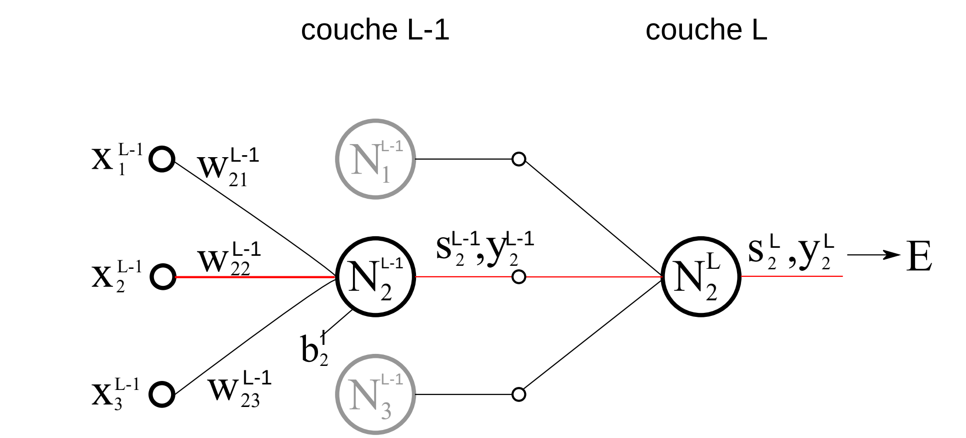
- First term: $\frac{\partial E\left(y_m^L\right)}{\partial y_i^{L-1}}$

- More complex: if y_i^{L-1} is modified, E is modified through s_i^{L-1} (and thus y_j^{L-1}) of layer L



ANNEX 2: Gradient descent in multi-layer networks

- **Gradient descent:**
  case of hidden layer L-1



$$\frac{\partial E(y_m^L)}{\partial w_{ij}^{L-1}} = \frac{\partial E(y_m^L)}{\partial y_i^{L-1}} \cdot \frac{\partial y_i^{L-1}}{\partial s_i^{L-1}} \cdot \frac{\partial s_i^{L-1}}{\partial w_{ij}^{L-1}}$$

$$\frac{\partial E(y_m^L)}{\partial y_i^{L-1}} = \frac{\partial E(y_m^L)}{\partial s_m^L} \cdot \frac{\partial s_m^L}{\partial y_i^{L-1}}$$

It is $-\delta_m^L$!

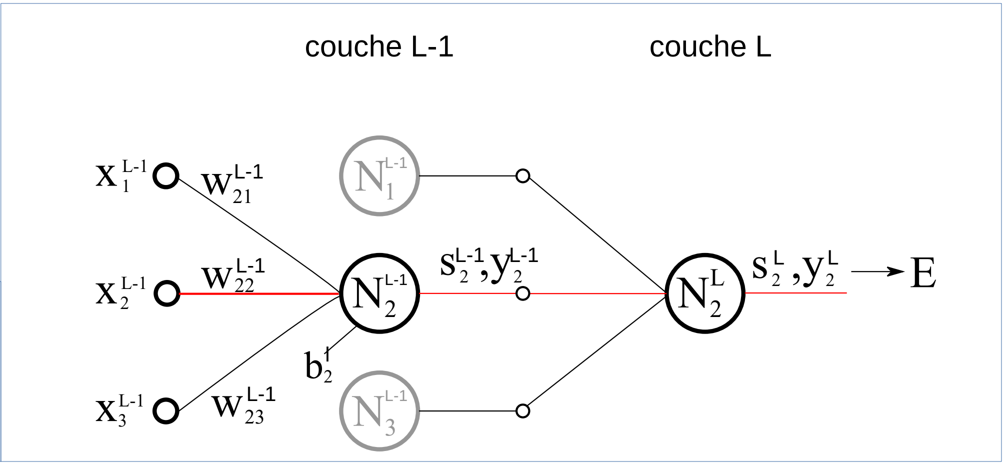
$$\frac{\partial s_m^L}{\partial y_i^{L-1}} = \frac{\partial (w_{m1}^L . x_1^L + w_{m2}^L . x_2^L + \ldots + w_{mi}^L . x_i^L + \ldots + b_m^L)}{\partial x_i^L} = w_{mi}^L$$

$$\boxed{\frac{\partial E(y_m^l)}{\partial y_i^{L-1}} = -\delta_m^L \; . \; w_{mi}^L}$$

ANNEX 2: Gradient descent in multi-layer networks

- **Gradient descent:**
  case of hidden layer L-1

$$\frac{\partial E(y_m^L)}{\partial w_{ij}^{L-1}} = \frac{\partial E(y_m^L)}{\partial y_i^{L-1}} \cdot \frac{\partial y_i^{L-1}}{\partial s_i^{L-1}} \cdot \frac{\partial s_i^{L-1}}{\partial w_{ij}^{L-1}}$$



- We summarize:

$$\frac{\partial E(y_i^L)}{\partial w_{ij}^{L-1}} = -\; f'_{(s_i^{L-1})} \;.\; x_j^{L-1} \;.\; \delta_m^L \;.\; w_{mi}^L$$

- Gradient descent: $w_{ij}^{L-1} \Leftarrow w_{ij}^{L-1} + \alpha . x_j^{L-1} . f'_{s_i^{L-1}} \;.\; \delta_m^L \;.\; w_{mi}^L$

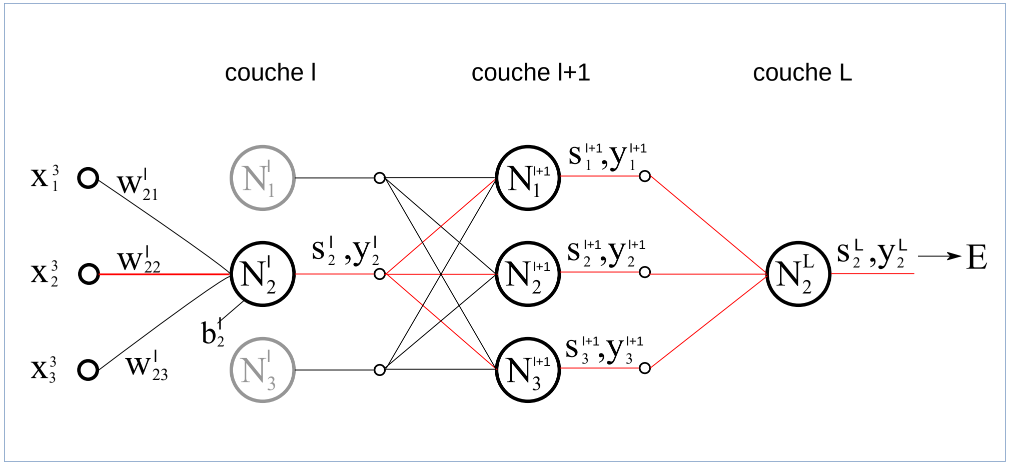
- We set: $\delta_i^{L-1} = f'_{s_i^{L-1}} \;.\; \delta_m^L \;.\; w_{mi}^L$

Note:

$$\delta_i^{L-1} = -\frac{\partial E(y_m^L)}{\partial s_i^{L-1}}$$

And again: $\boxed{w_{ij}^{L-1} \Leftarrow w_{ij}^{L-1} + \alpha . x_j^{L-1} . \delta_i^{L-1}}$

ANNEX 2: Gradient descent in multi-layer networks



- **Gradient descent:**

  Generalization for layers l

  - We must find:

$$\frac{\partial E\left(y_m^L\right)}{\partial w_{ij}^l}$$

  - We decompose again:

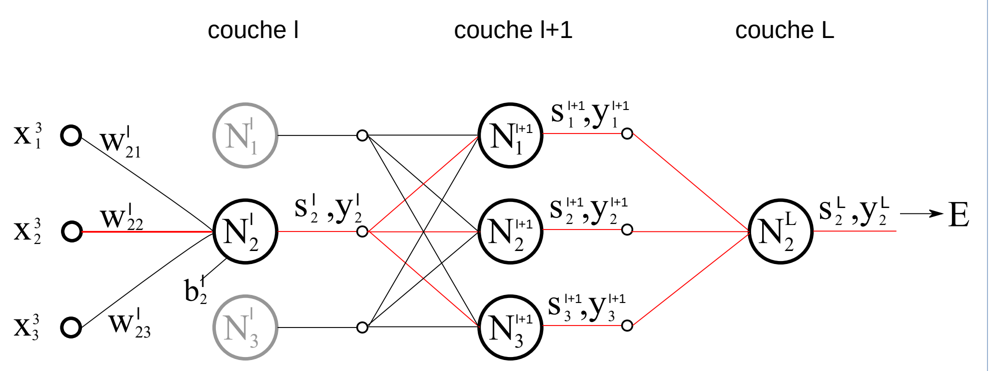
$$\frac{\partial E\left(y_m^L\right)}{\partial w_{ij}^l} = \frac{\partial E\left(y_m^L\right)}{\partial y_i^l} \cdot \frac{\partial y_i^l}{\partial s_i^l} \cdot \frac{\partial s_i^l}{\partial w_{ij}^l}$$

ANNEX 2: Gradient descent in multi-layer networks

- **Gradient descent:**

  Generalization for layers l

$$\frac{\partial E(y_m^L)}{\partial w_{ij}^l} = \frac{\partial E(y_m^L)}{\partial y_i^l} \cdot \frac{\partial y_i^l}{\partial s_i^L} \cdot \frac{\partial s_i^L}{\partial w_{ij}^l}$$



- Second and third terms are the same (again):

$$\boxed{\frac{\partial y_i^l}{\partial s_i^l} = f'_{(s_i^l)}}$$

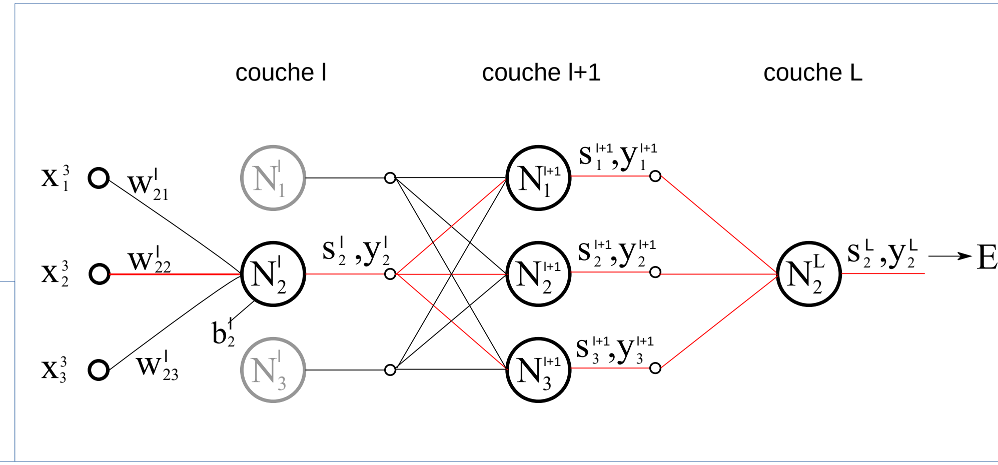
$$\boxed{\frac{\partial s_i^l}{\partial w_{ij}^l} = x_j^l}$$

ANNEX 2: Gradient descent in multi-layer networks

- **Gradient descent:**

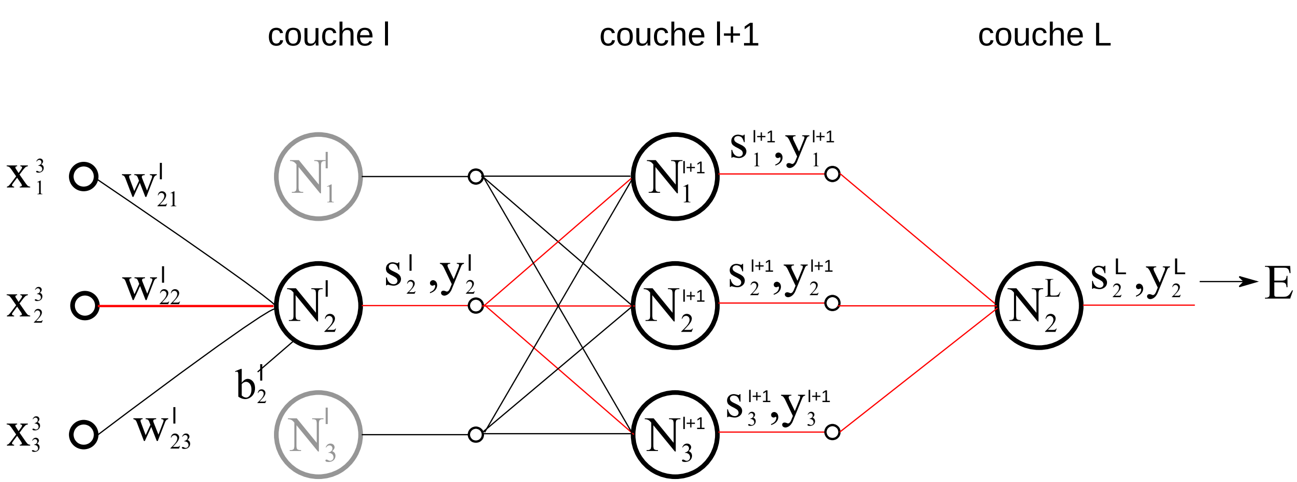
  Generalization for layers l

$$\frac{\partial E(y_m^L)}{\partial w_{ij}^l} = \frac{\partial E(y_m^L)}{\partial y_i^l} \cdot \frac{\partial y_i^l}{\partial s_i^L} \cdot \frac{\partial s_i^L}{\partial w_{ij}^l}$$



- First term: $\dfrac{\partial E(y_m^L)}{\partial y_i^l}$

- This time, if y_i^l is modified, E is modified through outputs s (and thus y_j^{l+1}) of next layer l+1

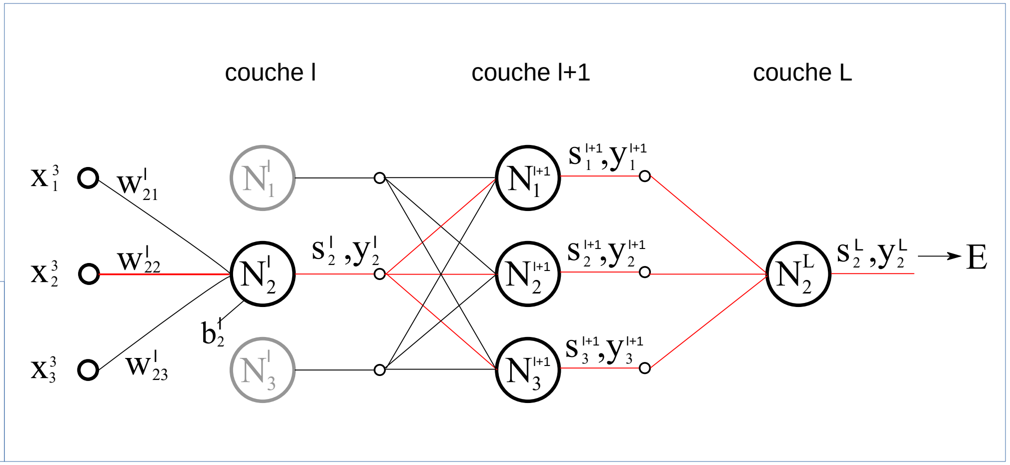


ANNEX 2: Gradient descent in multi-layer networks

- **Gradient descent:**

  Generalization for layers l

$$\frac{\partial E(y_m^L)}{\partial w_{ij}^l} = \frac{\partial E(y_m^L)}{\partial y_i^l} \cdot \frac{\partial y_i^l}{\partial s_i^L} \cdot \frac{\partial s_i^L}{\partial w_{ij}^l}$$



$$\frac{\partial E(y_m^L)}{\partial y_i^l} = \frac{\partial E(y_m^L)}{\partial s_1^{l+1}} \cdot \frac{\partial s_1^{l+1}}{\partial y_i^l} + \frac{\partial E(y_m^L)}{\partial s_2^{l+1}} \cdot \frac{\partial s_2^{l+1}}{\partial y_i^l} + \dots + \frac{\partial E(y_m^L)}{\partial s_n^{l+1}} \cdot \frac{\partial s_n^{l+1}}{\partial y_i^l}$$

$$\frac{\partial E(y_m^l)}{\partial y_i^l} = \sum_{k \in [1,n]} \frac{\partial E(y_m^L)}{\partial s_k^{l+1}} \cdot \frac{\partial s_k^{l+1}}{\partial y_i^l}$$

w_{ki}^{l+1}

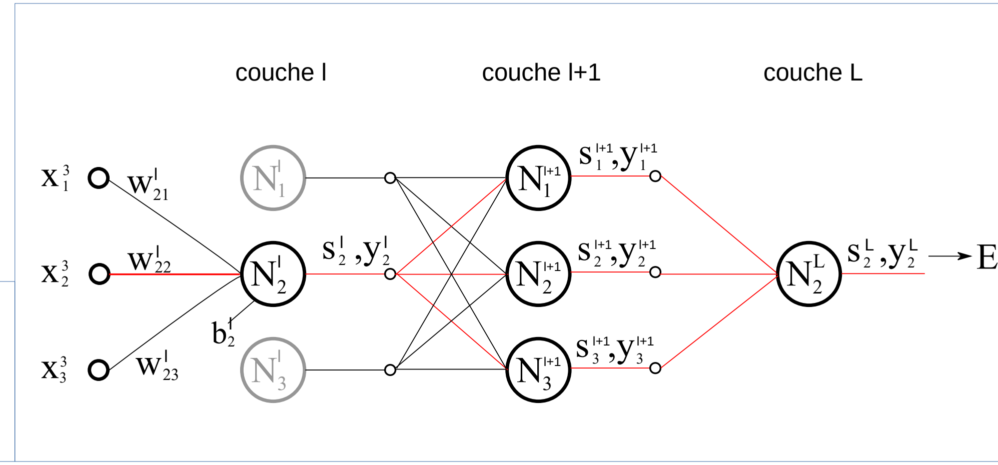
These are $-\delta_k^{l+1}$!

ANNEX 2: Gradient descent in multi-layer networks

- **Gradient descent:**

  Generalization for layers l

$$\frac{\partial E(y_m^L)}{\partial w_{ij}^l} = \frac{\partial E(y_m^L)}{\partial y_i^l} \cdot \frac{\partial y_i^l}{\partial s_i^L} \cdot \frac{\partial s_i^L}{\partial w_{ij}^l}$$



- We summarize (a last time):

$$\frac{\partial E(y_i^L)}{\partial w_{ij}^l} = -f'_{(s_i^l)} \cdot x_j^l \cdot \sum_{k\in[1,n]} \delta_k^{l+1} \cdot w_{ki}^{l+1}$$

- Gradient descent: $$w_{ij}^l \Leftarrow w_{ij}^l + \alpha . x_j^l . f'_{(s_i^l)} \cdot \sum_{k\in[1,n]} \delta_k^{l+1} . w_{ki}^{l+1}$$

- We set: $$\delta_i^l = f'_{s_i^l} \cdot \sum_{k\in[1,n]} \delta_k^{l+1} \cdot w_{ki}^{l+1}$$

And again: $$\boxed{w_{ij}^l \Leftarrow w_{ij}^l + \alpha . x_j^l . \delta_i^l}$$