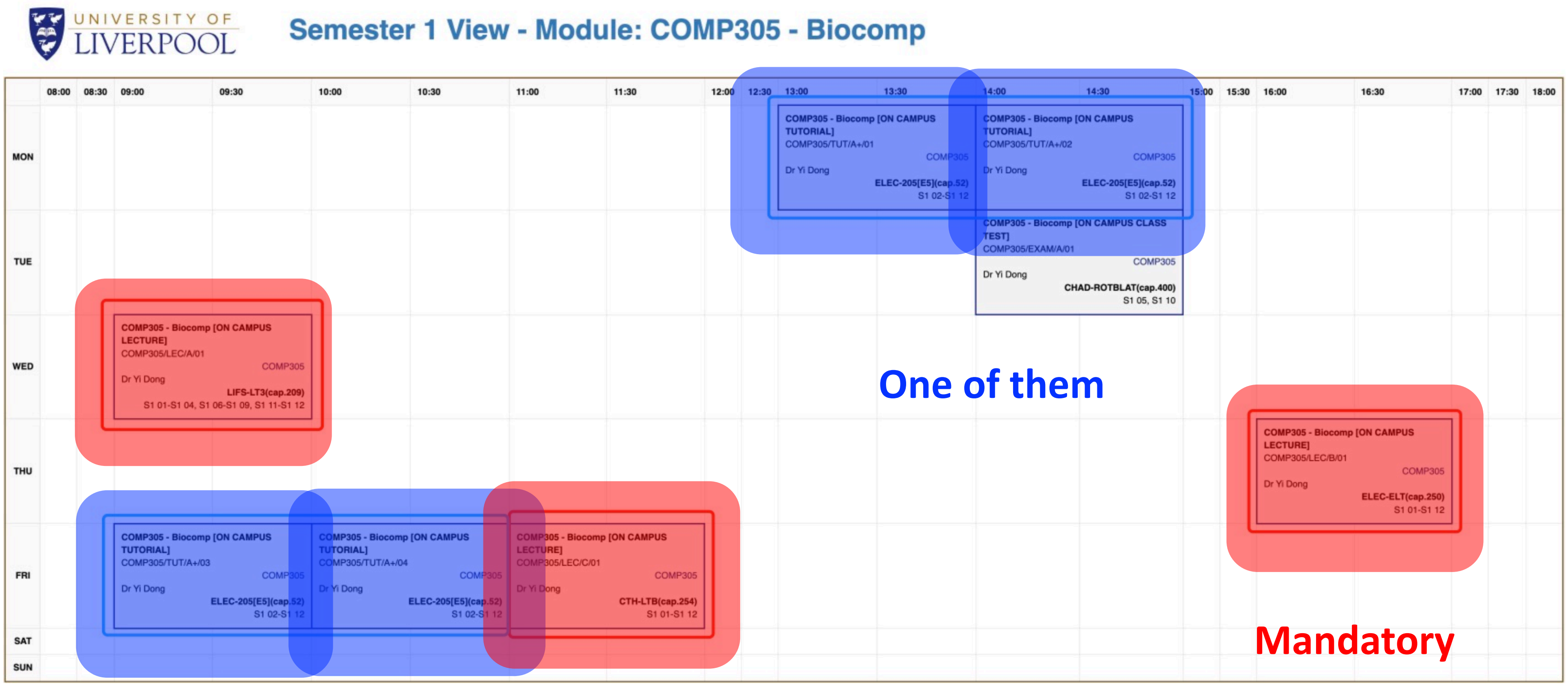
Comp305

Biocomputation

Lecturer: Yi Dong

Comp305 Module Timetable



There will be **26-30** lectures, thee per week. The lecture slides will appear on Canvas. Please use Canvas to access the lecture information. There will be **9** tutorials, one per week.

Lecture/Tutorial Rules

Questions are welcome as soon as they arise, because

1. Questions give feedback to the lecturer;

2. Questions help your understanding;

3. Your questions help your classmates, who might experience difficulties with formulating the same problems/doubts in the form of a question.

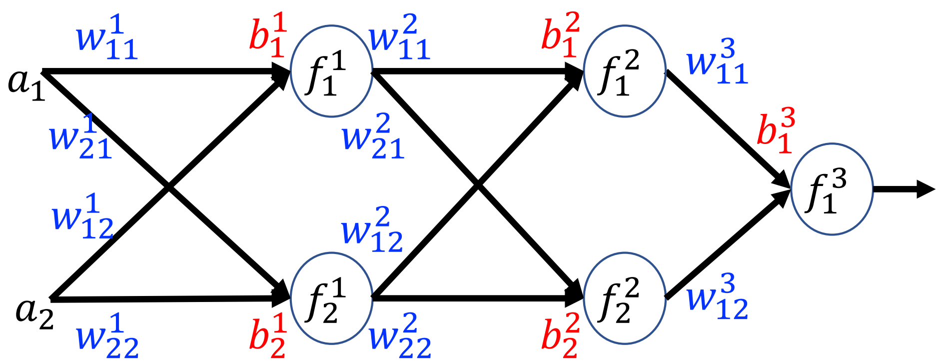
Comp305 Part I.

Artificial Neural Networks

Topic 5.

Multilayer Perceptron

Forward Propagation

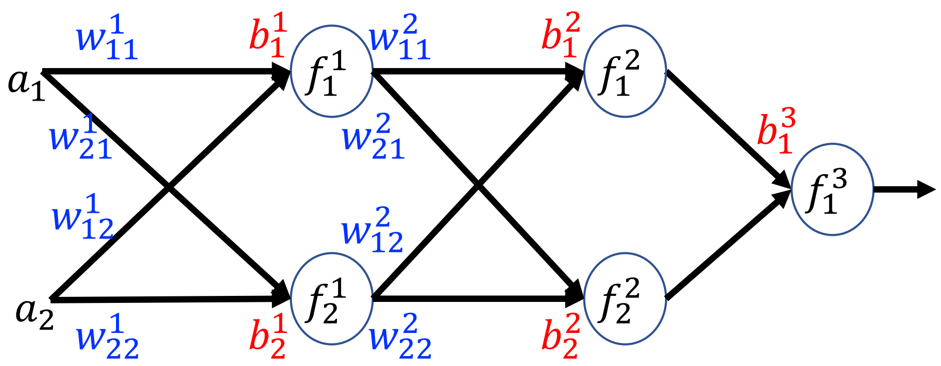


l: the number of layers,
n^h: the number of neurons in the h-th layer
$n = n^0$: the number of input neurons (0-th layer).
$m = n^l$: the number of output neurons (l-th layer).
X^h: the output value of the h-th layer.
$a = X^0$: the input value of the MLP.
$X = X^l$: the output value of the MLP.
$f^h\colon\mathbb{R}^{n_h} \to \mathbb{R}^{n_h}$: activation function of the h-th layer

Similarly, we can derive the relation for the following layers:

$$X^1 = F^1(w^1, X^0)$$
$$X^2 = F^2(w^2, X^1)$$
$$X^3 = F^3(w^3, X^2)$$
$$\cdots$$
$$X^l = F^l\left(w^l, X^{l-1}\right)$$

Learning of a Multilayer Perceptron



The output error function E^k for the k-th input pattern is:

$$E^k = \frac{1}{2}\sum_{j=1}^{m}\left(t_j^k - X_j^k\right)^2,$$

The **MLP error function** E **is :**

$$E = \frac{1}{2}\sum_{k=1}^{r}\sum_{j=1}^{m}\left(t_j^k - F_j\left(w^l, w^{l-1}, \cdots, w^1, a^k\right)\right)^2$$

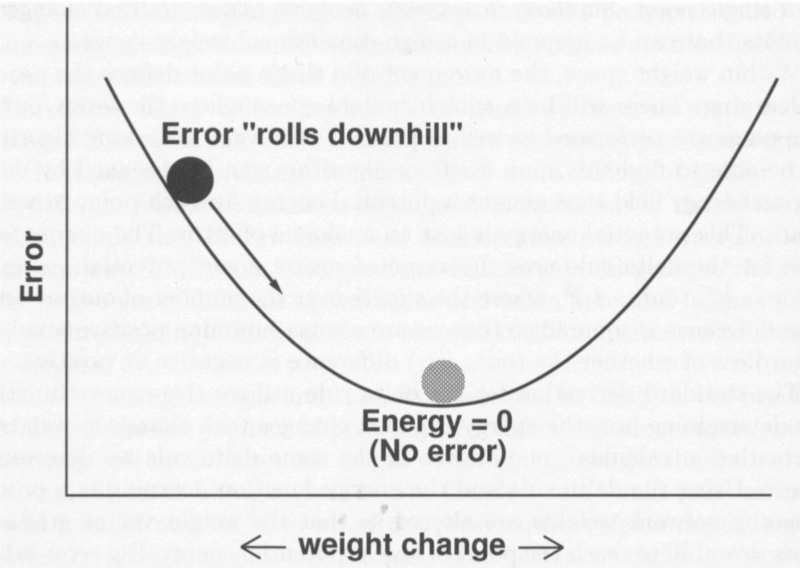
One of the most popular techniques is called

error backpropagation,

where the error of output neurons is propagated back to derive the weight adjustment of a given hidden neuron, based on how much the neuron contributes to the output error.

The ***backpropagation*** algorithm looks for the minimum of the error function E in the space of weights of connections w using the **method of *gradient descent.***

Learning of a Multilayer Perceptron



The **MLP error function** E **is :**

$$E = \frac{1}{2}\sum_{k=1}^{r}\sum_{j=1}^{m}\left(t_j^k - F_j\left(w^l, w^{l-1}, \cdots, w^1, a^k\right)\right)^2$$

Gradient descent method: a differentiable $F(x)$ decreases fastest if one goes from a in the direction of the negative gradient of F at a, $-\nabla F(a)$. It follows that, if

$$a' = a + \gamma\left(-\nabla F(a)\right) = a - \gamma \nabla F(a)$$

For a $\gamma \in \mathbb{R}_+$ small enough, then $F(a) \geq F(a')$

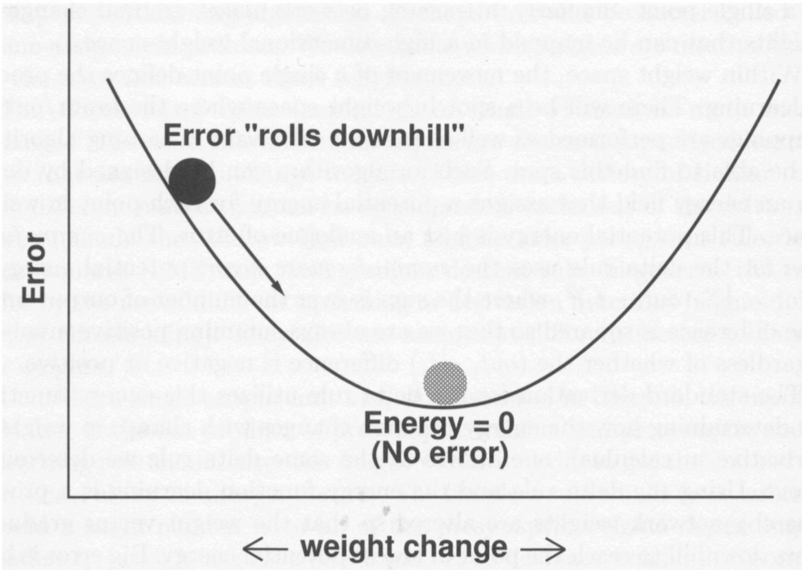
The ***gradient*** of E is:

$$\nabla E = \left(\frac{\partial E}{\partial w_{11}^1}, \cdots, \frac{\partial E}{\partial w_{n^1 n^0}^1}, \frac{\partial E}{\partial w_{11}^2}, \cdots, \frac{\partial E}{\partial w_{n^2 n^1}^2}, \cdots, \frac{\partial E}{\partial w_{11}^l}, \cdots, \frac{\partial E}{\partial w_{n^l n^{l-1}}^l}\right)$$

So based on the Gradient descent method, the weight updating policy should be

$$w = w - C\nabla E(w)$$

Learning of a Multilayer Perceptron



The **MLP error function** E **is :**

$$E = \frac{1}{2}\sum_{k=1}^{r}\sum_{j=1}^{m}\left(t_j^k - F_j\left(w^l, w^{l-1}, \cdots, w^1, a^k\right)\right)^2$$

Following calculus, a local minimum of a function of two or more variables is defined by equality to zero of its ***gradient:***

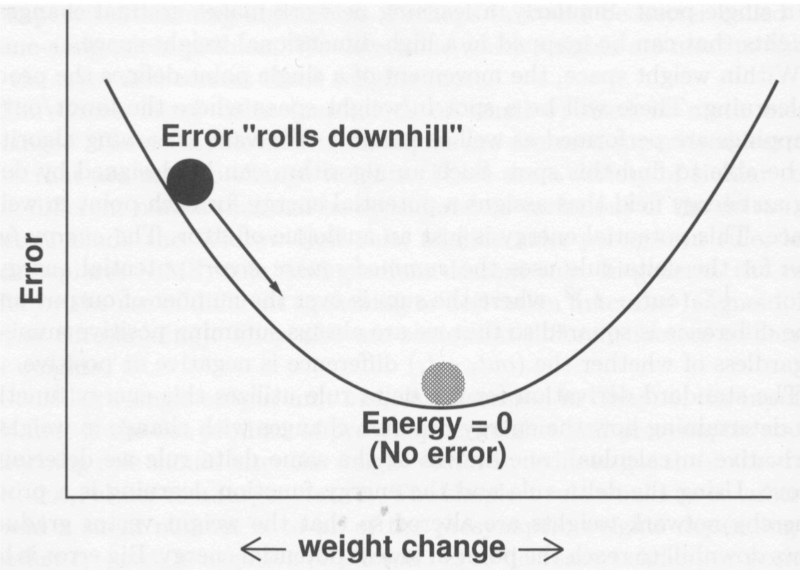
$$\nabla E = \left(\frac{\partial E}{\partial w_{11}^1}, \cdots, \frac{\partial E}{\partial w_{n^1 n^0}^1}, , \cdots, \frac{\partial E}{\partial w_{11}^l}, \cdots, \frac{\partial E}{\partial w_{n^l n^{l-1}}^l}\right)$$

Therefore, during the ***iterative process*** **of** ***gradient descent*** **each weight of connection, including the hidden ones, is updated:**

$$w_{ji}^h = w_{ji}^h + \Delta w_{ji}^h, \text{ where } \Delta w_{ji}^h = -C\frac{\partial E}{\partial w_{ji}^h}$$

Here C represents the learning rate as before.

Learning of a Multilayer Perceptron



This provides a powerful motivation for using ***continuous and differentiable activation functions f.***

Generic sigmoidal activation function :

$$f(S) = \frac{\alpha}{1 + e^{-\beta S + \gamma}} + \lambda$$

Its derivative is:

$$f'(S) = \frac{df}{dS} = \frac{\beta}{\alpha} \cdot (f(S) + \lambda)\big(\alpha + \lambda - f(S)\big)$$

Update rule:

$$w_{ji}^{h} = w_{ji}^{h} + \Delta w_{ji}^{h},$$

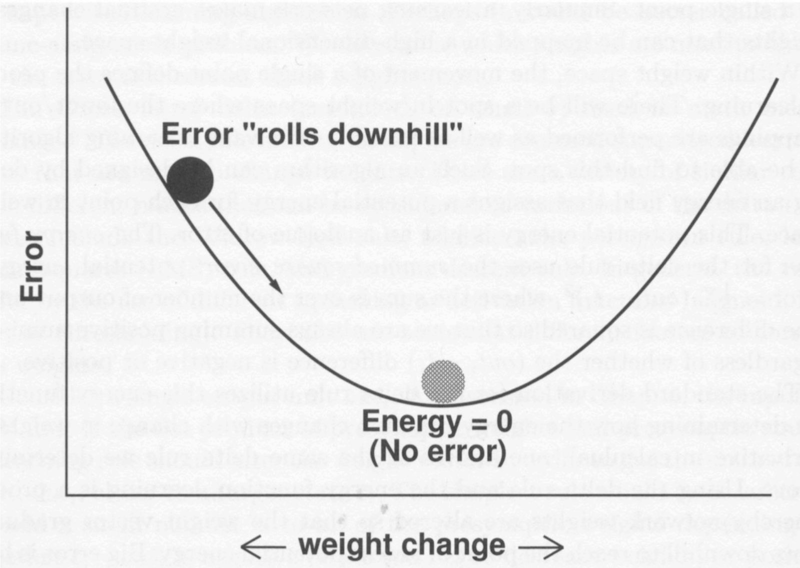
$$\text{where } \Delta w_{ji}^{h} = -C \frac{\partial E}{\partial w_{ji}^{h}}$$

If all activation functions $f(S)$ in the network are differentiable then, according to the *chain rule* of calculus, differentiating the error function E with respect to the weight of connection in consideration we can express the corresponding partial derivative of the error function.

Topic of Today's Lecture

Calculation of the partial derivative of the error function with respective to a specific weight.

Learning of a Multilayer Perceptron



Update rule:

$$w_{ji}^h = w_{ji}^h + \Delta w_{ji}^h,$$

$$\text{where } \Delta w_{ji}^h = -C\frac{\partial E}{\partial w_{ji}^h}$$

The **MLP error function** E **is :**

$$E = \frac{1}{2}\sum_{k=1}^{r}\sum_{j=1}^{m}\left(t_j^k - F_j\left(w^l, w^{l-1}, \cdots, w^1, a^k\right)\right)^2$$

The **error function** E **for a single input**:

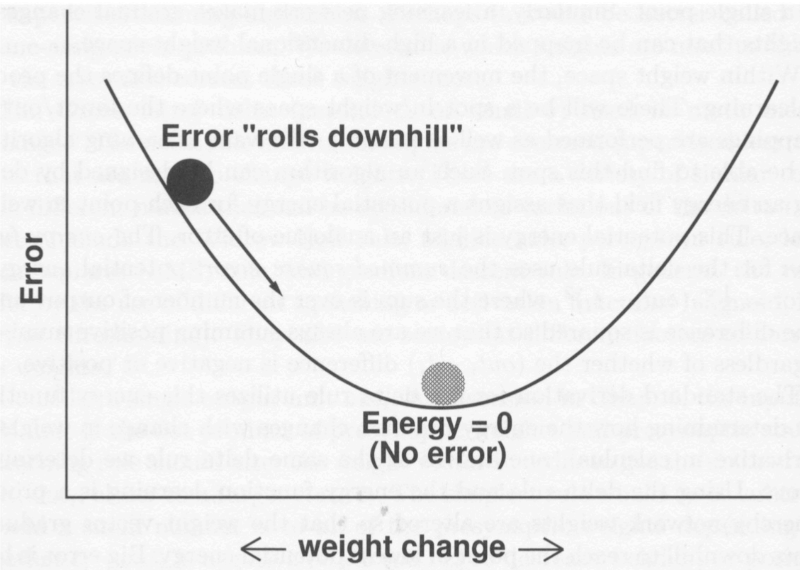
$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j\right)^2$$

$$= \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2$$

In practice, during the training,

- if we use the results of all the inputs within the data set to update weights, it is called batch gradient decent;
- if we use the result of a single input to update weights, it is called **stochastic gradient decent**.

Learning of a Multilayer Perceptron



Update rule:

$$w_{ji}^{h} = w_{ji}^{h} + \Delta w_{ji}^{h},$$

where $\Delta w_{ji}^{h} = -C\,\frac{\partial E}{\partial w_{ji}^{h}}$

The **MLP error function** E **is :**

$$E = \frac{1}{2}\sum_{k=1}^{r}\sum_{j=1}^{m}\left(t_j^k - F_j\left(w^l, w^{l-1}, \cdots, w^1, a^k\right)\right)^2$$

The **error function** E **for a single input**:

$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j\right)^2$$

$$= \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2$$

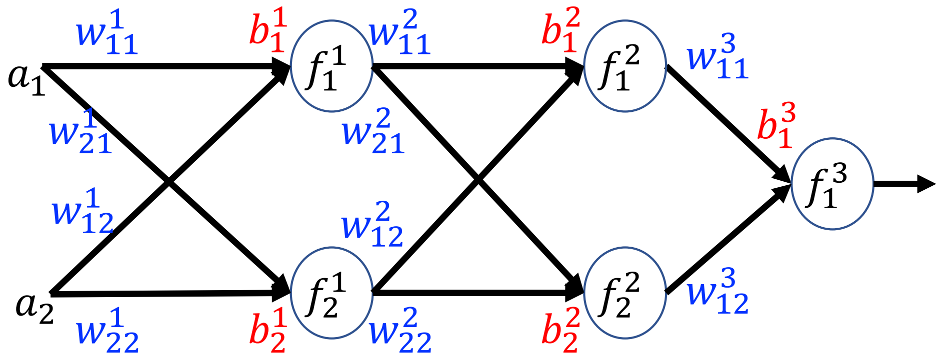
In practice, during the training,

- if we use the results of all the inputs within ... ed

We focus on Stochastic gradient decent in this module.

- if we use the result of a single input to update weights, it is called **stochastic gradient decent**.

Topic of Today's Lecture



Recall the learning rule:

$$w_{ji}^h = w_{ji}^h + \Delta w_{ji}^h, \text{ where } \Delta w_{ji}^h = -C\,\frac{\partial E}{\partial w_{ji}^h}$$

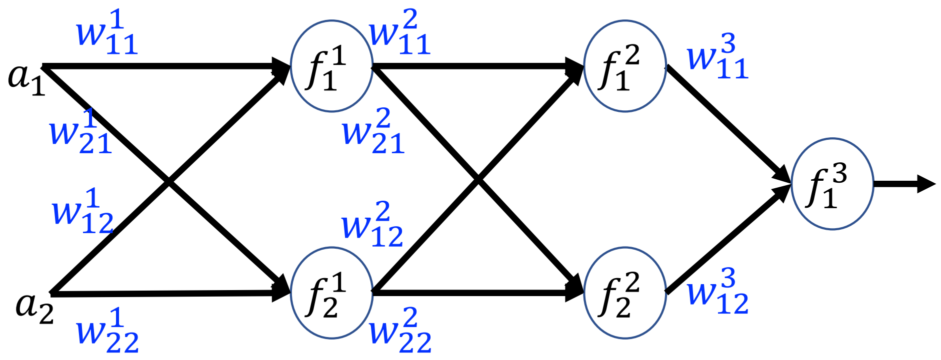
Here C represents the learning rate as before.

We consider the **error function** E **for a single input:**

$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}(t_j - X_j)^2$$
$$= \frac{1}{2}\sum_{j=1}^{m}(t_j - X_j^l)^2$$

The key issue is apparently how to compute the partial derivative $\frac{\partial E}{\partial w_{ji}^h}$.

Partial Derivative



We consider the **error function** E **for a single input:**

$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j\right)^2$$

$$= \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2$$

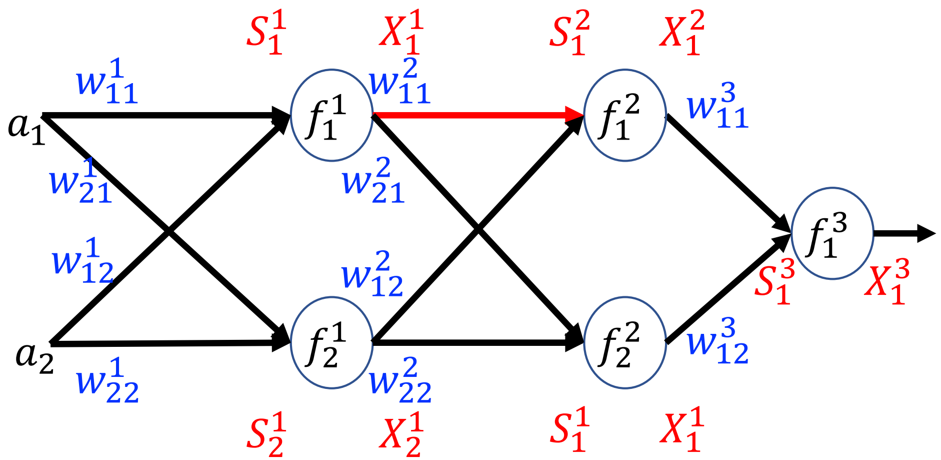
Now, assume we are interested to compute the partial derivative of a specific weight $w_{j_0 i_0}^{l_0}$

$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}}$$

of the connection between the j_0-th neuron in the l_0-th layer and the i_0-th neuron in the $(l_0 - 1)$-th layer.

The detailed deduction will be given in the following slides.

Partial Derivative

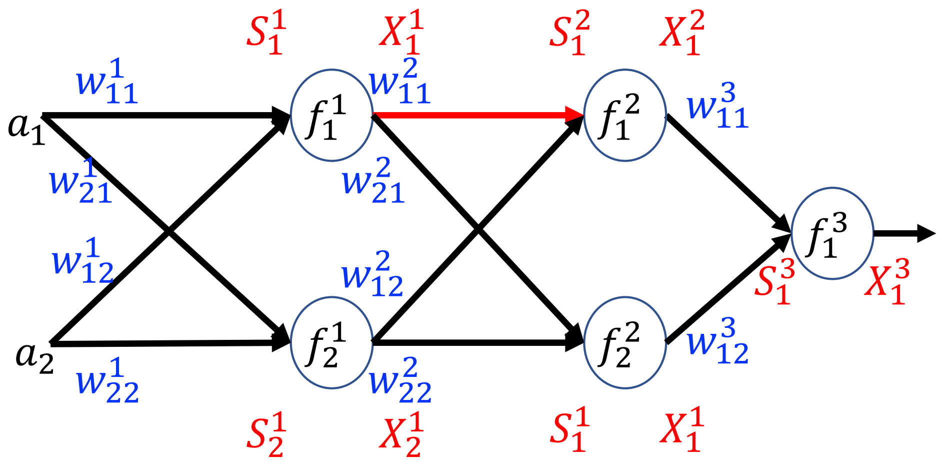


We consider the **error function** E **for a single input:**

$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j\right)^2$$

$$= \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2$$

Partial Derivative

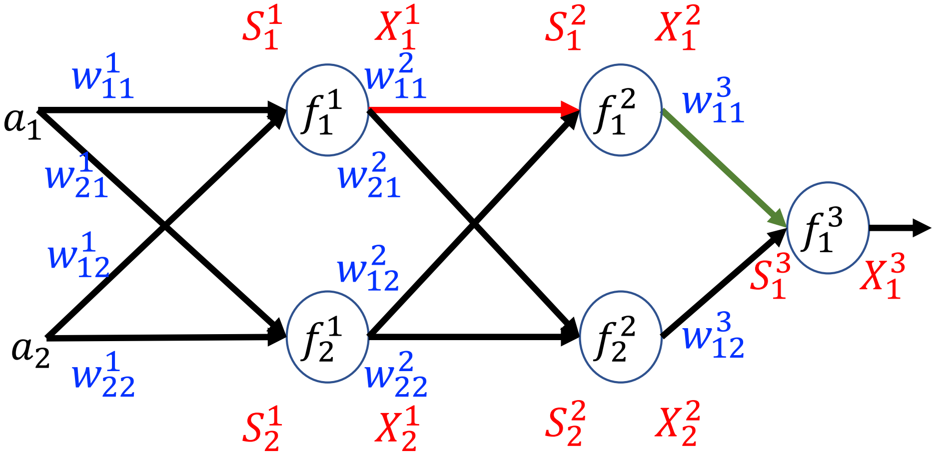


$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \frac{\partial \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2}{\partial w_{j_0 i_0}^{l_0}}$$

We consider the **error function** E **for a single input:**

$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j\right)^2$$
$$= \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2$$

Partial Derivative



We consider the **error function** E **for a single input:**

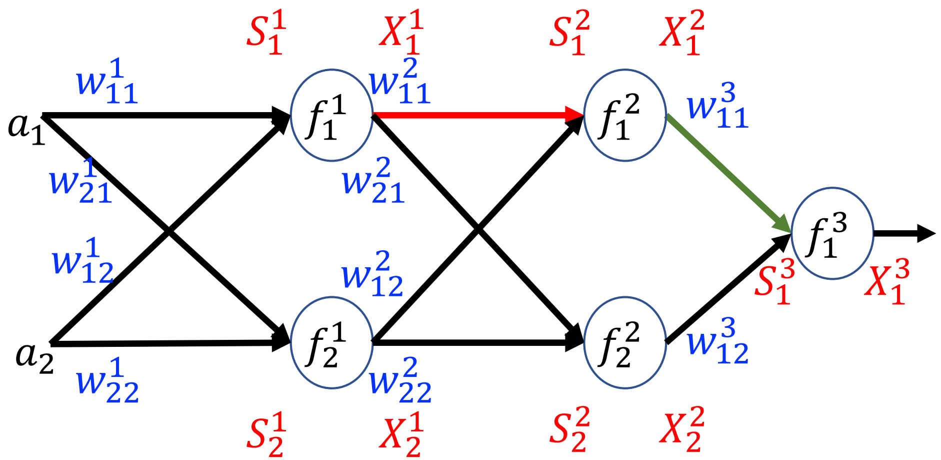
$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j\right)^2$$
$$= \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2$$

$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \frac{\partial \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2}{\partial w_{j_0 i_0}^{l_0}}$$

There are two difference cases:

1. Output layer: $l = l_0$.
2. Otherwise: $l \neq l_0$.

Partial Derivative



We consider the **error function** E **for a single input:**

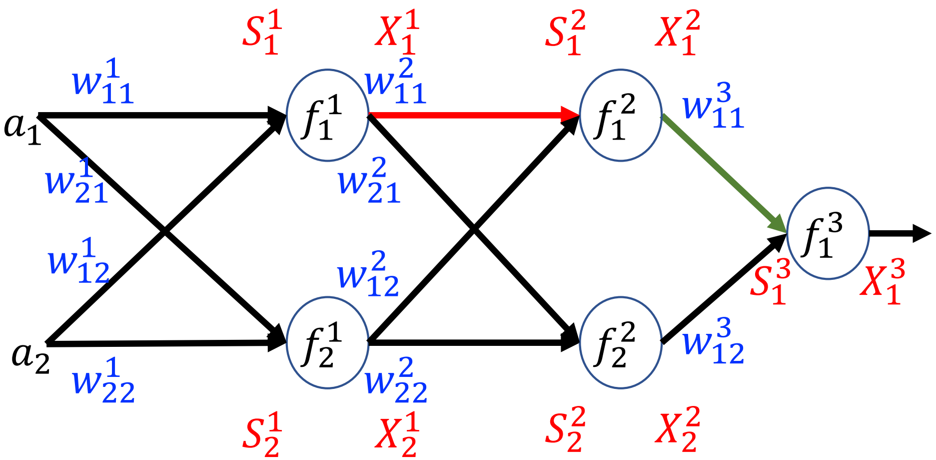
$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j\right)^2$$
$$= \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2$$

$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \frac{\partial \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2}{\partial w_{j_0 i_0}^{l_0}}$$

When $l = l_0$

$$= \frac{\partial \frac{1}{2}\left(\left(t_{j_0} - X_{j_0}^{l_0}\right)^2 + \sum_{j \neq j_0}\left(t_j - X_j^l\right)^2\right)}{\partial w_{j_0 i_0}^{l_0}}$$

Partial Derivative



We consider the **error function** E **for a single input:**

$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j\right)^2$$
$$= \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2$$

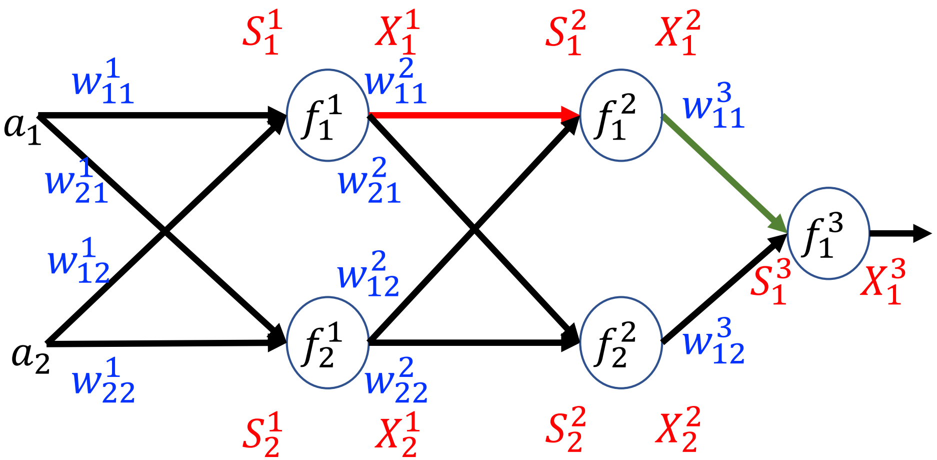
$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \frac{\partial \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2}{\partial w_{j_0 i_0}^{l_0}}$$

When $l = l_0$

$$= \frac{\partial \frac{1}{2}\left(\left(t_{j_0} - X_{j_0}^{l_0}\right)^2 + \sum_{j \neq j_0}\left(t_j - X_j^l\right)^2\right)}{\partial w_{j_0 i_0}^{l_0}}$$

$$= \frac{\partial \frac{1}{2}\left(\left(t_{j_0} - X_{j_0}^{l_0}\right)^2\right)}{\partial w_{j_0 i_0}^{l_0}}$$

Partial Derivative



We consider the **error function** E **for a single input:**

$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j\right)^2$$

$$= \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2$$

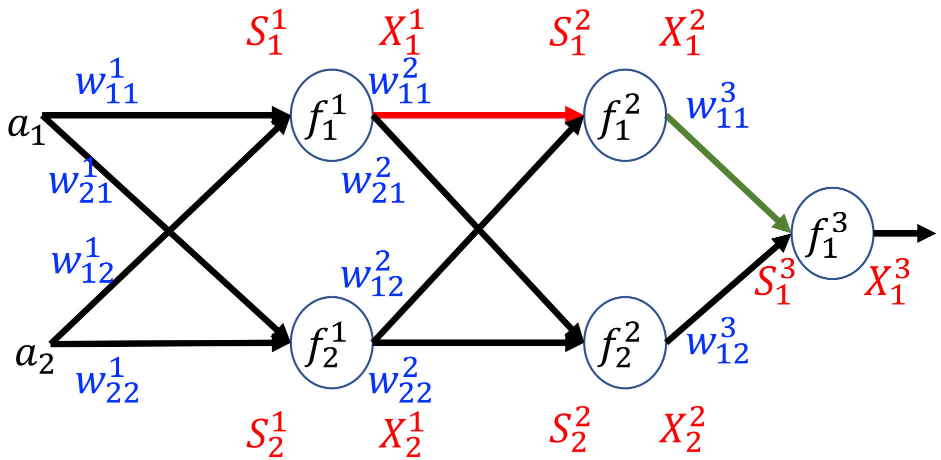
$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \frac{\partial \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2}{\partial w_{j_0 i_0}^{l_0}}$$

When $l = l_0$

$$= \frac{\partial \frac{1}{2}\left(\left(t_{j_0} - X_{j_0}^{l_0}\right)^2 + \sum_{j \neq j_0}\left(t_j - X_j^l\right)^2\right)}{\partial w_{j_0 i_0}^{l_0}}$$

$$= \frac{\partial \frac{1}{2}\left(\left(t_{j_0} - X_{j_0}^{l_0}\right)^2\right)}{\partial X_{j_0}^{l_0}} \cdot \frac{\partial X_{j_0}^{l_0}}{\partial w_{j_0 i_0}^{l_0}}$$

Partial Derivative



We consider the **error function** E **for a single input:**

$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j\right)^2$$

$$= \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2$$

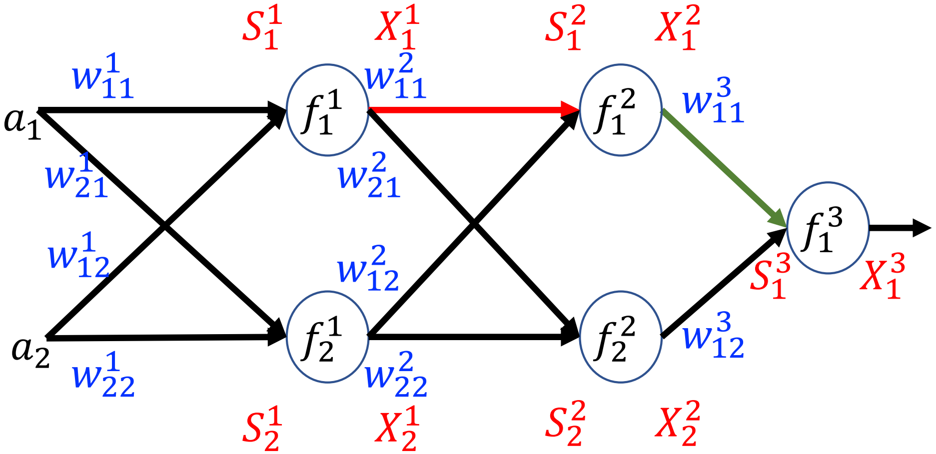
$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \frac{\partial \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2}{\partial w_{j_0 i_0}^{l_0}}$$

When $l = l_0$

$$= \frac{\partial \frac{1}{2}\left(\left(t_{j_0} - X_{j_0}^{l_0}\right)^2 + \sum_{j \neq j_0}\left(t_j - X_j^l\right)^2\right)}{\partial w_{j_0 i_0}^{l_0}}$$

$$= \frac{\partial \frac{1}{2}\left(\left(t_{j_0} - X_{j_0}^{l_0}\right)^2\right)}{\partial X_{j_0}^{l_0}} \cdot \frac{\partial X_{j_0}^{l_0}}{\partial S_{j_0}^{l_0}} \cdot \frac{\partial S_{j_0}^{l_0}}{\partial w_{j_0 i_0}^{l_0}}$$

Partial Derivative



We consider the **error function** E **for a single input:**

$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j\right)^2$$
$$= \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2$$

$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \frac{\partial \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2}{\partial w_{j_0 i_0}^{l_0}}$$

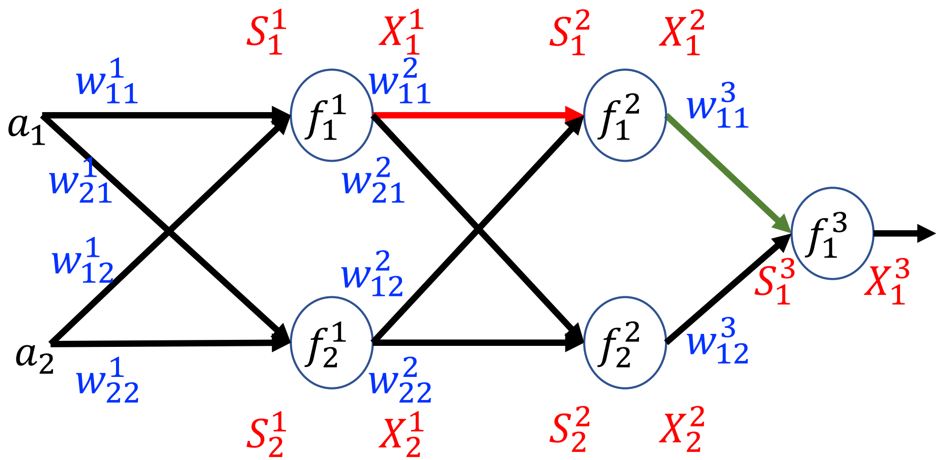
When $l = l_0$

$$= \frac{\partial \frac{1}{2}\left(\left(t_{j_0} - X_{j_0}^{l_0}\right)^2 + \sum_{j \neq j_0}\left(t_j - X_j^l\right)^2\right)}{\partial w_{j_0 i_0}^{l_0}}$$

$$= \frac{\partial \frac{1}{2}\left(\left(t_{j_0} - X_{j_0}^{l_0}\right)^2\right)}{\partial X_{j_0}^{l_0}} \cdot \frac{\partial X_{j_0}^{l_0}}{\partial S_{j_0}^{l_0}} \cdot \frac{\partial S_{j_0}^{l_0}}{\partial w_{j_0 i_0}^{l_0}}$$

$$= \left(X_{j_0}^{l_0} - t_{j_0}\right) \cdot \left(f_{j_0}^{l_0}\right)'\left(S_{j_0}^{l_0}\right) \cdot X_{i_0}^{l_0 - 1}$$

Partial Derivative



We consider the **error function** E **for a single input:**

$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j\right)^2$$

$$= \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2$$

$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \frac{\partial \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2}{\partial w_{j_0 i_0}^{l_0}}$$

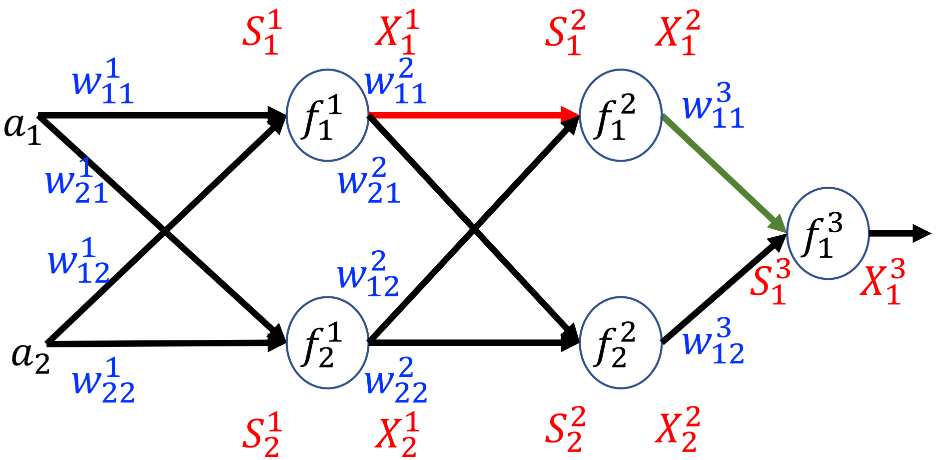
When $l = l_0$

$$= \frac{\partial \frac{1}{2}\left(\left(t_{j_0} - X_{j_0}^{l_0}\right)^2 + \sum_{j \neq j_0}\left(t_j - X_j^l\right)^2\right)}{\partial w_{j_0 i_0}^{l_0}}$$

$$= \frac{\partial \frac{1}{2}\left(\left(t_{j_0} - X_{j_0}^{l_0}\right)^2\right)}{\partial X_{j_0}^{l_0}} \cdot \frac{\partial X_{j_0}^{l_0}}{\partial S_{j_0}^{l_0}} \cdot \frac{\partial S_{j_0}^{l_0}}{\partial w_{j_0 i_0}^{l_0}}$$

$$= \left(X_{j_0}^{l_0} - t_{j_0}\right) \cdot \left(f_{j_0}^{l_0}\right)'\left(S_{j_0}^{l_0}\right) \cdot X_{i_0}^{l_0 - 1}$$

Partial Derivative



We consider the **error function** E **for a single input:**

$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j\right)^2$$

$$= \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2$$

$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \frac{\partial \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2}{\partial w_{j_0 i_0}^{l_0}} \qquad \text{When } l = l_0$$

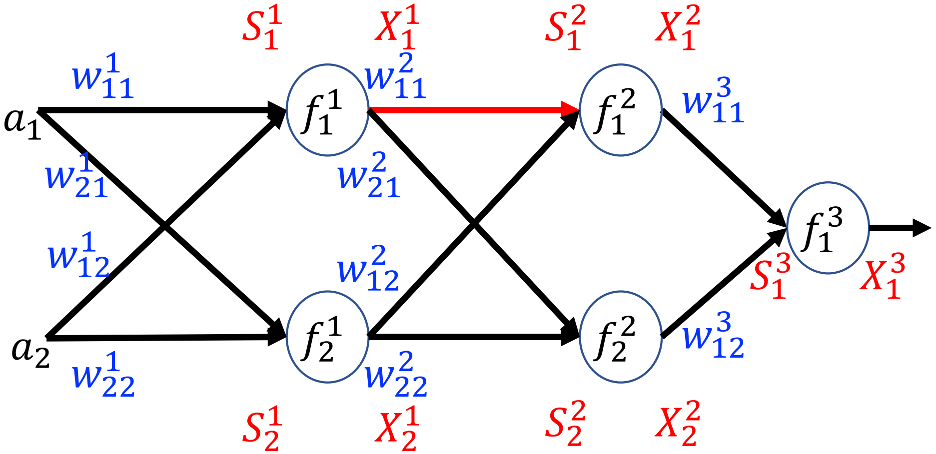
$$= \frac{\partial \frac{1}{2}\left(\left(t_{j_0} - X_{j_0}^{l_0}\right)^2 + \sum_{j\neq j_0}\left(t_j - X_j^l\right)^2\right)}{\partial w_{j_0 i_0}^{l_0}}$$

$$= \frac{\partial \frac{1}{2}\left(\left(t_{j_0} - X_{j_0}^{l_0}\right)^2\right)}{\partial X_{j_0}^{l_0}} \cdot \frac{\partial X_{j_0}^{l_0}}{\partial S_{j_0}^{l_0}} \cdot \frac{\partial S_{j_0}^{l_0}}{\partial w_{j_0 i_0}^{l_0}}$$

$$= \left(X_{j_0}^{l_0} - t_{j_0}\right) \cdot \left(f_{j_0}^{l_0}\right)'\left(S_{j_0}^{l_0}\right) \cdot X_{i_0}^{l_0-1}$$

$$\frac{\partial S_{j_0}^{l_0}}{\partial w_{j_0 i_0}^{l_0}} = \frac{\partial \sum_{i=1}^{l_0-1} w_{j_0 i}^{l_0} X_i^{l_0-1}}{\partial w_{j_0 i_0}^{l_0}} = \frac{\partial \left(w_{j_0 i_0}^{l_0} X_{i_0}^{l_0-1} + \sum_{i\neq i_0} w_{j_0 i}^{l_0} X_i^{l_0-1}\right)}{\partial w_{j_0 i_0}^{l_0}}$$

Partial Derivative



$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \frac{\partial \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2}{\partial w_{j_0 i_0}^{l_0}}$$

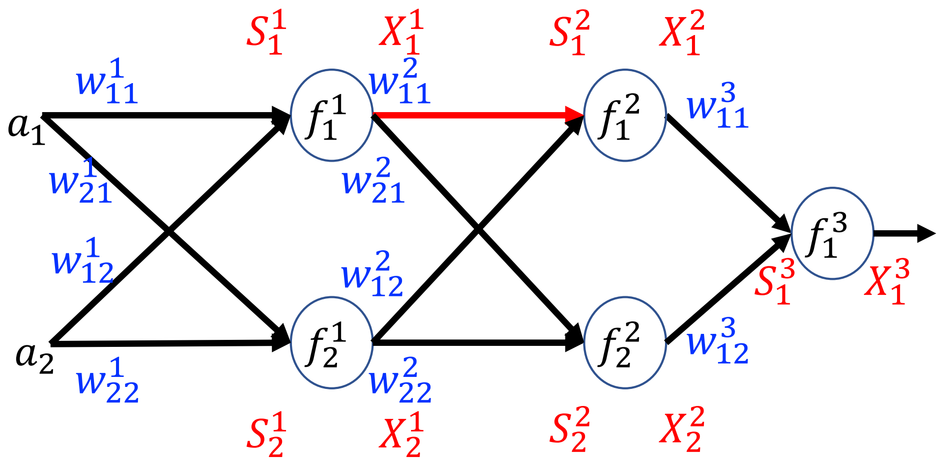
When $l \neq l_0$

We consider the **error function** E **for a single input:**

$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j\right)^2$$

$$= \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2$$

Partial Derivative



$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \frac{\partial \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2}{\partial w_{j_0 i_0}^{l_0}}$$

When $l \neq l_0$

$$= \sum_{j=1}^{m} \frac{\partial \frac{1}{2}\left(t_j - X_j^l\right)^2}{\partial w_{j_0 i_0}^{l_0}}$$

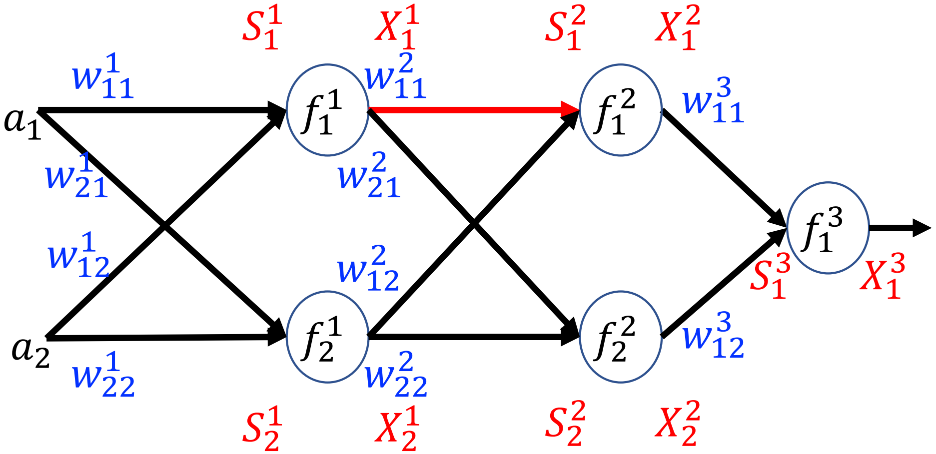
Sum rule

We consider the **error function** E **for a single input:**

$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j\right)^2$$

$$= \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2$$

Partial Derivative



We consider the **error function** E **for a single input:**

$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j\right)^2$$
$$= \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2$$

$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \frac{\partial \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2}{\partial w_{j_0 i_0}^{l_0}}$$

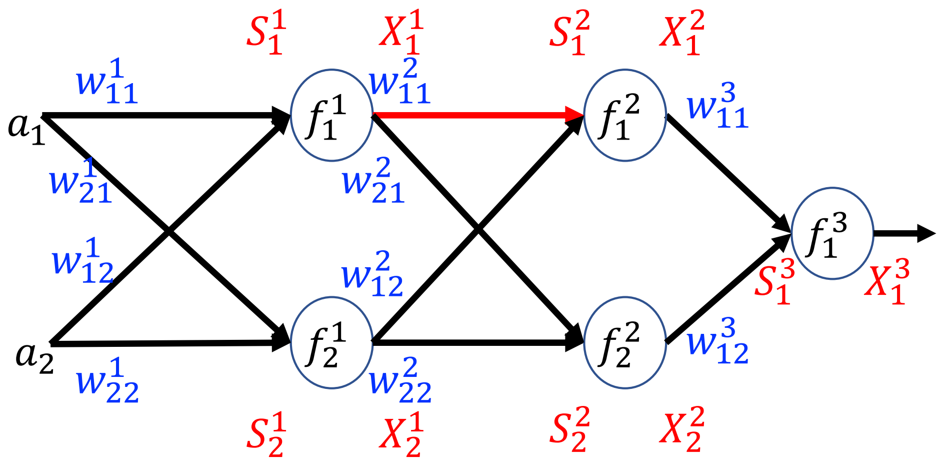
When $l \neq l_0$

$$= \sum_{j=1}^{m} \frac{\partial \frac{1}{2}\left(t_j - X_j^l\right)^2}{\partial X_j^l} \cdot \frac{\partial X_j^l}{\partial w_{j_0 i_0}^{l_0}}$$

Sum rule

Chain rule

Partial Derivative



We consider the **error function** E **for a single input:**

$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j\right)^2$$

$$= \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2$$

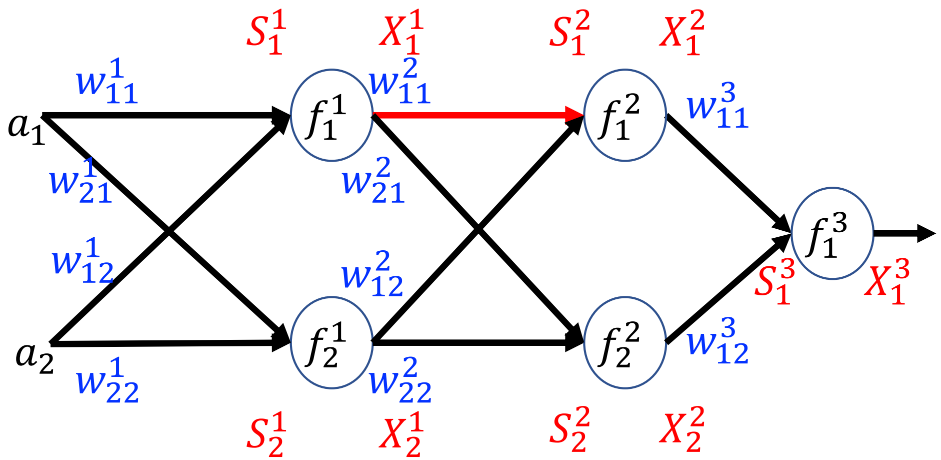
$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \frac{\partial \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2}{\partial w_{j_0 i_0}^{l_0}} \qquad \text{When } l \neq l_0$$

$$= \sum_{j=1}^{n^l} \frac{\partial \frac{1}{2}\left(t_j - X_j^l\right)^2}{\partial X_j^l} \cdot \frac{\partial X_j^l}{\partial w_{j_0 i_0}^{l_0}}$$

Sum rule

Chain rule

Partial Derivative



We consider the **error function** E **for a single input:**

$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}(t_j - X_j)^2$$
$$= \frac{1}{2}\sum_{j=1}^{m}(t_j - X_j^l)^2$$

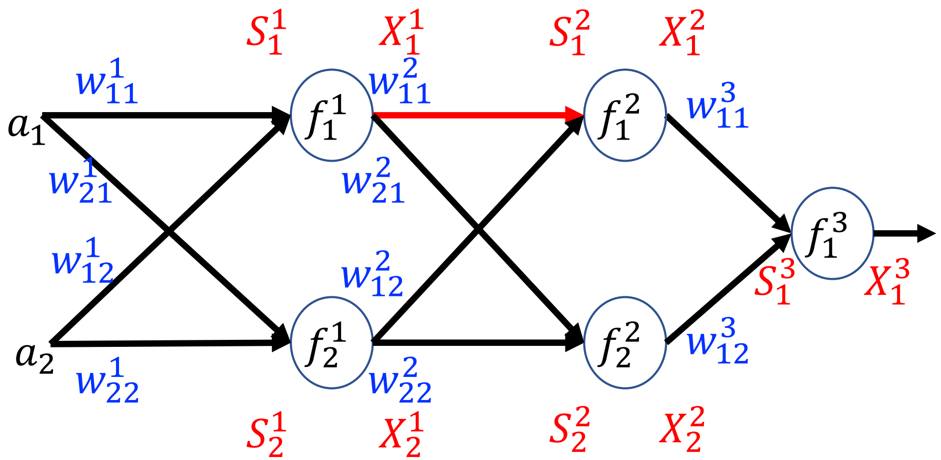
$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \frac{\partial \frac{1}{2}\sum_{j=1}^{m}(t_j - X_j^l)^2}{\partial w_{j_0 i_0}^{l_0}}$$

When $l \neq l_0$

$$= \sum_{j=1}^{n^l} \frac{\partial \frac{1}{2}(t_j - X_j^l)^2}{\partial X_j^l} \cdot \frac{\partial X_j^l}{\partial w_{j_0 i_0}^{l_0}}$$

$$= \sum_{j=1}^{n^l} (X_j^l - t_j) \cdot \frac{\partial X_j^l}{\partial w_{j_0 i_0}^{l_0}}$$

Partial Derivative

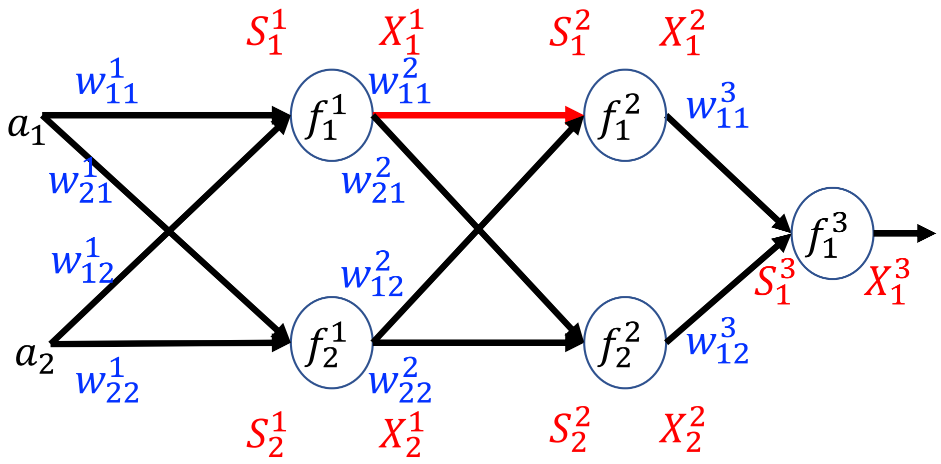


$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \sum_{j=1}^{n^l} \left(X_j^l - t_j\right) \cdot \frac{\partial X_j^l}{\partial w_{j_0 i_0}^{l_0}}$$

We consider the **error function** E **for a single input:**

$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j\right)^2$$
$$= \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2$$

Partial Derivative



We consider the **error function** E **for a single input:**

$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j\right)^2$$

$$= \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2$$

$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \sum_{j=1}^{n^l}\left(X_j^l - t_j\right) \cdot \frac{\partial X_j^l}{\partial w_{j_0 i_0}^{l_0}}$$

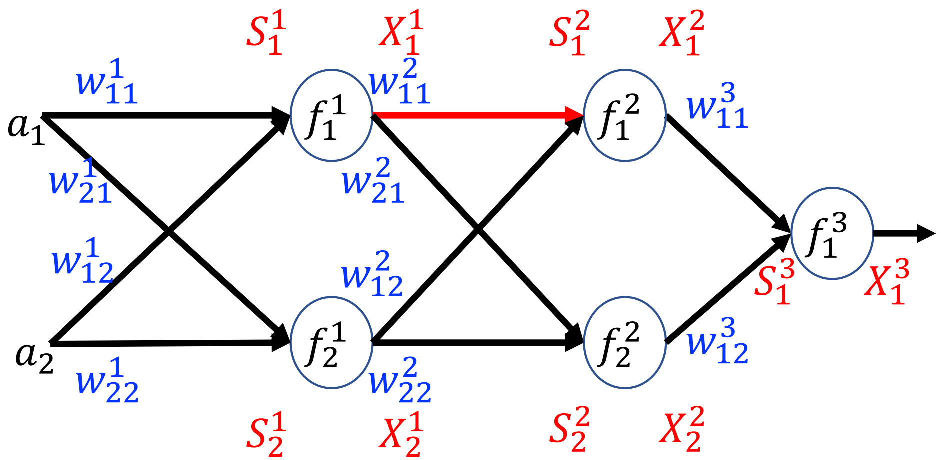
$$\frac{\partial X_j^l}{\partial w_{j_0 i_0}^{l_0}} = \frac{\partial X_j^l}{\partial S_j^l} \cdot \frac{\partial S_j^l}{\partial w_{j_0 i_0}^{l_0}}$$

When $l \neq l_0$

$$= \frac{\partial X_j^l}{\partial S_j^l} \cdot \frac{\partial \sum_{i=1}^{n^{l-1}} w_{ji}^{l-1} X_i^{l-1}}{\partial w_{j_0 i_0}^{l_0}}$$

$$= \frac{\partial X_j^l}{\partial S_j^l} \cdot \sum_{i=1}^{n^{l-1}} \frac{\partial w_{ji}^{l-1} X_i^{l-1}}{\partial w_{j_0 i_0}^{l_0}}$$

Partial Derivative



We consider the **error function** E **for a single input:**

$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j\right)^2$$

$$= \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2$$

$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \sum_{j=1}^{n^l}\left(X_j^l - t_j\right)\cdot\frac{\partial X_j^l}{\partial w_{j_0 i_0}^{l_0}}$$

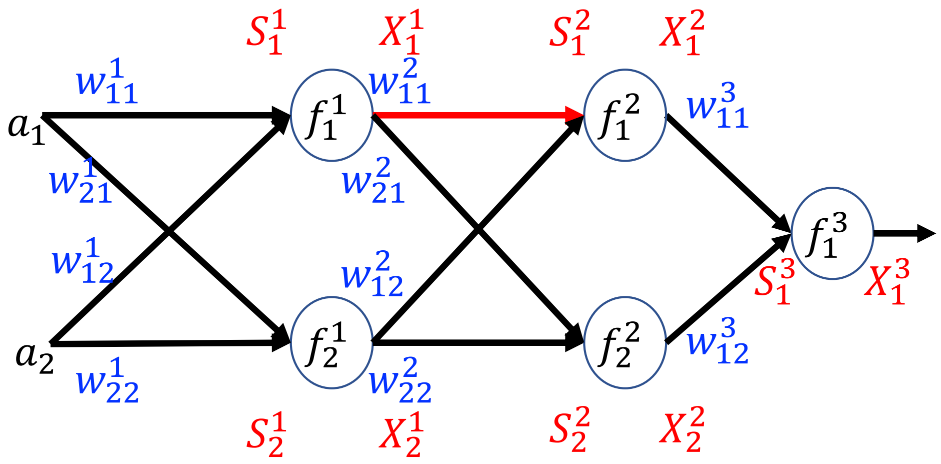
$$\frac{\partial X_j^l}{\partial w_{j_0 i_0}^{l_0}} = \frac{\partial X_j^l}{\partial S_j^l}\cdot\frac{\partial S_j^l}{\partial w_{j_0 i_0}^{l_0}}$$

When $l \neq l_0$

$$= \frac{\partial X_j^l}{\partial S_j^l}\cdot\frac{\partial \sum_{i=1}^{n^{l-1}} w_{ji}^{l-1} X_i^{l-1}}{\partial w_{j_0 i_0}^{l_0}}$$

$$= \frac{\partial X_j^l}{\partial S_j^l}\cdot\sum_{i=1}^{n^{l-1}}\frac{\partial w_{ji}^{l-1} X_i^{l-1}}{\partial w_{j_0 i_0}^{l_0}}$$

Partial Derivative



We consider the **error function** E **for a single input:**

$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j\right)^2$$

$$= \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2$$

$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \sum_{j=1}^{n^l}\left(X_j^l - t_j\right)\cdot\frac{\partial X_j^l}{\partial w_{j_0 i_0}^{l_0}}$$

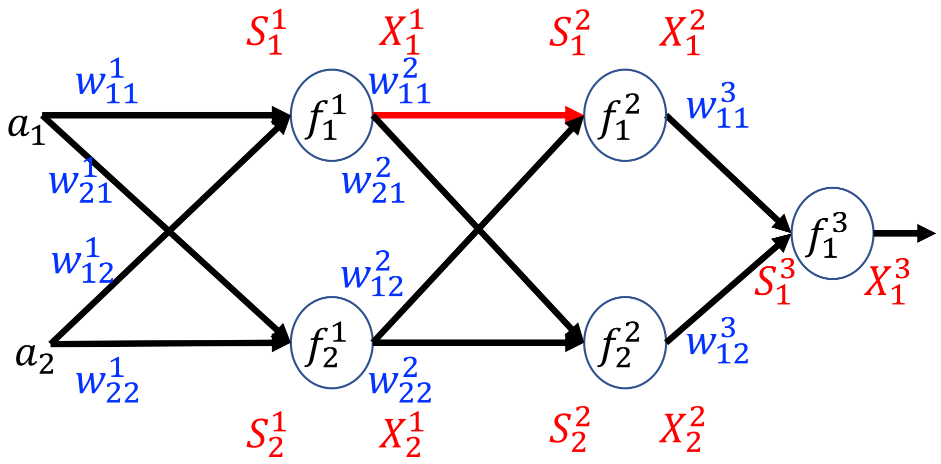
$$\frac{\partial X_j^l}{\partial w_{j_0 i_0}^{l_0}} = \frac{\partial X_j^l}{\partial S_j^l}\cdot\frac{\partial S_j^l}{\partial w_{j_0 i_0}^{l_0}}$$

When $l \neq l_0$

$$= \frac{\partial X_j^l}{\partial S_j^l}\cdot\frac{\partial \sum_{i=1}^{n^{l-1}} w_{ji}^{l-1} X_i^{l-1}}{\partial w_{j_0 i_0}^{l_0}}$$

$$= \frac{\partial X_j^l}{\partial S_j^l}\cdot\sum_{i=1}^{n^{l-1}} w_{ji}^{l-1}\frac{\partial X_i^{l-1}}{\partial w_{j_0 i_0}^{l_0}}$$

Partial Derivative



We consider the **error function** E **for a single input:**

$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j\right)^2$$
$$= \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2$$

$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \sum_{j=1}^{n^l}\left(X_j^l - t_j\right)\cdot\frac{\partial X_j^l}{\partial w_{j_0 i_0}^{l_0}}$$

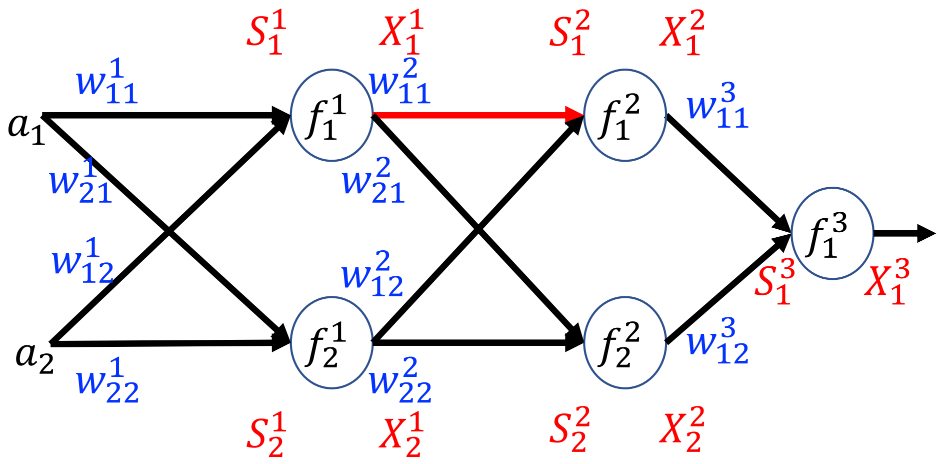
$$\frac{\partial X_j^l}{\partial w_{j_0 i_0}^{l_0}} = \frac{\partial X_j^l}{\partial S_j^l}\cdot\frac{\partial S_j^l}{\partial w_{j_0 i_0}^{l_0}}$$

When $l \neq l_0$

$$= \frac{\partial X_j^l}{\partial S_j^l}\cdot\frac{\partial \sum_{i=1}^{n^{l-1}} w_{ji}^{l-1} X_i^{l-1}}{\partial w_{j_0 i_0}^{l_0}}$$

$$= \left(f_j^l\right)'\left(S_j^l\right)\cdot\sum_{i=1}^{n^{l-1}} w_{ji}^{l-1}\frac{\partial X_i^{l-1}}{\partial w_{j_0 i_0}^{l_0}}$$

Partial Derivative



We consider the **error function** E **for a single input:**

$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}(t_j - X_j)^2$$

$$= \frac{1}{2}\sum_{j=1}^{m}(t_j - X_j^l)^2$$

$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \sum_{j=1}^{n^l}(X_j^l - t_j) \cdot \frac{\partial X_j^l}{\partial w_{j_0 i_0}^{l_0}}$$

When $l \neq l_0$

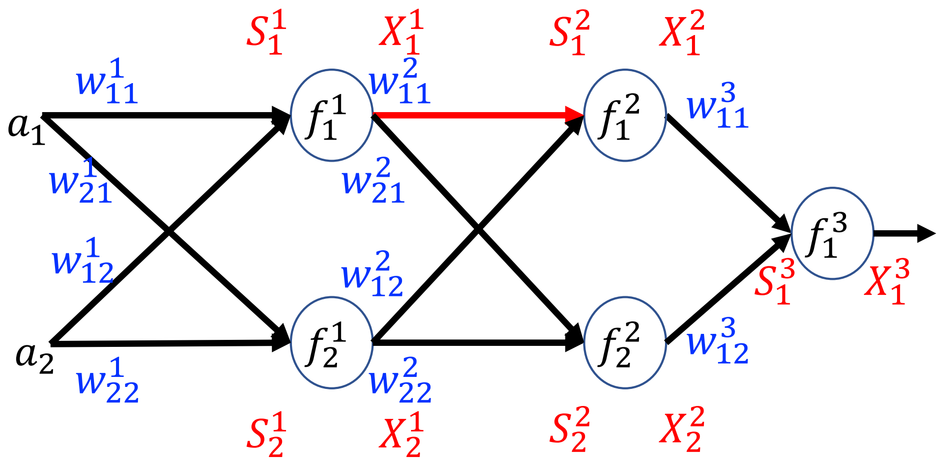
$$\frac{\partial X_j^l}{\partial w_{j_0 i_0}^{l_0}} = \frac{\partial X_j^l}{\partial S_j^l} \cdot \frac{\partial S_j^l}{\partial w_{j_0 i_0}^{l_0}}$$

$$= \frac{\partial X_j^l}{\partial S_j^l} \cdot \frac{\partial \sum_{i=1}^{n^{l-1}} w_{ji}^{l-1} X_i^{l-1}}{\partial w_{j_0 i_0}^{l_0}}$$

$$= (f_j^l)'(S_j^l) \cdot \sum_{i=1}^{n^{l-1}} w_{ji}^{l-1} \frac{\partial X_i^{l-1}}{\partial w_{j_0 i_0}^{l_0}}$$

Induction.

Partial Derivative



We consider the **error function** E **for a single input:**

$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j\right)^2$$

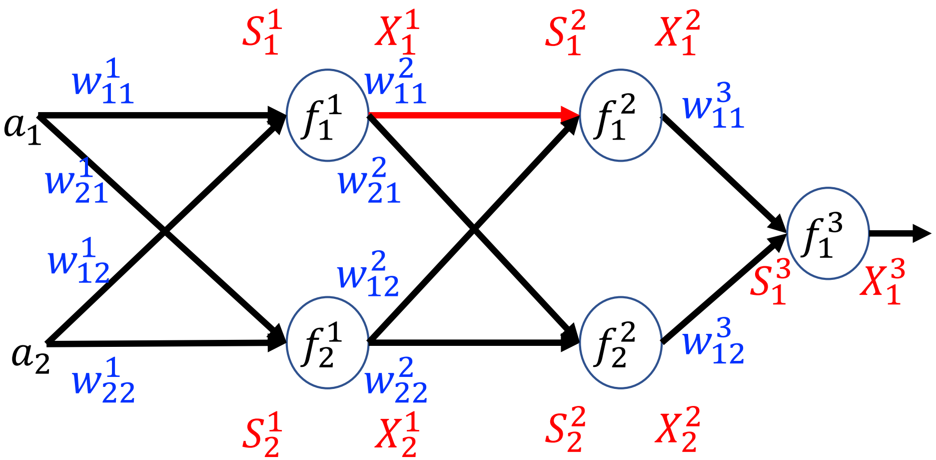
$$= \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2$$

$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \sum_{j=1}^{n^l}\left(X_j^l - t_j\right) \cdot \frac{\partial X_j^l}{\partial w_{j_0 i_0}^{l_0}}$$

$\boxed{l \text{ layer}}$ **Induction.**

$$\frac{\partial X_j^l}{\partial w_{j_0 i_0}^{l_0}} = \left(f_j^l\right)'\left(S_j^l\right) \cdot \sum_{i=1}^{n^{l-1}} w_{ji}^{l-1} \frac{\partial X_i^{l-1}}{\partial w_{j_0 i_0}^{l_0}}$$

Partial Derivative



We consider the **error function** E **for a single input:**

$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j\right)^2$$
$$= \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2$$

$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \sum_{j=1}^{n^l}\left(X_j^l - t_j\right)\cdot\frac{\partial X_j^l}{\partial w_{j_0 i_0}^{l_0}}$$

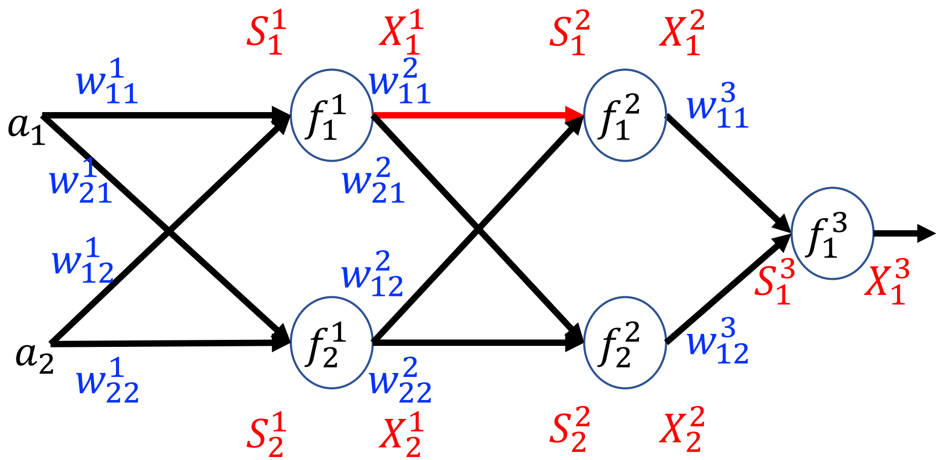
$\boxed{l \text{ layer}}$ **Induction.**

$$\frac{\partial X_j^l}{\partial w_{j_0 i_0}^{l_0}} = \left(f_j^l\right)'\left(S_j^l\right)\cdot\sum_{i=1}^{n^{l-1}} w_{ji}^{l-1}\frac{\partial X_i^{l-1}}{\partial w_{j_0 i_0}^{l_0}}$$

$\boxed{l_0 \text{ layer}, j \neq j_0}$ **Base case.**

$$\frac{\partial X_j^{l_0}}{\partial w_{j_0 i_0}^{l_0}} = \frac{\partial X_j^{l_0}}{\partial S_j^{l_0}}\cdot\frac{\partial S_j^{l_0}}{\partial w_{j_0 i_0}^{l_0}}$$

Partial Derivative



We consider the **error function** E **for a single input:**

$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j\right)^2$$
$$= \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2$$

$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \sum_{j=1}^{n^l}\left(X_j^l - t_j\right)\cdot\frac{\partial X_j^l}{\partial w_{j_0 i_0}^{l_0}}$$

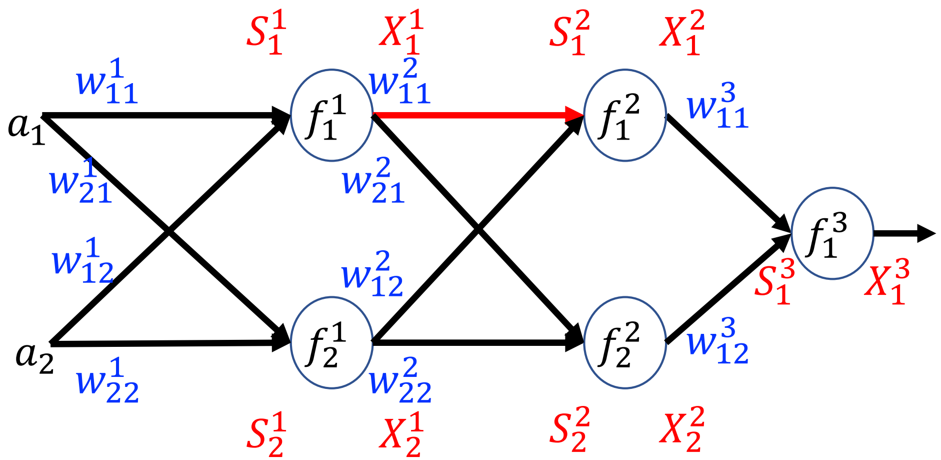
$\boxed{l \text{ layer}}$ **Induction.**

$$\frac{\partial X_j^l}{\partial w_{j_0 i_0}^{l_0}} = \left(f_j^l\right)'\left(S_j^l\right)\cdot\sum_{i=1}^{n^{l-1}} w_{ji}^{l-1}\frac{\partial X_i^{l-1}}{\partial w_{j_0 i_0}^{l_0}}$$

$\boxed{l_0 \text{ layer}, j \neq j_0}$ **Base case.**

$$\frac{\partial X_j^{l_0}}{\partial w_{j_0 i_0}^{l_0}} = \frac{\partial X_j^{l_0}}{\partial S_j^{l_0}}\cdot\frac{\partial \sum_{i=1}^{n^{l_0-1}} w_{ji}^{l_0} X_i^{l_0-1}}{\partial w_{j_0 i_0}^{l_0}}$$

Partial Derivative



We consider the **error function** E **for a single input:**

$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}(t_j - X_j)^2$$
$$= \frac{1}{2}\sum_{j=1}^{m}(t_j - X_j^l)^2$$

$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \sum_{j=1}^{n^l}(X_j^l - t_j)\cdot\frac{\partial X_j^l}{\partial w_{j_0 i_0}^{l_0}}$$

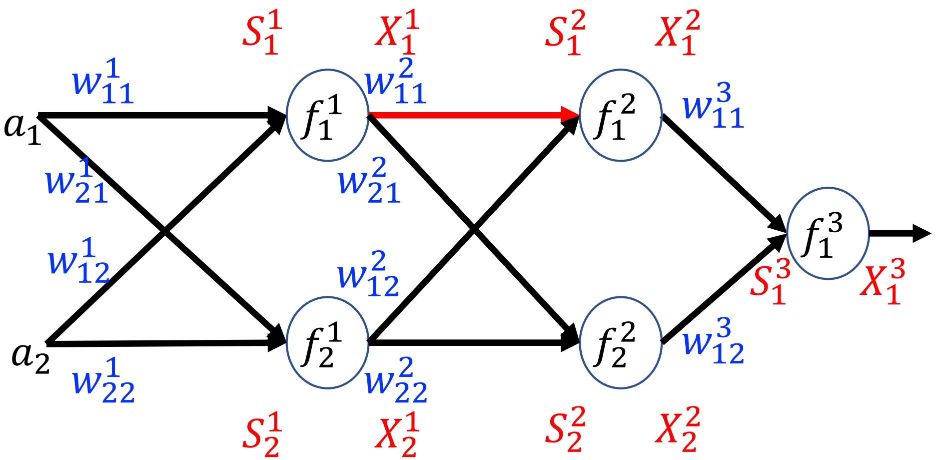
l layer **Induction.**

$$\frac{\partial X_j^l}{\partial w_{j_0 i_0}^{l_0}} = (f_j^l)'(S_j^l)\cdot\sum_{i=1}^{n^{l-1}} w_{ji}^{l-1}\frac{\partial X_i^{l-1}}{\partial w_{j_0 i_0}^{l_0}}$$

l_0 layer, $j \neq j_0$ **Base case.**

$$\frac{\partial X_j^{l_0}}{\partial w_{j_0 i_0}^{l_0}} = \frac{\partial X_j^{l_0}}{\partial S_j^{l_0}}\cdot\frac{\partial \sum_{i=1}^{n^{l_0-1}} w_{ji}^{l_0} X_i^{l_0-1}}{\partial w_{j_0 i_0}^{l_0}}$$

Partial Derivative



We consider the **error function** E **for a single input:**

$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j\right)^2$$

$$= \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2$$

$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \sum_{j=1}^{n^l}\left(X_j^l - t_j\right)\cdot\frac{\partial X_j^l}{\partial w_{j_0 i_0}^{l_0}}$$

l layer

Induction.

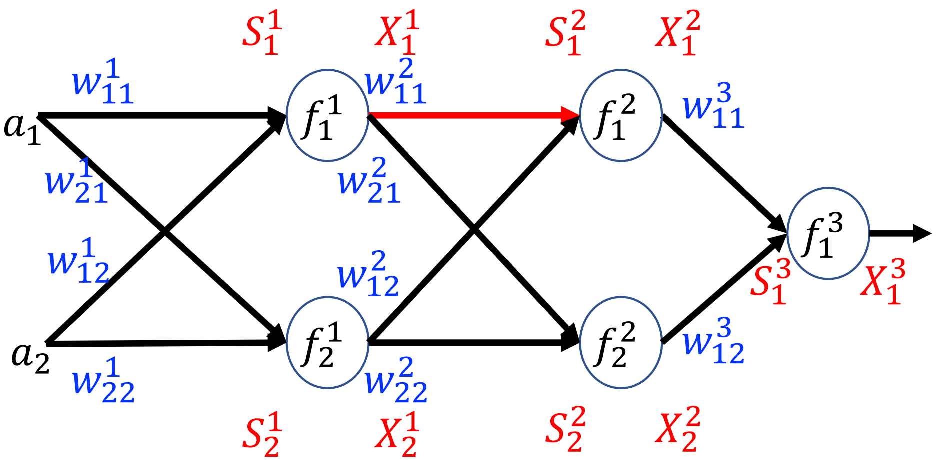
$$\frac{\partial X_j^l}{\partial w_{j_0 i_0}^{l_0}} = \left(f_j^l\right)'\left(S_j^l\right)\cdot\sum_{i=1}^{n^{l-1}} w_{ji}^{l-1}\frac{\partial X_i^{l-1}}{\partial w_{j_0 i_0}^{l_0}}$$

l_0 layer, $j \neq j_0$

Base case.

$$\frac{\partial X_j^{l_0}}{\partial w_{j_0 i_0}^{l_0}} = \frac{\partial X_j^{l_0}}{\partial S_j^{l_0}}\cdot\frac{\partial \sum_{i=1}^{n^{l_0-1}} w_{ji}^{l_0} X_i^{l_0-1}}{\partial w_{j_0 i_0}^{l_0}} = 0$$

Partial Derivative



We consider the **error function** E **for a single input:**

$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j\right)^2$$

$$= \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2$$

$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \sum_{j=1}^{n^l}\left(X_j^l - t_j\right)\cdot\frac{\partial X_j^l}{\partial w_{j_0 i_0}^{l_0}}$$

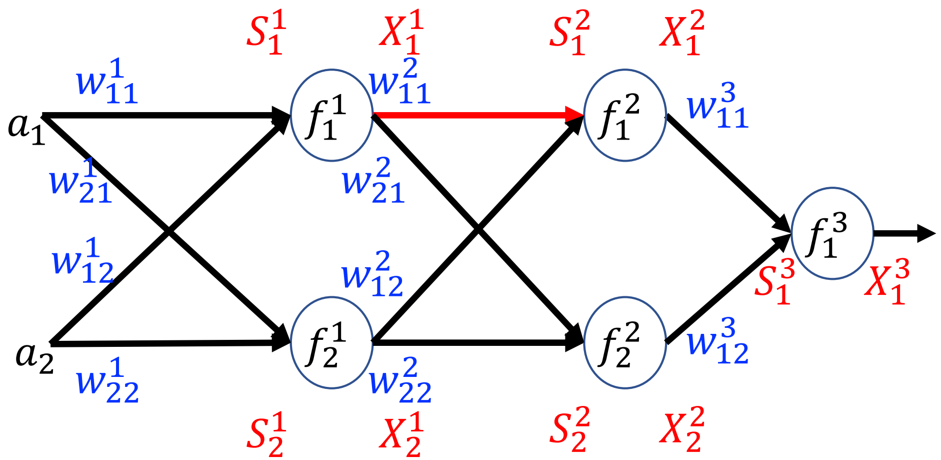
$\boxed{l \text{ layer}}$ **Induction.**

$$\frac{\partial X_j^l}{\partial w_{j_0 i_0}^{l_0}} = \left(f_j^l\right)'\left(S_j^l\right)\cdot\sum_{i=1}^{n^{l-1}} w_{ji}^{l-1}\,\frac{\partial X_i^{l-1}}{\partial w_{j_0 i_0}^{l_0}}$$

$\boxed{:l_0 \text{ layer}, j = j_0}$ **Base case.**

$$\frac{\partial X_j^{l_0}}{\partial w_{j_0 i_0}^{l_0}} = \frac{\partial X_{j_0}^{l_0}}{\partial S_{j_0}^{l_0}}\cdot\frac{\partial S_{j_0}^{l_0}}{\partial w_{j_0 i_0}^{l_0}} = \frac{\partial X_j^{l_0}}{\partial S_j^{l_0}}\cdot X_{i_0}^{l_0-1}$$

Partial Derivative



We consider the **error function** E **for a single input:**

$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}(t_j - X_j)^2$$

$$= \frac{1}{2}\sum_{j=1}^{m}(t_j - X_j^l)^2$$

$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \sum_{j=1}^{n^l}(X_j^l - t_j)\cdot\frac{\partial X_j^l}{\partial w_{j_0 i_0}^{l_0}}$$

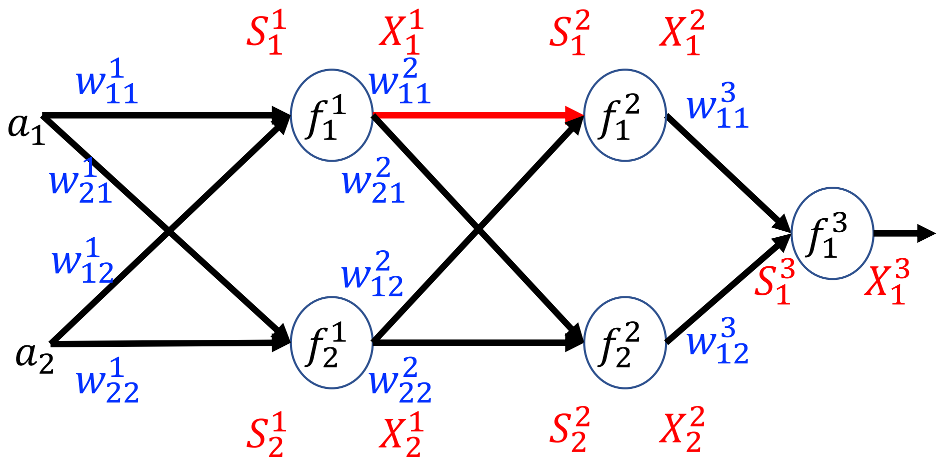
$\boxed{l \text{ layer}}$ **Induction.**

$$\frac{\partial X_j^l}{\partial w_{j_0 i_0}^{l_0}} = (f_j^l)'(S_j^l)\cdot\sum_{i=1}^{n^{l-1}} w_{ji}^{l-1}\frac{\partial X_i^{l-1}}{\partial w_{j_0 i_0}^{l_0}}$$

$\boxed{: l_0 \text{ layer}, j = j_0}$ **Base case.**

$$\frac{\partial X_j^{l_0}}{\partial w_{j_0 i_0}^{l_0}} = \left(f_{j_0}^{l_0}\right)'\left(S_j^{l_0}\right)\cdot X_{i_0}^{l_0-1}$$

Partial Derivative



We consider the **error function** E **for a single input:**

$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j\right)^2$$
$$= \frac{1}{2}\sum_{j=1}^{m}\left(t_j - X_j^l\right)^2$$

$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \sum_{j=1}^{n^l}\left(X_j^l - t_j\right)\cdot\frac{\partial X_j^l}{\partial w_{j_0 i_0}^{l_0}}$$

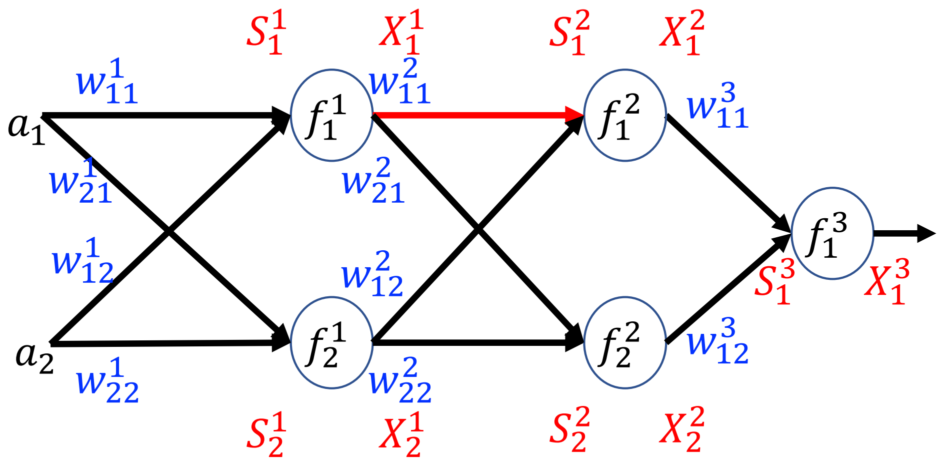
$\boxed{l \text{ layer}}$ **Induction.**

$$\frac{\partial X_j^l}{\partial w_{j_0 i_0}^{l_0}} = \left(f_j^l\right)'\left(S_j^l\right)\cdot\sum_{i=1}^{n^{l-1}} w_{ji}^{l-1}\frac{\partial X_i^{l-1}}{\partial w_{j_0 i_0}^{l_0}}$$

$\vdots$ $\boxed{l_0 \text{ layer}}$ **Base case.**

$$\frac{\partial X_j^{l_0}}{\partial w_{j_0 i_0}^{l_0}} = \begin{cases}\left(f_{j_0}^{l_0}\right)'\left(S_j^{l_0}\right)\cdot X_{i_0}^{l_0-1}, & j = j_0\\ 0, & j \neq j_0\end{cases}$$

Partial Derivative: Conclusion

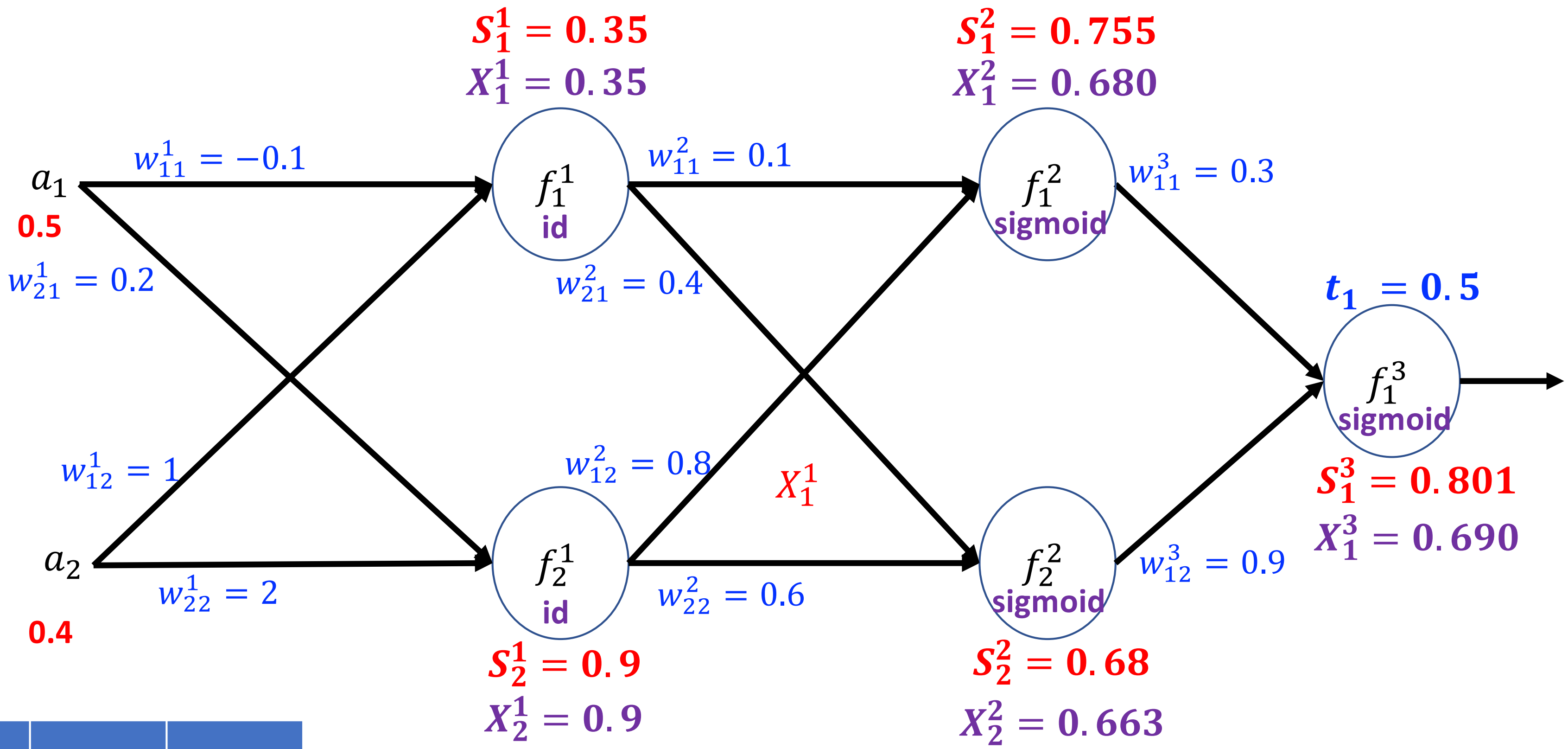


We consider the **error function** E **for a single input:**

$$E = \frac{1}{2}\sum_{j=1}^{m} e_j^2 = \frac{1}{2}\sum_{j=1}^{m} \left(t_j - X_j\right)^2$$
$$= \frac{1}{2}\sum_{j=1}^{m} \left(t_j - X_j^l\right)^2$$

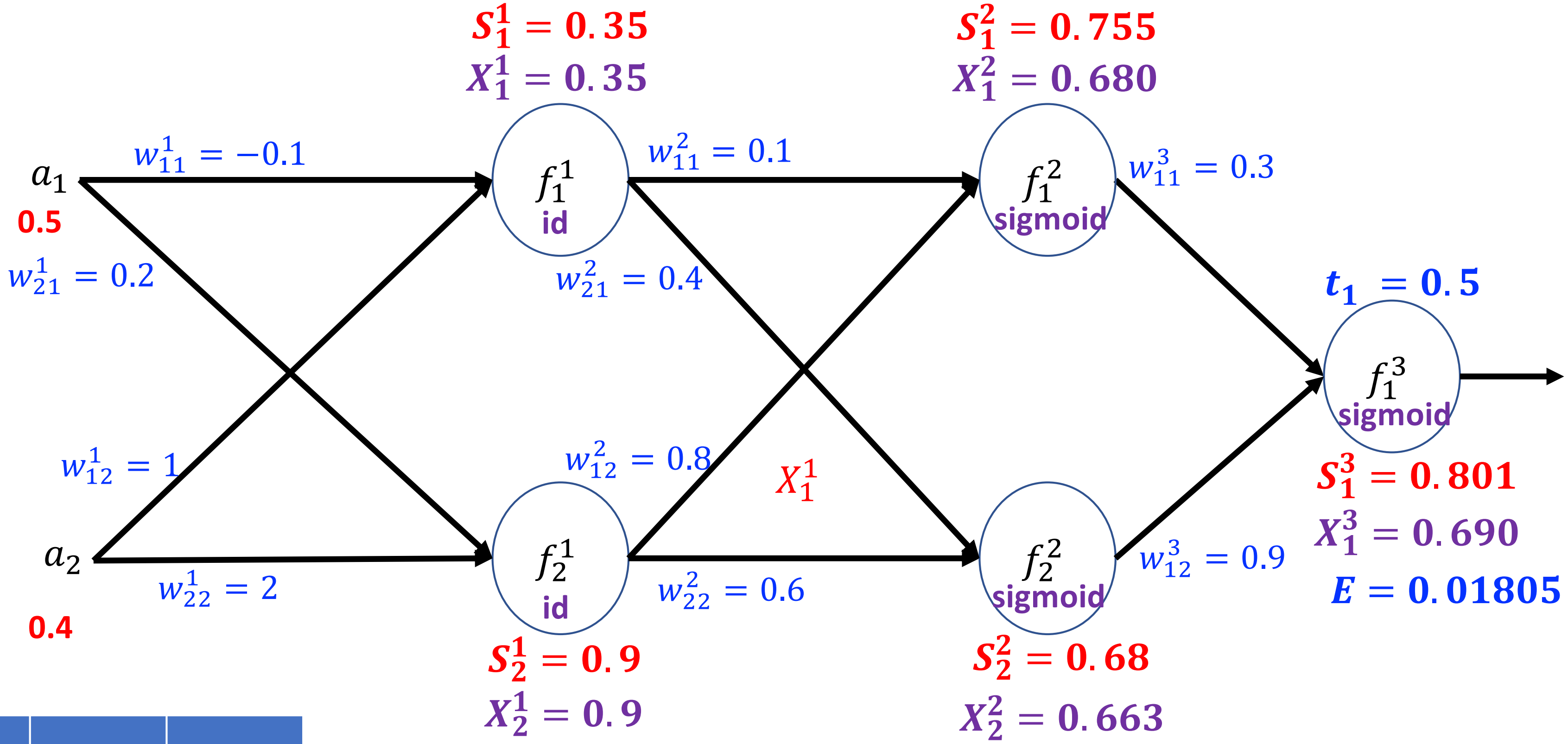
$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \begin{cases} \left(X_{j_0}^{l_0} - t_{j_0}\right) \cdot \left(f_{j_0}^{l_0}\right)' \left(S_{j_0}^{l_0}\right) \cdot X_{i_0}^{l_0-1} & \text{When } l = l_0 \\ \sum_{j=1}^{n^l} \left(X_j^l - t_j\right) \cdot \left(\left(f_j^l\right)' \left(S_j^l\right) \cdot \sum_{i=1}^{n^{l-1}} w_{ji}^{l-1} \left(\cdots \left(f_{j_0}^{l_0}\right)' \left(S_{j_0}^{l_0}\right) \cdot X_{i_0}^{l_0-1} \right) \right) & \text{When } l \neq l_0 \end{cases}$$

A Running Example



a_1	a_2	t_1
0.5	0.4	0.5

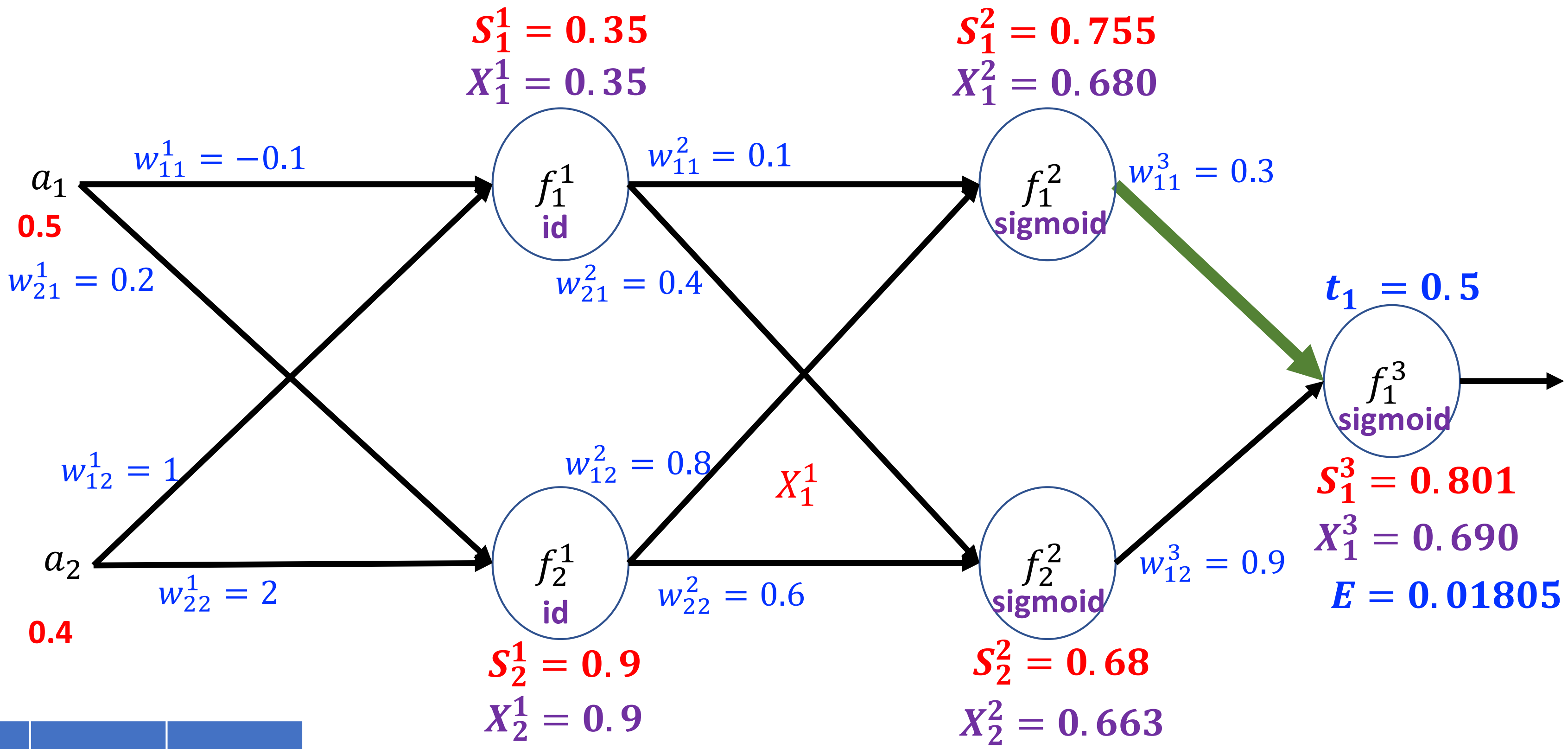
A Running Example



a_1	a_2	t_1
0.5	0.4	0.5

$$E = \frac{1}{2}\left(t_1 - X_1^3\right)^2 = 0.01805$$

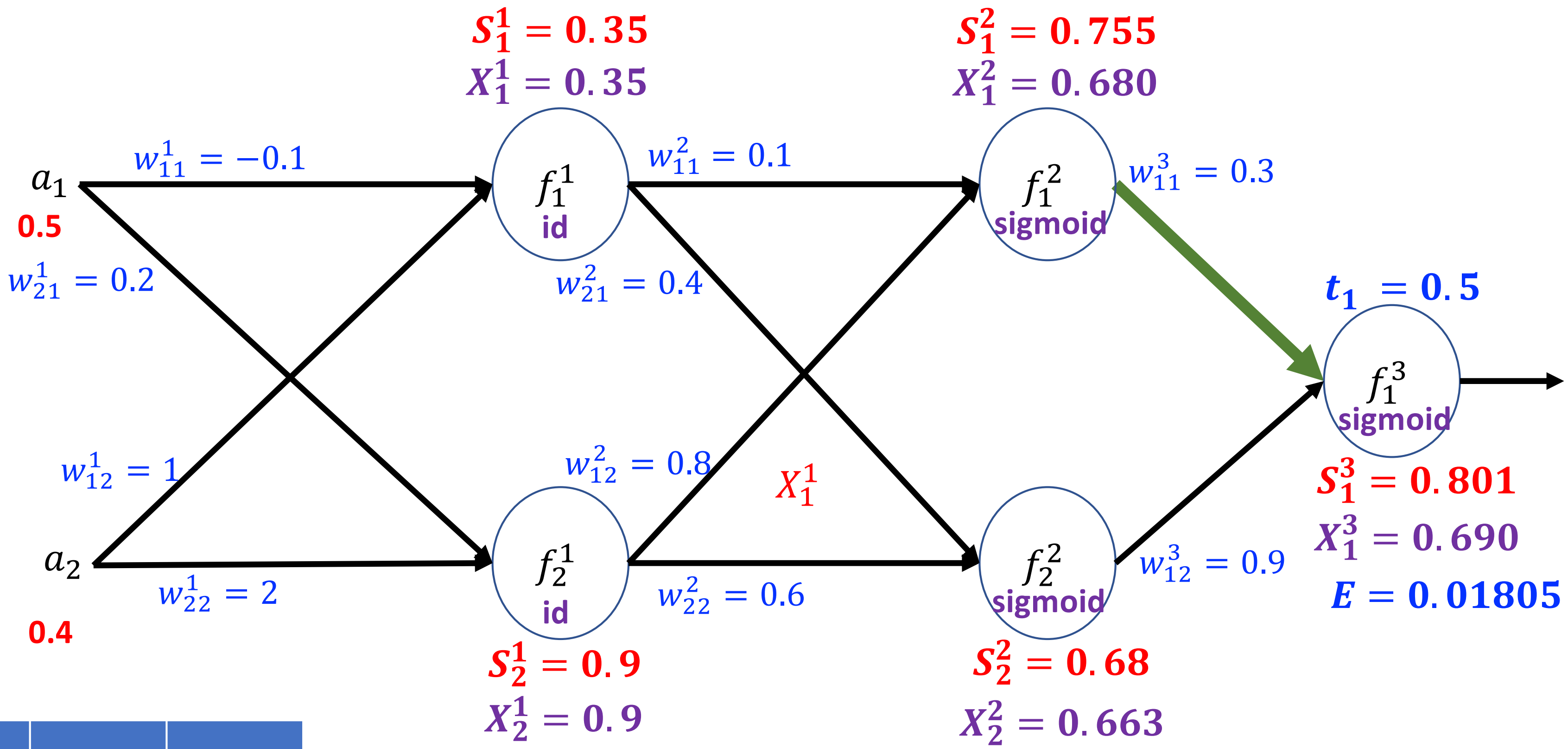
A Running Example: w_{11}^3



a_1	a_2	t_1
0.5	0.4	0.5

$$\frac{\partial E}{\partial w_{11}^3} = (X_1^3 - t_1) \cdot (sig)'(S_1^3) \cdot X_1^2$$

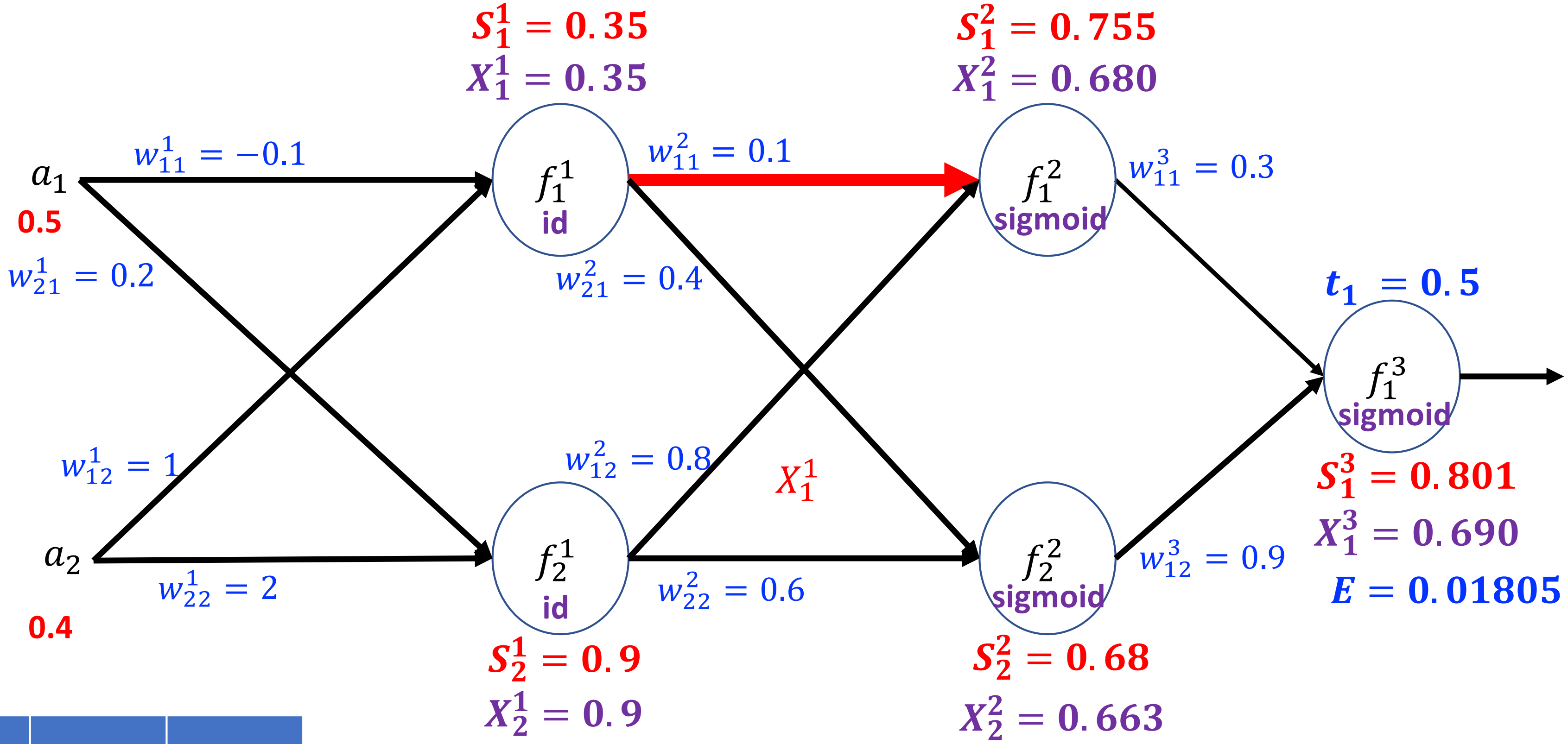
A Running Example: w_{11}^{3}



a_1	a_2	t_1
0.5	0.4	0.5

$$\frac{\partial E}{\partial w_{11}^{3}} = \underbrace{(X_1^3 - t_1)}_{0.19} \cdot \underbrace{(sig)'(S_1^3)}_{0.69\times(1-0.69)} \cdot \underbrace{X_1^2}_{0.68} = 0.02763$$

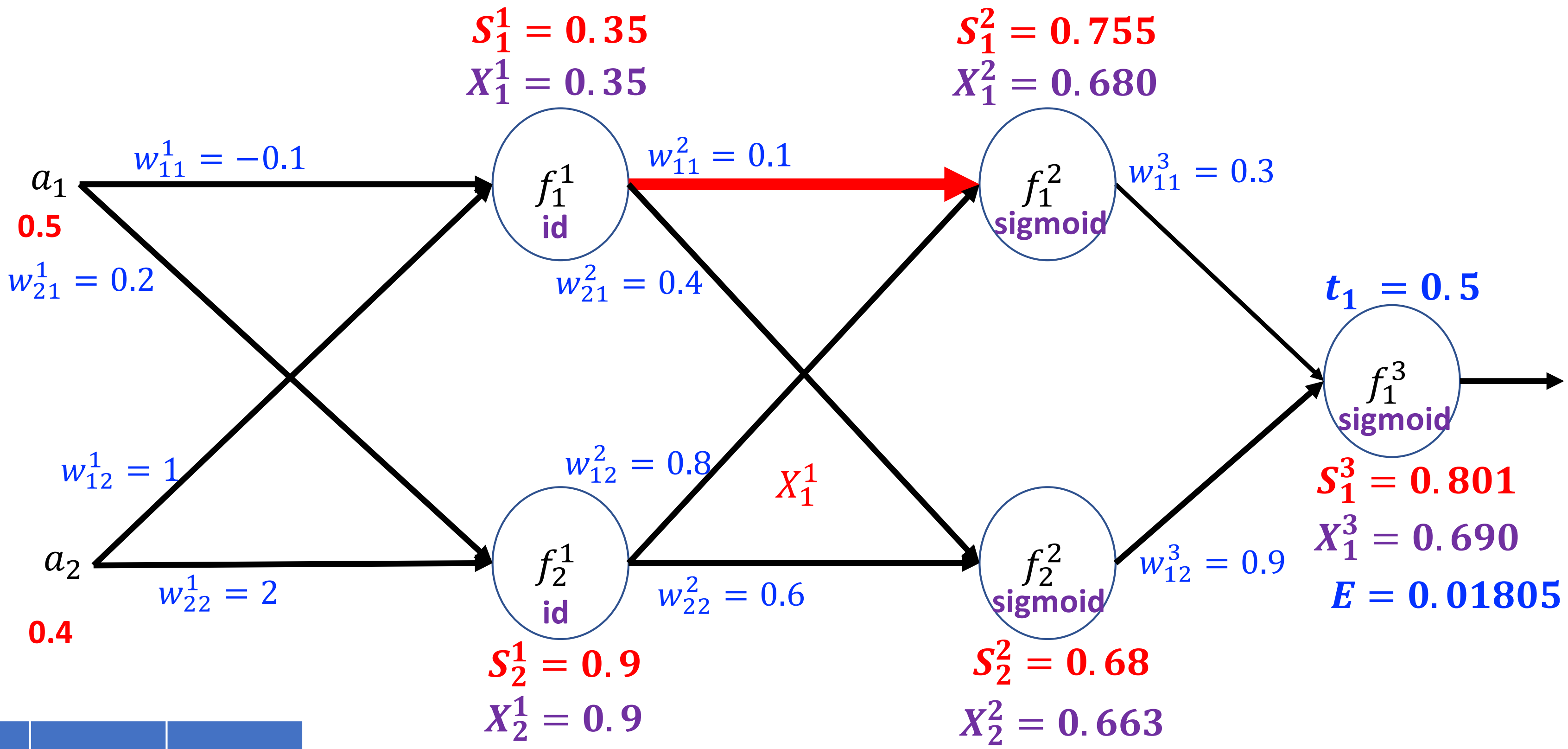
A Running Example: w_{11}^{2}



a_1	a_2	t_1
0.5	0.4	0.5

$$\frac{\partial E}{\partial w_{11}^2} = (X_1^3 - t_1) \cdot (sig)'(S_1^3) \cdot w_{11}^3 \cdot (sig)'(S_1^2) \cdot X_1^1$$

A Running Example: w_{11}^2



a_1	a_2	t_1
0.5	0.4	0.5

$$\frac{\partial E}{\partial w_{11}^2} = \underbrace{(X_1^3 - t_1)}_{0.19} \cdot \underbrace{(sig)'(S_1^3)}_{0.69\times(1-0.69)} \cdot \underbrace{w_{11}^3}_{0.3} \cdot \underbrace{(sig)'(S_1^2)}_{0.68\times(1-0.68)} \cdot \underbrace{X_1^1}_{0.35} = 0.0009$$