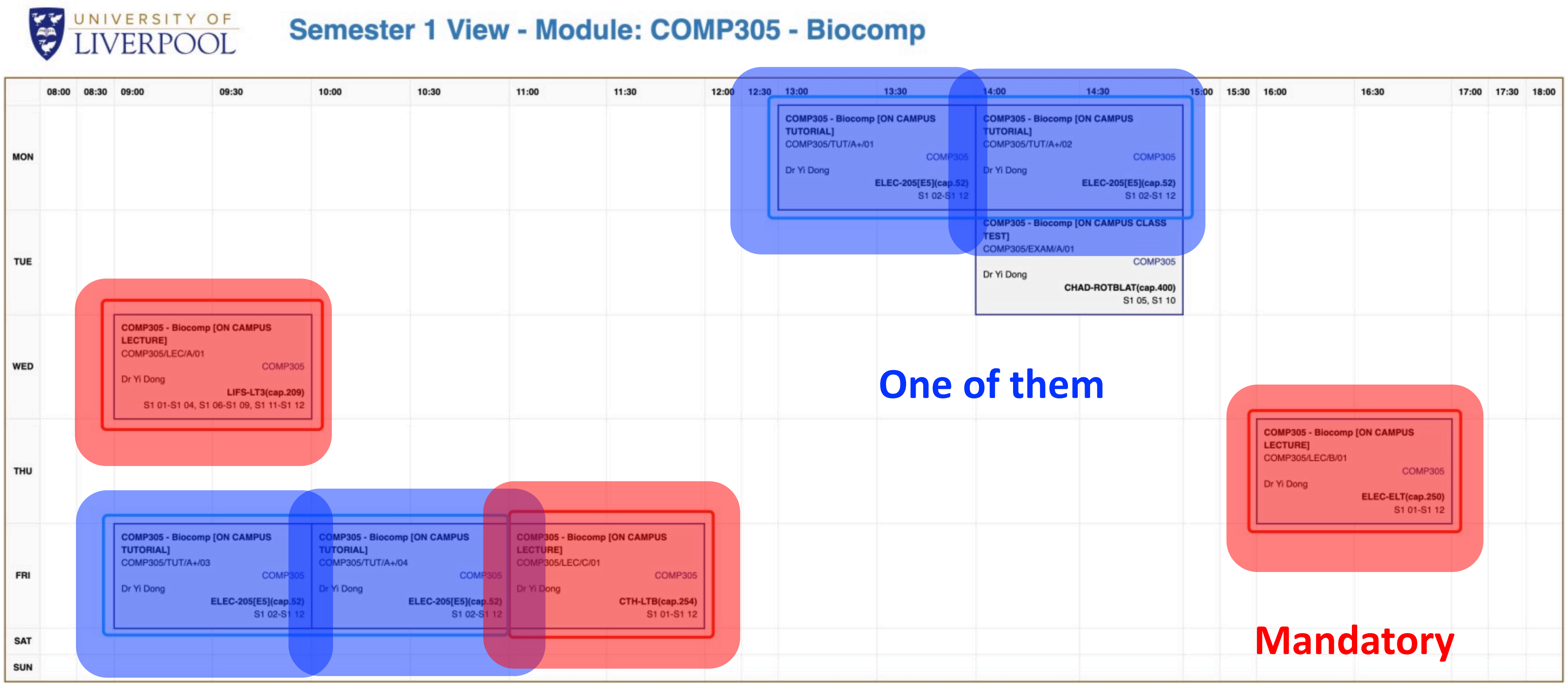
Comp305

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.

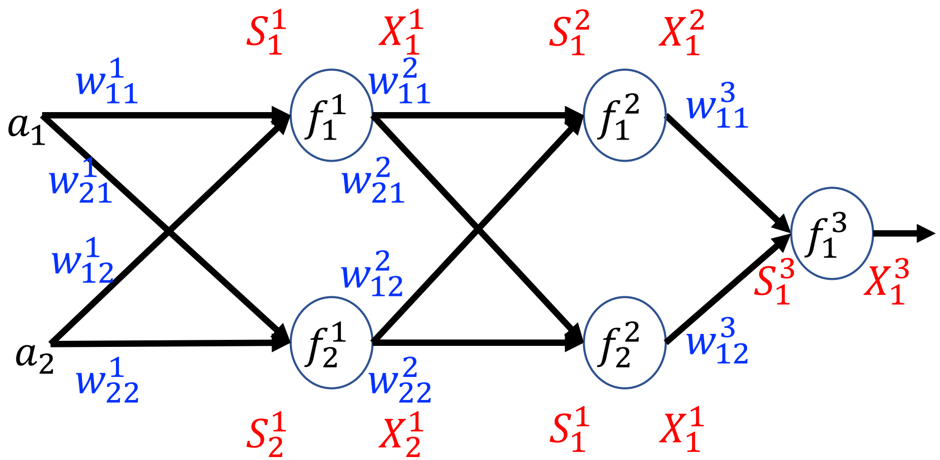
Comp305 Part I.

Artificial Neural Networks

Topic 5.

Multilayer Perceptron

Gradient of a Weight



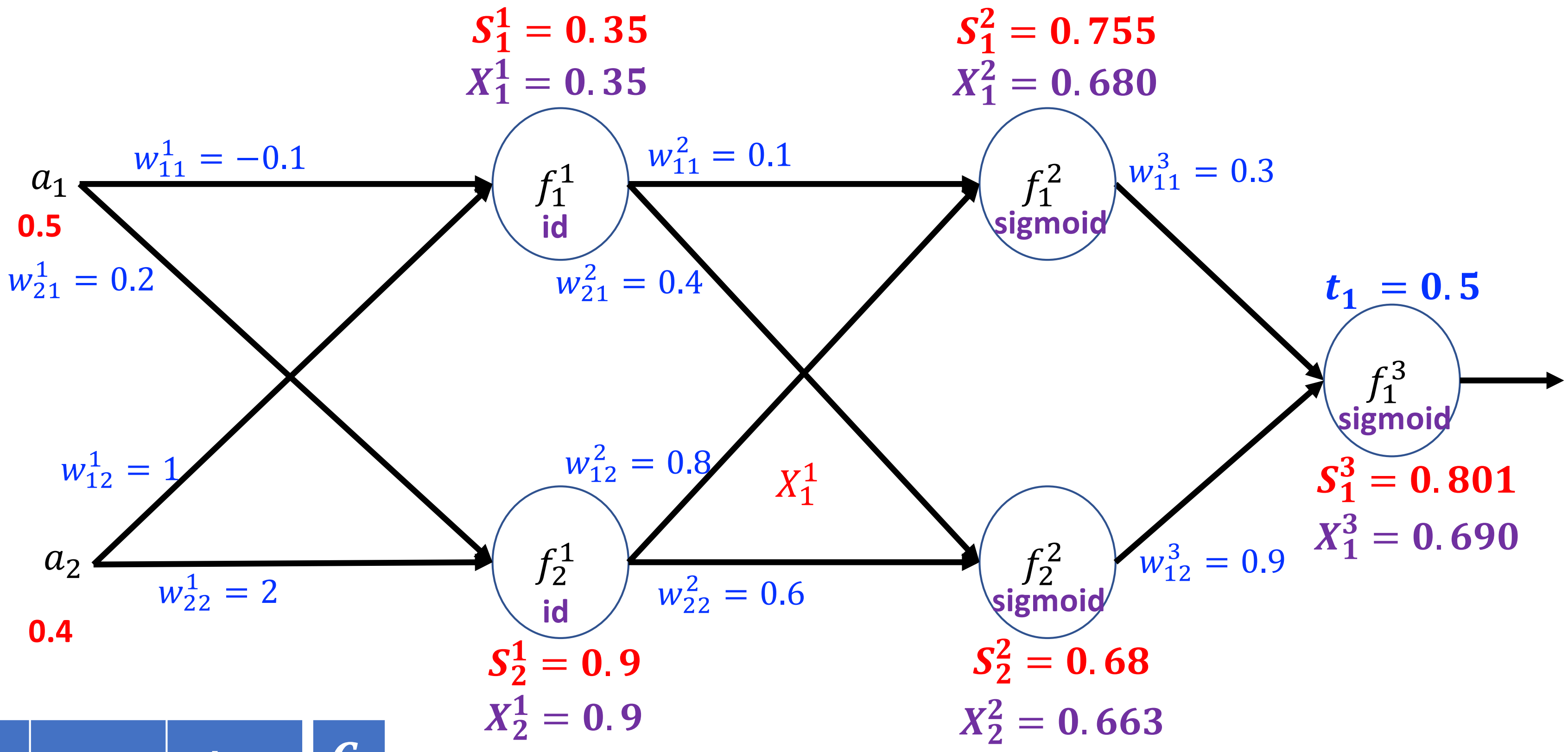
We consider the **error function** E **for a single input:**

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

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

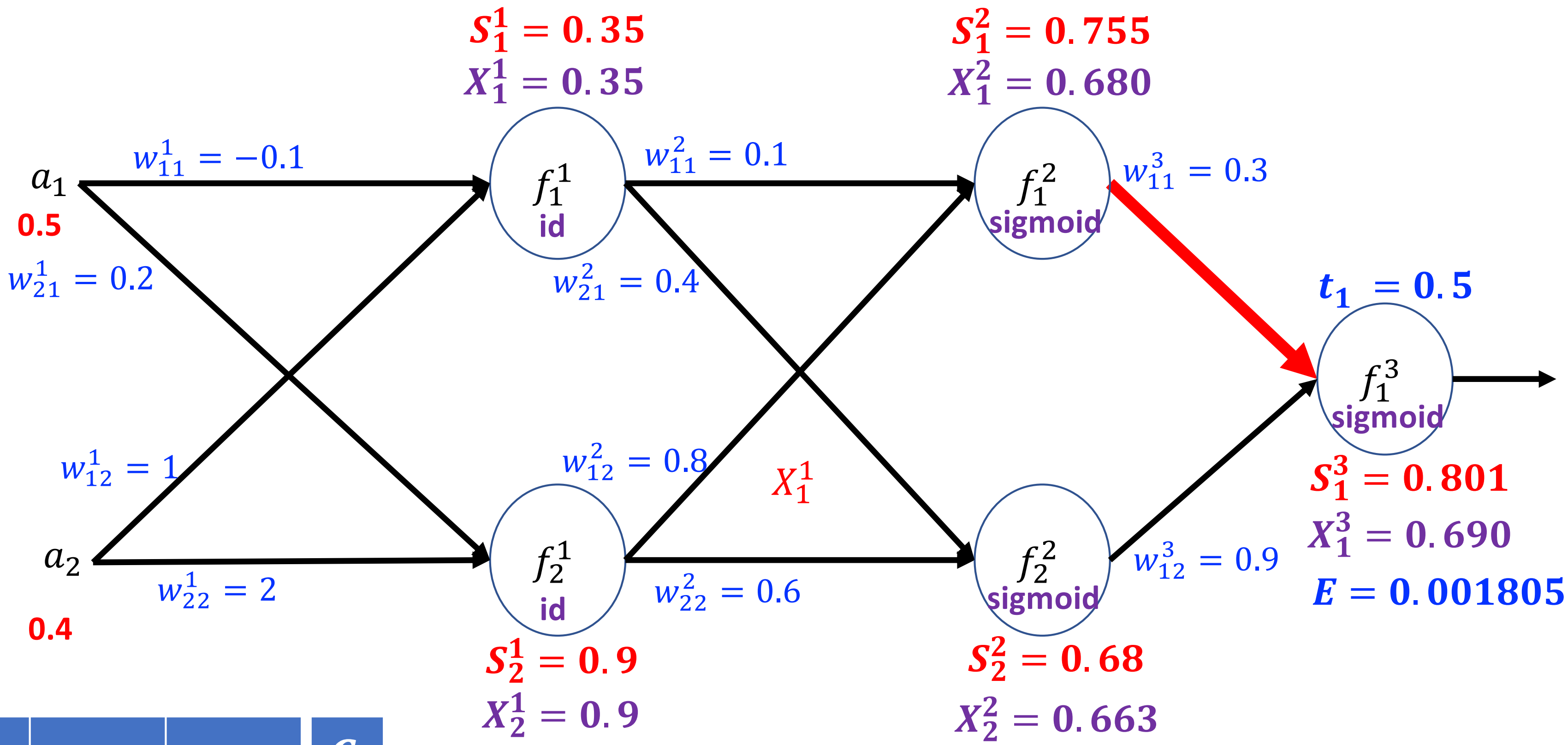
$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \begin{cases} \left(X_{j_0}^{l_0} - t_{j_0}\right)\cdot\left(f_{j_0}^{l_0}\right)'\left(S_{j_0}^{l_0}\right)\cdot X_{i_0}^{l_0-1} & \text{When } l = l_0 \\ \displaystyle\sum_{j=1}^{n^l}\left(X_j^l - t_j\right)\cdot\left(\left(f_j^l\right)'\left(S_j^l\right)\cdot\sum_{i=1}^{n^{l-1}}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

Gradient of a Weight



a_1	a_2	t_1	C
0.5	0.4	0.5	1

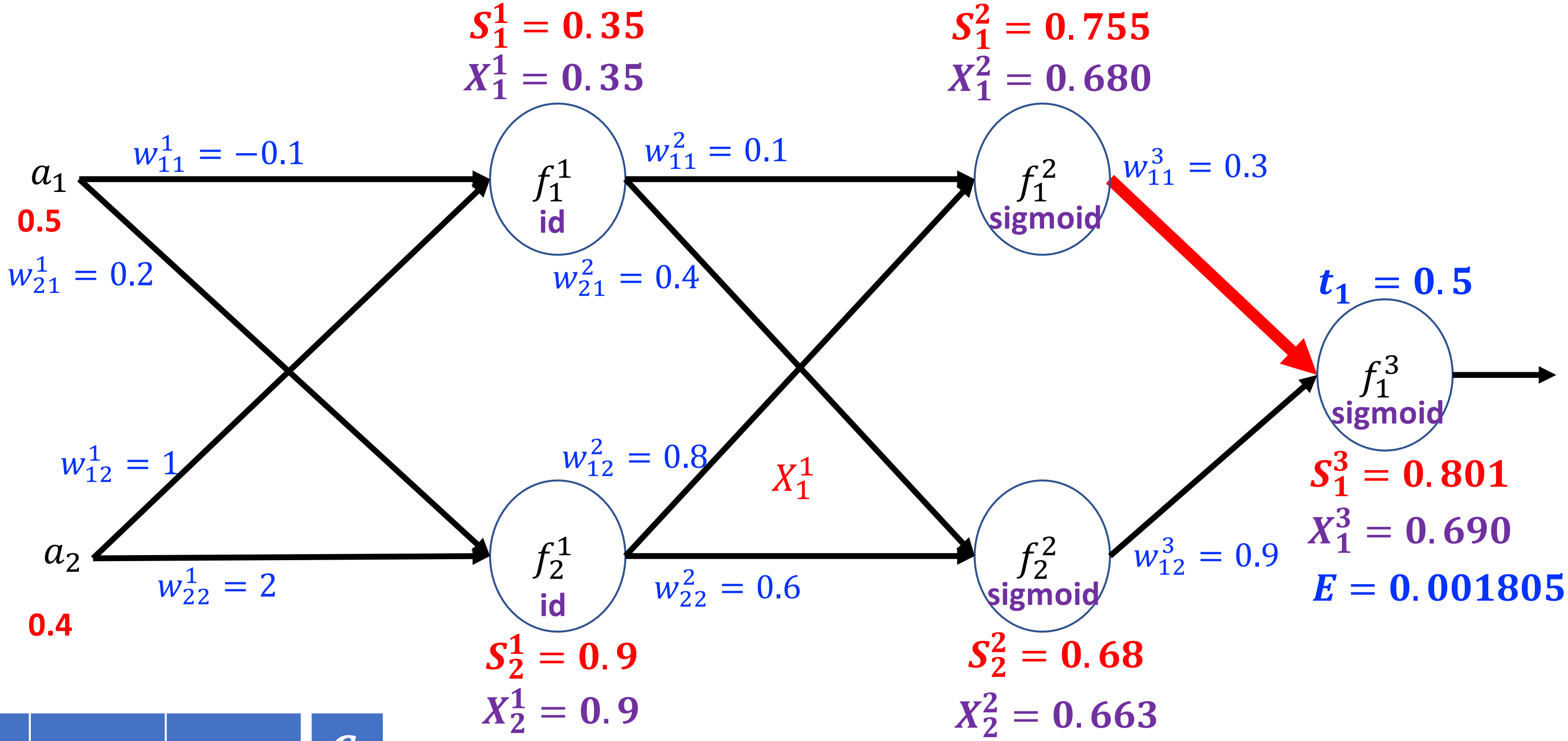
Gradient of a Weight



a_1	a_2	t_1	C
0.5	0.4	0.5	1

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

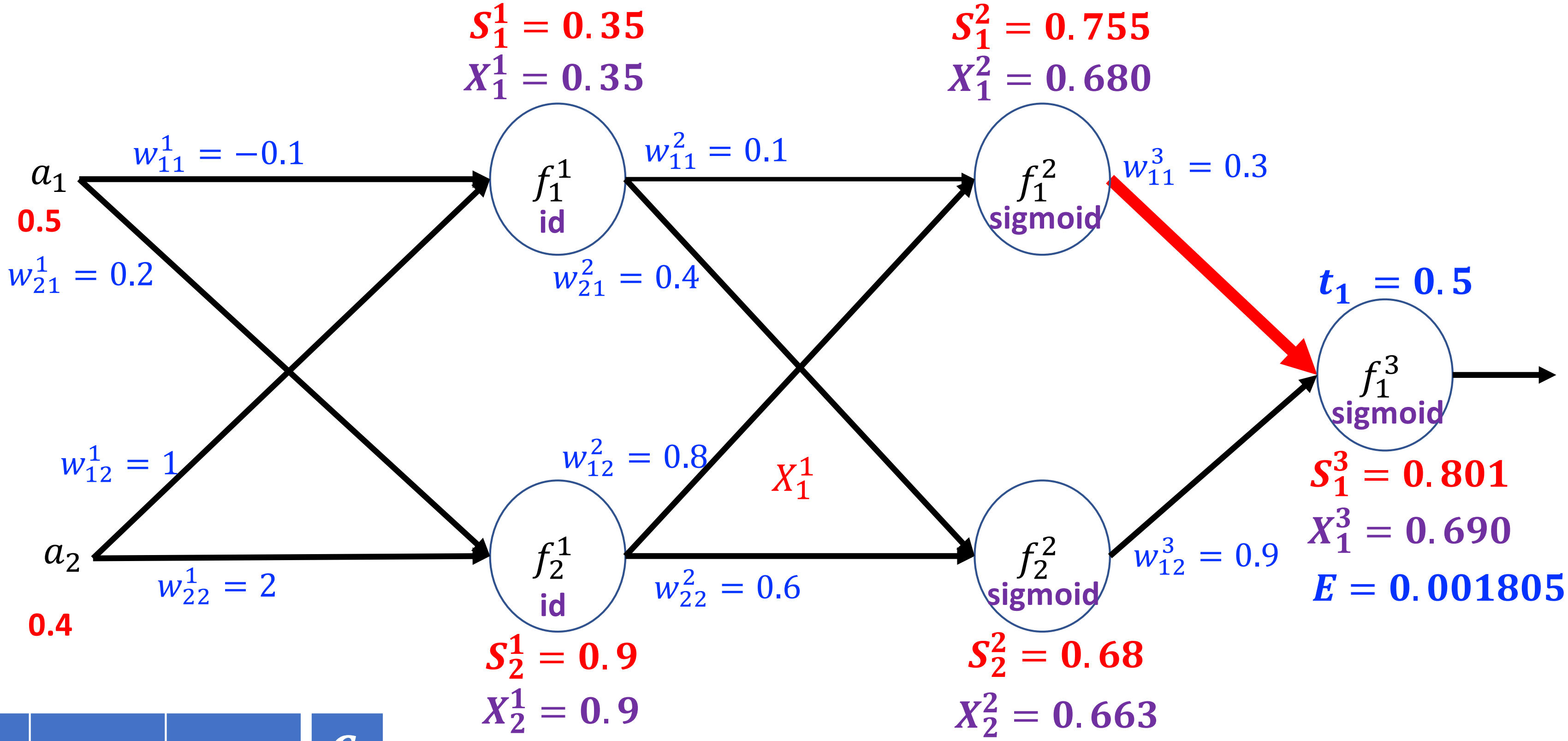
Gradient of a Weight



a_1	a_2	t_1	C
0.5	0.4	0.5	1

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

Gradient of a Weight



a_1	a_2	t_1
0.5	0.4	0.5

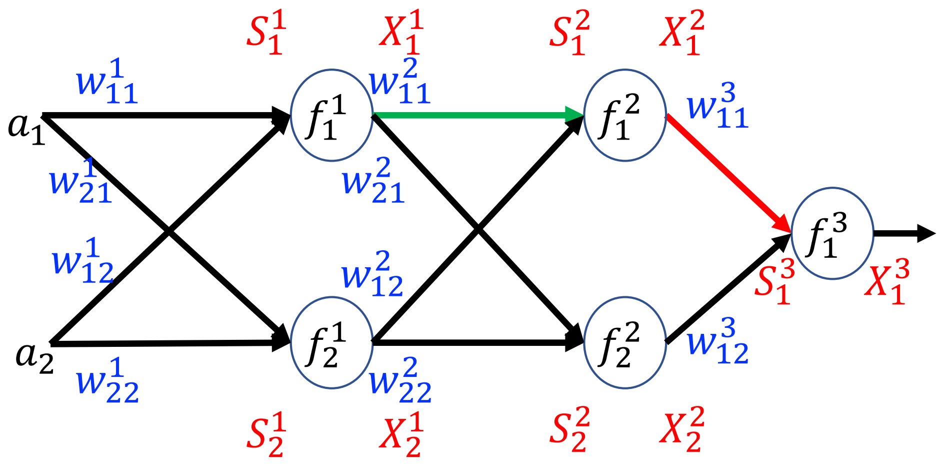
C
1

$$w_{11}^3 = w_{11}^3 - C\frac{\partial E}{\partial w_{11}^3} = 0.3 - 1 \times 0.02763 = 0.27237$$

Topic of Today’s Lecture

Backpropagation.

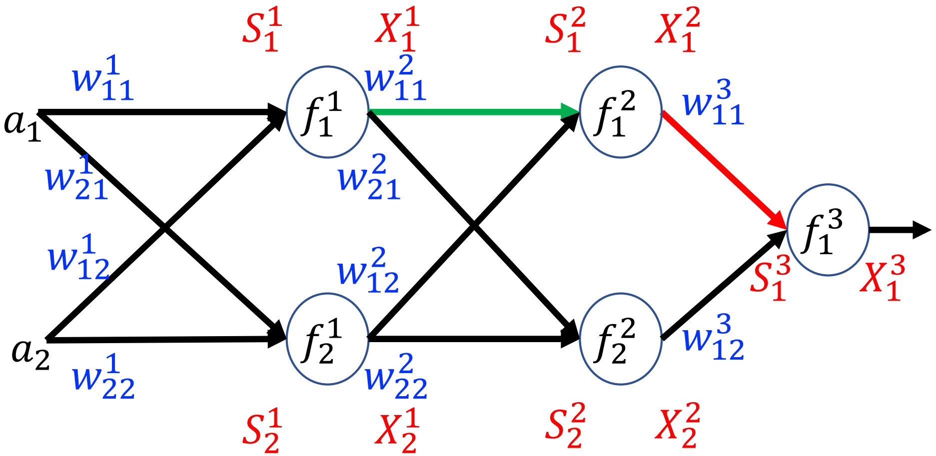
Backpropagation



$$\frac{\partial E}{\partial w_{11}^3} = (X_1^3 - t_1) \cdot (f_1^3)'(S_1^3) \cdot X_1^2$$

$$\frac{\partial E}{\partial w_{11}^2} = (X_1^3 - t_1) \cdot (f_1^3)'(S_1^3) \cdot w_{11}^3 \cdot (f_1^2)'(S_1^2) \cdot X_1^1$$

Backpropagation



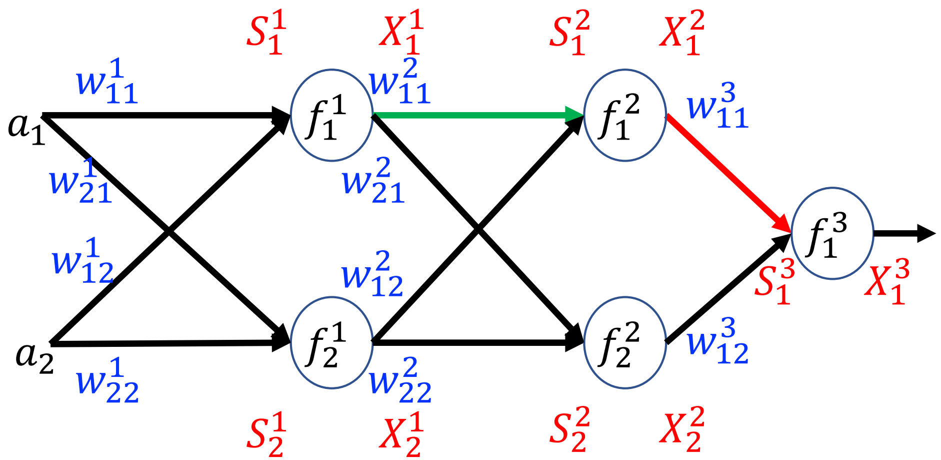
$$\frac{\partial E}{\partial w_{11}^3} = \underline{(X_1^3 - t_1) \cdot (f_1^3)'(S_1^3)} \cdot X_1^2$$

$$\frac{\partial E}{\partial w_{11}^2} = \underline{(X_1^3 - t_1) \cdot (f_1^3)'(S_1^3)} \cdot w_{11}^3 \cdot (f_1^2)'(S_1^2) \cdot X_1^1$$

- The closer the connection is to the output, the simpler the formula of the partial derivative of its weight is.

- There are some overlap between two formulas, indicating it is possible to store the overlap to avoid duplicate computation.

Backpropagation



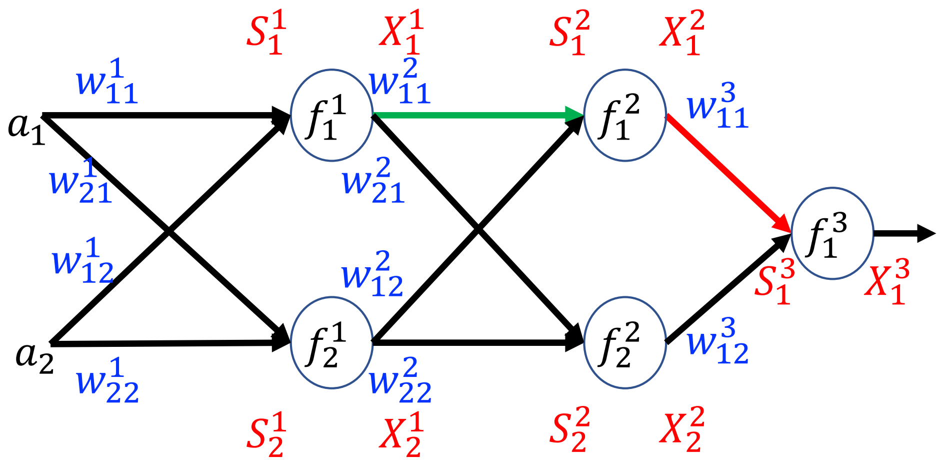
$$\frac{\partial E}{\partial w_{11}^3} = \underline{(X_1^3 - t_1) \cdot (f_1^3)'(S_1^3)} \cdot X_1^2$$

$$\frac{\partial E}{\partial w_{11}^2} = \underline{(X_1^3 - t_1) \cdot (f_1^3)'(S_1^3)} \cdot w_{11}^3 \cdot (f_1^2)'(S_1^2) \cdot X_1^1$$

First, we need to rewrite the following partial derivative formula we introduced previously.

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

Backpropagation



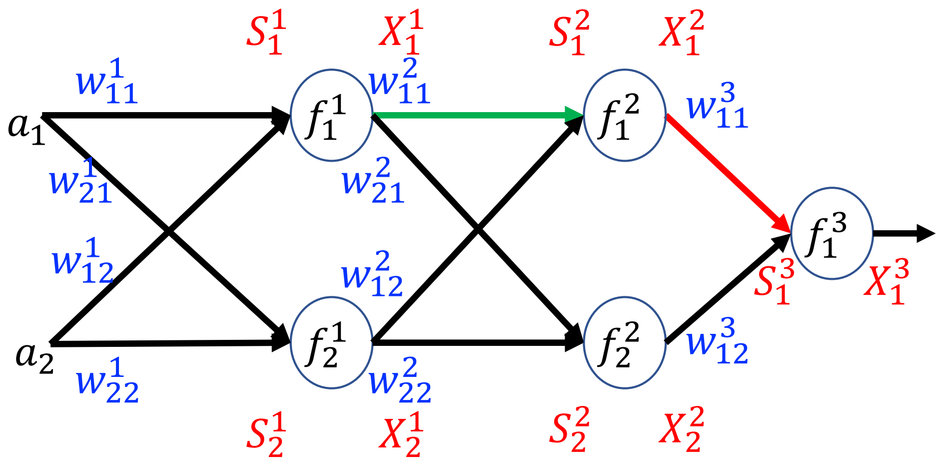
$$\frac{\partial E}{\partial w_{11}^3} = \underline{(X_1^3 - t_1) \cdot (f_1^3)'(S_1^3)} \cdot X_1^2$$

$$\frac{\partial E}{\partial w_{11}^2} = \underline{(X_1^3 - t_1) \cdot (f_1^3)'(S_1^3)} \cdot w_{11}^3 \cdot (f_1^2)'(S_1^2) \cdot X_1^1$$

Step 1. Rewrite the partial derivative.

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

Backpropagation



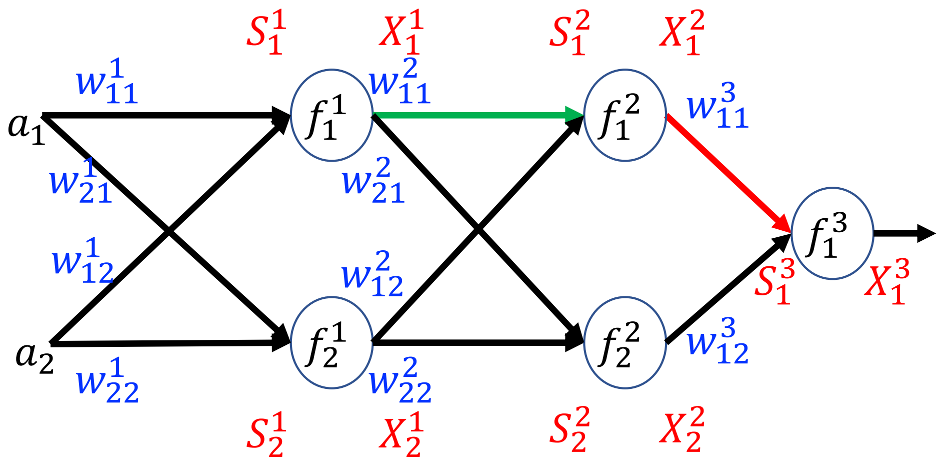
$$\frac{\partial E}{\partial w_{11}^3} = \underline{(X_1^3 - t_1)\cdot(f_1^3)'(S_1^3)}\cdot X_1^2$$

$$\frac{\partial E}{\partial w_{11}^2} = \underline{(X_1^3 - t_1)\cdot(f_1^3)'(S_1^3)}\cdot w_{11}^3\cdot(f_1^2)'(S_1^2)\cdot X_1^1$$

Step 1. Rewrite the partial derivative.

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

Backpropagation



$$\frac{\partial E}{\partial w_{11}^3} = \underline{(X_1^3 - t_1) \cdot (f_1^3)'(S_1^3)} \cdot X_1^2$$

$$\frac{\partial E}{\partial w_{11}^2} = \underline{(X_1^3 - t_1) \cdot (f_1^3)'(S_1^3)} \cdot w_{11}^3 \cdot (f_1^2)'(S_1^2) \cdot X_1^1$$

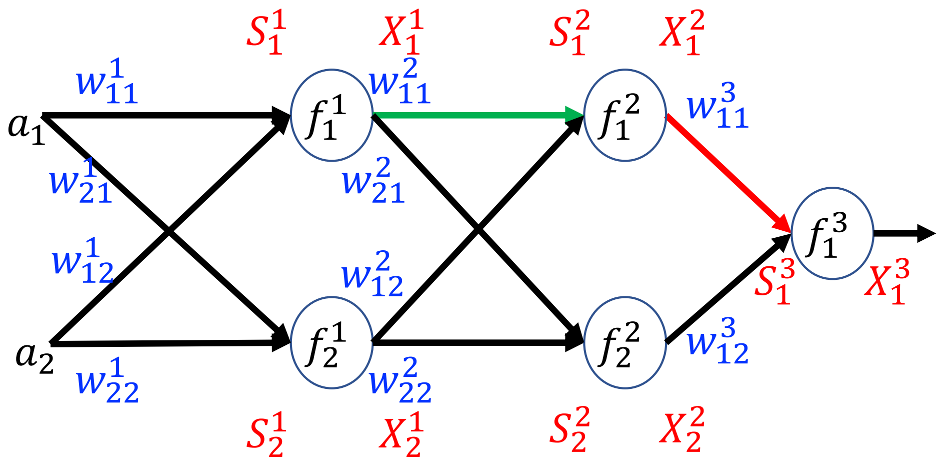
Why do we rewrite it?

Unify the representations in two cases.

Step 1. Rewrite the partial derivative.

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

Backpropagation



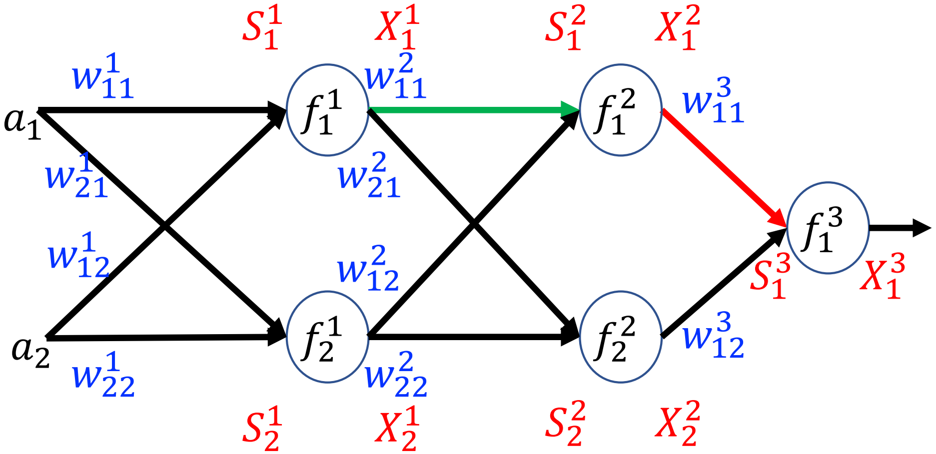
$$\frac{\partial E}{\partial w_{11}^3} = \underline{(X_1^3 - t_1) \cdot (f_1^3)'(S_1^3)} \cdot X_1^2$$

$$\frac{\partial E}{\partial w_{11}^2} = \underline{(X_1^3 - t_1) \cdot (f_1^3)'(S_1^3)} \cdot w_{11}^3 \cdot (f_1^2)'(S_1^2) \cdot X_1^1$$

Step 1. Rewrite the partial derivative.

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

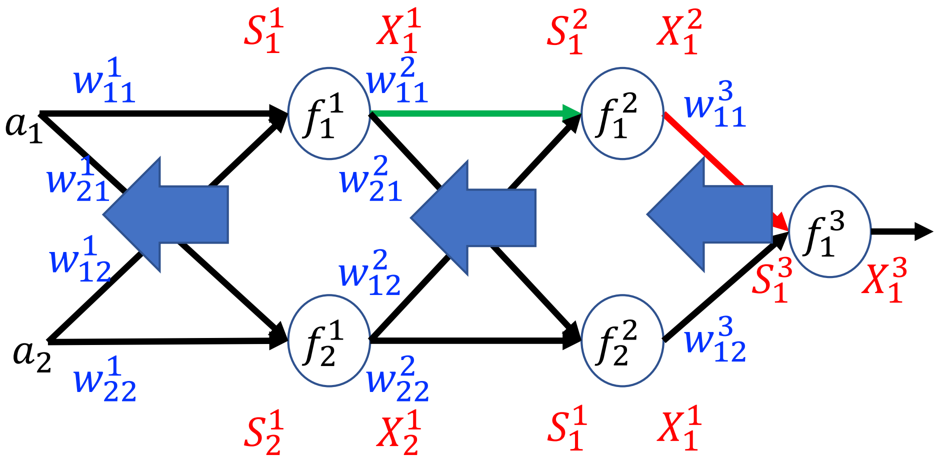
Backpropagation



$$\frac{\partial E}{\partial w^3_{11}} = \underline{(X^3_1 - t_1) \cdot (f^3_1)'(S^3_1)} \cdot X^2_1$$

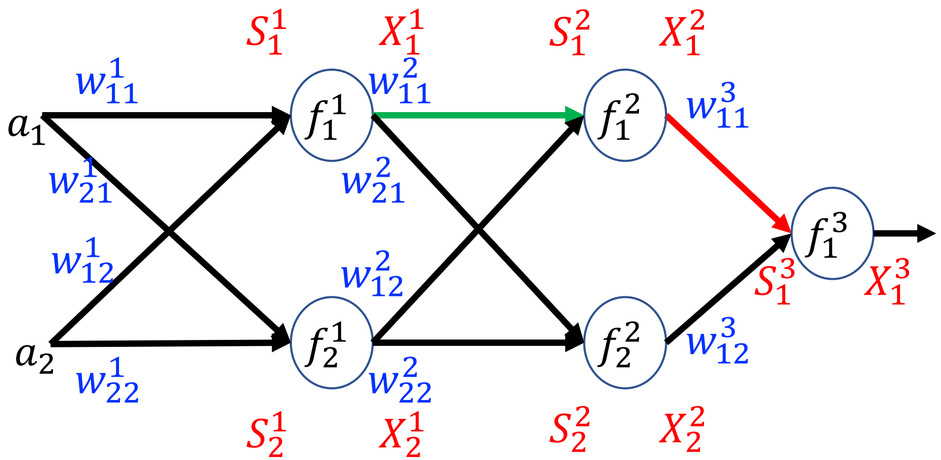
$$\frac{\partial E}{\partial w^2_{11}} = \underline{(X^3_1 - t_1) \cdot (f^3_1)'(S^3_1)} \cdot w^3_{11} \cdot (f^2_1)'(S^2_1) \cdot X^1_1$$

Step 2. Rewrite the formula in a matrix form.



The objective of this step is to derive the relation between the partial derivatives between two layers.

Backpropagation



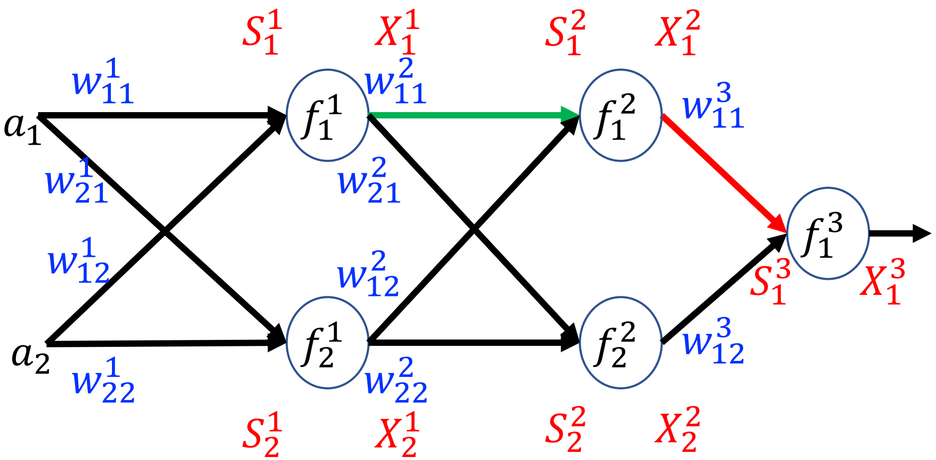
$$\frac{\partial E}{\partial w_{11}^3} = \underline{(X_1^3 - t_1) \cdot (f_1^3)'(S_1^3)} \cdot X_1^2$$

$$\frac{\partial E}{\partial w_{11}^2} = \underline{(X_1^3 - t_1) \cdot (f_1^3)'(S_1^3)} \cdot w_{11}^3 \cdot (f_1^2)'(S_1^2) \cdot X_1^1$$

Step 2. Rewrite the formula in a matrix form.

$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \underbrace{\sum_{j=1}^{n^l} (X_j^l - t_j)}_{1} \cdot \left((f_j^l)'(S_j^l) \cdot \sum_{i=1}^{n^{l-1}} w_{ji}^{l-1} \left(\cdots \left(f_{j'}^{l_0}\right)' \left(S_{j'}^{l_0}\right) \cdot \frac{\partial S_{j'}^{l_0}}{\partial w_{j_0 i_0}^{l_0}} \right) \right)$$

Backpropagation



$$\frac{\partial E}{\partial w^3_{11}} = \underline{(X^3_1 - t_1) \cdot (f^3_1)'(S^3_1)} \cdot X^2_1$$

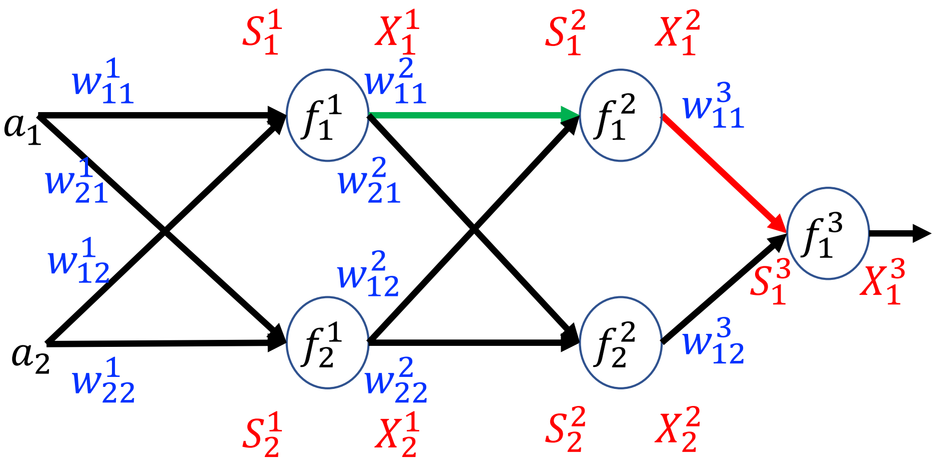
$$\frac{\partial E}{\partial w^2_{11}} = \underline{(X^3_1 - t_1) \cdot (f^3_1)'(S^3_1)} \cdot w^3_{11} \cdot (f^2_1)'(S^2_1) \cdot X^1_1$$

Step 2. Rewrite the formula in a matrix form. Let

$$\nabla_{X^l} E \triangleq \begin{pmatrix} \frac{\partial E}{\partial X^l_1} & \cdots & \frac{\partial E}{\partial X^l_{n^l}} \end{pmatrix} = \begin{pmatrix} X^l_1 - t_1 & \cdots & X^l_{n^l} - t_{n^l} \end{pmatrix}$$

be the gradient of the l-th layer.

Backpropagation



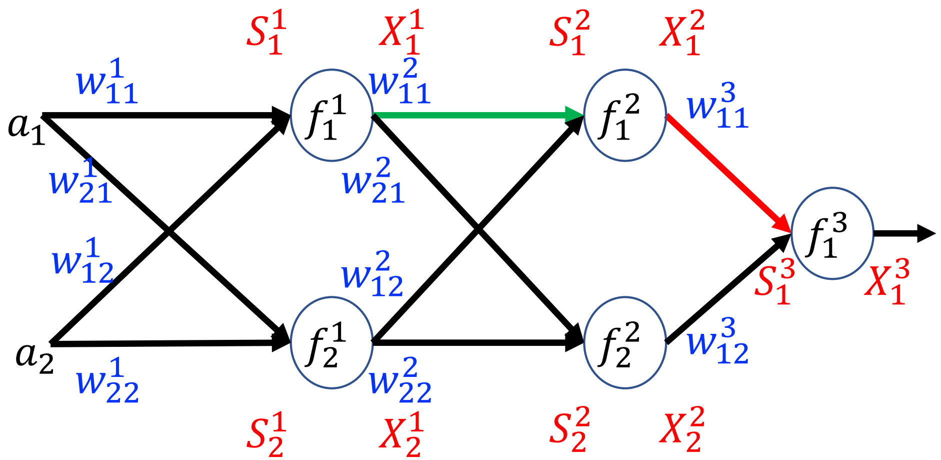
$$\frac{\partial E}{\partial w^3_{11}} = \underline{(X^3_1 - t_1) \cdot (f^3_1)'(S^3_1)} \cdot X^2_1$$

$$\frac{\partial E}{\partial w^2_{11}} = \underline{(X^3_1 - t_1) \cdot (f^3_1)'(S^3_1)} \cdot w^3_{11} \cdot (f^2_1)'(S^2_1) \cdot X^1_1$$

Step 2. Rewrite the formula in a matrix form.

$$\frac{\partial E}{\partial w^{l_0}_{j_0 i_0}} = \sum_{j=1}^{n^l} (X^l_j - t_j) \cdot \left(\underbrace{(f^l_j)'(S^l_j)}_{2} \cdot \sum_{i=1}^{n^{l-1}} w^{l-1}_{ji} \left(\cdots \left(f^{l_0}_{j'}\right)' \left(S^{l_0}_{j'}\right) \cdot \frac{\partial S^{l_0}_{j'}}{\partial w^{l_0}_{j_0 i_0}} \right) \right)$$

Backpropagation



$$\frac{\partial E}{\partial w_{11}^3} = \underline{(X_1^3 - t_1)\cdot (f_1^3)'(S_1^3)} \cdot X_1^2$$

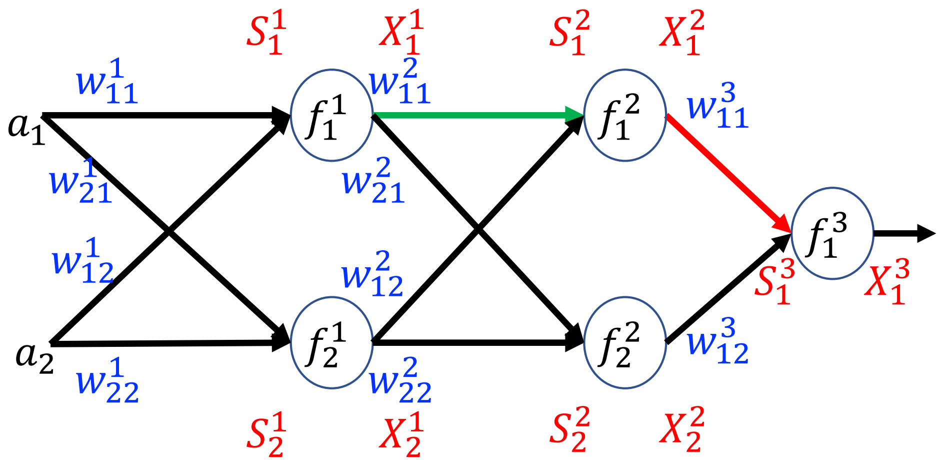
$$\frac{\partial E}{\partial w_{11}^2} = \underline{(X_1^3 - t_1)\cdot (f_1^3)'(S_1^3)} \cdot w_{11}^3 \cdot (f_1^2)'(S_1^2)\cdot X_1^1$$

Step 2. Rewrite the formula in a matrix form. Let

Diagonal Matrix!

$$\frac{dX^l}{dS^l} = \frac{df^l(S^l)}{dS^l} \triangleq \mathcal{J}_{S^l} f^l = \begin{pmatrix} \frac{df_1^l}{dS_1^l} & \cdots & \frac{df_1^l}{dS_{n^l}^l} \\ \vdots & \ddots & \vdots \\ \frac{df_{n^l}^l}{dS_1^l} & \cdots & \frac{df_{n^l}^l}{dS_{n^l}^l} \end{pmatrix} = \begin{pmatrix} (f_j^l)'(S_j^l) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & (f_{n^l}^l)'(S_{n^l}^l) \end{pmatrix}$$

Backpropagation



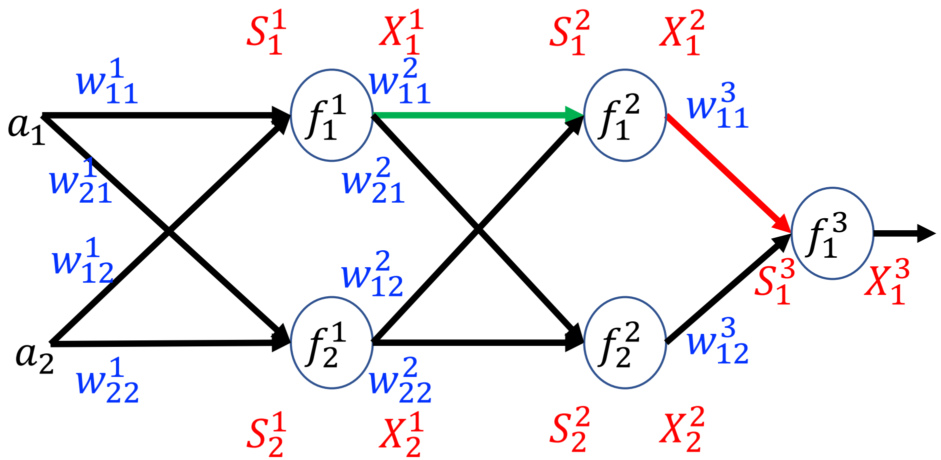
$$\frac{\partial E}{\partial w_{11}^3} = \underline{(X_1^3 - t_1) \cdot (f_1^3)'(S_1^3)} \cdot X_1^2$$

$$\frac{\partial E}{\partial w_{11}^2} = \underline{(X_1^3 - t_1) \cdot (f_1^3)'(S_1^3)} \cdot w_{11}^3 \cdot (f_1^2)'(S_1^2) \cdot X_1^1$$

Step 2. Rewrite the formula in a matrix form.

$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \sum_{j=1}^{n^l} (X_j^l - t_j) \cdot \left((f_j^l)'(S_j^l) \cdot \underbrace{\sum_{i=1}^{n^{l-1}} w_{ji}^{l-1}}_{3} \left(\cdots \left(f_{j'}^{l_0}\right)' \left(S_{j'}^{l_0}\right) \cdot \frac{\partial S_{j'}^{l_0}}{\partial w_{j_0 i_0}^{l_0}} \right) \right)$$

Backpropagation



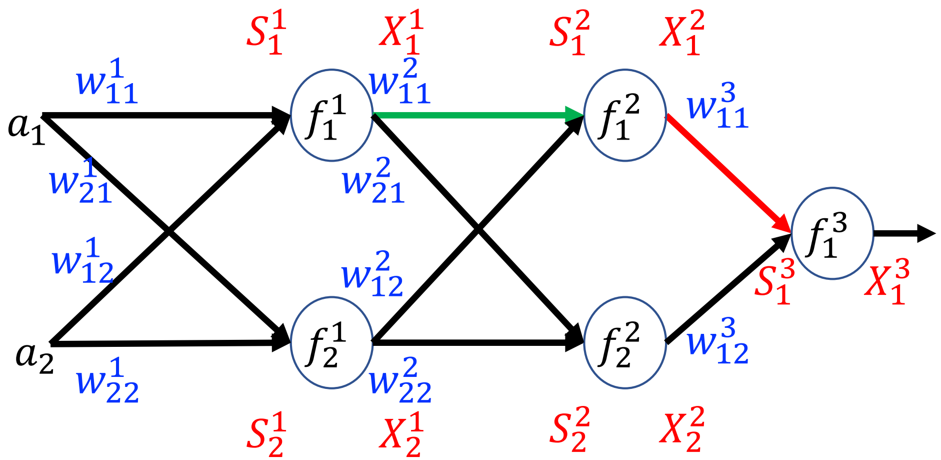
$$\frac{\partial E}{\partial w_{11}^3} = \underline{(X_1^3 - t_1)\cdot (f_1^3)'(S_1^3)} \cdot X_1^2$$

$$\frac{\partial E}{\partial w_{11}^2} = \underline{(X_1^3 - t_1)\cdot (f_1^3)'(S_1^3)} \cdot w_{11}^3 \cdot (f_1^2)'(S_1^2) \cdot X_1^1$$

Step 2. Rewrite the formula in a matrix form. Let

$$\frac{dS^l}{dX^{l-1}} = \begin{pmatrix} \frac{\partial S_1^l}{\partial X_1^{l-1}} & \cdots & \frac{\partial S_1^l}{\partial X_{n^{l-1}}^{l-1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial S_{n^l}^l}{\partial X_1^l} & \cdots & \frac{\partial S_{n^l}^l}{\partial X_{n^{l-1}}^{l-1}} \end{pmatrix} = \begin{pmatrix} w_{11}^l & \cdots & w_{1n^{l-1}}^l \\ \vdots & \ddots & \vdots \\ w_{n^l 1}^l & \cdots & w_{n^l n^{l-1}}^l \end{pmatrix} \triangleq w^l$$

Backpropagation



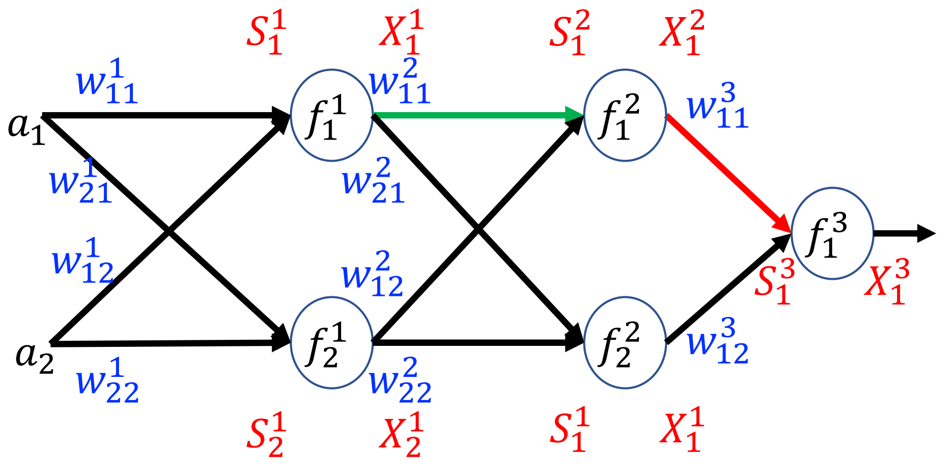
$$\frac{\partial E}{\partial w_{11}^3} = \underline{(X_1^3 - t_1) \cdot (f_1^3)'(S_1^3)} \cdot X_1^2$$

$$\frac{\partial E}{\partial w_{11}^2} = \underline{(X_1^3 - t_1) \cdot (f_1^3)'(S_1^3)} \cdot w_{11}^3 \cdot (f_1^2)'(S_1^2) \cdot X_1^1$$

Step 2. Rewrite the formula in a matrix form.

$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \sum_{j=1}^{n^l} (X_j^l - t_j) \cdot \left(\left(f_j^l\right)' \left(S_j^l\right) \cdot \sum_{i=1}^{n^{l-1}} w_{ji}^{l-1} \left(\cdots \left(f_{j'}^{l_0}\right)' \left(S_{j'}^{l_0}\right) \cdot \underbrace{\frac{\partial S_{j'}^{l_0}}{\partial w_{j_0 i_0}^{l_0}}}_{4} \right) \right)$$

Backpropagation



$$\frac{\partial E}{\partial w_{11}^3} = \underline{(X_1^3 - t_1) \cdot (f_1^3)'(S_1^3)} \cdot X_1^2$$

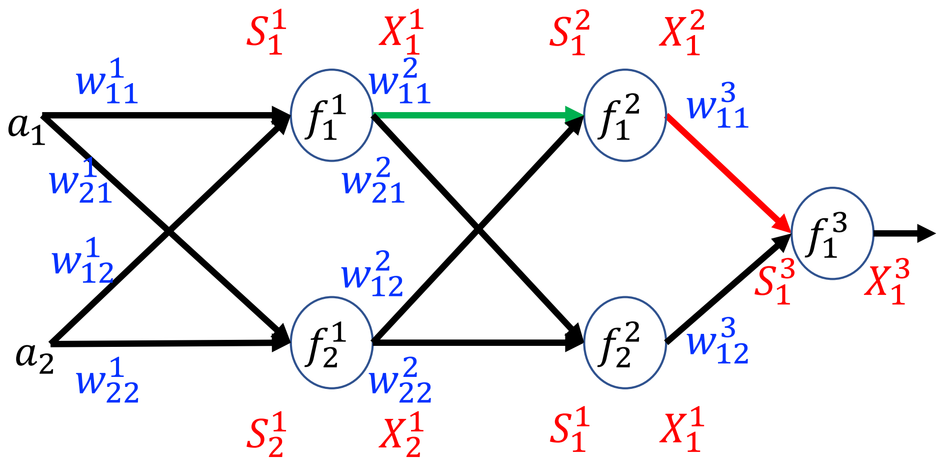
$$\frac{\partial E}{\partial w_{11}^2} = \underline{(X_1^3 - t_1) \cdot (f_1^3)'(S_1^3)} \cdot w_{11}^3 \cdot (f_1^2)'(S_1^2) \cdot X_1^1$$

Step 2. Rewrite the formula in a matrix form. Let

$$\frac{\partial S^{l_0}}{\partial w_{j_0 i_0}^{l_0}} \triangleq \begin{pmatrix} \frac{\partial S_1^{l_0}}{\partial w_{j_0 i_0}^{l_0}} & \cdots & \frac{\partial S_{n^{l_0}}^{l_0}}{\partial w_{j_0 i_0}^{l_0}} \end{pmatrix}^T = \begin{pmatrix} 0 & \cdots & \frac{\partial S_{j_0}^{l_0}}{\partial w_{j_0 i_0}^{l_0}} & \cdots & 0 \end{pmatrix}^T = \begin{pmatrix} 0 & \cdots & X_{i_0}^{l_0 - 1} & \cdots & 0 \end{pmatrix}^T$$

where $S^{l_0} = \begin{pmatrix} S_1^{l_0} & \cdots & S_{n^{l_0}}^{l_0} \end{pmatrix}^T$.

Backpropagation



$$\frac{\partial E}{\partial w_{11}^3} = \underline{(X_1^3 - t_1) \cdot (f_1^3)'(S_1^3)} \cdot X_1^2$$

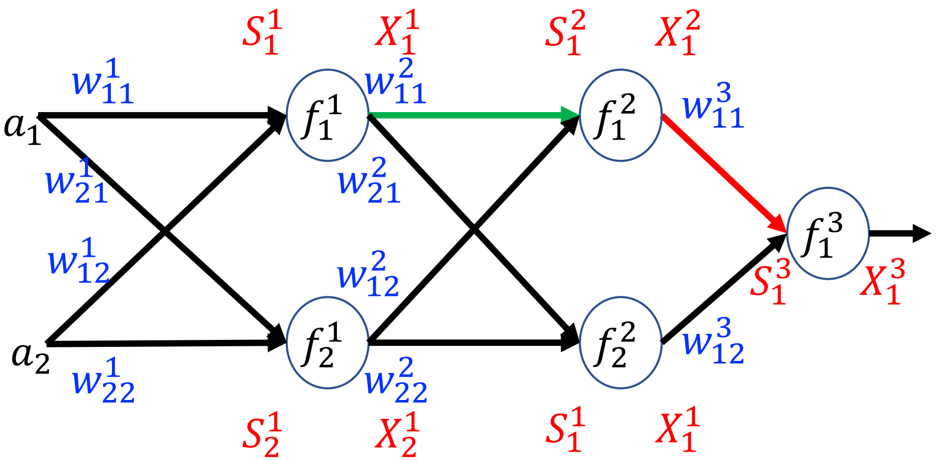
$$\frac{\partial E}{\partial w_{11}^2} = \underline{(X_1^3 - t_1) \cdot (f_1^3)'(S_1^3)} \cdot w_{11}^3 \cdot (f_1^2)'(S_1^2) \cdot X_1^1$$

Step 2. Rewrite the formula in a matrix form. Now we obtain:

$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \frac{dE}{dX^l} \cdot \frac{dX^l}{dS^l} \cdot \frac{dS^l}{dX^{l-1}} \cdot \frac{dX^{l-1}}{dS^{l-1}} \cdot \dots \cdot \frac{\partial S^{l_0}}{\partial w_{j_0 i_0}^{l_0}} = \nabla_{X^l} E \cdot \mathcal{J}_{S^l} f^l \cdot w^l \cdot \mathcal{J}_{S^{l-1}} f^{l-1} \cdot \dots \cdot \frac{\partial S^{l_0}}{\partial w_{j_0 i_0}^{l_0}}$$

This is the multivariate (vector) version of chain rule!

Backpropagation



$$\frac{\partial E}{\partial w_{11}^3} = \underline{(X_1^3 - t_1) \cdot (f_1^3)'(S_1^3)} \cdot X_1^2$$

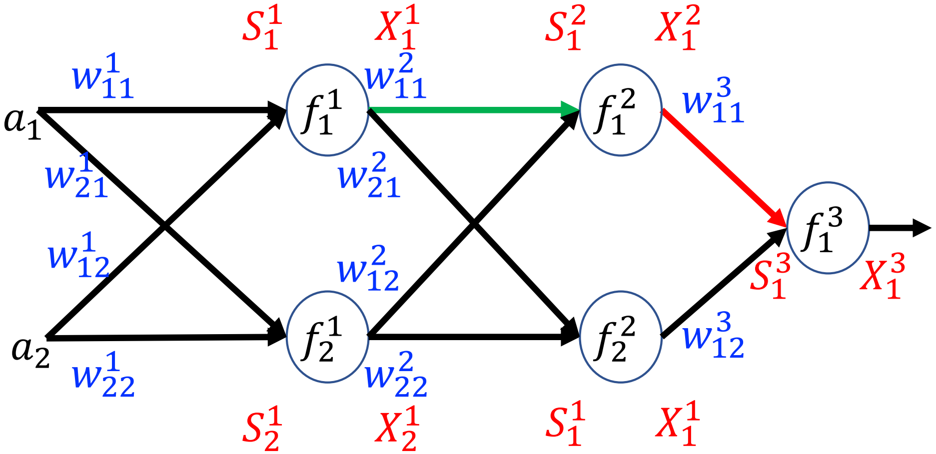
$$\frac{\partial E}{\partial w_{11}^2} = \underline{(X_1^3 - t_1) \cdot (f_1^3)'(S_1^3)} \cdot w_{11}^3 \cdot (f_1^2)'(S_1^2) \cdot X_1^1$$

$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \frac{dE}{dX^l} \cdot \frac{dX^l}{dS^l} \cdot \frac{dS^l}{dX^{l-1}} \cdot \frac{dX^{l-1}}{dS^{l-1}} \cdot \cdots \cdot \frac{\partial S^{l_0}}{\partial w_{j_0 i_0}^{l_0}} = \nabla_{X^l} E \cdot \mathcal{J}_{S^l} f^l \cdot w^l \cdot \mathcal{J}_{S^{l-1}} f^{l-1} \cdot \cdots \cdot \frac{\partial S^{l_0}}{\partial w_{j_0 i_0}^{l_0}}$$

Let $\delta^h = \nabla_{X^l} E \cdot \mathcal{J}_{S^l} f^l \cdot w^l \cdot \mathcal{J}_{S^{l-1}} f^{l-1} \cdot \cdots \cdot \mathcal{J}_{S^h} f^h$. Then we have

$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \delta^{l_0} \cdot \frac{\partial S^{l_0}}{\partial w_{j_0 i_0}^{l_0}} = \delta^{l_0} \cdot \begin{pmatrix} 0 & \cdots & X_{i_0}^{l_0 - 1} & \cdots & 0 \end{pmatrix}^T$$

Backpropagation



$$\frac{\partial E}{\partial w^3_{11}} = \underline{(X^3_1 - t_1) \cdot (f^3_1)'(S^3_1)} \cdot \boxed{X^2_1}$$

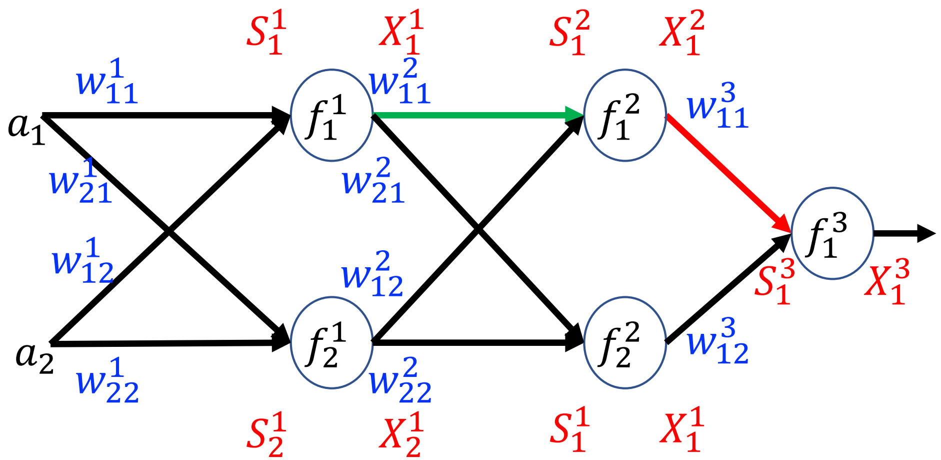
$$\frac{\partial E}{\partial w^2_{11}} = \underline{(X^3_1 - t_1) \cdot (f^3_1)'(S^3_1)} \cdot w^3_{11} \cdot (f^2_1)'(S^2_1) \cdot \boxed{X^1_1}$$

$$\frac{\partial E}{\partial w^{l_0}_{j_0 i_0}} = \frac{dE}{dX^l} \cdot \frac{dX^l}{dS^l} \cdot \frac{dS^l}{dX^{l-1}} \cdot \frac{dX^{l-1}}{dS^{l-1}} \cdot \dots \cdot \frac{\partial S^{l_0}}{\partial w^{l_0}_{j_0 i_0}} = \nabla_{X^l} E \cdot \mathcal{J}_{S^l} f^l \cdot w^l \cdot \mathcal{J}_{S^{l-1}} f^{l-1} \cdot \dots \cdot \frac{\partial S^{l_0}}{\partial w^{l_0}_{j_0 i_0}}$$

Let $\delta^h = \nabla_{X^l} E \cdot \mathcal{J}_{S^l} f^l \cdot w^l \cdot \mathcal{J}_{S^{l-1}} f^{l-1} \cdot \dots \cdot \mathcal{J}_{S^h} f^h$. Then we have

$$\frac{\partial E}{\partial w^{l_0}_{j_0 i_0}} = \delta^{l_0} \cdot \frac{\partial S^{l_0}}{\partial w^{l_0}_{j_0 i_0}} = \delta^{l_0} \cdot \boxed{\begin{pmatrix} 0 & \cdots & X^{l_0-1}_{i_0} & \cdots & 0 \end{pmatrix}^T}$$

Backpropagation



$$\frac{\partial E}{\partial w_{11}^3} = \underline{(X_1^3 - t_1) \cdot (f_1^3)'(S_1^3)} \cdot X_1^2$$

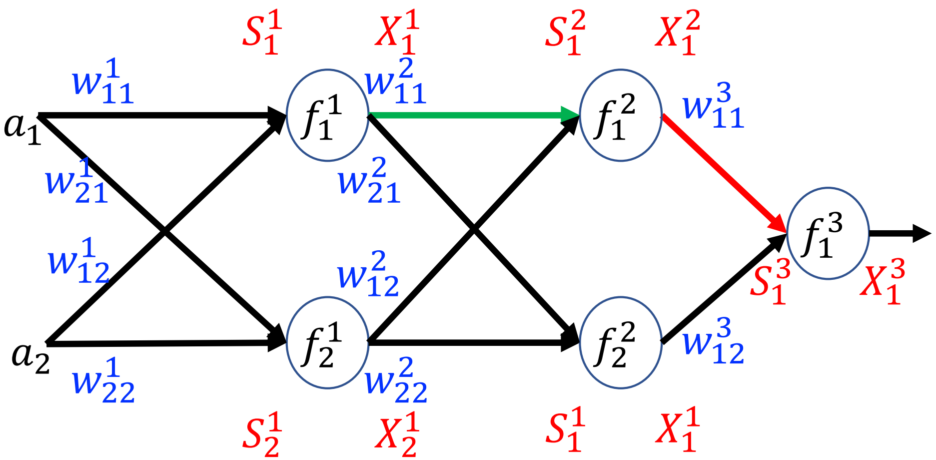
$$\frac{\partial E}{\partial w_{11}^2} = \underline{(X_1^3 - t_1) \cdot (f_1^3)'(S_1^3)} \cdot w_{11}^3 \cdot (f_1^2)'(S_1^2) \cdot X_1^1$$

$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \frac{dE}{dX^l} \cdot \frac{dX^l}{dS^l} \cdot \frac{dS^l}{dX^{l-1}} \cdot \frac{dX^{l-1}}{dS^{l-1}} \cdot \cdots \cdot \frac{\partial S^{l_0}}{\partial w_{j_0 i_0}^{l_0}} = \nabla_{X^l} E \cdot \mathcal{J}_{S^l} f^l \cdot w^l \cdot \mathcal{J}_{S^{l-1}} f^{l-1} \cdot \cdots \cdot \frac{\partial S^{l_0}}{\partial w_{j_0 i_0}^{l_0}}$$

Let $\delta^h = \nabla_{X^l} E \cdot \mathcal{J}_{S^l} f^l \cdot w^l \cdot \mathcal{J}_{S^{l-1}} f^{l-1} \cdot \cdots \cdot \mathcal{J}_{S^h} f^h$. Meanwhile, δ^h can be computed recursively:

$$\delta^{h-1} = \delta^h \cdot w^h \cdot \mathcal{J}_{S^{h-1}} f^{h-1}$$

Backpropagation



$$\frac{\partial E}{\partial w_{11}^3} = \underline{(X_1^3 - t_1)\cdot(f_1^3)'(S_1^3)}\cdot X_1^2$$

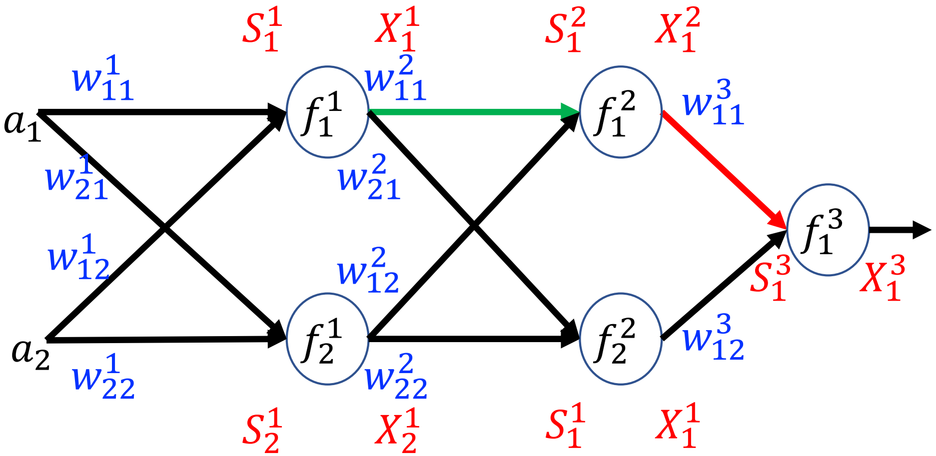
$$\frac{\partial E}{\partial w_{11}^2} = \underline{(X_1^3 - t_1)\cdot(f_1^3)'(S_1^3)}\cdot \boxed{w_{11}^3\cdot(f_1^2)'(S_1^2)}\cdot X_1^1$$

$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \frac{dE}{dX^l}\cdot\frac{dX^l}{dS^l}\cdot\frac{dS^l}{dX^{l-1}}\cdot\frac{dX^{l-1}}{dS^{l-1}}\cdot\cdots\cdot\frac{\partial S^{l_0}}{\partial w_{j_0 i_0}^{l_0}} = \nabla_{X^l}E\cdot\mathcal{J}_{S^l}f^l\cdot w^l\cdot\mathcal{J}_{S^{l-1}}f^{l-1}\cdot\cdots\cdot\frac{\partial S^{l_0}}{\partial w_{j_0 i_0}^{l_0}}$$

Let $\delta^h = \nabla_{X^l}E\cdot\mathcal{J}_{S^l}f^l\cdot w^l\cdot\mathcal{J}_{S^{l-1}}f^{l-1}\cdot\cdots\cdot\mathcal{J}_{S^h}f^h$. Meanwhile, δ^h can be computed recursively:

$$\delta^{h-1} = \delta^h\cdot\boxed{w^h\cdot\mathcal{J}_{S^{h-1}}f^{h-1}}$$

Backpropagation



$$\frac{\partial E}{\partial w_{11}^3} = (X_1^3 - t_1) \cdot (f_1^3)'(S_1^3) \cdot X_1^2$$

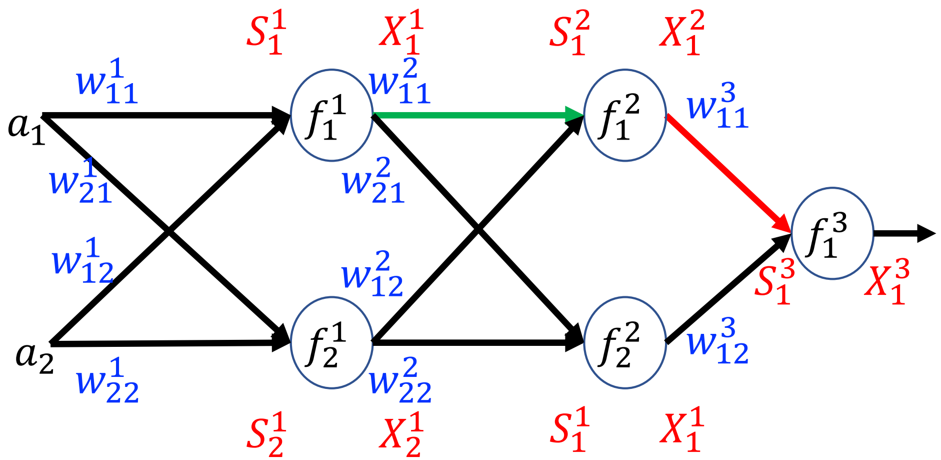
$$\frac{\partial E}{\partial w_{11}^2} = (X_1^3 - t_1) \cdot (f_1^3)'(S_1^3) \cdot \boxed{w_{11}^3 \cdot (f_1^2)'(S_1^2)} \cdot X_1^1$$

$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \frac{dE}{dX^l} \cdot \frac{dX^l}{dS^l} \cdot \frac{dS^l}{dX^{l-1}} \cdot \frac{dX^{l-1}}{dS^{l-1}} \cdot \cdots \cdot \frac{\partial S^{l_0}}{\partial w_{j_0 i_0}^{l_0}} = \nabla_{X^l} E \cdot \mathcal{J}_{S^l} f^l \cdot w^l \cdot \mathcal{J}_{S^{l-1}} f^{l-1} \cdot \cdots \cdot \frac{\partial S^{l_0}}{\partial w_{j_0 i_0}^{l_0}}$$

Let $\delta^h = \nabla_{X^l} E \cdot \mathcal{J}_{S^l} f^l \cdot w^l \cdot \mathcal{J}_{S^{l-1}} f^{l-1} \cdot \cdots \cdot \mathcal{J}_{S^h} f^h$. Meanwhile, δ^h can be computed recursively:

$$\delta^{h-1} = \boxed{\delta^h} \cdot \boxed{w^h \cdot \mathcal{J}_{S^{h-1}} f^{h-1}}$$

Backpropagation



$$\frac{\partial E}{\partial w_{11}^3} = \underline{(X_1^3 - t_1)\cdot(f_1^3)'(S_1^3)}\cdot X_1^2$$

$$\frac{\partial E}{\partial w_{11}^2} = \underline{(X_1^3 - t_1)\cdot(f_1^3)'(S_1^3)}\cdot \boxed{w_{11}^3\cdot(f_1^2)'(S_1^2)}\cdot X_1^1$$

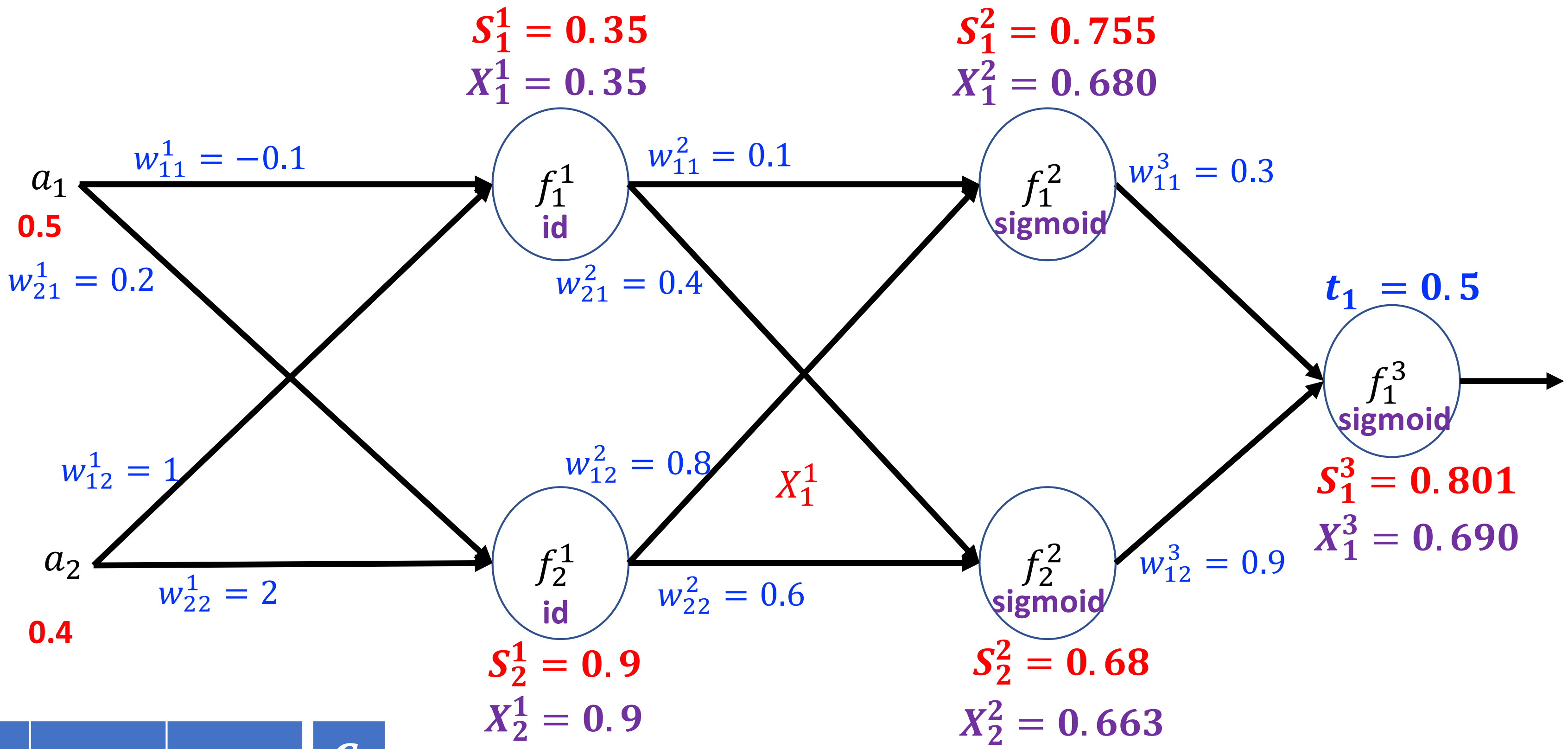
$$\frac{\partial E}{\partial w_{j_0 i_0}^{l_0}} = \frac{dE}{dX^l}\cdot\frac{dX^l}{dS^l}\cdot\frac{dS^l}{dX^{l-1}}\cdot\frac{dX^{l-1}}{dS^{l-1}}\cdot\cdots\cdot\frac{\partial S^{l_0}}{\partial w_{j_0 i_0}^{l_0}} = \nabla_{X^l}E\cdot\mathcal{J}_{S^l}f^l\cdot w^l\cdot\mathcal{J}_{S^{l-1}}f^{l-1}\cdot\cdots\cdot\frac{\partial S^{l_0}}{\partial w_{j_0 i_0}^{l_0}}$$

Let $\delta^h = \nabla_{X^l}E\cdot\mathcal{J}_{S^l}f^l\cdot w^l\cdot\mathcal{J}_{S^{l-1}}f^{l-1}\cdot\cdots\cdot\mathcal{J}_{S^h}f^h$. Meanwhile, δ^h can be computed recursively:

$$\delta^{h-1} = \boxed{\delta^h}\cdot\boxed{w^h\cdot\mathcal{J}_{S^{h-1}}f^{h-1}}$$

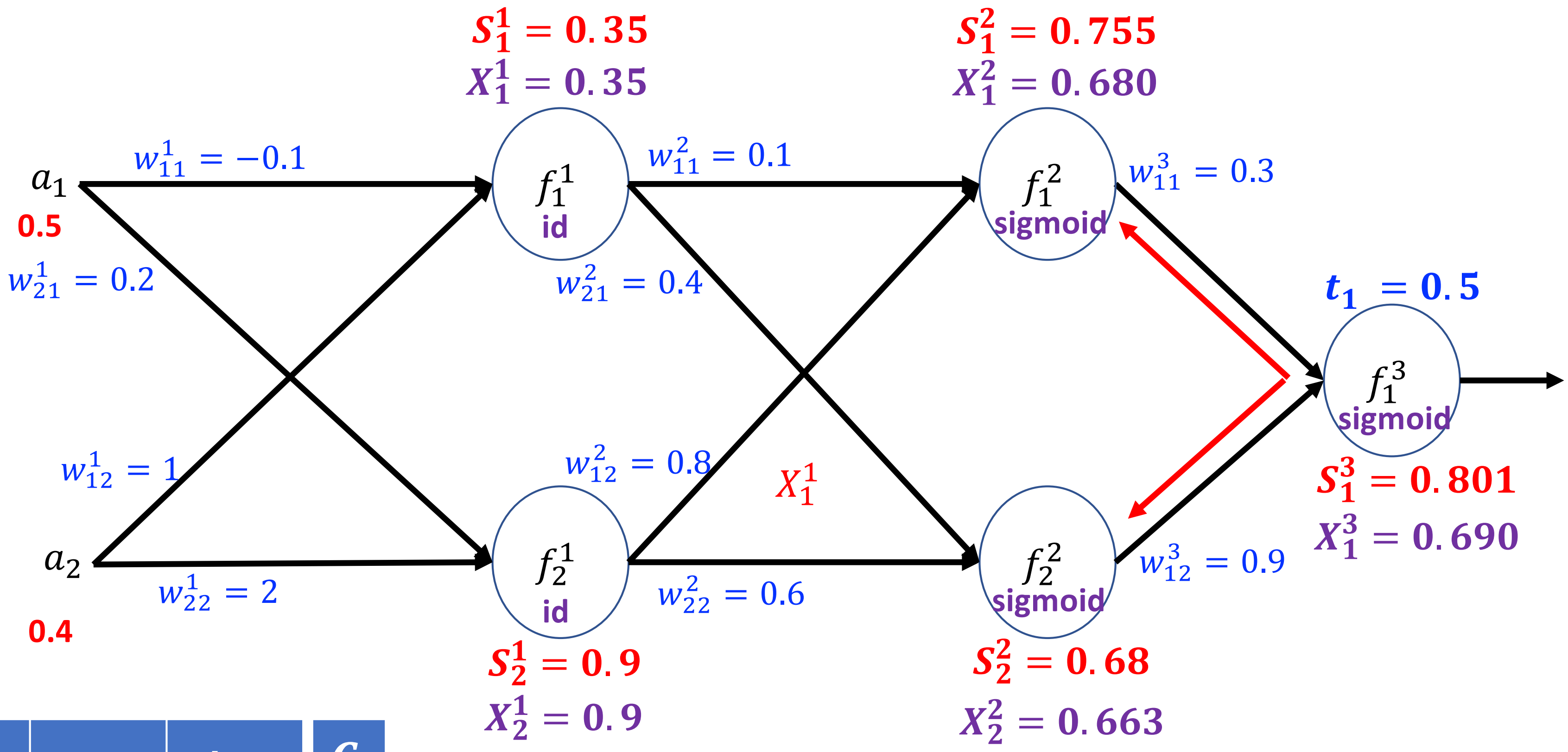
This recursive computing rule is called *backpropagation*!

Recall the Example



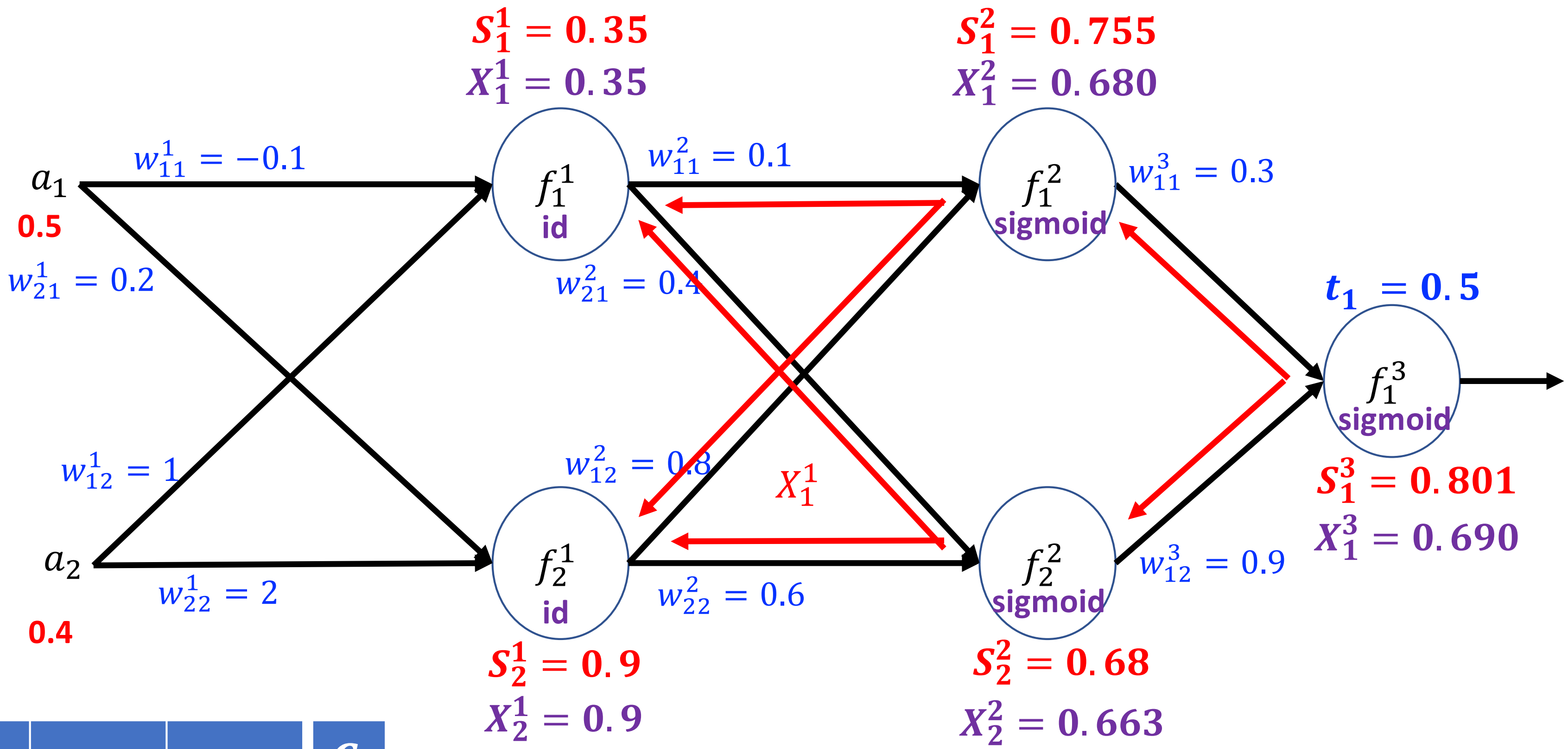
a_1	a_2	t_1	C
0.5	0.4	0.5	1

Recall the Example



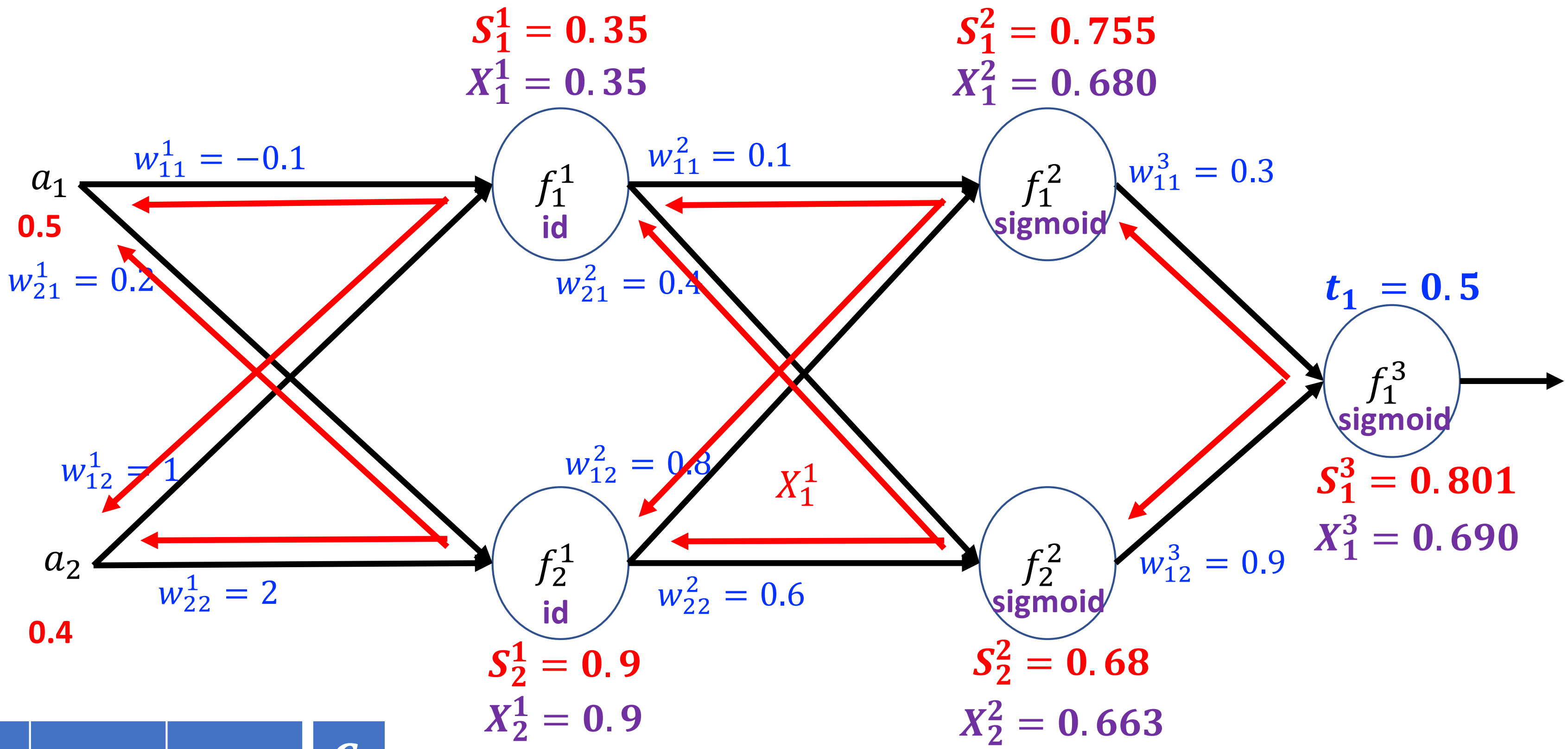
a_1	a_2	t_1	C
0.5	0.4	0.5	1

Recall the Example



a_1	a_2	t_1	C
0.5	0.4	0.5	1

Recall the Example



a_1	a_2	t_1	C
0.5	0.4	0.5	1

Comparing to the Naïve Computation

- Computing δ^{h-1} in terms of δ^{h} avoids obvious duplicate computation, more specifically multiplication, of layers h and beyond.

- Multiplying starting from the output layer – propagating the error ***backwards*** – means that each step simply multiplies a vector (δ^{h}) and a weight matrix (w^{h-1}), which makes matrix operations so important in neural network forward/backpropagation.

Table 3. Performance Improvements Optimizing C = AA^T Matrix Multiplication

Optimization	NVIDIA Tesla V100
No optimization	12.8 GB/s
Using shared memory to coalesce global reads	140.2 GB/s
Removing bank conflicts	199.4 GB/s

Source: Nvidia CUDA C++ Best Practices Guide

Learning Algorithm

Algorithm 3: Multilayer Perceptron Learning Algorithm

Data: Labelled data set D: r n-dimensional input points, each of which has m labels. Small positive real δ. Learning rate C.

Result: Weights w of all the connections.

Initialize weights w randomly;
while *!convergence* $(E \leq \delta)$ **do**
    Pick random $a \in D$;
    Compute the output X;
    Compute the output error E_a for this input a;
    $\delta^l = \nabla_{X^l} E_a \cdot \mathcal{J}_{S^l} f^l$;
    for $h = l, \cdots, 1$ **do**
        $\frac{\partial E_a}{\partial w^h} = \delta^h \cdot X^h$;
        $\delta^{h-1} = \delta^h \cdot w^h \cdot \mathcal{J}_{S^{h-1}} f^{h-1}$;
    end
    for $h = l, \cdots, 1$ **do**
        for $j = 1, \cdots, n^h$ **do**
            for $i = 1, \cdots, n^{h-1}$ **do**
                $w^h_{j,i} = w^h_{j,i} - C \frac{\partial E_a}{\partial w^h_{j,i}}$;
            end
        end
    end
end
return w;

- Randomly set initial values of parameters to be learnt.
- Compute the output error by forward propagation.
- The first FOR loop is to compute the partial derivative for each layer by backpropagation.
- We then update the weight of every connection in the network by gradient decent (lazy update).