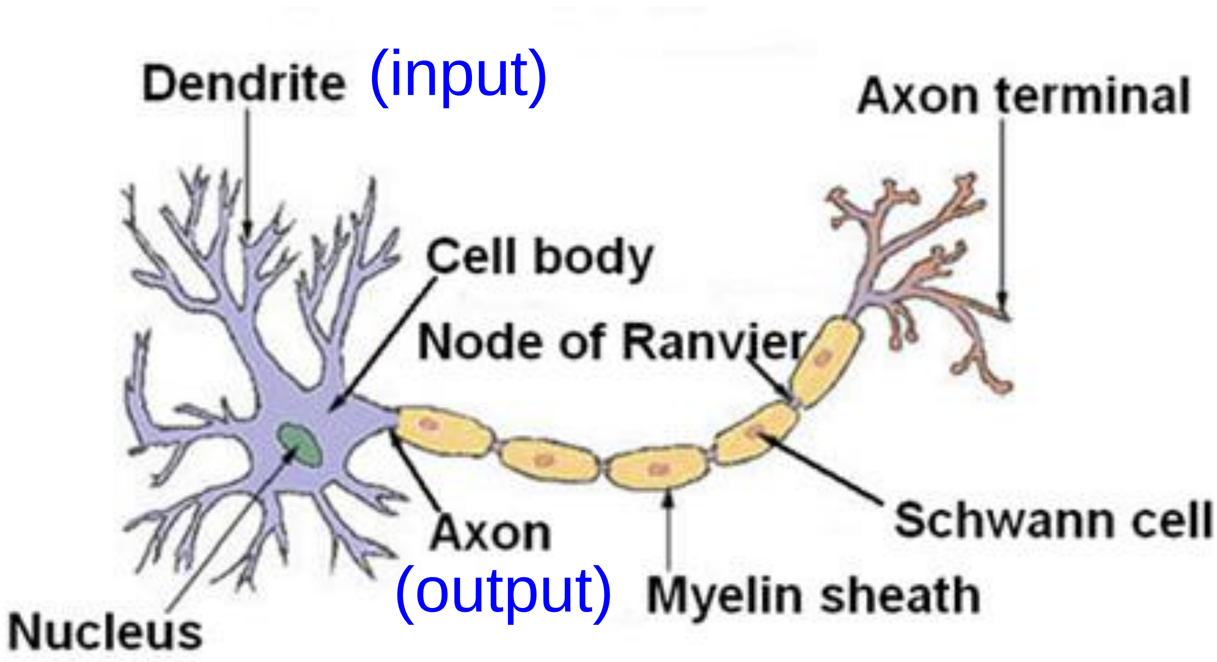
CENG 506 Deep Learning

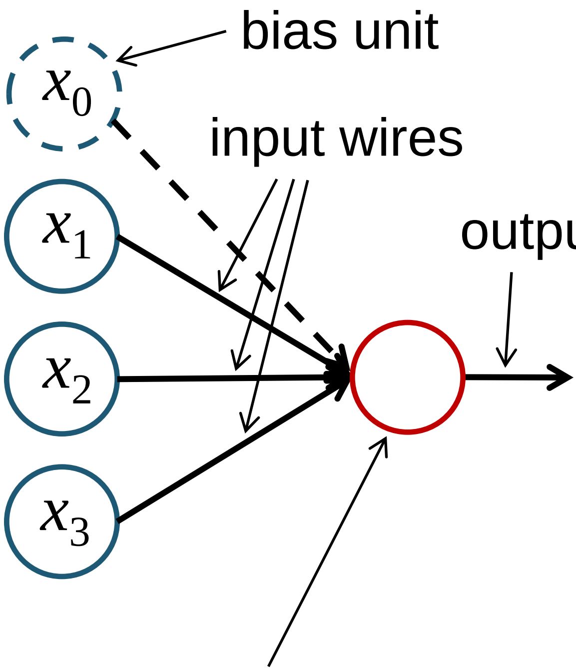
Lecture 3 – Neural Networks and Backpropagation

Slides were prepared using the course material of Stanford’s Machine Learning Course (CS229, Andrew Ng) and CNN Course (CS231n, Fei-Fei,Karpathy,Johnson)

Brain neurons



Neuron Model



$$f_w(x) = \frac{1}{1+e^{-w^T x}}$$

$$x = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix} \qquad w = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \end{bmatrix}$$

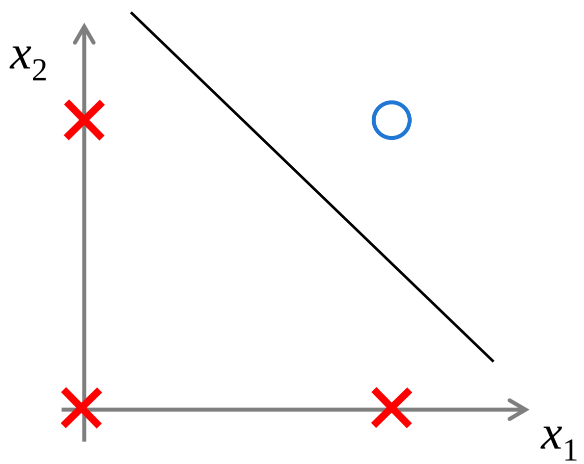
The parameters w are called ‘weights’.

This computation is determined by the *activation function*.
In this example, activation function is sigmoid function.

Example: AND

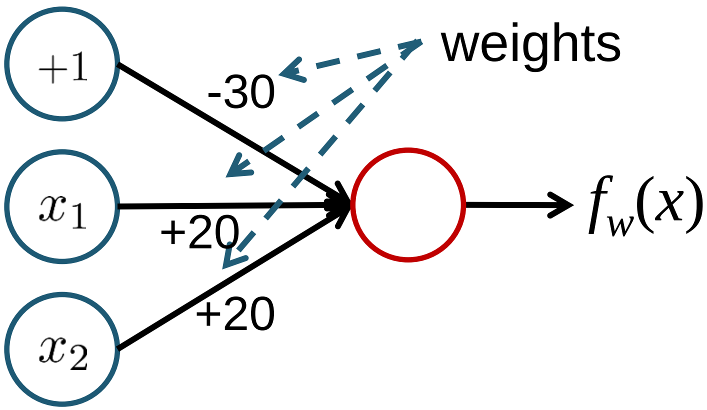
Let's build a linear classifier with a network with no hidden layers.

Logical AND operation



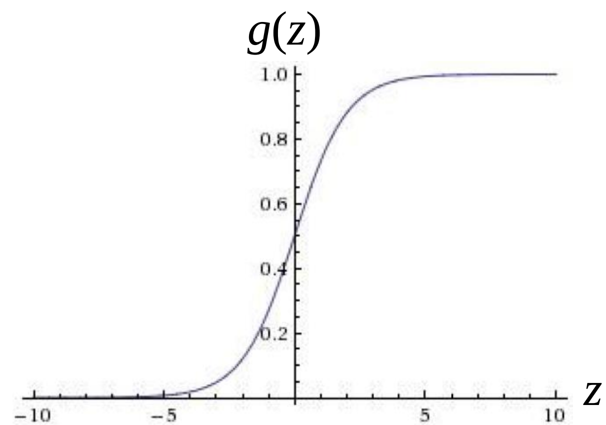
x_1, x_2 are binary (0 or 1).
y=1 if x_1 AND x_2
(So blue circle is the positive class)

Example: AND



$$f_w(x) = g(-30 + 20x_1 + 20x_2)$$

g: sigmoid function

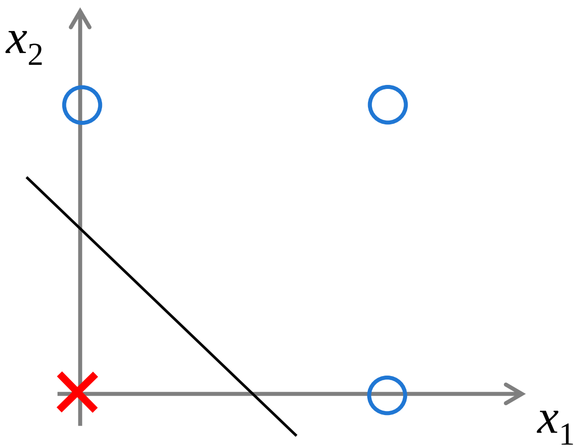


x_1	x_2	$f_w(x)$
0	0	$g(-30) \approx 0$
0	1	$g(-10) \approx 0$
1	0	$g(-10) \approx 0$
1	1	$g(10) \approx 1$

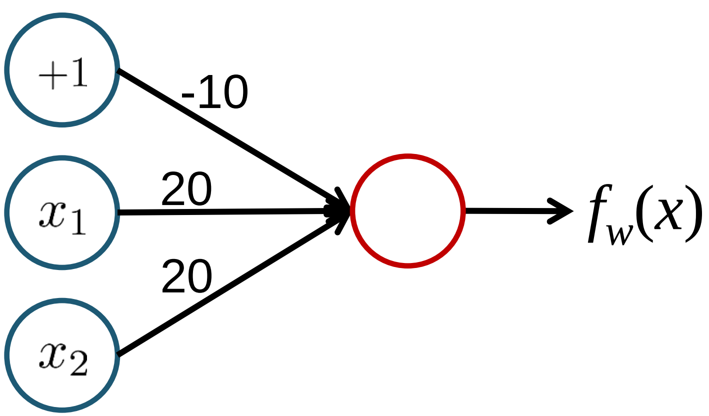
$$f_w(x) \approx x_1 \text{ AND } x_2$$

Example: OR

Another linear classifier for logical OR operation



x_1, x_2 are binary (0 or 1).
$y=1$ if x_1 OR x_2



x_1	x_2	$f_w(x)$
0	0	$g(-10) \approx 0$
0	1	$g(10) \approx 1$
1	0	$g(10) \approx 1$
1	1	$g(30) \approx 1$

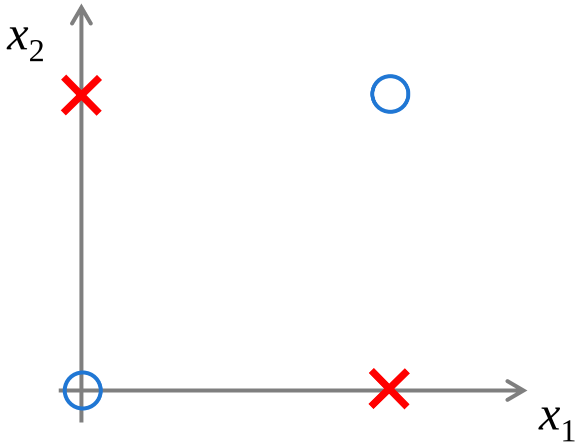
$$f_w(x) = g(-10+20x_1+20x_2)$$

$$f_w(x) \approx x_1 \text{ OR } x_2$$

Nonlinear classification example

Wait a minute! We said before that NNs are good for nonlinear classification.
This is done by adding more layers.

A non-linear classification example: XNOR

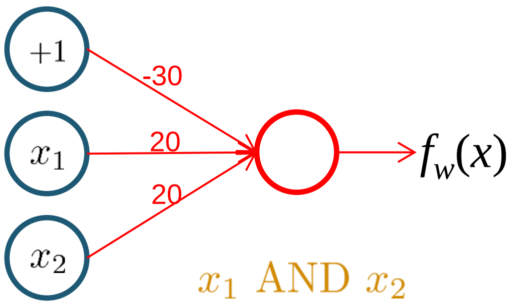


y=0 if x_1 XOR x_2
y=1 if NOT (x_1 XOR x_2)= x_1 XNOR x_2

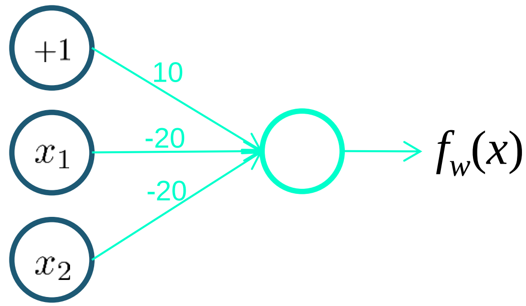
(blue circles are the positive class)

Hint: x_1 XNOR x_2 = (NOT x_1 AND NOT x_2) OR (x_1 AND x_2)

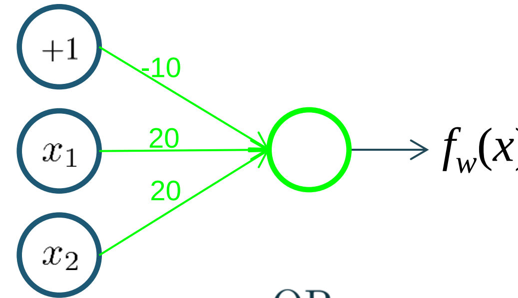
Example: XNOR



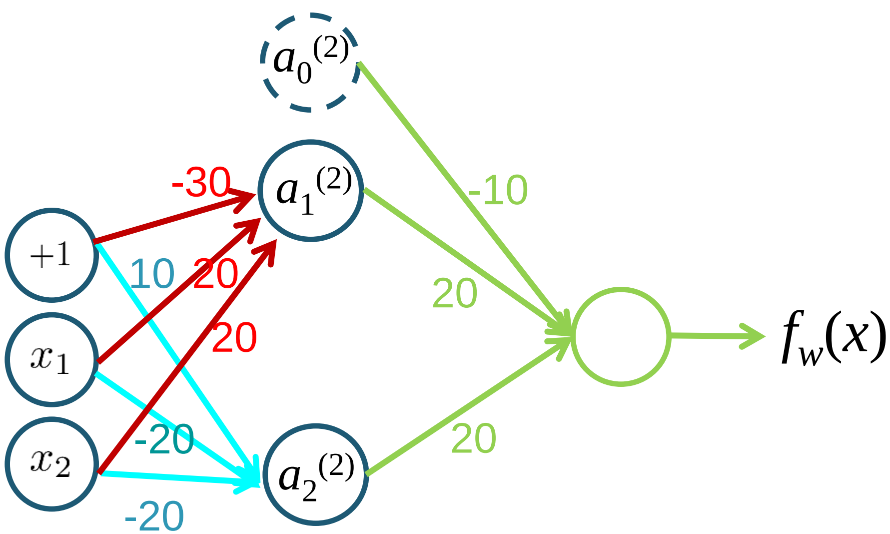
x_1 AND x_2



(NOT x_1) AND (NOT x_2)

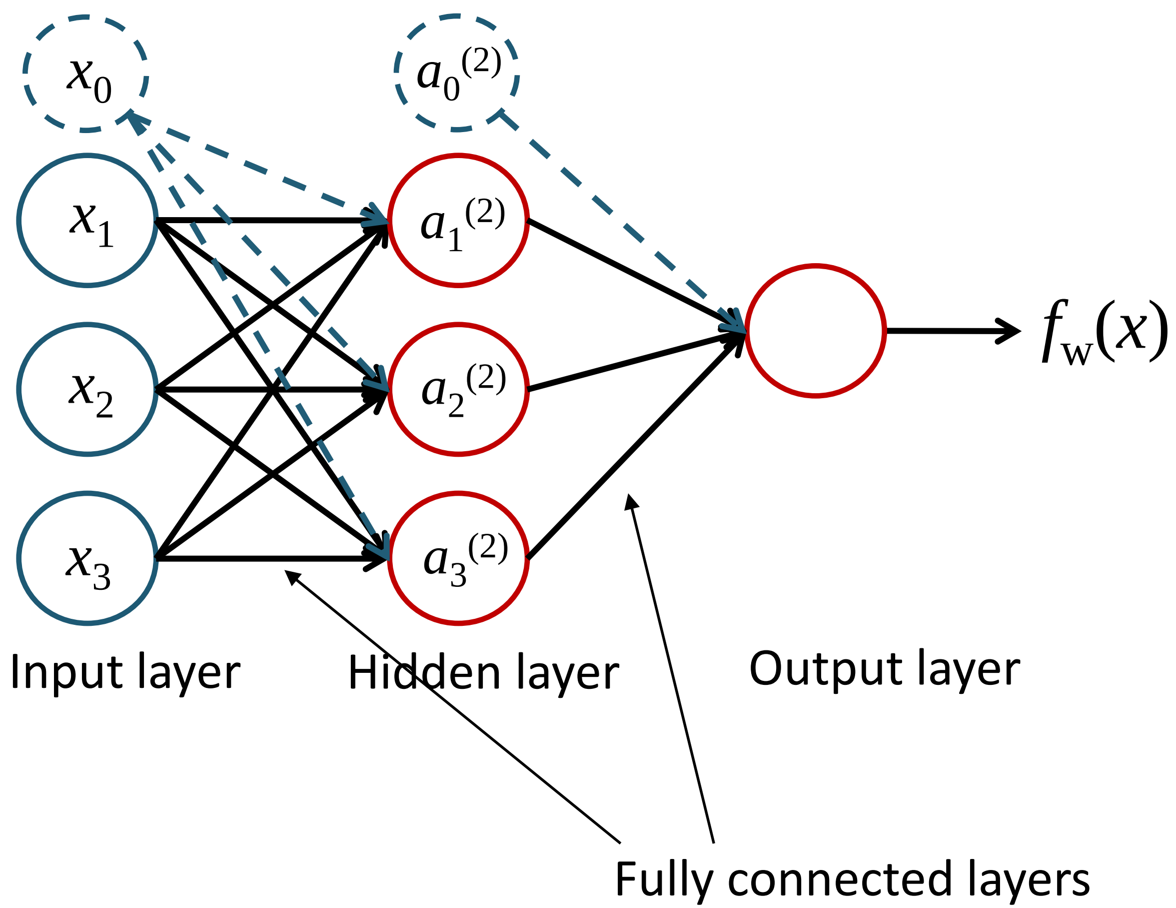


x_1 OR x_2



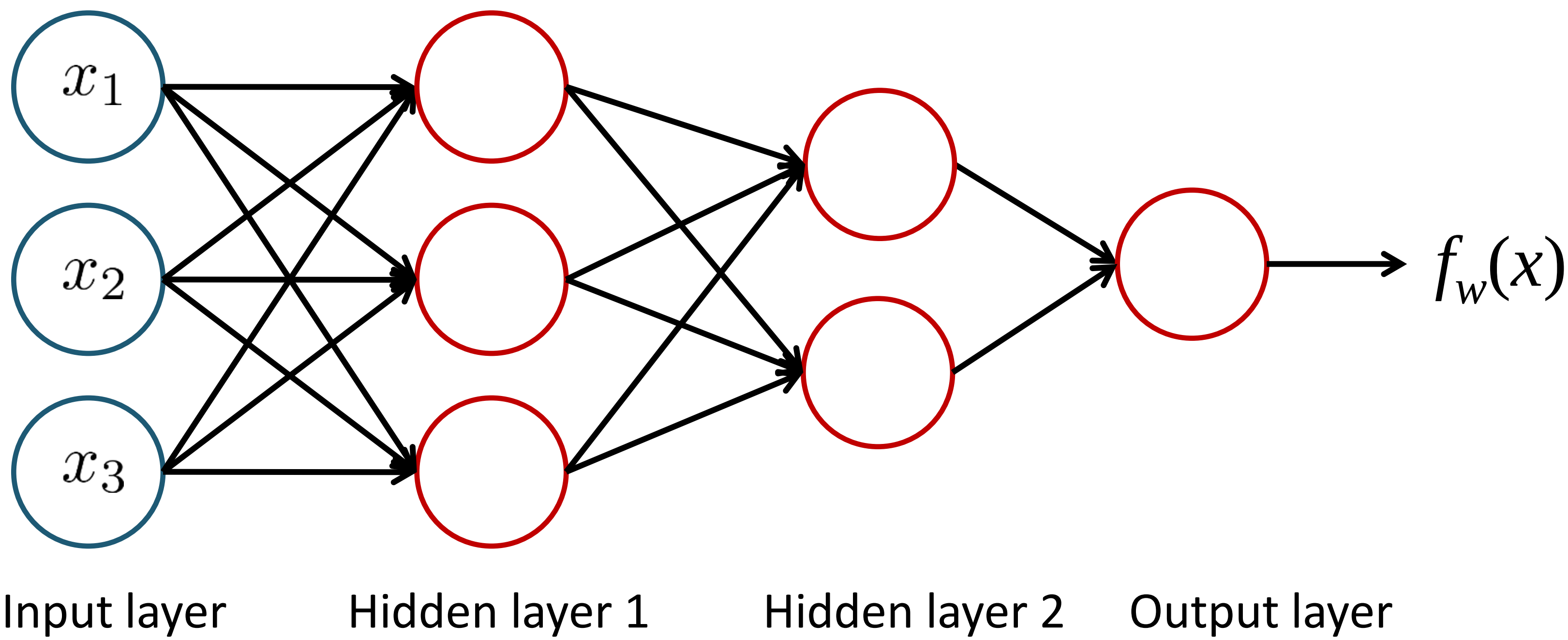
x_1	x_2	$a_1^{(2)}$	$a_2^{(2)}$	$f_w(x)$
0	0	0	1	1
0	1	0	0	0
1	0	0	0	0
1	1	1	0	1

2-layer Neural Network
(or 1-hidden-layer neural network)



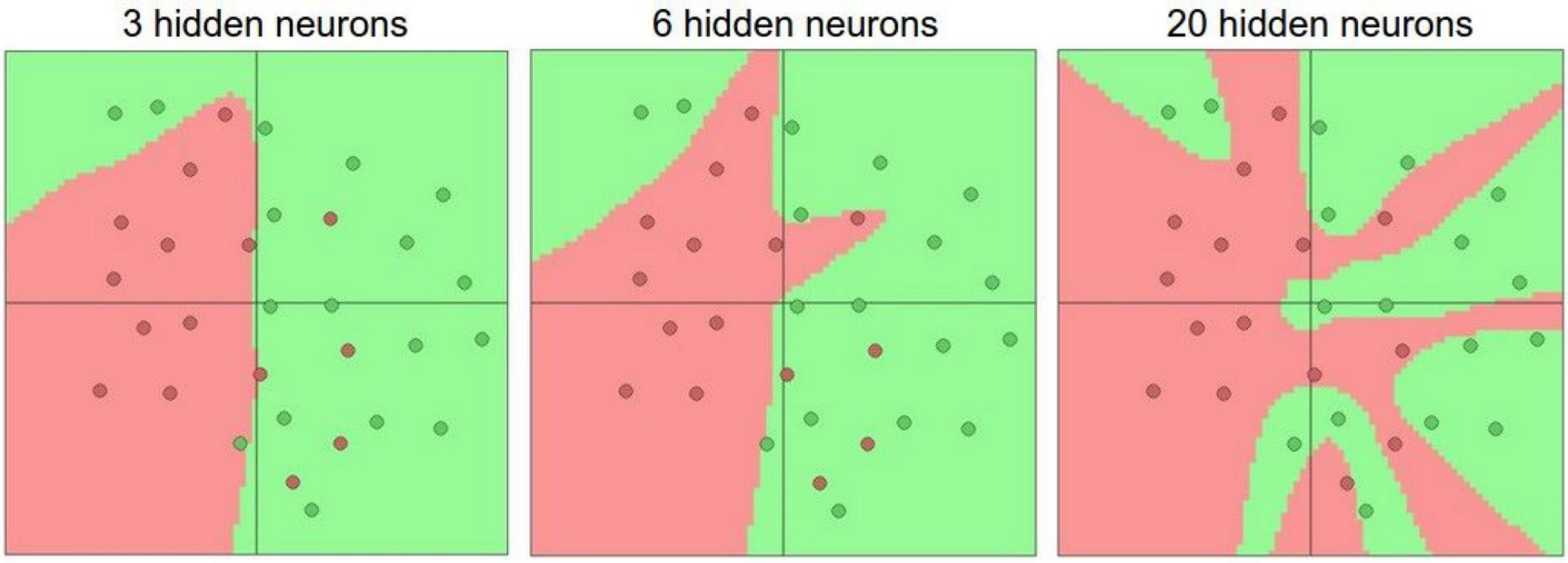
3-layer Neural Network
(or 2-hidden-layer neural network)

As we go further in layers, more complex functions are modeled.



Effect of the number of hidden neurons/layers

More neurons, more capacity to learn complex boundaries



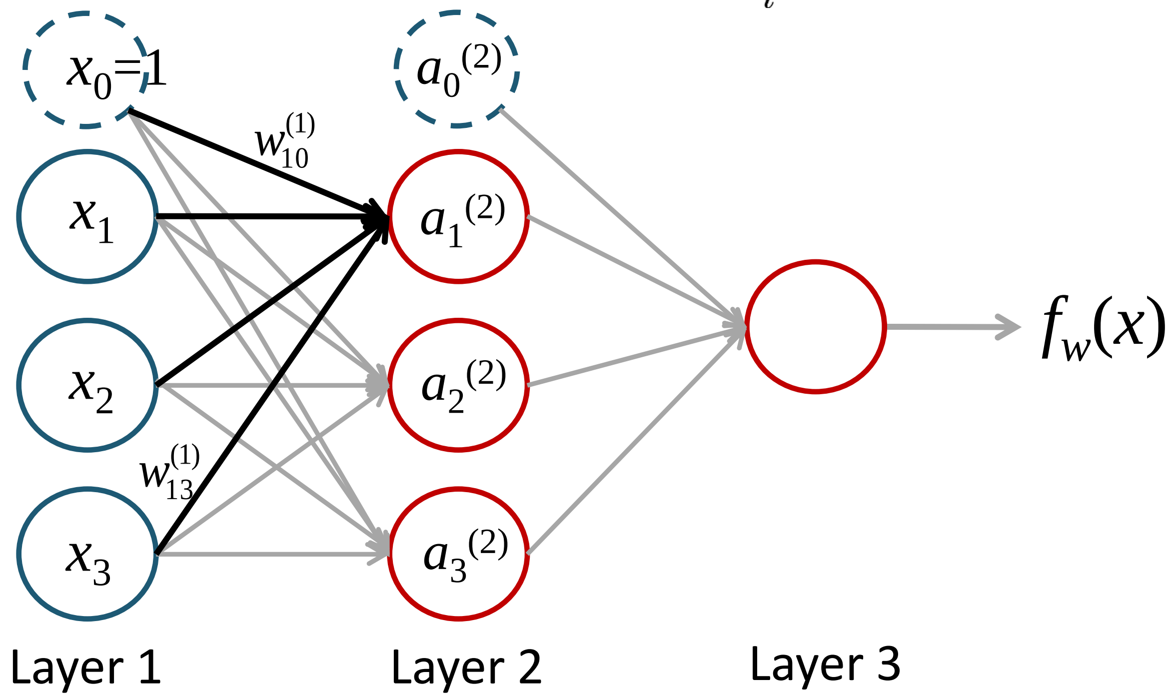
Forward Propagation

$$z_1^{(2)} = w_{10}^{(1)} + w_{11}^{(1)} x_1 + w_{12}^{(1)} x_2 + w_{13}^{(1)} x_3$$

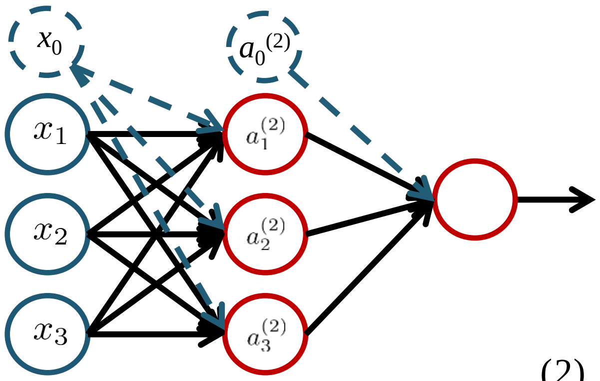
$$a_1^{(2)} = g(z_1^{(2)})$$

Remember g is the sigmoid function

$a_i^{(j)} =$ "activation" of unit i in layer j



Forward Propagation



$$a_1^{(2)} = g(\underbrace{w_{10}^{(1)}x_0 + w_{11}^{(1)}x_1 + w_{12}^{(1)}x_2 + w_{13}^{(1)}x_3}_{z_1^{(2)}})$$

$$a_2^{(2)} = g(\underbrace{w_{20}^{(1)}x_0 + w_{21}^{(1)}x_1 + w_{22}^{(1)}x_2 + w_{23}^{(1)}x_3}_{z_2^{(2)}})$$

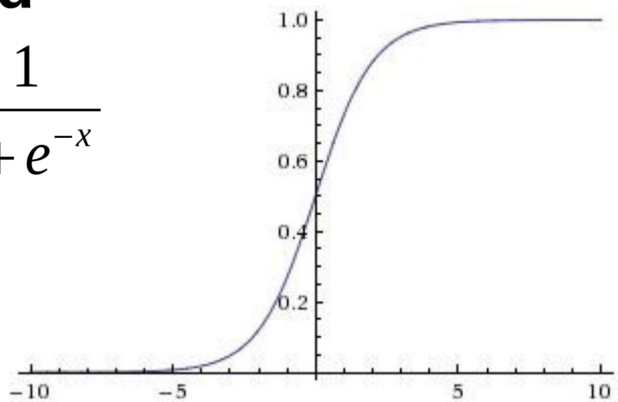
$$a_3^{(2)} = g(\underbrace{w_{30}^{(1)}x_0 + w_{31}^{(1)}x_1 + w_{32}^{(1)}x_2 + w_{33}^{(1)}x_3}_{z_3^{(2)}})$$

$$f_w(x) = a_1^{(3)} = g(\underbrace{w_{10}^{(2)}a_0^{(2)} + w_{11}^{(2)}a_1^{(2)} + w_{12}^{(2)}a_2^{(2)} + w_{13}^{(2)}a_3^{(2)}}_{z_1^{(3)}})$$

Activation functions

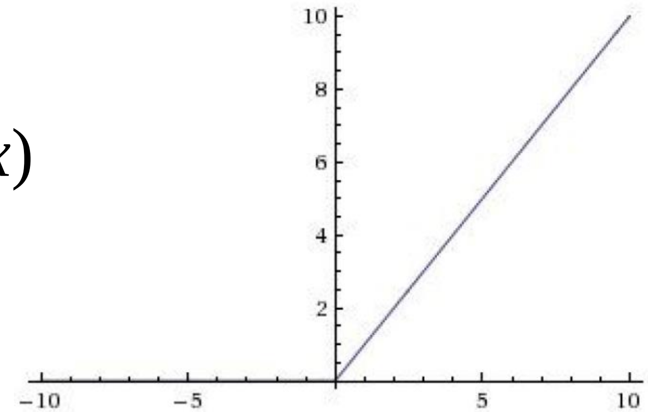
Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



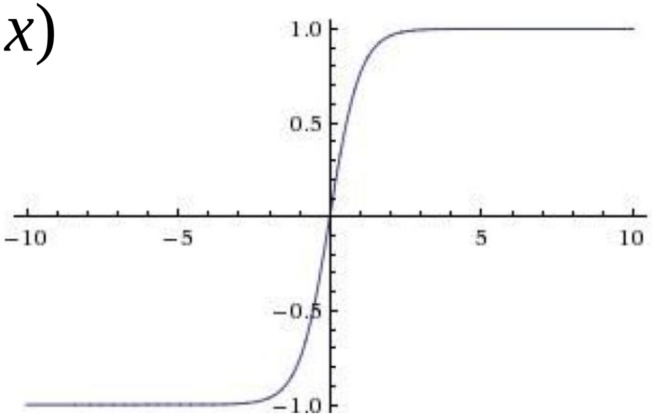
ReLU

$$\max(0, x)$$



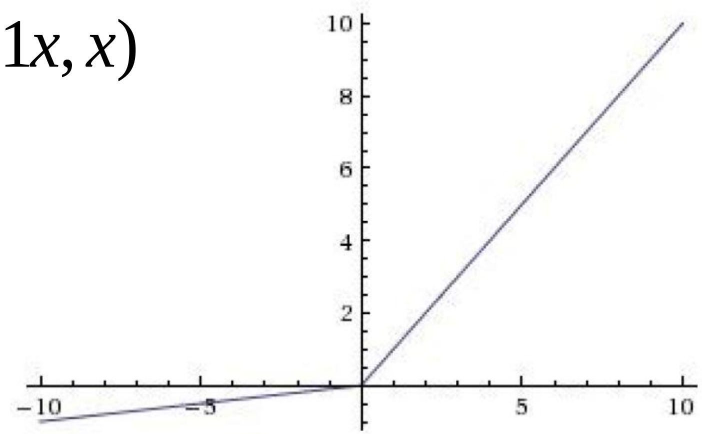
tanh

$$\tanh(x)$$



Leaky ReLU

$$\max(0.1x, x)$$



Neural Network Learning

We have learned about:

- Analogy with the human brain
- Layers of neural networks
- Forward propagation

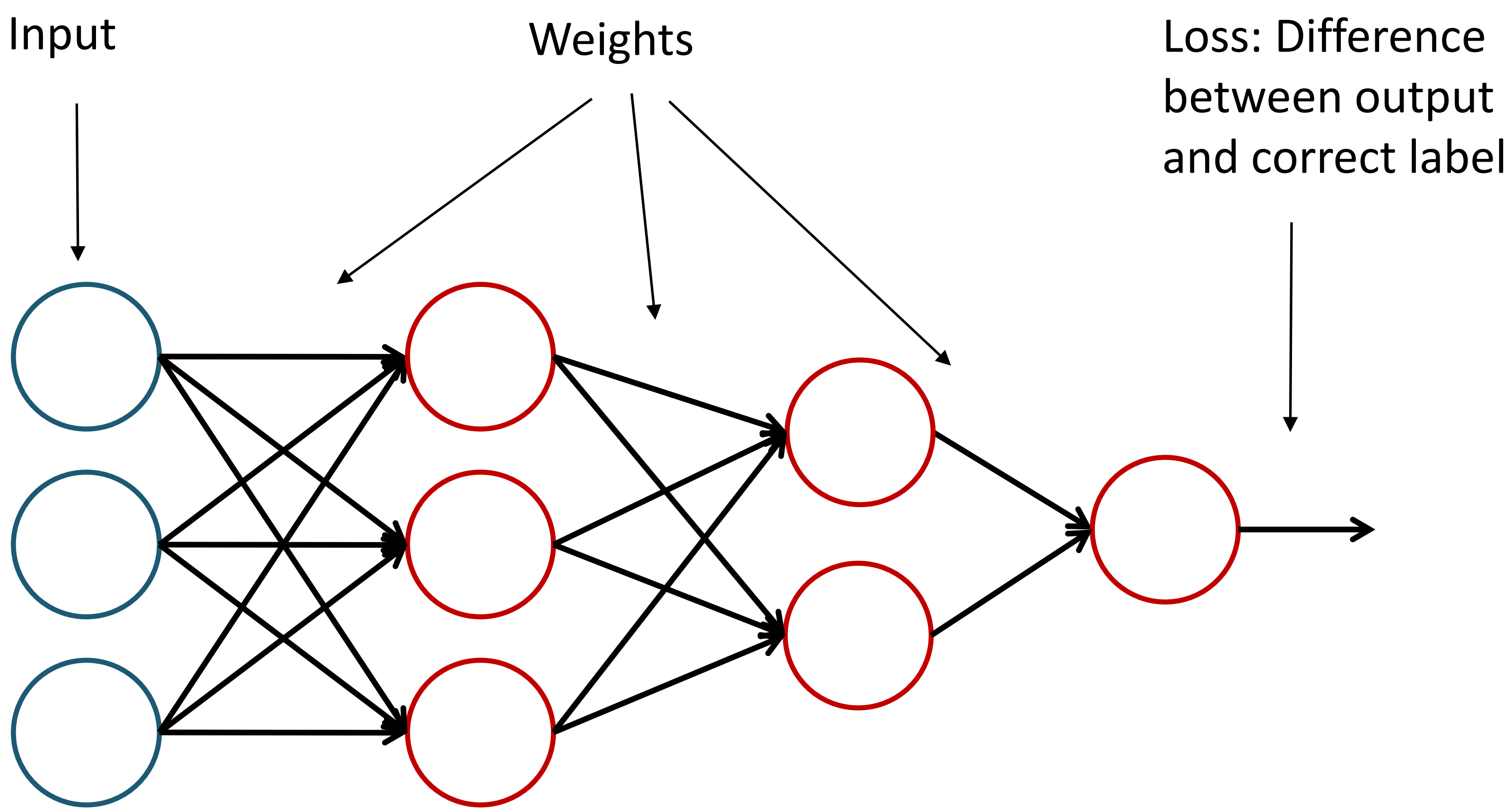
But how do neural networks ‘learn’ actually?

Learning corresponds to determining the ‘weights’ of units.

These weights are optimized using a cost function.

We define ‘loss’ here.

Loss

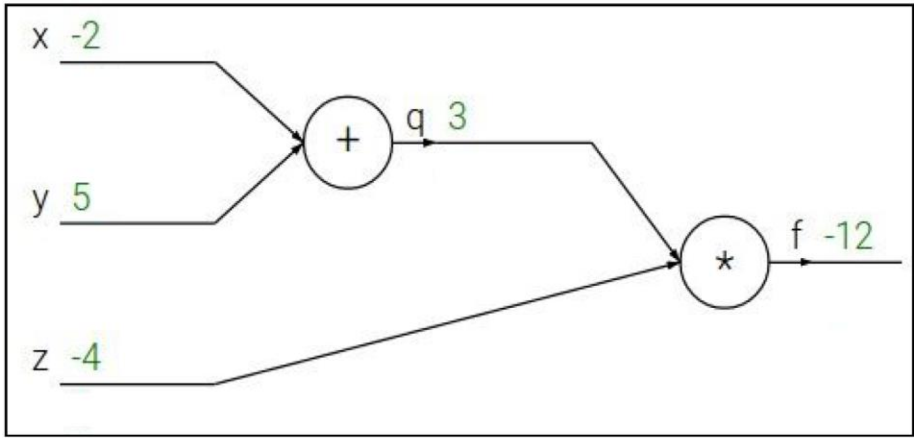


To update the weights, we ‘backpropagate’ the loss to previous layers.

Derivatives for backpropagation

$$f(x, y, z) = (x + y)z$$

E.g. x=-2, y=5, z=-4



$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

Want: $\dfrac{\partial f}{\partial x}, \dfrac{\partial f}{\partial y}, \dfrac{\partial f}{\partial z}$

Derivatives for backpropagation

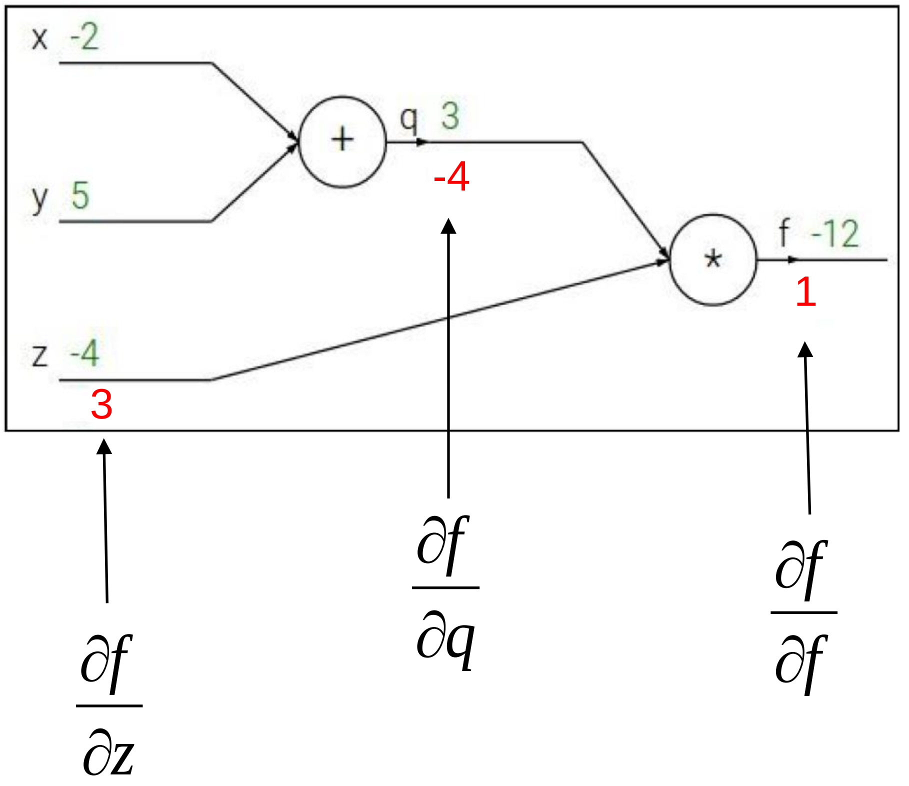
$f(x, y, z) = (x + y)z$

E.g. x=-2, y=5, z=-4

$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



Derivatives for backpropagation

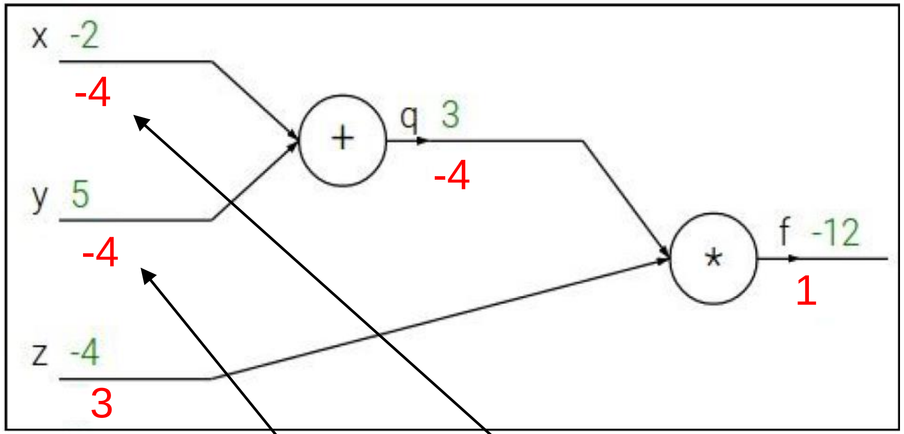
$$f(x,y,z) = (x+y)z$$

E.g. x=-2, y=5, z=-4

$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

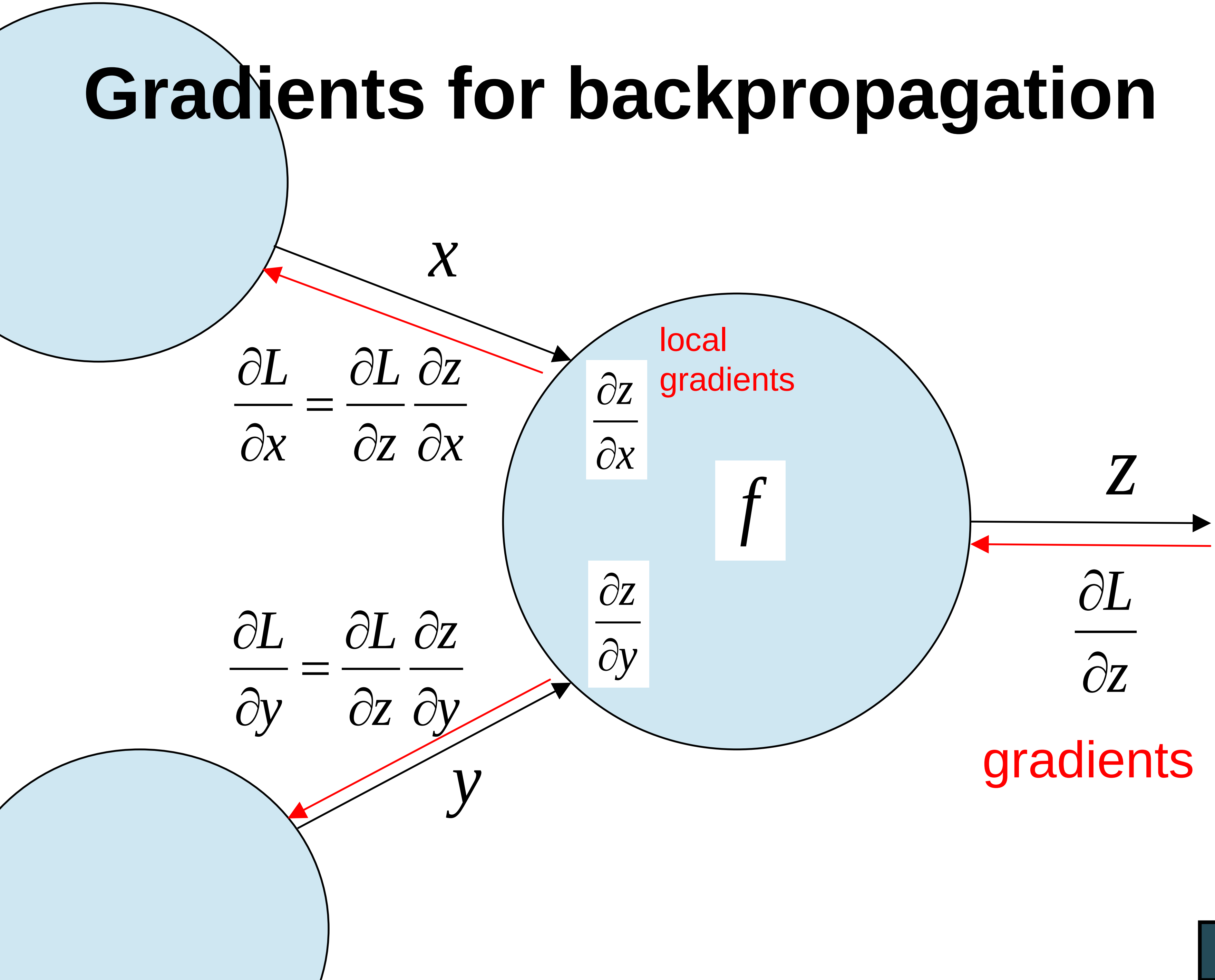
Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



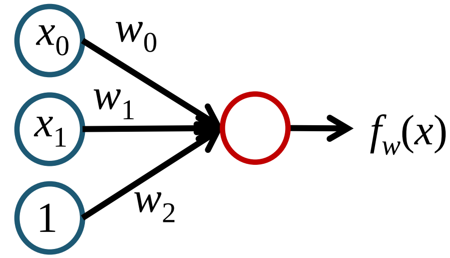
$$\frac{\partial f}{\partial y} \qquad \frac{\partial f}{\partial x}$$

Chain rule: $\frac{\partial f}{\partial y} = \frac{\partial f}{\partial q}\frac{\partial q}{\partial y} \qquad \frac{\partial f}{\partial x} = \frac{\partial f}{\partial q}\frac{\partial q}{\partial x}$

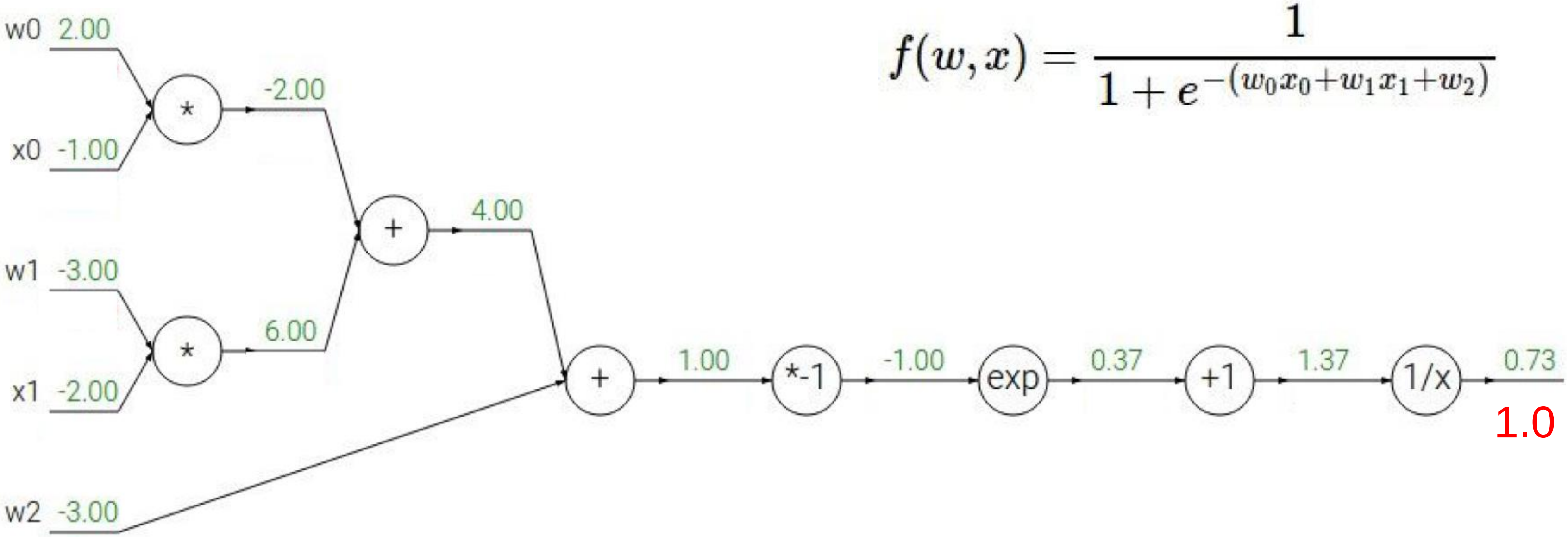
Gradients for backpropagation



Another example



$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

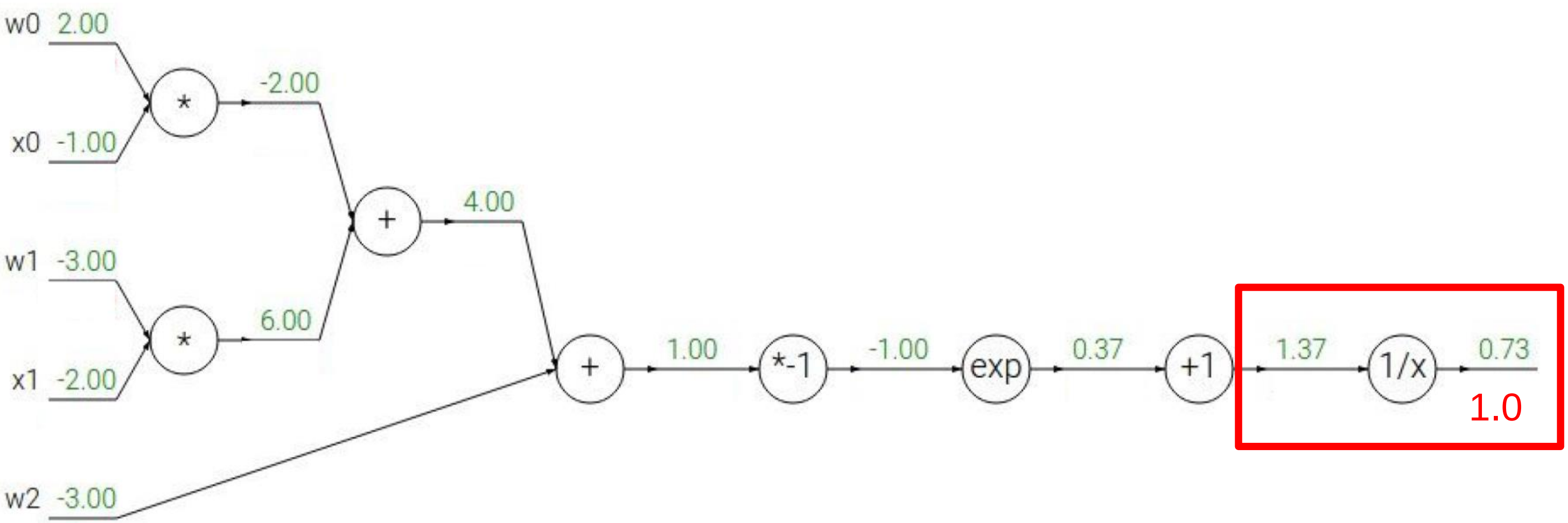


$$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x \qquad\qquad f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$$

$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a \qquad\qquad f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$$

Another example

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$



$$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$$

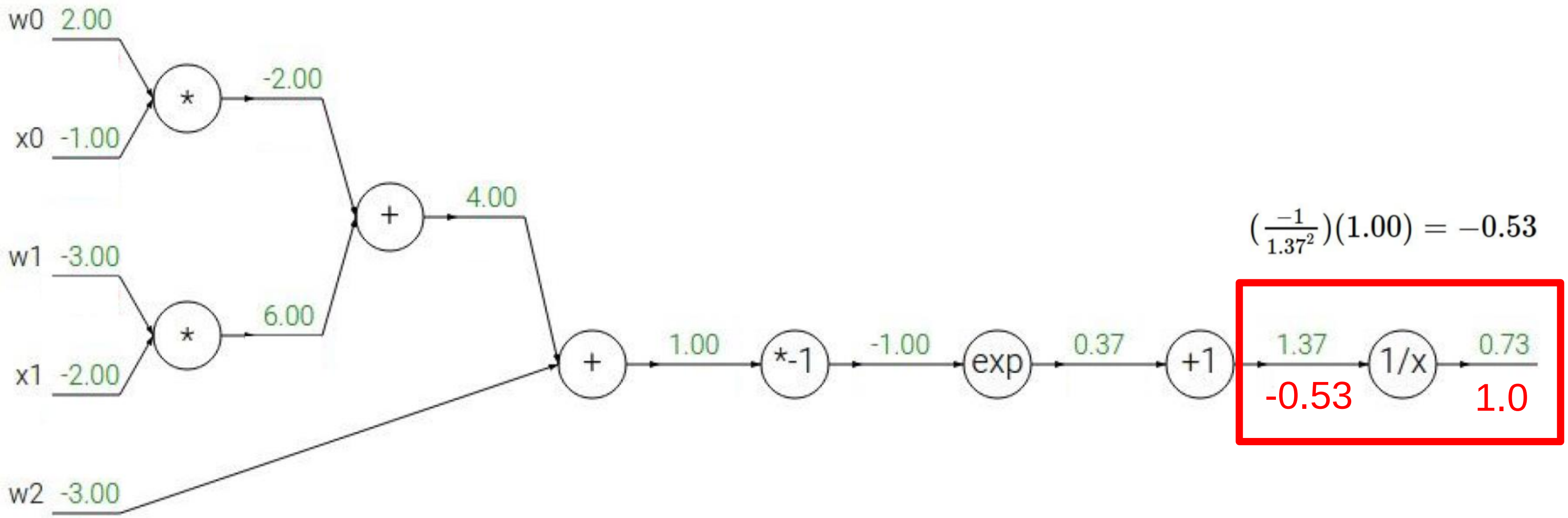
$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$$

$$\boxed{f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2}$$

$$f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$$

Another example

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$



$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$

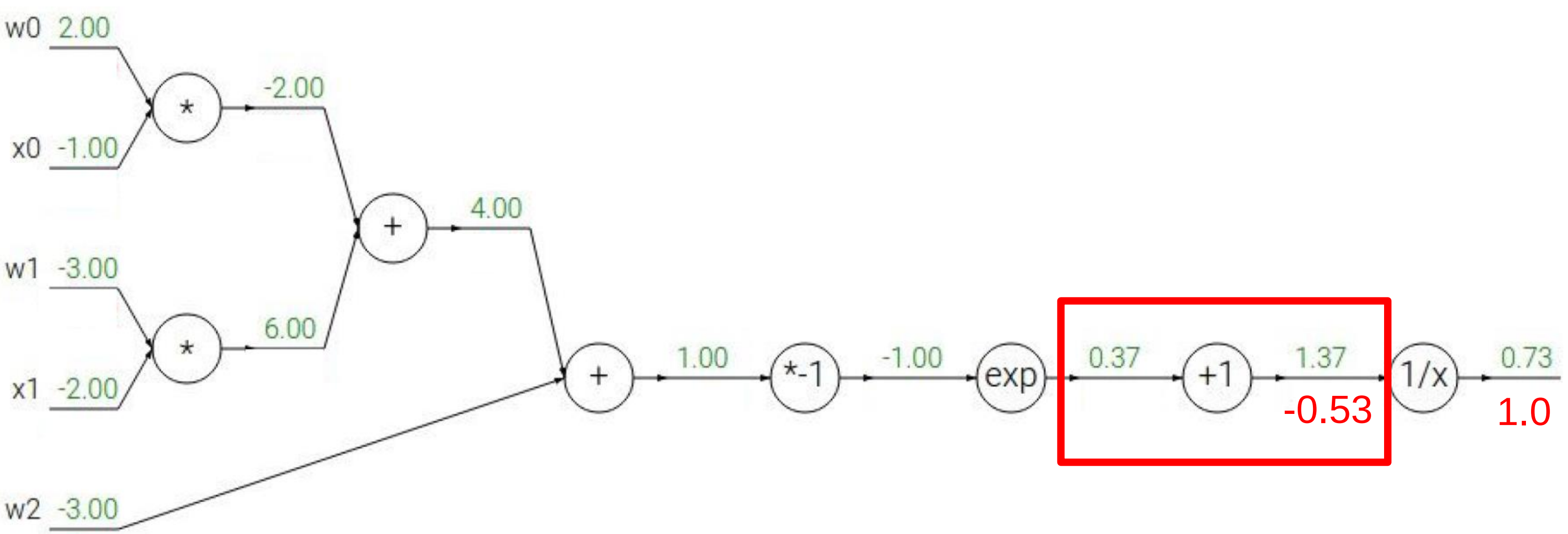
$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$

$\boxed{f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2}$

$f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$

Another example

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$



$$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$$

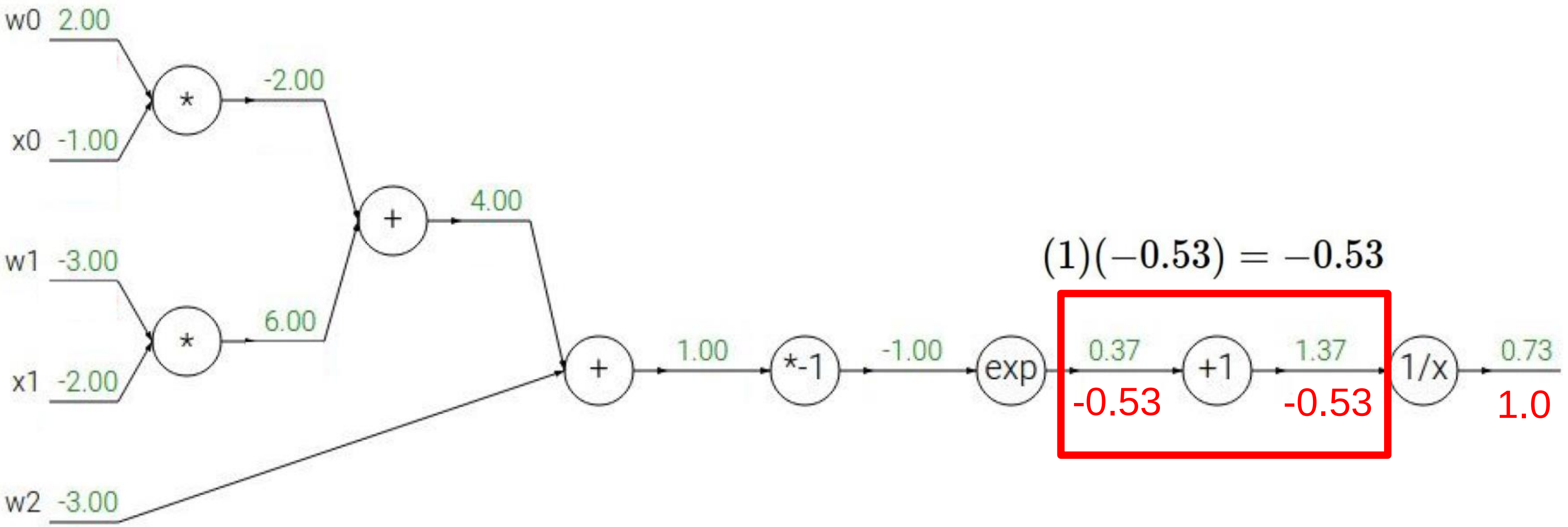
$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$$

$$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$$

$$\boxed{f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1}$$

Another example

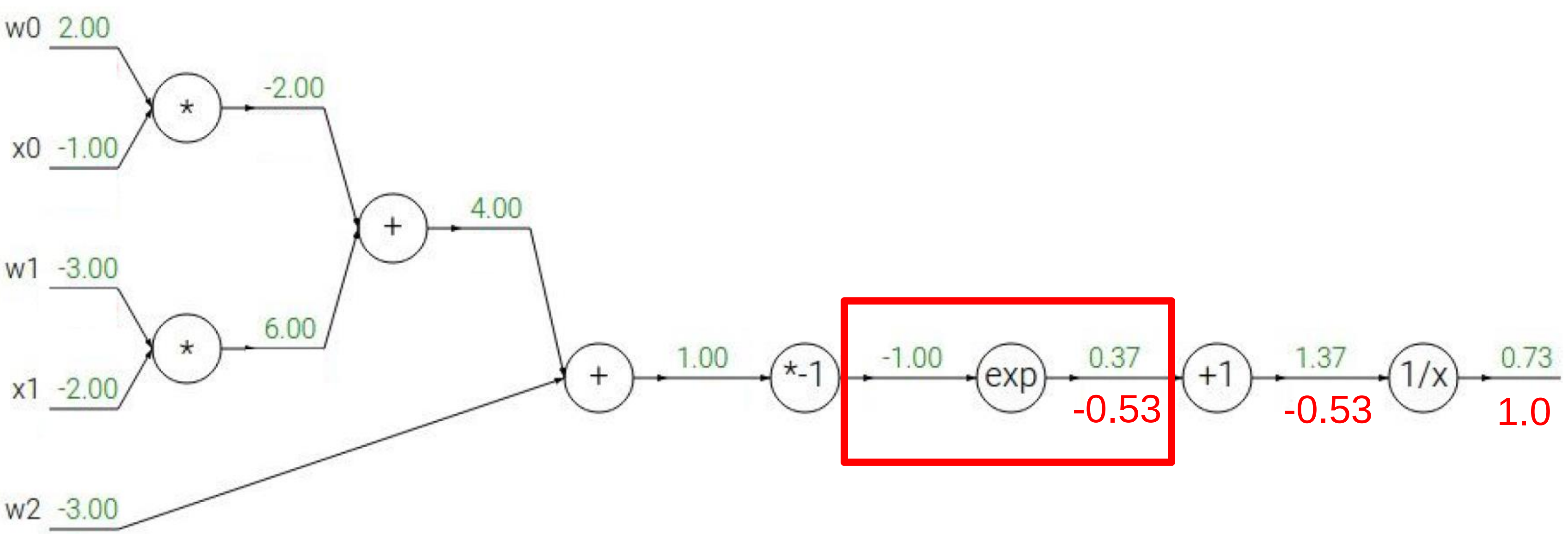
$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$



$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$		$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$	
$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$		$\boxed{f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1}$	

Another example

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$



$$\boxed{f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x}$$

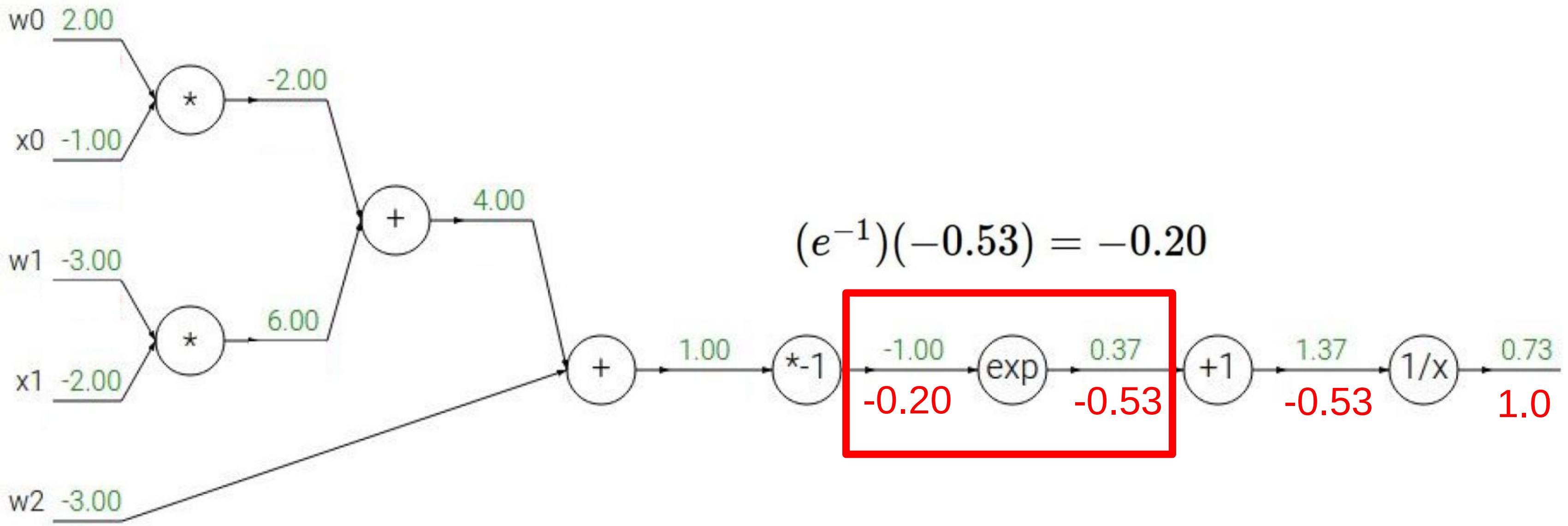
$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$$

$$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$$

$$f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$$

Another example

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$



$$\boxed{f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x}$$

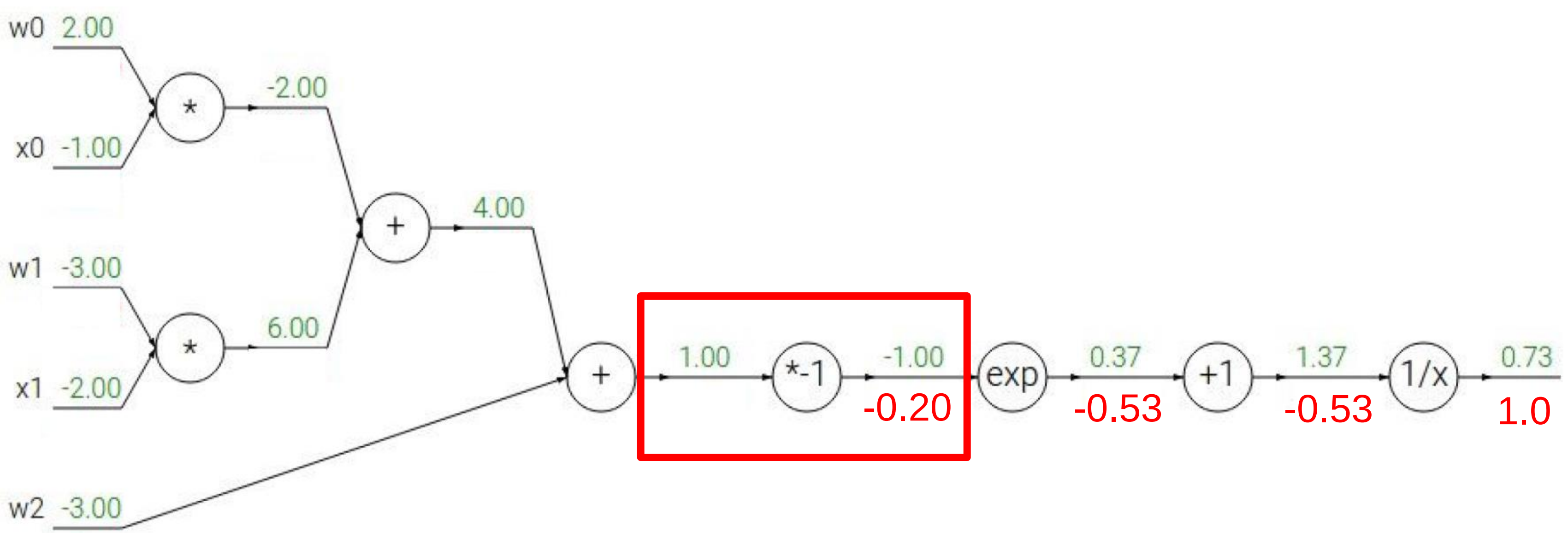
$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$$

$$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$$

$$f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$$

Another example

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$



$$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$$

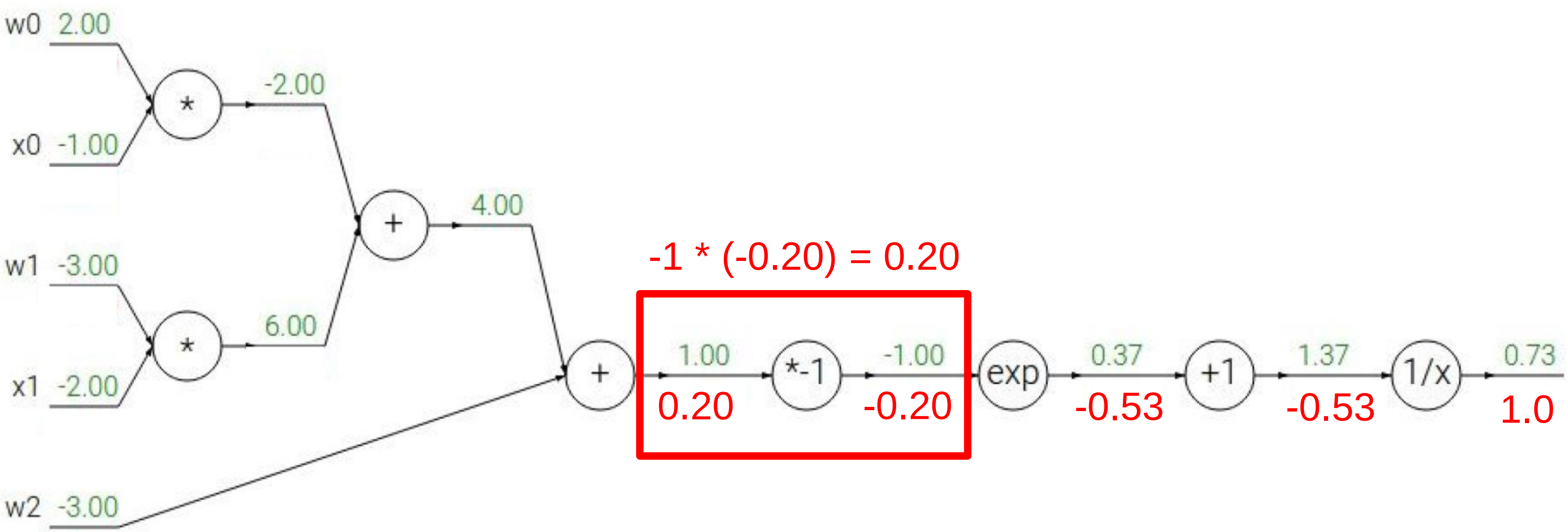
$$\boxed{f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a}$$

$$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$$

$$f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$$

Another example

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$



$$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$$

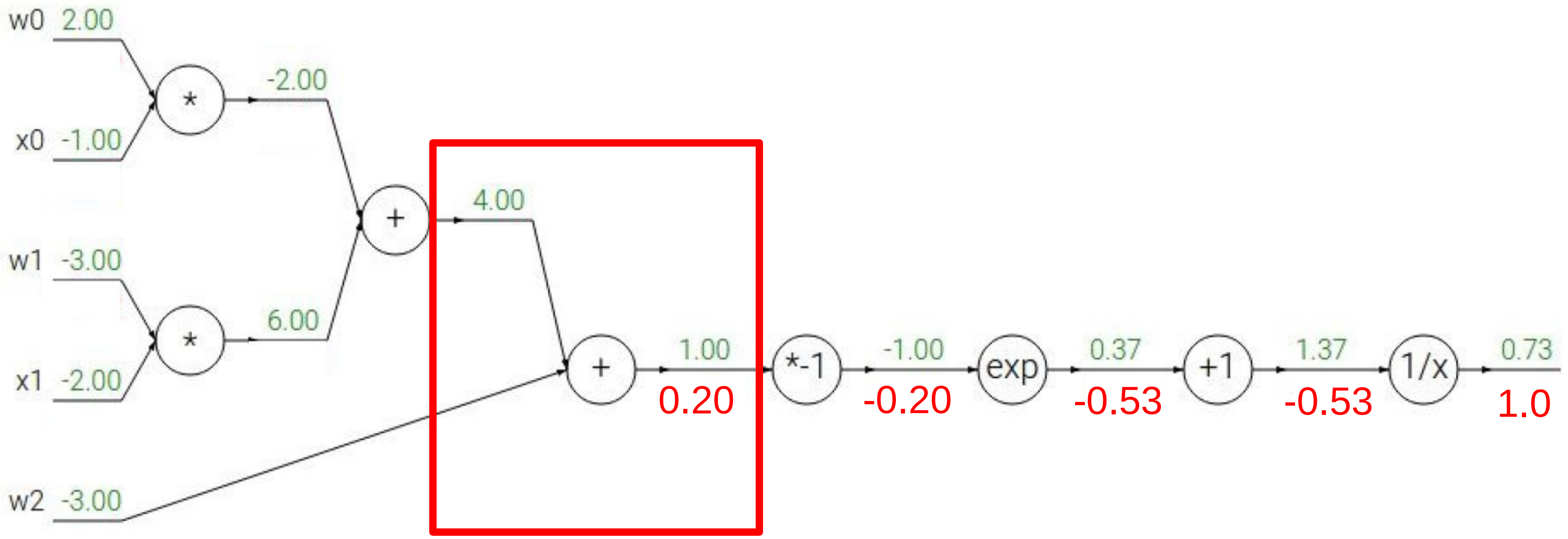
$$\boxed{f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a}$$

$$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$$

$$f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$$

Another example

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$



$$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$$

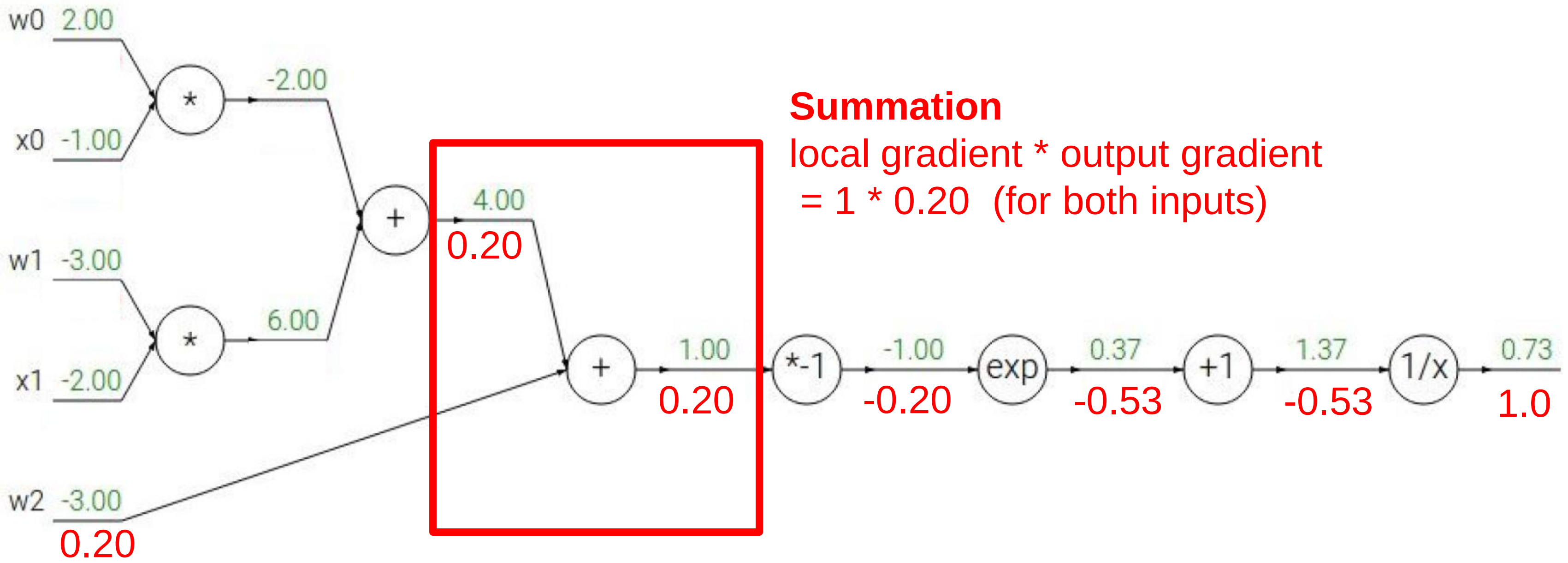
$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$$

$$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$$

$$f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$$

Another example

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

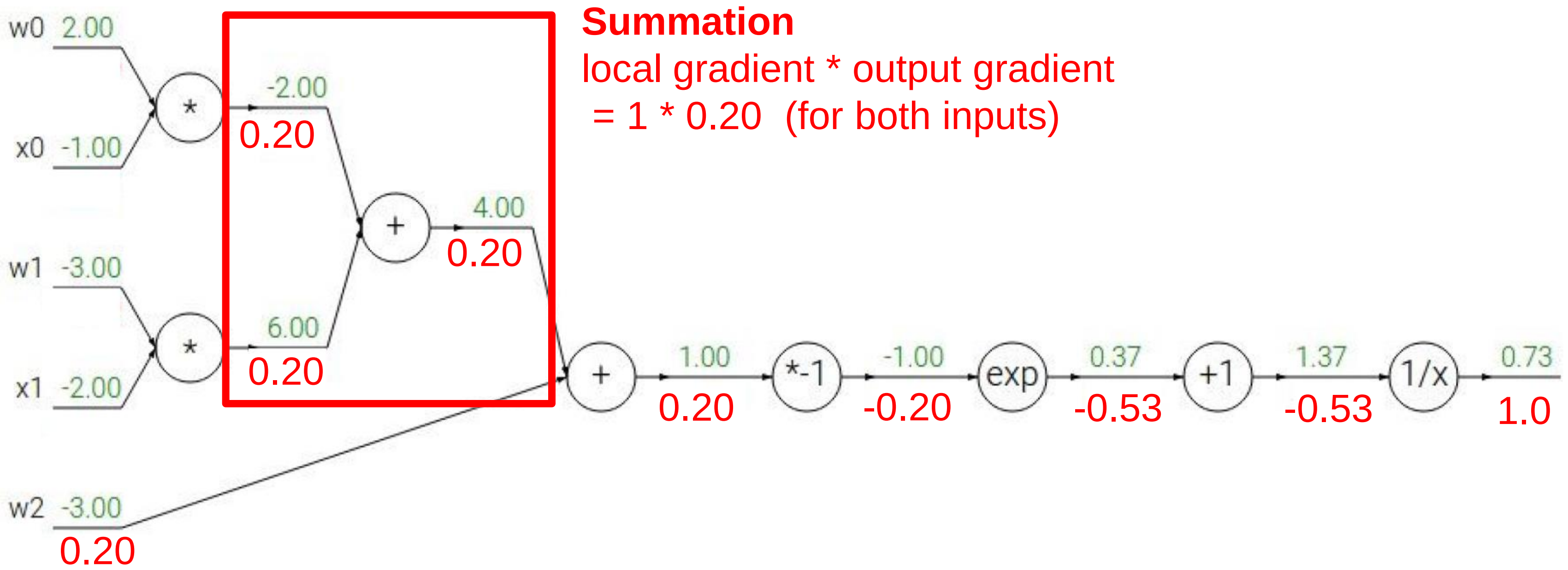


$$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x \qquad\qquad f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$$

$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a \qquad\qquad f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$$

Another example

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

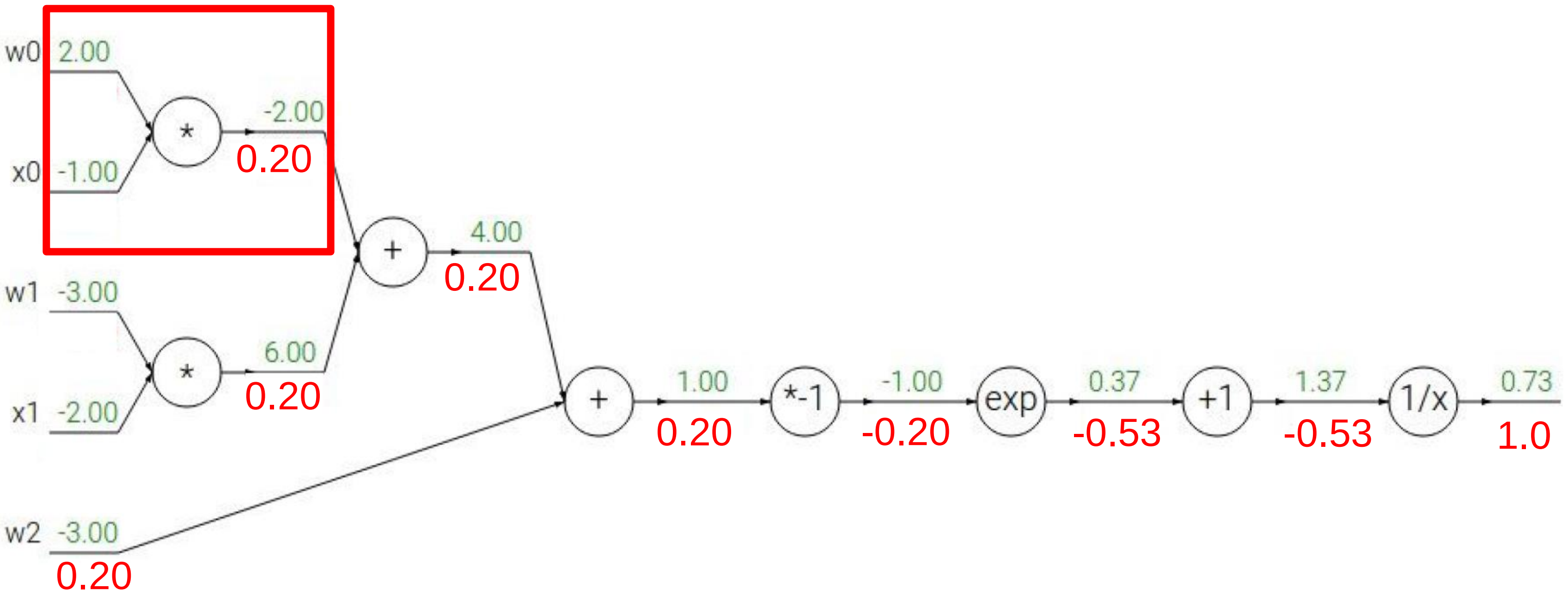


$$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x \qquad\qquad f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$$

$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a \qquad\qquad f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$$

Another example

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$



$$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$$

$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$$

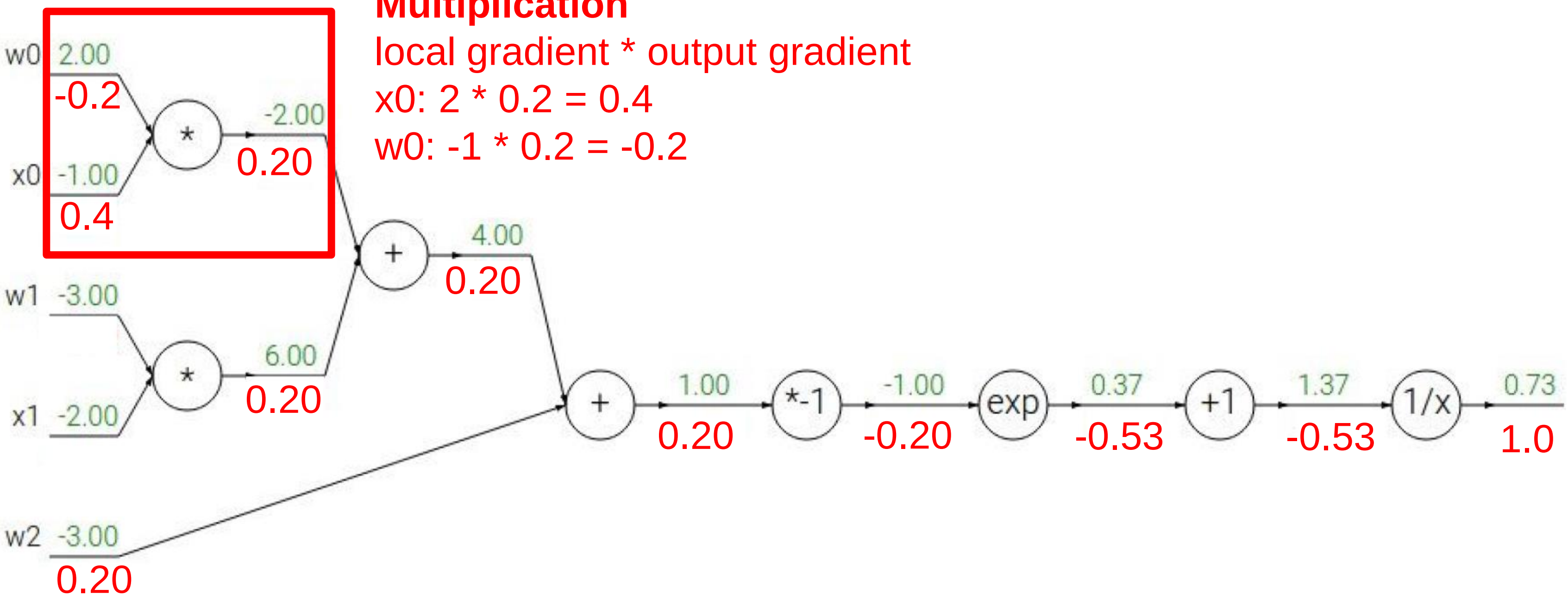
$$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$$

$$f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$$

Another example

$$f(w,x) = \frac{1}{1 + e^{-(w_0 x_0 + w_1 x_1 + w_2)}}$$

Multiplication
local gradient * output gradient
x0: 2 * 0.2 = 0.4
w0: -1 * 0.2 = -0.2



$$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$$

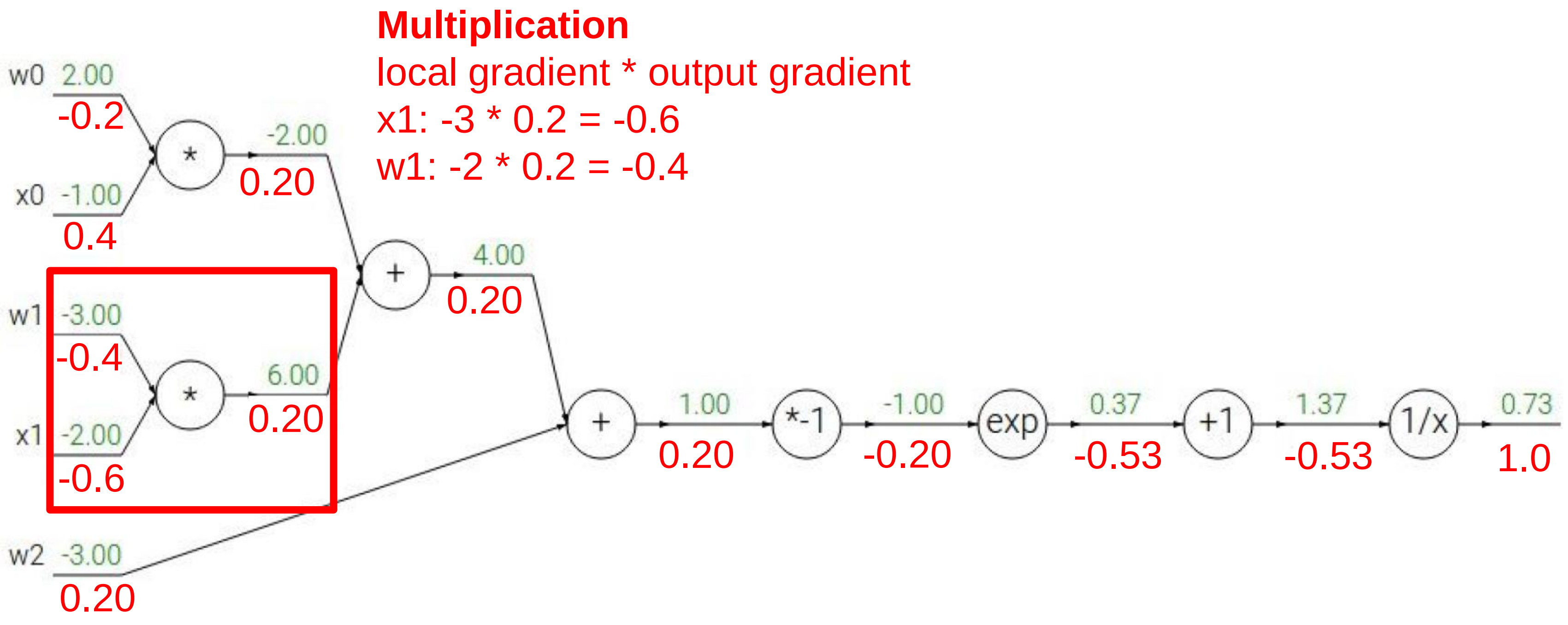
$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$$

$$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$$

$$f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$$

Another example

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$



$$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x \qquad\qquad f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$$

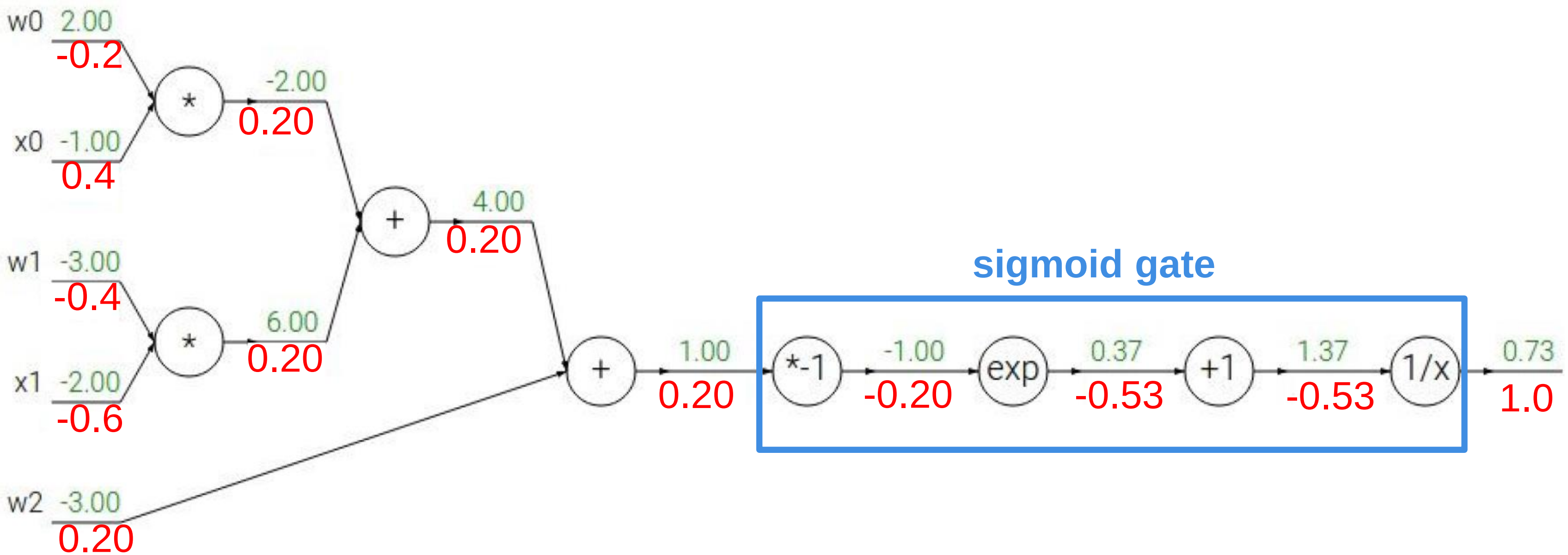
$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a \qquad\qquad f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$$

Another example

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

$$\boxed{\sigma(x) = \frac{1}{1+e^{-x}}}$$ sigmoid function

$$\frac{d\sigma(x)}{dx} = \frac{e^{-x}}{(1+e^{-x})^2} = \left(\frac{1+e^{-x}-1}{1+e^{-x}}\right)\left(\frac{1}{1+e^{-x}}\right) = (1-\sigma(x))\,\sigma(x)$$



Another example

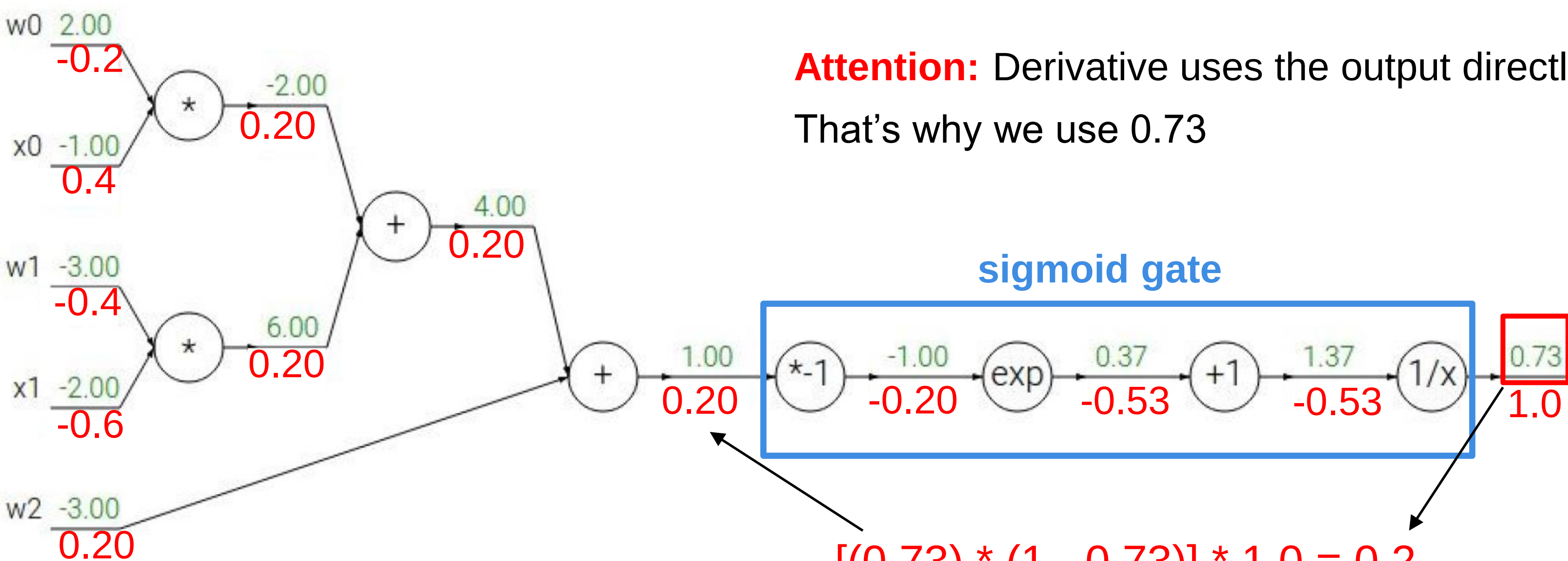
$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

$$\boxed{\sigma(x) = \frac{1}{1+e^{-x}}}$$ sigmoid function

$$\frac{d\sigma(x)}{dx} = \frac{e^{-x}}{(1+e^{-x})^2} = \left(\frac{1+e^{-x}-1}{1+e^{-x}}\right)\left(\frac{1}{1+e^{-x}}\right) = (1-\sigma(x))\,\sigma(x)$$

Attention: Derivative uses the output directly

That's why we use 0.73



$$\underbrace{[(0.73) * (1 - 0.73)]}_{\text{local gradient}} * \underbrace{1.0}_{\text{output gradient}} = 0.2$$

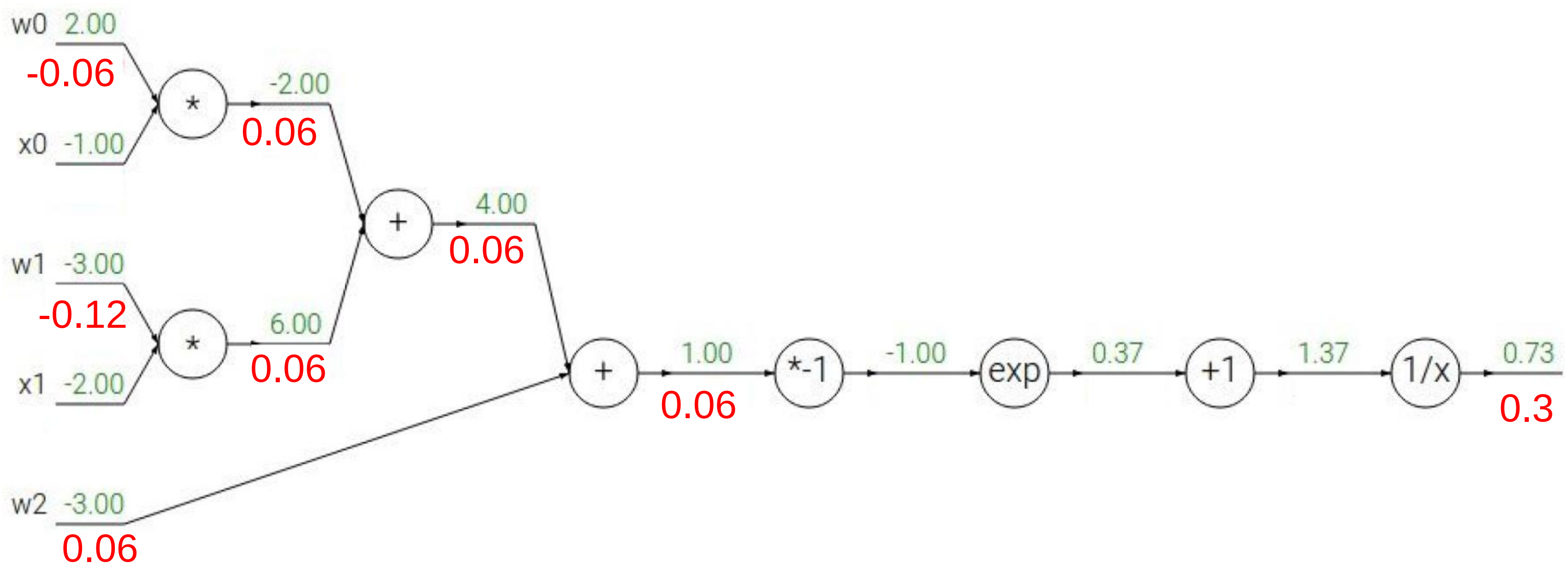
Example of weight update

Gradients, when applied to weights, decreases the loss at each iteration.

Let’s say that correct value for a training sample is 0.43 whereas the neuron gives 0.73 at the beginning (loss is 0.3).

Update in the 1st round (we subtract the gradients):

weights: w_0=2.06 w_1=-2.88 w_2=-3.06 function value: 0.66



Example of weight update

Gradients, when applied to weights, decreases the loss at each iteration.

Let's say that correct value for a training sample is 0.43 whereas the neuron gives 0.73 at the beginning (loss is 0.3).

Update in the,

1st round, weights: w_0=2.06 w_1=-2.88 w_2=-3.06 function value: 0.66

2nd round, weights: w_0=2.11 w_1=-2.78 w_2=-3.11 function value: 0.58

3rd round, weights: w_0=2.14 w_1=-2.70 w_2=-3.15 function value: 0.53

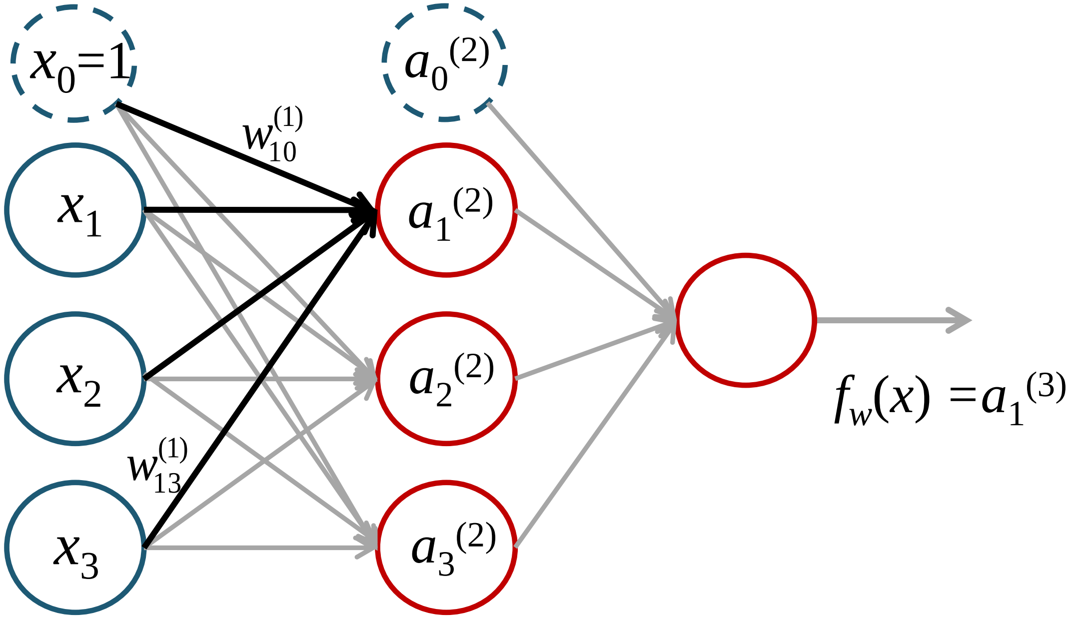
..

..

..

19th round, weights: w_0=2.21 w_1=-2.57 w_2=-3.21 function value: **0.43**

Forward Propagation Refresher

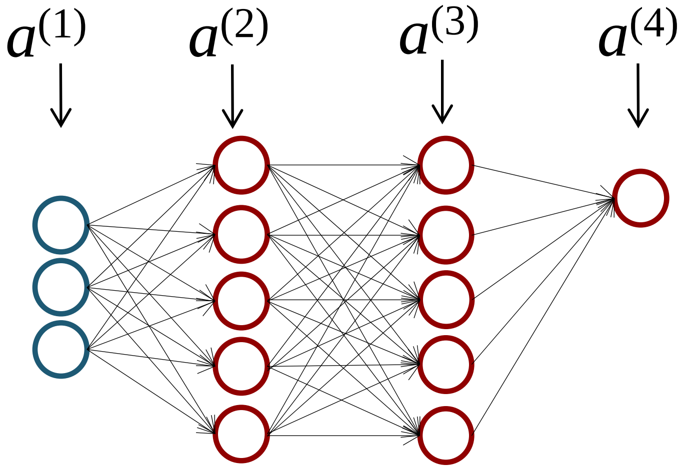


$$z_1^{(2)} = w_{10}^{(1)} + w_{11}^{(1)} x_1 + w_{12}^{(1)} x_2 + w_{13}^{(1)} x_3$$

$$a_1^{(2)} = g(z_1^{(2)})$$

Backpropagation: multiple layers

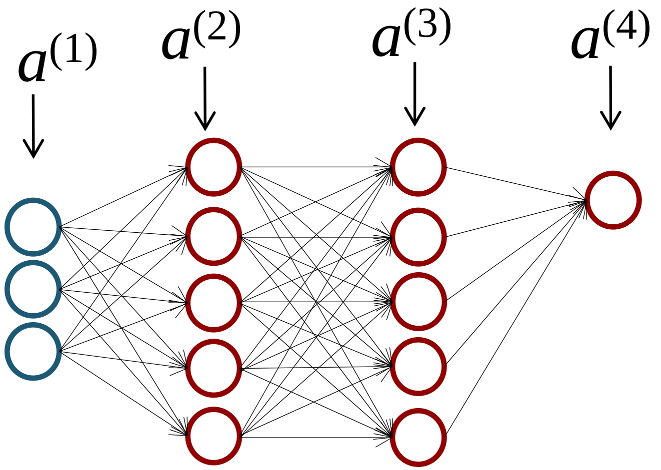
Need to compute $\dfrac{\partial E}{\partial w_{ij}^{(l)}}$ for weight update



Backpropagation: multiple layers

$\delta_j^{(l)} = \dfrac{\partial E}{\partial z_j^{(l)}}$, aka node delta, is the error of unit j in layer l .

For output unit (layer 4 here) error is:



$$\delta_1^{(4)} = \underbrace{(a_1^{(4)} - y)}_{\substack{\textit{loss,}\\ \textit{output gradient}}} \cdot \underbrace{\frac{\partial a_1^{(4)}}{\partial z_1^{(4)}}}_{\substack{\textit{local}\\ \textit{gradient}}}$$

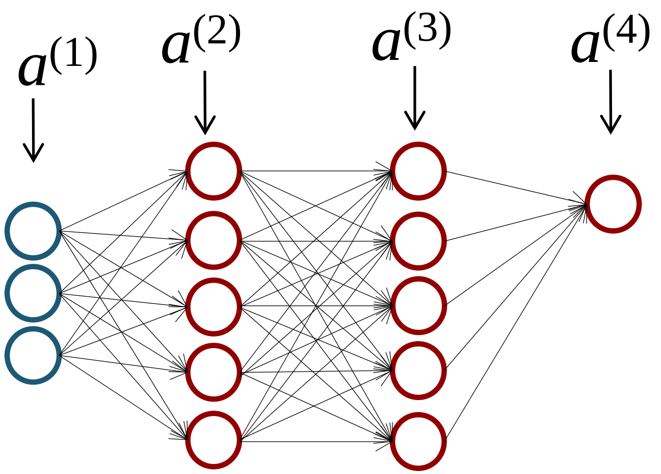
$$\delta_1^{(4)} = (a_1^{(4)} - y) \cdot \underbrace{a_1^{(4)} \cdot (1 - a_1^{(4)})}_{\textit{local gradient}}$$

Note: We assumed that we used sigmoid in the output unit $\delta_1^{(4)}$

Backpropagation: multiple layers

$\delta_j^{(l)} = \frac{\partial E}{\partial z_j^{(l)}}$, aka node delta, is the error of unit j in layer l .

For output unit (layer 4 here) error is:



$$\delta_1^{(4)} = (a_1^{(4)} - y) \cdot \frac{\partial a_1^{(4)}}{\partial z_1^{(4)}}$$

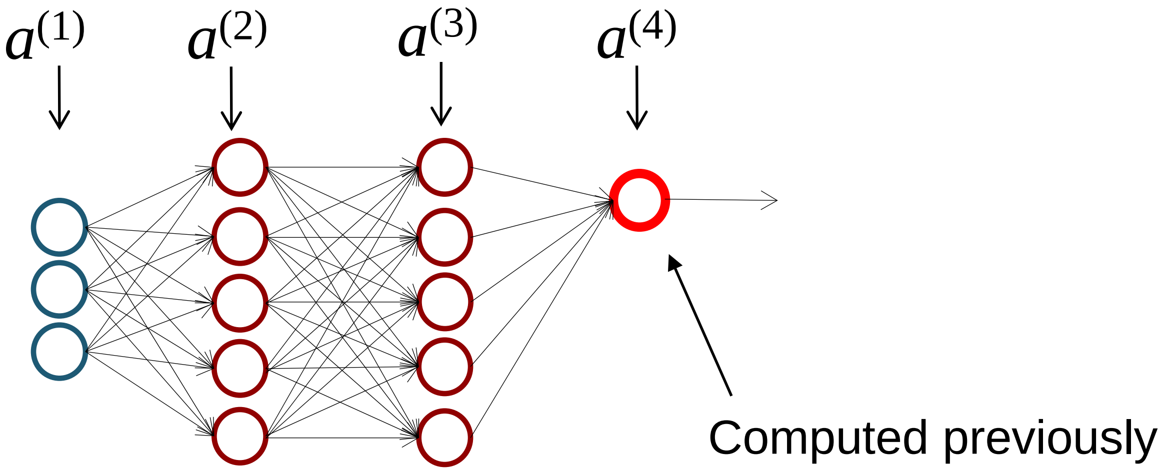
$$\delta^{(3)} = \left(w^{(3)}\right)^T \cdot \delta^{(4)} .* \frac{\partial a^{(3)}}{\partial z^{(3)}} \leftarrow a^{(3)} .* (1 - a^{(3)})$$

$$\delta^{(2)} = \left(w^{(2)}\right)^T \cdot \delta^{(3)} .* \frac{\partial a^{(2)}}{\partial z^{(2)}} \leftarrow a^{(2)} .* (1 - a^{(2)})$$

We ‘backpropagate’ the error.

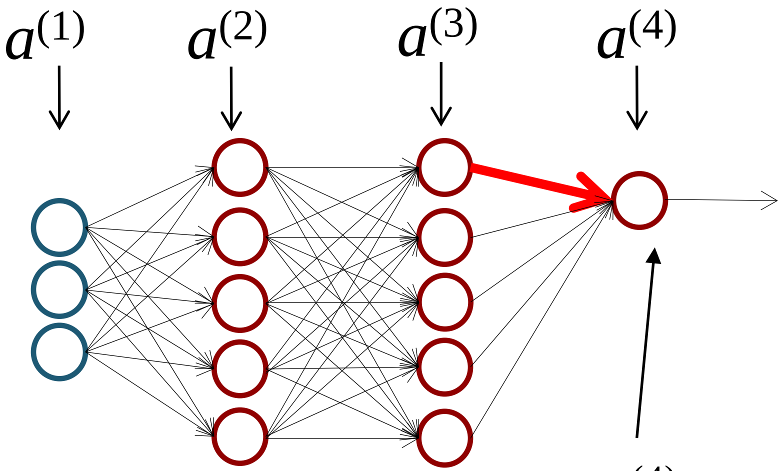
Backpropagation: example

$$\delta_1^{(3)} = \frac{\partial E}{\partial z_1^{(3)}} = a_1^{(3)} \cdot (1 - a_1^{(3)}) \cdot w_{11}^{(3)} \cdot \boxed{\delta_1^{(4)}}$$



Backpropagation: example

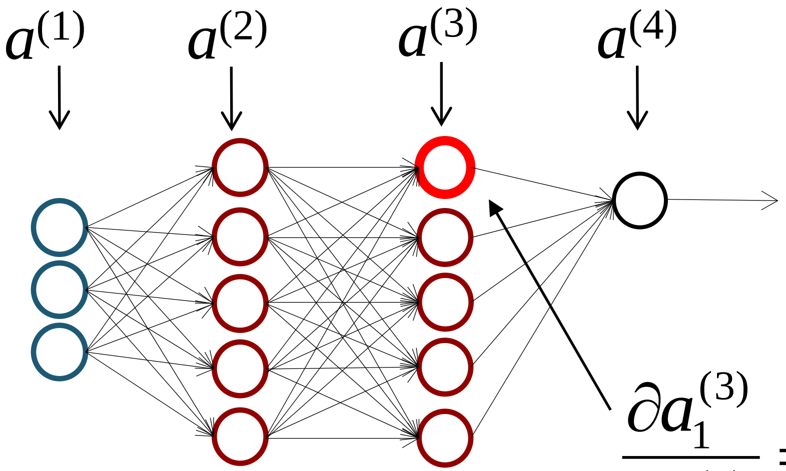
$$\delta_1^{(3)} = \frac{\partial E}{\partial z_1^{(3)}} = a_1^{(3)} \cdot (1 - a_1^{(3)}) \cdot \boxed{w_{11}^{(3)}} \cdot \delta_1^{(4)}$$



$$\frac{\partial z_1^{(4)}}{\partial a_1^{(3)}} = \frac{\partial (w_{11}^{(3)} a_1^{(3)} + w_{12}^{(3)} a_2^{(3)} + w_{13}^{(3)} a_3^{(3)} + ..)}{\partial a_1^{(3)}}$$

Backpropagation: example

$$\delta_1^{(3)} = \frac{\partial E}{\partial z_1^{(3)}} = \boxed{a_1^{(3)} \cdot (1 - a_1^{(3)})} \cdot w_{11}^{(3)} \cdot \delta_1^{(4)}$$



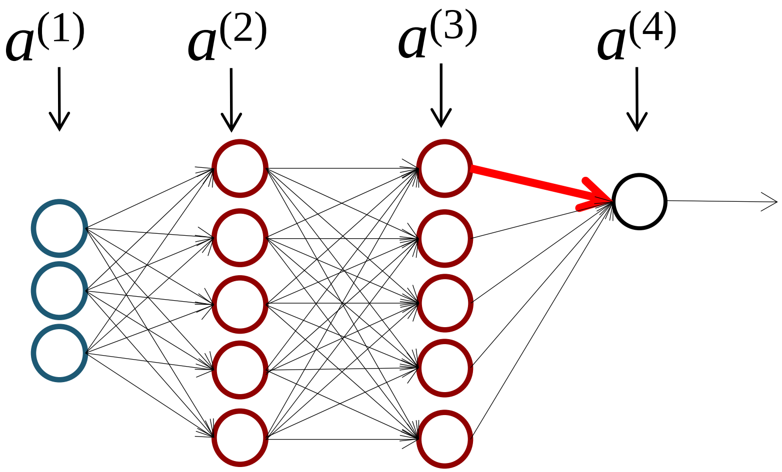
$$\frac{\partial a_1^{(3)}}{\partial z_1^{(3)}} = g'(z_1^{(3)})$$

Backpropagation: weight gradient

$$\frac{\partial E}{\partial w_{ij}^{(l)}} = \delta_i^{(l+1)} a_j^{(l)}$$

$$\frac{\partial (w_{11}^{(3)} a_1^{(3)} + w_{12}^{(3)} a_2^{(3)} + w_{13}^{(3)} a_3^{(3)} + ..)}{\partial w_{11}^{(3)}}$$

E.g. $$\frac{\partial E}{\partial w_{11}^{(3)}} = \frac{\partial E}{\partial z_1^{(4)}} \frac{\partial z_1^{(4)}}{\partial w_{11}^{(3)}} = \delta_1^{(4)} a_1^{(3)}$$

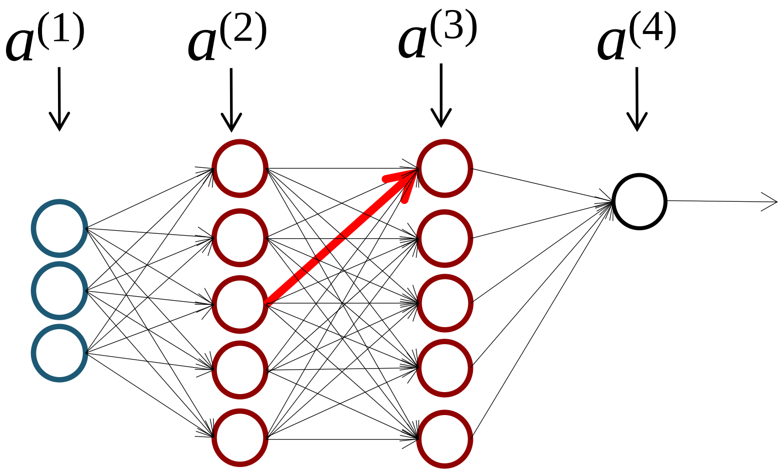


Backpropagation: weight gradient

$$\frac{\partial E}{\partial w_{ij}^{(l)}} = \delta_i^{(l+1)} a_j^{(l)}$$

$$\frac{\partial (w_{11}^{(2)} a_1^{(2)} + w_{12}^{(2)} a_2^{(2)} + w_{13}^{(2)} a_3^{(2)} + ..)}{\partial w_{13}^{(2)}}$$

Another E.g. $$\frac{\partial E}{\partial w_{13}^{(2)}} = \frac{\partial E}{\partial z_1^{(3)}} \frac{\partial z_1^{(3)}}{\partial w_{13}^{(2)}} = \delta_1^{(3)} a_3^{(2)}$$



A example of backpropagation with actual numbers

Please see the example given in https://mattmazur.com/2015/03/17/a-step-by-step-backpropagation-example/comment-page-5/#comments