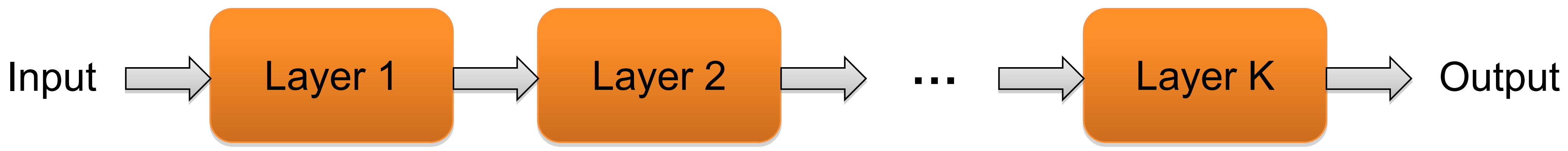
Backpropagation

Overview

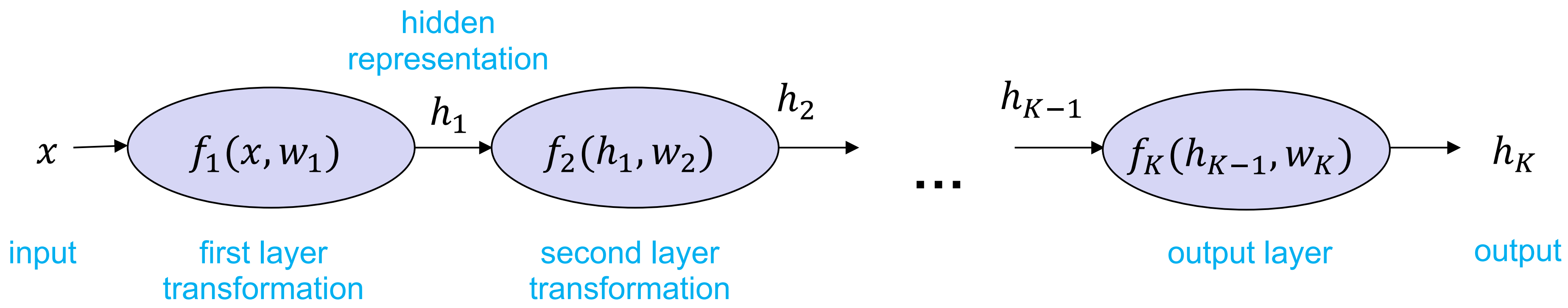
- Computation graphs
- Using the chain rule
- General backpropagation algorithm
- Toy examples of backward pass
- Matrix-vector calculations: ReLU, linear layer

Recall: Multi-layer neural networks

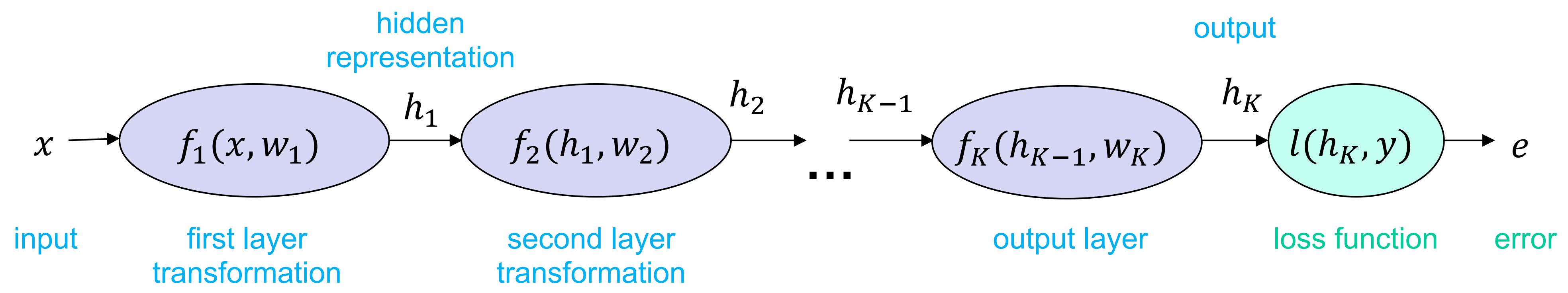
- The function computed by the network is a composition of the functions computed by individual layers (e.g., linear layers and nonlinearities):



- More precisely:



Training a multi-layer network

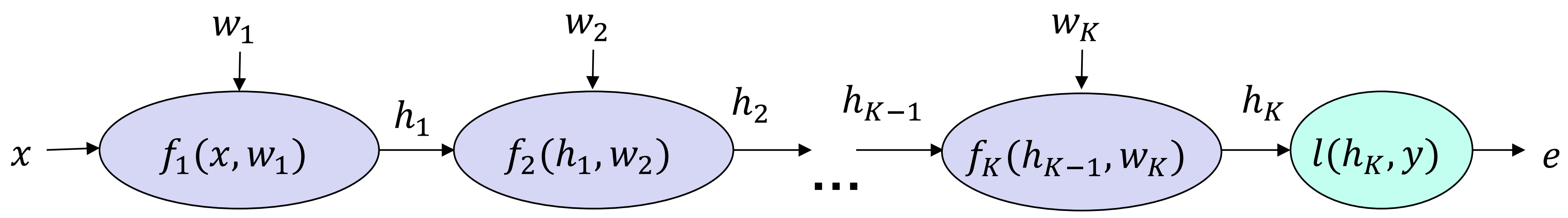


- What is the SGD update for the parameters w_k of the kth layer?

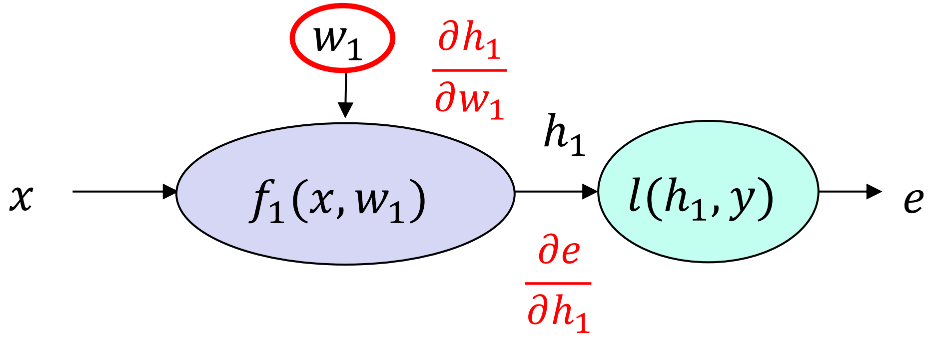
$$w_k \leftarrow w_k - \eta \frac{\partial e}{\partial w_k}$$

- To train the network, we need to find the gradient of the error w.r.t. the parameters of each layer, $\frac{\partial e}{\partial w_k}$

Computation graph



Chain rule



Let's start with $k = 1$

$$e = l(f_1(x, w_1), y)$$

Example: $e = (y - w_1^T x)^2$

$$h_1 = f_1(x, w_1) = w_1^T x$$

$$e = l(h_1, y) = (y - h_1)^2$$

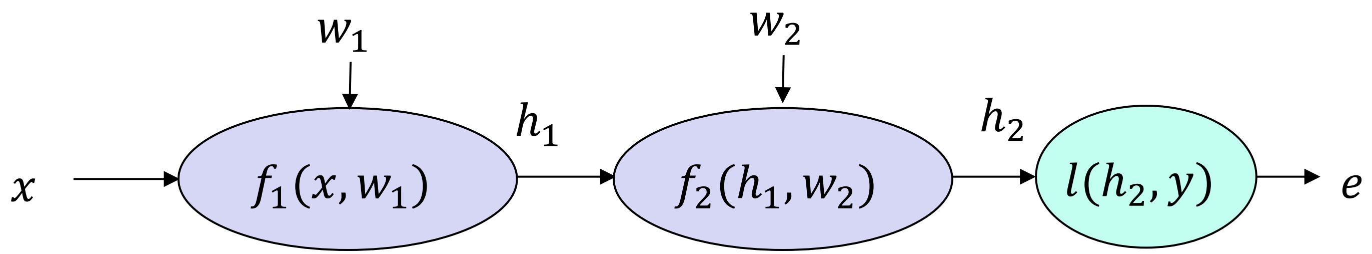
$$\frac{\partial h_1}{\partial w_1} =$$

$$\frac{\partial e}{\partial h_1} =$$

$$\frac{\partial e}{\partial w_1} =$$

Chain rule

$k = 2$

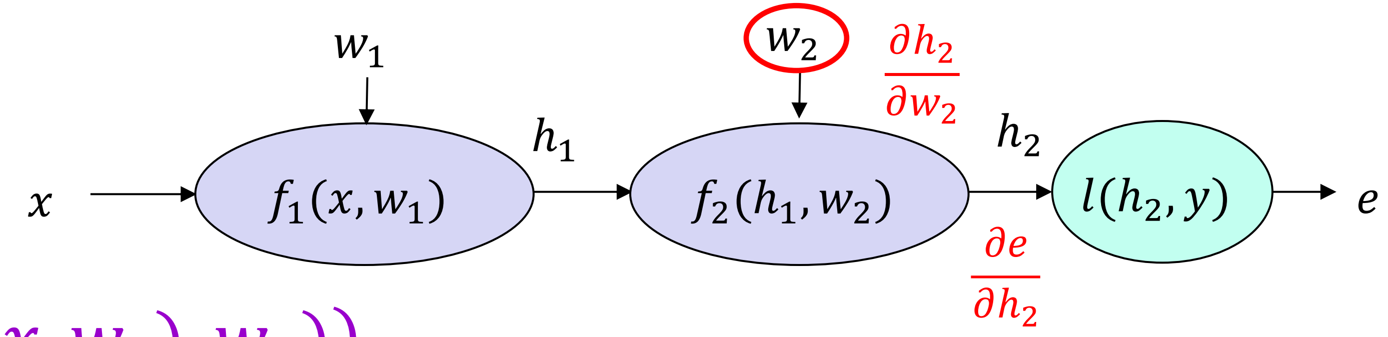


$$e = l(f_2(f_1(x, w_1), w_2))$$

Example: $e = -\log\left(\sigma(w_1^T x)\right)$ (assume $y = 1$)

$$h_1 = f_1(x, w_1) = w_1^T x$$
$$h_2 = f_2(h_1) = \sigma(h_1)$$
$$e = l(h_2, 1) = -\log(h_2)$$

Chain rule



$$k = 2$$

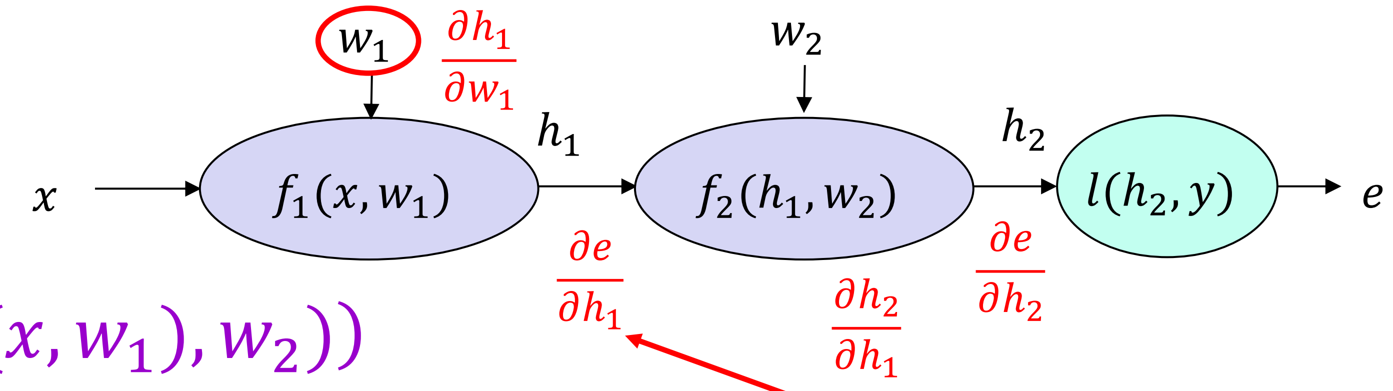
$$e = l(f_2(f_1(x, w_1), w_2))$$

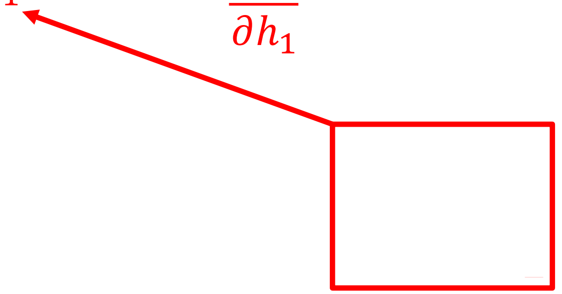
$$\frac{\partial e}{\partial w_2} =$$

Chain rule

$k = 2$

$e = l(f_2(f_1(x, w_1), w_2))$





$$\frac{\partial e}{\partial w_2} = \frac{\partial e}{\partial h_2}\frac{\partial h_2}{\partial w_2}$$

Example: $e = -\log\left(\sigma(w_1^T x)\right)$ (assume $y = 1$)

$$h_1 = f_1(x, w_1) = w_1^T x$$
$$h_2 = f_2(h_1) = \sigma(h_1)$$
$$e = l(h_2, 1) = -\log(h_2)$$

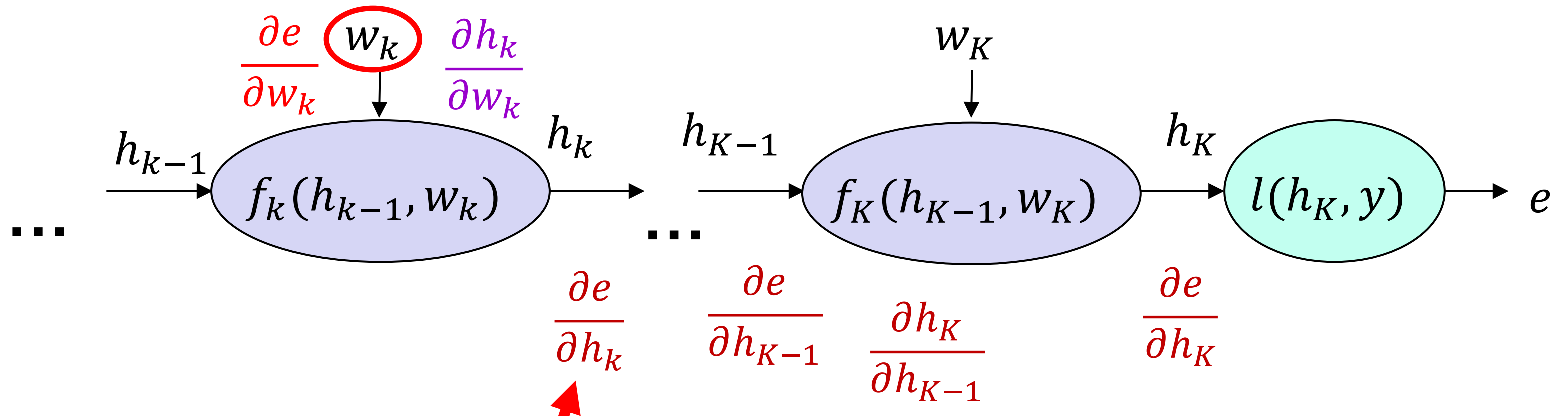
$$\frac{\partial h_1}{\partial w_1} =$$

$$\frac{\partial h_2}{\partial h_1} =$$

$$\frac{\partial e}{\partial h_2} =$$

$$\frac{\partial e}{\partial w_1} = \frac{\partial e}{\partial h_2}\frac{\partial h_2}{\partial h_1}\frac{\partial h_1}{\partial w_1} =$$

Chain rule



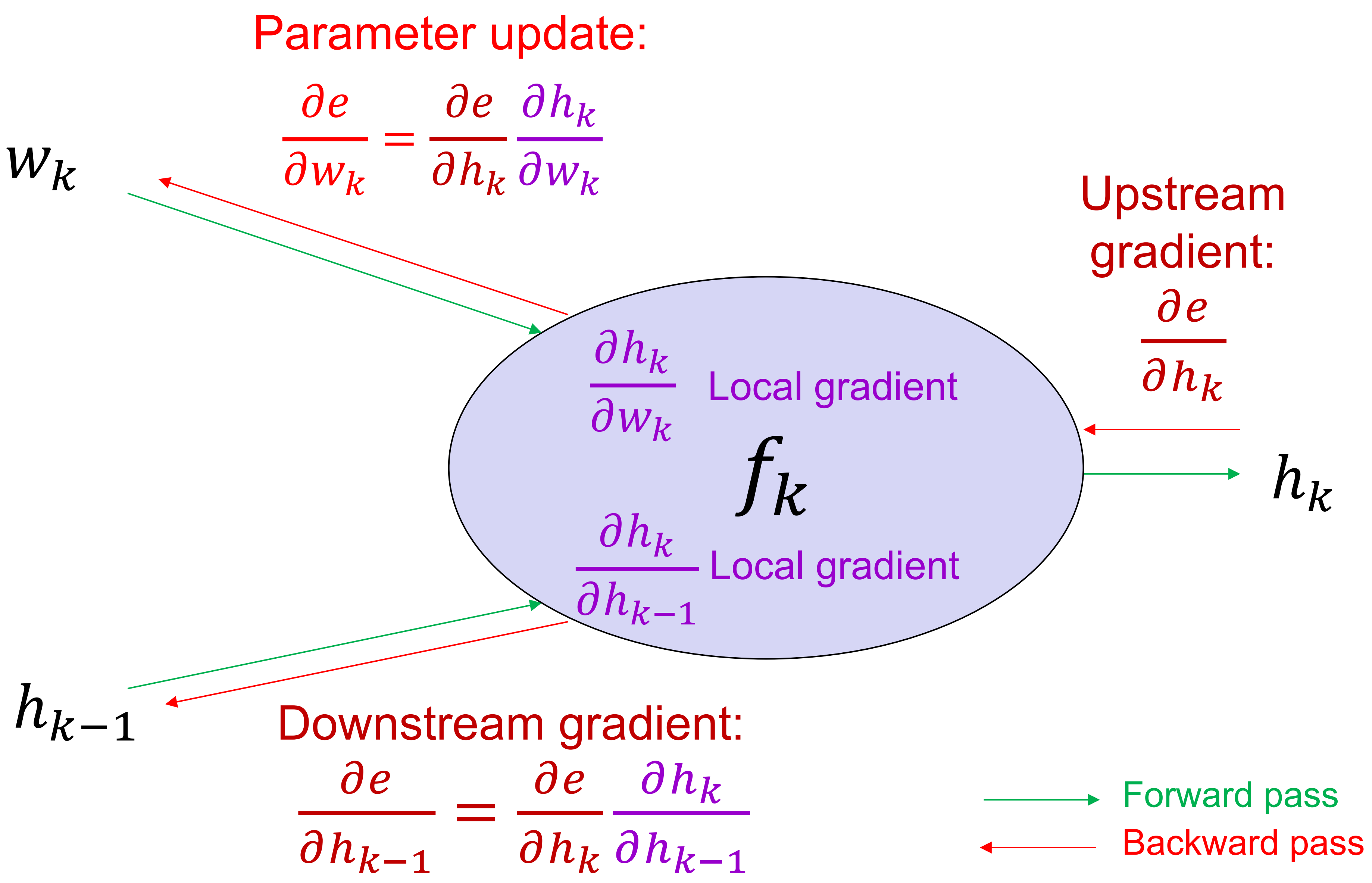
General case:

$$\frac{\partial e}{\partial w_k} = \boxed{\frac{\partial e}{\partial h_K} \frac{\partial h_K}{\partial h_{K-1}} \cdots \frac{\partial h_{k+1}}{\partial h_k}} \frac{\partial h_k}{\partial w_k}$$

Upstream gradient $\frac{\partial e}{\partial h_k}$

Local gradient

Backpropagation summary



What about more general computation graphs?

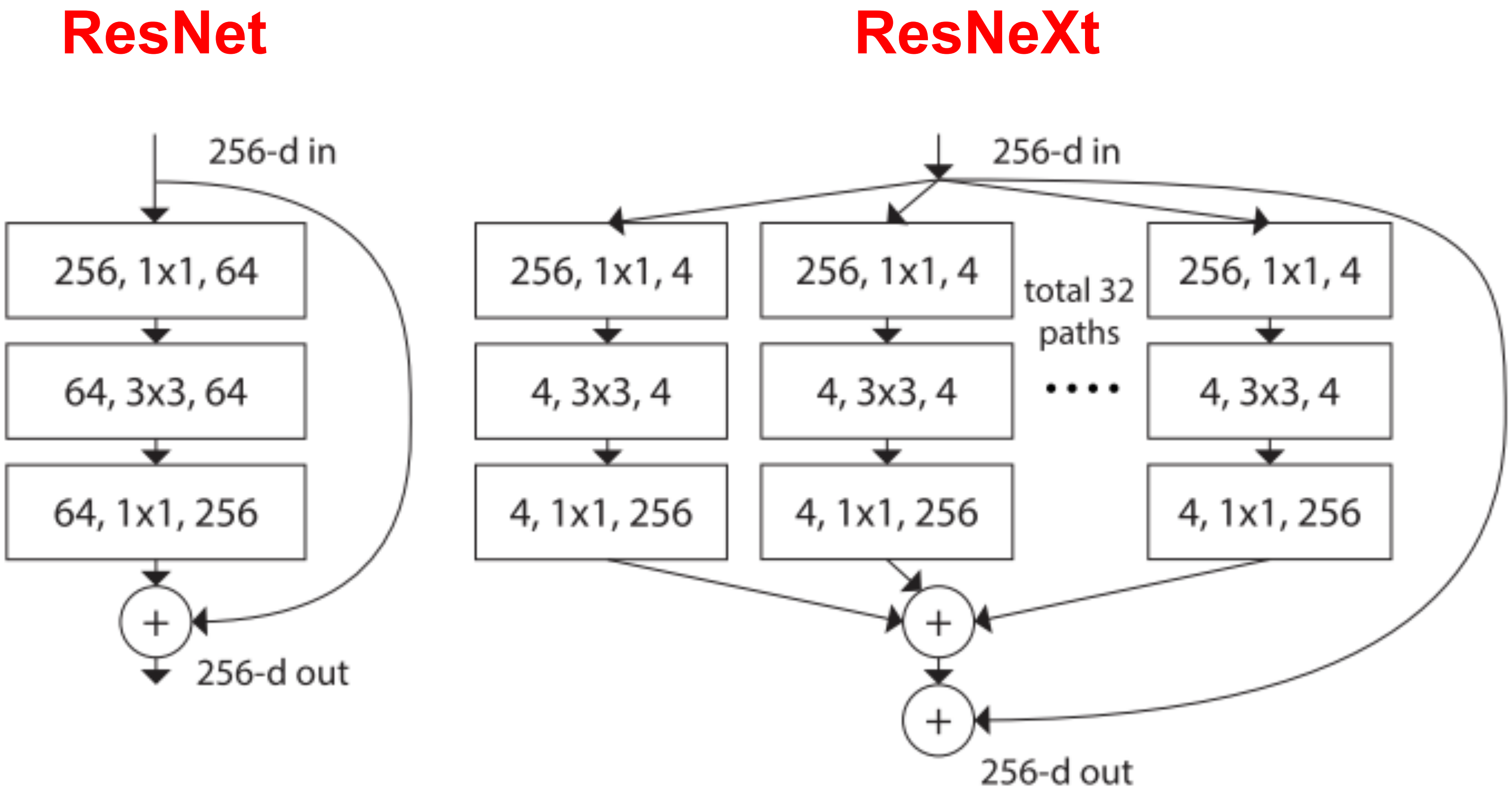
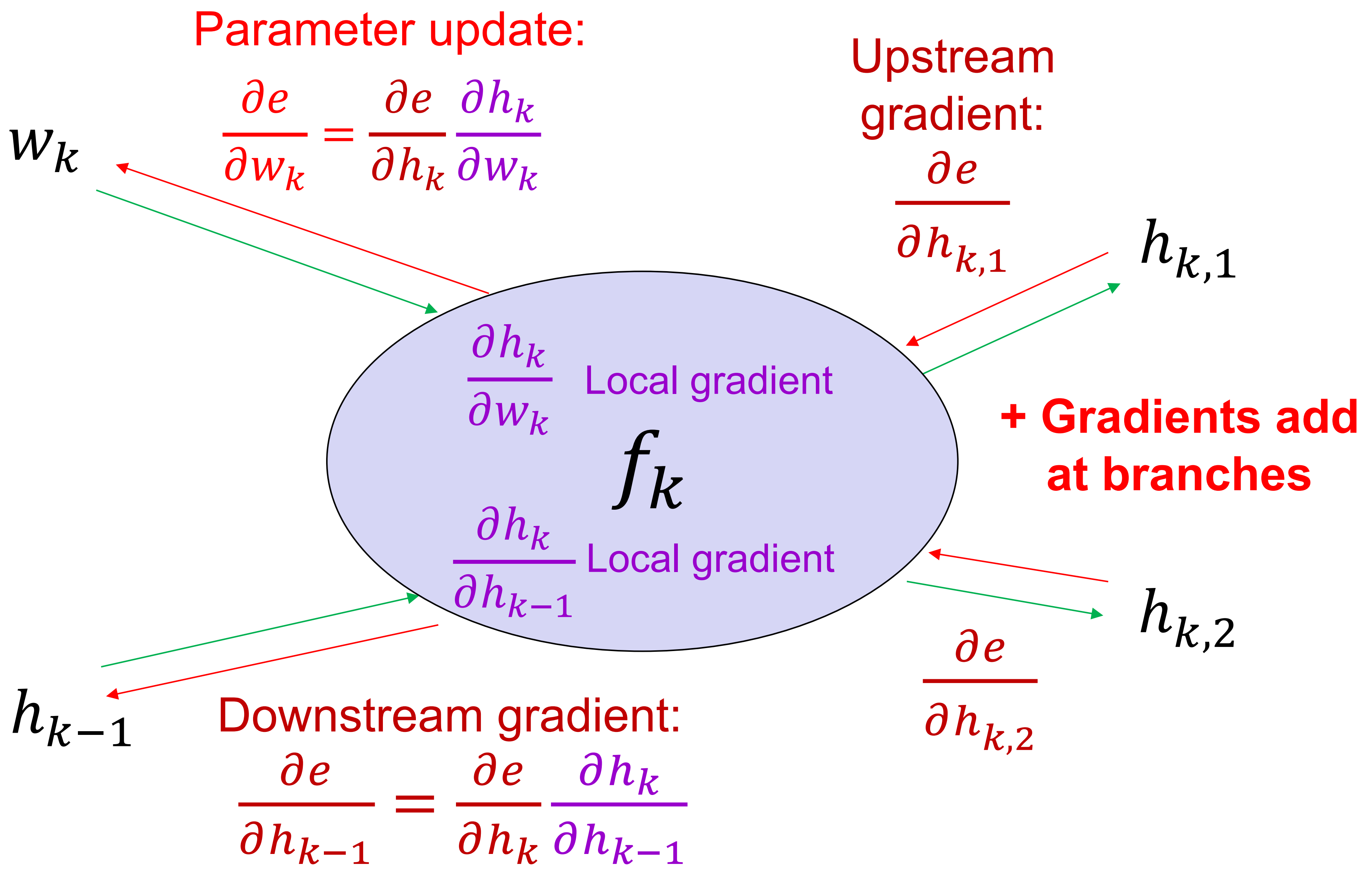


Figure source

What about more general computation graphs?

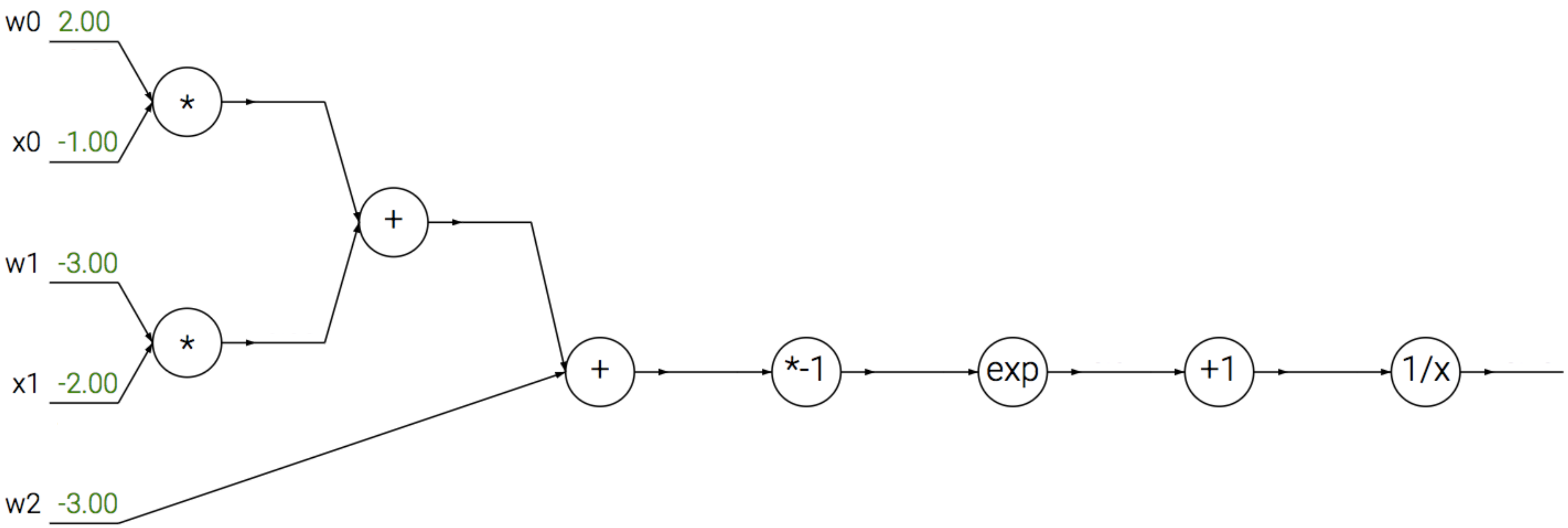


Overview

- Computation graphs
- Using the chain rule
- General backprop algorithm
- Toy examples of backward pass

A detailed example

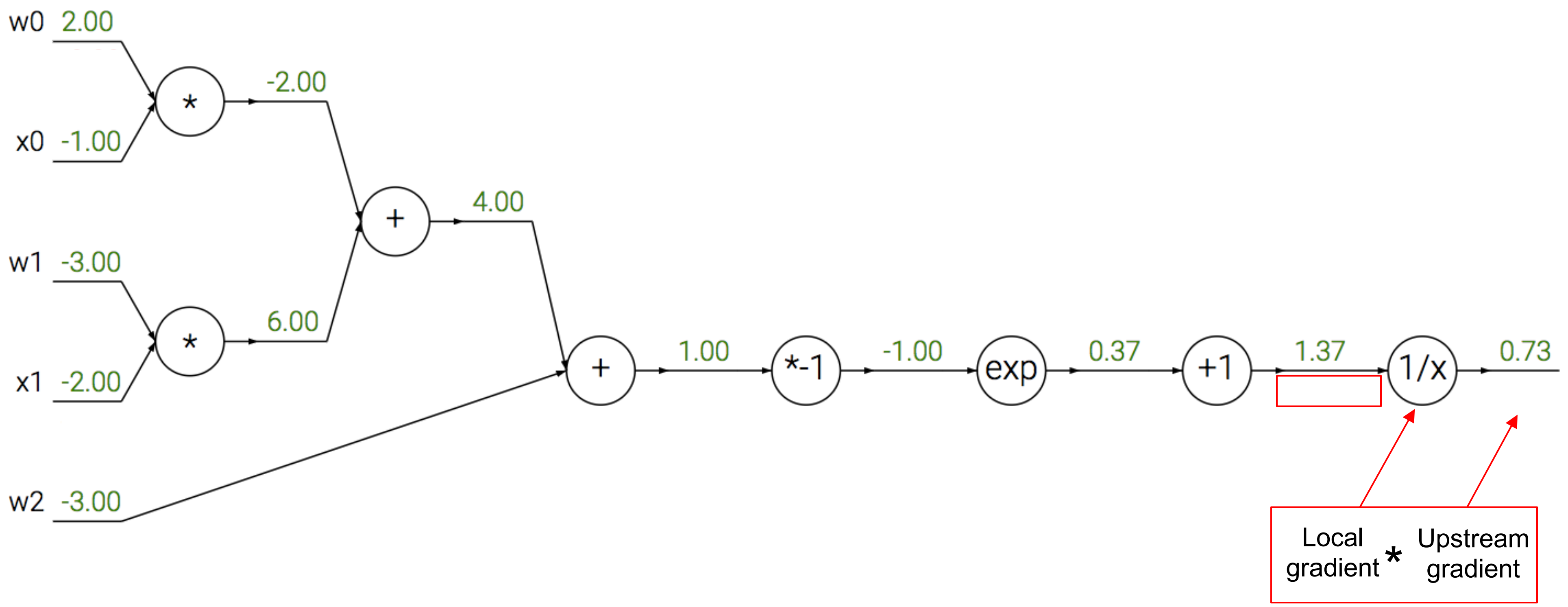
$$f(x, w) = \frac{1}{1 + \exp[-(w^{(0)}x^{(0)} + w^{(1)}x^{(1)} + w^{(2)})]}$$



Source: [Stanford 231n]()

A detailed example

$$f(x,w) = \frac{1}{1 + \exp[-(w^{(0)}x^{(0)} + w^{(1)}x^{(1)} + w^{(2)})]}$$



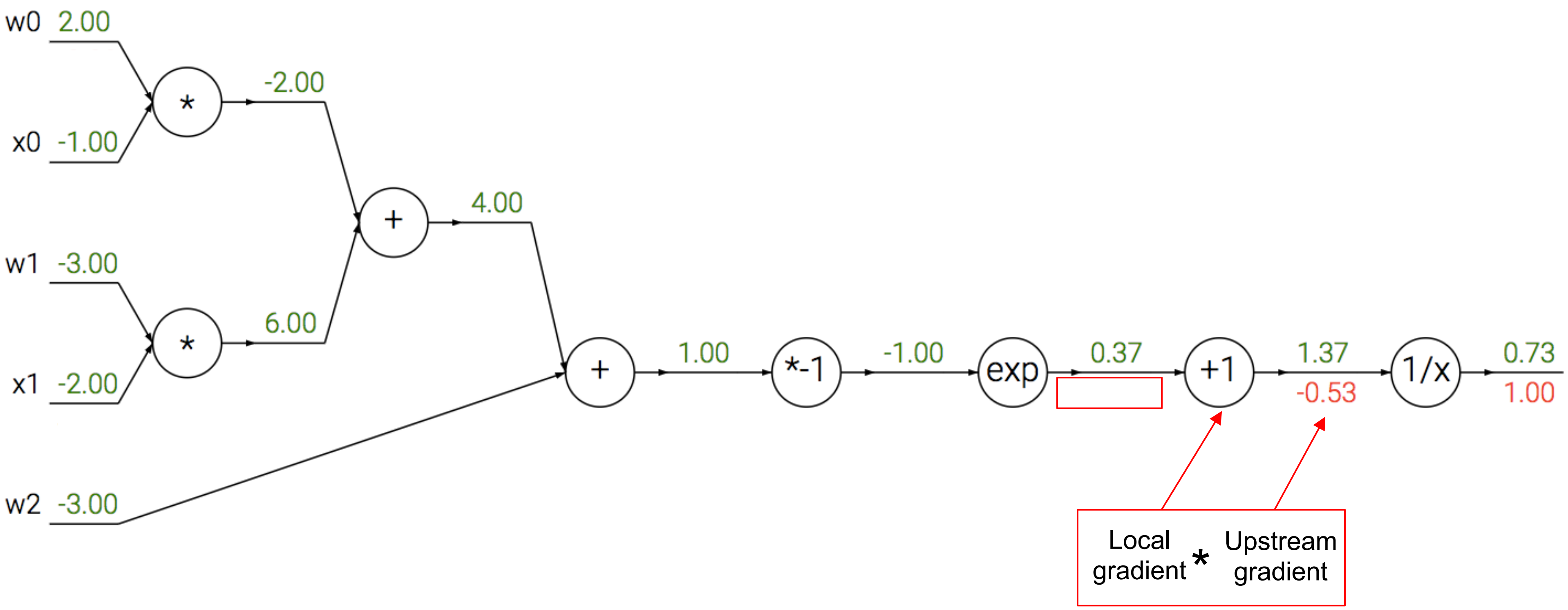
$$(1/x)' = -1/x^2$$

$$-\frac{1}{1.37^2} * 1 = -0.53$$

Source: [Stanford 231n]

A detailed example

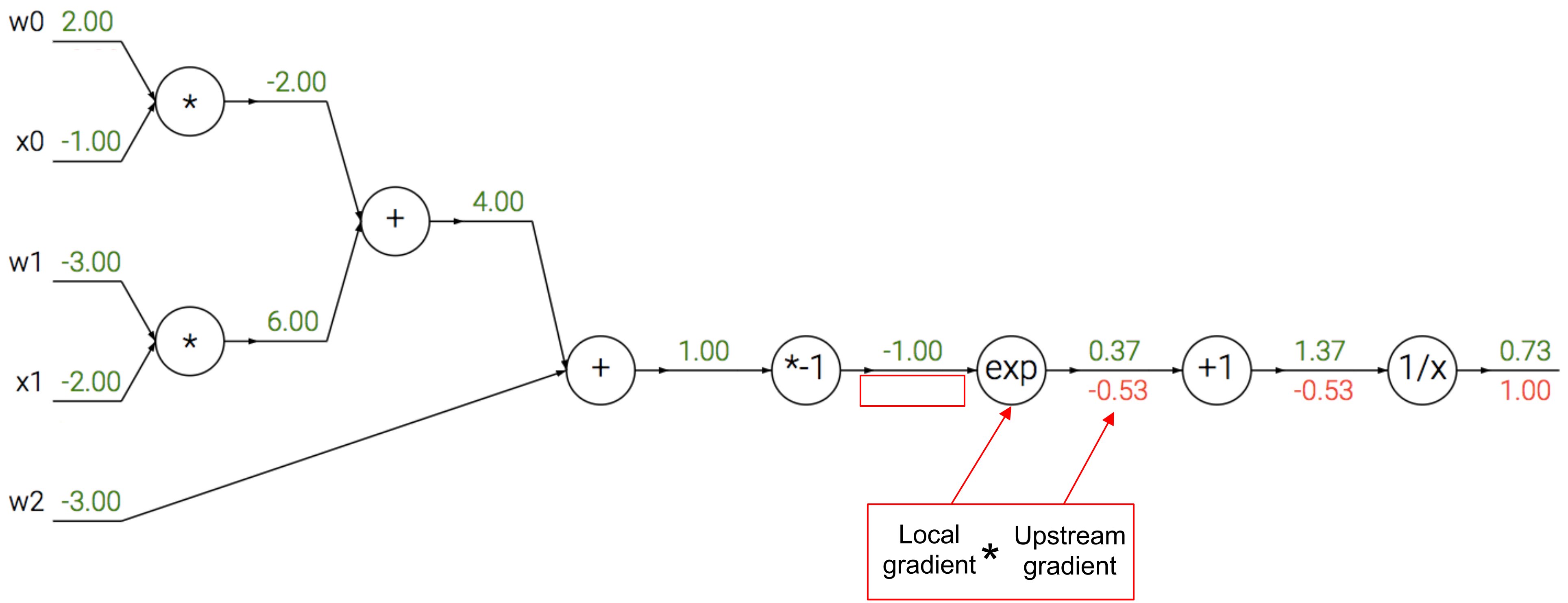
$$f(x,w) = \frac{1}{1 + \exp[-(w^{(0)}x^{(0)} + w^{(1)}x^{(1)} + w^{(2)})]}$$



Source: [Stanford 231n]()

A detailed example

$$f(x,w) = \frac{1}{1 + \exp[-(w^{(0)}x^{(0)} + w^{(1)}x^{(1)} + w^{(2)})]}$$

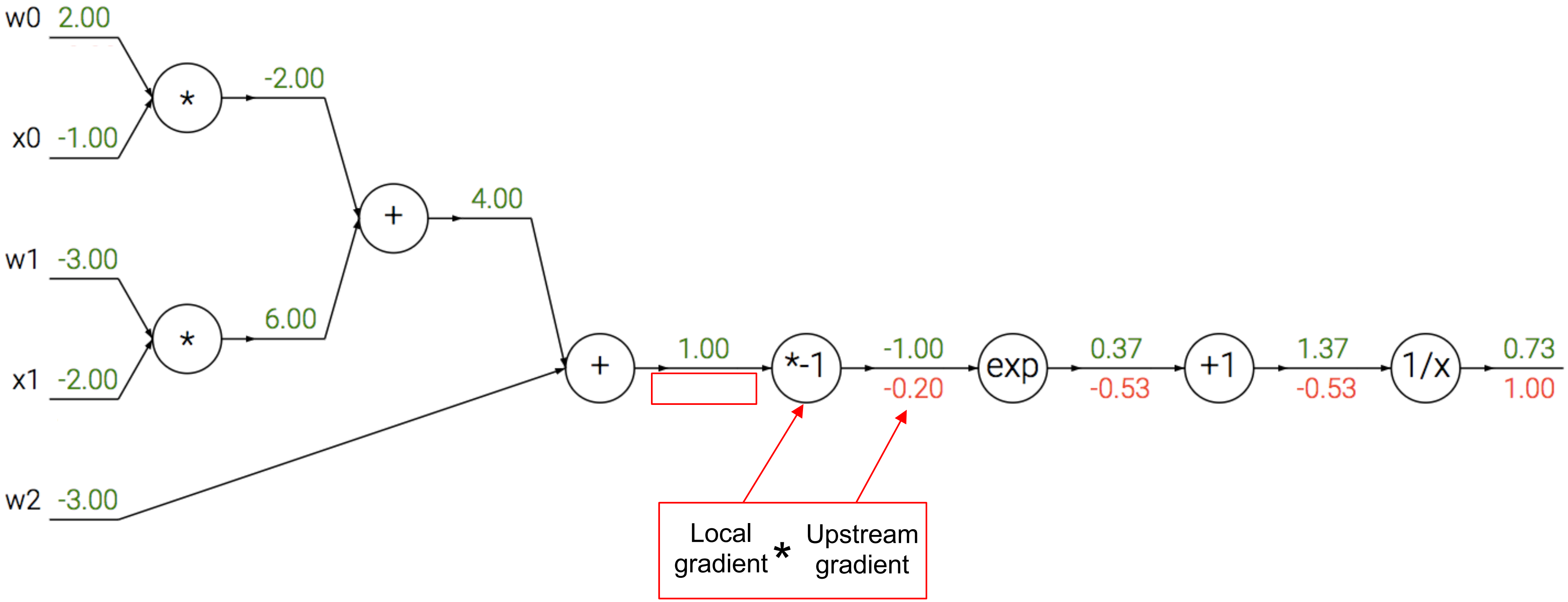


$$\exp(-1) * (-0.53) = -0.20$$

Source: [Stanford 231n]()

A detailed example

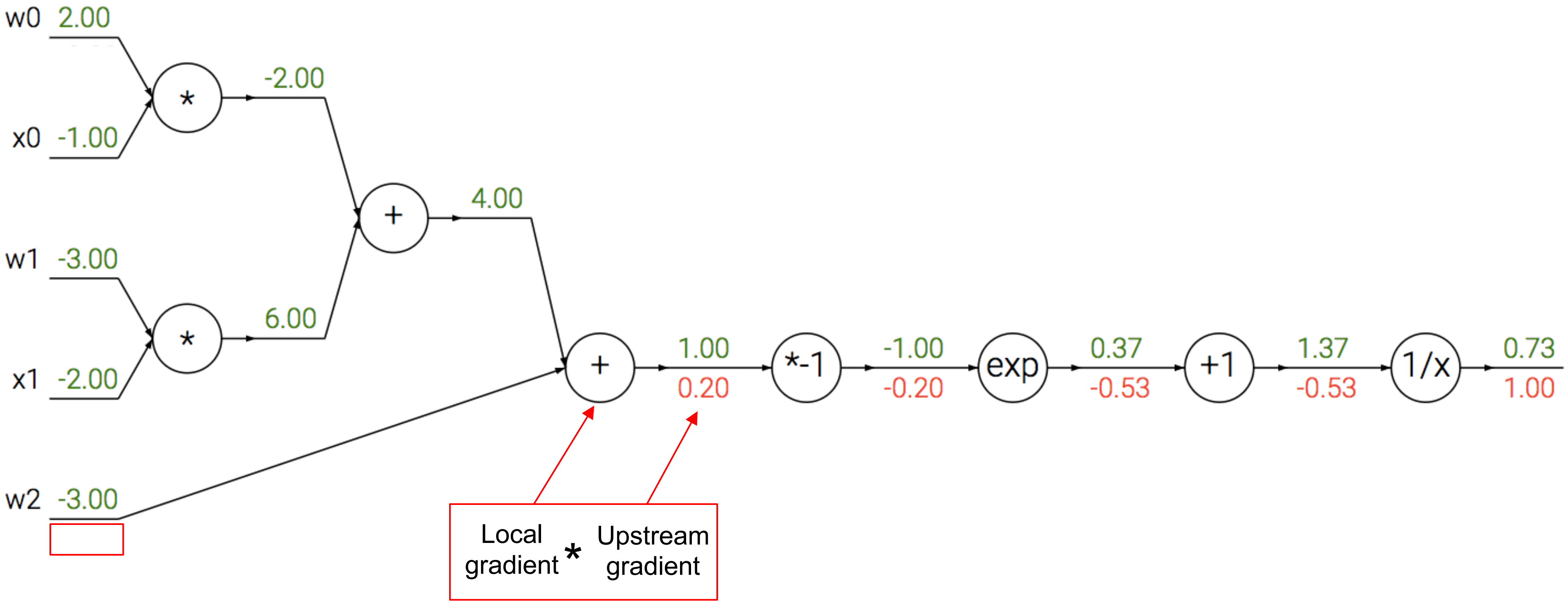
$$f(x,w) = \frac{1}{1+\exp[-(w^{(0)}x^{(0)}+w^{(1)}x^{(1)}+w^{(2)})]}$$



Source: [Stanford 231n]()

A detailed example

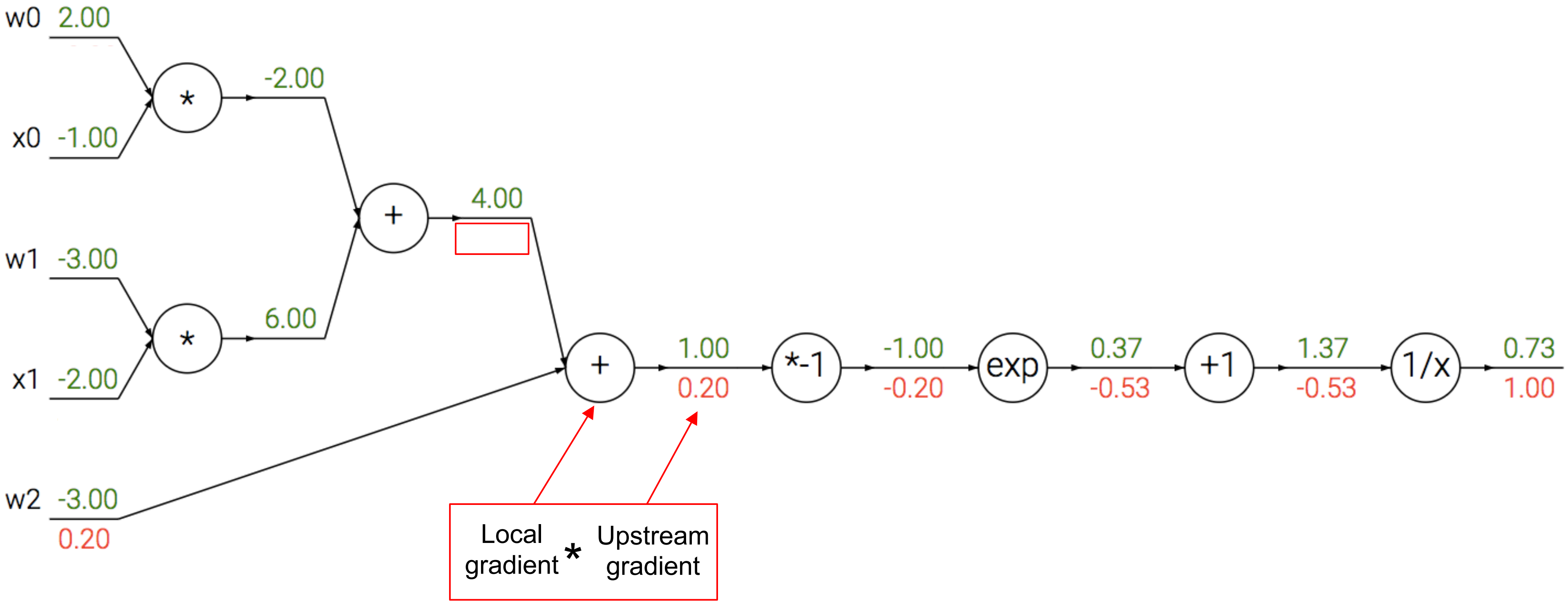
$$f(x,w) = \frac{1}{1+\exp[-(w^{(0)}x^{(0)} + w^{(1)}x^{(1)} + w^{(2)})]}$$



Source: [Stanford 231n]()

A detailed example

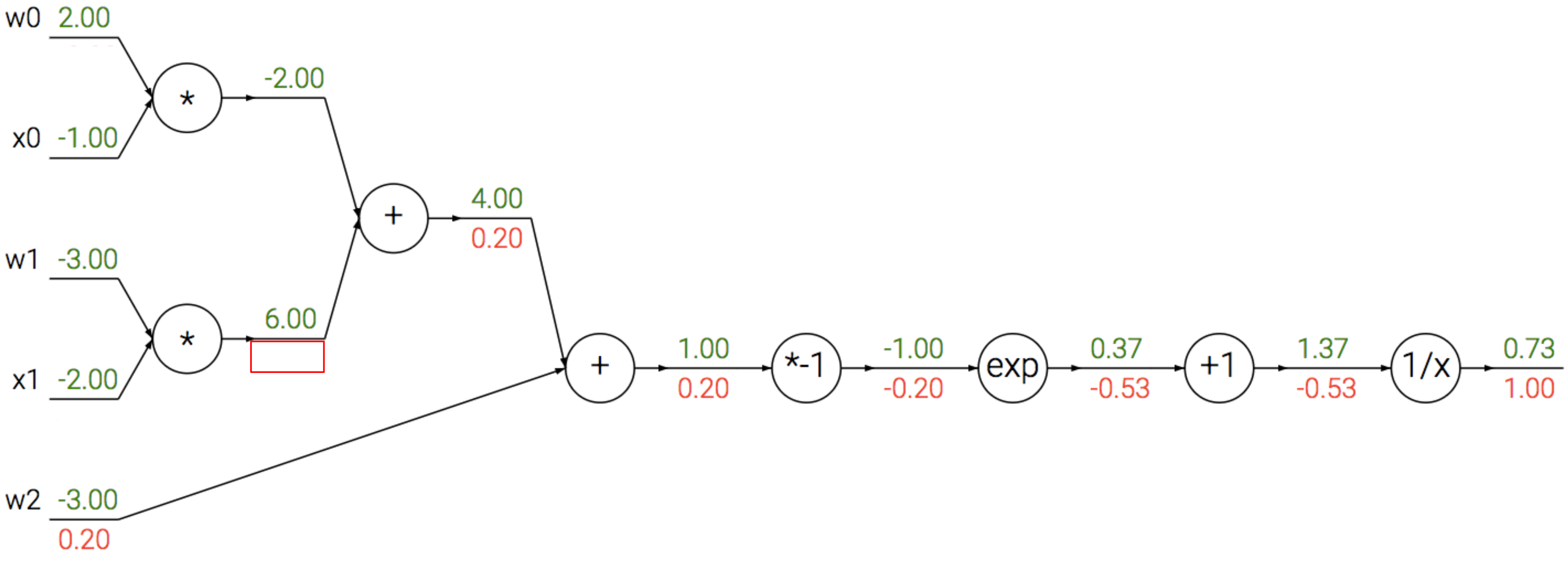
$$f(x,w) = \frac{1}{1+\exp[-(w^{(0)}x^{(0)} + w^{(1)}x^{(1)} + w^{(2)})]}$$



Source: [Stanford 231n]()

A detailed example

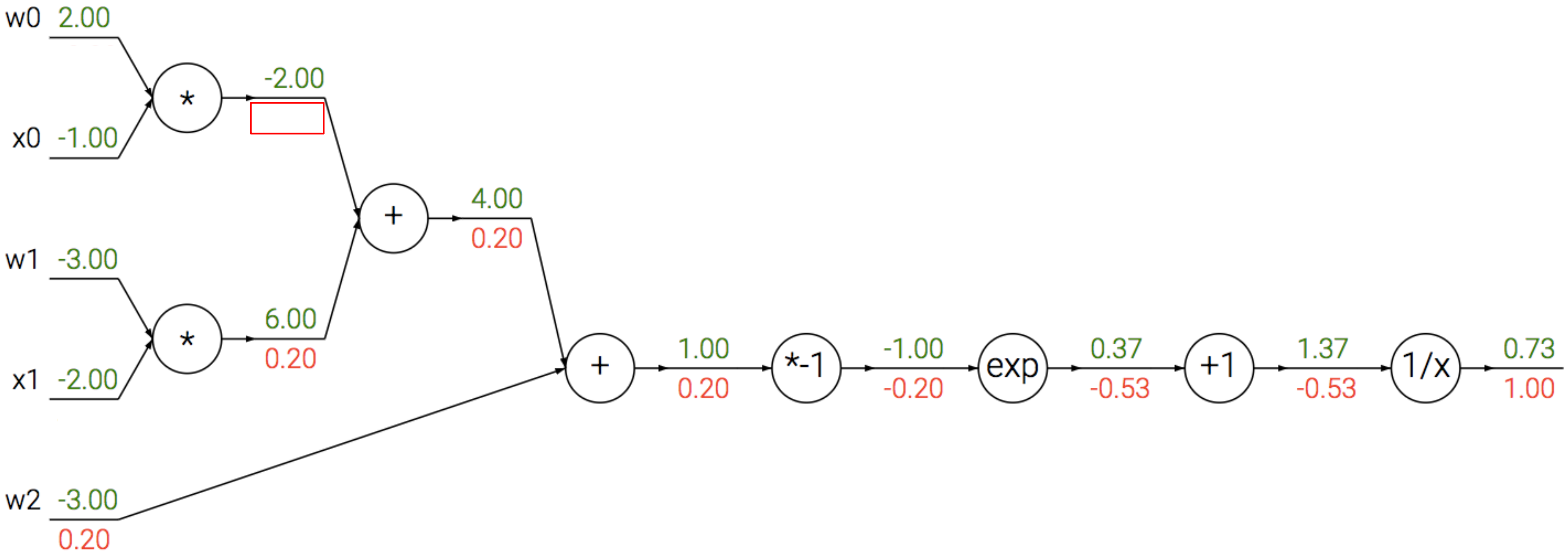
$$f(x,w) = \frac{1}{1+\exp[-(w^{(0)}x^{(0)} + w^{(1)}x^{(1)} + w^{(2)})]}$$



Source: [Stanford 231n]

A detailed example

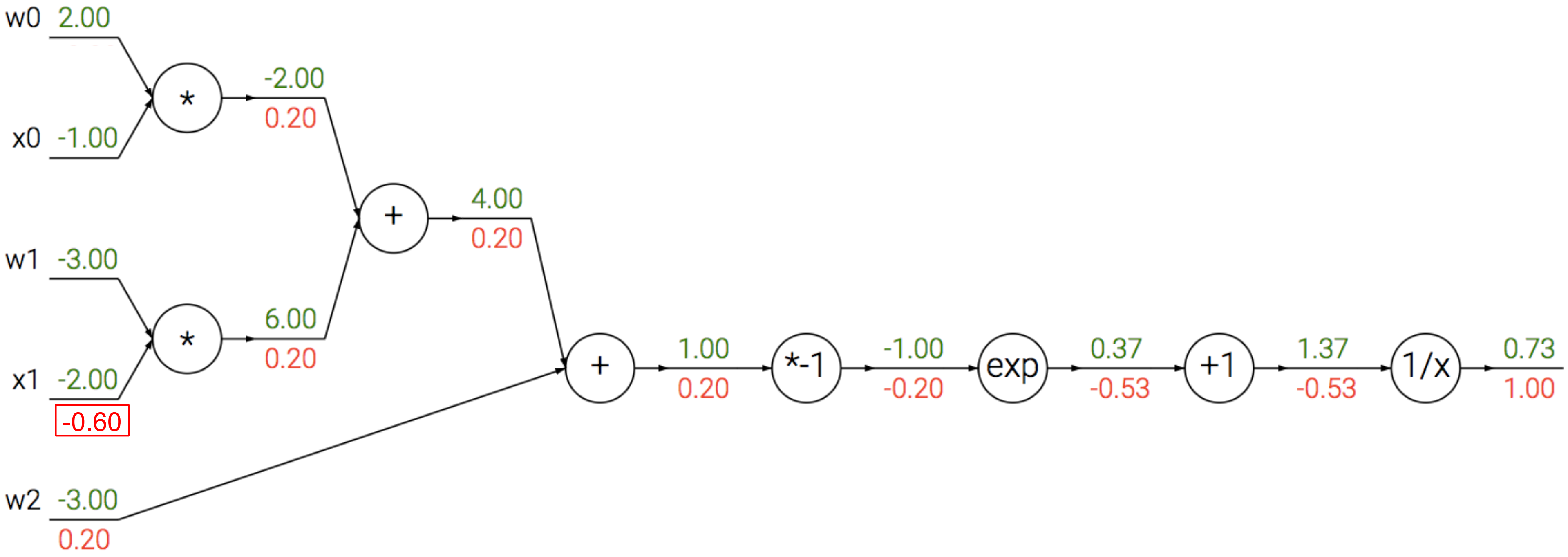
$$f(x,w) = \frac{1}{1+\exp[-(w^{(0)}x^{(0)} + w^{(1)}x^{(1)} + w^{(2)})]}$$



Source: [Stanford 231n]()

A detailed example

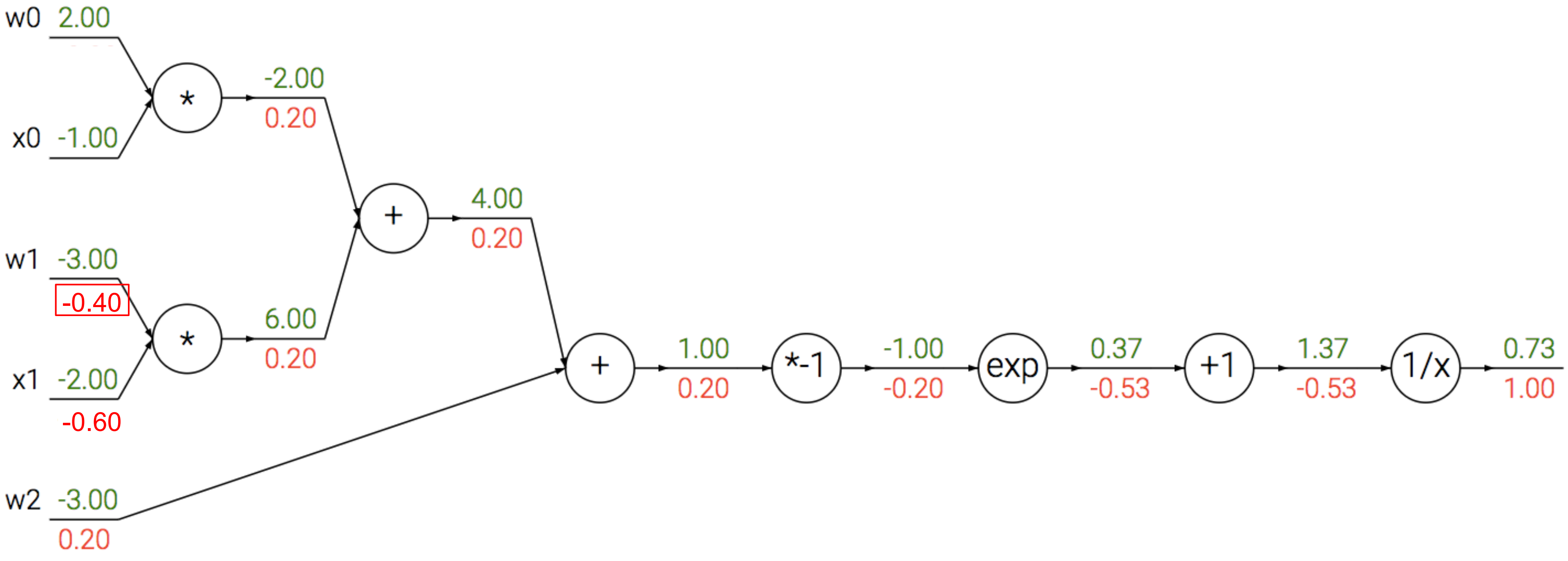
$$f(x,w) = \frac{1}{1+\exp[-(w^{(0)}x^{(0)} + w^{(1)}x^{(1)} + w^{(2)})]}$$



Source: [Stanford 231n]()

A detailed example

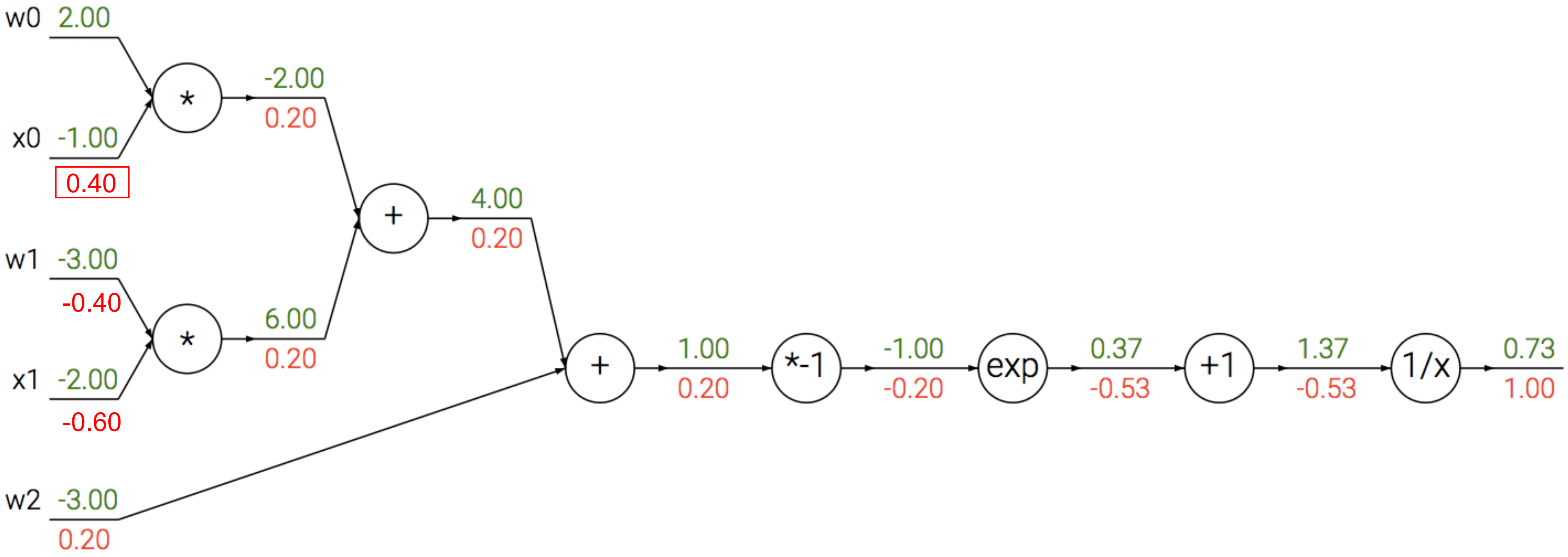
$$f(x,w) = \frac{1}{1+\exp[-(w^{(0)}x^{(0)} + w^{(1)}x^{(1)} + w^{(2)})]}$$



Source: [Stanford 231n]()

A detailed example

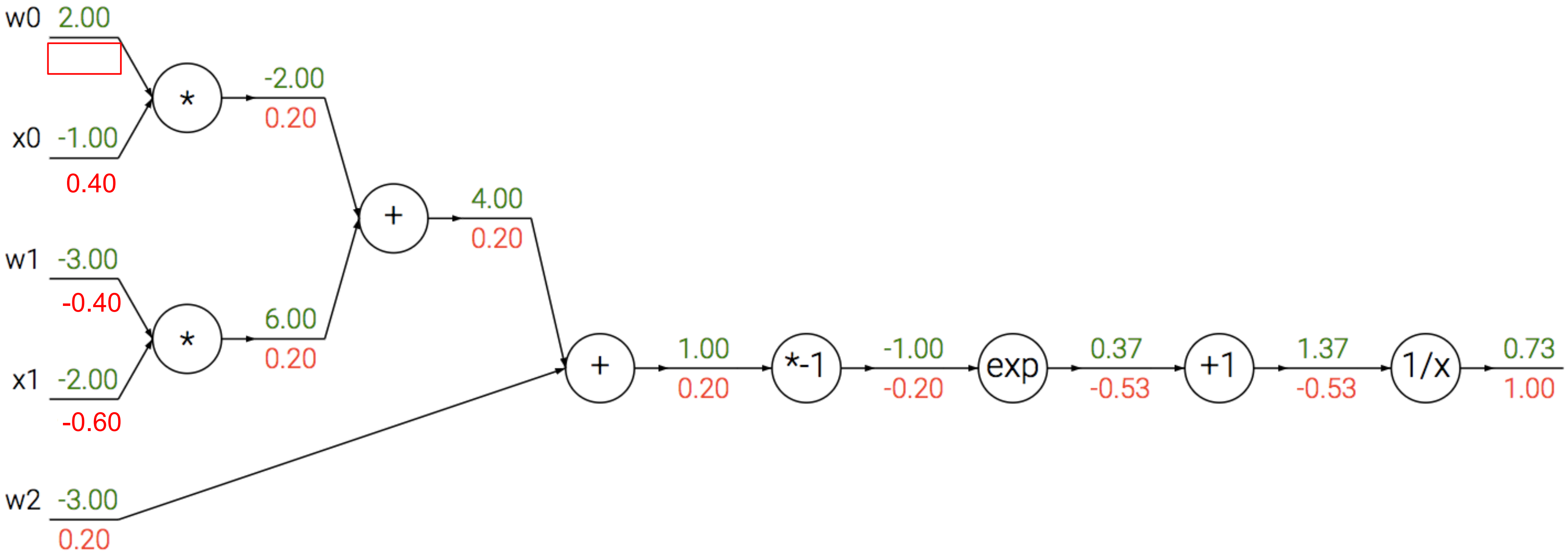
$$f(x,w) = \frac{1}{1+\exp[-(w^{(0)}x^{(0)} + w^{(1)}x^{(1)} + w^{(2)})]}$$



Source: [Stanford 231n]()

A detailed example

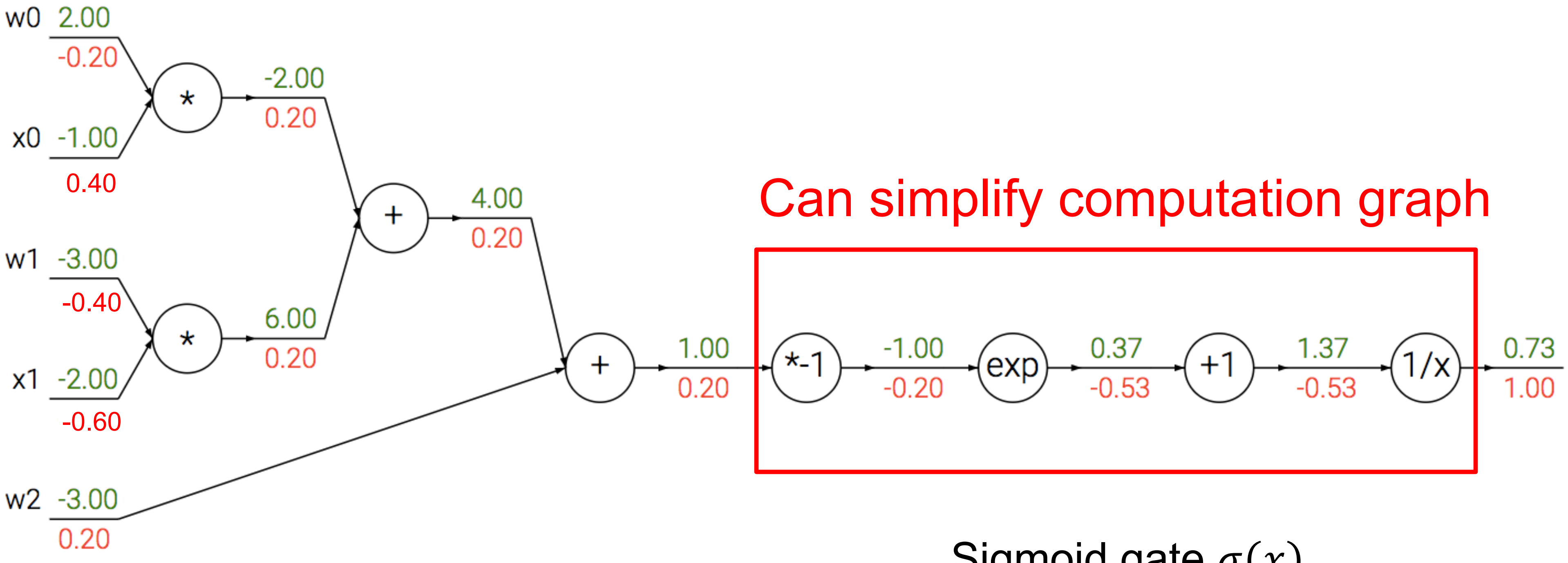
$$f(x,w) = \frac{1}{1 + \exp[-(w^{(0)}x^{(0)} + w^{(1)}x^{(1)} + w^{(2)})]}$$



Source: [Stanford 231n]()

A detailed example

$$f(x, w) = \frac{1}{1 + \exp[-(w^{(0)}x^{(0)} + w^{(1)}x^{(1)} + w^{(2)})]}$$



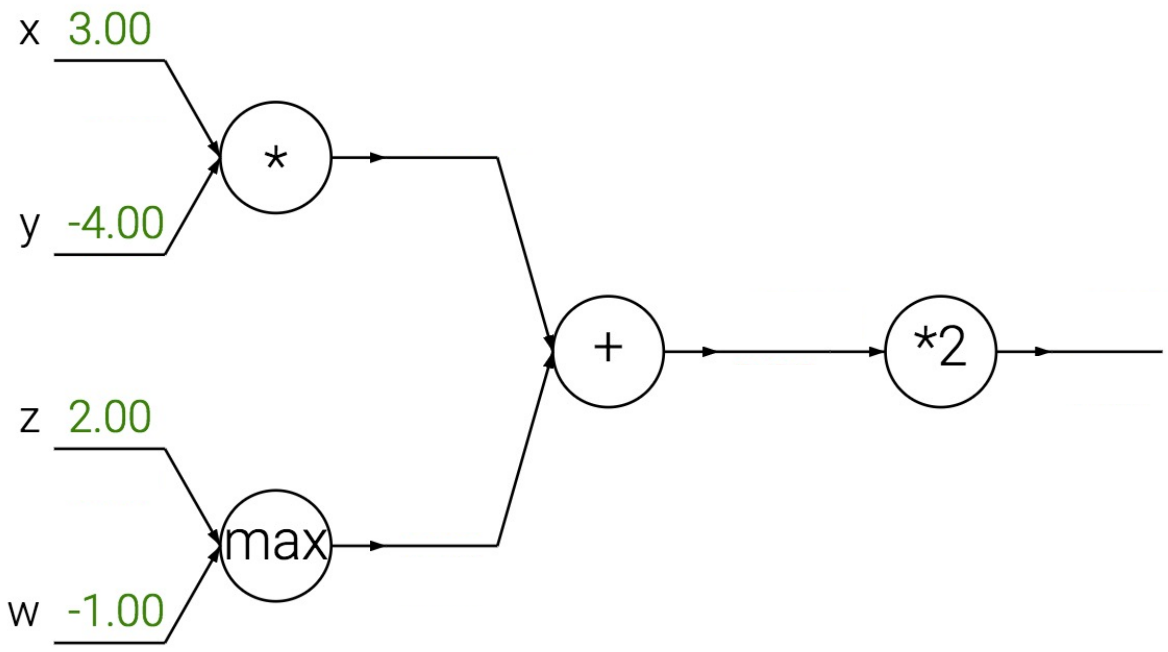
Sigmoid gate $\sigma(x)$

$$\sigma'(x) = \sigma(x)\big(1 - \sigma(x)\big)$$

$$\sigma(1)\big(1 - \sigma(1)\big) = 0.73 * (1 - 0.73) = 0.20$$

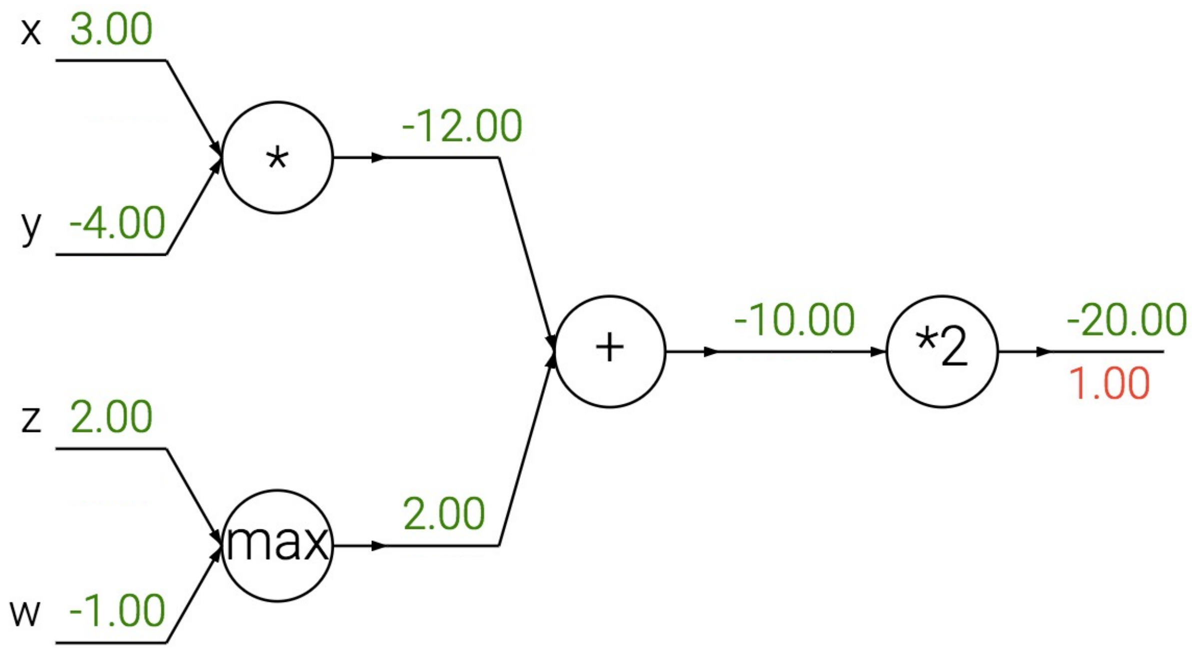
Source: [Stanford 231n]

Another example



Source: [Stanford 231n]

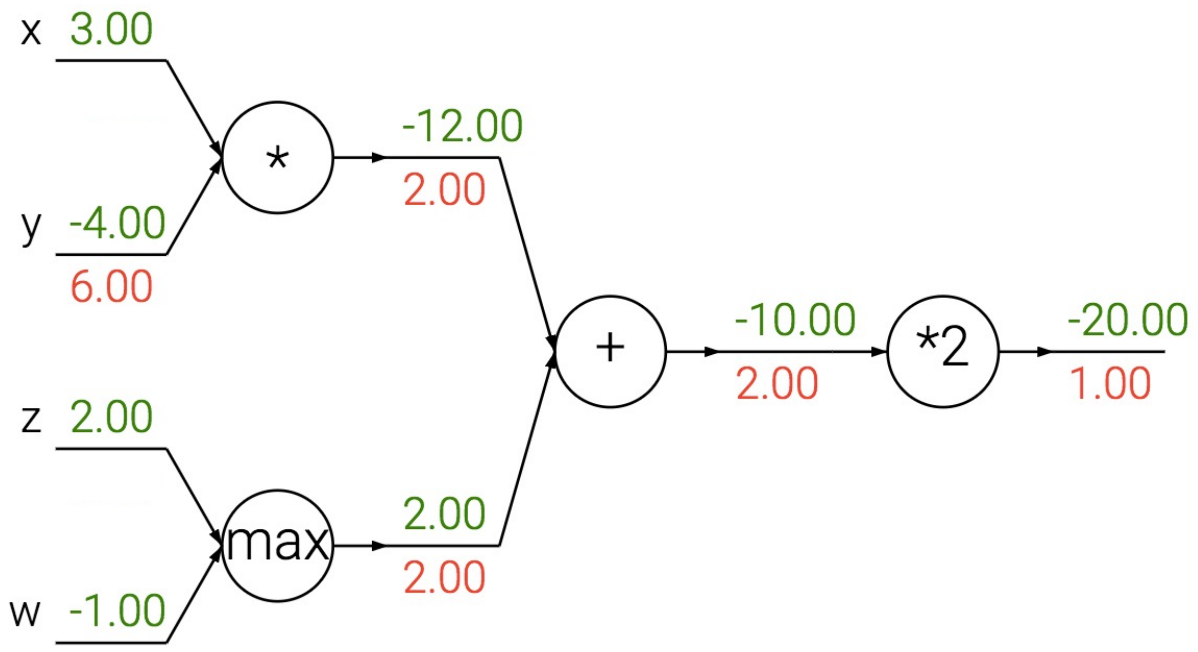
Another example



Add gate: “gradient distributor”

Source: [Stanford 231n]

Another example

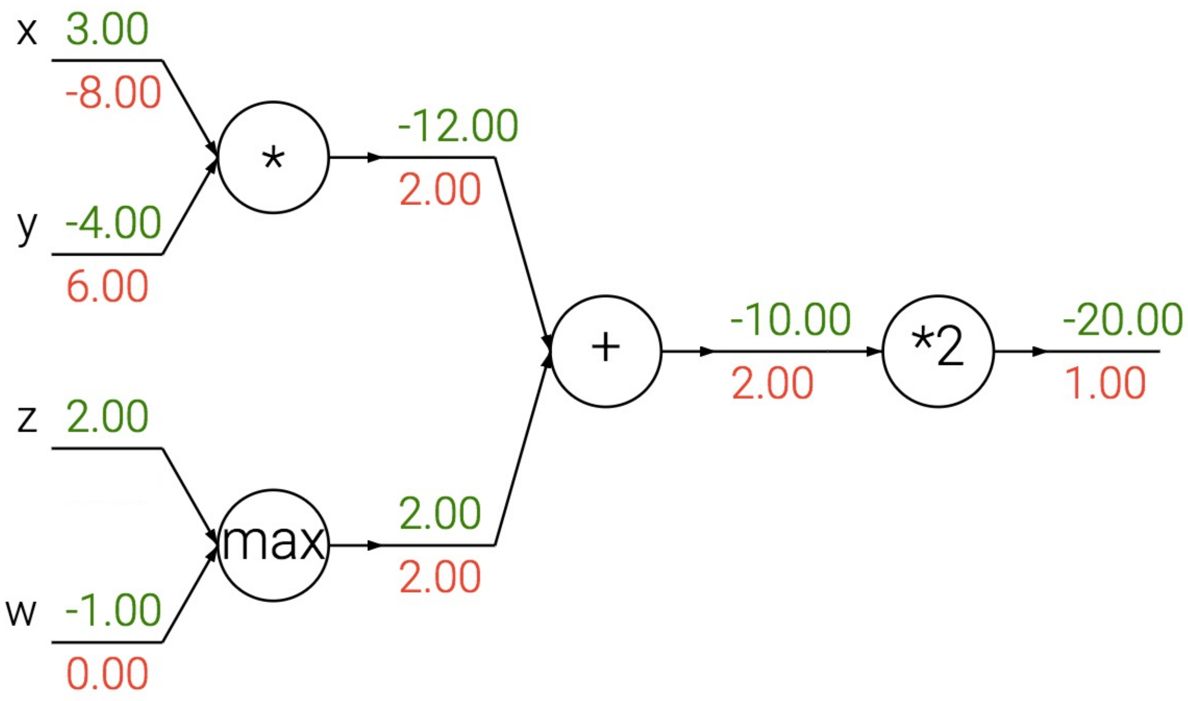


Add gate: “gradient distributor”

Multiply gate: “gradient switcher”

Source: Stanford 231n

Another example



Add gate: “gradient distributor”

Multiply gate: “gradient switcher”

Max gate: “gradient router”

Source: [Stanford 231n]

General tips

- Derive error signal (upstream gradient) directly, avoid explicit computation of huge local derivatives
- Write out expression for a single element of the Jacobian, then deduce the overall formula
- Keep consistent indexing conventions, order of operations
- Use dimension analysis

- **For further reading:**
  - Lecture 4 of Stanford 231n and associated links in the syllabus
  - Yes you should understand backprop by Andrej Karpathy