

Backpropagation

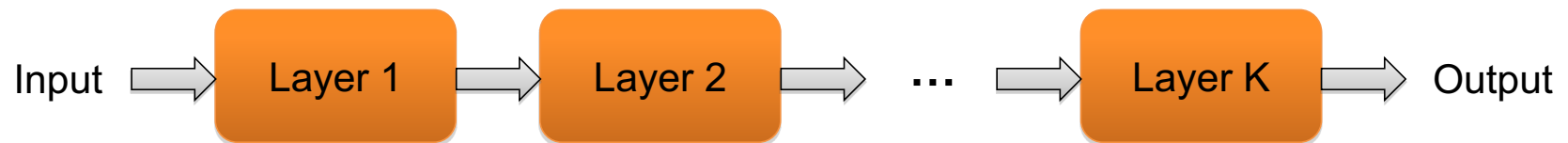


Overview

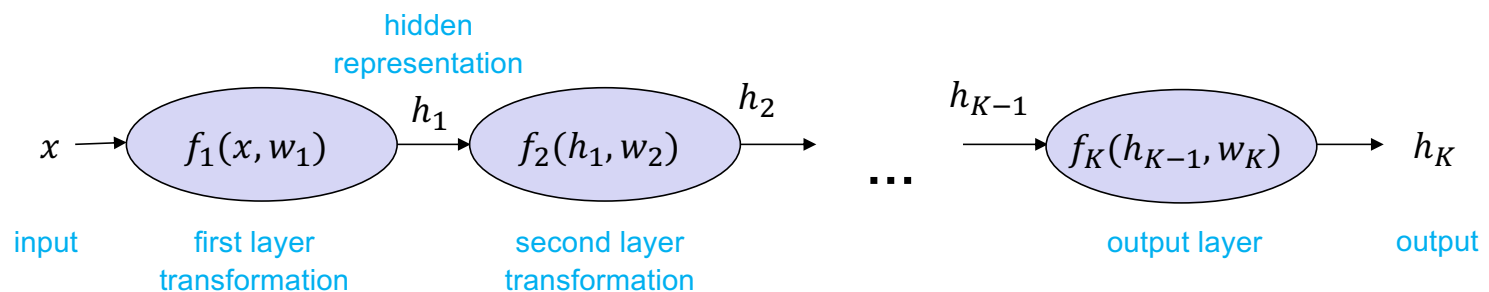
- Computation graphs
- Using the chain rule
- General backpropagation algorithm
- Toy examples of backward pass
- Matrix-vector calculations: ReLU, linear layer

Last time: Multi-layer neural networks

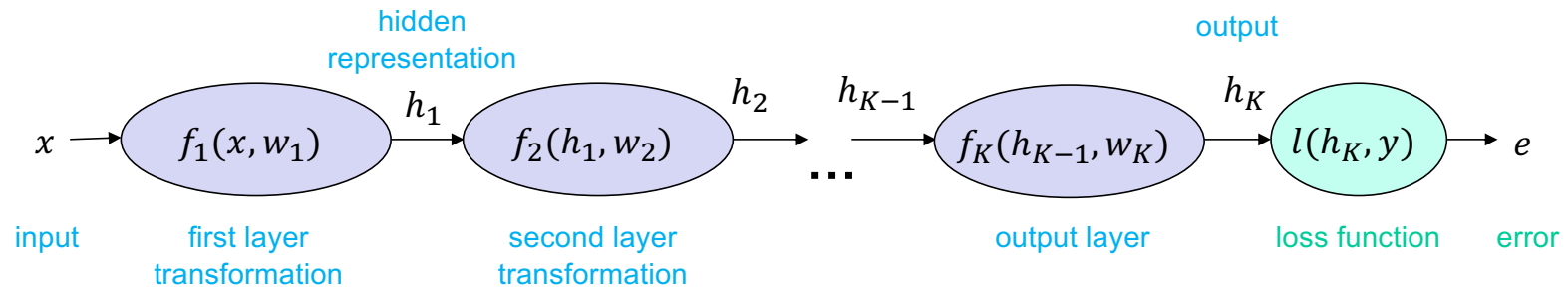
- The function computed by the network is a composition of the functions computed by individual layers (e.g., linear layers and nonlinearities):



- More precisely:



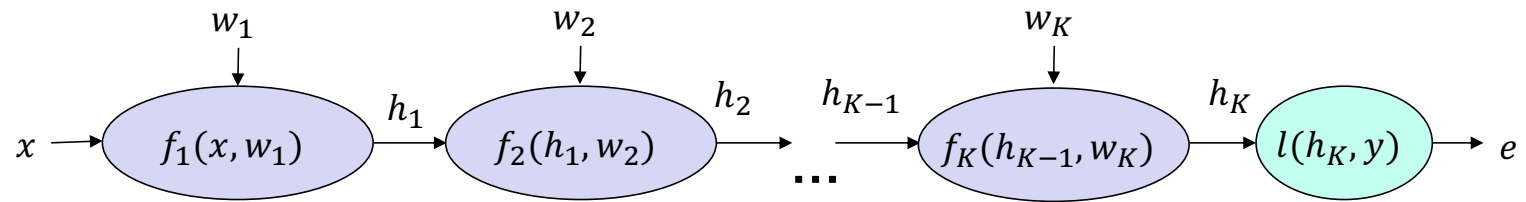
Training a multi-layer network



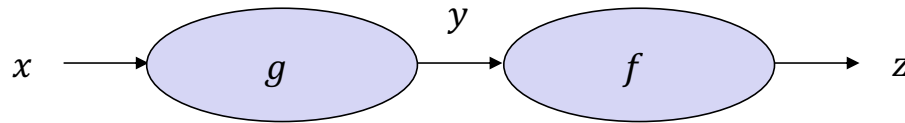
- To train the network, we need to find the **gradient of the error w.r.t. the parameters of each layer**, $\frac{\partial e}{\partial w_k}$, and then apply the SGD update

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

Computation graph



The chain rule



$$\frac{dz}{dx} = \frac{dz}{dy} \frac{dy}{dx}$$

In **calculus**, the **chain rule** is a **formula** that expresses the **derivative** of the **composition** of two **differentiable functions** f and g in terms of the derivatives of f and g . More precisely, if $h = f \circ g$ is the function such that $h(x) = f(g(x))$ for every x , then the chain rule is, in **Lagrange's notation**,

$$h'(x) = f'(g(x))g'(x).$$

or, equivalently,

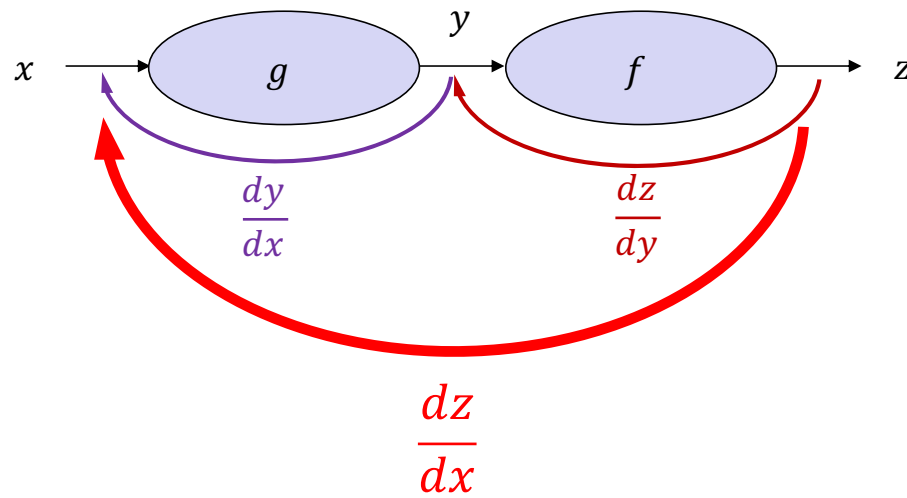
$$h' = (f \circ g)' = (f' \circ g) \cdot g'.$$

The chain rule may also be expressed in **Leibniz's notation**. If a variable z depends on the variable y , which itself depends on the variable x (that is, y and z are **dependent variables**), then z depends on x as well, via the intermediate variable y . In this case, the chain rule is expressed as

$$\frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx}$$

https://en.wikipedia.org/wiki/Chain_rule

The chain rule



$$\frac{dz}{dx} = \frac{dz}{dy} \frac{dy}{dx}$$

Applying the chain rule

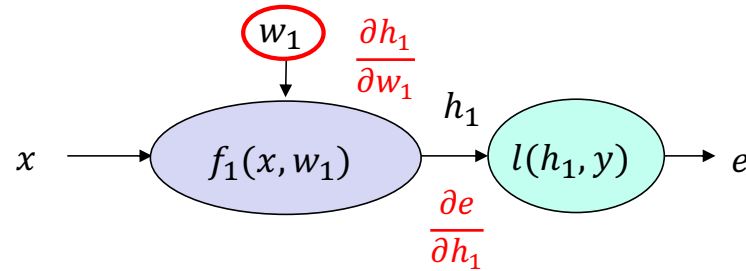
Let's start with $K = 1$

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

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

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

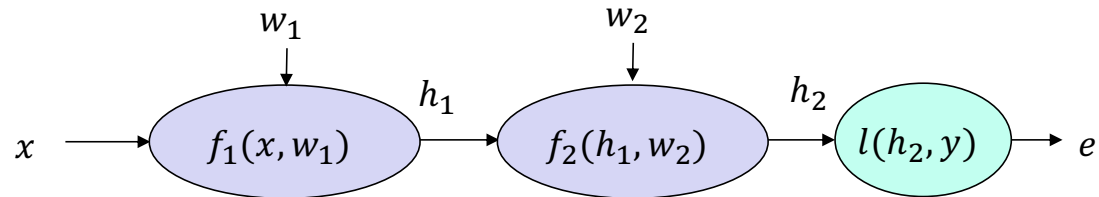
$$e = l(h_1, y) = (y - h_1)^2$$



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

Applying the chain rule

$$K = 2$$



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

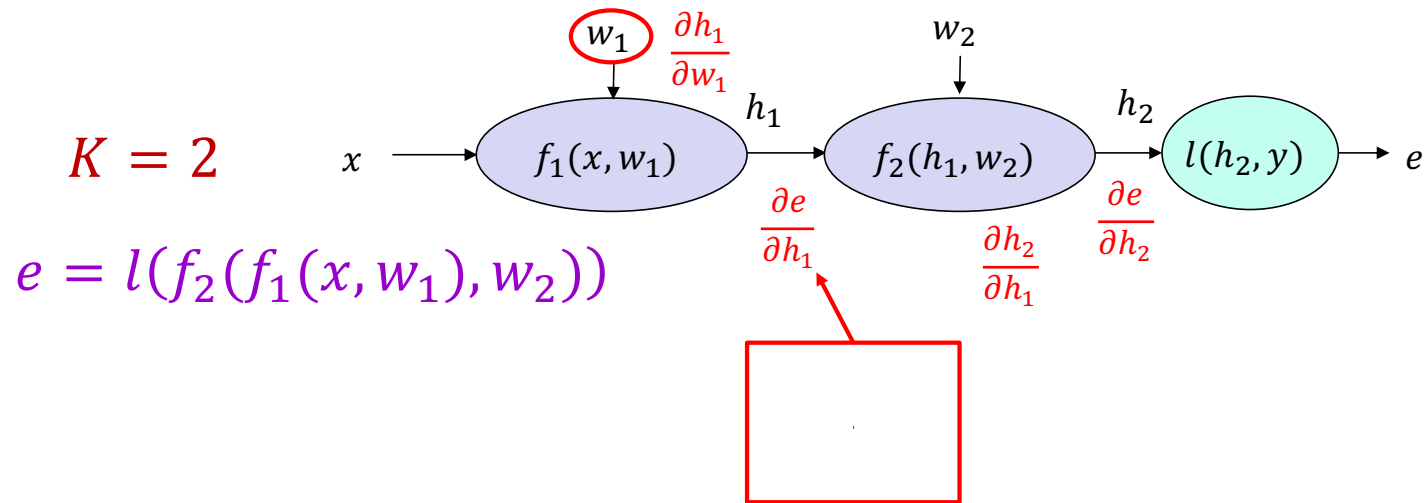
Example: $e = -\log(\sigma(w_1^T x))$ (assume $y = 1$)

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

$$h_2 = f_2(h_1) = \sigma(h_1)$$

$$e = l(h_2, 1) = -\log(h_2)$$

Applying the chain rule



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

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

$$h_2 = f_2(h_1) = \sigma(h_1)$$

$$e = l(h_2, 1) = -\log(h_2)$$

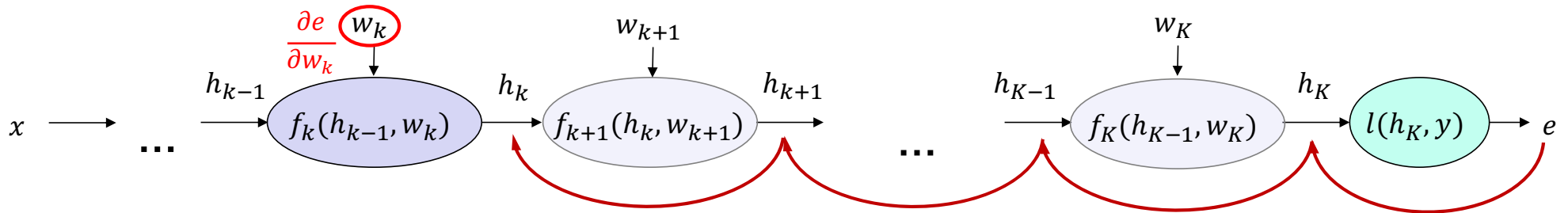
$$\frac{\partial h_1}{\partial w_1} =$$

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

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

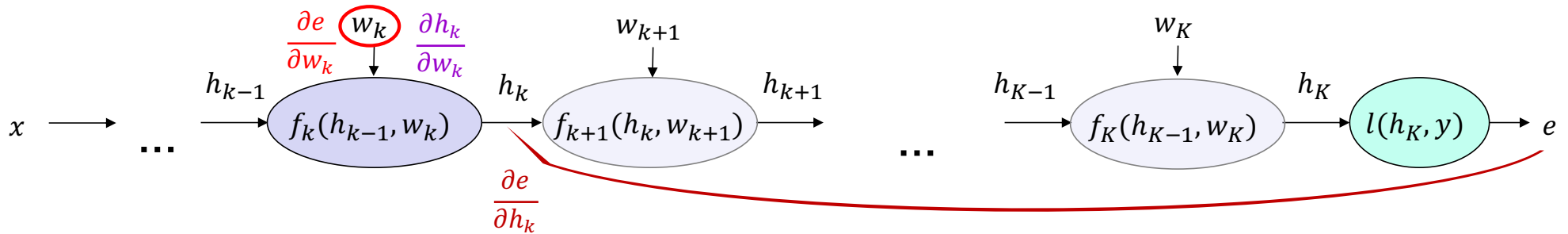
$$\frac{\partial e}{\partial w_1} = \frac{\partial e}{\partial h_2} \frac{\partial h_2}{\partial h_1} \frac{\partial h_1}{\partial w_1} =$$

Chain rule: General case



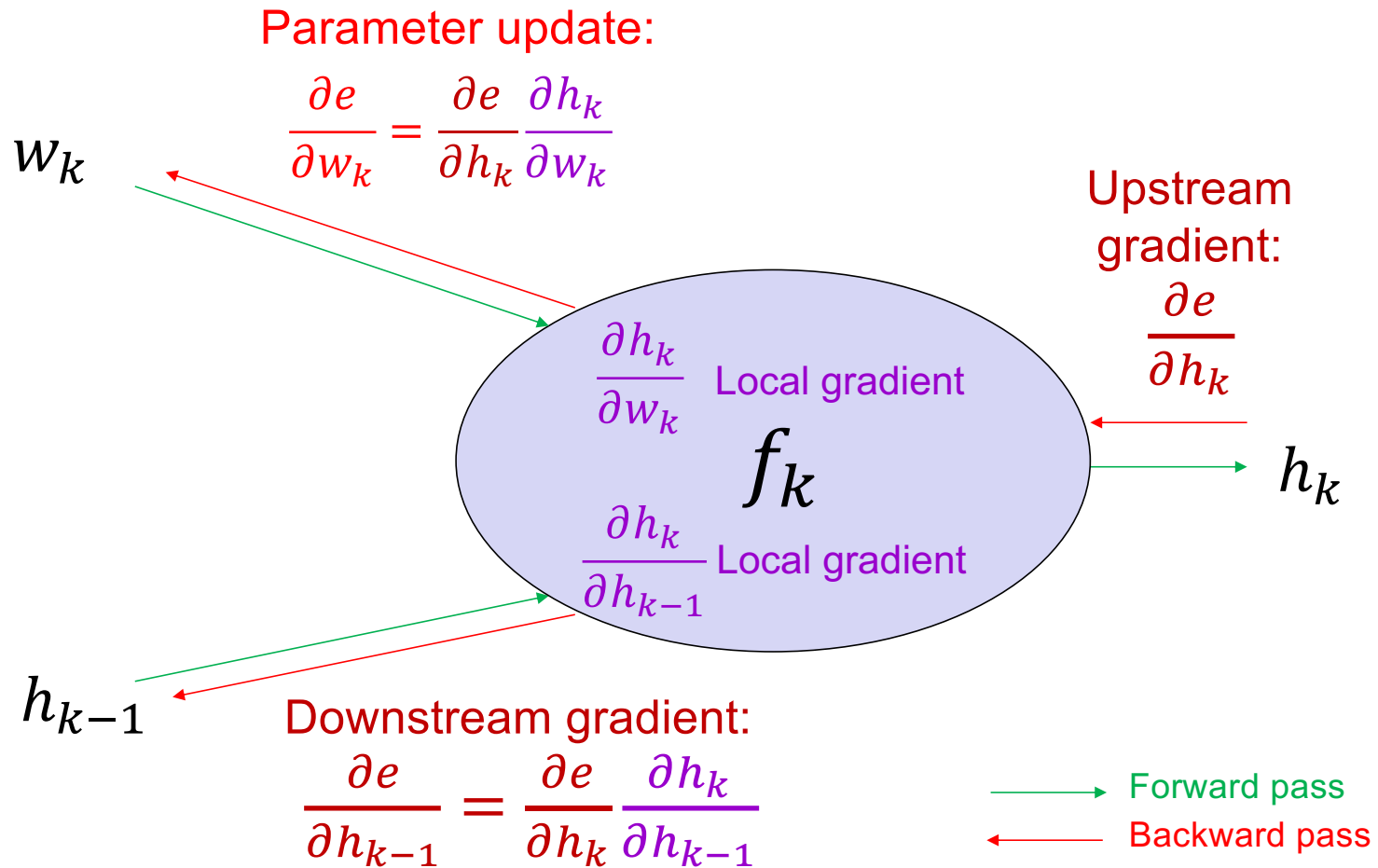
$$\frac{\partial e}{\partial w_k} = \frac{\partial e}{\partial h_K} \frac{\partial h_K}{\partial h_{K-1}} \dots \frac{\partial h_{k+1}}{\partial h_k}$$

Chain rule: General case



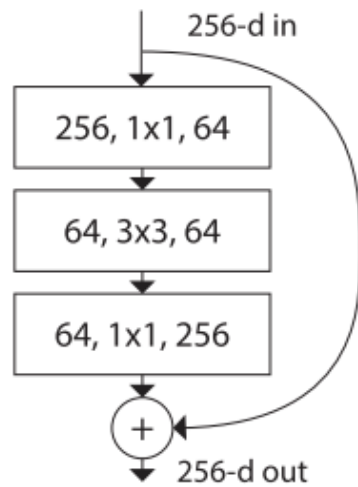
$$\frac{\partial e}{\partial w_k} = \underbrace{\frac{\partial e}{\partial h_K} \frac{\partial h_K}{\partial h_{K-1}} \dots \frac{\partial h_{k+1}}{\partial h_k}}_{\text{Upstream gradient } \frac{\partial e}{\partial h_k}} \underbrace{\frac{\partial h_k}{\partial w_k}}_{\text{Local gradient}}$$

Backpropagation summary

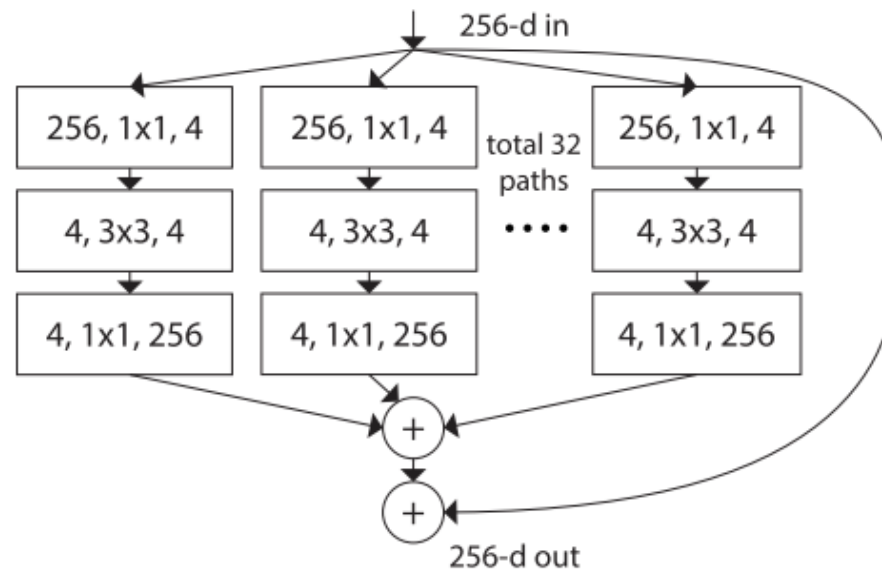


What about more general computation graphs?

ResNet

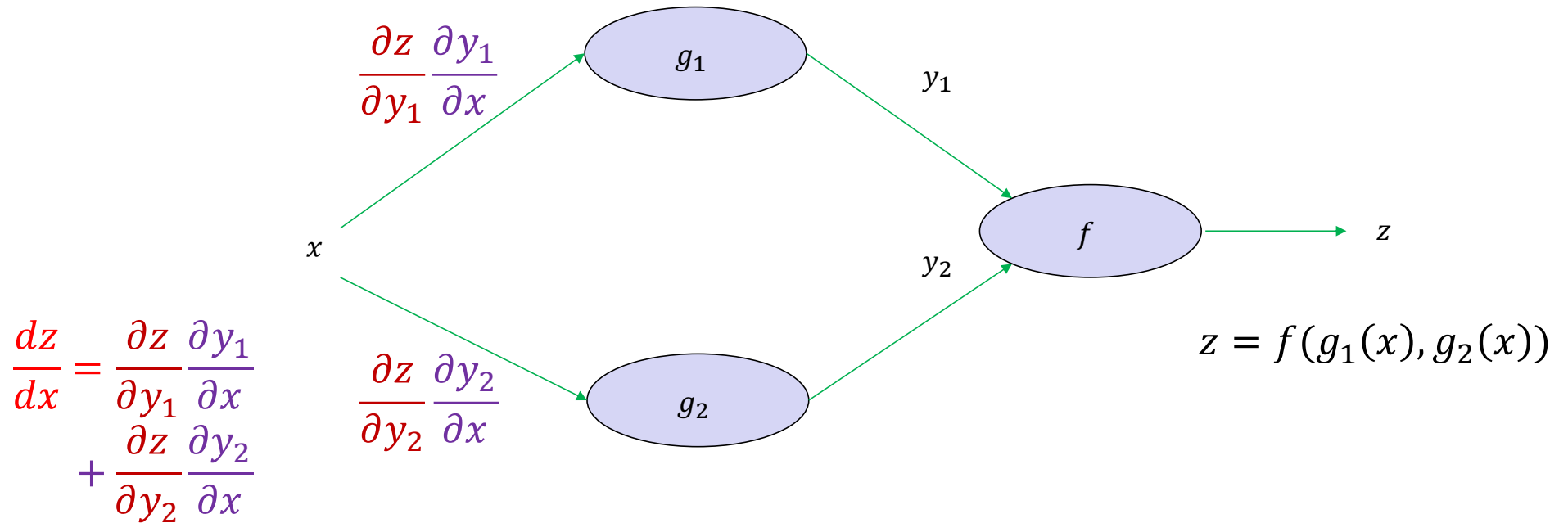


ResNeXt



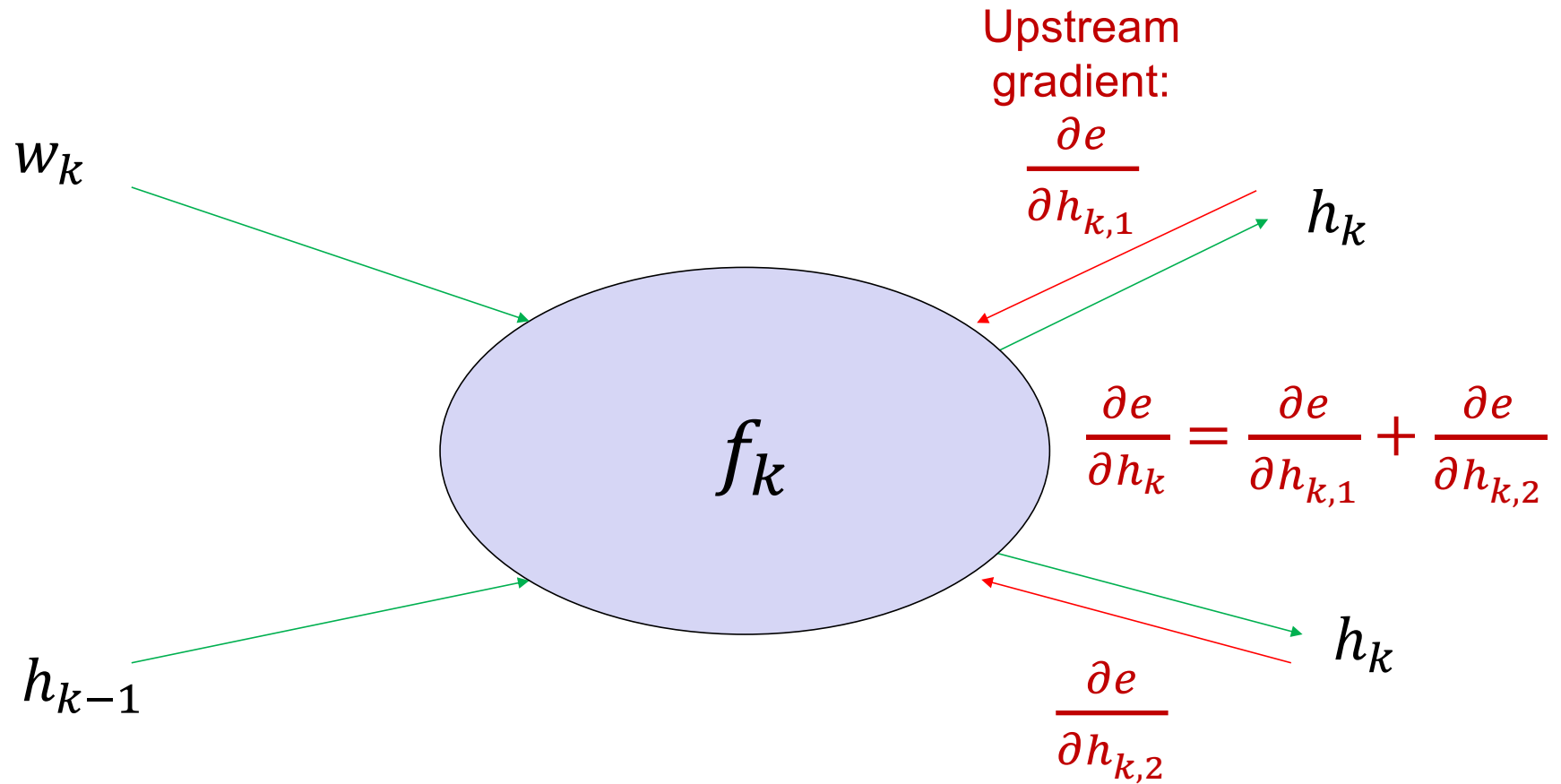
[Figure source](#)

The chain rule: Multiple paths



https://en.wikipedia.org/wiki/Chain_rule

The chain rule: Multiple paths

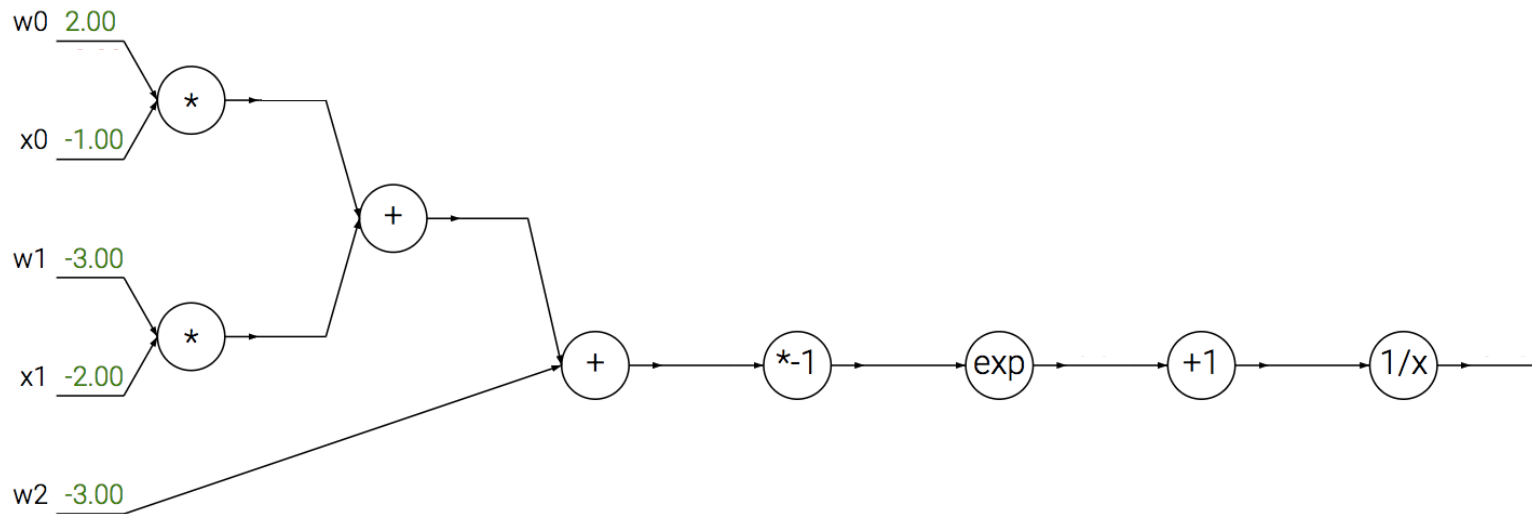


Overview

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

A detailed example

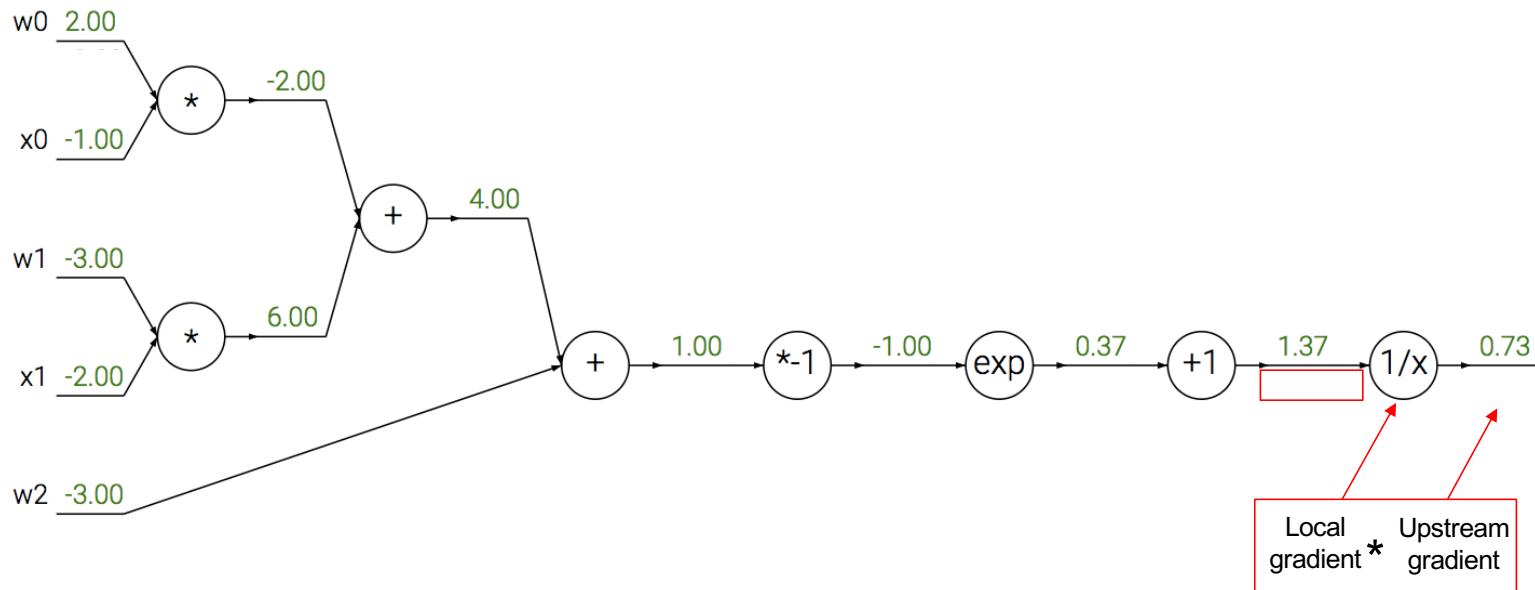
$$f(x, w) = \frac{1}{1 + \exp[-(w^{(0)}x^{(0)} + w^{(1)}x^{(1)} + w^{(2)})]}$$



Source: [Stanford 231n](#)

A detailed example

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



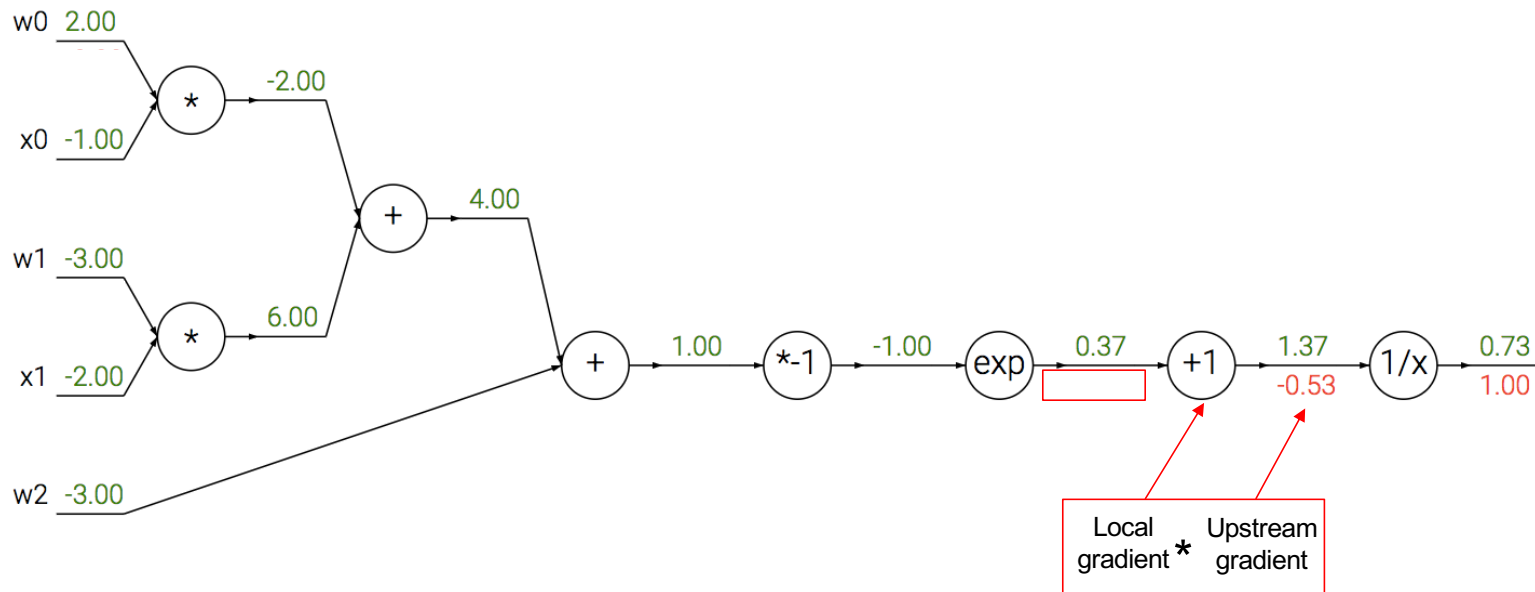
$$(1/x)' = -1/x^2$$

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

Source: [Stanford 231n](#)

A detailed example

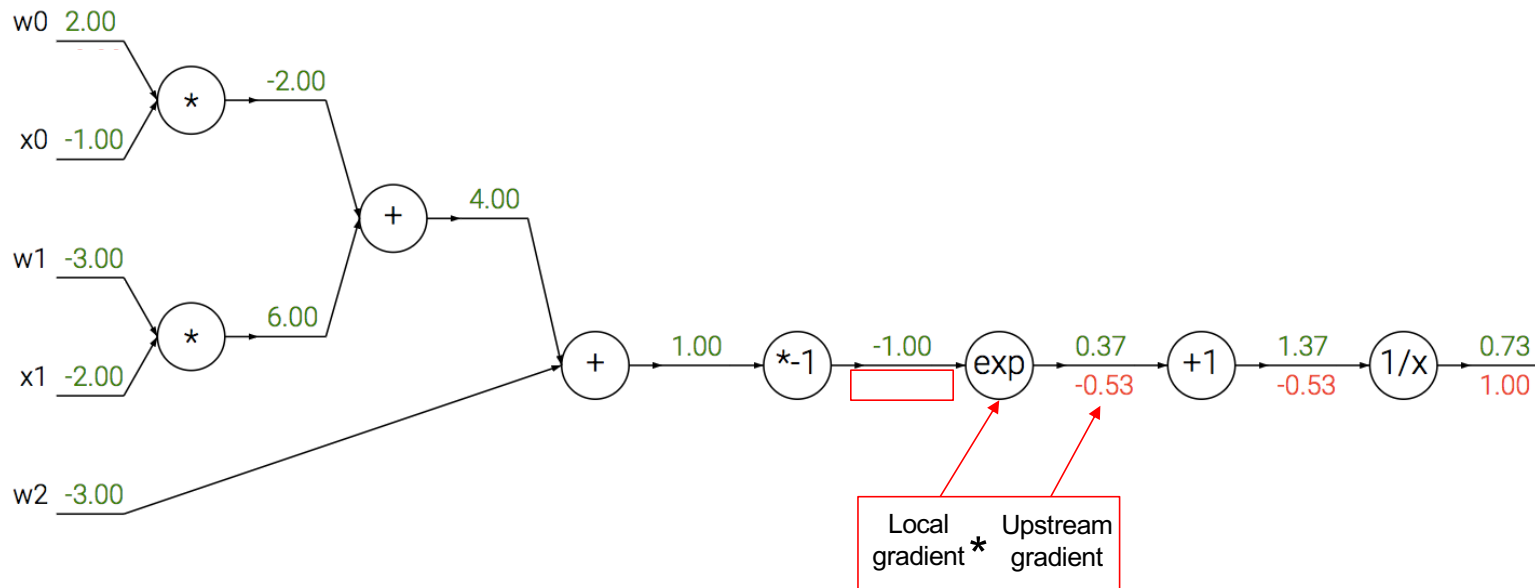
$$f(x, w) = \frac{1}{1 + \exp[-(w^{(0)}x^{(0)} + w^{(1)}x^{(1)} + w^{(2)})]}$$



Source: [Stanford 231n](#)

A detailed example

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

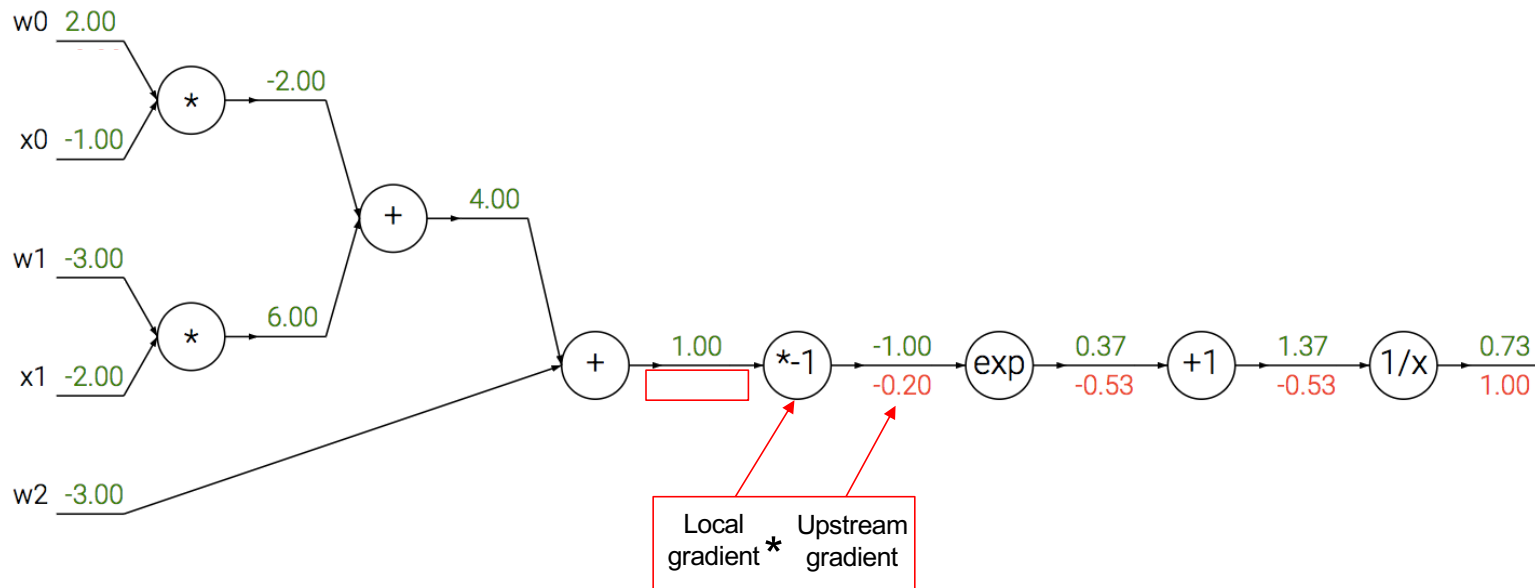


$$\exp(-1) * (-0.53) = -0.20$$

Source: [Stanford 231n](#)

A detailed example

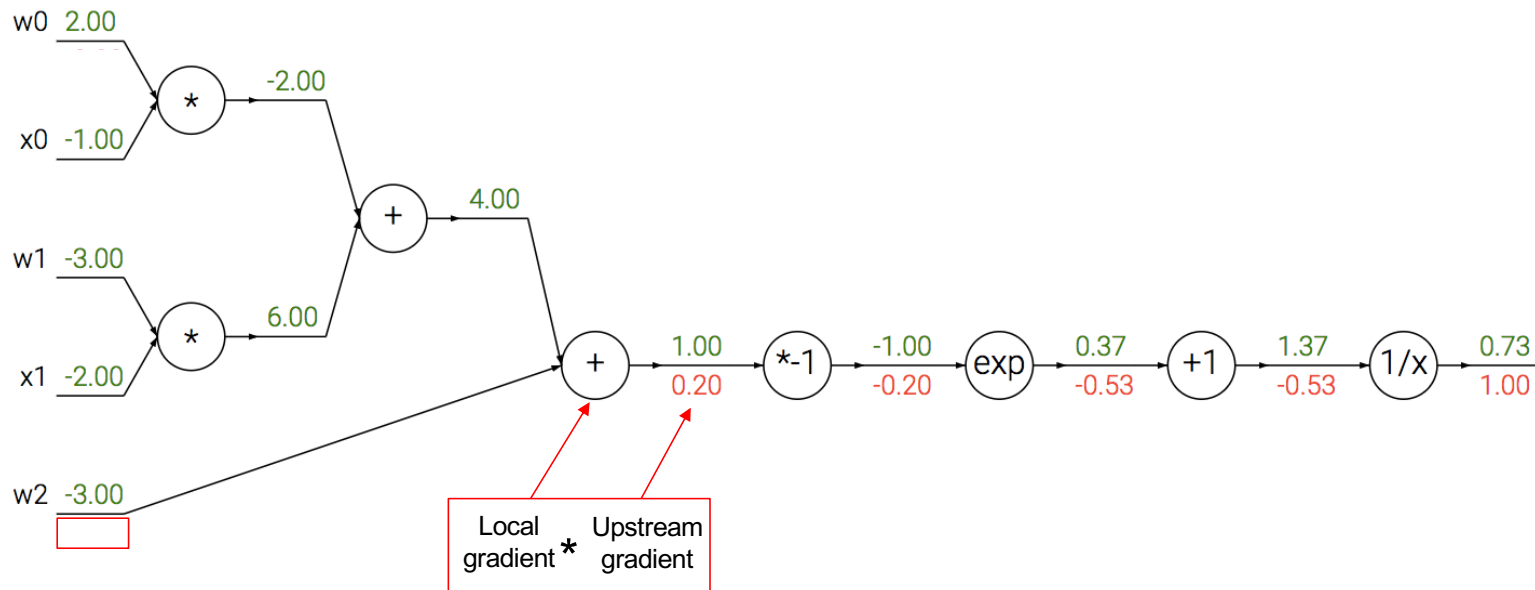
$$f(x, w) = \frac{1}{1 + \exp[-(w^{(0)}x^{(0)} + w^{(1)}x^{(1)} + w^{(2)})]}$$



Source: [Stanford 231n](#)

A detailed example

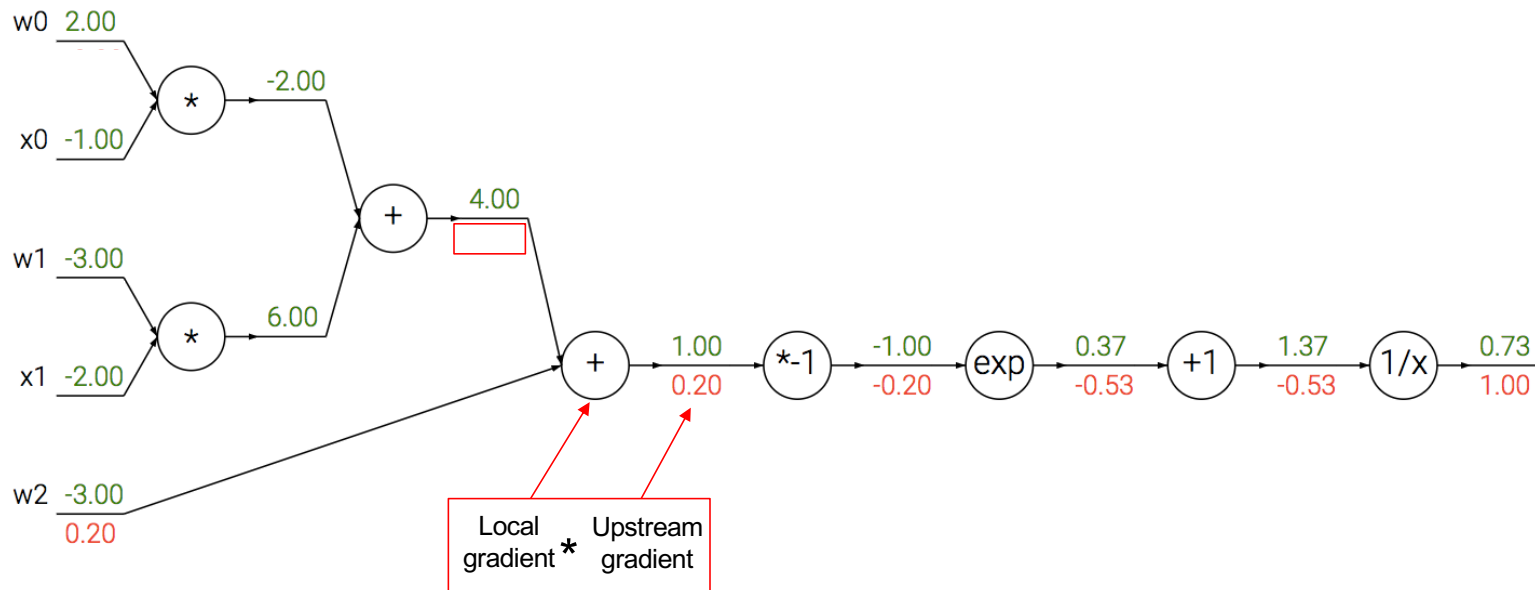
$$f(x, w) = \frac{1}{1 + \exp[-(w^{(0)}x^{(0)} + w^{(1)}x^{(1)} + w^{(2)})]}$$



Source: [Stanford 231n](#)

A detailed example

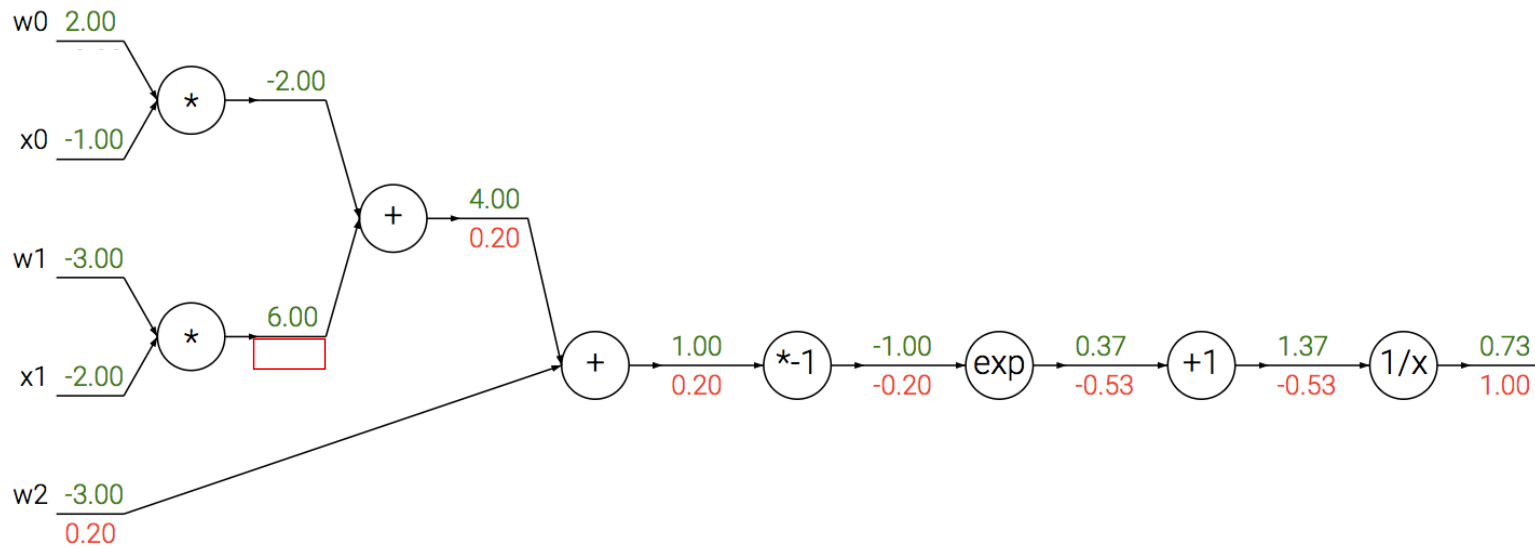
$$f(x, w) = \frac{1}{1 + \exp[-(w^{(0)}x^{(0)} + w^{(1)}x^{(1)} + w^{(2)})]}$$



Source: [Stanford 231n](#)

A detailed example

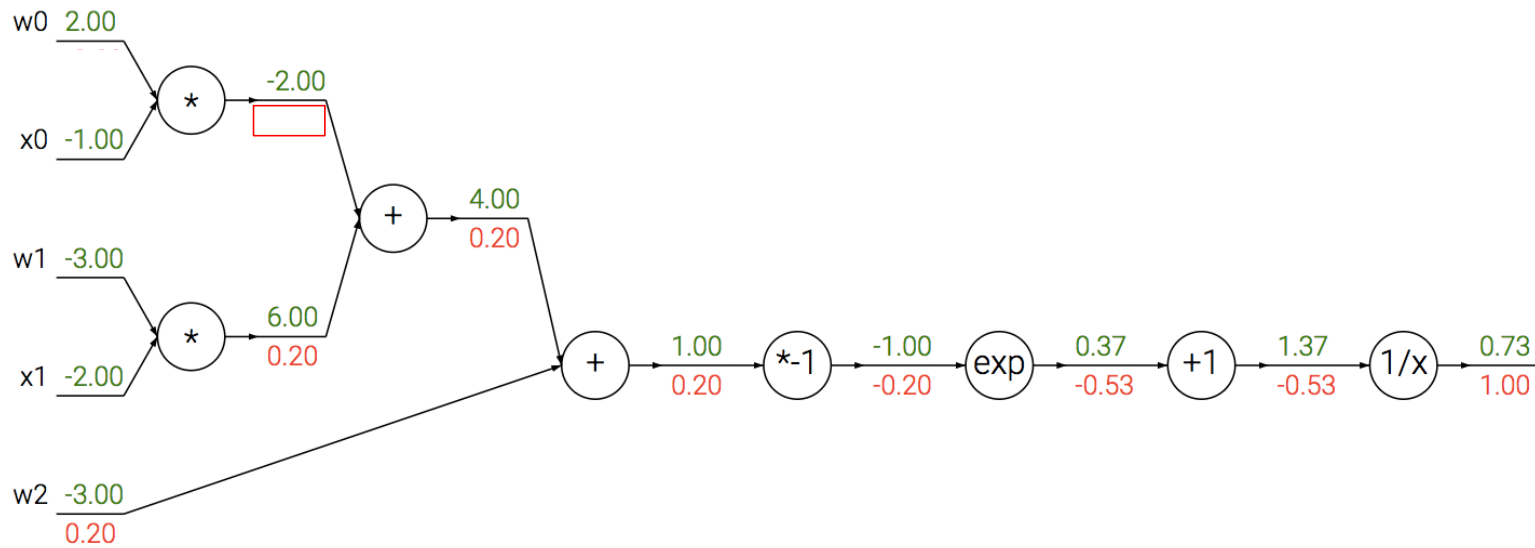
$$f(x, w) = \frac{1}{1 + \exp[-(w^{(0)}x^{(0)} + w^{(1)}x^{(1)} + w^{(2)})]}$$



Source: [Stanford 231n](#)

A detailed example

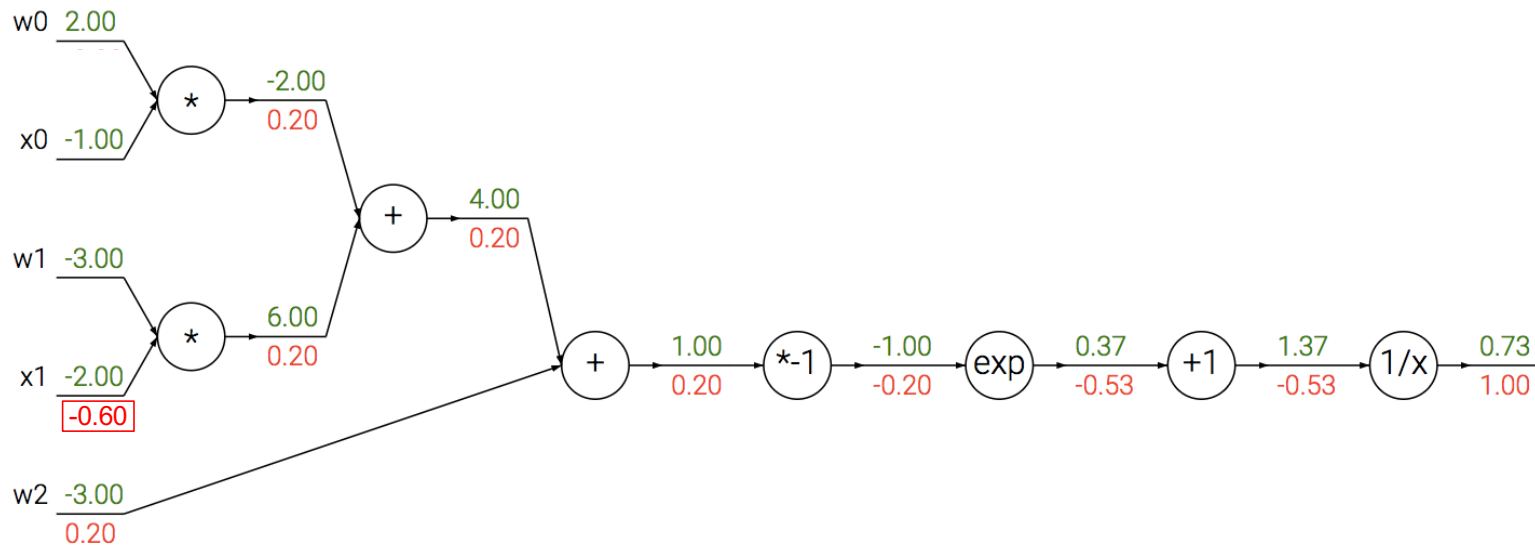
$$f(x, w) = \frac{1}{1 + \exp[-(w^{(0)}x^{(0)} + w^{(1)}x^{(1)} + w^{(2)})]}$$



Source: [Stanford 231n](#)

A detailed example

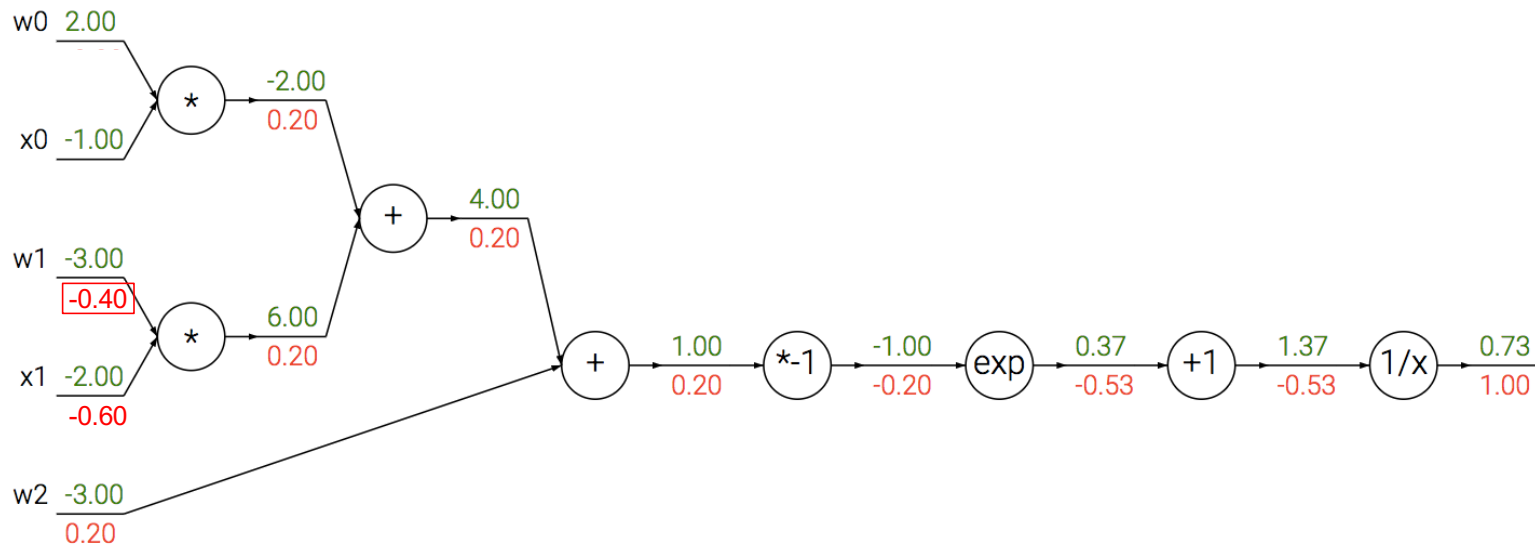
$$f(x, w) = \frac{1}{1 + \exp[-(w^{(0)}x^{(0)} + w^{(1)}x^{(1)} + w^{(2)})]}$$



Source: [Stanford 231n](#)

A detailed example

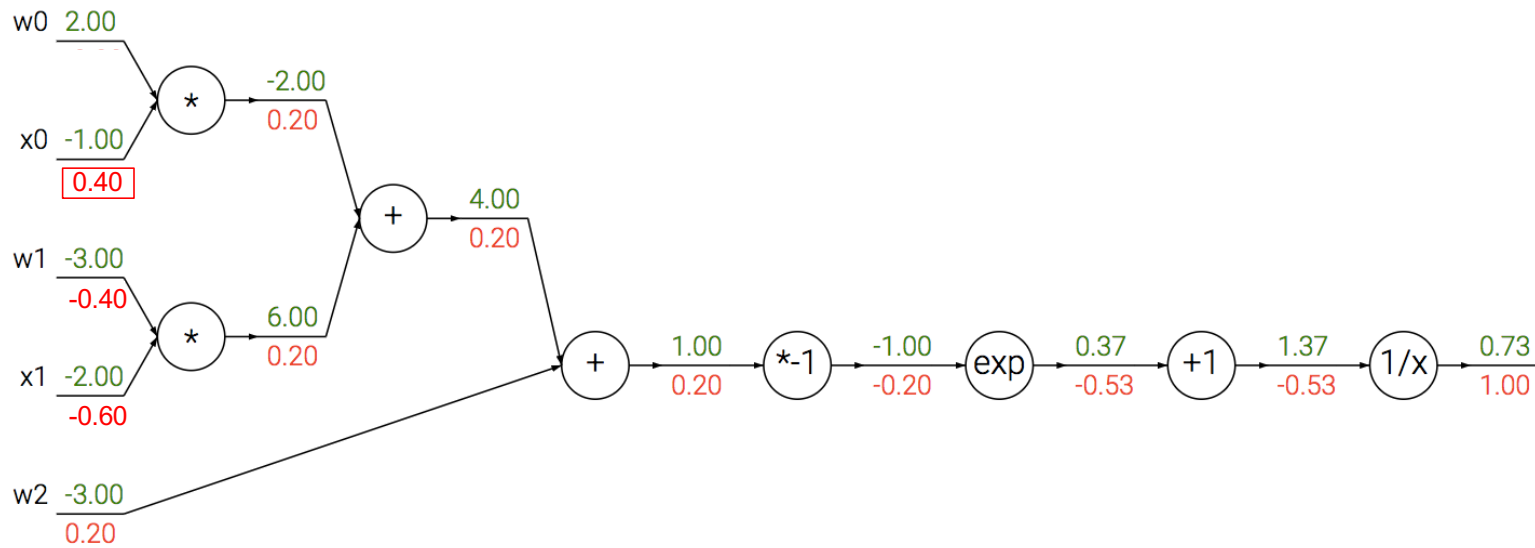
$$f(x, w) = \frac{1}{1 + \exp[-(w^{(0)}x^{(0)} + w^{(1)}x^{(1)} + w^{(2)})]}$$



Source: [Stanford 231n](#)

A detailed example

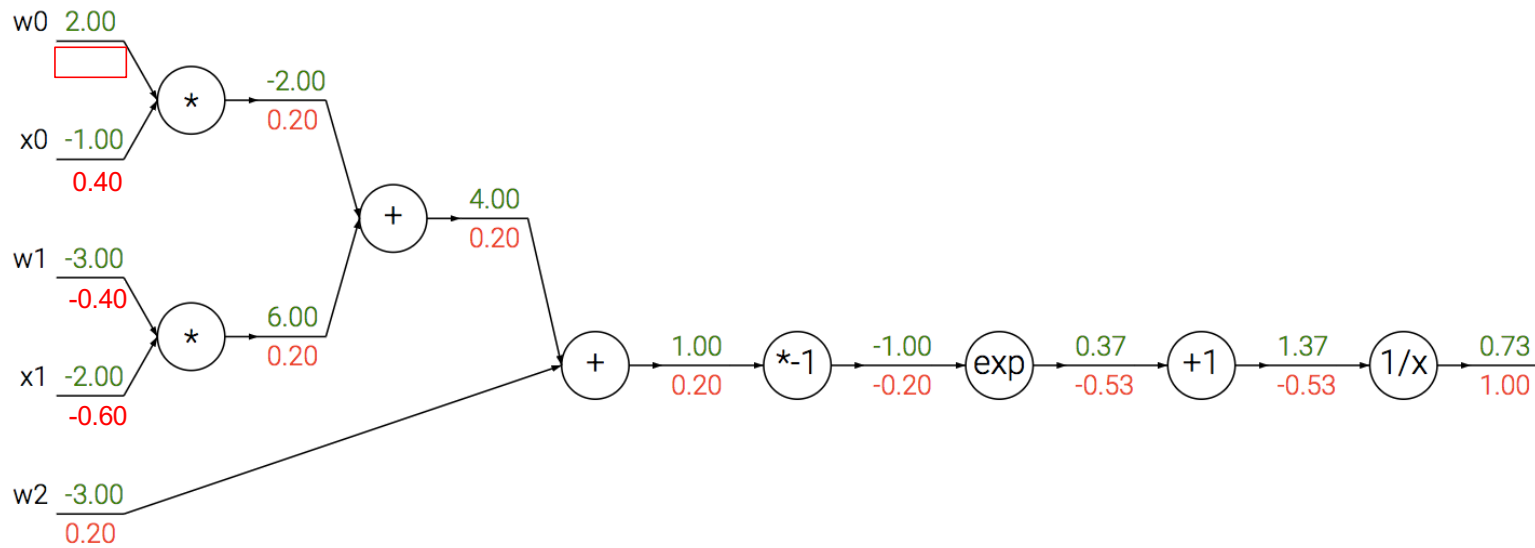
$$f(x, w) = \frac{1}{1 + \exp[-(w^{(0)}x^{(0)} + w^{(1)}x^{(1)} + w^{(2)})]}$$



Source: [Stanford 231n](#)

A detailed example

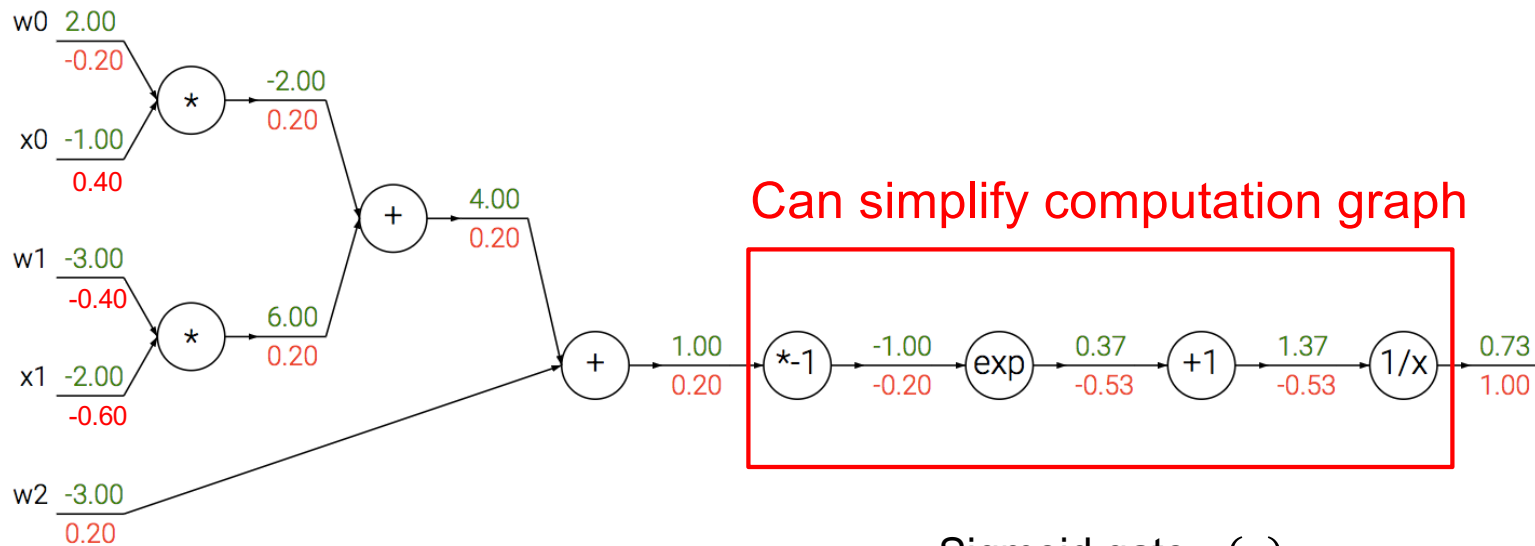
$$f(x, w) = \frac{1}{1 + \exp[-(w^{(0)}x^{(0)} + w^{(1)}x^{(1)} + w^{(2)})]}$$



Source: [Stanford 231n](#)

A detailed example

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



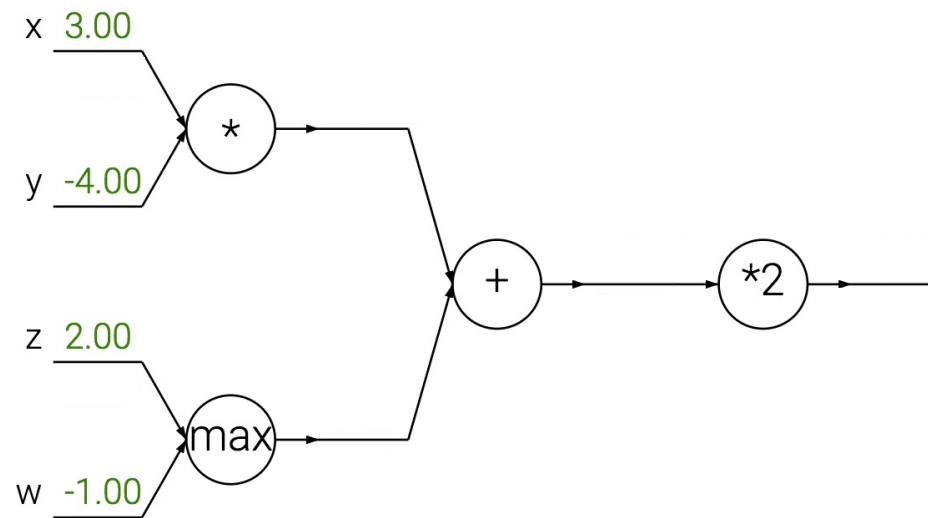
Sigmoid gate $\sigma(x)$

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

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

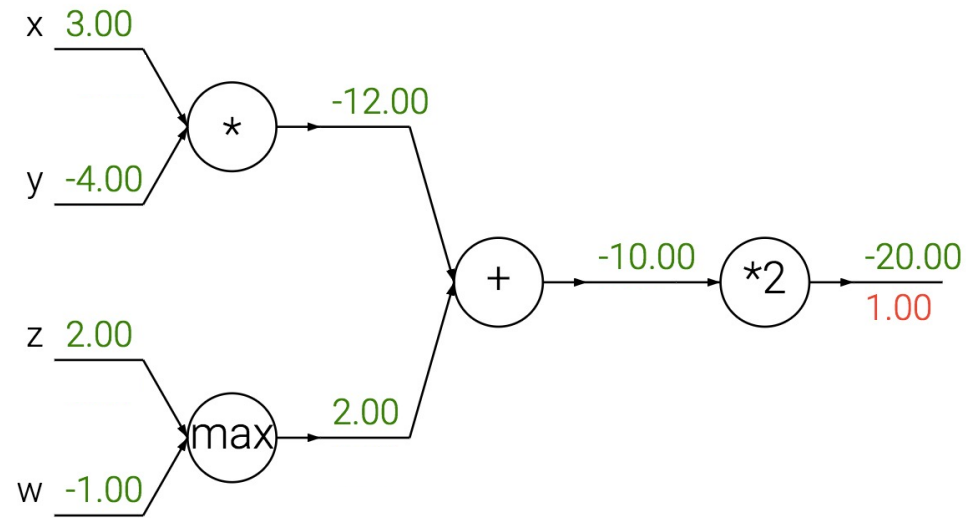
Source: [Stanford 231n](#)

Example 2



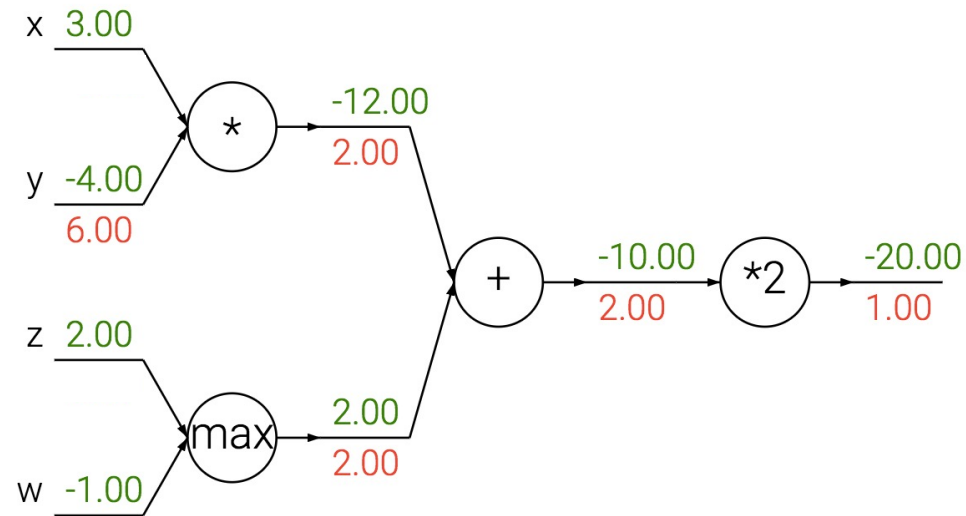
Source: [Stanford 231n](#)

Example 2



Add gate: “gradient distributor”

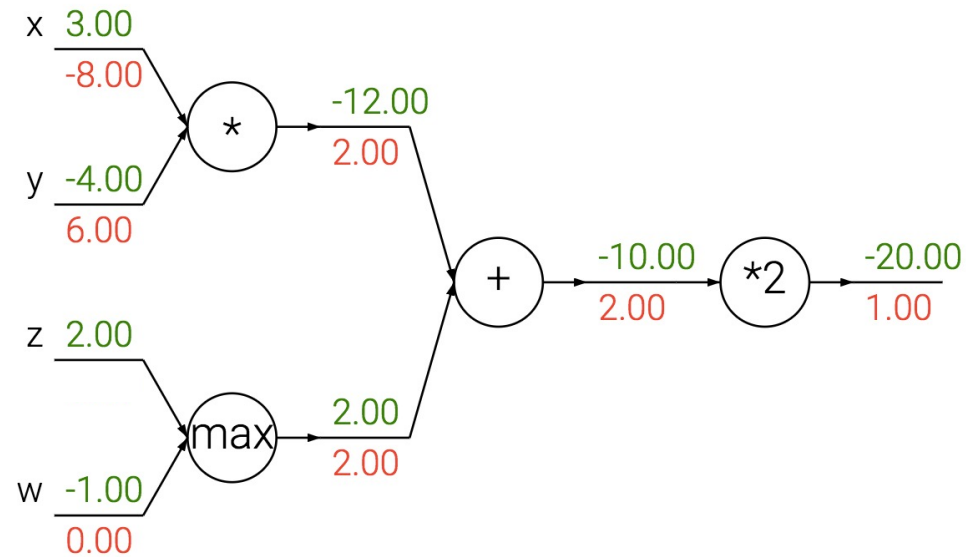
Example 2



Add gate: “gradient distributor”

Multiply gate: “gradient switcher”

Example 2



Add gate: “gradient distributor”

Multiply gate: “gradient switcher”

Max gate: “gradient router”