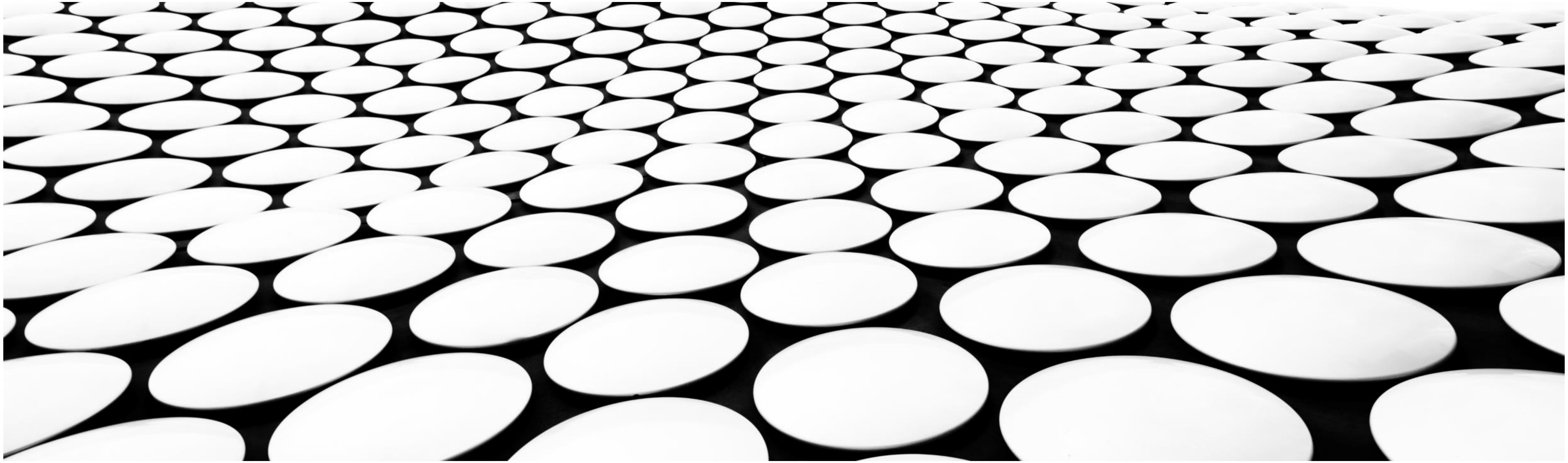
深度学习

邱怡轩



今天的主题

- 反向传播算法
- PyTorch 自动微分

反向传播算法

反向传播算法

- 反向传播算法（Backpropagation，BP）
- 并不神秘
- 本质上是一种聪明、高效地求导数的办法

反向传播算法

- BP的地位比较尴尬
- 首先它确实很重要，甚至一度推动了人工智能的发展进程
- 但如今几乎已经被更通用的自动微分取代
- 现实中你以99.9%的可能不需要自己实现

反向传播算法

- 为什么还要学习BP?

- 魔鬼在细节中
- 理解工作机制
- 指导如何调参
- 辅助Debugging

反向传播算法

- 为什么还要学习BP?
- 魔鬼在细节中
- 理解工作机制
- 指导如何调参
- 辅助Debugging
- 顺便学习了一些其他有用的知识

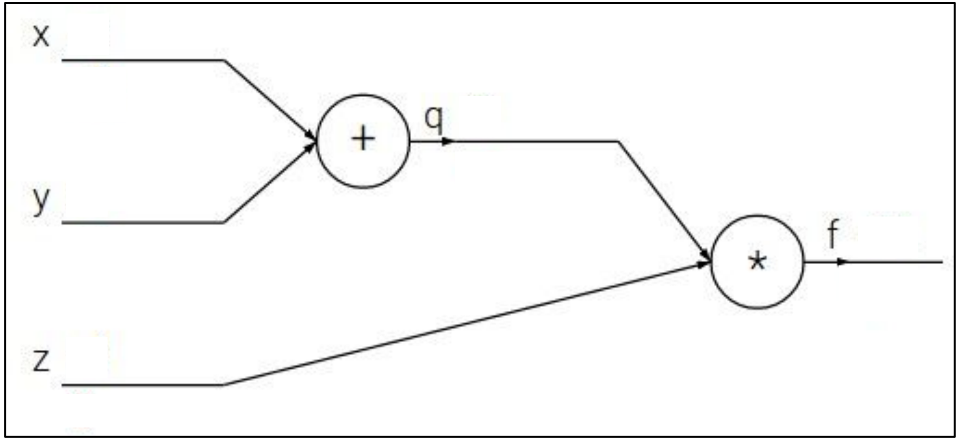
以下内容来自
http://cs231n.stanford.edu

Backpropagation: a simple example

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

Backpropagation: a simple example

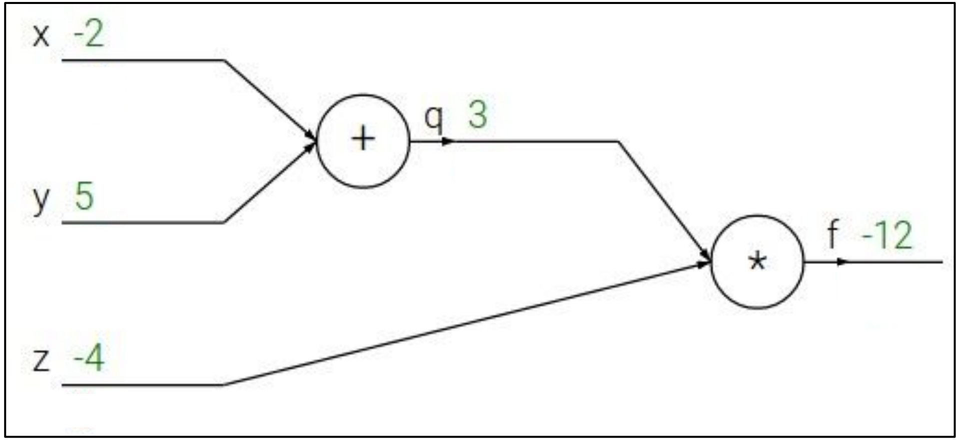
$$f(x, y, z) = (x + y)z$$



Backpropagation: a simple example

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

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

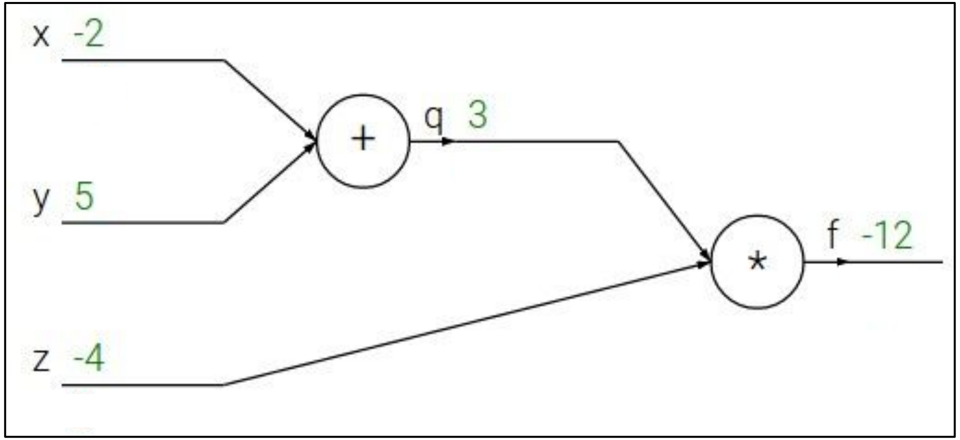


Backpropagation: a simple example

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

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

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



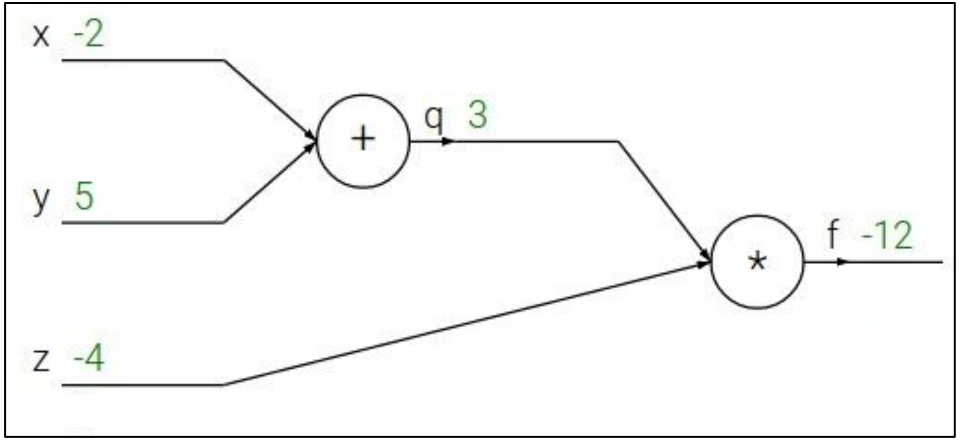
Backpropagation: a simple example

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

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

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

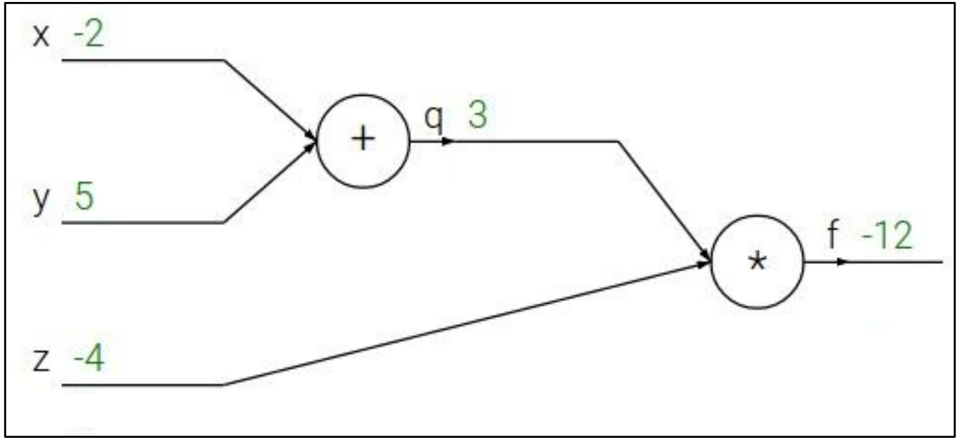
$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$



Backpropagation: a simple example

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

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

Backpropagation: a simple example

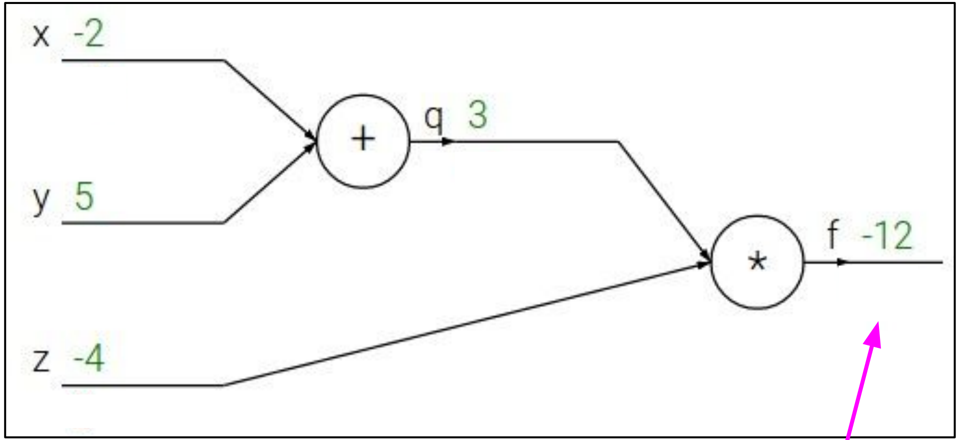
$$f(x,y,z) = (x+y)z$$

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



$\frac{\partial f}{\partial f}$

Backpropagation: a simple example

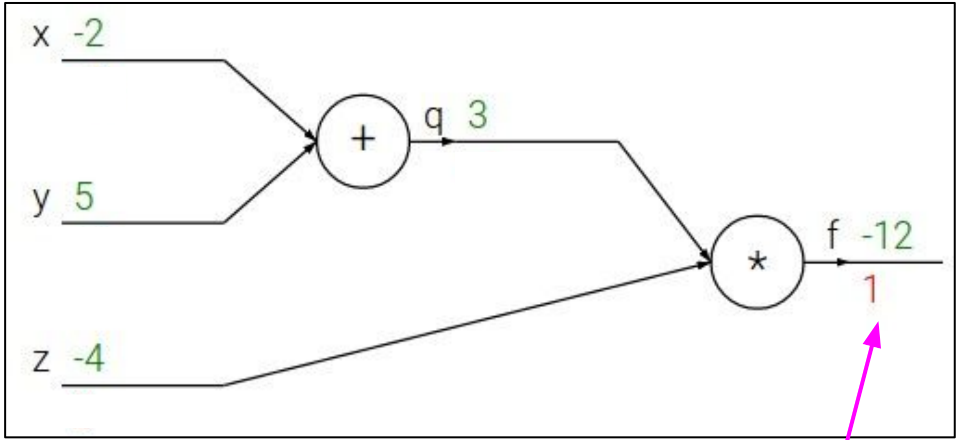
$$f(x, y, z) = (x + y)z$$

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



$$\frac{\partial f}{\partial f}$$

Backpropagation: a simple example

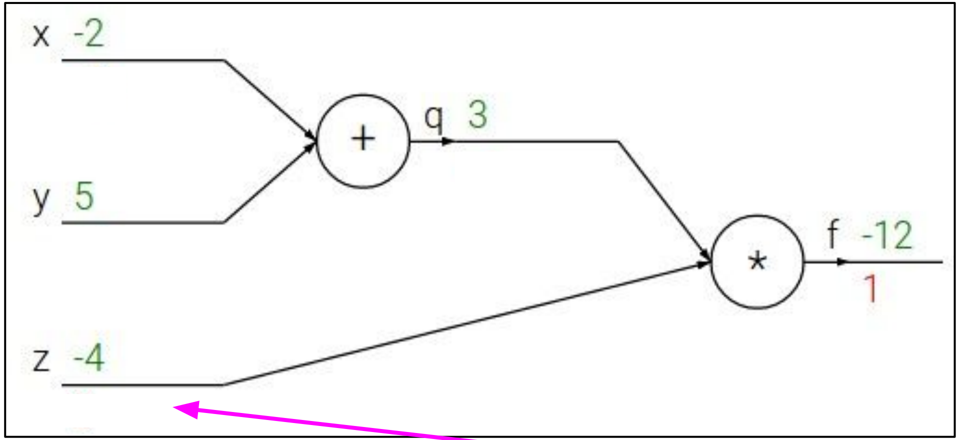
$$f(x,y,z) = (x+y)z$$

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



$\frac{\partial f}{\partial z}$

Backpropagation: a simple example

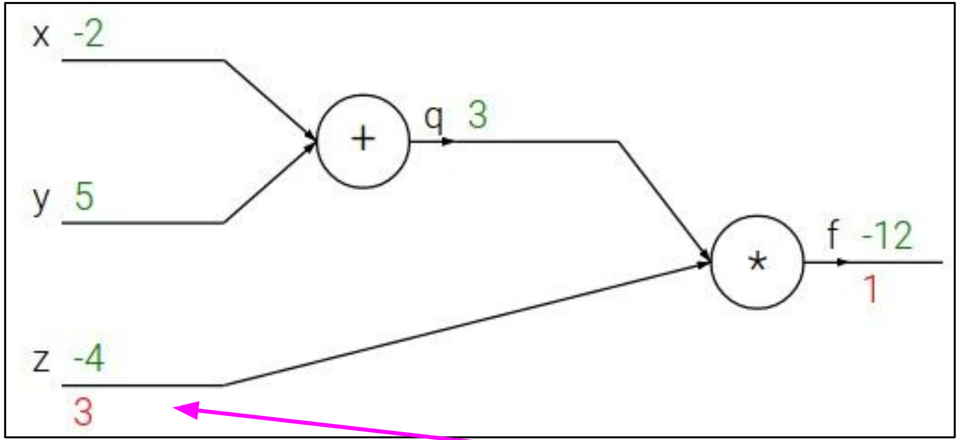
$$f(x, y, z) = (x + y)z$$

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



$$\frac{\partial f}{\partial z}$$

Backpropagation: a simple example

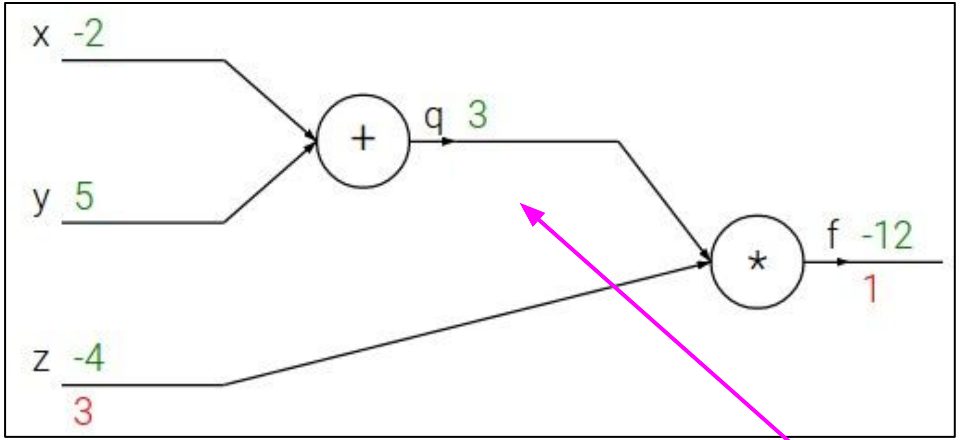
$$f(x, y, z) = (x + y)z$$

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



$$\frac{\partial f}{\partial q}$$

Backpropagation: a simple example

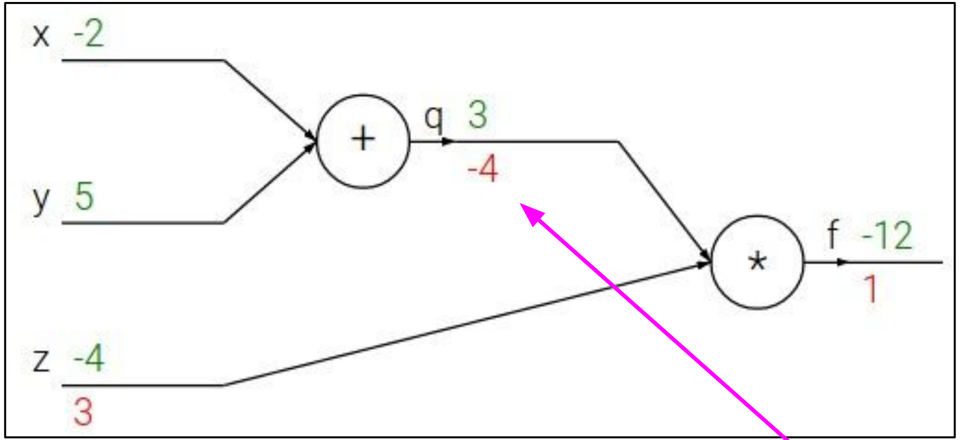
$$f(x, y, z) = (x + y)z$$

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



$$\frac{\partial f}{\partial q}$$

Backpropagation: a simple example

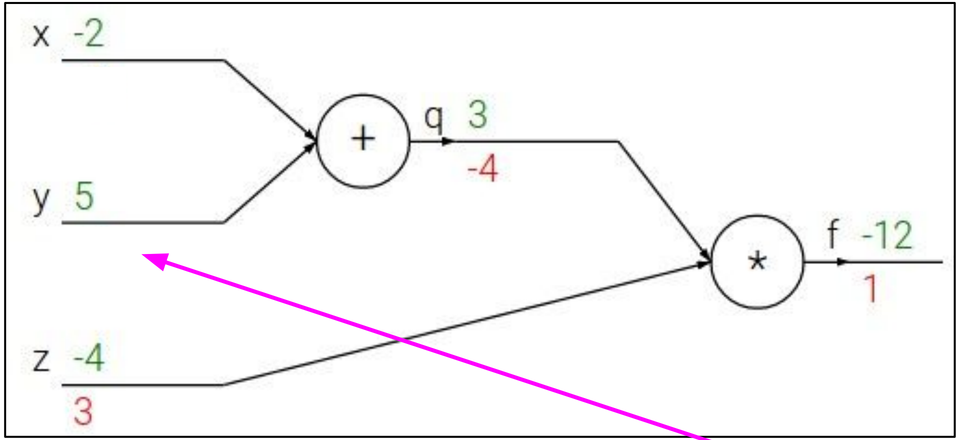
$$f(x, y, z) = (x + y)z$$

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



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

Chain rule:

$$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial q} \frac{\partial q}{\partial y}$$

Upstream gradient

Local gradient

Backpropagation: a simple example

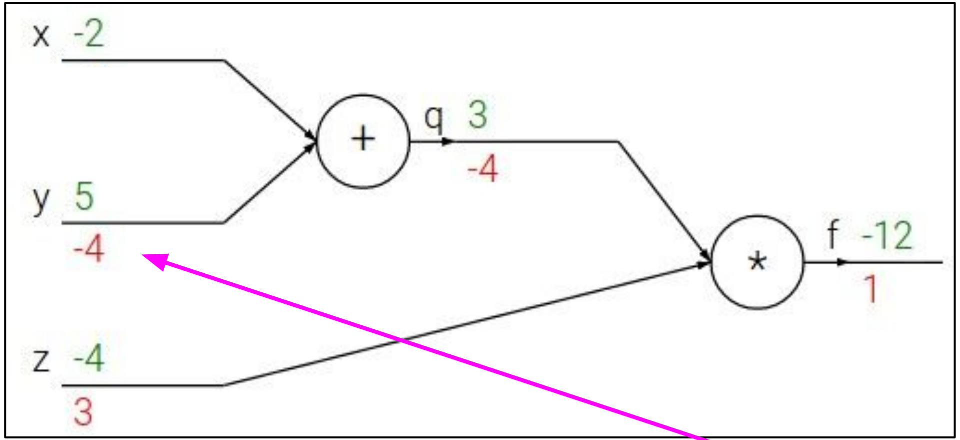
$$f(x, y, z) = (x + y)z$$

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



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

Chain rule:

$$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial q} \frac{\partial q}{\partial y}$$

Upstream gradient

Local gradient

Backpropagation: a simple example

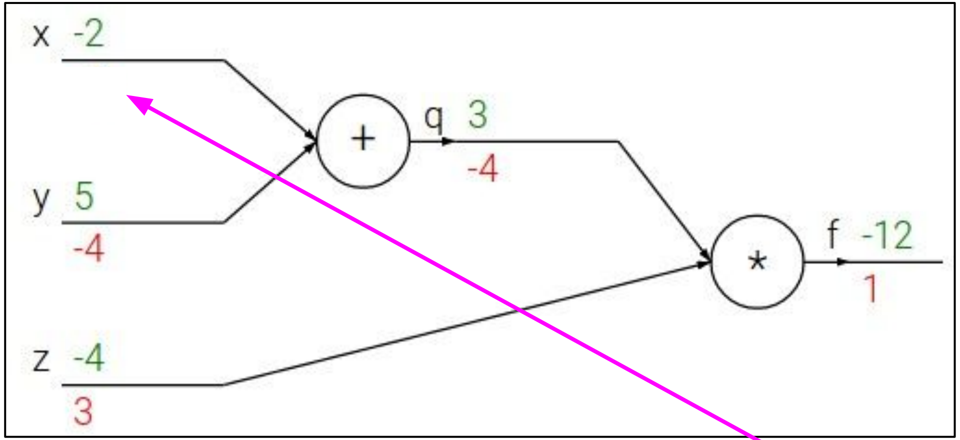
$$f(x,y,z) = (x+y)z$$

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



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

Chain rule:

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial q}\frac{\partial q}{\partial x}$$

Upstream gradient

Local gradient

Backpropagation: a simple example

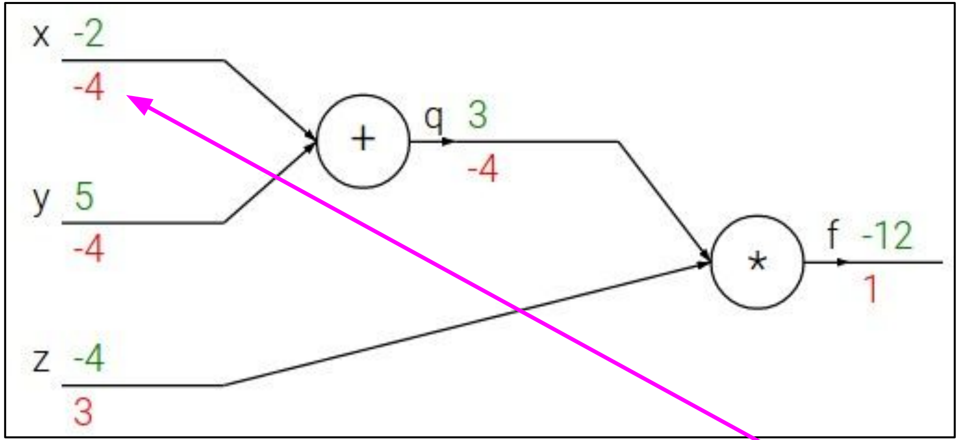
$$f(x, y, z) = (x + y)z$$

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



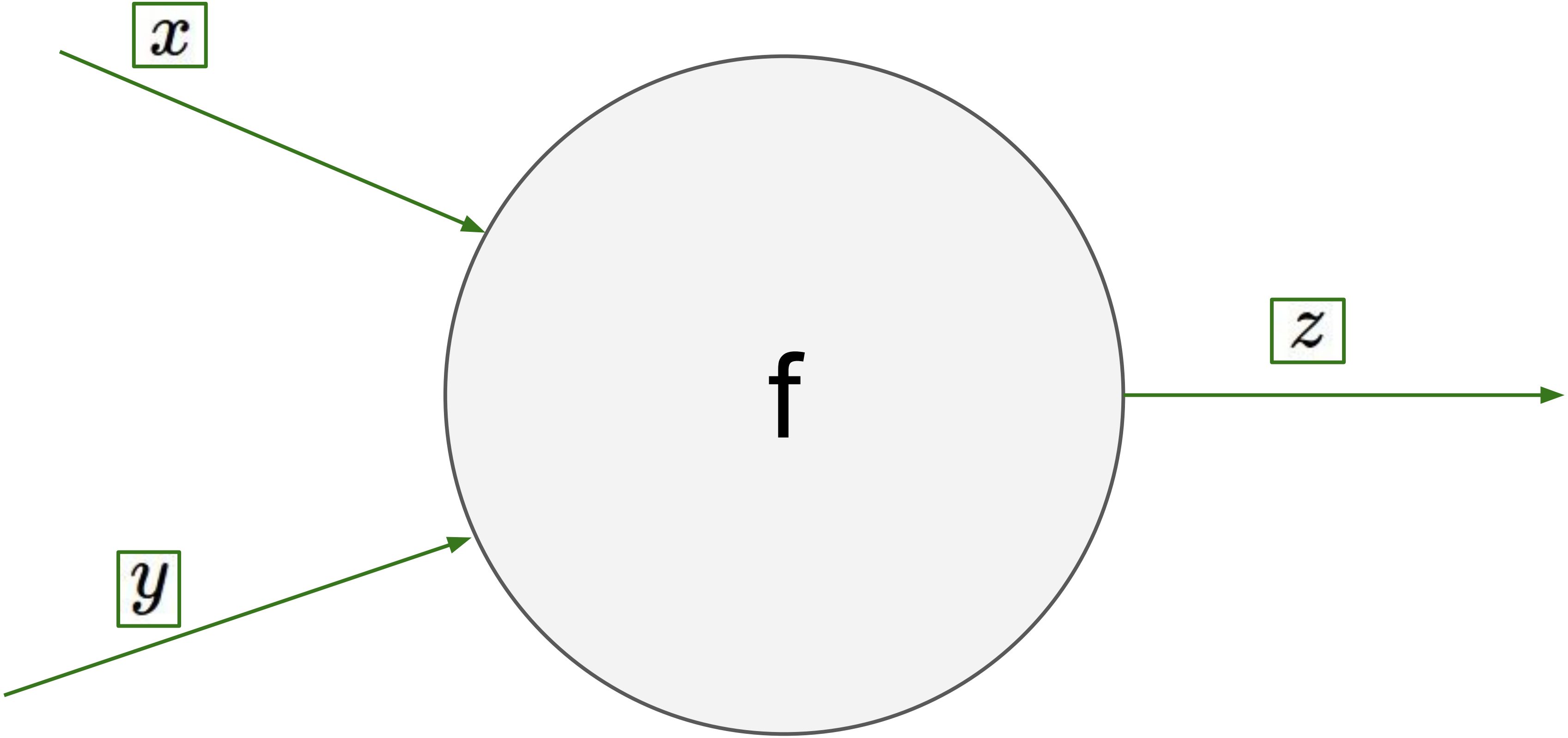
$$\frac{\partial f}{\partial x}$$

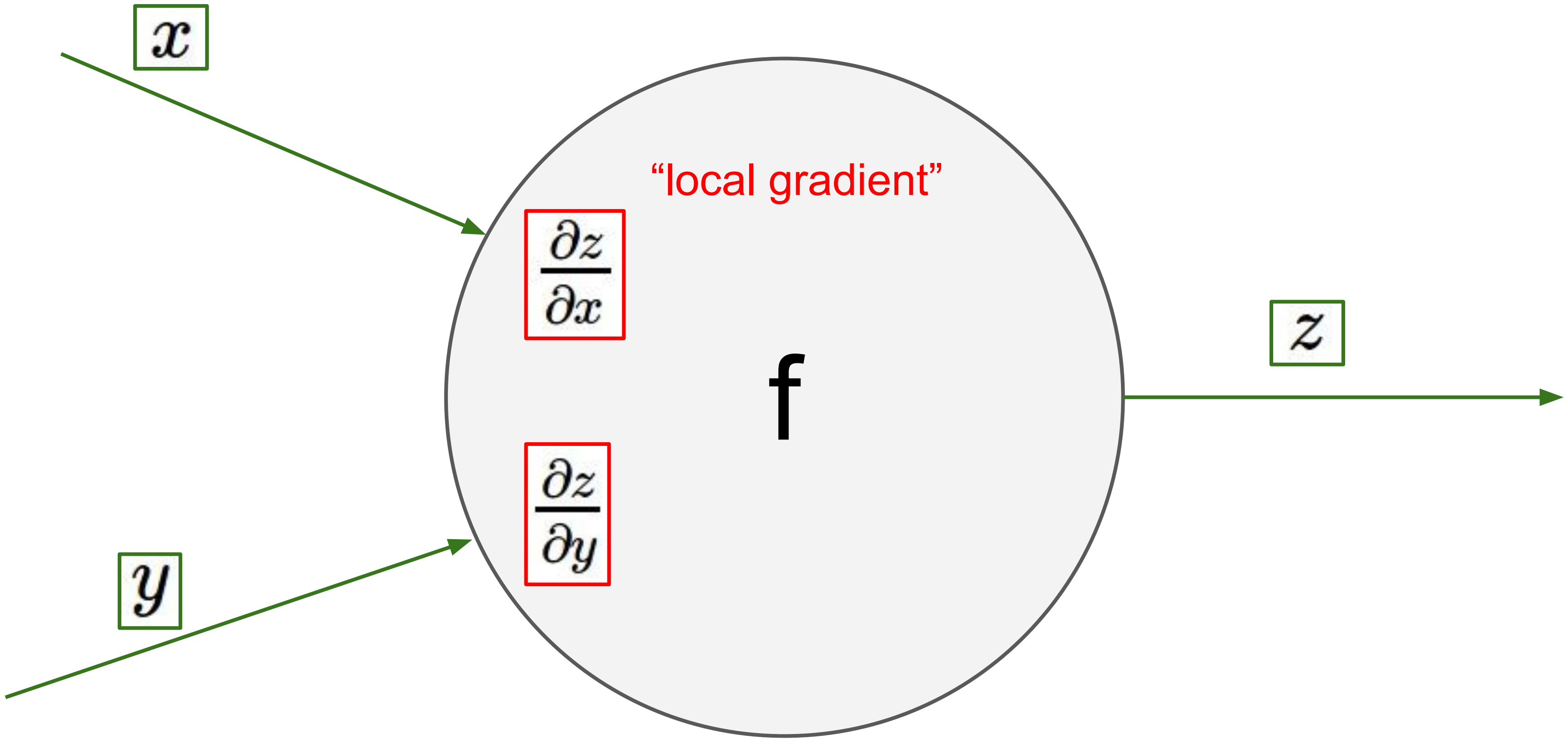
Chain rule:

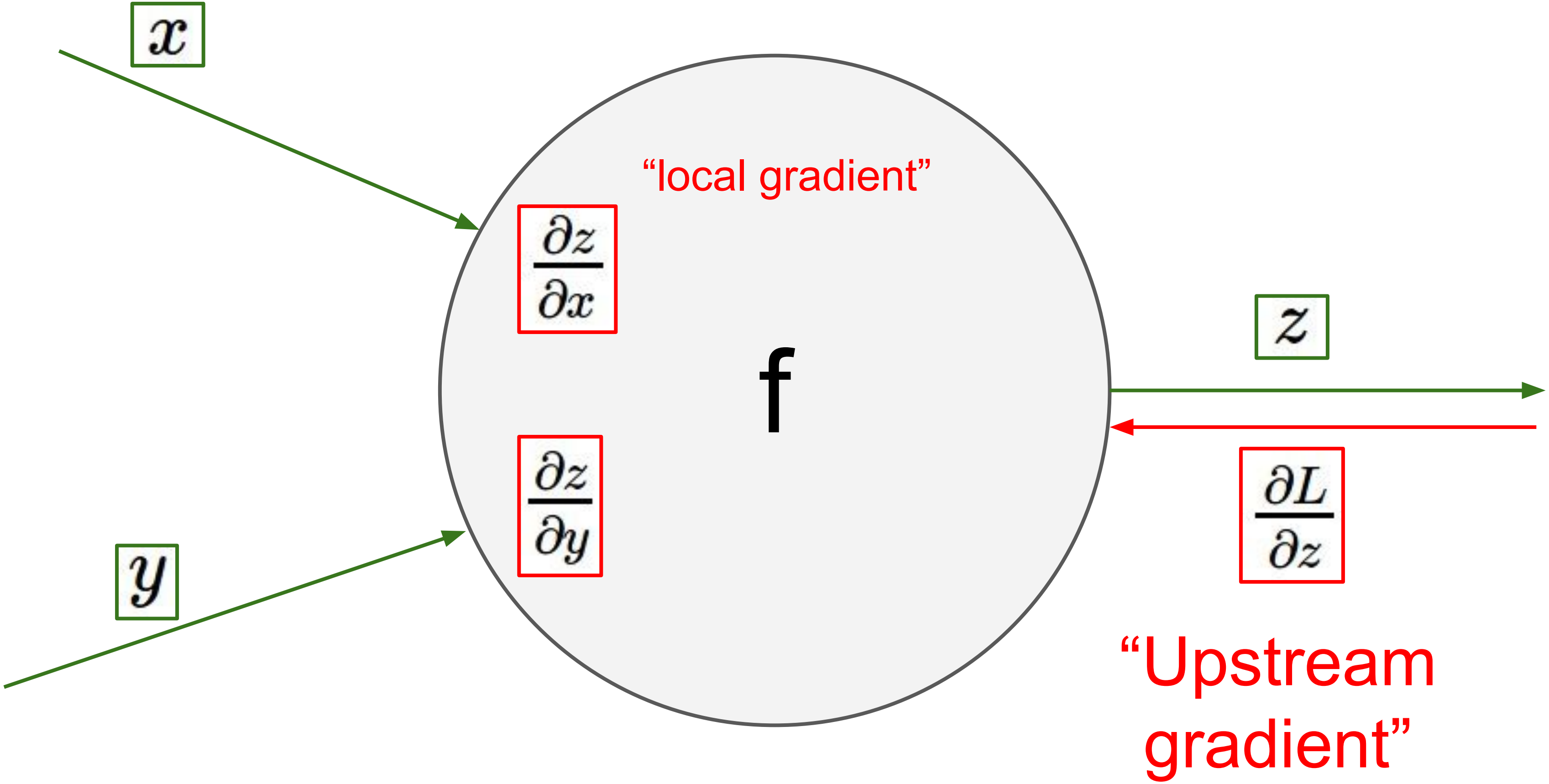
$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial q} \frac{\partial q}{\partial x}$$

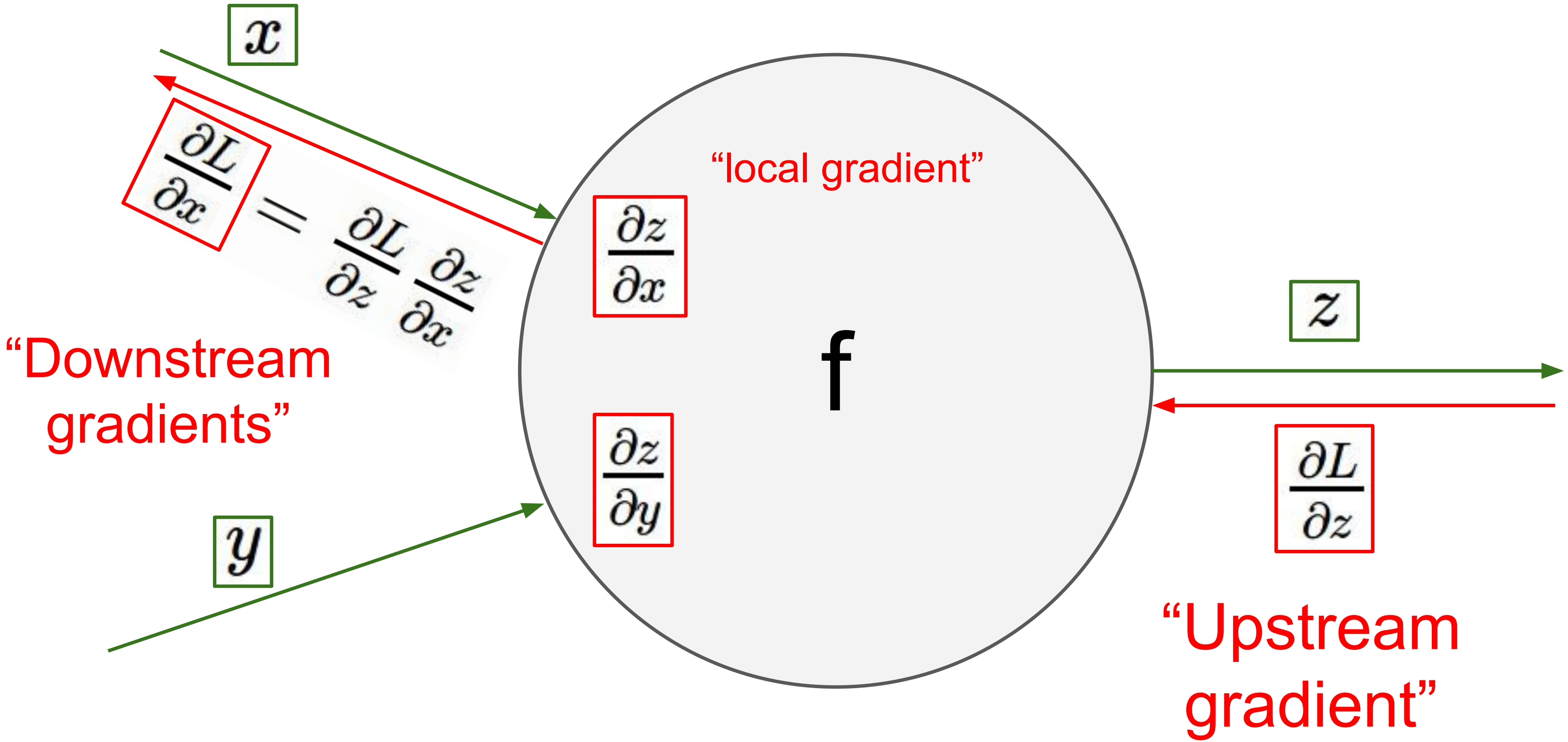
Upstream gradient

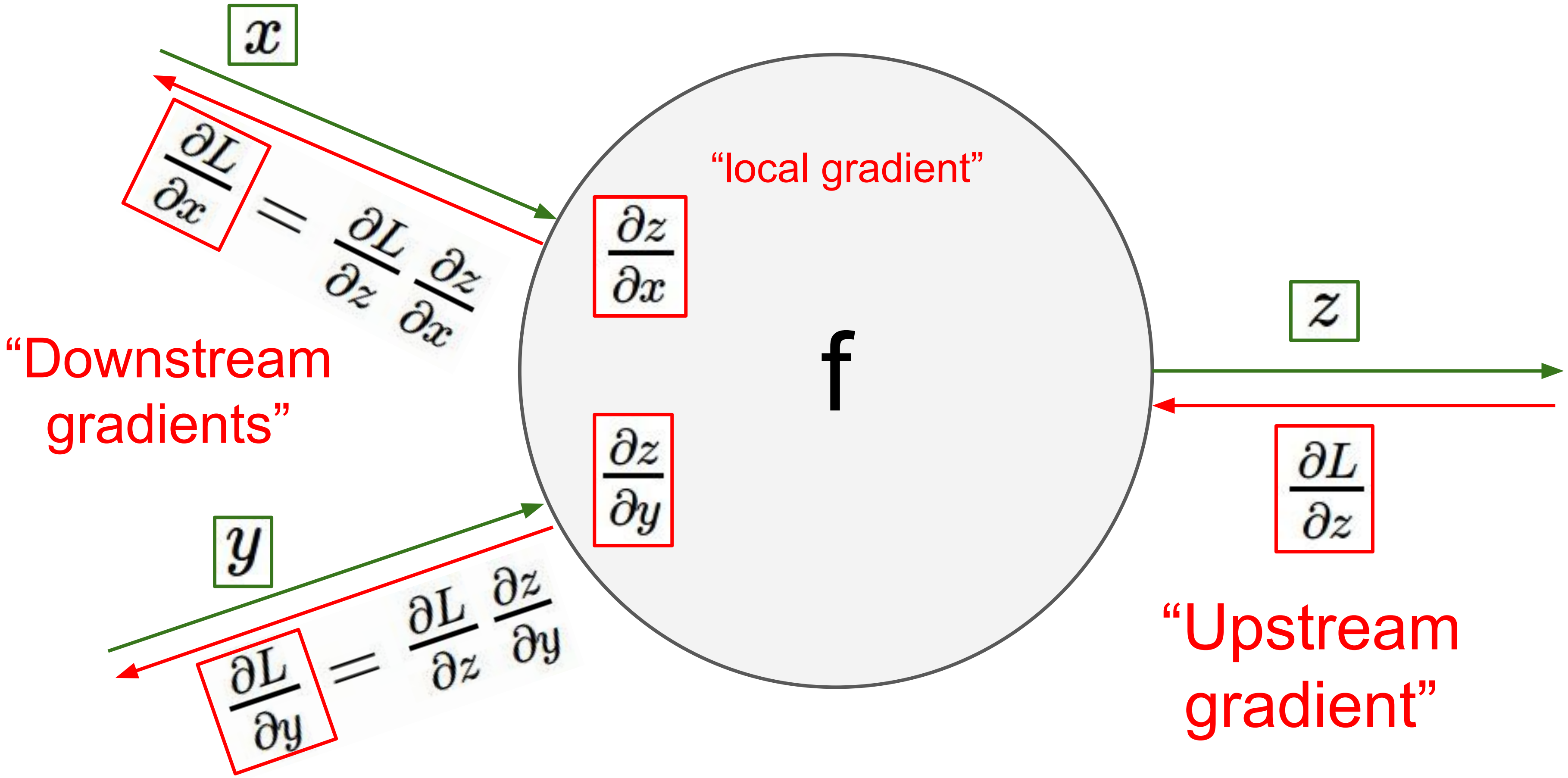
Local gradient


x
y
f
z

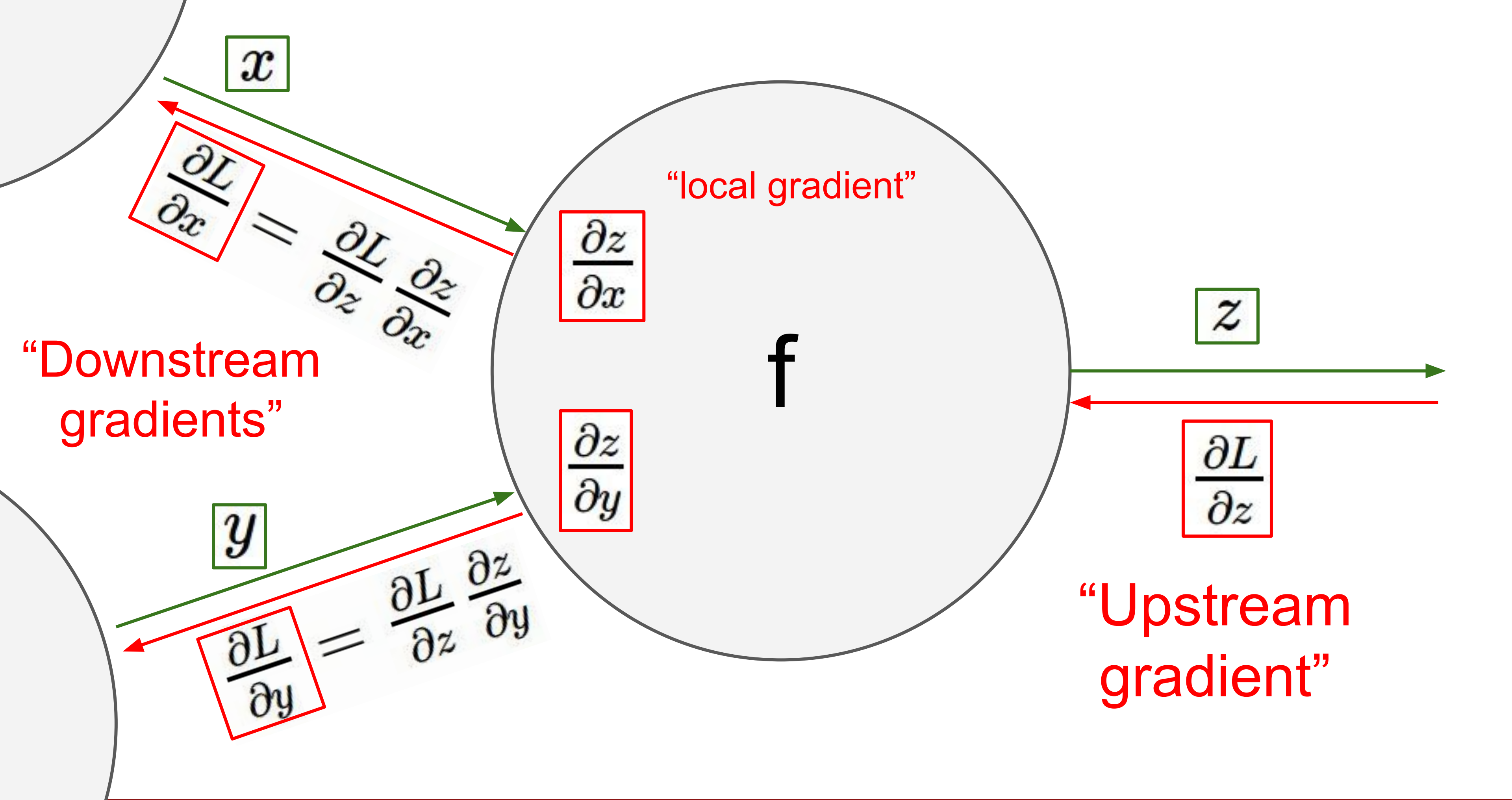

x
"local gradient"
$\frac{\partial z}{\partial x}$
z
f
$\frac{\partial z}{\partial y}$
y


x
y
z
f
"local gradient"
$\frac{\partial z}{\partial x}$
$\frac{\partial z}{\partial y}$
$\frac{\partial L}{\partial z}$
"Upstream gradient"

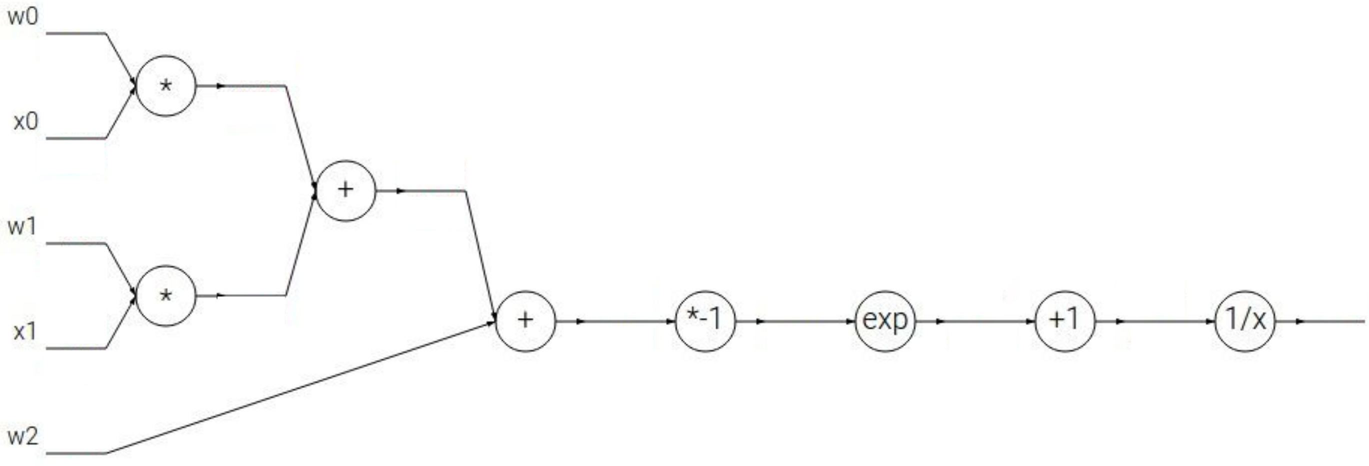

x
$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial z}\frac{\partial z}{\partial x}$
"Downstream gradients"
"local gradient"
$\frac{\partial z}{\partial x}$
$\frac{\partial z}{\partial y}$
f
z
$\frac{\partial L}{\partial z}$
"Upstream gradient"
y



x
$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial z}\frac{\partial z}{\partial x}$
"Downstream gradients"
y
$\frac{\partial L}{\partial y} = \frac{\partial L}{\partial z}\frac{\partial z}{\partial y}$
"local gradient"
$\frac{\partial z}{\partial x}$
$\frac{\partial z}{\partial y}$
f
z
$\frac{\partial L}{\partial z}$
"Upstream gradient"

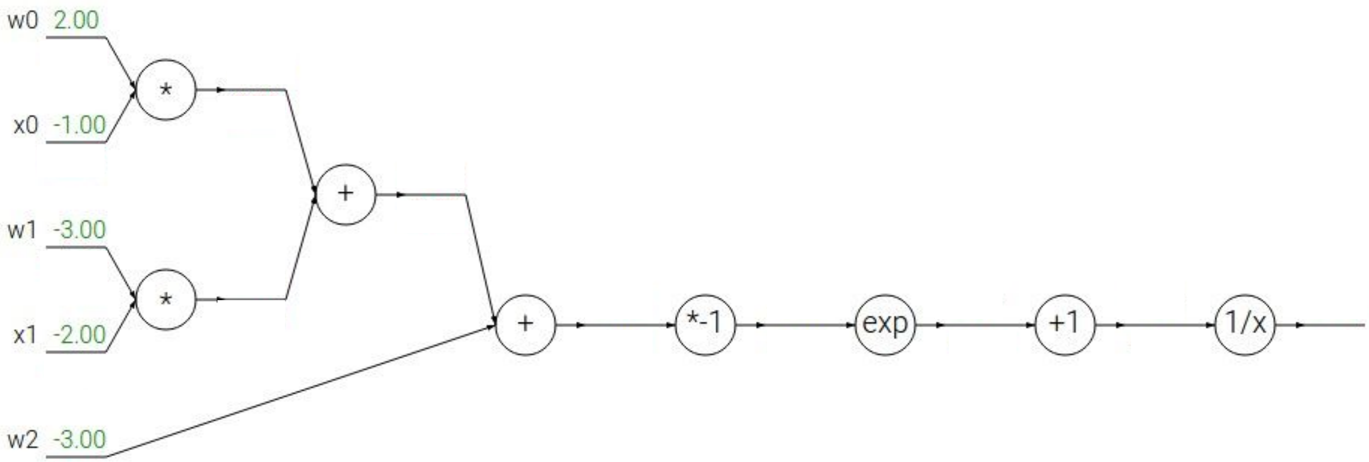

x
$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial z} \frac{\partial z}{\partial x}$
"Downstream gradients"
"local gradient"
$\frac{\partial z}{\partial x}$
$\frac{\partial z}{\partial y}$
f
z
$\frac{\partial L}{\partial z}$
"Upstream gradient"
y
$\frac{\partial L}{\partial y} = \frac{\partial L}{\partial z} \frac{\partial z}{\partial y}$

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



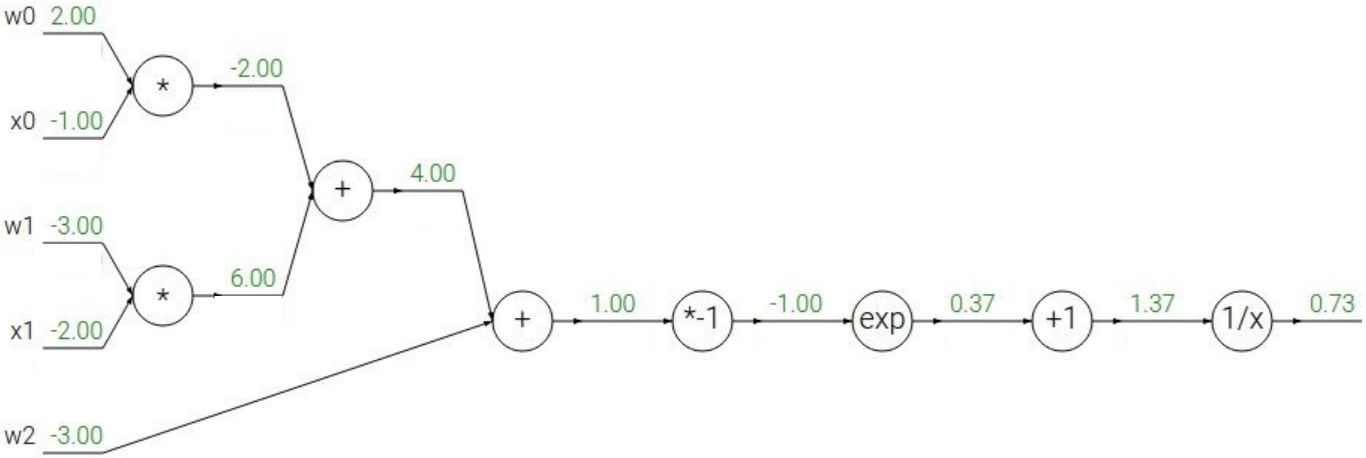
Another example:

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



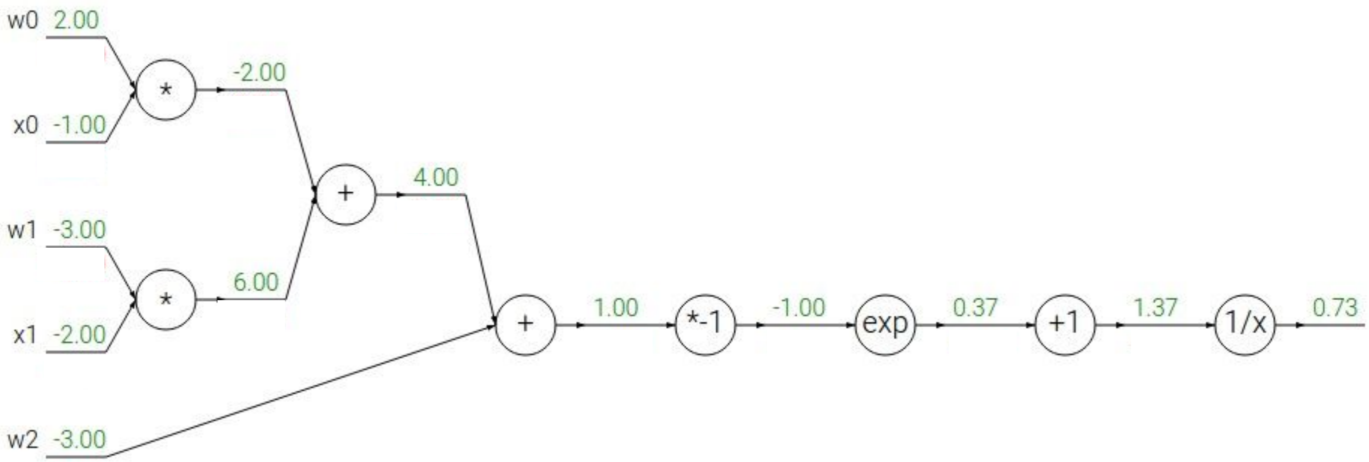
Another example:

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



Another example:

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



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

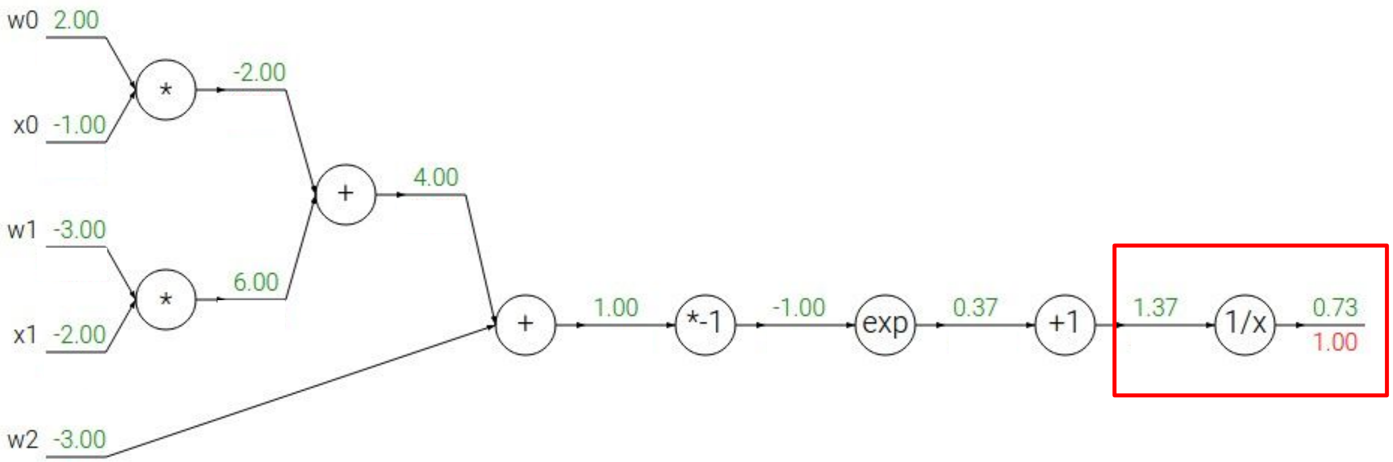
$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$$

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

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

Another example:

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



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

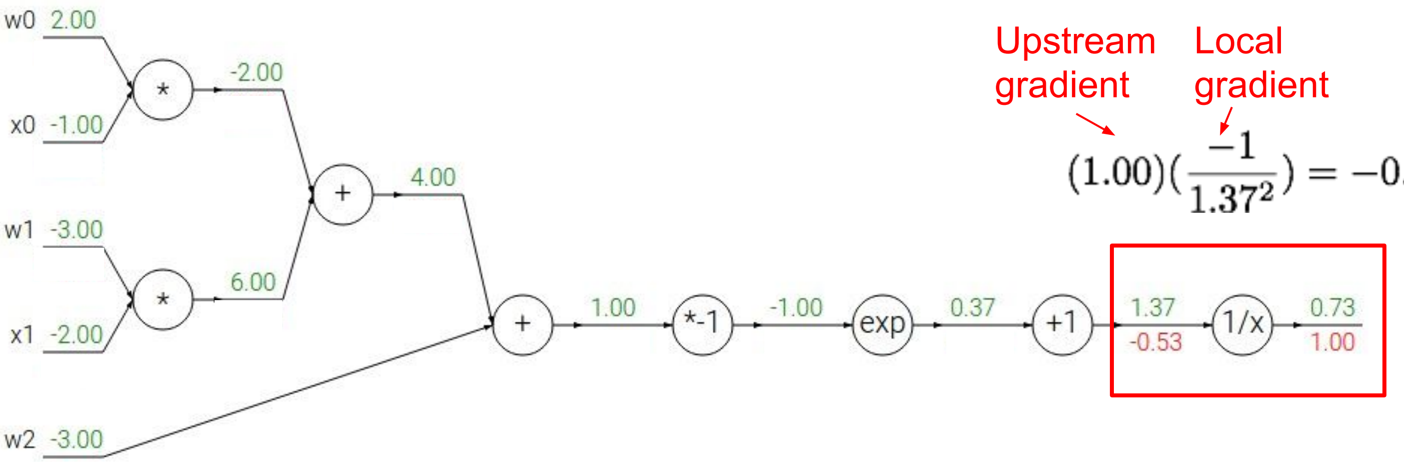
$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$$

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

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

Another example:

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



$$(1.00)\left(\frac{-1}{1.37^2}\right) = -0.53$$

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

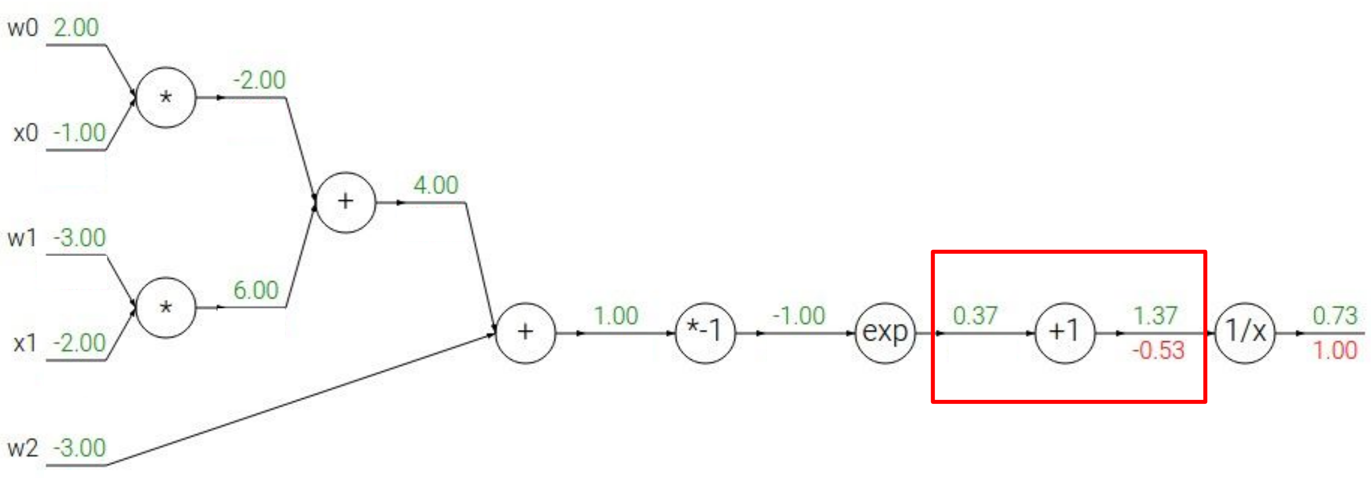
$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$$

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

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

Another example:

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



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

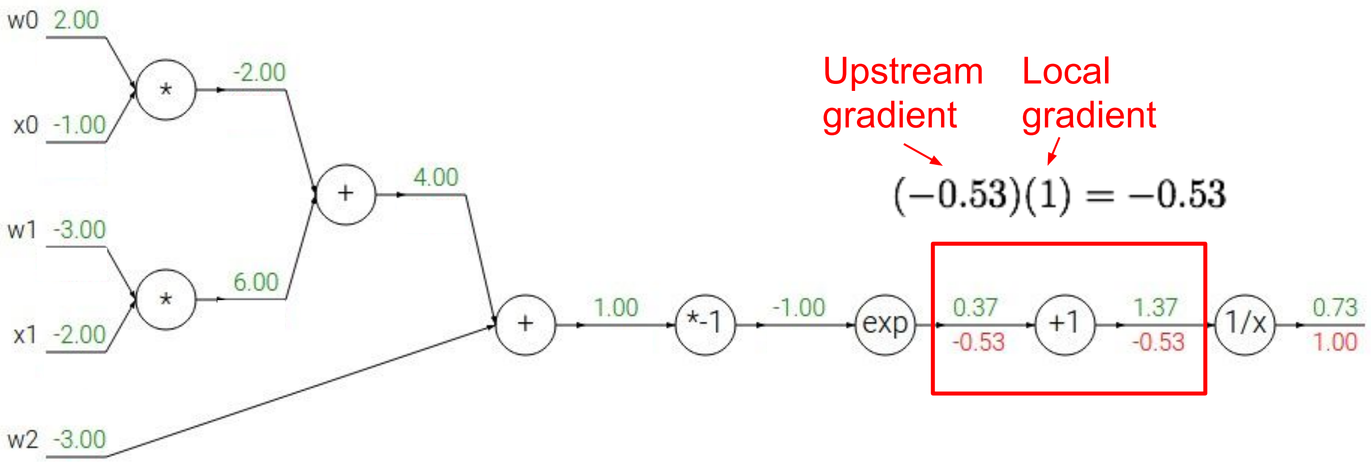
$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$$

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

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

Another example:

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



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

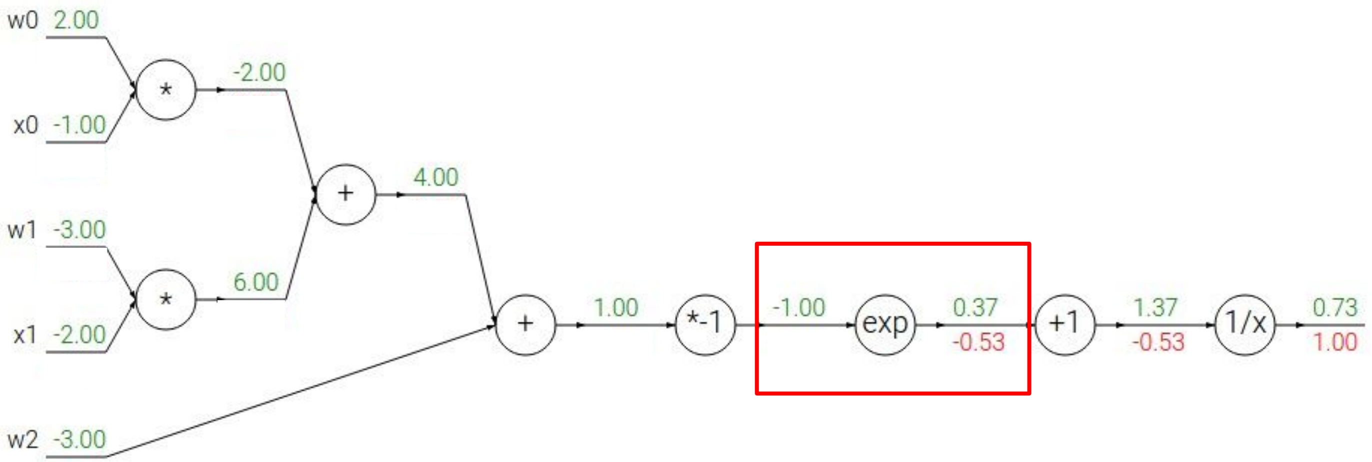
$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$

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

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

Another example:

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



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

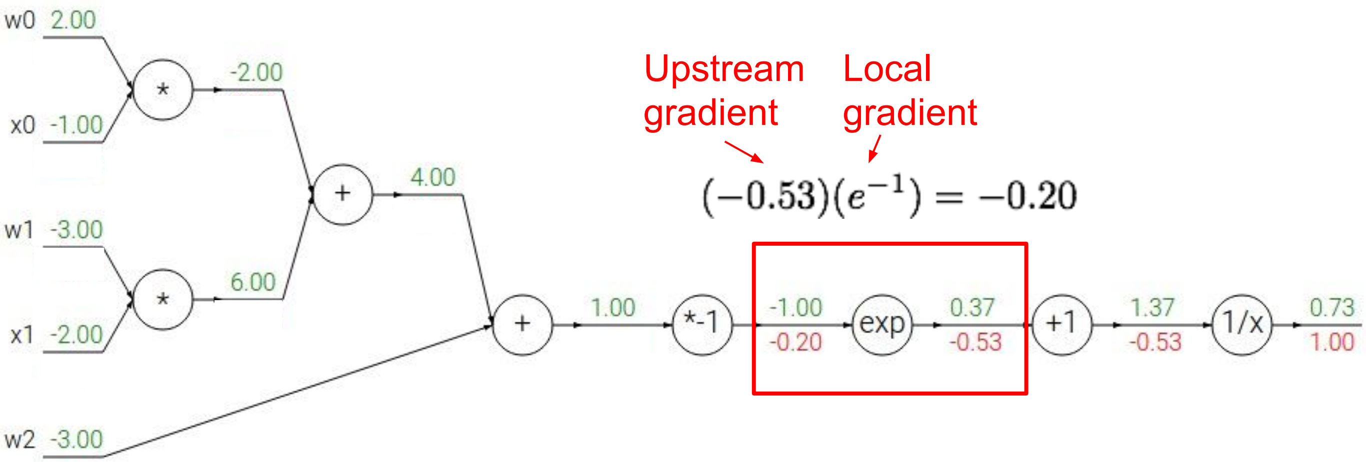
$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$$

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

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

Another example:

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



$$(-0.53)(e^{-1}) = -0.20$$

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

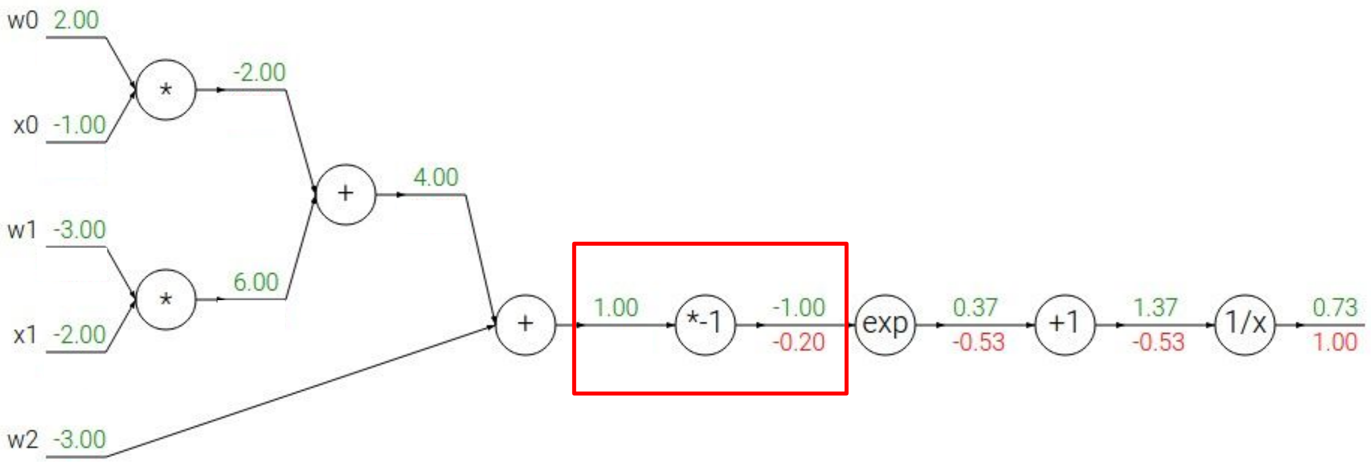
$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$$

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

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

Another example:

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



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

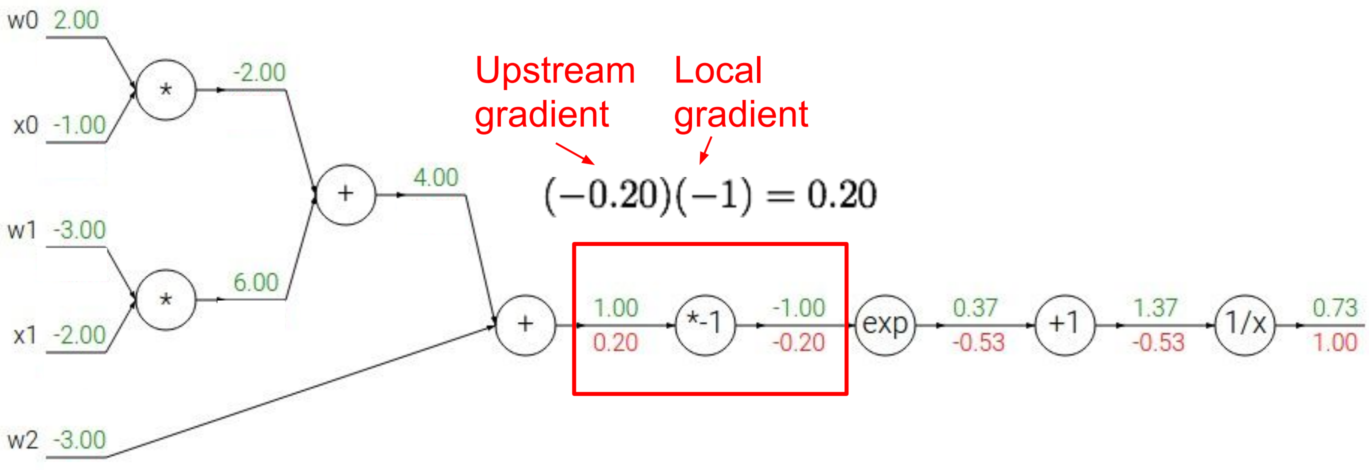
$$\boxed{f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a}$$

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

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

Another example:

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



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

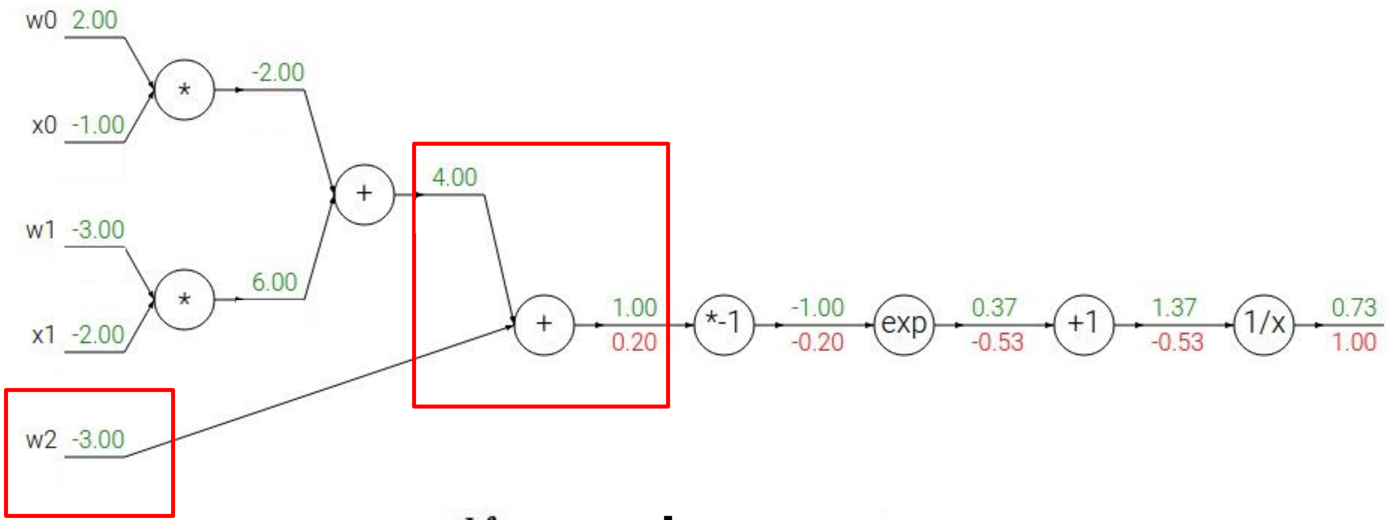
$$\boxed{f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a}$$

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

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

Another example:

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



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

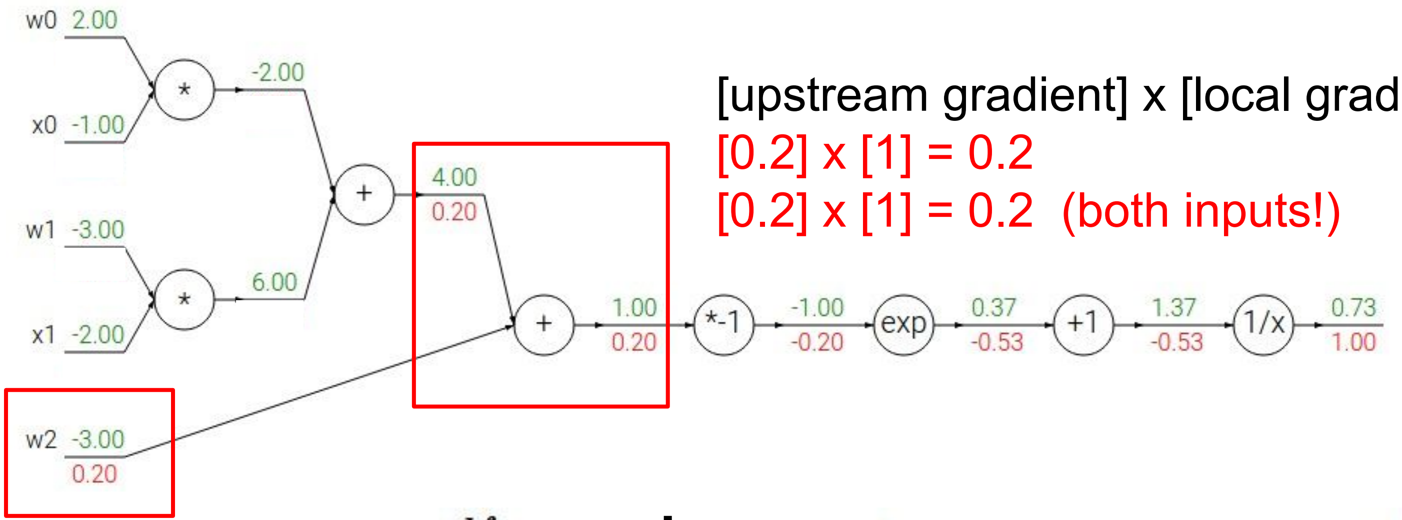
$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$$

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

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

Another example:

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



[upstream gradient] x [local gradient]

[0.2] x [1] = 0.2

[0.2] x [1] = 0.2 (both inputs!)

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

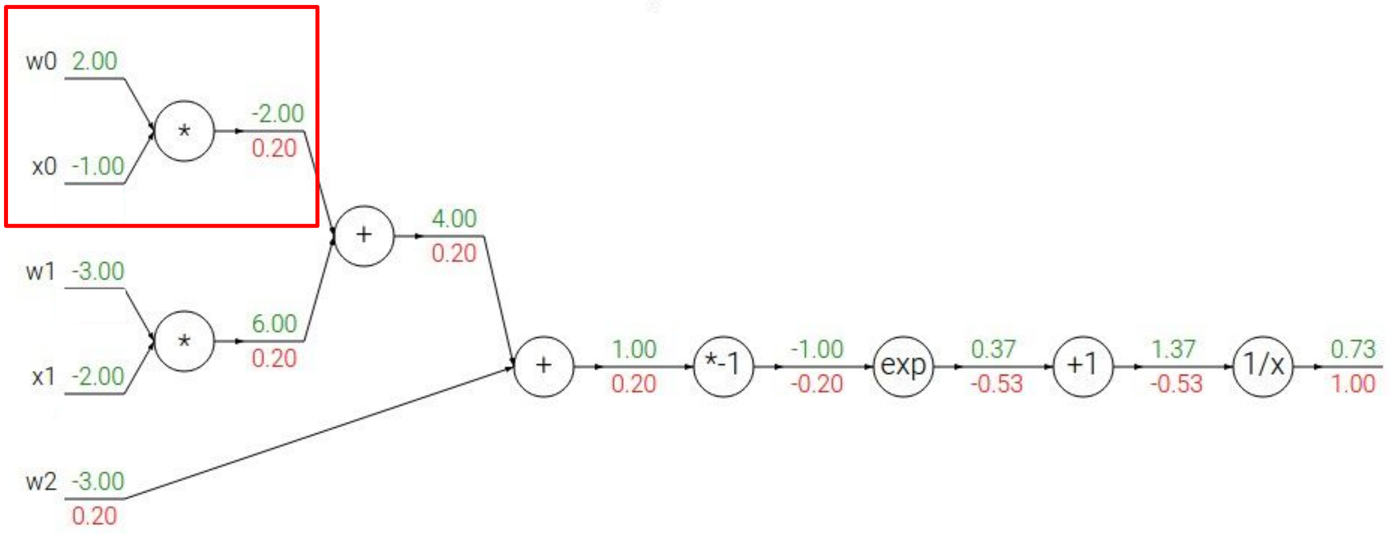
$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$$

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

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

Another example:

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



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

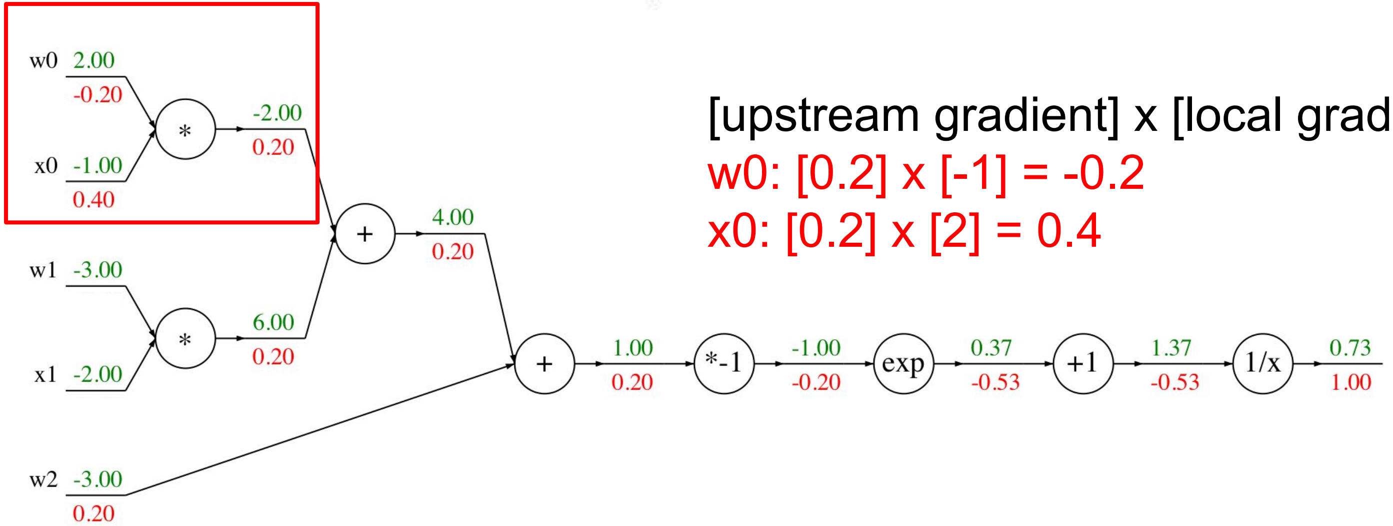
$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$$

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

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

Another example:

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



[upstream gradient] x [local gradient]

w0: [0.2] x [-1] = -0.2

x0: [0.2] x [2] = 0.4

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

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

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

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

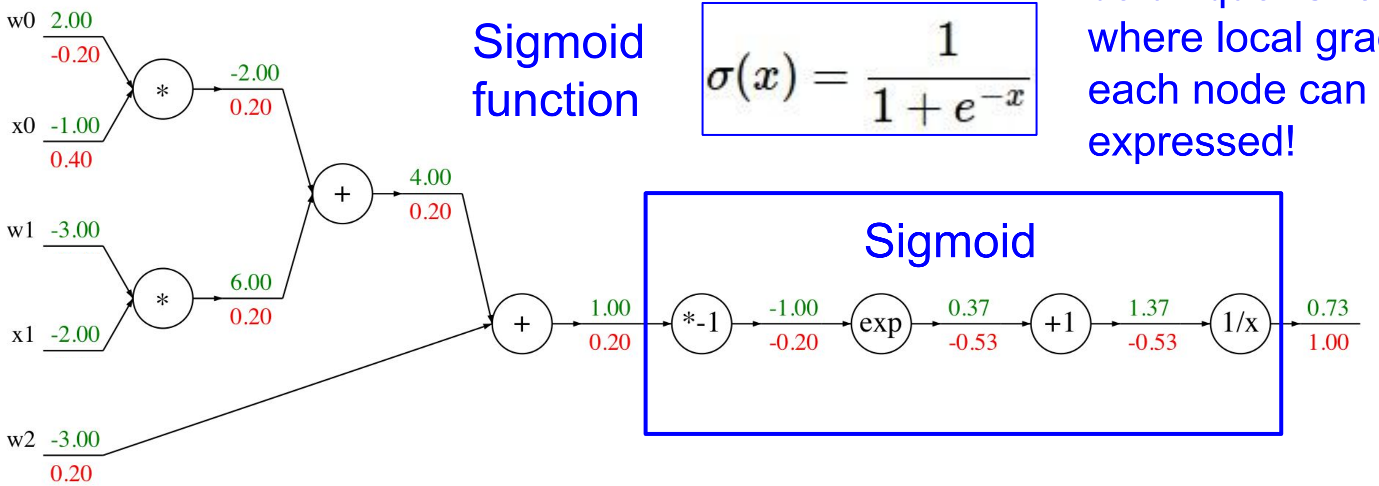
Another example:

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

Computational graph representation may not be unique. Choose one where local gradients at each node can be easily expressed!

Sigmoid function

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



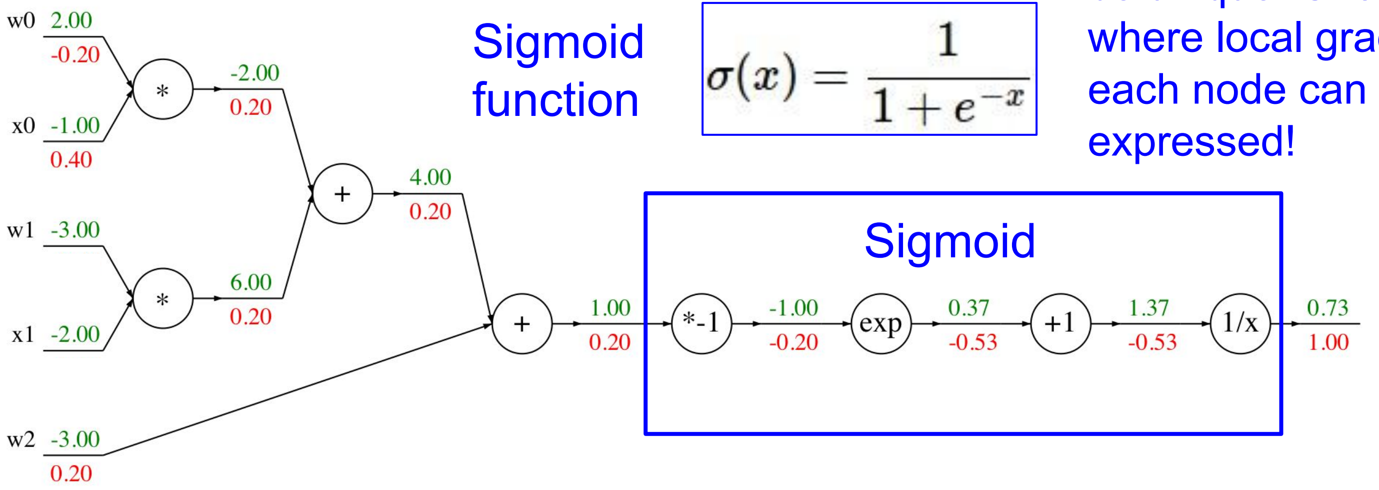
Another example:

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

Computational graph representation may not be unique. Choose one where local gradients at each node can be easily expressed!

Sigmoid function

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



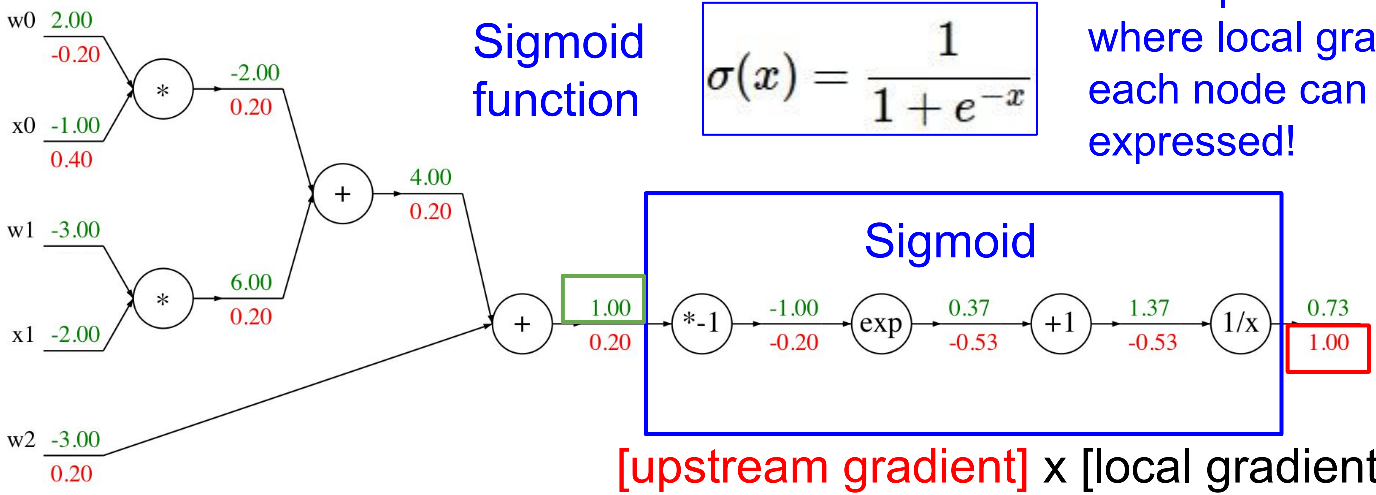
Sigmoid local gradient:

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

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

Computational graph representation may not be unique. Choose one where local gradients at each node can be easily expressed!

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



[upstream gradient] x [local gradient]

[1.00] x [(1 - 1/(1+e^1)) (1/(1+e^1))] = 0.2

Sigmoid local gradient:

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

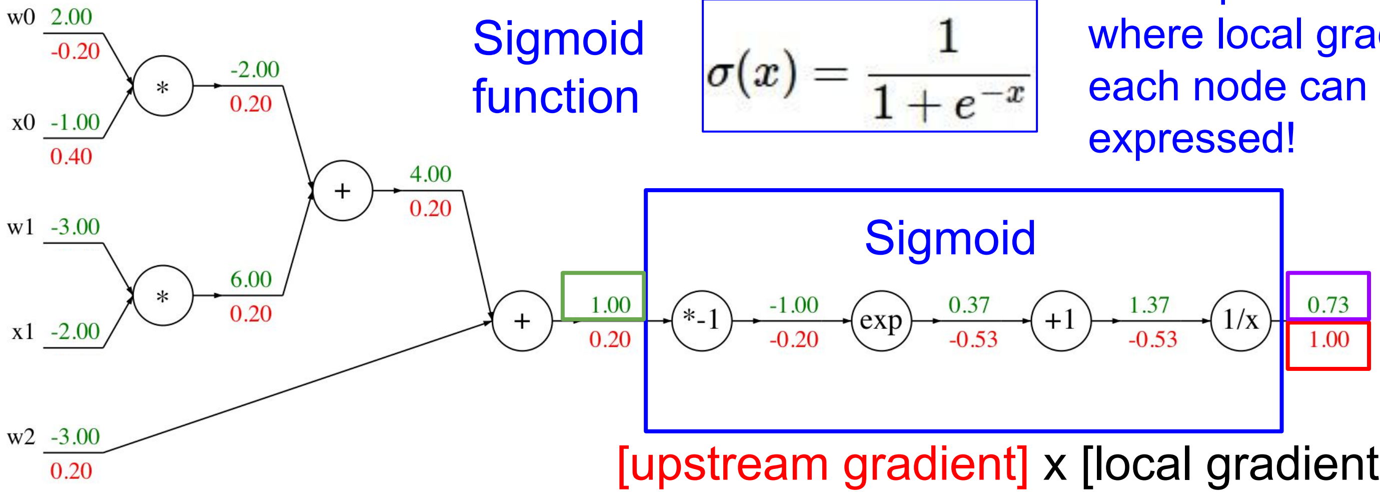
Another example:

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

Computational graph representation may not be unique. Choose one where local gradients at each node can be easily expressed!

Sigmoid function

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



[upstream gradient] x [local gradient]

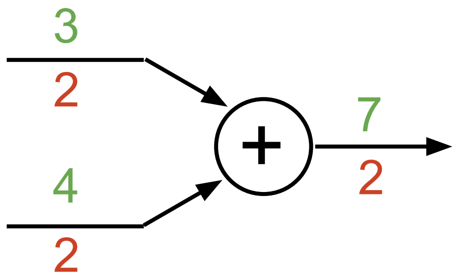
[1.00] x [(1 - 0.73) (0.73)] = 0.2

Sigmoid local gradient:

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

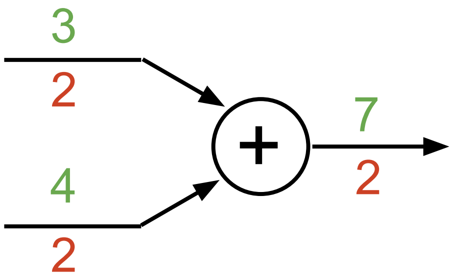
Patterns in gradient flow

add gate: gradient distributor

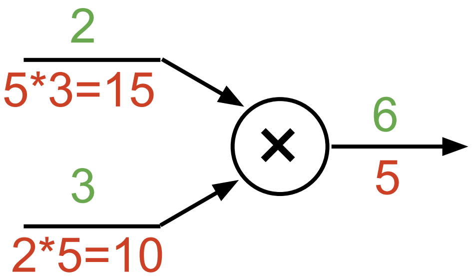


Patterns in gradient flow

add gate: gradient distributor

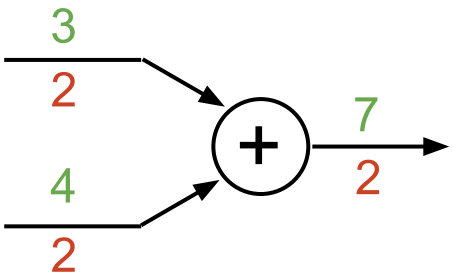


mul gate: “swap multiplier”

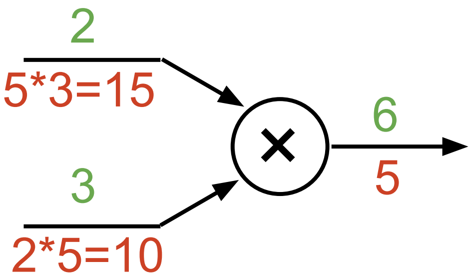


Patterns in gradient flow

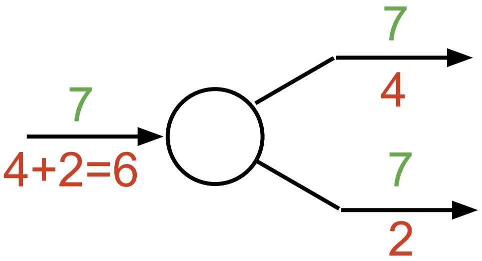
add gate: gradient distributor



mul gate: "swap multiplier"

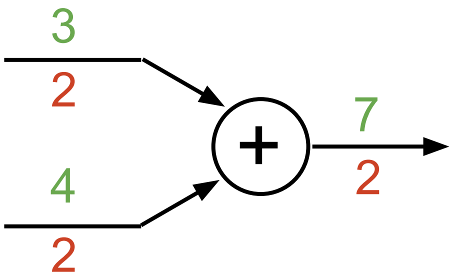


copy gate: gradient adder

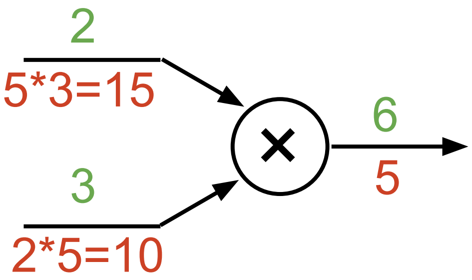


Patterns in gradient flow

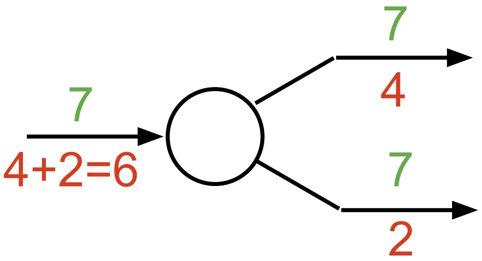
add gate: gradient distributor



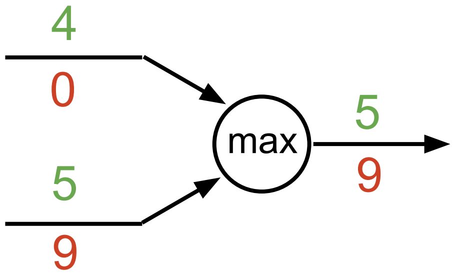
mul gate: “swap multiplier”



copy gate: gradient adder



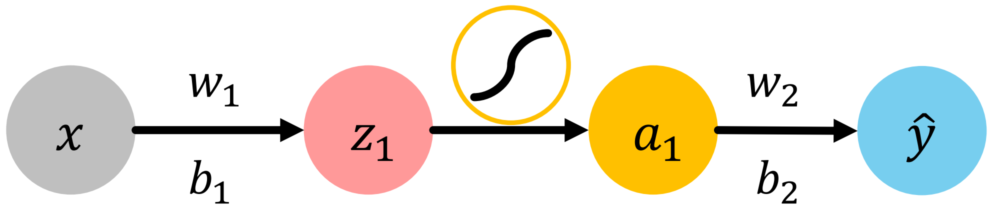
max gate: gradient router



回到神经网络

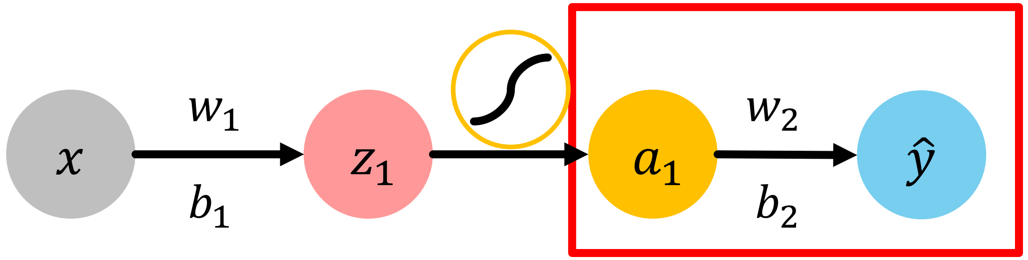
- 利用之前介绍的方法，考虑一个简单的前馈神经网络
- $a_0 = x, \; z_1 = w_1 a_0 + b_1, \; a_1 = \sigma(z_1)$

  Sigmoid函数
- $\hat{y} = z_2 = w_2 a_1 + b_2$
- $l = (y - \hat{y})^2$
- 需要计算 $\frac{\mathrm{d}l}{\mathrm{d}w_1}, \; \frac{\mathrm{d}l}{\mathrm{d}b_1}, \; \frac{\mathrm{d}l}{\mathrm{d}w_2}, \; \frac{\mathrm{d}l}{\mathrm{d}b_2}$



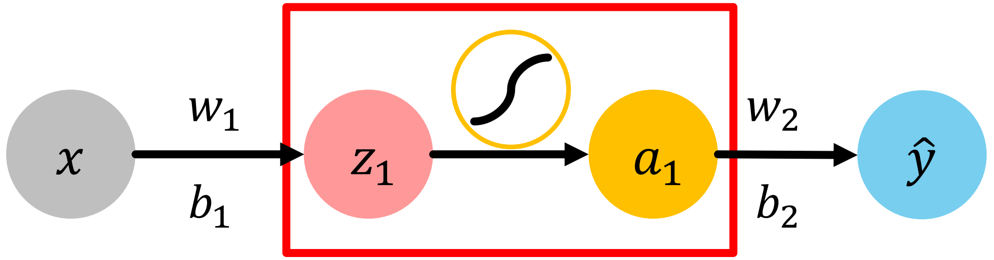
回到神经网络

- $\hat{y} = z_2$ 的上游导数
- $\dfrac{\mathrm{d}l}{\mathrm{d}z_2} = -2(y - z_2)$
- z_2 对 a_1, w_2, b_2 的局部导数
- $\dfrac{\mathrm{d}z_2}{\mathrm{d}a_1} = w_2, \quad \dfrac{\mathrm{d}z_2}{\mathrm{d}w_2} = a_1, \quad \dfrac{\mathrm{d}z_2}{\mathrm{d}b_2} = 1$
- 因此 $\dfrac{\mathrm{d}l}{\mathrm{d}a_1} = \dfrac{\mathrm{d}l}{\mathrm{d}z_2} \cdot w_2, \quad \dfrac{\mathrm{d}l}{\mathrm{d}w_2} = \dfrac{\mathrm{d}l}{\mathrm{d}z_2} \cdot a_1, \quad \dfrac{\mathrm{d}z_2}{\mathrm{d}b_2} = \dfrac{\mathrm{d}l}{\mathrm{d}z_2}$



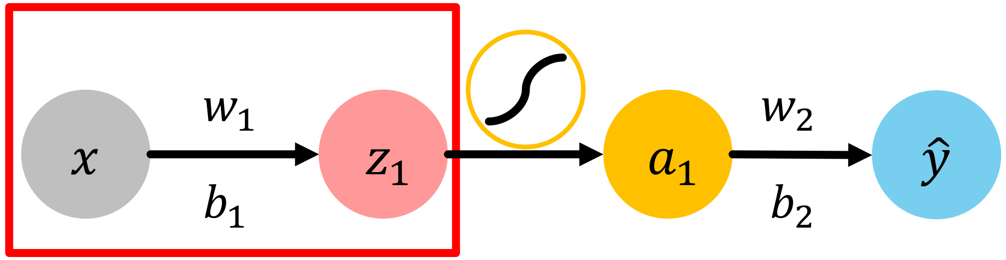
回到神经网络

- a_1 的上游导数 $\frac{\mathrm{d}l}{\mathrm{d}a_1}$ 已经在上一层计算好
- a_1 对 z_1 的局部导数
- $\frac{\mathrm{d}a_1}{\mathrm{d}z_1} = \sigma'(z_1) = \sigma(z_1)[1-\sigma(z_1)] = a_1(1-a_1)$
- 因此 $\frac{\mathrm{d}l}{\mathrm{d}z_1} = \frac{\mathrm{d}l}{\mathrm{d}a_1} \cdot a_1(1-a_1)$



回到神经网络

- z_1 的上游导数 $\frac{\mathrm{d}l}{\mathrm{d}z_1}$ 已经在上一层计算好
- z_1 对 w_1, b_1 的局部导数
- $\frac{\mathrm{d}z_1}{\mathrm{d}w_1} = x, \quad \frac{\mathrm{d}z_1}{\mathrm{d}b_1} = 1$
- 因此 $\frac{\mathrm{d}l}{\mathrm{d}w_1} = \frac{\mathrm{d}l}{\mathrm{d}z_1} \cdot x, \quad \frac{\mathrm{d}l}{\mathrm{d}b_1} = \frac{\mathrm{d}l}{\mathrm{d}z_1}$



扩展

- 这一方法可以很容易扩展到任意层数的情形
- 但每层有多个神经元的情况怎么办?
- 此时 W_i 是矩阵，b_i 是向量

先上结论

规则0

- 参数导数的维度和参数的维度保持一致

规则1

- 对于逐元素计算的激活函数，$a=\sigma(z)$
- $a=(a^1,\dots,a^m)^T$，$z=(z^1,\dots,z^m)^T$，$a^i=\sigma(z^i)$
- 导数也可逐元素计算

- 假设 a 的上游导数为 $\frac{\mathrm{d}l}{\mathrm{d}a}=\left(\frac{\mathrm{d}l}{\mathrm{d}a^1},\dots,\frac{\mathrm{d}l}{\mathrm{d}a^m}\right)^T$，则
- $\frac{\mathrm{d}l}{\mathrm{d}z}=\left(\frac{\mathrm{d}l}{\mathrm{d}a^1}\cdot\sigma'(z^1),\dots,\frac{\mathrm{d}l}{\mathrm{d}a^m}\cdot\sigma'(z^m)\right)^T=\frac{\mathrm{d}l}{\mathrm{d}a}\circ\sigma'(z)$

规则2

- 对于线性变换，$z = Wa + b$，$W \in \mathrm{R}^{n\times m}$
- $a = (a^1, \dots, a^m)^T$，$z = (z^1, \dots, z^n)^T$
- 假设 z 的上游导数为 $\frac{\mathrm{d}l}{\mathrm{d}z} = \left(\frac{\mathrm{d}l}{\mathrm{d}z^1}, \dots, \frac{\mathrm{d}l}{\mathrm{d}z^n}\right)^T$
- 那么
- $\frac{\mathrm{d}l}{\mathrm{d}W} = \frac{\mathrm{d}l}{\mathrm{d}z} a^T$，$\frac{\mathrm{d}l}{\mathrm{d}b} = \frac{\mathrm{d}l}{\mathrm{d}z}$，$\frac{\mathrm{d}l}{\mathrm{d}a} = W^T \frac{\mathrm{d}l}{\mathrm{d}z}$

规则2

- 对于线性变换，$z = Wa + b$，$W \in \mathrm{R}^{n\times m}$
- $a = (a^1, \ldots, a^m)^T$，$z = (z^1, \ldots, z^n)^T$
- 假设 z 的上游导数为 $\frac{\mathrm{d}l}{\mathrm{d}z} = \left(\frac{\mathrm{d}l}{\mathrm{d}z^1}, \ldots, \frac{\mathrm{d}l}{\mathrm{d}z^n}\right)^T$
- 那么
- $\underbrace{\boxed{\frac{\mathrm{d}l}{\mathrm{d}W}}}_{n\times m} = \underbrace{\boxed{\frac{\mathrm{d}l}{\mathrm{d}z}}}_{n\times 1} \underbrace{\boxed{a^T}}_{1\times m}$, $\frac{\mathrm{d}l}{\mathrm{d}b} = \frac{\mathrm{d}l}{\mathrm{d}z}$，$\frac{\mathrm{d}l}{\mathrm{d}a} = W^T \frac{\mathrm{d}l}{\mathrm{d}z}$

规则2

- 对于线性变换，$z = Wa + b$，$W \in \mathrm{R}^{n\times m}$
- $a = (a^1, \dots, a^m)^T$，$z = (z^1, \dots, z^n)^T$
- 假设 z 的上游导数为 $\frac{\mathrm{d}l}{\mathrm{d}z} = \left(\frac{\mathrm{d}l}{\mathrm{d}z^1}, \dots, \frac{\mathrm{d}l}{\mathrm{d}z^n}\right)^T$
- 那么
- $\frac{\mathrm{d}l}{\mathrm{d}W} = \frac{\mathrm{d}l}{\mathrm{d}z} a^T$, $\underbrace{\boxed{\frac{\mathrm{d}l}{\mathrm{d}b}}}_{n\times 1} = \underbrace{\boxed{\frac{\mathrm{d}l}{\mathrm{d}z}}}_{n\times 1}$, $\frac{\mathrm{d}l}{\mathrm{d}a} = W^T \frac{\mathrm{d}l}{\mathrm{d}z}$

规则2

- 对于线性变换，$z = Wa + b$，$W \in \mathrm{R}^{n\times m}$
- $a = (a^1, \ldots, a^m)^T$，$z = (z^1, \ldots, z^n)^T$
- 假设 z 的上游导数为 $\frac{\mathrm{d}l}{\mathrm{d}z} = \left(\frac{\mathrm{d}l}{\mathrm{d}z^1}, \ldots, \frac{\mathrm{d}l}{\mathrm{d}z^n}\right)^T$
- 那么
- $\frac{\mathrm{d}l}{\mathrm{d}W} = \frac{\mathrm{d}l}{\mathrm{d}z} a^T$，$\frac{\mathrm{d}l}{\mathrm{d}b} = \frac{\mathrm{d}l}{\mathrm{d}z}$，$\underbrace{\boxed{\frac{\mathrm{d}l}{\mathrm{d}a}}}_{m\times 1} = \underbrace{\boxed{W^T}}_{m\times n} \underbrace{\boxed{\frac{\mathrm{d}l}{\mathrm{d}z}}}_{n\times 1}$

规则3

- 不要照搬数学表达式
- 实际编程中要考虑到数据的存储方式

理论 VS 实际

规则0

理论

- 参数导数的维度和参数的维度保持一致

实际

- 参数导数的维度和参数的维度保持一致

规则1

理论

- 对于逐元素计算的激活函数，导数也可逐元素计算
- $a: [m \times 1]$, $z: [m \times 1]$, $a_i = \sigma(z_i)$
- $\frac{\mathrm{d}l}{\mathrm{d}z} = \frac{\mathrm{d}l}{\mathrm{d}a} \circ \sigma'(z)$

实际

- 对于逐元素计算的激活函数，导数也可逐元素计算
- $A: [n \times m]$, $Z: [n \times m]$, $A_{ij} = \sigma(Z_{ij})$
- $\frac{\mathrm{d}l}{\mathrm{d}Z} = \frac{\mathrm{d}l}{\mathrm{d}A} \circ \sigma'(Z)$

规则1.5

理论

- 线性变换，$z = Wa + b$，$W \in \mathrm{R}^{p \times m}$
- a: $[m \times 1]$，z: $[p \times 1]$
- W: $[p \times m]$，b: $[p \times 1]$

- a 代表上一层神经元（m个）
- z 代表下一层神经元（p个）

实际

- $Z = AW + \mathbf{1}_n b^T$，$W \in \mathrm{R}^{m \times p}$
- A: $[n \times m]$，Z: $[n \times p]$
- W: $[m \times p]$，b: $[p \times 1]$

- n 代表样本量
- A 的第 i 行代表第 i 个观测在上一层的神经元
- Z 的第 i 行代表第 i 个观测在下一层的神经元

规则2

理论

- $\overset{p\times m}{\frac{\mathrm{d}l}{\mathrm{d}W}} = \overset{p\times 1}{\frac{\mathrm{d}l}{\mathrm{d}z}} \overset{1\times m}{a^T}$
- $\underset{p\times 1}{\frac{\mathrm{d}l}{\mathrm{d}b}} = \underset{p\times 1}{\frac{\mathrm{d}l}{\mathrm{d}z}}$
- $\underset{m\times 1}{\frac{\mathrm{d}l}{\mathrm{d}a}} = \underset{m\times p}{W^T} \underset{p\times 1}{\frac{\mathrm{d}l}{\mathrm{d}z}}$

实际

- $\overset{m\times p}{\frac{\mathrm{d}l}{\mathrm{d}W}} = \overset{m\times n}{A^T} \overset{n\times p}{\frac{\mathrm{d}l}{\mathrm{d}Z}}$
- $\underset{p\times 1}{\frac{\mathrm{d}l}{\mathrm{d}b}} = \underset{p\times n}{\left(\frac{\mathrm{d}l}{\mathrm{d}Z}\right)^T} \underset{n\times 1}{\mathbf{1}_n}$
- $\underset{n\times m}{\frac{\mathrm{d}l}{\mathrm{d}A}} = \underset{n\times p}{\frac{\mathrm{d}l}{\mathrm{d}Z}} \underset{p\times m}{W^T}$

自动微分与优化

- 参见 `lec5-autograd-optimize.ipynb`

前馈神经网络实现

- 参见 `lec5-fnn.ipynb`