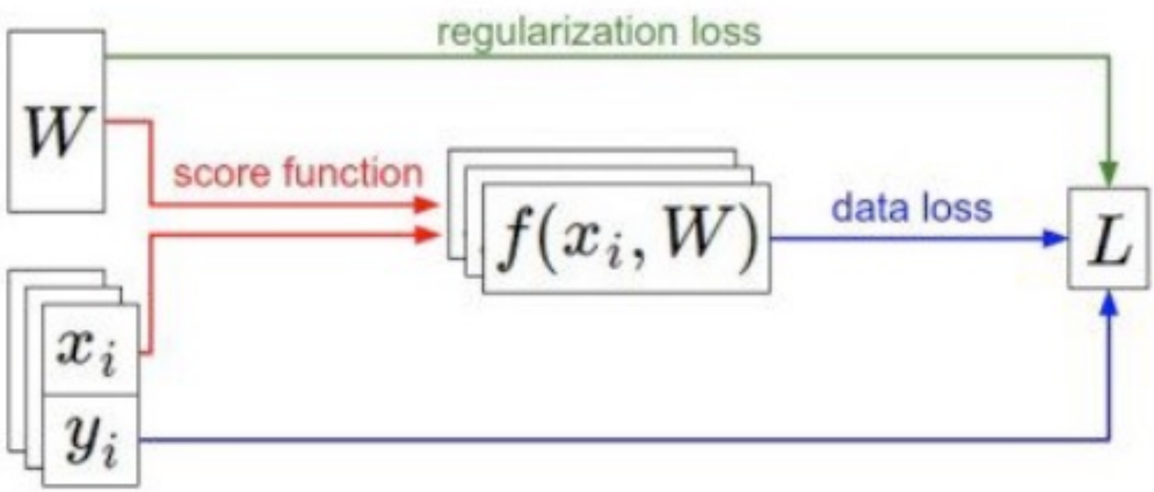
14.332.435/16.332.530
Introduction to Deep Learning

Lecture 4
Backpropagation

Yuqian Zhang

Department of Electrical and Computer Engineering

Recall Last Time

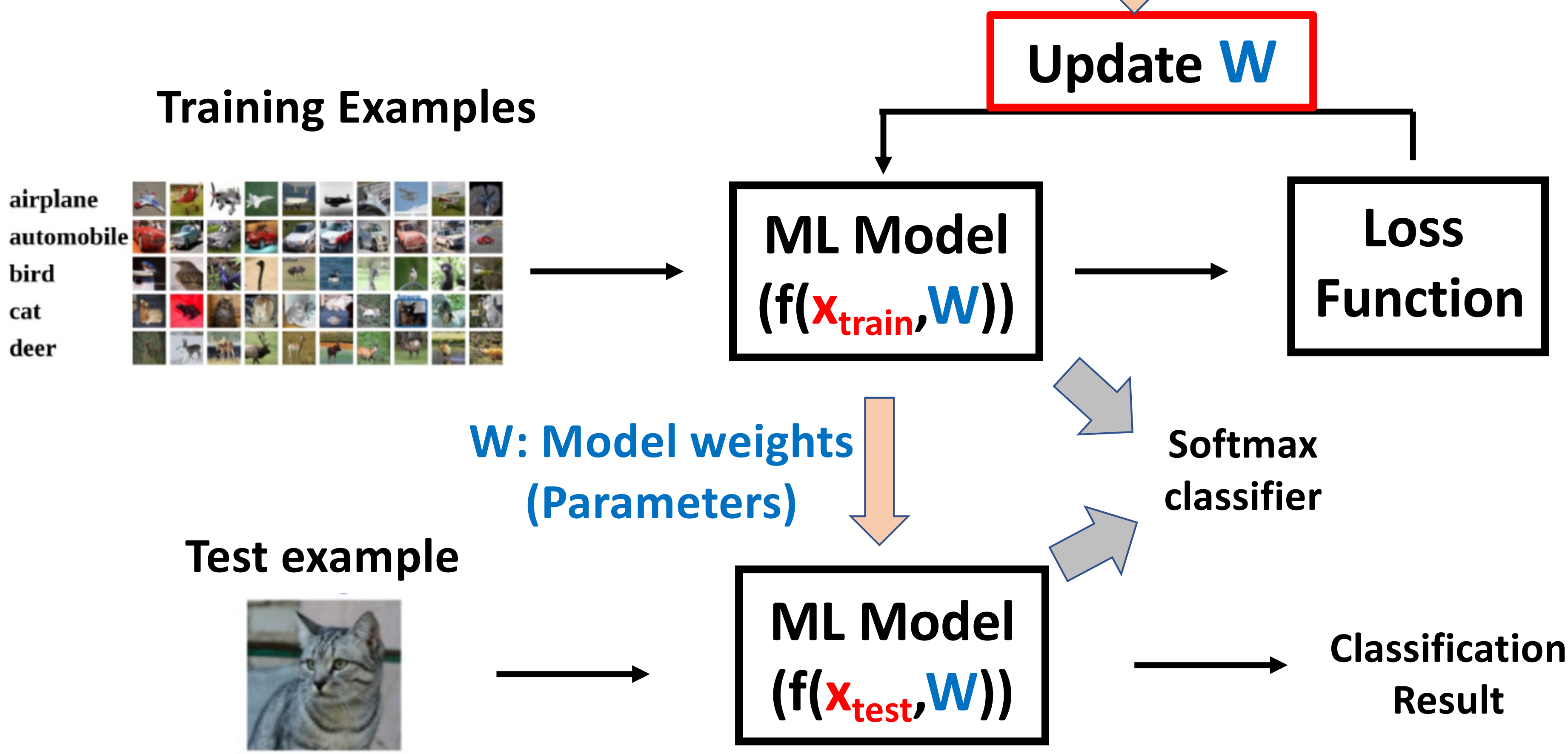


$$w_i = w_i - step_{size} * \boxed{\frac{\partial L}{\partial w_i}}$$

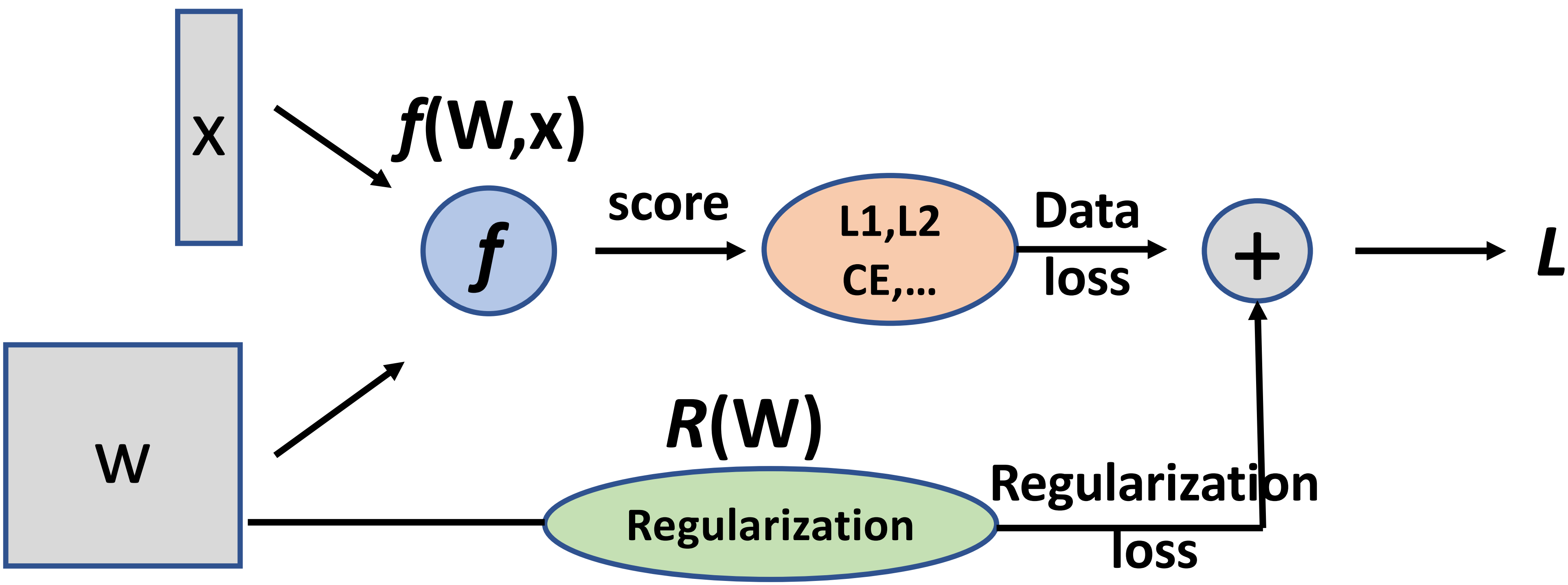
Use numerical method to compute

Today’s Agenda

Use **backpropagation to calculate** $\frac{\partial L}{\partial W}$



Computational Graph



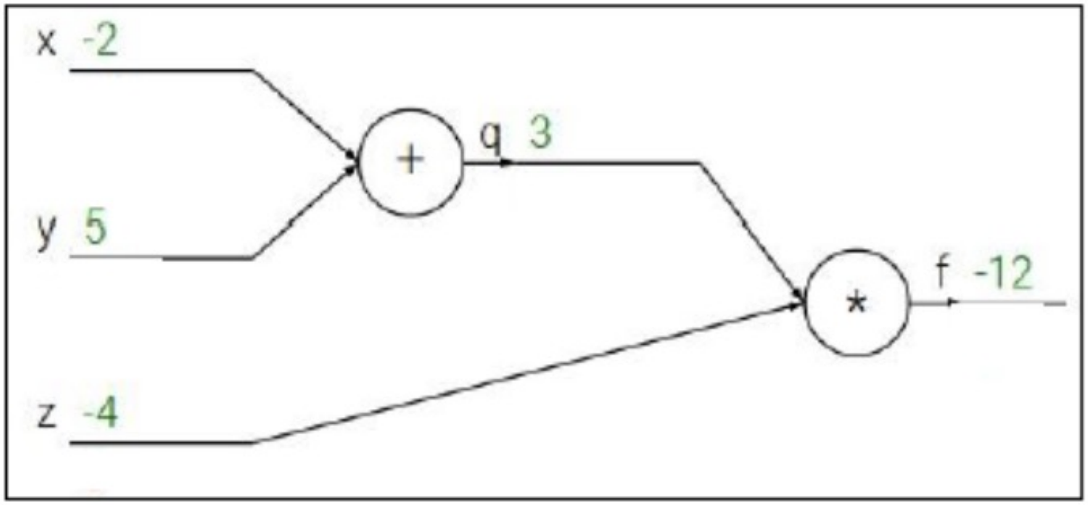
CE: Cross Entropy

Example

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

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

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

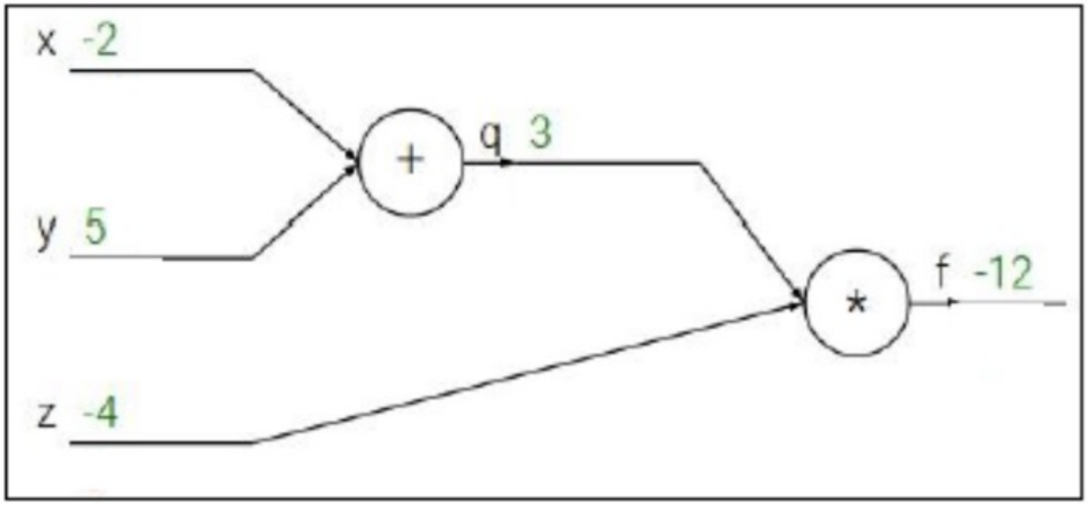


Example

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

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

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



Solution: $q = x + y$

$$f = qz$$

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

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

Example

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

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

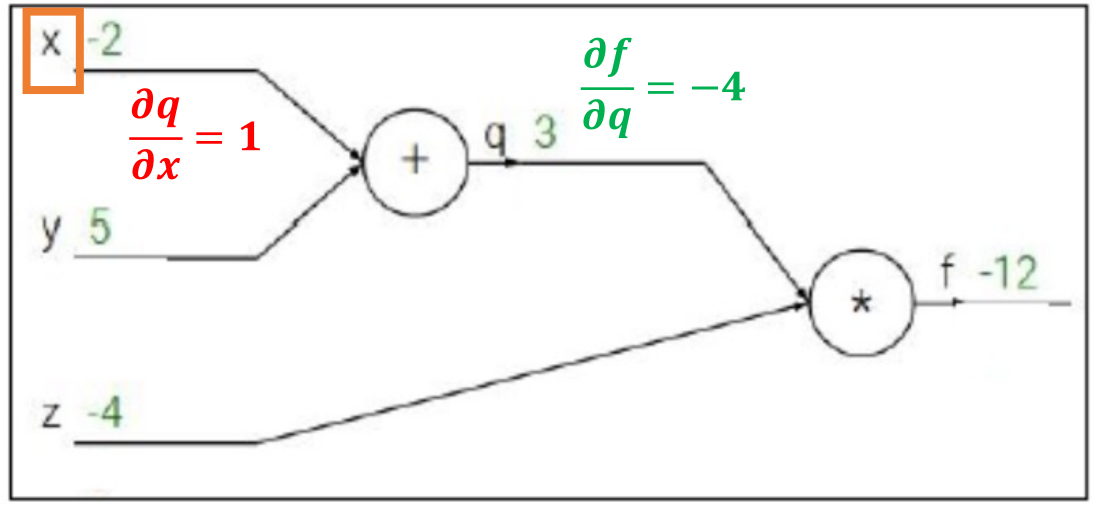
Want: $\boxed{\frac{\partial f}{\partial x}}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$

Solution: $q = x + y$

$$f = qz$$

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

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



Chain Rule: for $f(q(x))$, $\frac{\partial f}{\partial x} = \underbrace{\frac{\partial f}{\partial q}}_{\text{Upstream gradient}} \underbrace{\frac{\partial q}{\partial x}}_{\text{Local gradient}}$

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial q}\frac{\partial q}{\partial x} = z \times 1 = -4$$

Example

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

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

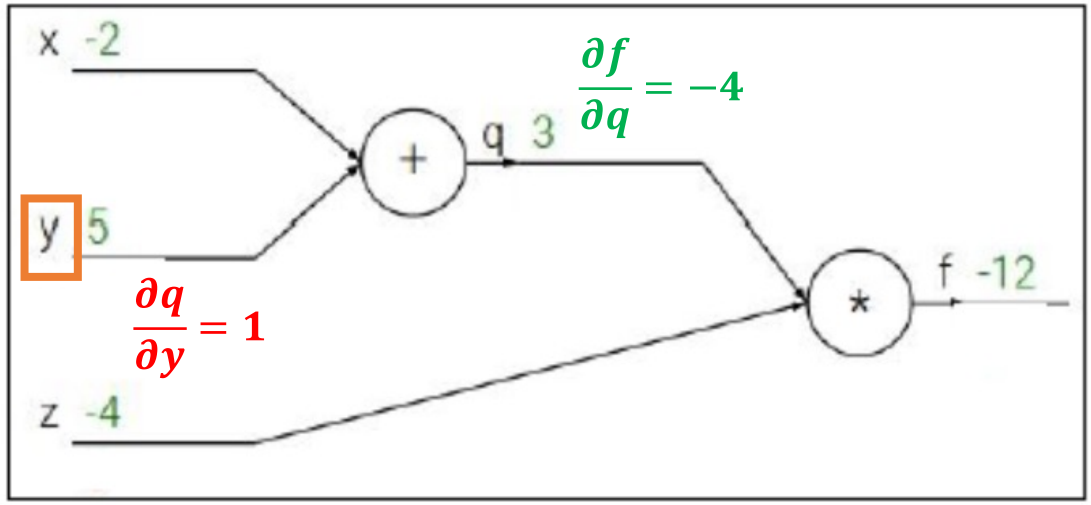
Want: $\frac{\partial f}{\partial x}, \boxed{\frac{\partial f}{\partial y}}, \frac{\partial f}{\partial z}$

Solution: $q = x + y$

$$f = qz$$

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

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



Chain Rule: for $f(q(y))$, $\frac{\partial f}{\partial y} = \frac{\partial f}{\partial q}\frac{\partial q}{\partial y}$

Upstream gradient

Local gradient

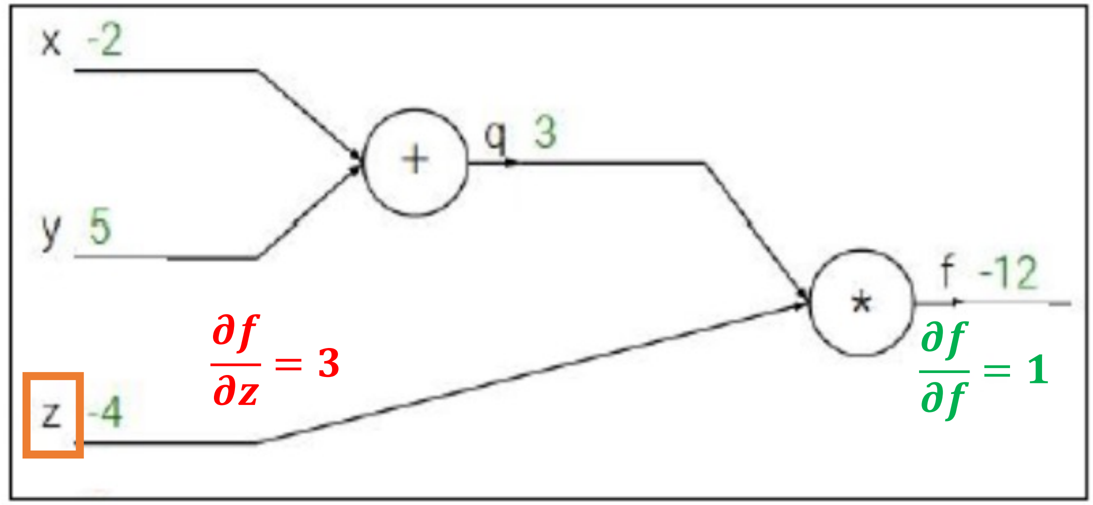
$$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial q}\frac{\partial q}{\partial y} = z \times 1 = -4$$

Example

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

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

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



Solution: $q = x + y$

$$f = qz$$

Chain Rule: for $f(z)$, $\dfrac{\partial f}{\partial z} = \dfrac{\partial f}{\partial f}\dfrac{\partial f}{\partial z}$

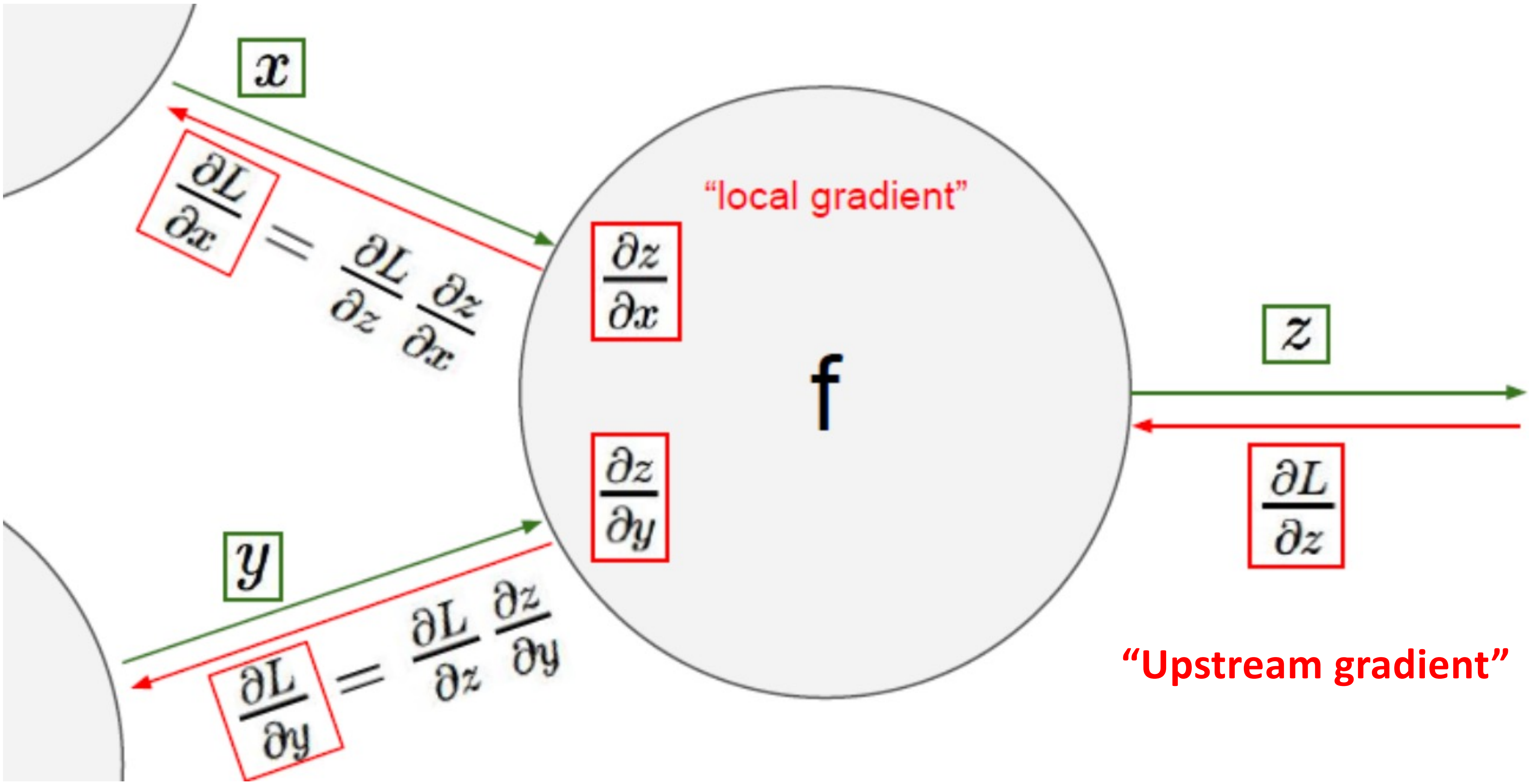
Upstream gradient

Local gradient

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

$$\frac{\partial f}{\partial z} = \frac{\partial f}{\partial f}\frac{\partial f}{\partial z} = 1 \times 3 = 3$$

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

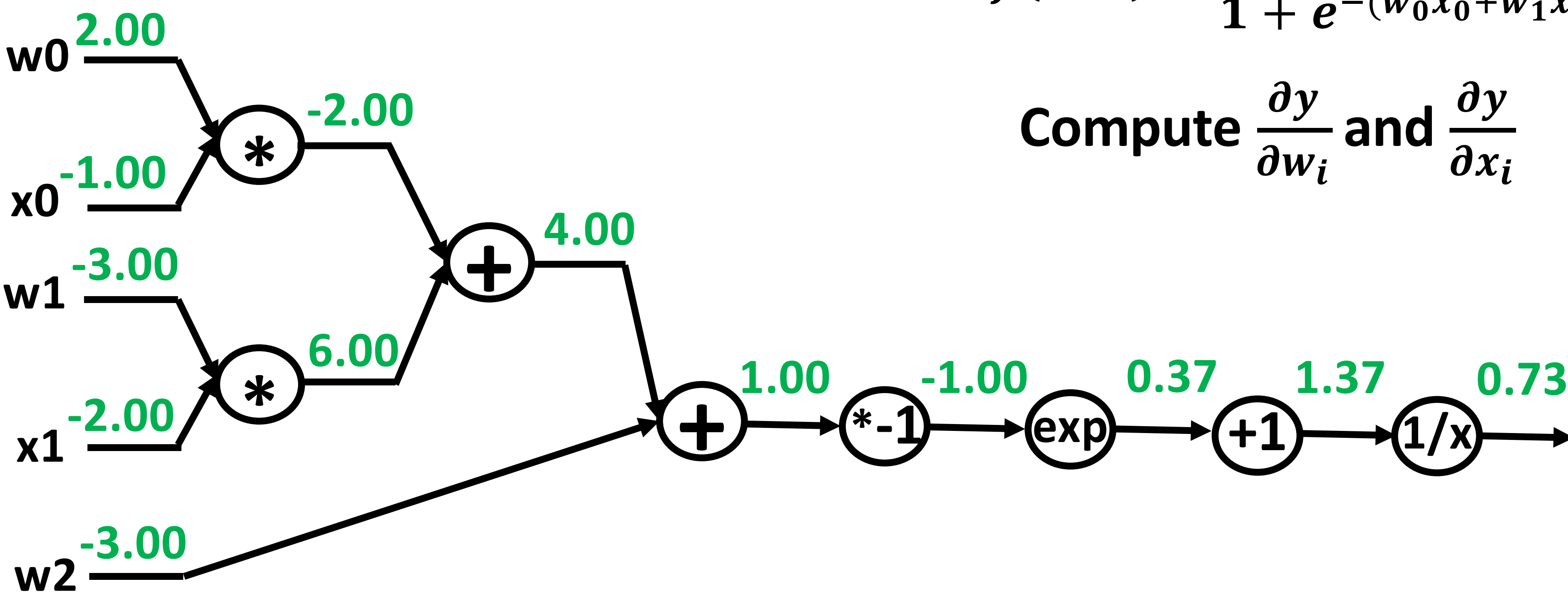


x
$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial z}\frac{\partial z}{\partial x}$
"local gradient"
$\frac{\partial z}{\partial x}$
f
$\frac{\partial z}{\partial y}$
z
$\frac{\partial L}{\partial z}$
y
$\frac{\partial L}{\partial y} = \frac{\partial L}{\partial z}\frac{\partial z}{\partial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

Example

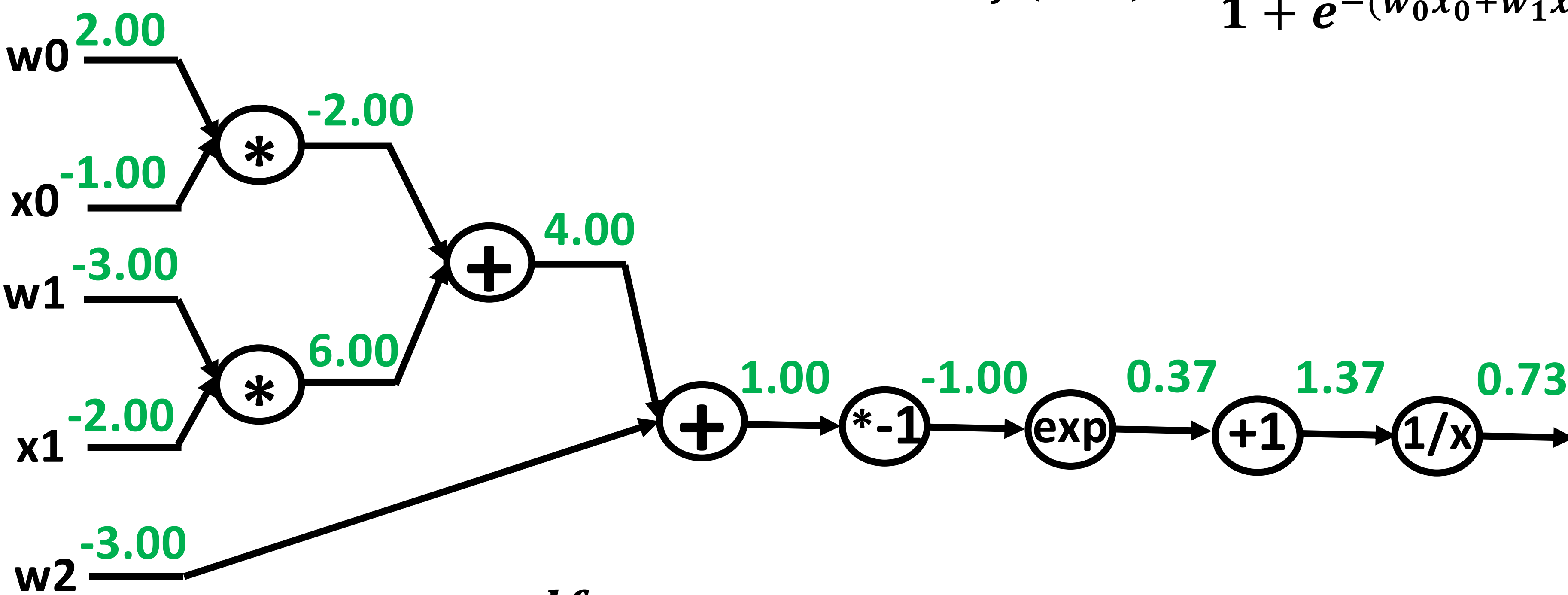
$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

Compute $\frac{\partial y}{\partial w_i}$ and $\frac{\partial y}{\partial x_i}$



Example

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



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

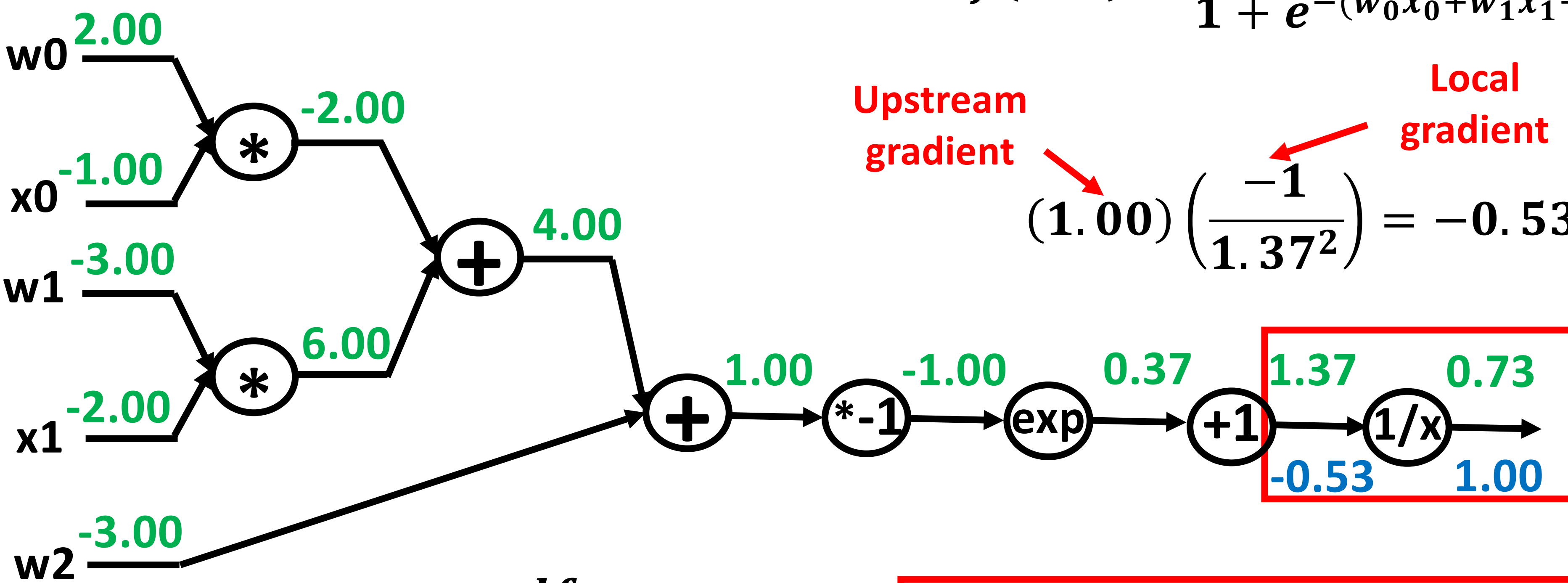
$$f(x) = \frac{1}{x} \quad \Longleftrightarrow \quad \frac{df}{dx} = \frac{-1}{x^2}$$

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

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

Example

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



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

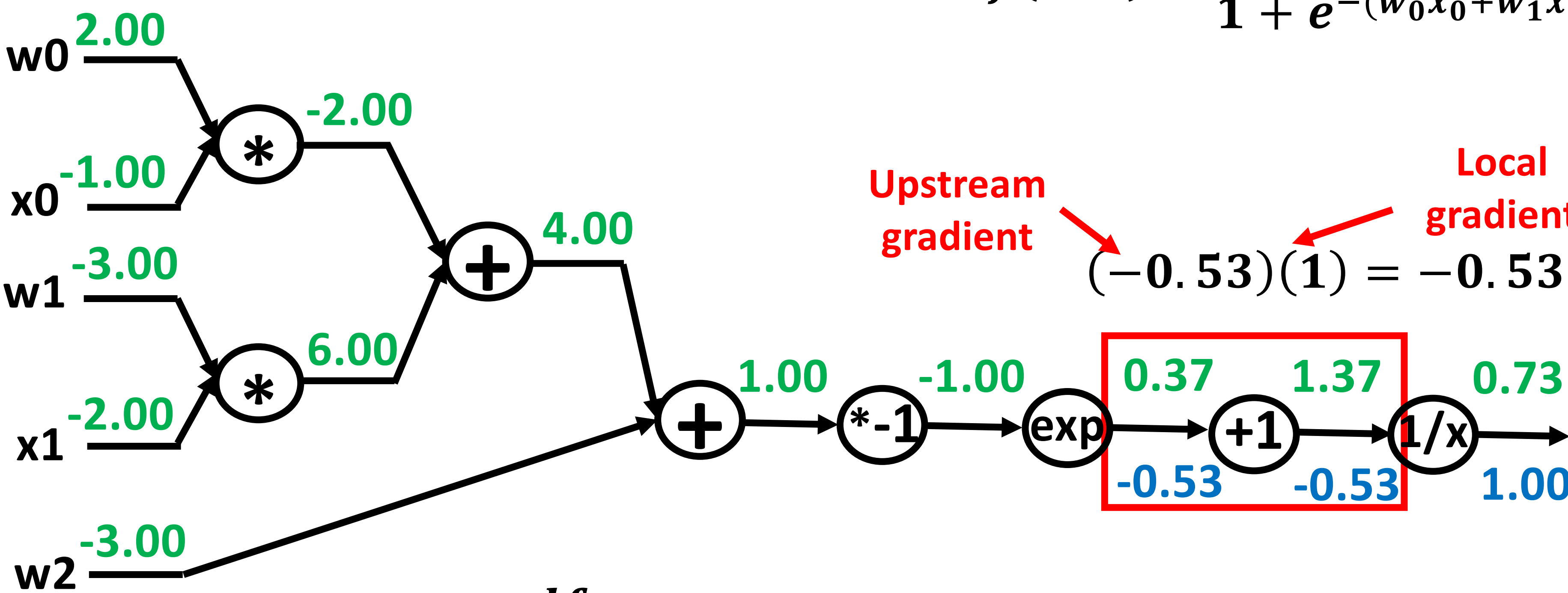
$\boxed{f(x) = \frac{1}{x} \quad \Longleftrightarrow \quad \frac{df}{dx} = \frac{-1}{x^2}}$

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

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

Example

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



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

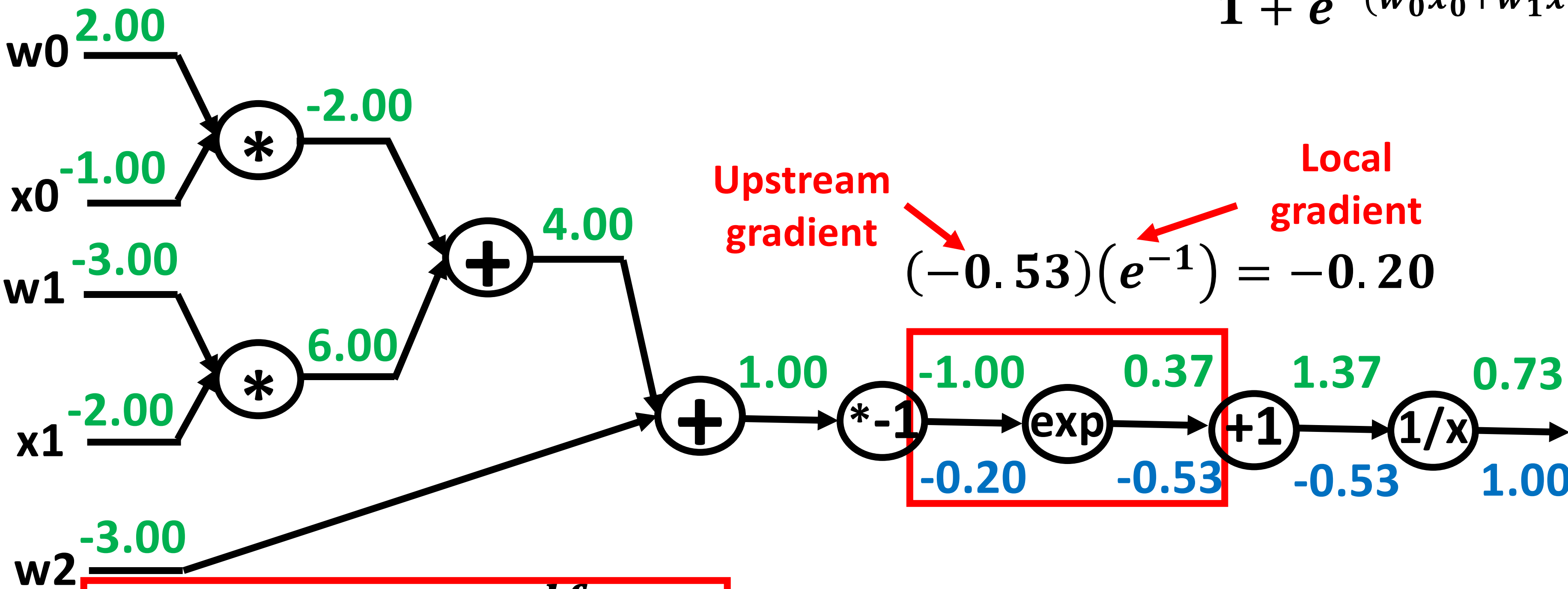
$$f(x) = \frac{1}{x} \quad \Longleftrightarrow \quad \frac{df}{dx} = \frac{-1}{x^2}$$

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

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

Example

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



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

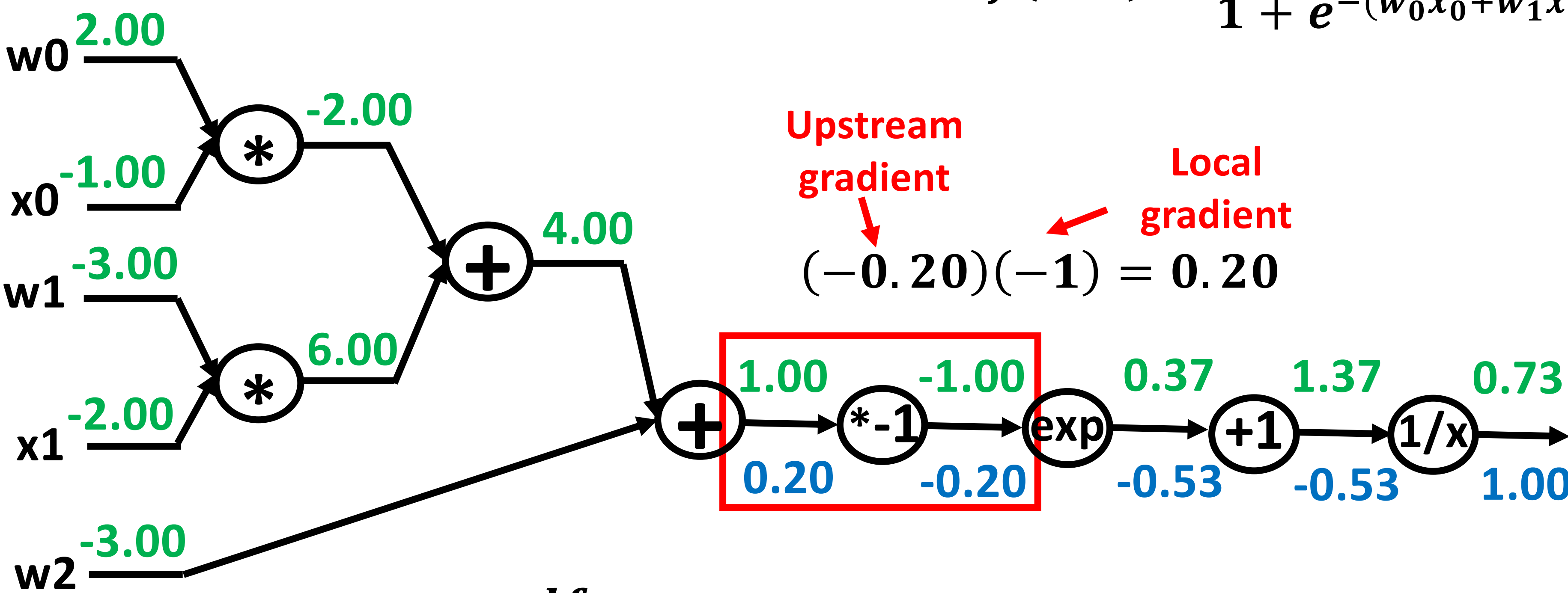
$$f(x) = \frac{1}{x} \quad \Longleftrightarrow \quad \frac{df}{dx} = \frac{-1}{x^2}$$

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

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

Example

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



$$(-0.20)(-1) = 0.20$$

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

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

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

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

ADD acts as *distributor*

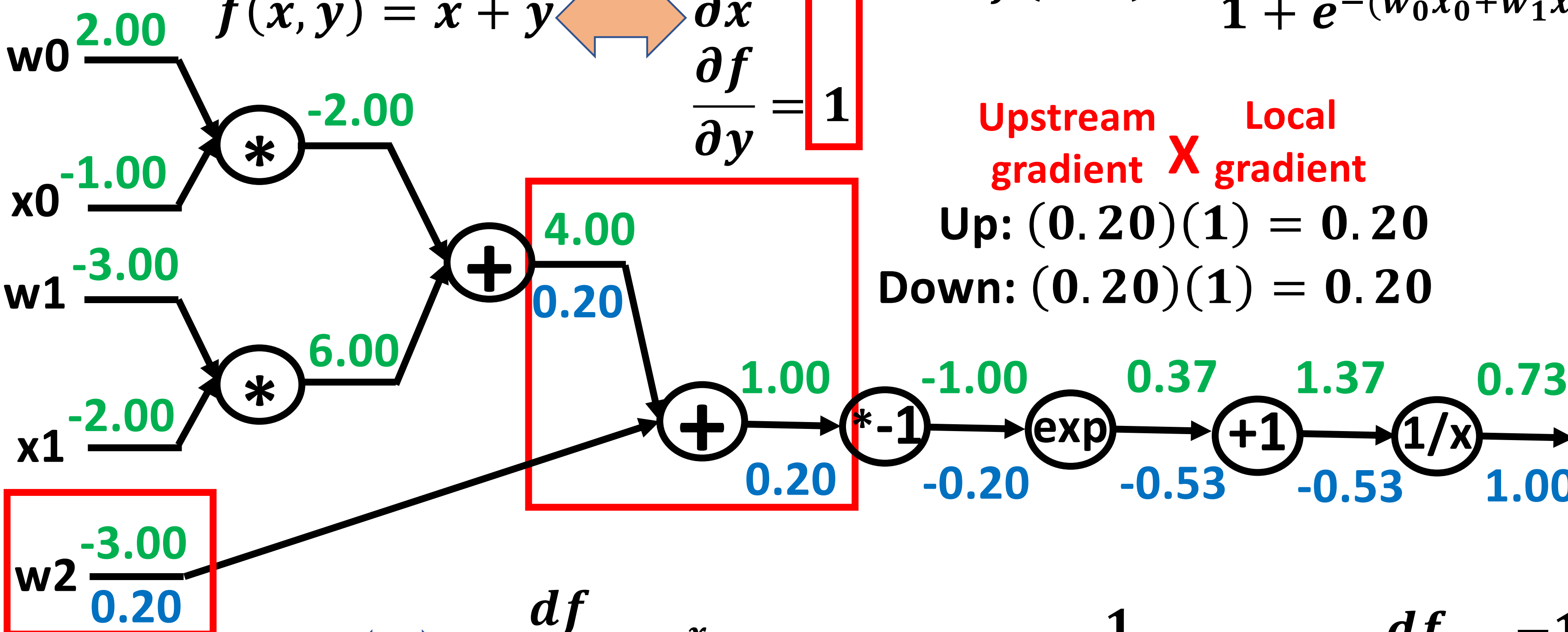
$$f(x,y) = x + y \Leftrightarrow \frac{\partial f}{\partial x} = \boxed{1} \qquad \frac{\partial f}{\partial y} = \boxed{1}$$

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

Upstream gradient X Local gradient

Up: $(0.20)(1) = 0.20$

Down: $(0.20)(1) = 0.20$



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

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

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

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

MUL acts as *switcher*

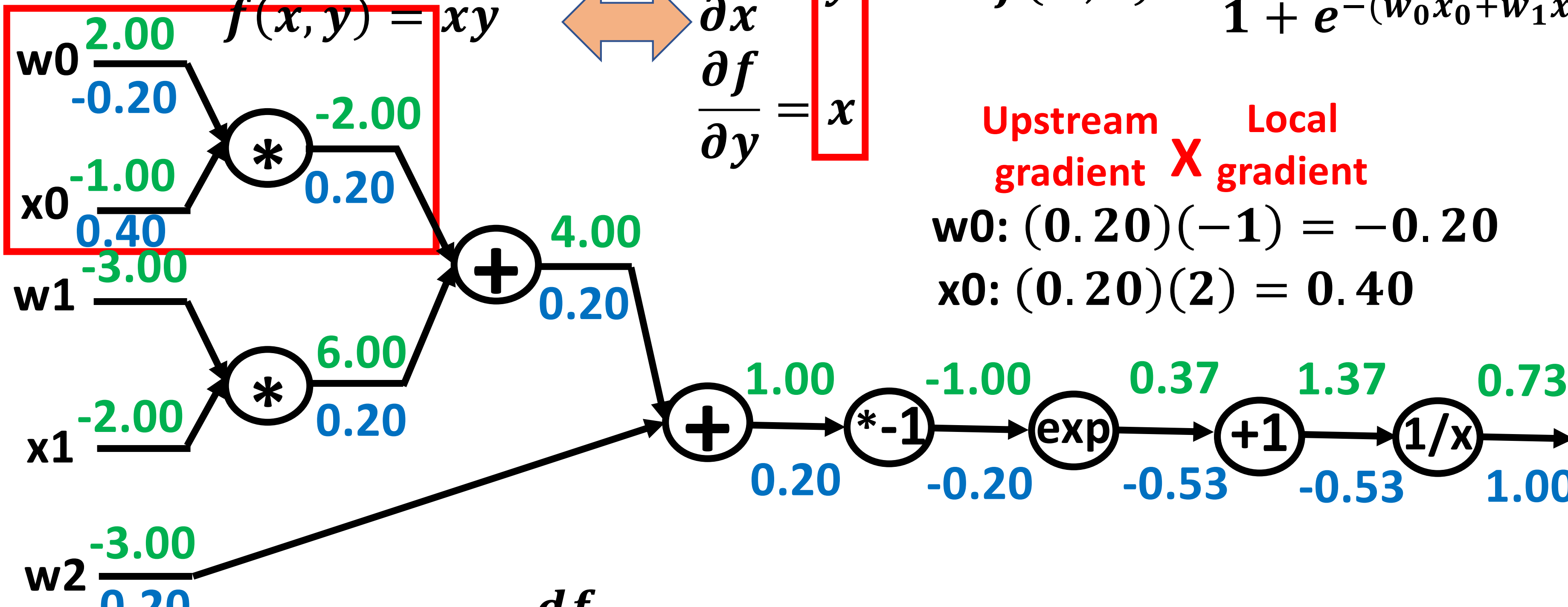
$$f(x,y) = xy \iff \frac{\partial f}{\partial x} = \boxed{y} \qquad \frac{\partial f}{\partial y} = \boxed{x}$$

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

Upstream gradient X Local gradient

w0: $(0.20)(-1) = -0.20$

x0: $(0.20)(2) = 0.40$



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

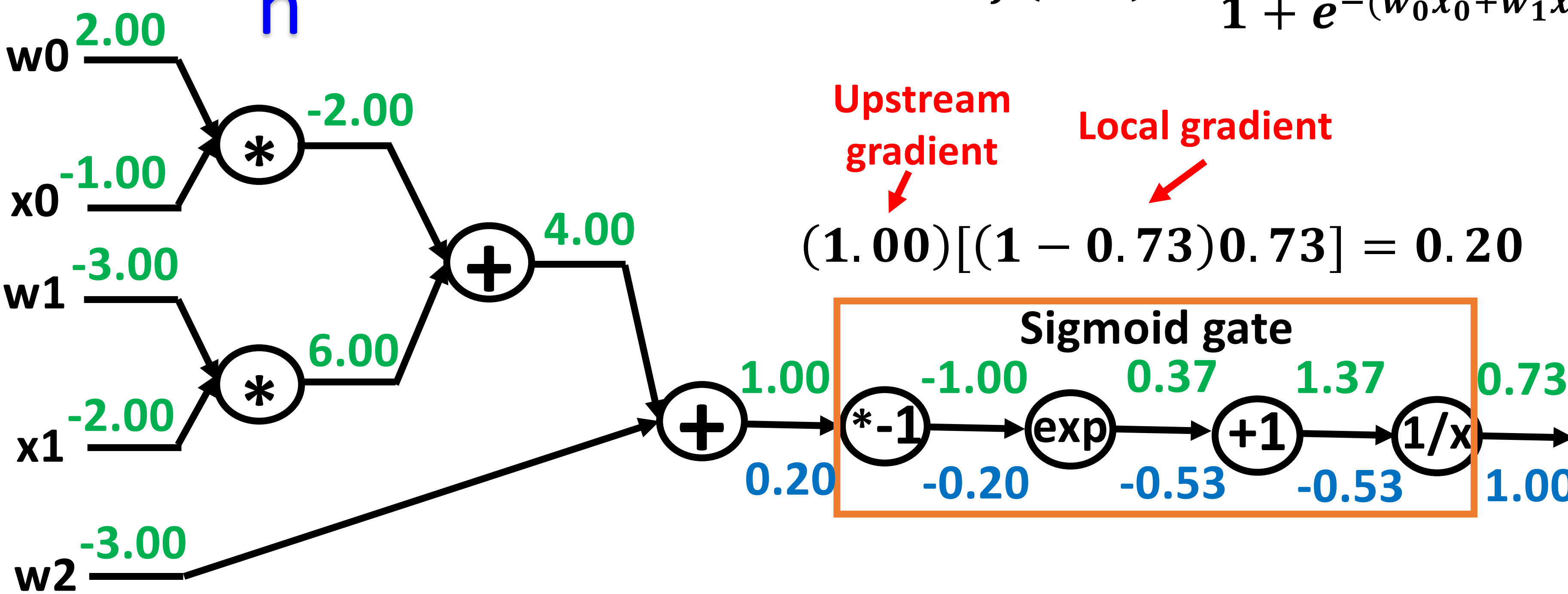
$$f(x) = \frac{1}{x} \iff \frac{df}{dx} = \frac{-1}{x^2}$$

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

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

Simplificatio n

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



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

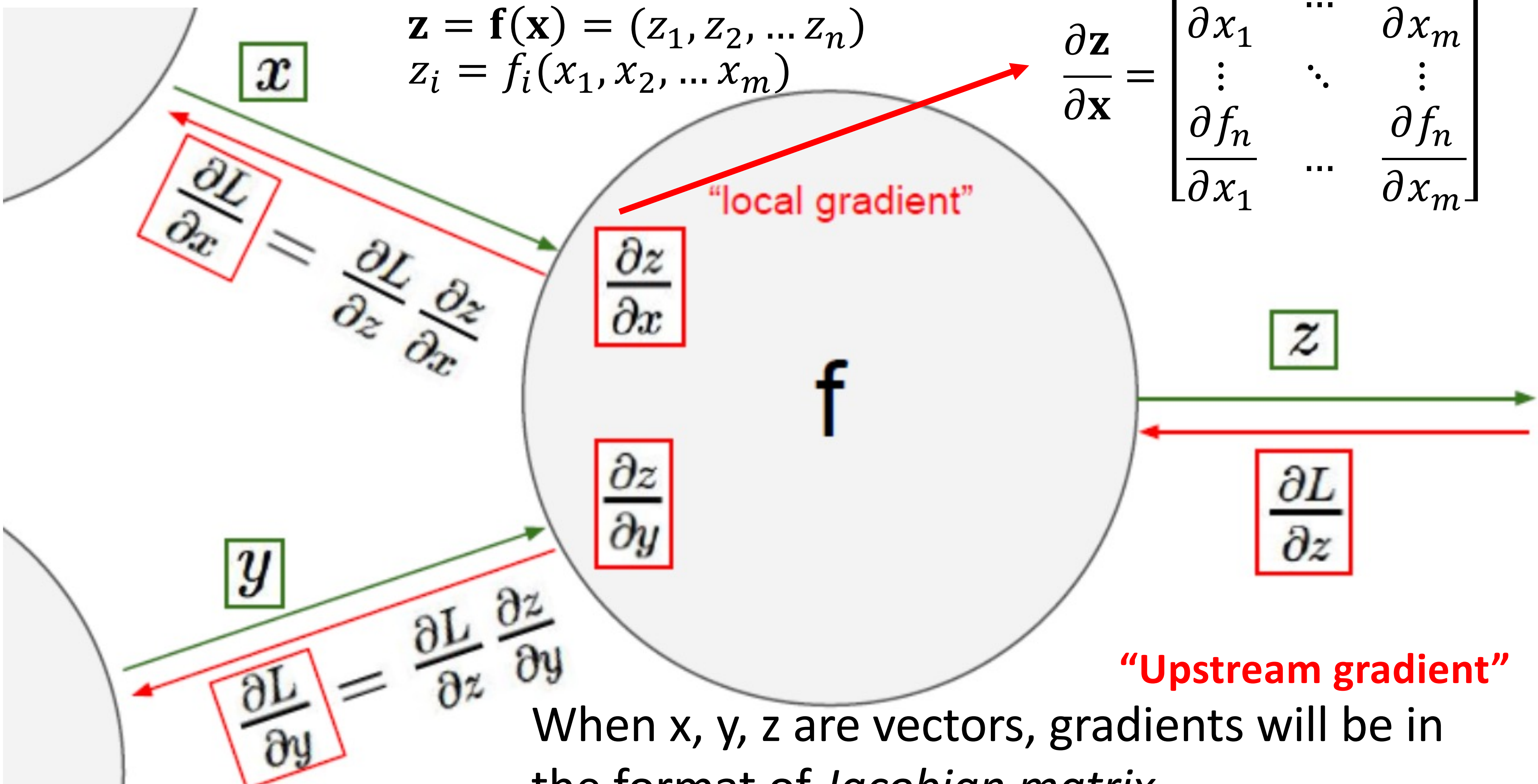
Sigmoid function

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

$$= \boxed{(1-\sigma(x))\sigma(x)}$$

$$\mathbf{x} = (x_1, x_2, \dots x_m)$$
$$\mathbf{z} = \mathbf{f}(\mathbf{x}) = (z_1, z_2, \dots z_n)$$
$$z_i = f_i(x_1, x_2, \dots x_m)$$

$$\frac{\partial \mathbf{z}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_m} \end{bmatrix}$$



When x, y, z are vectors, gradients will be in the format of *Jacobian matrix*

Matrix Calculus Primer

Scalar-by-Vector

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} & \frac{\partial y}{\partial x_2} & \cdots & \frac{\partial y}{\partial x_n} \end{bmatrix}$$

Vector-by-Vector

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

Scalar-by-Matrix

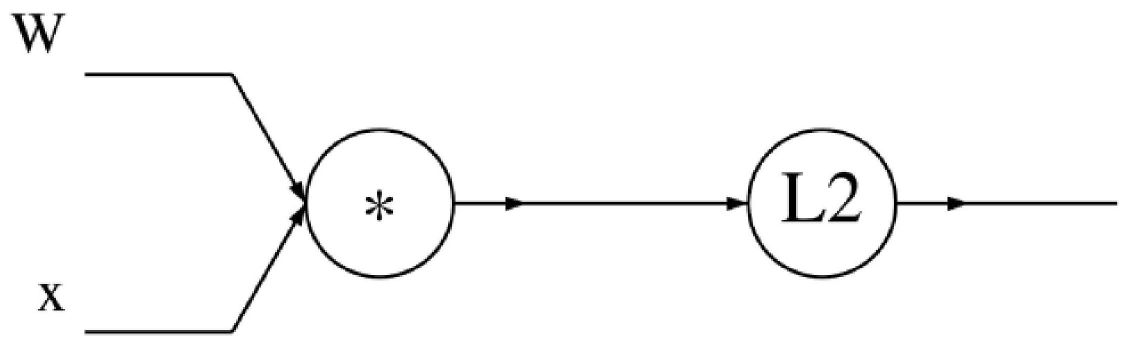
$$\frac{\partial y}{\partial A} = \begin{bmatrix} \frac{\partial y}{\partial A_{11}} & \frac{\partial y}{\partial A_{12}} & \cdots & \frac{\partial y}{\partial A_{1n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial A_{m1}} & \frac{\partial y}{\partial A_{m2}} & \cdots & \frac{\partial y}{\partial A_{mn}} \end{bmatrix}$$

Vectorized Example

$$f(x, W) = ||W \cdot x||^2 = \sum_{i=1}^{n} (W \cdot x)_i^2$$

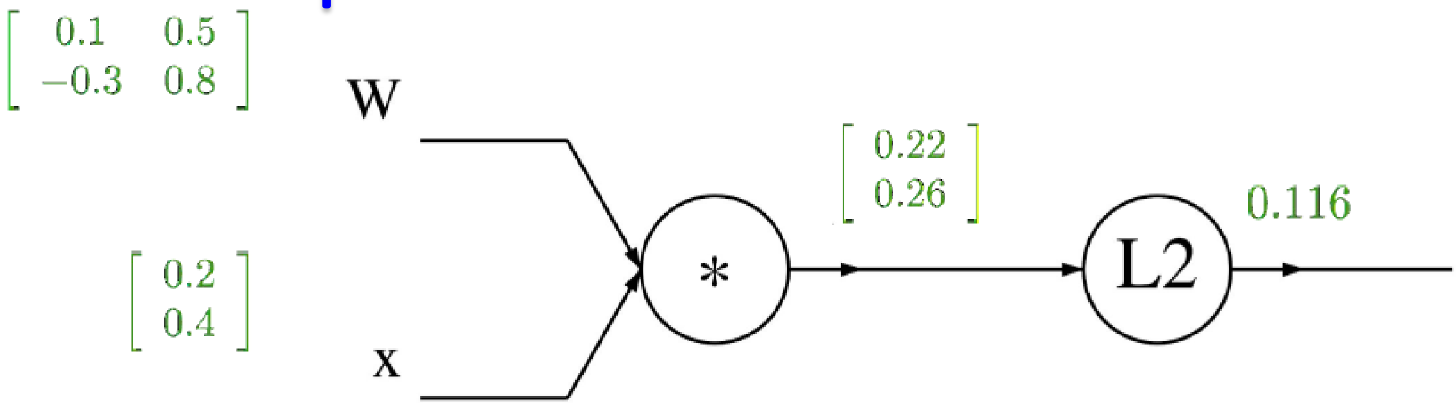
Vectorized Example

$$f(x, W) = ||W \cdot x||^2 = \sum_{i=1}^{n} (W \cdot x)_i^2$$



Vectorized Example

$$f(x, W) = ||W \cdot x||^2 = \sum_{i=1}^{n} (W \cdot x)_i^2$$

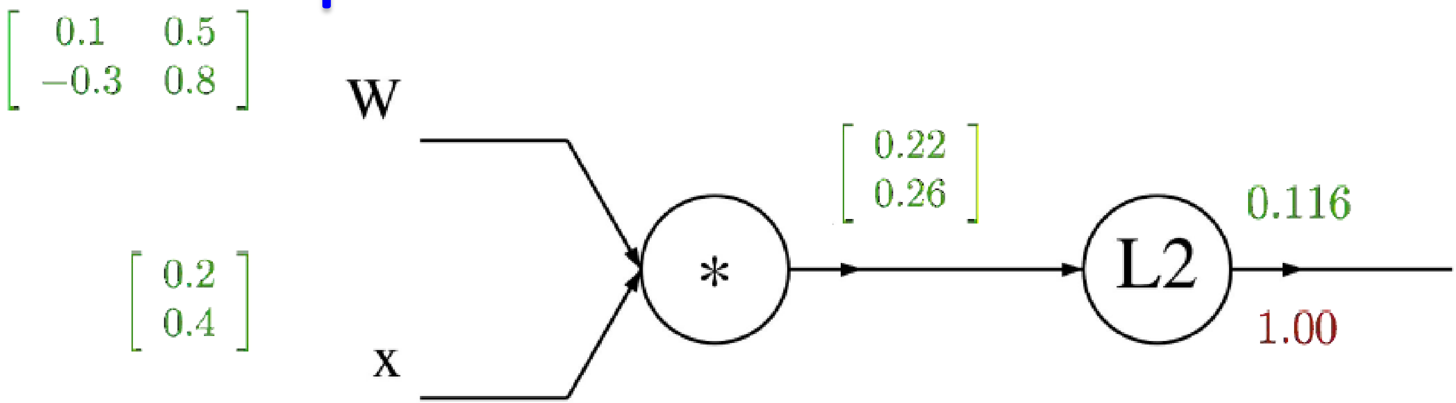


$$q = W \cdot x = \begin{pmatrix} W_{1,1}x_1 + \cdots + W_{1,n}x_n \\ \vdots \\ W_{n,1}x_1 + \cdots + W_{n,n}x_n \end{pmatrix}$$

$$f(q) = ||q||^2 = q_1^2 + \cdots + q_n^2$$

Vectorized Example

$$f(x, W) = ||W \cdot x||^2 = \sum_{i=1}^{n} (W \cdot x)_i^2$$

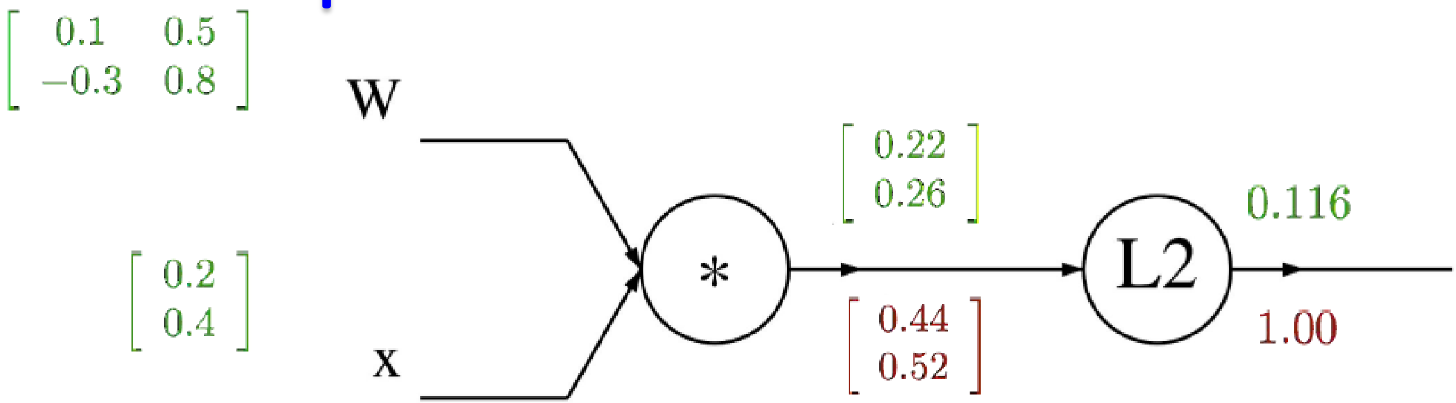


$$q = W \cdot x = \begin{pmatrix} W_{1,1}x_1 + \cdots + W_{1,n}x_n \\ \vdots \\ W_{n,1}x_1 + \cdots + W_{n,n}x_n \end{pmatrix}$$

$$f(q) = ||q||^2 = q_1^2 + \cdots + q_n^2$$

Vectorized Example

$$f(x, W) = ||W \cdot x||^2 = \sum_{i=1}^{n} (W \cdot x)_i^2$$



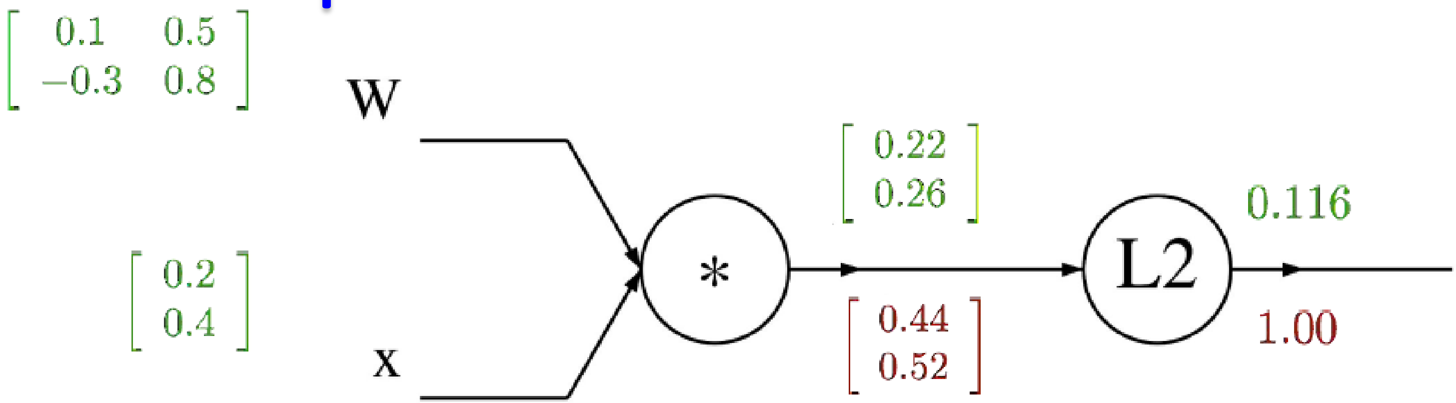
$$\frac{\partial f}{\partial q_i} = 2q_i \qquad \boxed{\nabla_q f = 2q}$$

$$q = W \cdot x = \begin{pmatrix} W_{1,1}x_1 + \cdots + W_{1,n}x_n \\ \vdots \\ W_{n,1}x_1 + \cdots + W_{n,n}x_n \end{pmatrix}$$

$$\boxed{f(q) = ||q||^2 = q_1^2 + \cdots + q_n^2}$$

Vectorized Example

$$f(x, W) = ||W \cdot x||^2 = \sum_{i=1}^{n} (W \cdot x)_i^2$$



$$\frac{\partial f}{\partial q_i} = 2q_i \qquad \boxed{\nabla_q f = 2q}$$

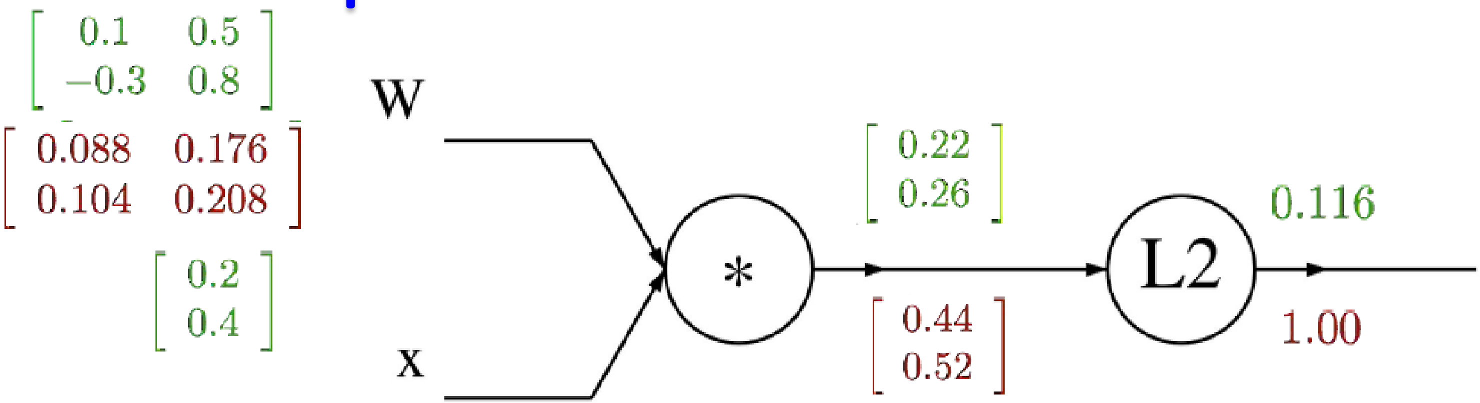
$$\frac{\partial q_k}{\partial W_{i,j}} = \mathbf{1}_{k=i} x_j$$

$$\boxed{q = W \cdot x = \begin{pmatrix} W_{1,1}x_1 + \cdots + W_{1,n}x_n \\ \vdots \\ W_{n,1}x_1 + \cdots + W_{n,n}x_n \end{pmatrix}}$$

$$f(q) = ||q||^2 = q_1^2 + \cdots + q_n^2$$

Vectorized Example

$$f(x, W) = ||W \cdot x||^2 = \sum_{i=1}^{n} (W \cdot x)_i^2$$



$$\frac{\partial f}{\partial q_i} = 2q_i \qquad \boxed{\nabla_q f = 2q}$$

$$\frac{\partial q_k}{\partial W_{i,j}} = \mathbf{1}_{k=i} x_j$$

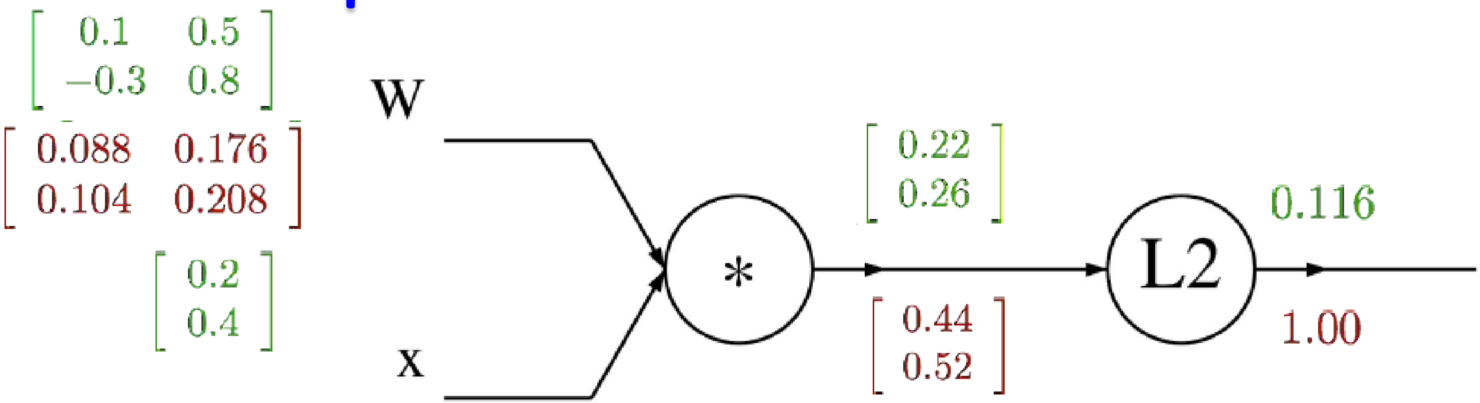
$$q = W \cdot x = \begin{pmatrix} W_{1,1}x_1 + \cdots + W_{1,n}x_n \\ \vdots \\ W_{n,1}x_1 + \cdots + W_{n,n}x_n \end{pmatrix}$$

$$\begin{aligned} \frac{\partial f}{\partial W_{i,j}} &= \sum_k \frac{\partial f}{\partial q_k} \frac{\partial q_k}{\partial W_{i,j}} \\ &= \sum_k (2q_k)(\mathbf{1}_{k=i} x_j) \\ &= 2q_i x_j \end{aligned} \qquad \boxed{\nabla_W f = 2q \cdot x^T}$$

$$f(q) = ||q||^2 = q_1^2 + \cdots + q_n^2$$

Vectorized Example

$$f(x, W) = ||W \cdot x||^2 = \sum_{i=1}^{n} (W \cdot x)_i^2$$



$$\frac{\partial f}{\partial q_i} = 2q_i \qquad \boxed{\nabla_q f = 2q}$$

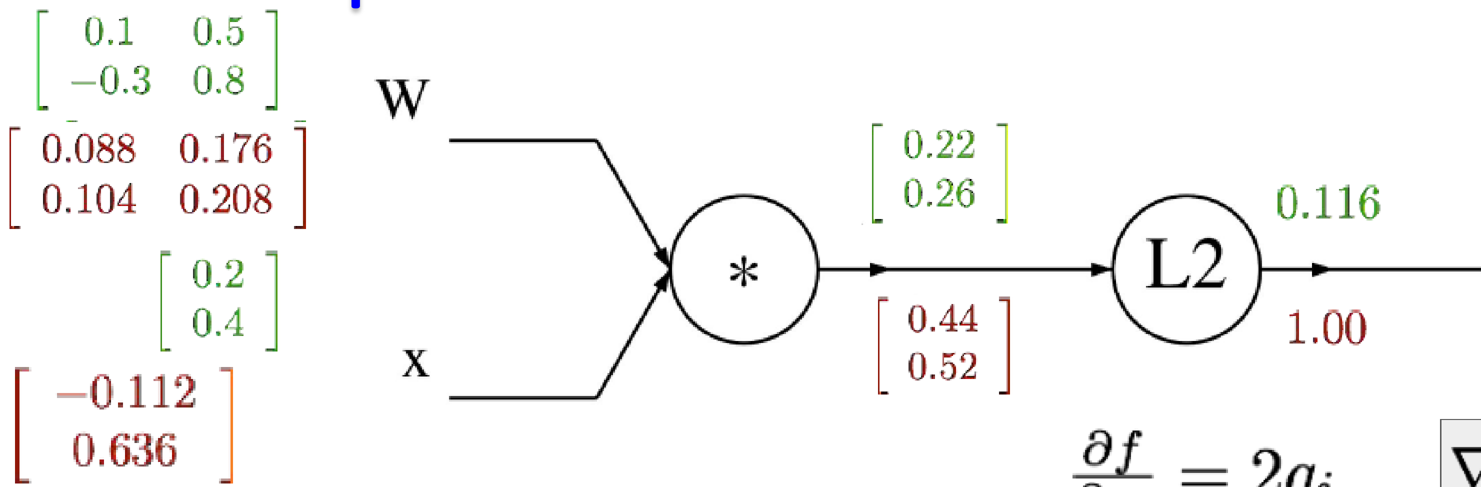
$$\frac{\partial q_k}{\partial x_i} = W_{k,i}$$

$$\boxed{q = W \cdot x = \begin{pmatrix} W_{1,1}x_1 + \cdots + W_{1,n}x_n \\ \vdots \\ W_{n,1}x_1 + \cdots + W_{n,n}x_n \end{pmatrix}}$$

$$f(q) = ||q||^2 = q_1^2 + \cdots + q_n^2$$

Vectorized Example

$$f(x, W) = ||W \cdot x||^2 = \sum_{i=1}^{n} (W \cdot x)_i^2$$



$$q = W \cdot x = \begin{pmatrix} W_{1,1}x_1 + \cdots + W_{1,n}x_n \\ \vdots \\ W_{n,1}x_1 + \cdots + W_{n,n}x_n \end{pmatrix}$$

$$f(q) = ||q||^2 = q_1^2 + \cdots + q_n^2$$

$$\frac{\partial f}{\partial q_i} = 2q_i \qquad \boxed{\nabla_q f = 2q}$$

$$\frac{\partial q_k}{\partial x_i} = W_{k,i}$$

$$\frac{\partial f}{\partial x_i} = \sum_k \frac{\partial f}{\partial q_k} \frac{\partial q_k}{\partial x_i} = \sum_k 2q_k W_{k,i} \qquad \boxed{\nabla_x f = 2W^T \cdot q}$$

Acknowledgement

Many materials of the slides of this course are adopted and re-produced from several deep learning courses and tutorials.

- Prof. Fei-fei Li, Stanford, CS231n: Convolutional Neural Networks for Visual Recognition (online available)
- Prof. Andrew Ng, Stanford, CS230: Deep learning (online available)
- Prof. Yanzhi Wang, Northeastern, EECE7390: Advance in deep learning
- Prof. Jianting Zhang, CUNY, CSc G0815 High-Performance Machine Learning: Systems and Applications
- Prof. Vivienne Sze, MIT, "Tutorial on Hardware Architectures for Deep Neural Networks"
- Pytorch official tutorial https://pytorch.org/tutorials/