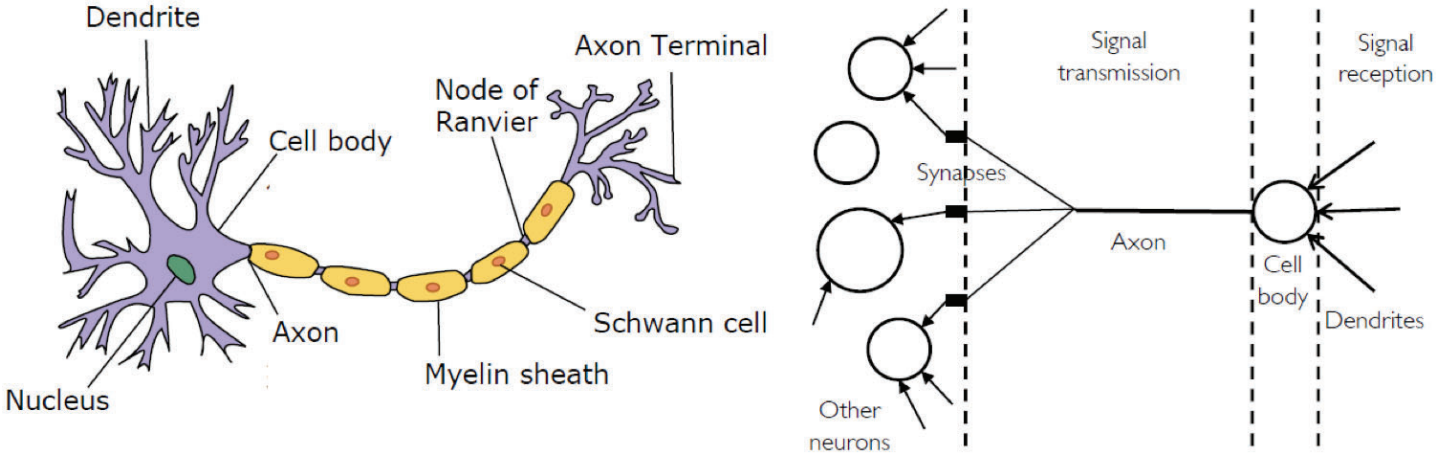
BUSN 20800: Big Data

Lecture 7: Foundations of Neural Network and Deep Learning

Dacheng Xiu

University of Chicago Booth School of Business

Origins of Neural Networks

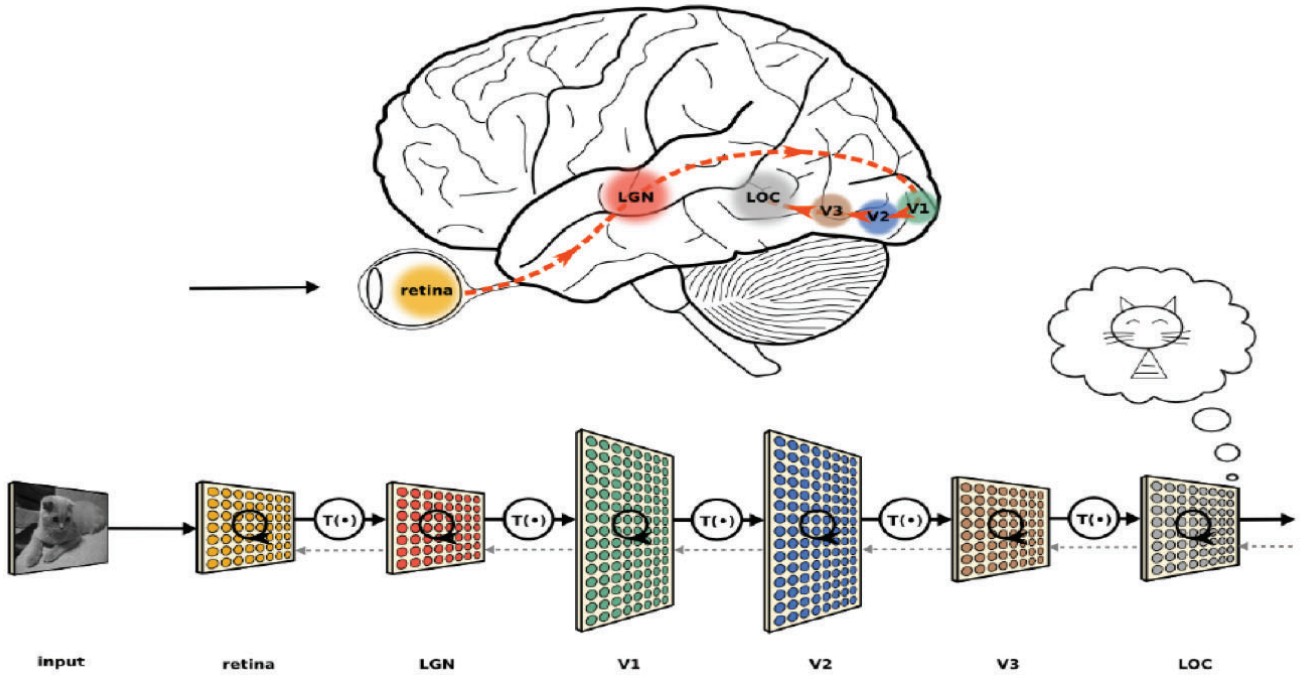


Human brain, through billions of interactions among hundreds of millions of **neurons**, learns to solve (some) complex problems nearly instantly (like face recognition)

- They receive information from other neurons through their dendrites
- They process the information in their cell body (soma)
- They send information through a 'cable' called an axon
- The point of connection between the axon and other neurons' dendrites are called synapses

Neural Networks

Neural networks (aka deep learning) model high-level abstractions in data using architectures consisting of multiple nonlinear transformations.



Why Neural Networks?

- Theoretical underpinnings as "universal approximators" for any smooth predictive association (e.g., Hornik, 1991)
  - though many other nonparametric methods can do the same.
- Flexibility draws from ability to intertwine many telescoping layers of nonlinear predictor interactions (hence synonym "deep learning")
- At same time, their complexity ranks them least transparent, least interpretable, and most highly parameterized ML tools
- Honest answer: They work and it's not entirely clear why

Question: *Are you concerned about research in NN and Deep Learning, being too much results driven without backing with a strong theoretical explanation?*

Yann LeCun's Q&A session on Quora (link)

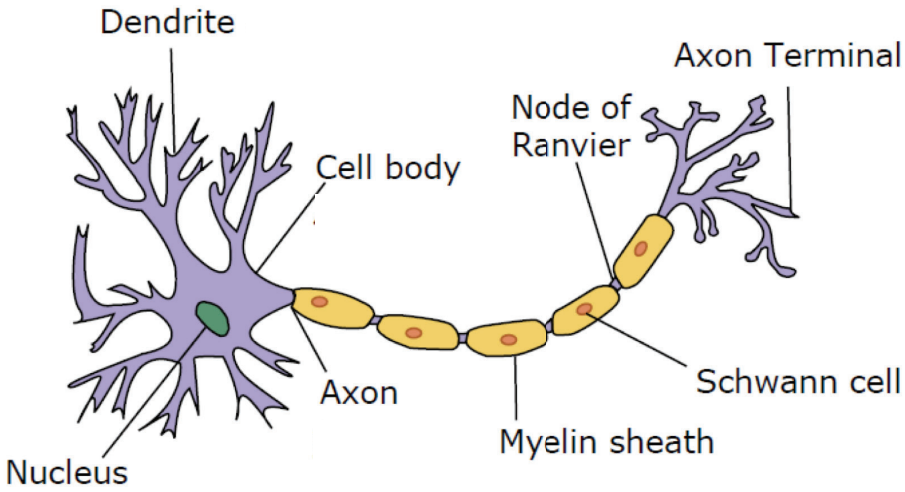
> **Answer:** *I do think that there is a need for better theoretical understanding of deep learning.*
>
> *But if a method works, it* **should not** *be abandoned nor dismissed just because theorists haven't yet figured out how to explain it. The field of machine learning made that mistake in the mid 1990s...*

Michael Nielsen (on neural networks)

> *"I've painted a picture of uncertainty here, stressing that we do not yet have a solid theory of how [models] should be chosen.*
>
> *...*
>
> *On the other hand, we* **shouldn't** *let the lack of a full theory stop us! We have powerful tools already at hand, and can make a lot of progress with those tools."*


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

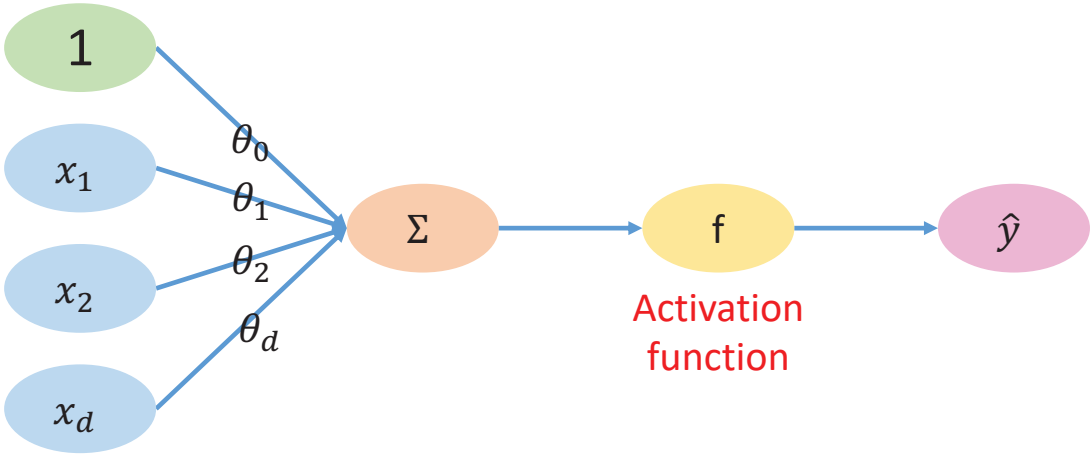
Perceptron

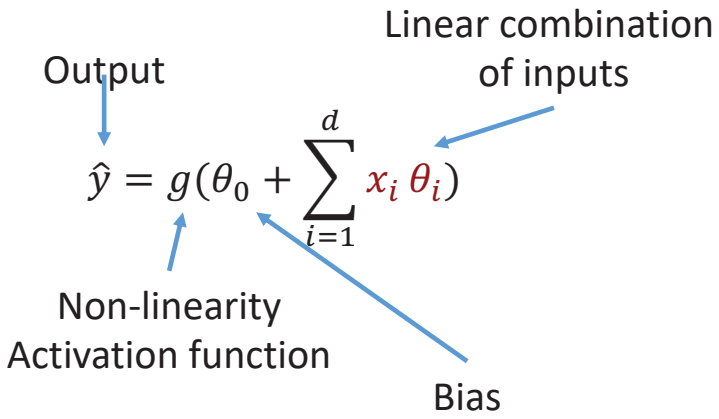


Rosenblatt 1958

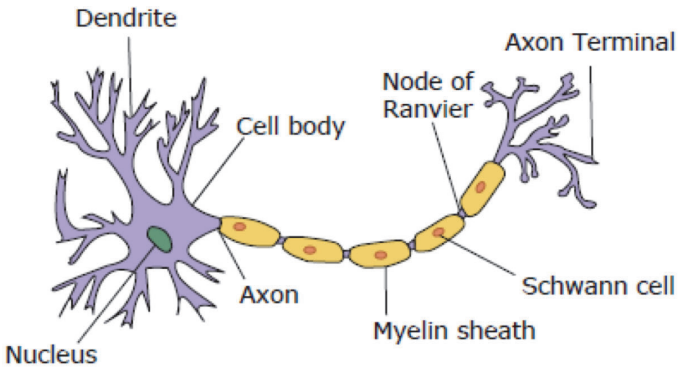
An psychologist

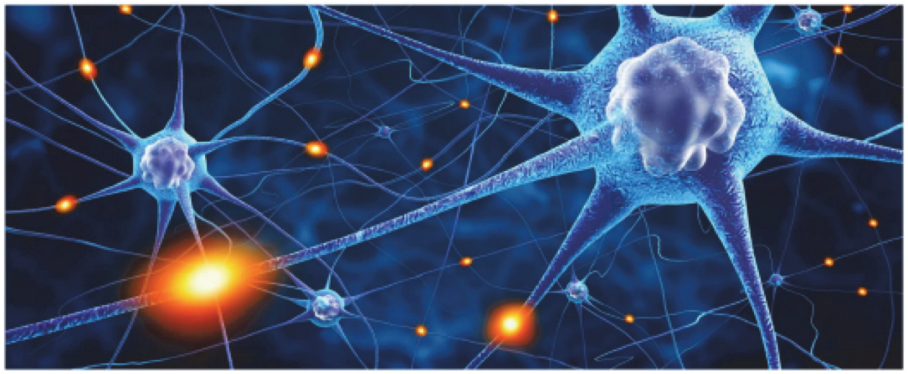
Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms

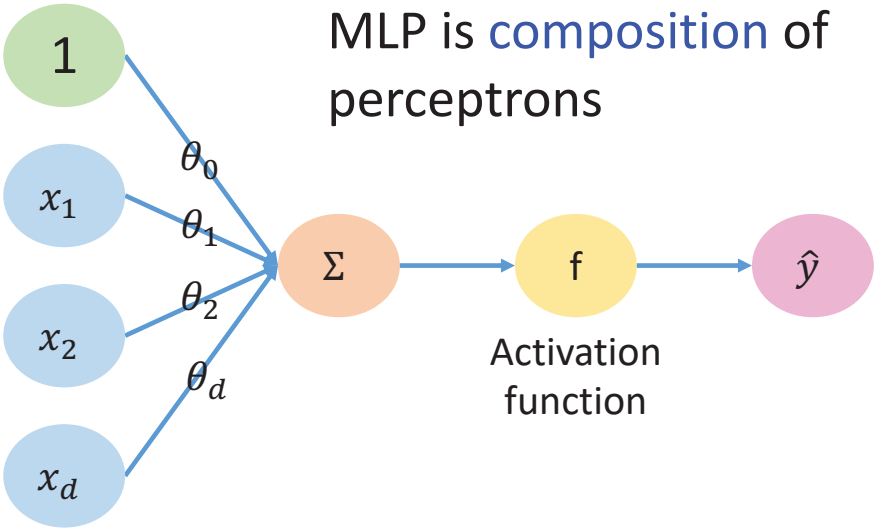




Multilayer Perceptrons (MLP)

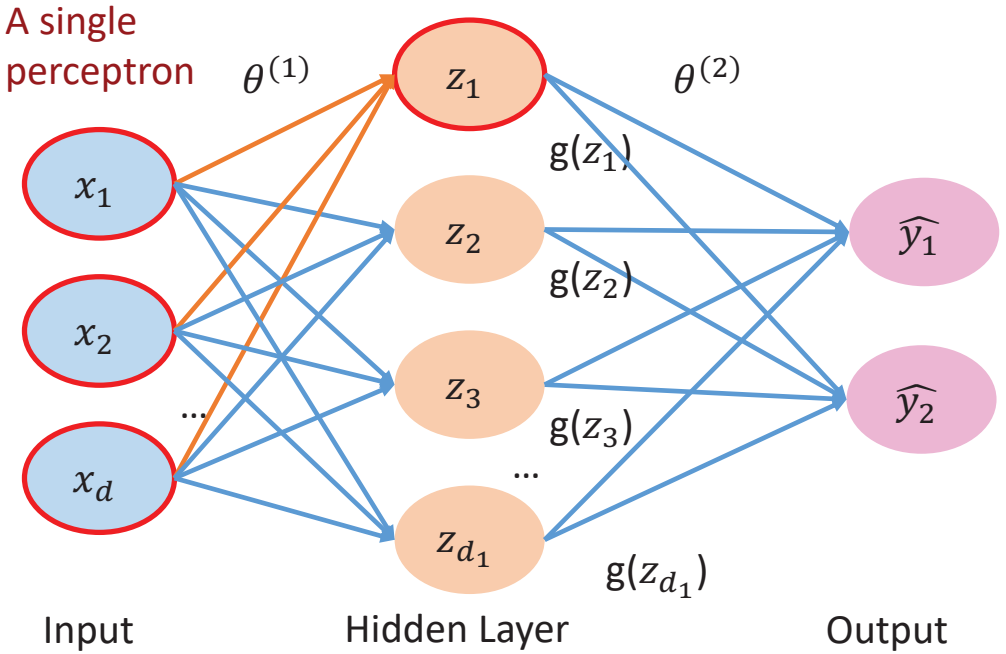




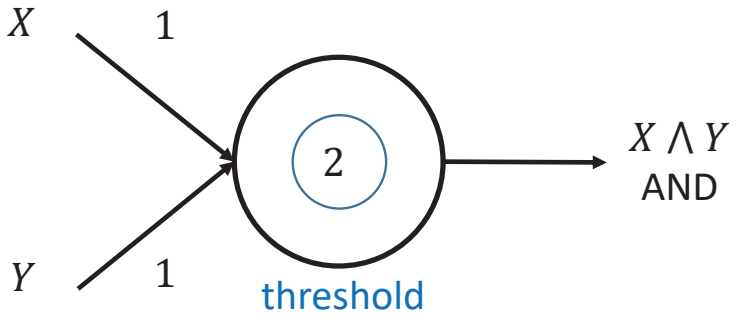


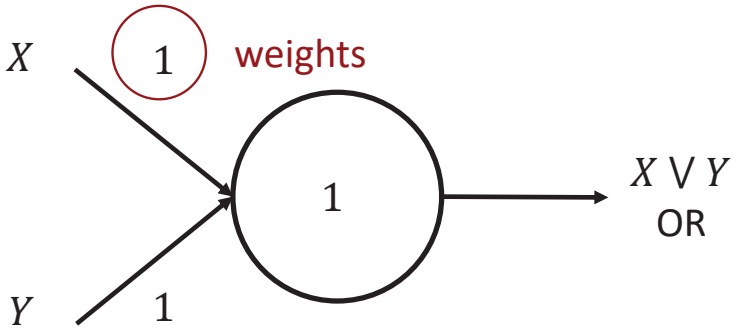
MLP is composition of perceptrons

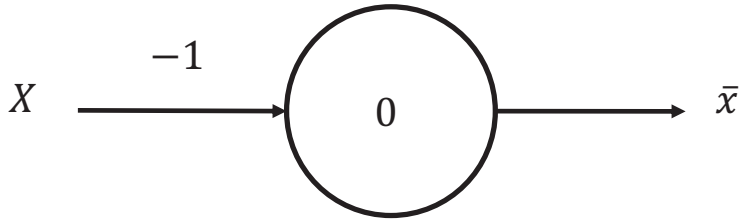
Activation function

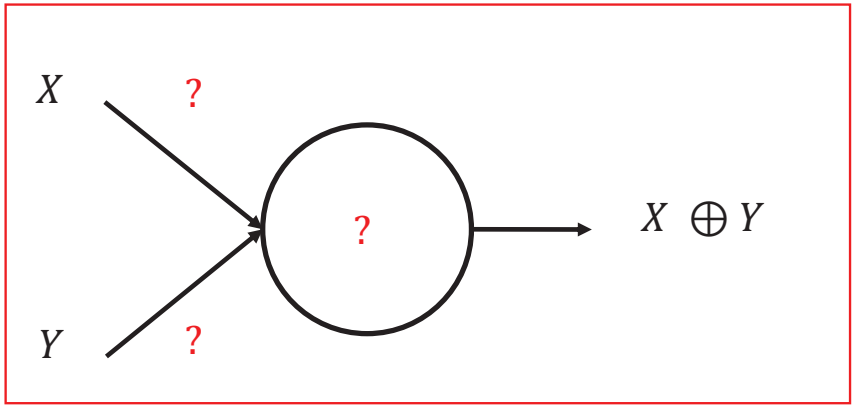


Perceptron: Boolean Functions





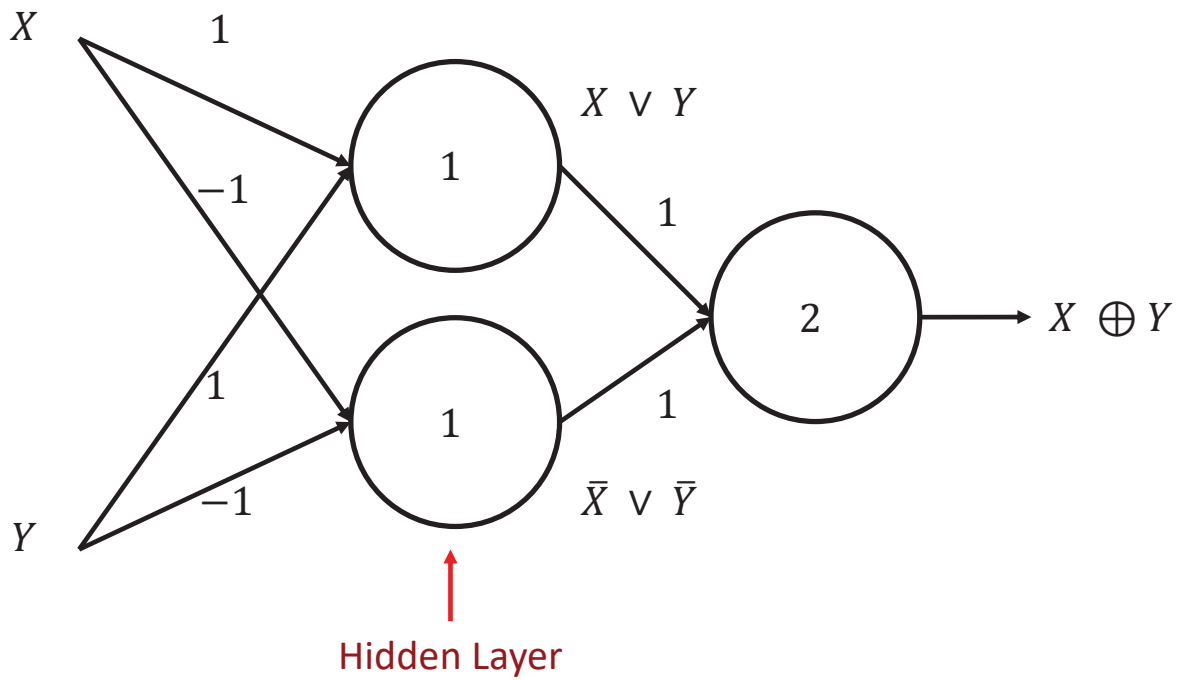




- **Rosenblatt**: Perceptron can represent any Boolean logics
- **Minsky**: Wait, just want to know how to represent XOR?

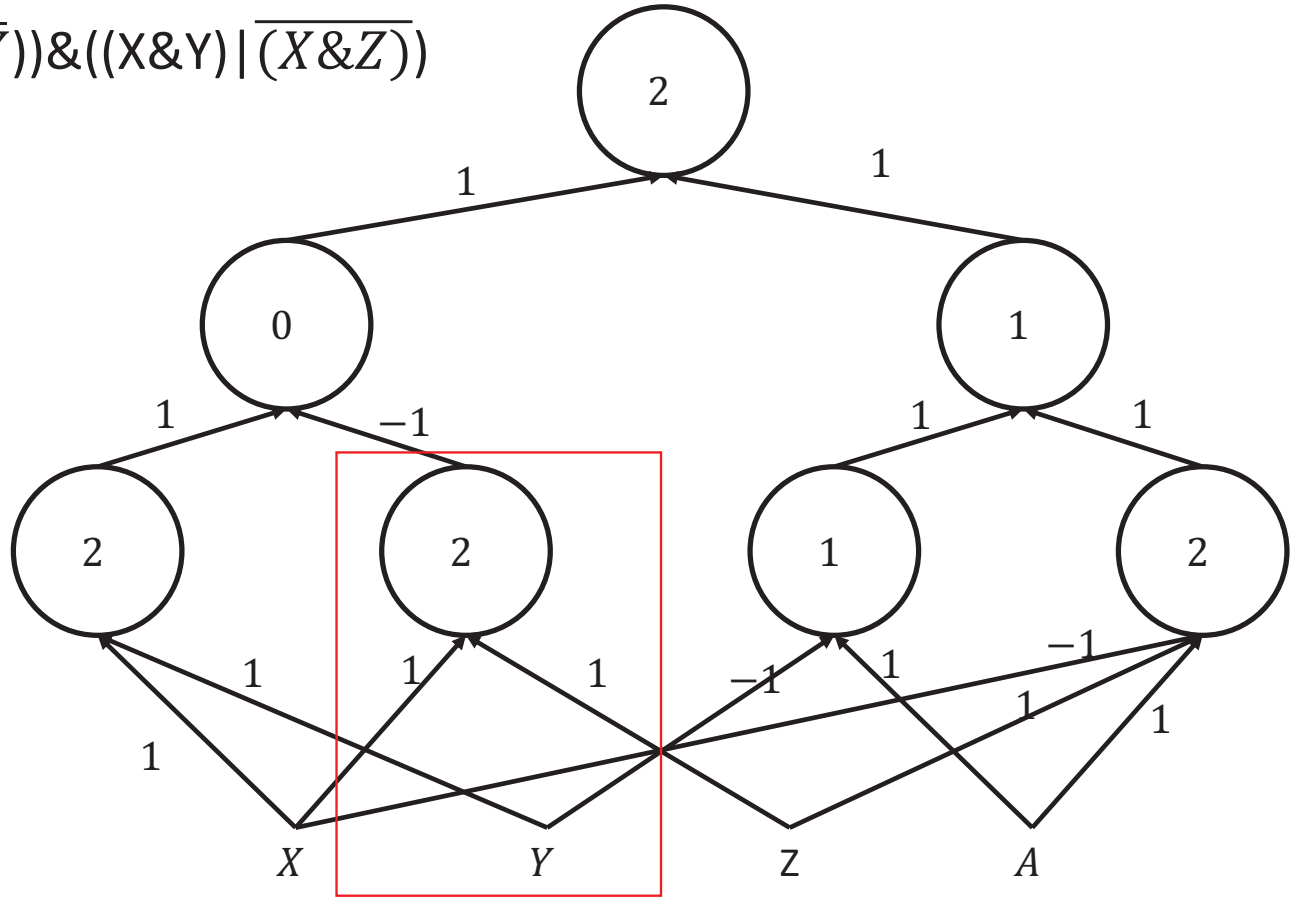
MLP for XOR

$$X \oplus Y = (X \vee Y) \,\&\, (\bar{X} \vee \bar{Y})$$



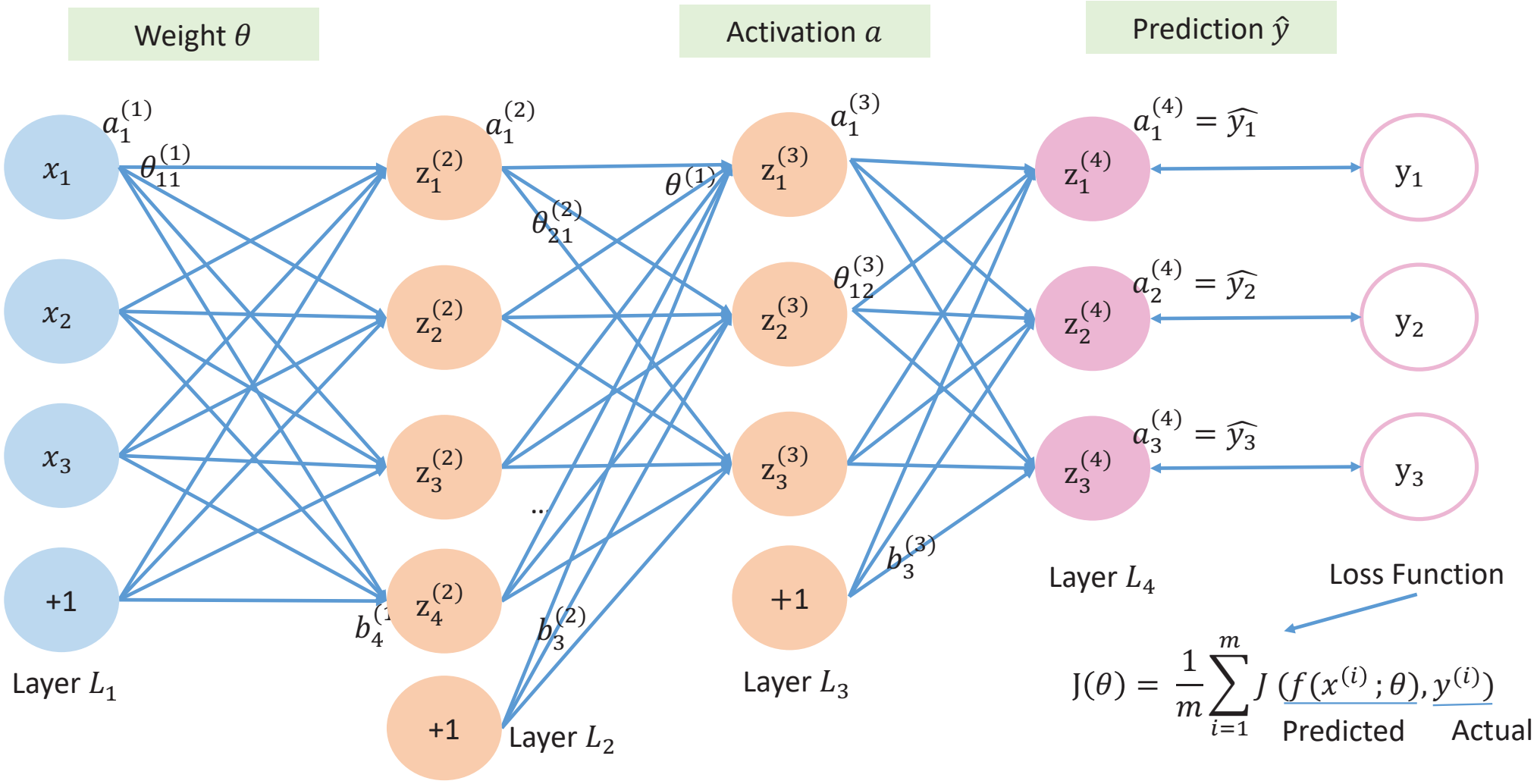
MLP: Boolean Functions

$((A\&\bar{X}\&Z)|A\&\bar{Y}))\&((X\&Y)|\overline{(X\&Z)})$



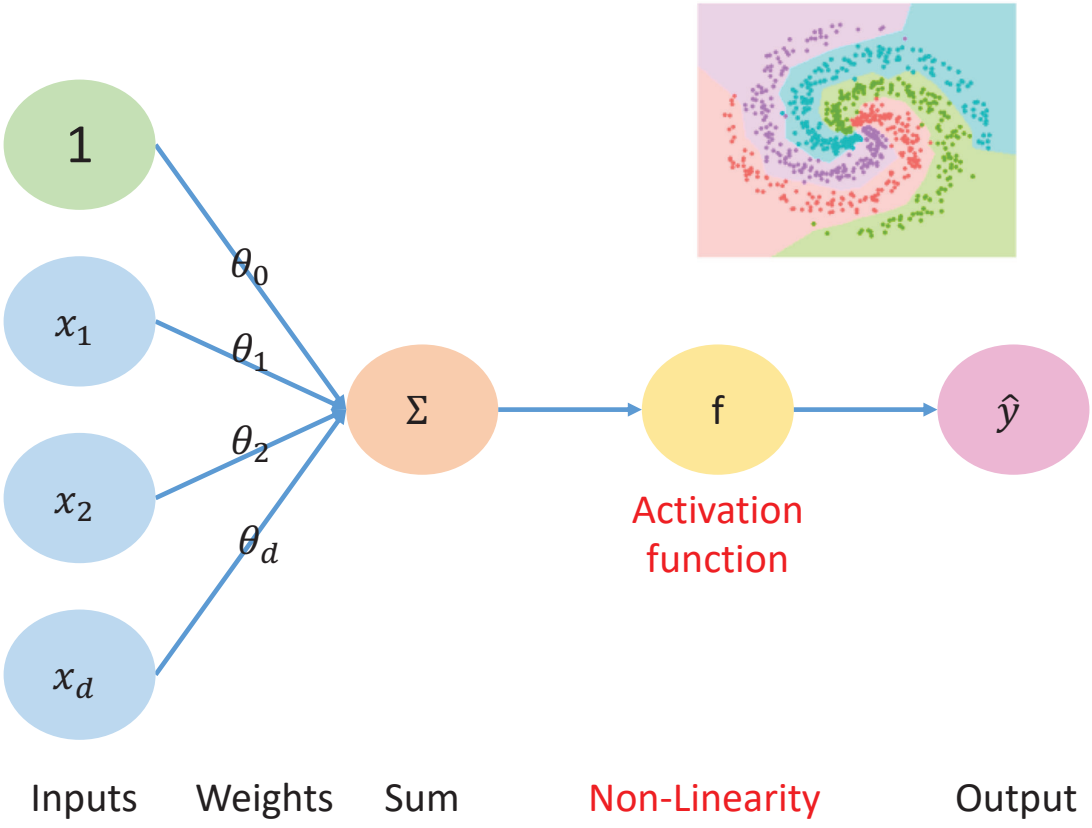
- MLP is universal to represent arbitrarily complex Boolean functions
- The connections in MLP can be sparse - a phenomenon in the Brain

Multilayer Perceptrons (MLP)



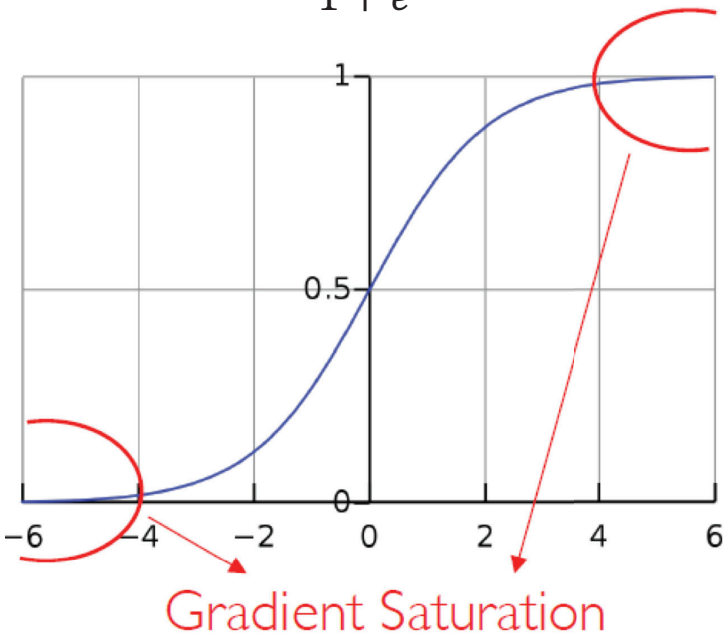
Activation Functions

Non-linearity through activation functions $\hat{y} = g(\theta_0 + X^T\theta)$



Sigmoid function

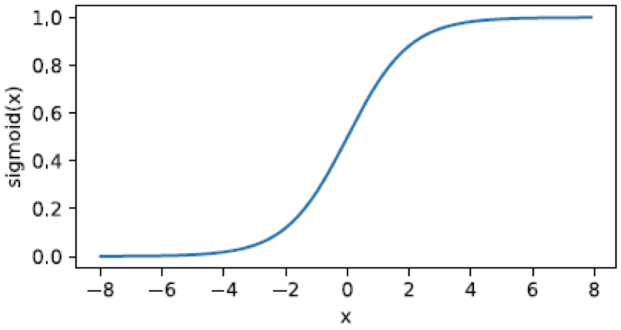
$$g(z) = \sigma(z) = \frac{1}{1 + e^{-z}}$$



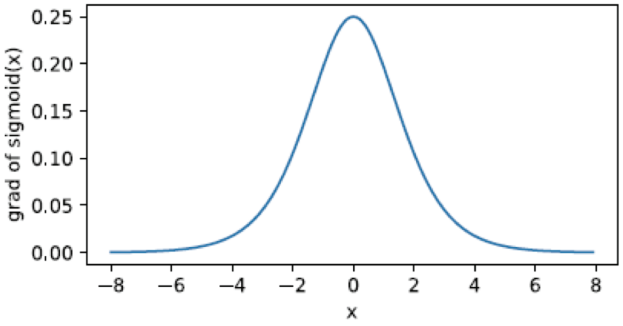
Activation Functions

Sigmoid

$$g(z) = \frac{1}{1 + e^{-z}}$$



$$g'(z) = g(z)(1 - g(z))$$

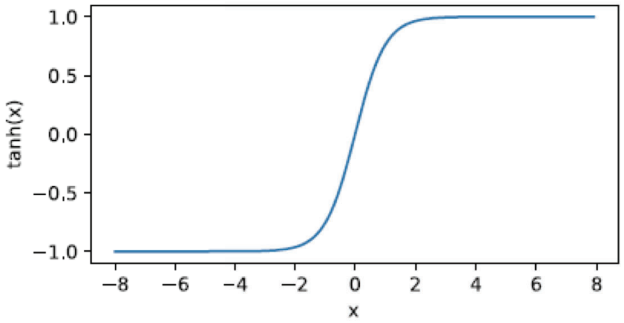


Range in (0,1), probability

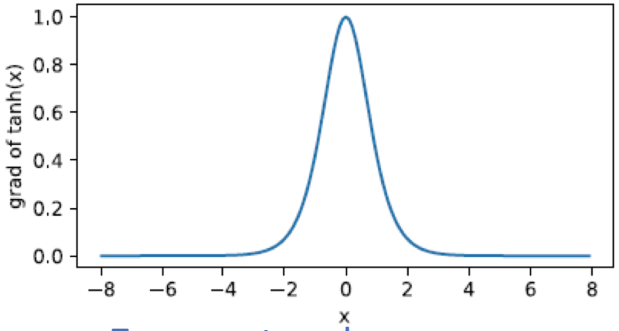
Relative smaller gradient

Hyperbolic Tangent (tanh)

$$g(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$$



$$g'(z) = 1 - g(z)^2$$

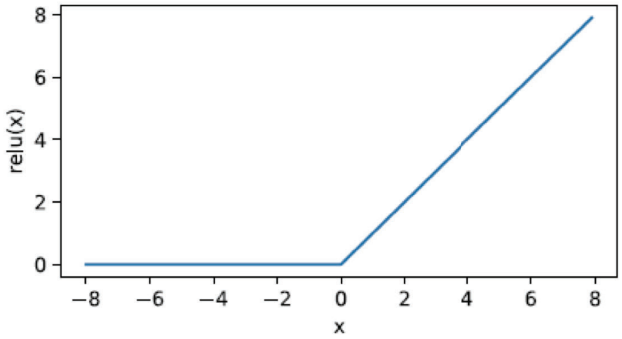


Zero-centered

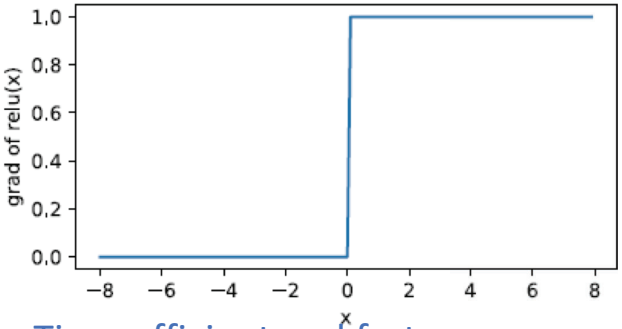
Relatively larger saturation region

Rectified Linear Unit (ReLU)

$$g(z) = \max(0, z)$$



$$g'(z) = sgn(z)$$



Time-efficient and faster convergence,

But neurons are prone to death

Activation Functions

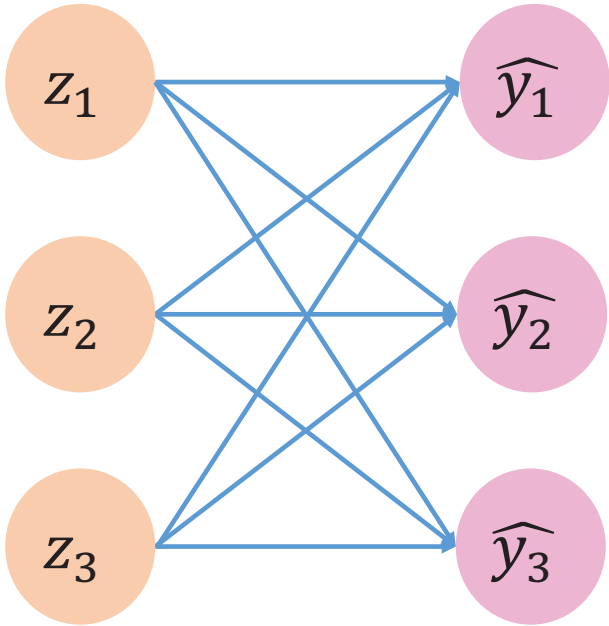
Softmax function

- $g(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}}$

Multiple output
Categorical distribution

Rooster
Cat
Dog
Donkey

- Winner takes all
- The biggest one dominates the outputs



Exponent arithmetic

- May cause numerical overflow
- For numerical stability, implementation is

$$g(z)_i = \frac{e^{z_i - z_m}}{\sum_{j=1}^{k} e^{z_j - z_m}}, m = argmax(z_j)$$

$$z_j \in (-\infty, +\infty), \hat{y}_j \in [0,1], \sum_{j=1}^{k} \hat{y}_j = 1.$$

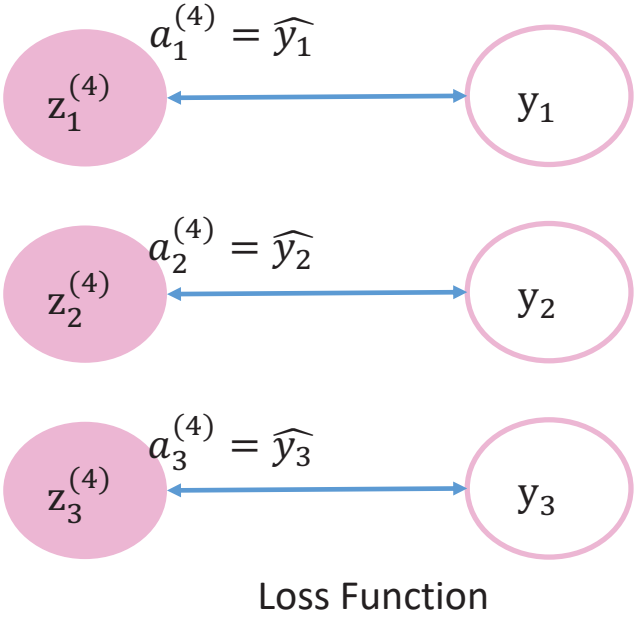
Cost Function

- Softmax function in the output layer

$$\hat{y} = a^{(n_l)} = f_\theta\left(x^{(i)}\right) = \begin{bmatrix} p\left(y^{(i)} = 1 \middle| x^{(i)}; \theta\right) \\ p\left(y^{(i)} = 2 \middle| x^{(i)}; \theta\right) \\ \dots \\ p\left(y^{(i)} = k \middle| x^{(i)}; \theta\right) \end{bmatrix} = \frac{1}{\sum_{j=1}^{k} \exp(z_j^{(n_l)})} \begin{bmatrix} \exp(z_1^{(n_l)}) \\ \exp(z_2^{(n_l)}) \\ \dots \\ \exp(z_k^{(n_l)}) \end{bmatrix}$$

- Cross-entropy $\mathrm{J}(\mathrm{q}, \mathrm{p}) = -\sum_{j=1}^{k} q_j \log p_j$
  - Loss $\mathrm{J}(\mathrm{y}, \hat{y}) = -\sum_{j=1}^{k} y_j \log \hat{y}_j$
  - y follows one-hot coding
  - $y_j = 1 \; if \; y^{(i)} = j \; and \; y_j$ = 0 otherwise
- Cost function

$$\min J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} [\sum_{j=1}^{k} \mathrm{I}\{y^{(i)} = j\} log \frac{\exp(z_j^{(n_l)})}{\sum_{j'=1}^{k} \exp(z_{j'}^{(n_l)})}]$$



Loss Function

Gradient-Based Training

$$argmin\ \mathrm{O}(D;\theta) = \sum_{i=1}^{m} L(y_i, f(x_i);\theta) + \Omega(\theta)$$

Structural Risk Minimization

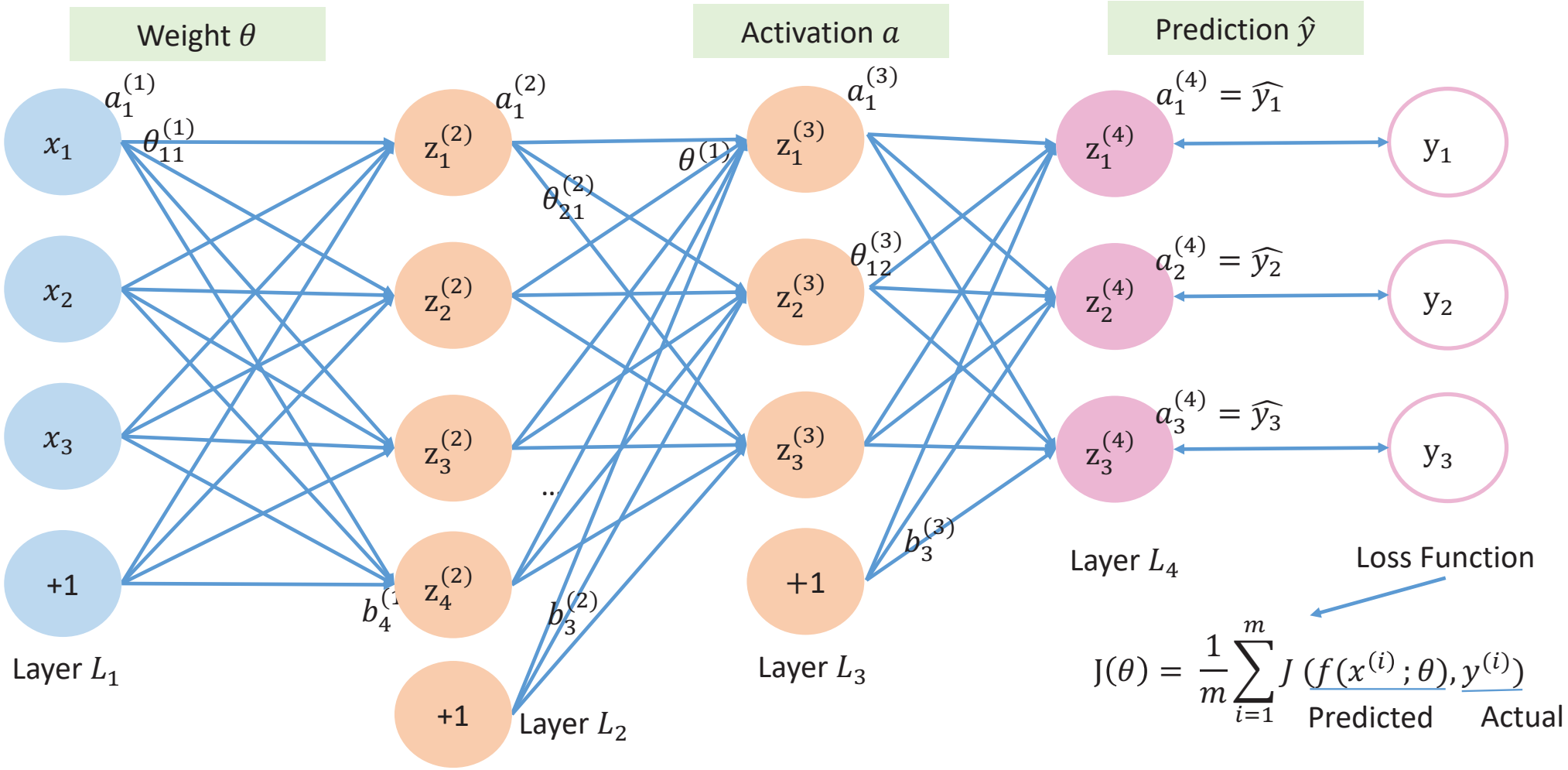
Model

Data

Hypothesis

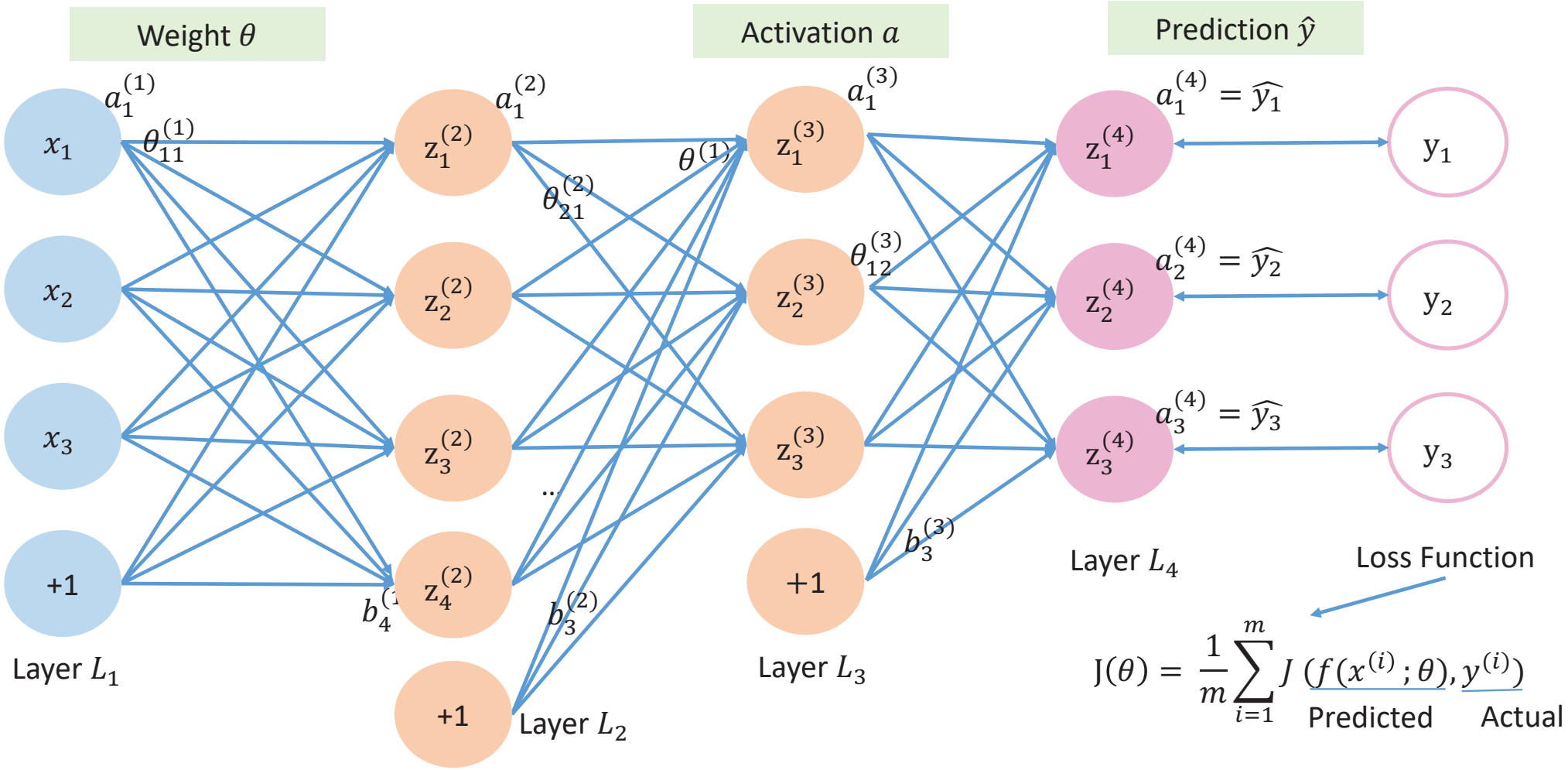
Parameters

Forward Propagation



Forward propagate all (x, y)'s to compute activations and objective $J(\theta, b)$

Backward Propagation

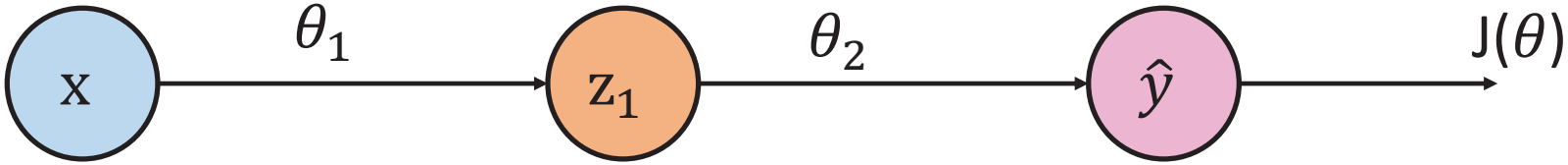


Backward propagate the gradients $\nabla J(\theta, b)$ to update parameters in all layers

Backpropagation (BP) Algorithm

- The backpropagation algorithm applies the **chain rule** of calculus to the parameters θ of each layer

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y}\frac{\partial y}{\partial x}$$



$$\nabla_x z = \left(\frac{\partial y}{\partial x}\right)^T \nabla_y z$$

Jacobian matrix

$$\frac{\partial J(\theta)}{\partial \theta_1} = \frac{\partial J(\theta)}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z_1}\frac{\partial z_1}{\partial \theta_1}$$

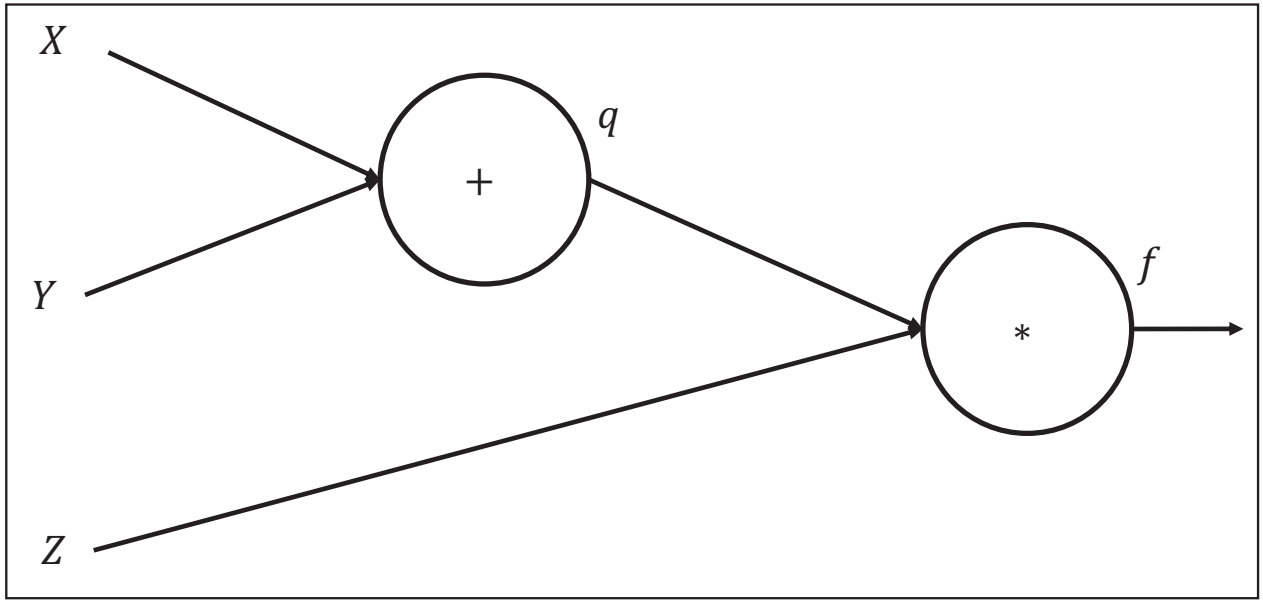
- Back propagation is an efficient implementation of the chain rule
  - Uses dynamic programming (table filling)
  - Avoids re-computing repeated subexpressions (in dependency)
  - Speed vs memory tradeoff (sensitive to #samples m)

Backpropagation: a simple example

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

Backpropagation: a simple example

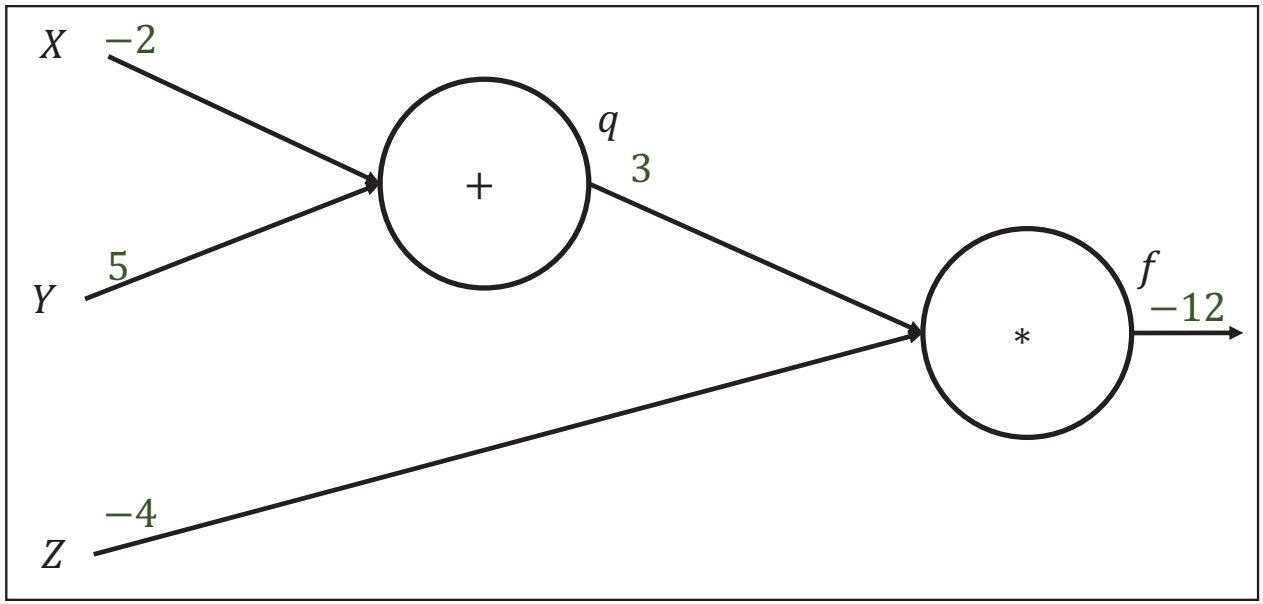
$$f(x, y, z) = (x + y)z$$



Backpropagation: a simple example

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

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

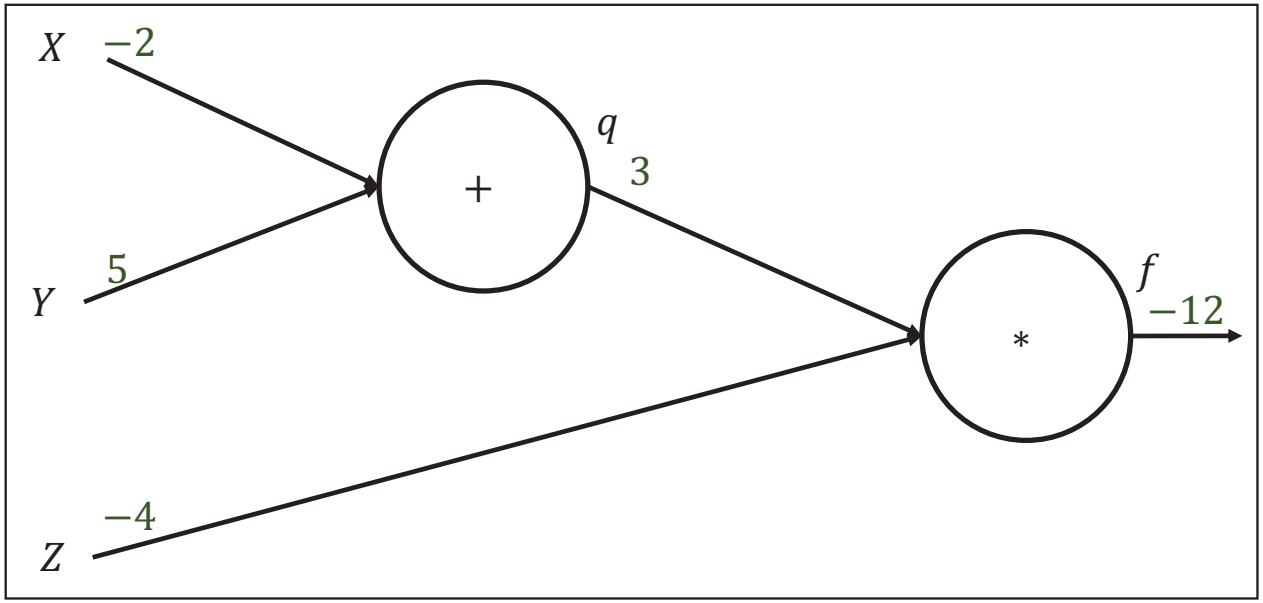


Backpropagation: a simple example

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

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

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



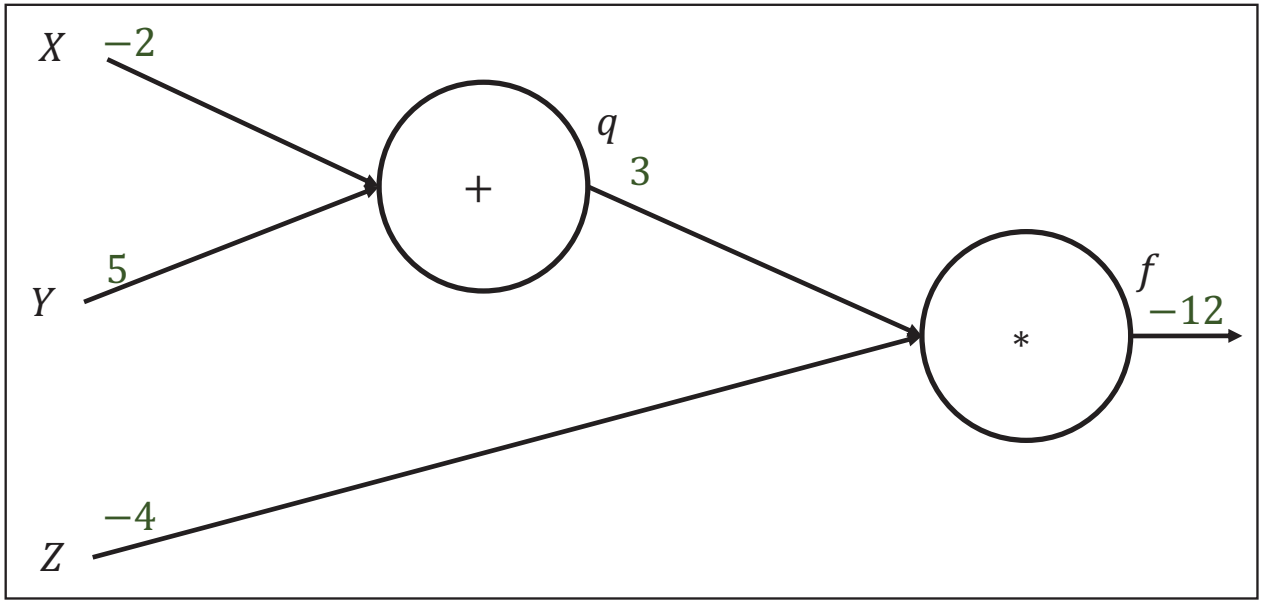
Backpropagation: a simple example

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

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

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

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



Backpropagation: a simple example

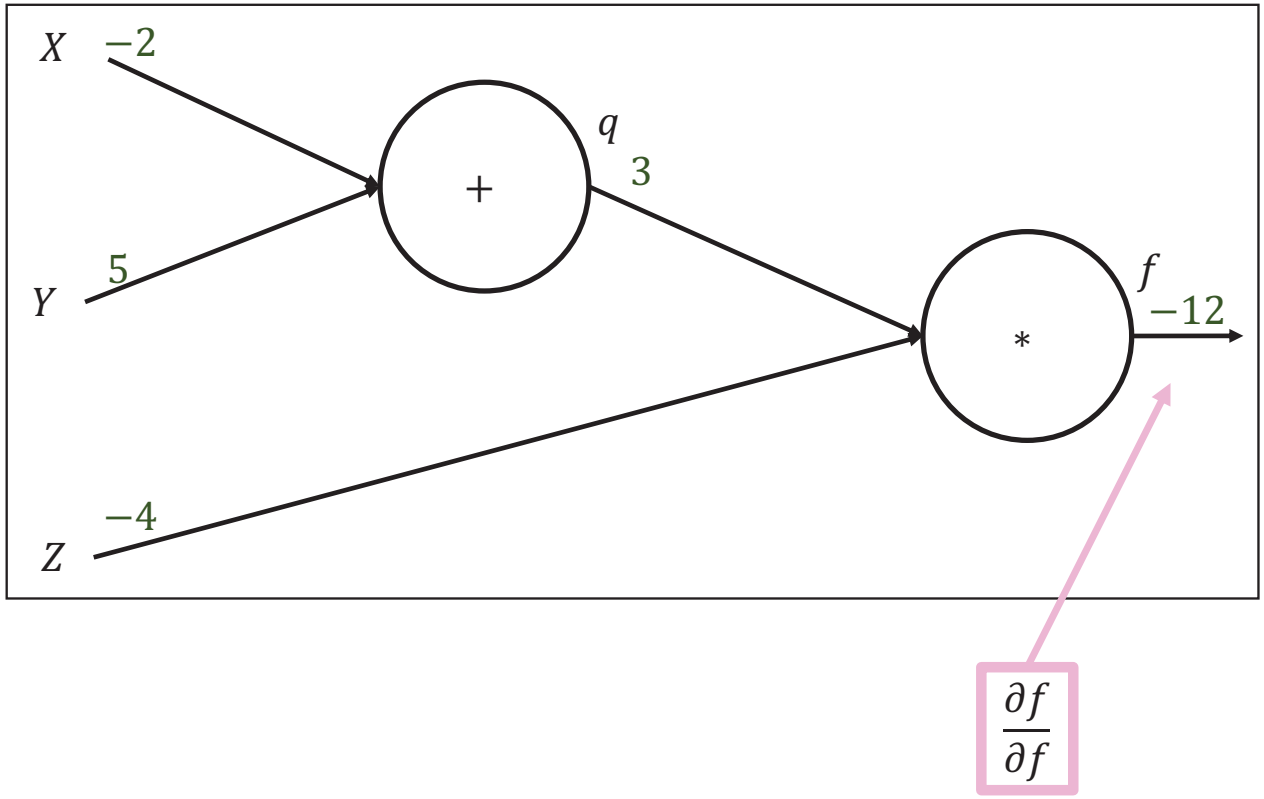
$$f(x, y, z) = (x + y)z$$

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

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

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

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



Backpropagation: a simple example

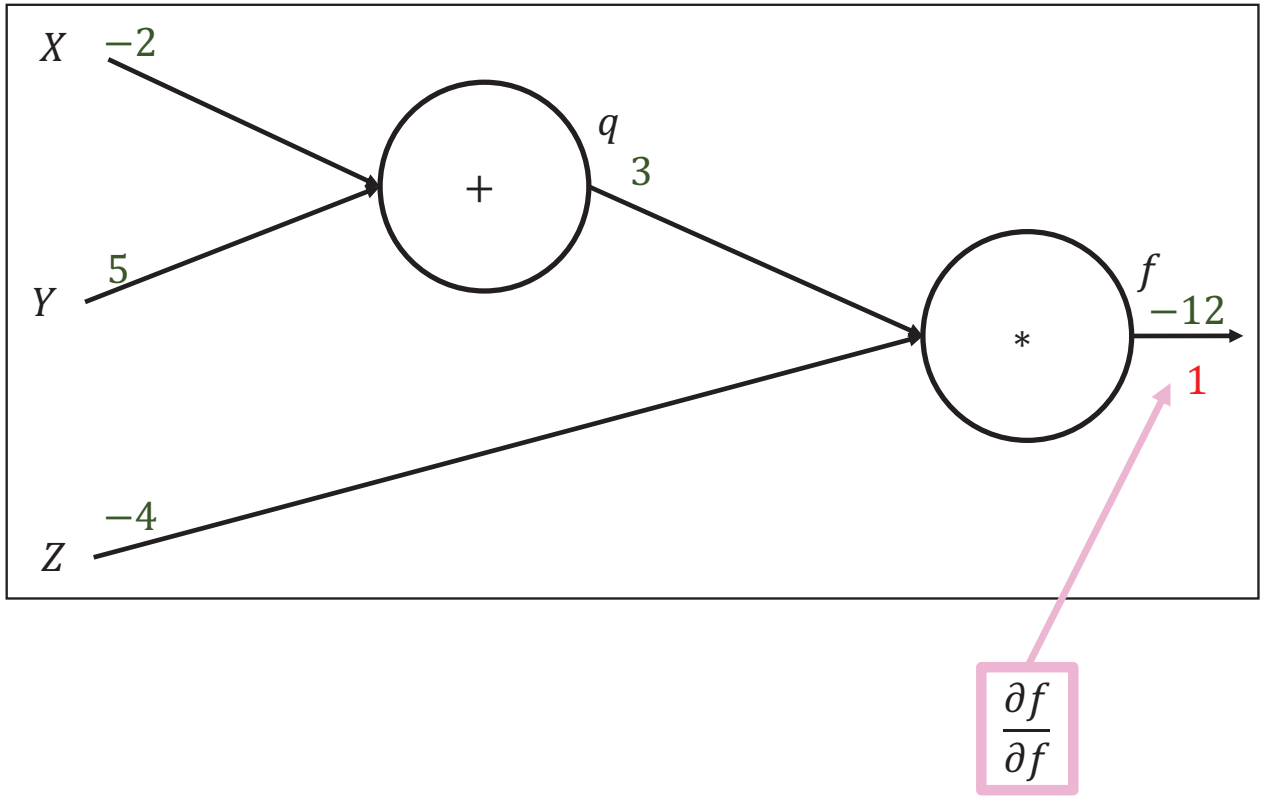
$$f(x, y, z) = (x + y)z$$

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

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

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

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



Backpropagation: a simple example

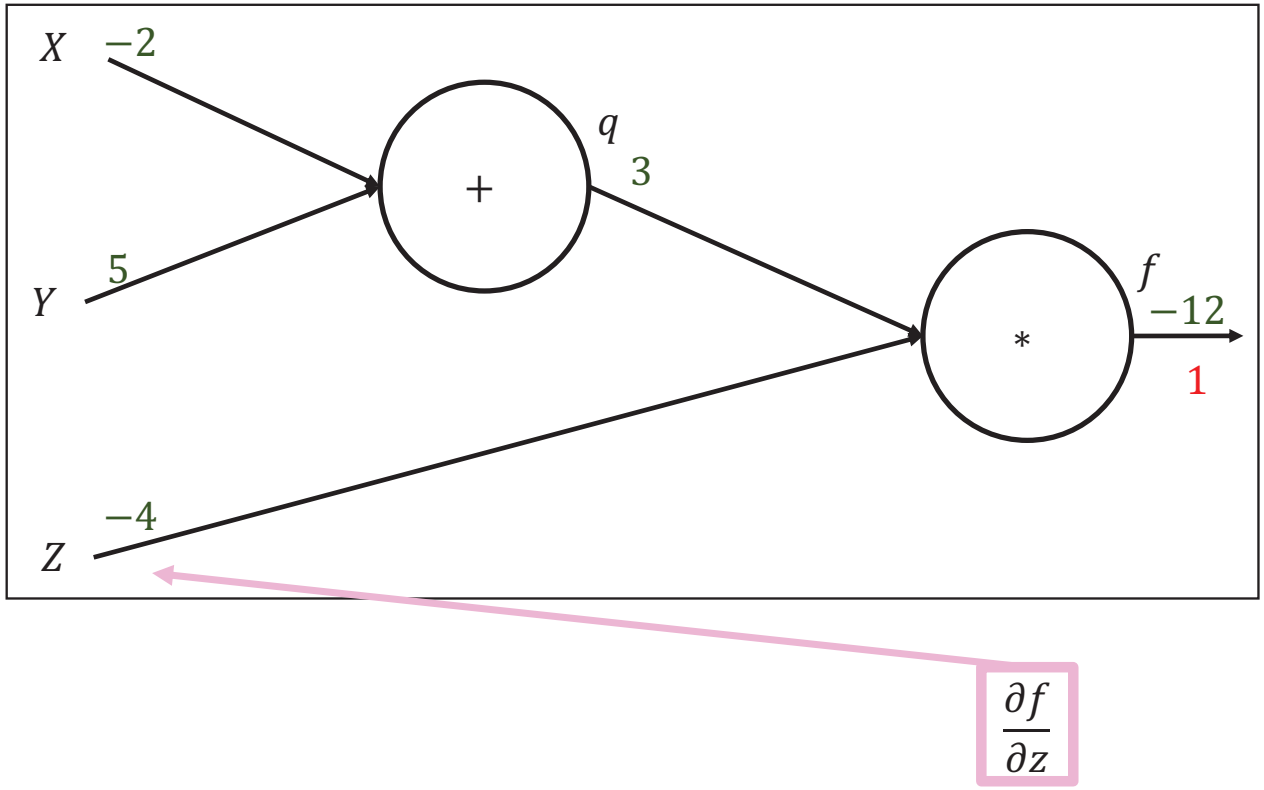
$$f(x, y, z) = (x + y)z$$

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

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

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

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



Backpropagation: a simple example

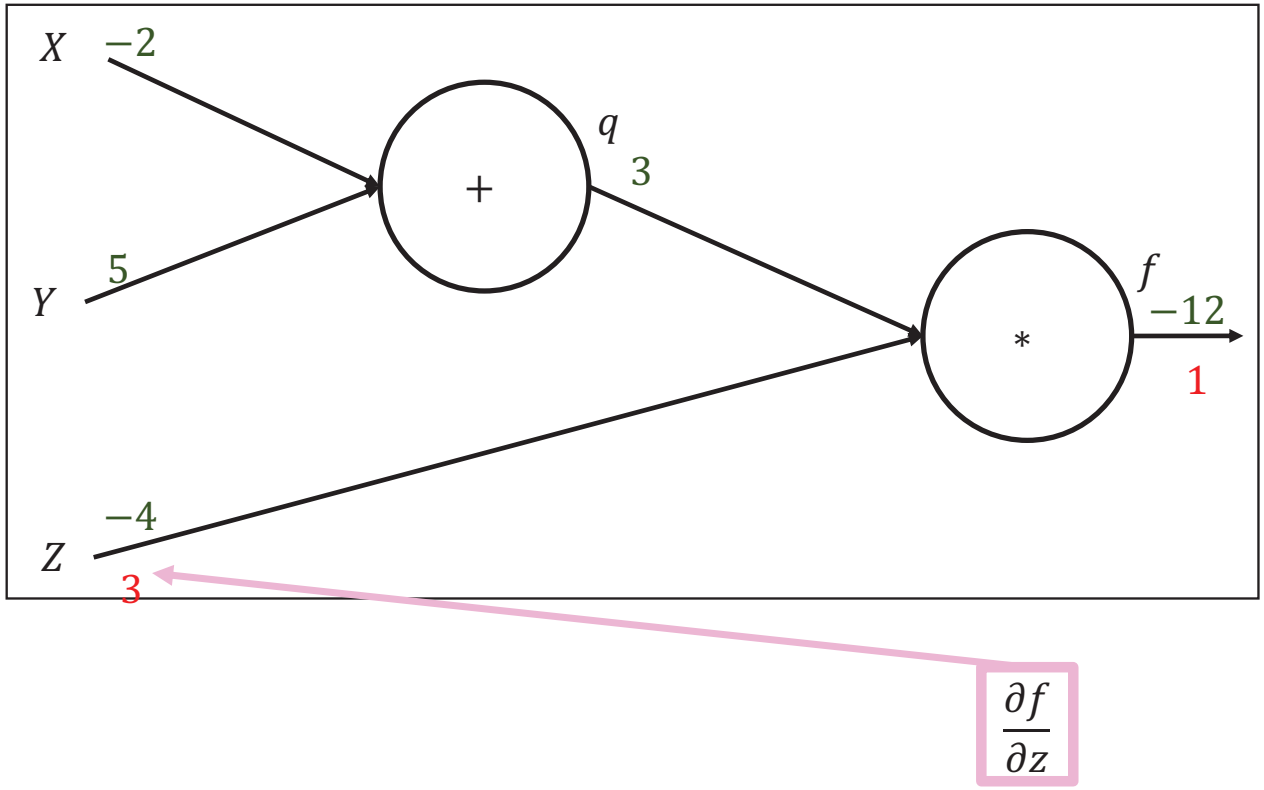
$$f(x, y, z) = (x + y)z$$

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

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

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

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



Backpropagation: a simple example

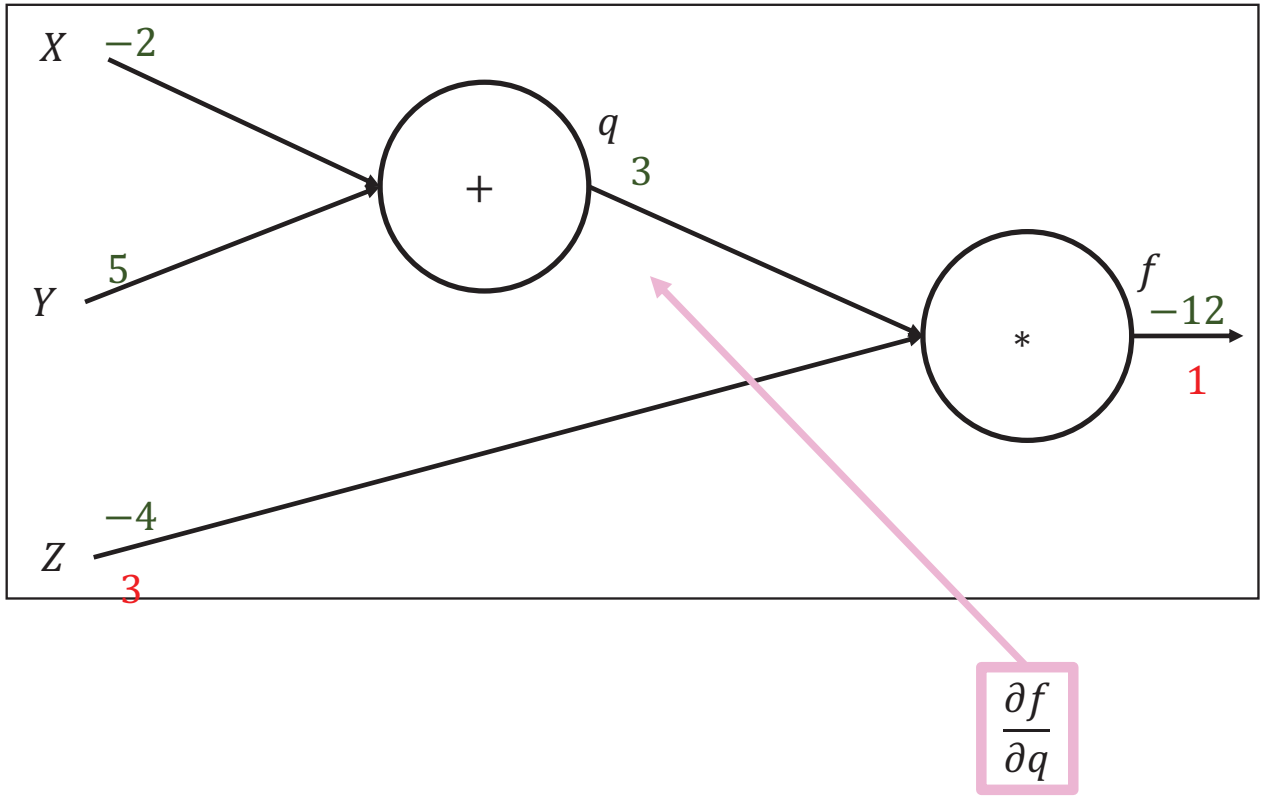
$$f(x, y, z) = (x + y)z$$

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

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

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

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



Backpropagation: a simple example

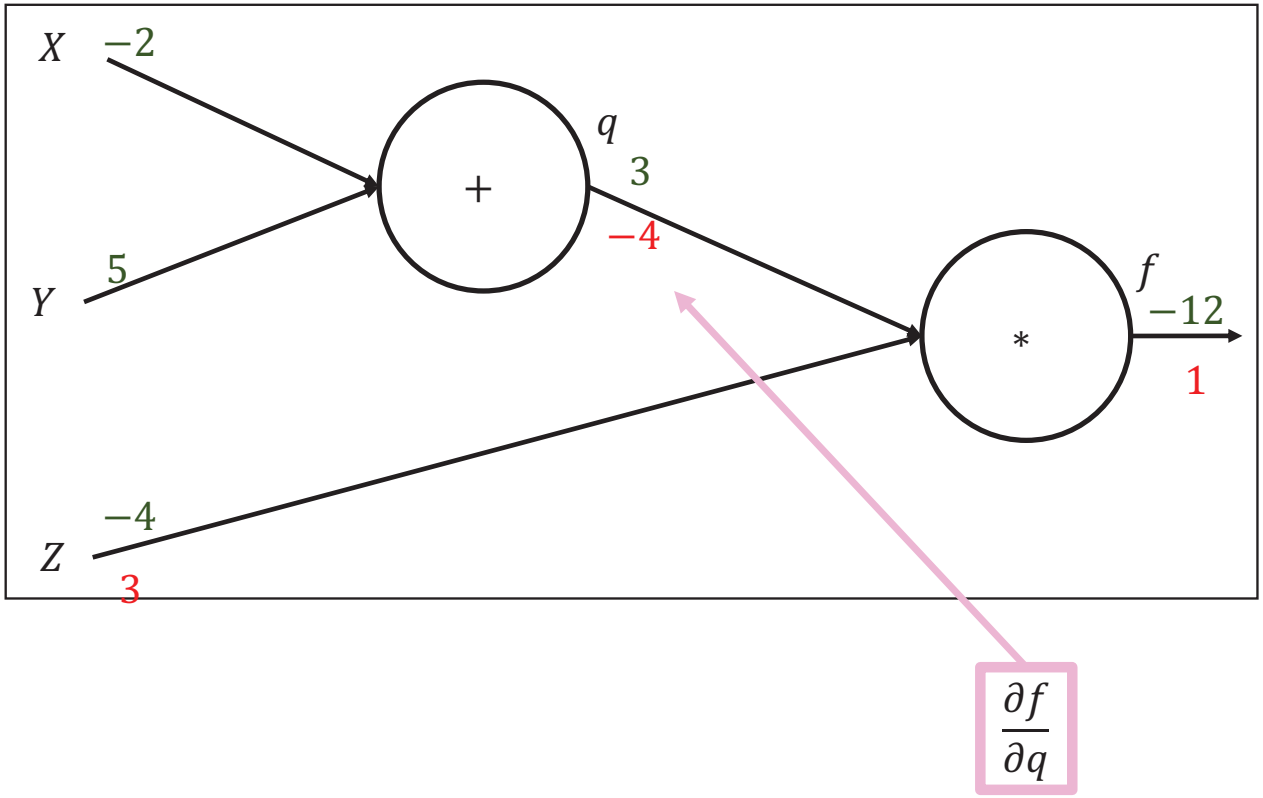
$$f(x, y, z) = (x + y)z$$

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

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

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

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



Backpropagation: a simple example

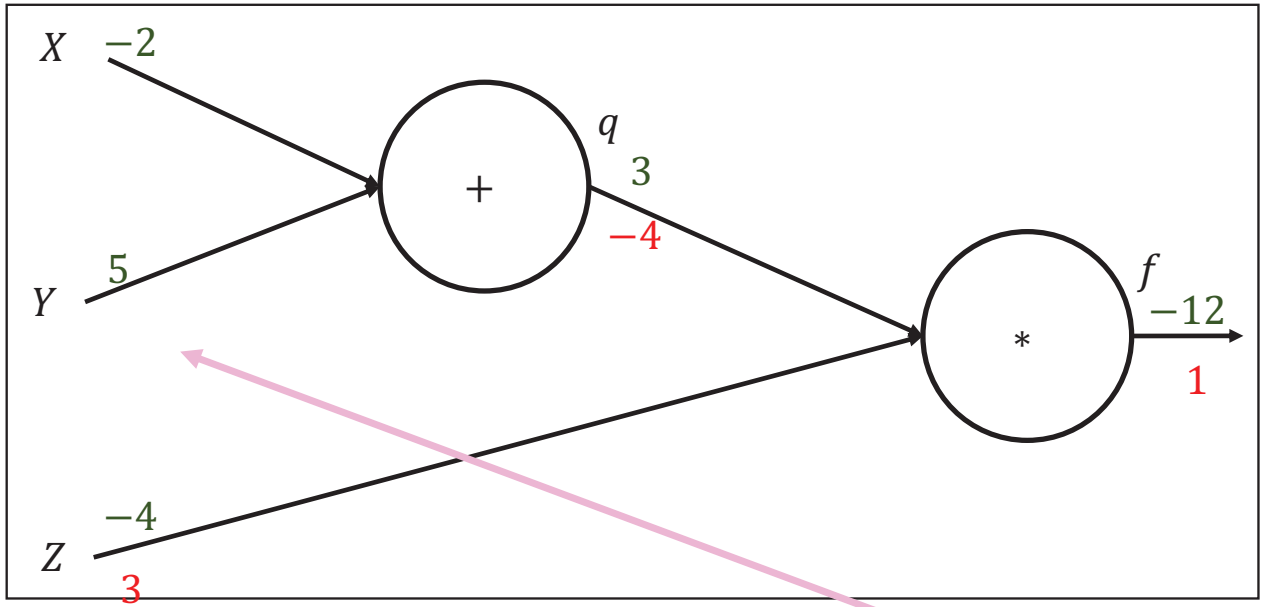
$$f(x, y, z) = (x + y)z$$

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

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

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

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



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

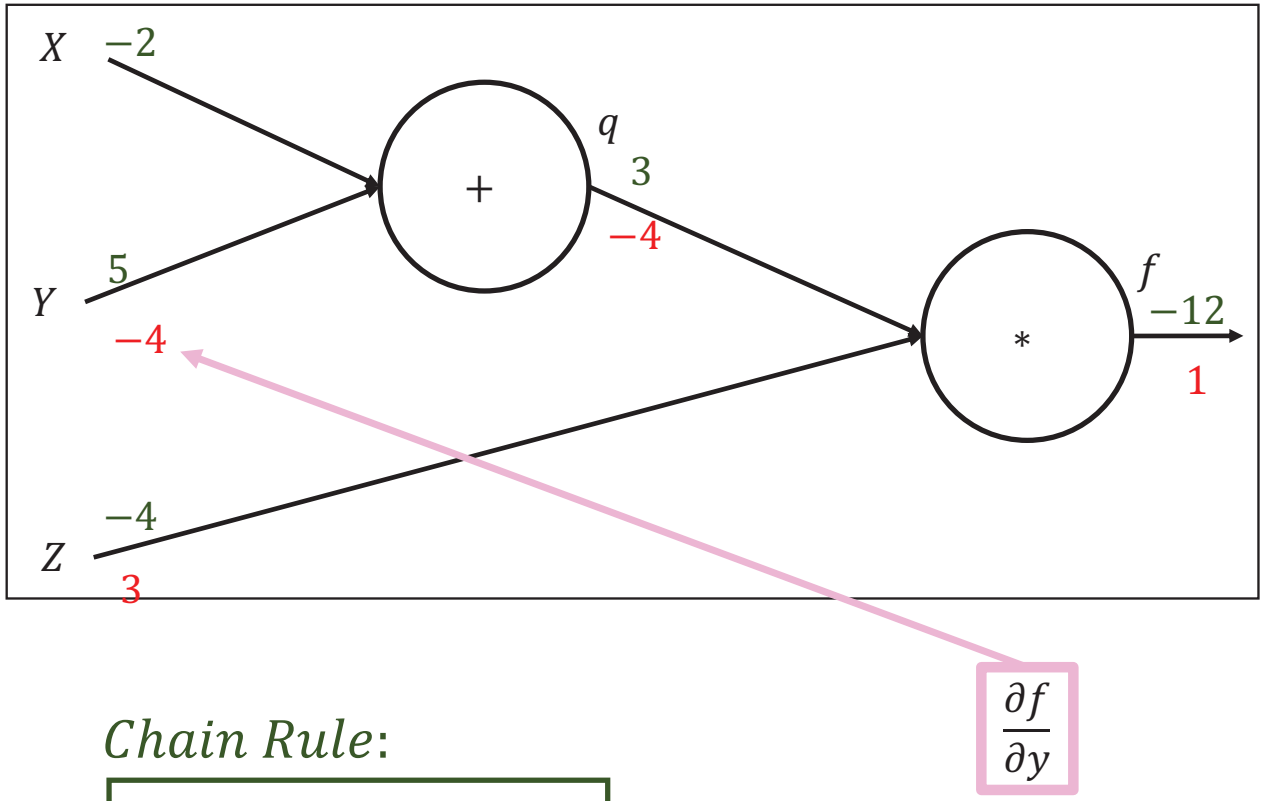
$$f(x, y, z) = (x + y)z$$

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

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

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

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



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

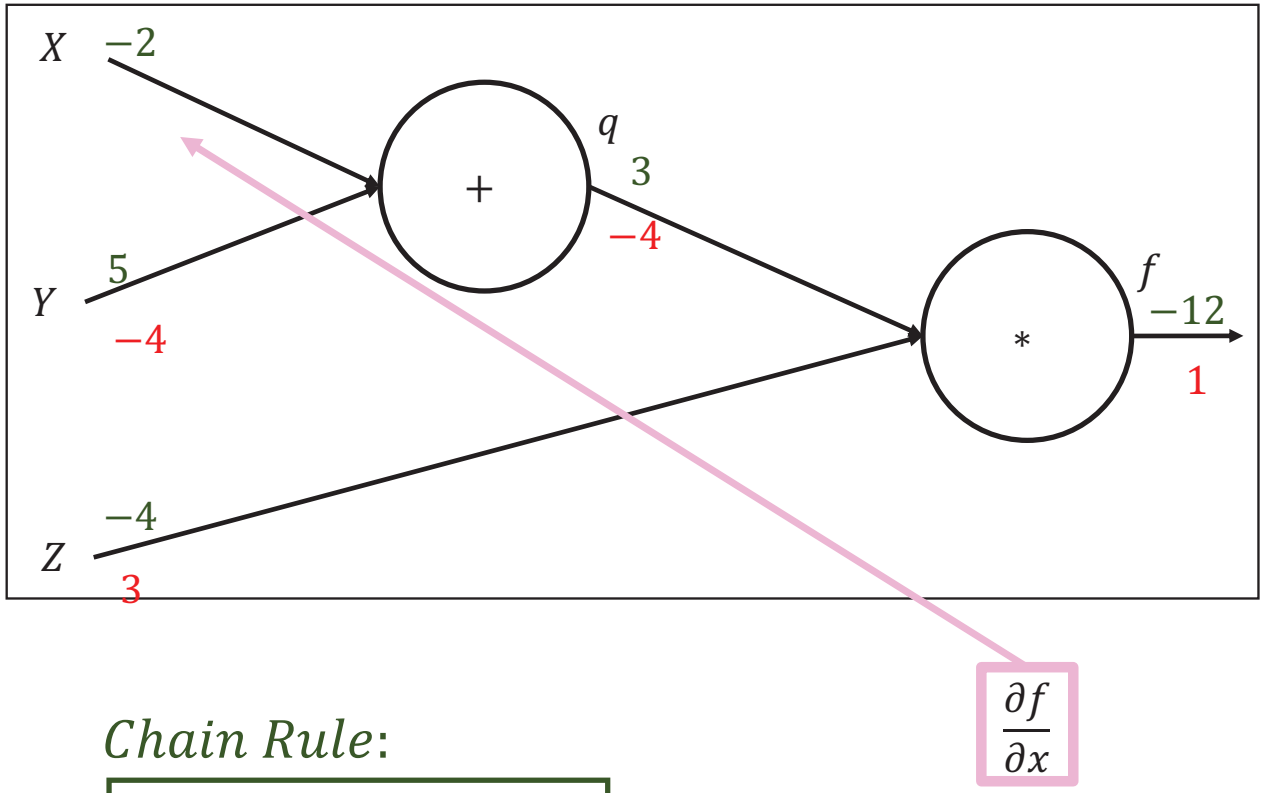
$$f(x, y, z) = (x + y)z$$

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

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

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

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



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

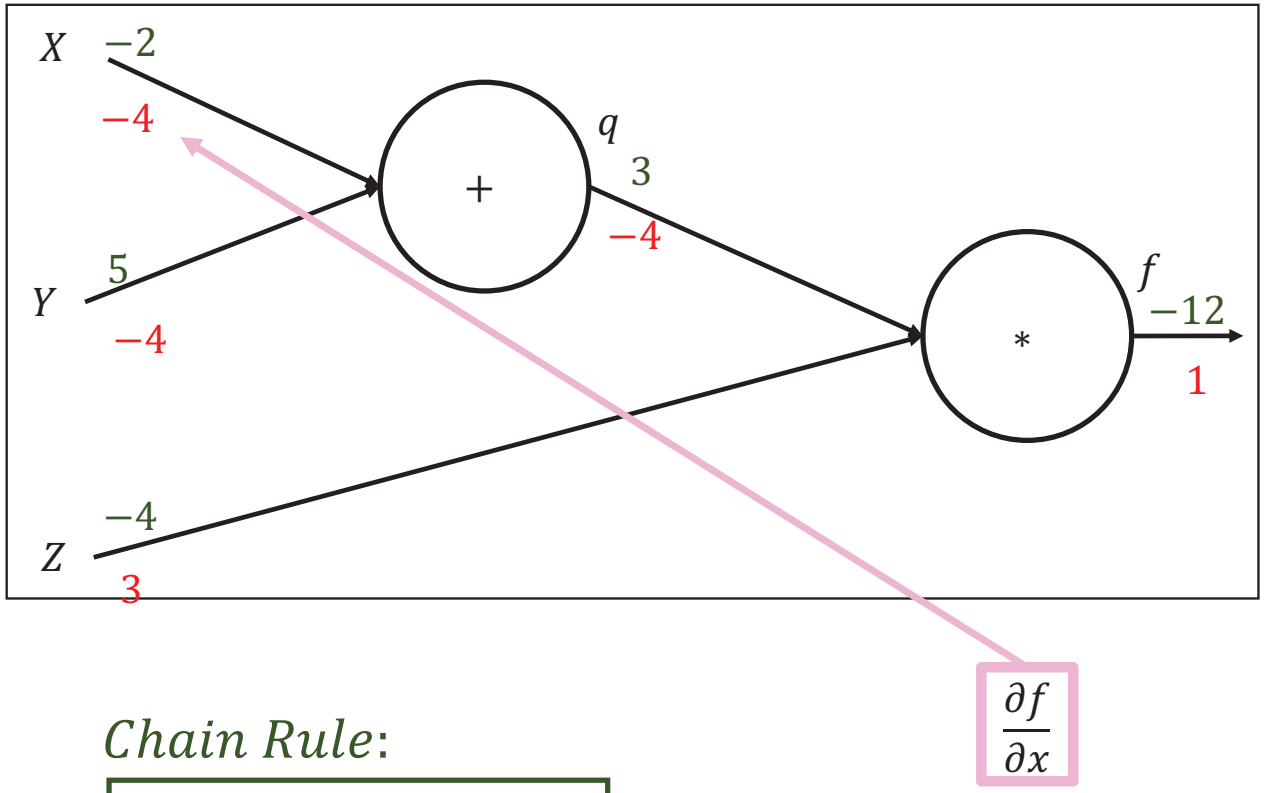
$$f(x, y, z) = (x + y)z$$

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

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

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

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



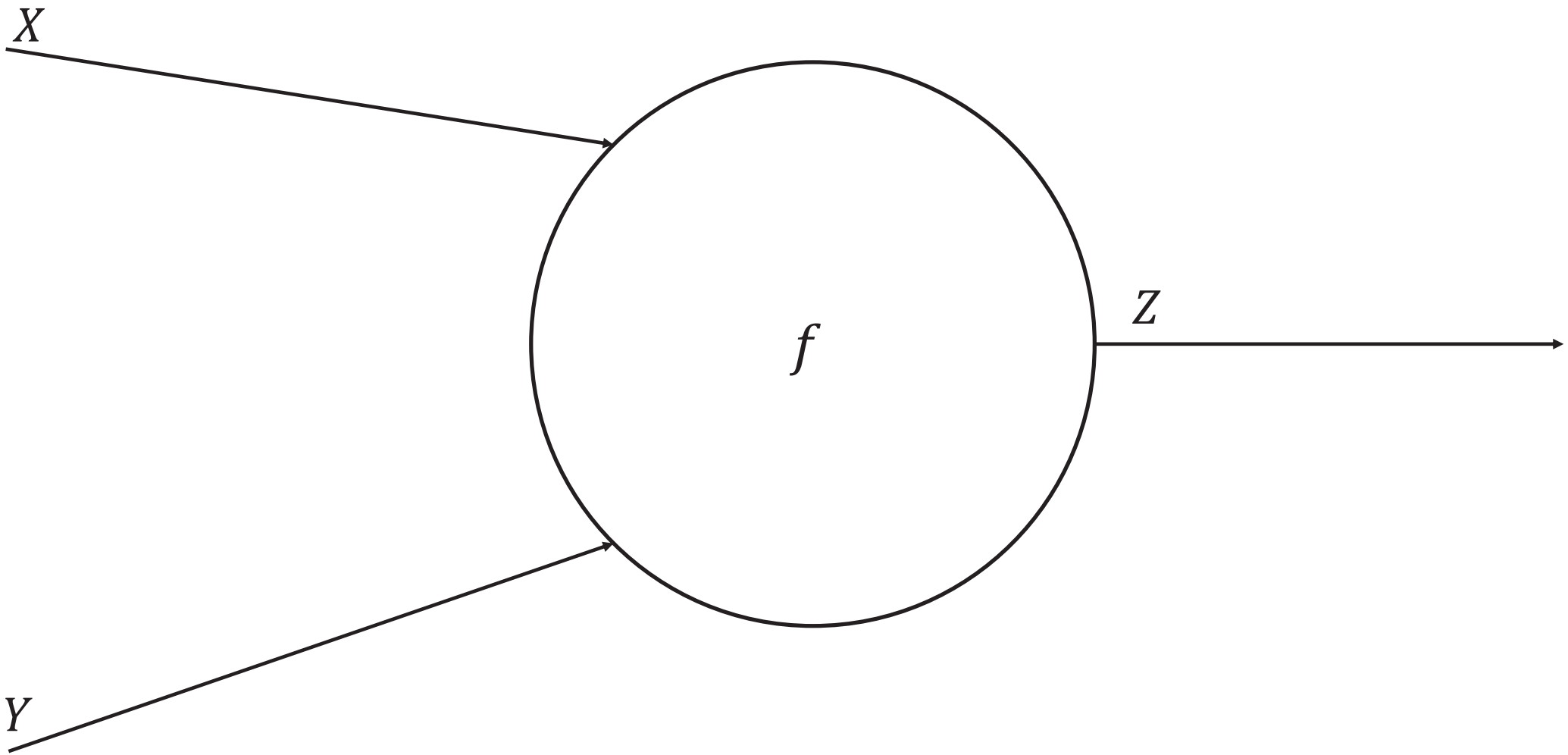
Chain Rule:

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

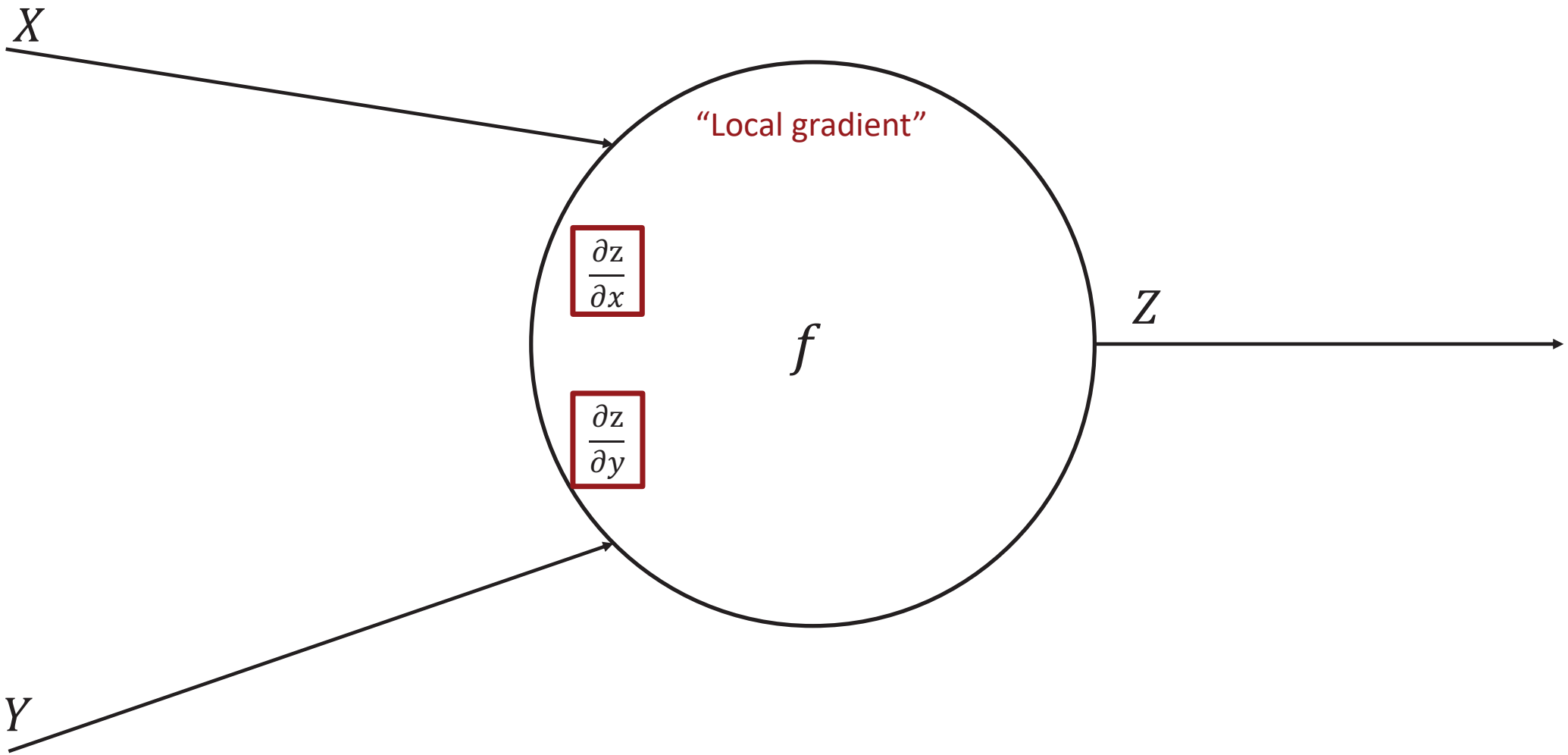
Upstream gradient

Local gradient

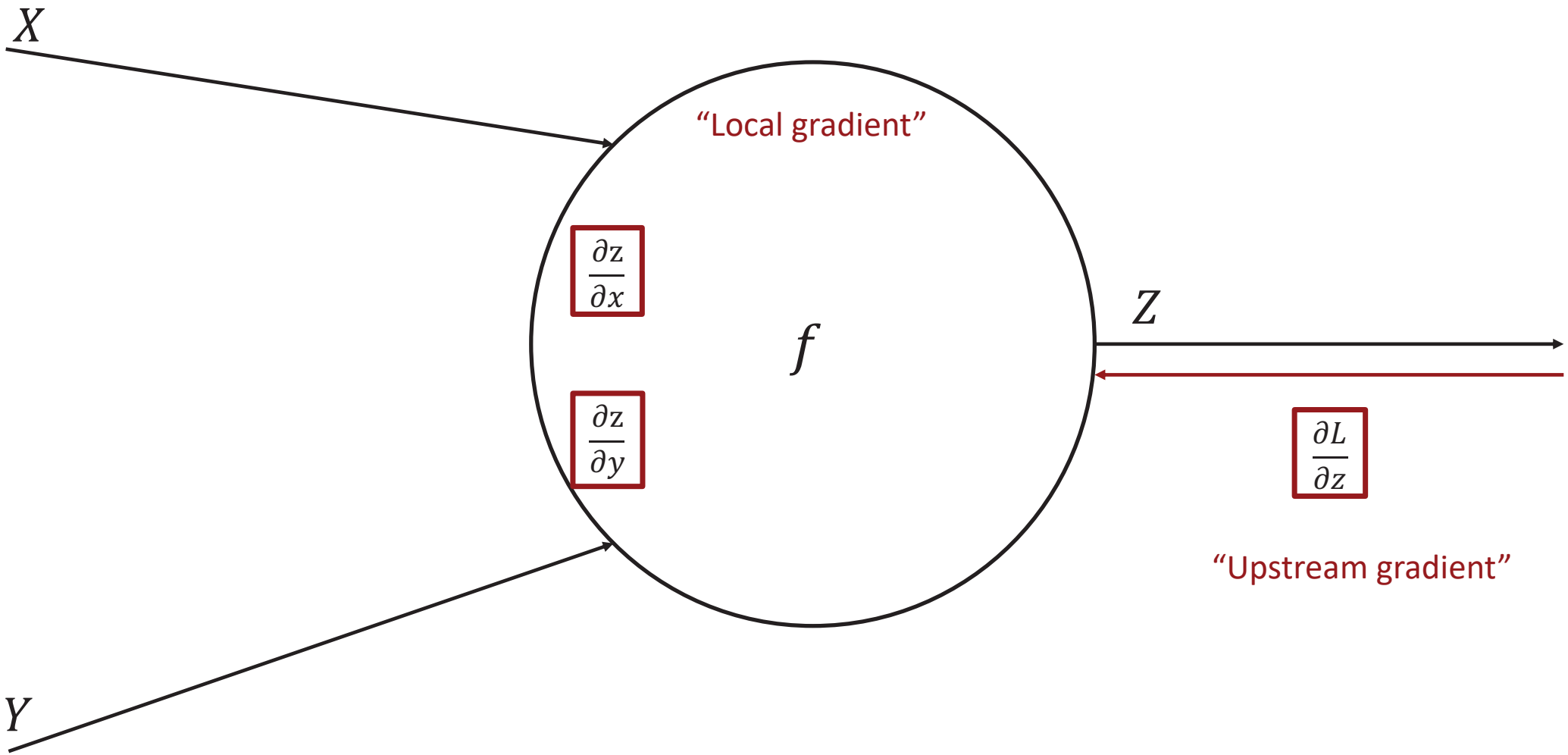
Backpropagation



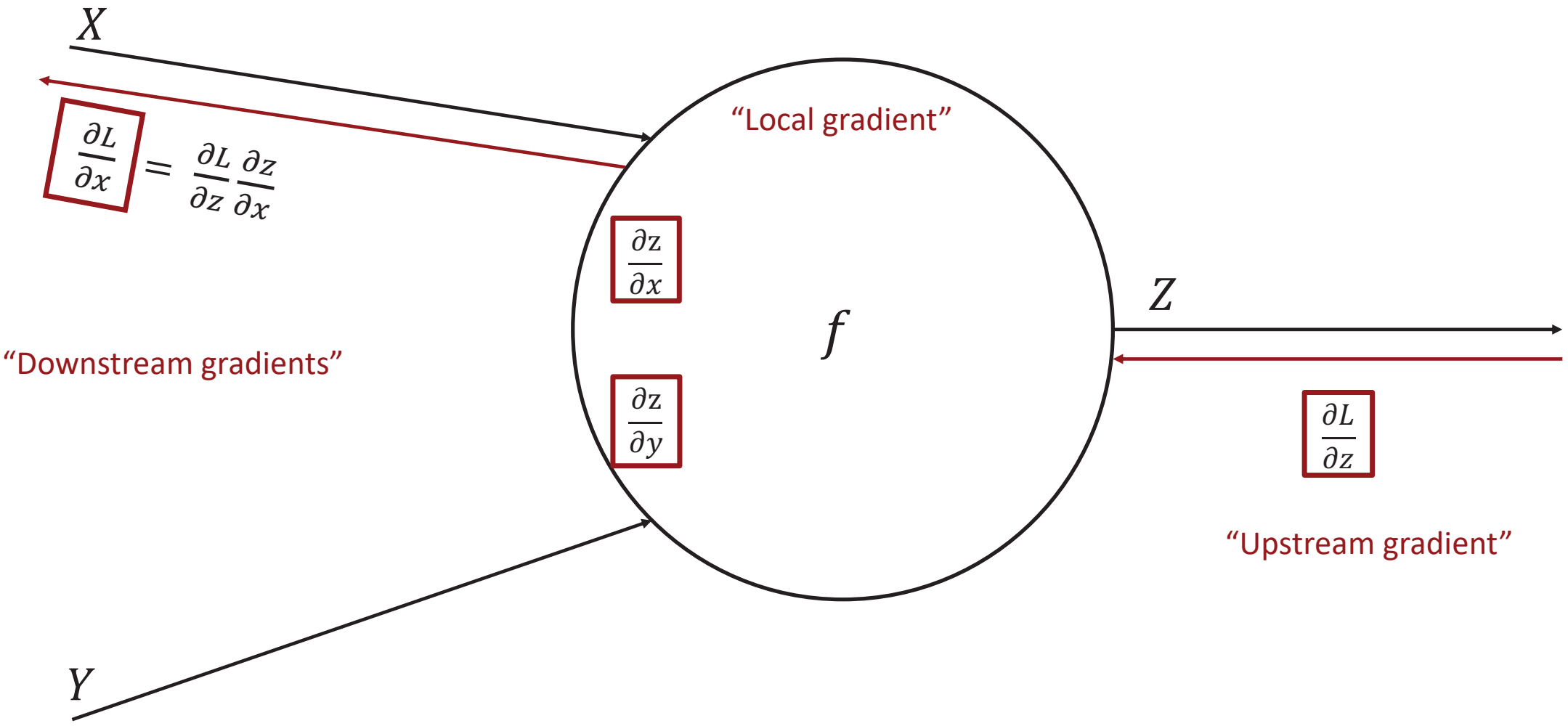
Backpropagation



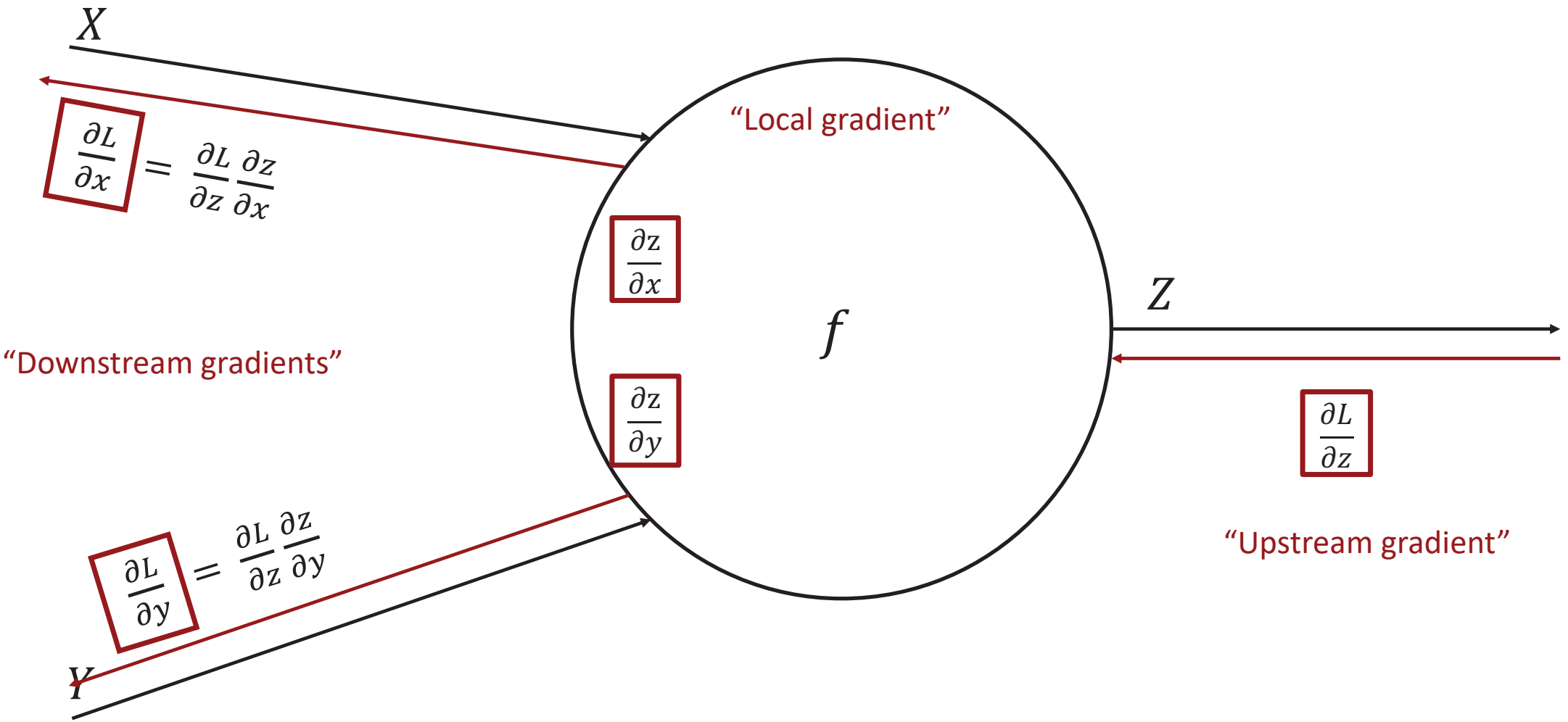
Backpropagation



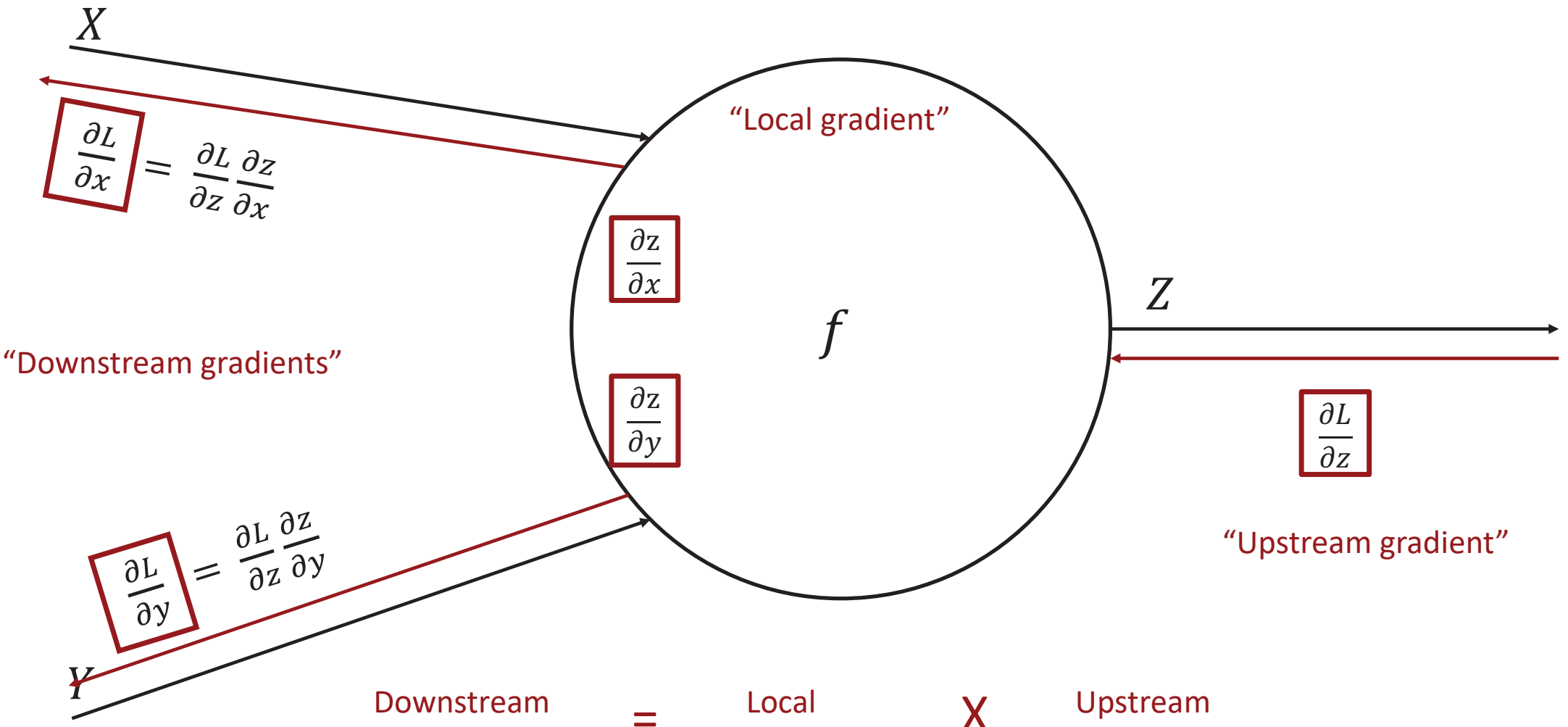
Backpropagation



Backpropagation



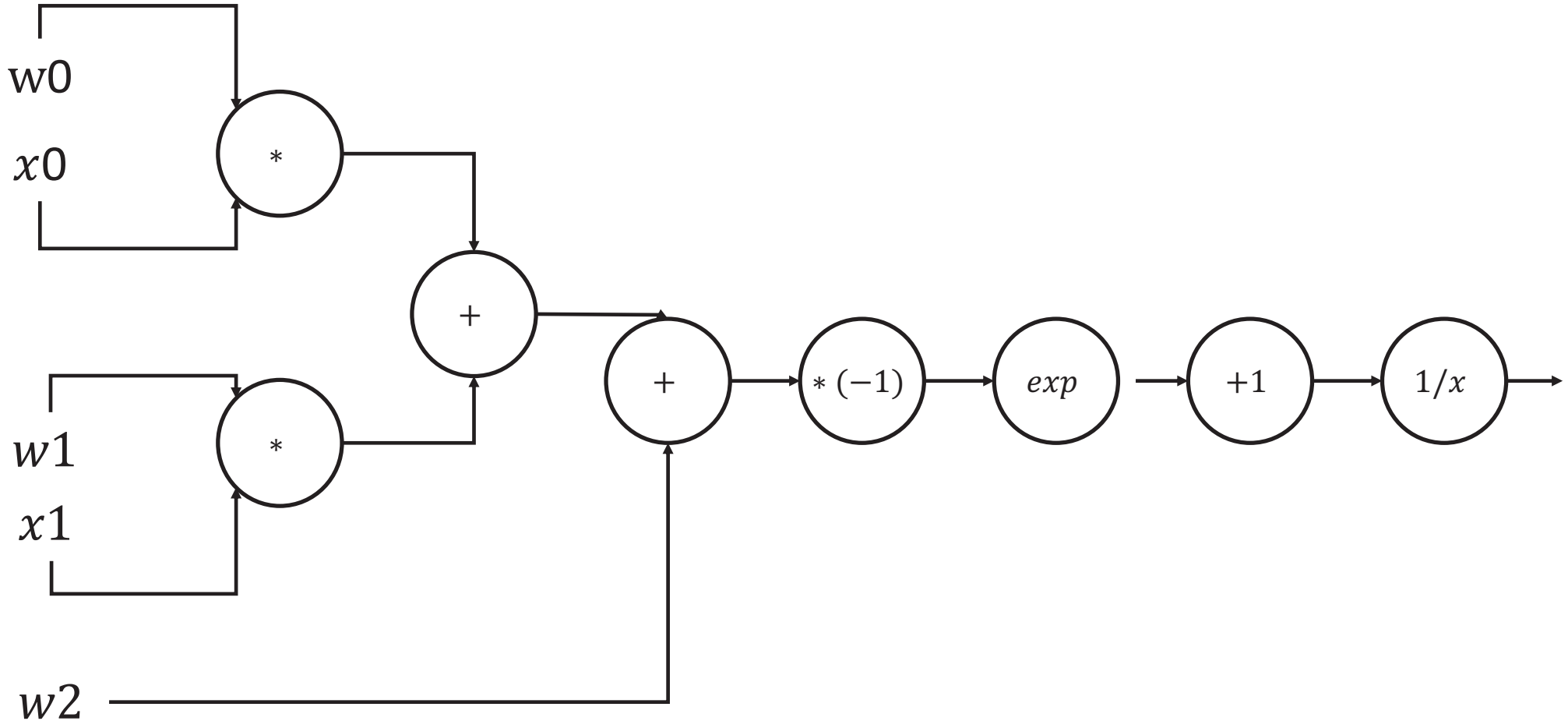
Backpropagation



Downstream Gradients = Local Gradient X Upstream Gradient

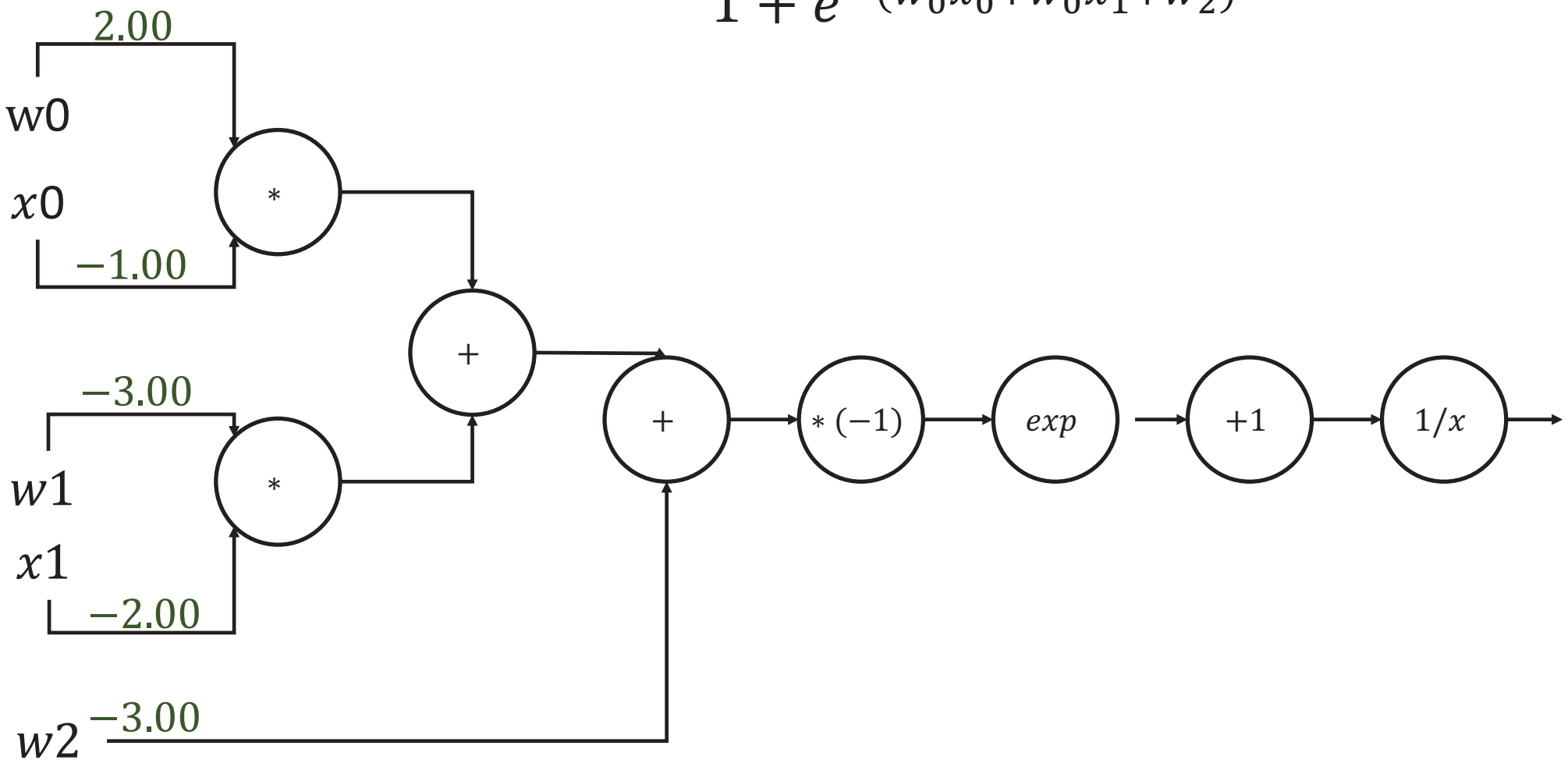
Another example:

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



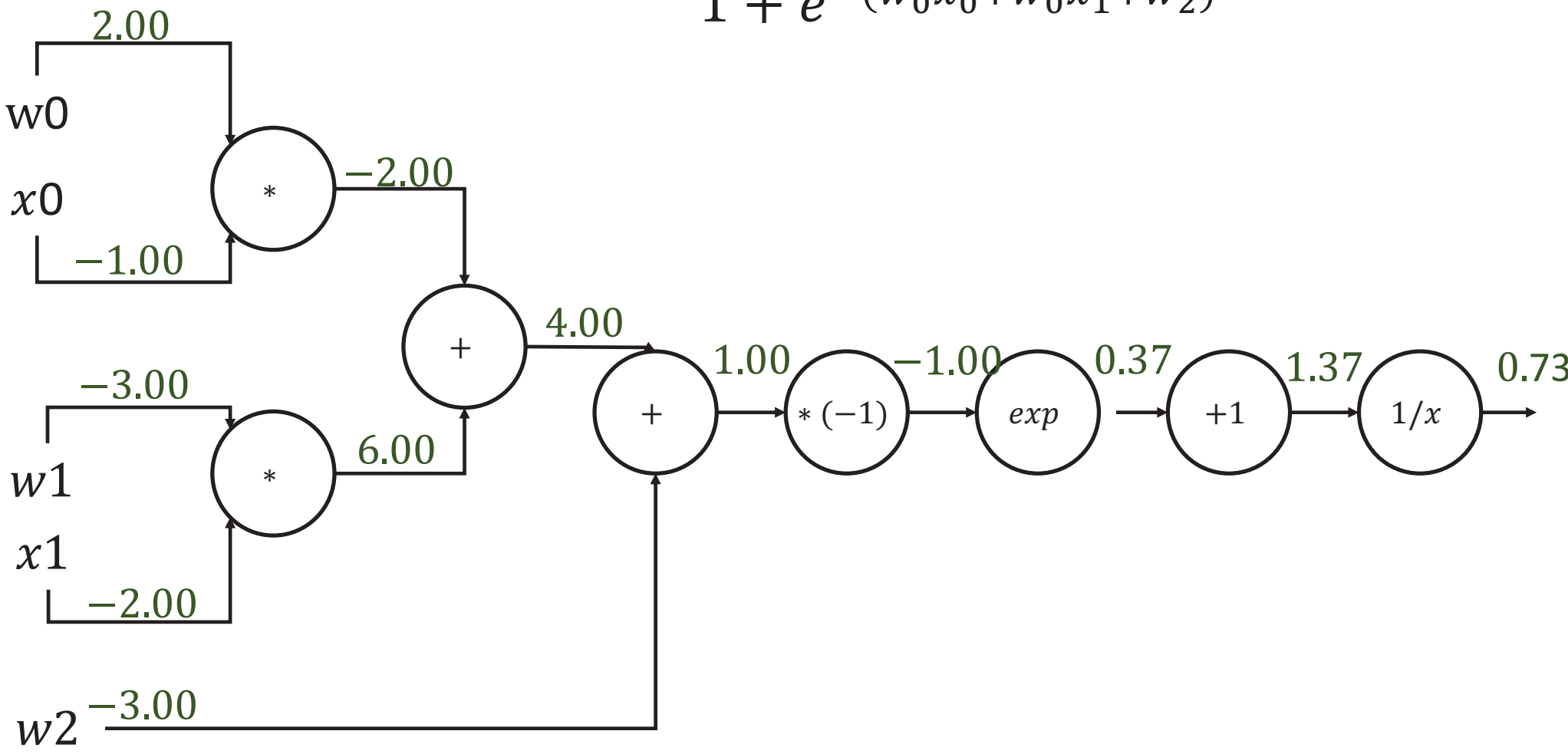
Another example:

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_0x_1+w_2)}}$$



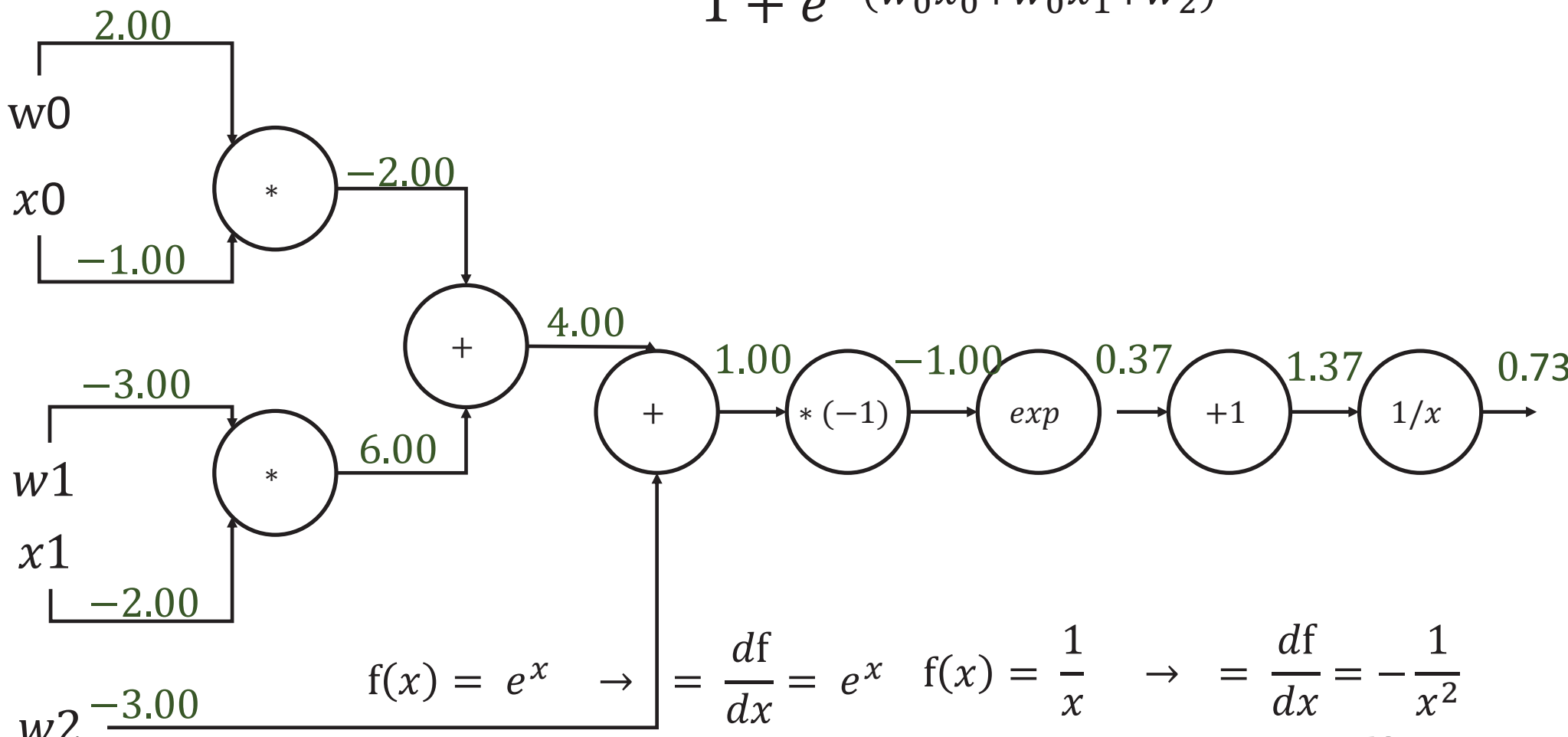
Another example:

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



Another example:

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

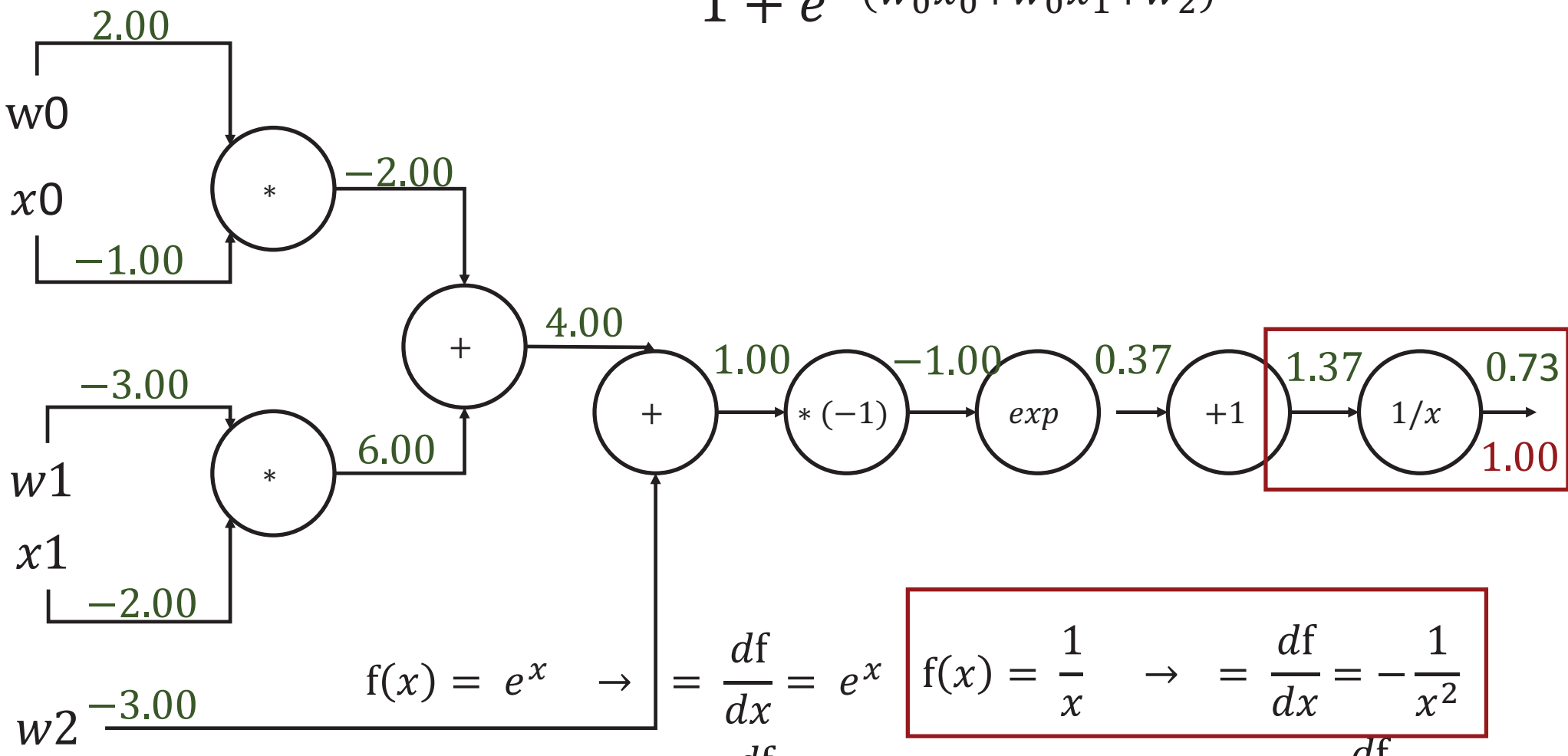


$$\mathrm{f}(x) = e^x \quad \rightarrow \quad = \frac{d\mathrm{f}}{dx} = e^x \qquad \mathrm{f}(x) = \frac{1}{x} \quad \rightarrow \quad = \frac{d\mathrm{f}}{dx} = -\frac{1}{x^2}$$

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

Another example:

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



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

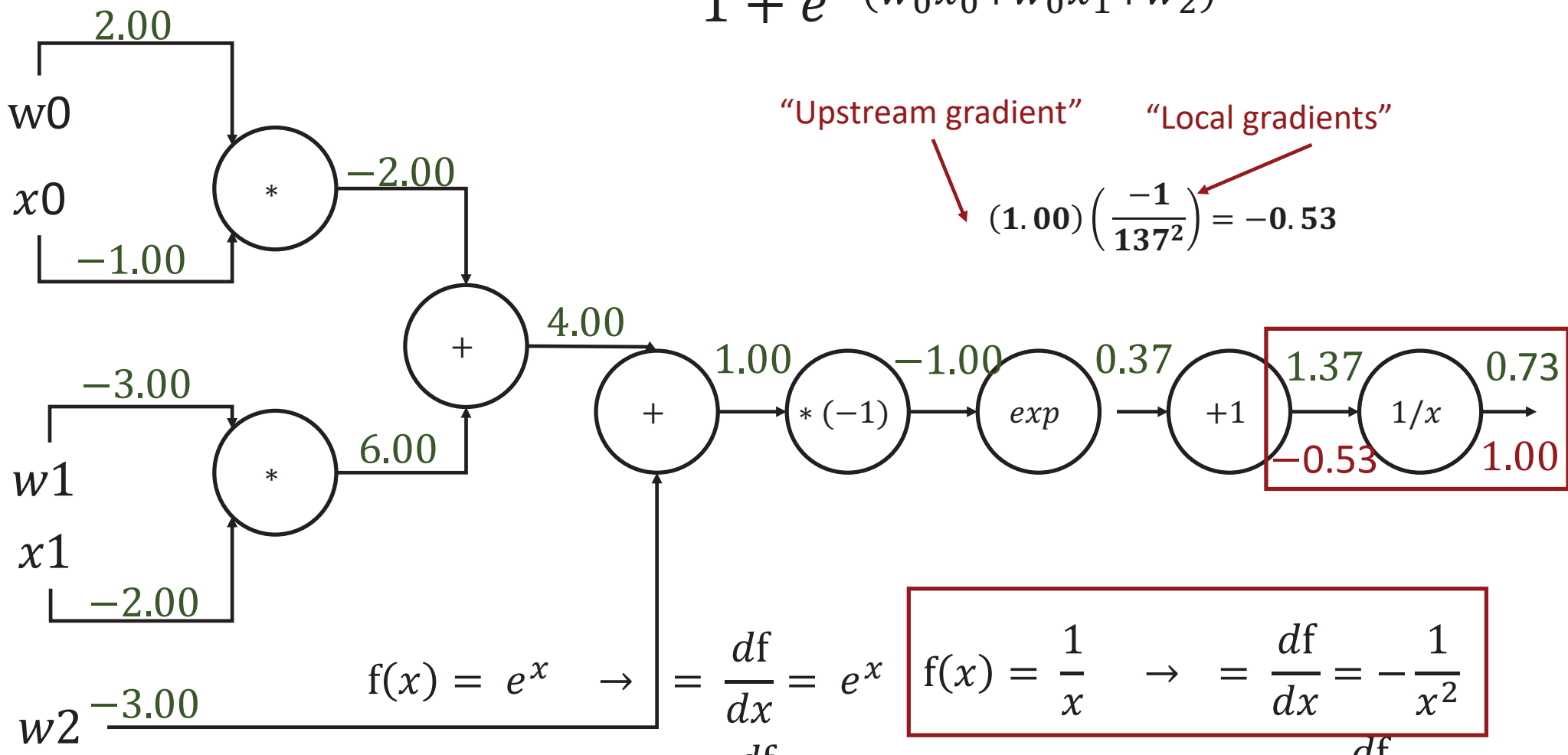
$$\mathrm{f}(x) = \frac{1}{x} \quad \rightarrow \quad = \frac{df}{dx} = -\frac{1}{x^2}$$

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

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

Another example:

$$f(w,x) = \frac{1}{1 + e^{-(w_0x_0 + w_0x_1 + w_2)}}$$



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

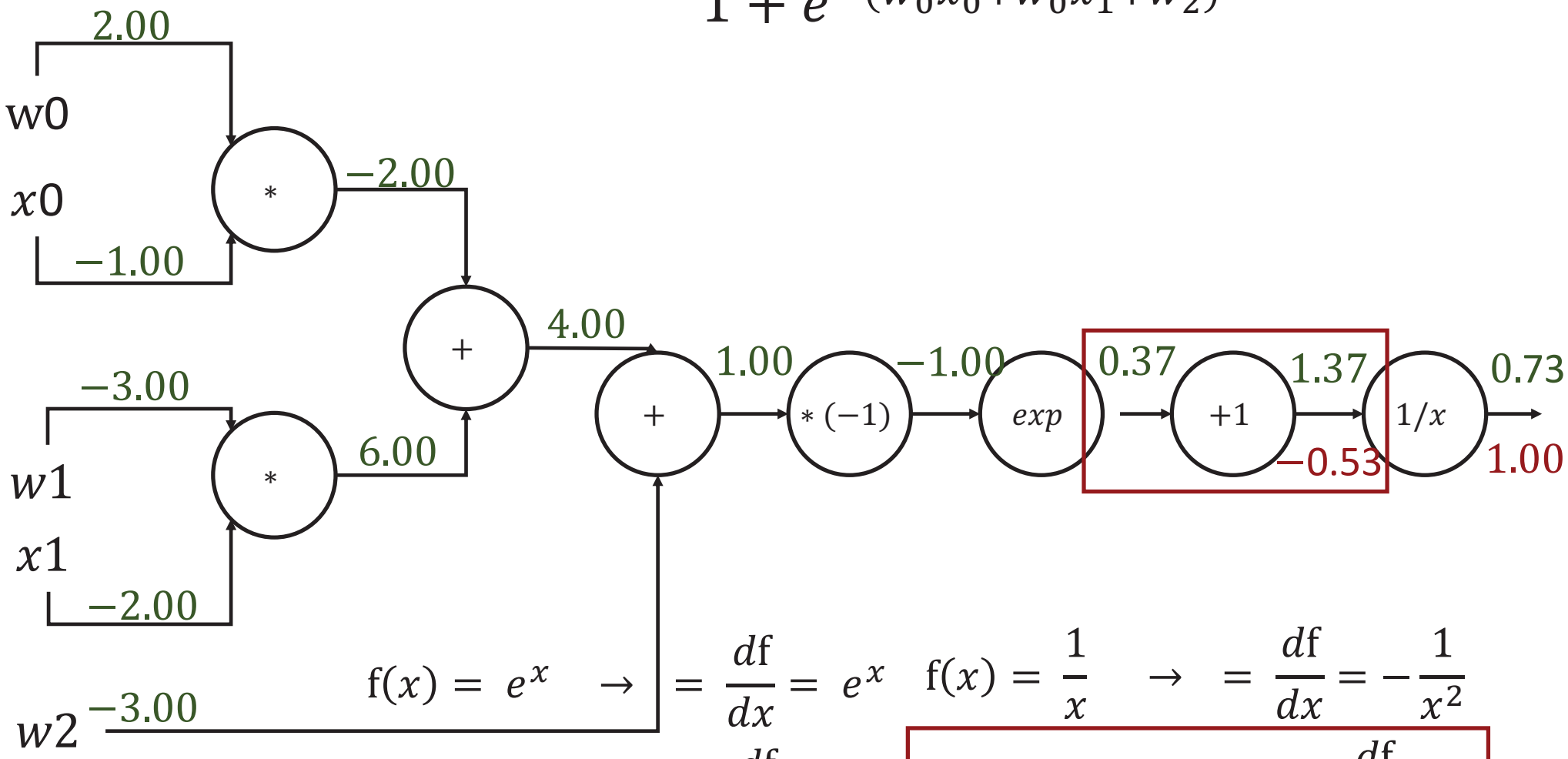
$$\mathrm{f}(x) = \frac{1}{x} \quad \rightarrow \quad = \frac{df}{dx} = -\frac{1}{x^2}$$

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

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

Another example:

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



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

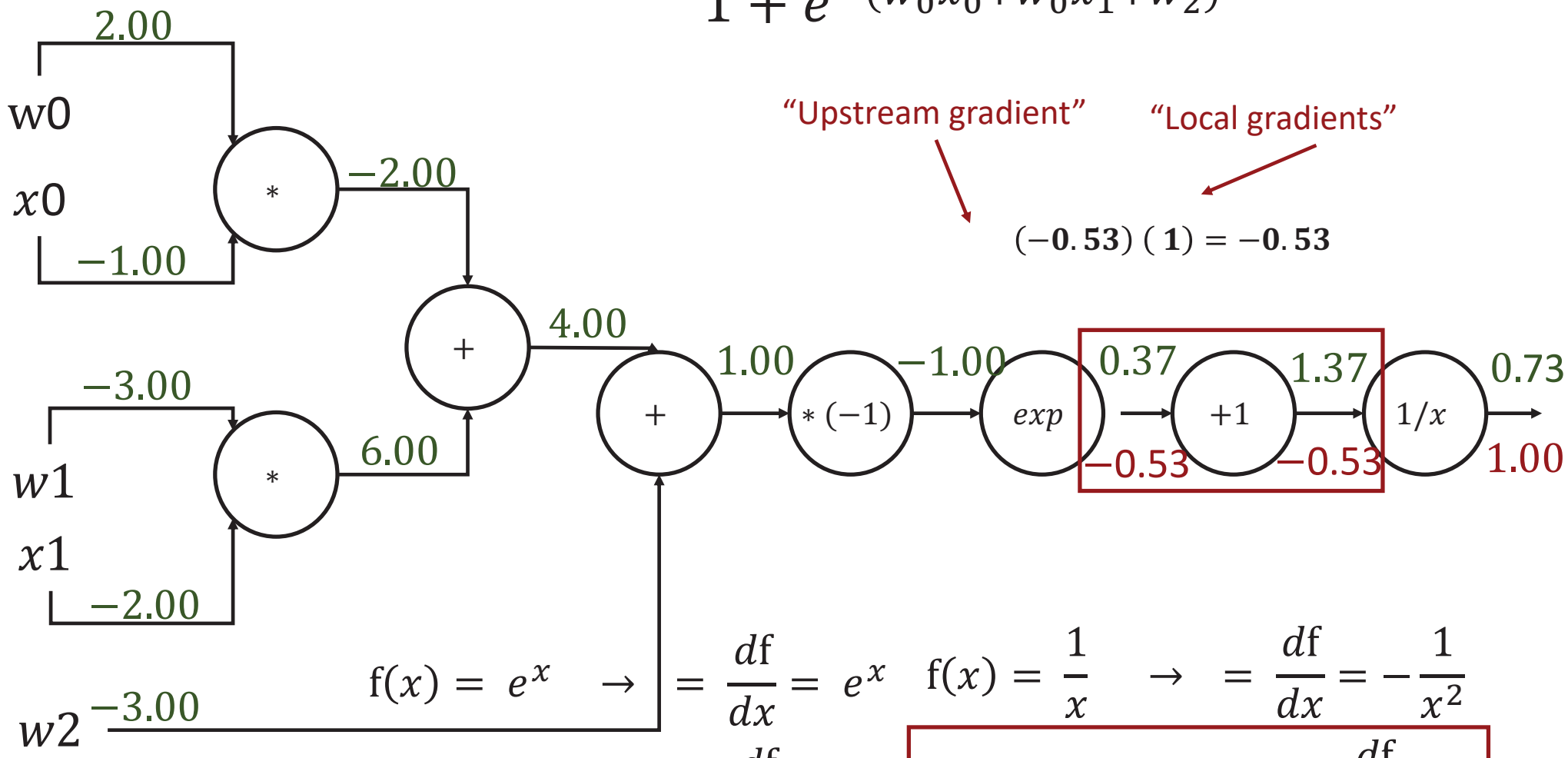
$$\mathrm{f}(x) = \frac{1}{x} \quad \rightarrow \quad = \frac{d\mathrm{f}}{dx} = -\frac{1}{x^2}$$

$$f_a(x) = ax \quad \rightarrow \quad = \frac{d\mathrm{f}}{dx} = a$$

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

Another example:

$$f(w,x) = \frac{1}{1 + e^{-(w_0x_0 + w_0x_1 + w_2)}}$$



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

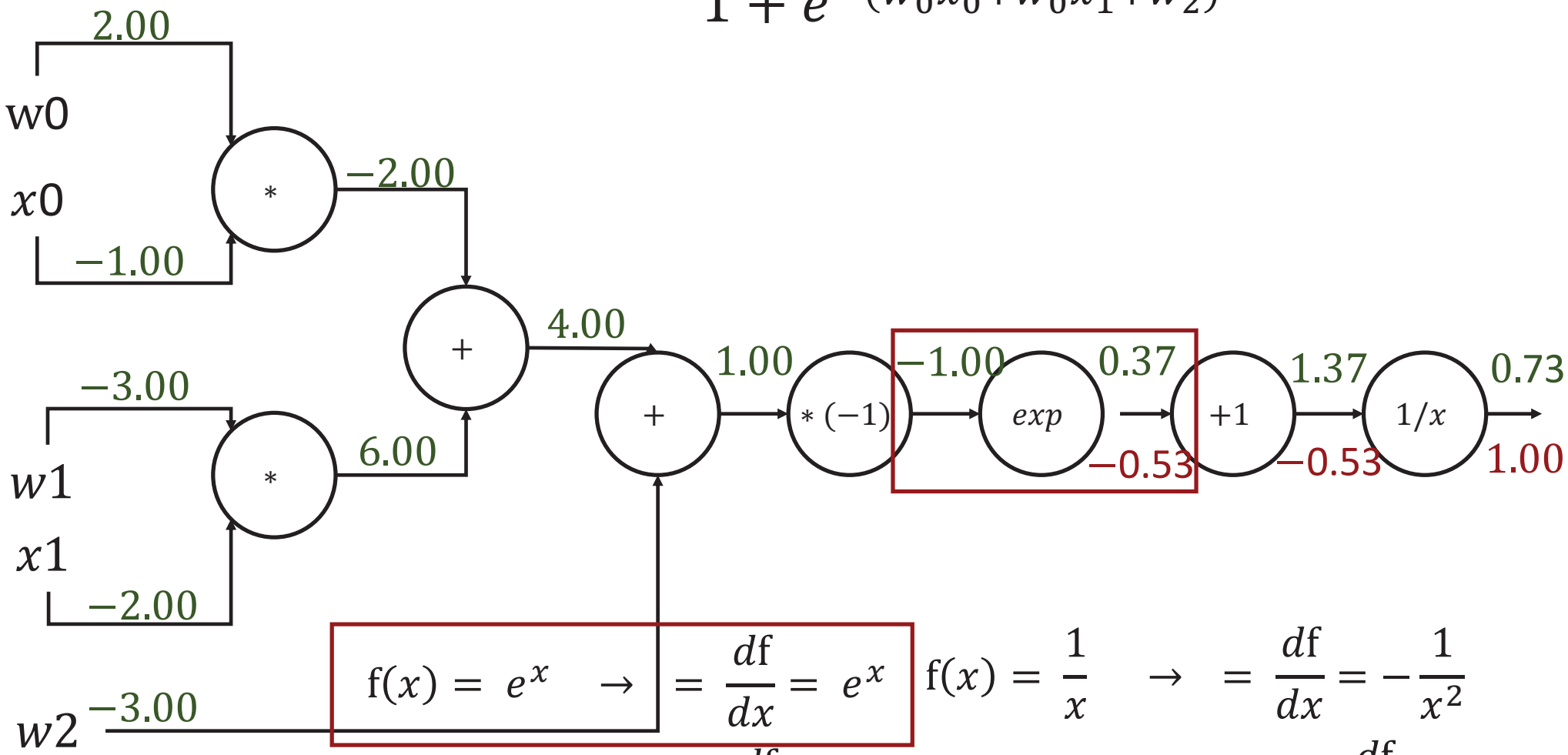
$$\mathrm{f}(x) = \frac{1}{x} \quad \rightarrow \quad = \frac{df}{dx} = -\frac{1}{x^2}$$

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

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

Another example:

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



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

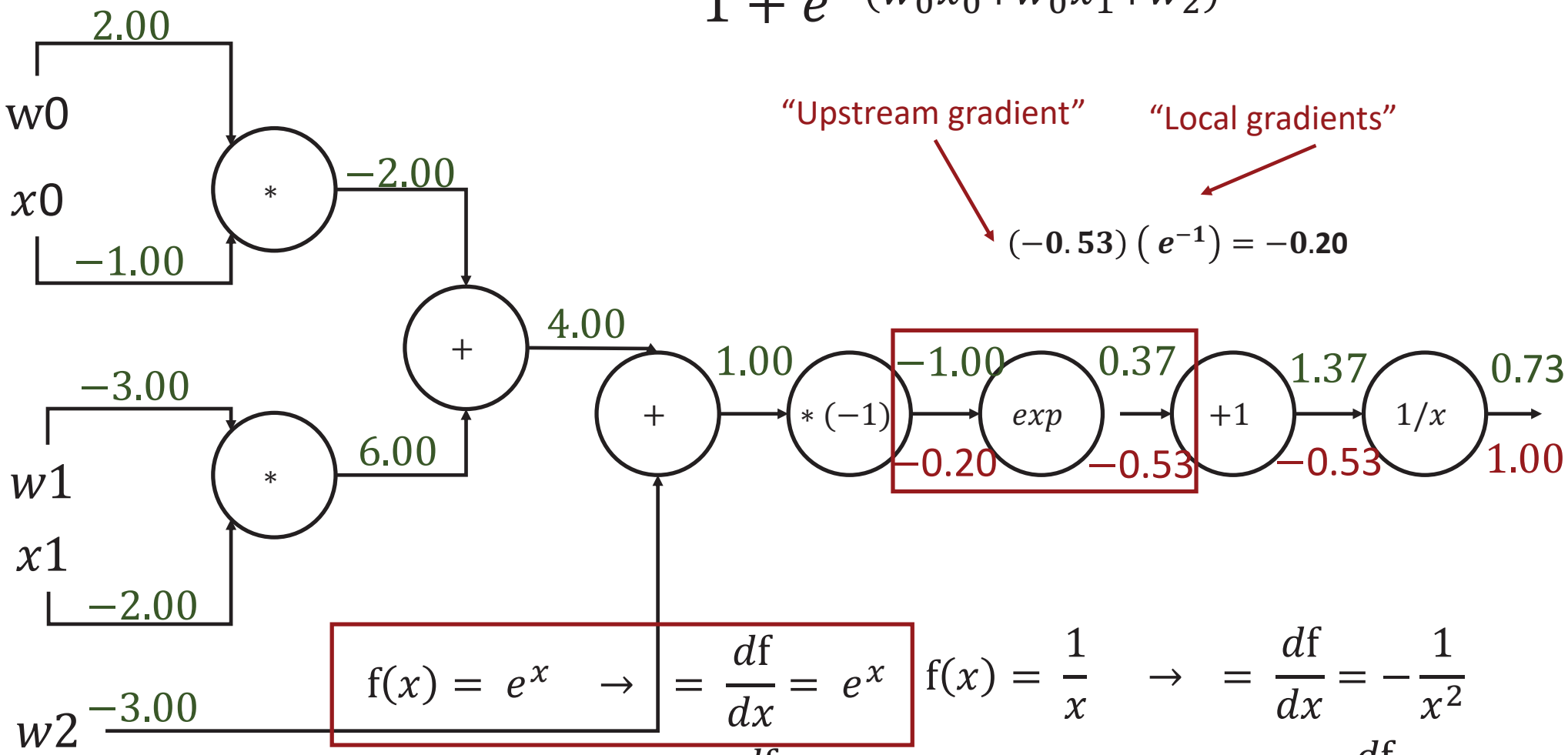
$$\mathrm{f}(x) = \frac{1}{x} \quad \rightarrow \quad = \frac{df}{dx} = -\frac{1}{x^2}$$

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

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

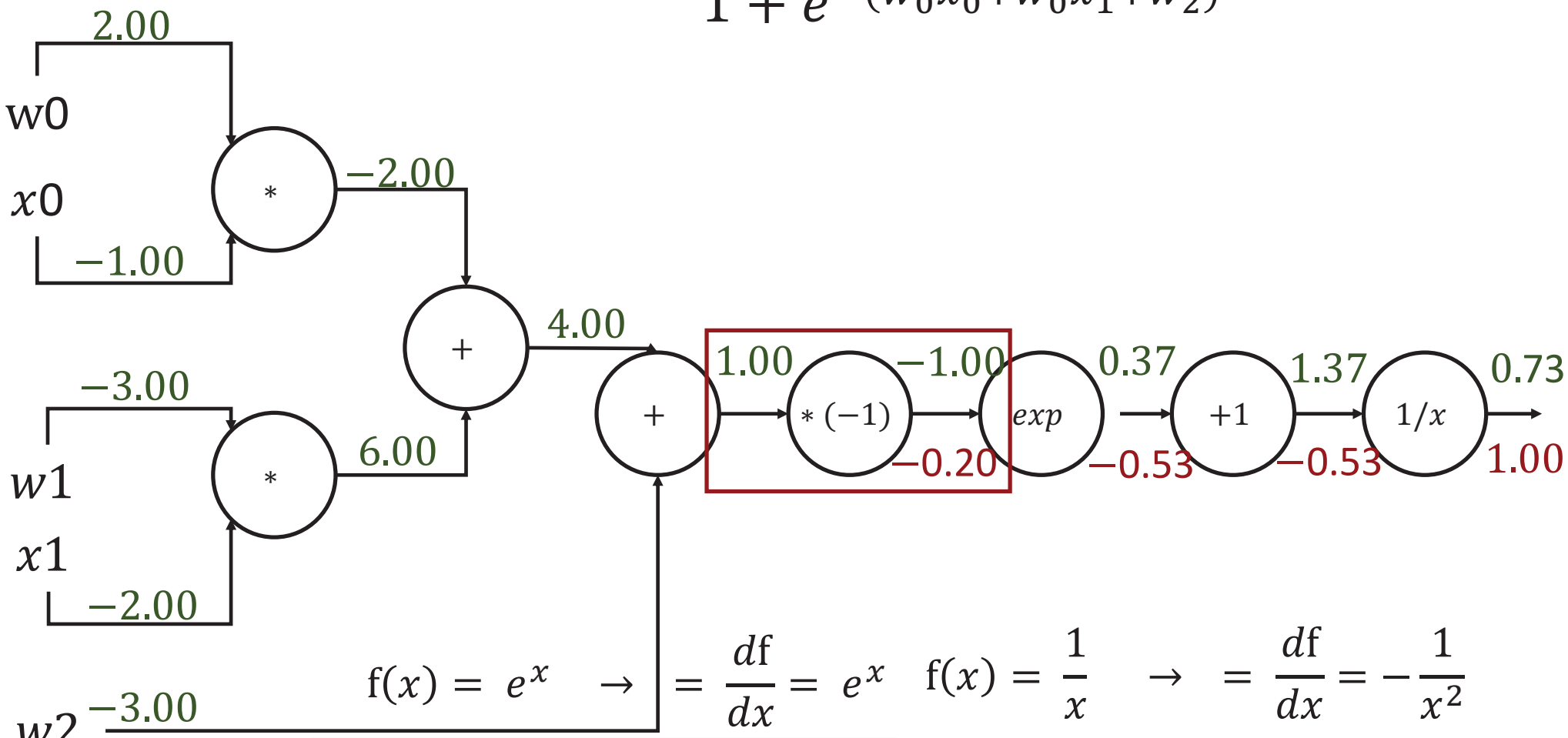
Another example:

$$f(w,x) = \frac{1}{1 + e^{-(w_0x_0 + w_0x_1 + w_2)}}$$



Another example:

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



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

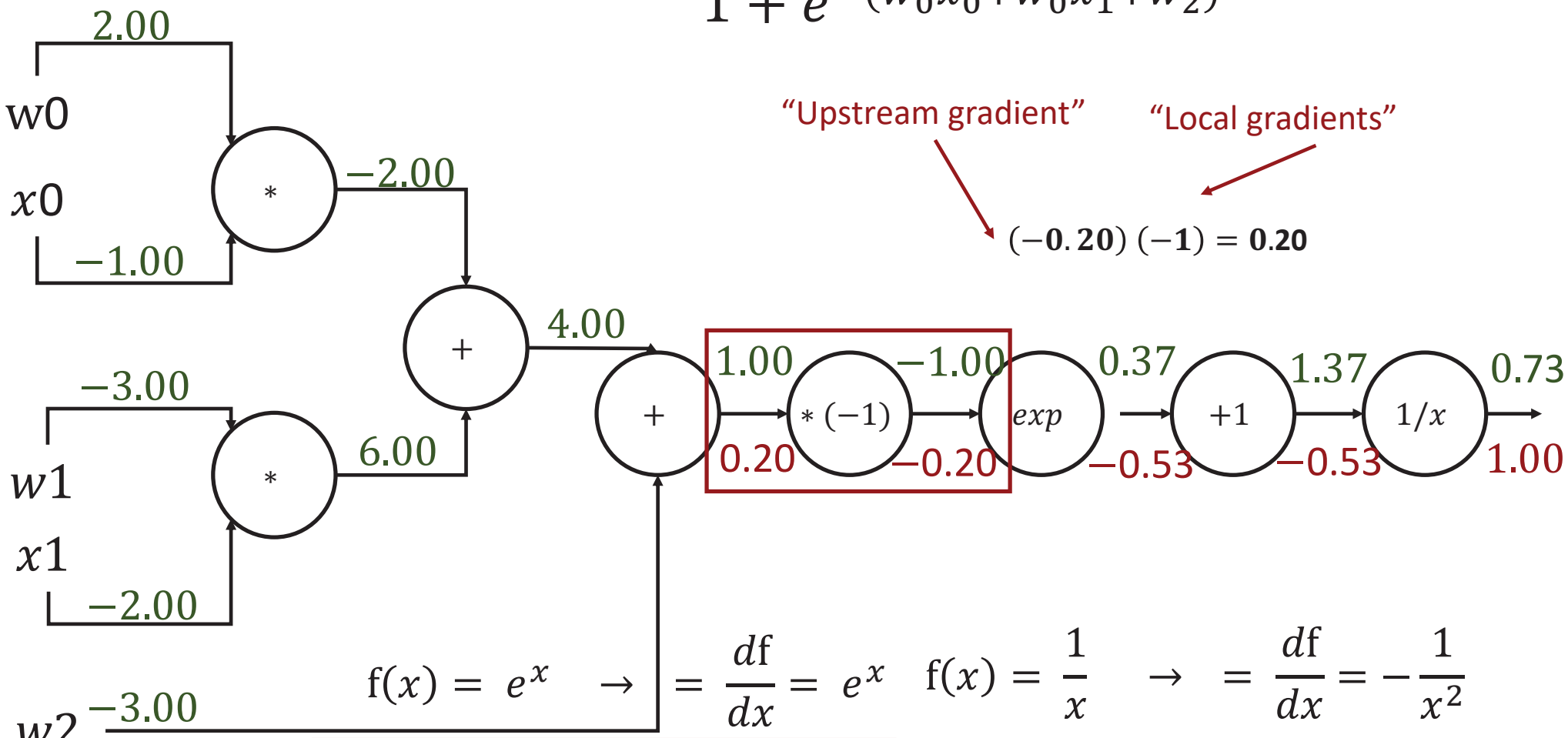
$$\mathrm{f}(x) = \frac{1}{x} \quad \rightarrow \quad = \frac{df}{dx} = -\frac{1}{x^2}$$

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

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

Another example:

$$f(w,x) = \frac{1}{1 + e^{-(w_0x_0 + w_0x_1 + w_2)}}$$

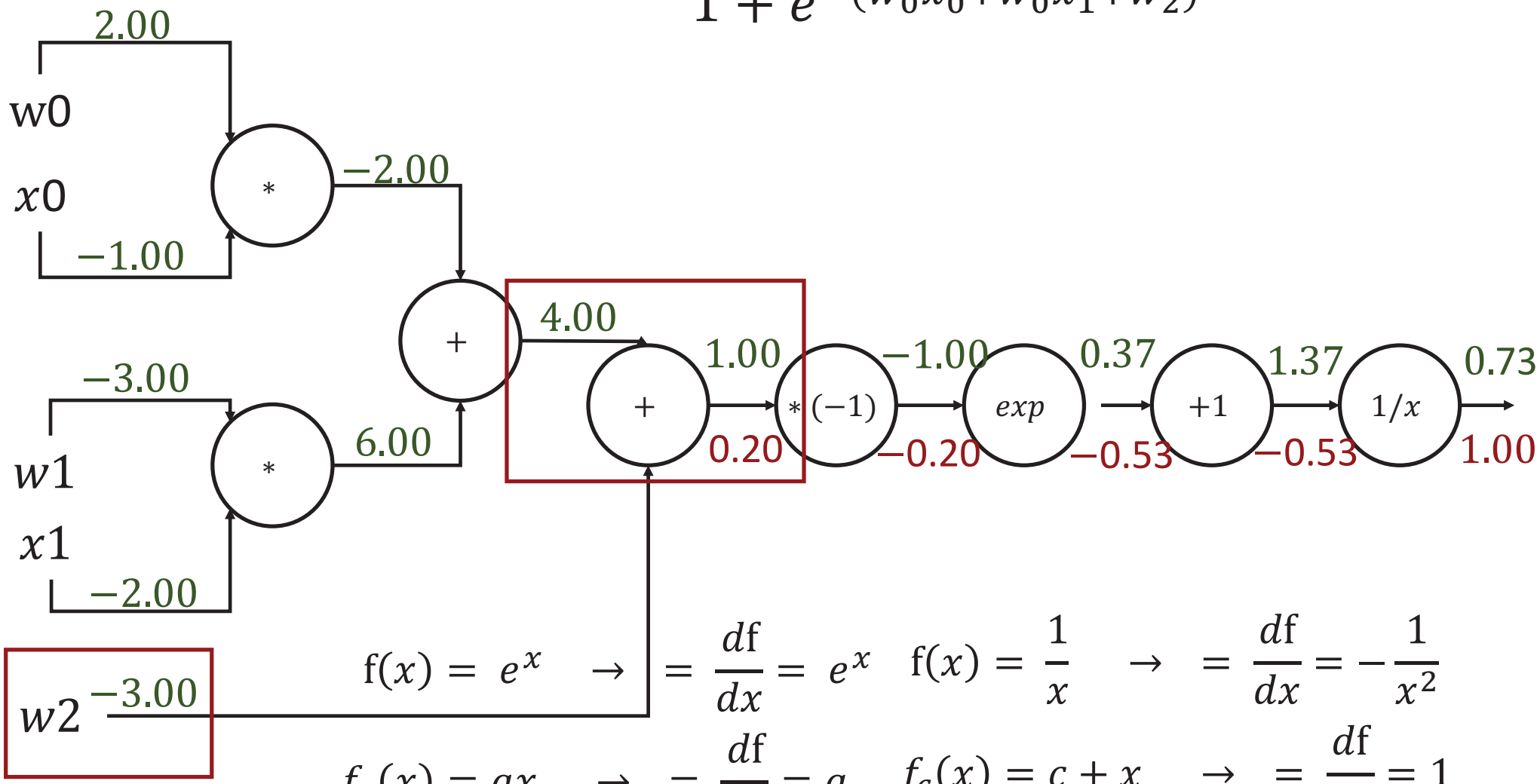


$$\mathrm{f}(x) = e^x \quad \rightarrow \quad = \frac{d\mathrm{f}}{dx} = e^x \qquad \mathrm{f}(x) = \frac{1}{x} \quad \rightarrow \quad = \frac{d\mathrm{f}}{dx} = -\frac{1}{x^2}$$

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

Another example:

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

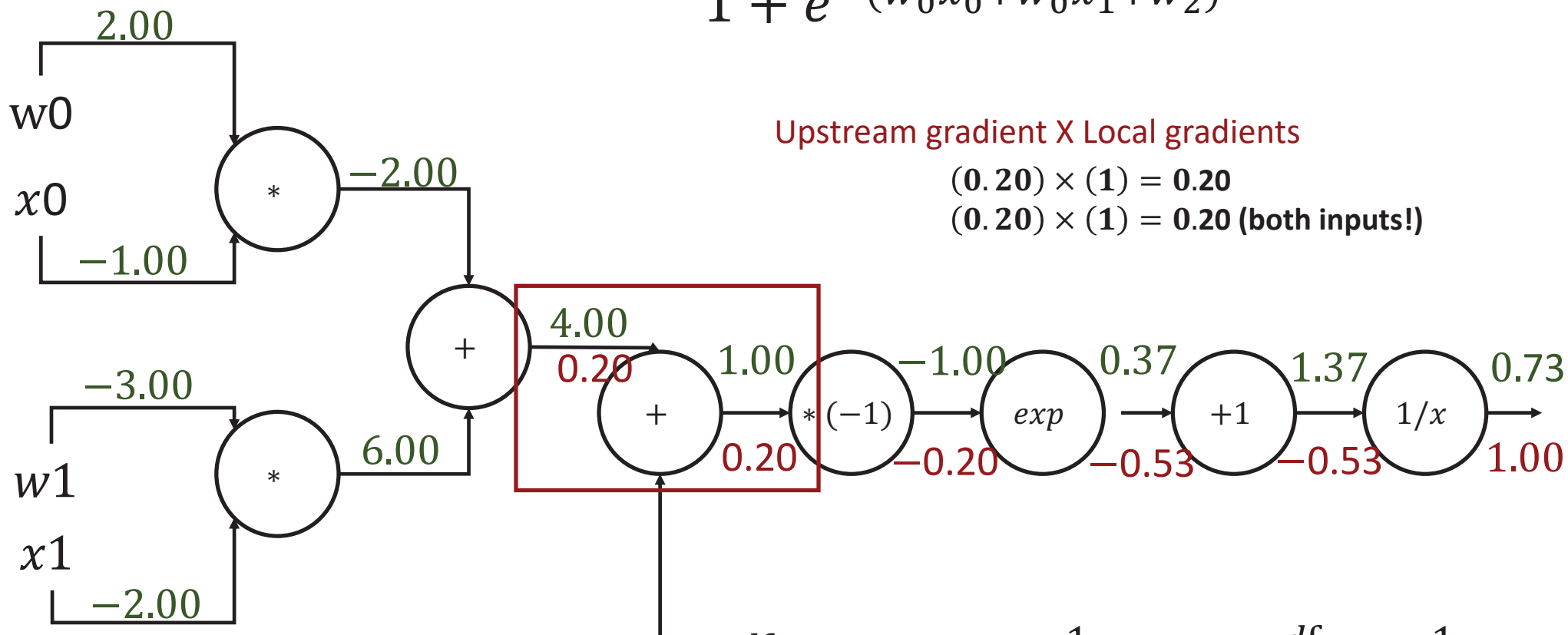


$$\mathrm{f}(x) = e^x \quad \rightarrow \quad = \frac{d\mathrm{f}}{dx} = e^x \qquad \mathrm{f}(x) = \frac{1}{x} \quad \rightarrow \quad = \frac{d\mathrm{f}}{dx} = -\frac{1}{x^2}$$

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

Another example:

$$f(w,x) = \frac{1}{1 + e^{-(w_0x_0 + w_0x_1 + w_2)}}$$



Upstream gradient X Local gradients

$(\mathbf{0.20}) \times (\mathbf{1}) = \mathbf{0.20}$

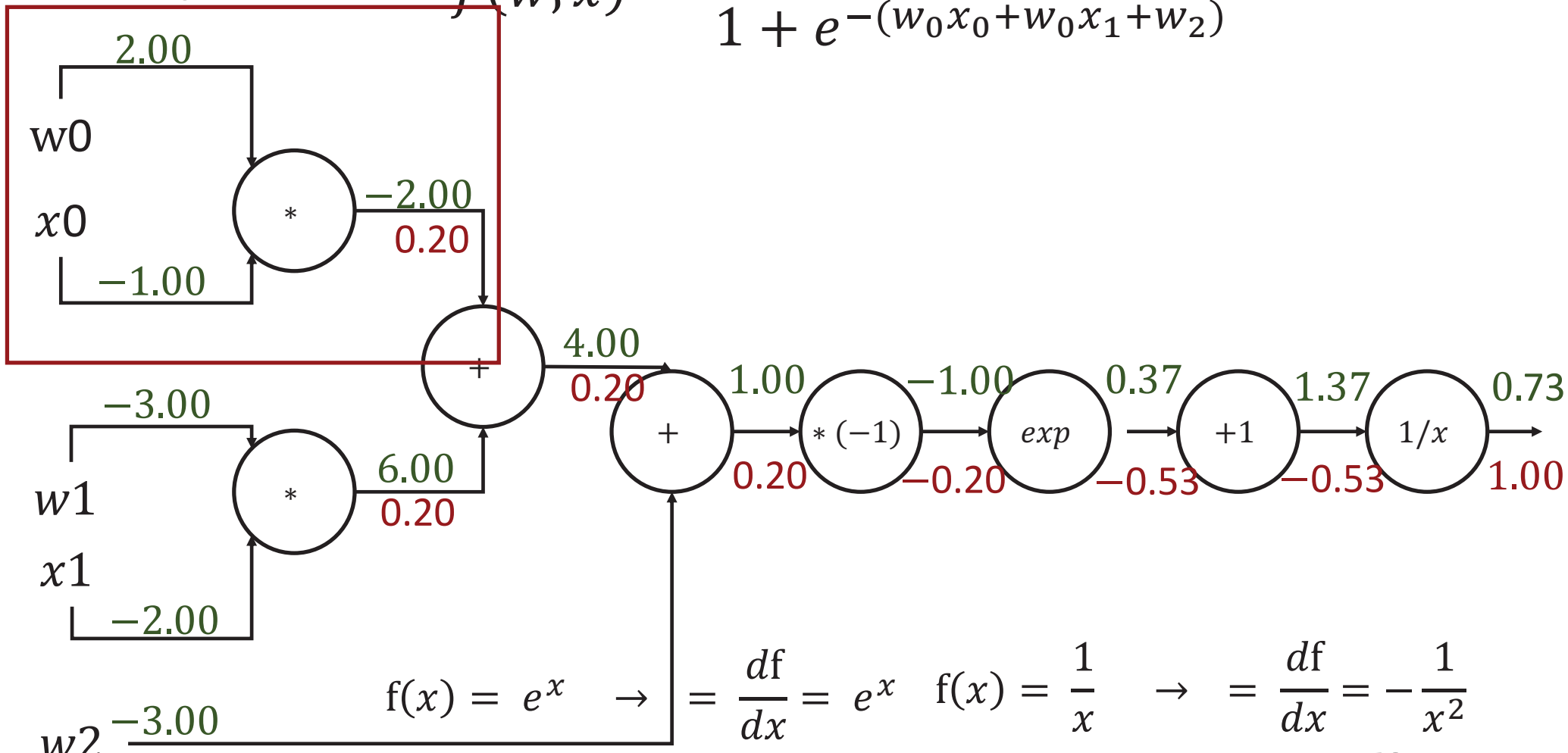
$(\mathbf{0.20}) \times (\mathbf{1}) = \mathbf{0.20}$ **(both inputs!)**

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

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

Another example:

$$f(w,x) = \frac{1}{1 + e^{-(w_0x_0 + w_0x_1 + w_2)}}$$

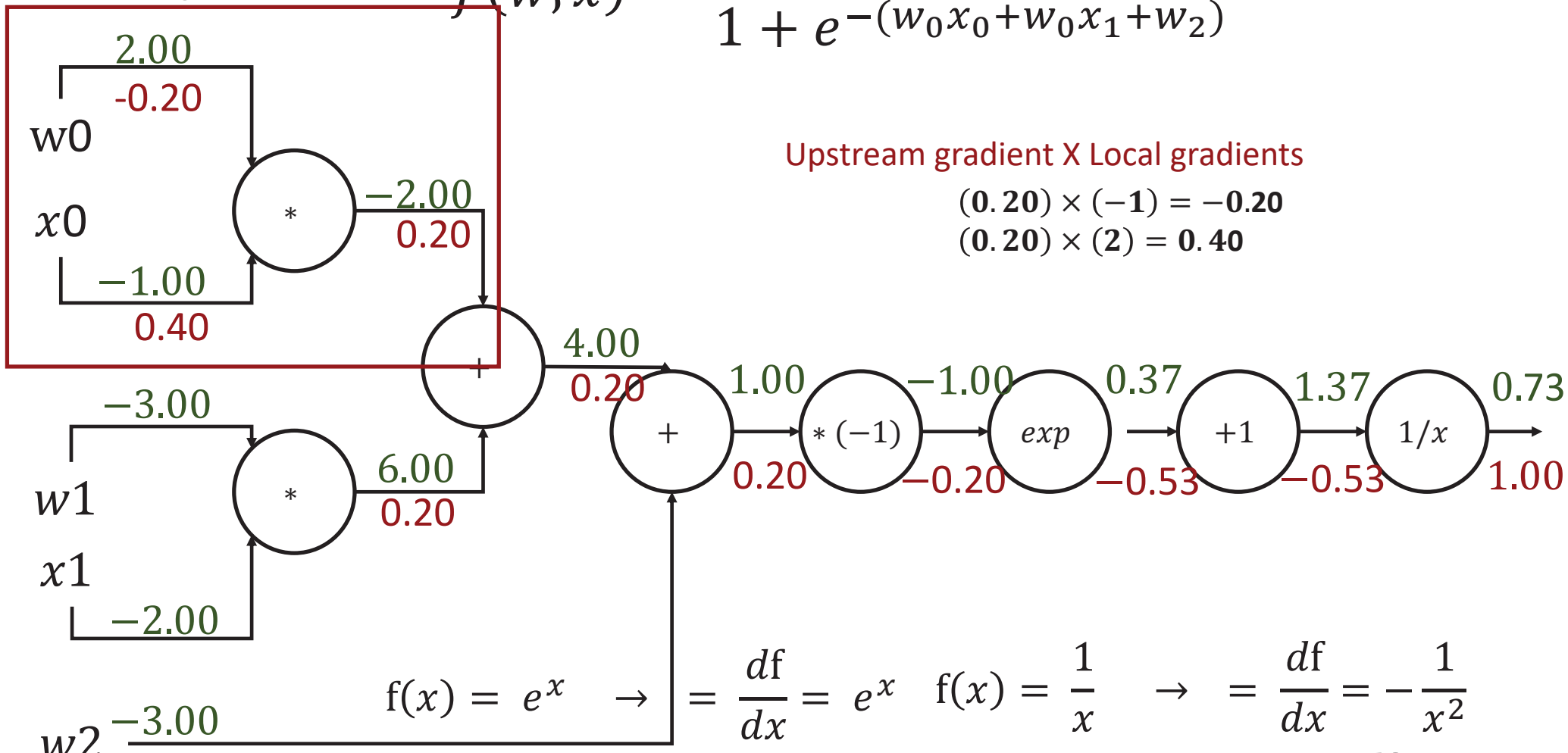


$$\mathrm{f}(x) = e^x \quad \rightarrow \quad = \frac{d\mathrm{f}}{dx} = e^x \qquad \mathrm{f}(x) = \frac{1}{x} \quad \rightarrow \quad = \frac{d\mathrm{f}}{dx} = -\frac{1}{x^2}$$

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

Another example:

$$f(w,x) = \frac{1}{1 + e^{-(w_0x_0 + w_0x_1 + w_2)}}$$



Upstream gradient X Local gradients

$(\mathbf{0.20}) \times (\mathbf{-1}) = \mathbf{-0.20}$

$(\mathbf{0.20}) \times (\mathbf{2}) = \mathbf{0.40}$

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

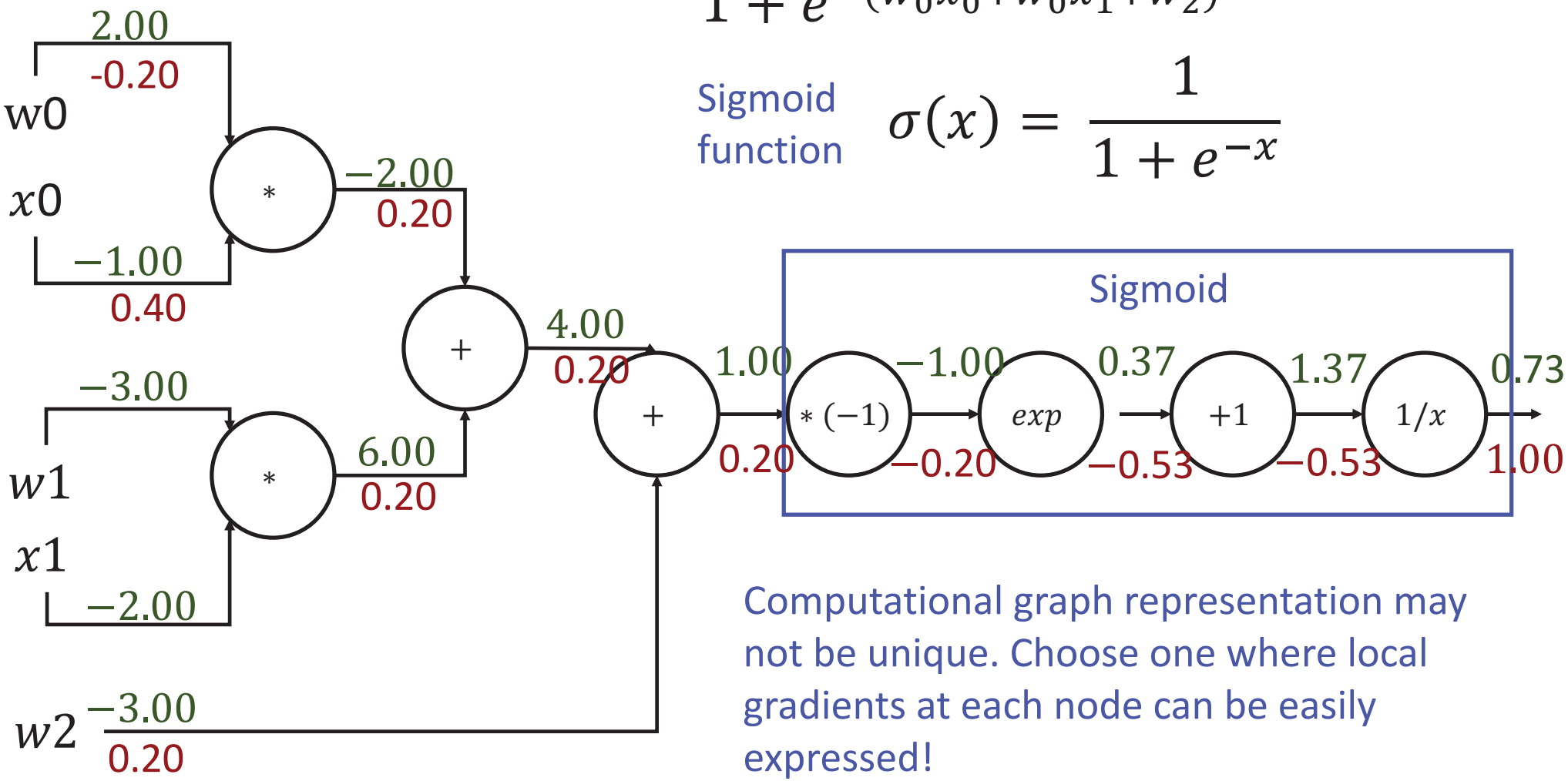
$$f_a(x) = ax \quad \rightarrow \quad = \frac{df}{dx} = a \qquad f_c(x) = c + x \quad \rightarrow \quad = \frac{df}{dx} = 1$$

Another example:

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_0x_1+w_2)}}$$

Sigmoid function

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

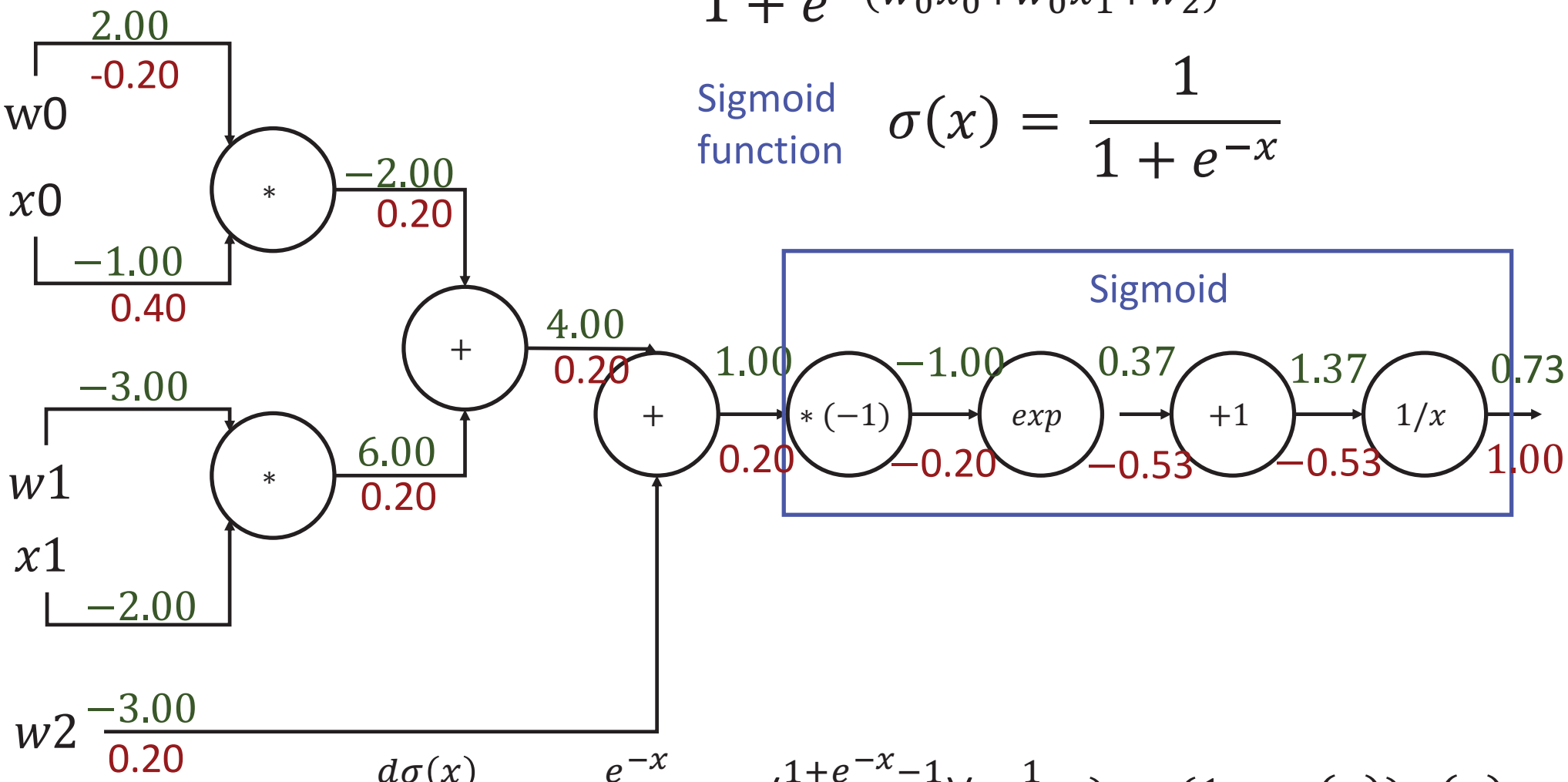


Computational graph representation may not be unique. Choose one where local gradients at each node can be easily expressed!

Another example:

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_0x_1+w_2)}}$$

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



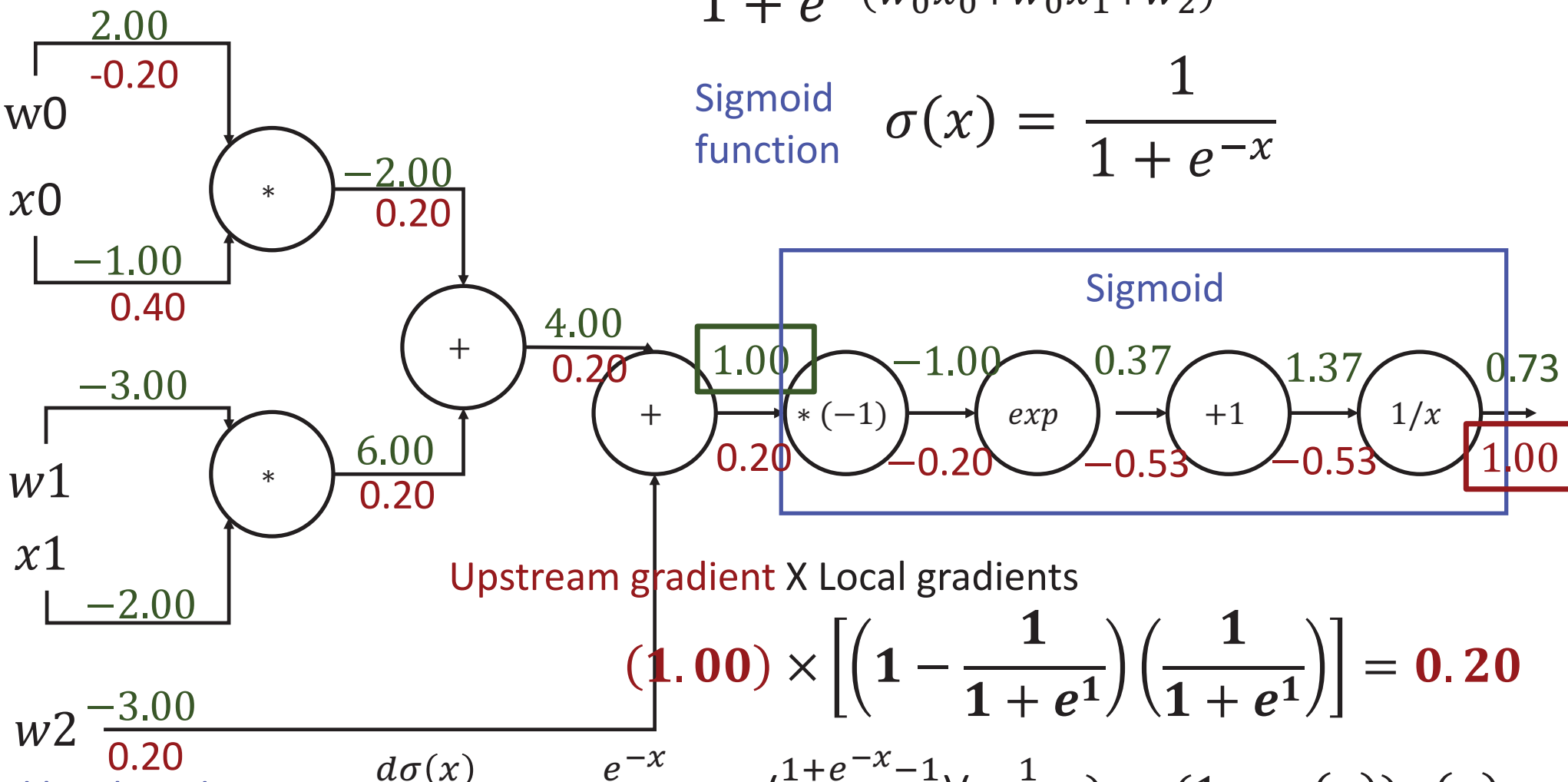
Sigmoid local gradient:

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

Another example:

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_0x_1+w_2)}}$$

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



Upstream gradient X Local gradients

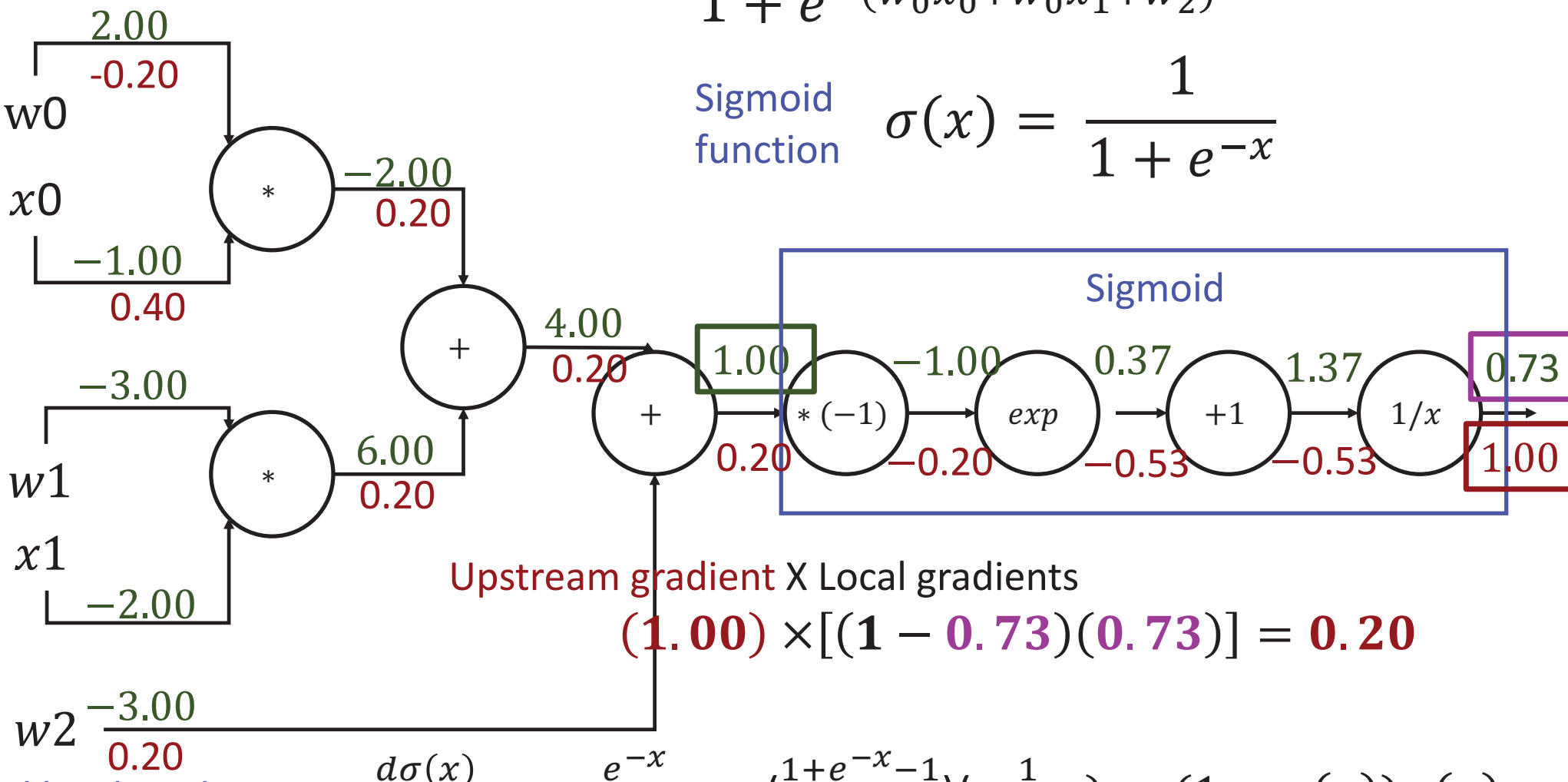
$$(\mathbf{1.00}) \times \left[\left(\mathbf{1} - \frac{\mathbf{1}}{\mathbf{1}+\boldsymbol{e}^{\mathbf{1}}}\right)\left(\frac{\mathbf{1}}{\mathbf{1}+\boldsymbol{e}^{\mathbf{1}}}\right)\right] = \mathbf{0.20}$$

Sigmoid local gradient: $\frac{d\sigma(x)}{dx} = \frac{e^{-x}}{(1+e^{-x})^2} = (\frac{1+e^{-x}-1}{1+e^{-x}})(\frac{1}{1+e^{-x}}) = (1-\sigma(x))\sigma(x)$

Another example:

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_0x_1+w_2)}}$$

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



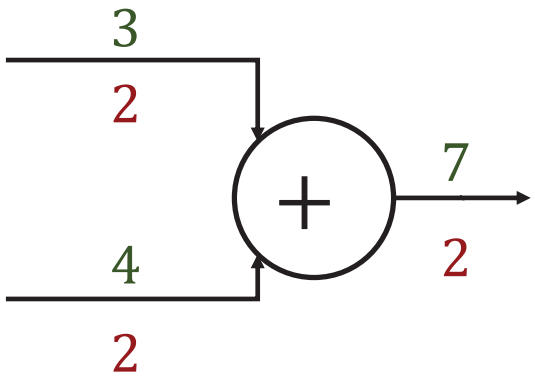
Upstream gradient X Local gradients

$$(\mathbf{1.00}) \times [(\mathbf{1} - \mathbf{0.73})(\mathbf{0.73})] = \mathbf{0.20}$$

Sigmoid local gradient: $\frac{d\sigma(x)}{dx} = \frac{e^{-x}}{(1+e^{-x})^2} = (\frac{1+e^{-x}-1}{1+e^{-x}})(\frac{1}{1+e^{-x}}) = (1-\sigma(x))\sigma(x)$

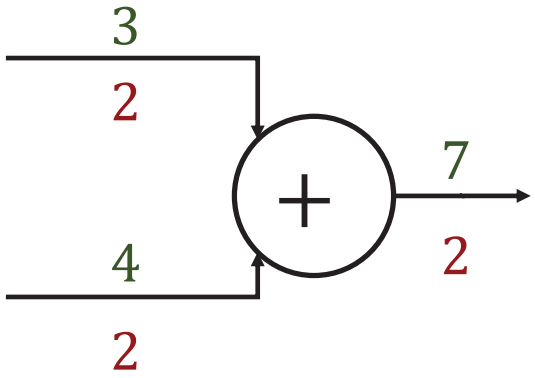
Patterns in gradient flow

add gate: gradient distributor

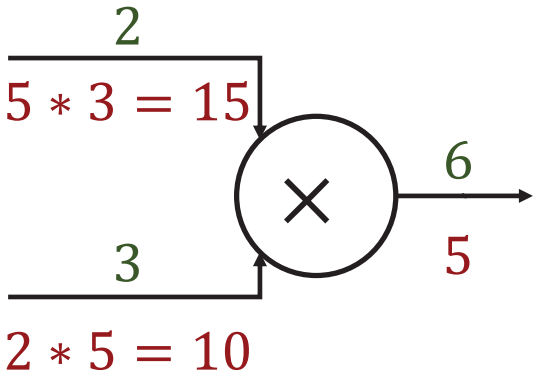


Patterns in gradient flow

add gate: gradient distributor

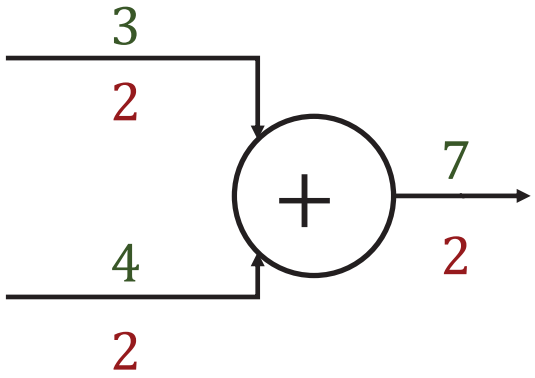


mul gate: "swap multiplier"

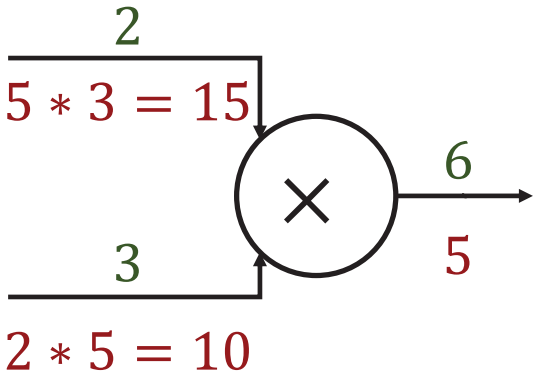


Patterns in gradient flow

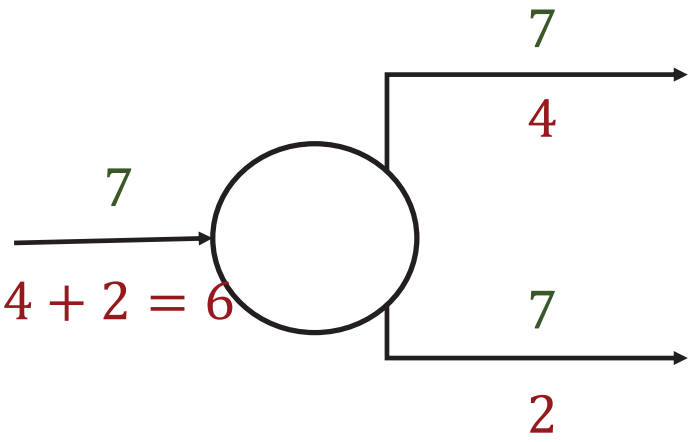
add gate: gradient distributor



mul gate: "swap multiplier"

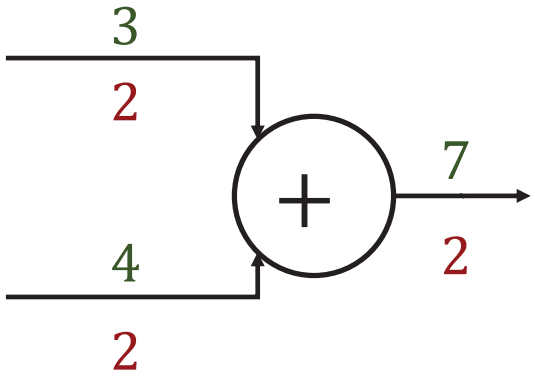


copy gate: gradient adder

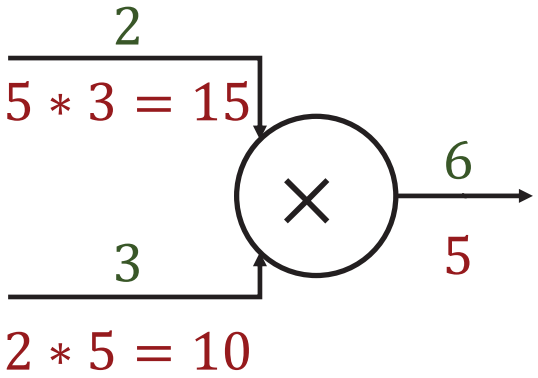


Patterns in gradient flow

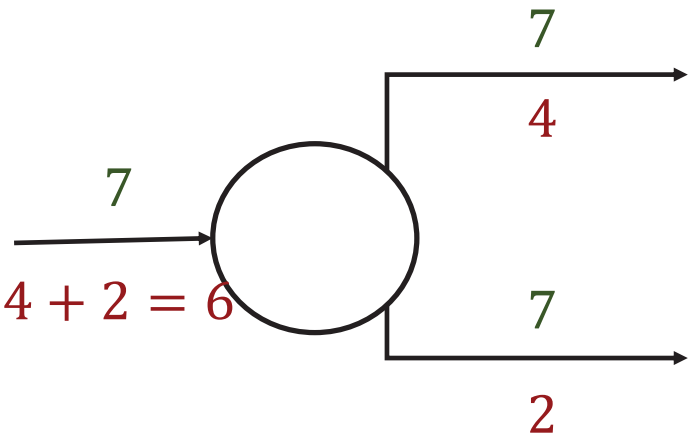
add gate: gradient distributor



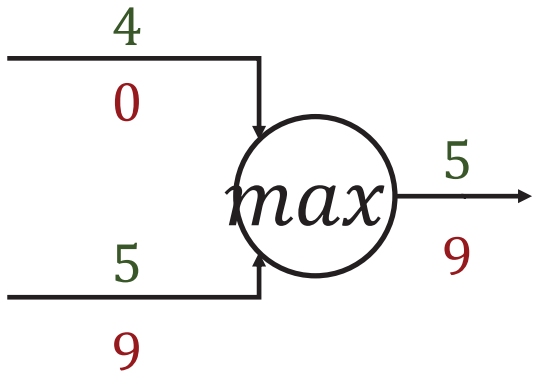
mul gate: “swap multiplier”



copy gate: gradient adder

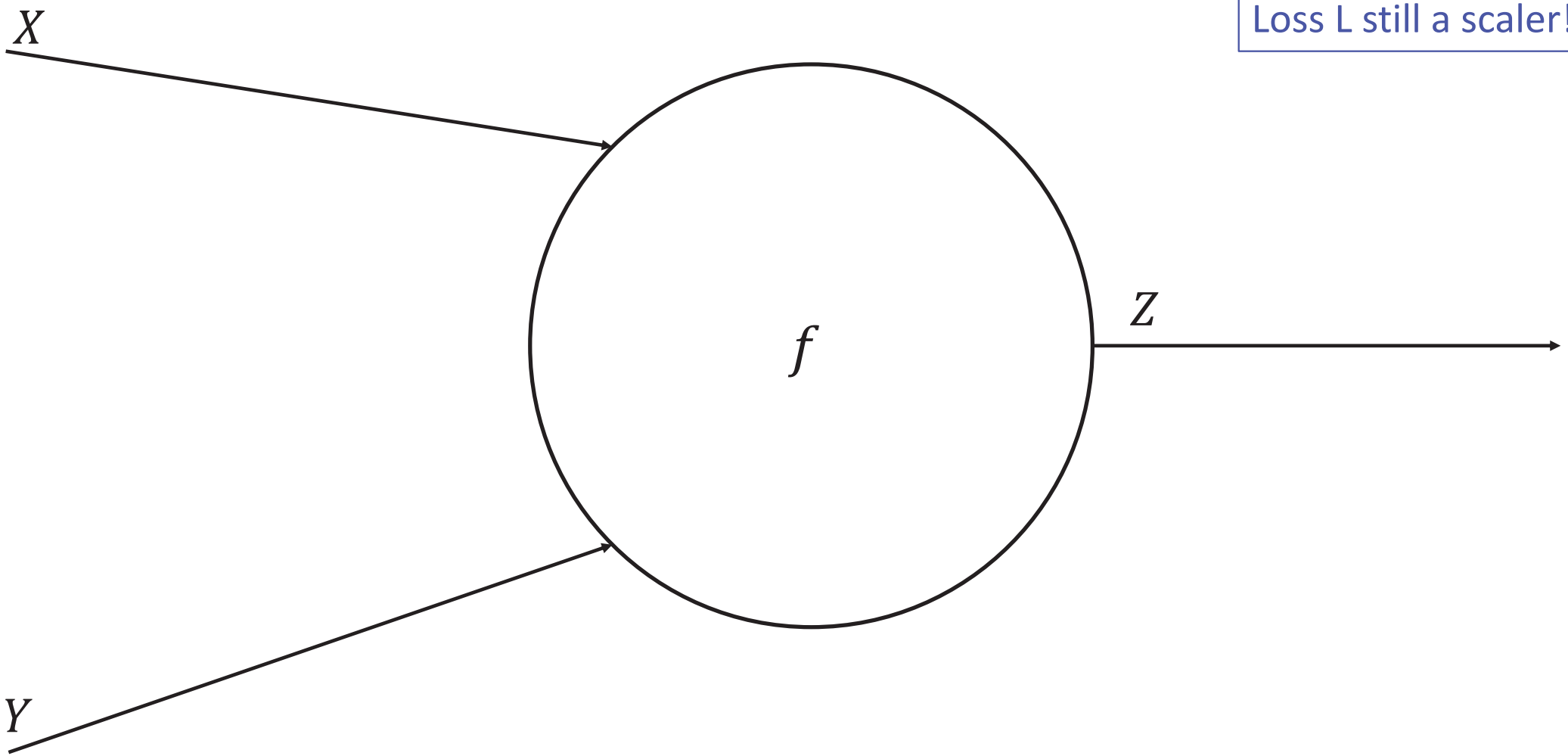


max gate: gradient router

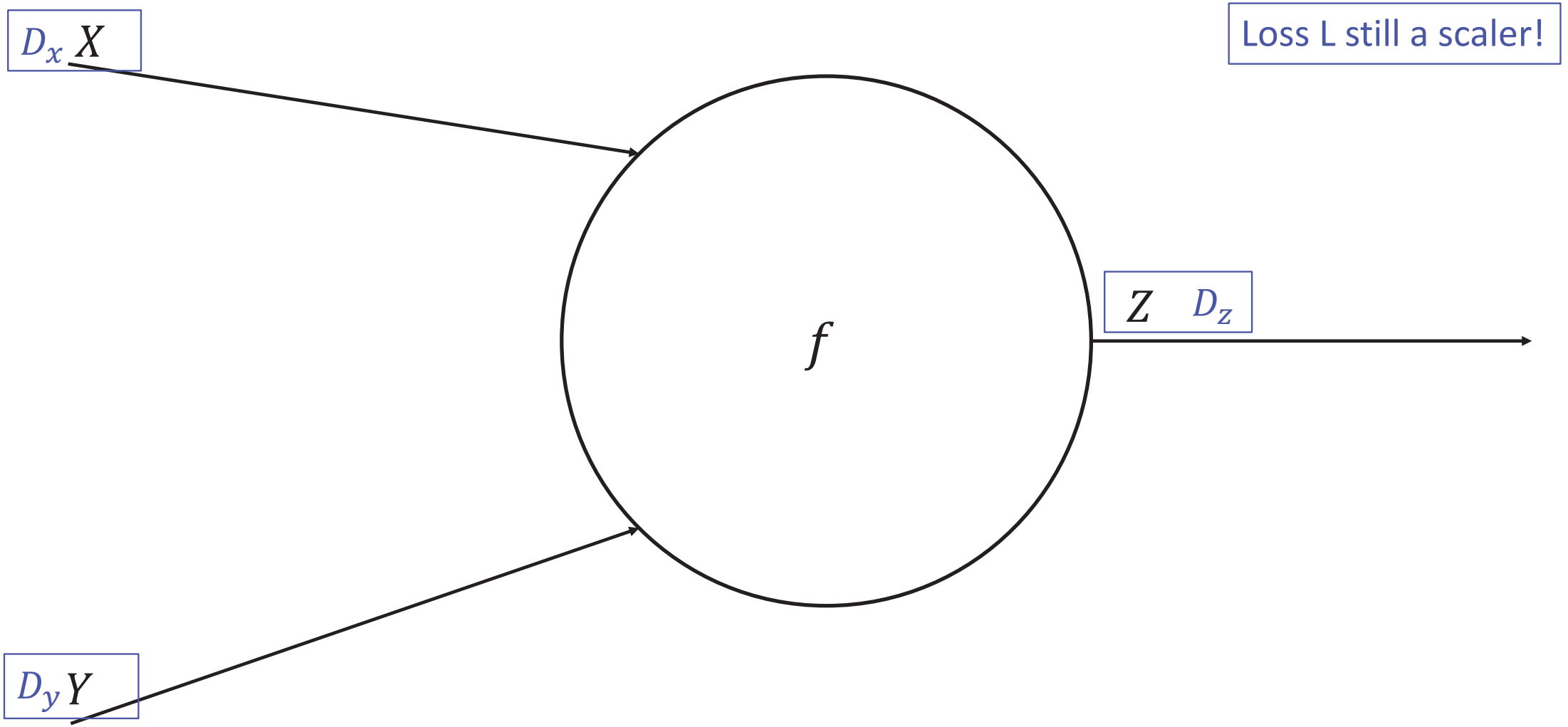


Backpropagation with vectors

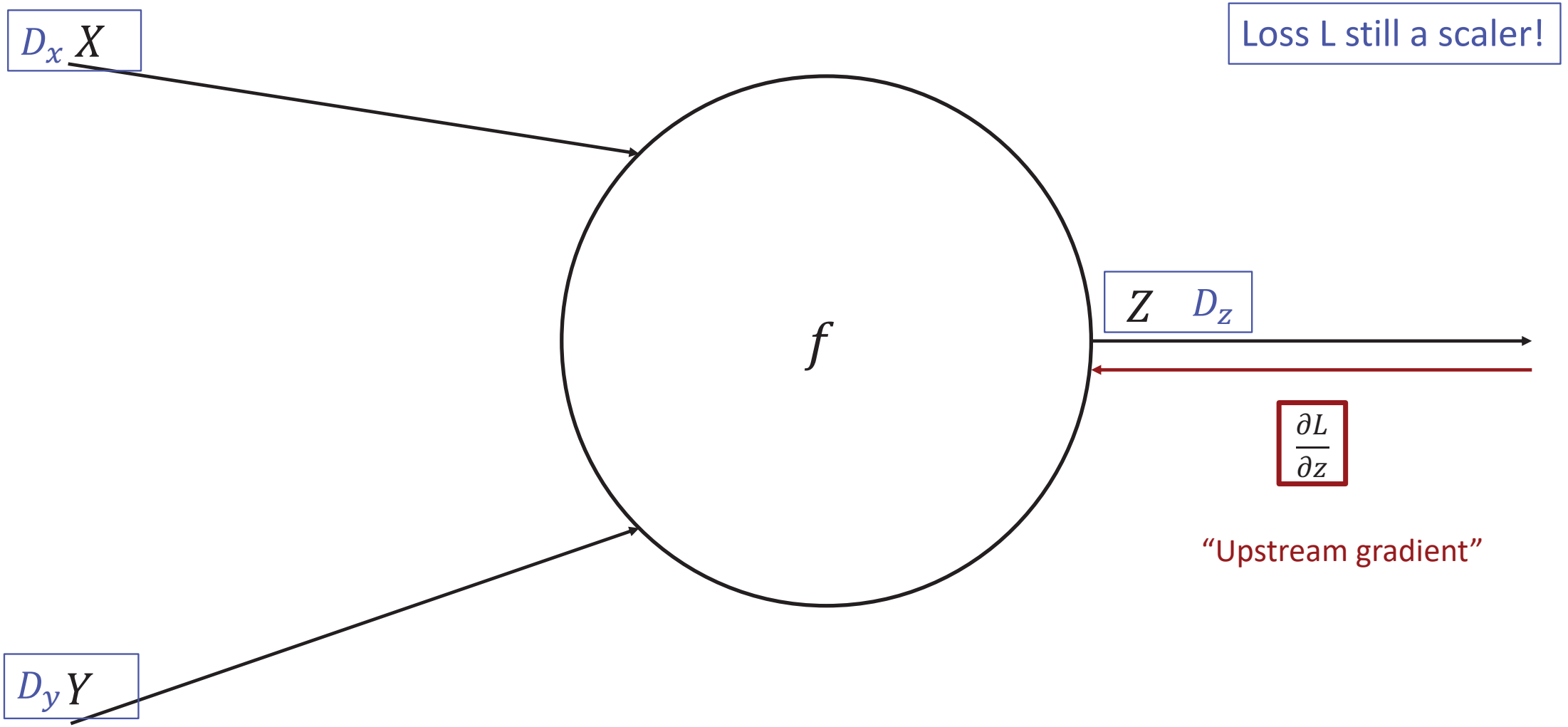
Loss L still a scaler!



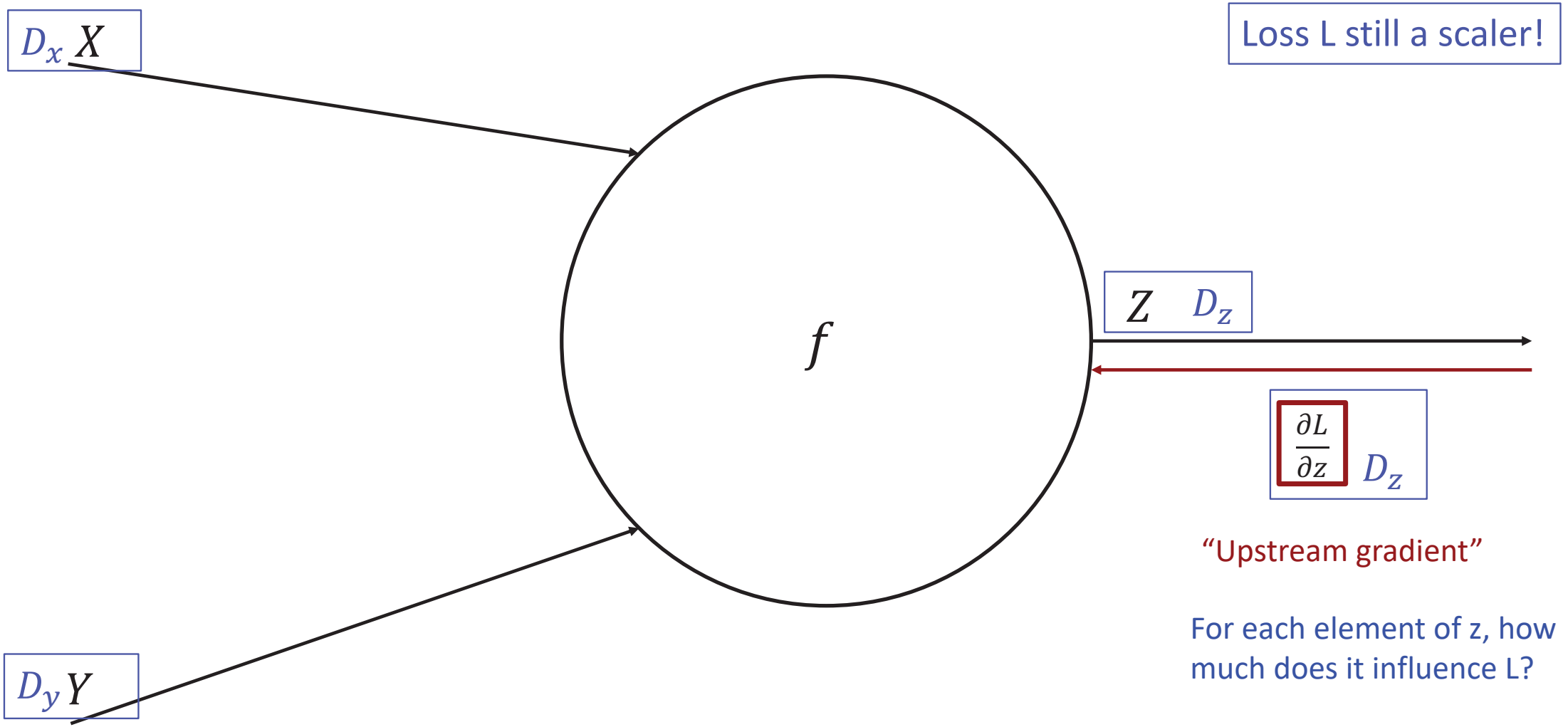
Backpropagation with vectors



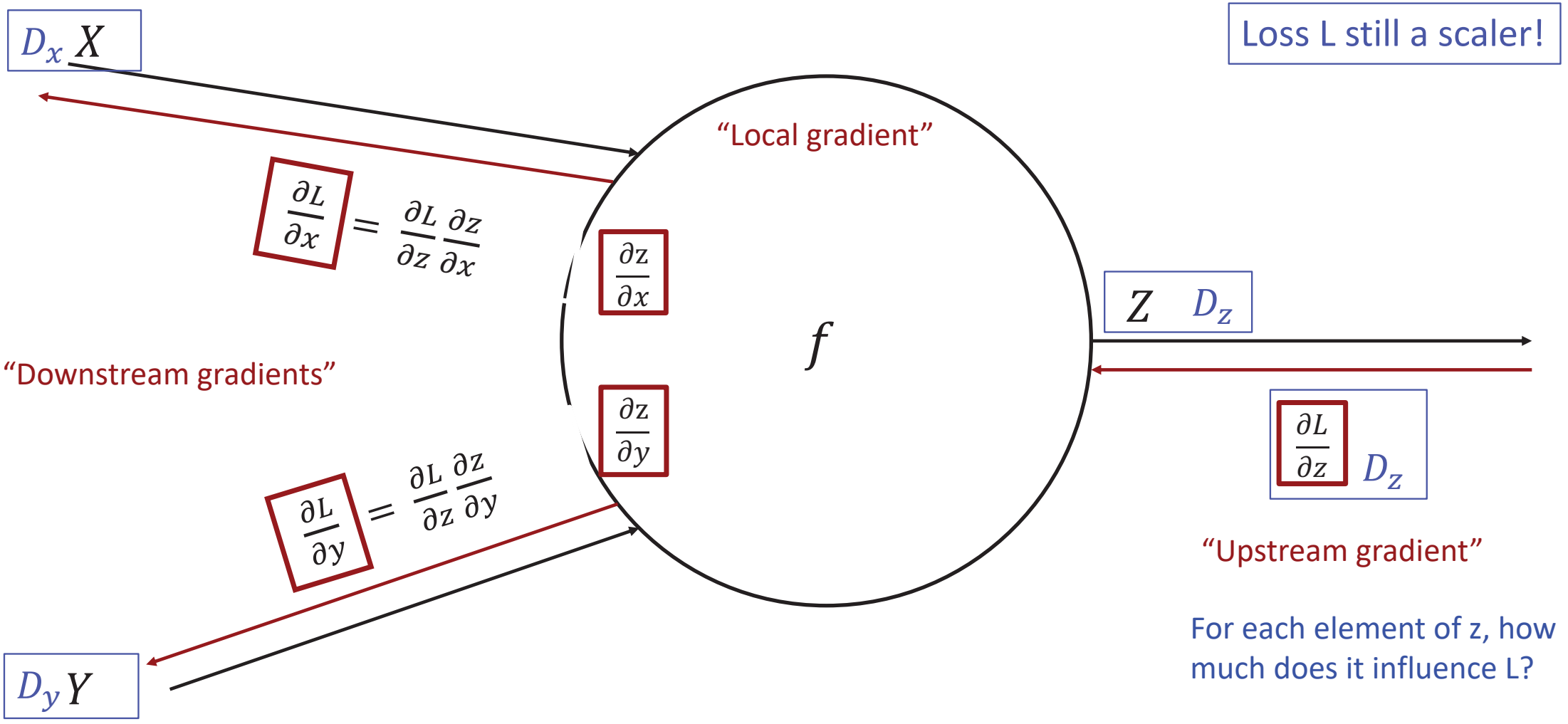
Backpropagation with vectors



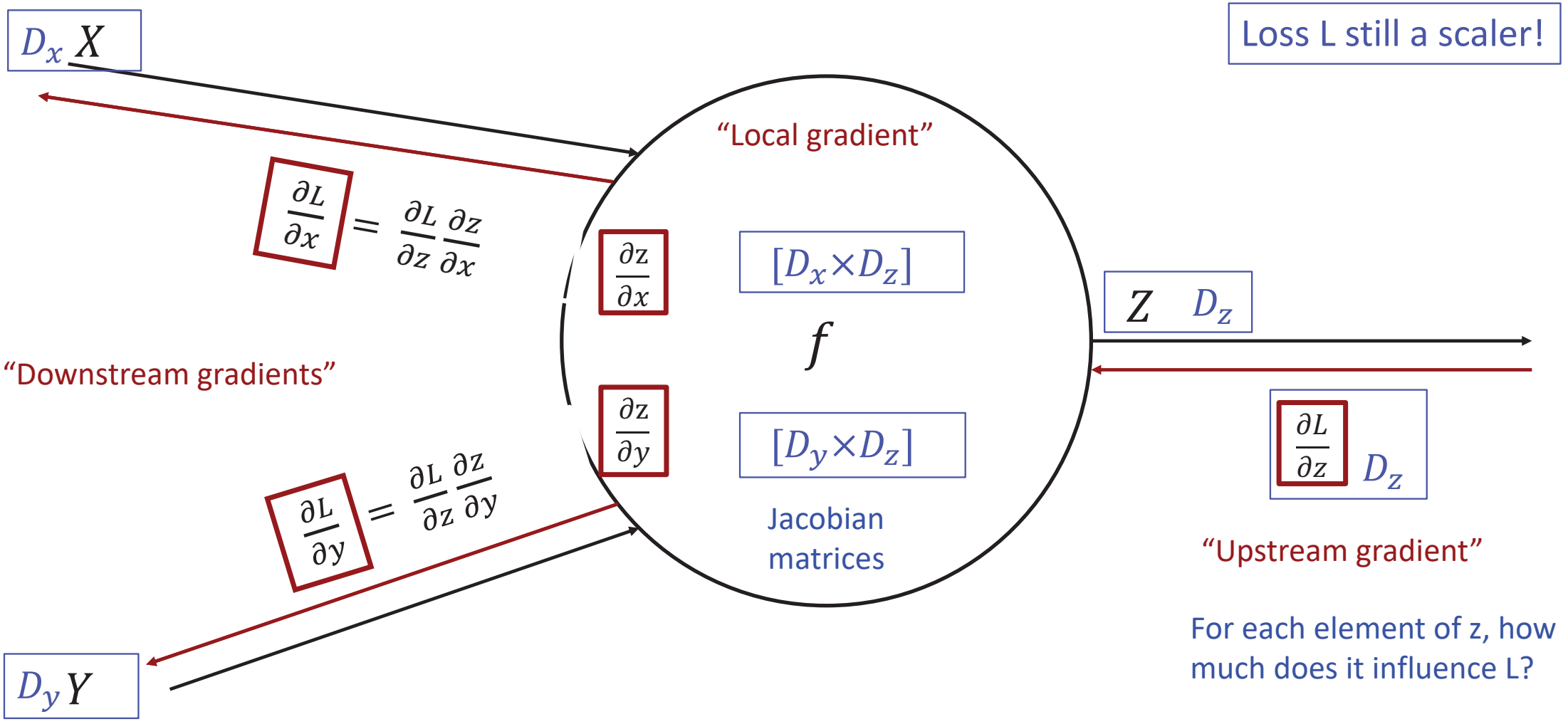
Backpropagation with vectors



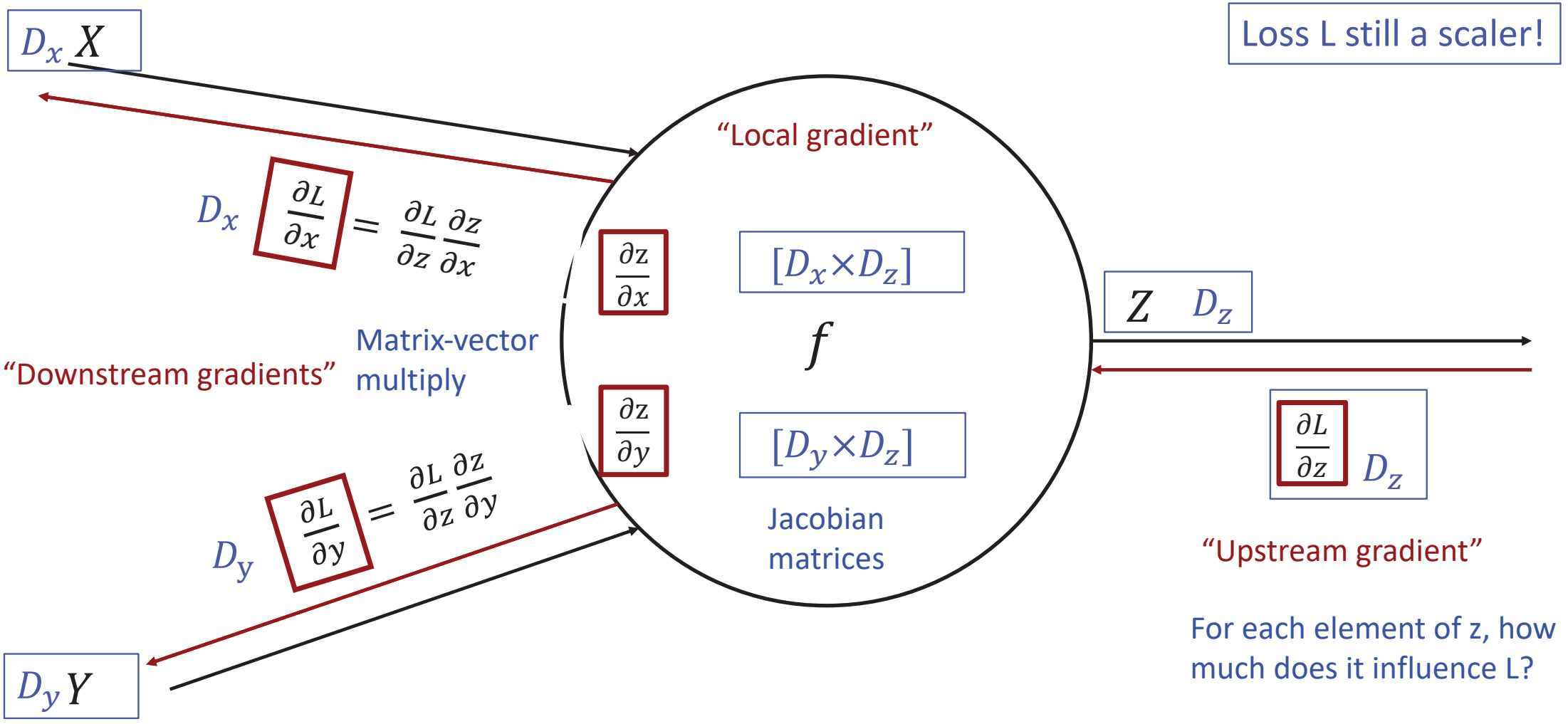
Backpropagation with vectors



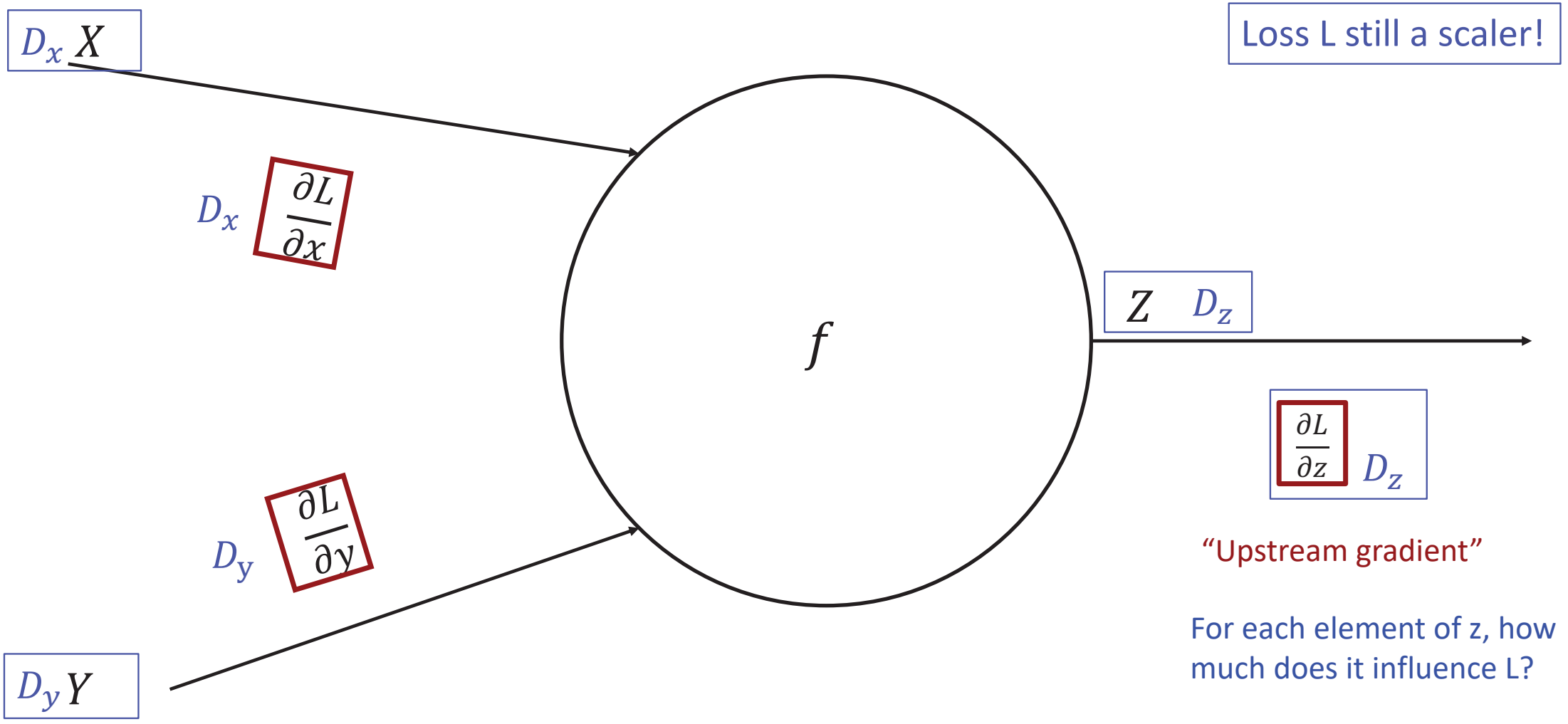
Backpropagation with vectors



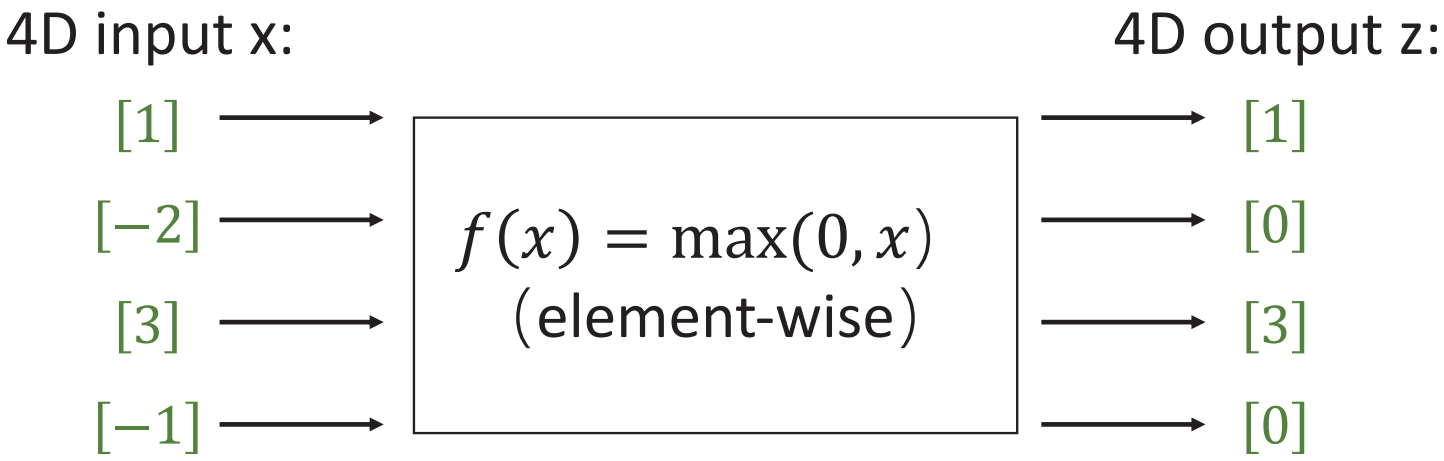
Backpropagation with vectors



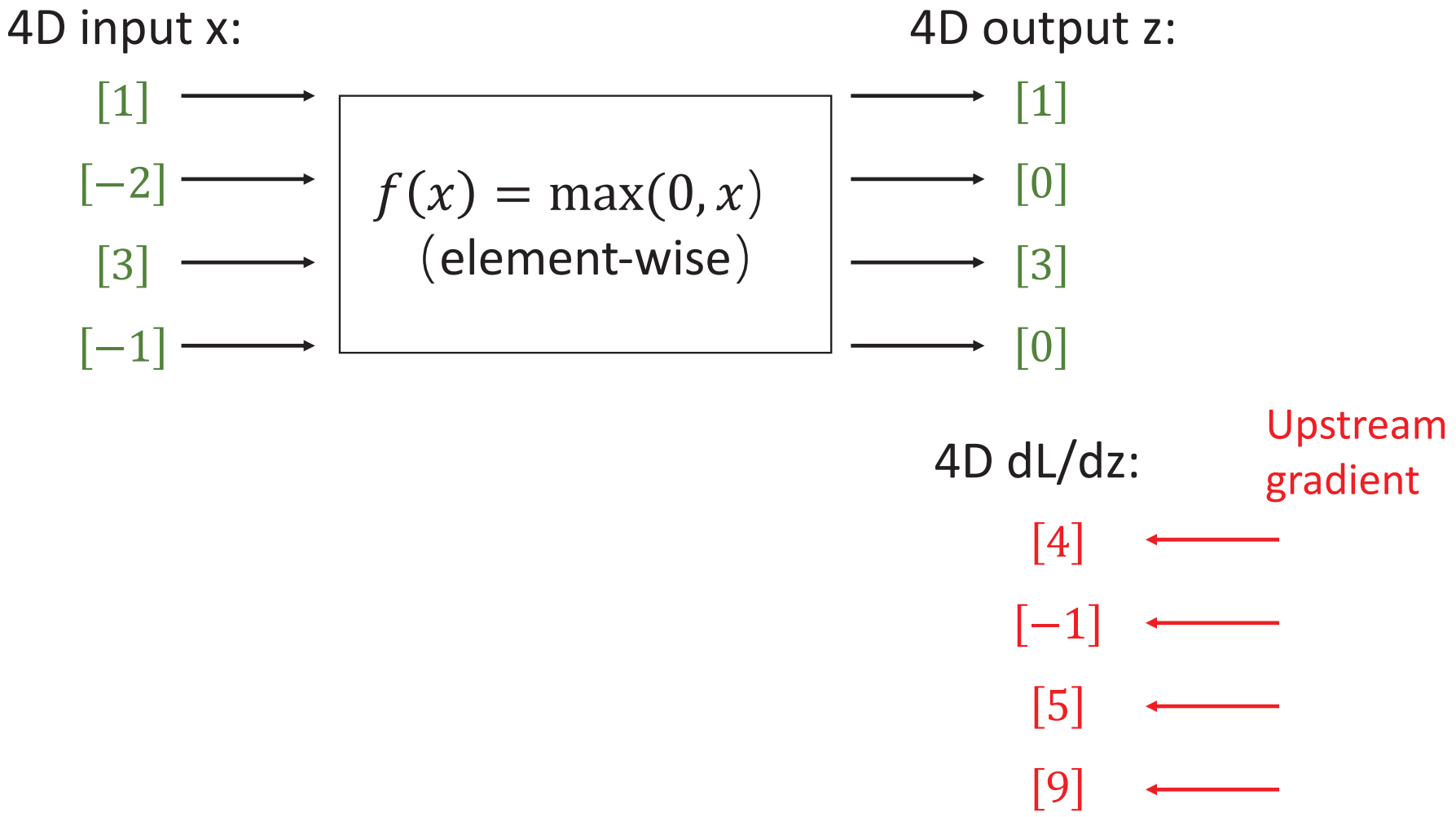
Gradients of variables wrt loss have same dims as the original variable



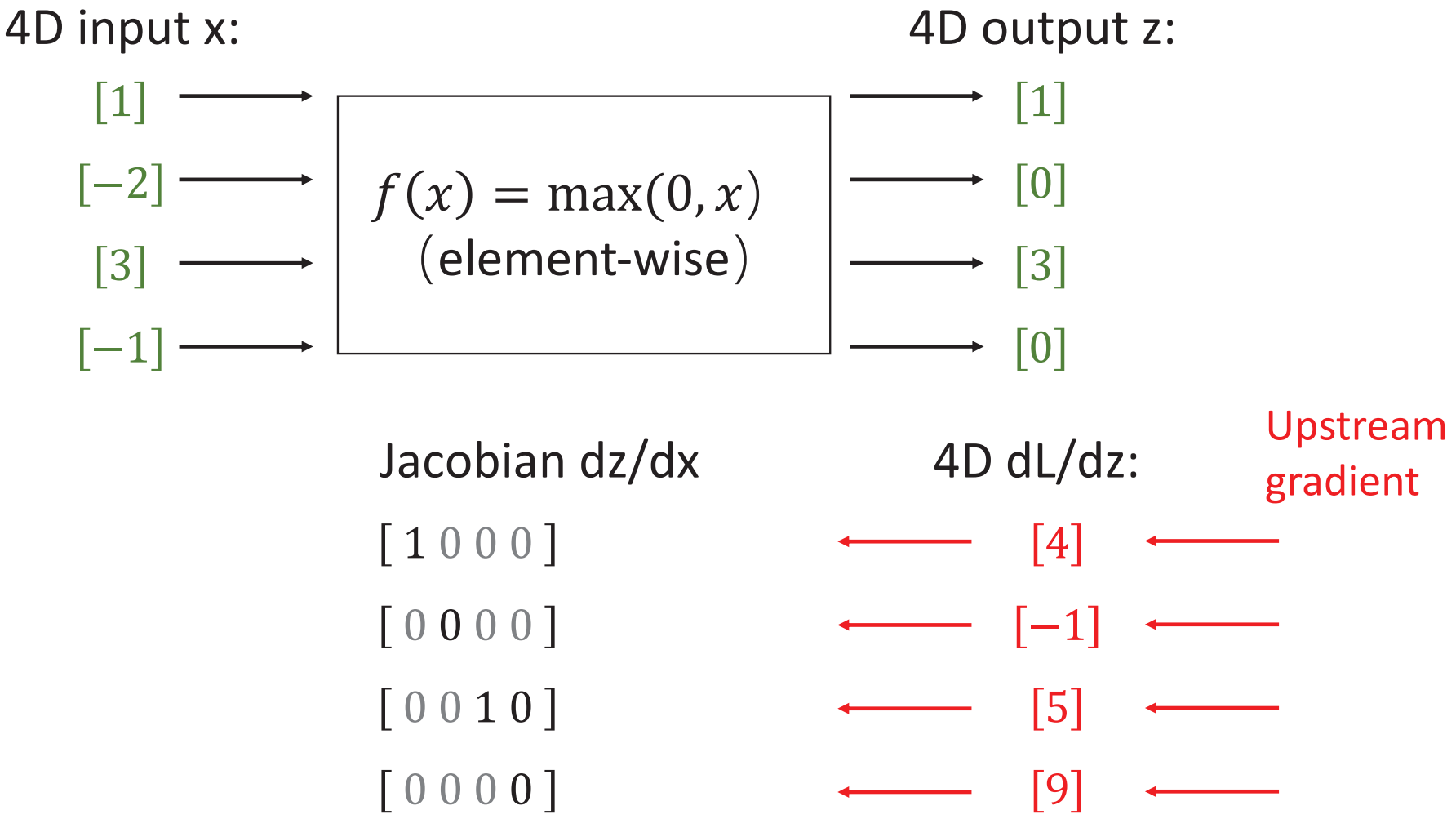
Backpropagation with vectors



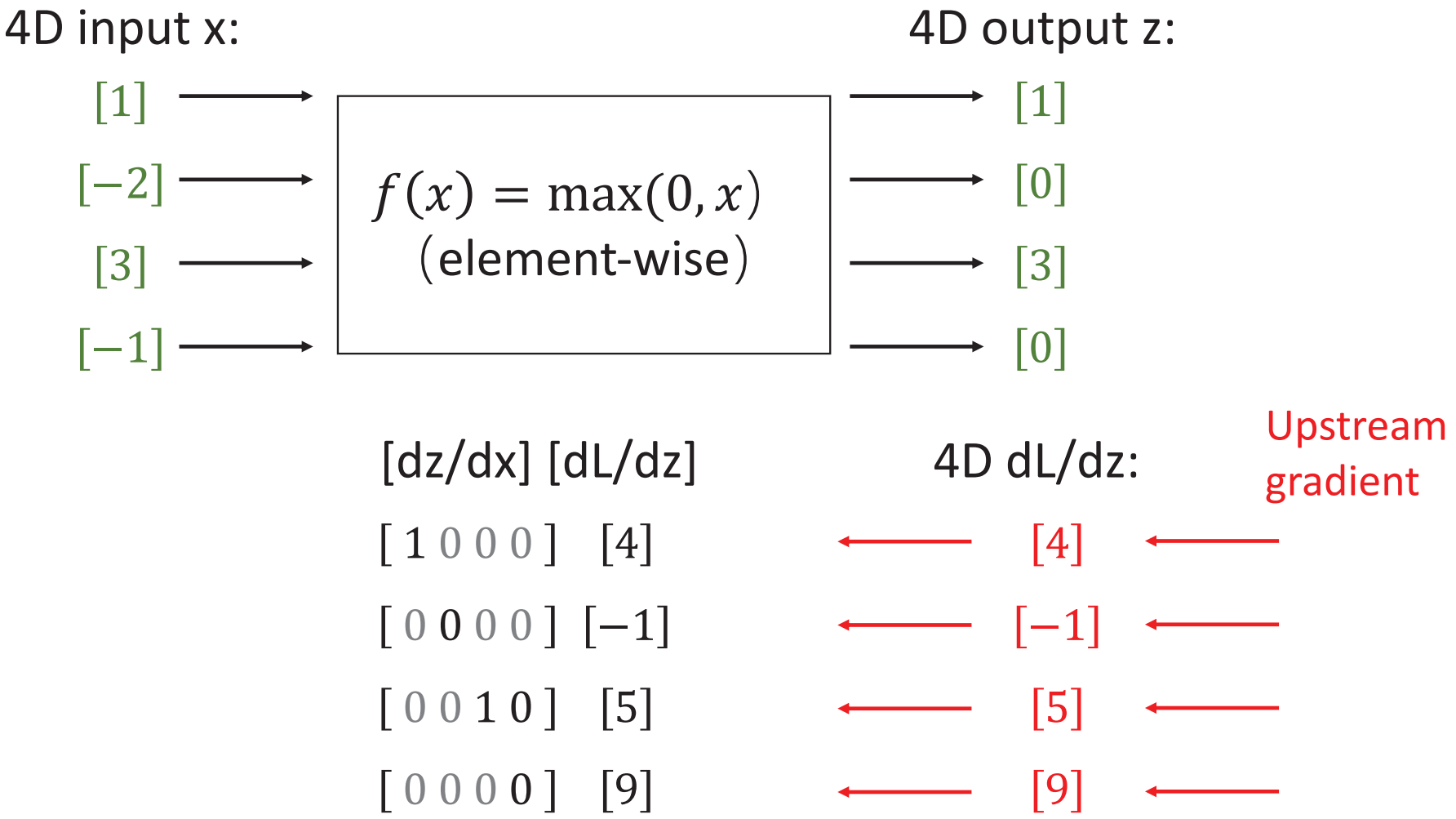
Backpropagation with vectors



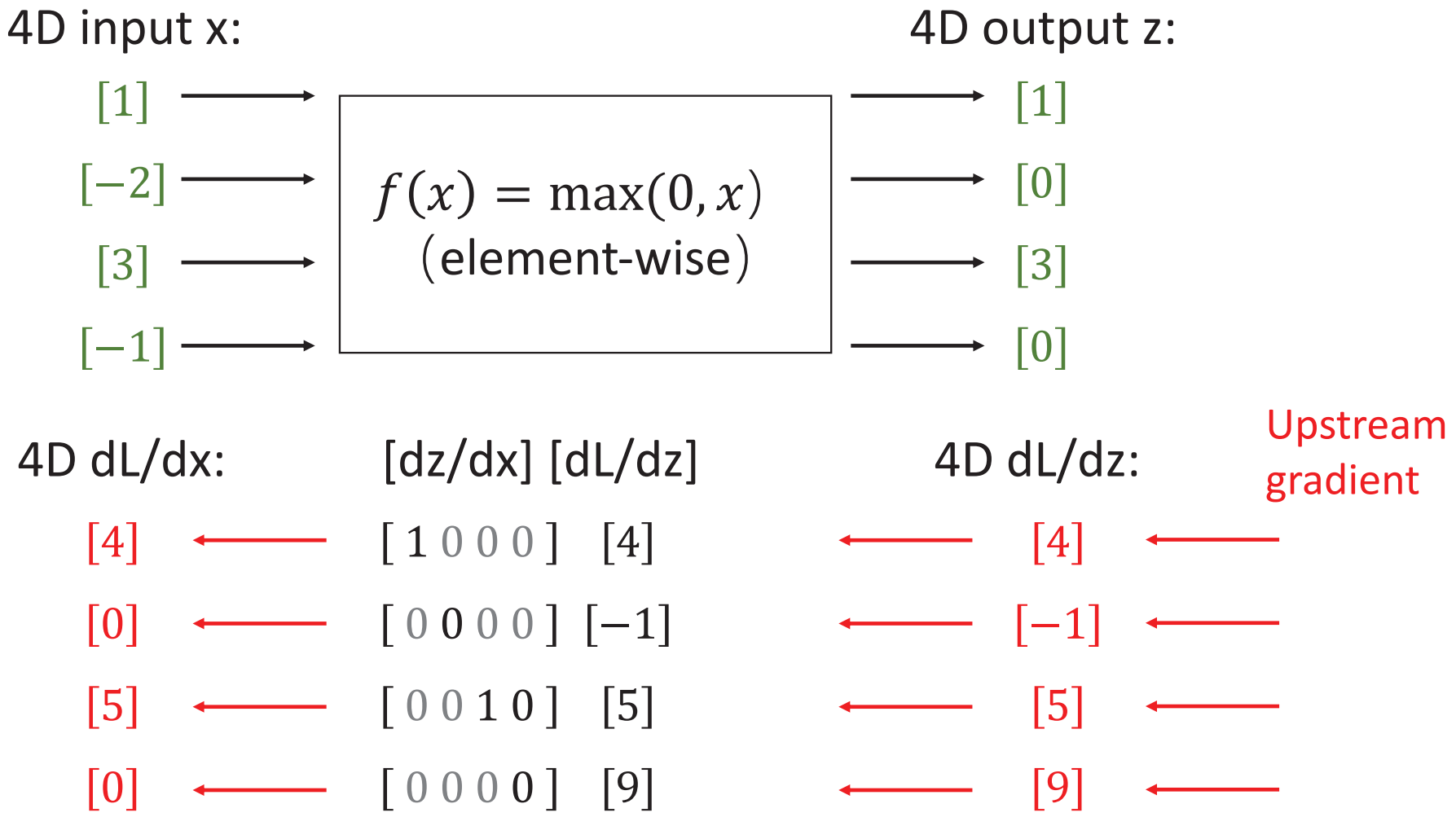
Backpropagation with vectors



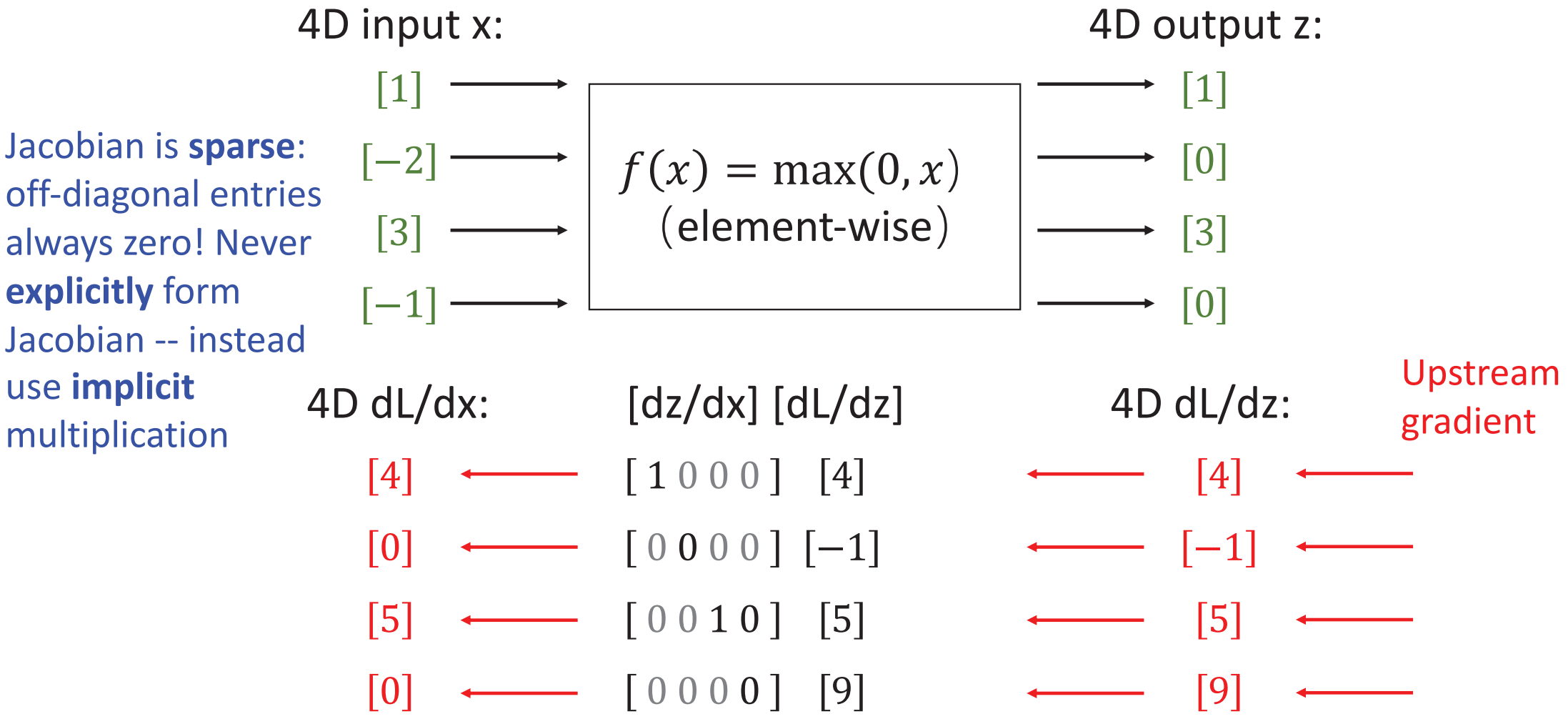
Backpropagation with vectors



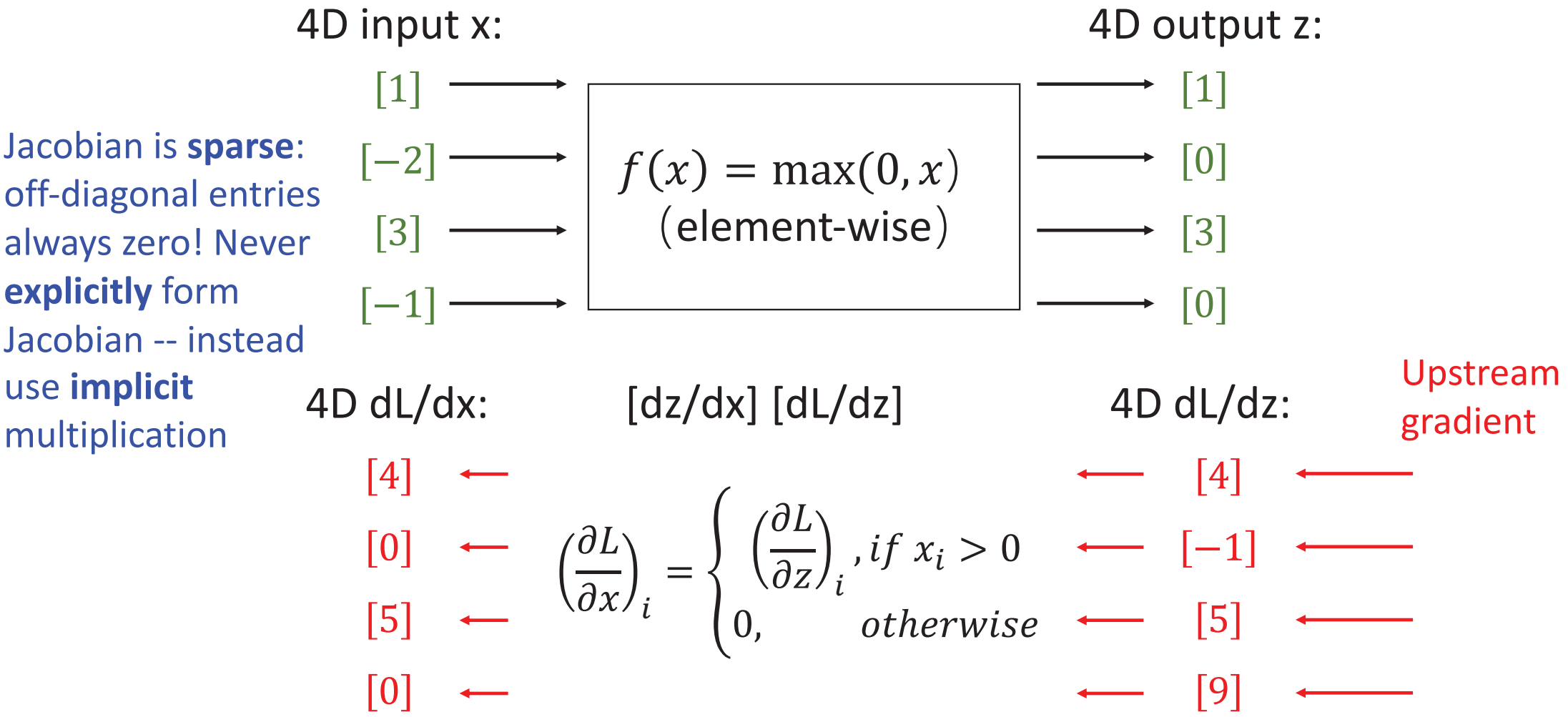
Backpropagation with vectors



Backpropagation with vectors



Backpropagation with vectors



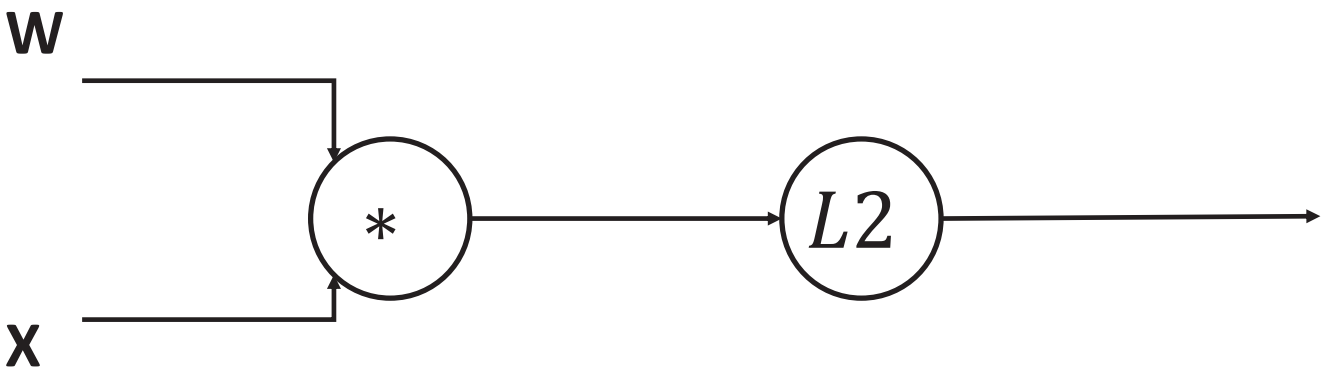
A vectorized example: $$f(x, W) = ||W \cdot x||^2 = \sum_{i=1}^{n} (W \cdot x)_i^2$$

A vectorized example:

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

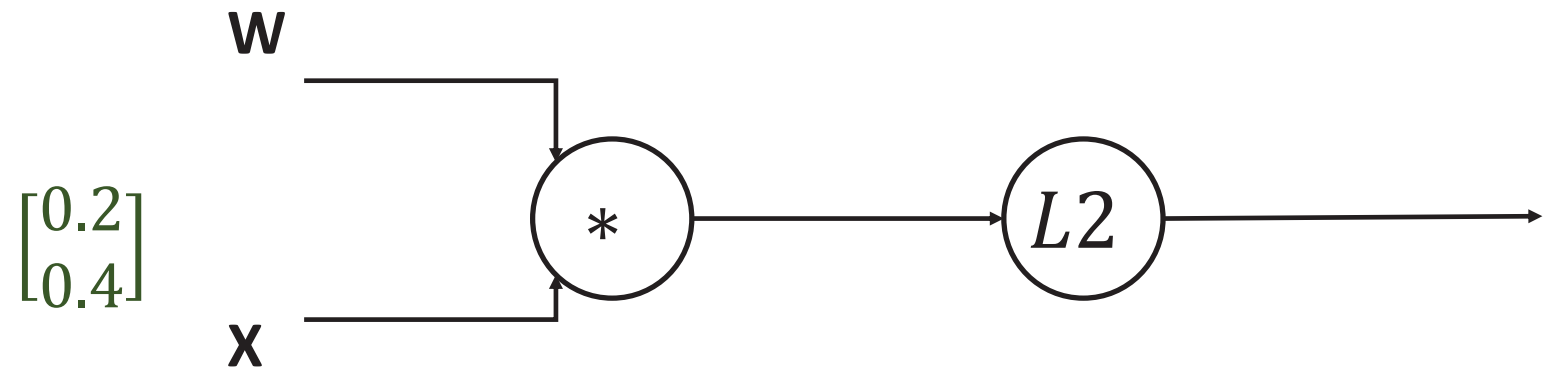
$x \in \mathbb{R}^n$, $W \in \mathbb{R}^{n \times n}$

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



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

$$\begin{bmatrix} 0.1 & 0.5 \\ -0.3 & 0.8 \end{bmatrix}$$



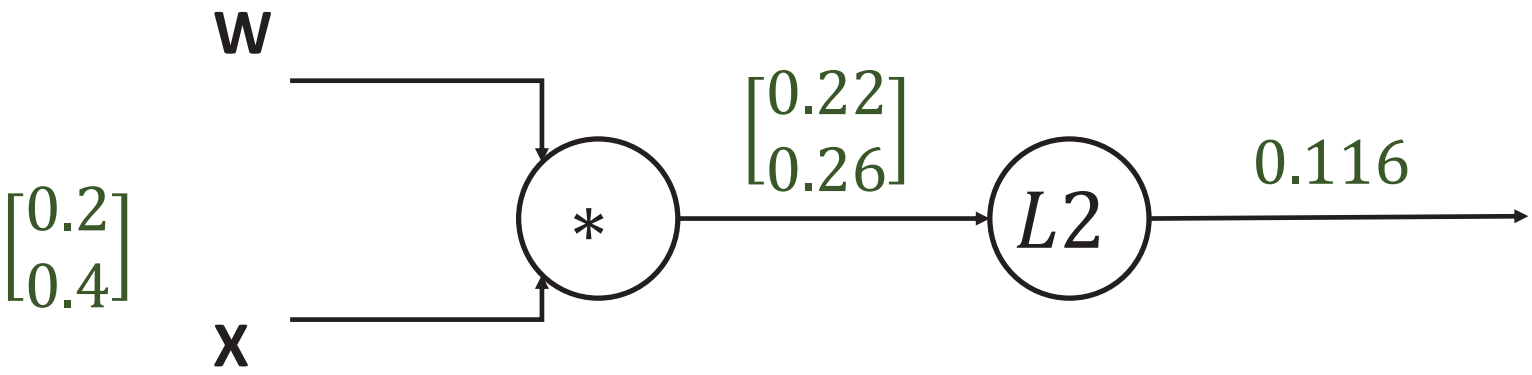
$$\begin{bmatrix} 0.2 \\ 0.4 \end{bmatrix}$$

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

$$f(q) = ||q||^2 = \sum_{i=1}^{n} (q)_i^2$$

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

$$\begin{bmatrix} 0.1 & 0.5 \\ -0.3 & 0.8 \end{bmatrix}$$



$$\begin{bmatrix} 0.2 \\ 0.4 \end{bmatrix}$$

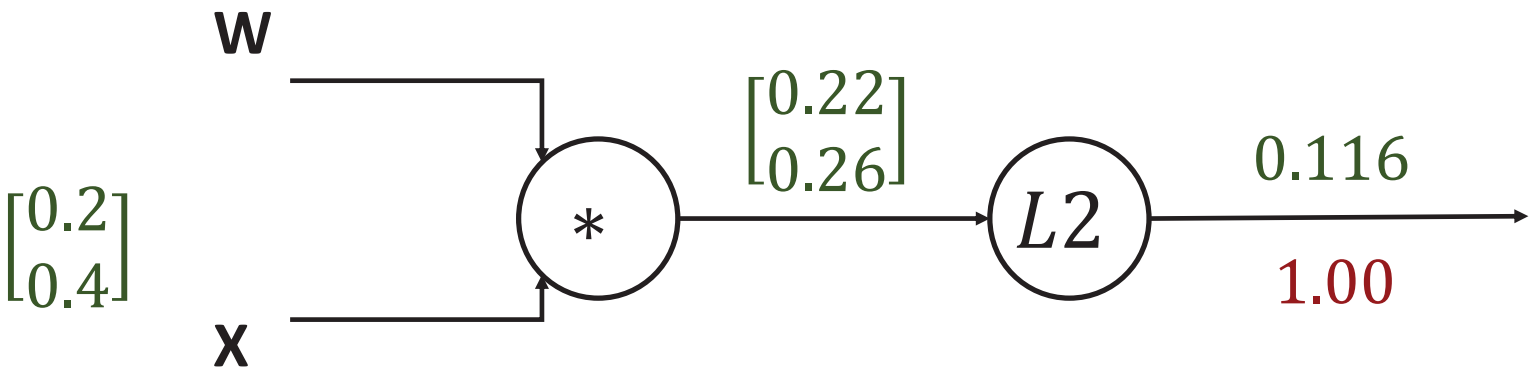
$$\begin{bmatrix} 0.22 \\ 0.26 \end{bmatrix}$$

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

$$f(q) = ||q||^2 = \sum_{i=1}^{n} (q)_i^2$$

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

$$\begin{bmatrix} 0.1 & 0.5 \\ -0.3 & 0.8 \end{bmatrix}$$

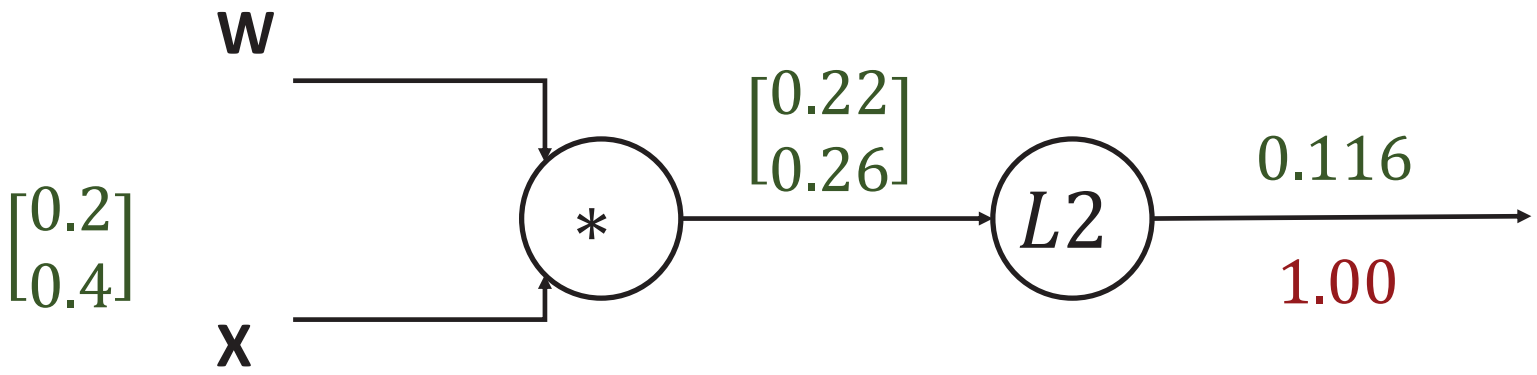


$$q = W \cdot x = \begin{pmatrix} W_{1,1}x_1 + & \cdots & +W_{1,n}x_n \\ \vdots & & \vdots \\ W_{n,1}x_1 + & \cdots & +W_{n,n}x_n \end{pmatrix}$$

$$f(q) = ||q||^2 = \sum_{i=1}^{n} (q)_i^2$$

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

$$\begin{bmatrix} 0.1 & 0.5 \\ -0.3 & 0.8 \end{bmatrix}$$



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

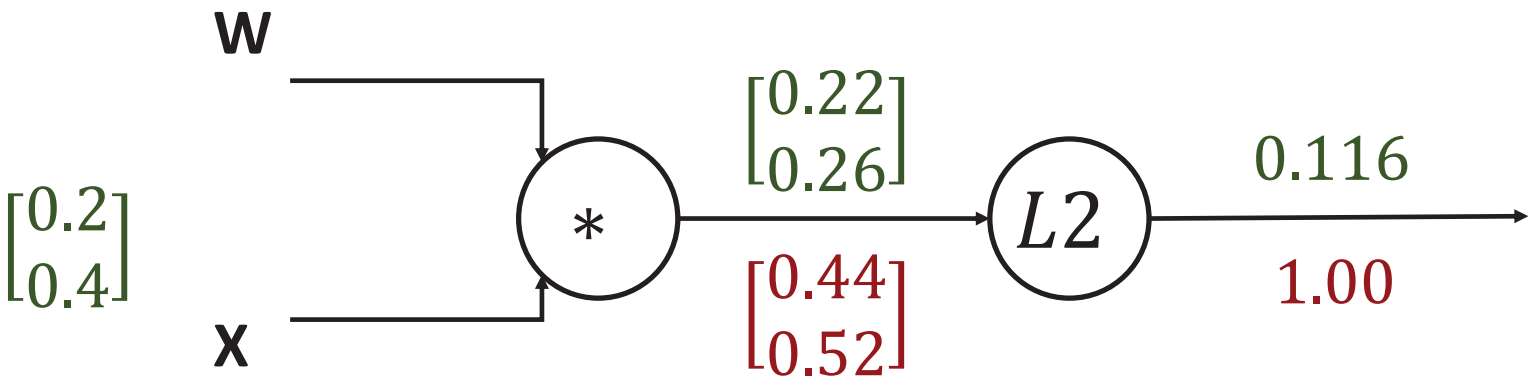
$$f(q) = ||q||^2 = \sum_{i=1}^{n} (q)_i^2$$

$$\frac{\partial f}{\partial q_i} = 2q_i$$

$$\nabla_q f = 2q$$

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

$$\begin{bmatrix} 0.1 & 0.5 \\ -0.3 & 0.8 \end{bmatrix}$$



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

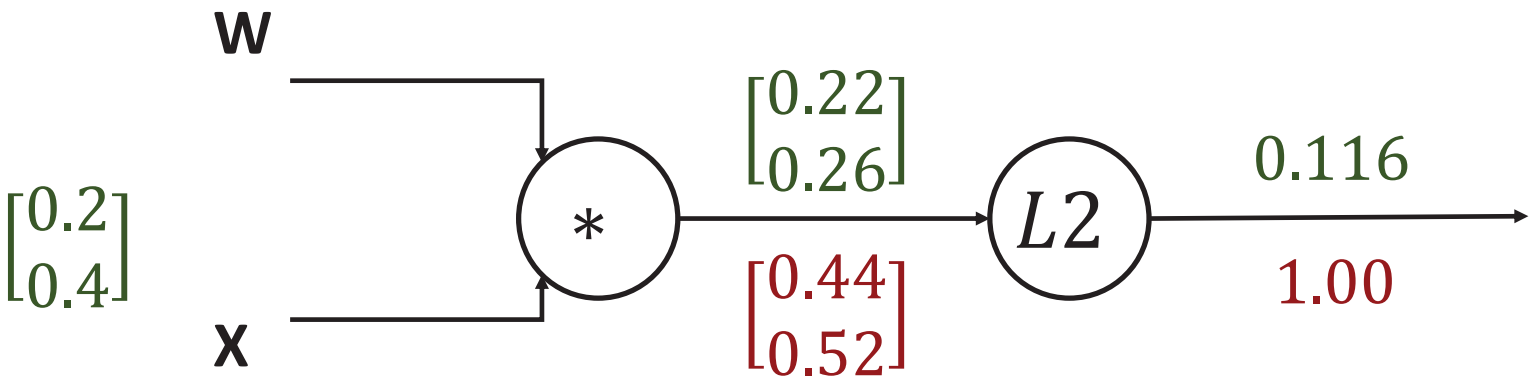
$$f(q) = ||q||^2 = \sum_{i=1}^{n} (q)_i^2$$

$$\frac{\partial f}{\partial q_i} = 2q_i$$

$$\nabla_q f = 2q$$

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

$$\begin{bmatrix} 0.1 & 0.5 \\ -0.3 & 0.8 \end{bmatrix}$$



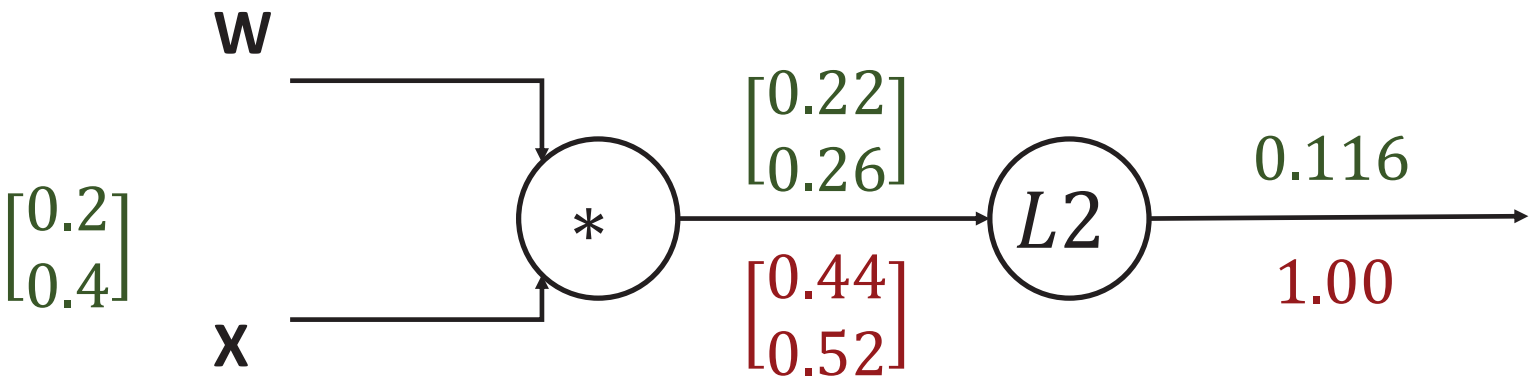
$$\frac{\partial q_k}{\partial W_{i,j}} = I_{k=i} x_j$$

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

$$f(q) = ||q||^2 = \sum_{i=1}^{n} (q)_i^2$$

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

$$\begin{bmatrix} 0.1 & 0.5 \\ -0.3 & 0.8 \end{bmatrix}$$



$$\begin{bmatrix} 0.2 \\ 0.4 \end{bmatrix}$$

$$\begin{bmatrix} 0.22 \\ 0.26 \end{bmatrix}$$

$$\begin{bmatrix} 0.44 \\ 0.52 \end{bmatrix}$$

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

$$f(q) = ||q||^2 = \sum_{i=1}^{n} (q)_i^2$$

$$\frac{\partial q_k}{\partial W_{i,j}} = I_{k=i} x_j$$

$$\frac{\partial f}{\partial W_{i,j}} = \sum_k \frac{\partial f}{\partial q_k} \frac{\partial q_k}{\partial W_{i,j}}$$

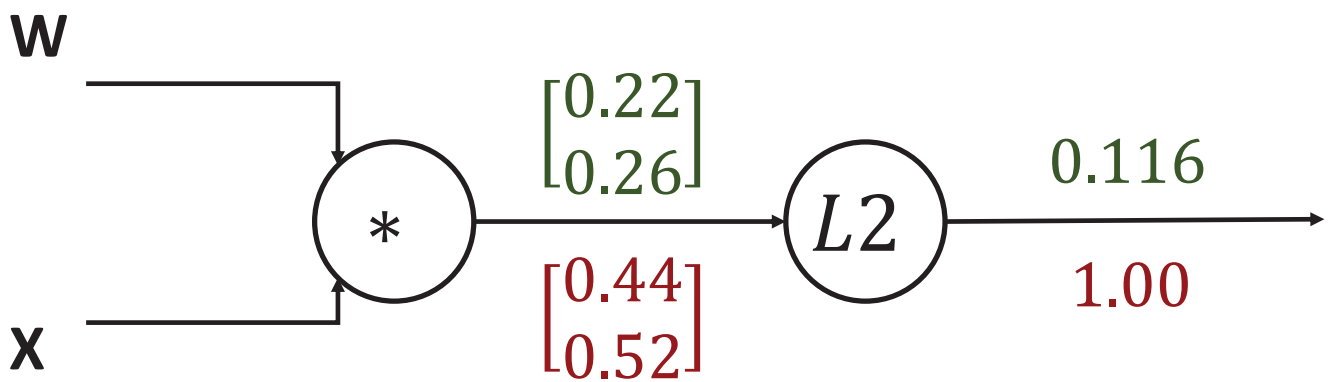
$$= \sum_k (2q_k)\left(I_{k=i} x_j\right) = 2q_i x_j$$

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

$$\begin{bmatrix} 0.1 & 0.5 \\ -0.3 & 0.8 \end{bmatrix}$$

$$\begin{bmatrix} 0.088 & 0.176 \\ 0.104 & 0.208 \end{bmatrix}$$

$$\begin{bmatrix} 0.2 \\ 0.4 \end{bmatrix}$$



$$\begin{bmatrix} 0.22 \\ 0.26 \end{bmatrix}$$

$$\begin{bmatrix} 0.44 \\ 0.52 \end{bmatrix}$$

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

$$f(q) = ||q||^2 = \sum_{i=1}^{n} (q)_i^2$$

$$\frac{\partial q_k}{\partial W_{i,j}} = I_{k=i} x_j$$

$$\frac{\partial f}{\partial W_{i,j}} = \sum_k \frac{\partial f}{\partial q_k} \frac{\partial q_k}{\partial W_{i,j}}$$

$$= \sum_k (2q_k)\left(I_{k=i} x_j\right) = 2q_i x_j$$

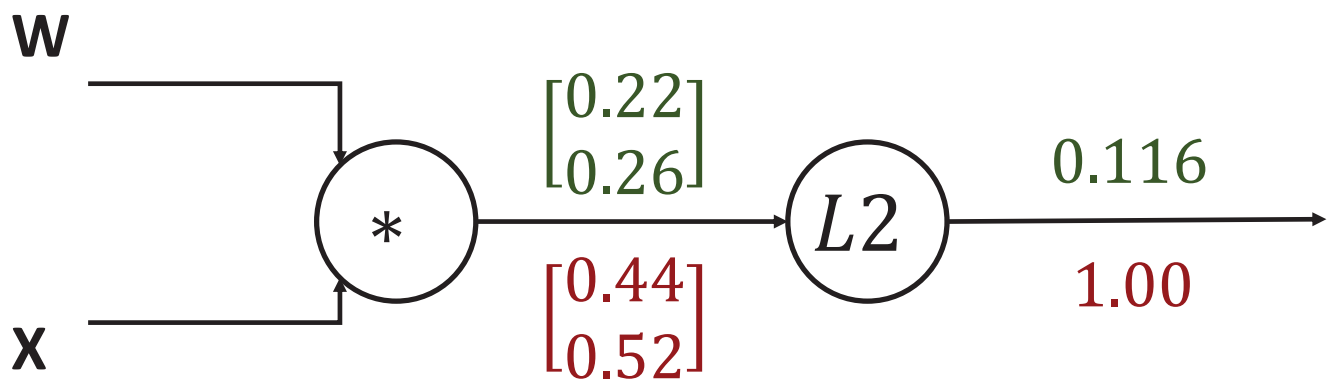
A vectorized example: $f(x, W) = ||W \cdot x||^2 = \sum_{i=1}^{n} (W \cdot x)_i^2$

$$\nabla_W f = 2q \cdot x^T$$

$$\begin{bmatrix} 0.1 & 0.5 \\ -0.3 & 0.8 \end{bmatrix}$$

$$\begin{bmatrix} 0.088 & 0.176 \\ 0.104 & 0.208 \end{bmatrix}$$

$$\begin{bmatrix} 0.2 \\ 0.4 \end{bmatrix}$$



$$\begin{bmatrix} 0.22 \\ 0.26 \end{bmatrix}$$

$$\begin{bmatrix} 0.44 \\ 0.52 \end{bmatrix}$$

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

$$f(q) = ||q||^2 = \sum_{i=1}^{n} (q)_i^2$$

$$\frac{\partial q_k}{\partial W_{i,j}} = I_{k=i} x_j$$

$$\frac{\partial f}{\partial W_{i,j}} = \sum_k \frac{\partial f}{\partial q_k} \frac{\partial q_k}{\partial W_{i,j}}$$

$$= \sum_k (2q_k)\left(I_{k=i} x_j\right) = 2q_i x_j$$

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

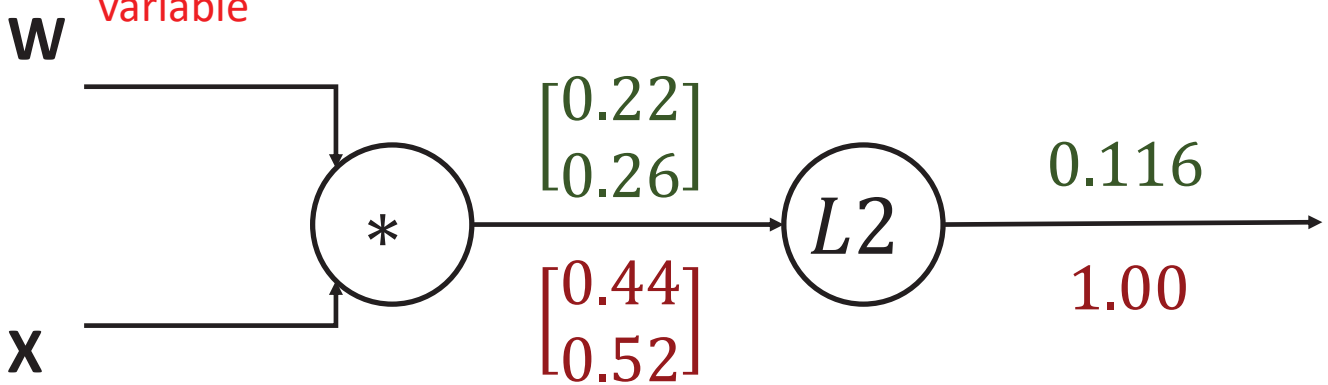
$$\begin{bmatrix} 0.1 & 0.5 \\ -0.3 & 0.8 \end{bmatrix}$$

$$\begin{bmatrix} 0.088 & 0.176 \\ 0.104 & 0.208 \end{bmatrix}$$

$$\begin{bmatrix} 0.2 \\ 0.4 \end{bmatrix}$$

Always check: The gradient with respect to a variable should have the same shape as the variable

$$\nabla_W f = 2q \cdot x^T$$



$$\begin{bmatrix} 0.22 \\ 0.26 \end{bmatrix}$$

$$\begin{bmatrix} 0.44 \\ 0.52 \end{bmatrix}$$

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

$$f(q) = ||q||^2 = \sum_{i=1}^{n} (q)_i^2$$

$$\frac{\partial q_k}{\partial W_{i,j}} = I_{k=i} x_j$$

$$\frac{\partial f}{\partial W_{i,j}} = \sum_k \frac{\partial f}{\partial q_k} \frac{\partial q_k}{\partial W_{i,j}}$$

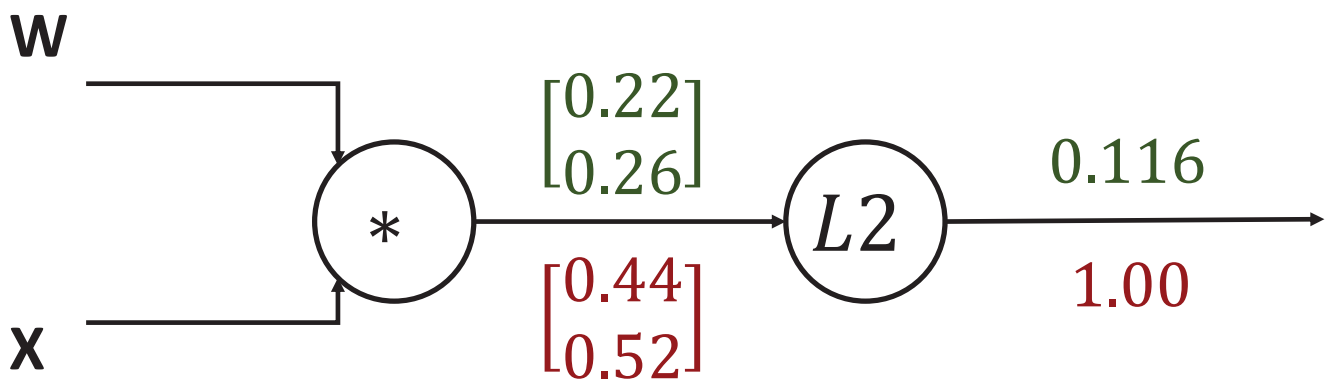
$$= \sum_k (2q_k)\left(I_{k=i} x_j\right) = 2q_i x_j$$

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

$$\begin{bmatrix} 0.1 & 0.5 \\ -0.3 & 0.8 \end{bmatrix}$$

$$\begin{bmatrix} 0.088 & 0.176 \\ 0.104 & 0.208 \end{bmatrix}$$

$$\begin{bmatrix} 0.2 \\ 0.4 \end{bmatrix}$$



$$\begin{bmatrix} 0.22 \\ 0.26 \end{bmatrix}$$

$$\begin{bmatrix} 0.44 \\ 0.52 \end{bmatrix}$$

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

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

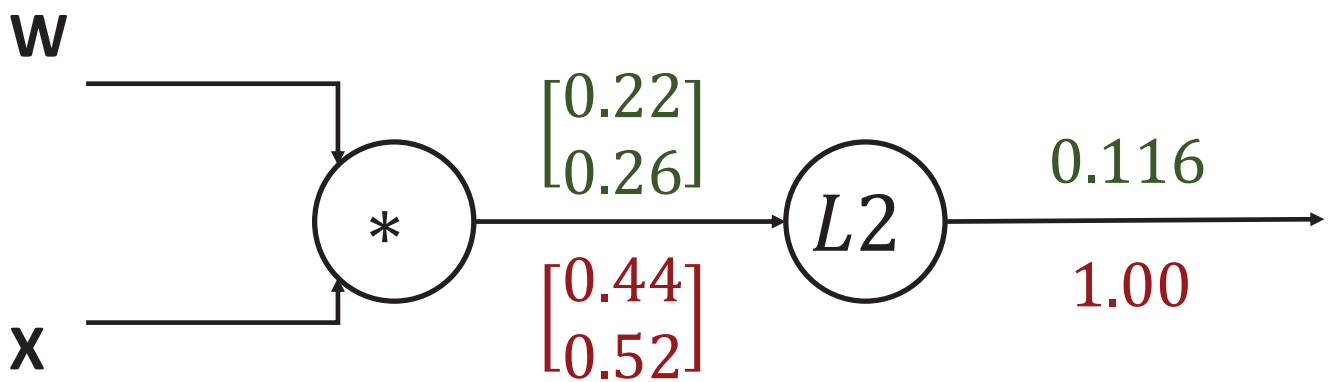
$$f(q) = ||q||^2 = \sum_{i=1}^{n} (q)_i^2$$

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

$$\begin{bmatrix} 0.1 & 0.5 \\ -0.3 & 0.8 \end{bmatrix}$$

$$\begin{bmatrix} 0.088 & 0.176 \\ 0.104 & 0.208 \end{bmatrix}$$

$$\begin{bmatrix} 0.2 \\ 0.4 \end{bmatrix}$$



$$\begin{bmatrix} 0.22 \\ 0.26 \end{bmatrix}$$

$$\begin{bmatrix} 0.44 \\ 0.52 \end{bmatrix}$$

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

$$f(q) = ||q||^2 = \sum_{i=1}^{n} (q)_i^2$$

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

$$\frac{\partial f}{\partial x_i} = \sum_k \frac{\partial f}{\partial q_k} \frac{\partial q_k}{\partial x_i}$$

$$= \sum_k (2q_k)(W_{k,i})$$

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

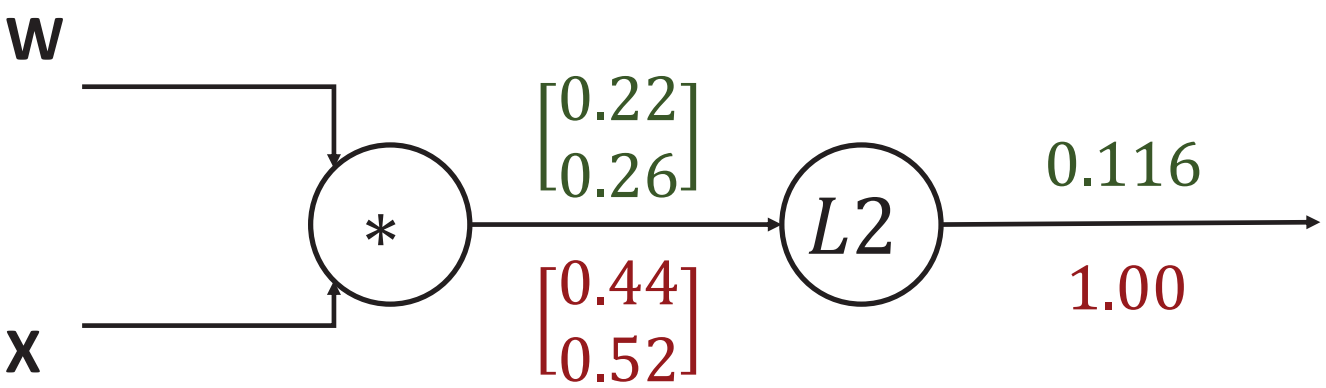
$$\nabla_x f = 2W^T \cdot q$$

$$\begin{bmatrix} 0.1 & 0.5 \\ -0.3 & 0.8 \end{bmatrix}$$

$$\begin{bmatrix} 0.088 & 0.176 \\ 0.104 & 0.208 \end{bmatrix}$$

$$\begin{bmatrix} 0.2 \\ 0.4 \end{bmatrix}$$

$$\begin{bmatrix} -0.112 \\ 0.636 \end{bmatrix}$$



$$\begin{bmatrix} 0.22 \\ 0.26 \end{bmatrix}$$

$$\begin{bmatrix} 0.44 \\ 0.52 \end{bmatrix}$$

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

$$f(q) = ||q||^2 = \sum_{i=1}^{n} (q)_i^2$$

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

$$\frac{\partial f}{\partial x_i} = \sum_k \frac{\partial f}{\partial q_k} \frac{\partial q_k}{\partial x_i}$$

$$= \sum_k (2q_k)(W_{k,i})$$

Difficult Choices

- Many choices must be made regarding architecture of the NN, including
  - Number of hidden layers
  - Number of neurons in each layer
  - Which units are connected
- While "univ. approx." result suggests sufficiency of 1 hidden layer, deeper networks often achieve same accuracy with fewer parameters
- Eldan and Schamir (2016) formally demonstrate that depth — even if increased by one layer (from 2 to 3) — can be exponentially more valuable than increasing width in standard feed-forward neural networks
- However, deeper networks are difficult to train...e.g., vanishing gradients.
- An intuitive design follows the geometric pyramid rule Masters (1993).
- Do not optimize over all possible architectures.

Stochastic Gradient Descent

- Gradient descent
  - $\theta_{t+1} = \theta_t - \eta_t \nabla MSE(\theta_t)$: $\eta_t > 0$ is the **learning rate**
  - η too big is unstable, η too small is too slow
- "Backpropagation"
  - Crucial algo for calculating gradient
  - Essentially uses a version of chain rule for more efficient computation
- "Mini-batches"
  - Because dataset are typically huge, full descent routine too slow
  - Rather than calculating $\nabla MSE = \nabla \frac{1}{N} \sum_i (y_i - g(x_i; \theta))^2$ ...
  - Calculate $\nabla MSE_m = \nabla \frac{1}{N_m} \sum_{i \in m} (y_i - g(x_i; \theta))^2$ for random subset m
  - Cycle through subsets (without replacement). A single pass through of the entire training sample is called an "epoch."

Optimization "Tricks"

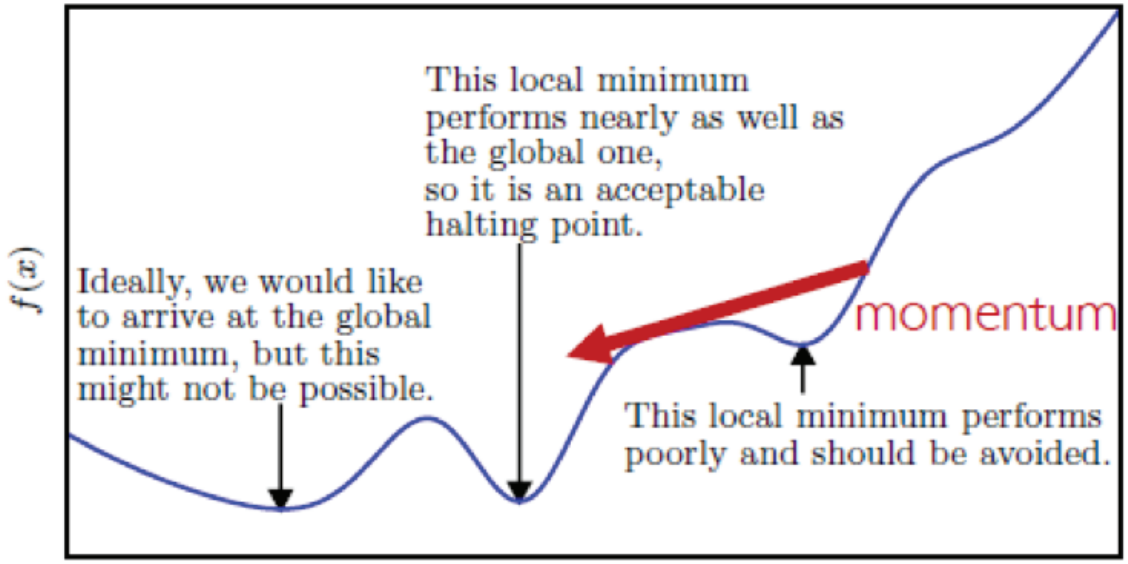
"Learning-rate decay"

- Algos to shrink learning rate toward zero as gradient approaches zero because the true gradient, when it is small, will be dominated by SDG noise.

$$\text{e.g., } \eta = \eta_0 \exp(-\kappa t); \quad \eta = \eta_0 (1 + \kappa t)^{-1}.$$

- η can be adaptive for individual parameters
- η is one of the most difficult to set hyperparameters!

- Modern algorithms also incorporate momentum, Nesterov acceleration, ... e.g., "Adam" Cf. Kingma and Ba (2015, ICLR) Cited: 34,611 as of last week. An Overview link Cf. Ruder (2016)



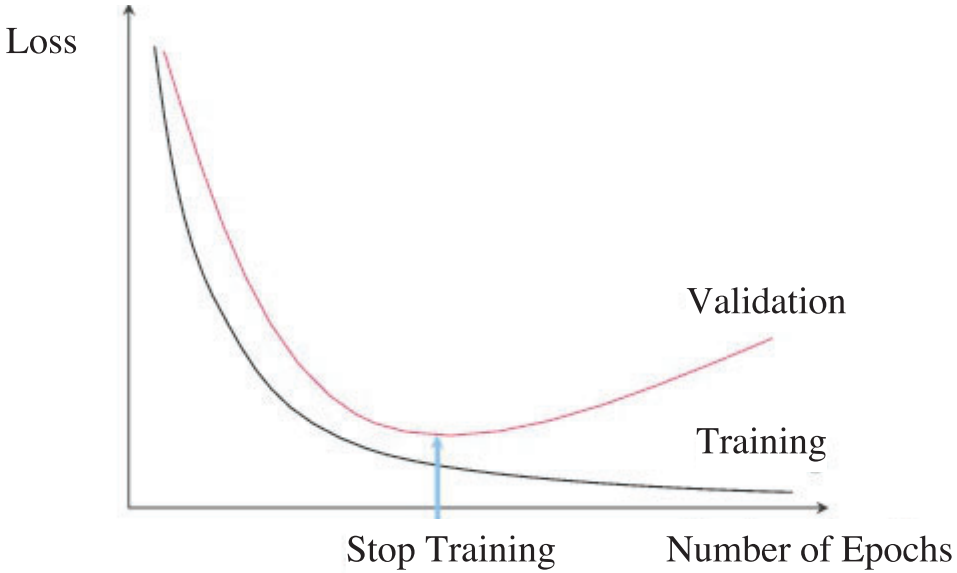
Improving NN algos is continuously and intensively researched!

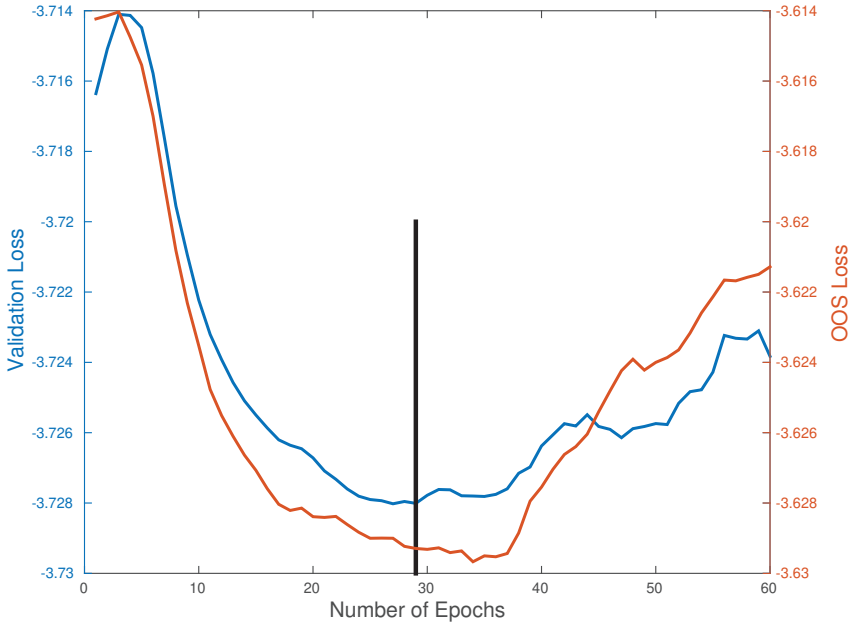
- "Dropout" Cf. Srivastava, Hinton, Krizhevsky, Sutskever, and Salakhutdinov (2014, JMLR) Cited: 16,374 as of last week.
- "Batch Normalization" Cf. Ioffe and Szegedy (2015, ICML) Cited: 14,654 as of last week.
- These algos are very specific to training deep neural networks, and their theoretical justification is undergoing, hence we omit the details.

Regularization "Tricks"

- L_1 and L_2 penalization
- Ensemble: randomness in SGD, parameter initialization, etc
- "Early stopping"
  - Also an optimization trick: ending the parameter search early by monitoring the loss function in the validation sample.
  - Parameter estimates are shrunken toward the initial guess, and this is how early stopping regularizes against overfit.

Early Stopping





Note: an epoch is a complete pass through a given dataset.