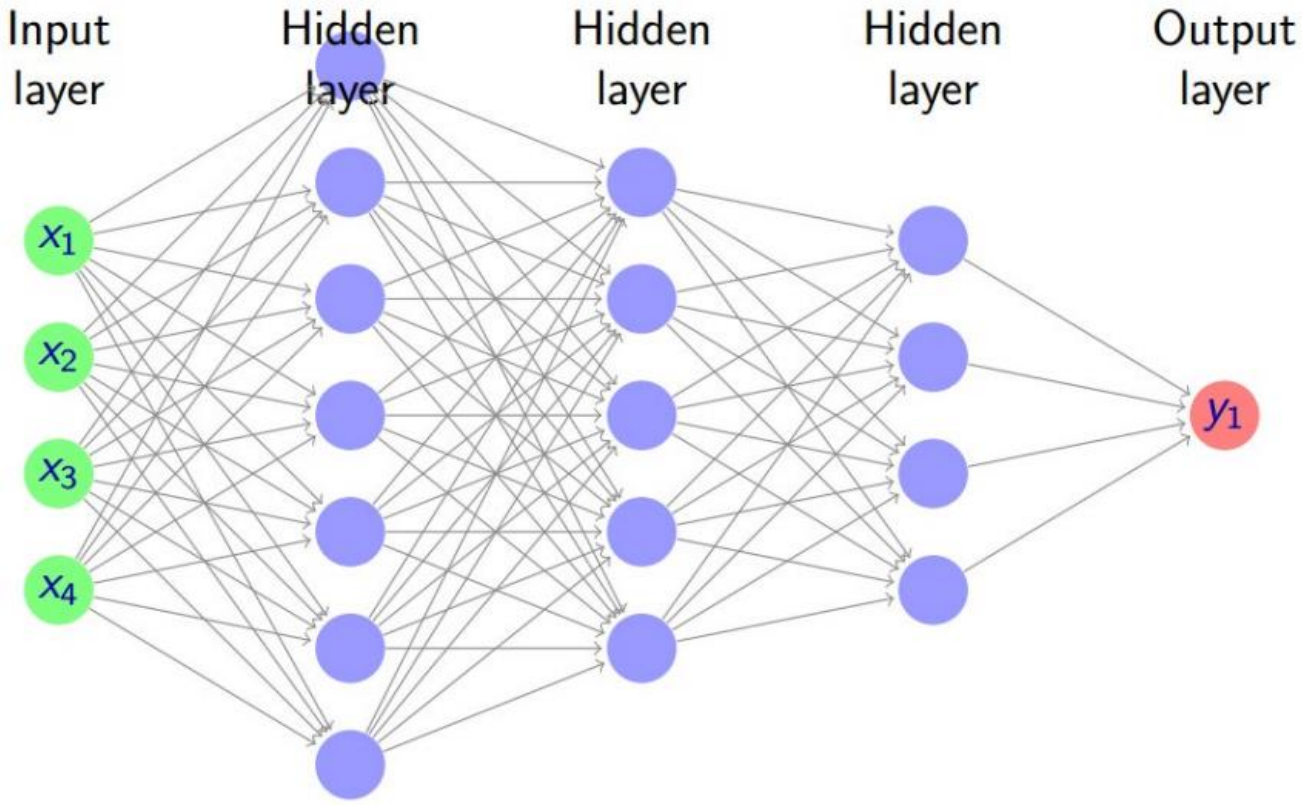
Feedforward Neural Network, Backpropagation

HESAM HOSSEINI

SUMMER 2024

Feedforward Neural Network



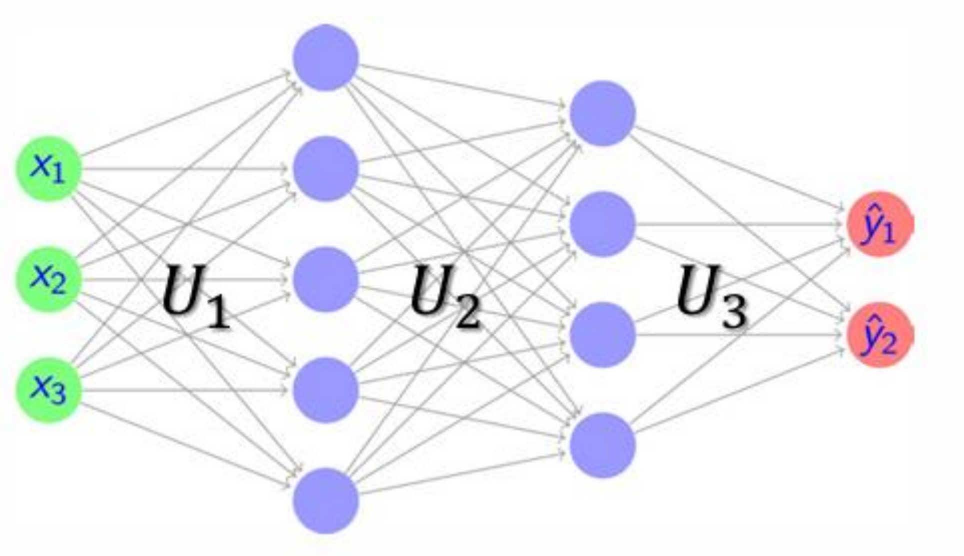
Feedforward Neural Network

$$U_1 = \begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} & u_{15} \\ u_{21} & u_{22} & u_{23} & u_{24} & u_{25} \\ u_{31} & u_{32} & u_{33} & u_{34} & u_{35} \end{bmatrix}_{3\times5}$$

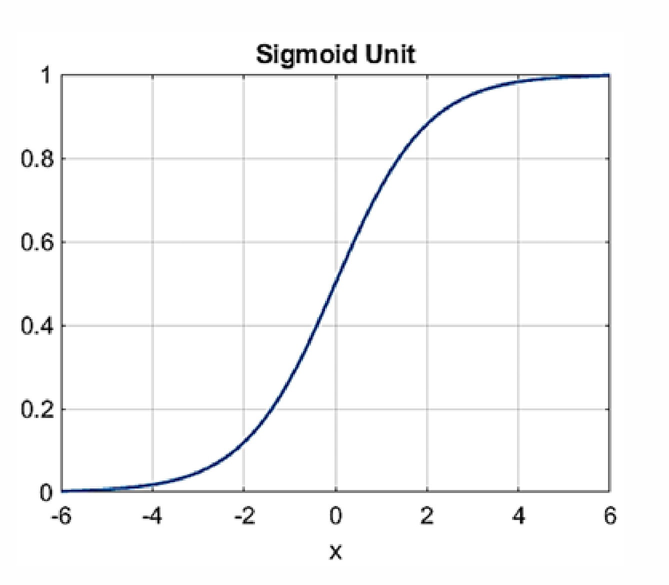
$$U_2 = \begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ u_{21} & u_{22} & u_{23} & u_{24} \\ u_{31} & u_{32} & u_{33} & u_{34} \\ u_{41} & u_{42} & u_{43} & u_{44} \\ u_{51} & u_{52} & u_{53} & u_{54} \end{bmatrix}_{5\times4}$$

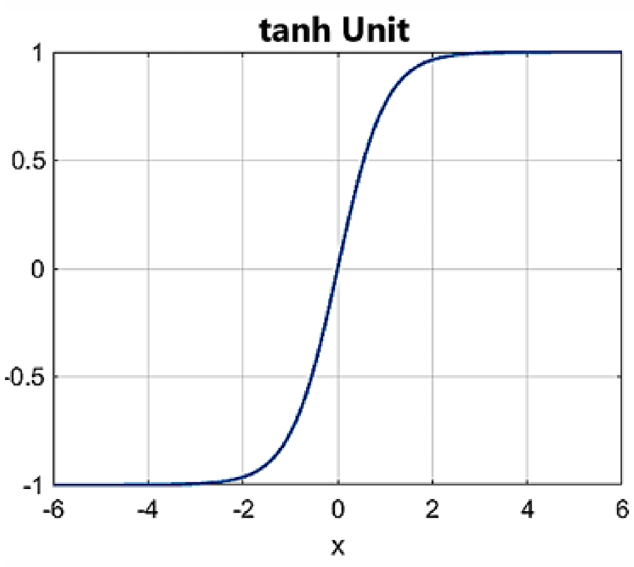
$$U_3 = \begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \\ u_{31} & u_{32} \end{bmatrix}_{4\times2}$$

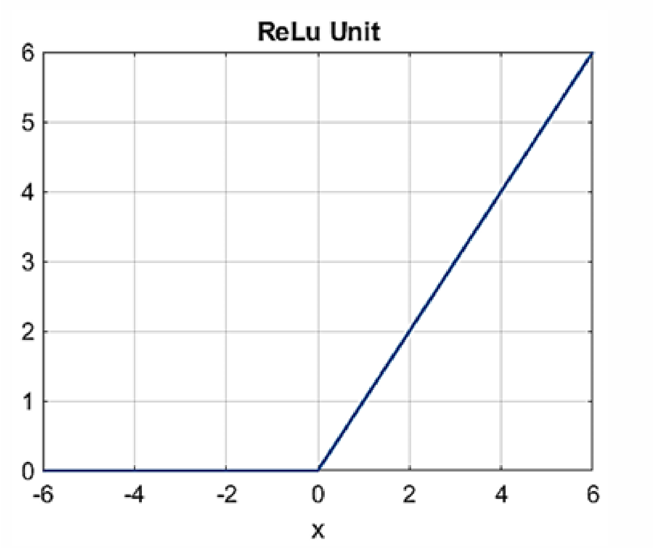
$$X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$



Activation functions



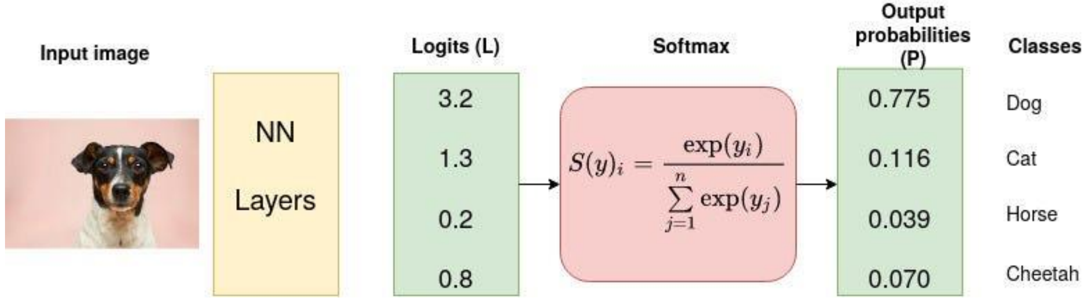




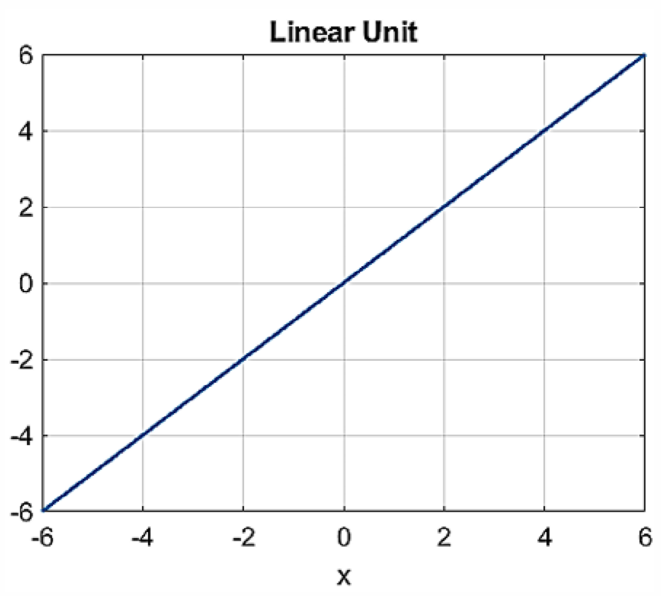
Final Layer

CLASSIFICATION

$$f_i(\boldsymbol{x}) = \frac{e^{x_i}}{\sum_{k=1}^{m} e^{x_k}},$$



REGRESSION



Regression loss function

N: # of $samples$

$\boldsymbol{y}_i \subset \mathcal{R}^D$: $Desired\ output\ (target)$

$\widehat{\boldsymbol{y}}_i \subset \mathcal{R}^D$: $Actual\ output$

$\boldsymbol{e}_i \subset \mathcal{R}^D$: $(\boldsymbol{y}_i - \widehat{\boldsymbol{y}}_i)\ error$

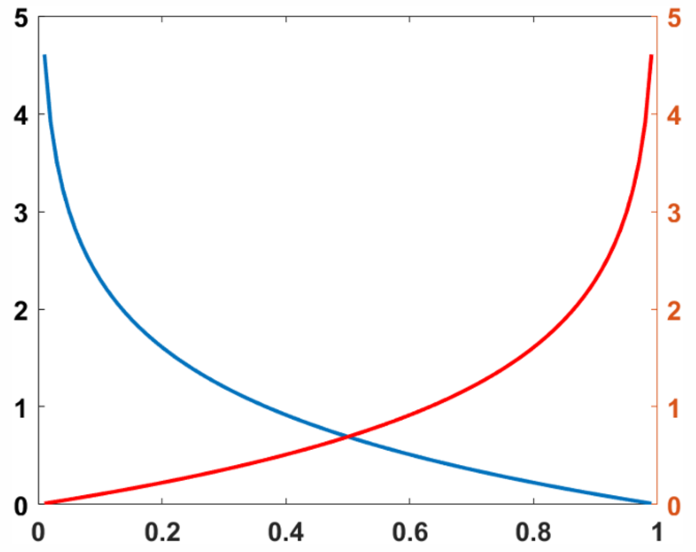
$$MSE = \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{e}_i\|_2^2$$

Binary Classification Loss function

- Binary Cross Entropy (BCE or log-loss)

$$BCE = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i log\hat{y}_i + (1-y_i)log(1-\hat{y}_i)\right), \quad y_i \in \{0,1\}$$

  - BCE plot for yi = 0 and yi = 1



Multi-Class Classification

N: # $of\ samples$

$\boldsymbol{y}_i \in \{0,1\}^M$: $Desired\ output\ (target)$

$\boldsymbol{y}_i$ is one-hot vector, example (for M=5):

$$\boldsymbol{y}_i = (0 \quad 0 \quad 1 \quad 0 \quad 0)^T$$

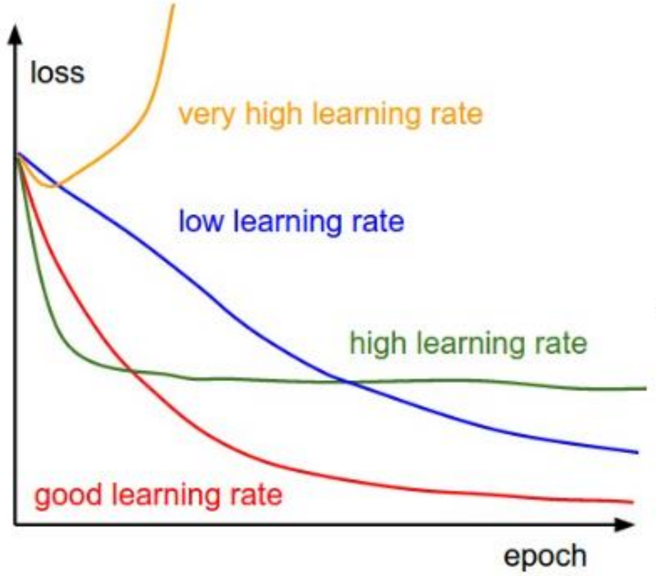
$\hat{\boldsymbol{y}}_i \in [0\ 1]^M$: $Actual\ outputs$

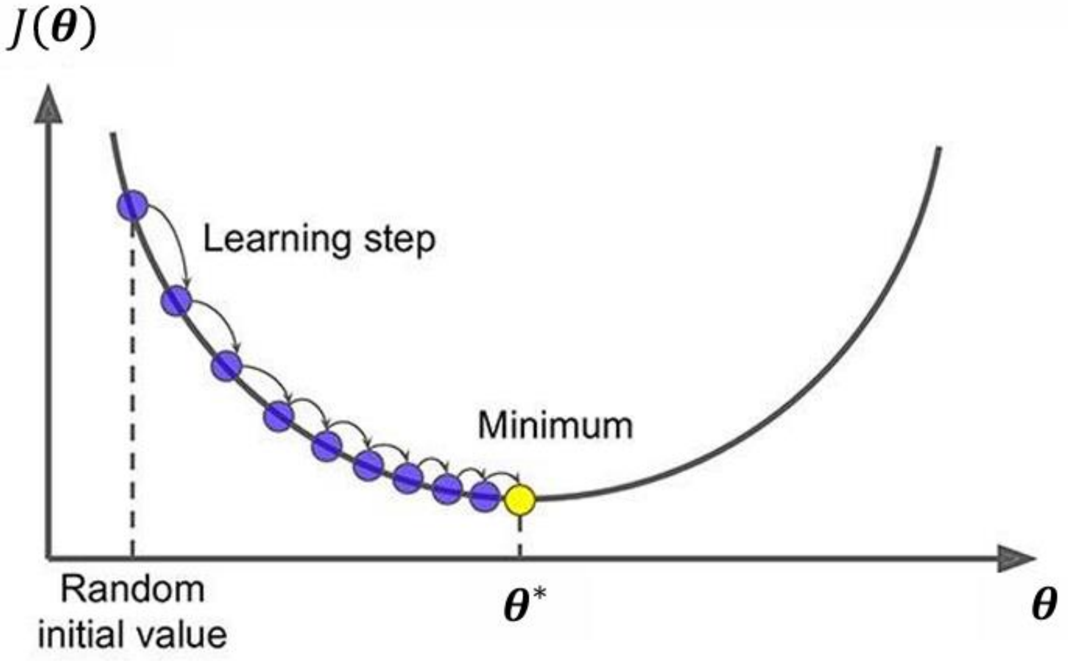
Cross Entropy (CE):

$$CE = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{M} y_{i,k} log\hat{y}_{i,k}\,, \quad y_{i,k} \in \{0,1\}$$

Review : Gradient Descent

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \epsilon^{(t)} \left.\frac{\partial Loss(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}$$





How do we compute gradients?

- **Analytic or “Manual” Differentiation:** fast , exact , error-pron
- **Numerical Differentiation:** slow, approximate, easy to write

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}.$$

- **Problems for Analytic/symbolic gradient**:
  - for complex functions, expressions can be exponentially large
  - Need to re-derive from scratch for any minor changes. Not modular!
  - Difficult to deal with piece-wise functions (require many symbolic cases)

Automatic Differentiation (AutoDiff)

Intuition: Interleave symbolic differentiation and simplification

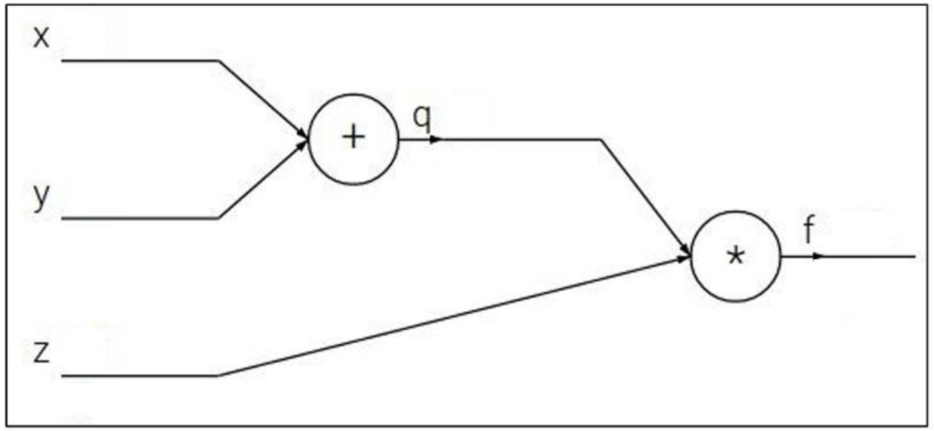
Key Idea: Apply symbolic differentiation at the elementary operation level, evaluate and keep intermediate results

Success of deep learning owes A LOT to success of AutoDiff algorithms (also to advances in parallel architectures, and large datasets, ...)

Backpropagation: a simple example

Backpropagation: a simple example

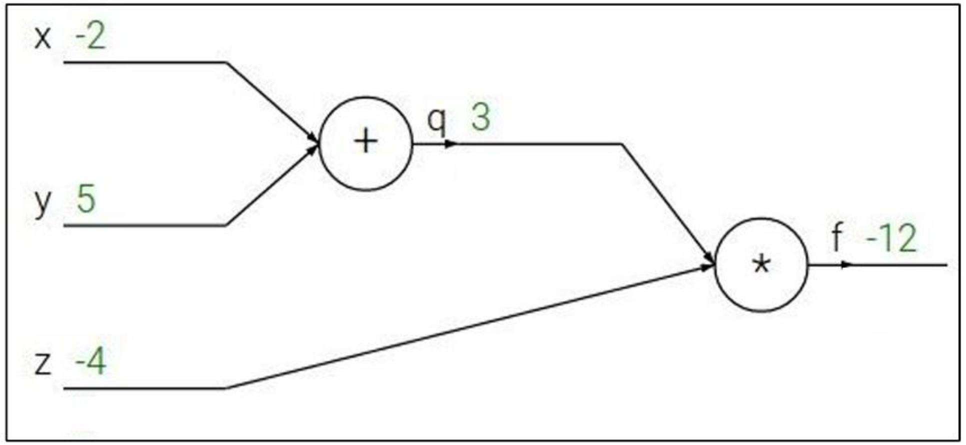
$$f(x, y, z) = (x + y)z$$



Backpropagation: a simple example

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

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



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

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

Backpropagation: a simple example

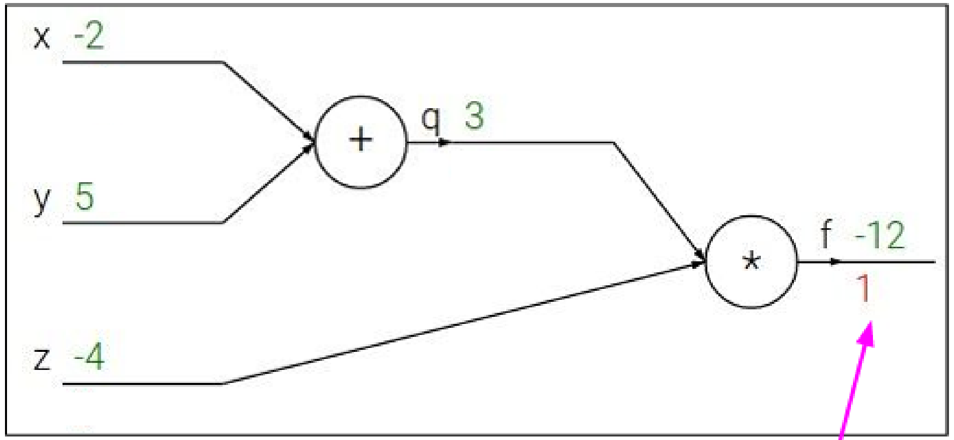
$$f(x, y, z) = (x + y)z$$

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



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

Backpropagation: a simple example

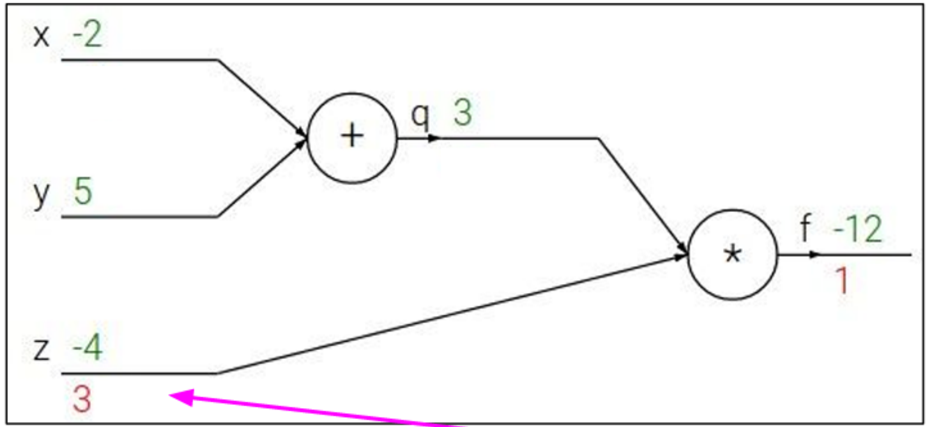
$$f(x, y, z) = (x + y)z$$

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

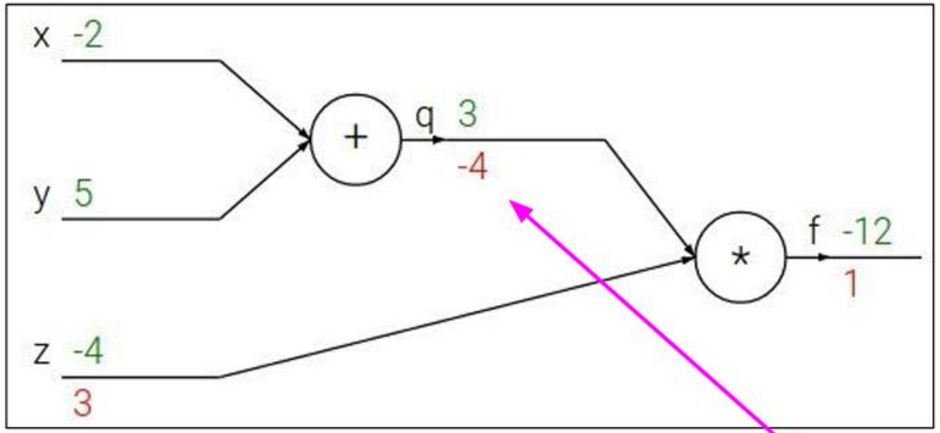
$$f(x, y, z) = (x + y)z$$

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

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

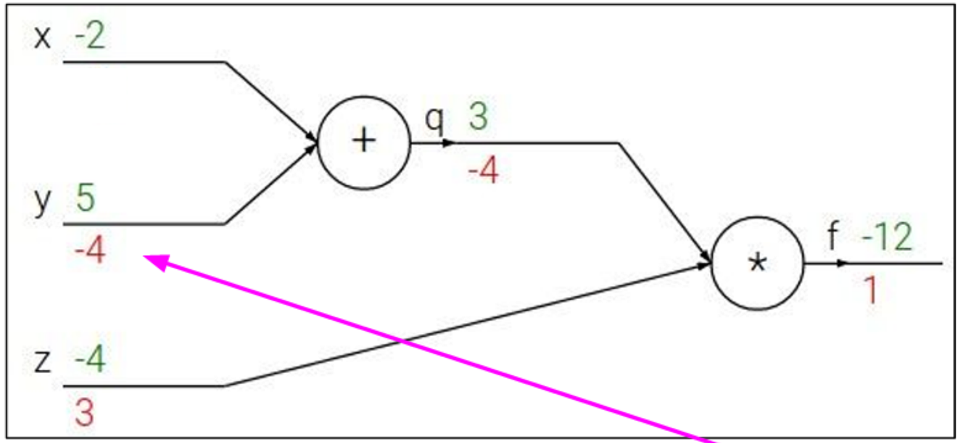
$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

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



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

Backpropagation: a simple example



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

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

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

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

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

Chain rule:

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

Upstream gradient

Local gradient

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

Backpropagation: a simple example

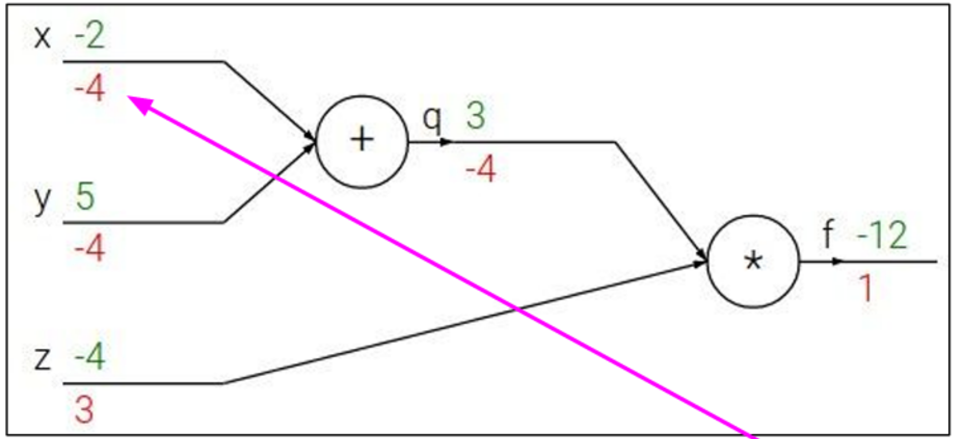
$$f(x, y, z) = (x + y)z$$

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



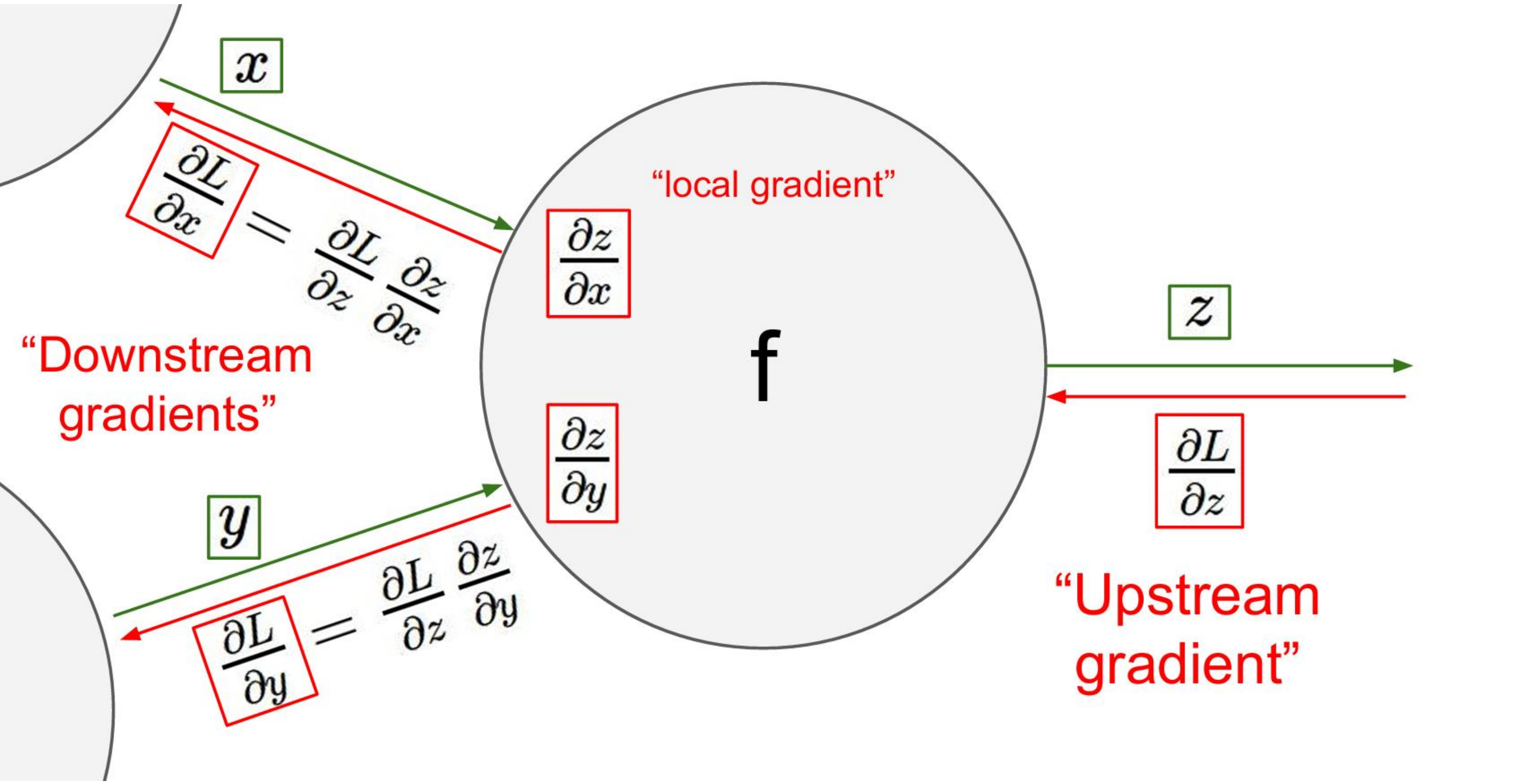
$$\frac{\partial f}{\partial x}$$

Chain rule:

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

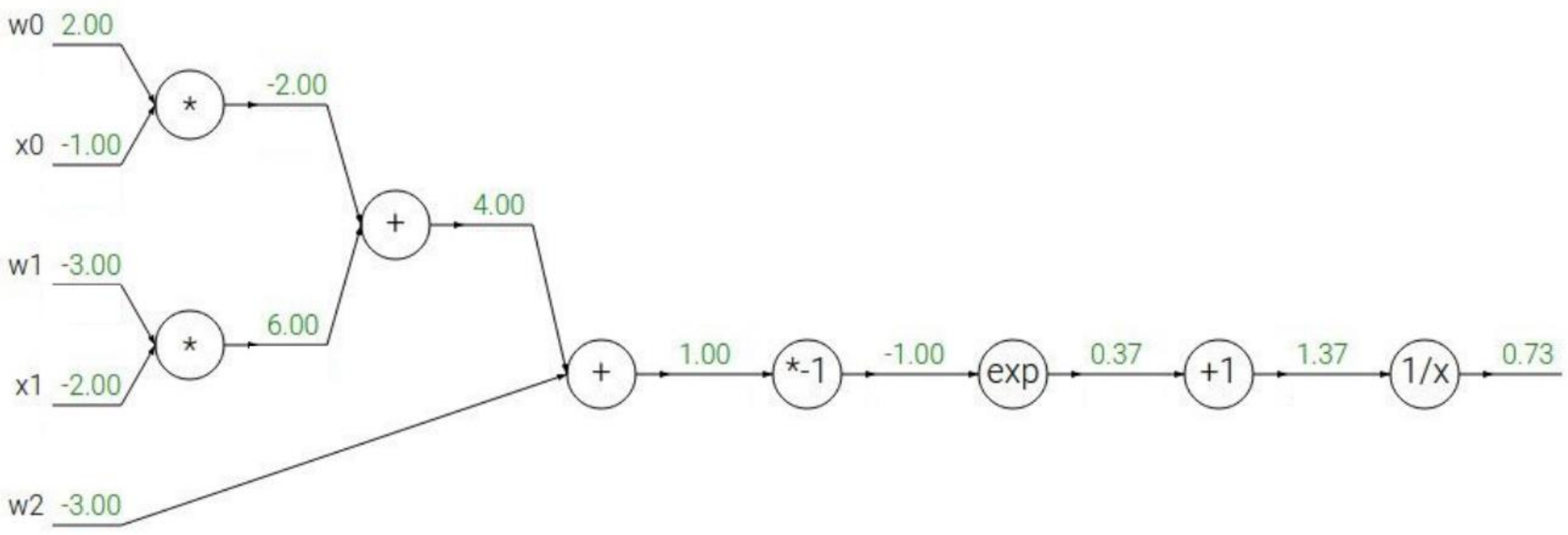
Upstream gradient

Local gradient


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"

Another example:

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



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

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

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

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

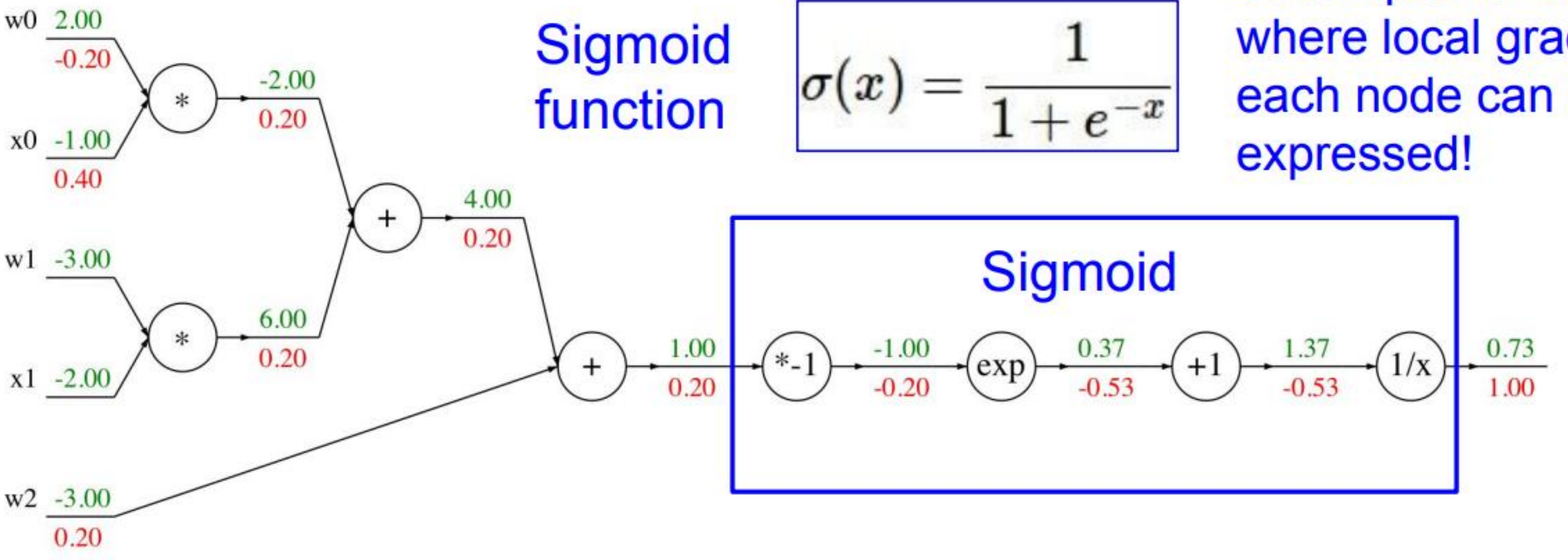
Another example:

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

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:

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

Patterns in gradient flow