TBSI 清华-伯克利深圳学院
Tsinghua-Berkeley Shenzhen Institute

Learning From Data

Lecture ~~6~~ 7: Backpropagation and Neural Networks

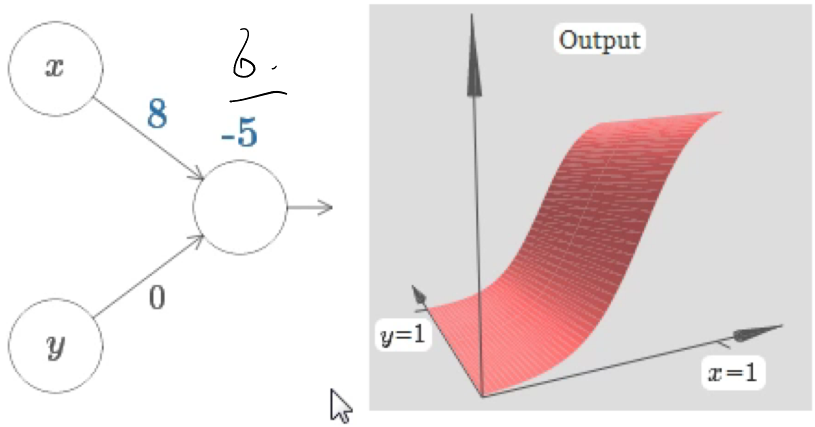
Yang Li

Slides by Lichen Wang

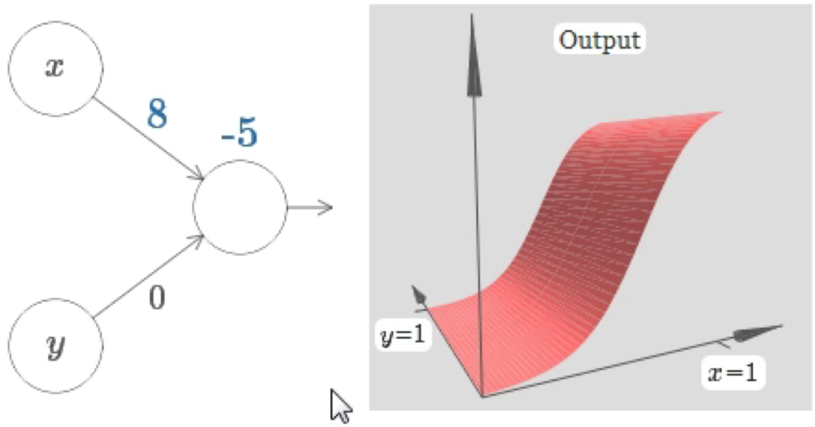
10/29/2021

Learning From Data

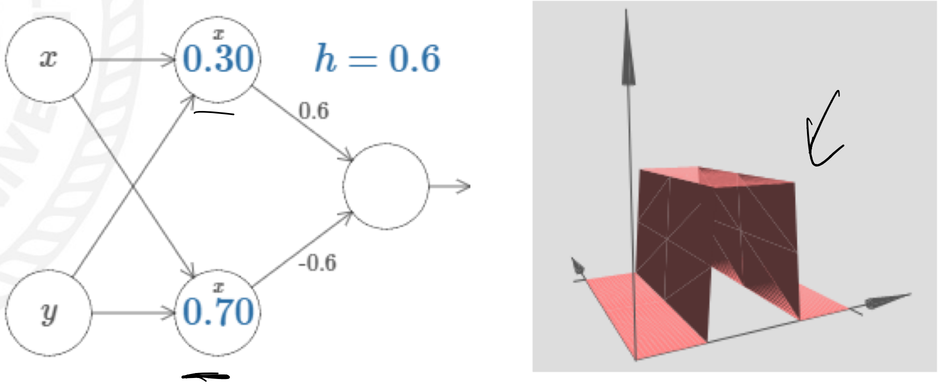
Power of single neuron



Power of single neuron

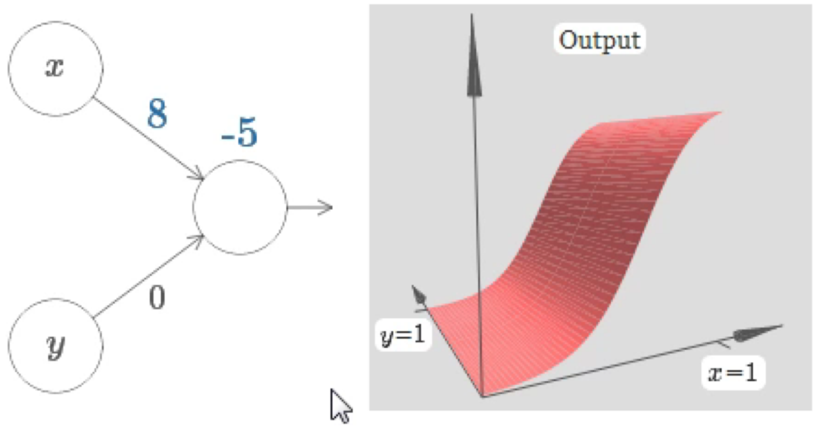


Two hidden units

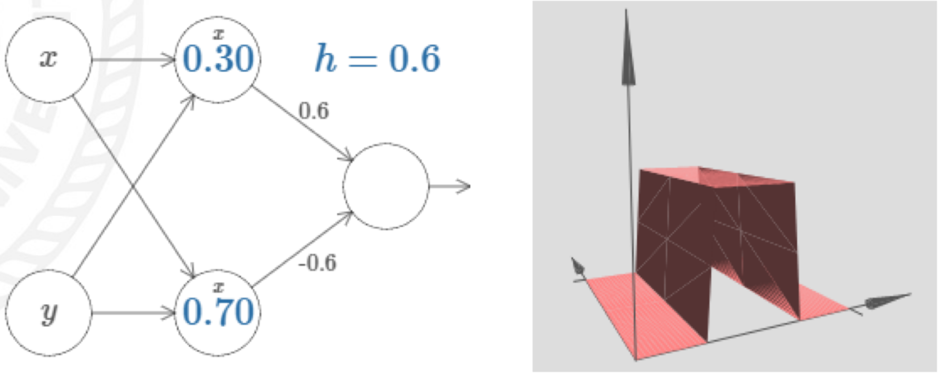


Learning From Data

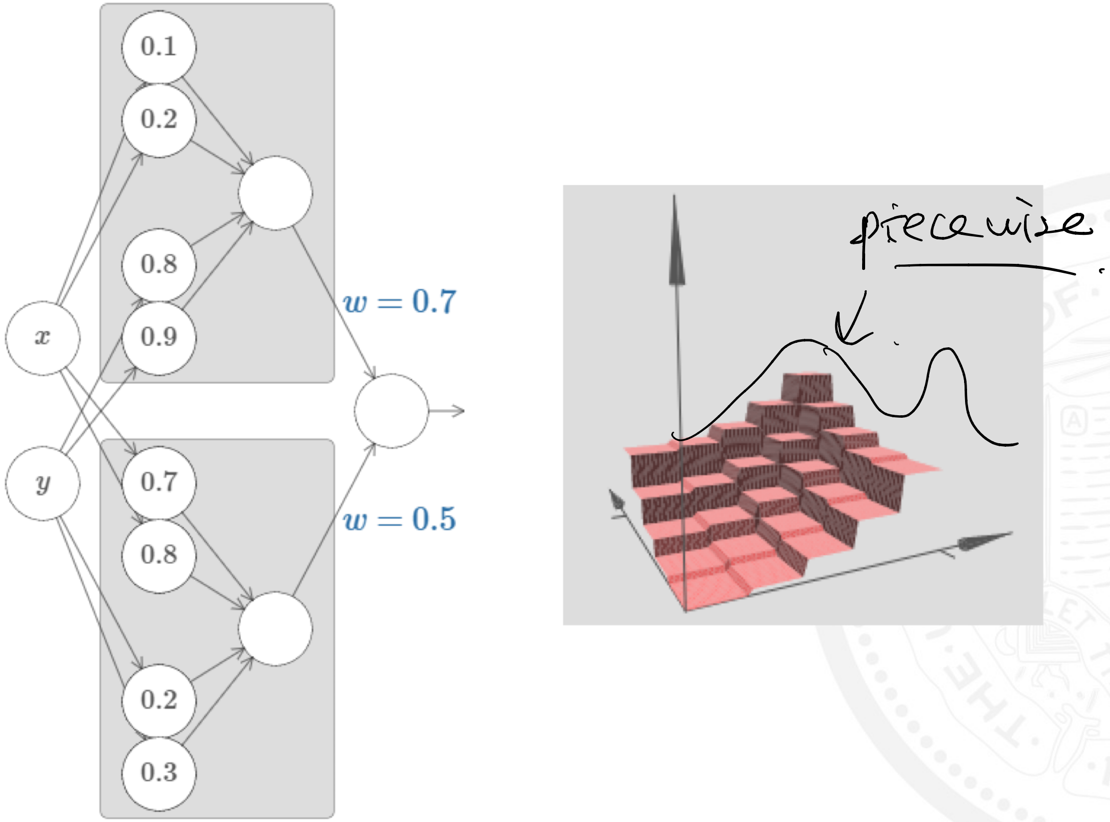
Power of single neural



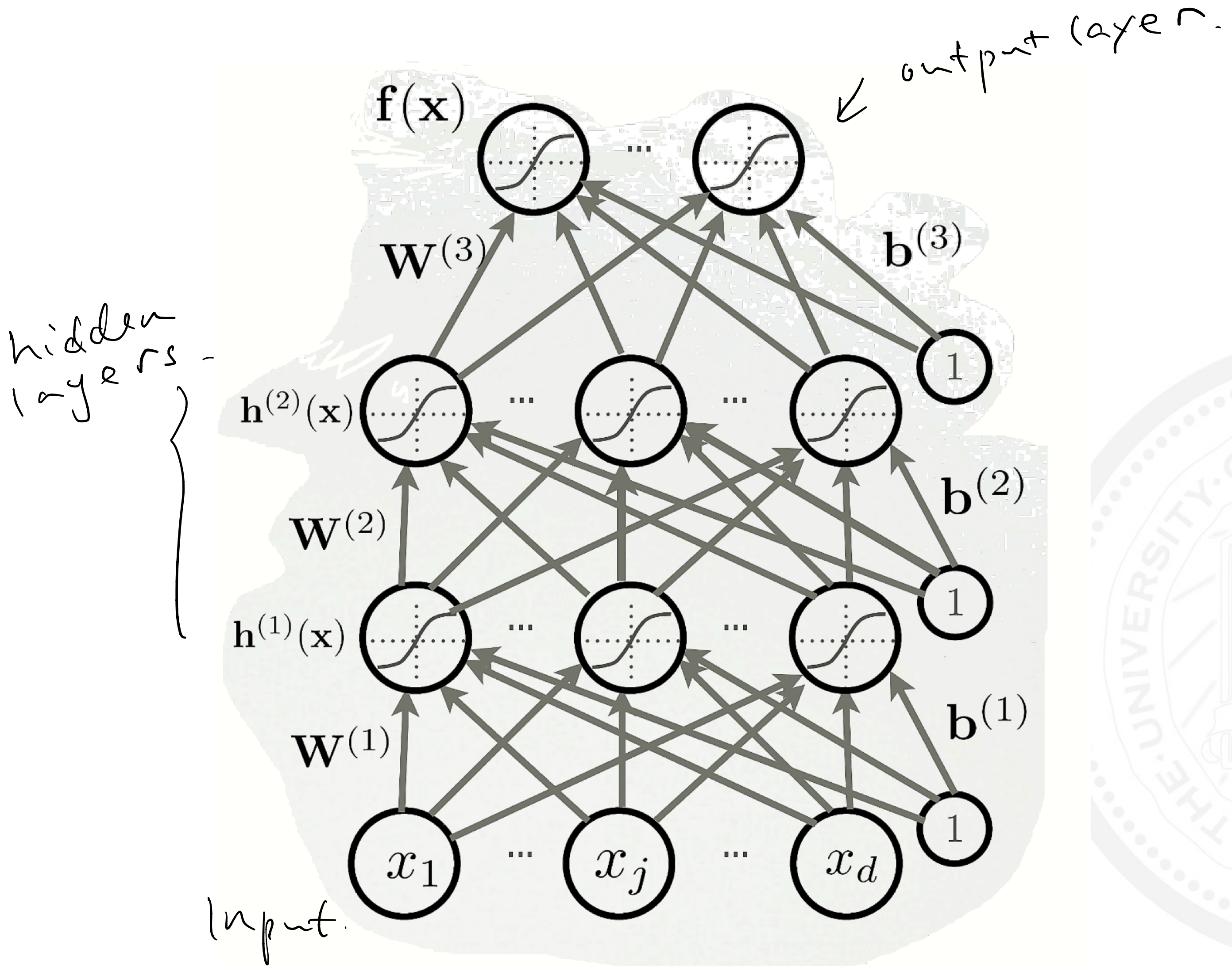
Two hidden units



Many hidden units



Could have L hidden layers



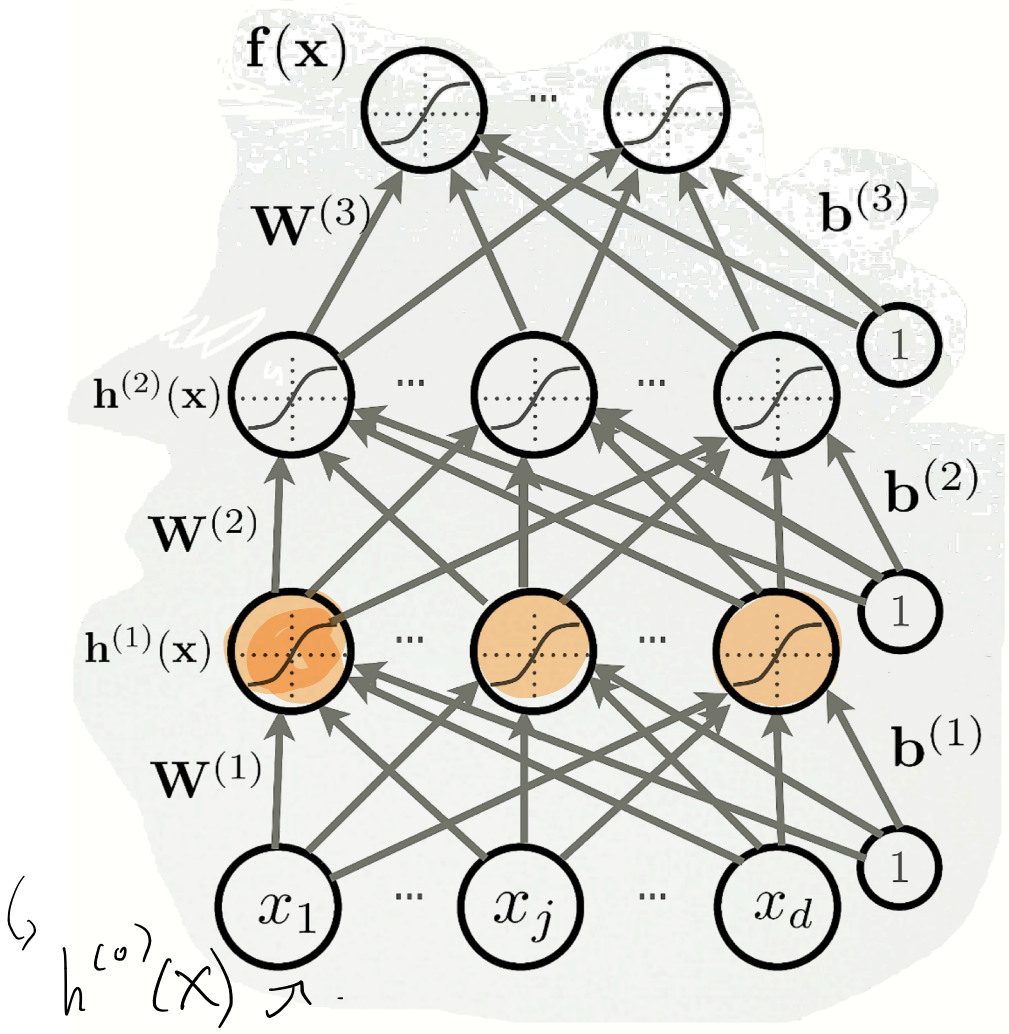
Multilayer Neural Network

Could have L hidden layers

k-th layer.

- layer input activation for $k > 0$, $\mathbf{h}^{(0)}(\mathbf{x}) = \mathbf{x}$

$$\mathbf{a}^{(k)}(\mathbf{x}) = \mathbf{b}^{(k)} + \mathbf{W}^{(k)}\mathbf{h}^{(k-1)}(\mathbf{x})$$



Multilayer Neural Network

Could have L hidden layers

- layer input activation for $k>0$, $\mathbf{h}^{(0)}(\mathbf{x})=\mathbf{x}$

$$\mathbf{a}^{(k)}(\mathbf{x})=\mathbf{b}^{(k)}+\mathbf{W}^{(k)}\mathbf{h}^{(k-1)}(\mathbf{x})$$

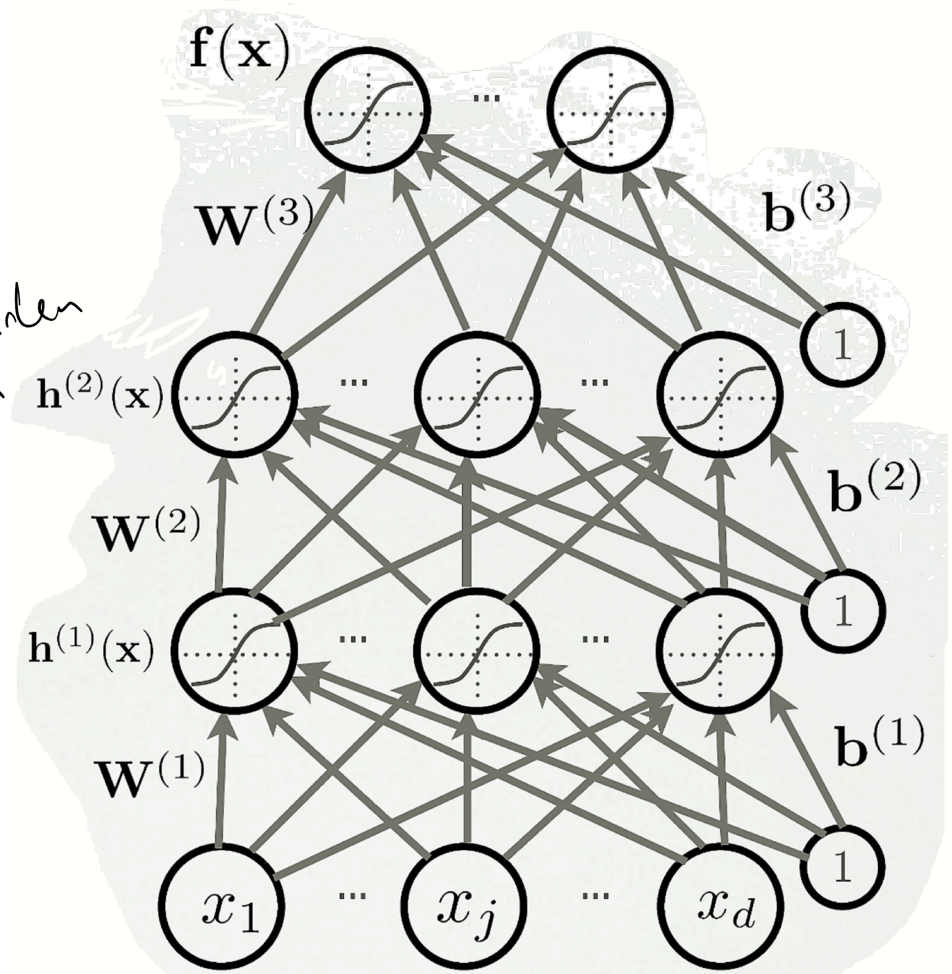
one hidden layer

- hidden layer activation for $1\le k\le L$

$$\mathbf{h}^{(k)}(\mathbf{x})=\mathbf{g}(\mathbf{a}^{(k)}(\mathbf{x}))$$

activation

Repeat L times



Could have L hidden layers

- layer input activation for $k>0$, $\mathbf{h}^{(0)}(\mathbf{x})=\mathbf{x}$

$$\mathbf{a}^{(k)}(\mathbf{x})=\underbrace{\mathbf{b}^{(k)}+\mathbf{W}^{(k)}\mathbf{h}^{(k-1)}(\mathbf{x})}_{\text{linear}}$$

- hidden layer activation for $1\le k\le L$

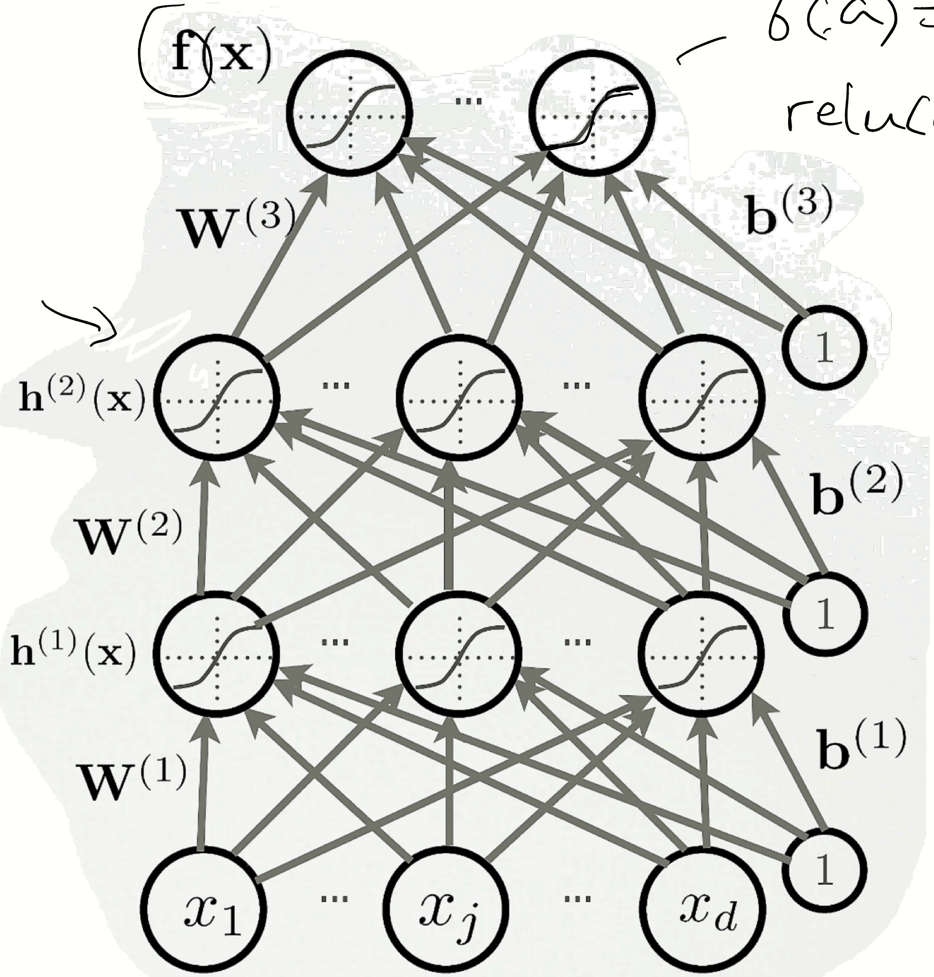
$$\mathbf{h}^{(k)}(\mathbf{x})=\mathbf{g}(\mathbf{a}^{(k)}(\mathbf{x}))$$

threshold function (non linear)

- output layer activation for $k=L+1$

$$\mathbf{h}^{(L+1)}(\mathbf{x})=\mathbf{g}(\mathbf{a}^{(L+1)}(\mathbf{x}))=\mathbf{f}(\mathbf{x})$$

output function (prediction function)



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

$\mathrm{relu}(a)=\max(0,a)$

Empirical risk minimization

Empirical risk

output. ith label.

$$\operatorname*{argmin}_{w} \frac{1}{m} \sum_{i=1}^{m} \underline{L(f(x^{(i)}; W), y^{(i)})} + \underline{\lambda \Omega(W)}$$

- $L(f(x^{(i)}; W), y^{(i)})$ is the loss function
- $\lambda\Omega(\mathbf{W})$ is the regularizer

Empirical risk

$$\underset{w}{\operatorname{argmin}} \frac{1}{m} \sum_{i=1}^{m} L(f(x^{(i)}; W), y^{(i)}) + \lambda\Omega(W)$$

- $L(f(x^{(i)}; W), y^{(i)})$ is the loss function
- $\lambda\Omega(\mathbf{W})$ is the regularizer

$\in \{1, \dots, k\}$

Softmax loss for sample $(x^{(i)}, y^{(i)})$

$f(x)$

$$L_i = -\log\left(\frac{e^{f_{y^{(i)}}}}{\sum_{j=1} e^{f_j}}\right)$$

$\theta^T x$

$y^{(i)}$

$f_j := j$-th element of class score vector $f(x^{(i)}; W)$

Softmax example:

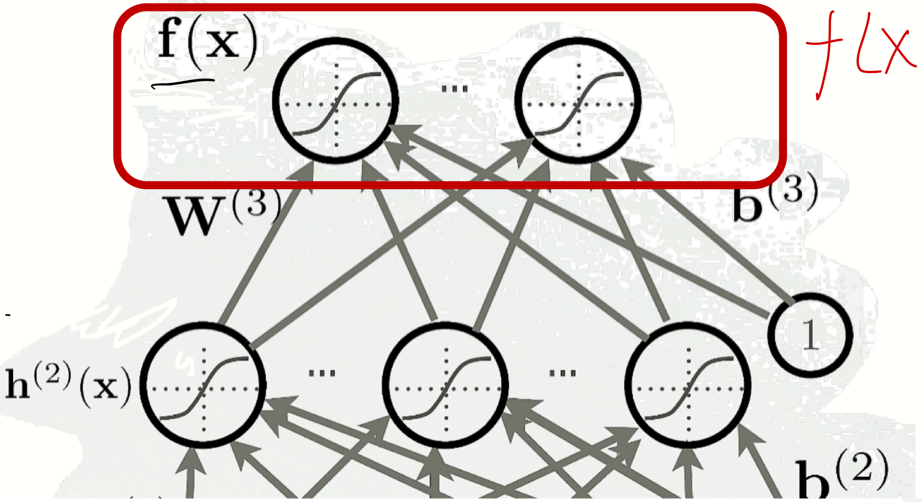
Unnormalized class probability of K classes

$P(y=1|x)$

$P(y=k|x)$

class score vector

$f(x^{(i)}; W)$



Empirical risk minimization

Empirical risk

$$\underset{w}{\operatorname{argmin}} \frac{1}{m} \sum_{i=1}^{m} L(f(x^{(i)}; W), y^{(i)}) + \lambda\Omega(W)$$

$y^{(i)}$

- $L(f(x^{(i)}; W), y^{(i)})$ is the loss function
- $\lambda\Omega(\mathbf{W})$ is the regularizer

Softmax loss for sample $(x^{(i)}, y^{(i)})$

$$L_i = -\log\left(\frac{e^{f_{y^{(i)}}}}{\sum_j e^{f_j}}\right)$$

f_j: $=$ j-th element of class score vector $f(x^{(i)}; W)$

Softmax example:

$x^{(i)}$	Beckham (f_1)	Yanzu (f_2)	Lichen (f_3)	Loss L_i
Beckham 	4.9	1.1	-0.9	$-\log\left(\frac{e^{4.9}}{e^{4.9}+e^{1.1}+e^{-0.9}}\right) =$
Yanzu 	-2.6	1.7	1.2	$-\log\left(\frac{e^{1.7}}{e^{1.1}+e^{1.7}+e^{2.2}}\right)$ $e^{-2.6}+e^{1.7}+e^{1.2}$
Lichen 	0.2	1.2	2.2	$-\log\left(\frac{e^{2.2}}{e^{0.2}+e^{1.2}+e^{2.2}}\right)$

$f(x^{(i)})$

f_{y_i}

$f(x^{(i)}; W)$

given some weight W.

how to evaluate loss given $(x^i, y^i; W)$

Optimization

- Find the optimal parameter

$$\operatorname*{argmin}_{w} \frac{1}{m} \sum_{i=1}^{m} L\left(f\left(x^{(i)}; W\right), y^{(i)}\right) + \lambda \Omega(W)$$

To apply this algorithm, we need:

1. A procedure to compute the parameter gradient
2. The regularizer (and its gradient)
3. Updating rule
4. Initialization method

- Find the optimal parameter

$$\underset{w}{\operatorname{argmin}} \frac{1}{m} \sum_{i=1}^{m} L\left(f\left(x^{(i)}; W\right), y^{(i)}\right) + \lambda\Omega(W)$$

- Stochastic Gradient Descent (SGD)

Algorithm

1. Initialize **W**

$\lambda \geq 0.$

repeat: for each training example ($\mathbf{x}^{(i)}$, $y^{(i)}$)

2a. $\Delta := -\left(\nabla_W L\left(f(x^{(i)}; W), y^{(i)}\right) + \lambda \nabla_W \Omega(W)\right)$ gradient update (negative gradient direction)

2b. $\mathbf{W} \leftarrow \mathbf{W} + \alpha\Delta$

learning rate $\alpha > 0.$

Training epoch = Iterating over all examples

To apply this algorithm, we need:

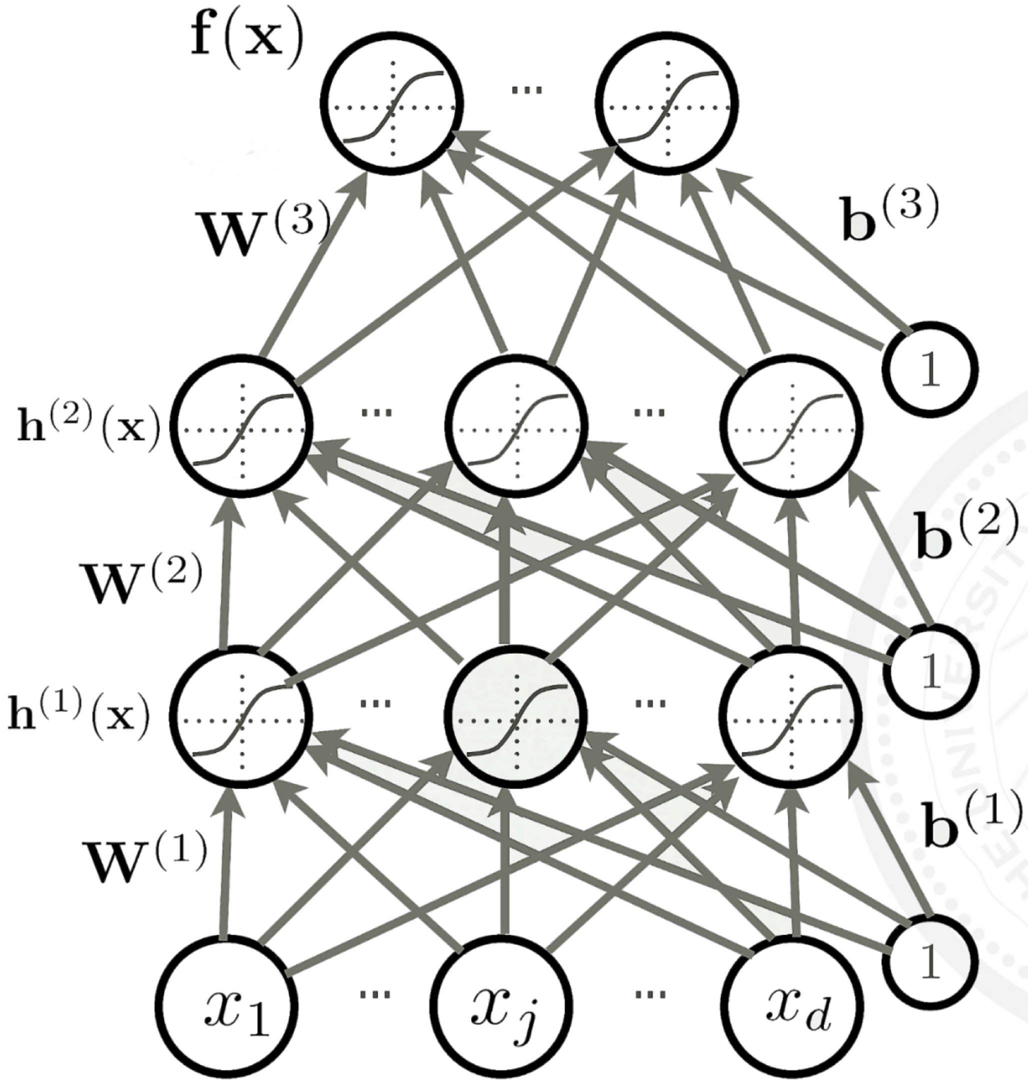
1. A procedure to compute the parameter gradient
2. The regularizer (and its gradient)
3. Updating rule
4. Initialization method

Animation: (for SVM loss)

http://vision.stanford.edu/teaching/cs231n-demos/linear-classify/

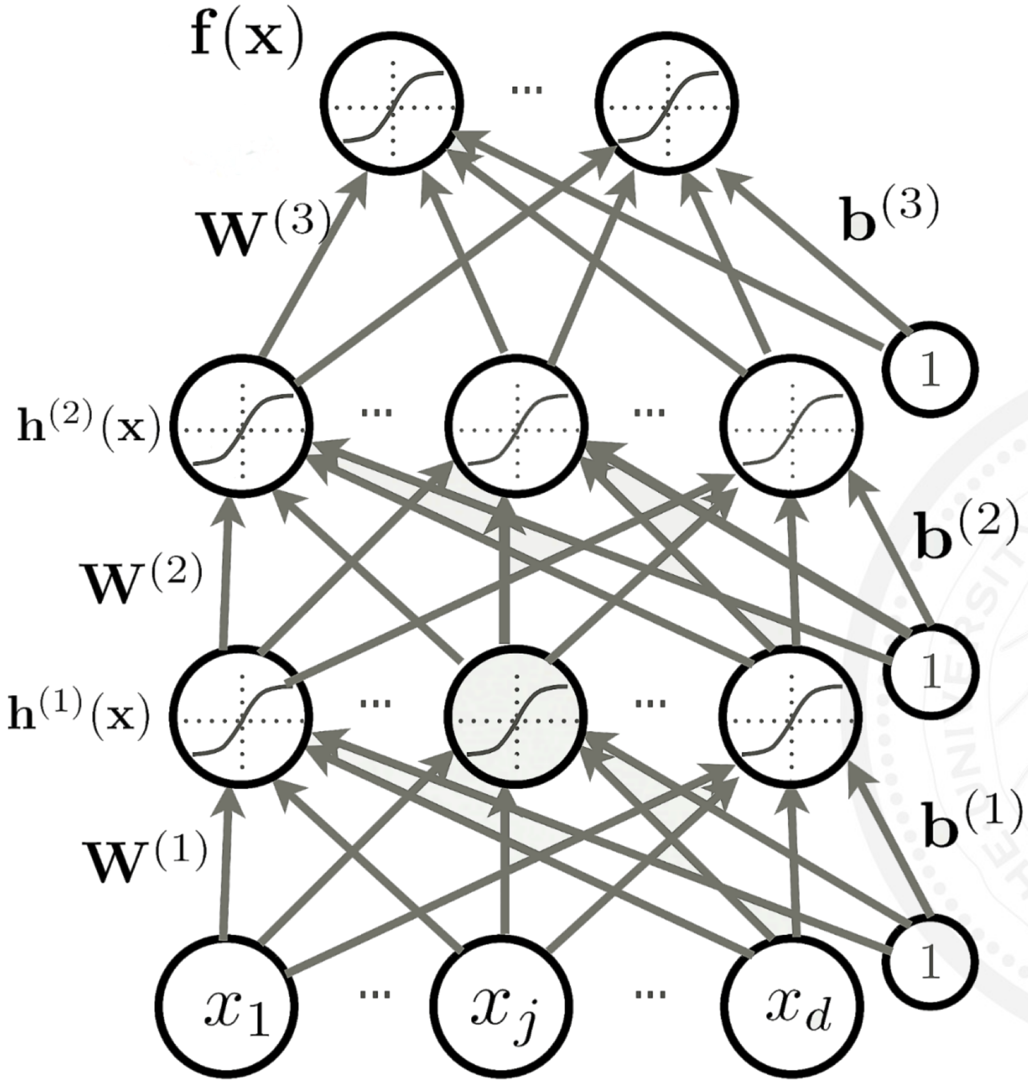
Follow the slope

How many parameter do we have?



How many parameter do we have?

VGGNet *[Simonyan and Zisserman, 2014]* used 138M parameters

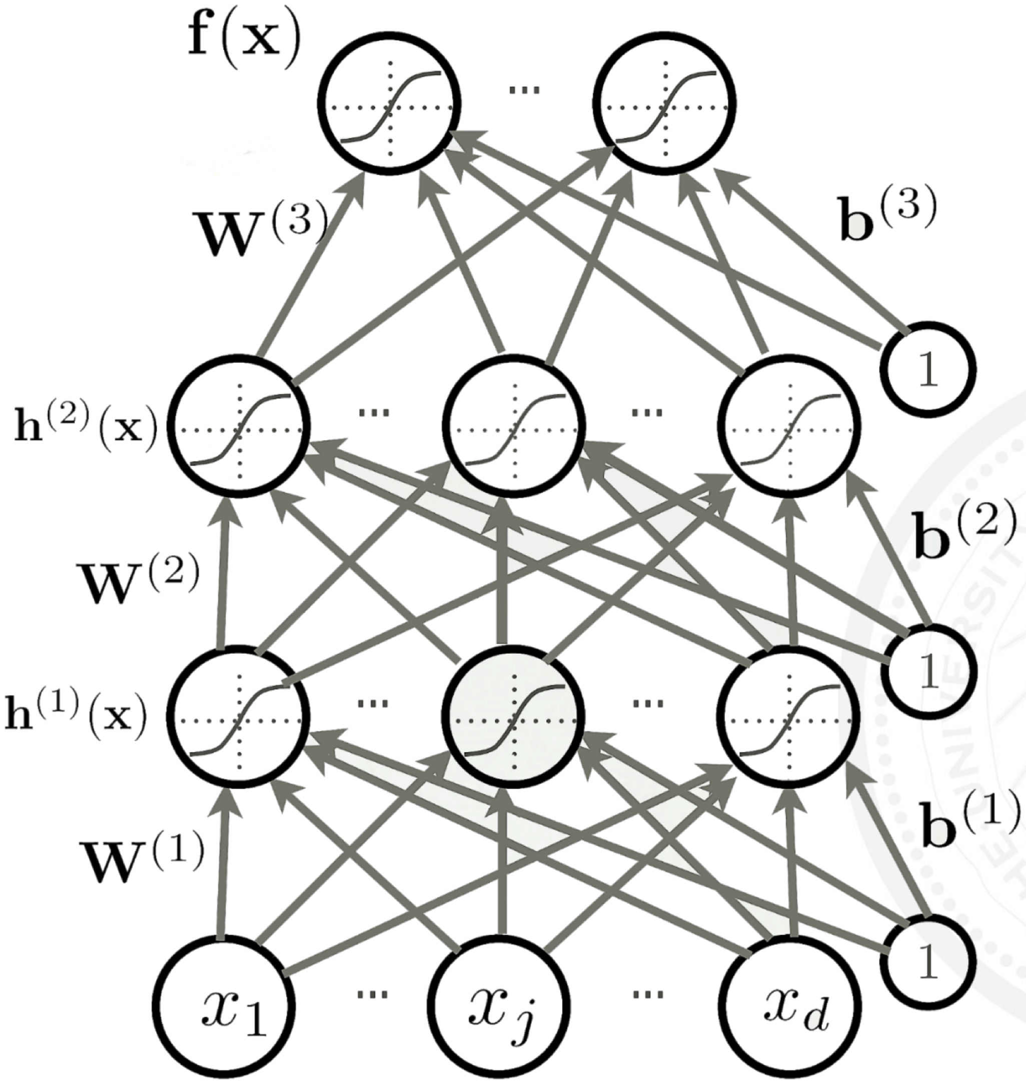


How many parameter do we have?

VGGNet *[Simonyan and Zisserman, 2014]* used 138M parameters

In 1-dimension, the derivative of a function:

$$\frac{df(x)}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$



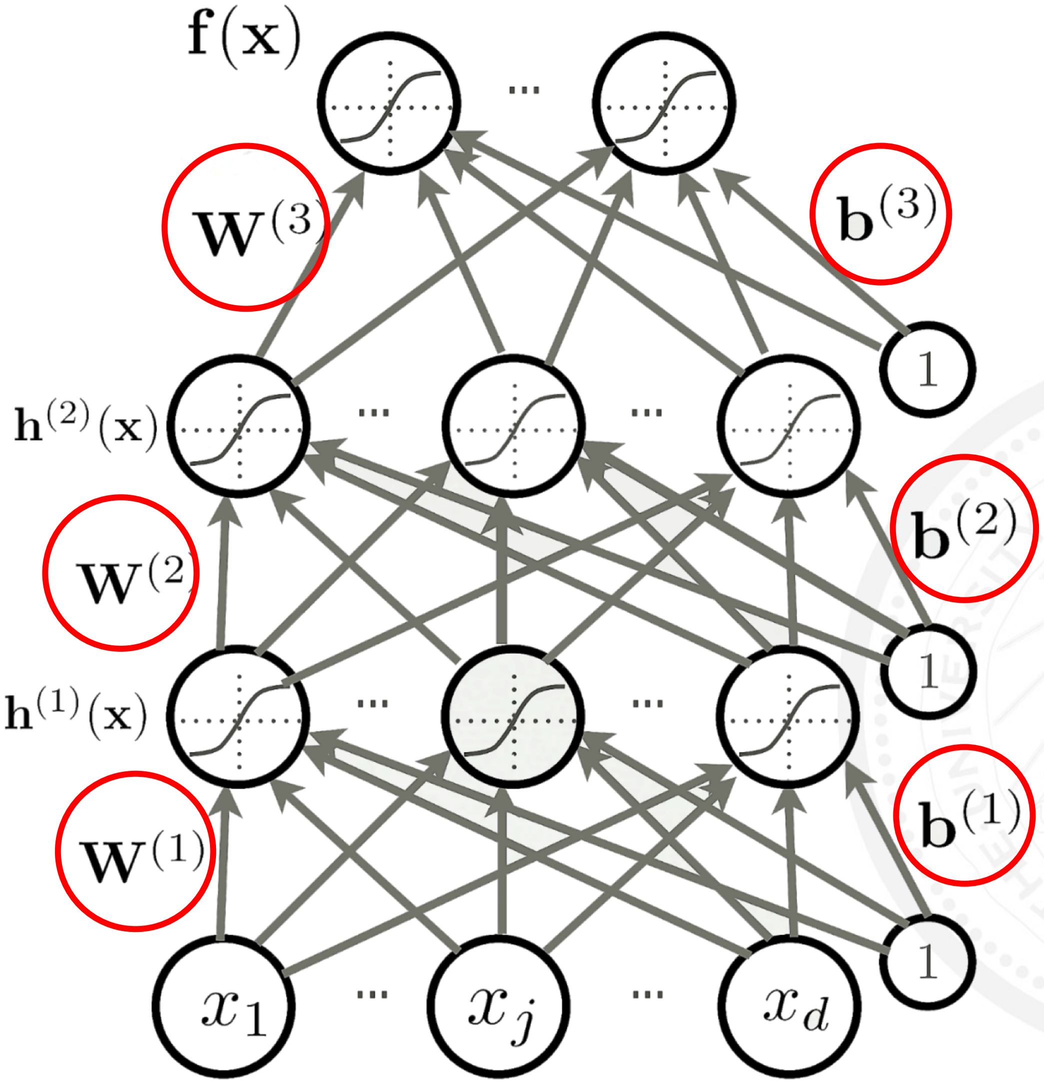
Follow the slope

How many parameter do we have?

VGGNet *[Simonyan and Zisserman, 2014]* used 138M parameters

In 1-dimension, the derivative of a function:

$$\frac{df(x)}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$



Current W:

Gradient dW:

Numerical Gradient

Current W:

[0.25,
-1.56,
0.55,
3.8,
0.98,
0.77,
-0.11,
-2.9,…]

Loss 1.25742

Gradient dW:

[? ,
?,
?,
?,
?,
?,
?,
?,…]

Numerical Gradient

TBSI 清华-伯克利深圳学院 Tsinghua-Berkeley Shenzhen Institute

Current W:	W + h (third dim):	Gradient dW:
[0.25,	[0.25 + 0.0001,	[? ,
-1.56,	-1.56,	?,
0.55,	0.55,	?,
3.8,	3.8,	?,
0.98,	0.98,	?,
0.77,	0.77,	?,
-0.11,	-0.11,	?,
-2.9,…]	-2.9,…]	?,…]
Loss 1.25742	**Loss 1.25763**	

Numerical Gradient

Current W:	W + h (third dim):	Gradient dW:
[0.25,	[0.25 + 0.0001,	[2.1,
-1.56,	-1.56,	?,
0.55,	0.55,	?,
3.8,	3.8,	?,
0.98,	0.98,	?,
0.77,	0.77,	?,
-0.11,	-0.11,	?,
-2.9,…]	-2.9,…]	?,…]
Loss 1.25742	**Loss 1.25763**	

$$\lim_{h\to 0}\frac{f(x+h)-f(x)}{h}$$

$$= \frac{(1.25763 - 1.25742)}{0.0001}$$

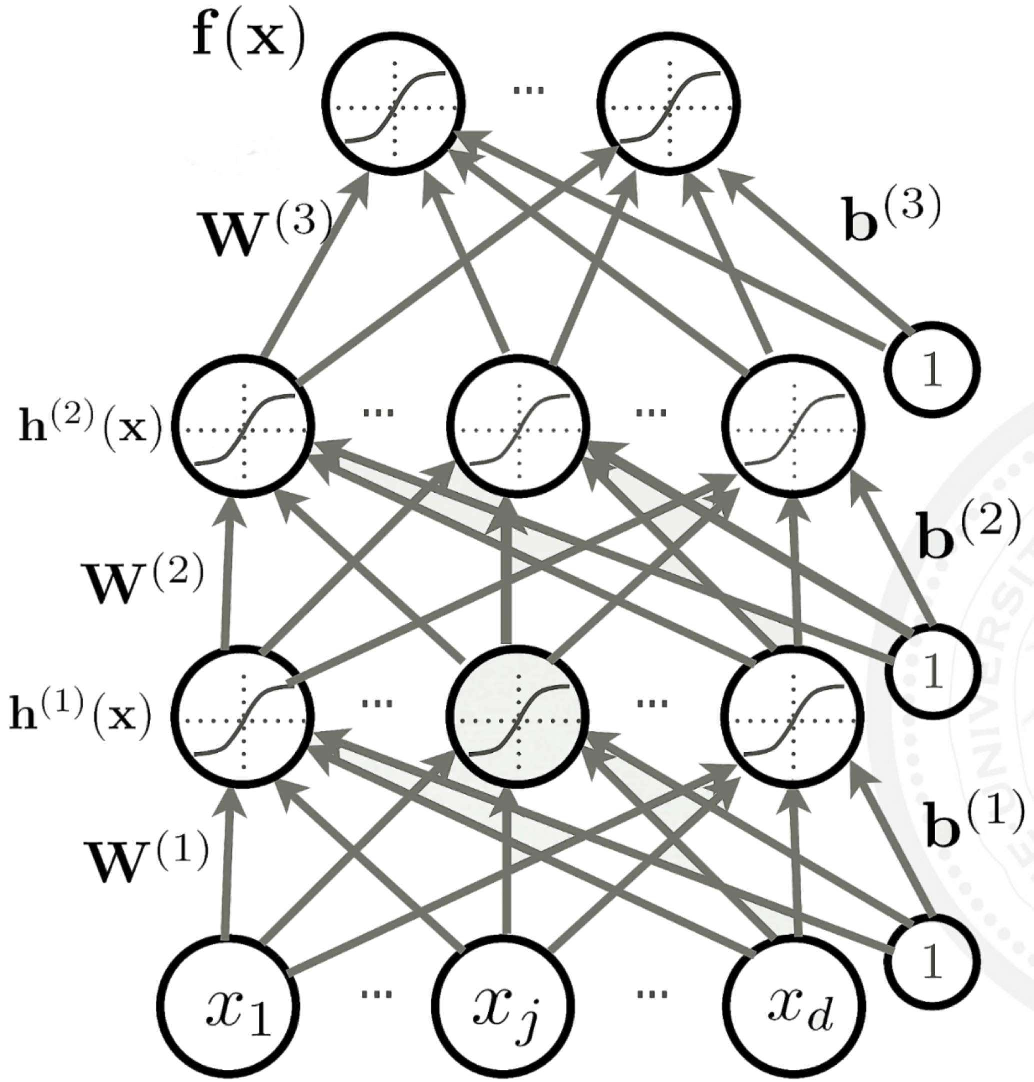
Follow the slope

How many parameter do we have?

VGGNet *[Simonyan and Zisserman, 2014]* used 138M parameters

In 1-dimension, the derivative of a function:

$$\frac{df(x)}{dx} = \lim_{h\to 0} \underbrace{\frac{f(x+h)-f(x)}{h}}$$



Follow the slope

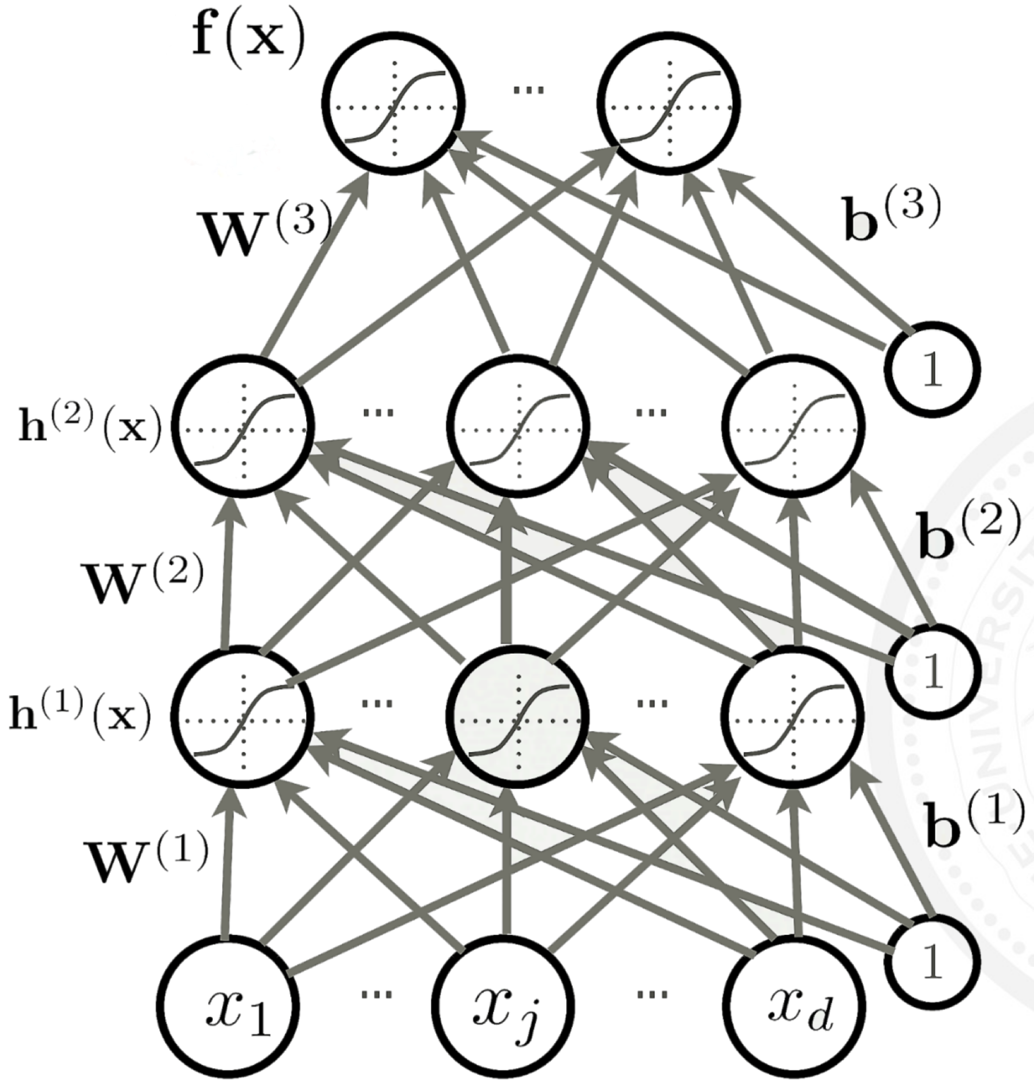
How many parameter do we have?

VGGNet *[Simonyan and Zisserman, 2014]* used 138M parameters

In 1-dimension, the derivative of a function:

$$\frac{df(x)}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

Numerical gradient: approximate, slow, easy to write



Follow the slope

How many parameter do we have?

VGGNet *[Simonyan and Zisserman, 2014]* used 138M parameters

In 1-dimension, the derivative of a function:

$$\frac{df(x)}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

Numerical gradient: approximate, slow, easy to write

Calculus! $L(i)$ (i)

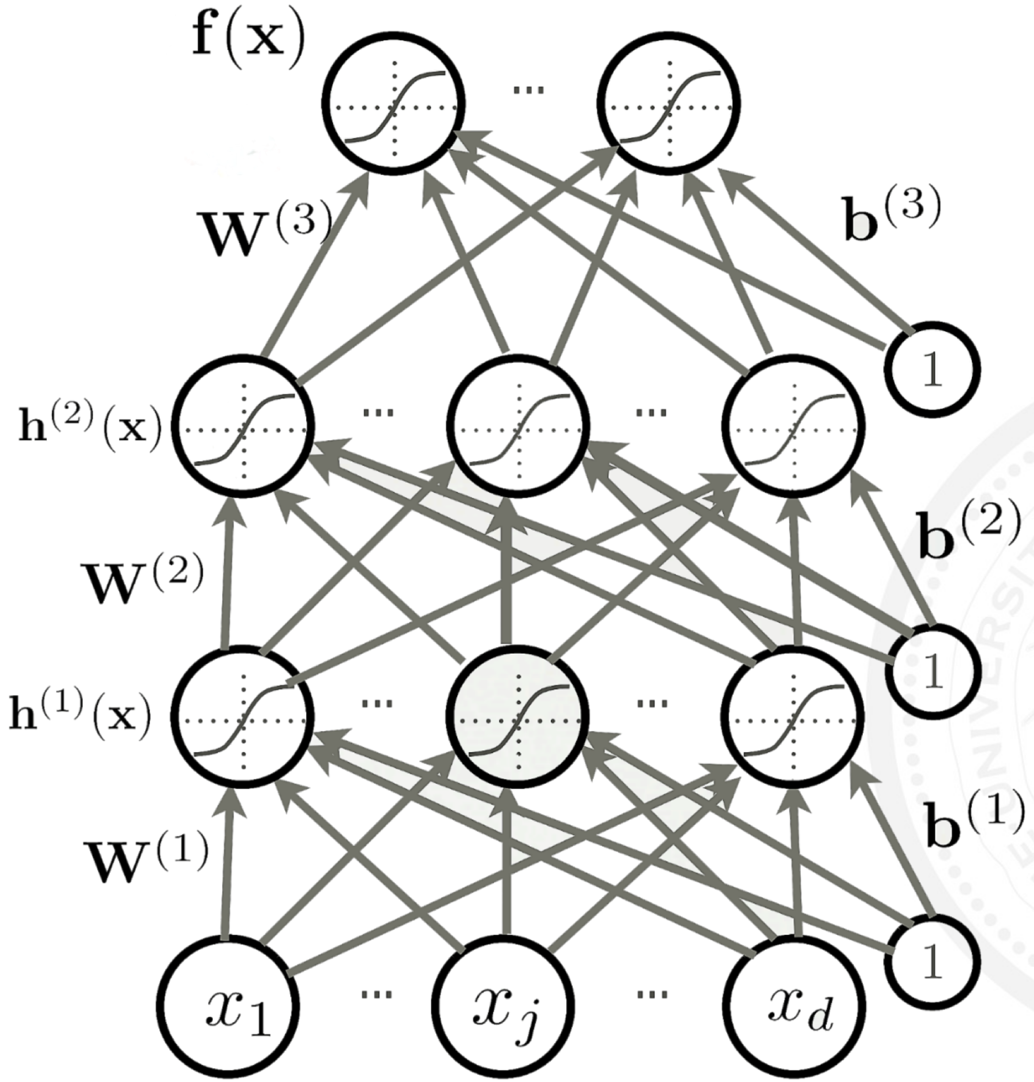
$$L = \frac{1}{N} \sum_t \underline{L_t\left(f\left(\mathbf{x}^{(t)}; \mathbf{W}\right), y^{(t)}\right) + \lambda\Omega(\mathbf{W})}$$

$$\mathbf{h}^{(L+1)}(\mathbf{x}) = \mathbf{g}(\underline{\mathbf{a}^{(L+1)}(\mathbf{x})}) = \mathbf{f}(\mathbf{x})$$

$$\underline{\mathbf{h}^{(k)}}(\mathbf{x}) = \underline{\mathbf{g}(\mathbf{a}^{(k)}(\mathbf{x}))}$$

$$\underline{\mathbf{a}^{(k)}}(\mathbf{x}) = \underline{\mathbf{b}^{(k)} + \mathbf{W}^{(k)}\mathbf{h}^{(k-1)}(\mathbf{x})}$$

Analytic gradient: exact, fast, error-prone

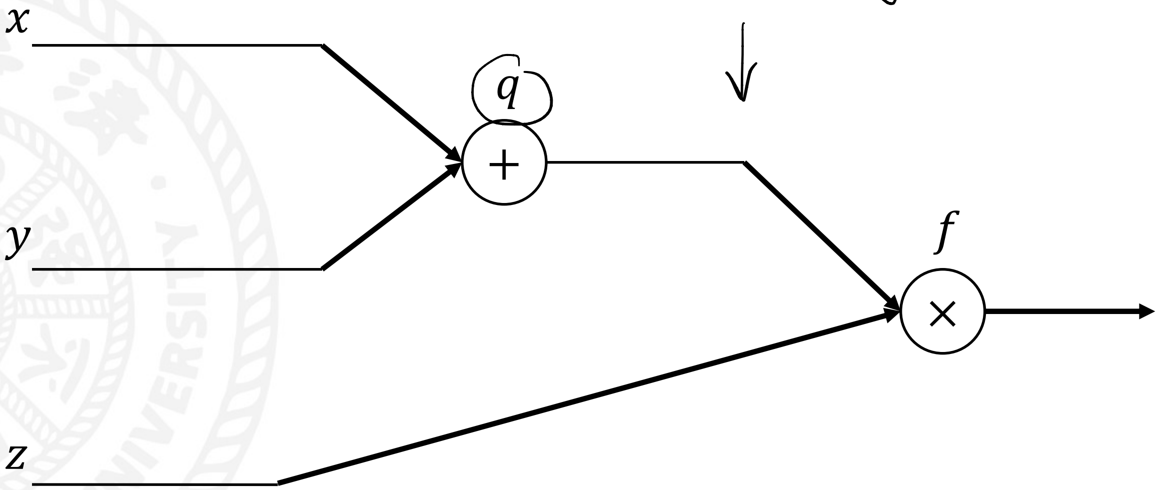


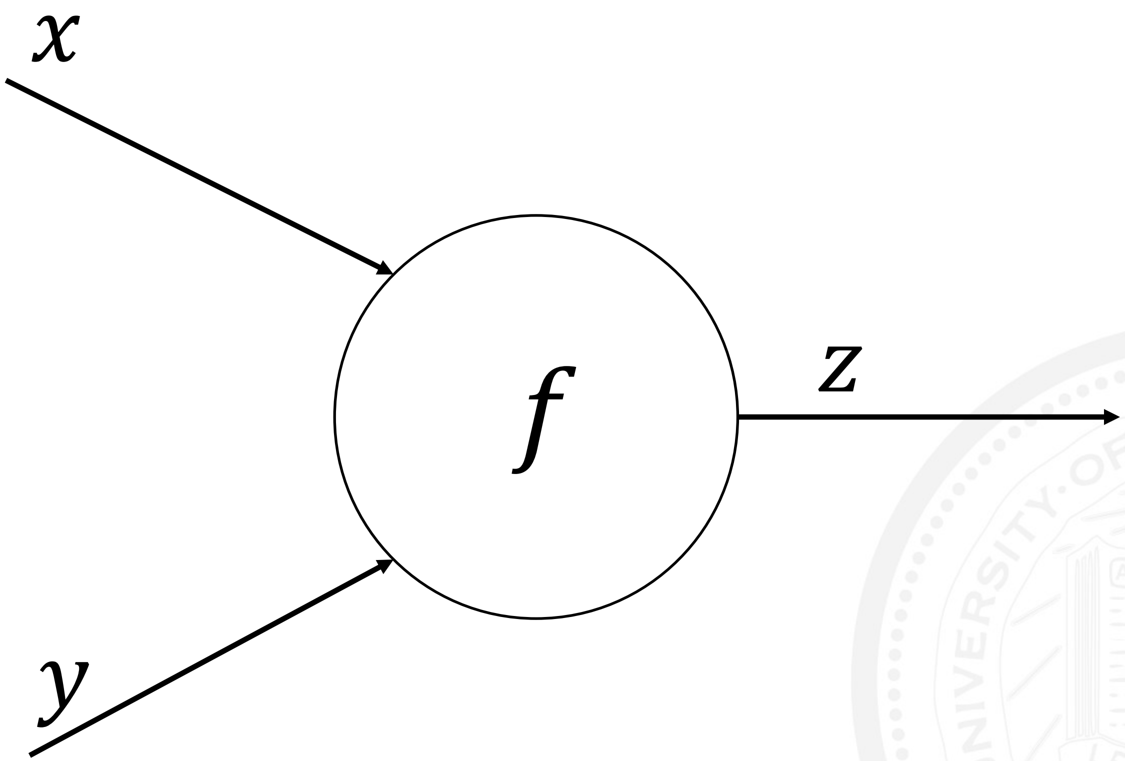
Backpropagation

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

We want $\frac{df}{dx}, \frac{df}{dy}, \frac{df}{dz}$

Computation graph.

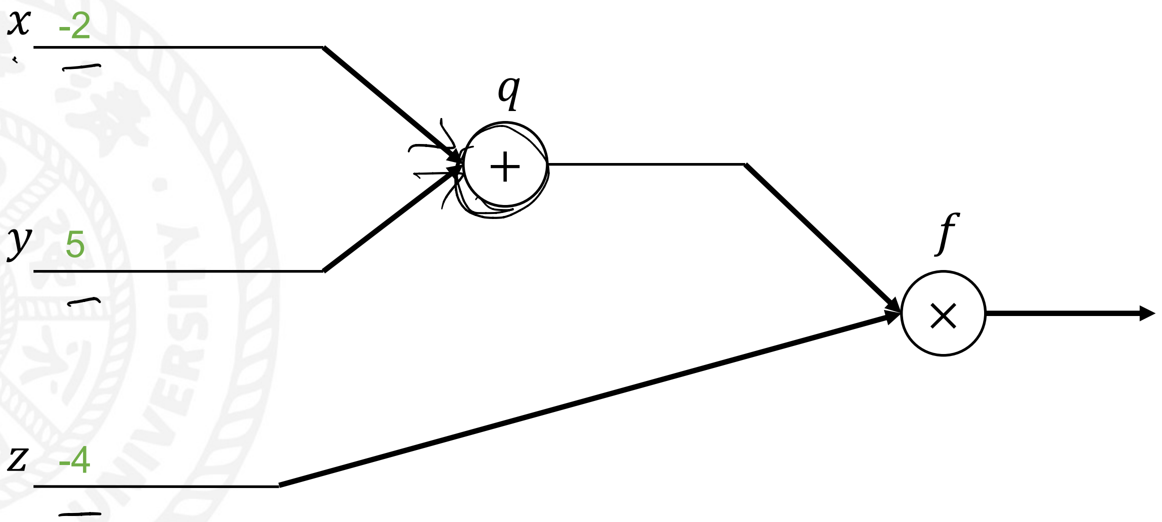


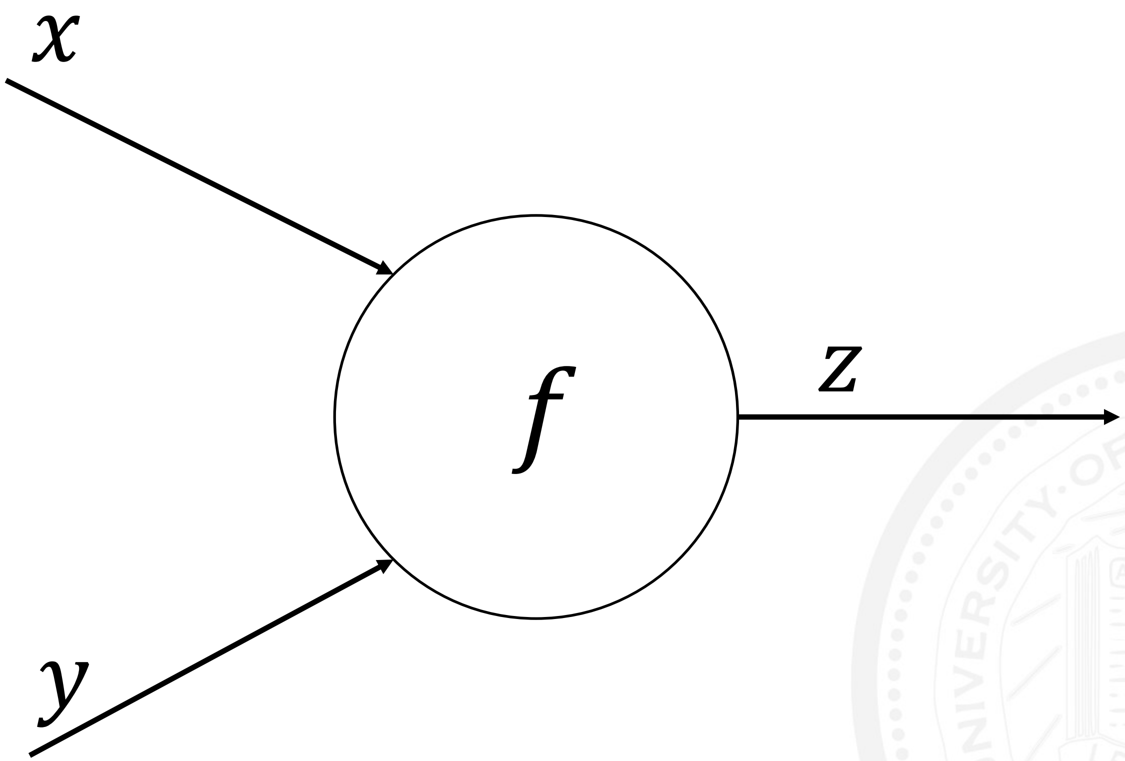


Backpropagation

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

We want $\frac{df}{dx}, \frac{df}{dy}, \frac{df}{dz}$

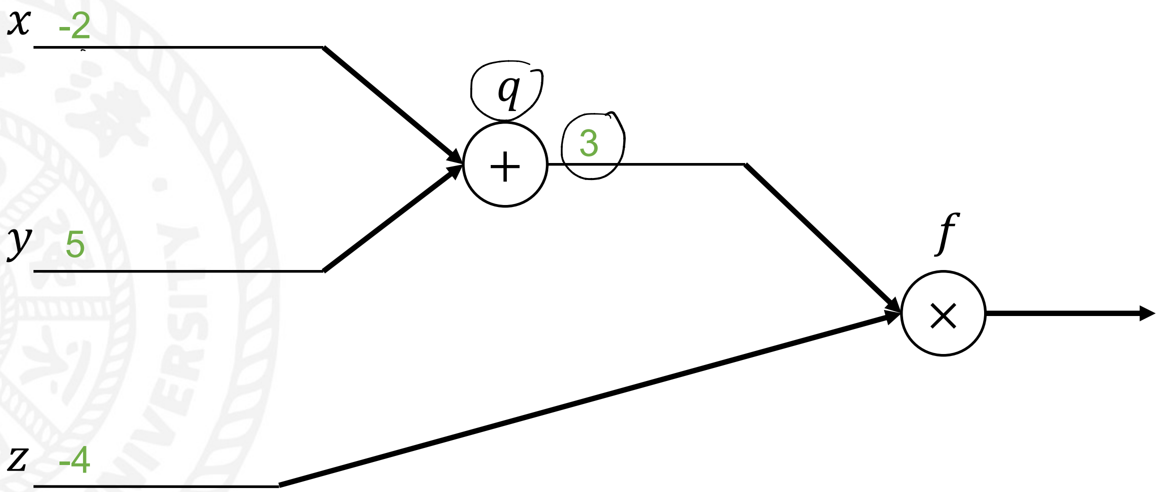


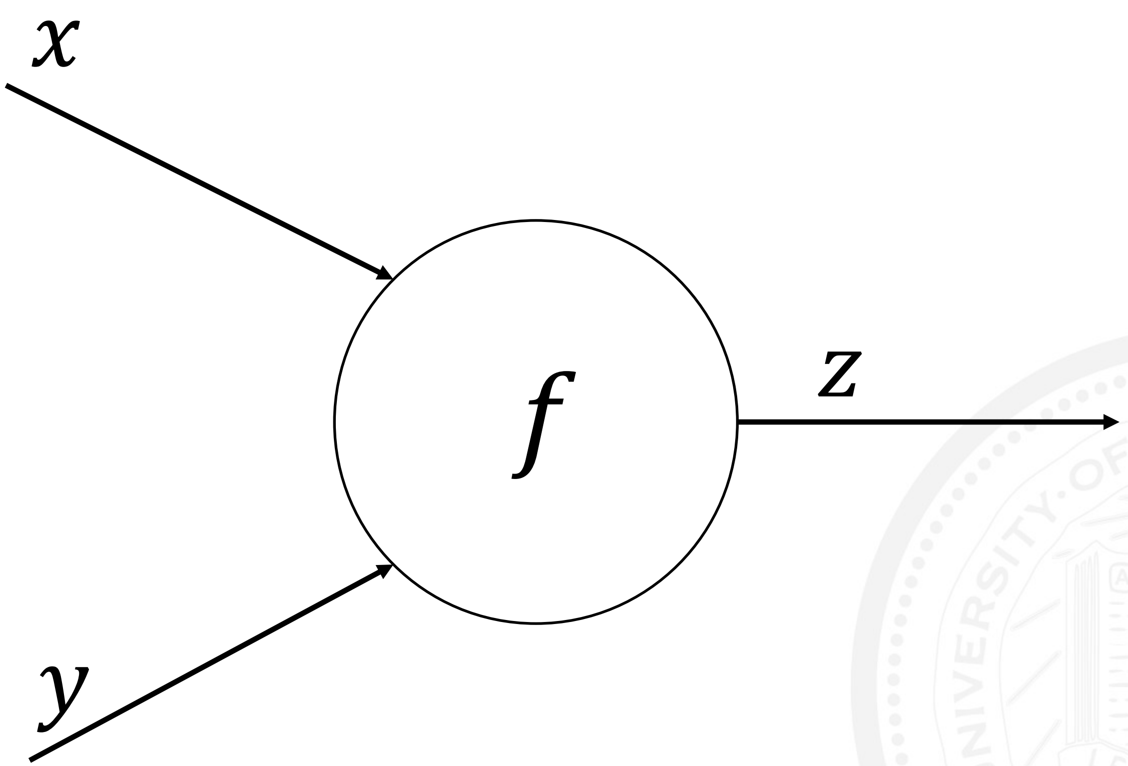


Backpropagation

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

We want $\frac{df}{dx}, \frac{df}{dy}, \frac{df}{dz}$

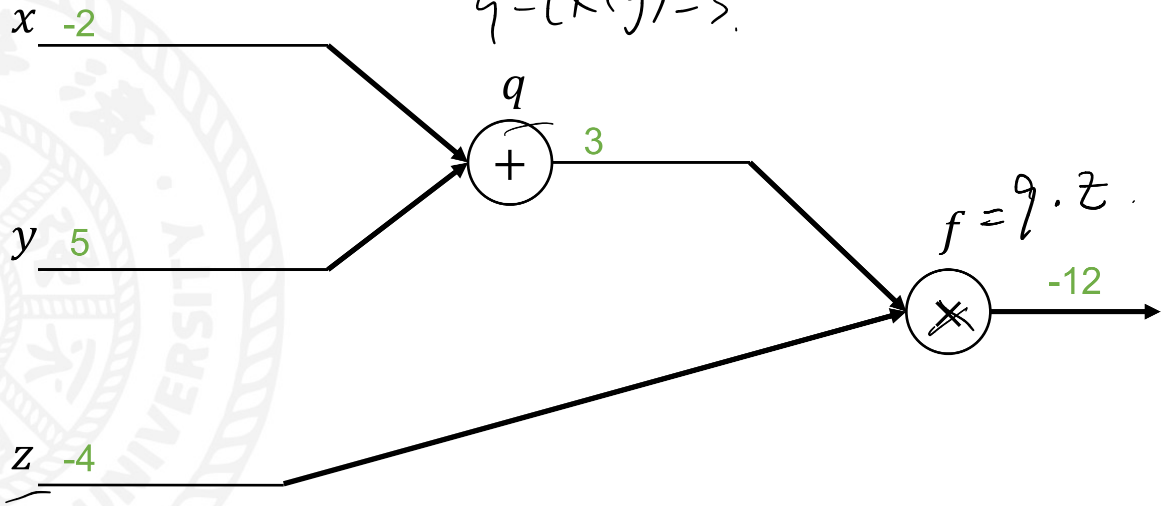


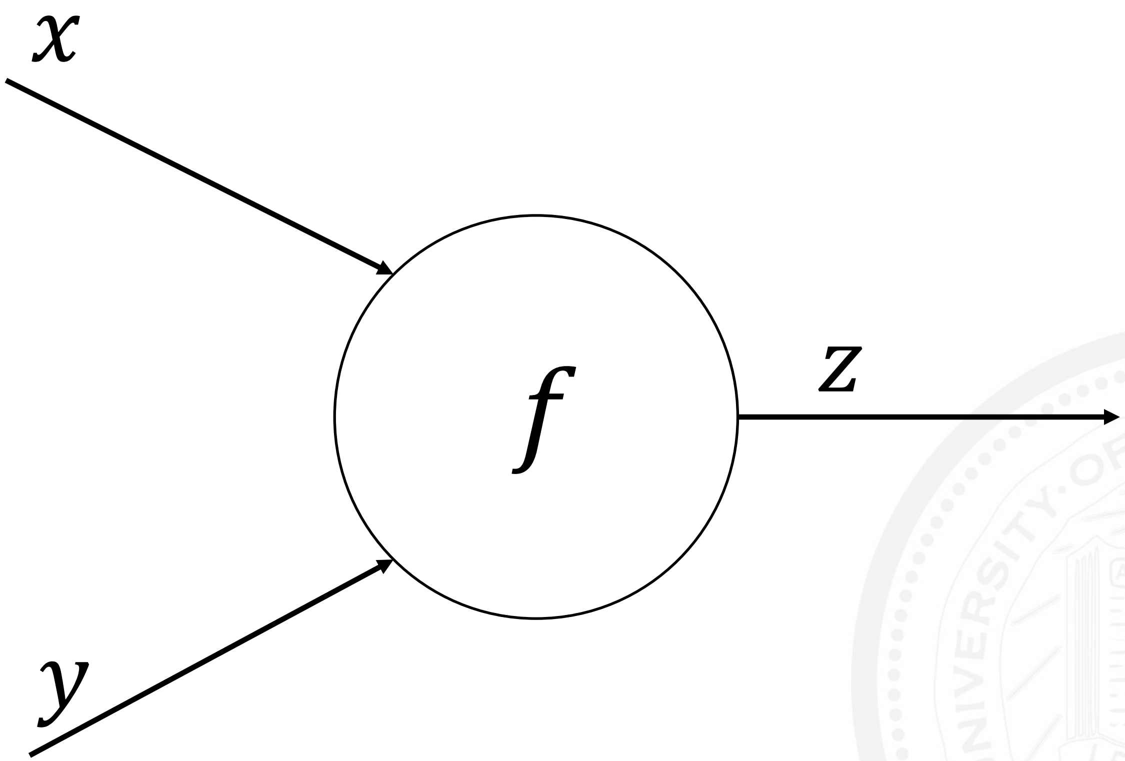


$$f(x, y, z) = (x + y)z$$

We want $\dfrac{df}{dx}, \dfrac{df}{dy}, \dfrac{df}{dz}$

$q = (x+y) = 3.$



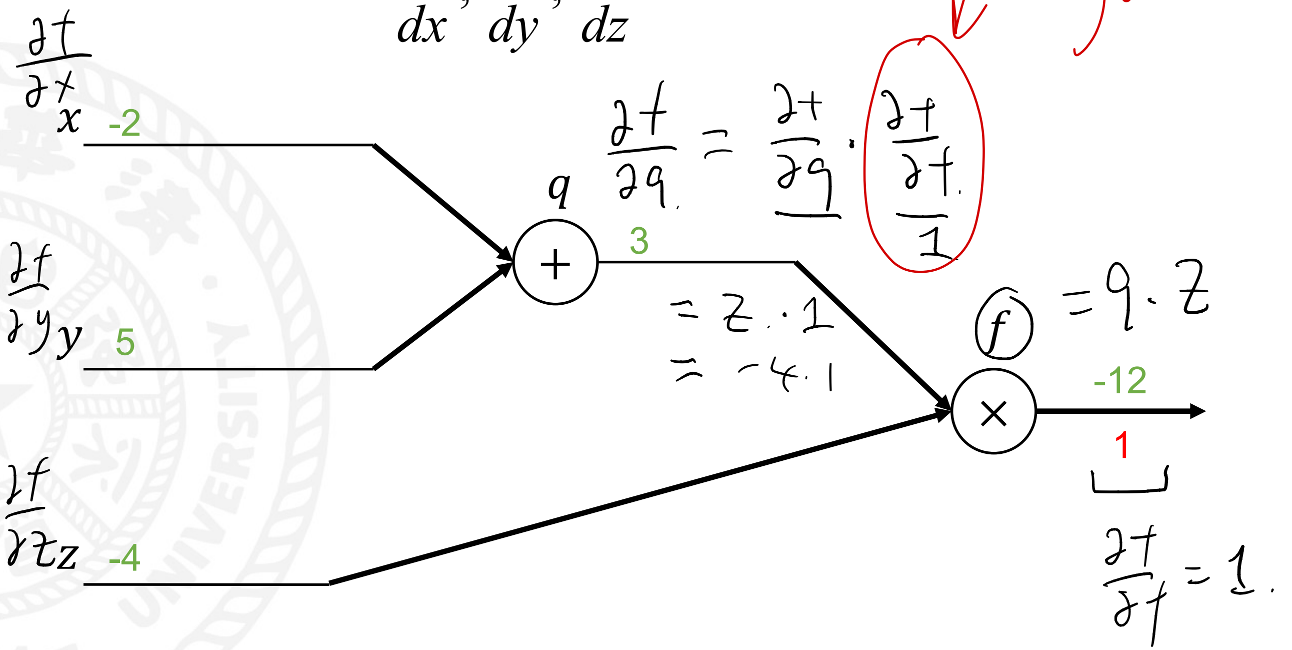


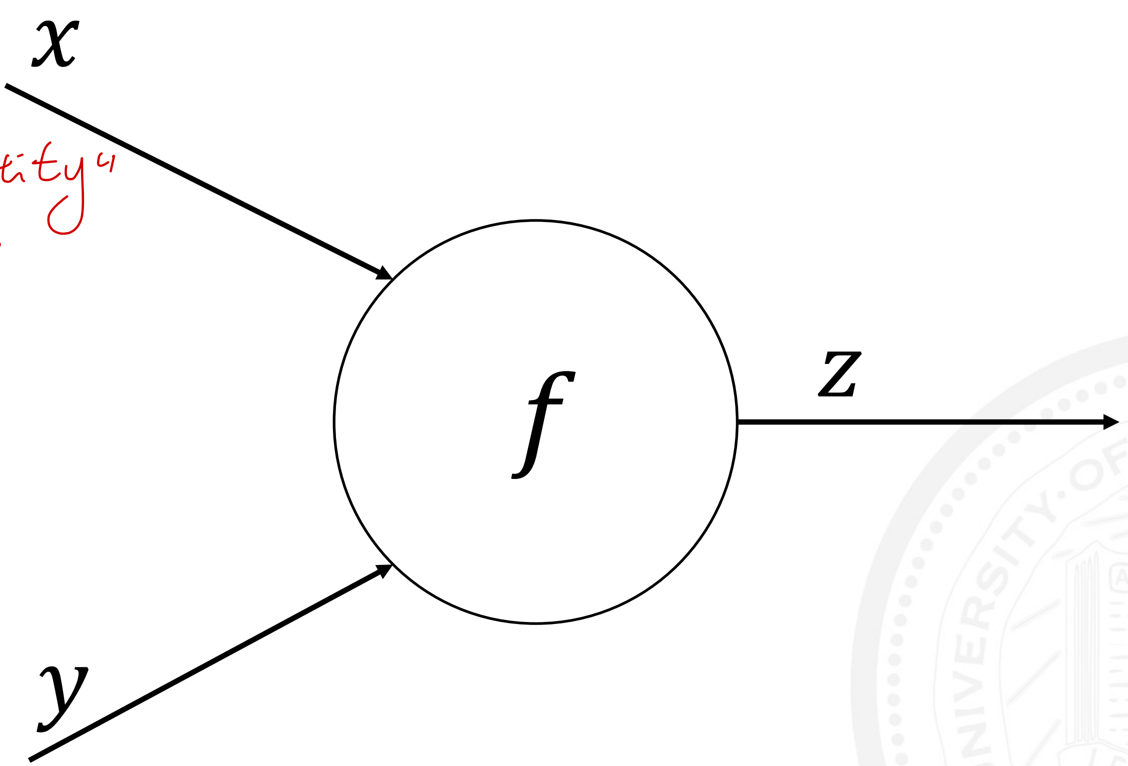
Backpropagation

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

We want $\dfrac{df}{dx}, \dfrac{df}{dy}, \dfrac{df}{dz}$

derivative of the "identity" function

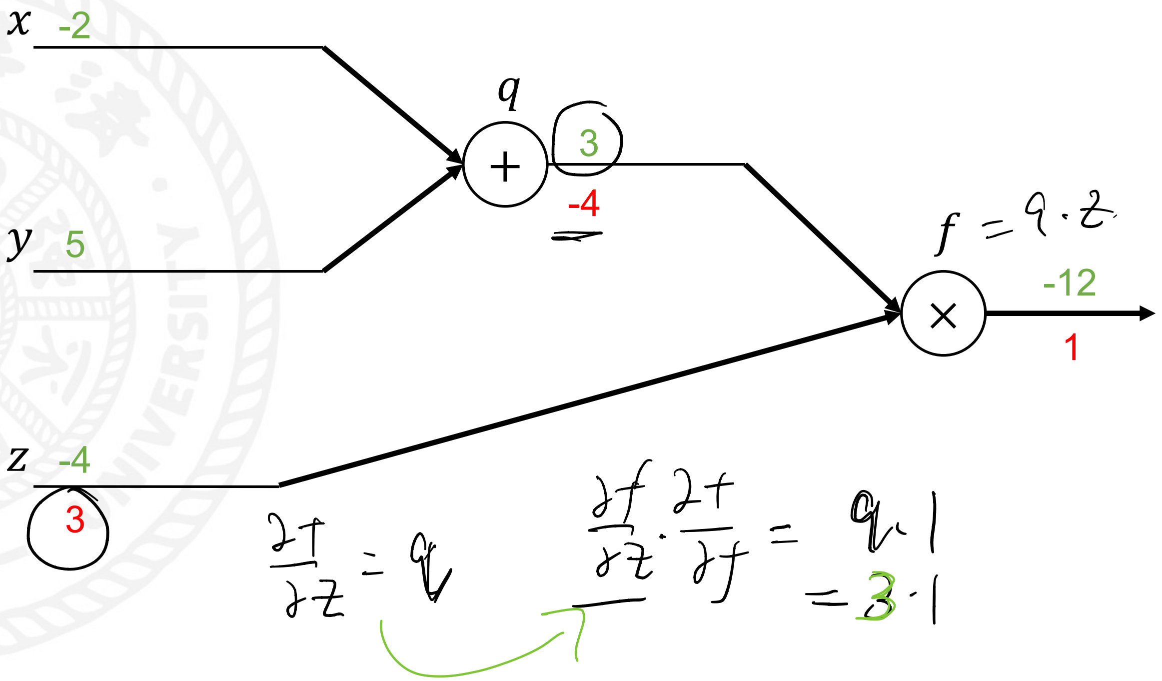


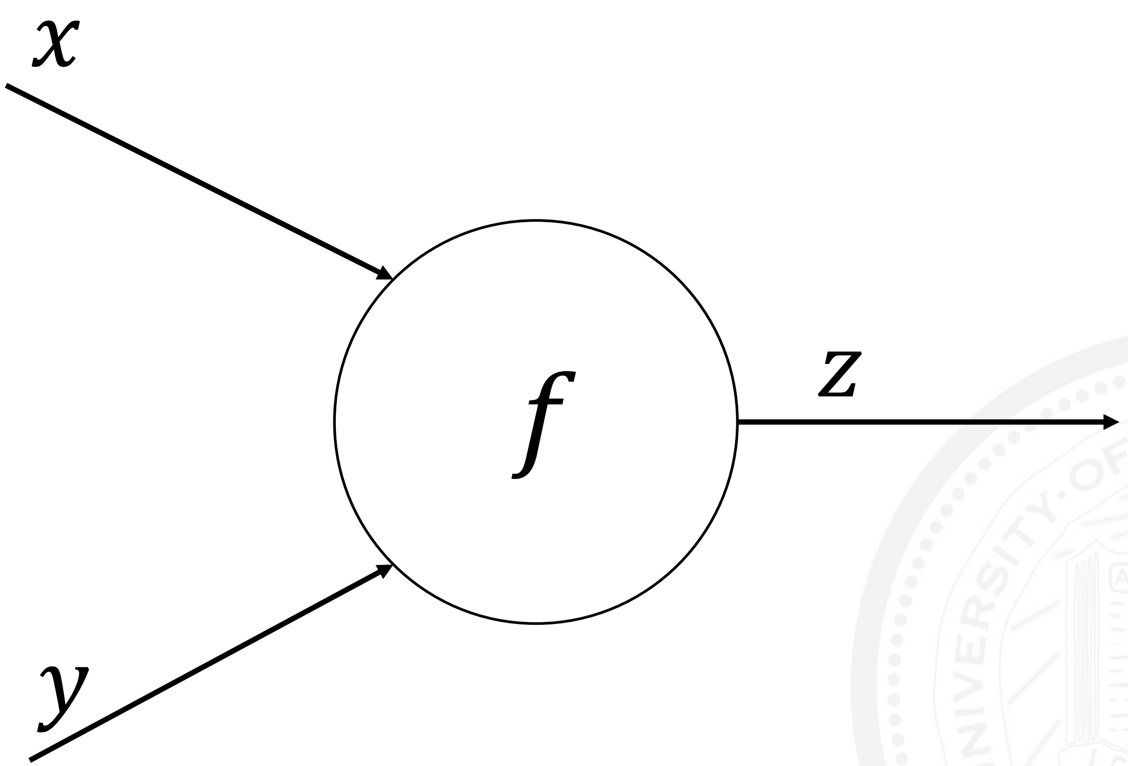


Backpropagation

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

We want $\frac{df}{dx}, \frac{df}{dy}, \frac{df}{dz}$





Backpropagation

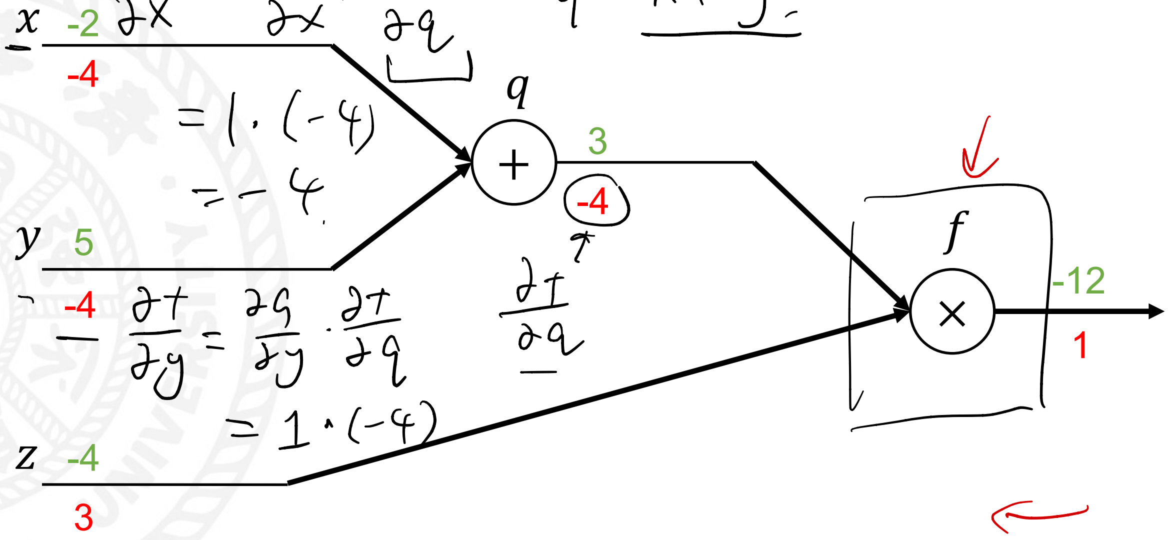
$$f(x,y,z) = (x+y)z$$

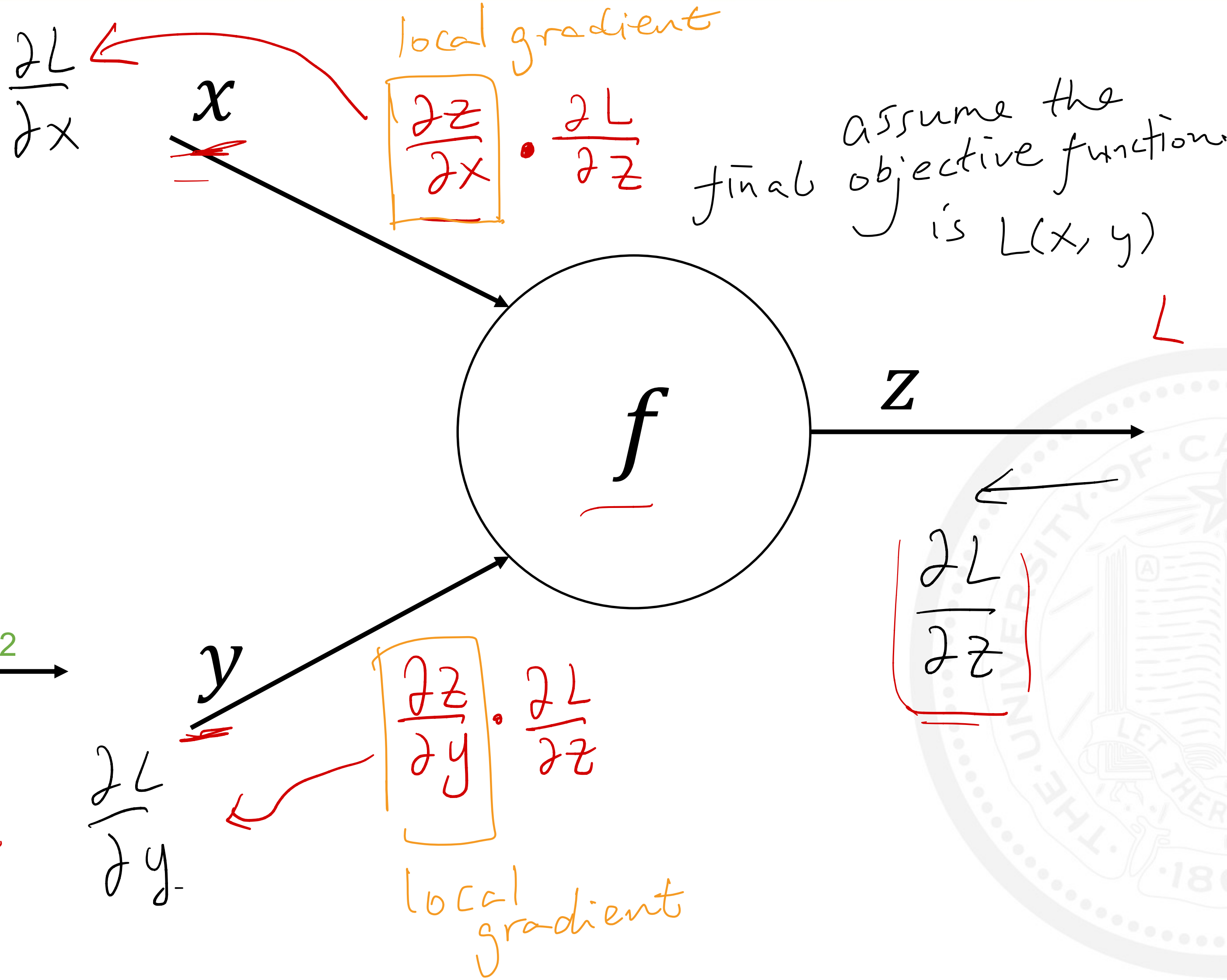
We want $\frac{df}{dx}, \frac{df}{dy}, \frac{df}{dz}$

$\frac{\partial f}{\partial x} = \frac{\partial q}{\partial x} \cdot \frac{\partial f}{\partial q} = 1 \cdot (-4) = -4$

$q = x + y$

$\frac{\partial f}{\partial y} = \frac{\partial q}{\partial y} \cdot \frac{\partial f}{\partial q} = 1 \cdot (-4)$

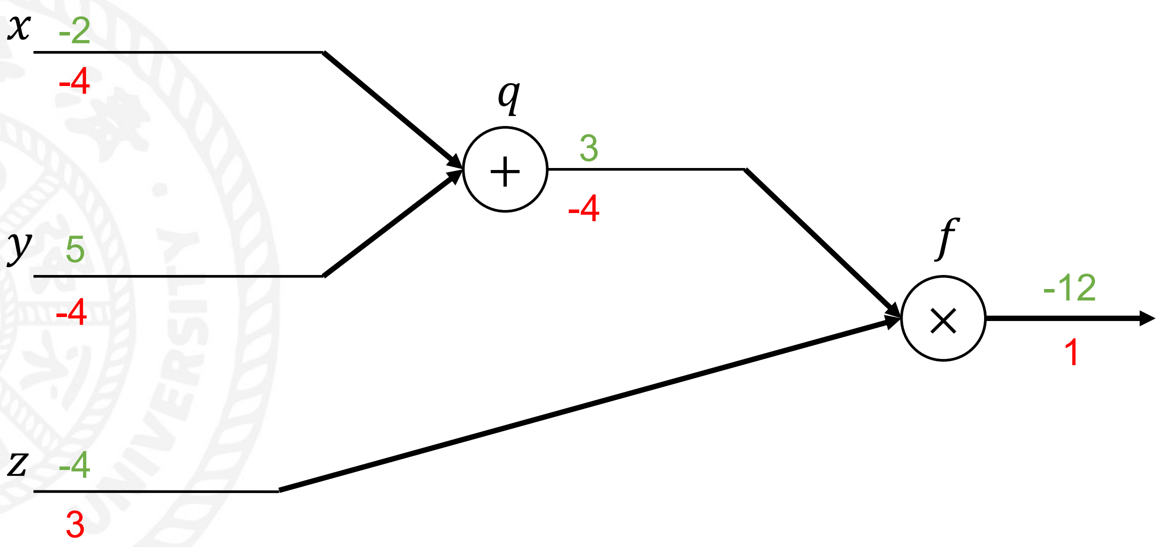


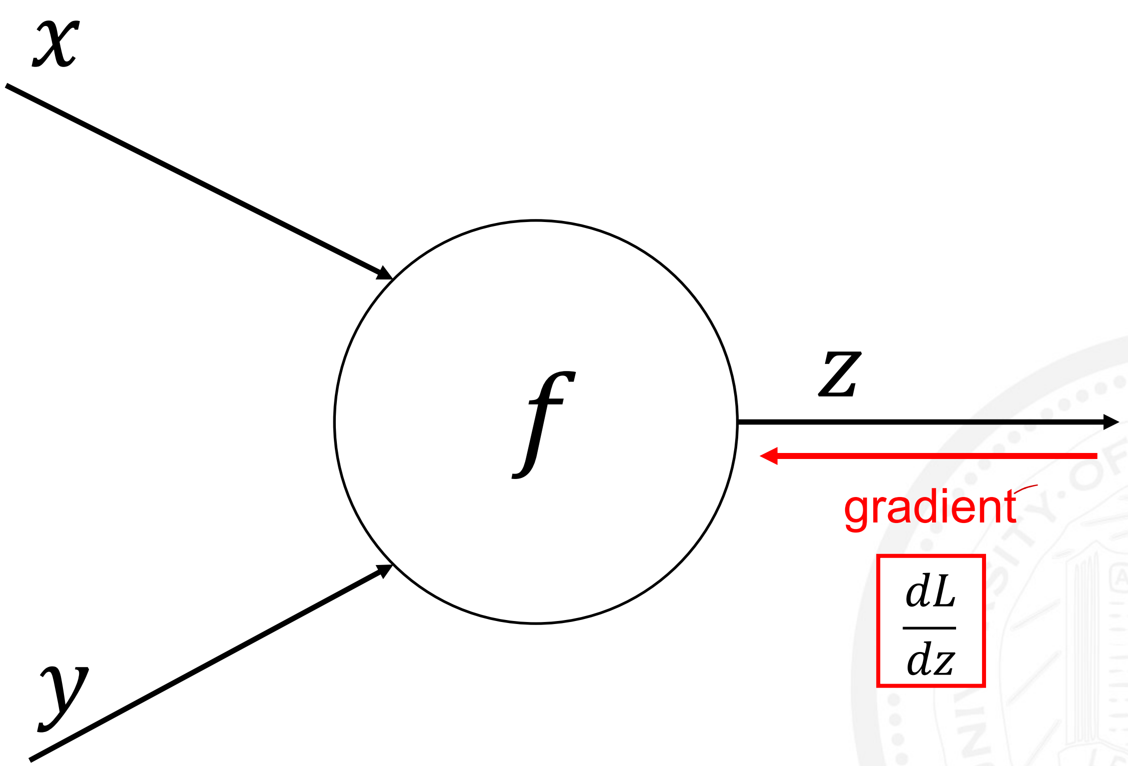


Backpropagation

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

We want $\dfrac{df}{dx}, \dfrac{df}{dy}, \dfrac{df}{dz}$

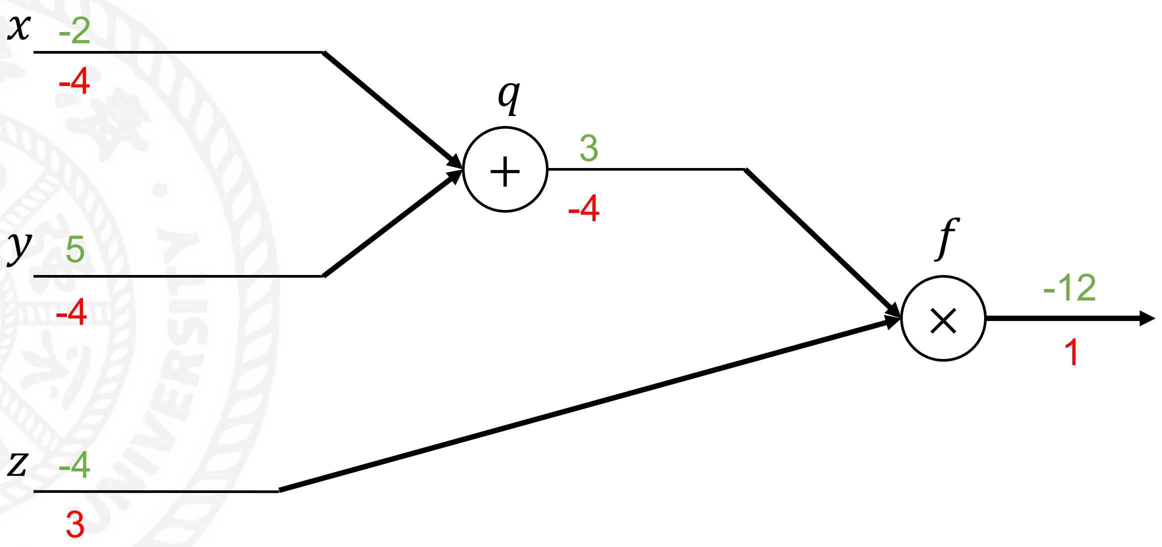


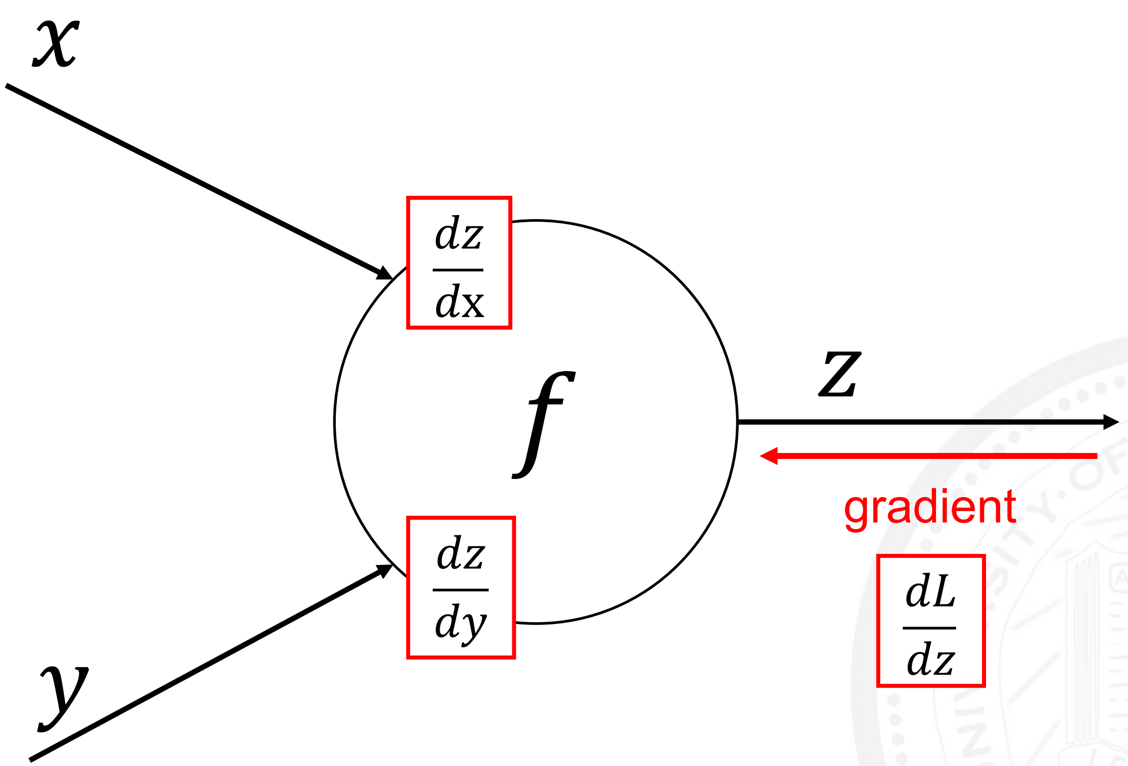


Backpropagation

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

We want $\dfrac{df}{dx}, \dfrac{df}{dy}, \dfrac{df}{dz}$

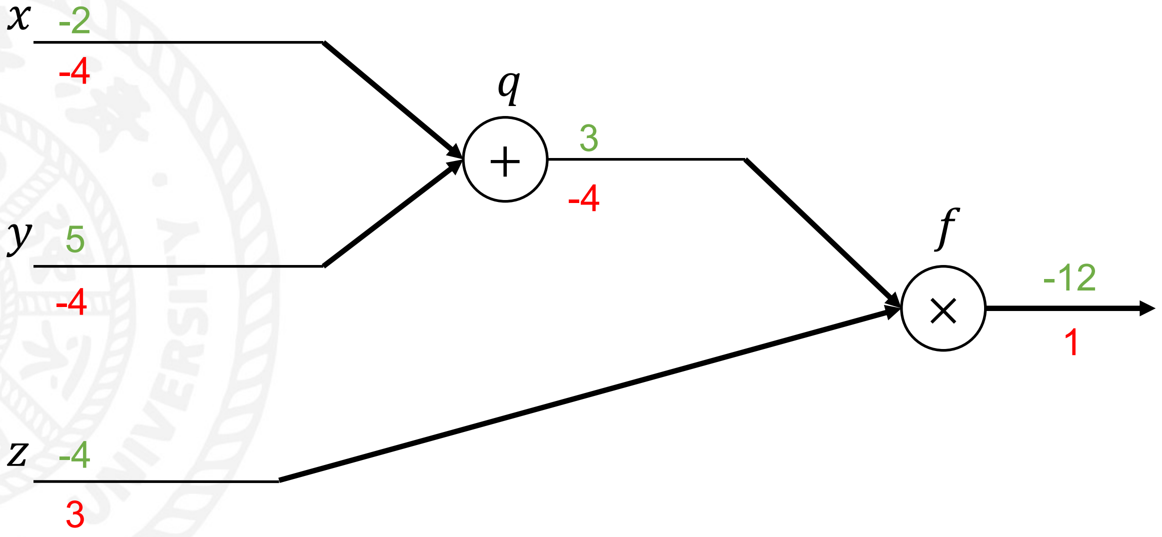


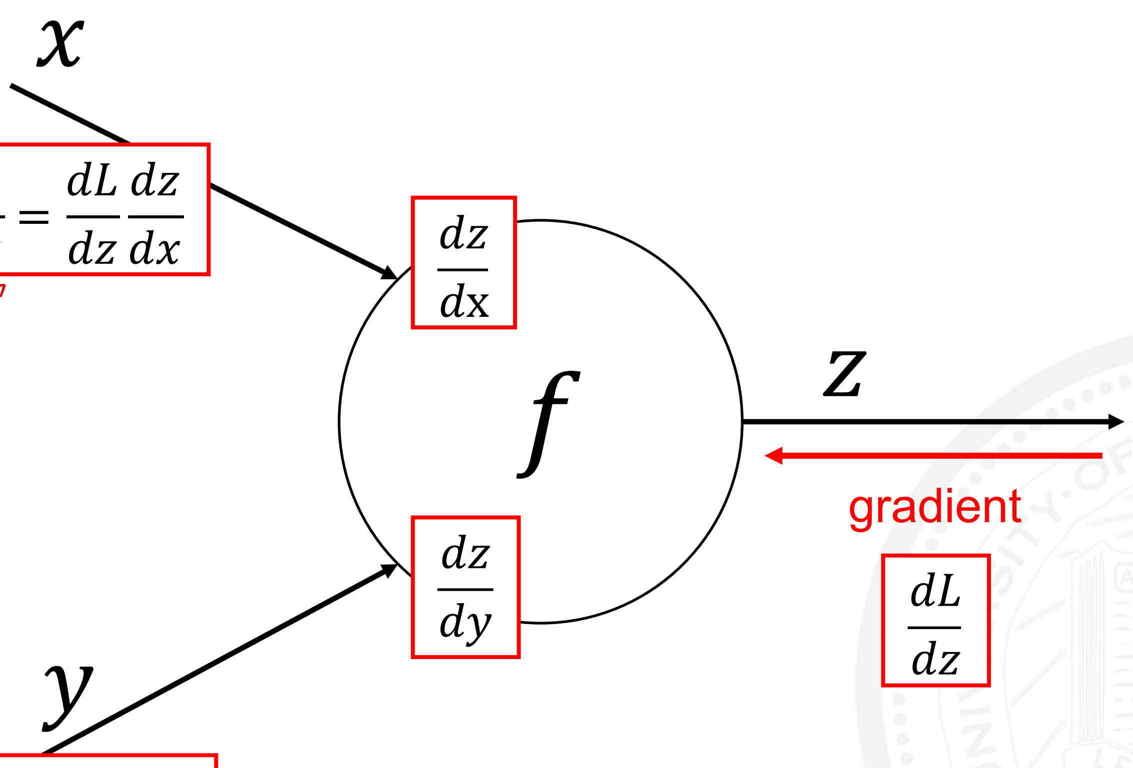


Backpropagation

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

We want $\frac{df}{dx}, \frac{df}{dy}, \frac{df}{dz}$





$$\boxed{\frac{dL}{dx}=\frac{dL}{dz}\frac{dz}{dx}}$$

$$\boxed{\frac{dz}{dx}}$$

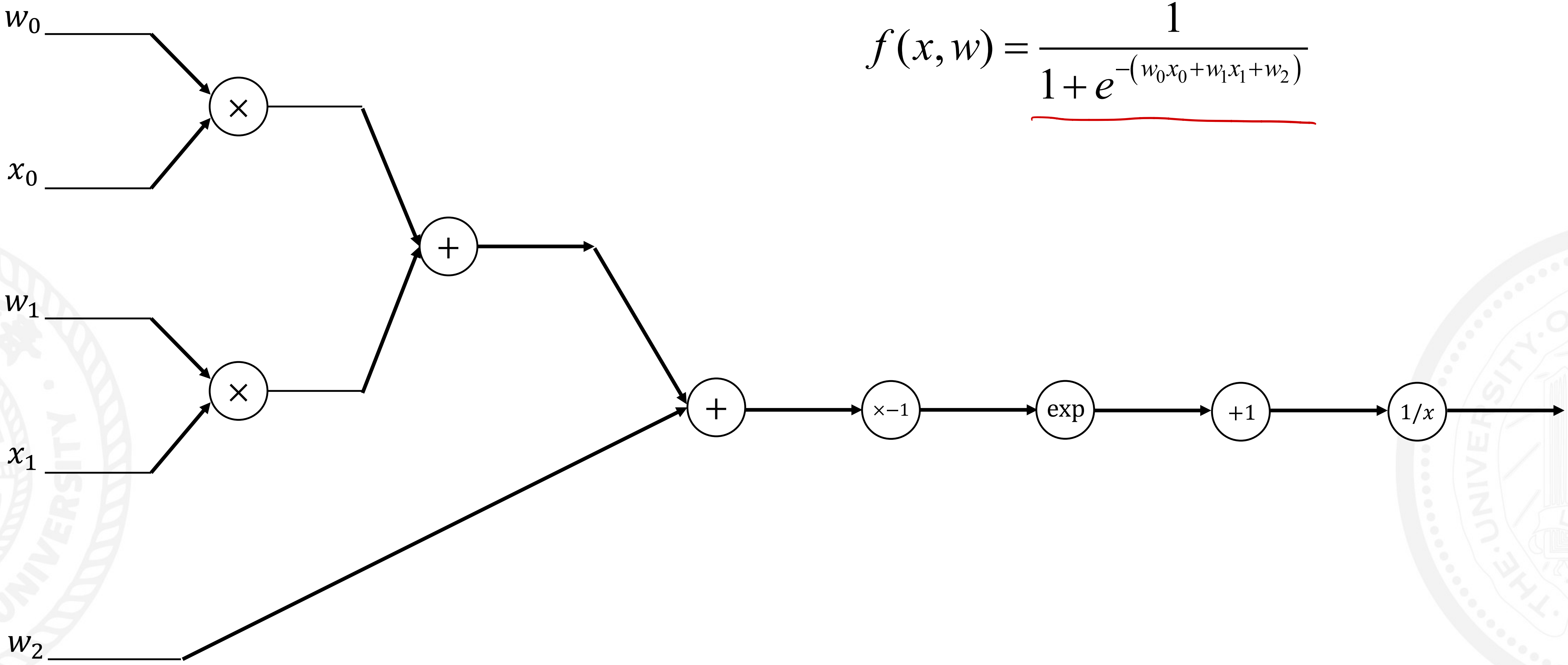
$$\boxed{\frac{dz}{dy}}$$

$$\boxed{\frac{dL}{dz}}$$

$$\boxed{\frac{dL}{dy}=\frac{dL}{dz}\frac{dz}{dy}}$$

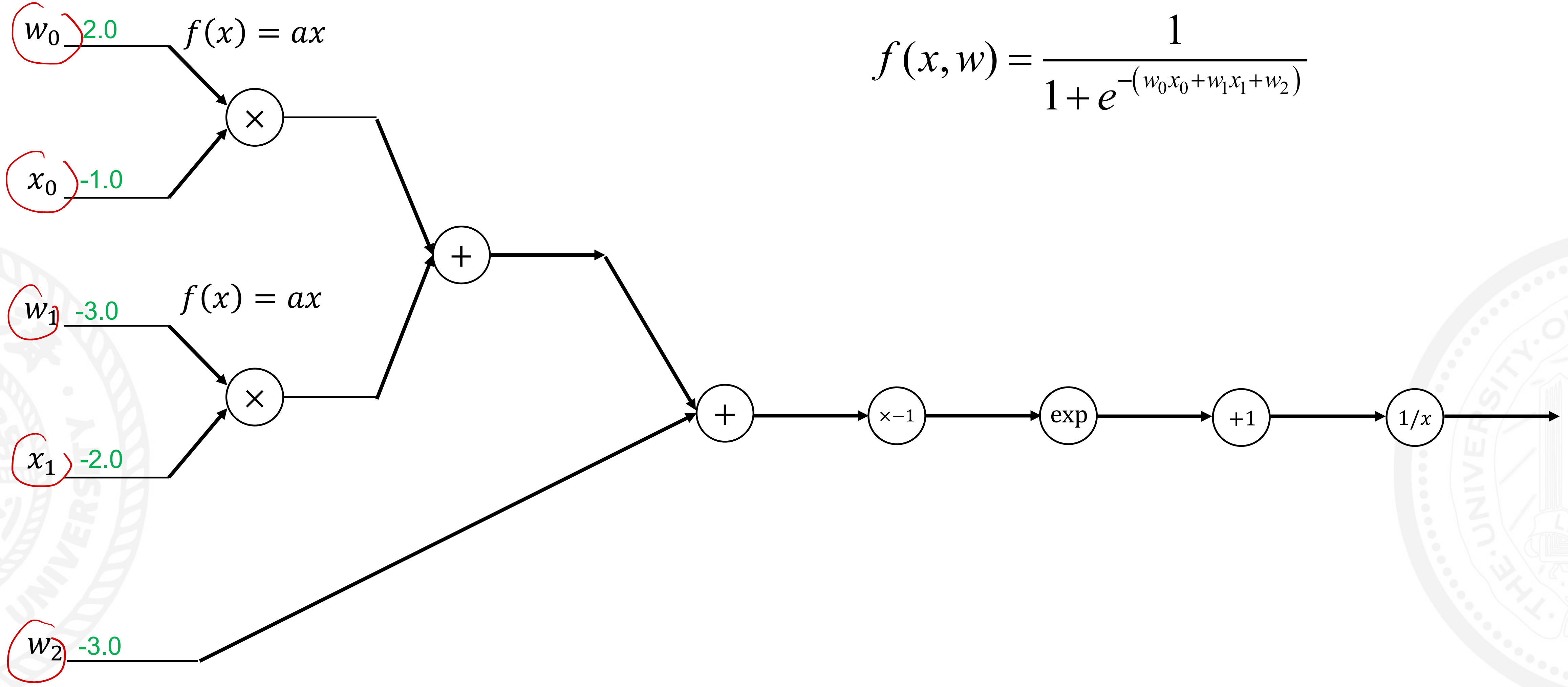
Backpropagation

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



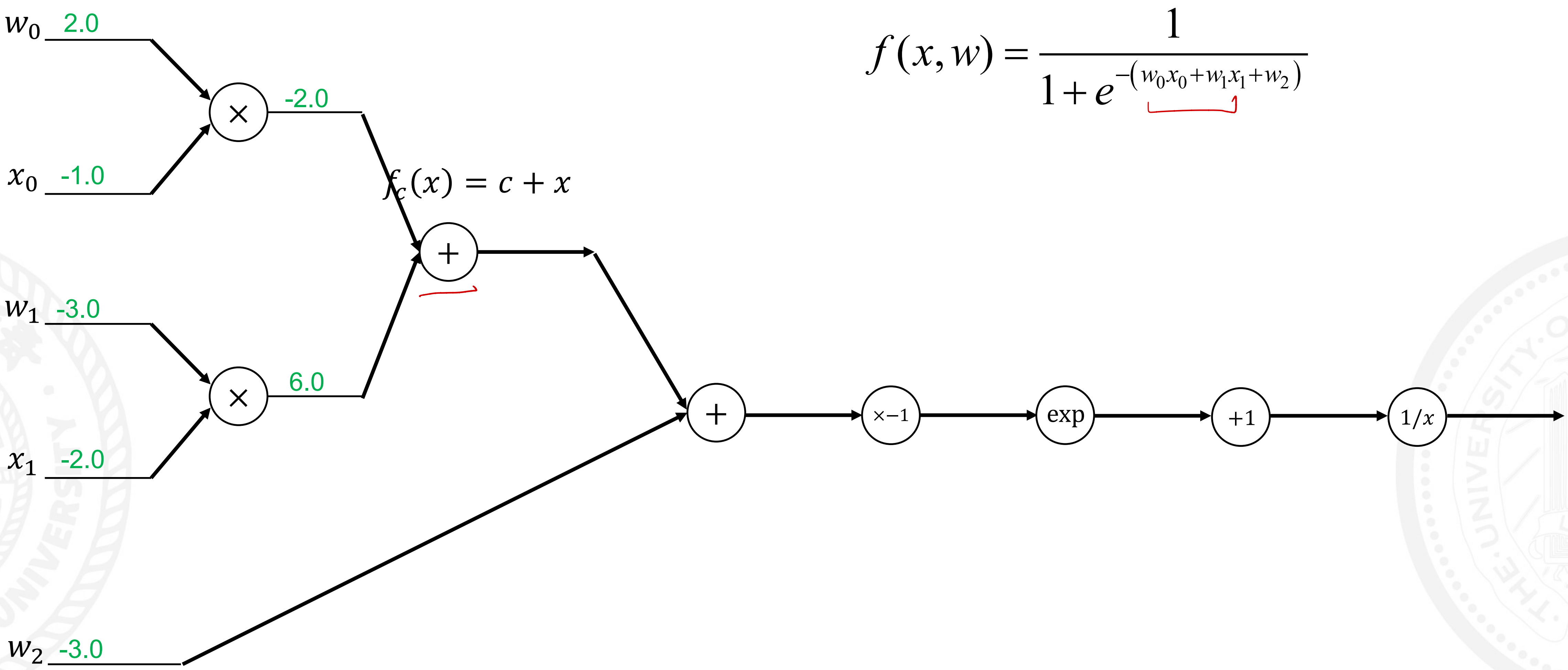
Backpropagation

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

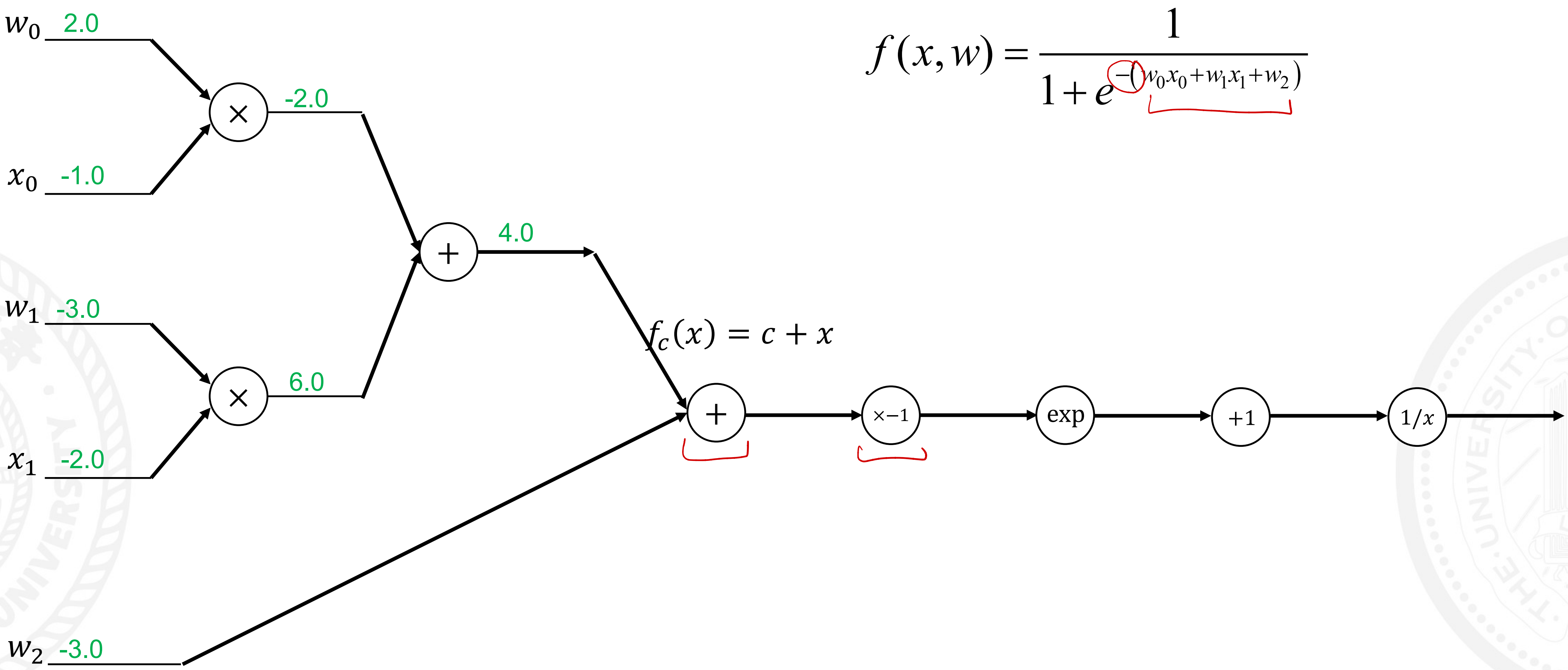


Backpropagation

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

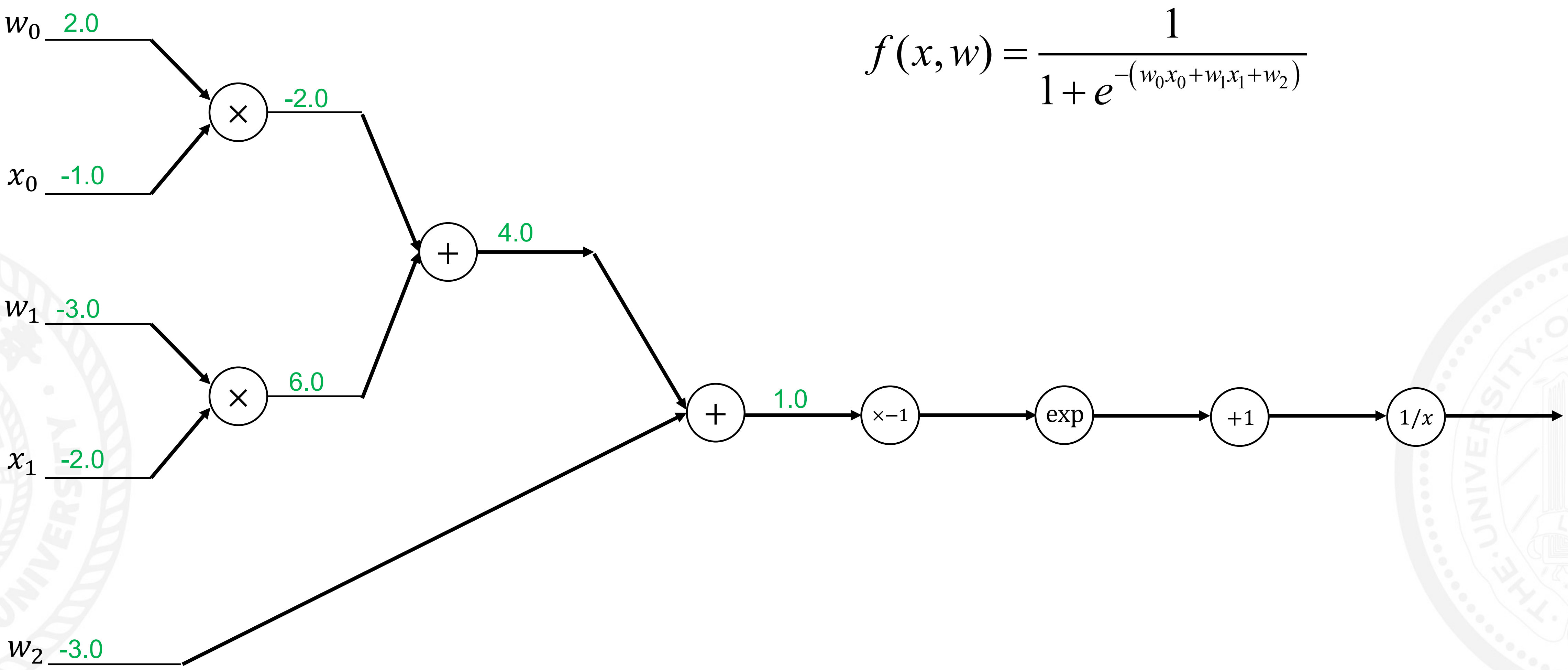


$$f(x,w) = \frac{1}{1+e^{-(\underbrace{w_0x_0+w_1x_1+w_2}_{})}}$$



Backpropagation

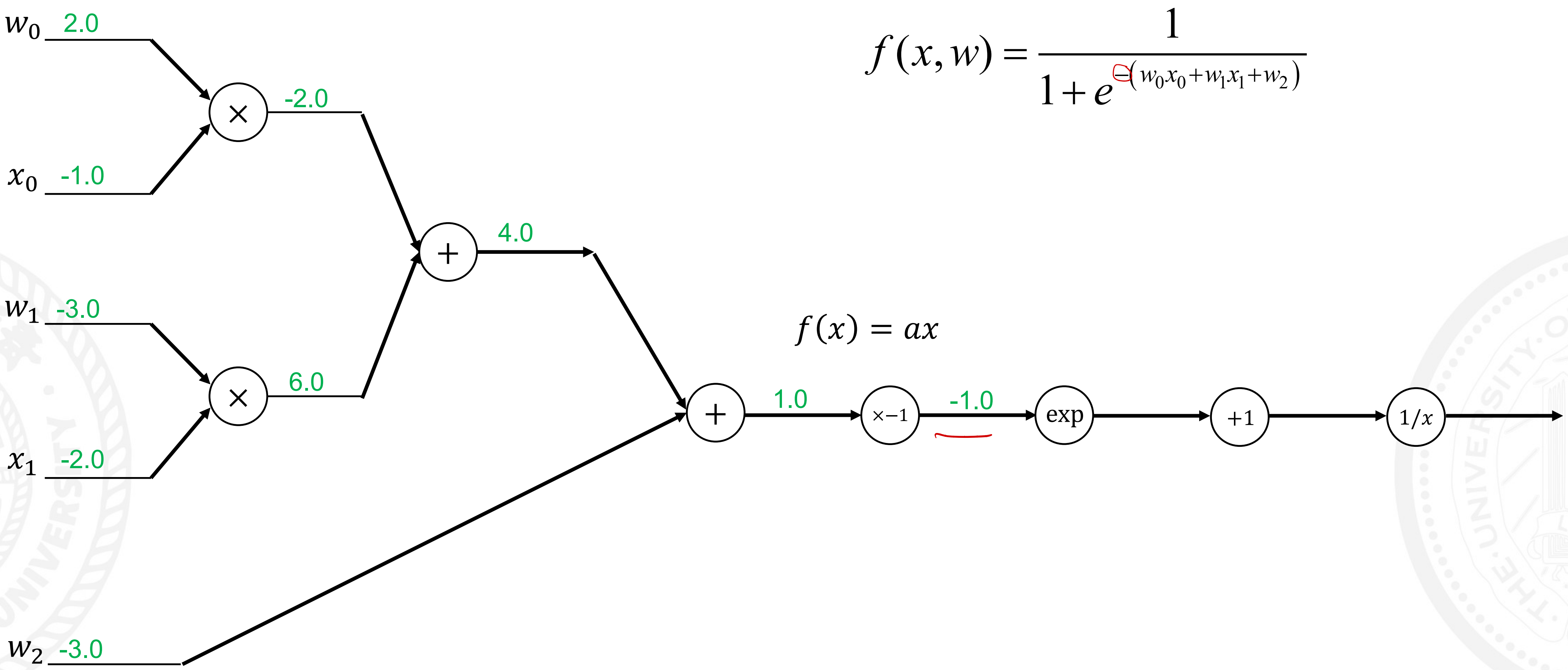
$$f(x,w)=\frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$



Backpropagation

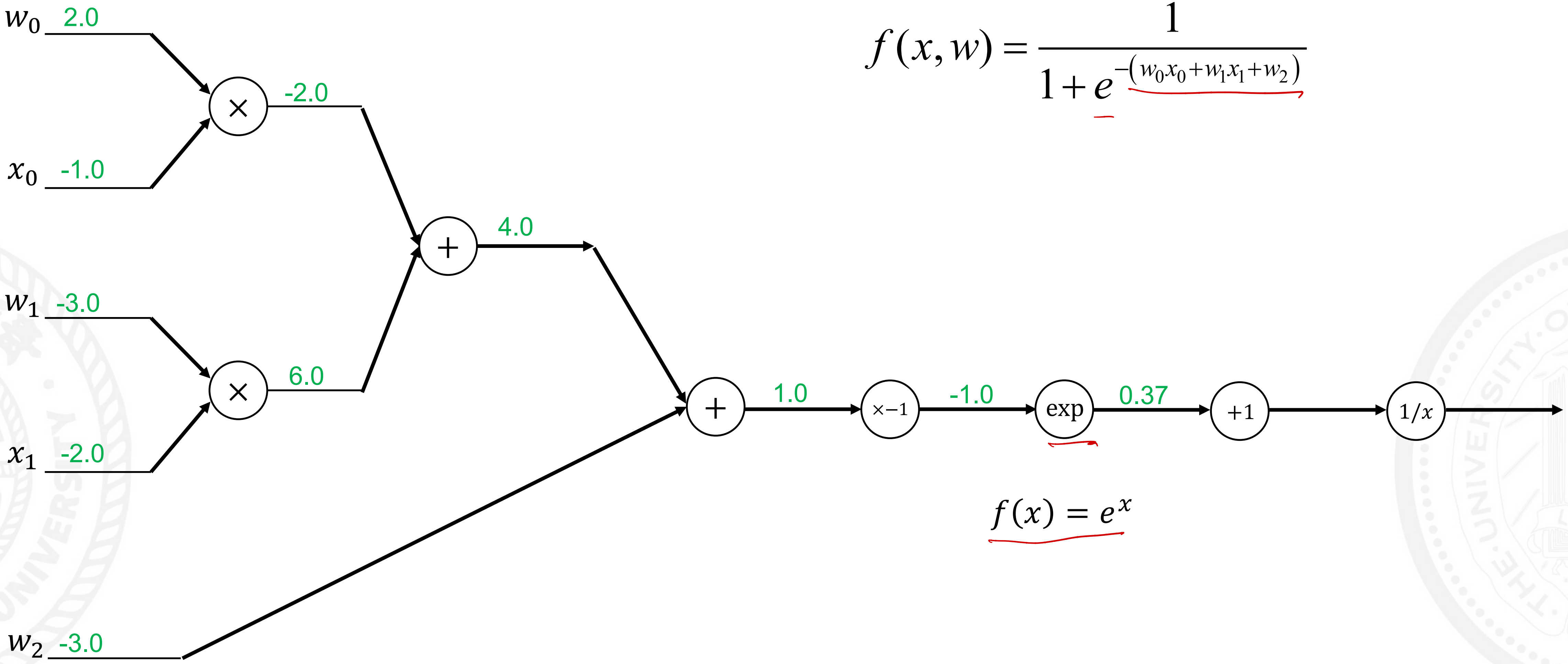
$$f(x,w) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

$$f(x) = ax$$



Backpropagation

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

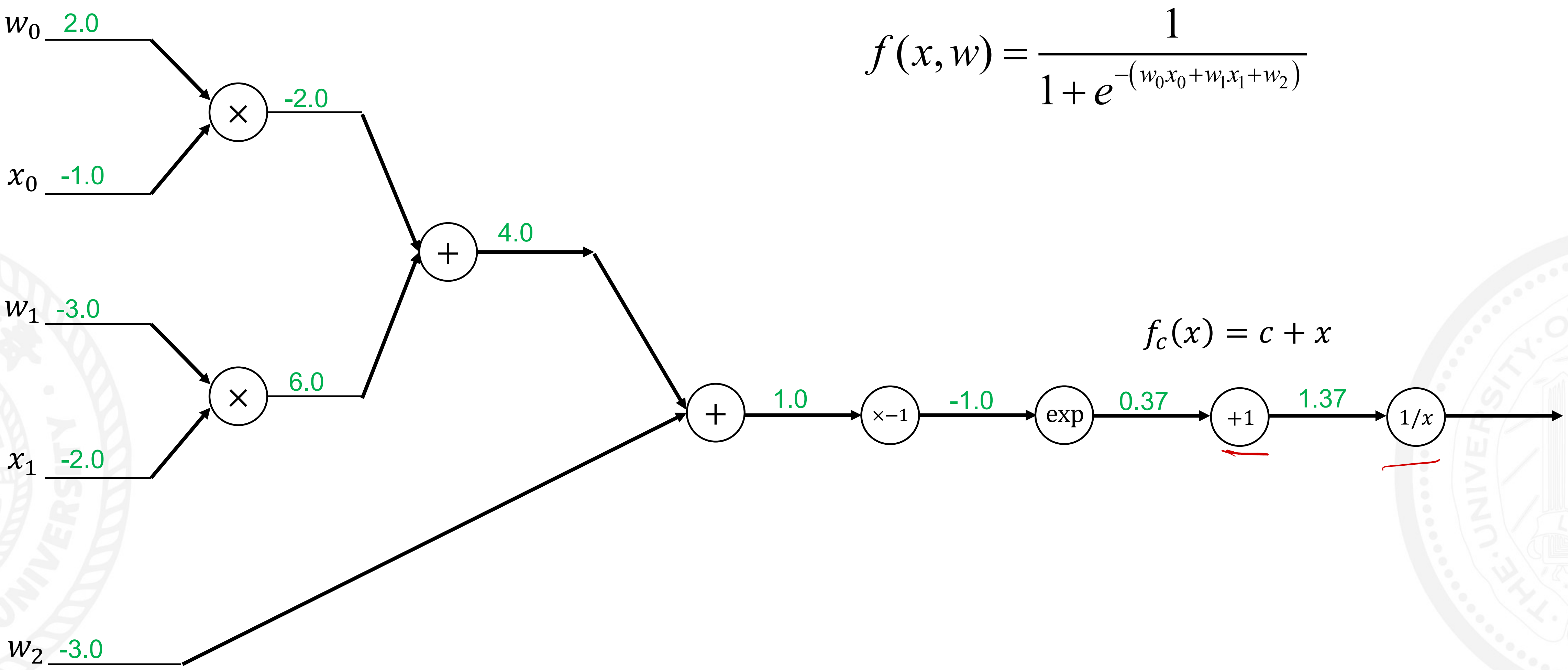


$$f(x) = e^x$$

Backpropagation

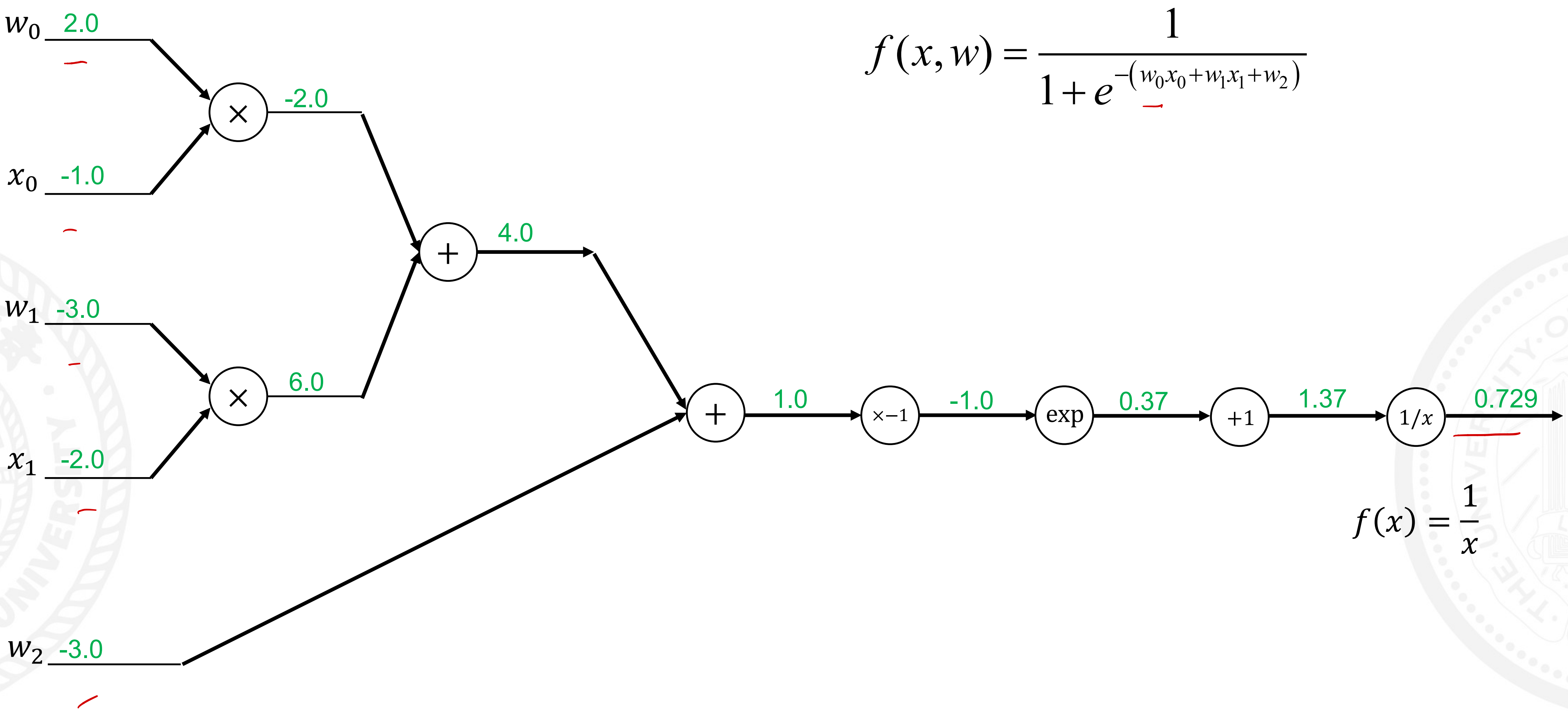
$$f(x,w) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

$$f_c(x) = c + x$$



Backpropagation

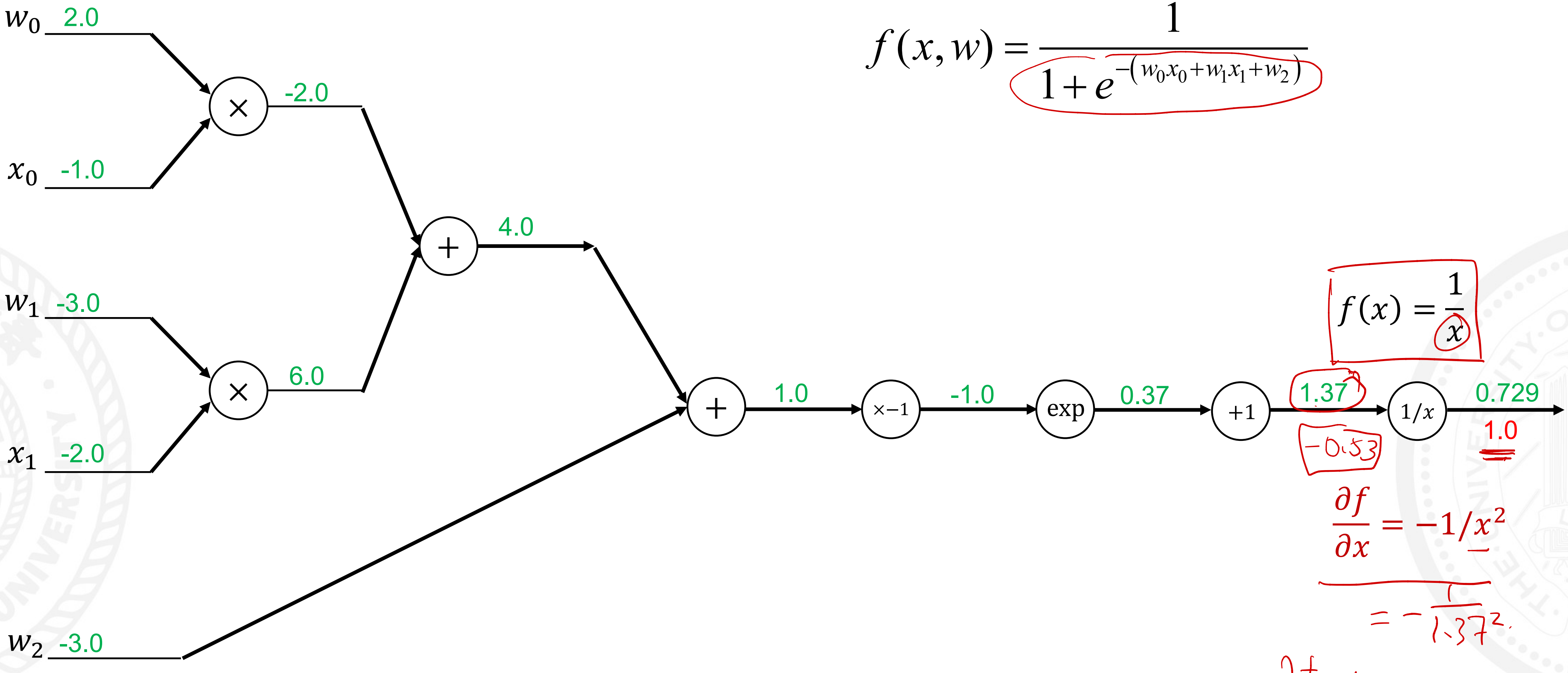
$$f(x,w) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$



$$f(x) = \frac{1}{x}$$

Backpropagation

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



$$f(x) = \frac{1}{x}$$

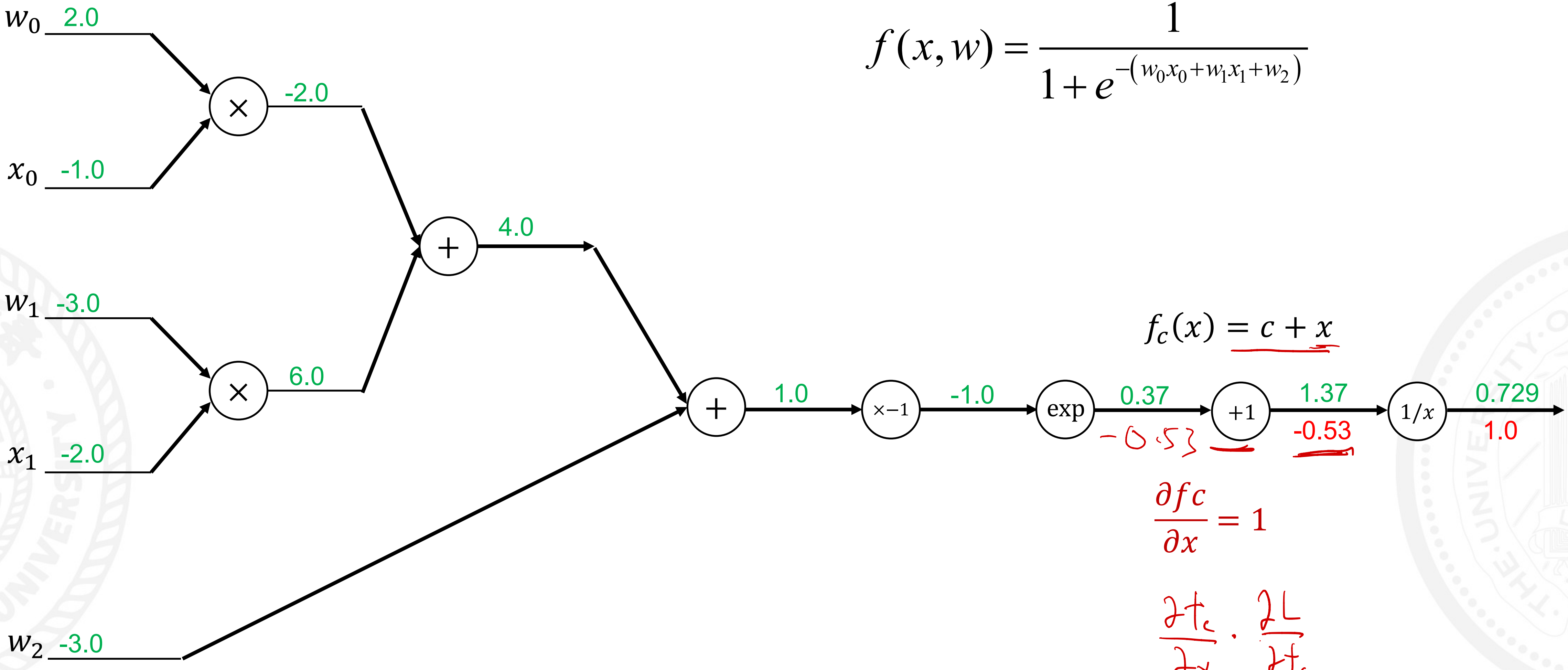
$$\frac{\partial f}{\partial x} = -1/x^2$$

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

$$\frac{\partial f}{\partial x} \cdot 1$$

Backpropagation

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



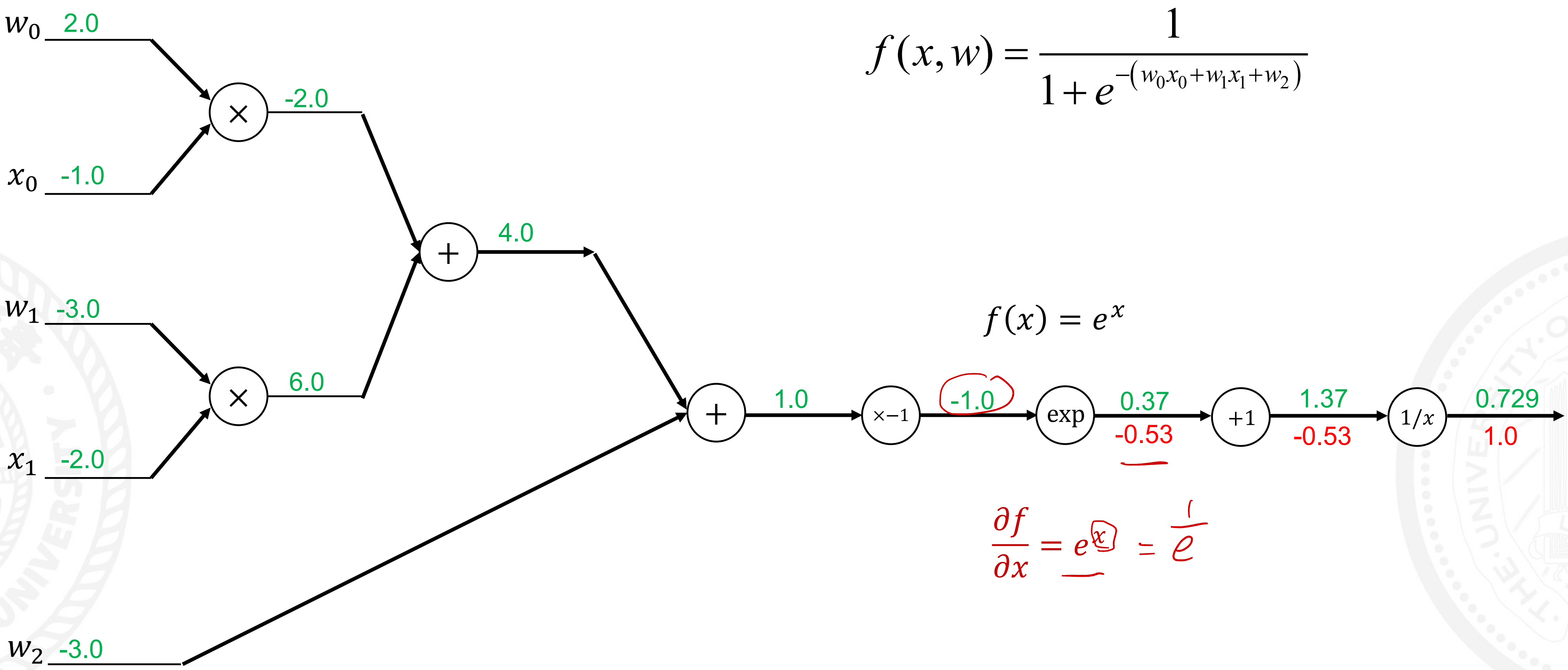
$$f_c(x) = c + x$$

$$\frac{\partial fc}{\partial x} = 1$$

$$\underbrace{\frac{\partial f_c}{\partial x}}_{1} \cdot \underbrace{\frac{\partial L}{\partial f_c}}_{-0.53}$$

Backpropagation

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



$$f(x) = e^x$$

$$\frac{\partial f}{\partial x} = e^{x} = \frac{1}{e}$$

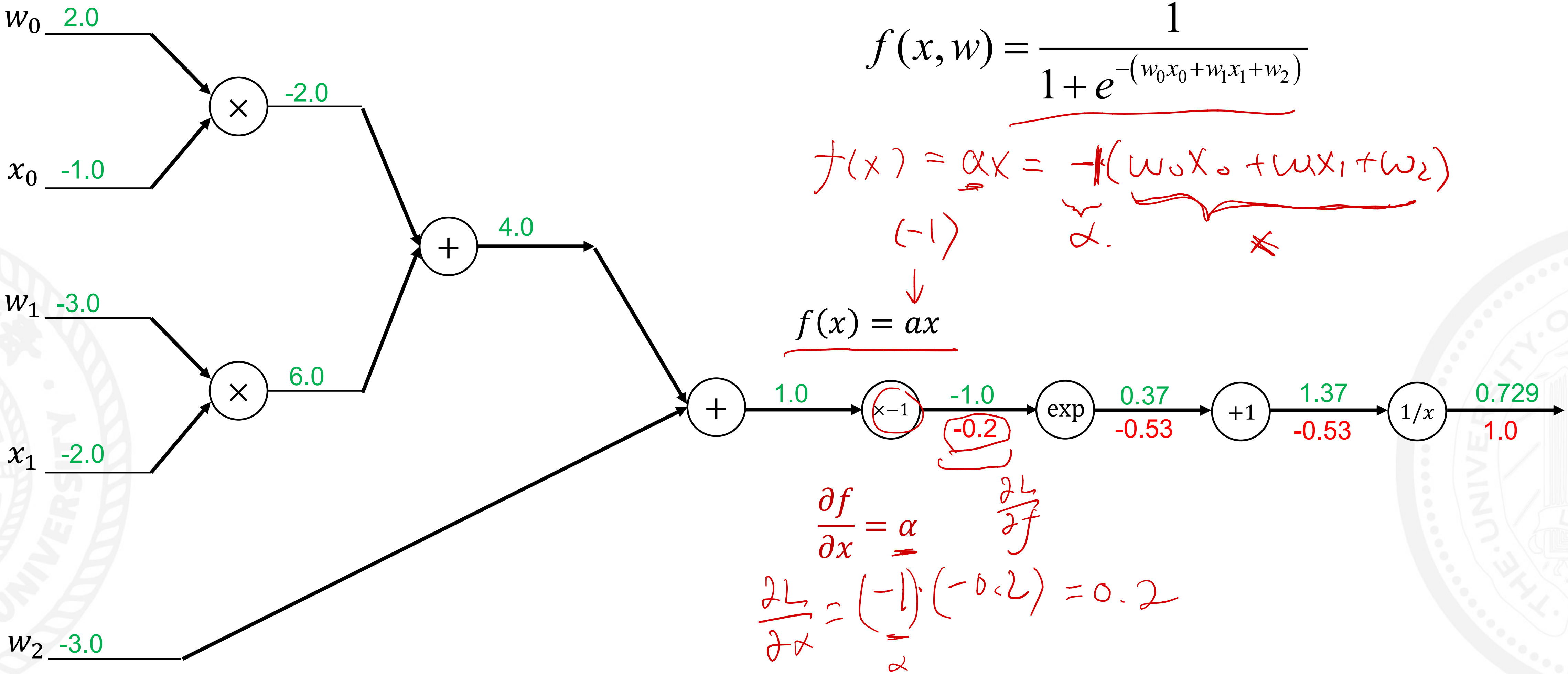
Backpropagation

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

$$f(x) = ax = \underbrace{-1}_{\alpha}\underbrace{(w_0x_0+w_1x_1+w_2)}_{x}$$

(-1)

$$f(x) = ax$$

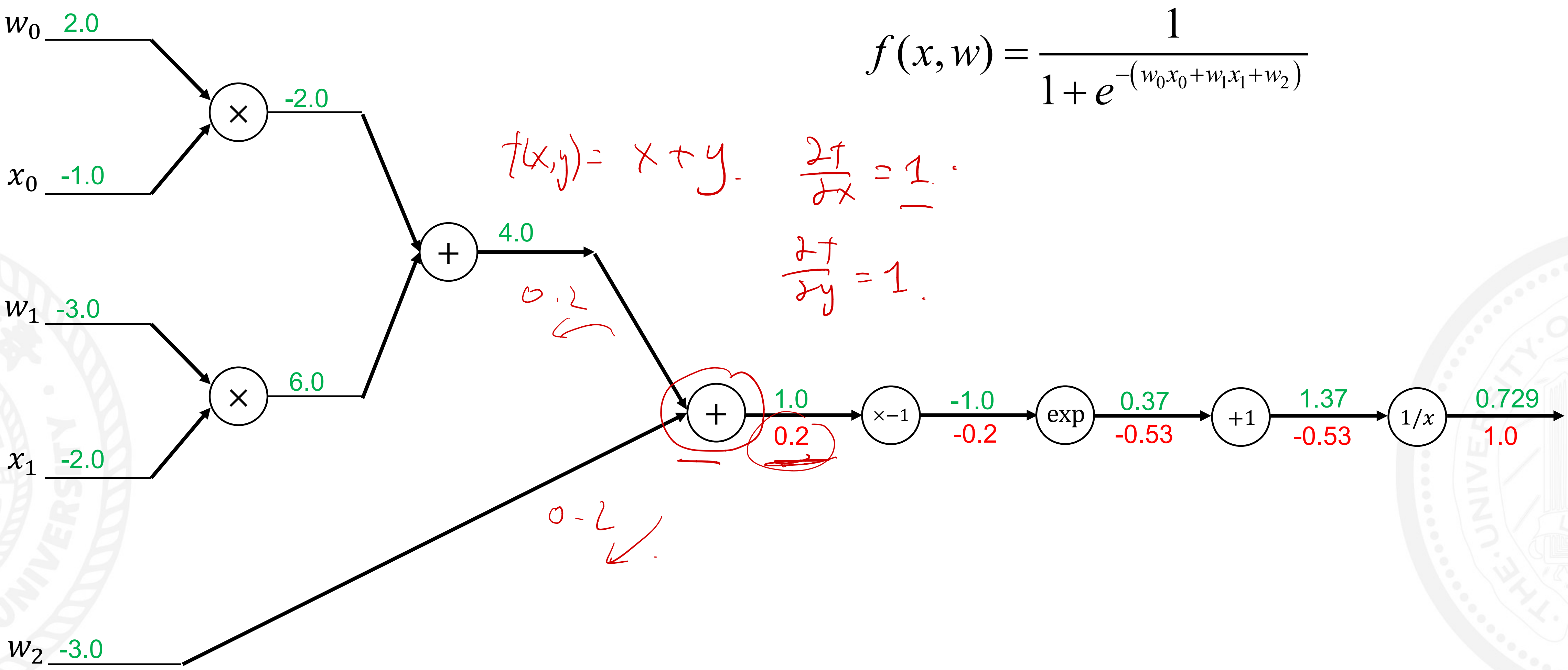


$$\frac{\partial f}{\partial x} = \alpha \qquad \frac{\partial L}{\partial f}$$

$$\frac{\partial L}{\partial \alpha} = \underbrace{(-1)}_{\alpha}(-0.2) = 0.2$$

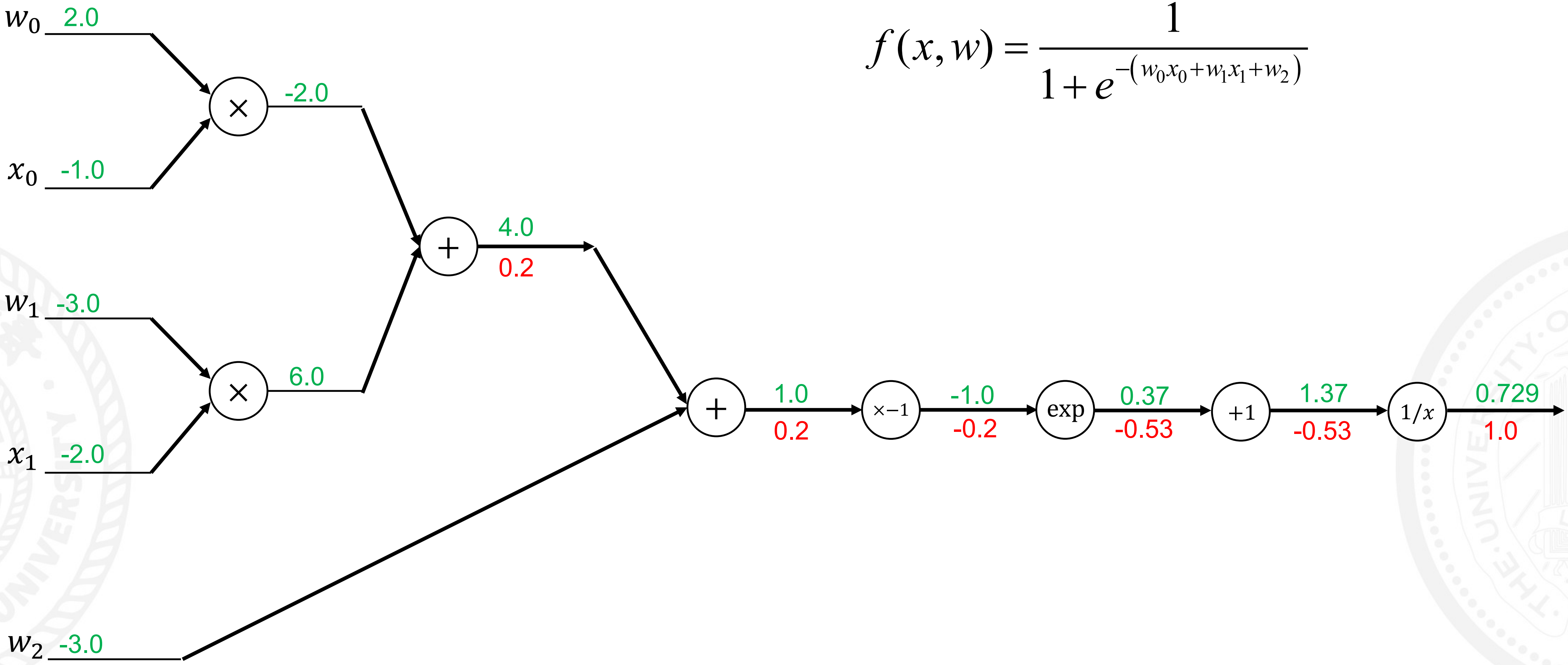
Backprogation

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



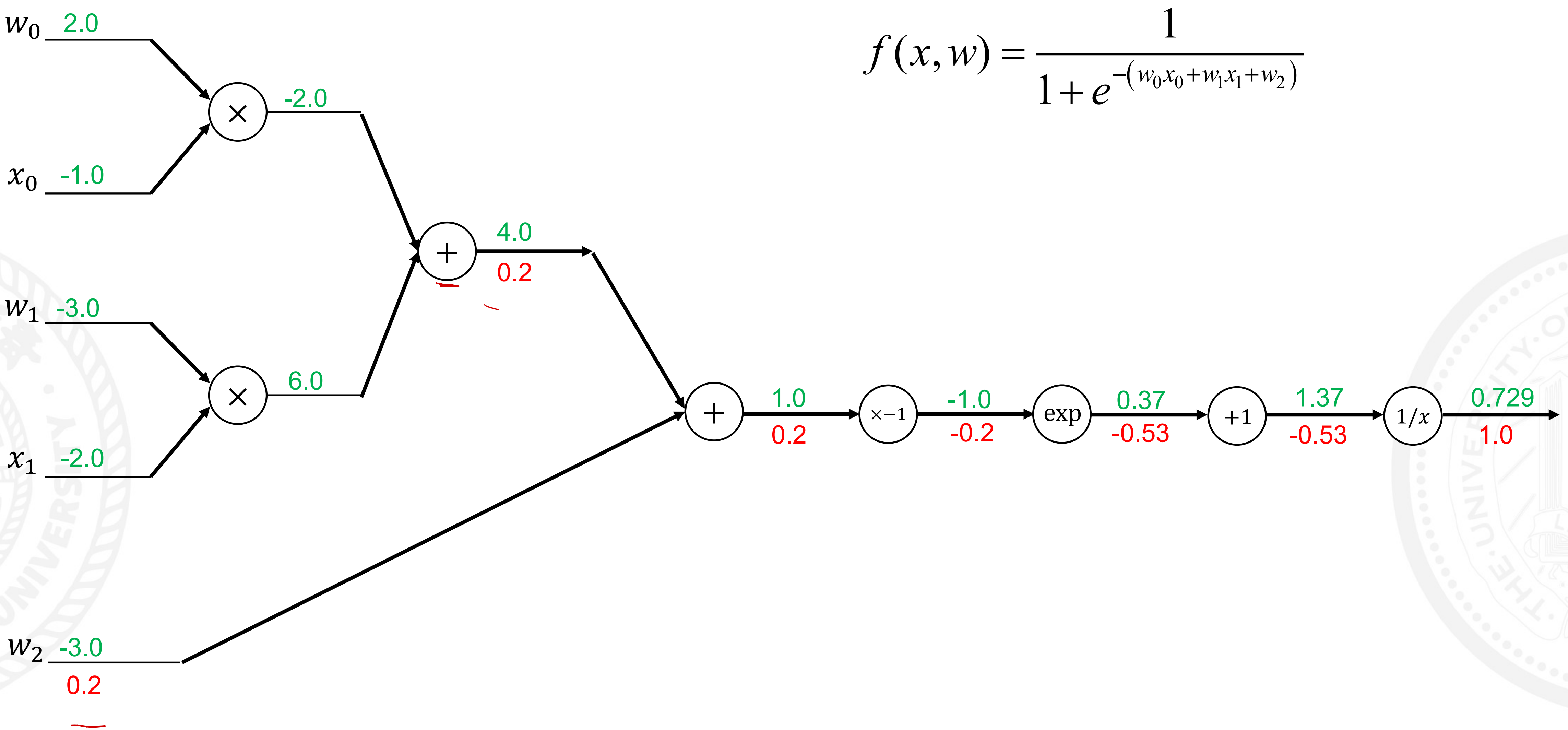
Backpropagation

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



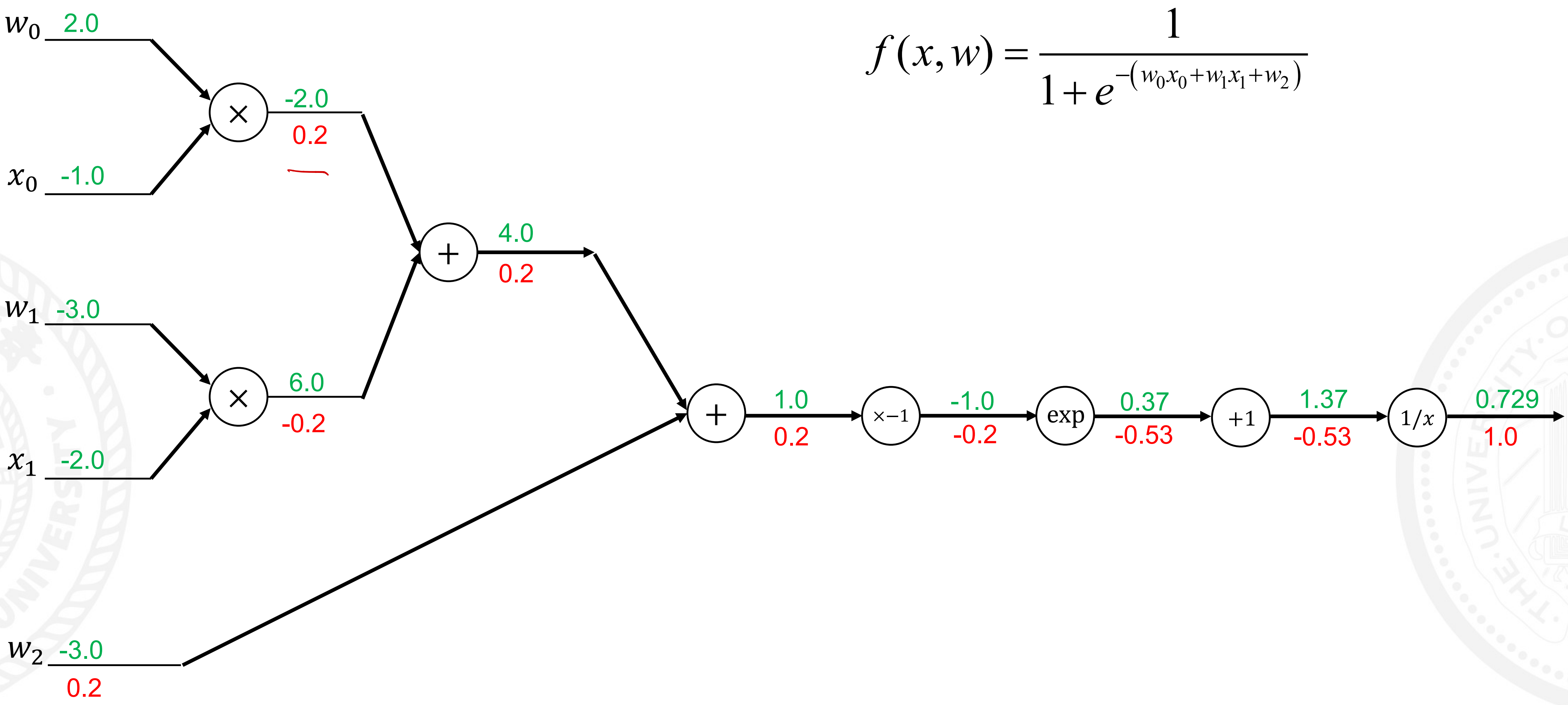
Backpropagation

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



Backprop

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



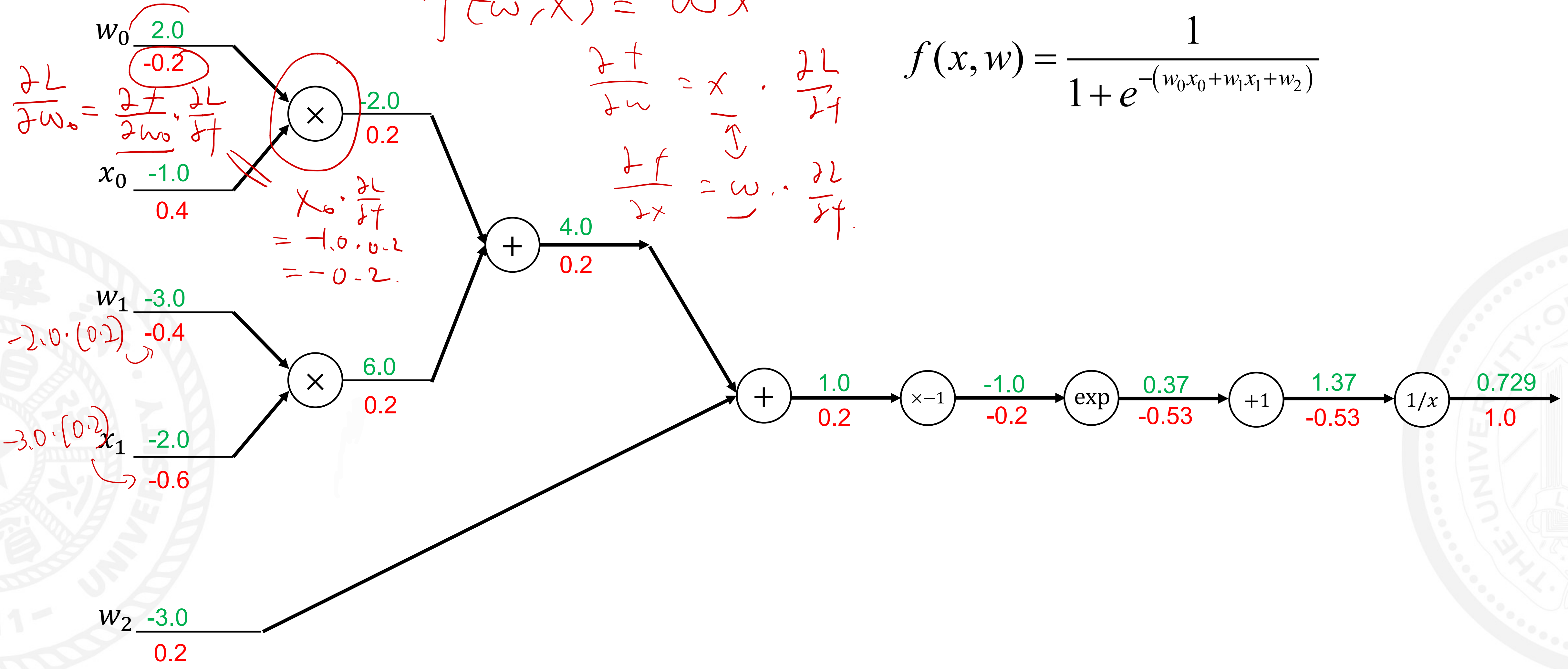
Backpropagation

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

$f(w,x) = wx$

$\frac{\partial f}{\partial w} = x \cdot \frac{\partial L}{\partial f}$

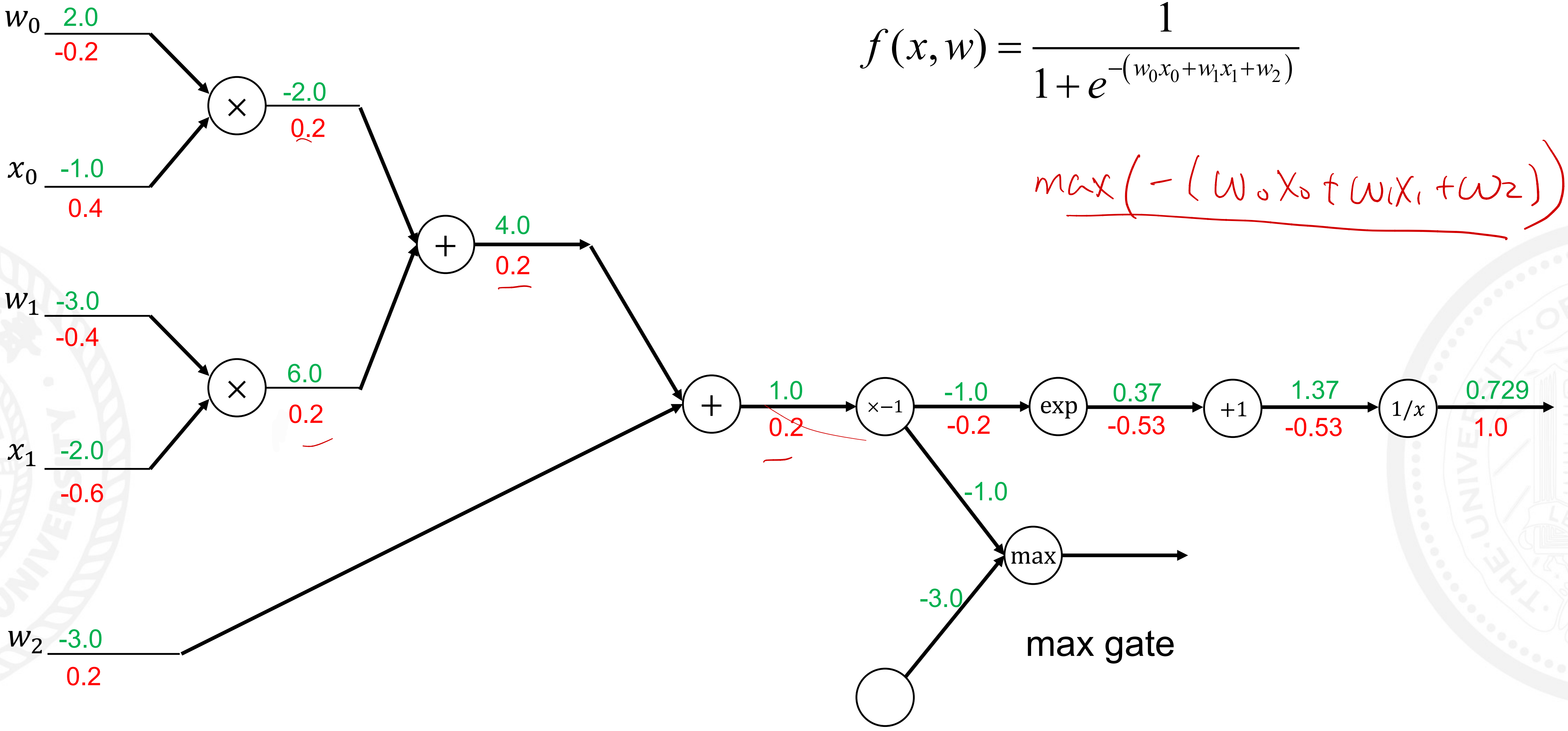
$\frac{\partial f}{\partial x} = w \cdot \frac{\partial L}{\partial f}$



Backpropagation

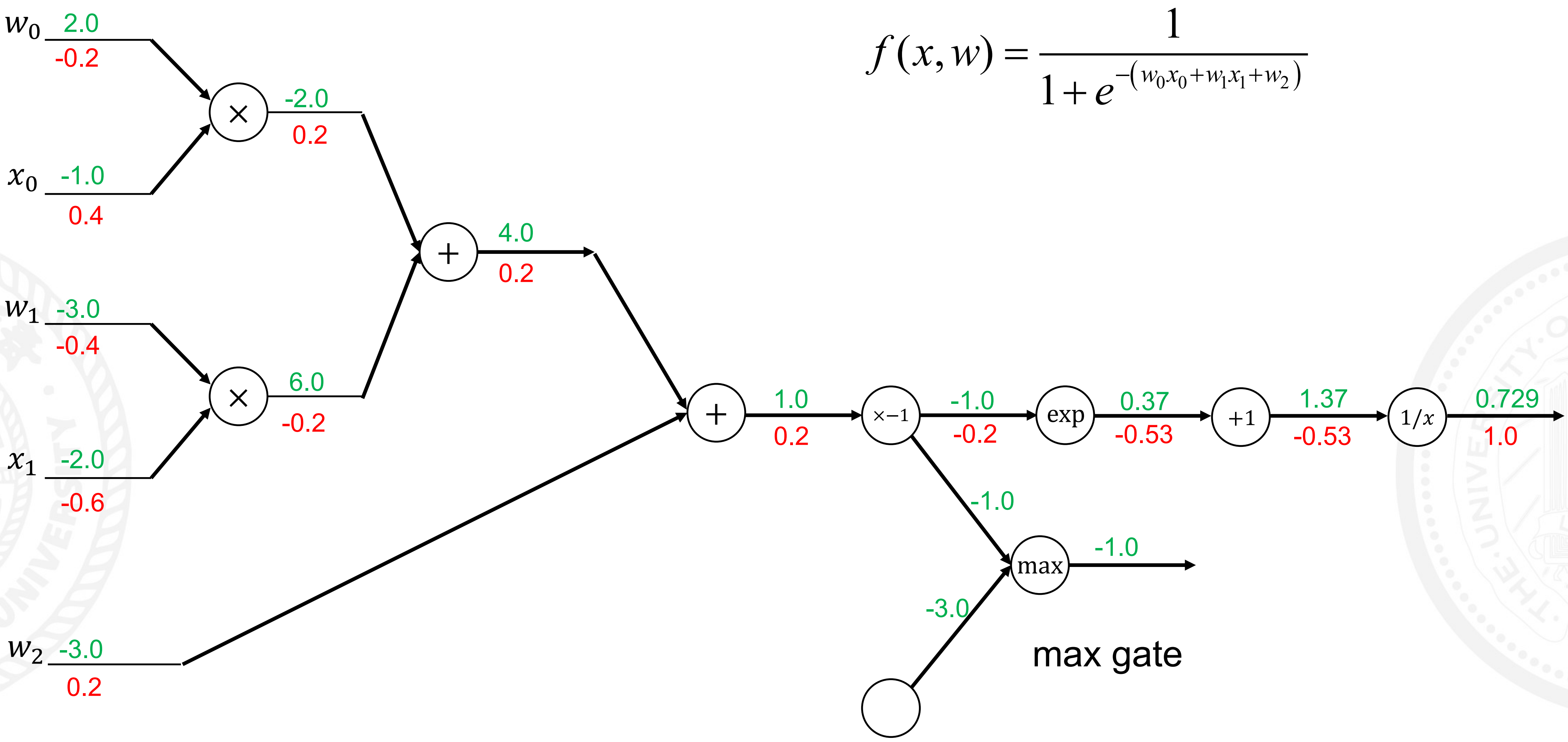
$$f(x,w)=\frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

max(−(w₀x₀ + w₁x₁ + w₂))



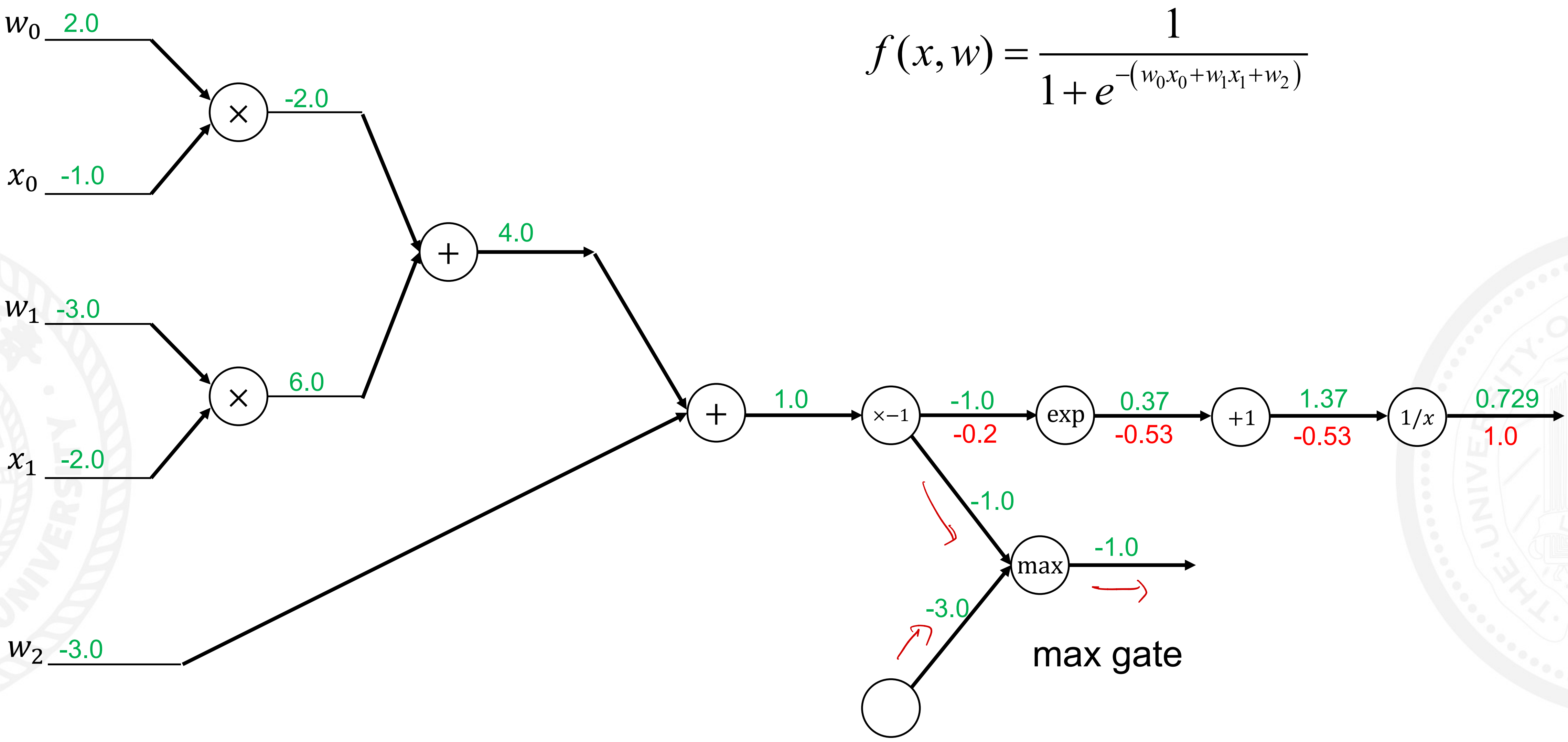
Backpropagation

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



Backpropagation

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



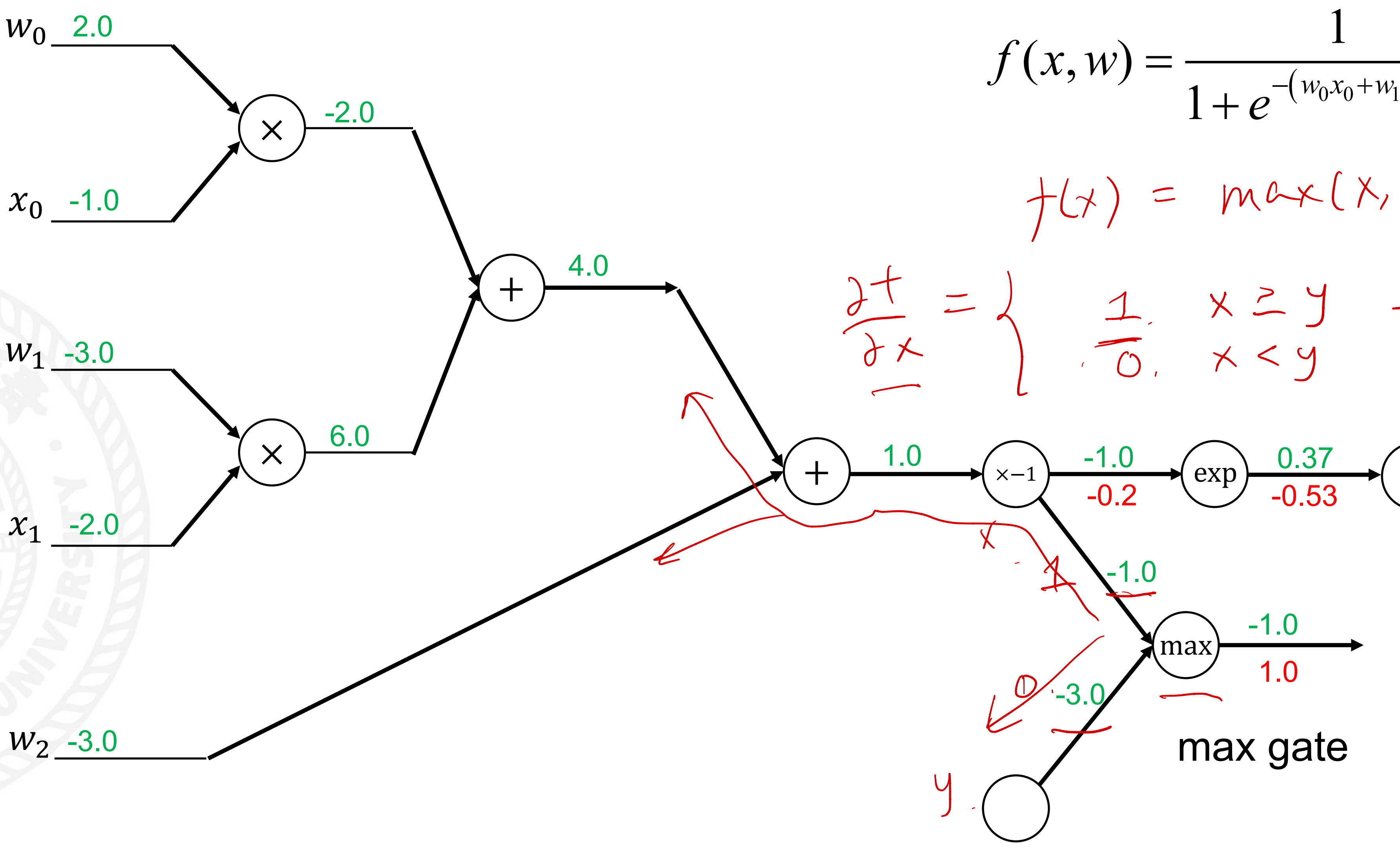
max gate

Backpropagation

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

$$f(x) = \max(x,y) = \begin{cases} x & x \geq y \\ y & x < y \end{cases}$$

$$\frac{\partial f}{\partial x} = \begin{cases} 1 & x \geq y \\ 0 & x < y \end{cases} \qquad \frac{\partial f}{\partial y} = \begin{cases} 0 & x \geq y \\ 1 & x < y \end{cases}$$

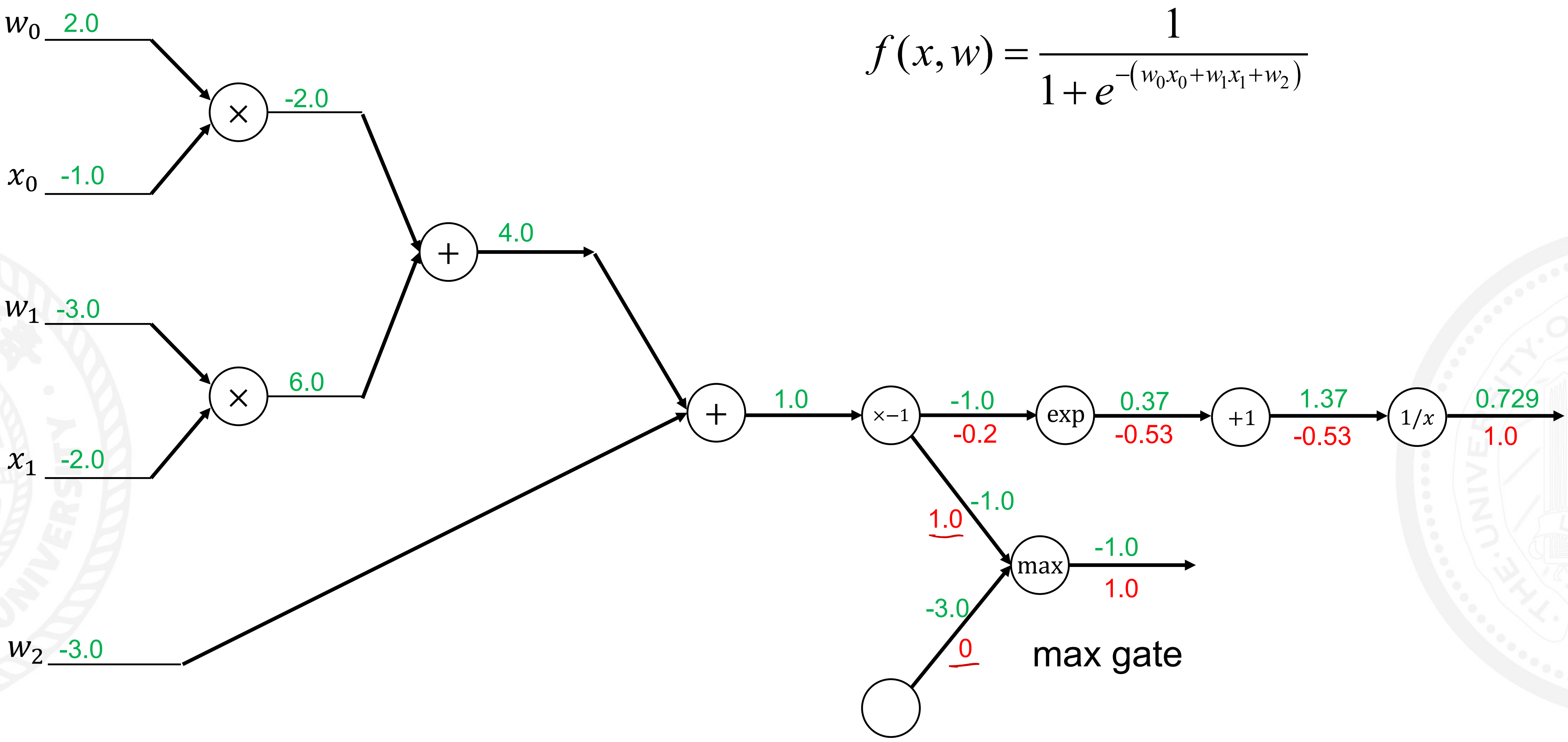


$$\frac{\partial f}{\partial x} = 1$$

$$\frac{\partial L}{\partial x} = \frac{\partial f}{\partial x} \frac{\partial L}{\partial f} = 1 \cdot 1$$

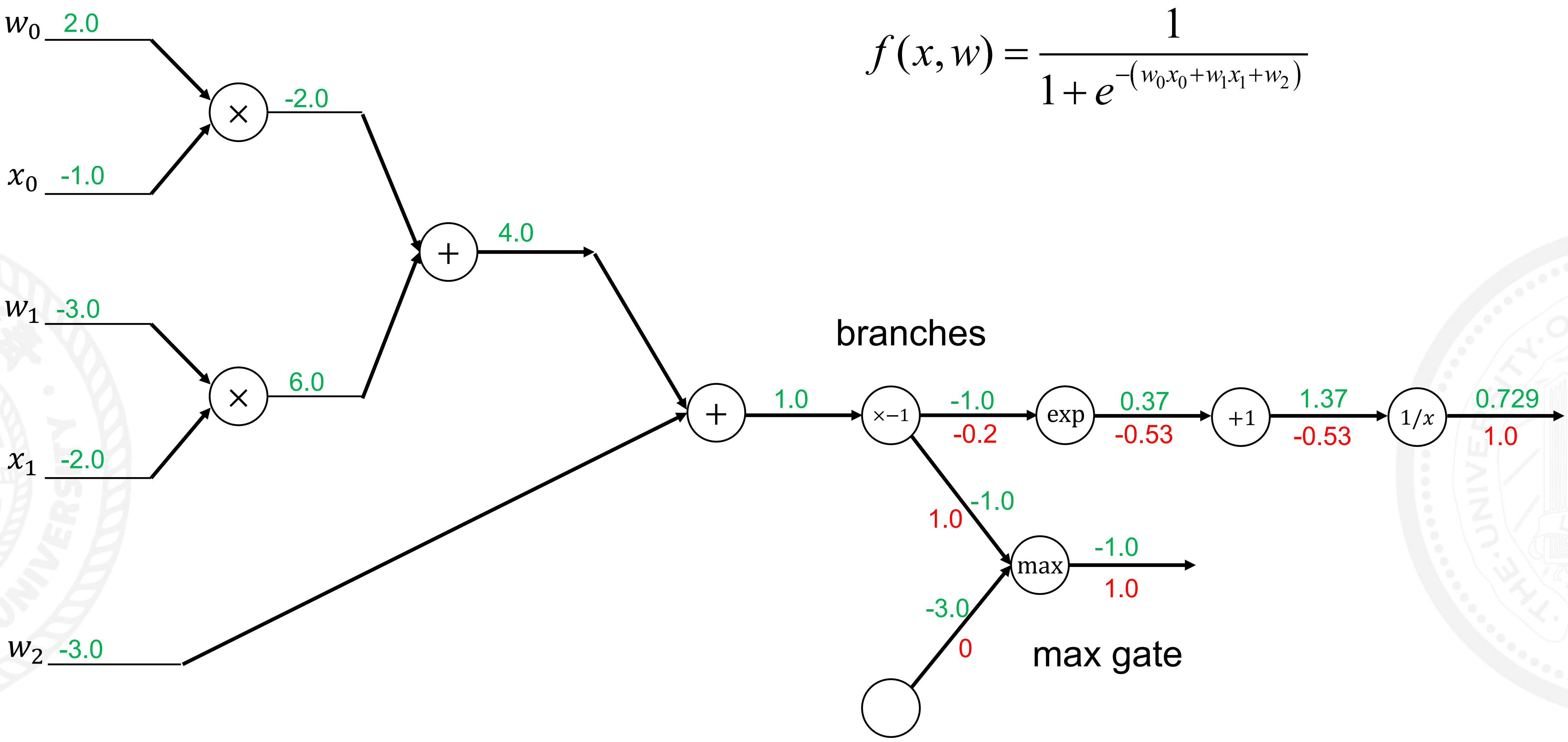
Backpropagation

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



Backpropagation

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



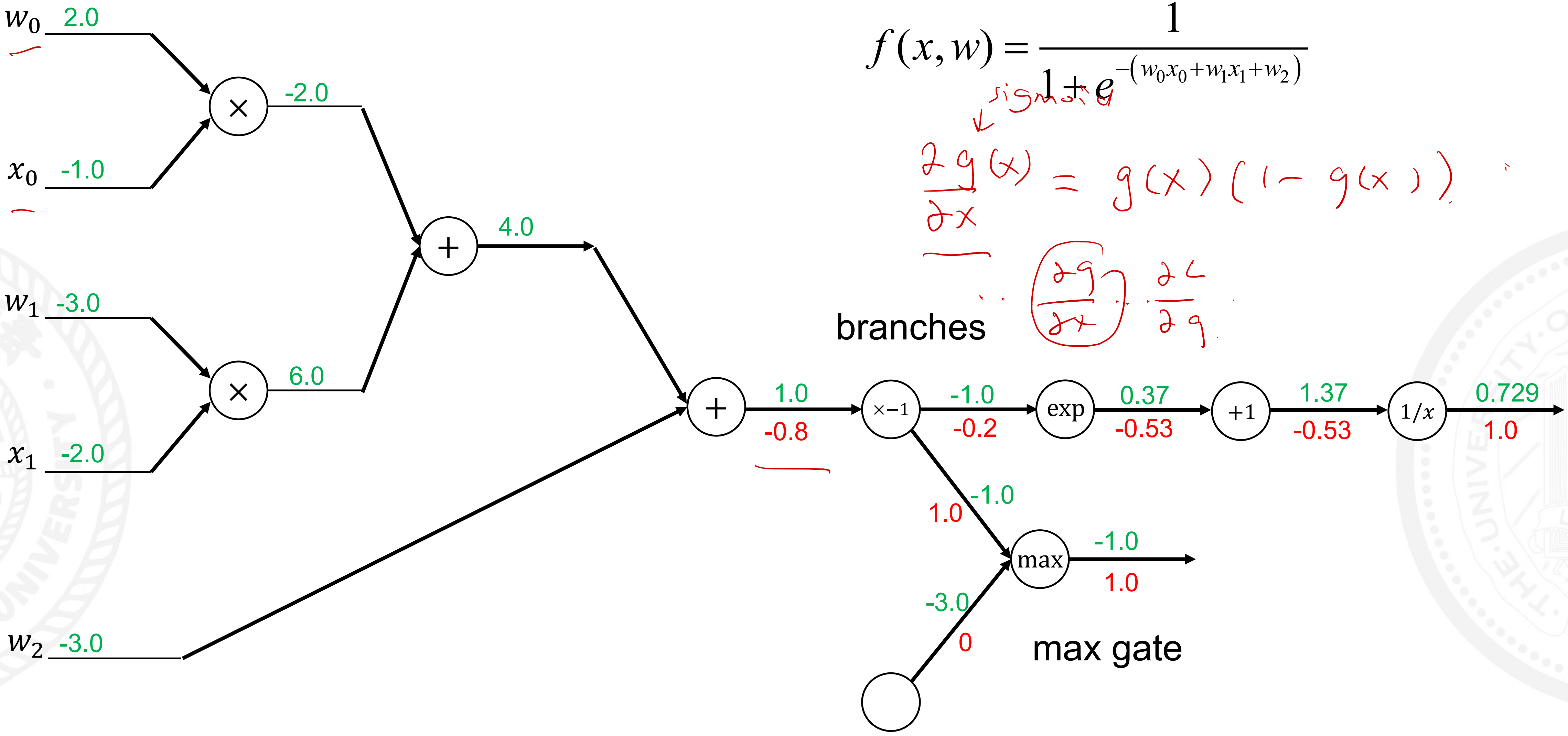
Backpropagation

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

sigmoid

$$\frac{\partial g(x)}{\partial x} = g(x)\,(1-g(x))$$

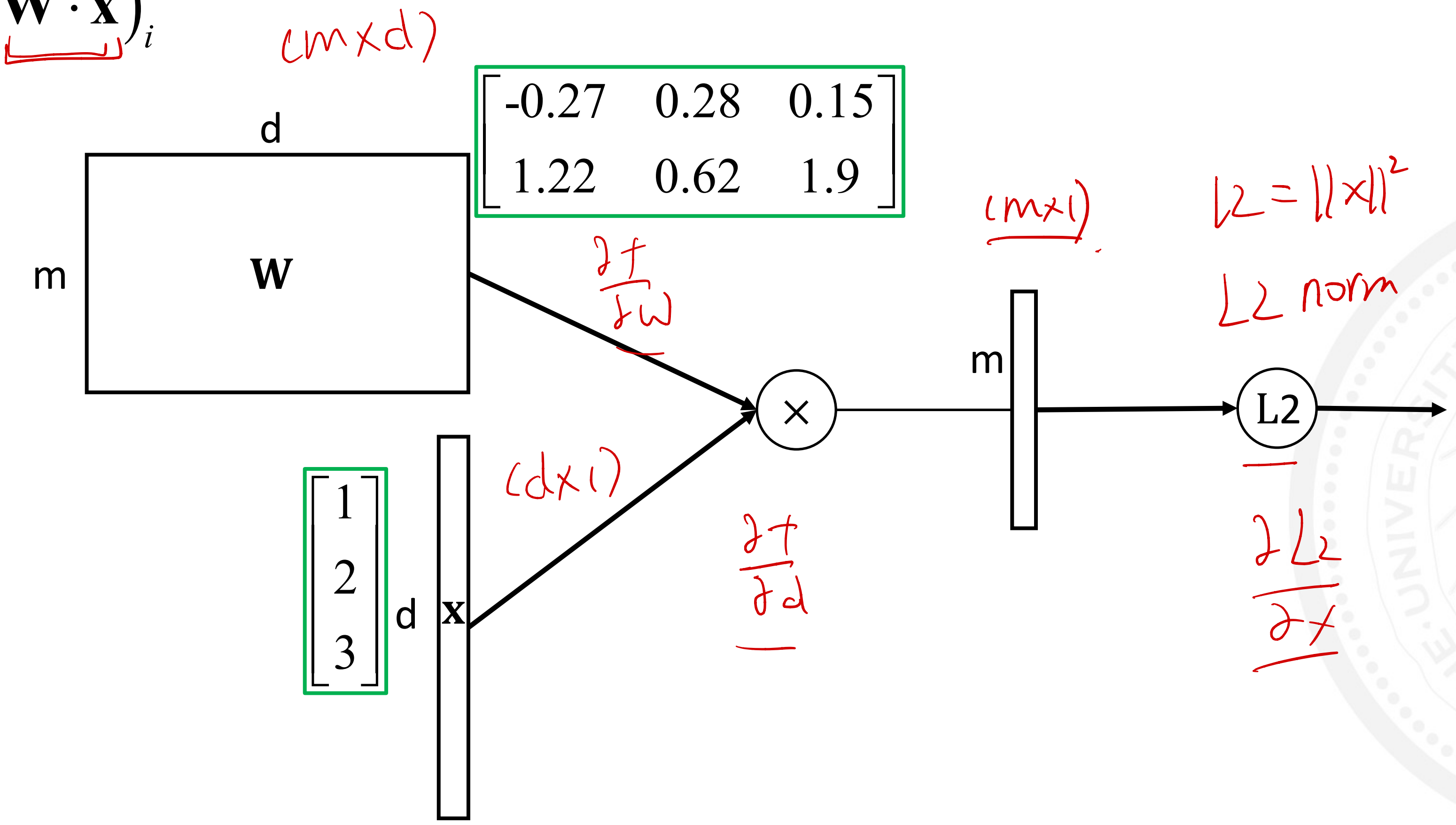
$$\frac{\partial g}{\partial x} \cdot \frac{\partial L}{\partial g}$$



Backpropagation

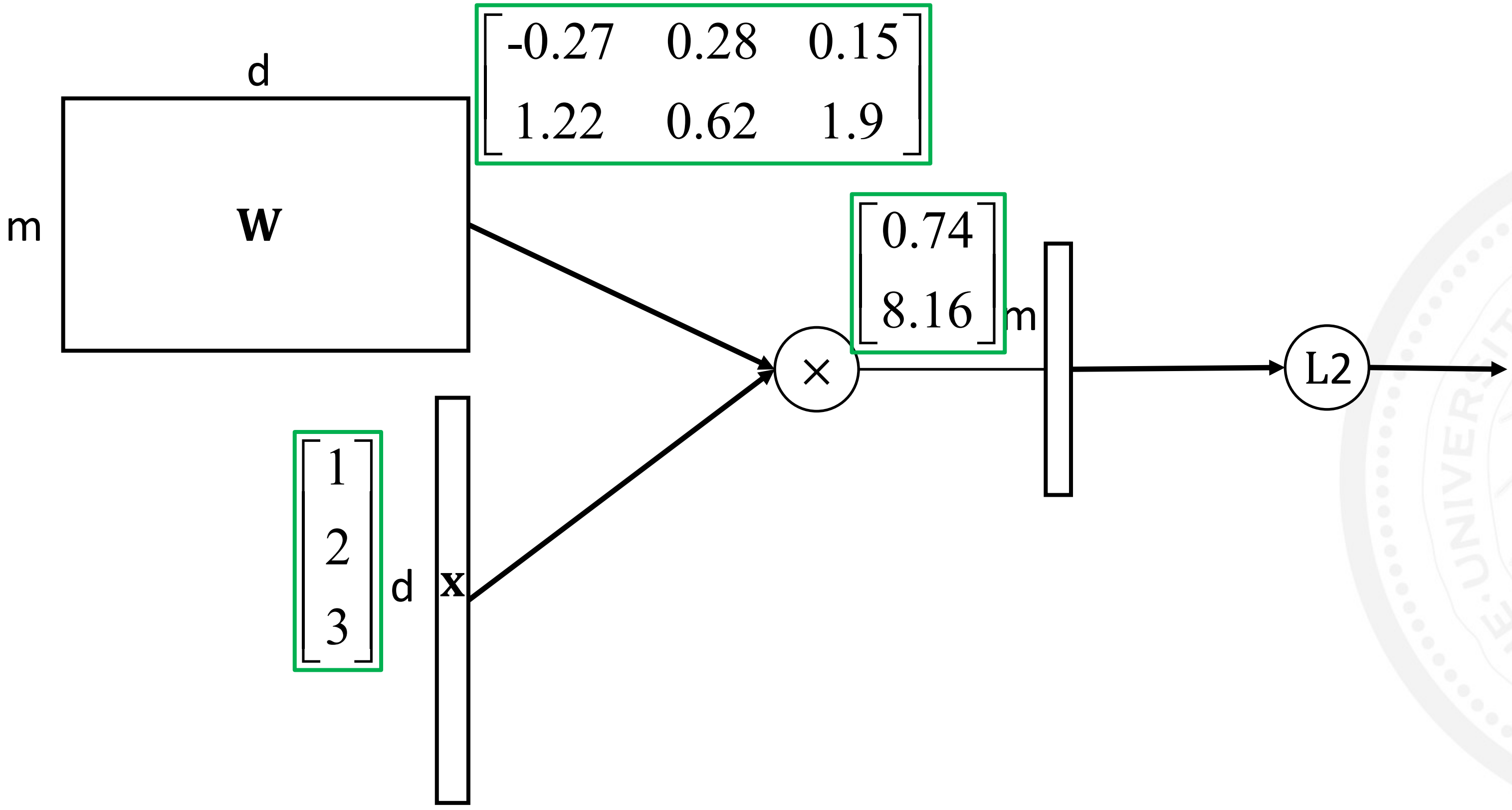
Vectorized example

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



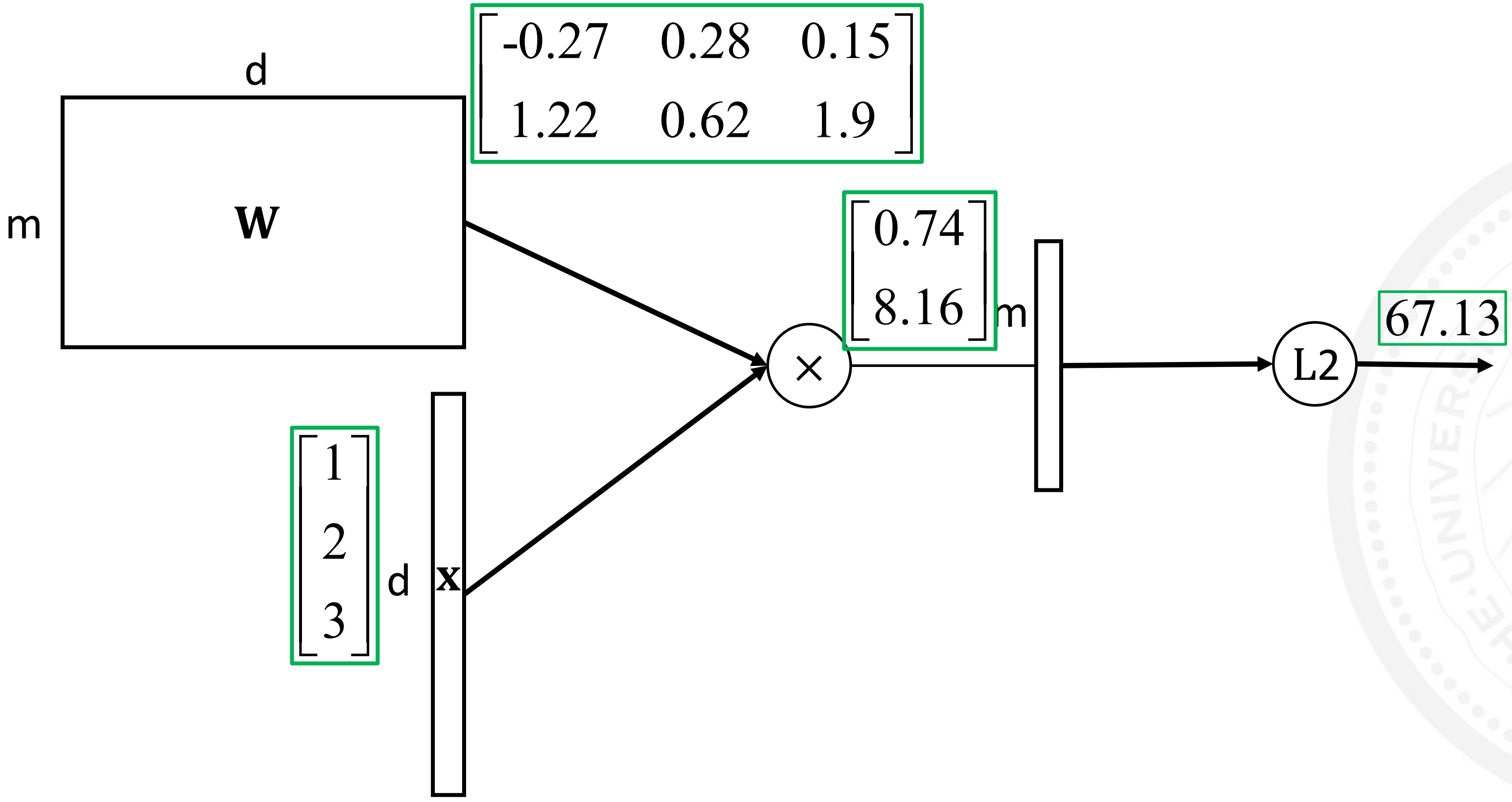
Vectorized example

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



Vectorized example

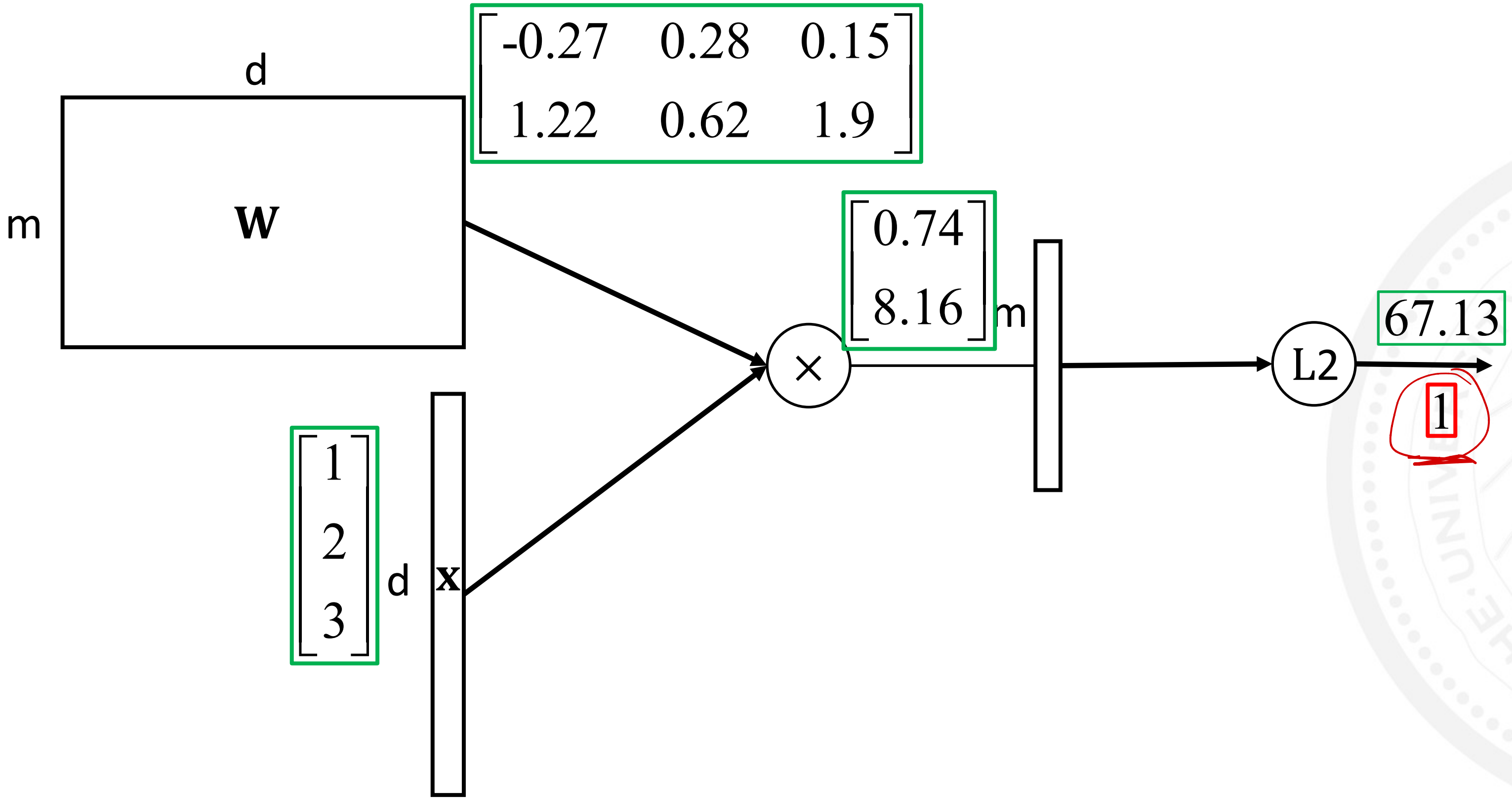
$$f(\mathbf{x}, \mathbf{W}) = \|\mathbf{W} \cdot \mathbf{x}\|^2 = \sum_{i=1}^{n} (\mathbf{W} \cdot \mathbf{x})_i^2$$



Backpropagation

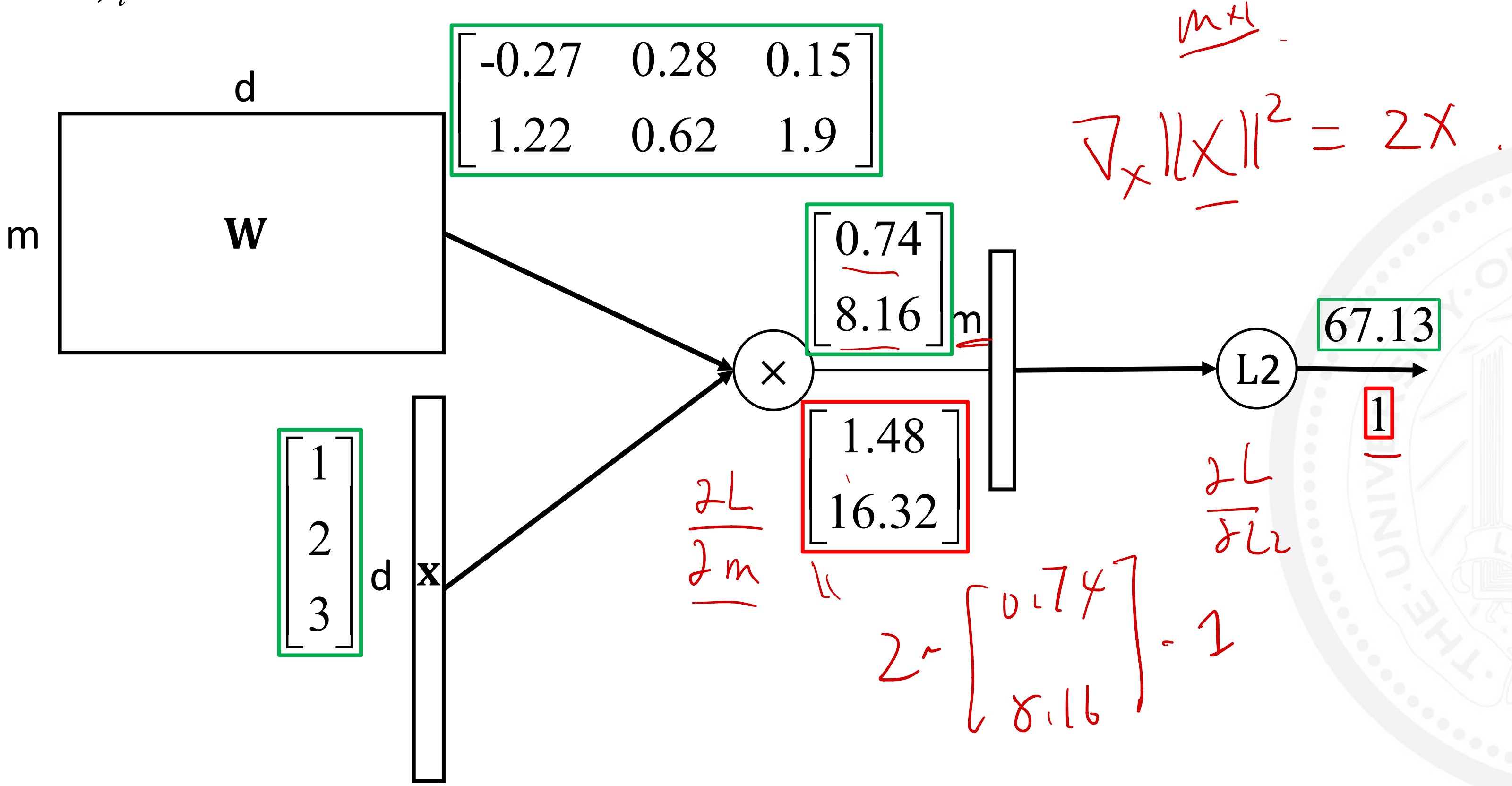
Vectorized example

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



Vectorized example

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



Backpropagation

Vectorized example

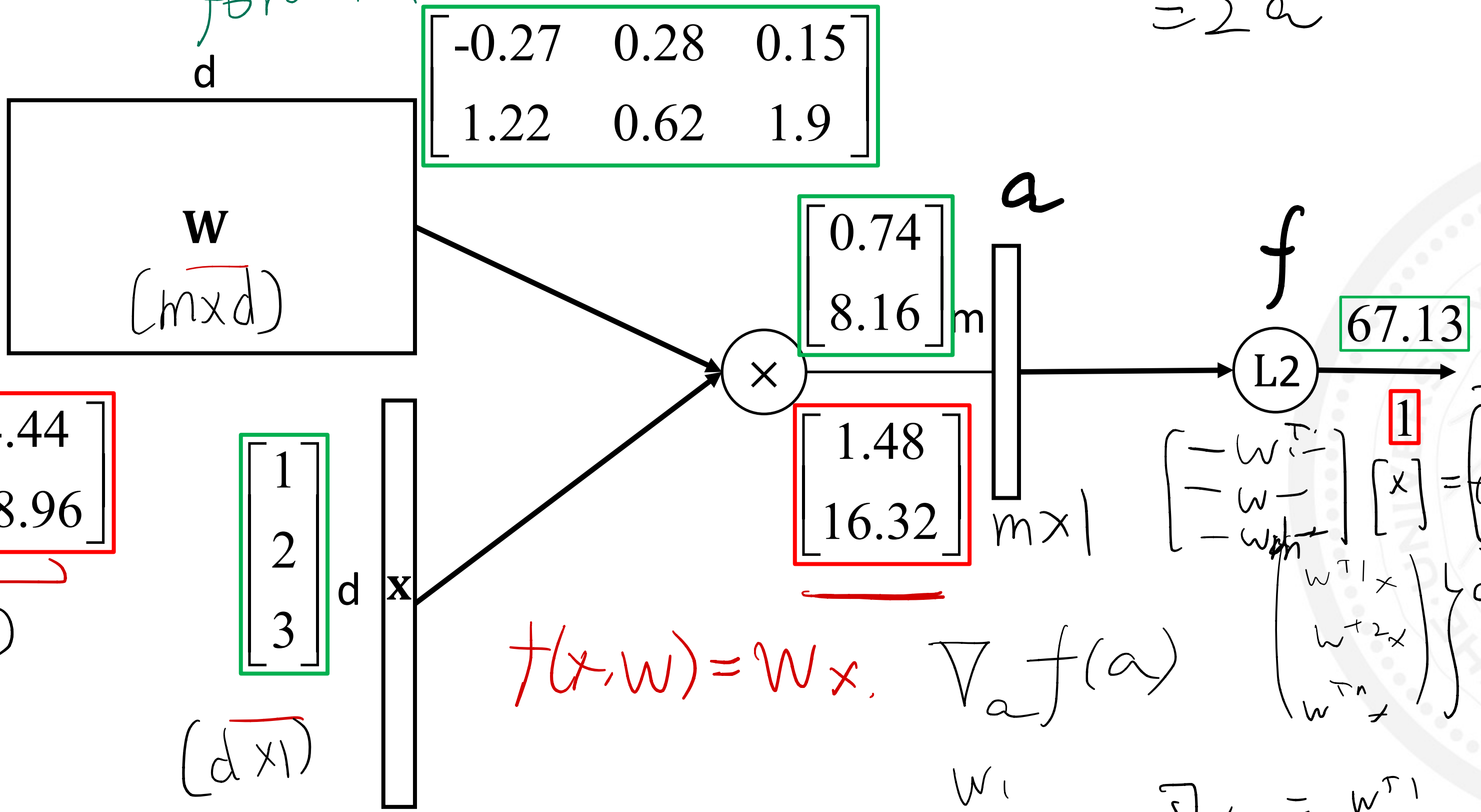
$$f(\mathbf{x},\mathbf{W}) = \|\mathbf{W}\cdot\mathbf{x}\|^2 = \sum_{i=1}^{n}(\mathbf{W}\cdot\mathbf{x})_i^2$$

① $a(x,W) = W^T x = W \cdot x$ $\left\{ \begin{array}{l} \nabla_W a(x,W) = x^T \\ \nabla_x a(x,W) = W \end{array} \right.$

② $f(a) = \|a\|^2 = a^T a$ $\left\{ \nabla_a f(a) = 2a \right.$

local gradient

forward



(m×1) (1×d) $\nabla_W a(W,x) = x^T$ (m×d)

$$\underbrace{\begin{bmatrix}1.48\\16.32\end{bmatrix}}\times\underbrace{\begin{bmatrix}1 & 2 & 3\end{bmatrix}}=\underbrace{\begin{bmatrix}1.48 & 2.96 & 4.44\\16.32 & 32.64 & 48.96\end{bmatrix}}$$

$\nabla_a f(a) \cdot \underbrace{\nabla_W a(x,W)}_{x^T} = \nabla_W f(x,W)$

$f(x,W) = Wx$, $\nabla_a f(a)$

$\begin{bmatrix} -w_1^T- \\ -w- \\ -w_m^T- \end{bmatrix}\begin{bmatrix}x\end{bmatrix} = \begin{bmatrix} w_1^T x \\ \vdots \\ w_m^T x \end{bmatrix}$

$\begin{pmatrix} w^{T1}x \\ w^{T2}x \\ w^{Tn}x \end{pmatrix}$ d×1

$W_1 \vdots V_m$

$\nabla x_1 = w^{T1}$