SDU Summer School

Deep Learning

Summer 2022

Welcome to the Summer School

Backpropagation

- **Function Principle**
- **Generalization to Vectors**

Chain Rule of Calculus

- If g is differentiable at x and f is differentiable at $g(x)$, then the composite function $F = f \circ g$ defined by $F(x) = f\big(g(x)\big)$ is differentiable at x and F' is given by

$$F' = f'\big(g(x)\big) \cdot g'(x)$$

- In Leibniz notation, if $y = f(u)$ and $u = g(x)$, then

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx}$$

Forward vs. Backward Mode

- **Forward Accumulation**:
  - One first fixes the **independent variable** with respect to which differentiation is performed and computes the derivative of each sub-expression recursively

$$\frac{dy}{dx} = \frac{dy}{dw_{n-1}}\frac{dw_{i-1}}{dx} = \frac{dy}{dw_{n-1}}\left(\frac{dw_{n-1}}{dw_{n-2}}\frac{dw_{i-2}}{dx}\right) = \ \dots$$

- **Backward Accumulation**:
  - One first fixes the **dependent variable** to be differentiated and computes the derivative with respect to each sub-expression recursively

$$\frac{dy}{dx} = \frac{dy}{dw_1}\frac{dw_1}{dx} = \left(\frac{dy}{dw_2}\frac{dw_2}{dw_1}\right)\frac{dw_1}{dx} = \ \dots$$

Forward Accumulation Example

$$
\begin{aligned}
z &= f(x_1, x_2) \\
&= x_1 x_2 + \sin x_1 \\
&= w_1 w_2 + \sin w_1 \\
&= w_3 + w_4 \\
&= w_5
\end{aligned}
$$

The choice of the independent variable defines the used seed. For example, we want to differentiate with respect to x_1:

$$\dot{w}_1 = \frac{dx_1}{dx_1} = 1 \text{ and } \dot{w}_2 = \frac{dx_2}{dx_1} = 0$$

$$\frac{\boldsymbol{df}}{\boldsymbol{dx_1}}:$$

Compute Value	Compute derivative
$w_1 = x_1$	$\dot{w}_1 = 1$
$w_2 = x_2$	$\dot{w}_2 = 0$
$w_3 = w_1 \cdot w_2$	$\dot{w}_3 = w_2 \cdot \dot{w}_1 + w_1 \cdot \dot{w}_2$
$w_4 = \sin w_1$	$\dot{w}_4 = \cos w_1 \cdot \dot{w}_1$
$w_5 = w_3 + w_4$	$\dot{w}_5 = \dot{w}_3 + \dot{w}_4$

Observations

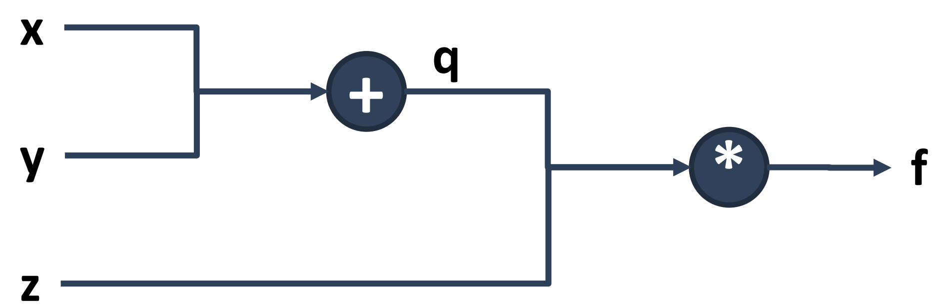
- In order to also get a derivative $\frac{df}{dx_2}$ another run would be required to receive the gradient

- Forward accumulation is good for functions
$$f: \mathbb{R}^n \to \mathbb{R}^m$$
with $\mathrm{m} \gg n$

- In deep learning: Normally millions of weights (n) to optimize with only one output ($m = 1$), the costs
  - => Therefore, Backward accumulation, or Backpropagation

Backpropagation: A Simple Example

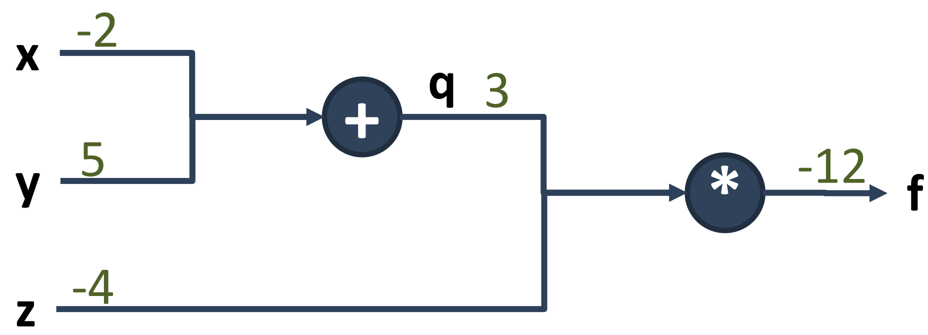
$$f(x, y, z) = (x + y)z$$



Backpropagation: A Simple Example

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

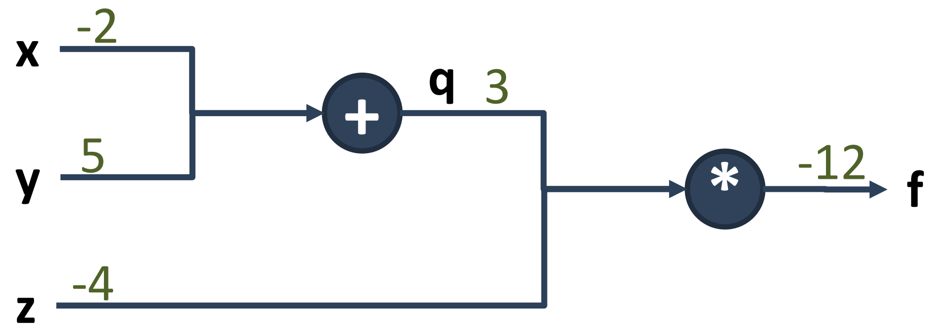
- With
  - $x = -2$
  - $y = 5$
  - $z = -4$



Backpropagation: A Simple Example

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

- With
  - $x = -2$
  - $y = 5$
  - $z = -4$



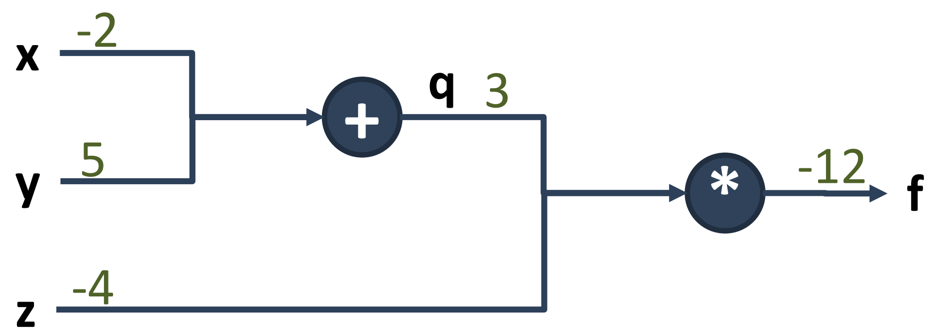
$$q = x + y; \quad \frac{\partial q}{\partial x} = 1; \ \frac{\partial q}{\partial y} = 1;$$

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

Backpropagation: A Simple Example

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

- With
  - $x = -2$
  - $y = 5$
  - $z = -4$



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

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

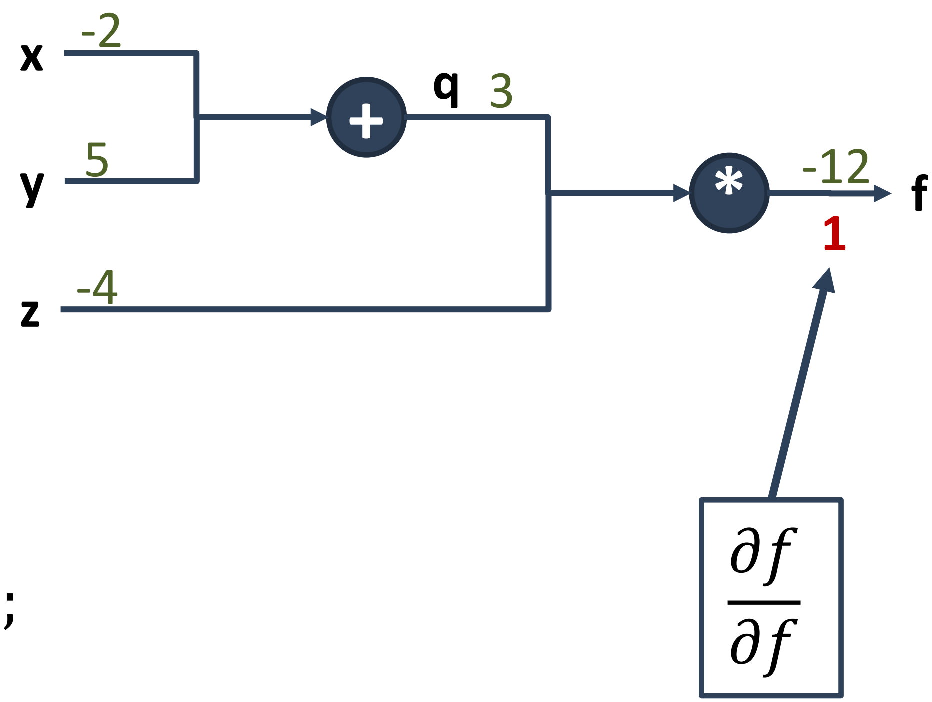
Looking for:

$$\frac{\partial f}{\partial x} \qquad \frac{\partial f}{\partial y} \qquad \frac{\partial f}{\partial z}$$

Backpropagation: A Simple Example

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

- With
  - $x = -2$
  - $y = 5$
  - $z = -4$



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

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

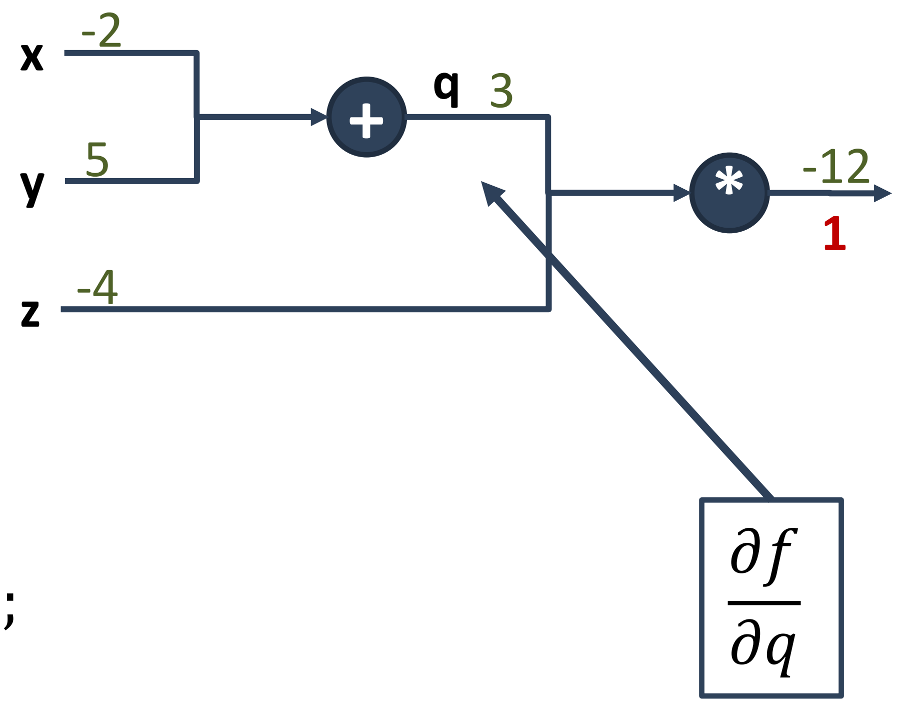
Looking for:

$$\frac{\partial f}{\partial x} \qquad \frac{\partial f}{\partial y} \qquad \frac{\partial f}{\partial z}$$

Backpropagation: A Simple Example

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

- With
  - $x = -2$
  - $y = 5$
  - $z = -4$



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

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

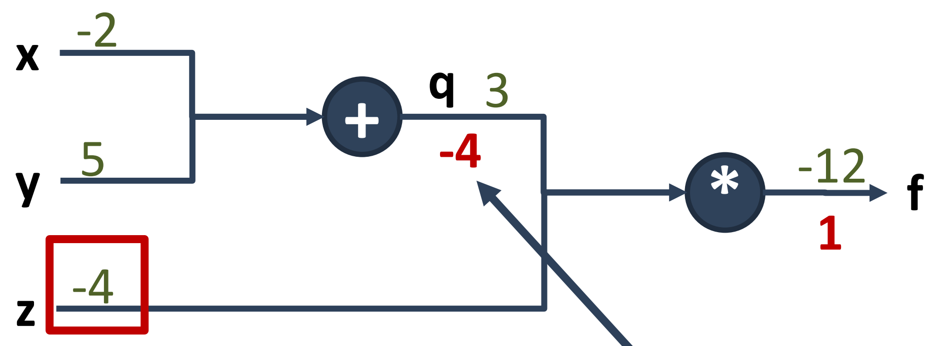
Looking for:

$$\frac{\partial f}{\partial x} \quad \frac{\partial f}{\partial y} \quad \frac{\partial f}{\partial z}$$

Backpropagation: A Simple Example

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

- With
  - $x = -2$
  - $y = 5$
  - $z = -4$



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

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

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

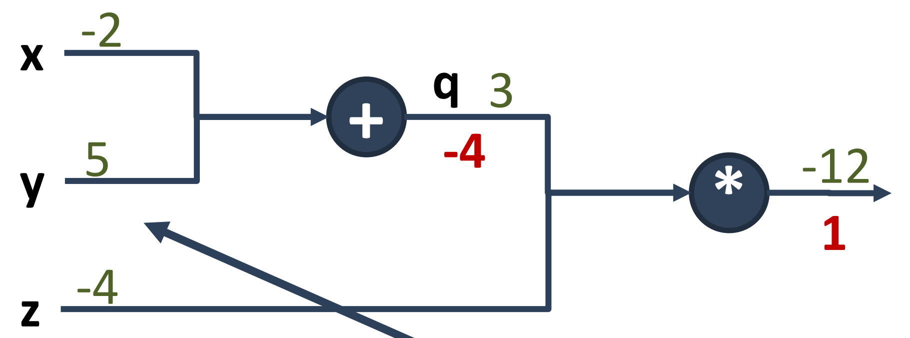
Looking for:

$$\frac{\partial f}{\partial x} \quad \frac{\partial f}{\partial y} \quad \frac{\partial f}{\partial z}$$

Backpropagation: A Simple Example

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

- With
  - $x = -2$
  - $y = 5$
  - $z = -4$



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

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

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

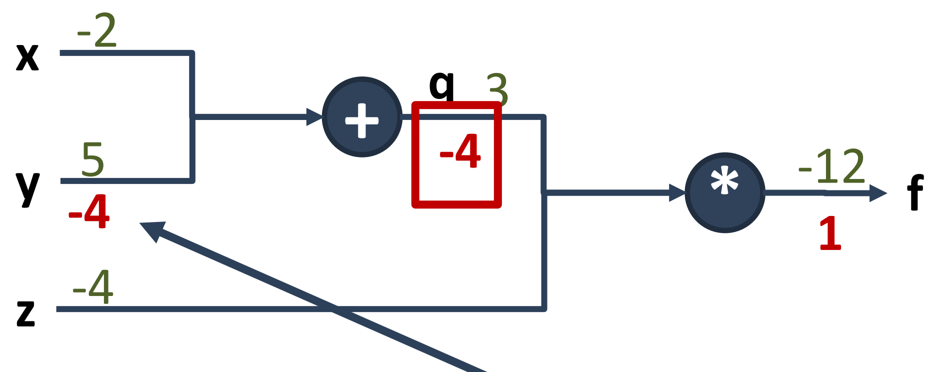
Looking for:

$$\frac{\partial f}{\partial x} \quad \frac{\partial f}{\partial y} \quad \frac{\partial f}{\partial z}$$

Backpropagation: A Simple Example

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

- With
  - $x = -2$
  - $y = 5$
  - $z = -4$



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

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

Chain Rule:

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

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

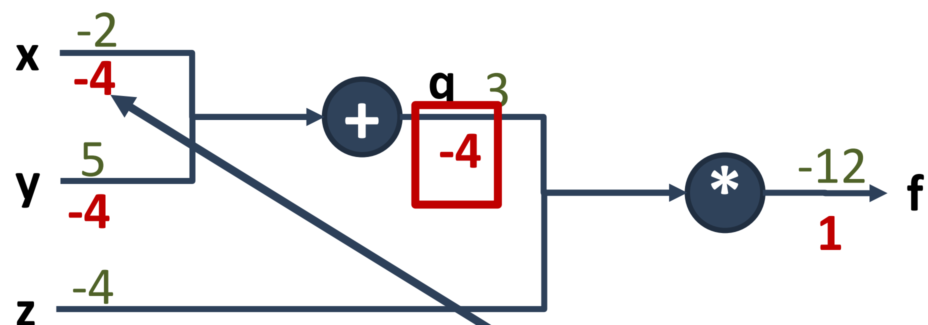
Looking for:

$$\frac{\partial f}{\partial x} \quad \frac{\partial f}{\partial y} \quad \frac{\partial f}{\partial z}$$

Backpropagation: A Simple Example

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

- With
  - $x = -2$
  - $y = 5$
  - $z = -4$



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

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

Chain Rule:

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

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

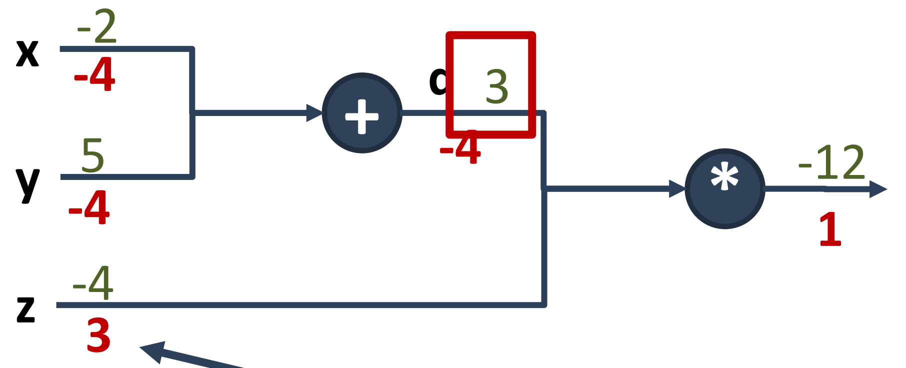
Looking for:

$$\frac{\partial f}{\partial x} \quad \frac{\partial f}{\partial y} \quad \frac{\partial f}{\partial z}$$

Backpropagation: A Simple Example

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

- With
  - $x = -2$
  - $y = 5$
  - $z = -4$



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

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

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

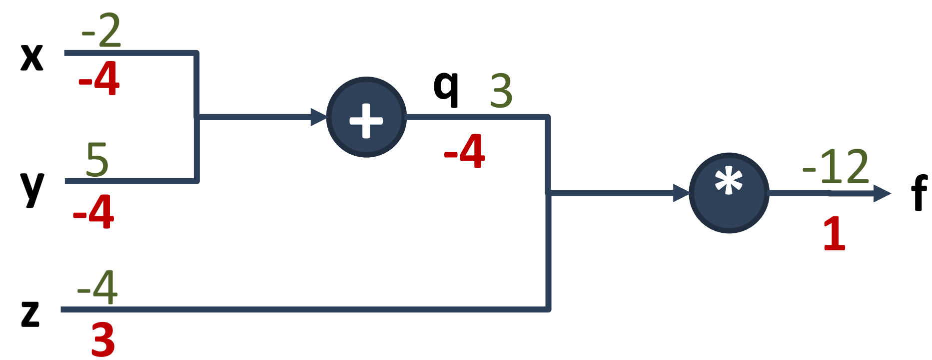
Looking for:

$$\frac{\partial f}{\partial x} \quad \frac{\partial f}{\partial y} \quad \frac{\partial f}{\partial z}$$

Backpropagation: A Simple Example

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

- With
  - $x = -2$
  - $y = 5$
  - $z = -4$



$$\nabla_{x,y,z} f = \begin{pmatrix} -4 \\ -4 \\ 3 \end{pmatrix}$$

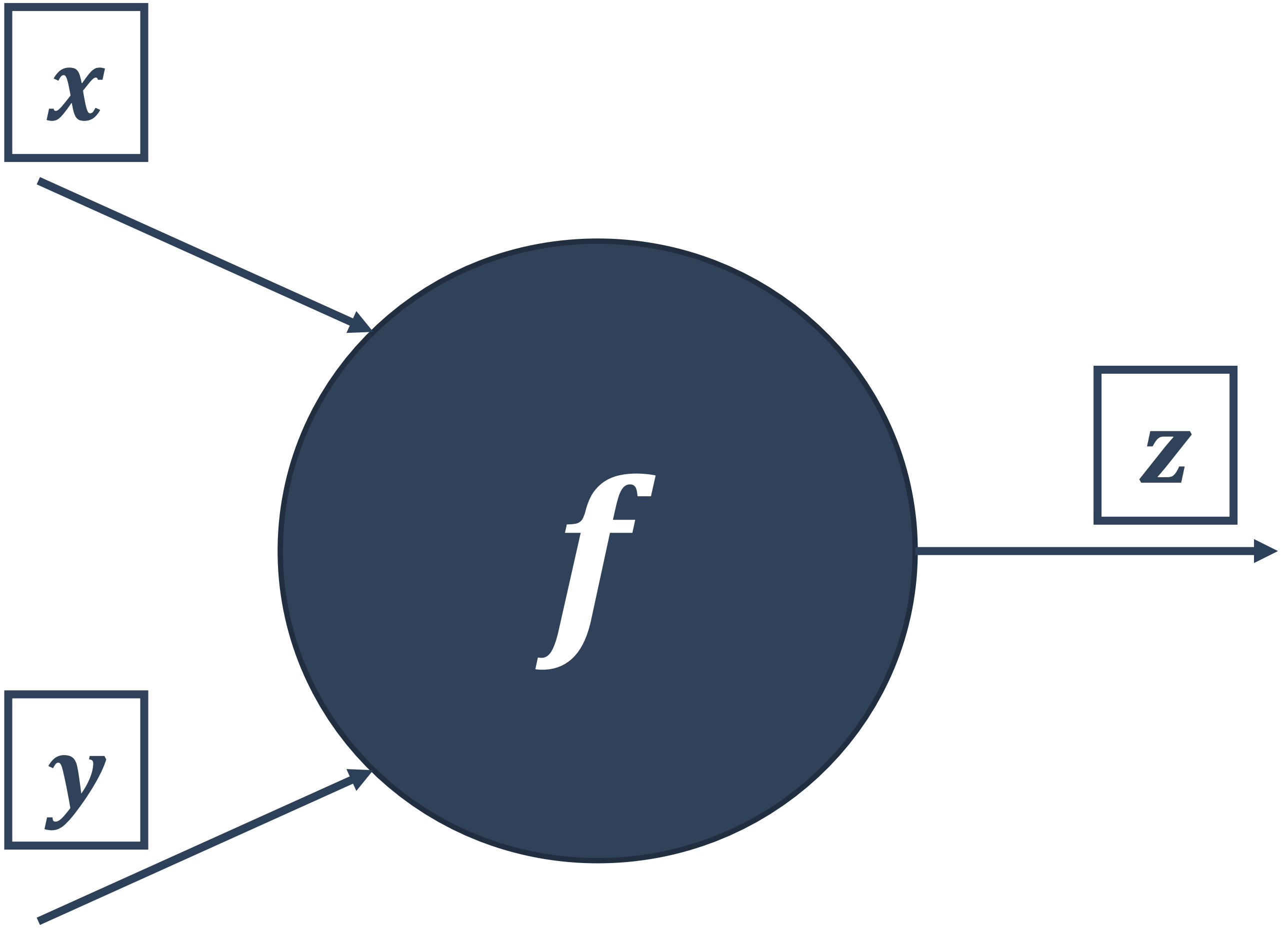
$$q = x + y; \quad \frac{\partial q}{\partial x} = 1; \ \frac{\partial q}{\partial y} = 1;$$

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

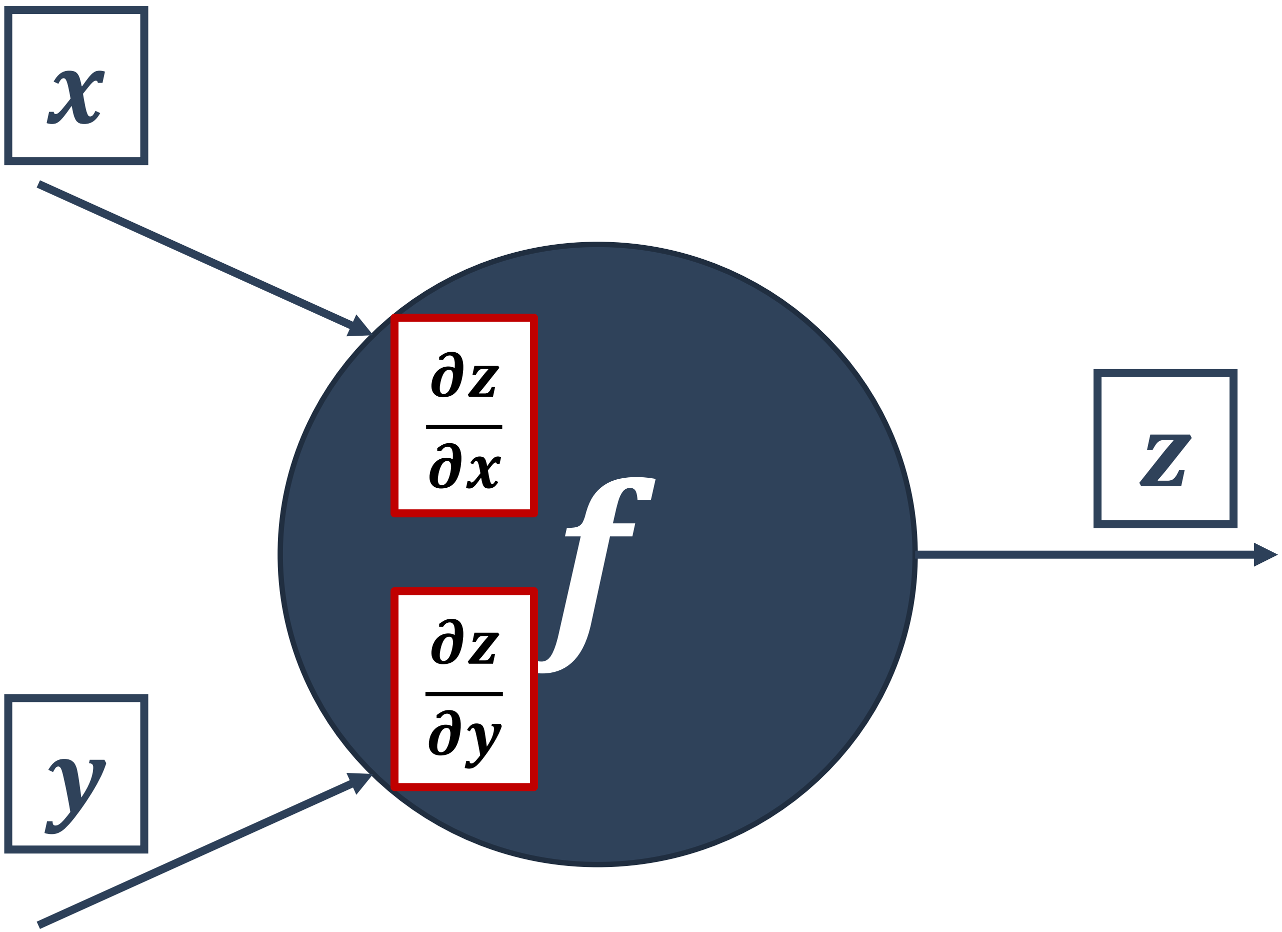
Looking for:

$$\frac{\partial f}{\partial x} \quad \frac{\partial f}{\partial y} \quad \frac{\partial f}{\partial z}$$

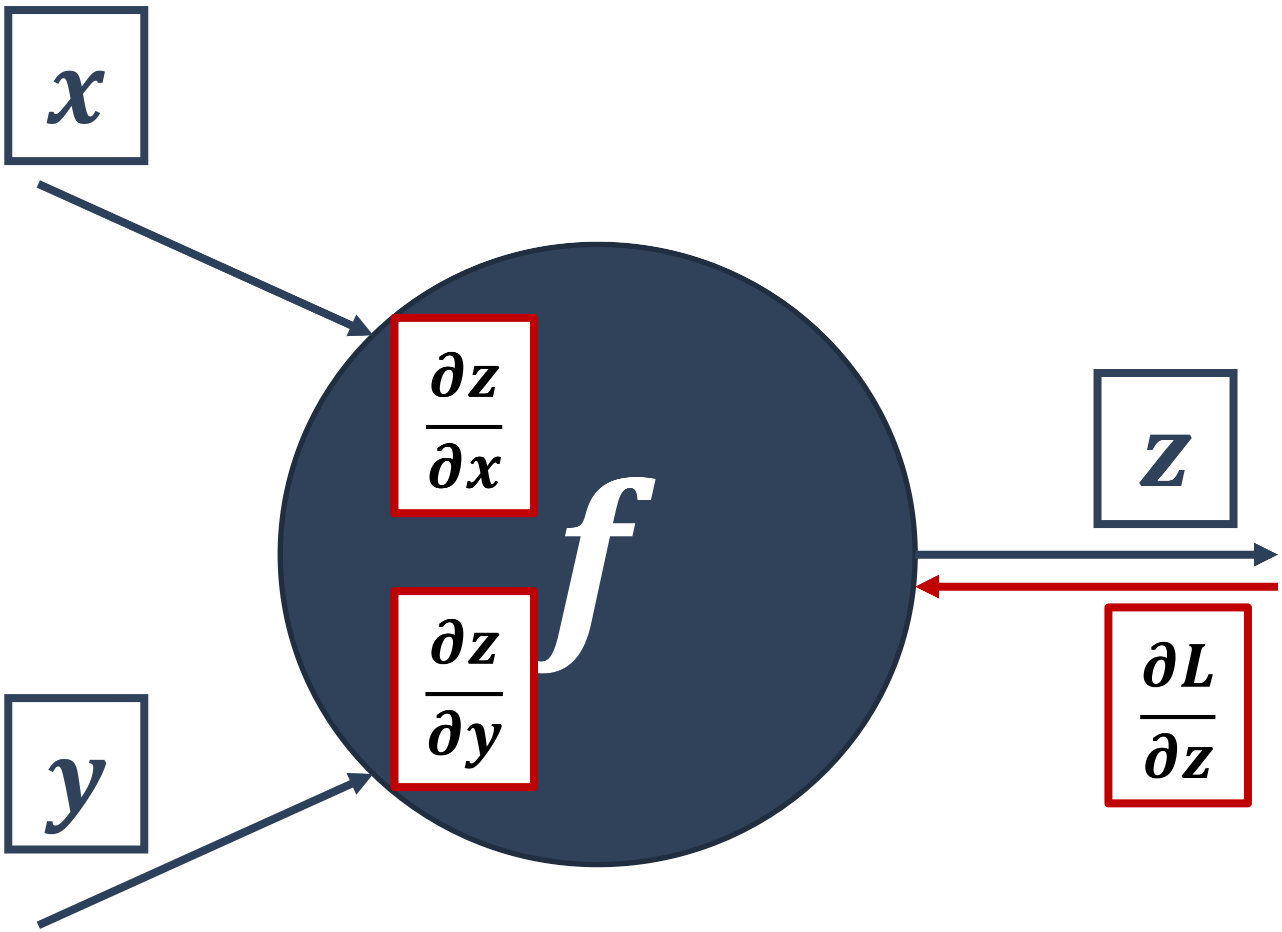
General Principle



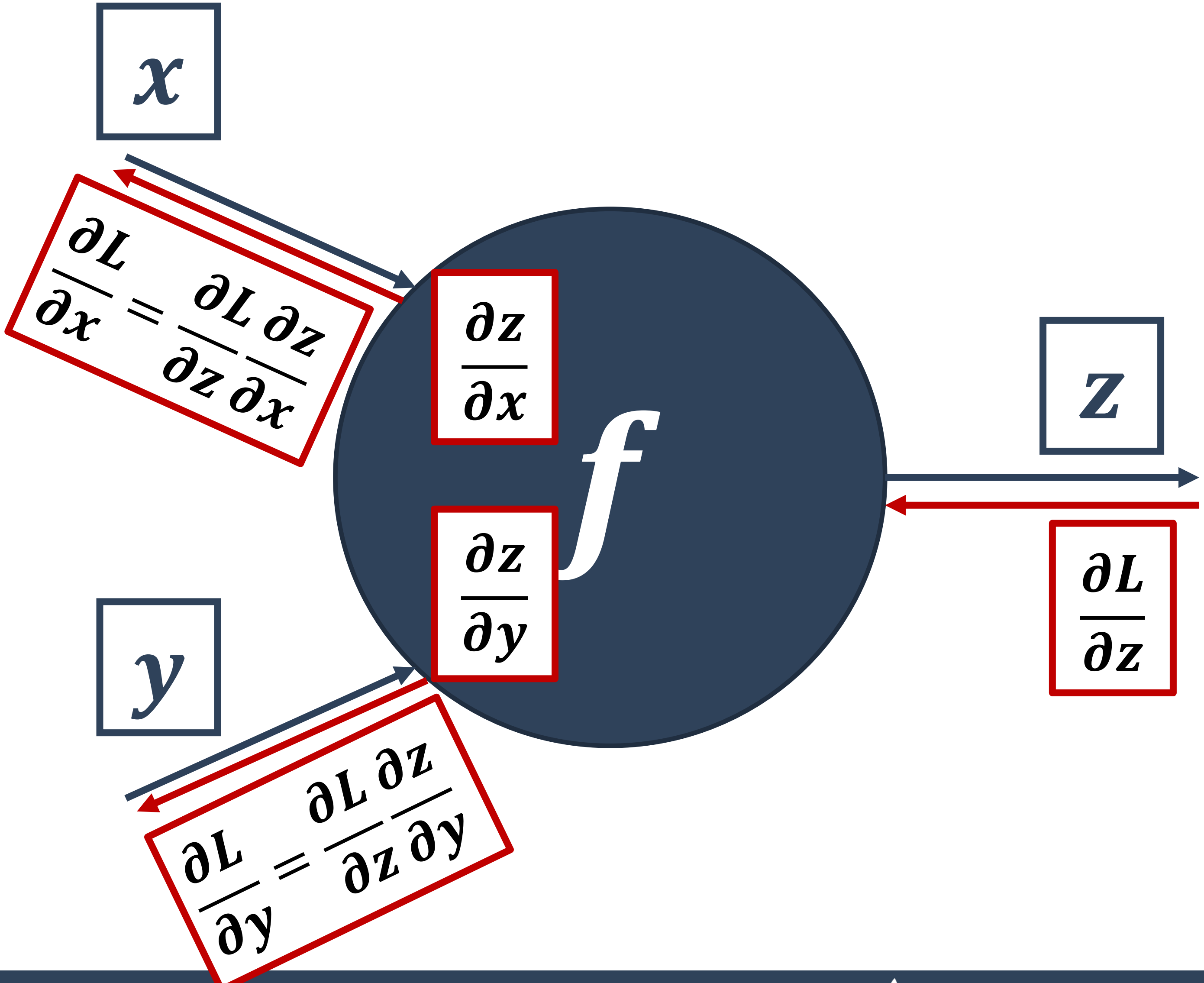
General Principle



General Principle

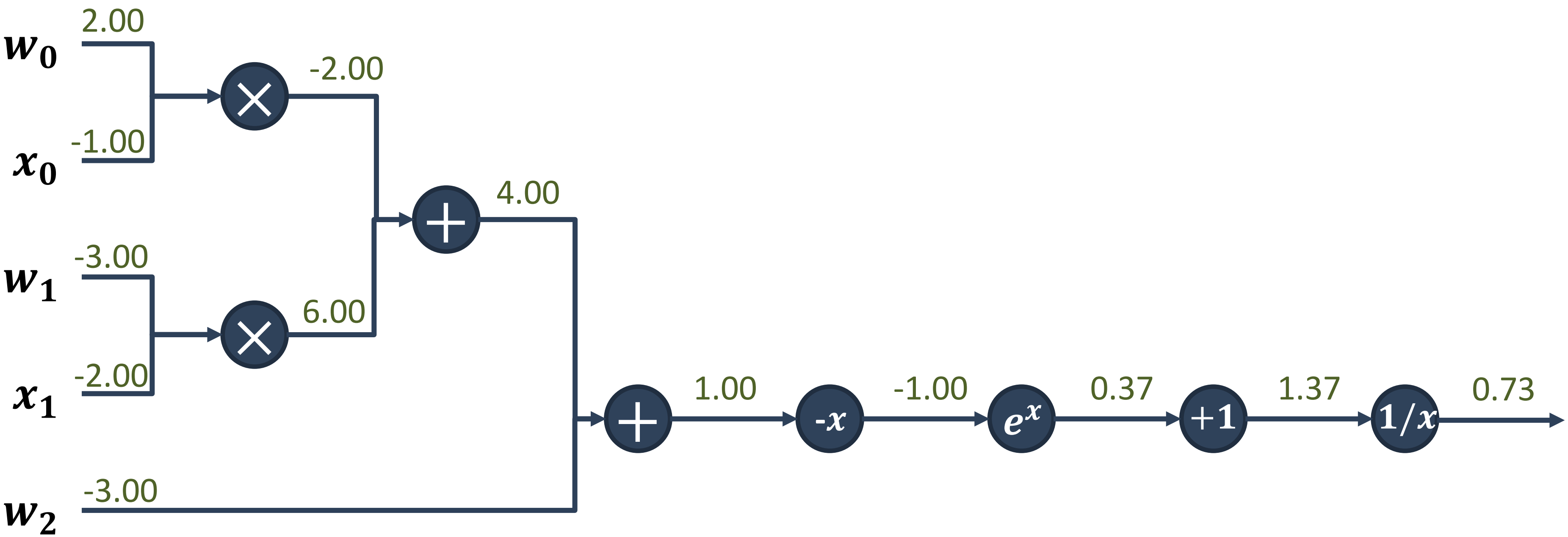


General Principle



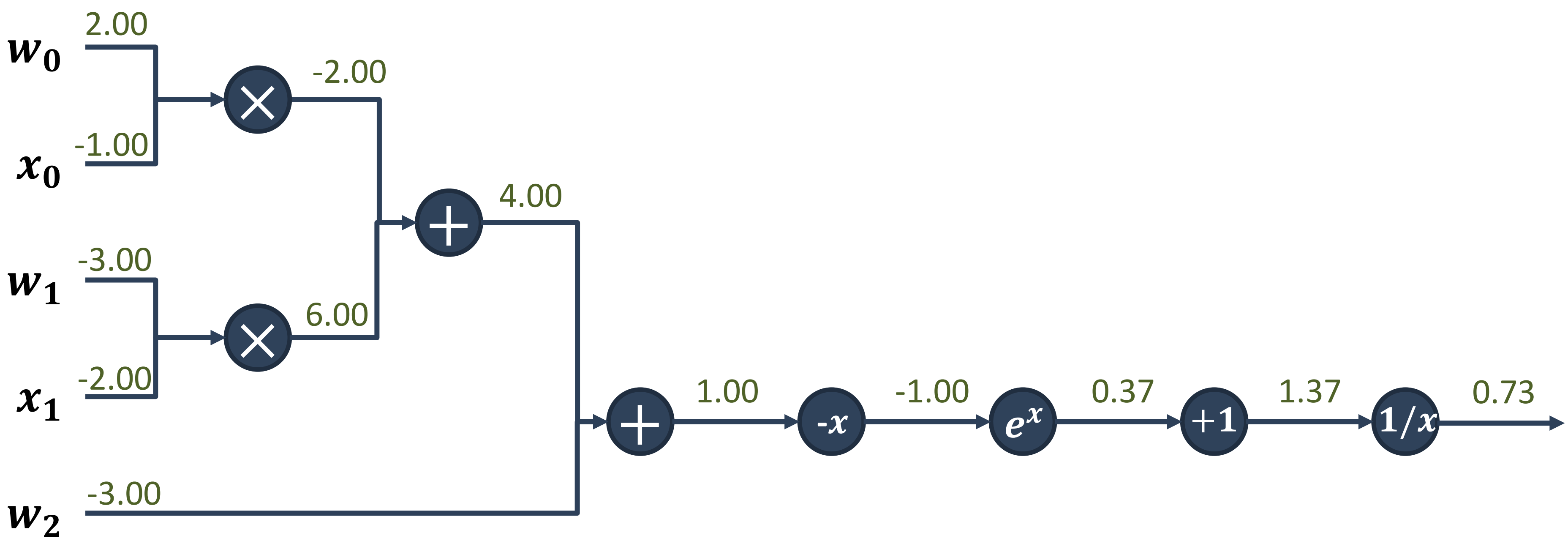
Different Example

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



Different Example

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



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

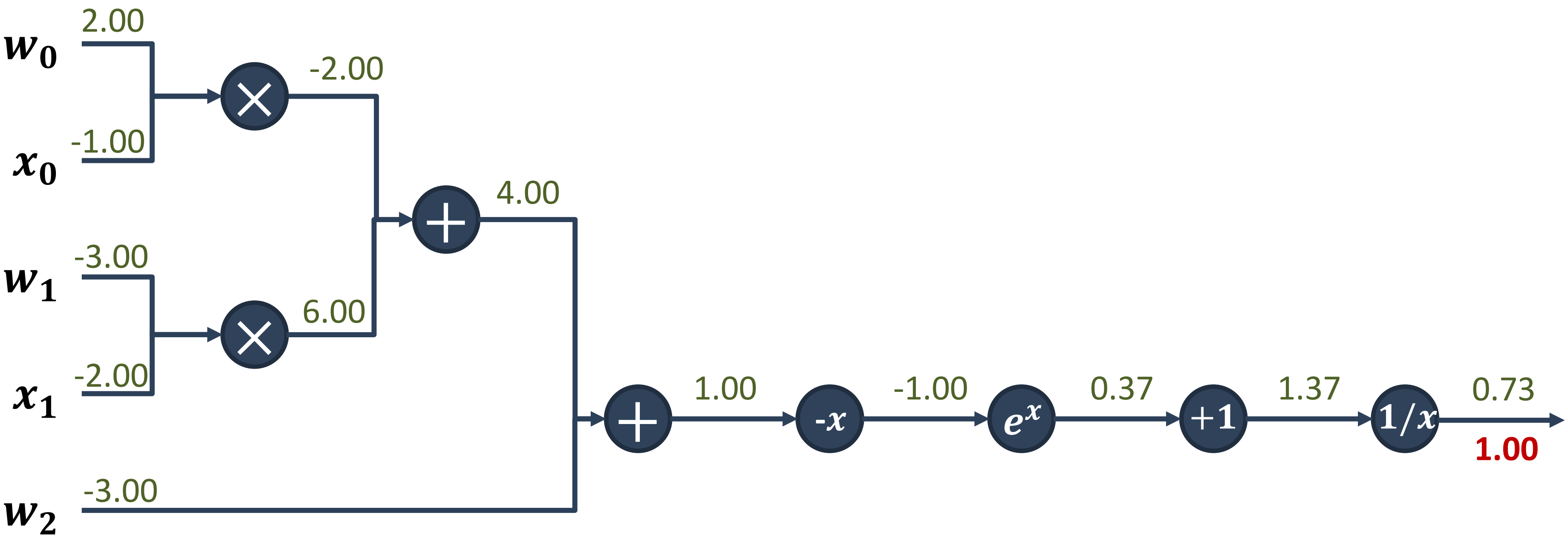
$$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{\partial f}{\partial x} = -\frac{1}{x^2}$$

$$f_a(x) = ax \quad \rightarrow \quad \frac{\partial f}{\partial x} = a$$

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

Different Example

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

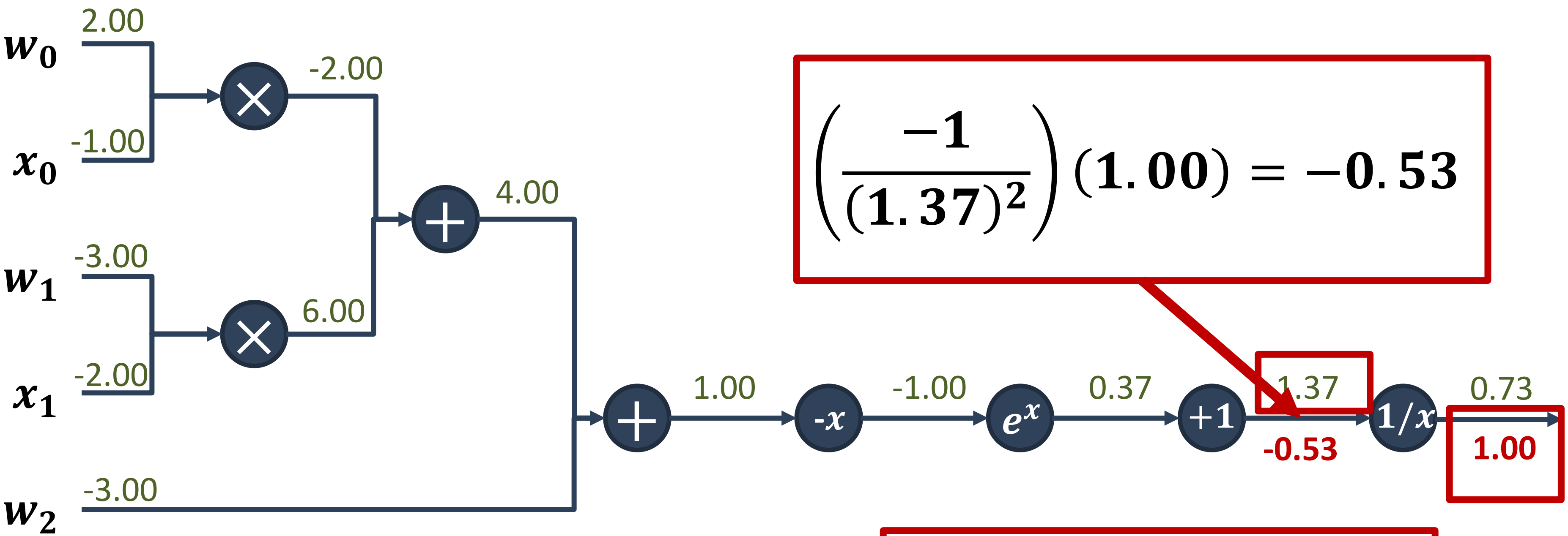


$$f(x) = e^x \quad \rightarrow \quad \frac{\partial f}{\partial x} = e^x \qquad\qquad f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{\partial f}{\partial x} = -\frac{1}{x^2}$$

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

Different Example

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



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

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

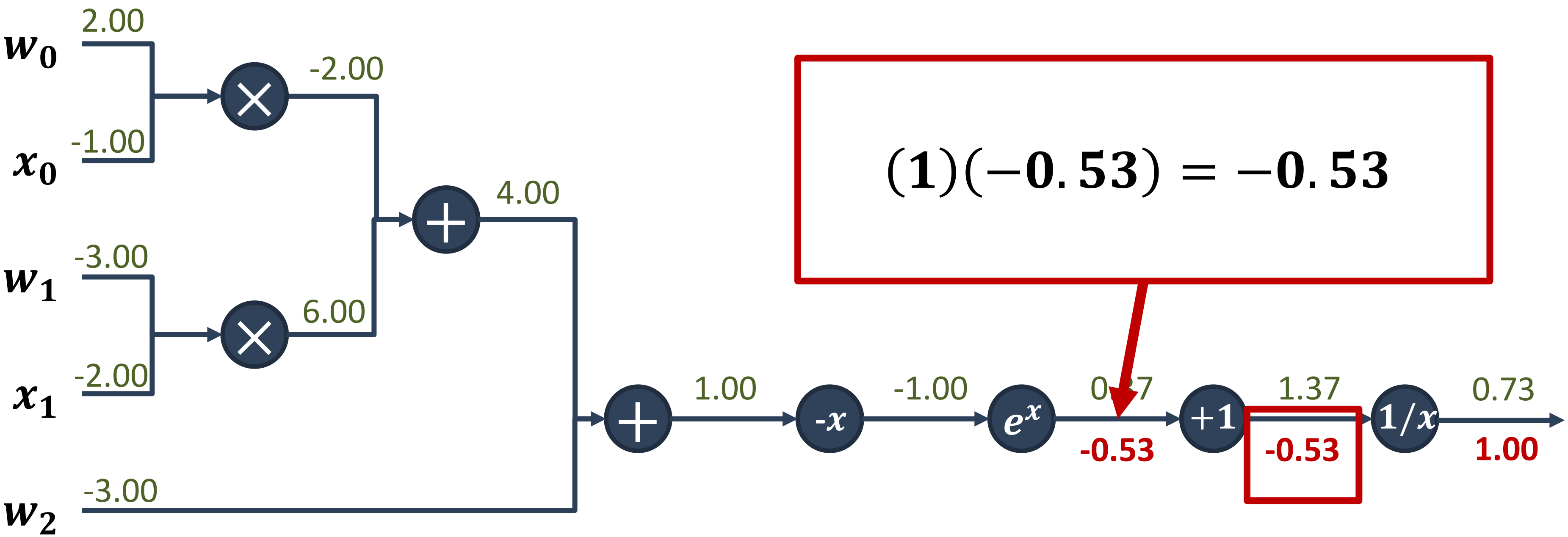
$$\boxed{f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{\partial f}{\partial x} = -\frac{1}{x^2}}$$

$$f_a(x) = ax \quad \rightarrow \quad \frac{\partial f}{\partial x} = a$$

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

Different Example

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



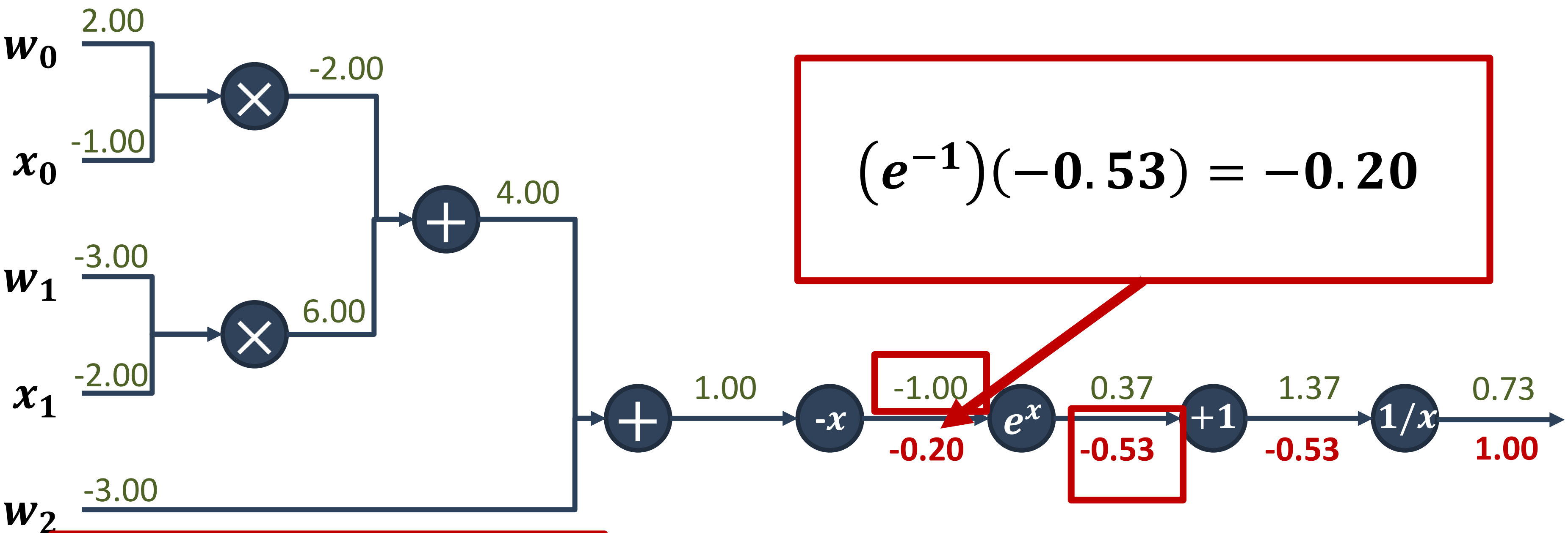
$$\boxed{(1)(-0.53) = -0.53}$$

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

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

Different Example

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



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

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

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

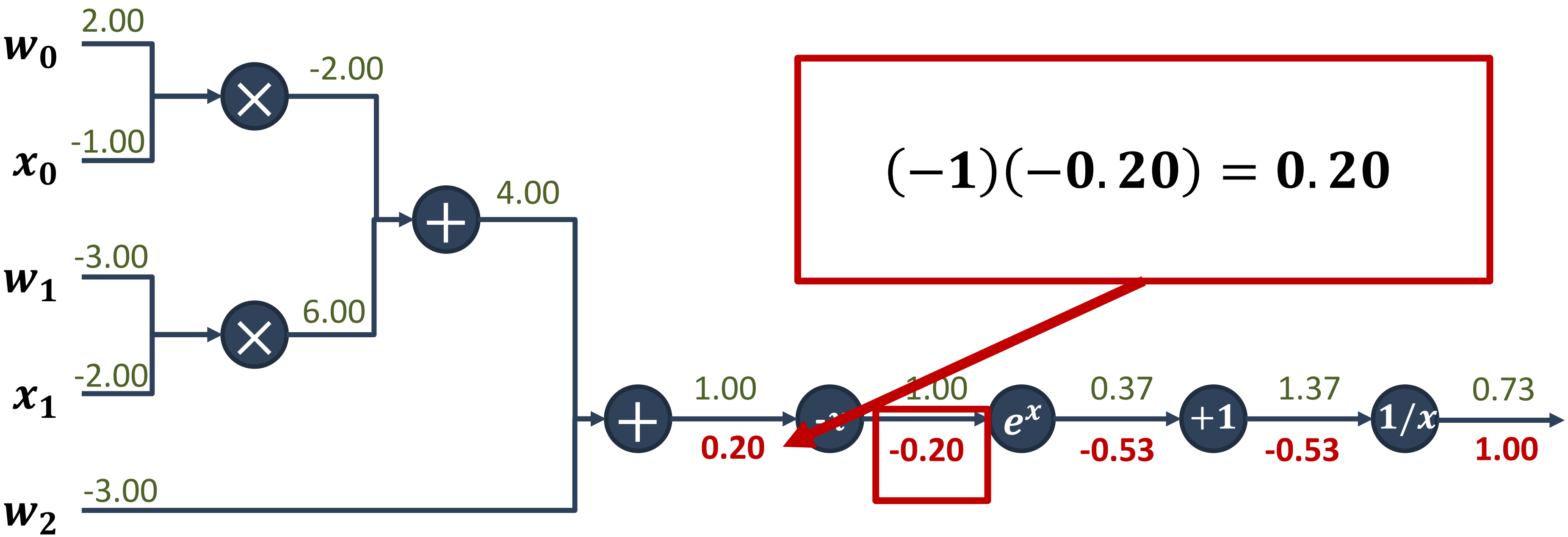
$$f_a(x) = ax \quad \rightarrow \quad \frac{\partial f}{\partial x} = a$$

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

Different Example

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

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



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

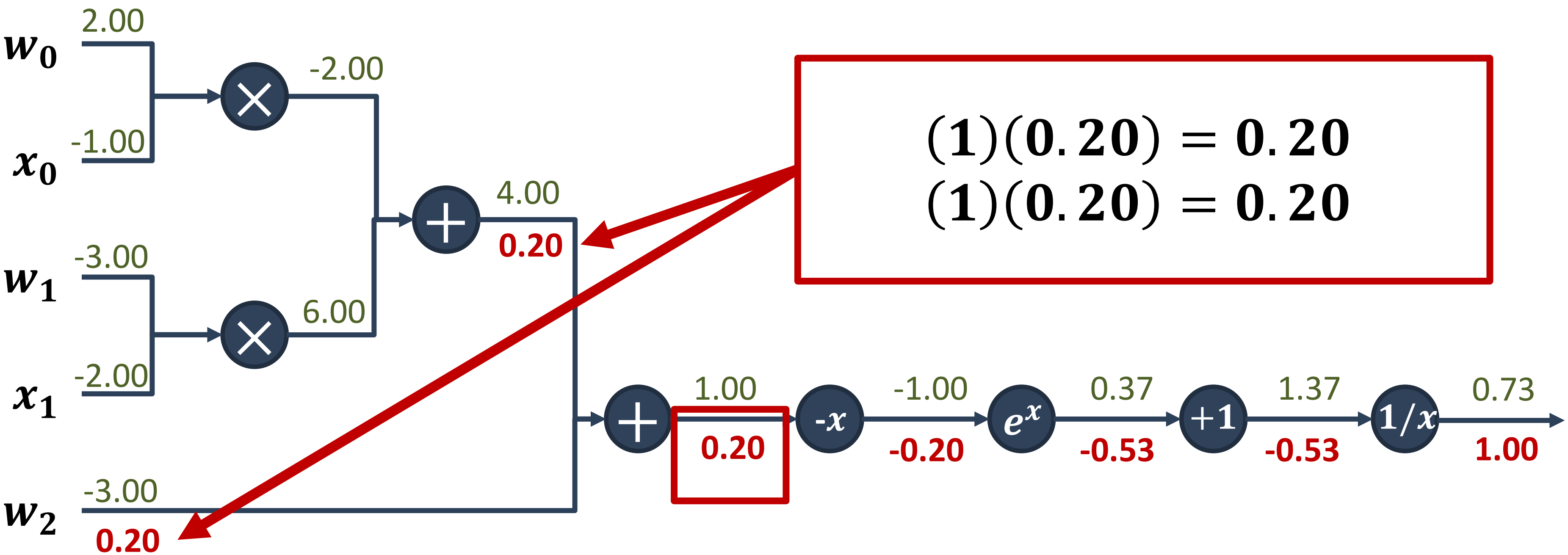
$$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{\partial f}{\partial x} = -\frac{1}{x^2}$$

$$\boxed{f_a(x) = ax \quad \rightarrow \quad \frac{\partial f}{\partial x} = a}$$

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

Different Example

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



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

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

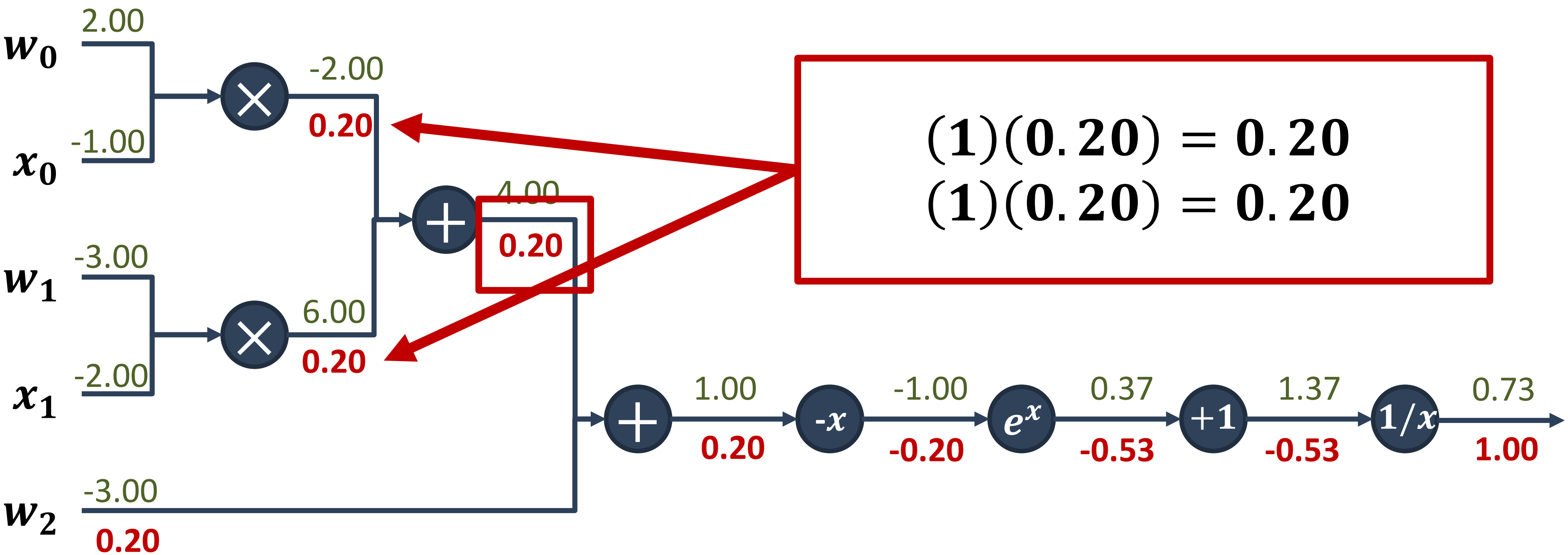
$$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{\partial f}{\partial x} = -\frac{1}{x^2}$$

$$f_a(x) = ax \quad \rightarrow \quad \frac{\partial f}{\partial x} = a$$

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

Different Example

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



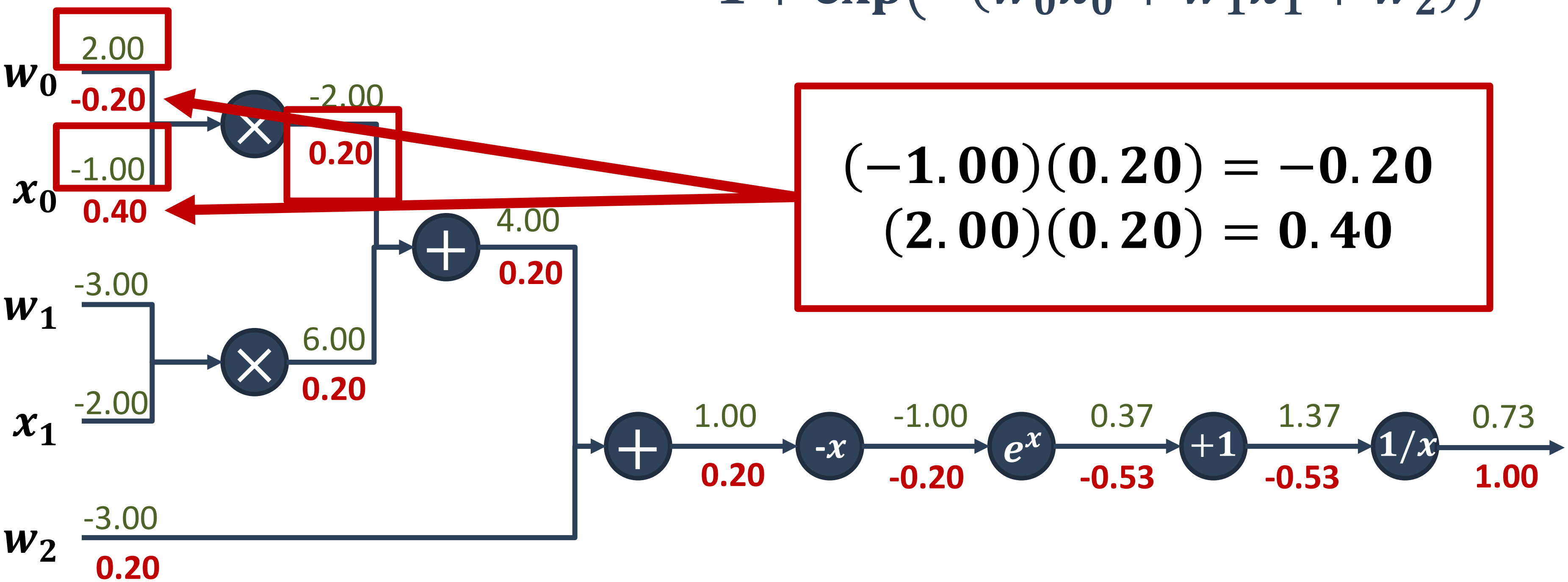
$$(1)(0.20) = 0.20$$
$$(1)(0.20) = 0.20$$

$$f(x) = e^x \rightarrow \frac{\partial f}{\partial x} = e^x \qquad f(x) = \frac{1}{x} \rightarrow \frac{\partial f}{\partial x} = -\frac{1}{x^2}$$

$$f_a(x) = ax \rightarrow \frac{\partial f}{\partial x} = a \qquad \boxed{f_c(x) = c + x \rightarrow \frac{\partial f}{\partial x} = 1}$$

Different Example

$$f(w,x)=\frac{1}{1+\exp\left(-(w_0x_0+w_1x_1+w_2)\right)}$$



$$(-1.00)(0.20)=-0.20$$
$$(2.00)(0.20)=0.40$$

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

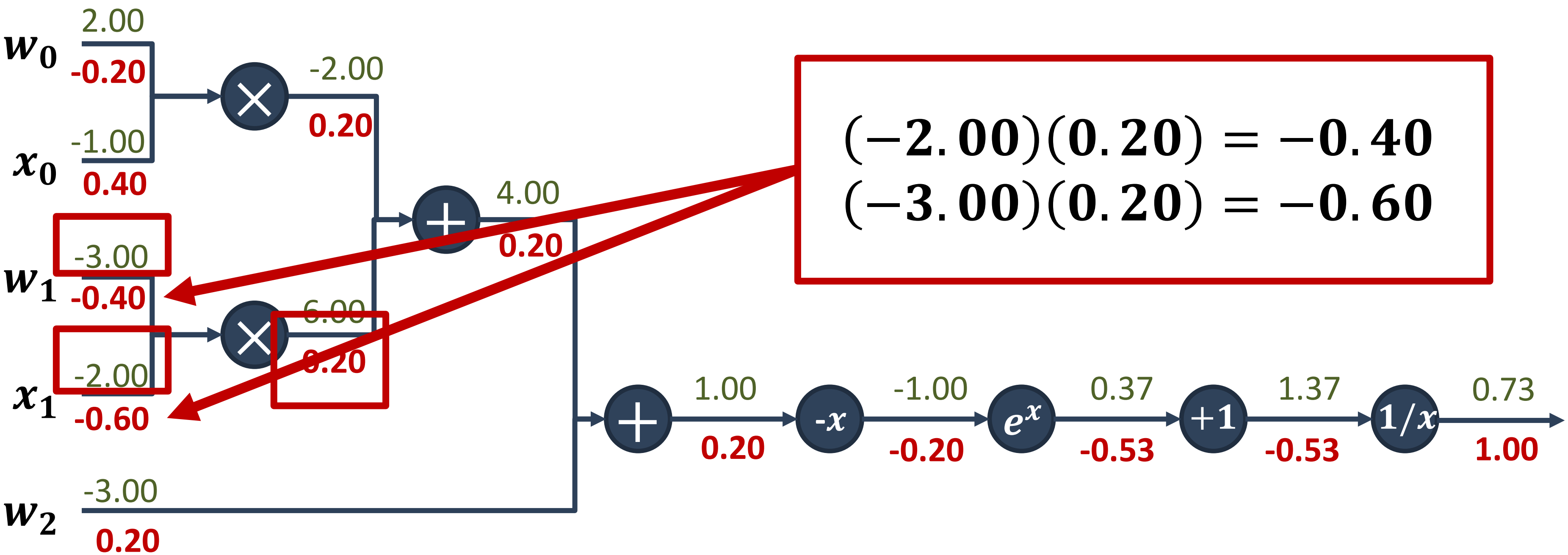
$$f(x)=\frac{1}{x} \quad\rightarrow\quad \frac{\partial f}{\partial x}=-\frac{1}{x^2}$$

$$\boxed{f_a(x)=ax \quad\rightarrow\quad \frac{\partial f}{\partial x}=a}$$

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

Different Example

$$f(w,x) = \frac{1}{1+\exp\left(-(w_0x_0 + w_1x_1 + w_2)\right)}$$



$$(-2.00)(0.20) = -0.40$$
$$(-3.00)(0.20) = -0.60$$

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

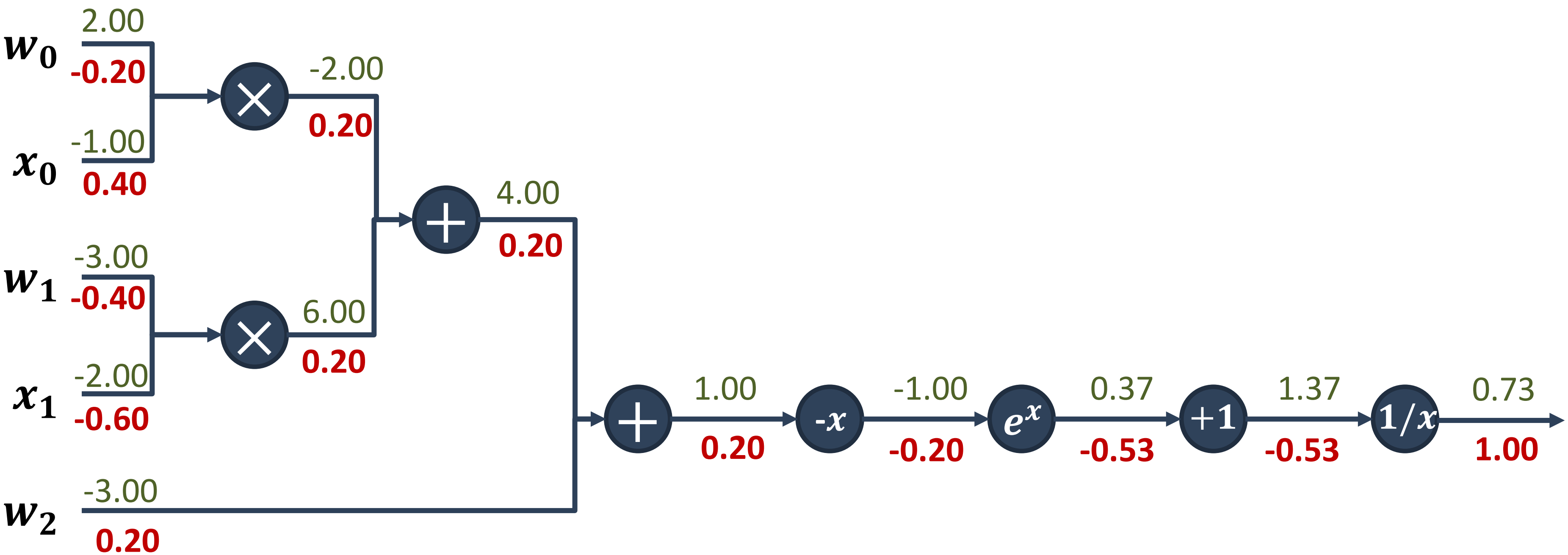
$$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{\partial f}{\partial x} = -\frac{1}{x^2}$$

$$\boxed{f_a(x) = ax \quad \rightarrow \quad \frac{\partial f}{\partial x} = a}$$

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

Different Example

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



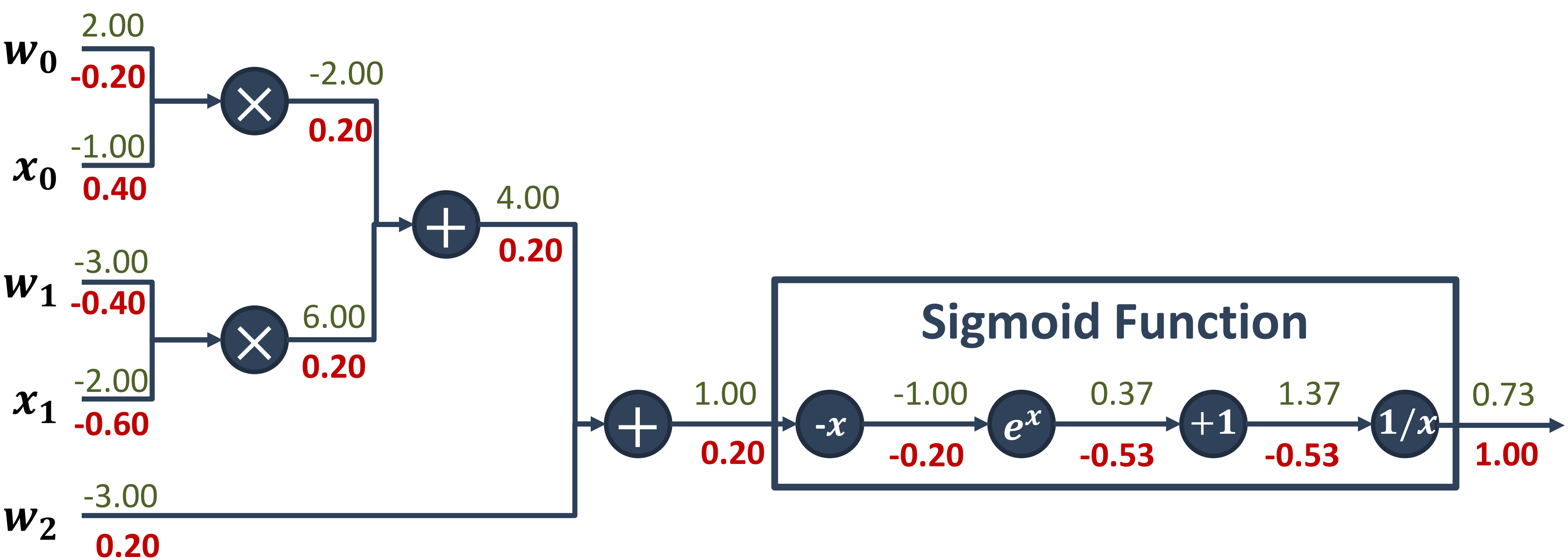
$$f(x) = e^x \quad \rightarrow \quad \frac{\partial f}{\partial x} = e^x \qquad\qquad f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{\partial f}{\partial x} = -\frac{1}{x^2}$$

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

Different Example: Sigmoid Function

$$f(w,x) = \frac{1}{1+\exp\left(-(w_0x_0+w_1x_1+w_2)\right)}$$

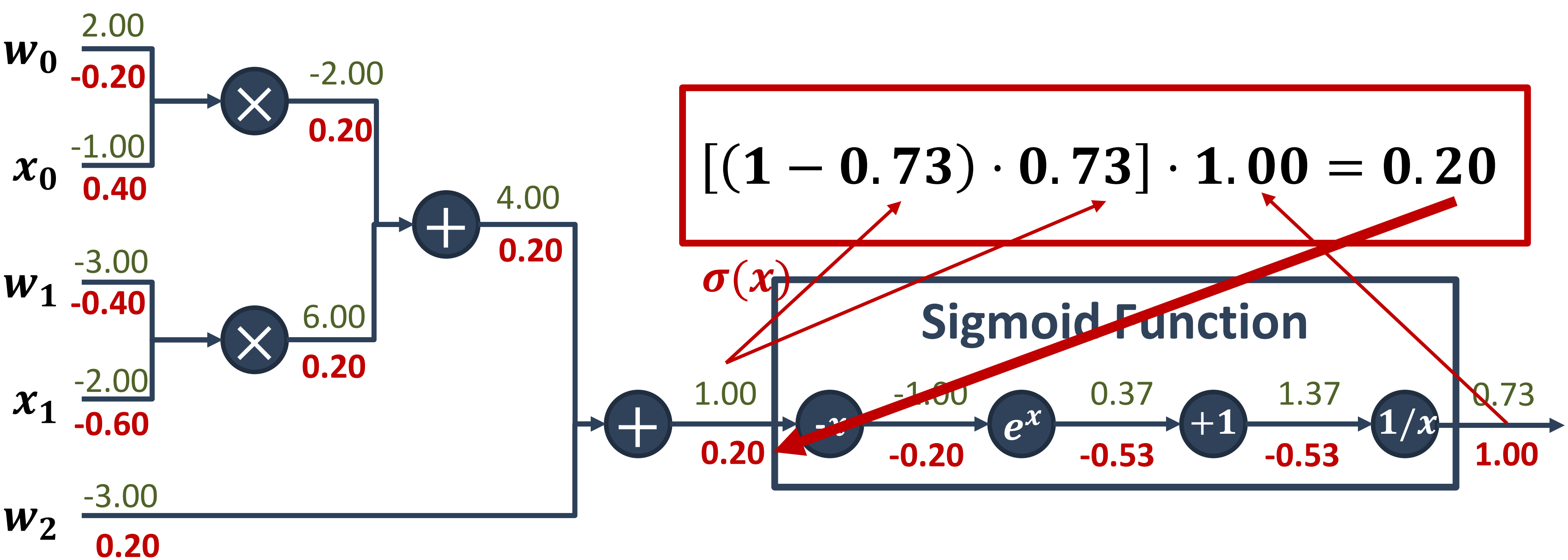
$$\sigma(x) = \frac{1}{1+\exp(-x)} \qquad \frac{\partial\sigma}{\partial x} = \left(1-\sigma(x)\right)\sigma(x)$$



Different Example: Sigmoid Function

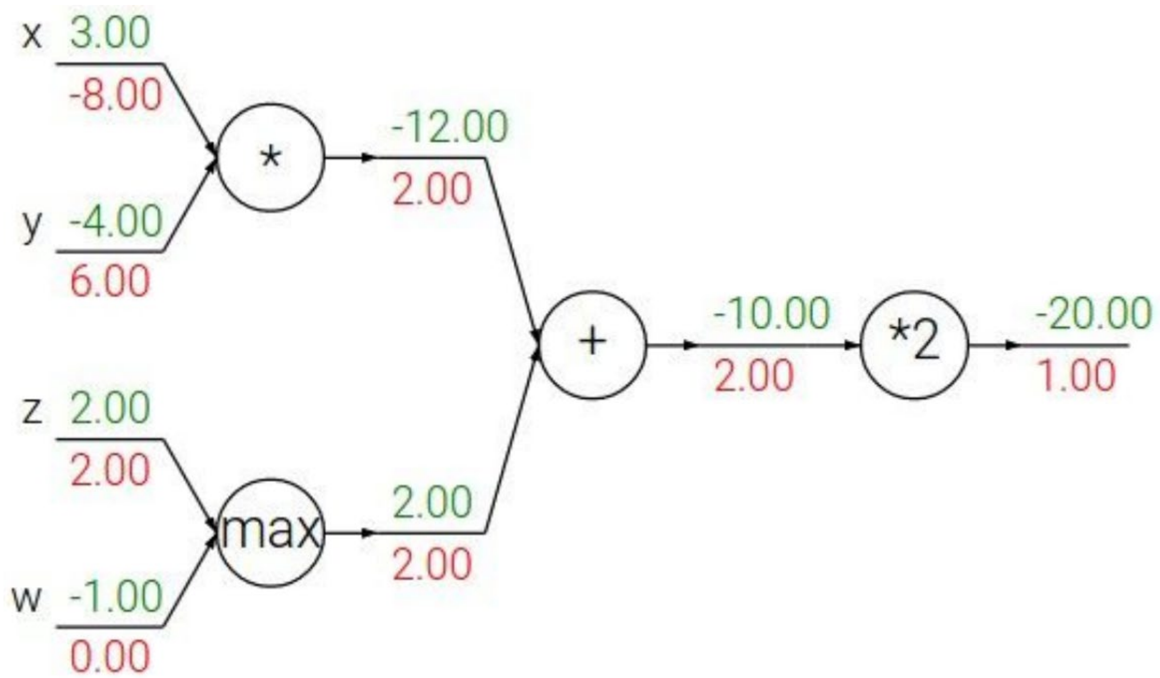
$$f(w,x) = \frac{1}{1+\exp\left(-(w_0x_0 + w_1x_1 + w_2)\right)}$$

$$\sigma(x) = \frac{1}{1+\exp(-x)} \qquad \frac{\partial\sigma}{\partial x} = \left(1-\sigma(x)\right)\sigma(x)$$



Patterns in Backflow of the Gradient

- **add**
  - Gradient distributor
- **max**
  - Gradient router
- **mul**
  - Gradient switcher



Backpropagation

- Function Principle
- Generalization to Vectors

Generalization to Vectors

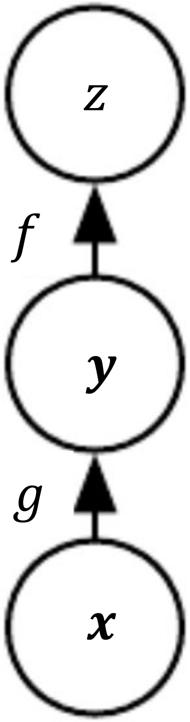
- Suppose $\boldsymbol{x} \in \mathbb{R}^m, \boldsymbol{y} \in \mathbb{R}^n$
  - g maps from $\mathbb{R}^m$ to $\mathbb{R}^n$ and
  - f maps from $\mathbb{R}^n$ to $\mathbb{R}$
- If $\boldsymbol{y} = g(\boldsymbol{x})$ and $z = f(\boldsymbol{y})$, then

$$\frac{\partial z}{\partial x_i} = \sum_j \frac{\partial z}{\partial y_i} \cdot \frac{\partial y}{\partial x_i}$$

- Or, in vector notation:

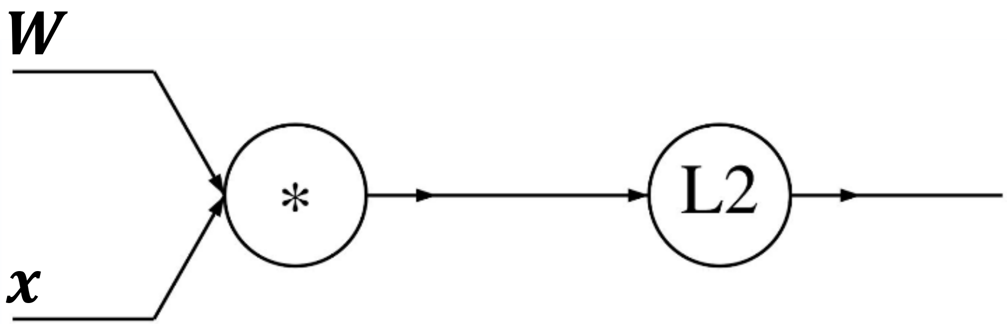
$$\nabla_x z = \left(\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right)^T \nabla_{\boldsymbol{y}} z$$

- That is the product of the Jacobian matrix $\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{y}}$ and the gradient vector $\nabla_{\boldsymbol{y}} z$.



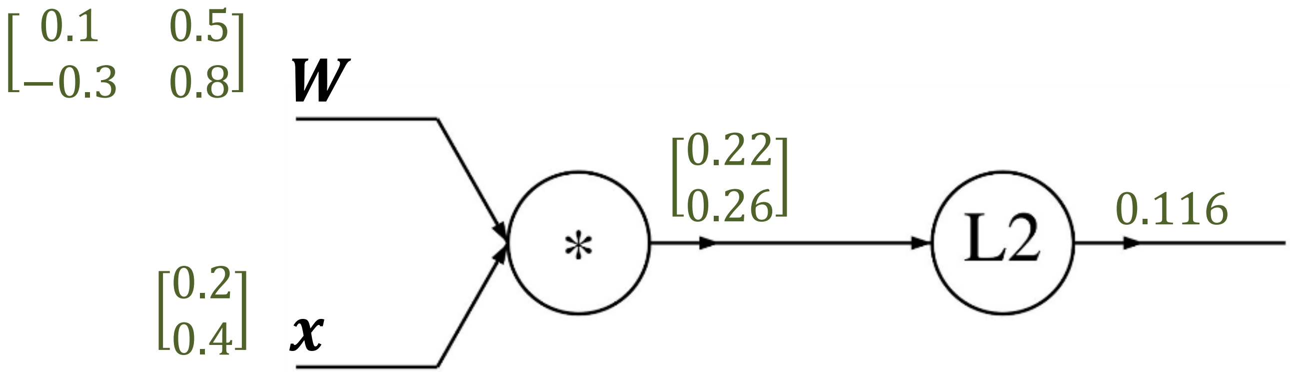
Vectorized Example

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



Vectorized Example

$$f(\boldsymbol{x}, \boldsymbol{W}) = \|\boldsymbol{W}\boldsymbol{x}\|^2 = \sum_i^n (\boldsymbol{W}\boldsymbol{x})_i^2$$

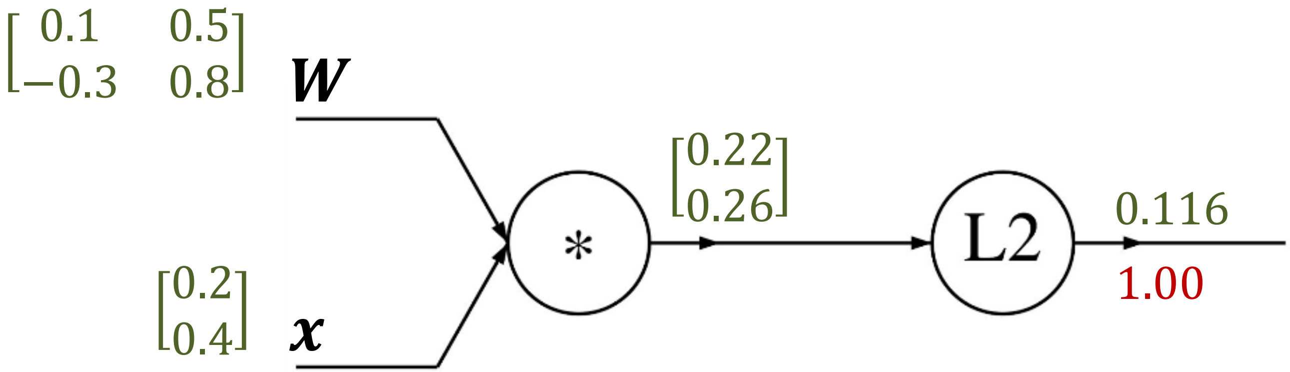


$$q = Wx = \begin{pmatrix} W_{1,1}x_1 & \cdots & W_{1,n}x_n \\ \vdots & \ddots & \vdots \\ W_{n,1}x_1 & \cdots & W_{n,n}x_n \end{pmatrix}$$

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

Vectorized Example

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

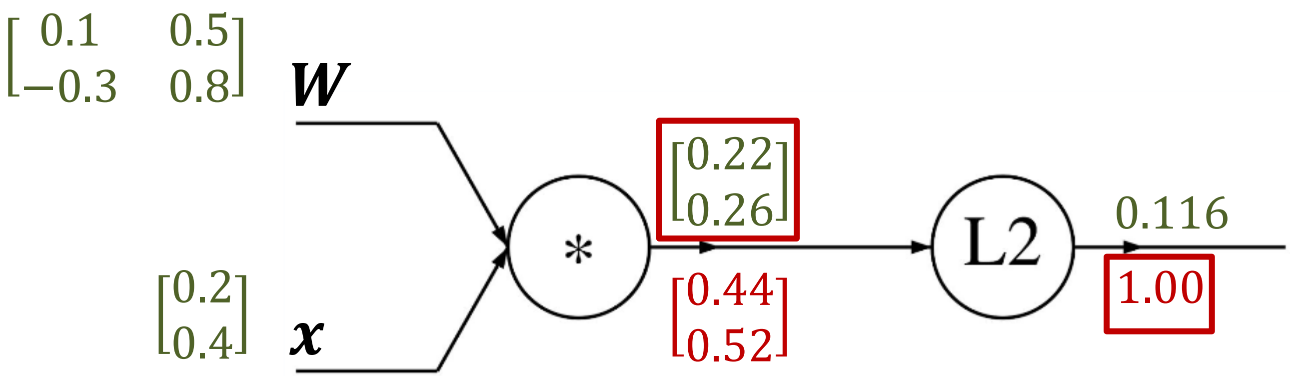


$$q = Wx = \begin{pmatrix} W_{1,1}x_1 & \cdots & W_{1,n}x_n \\ \vdots & \ddots & \vdots \\ W_{n,1}x_1 & \cdots & W_{n,n}x_n \end{pmatrix}$$

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

Vectorized Example

$$f(\boldsymbol{x}, \boldsymbol{W}) = \|\boldsymbol{W}\boldsymbol{x}\|^2 = \sum_i^n (\boldsymbol{W}\boldsymbol{x})_i^2$$



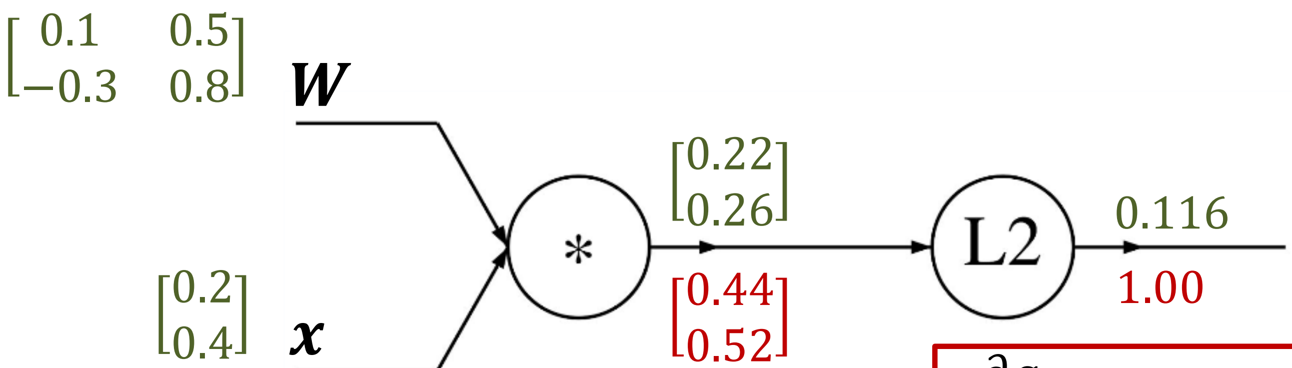
$$q = Wx = \begin{pmatrix} W_{1,1}x_1 & \cdots & W_{1,n}x_n \\ \vdots & \ddots & \vdots \\ W_{n,1}x_1 & \cdots & W_{n,n}x_n \end{pmatrix}$$

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

$$\boxed{\begin{aligned} \frac{\partial f}{\partial q_i} &= 2q_i \\ \nabla_q f &= 2q \end{aligned}}$$

Vectorized Example

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



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

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

$$= \sum_k (2q_k)(\mathbf{1}_{k=i} x_j)$$

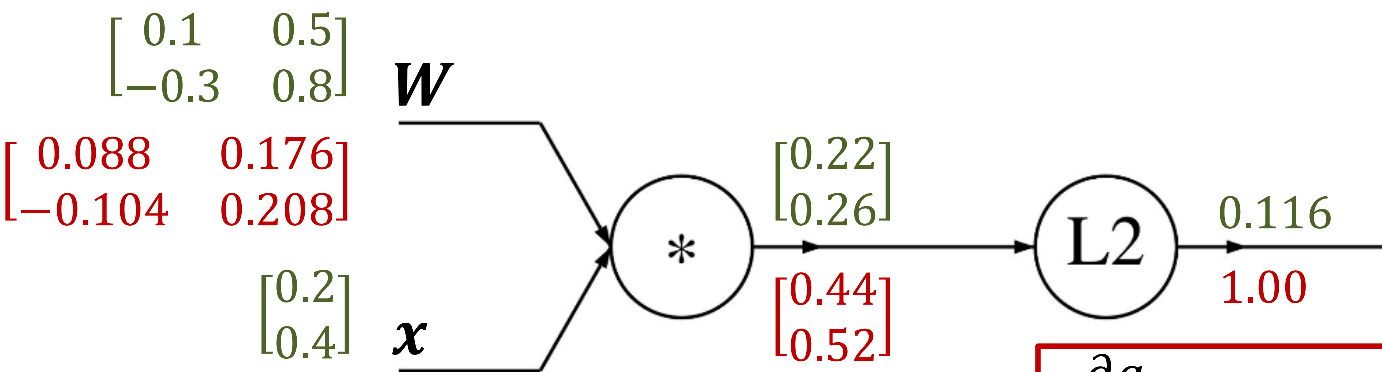
$$= 2q_i x_j$$

$$q = Wx = \begin{pmatrix} W_{1,1}x_1 & \cdots & W_{1,n}x_n \\ \vdots & \ddots & \vdots \\ W_{n,1}x_1 & \cdots & W_{n,n}x_n \end{pmatrix}$$

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

Vectorized Example

$$f(\boldsymbol{x}, \boldsymbol{W}) = \|\boldsymbol{W}\boldsymbol{x}\|^2 = \sum_i^n (\boldsymbol{W}\boldsymbol{x})_i^2$$



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

$$\frac{\partial f}{\partial W} = \sum \frac{\partial f}{} \frac{\partial q_k}{}$$

$$\boxed{\nabla_W f = 2\boldsymbol{q} \cdot \boldsymbol{x}^T}$$

$$= \sum_k (2q_k)(1_{k=i} x_j)$$

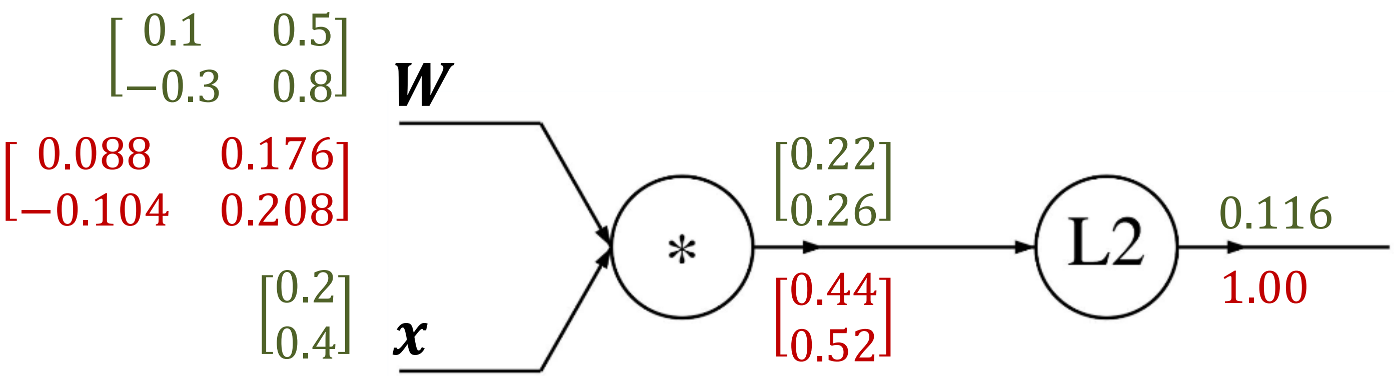
$$= 2 q_i x_j$$

$$q = Wx = \begin{pmatrix} W_{1,1}x_1 & \cdots & W_{1,n}x_n \\ \vdots & \ddots & \vdots \\ W_{n,1}x_1 & \cdots & W_{n,n}x_n \end{pmatrix}$$

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

Vectorized Example

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



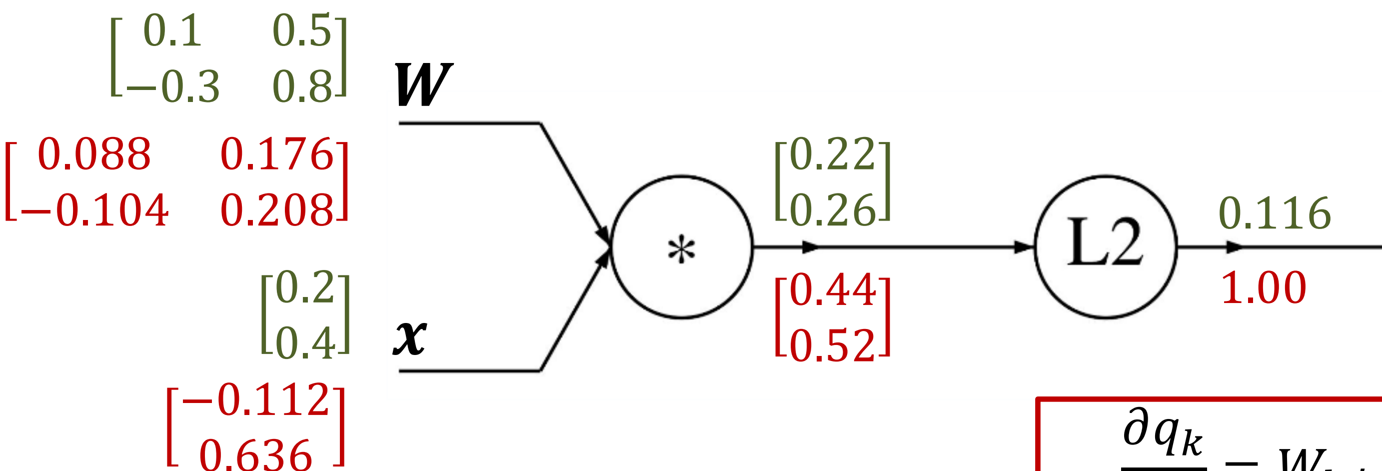
$$q = Wx = \begin{pmatrix} W_{1,1}x_1 & \cdots & W_{1,n}x_n \\ \vdots & \ddots & \vdots \\ W_{n,1}x_1 & \cdots & W_{n,n}x_n \end{pmatrix}$$

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

$$\boxed{\begin{aligned} \frac{\partial q_k}{\partial x_i} &= W_{k,i} \\ \frac{\partial f}{\partial x_i} &= \sum_k \frac{\partial f}{\partial q_k} \frac{\partial q_k}{\partial x_i} \\ &= \sum_k 2 q_k W_{k,i} \end{aligned}}$$

Vectorized Example

$$f(\boldsymbol{x}, \boldsymbol{W}) = \|\boldsymbol{W}\boldsymbol{x}\|^2 = \sum_i^n (\boldsymbol{W}\boldsymbol{x})_i^2$$



$$q = Wx = \begin{pmatrix} W_{1,1}x_1 & \cdots & W_{1,n}x_n \\ \vdots & \ddots & \vdots \\ W_{n,1}x_1 & \cdots & W_{n,n}x_n \end{pmatrix}$$

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

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

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

$$\boxed{\nabla_{\boldsymbol{x}} f = 2\boldsymbol{W}^T\boldsymbol{q}}$$

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

Two approaches to backpropagation

1. **Symbol-to-number differentiation**
   - Take a computational graph and a set of numerical values for inputs to the graph
   - Return a set of numerical values describing gradient at those input values
   - Used by libraries: Torch and Caffe

2. **Symbol-to-symbol differentiation**
   - Take a computational graph
   - Add additional nodes to the graph that provide a symbolic description of desired derivatives
   - Used by libraries: Theano and Tensorflow

Symbol-to-symbol Derivatives

- To compute derivative using this approach, backpropagation does not need to ever access any actual numerical values

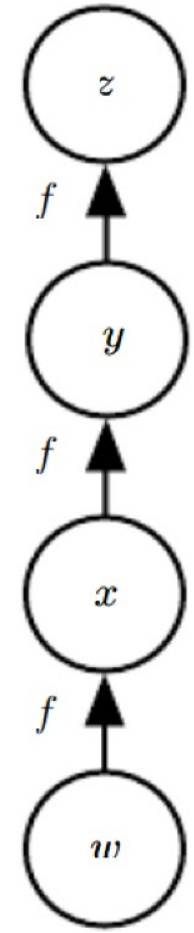
- Instead it adds nodes to a computational graph describing how to compute the derivatives

- A generic graph evaluation engine can later compute derivatives for any specific numerical values

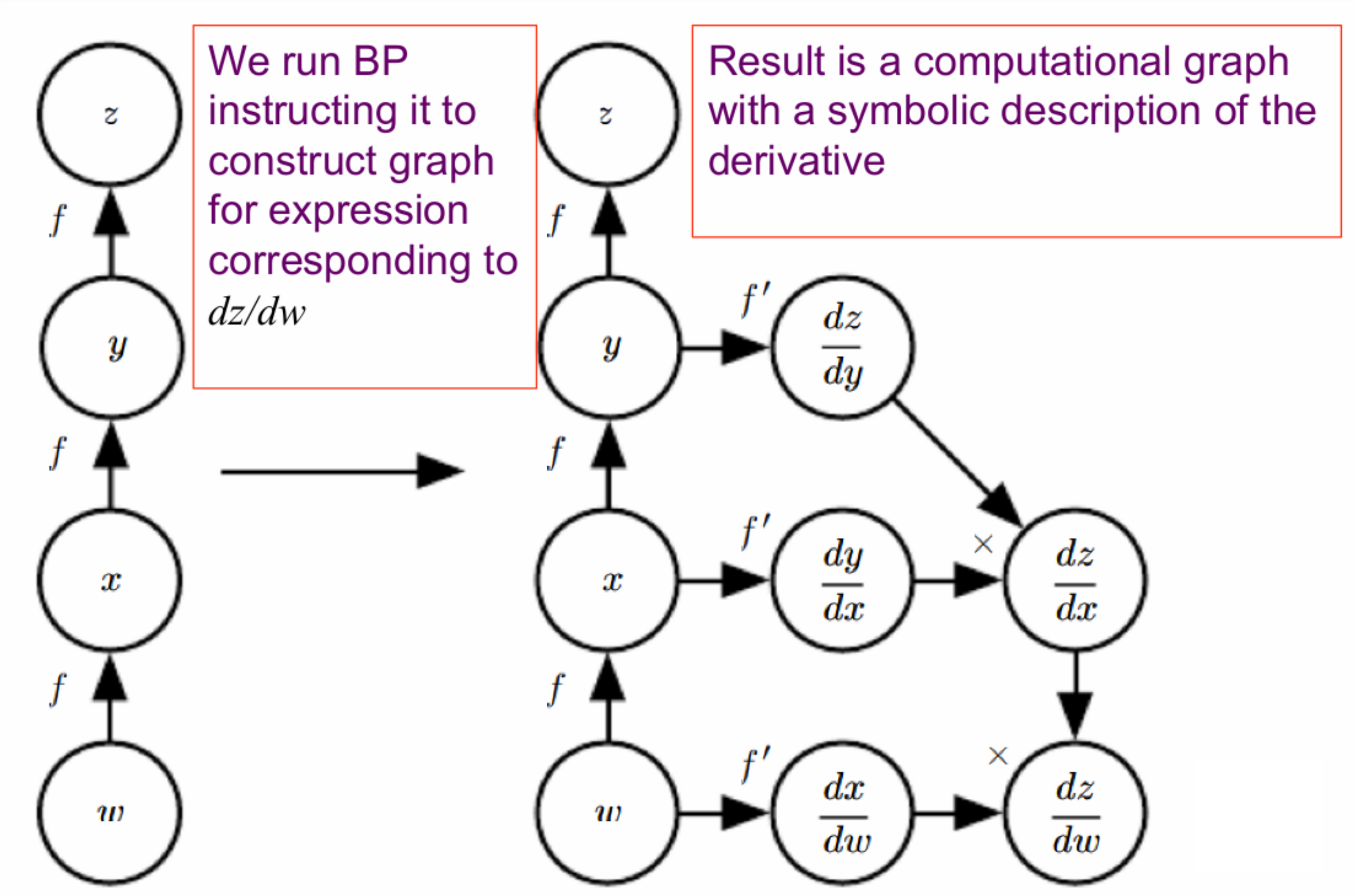
Example

- Consider the following function:

$$z = f\left(f\left(f(w)\right)\right)$$



Symbol-to-Symbol Derivative Computation



Advantages of Approach

- Derivatives are described in the same language as the original expression
- Because the derivatives are just another computational graph, it is possible to run back-propagation again
  - Differentiating the derivatives
  - Yields higher-order derivatives

What is Back-Propagation and what not!

- Often simply called backprop
  - Allows information from the cost to flow back through network to compute gradient

- The backpropagation algorithm does this using a simple and inexpensive procedure (and some optimizations, like dynamic programming to avoid evaluating the same expression twice)

- Backpropagation **is not Learning**
  - Only refers to the method for computing gradients
  - Needs to be coupled with a learning algorithm, e.g., stochastic gradient descent
  - Backprob is NOT specific to Deep Learning