SDU Summer School

Deep Learning

Summer 2018

Deep Feedforward Networks

Deep Feedforward Networks

- **PART I**
  - Feedforward Networks
  - Output Units
  - Hidden Units
  - Architecture Design
- **PART II**
  - **Gradient-Based Learning**
  - Backpropagation

Training Feedforward Networks

- We have seen how to construct a FNN
- We can input a data point into the FNN and receive a prediction
- We now need to define a function which judges the quality of our predictions and allows us to optimize the network, i.e., train the network.

Training Feedforward Networks

- We already two such error functions:

- Mean-Squared-Error (minimize):

$$J(\theta) = \frac{1}{n} \sum_{i}^{n} \left\| y^{(i)} - \hat{y}^{(i)} \right\|^2$$

- For Logistic Regression we have seen the MLE (maximize)

$$L(X, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(y^{(i)} | \boldsymbol{x}^{(i)}; \boldsymbol{\theta})$$

Cross Entropy Loss

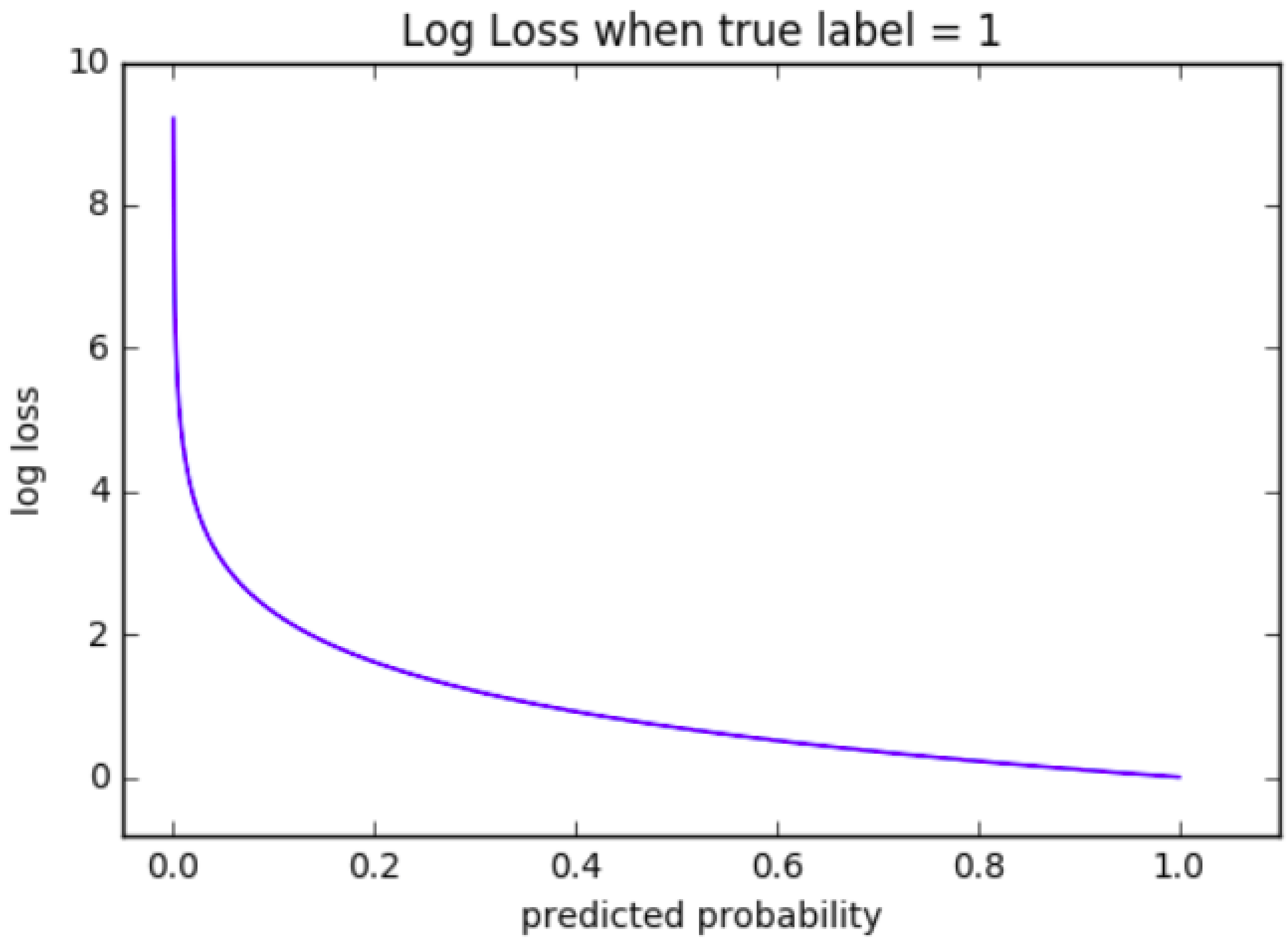
- For Neural Networks, we usually us the cross-entropy loss
- Minimizing the cross-entropy loss corresponds to maximize the log likelihood:

$$J(\boldsymbol{\theta}) = -\mathbb{E}[p(y|\boldsymbol{x};\boldsymbol{\theta})]$$

- In case of our 2 logistic regression:

$$-\frac{1}{n}\sum_{i=1}^{n}\left[y^{(i)}\log h_{\boldsymbol{\theta}}(\boldsymbol{x}) \;+\left(1-y^{(i)}\right)\log\left(1-h_{\boldsymbol{\theta}}(\boldsymbol{x})\right)\right]$$

Cross Entropy Loss



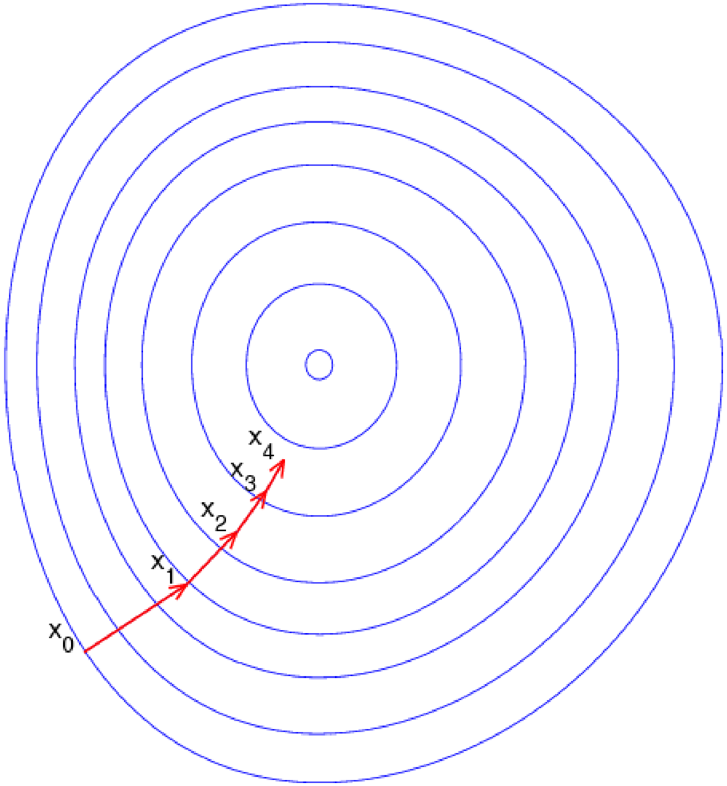
Our Recipe

- We have the model defined
  - By constructing the neural network
- We have a loss function
  - For example the cross-entropy
- We have a goal:
  - Modify the model parameters such that we minimize the loss function
- How to minimize a function?
  - Find the nulls of the derivate?
  - Not possible, function too complicated.

Gradient Based Learning

The Central Idea

- Update the model parameters following the steepest slope



More Mathematically

- Suppose function $y = f(x)$
- Derivative of function denoted: $f'(x)$ or as dy/dx
  - Derivative $f'(x)$ gives the slope of $f(x)$ at point x
  - It specifies how to scale a small change in input to obtain a corresponding change in the output:

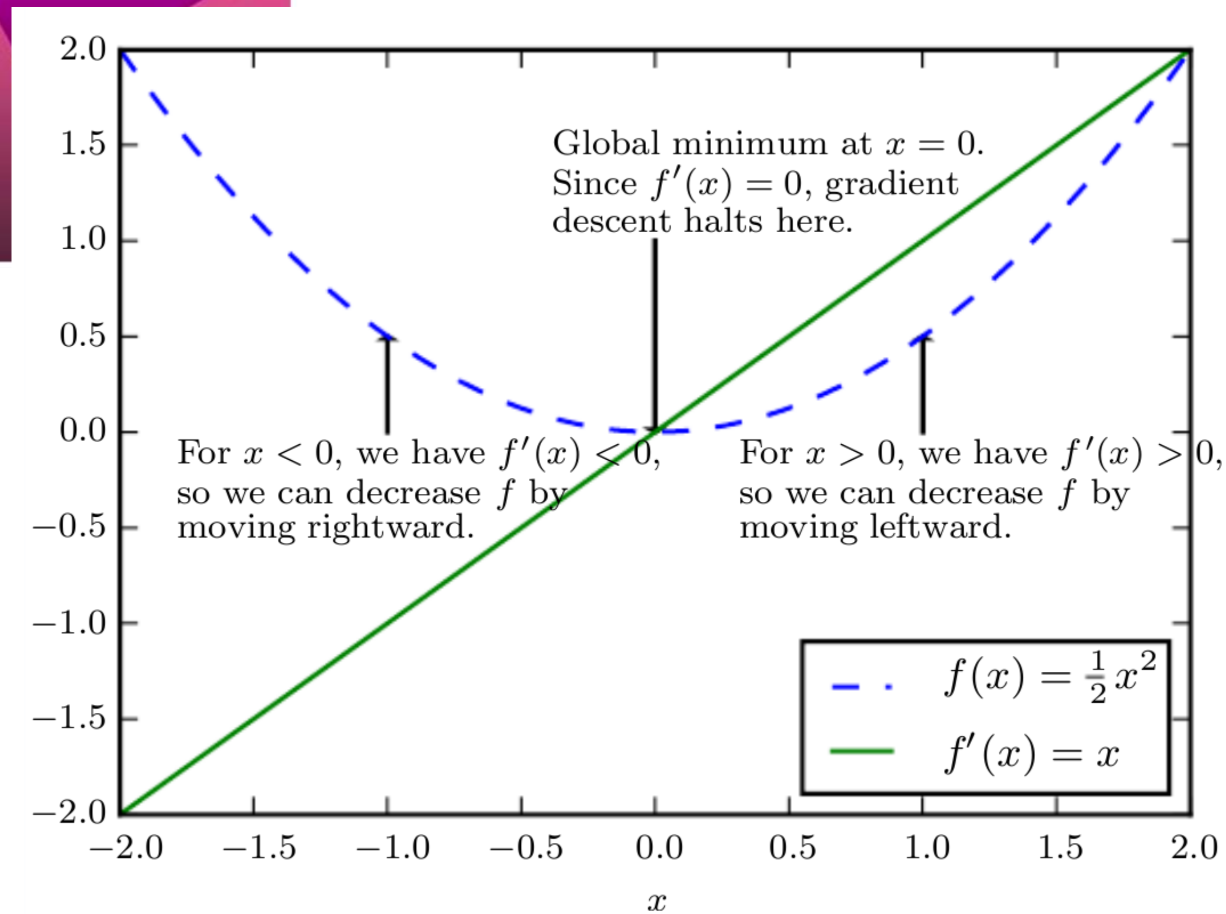
$$f(x+\varepsilon) \approx f(x) + \varepsilon f'(x)$$

- We know that

$$f(x - \varepsilon \operatorname{sign}(f'(x)))$$

is less than $f(x)$ for small ε.
  - Thus we can reduce f (x) by moving x in small steps with opposite sign of derivative
  - This technique is called gradient descent (Cauchy 1847)

Usage of Derivates



Functions with multiple inputs

- Need partial derivatives

$$\frac{\partial}{\partial x_i} f(\boldsymbol{x})$$

- Measures how f changes as only variable x_i increases at point x

- Gradient is vector containing all of the partial derivatives denoted with $\nabla_{\boldsymbol{x}} f(\boldsymbol{x})$
  - Element i of the gradient is the partial derivative of f wrt x_i
  - Critical points are where every element of the gradient is equal to zero
  - A function can be minimized when moving in the direction opposite to the gradient

Functions with multiple inputs

- We can decrease f by moving in the direction of the negative gradient vector

- Steepest descent proposes a new point

$$\boldsymbol{x}' = \boldsymbol{x} - \epsilon \nabla_{\boldsymbol{x}} f(\boldsymbol{x})$$

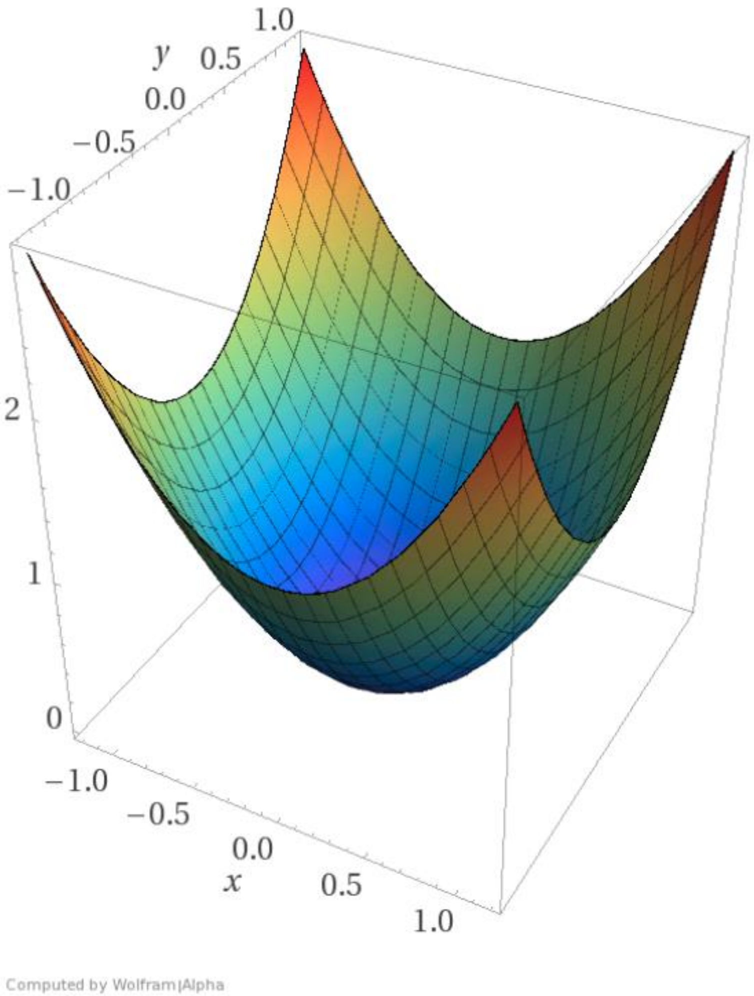
- With ϵ being the learning rate (there are many methods of defining ϵ)
- Ascending an objective function of discrete parameters is called hill climbing

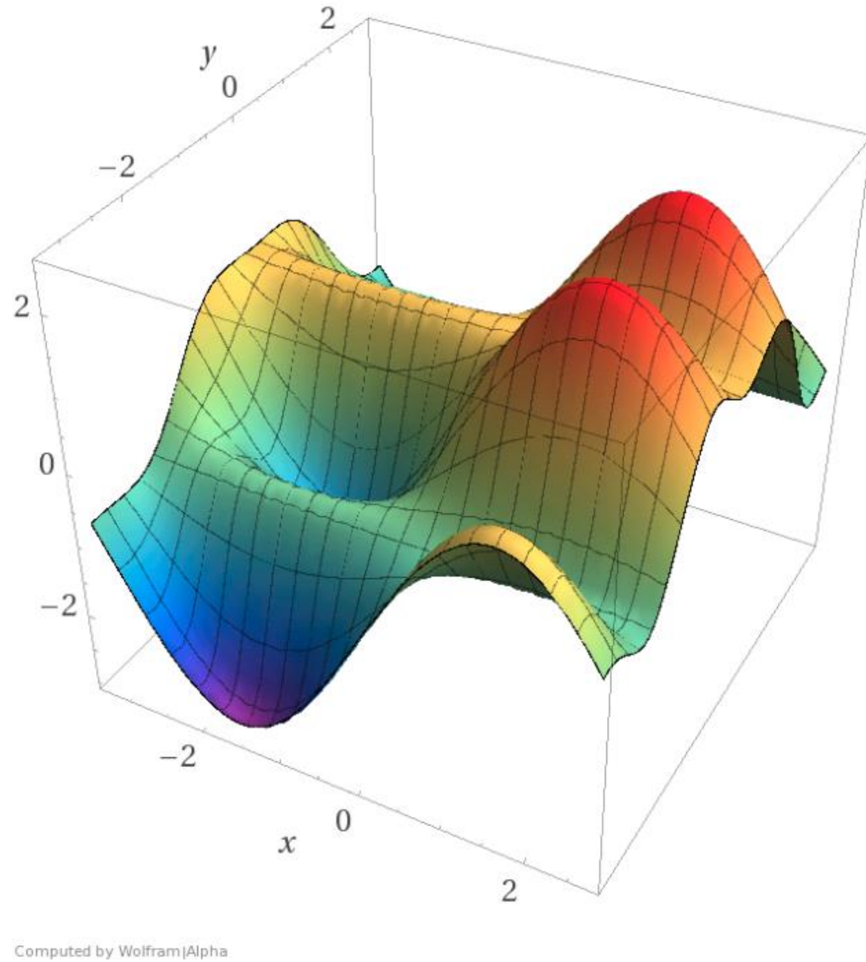
Specialties of Deep Learning

- Neural Network training not different from ML models with gradient descent. The components are needed:
  1. optimization procedure, e.g., gradient descent
  2. cost function, e.g., MLE
  3. model family, e.g., linear with basis functions

- Difference: **nonlinearity causes non-convex loss**
  - Use iterative gradient-based optimizers that merely drives cost to low value
  - No guarantees in comparison to convex optimizations
  - The initialization matters

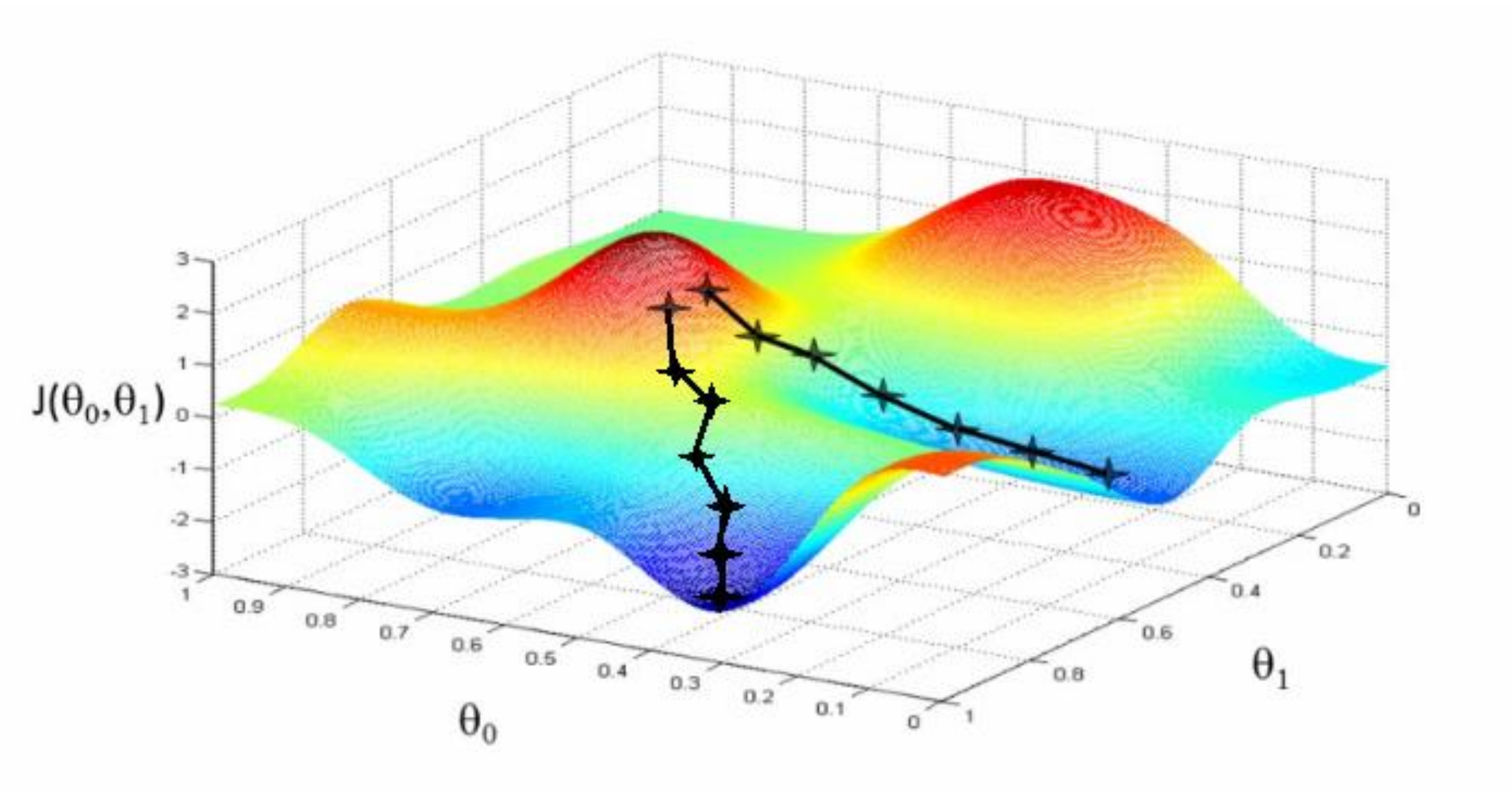
Convex vs. Non-Convex





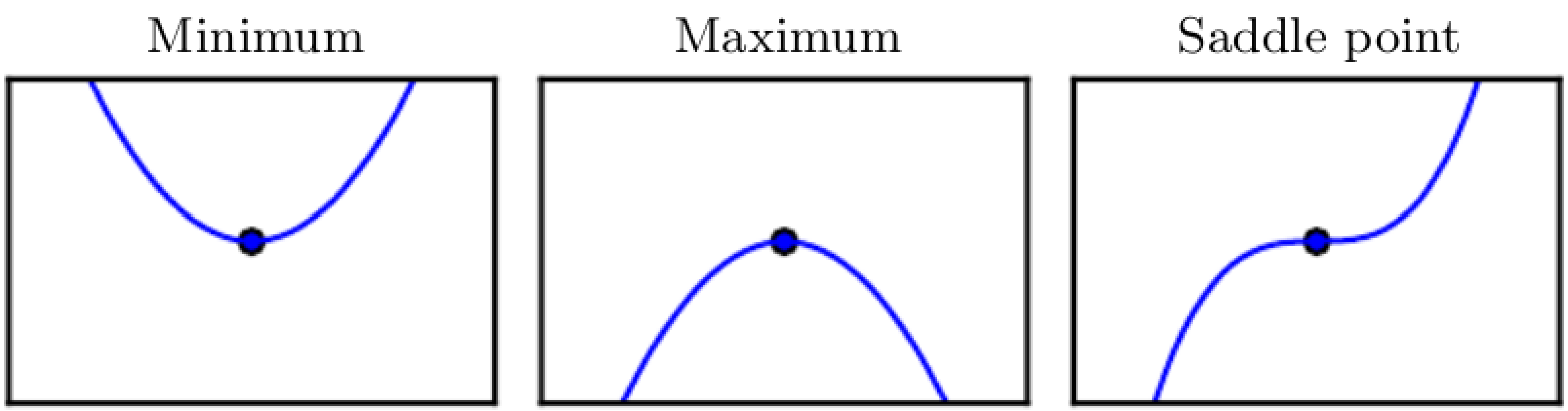
- https://www.matroid.com/blog/post/the-hard-thing-about-deep-learning

Problem: We can end-up in local minima

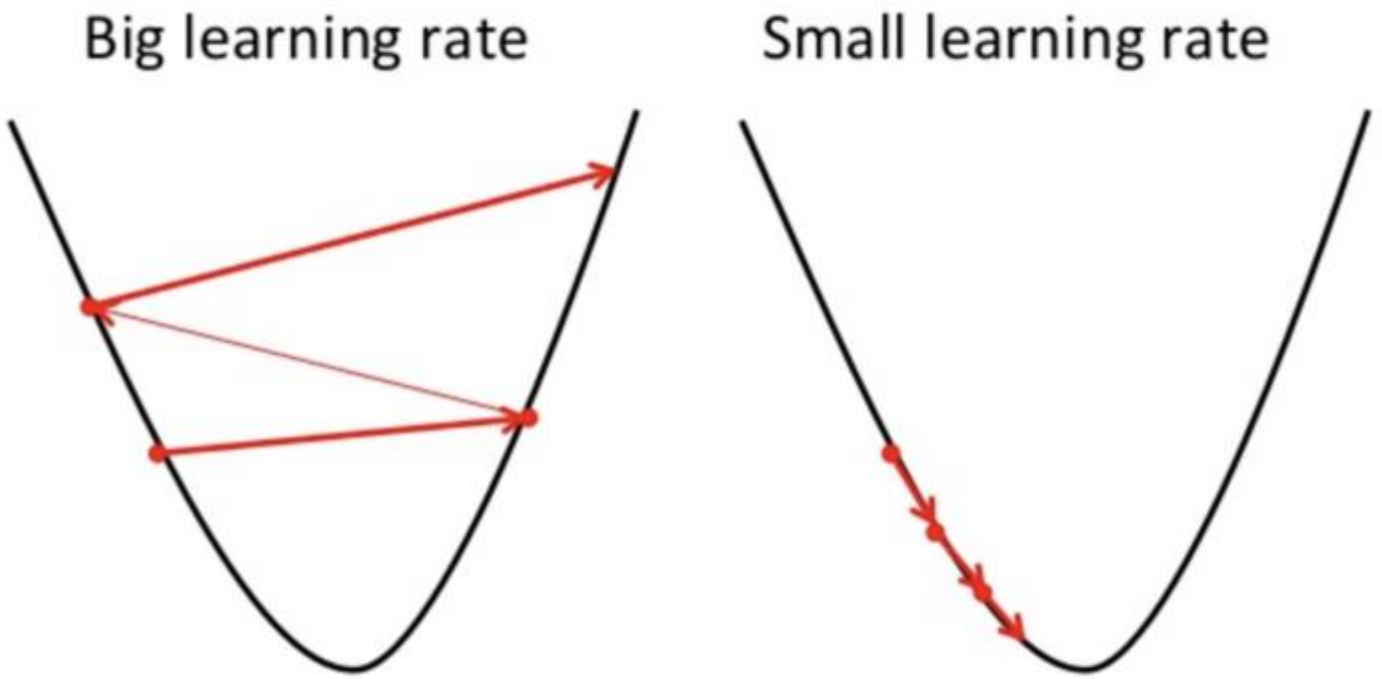


Problem: Stationary points, Local Optima

- When $f'(x) = 0$ derivative provides no information about direction of move
- Points where $f'(x) = 0$ are known as stationary or critical points
  - **Local minimum/maximum**: a point where $f(x)$ lower/ higher than all its neighbors
  - **Saddle Points**: neither maxima nor minima



Problem: The Learning Rate



Convergence of Steepest Descent

- Steepest descent converges when every element of the gradient is zero

- Pure math way of life:
  - Find literally the smallest value of $f(x)$
  - Or maybe: find some critical point of $f(x)$ where the value is locally smallest

- **Deep learning way of life:**
  - **Decrease the value of $f(x)$ a lot**
  - **But we have a highly non-convex problem (because of the activation functions) => No guarantees!**

About the Gradient

- Gradient must be large and predictable enough to serve as good guide to the learning algorithm
- Functions that **saturate** (become very flat) undermine this
  - Because the gradient becomes very small
  - Happens when activation functions producing output of hidden/output units saturate

- **Negative log-likelihood** helps avoid saturation problem for many models
  - Many output units involve exp functions that saturate when its argument is very negative
  - Log function in Negative log likelihood cost function undoes exp of some units

Stochastic Gradient Descent (SGD)

- A recurring problem in machine learning:
  - large training sets are necessary for good generalization
  - but large training sets are also computationally expensive

- Nearly all deep learning is powered by one very important algorithm: **Stochastic Gradient Descent**

Insight of SGD

- Insight: Gradient descent based on only a sample (we don't have the universe as data) is an **expectation**
  - Expectation may be approximated using small set of samples

- In each step of SGD we can sample a minibatch of examples
$$B = \{\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(m')}\}$$
  - drawn uniformly from the training set
  - Minibatch size m' is typically chosen to be small: 1 to a hundred
  - Crucially m' is held fixed even if sample set is in billions
  - We may fit a training set with billions of examples using updates computed on only a hundred examples

SGD Estimate on minibatch

- Estimate of gradient is formed as

$$\boldsymbol{g} = \frac{1}{m'} \nabla_{\boldsymbol{\theta}} \sum_{i=1}^{m'} L\left(\boldsymbol{x}^{(i)}, y^{(i)}, \boldsymbol{\theta}\right)$$

using only the examples of the minibatch

- SDG then simply follows the estimated gradient downhill

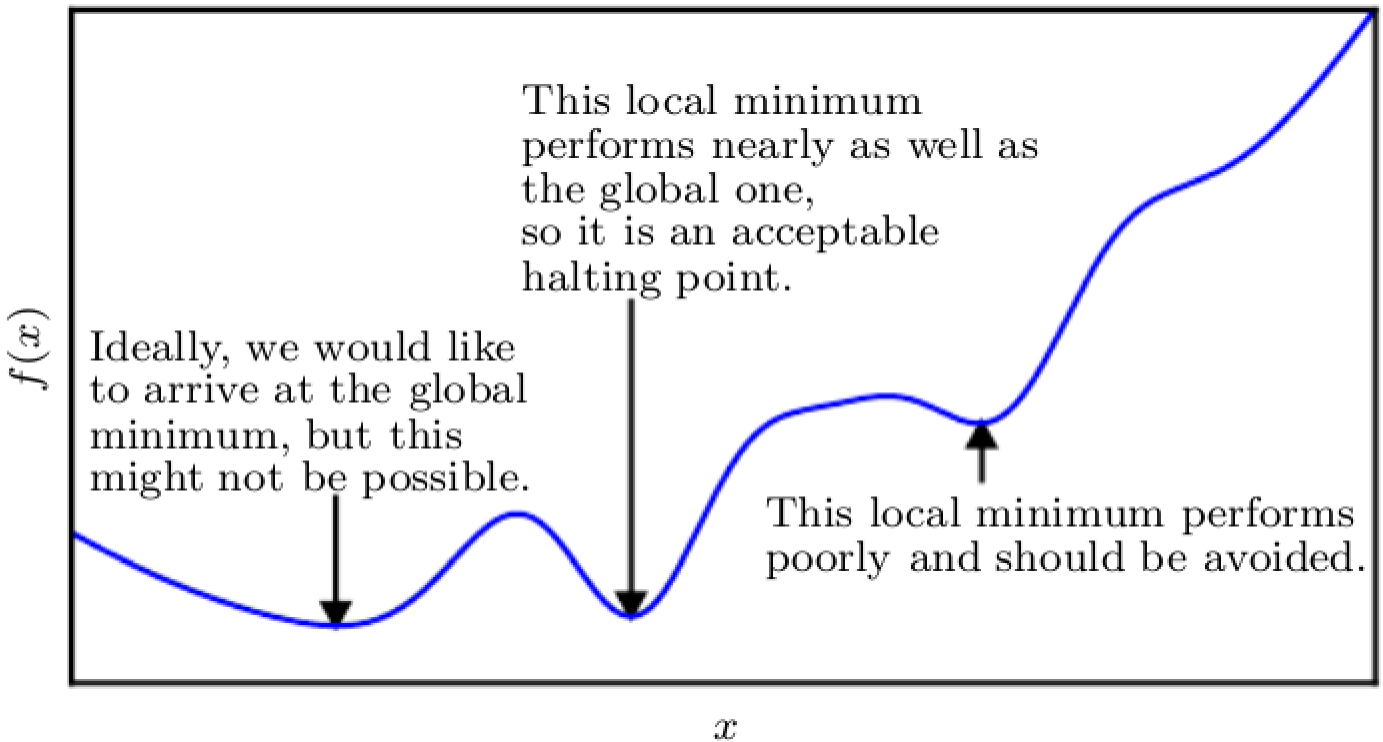
$$\boldsymbol{\theta}' = \boldsymbol{\theta} - \epsilon \boldsymbol{g}$$

How good is SGD?

- In the past gradient descent was regarded as slow and unreliable
- Application of gradient descent to non-convex optimization problems was regarded as unprincipled
- SGD is not guaranteed to arrive at even a local minimum in reasonable time

- **But it often finds a very low value of the cost function quickly enough**
- As $m \to \infty$ the model will eventually converge to its best possible test error before SGD has sampled every example

Good Enough in Practice



Different Optimizers

- Delta-bar-delta Algorithm
  - (Applicable to only full batch optimization)
  - If partial derivative of the loss wrt to a parameter remains the same sign, the learning rate should increase
  - If that partial derivative changes sign, the learning rate should decrease
- AdaGrad
  - Individually adapts learning rates of all params
  - By scaling them inversely proportional to the sum of the historical squared values of the gradient
- RMSProp
  - Modifies AdaGrad for a nonconvex setting
  - Change gradient accumulation into exponentially weighted moving average
  - Converges rapidly when applied to convex function

Do we have everything?

- We have the model defined
- We have a loss function
- We have the optimization goal
- We the optimization algorithm
  - Stochastic Gradient Descent

But what about the gradient?

Deep Feedforward Networks

- **PART I**
  - Feedforward Networks
  - Output Units
  - Hidden Units
  - Architecture Design
- **PART II**
  - Gradient-Based Learning
  - **Backpropagation**

Chain Rule of Calculus

- Formula for computing derivatives of functions formed by composing other functions whose derivatives are known

- For example:

$$y = f\left(g\big(h(x)\big)\right) = f\left(g\big(h(w_0)\big)\right) = f\big(g(w_1)\big) = f(w_2) = w_3$$

- The chain rule gives:

$$\frac{dy}{dx} = \frac{dy}{dw_2}\frac{dw_2}{d_{w1}}\frac{dw_1}{dx}$$

Forward vs. Backward Mode

$$y = f\left(g\big(h(x)\big)\right) = f\left(g\big(h(w_0)\big)\right) = f\big(g(w_1)\big) = f(w_2) = w_3$$

- Forward Accumulation:

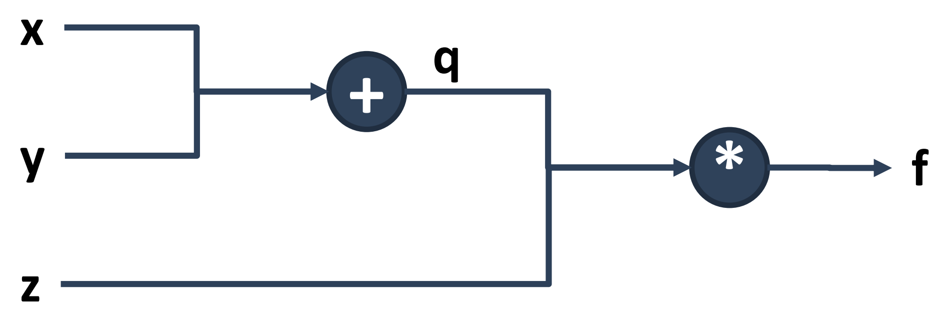
$$\frac{dw_i}{dx} = \frac{dw_i}{dw_{i-1}} \frac{dw_{i-1}}{dx} \quad with \;\; w_3 = y$$

- Reverse Accumulation:

$$\frac{dy}{dw_i} = \frac{dy}{dw_{i+1}} \frac{dw_{i+1}}{dw_i} \quad with \;\; w_0 = \mathrm{x}$$

Backpropagation: A Simple Example

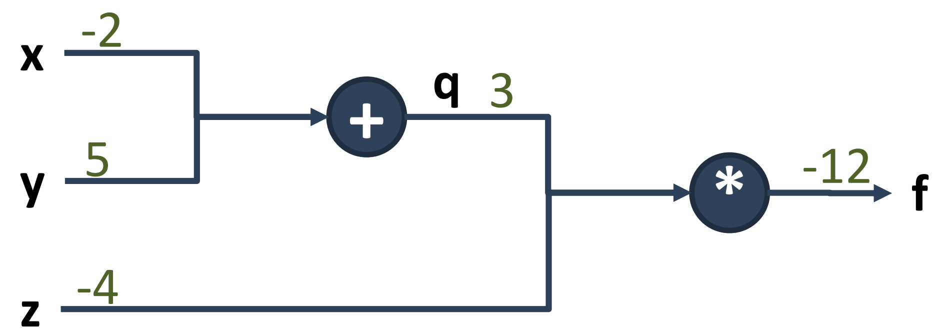
$$f(x, y, z) = (x + y)z$$



Backpropagation: A Simple Example

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

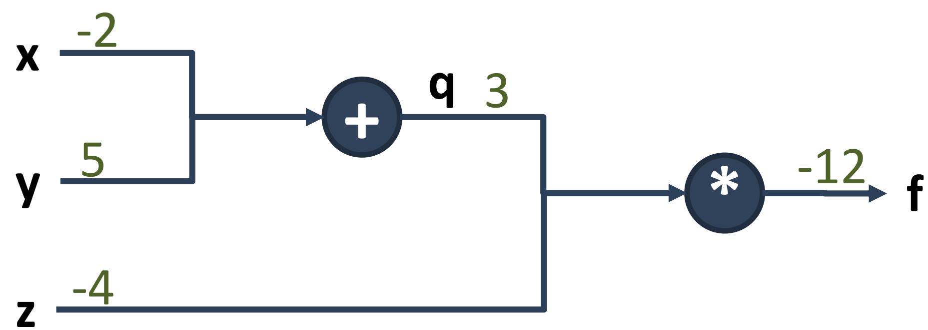
- With
  - $x = -2$
  - $y = 5$
  - $z = -4$



Backpropagation: A Simple Example

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

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



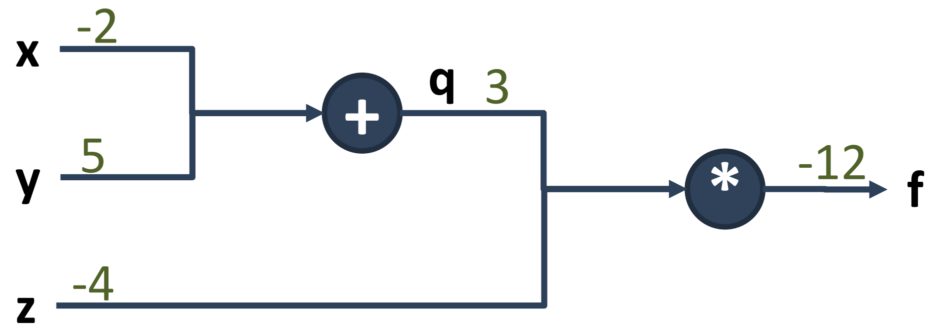
$$q = x + y; \quad \frac{\partial q}{\partial x} = 1; \ \frac{\partial q}{\partial y} = 1;$$

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

Backpropagation: A Simple Example

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

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



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

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

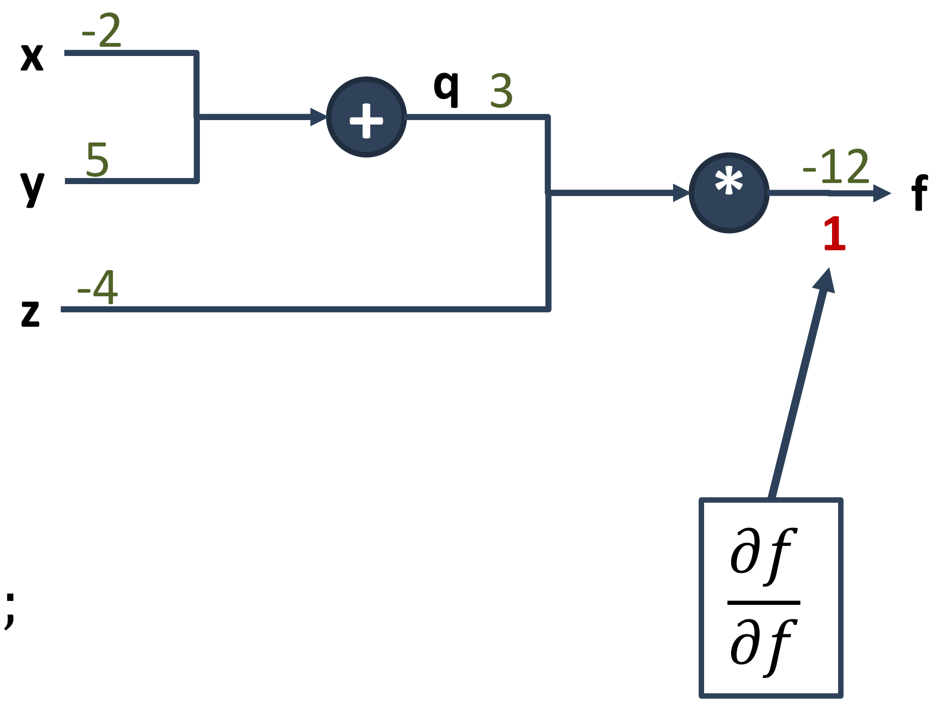
Looking for:

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

Backpropagation: A Simple Example

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

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



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

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

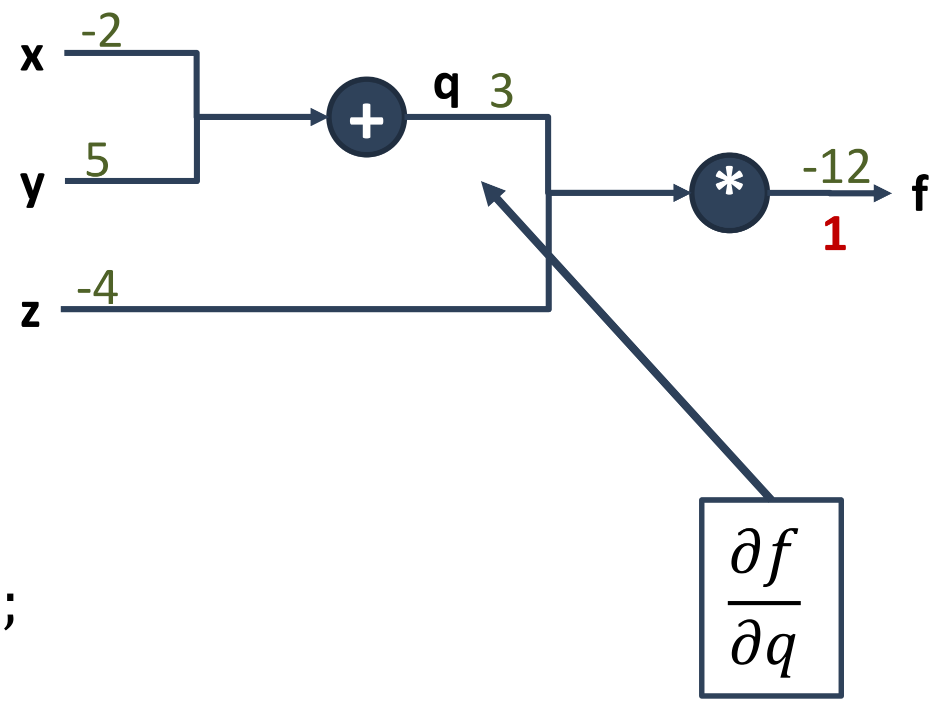
Looking for:

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

Backpropagation: A Simple Example

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

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



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

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

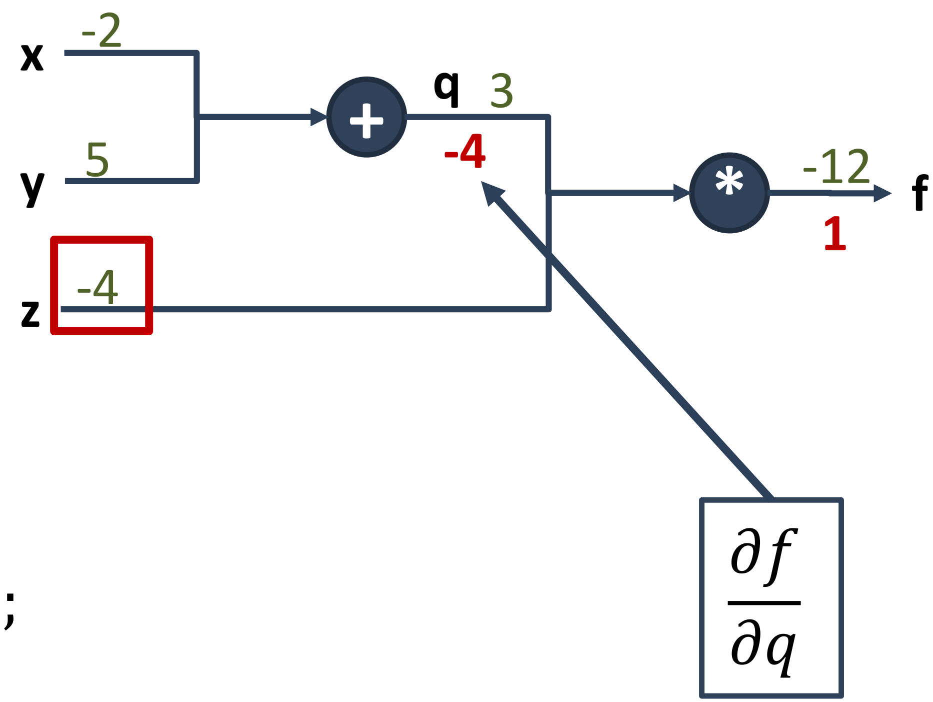
Looking for:

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

Backpropagation: A Simple Example

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

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



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

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

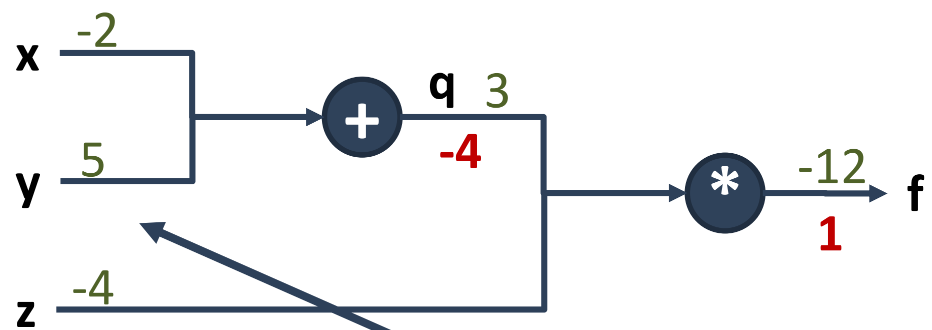
Looking for:

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

Backpropagation: A Simple Example

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

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



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

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

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

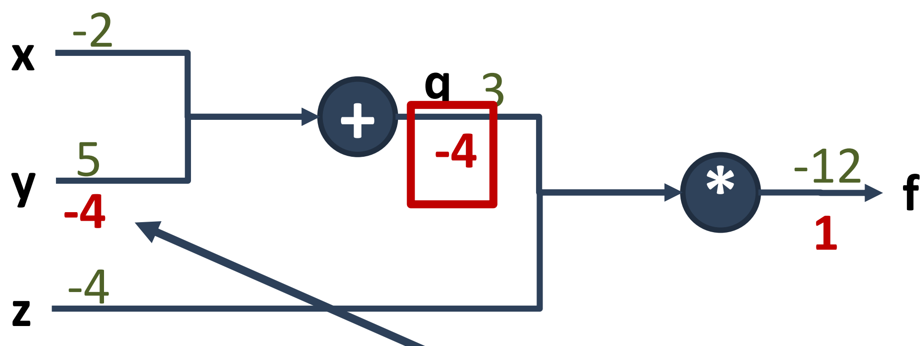
Looking for:

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

Backpropagation: A Simple Example

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

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



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

Chain Rule:

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

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

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

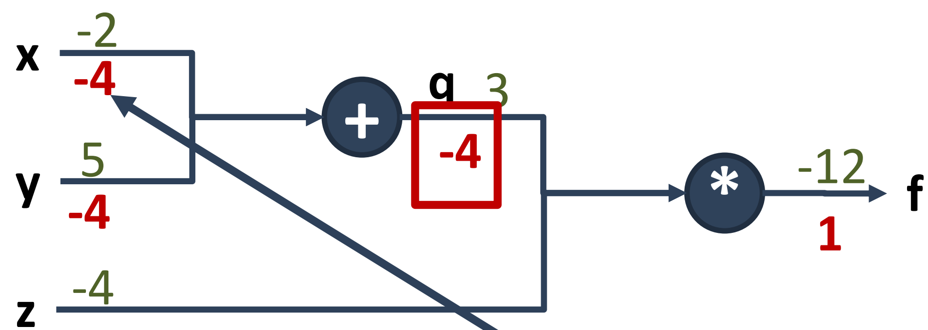
Looking for:

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

Backpropagation: A Simple Example

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

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



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

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

Chain Rule:

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

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

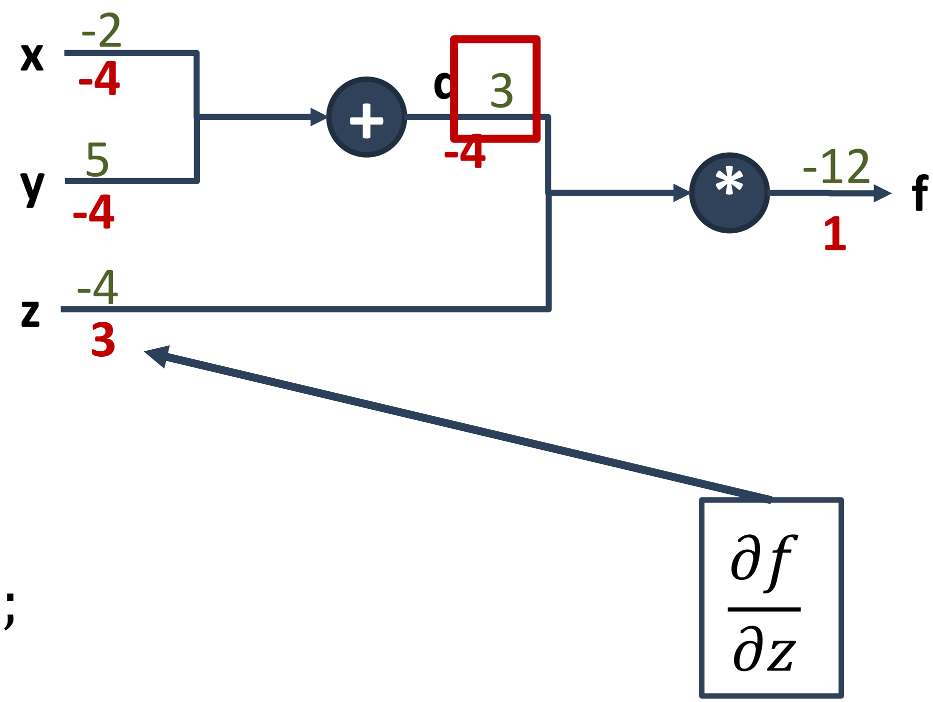
Looking for:

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

Backpropagation: A Simple Example

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

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



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

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

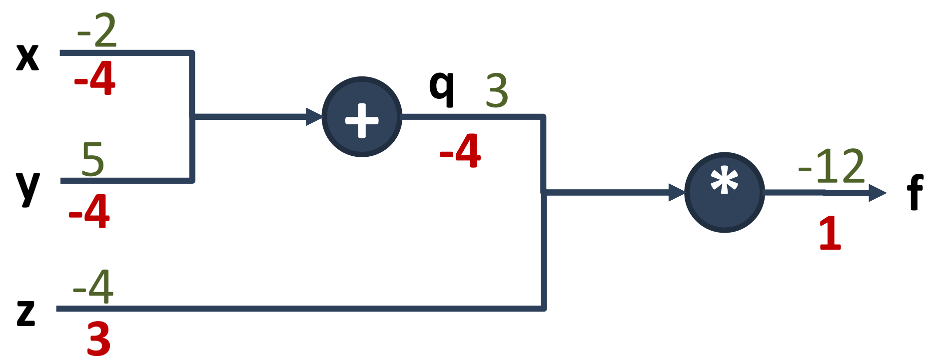
Looking for:

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

Backpropagation: A Simple Example

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

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



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

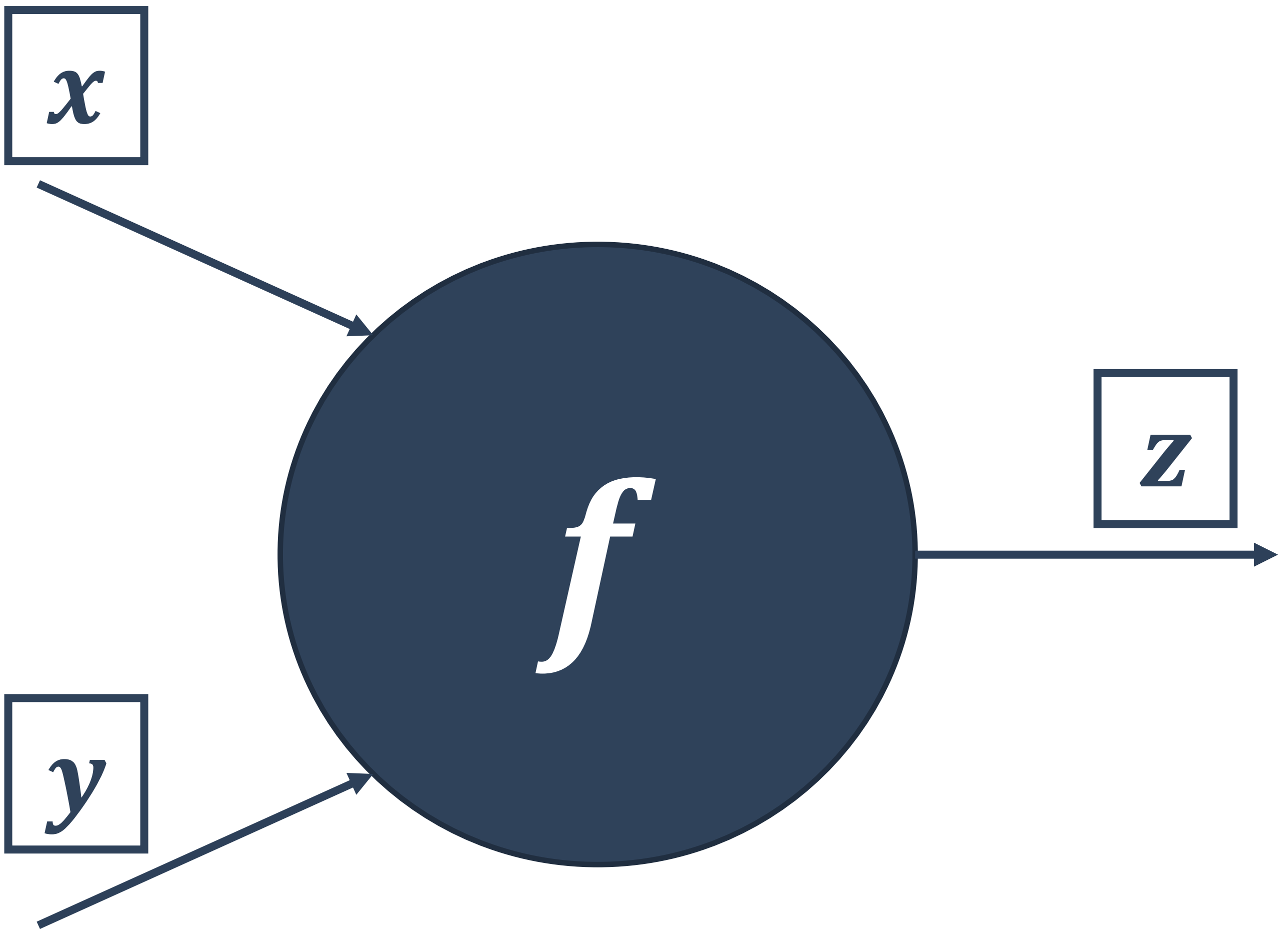
$$f = qz; \quad \frac{\partial f}{\partial q} = z; \ \frac{\partial f}{\partial z} = q;$$

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

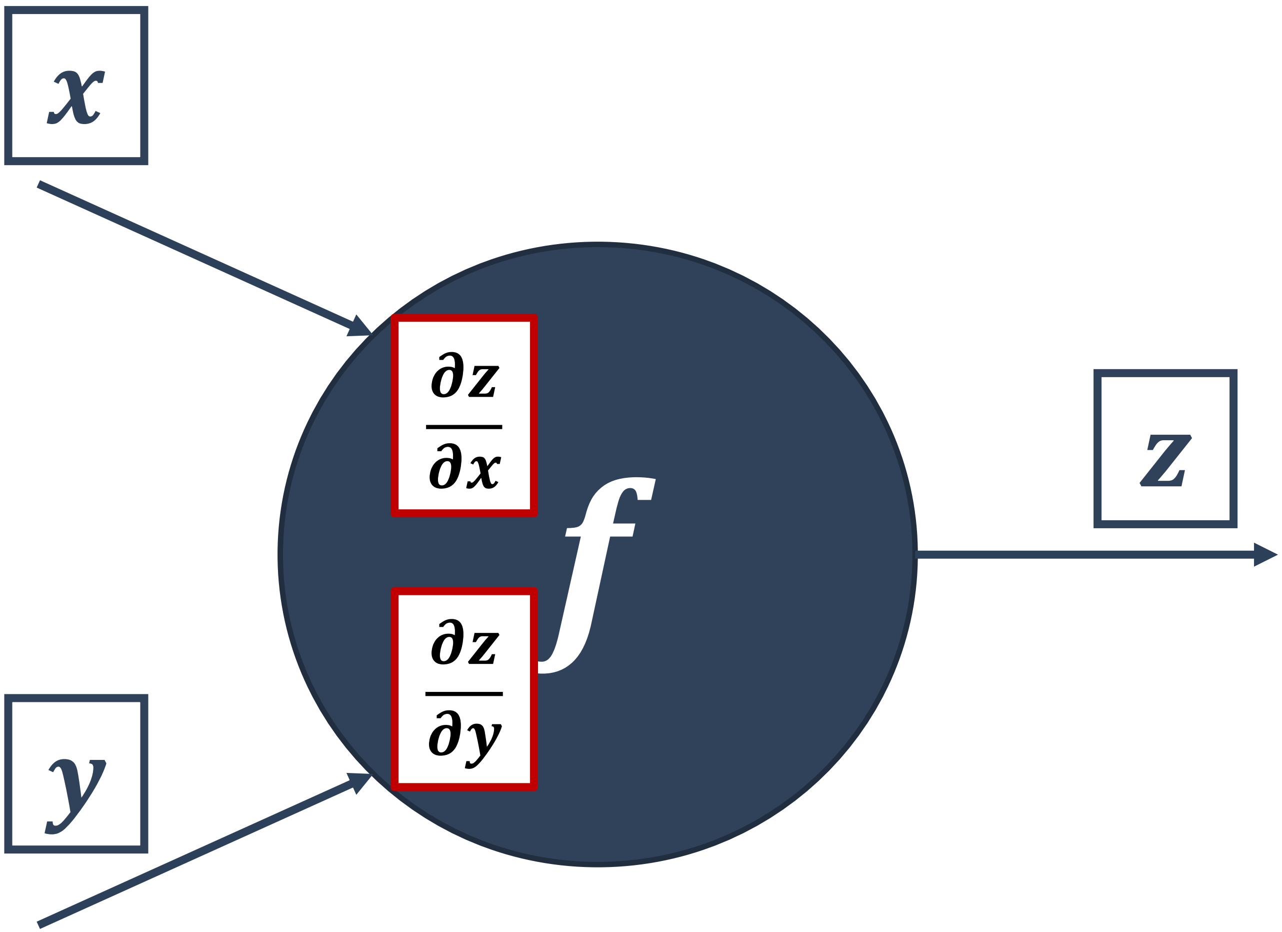
Looking for:

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

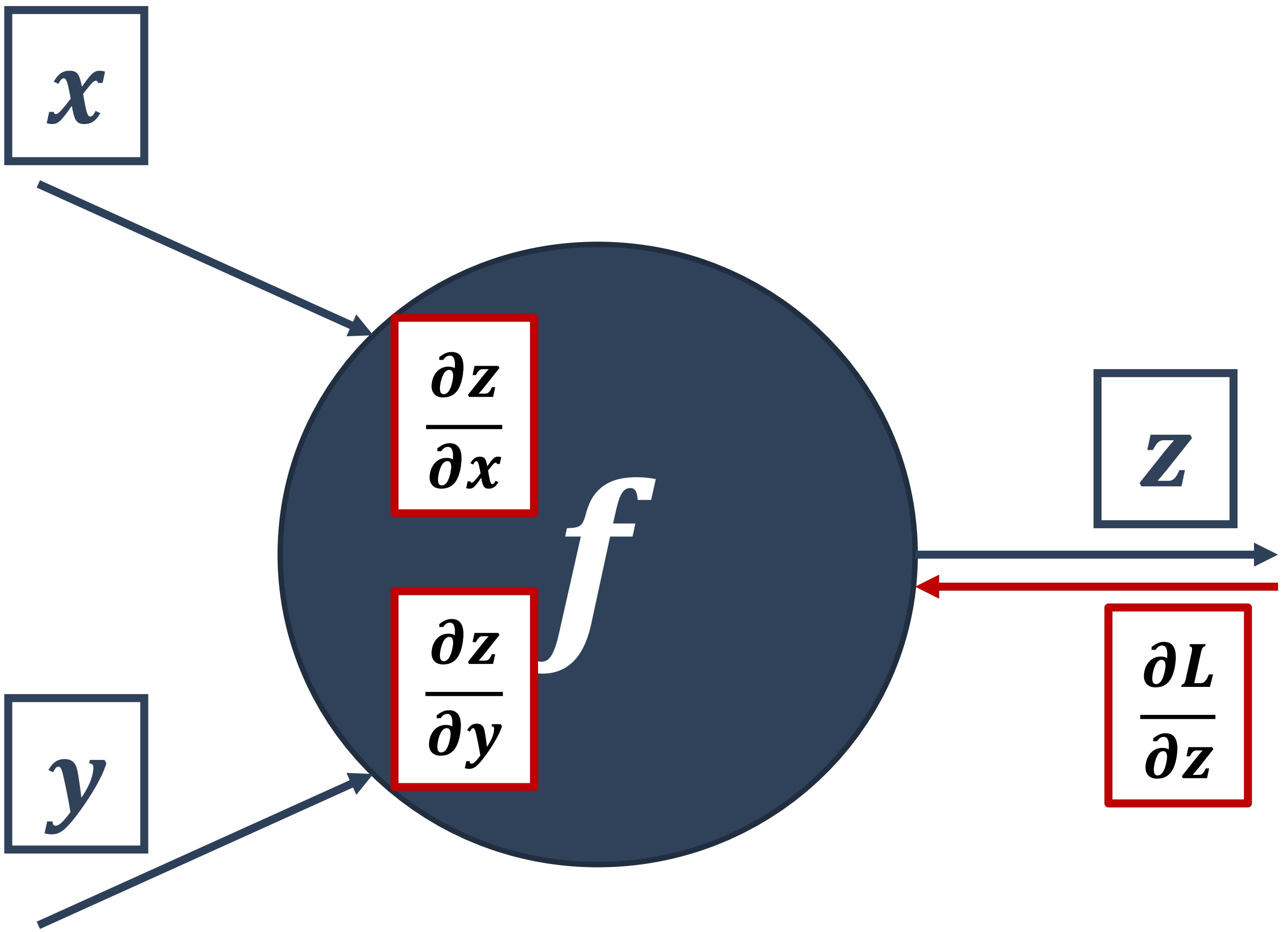
General Principle



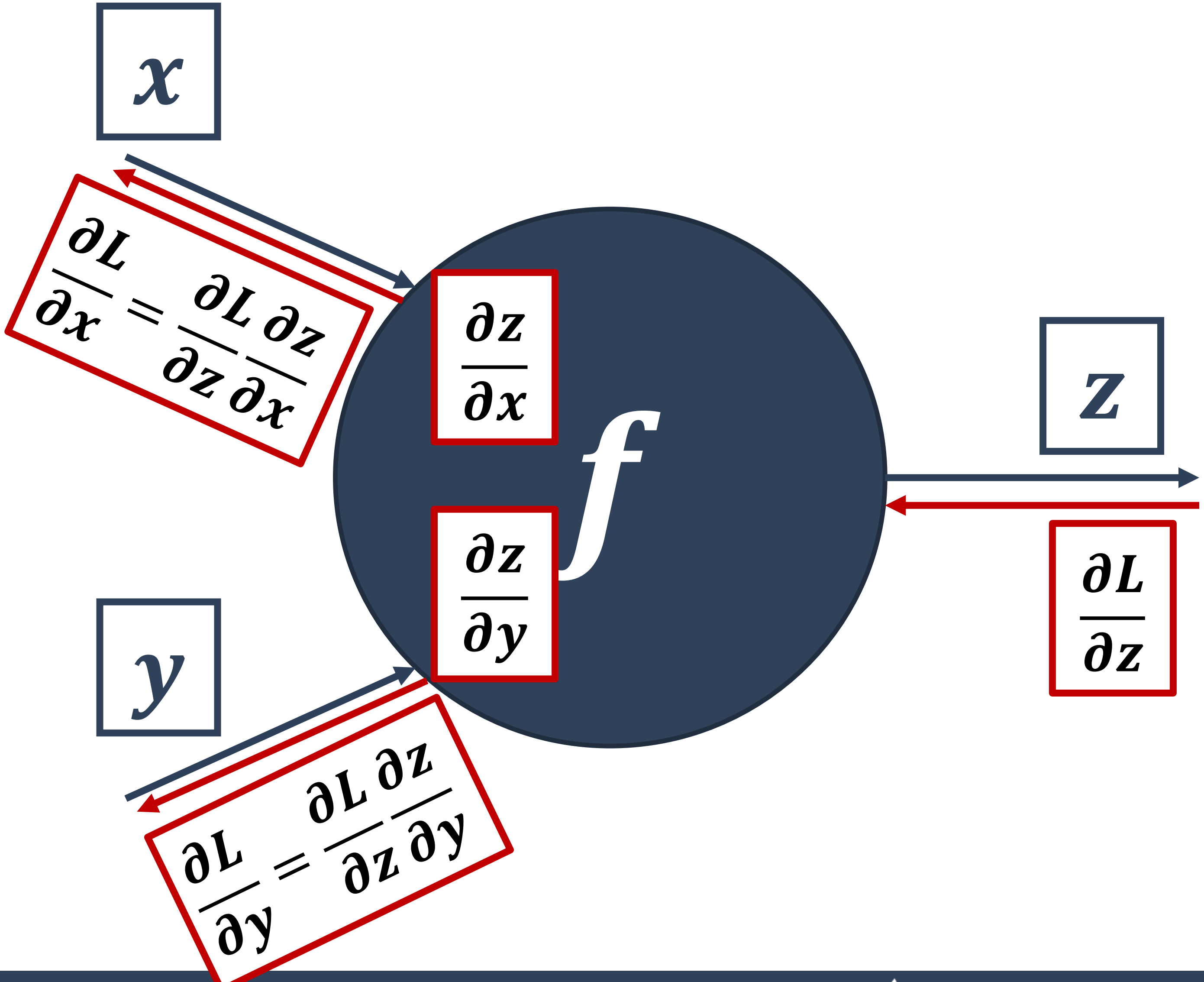
General Principle



General Principle

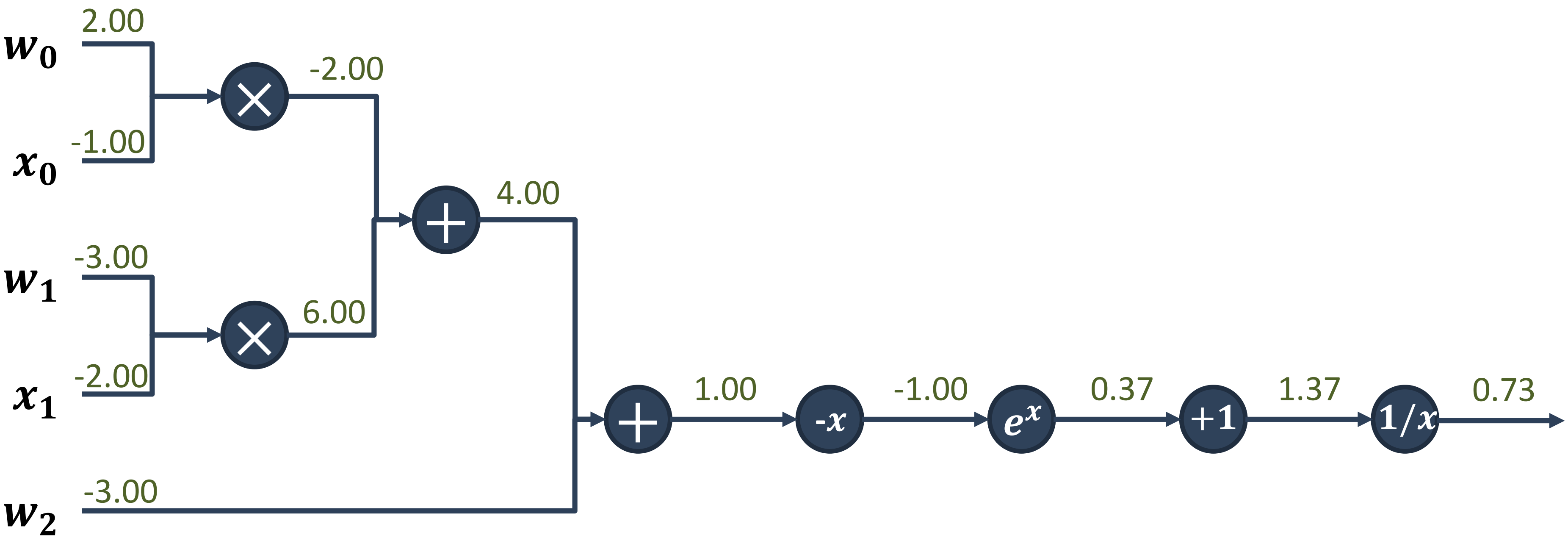


General Principle



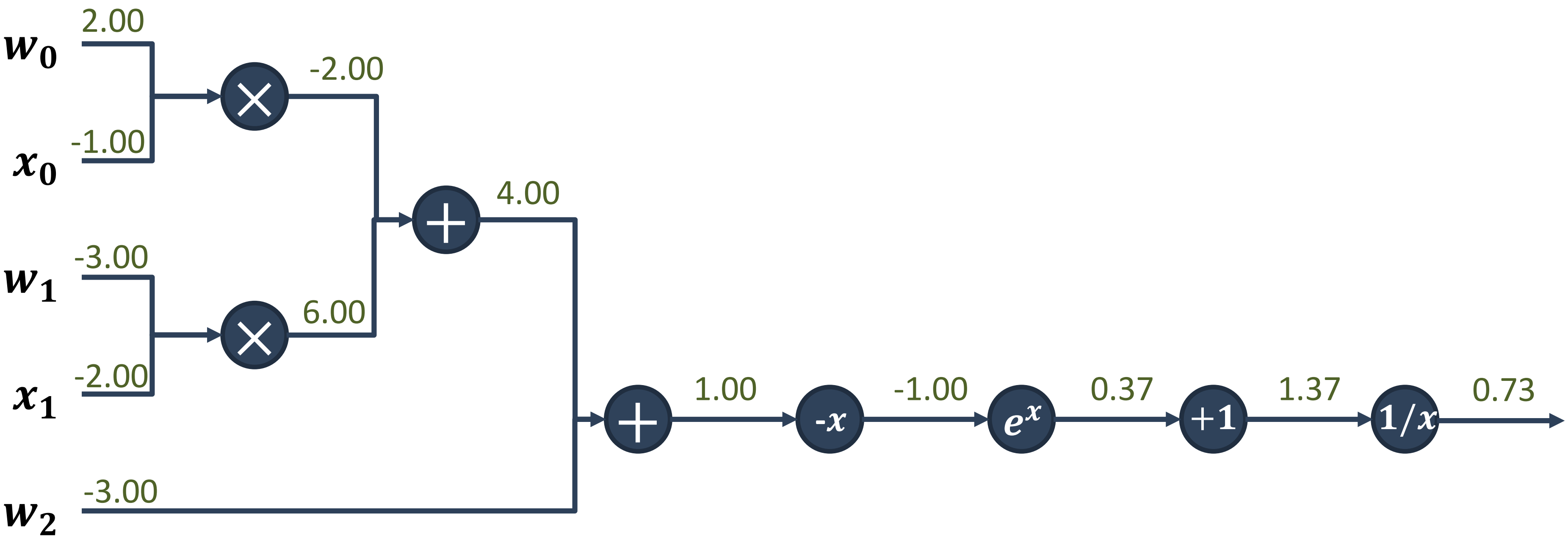
Different Example

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



Different Example

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



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

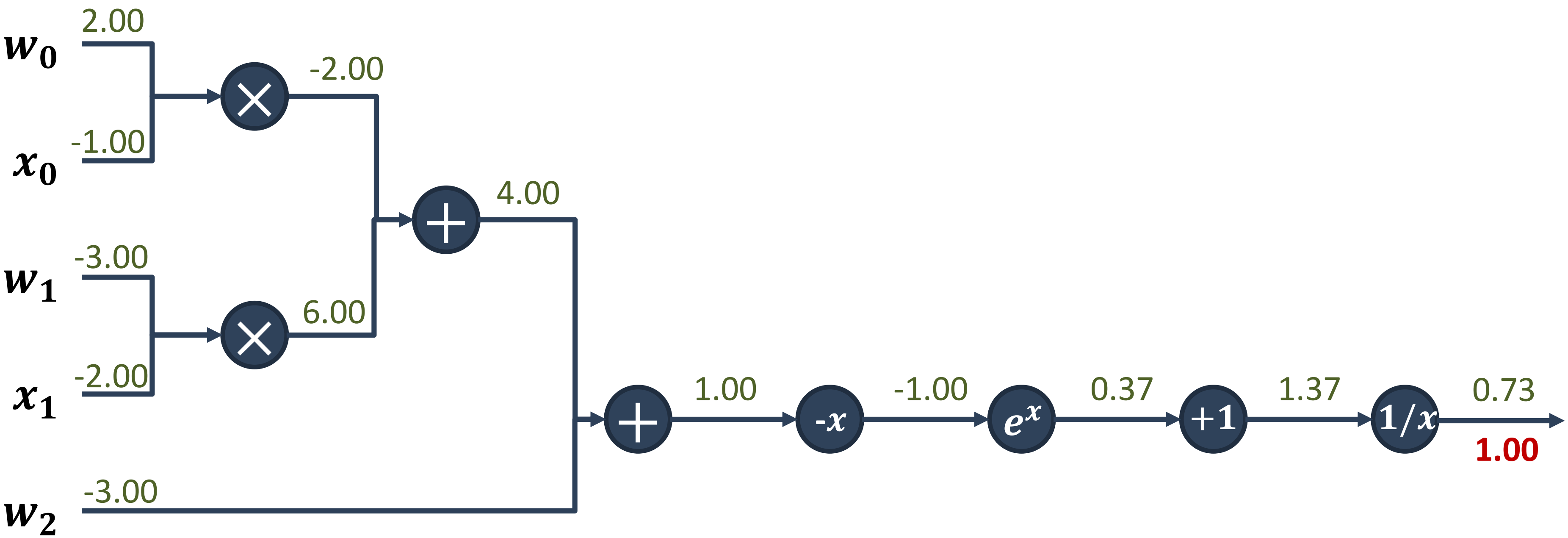
$$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{\partial f}{\partial x} = -\frac{1}{x^2}$$

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

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

Different Example

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

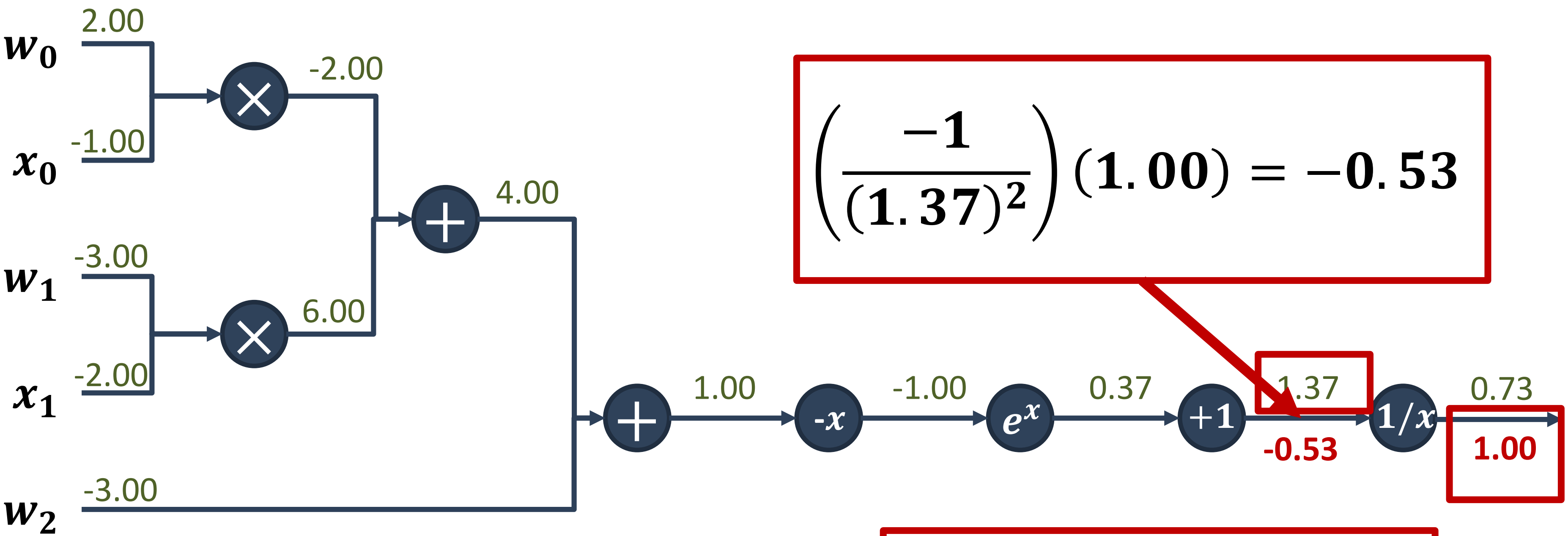


$$f(x) = e^x \quad \rightarrow \quad \frac{\partial f}{\partial x} = e^x \qquad\qquad f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{\partial f}{\partial x} = -\frac{1}{x^2}$$

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

Different Example

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



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

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

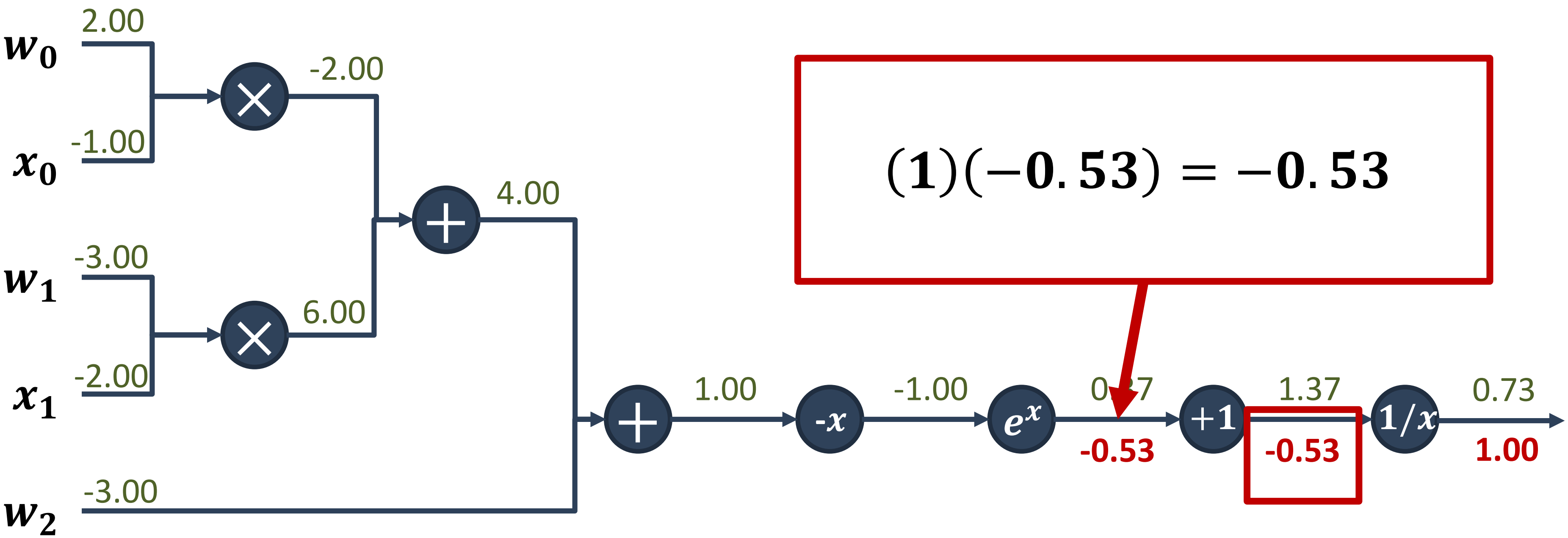
$$\boxed{f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{\partial f}{\partial x} = -\frac{1}{x^2}}$$

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

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

Different Example

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



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

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

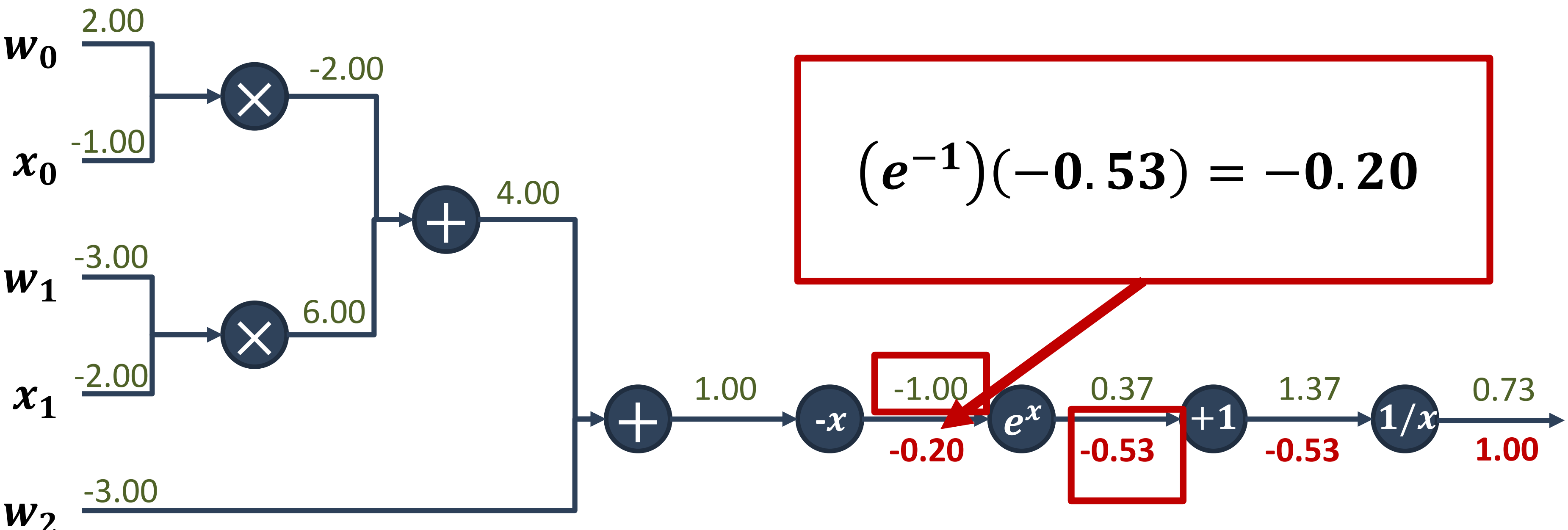
$$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{\partial f}{\partial x} = -\frac{1}{x^2}$$

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

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

Different Example

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



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

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

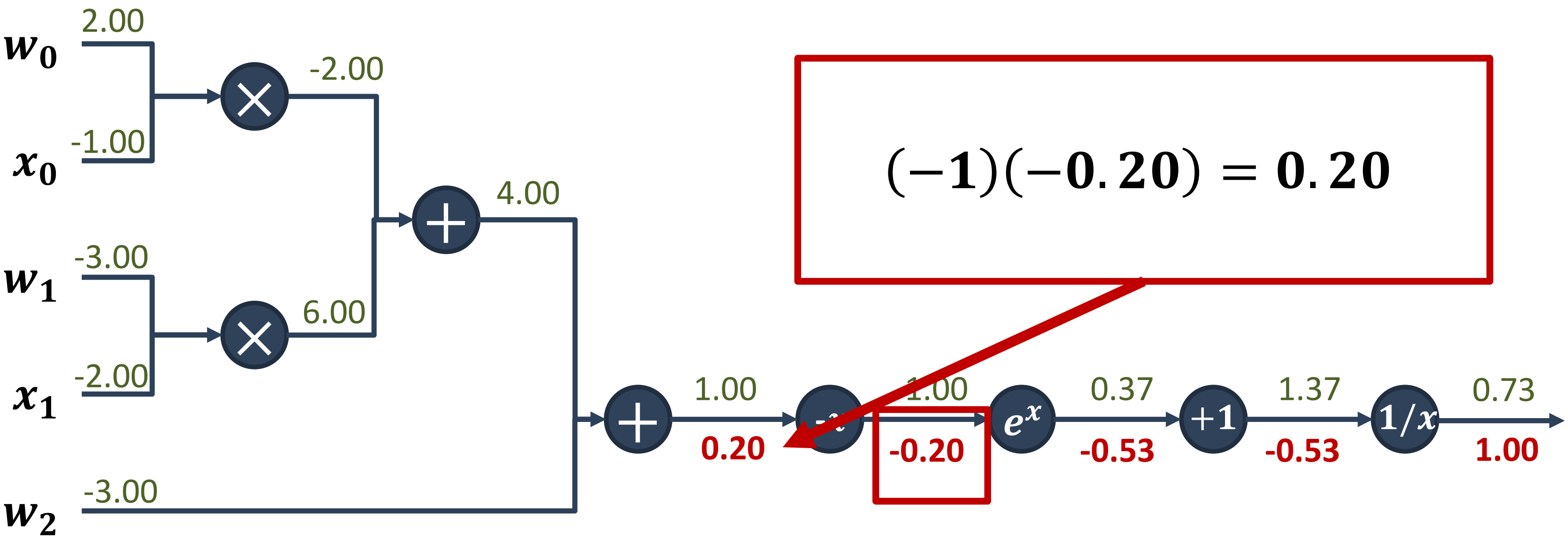
$$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{\partial f}{\partial x} = -\frac{1}{x^2}$$

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

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

Different Example

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



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

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

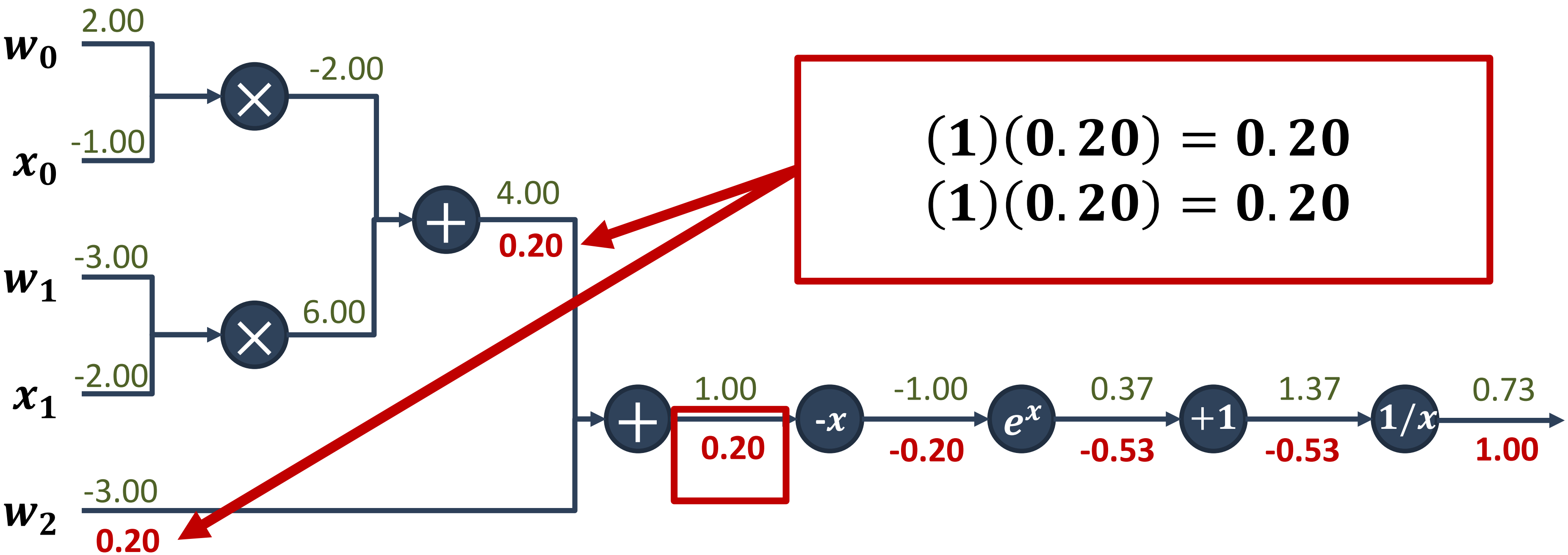
$$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{\partial f}{\partial x} = -\frac{1}{x^2}$$

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

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

Different Example

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



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

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

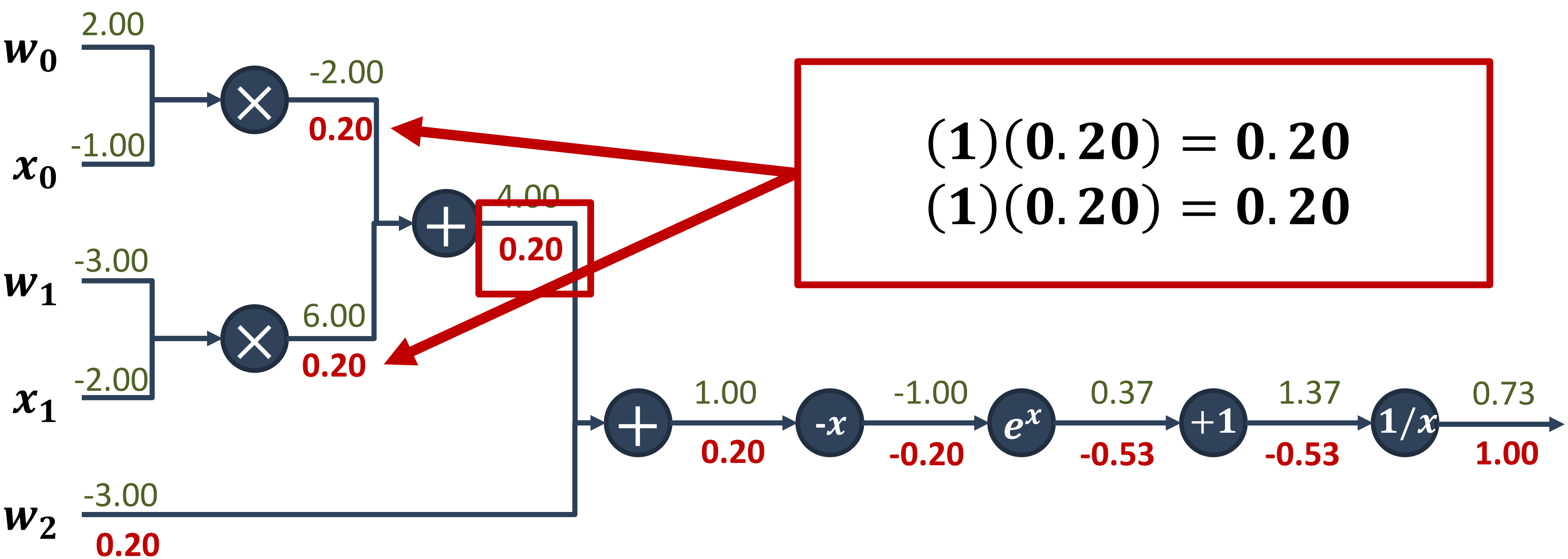
$$f_a(x) = ax \quad \rightarrow \quad \frac{\partial f}{\partial x} = a$$

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

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

Different Example

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



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

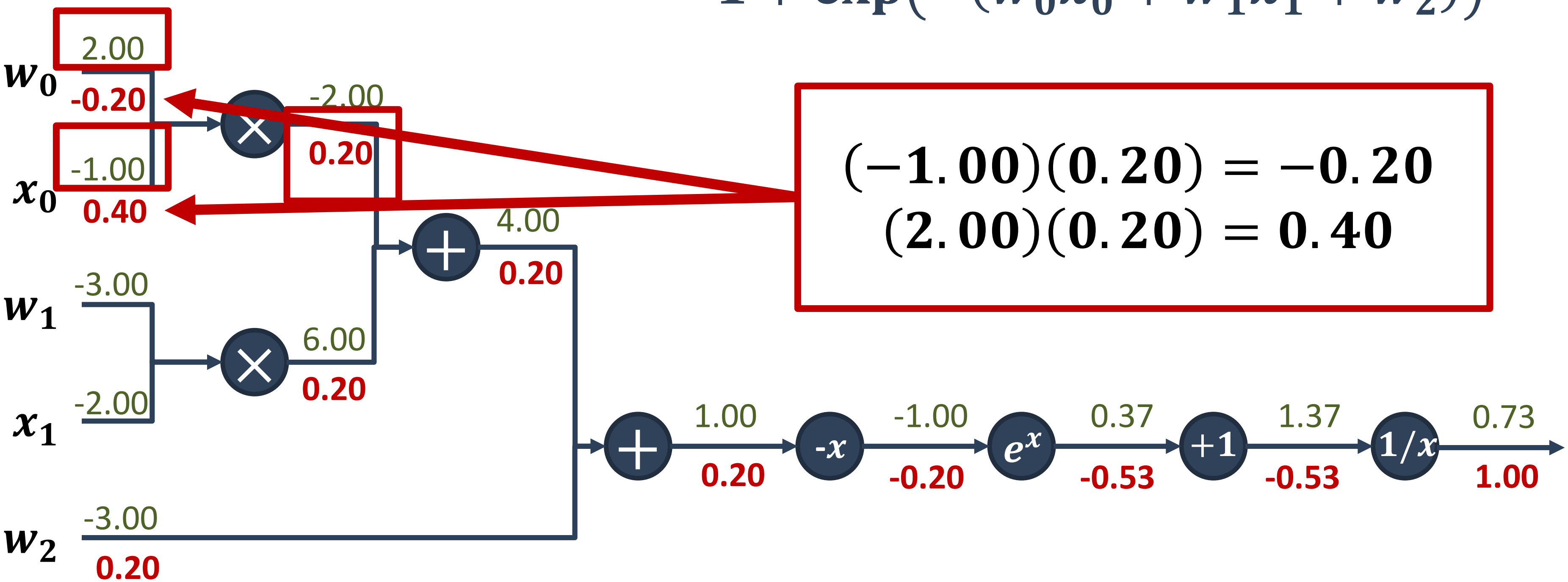
$$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{\partial f}{\partial x} = -\frac{1}{x^2}$$

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

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

Different Example

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



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

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

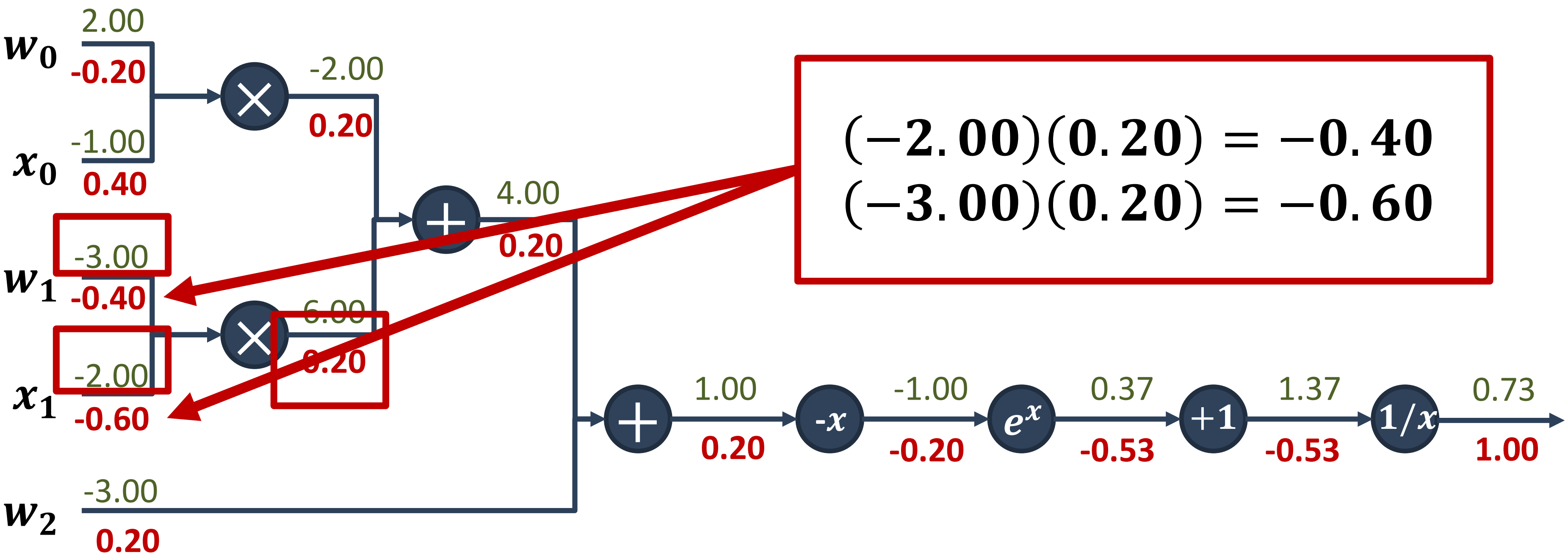
$$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{\partial f}{\partial x} = -\frac{1}{x^2}$$

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

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

Different Example

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



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

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

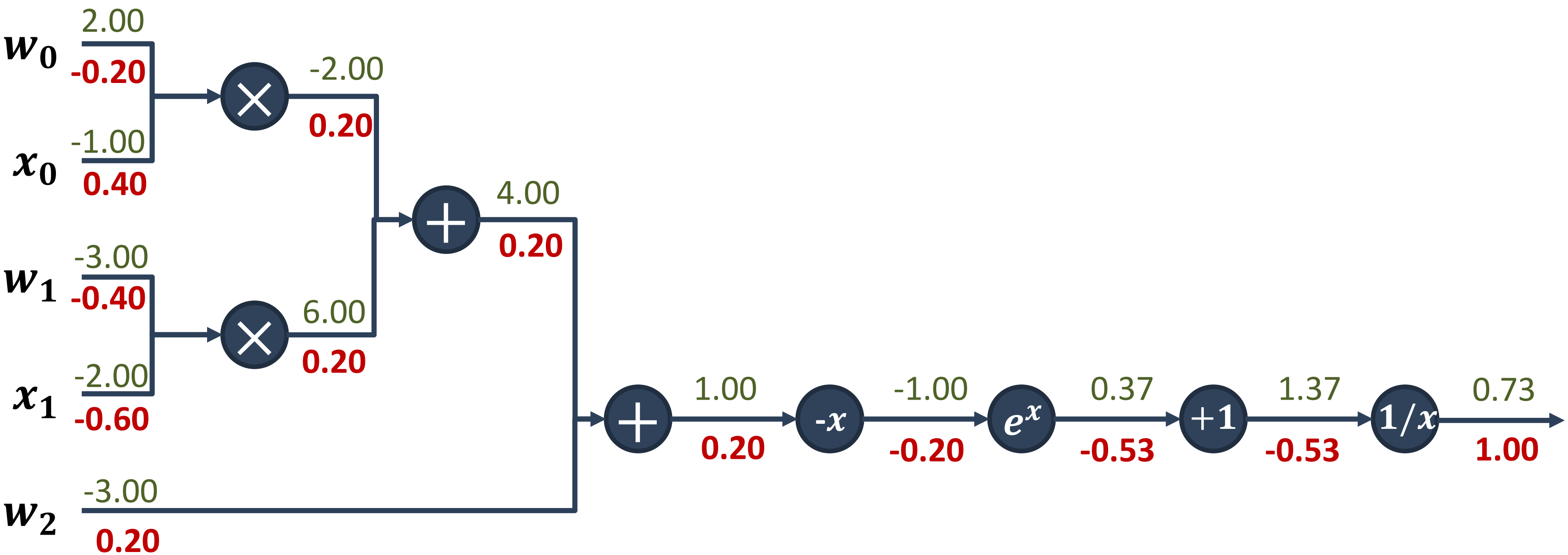
$$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{\partial f}{\partial x} = -\frac{1}{x^2}$$

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

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

Different Example

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



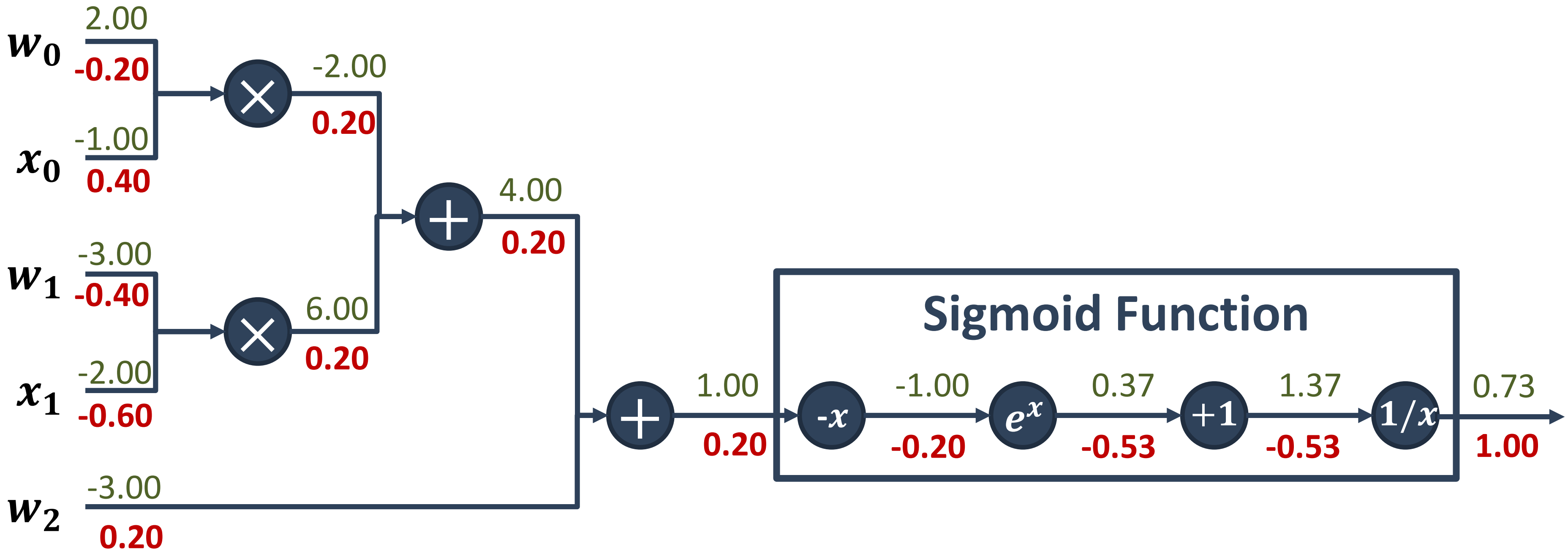
$$f(x) = e^x \quad \rightarrow \quad \frac{\partial f}{\partial x} = e^x \qquad\qquad f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{\partial f}{\partial x} = -\frac{1}{x^2}$$

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

Different Example: Sigmoid Function

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

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

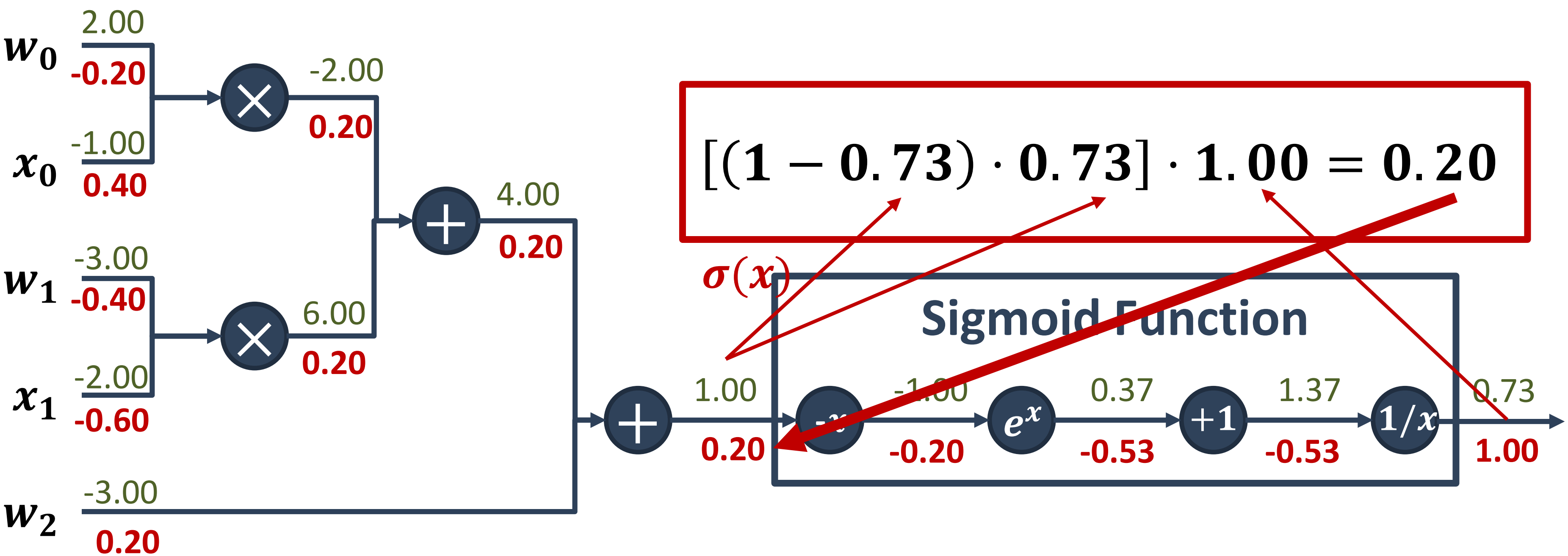


Different Example: Sigmoid Function

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

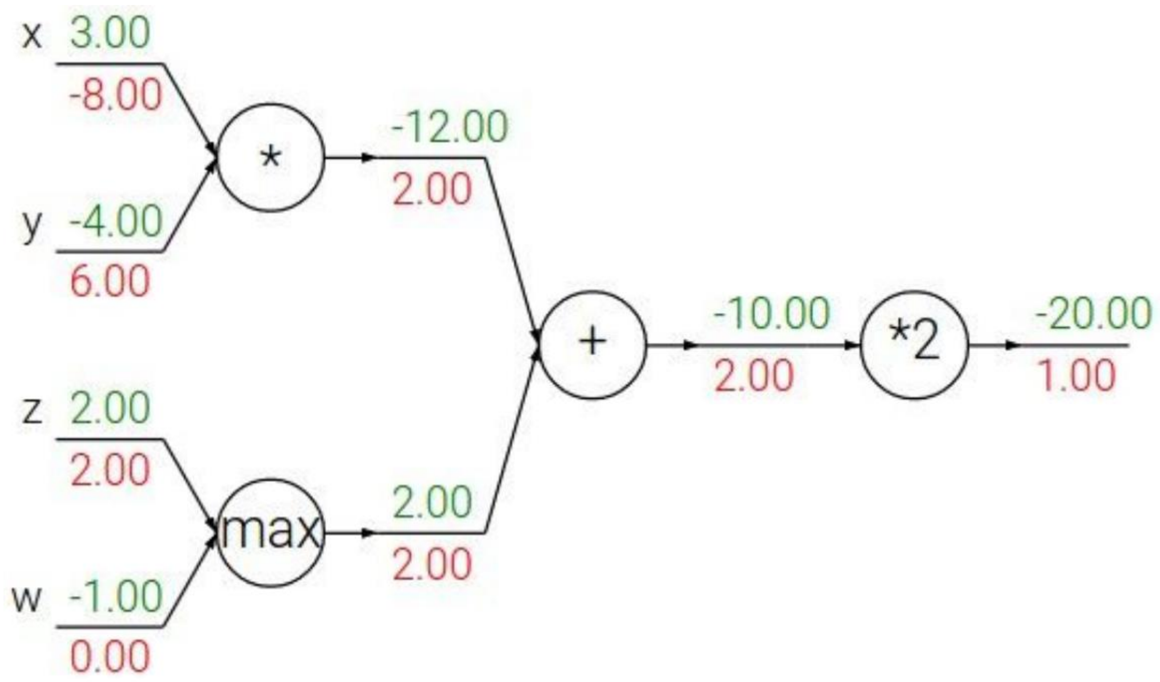
$$\sigma(x) = \frac{1}{1+\exp(-x)} \qquad \frac{\partial \sigma}{\partial x} = \left(1-\sigma(x)\right)\sigma(x)$$

$$[(1-0.73)\cdot 0.73]\cdot 1.00 = 0.20$$



Patterns in Backflow of the Gradient

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



Generalization to Vectors

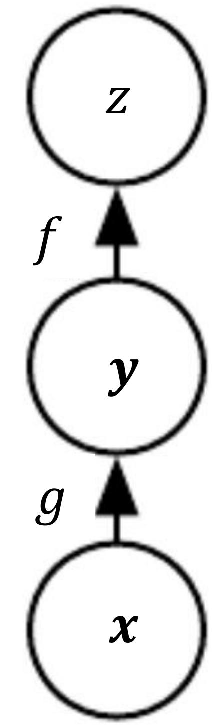
- Suppose $\boldsymbol{x} \in \mathbb{R}^m, \boldsymbol{y} \in \mathbb{R}^n$
  - g maps from $\mathbb{R}^m$ to $\mathbb{R}^n$ and
  - f maps from $\mathbb{R}^n$ to $\mathbb{R}$
- If $\boldsymbol{y} = g(\boldsymbol{x})$ and $z = f(\boldsymbol{y})$, then

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

- Or, in vector notation:

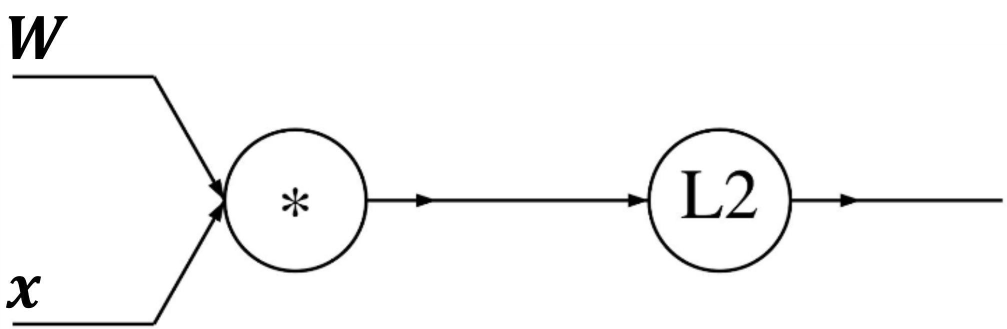
$$\nabla_x z = \left(\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right)^T \nabla_{\boldsymbol{y}} z$$

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



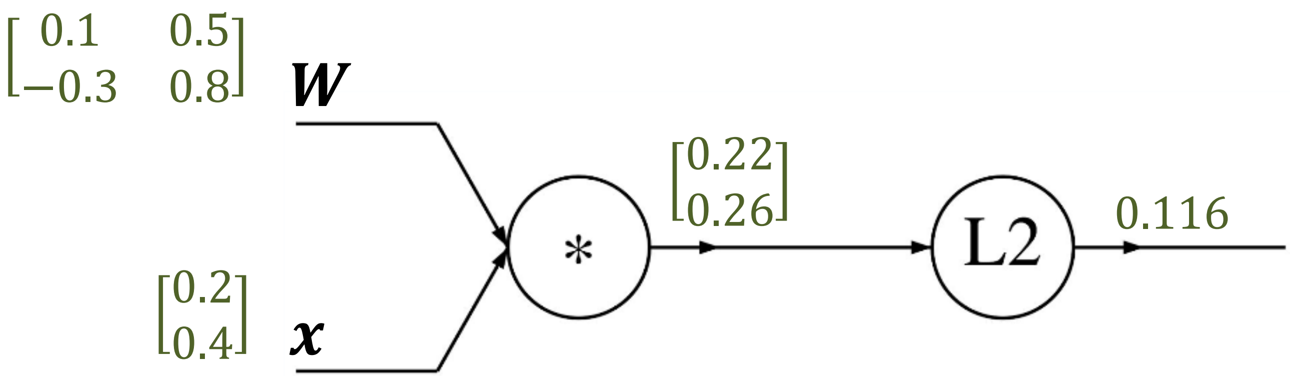
Vectorized Example

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



Vectorized Example

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

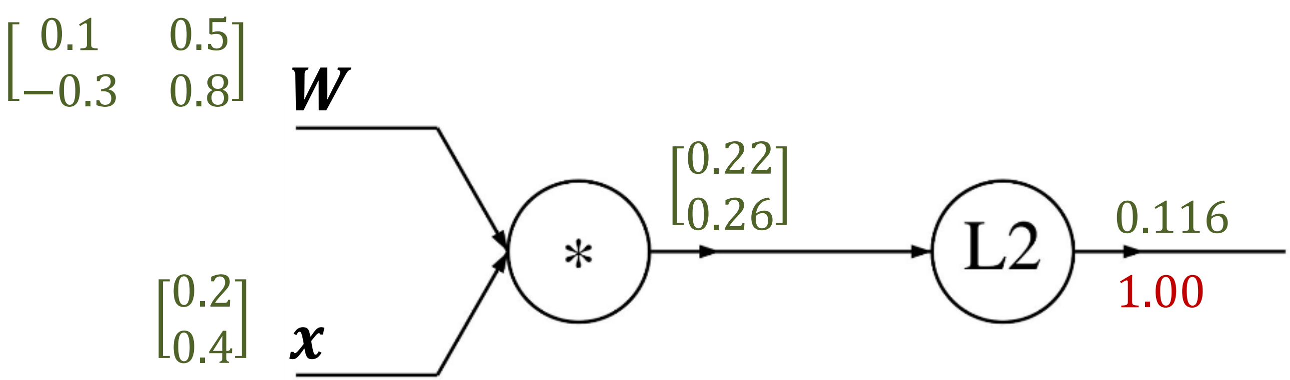


$$q = Wx = \begin{pmatrix} W_{1,1}x_1 & \cdots & W_{1,n}x_n \\ \vdots & \ddots & \vdots \\ W_{n,1}x_1 & \cdots & W_{n,n}x_n \end{pmatrix}$$

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

Vectorized Example

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

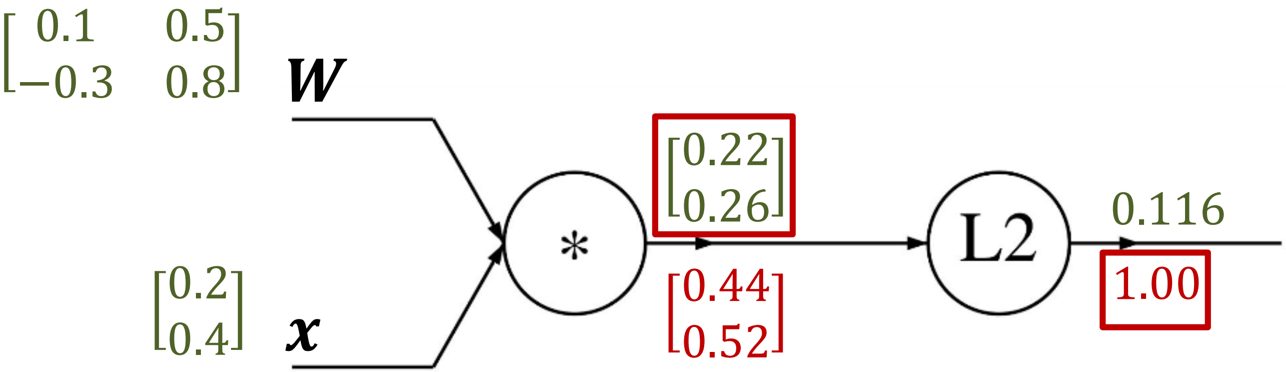


$$q = Wx = \begin{pmatrix} W_{1,1}x_1 & \cdots & W_{1,n}x_n \\ \vdots & \ddots & \vdots \\ W_{n,1}x_1 & \cdots & W_{n,n}x_n \end{pmatrix}$$

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

Vectorized Example

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



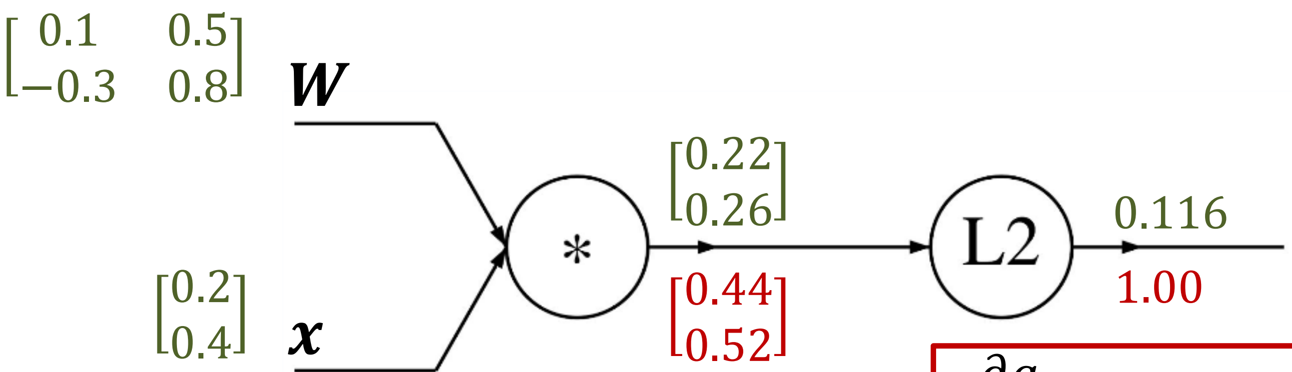
$$q = Wx = \begin{pmatrix} W_{1,1}x_1 & \cdots & W_{1,n}x_n \\ \vdots & \ddots & \vdots \\ W_{n,1}x_1 & \cdots & W_{n,n}x_n \end{pmatrix}$$

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

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

Vectorized Example

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



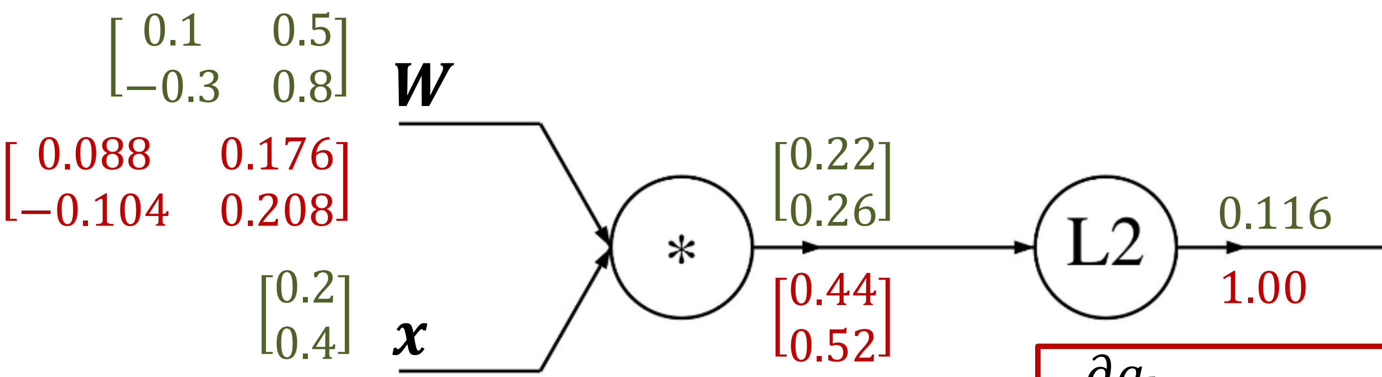
$$\boxed{\begin{aligned} \frac{\partial q_k}{\partial W_{i,j}} &= \mathbf{1}_{k=i} x_j \\ \frac{\partial f}{\partial W_{i,j}} &= \sum_k \frac{\partial f}{\partial q_k} \frac{\partial q_k}{\partial W_{i,j}} \\ &= \sum_k (2q_k)(\mathbf{1}_{k=i} x_j) \\ &= 2q_i x_j \end{aligned}}$$

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

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

Vectorized Example

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



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

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

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

$$= 2q_i x_j$$

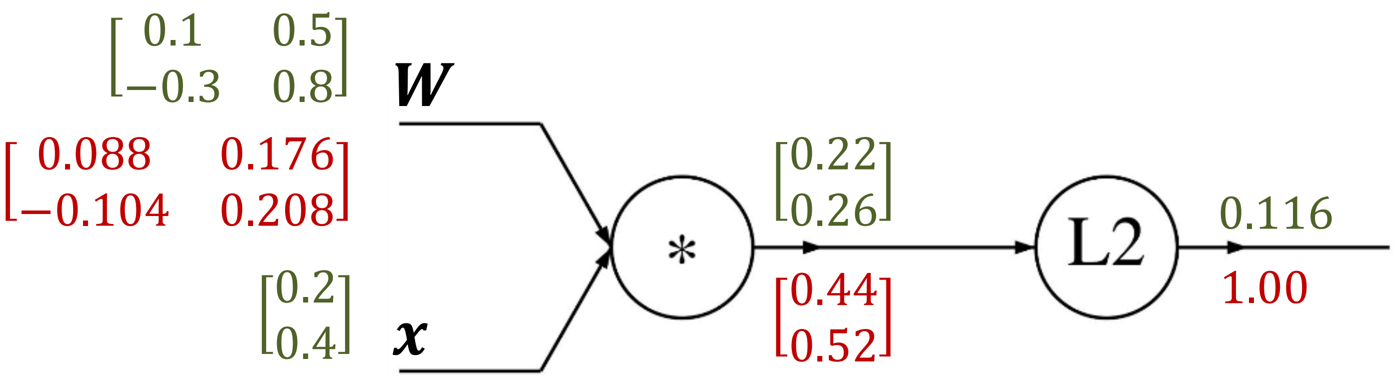
$$\boxed{\nabla_W f = 2\boldsymbol{q} \cdot \boldsymbol{x}^T}$$

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

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

Vectorized Example

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



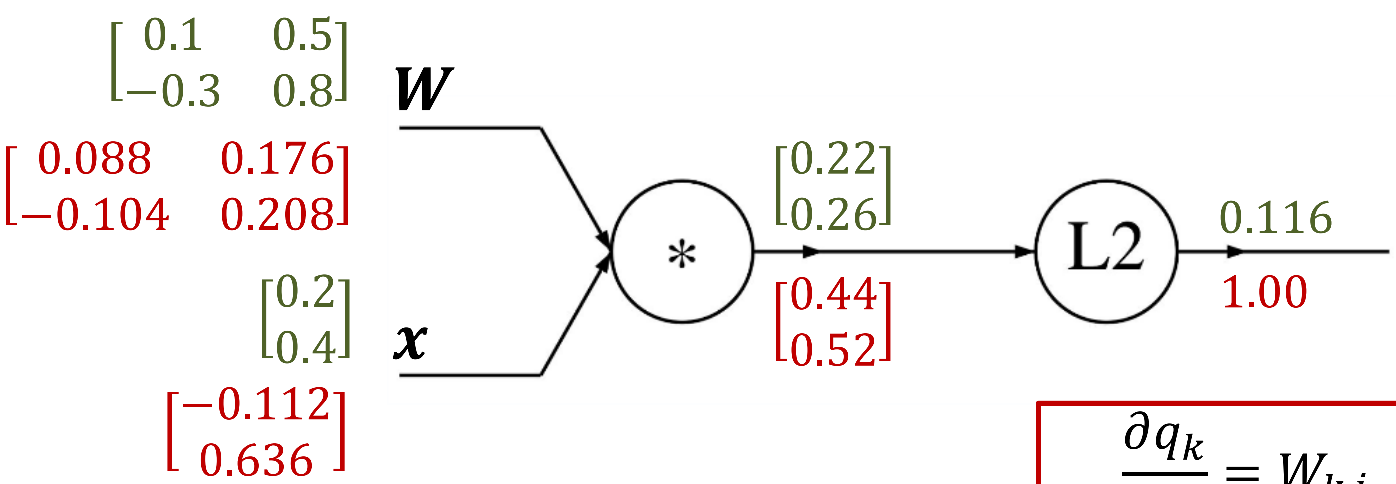
$$q = Wx = \begin{pmatrix} W_{1,1}x_1 & \cdots & W_{1,n}x_n \\ \vdots & \ddots & \vdots \\ W_{n,1}x_1 & \cdots & W_{n,n}x_n \end{pmatrix}$$

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

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

Vectorized Example

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



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

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

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

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

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

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

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