성균관대학교
SUNG KYUN KWAN UNIVERSITY

Deep Learning

- Backpropagation and Automatic Differentiation -

Eunbyung Park

Assistant Professor

School of Electronic and Electrical Engineering

Eunbyung Park (silverbottlep.github.io)

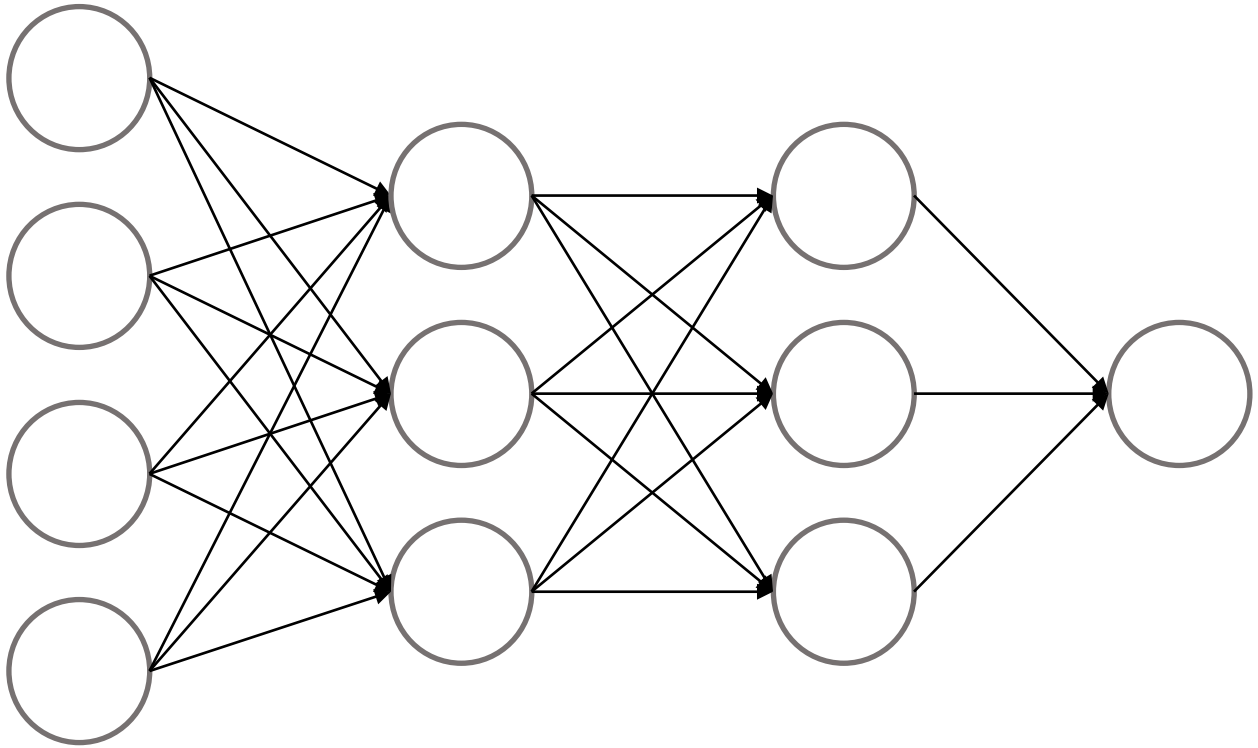
The Chain Rule

Gradient Descent

- We are using *gradient descent* for training deep neural networks
- An algorithm *to compute the derivatives of a loss function* for deep neural networks

$$W := W - \alpha \left(\frac{\partial L}{\partial W} \right)$$

(descent) (step-size) (gradient)



The Chain Rule

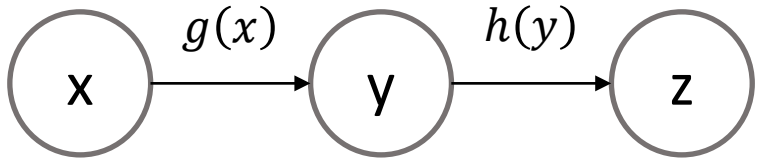
- A single variable chain rule

$$f, g, h: \mathbb{R} \to \mathbb{R}$$

$$f: h \circ g$$

$$f'(x) = h'\big(g(x)\big)g'(x)$$

$$\frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx} \qquad y = g(x), z = h(y)$$



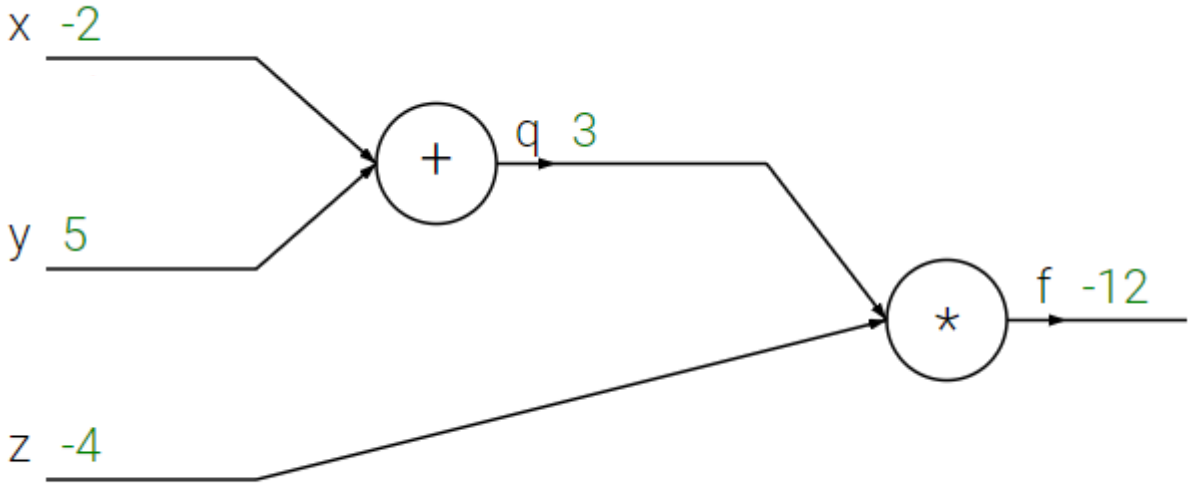
Simple Example

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

$$q = x + y, \qquad f = qz$$

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

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



Simple Example

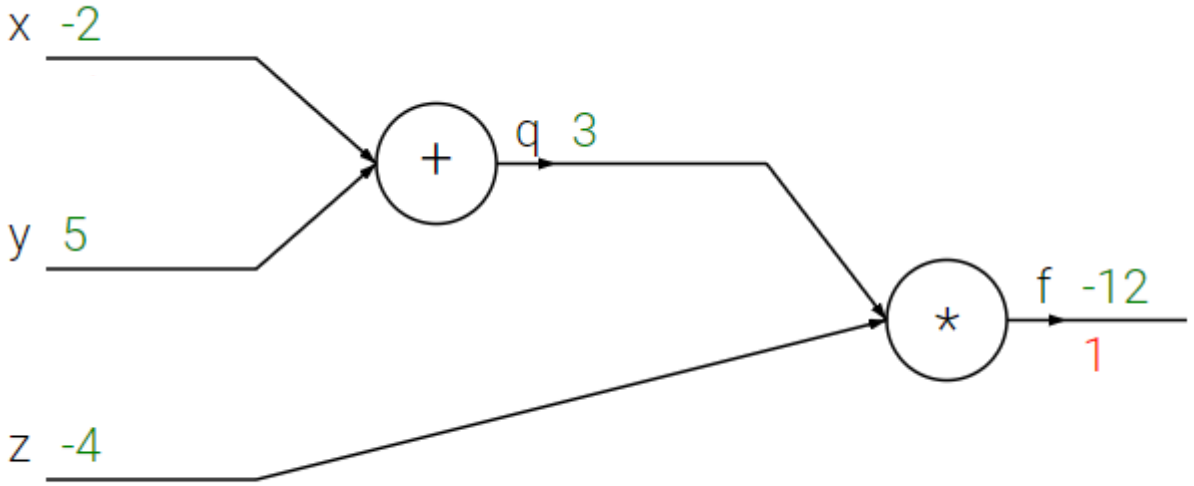
$$f(x, y, z) = (x + y)z$$

$$q = x + y, \qquad f = qz$$

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

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

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



Simple Example

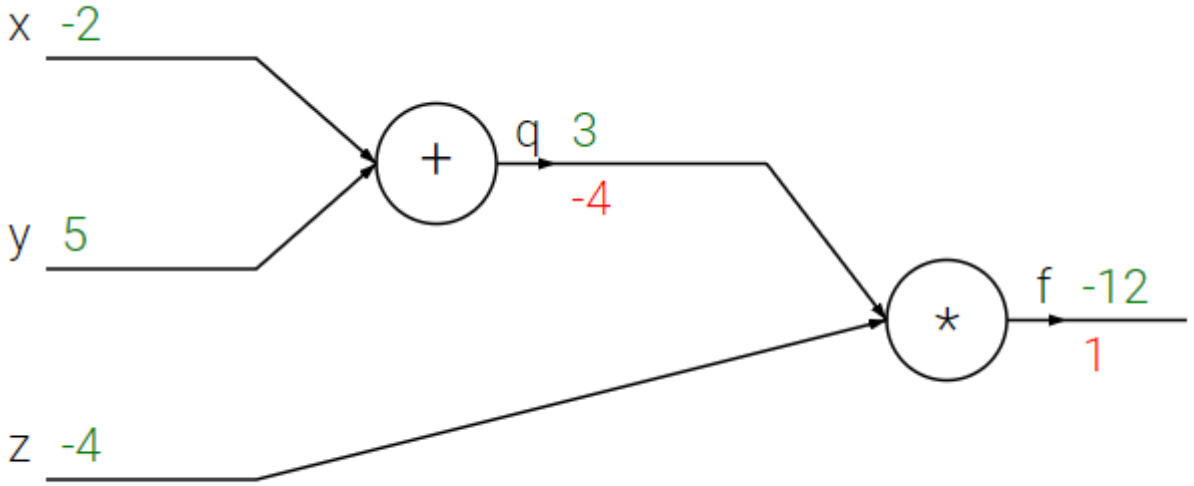
$$f(x, y, z) = (x + y)z$$

$$q = x + y, \qquad f = qz$$

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

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

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



Simple Example

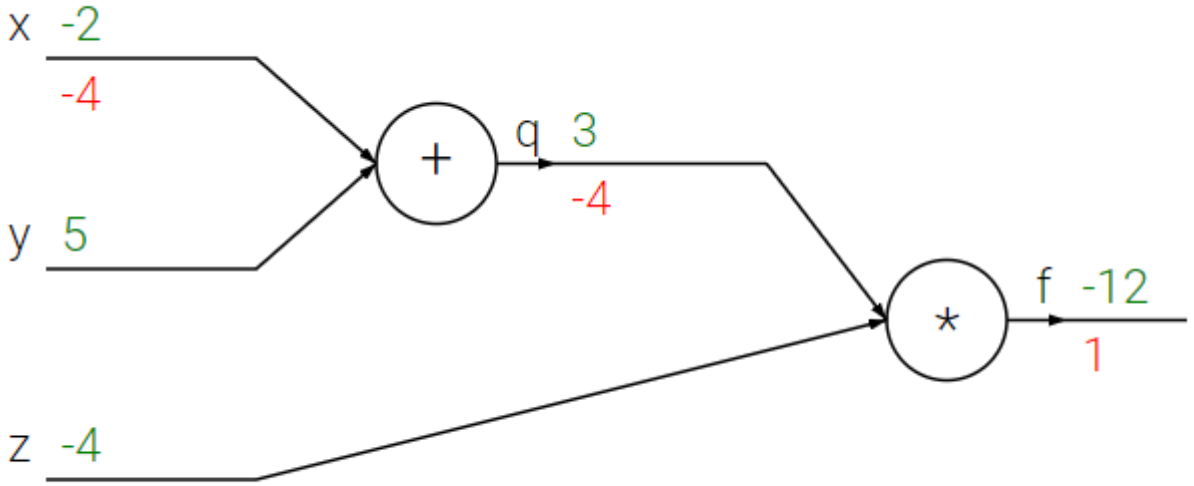
$$f(x, y, z) = (x + y)z$$

$$q = x + y, \qquad f = qz$$

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

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

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



Simple Example

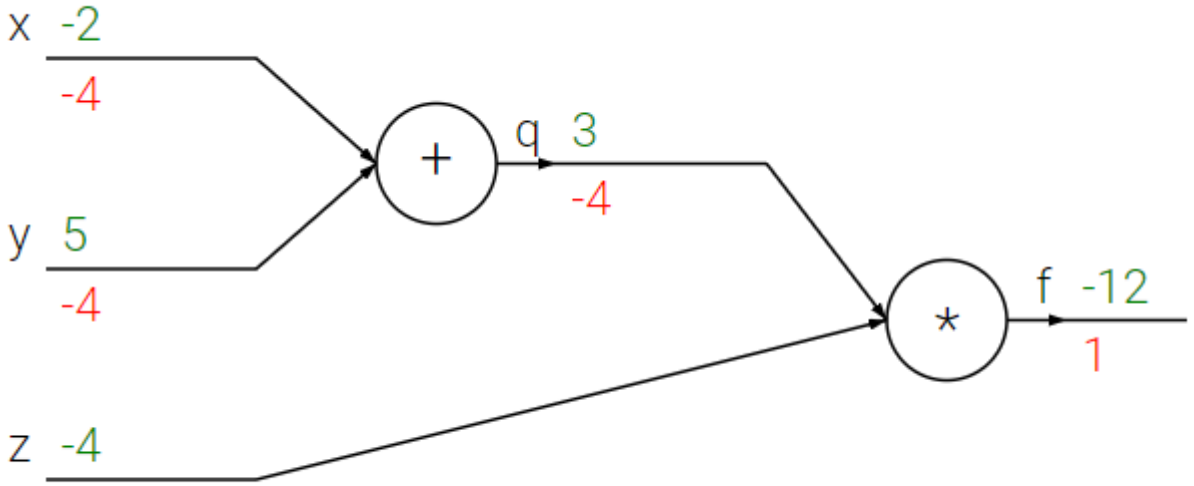
$$f(x, y, z) = (x + y)z$$

$$q = x + y, \qquad f = qz$$

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

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

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



Simple Example

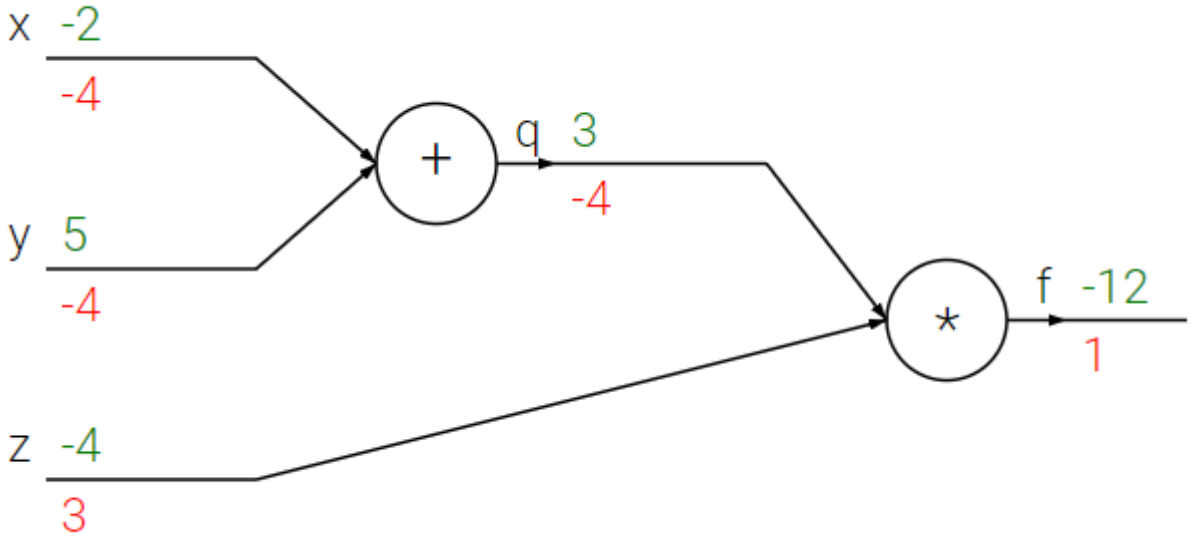
$$f(x, y, z) = (x + y)z$$

$$q = x + y, \qquad f = qz$$

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

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

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



Sigmoid Example

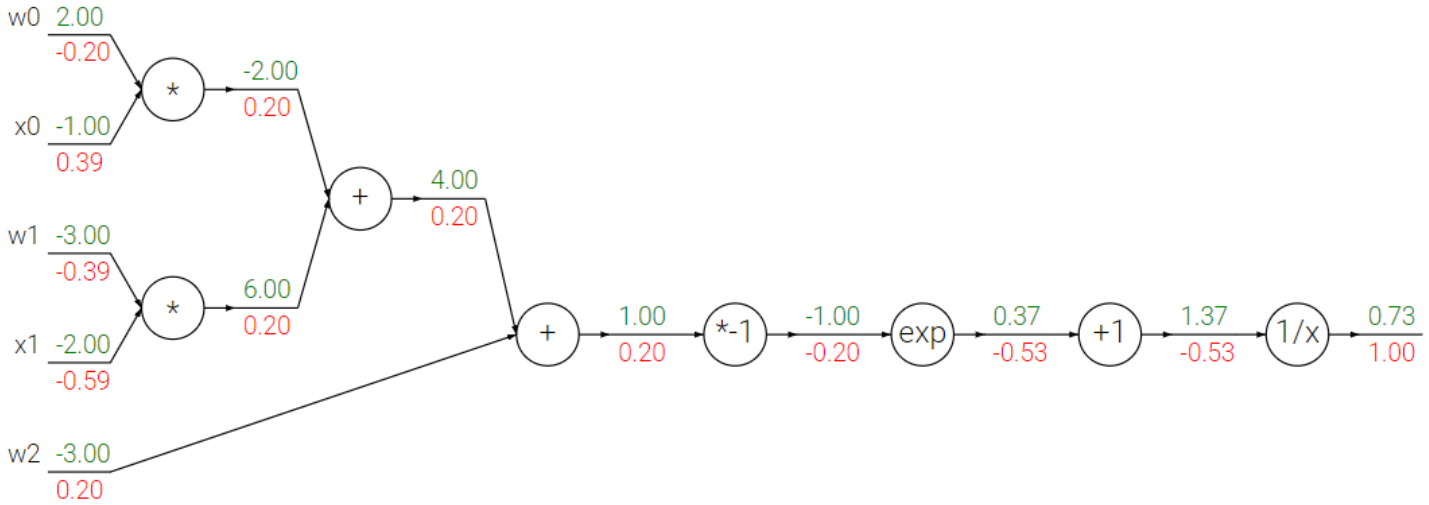
$$\sigma(x, w) = \frac{1}{1 + e^{-(w_0 x_0 + w_1 x_1 + w_2)}}$$

$$f(x) = \frac{1}{x}, \qquad g(x) = 1 + x, \qquad h(x) = e^{-x}, \qquad i(x) = w_0 x_0 + w_1 x_1 + w_2$$

Sigmoid Example

$$\sigma(x, w) = \frac{1}{1 + e^{-(w_0 x_0 + w_1 x_1 + w_2)}}$$

$$f(x) = \frac{1}{x}, \qquad g(x) = 1 + x, \qquad h(x) = e^{-x}, \qquad i(x) = w_0 x_0 + w_1 x_1 + w_2$$



Backpropagation Algorithm

Gradient

- In vector calculus, the *gradient* of a *scalar-valued* differentiable function $f: \mathbb{R}^n \to \mathbb{R}$ at the point x

$$\nabla f: \mathbb{R}^n \to \mathbb{R}^n \qquad \nabla f = \frac{\partial f}{\partial x} = \left[\frac{\partial f}{\partial x_1}, \cdots, \frac{\partial f}{\partial x_n}\right]$$

Jacobian

- In vector calculus, the *Jacobian* of a *vector-valued* differentiable function is the matrix of all its first-order partial derivatives.

$$f : \mathbb{R}^n \to \mathbb{R}^m$$

$$\mathbf{J}_{ij} = \frac{\partial f_i}{\partial x_j} \qquad \mathbf{J} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

Matrix Calculus

$X \in \mathbb{R}^{n \times m}, y \in \mathbb{R}$

$f: \mathbb{R}^{n \times m} \to \mathbb{R}$

$y = f(x)$

$$\frac{\partial y}{\partial X} = \begin{bmatrix} \frac{\partial y}{\partial X_{11}} & \cdots & \frac{\partial y}{\partial X_{1m}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial X_{n1}} & \cdots & \frac{\partial y}{\partial X_{nm}} \end{bmatrix} \in \mathbb{R}^{n \times m}$$

Matrix Calculus

$X \in \mathbb{R}^{n \times m}, y \in \mathbb{R}^{l}$

$f: \mathbb{R}^{n \times m} \to \mathbb{R}^{l}$

$y = f(x)$

$$\frac{\partial y_1}{\partial X} = \begin{bmatrix} \frac{\partial y_1}{\partial X_{11}} & \cdots & \frac{\partial y_1}{\partial X_{1m}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial X_{n1}} & \cdots & \frac{\partial y_1}{\partial X_{nm}} \end{bmatrix} \in \mathbb{R}^{n \times m}$$

$$\frac{\partial y}{\partial X} \in \mathbb{R}^{l \times n \times m}$$ (3 dim tensor)

Finite Difference

- Numerical method to compute the gradients based on the definition of gradients

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

$$\frac{df}{dx} \approx \frac{f(x+\Delta x) - f(x)}{\Delta x}$$

Forward difference

$$\frac{df}{dx} \approx \frac{f(x) - f(x-\Delta x)}{\Delta x}$$

Backward difference

$$\frac{df}{dx} \approx \frac{f(x+\Delta x) - f(x-\Delta x)}{2\Delta x}$$

Central difference

Finite Difference

- Numerical method to compute the gradients based on the definition of gradients

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

What's wrong with this approach?

$$\frac{df}{dx} \approx \frac{f(x+\Delta x) - f(x)}{\Delta x}$$

Forward difference

$$\frac{df}{dx} \approx \frac{f(x) - f(x-\Delta x)}{\Delta x}$$

Backward difference

$$\frac{df}{dx} \approx \frac{f(x+\Delta x) - f(x-\Delta x)}{2\Delta x}$$

Central difference

The Chain Rule

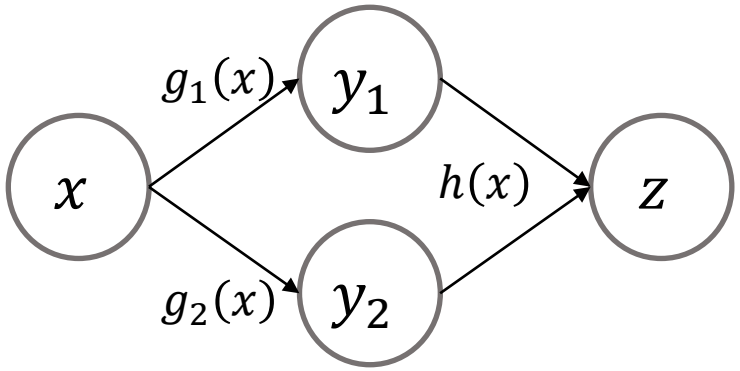
- Multi-variable chain rule

$$f, g_1, g_2 : \mathbb{R} \to \mathbb{R}, \qquad h : \mathbb{R}^2 \to \mathbb{R}$$

$$y_1 = g_1(x), \qquad y_2 = g_2(y)$$

$$z = h(y_1, y_2)$$

$$\frac{dz}{dx} = \frac{dz}{dy_1}\frac{dy_1}{dx} + \frac{dz}{dy_2}\frac{dy_2}{dx}$$



The Chain Rule

- Multi-variable chain rule

$x \in \mathbb{R}, y \in \mathbb{R}^n, z \in \mathbb{R}$

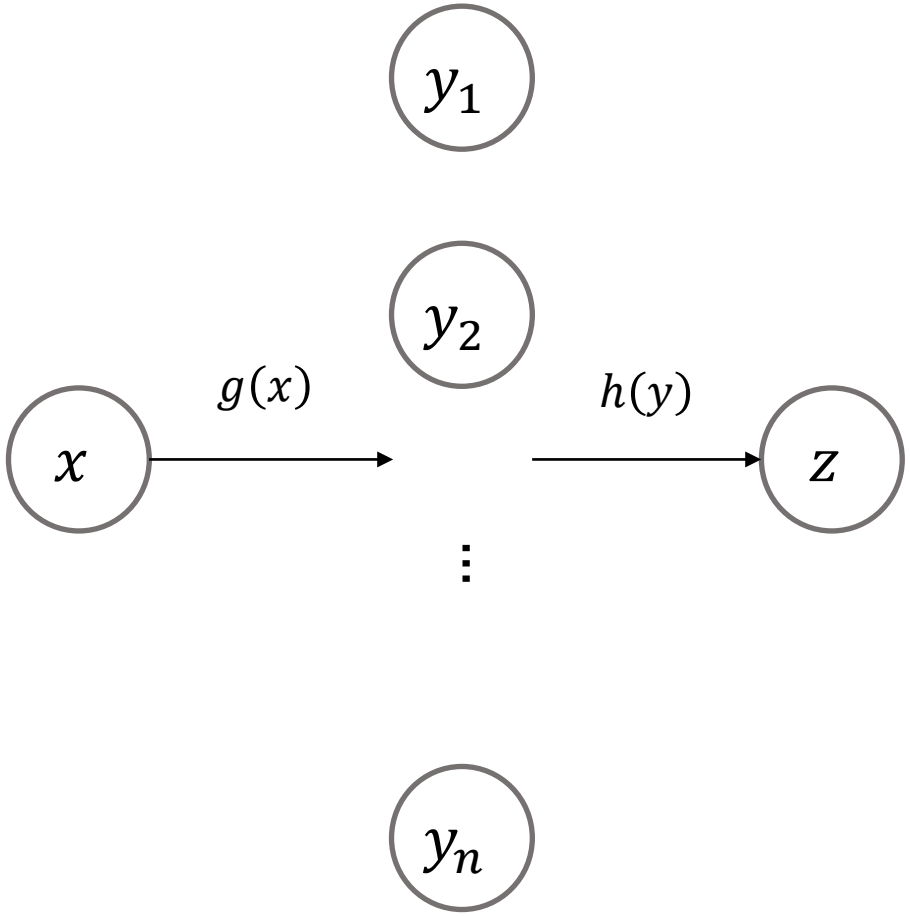
$g: \mathbb{R} \to \mathbb{R}^n, \qquad y = g(x)$

$h: \mathbb{R}^n \to \mathbb{R}, \qquad z = h(y)$

$$\frac{\partial z}{\partial x} = \sum_{i=1}^{n} \frac{\partial z}{\partial y_i} \frac{dy_i}{dx} = \frac{\partial z}{\partial y} \frac{\partial y}{\partial x}$$

$\frac{\partial y}{\partial x} \in \mathbb{R}^{n \times 1}$

$\frac{\partial z}{\partial y} \in \mathbb{R}^{1 \times n}$



The Chain Rule

- Multi-variable chain rule

$x \in \mathbb{R}^n, y \in \mathbb{R}^m, z \in \mathbb{R}$

$g: \mathbb{R}^n \to \mathbb{R}^m, \qquad y = g(x)$

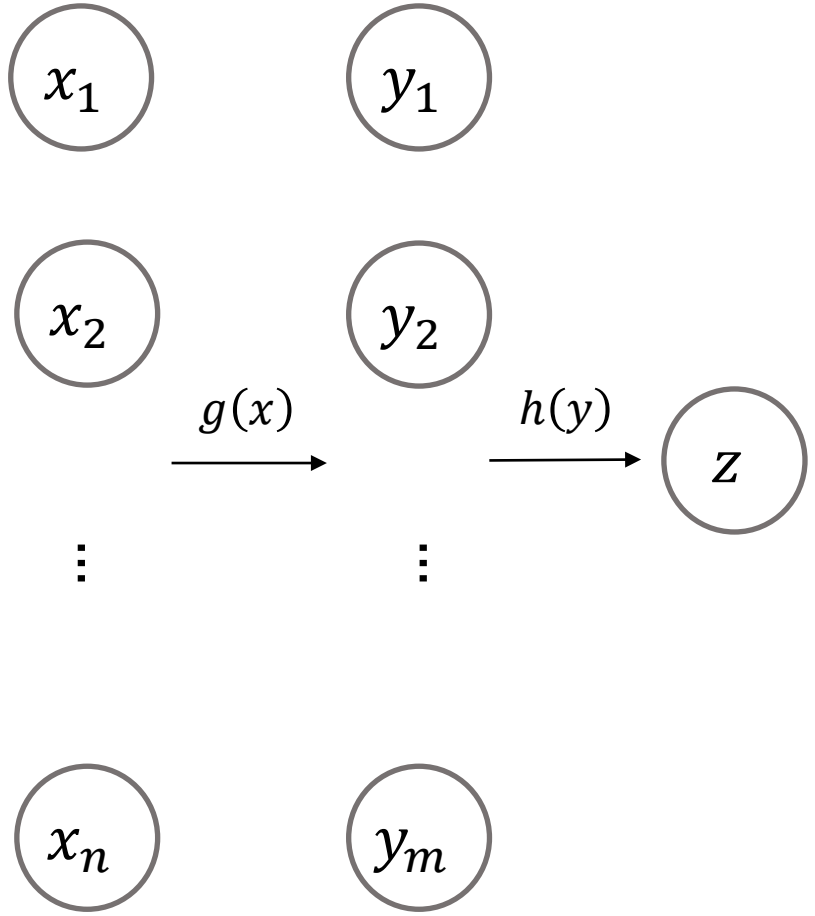
$h: \mathbb{R}^m \to \mathbb{R}, \qquad z = h(y)$

$$\frac{\partial z}{\partial x_j} = \sum_{i=1}^{m} \frac{\partial z}{\partial y_i}\frac{\partial y_i}{\partial x_j} = \frac{\partial z}{\partial y}\frac{\partial y}{\partial x_j}$$

$\frac{\partial z}{\partial y} \in \mathbb{R}^{1\times m}$, $\frac{\partial y}{\partial x_j} \in \mathbb{R}^{m\times 1}$

$$\frac{\partial z}{\partial x} = \left[\sum_{i=1}^{m} \frac{\partial z}{\partial y_i}\frac{\partial y_i}{\partial x_1}, \dots, \sum_{i=1}^{m} \frac{\partial z}{\partial y_i}\frac{\partial y_i}{\partial x_n}\right] = \frac{\partial z}{\partial y}\frac{\partial y}{\partial x}$$

$\frac{\partial z}{\partial y} \in \mathbb{R}^{1\times m}$, $\frac{\partial y}{\partial x} \in \mathbb{R}^{m\times n}$



Two Layers MLP

$x \in \mathbb{R}^n, y \in \mathbb{R}^m, z \in \mathbb{R}, W \in \mathbb{R}^{m \times n}, v \in \mathbb{R}^m$

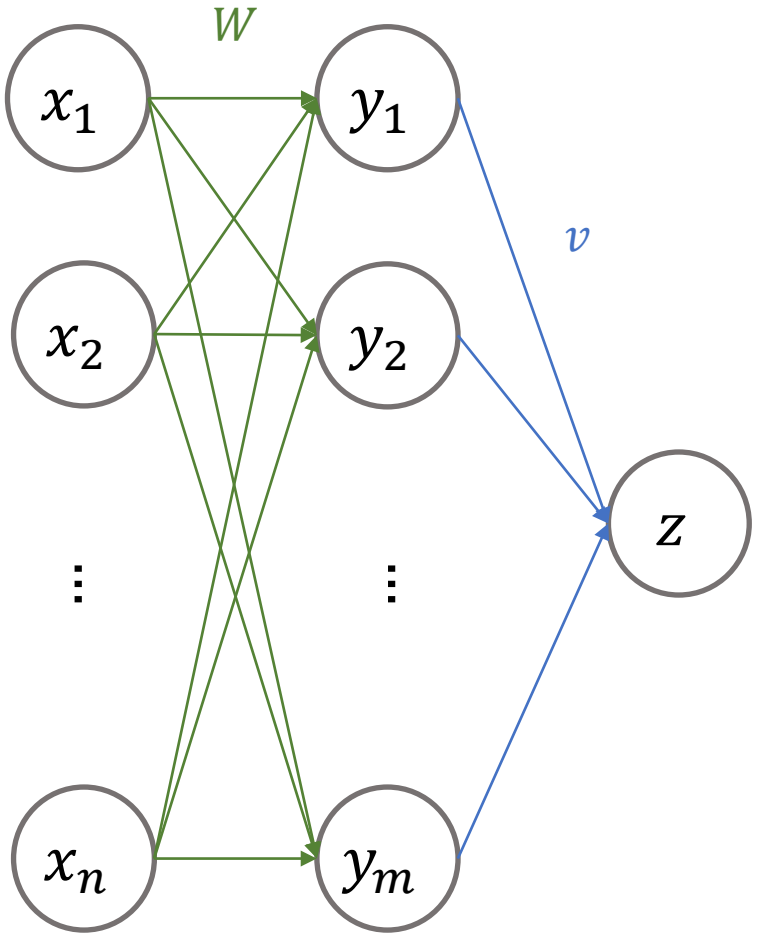
$$y = Wx \qquad z = \sum_{i=1}^{m} v_i y_i = v^\top y$$

$$\frac{\partial z}{\partial x_j} = \sum_{i=1}^{m} \frac{\partial z}{\partial y_i}\frac{\partial y_i}{\partial x_j} = \frac{\partial z}{\partial y}\frac{\partial y}{\partial x_j}$$

$\frac{\partial z}{\partial y} \in \mathbb{R}^{1 \times m}$, $\frac{\partial y}{\partial x_j} \in \mathbb{R}^{m \times 1}$

$$\frac{\partial z}{\partial x} = \left[\sum_{i=1}^{m} \frac{\partial z}{\partial y_i}\frac{\partial y_i}{\partial x_1}, \cdots, \sum_{i=1}^{m} \frac{\partial z}{\partial y_i}\frac{\partial y_i}{\partial x_n}\right] = \frac{\partial z}{\partial y}\frac{\partial y}{\partial x} = v^\top W$$

$\frac{\partial z}{\partial y} \in \mathbb{R}^{1 \times m}$, $\frac{\partial y}{\partial x} \in \mathbb{R}^{m \times n}$



Derivatives of Linear Layer

$$y = Wx \qquad z = \sum_{i=1}^{m} v_i y_i = v^\top y$$

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

Two Layers MLP

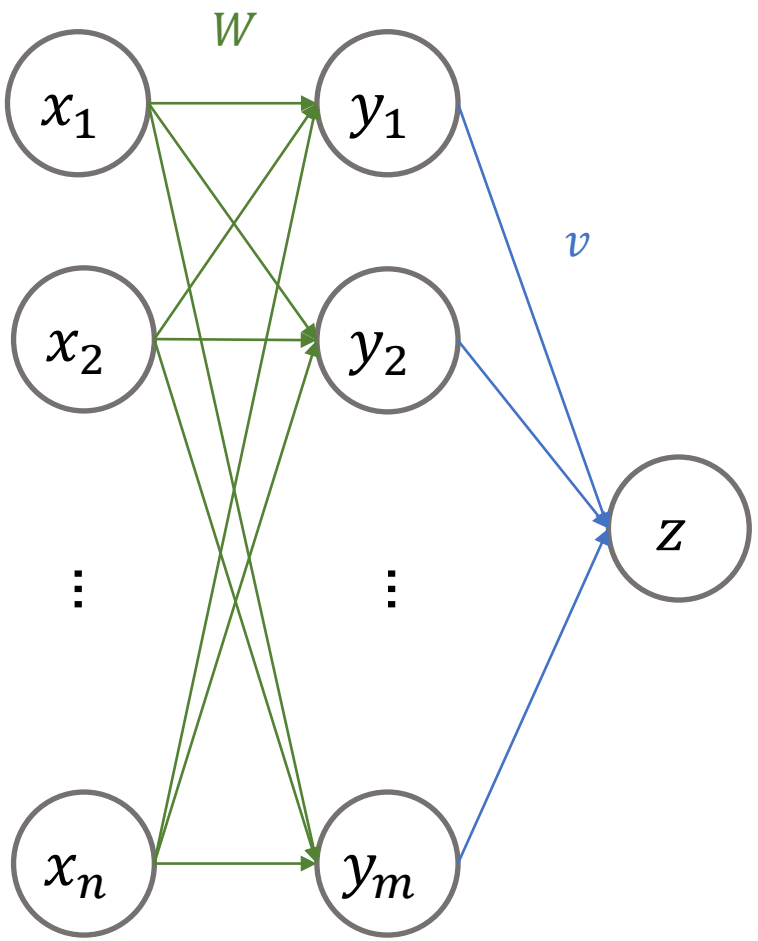
$$x \in \mathbb{R}^n, y \in \mathbb{R}^m, z \in \mathbb{R}, W \in \mathbb{R}^{m \times n}, v \in \mathbb{R}^m$$

$$y = Wx \qquad z = \sum_{i=1}^{m} v_i y_i = v^\top y$$

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

$\frac{\partial z}{\partial W} \in \mathbb{R}^{m \times n}$, $\frac{\partial y}{\partial W} \in \mathbb{R}^{m \times m \times n}$, $\frac{\partial z}{\partial y} \in \mathbb{R}^{1 \times m}$

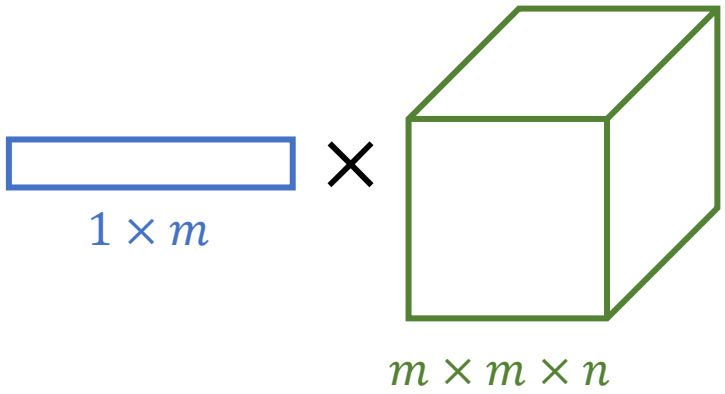
Tensor Product
(n-mode product)

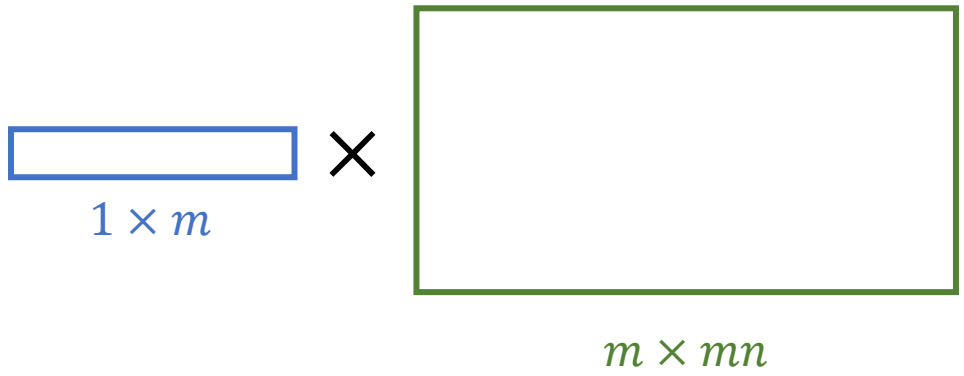


Tensor Product

- N-mode product
  - Matricization -> matrix multiplication

$$\underset{\in \mathbb{R}^{m\times n}}{\frac{\partial z}{\partial W}} = \underset{\in \mathbb{R}^{1\times m}}{\frac{\partial z}{\partial y}} \underset{\in \mathbb{R}^{m\times m\times n}}{\frac{\partial y}{\partial W}}$$





Vector Jacobian Product (VJP)

- Jacobian is very sparse and explicit formation of it is too expensive

$x \in \mathbb{R}^n, y \in \mathbb{R}^m, z \in \mathbb{R}, W \in \mathbb{R}^{m \times n}, v \in \mathbb{R}^m$

$y = Wx$

$$\frac{\partial y_1}{\partial W} = \begin{bmatrix} \frac{\partial y_1}{\partial W_{11}} & \cdots & \frac{\partial y_1}{\partial W_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial W_{m1}} & \cdots & \frac{\partial y_1}{\partial W_{mn}} \end{bmatrix} = \begin{bmatrix} \frac{\partial y_1}{\partial W_{11}} & \cdots & \frac{\partial y_1}{\partial W_{1n}} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$\in \mathbb{R}^{m \times n}$

$$\frac{\partial y_2}{\partial W} = \begin{bmatrix} \frac{\partial y_2}{\partial W_{11}} & \cdots & \frac{\partial y_2}{\partial W_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_2}{\partial W_{m1}} & \cdots & \frac{\partial y_2}{\partial W_{mn}} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ \frac{\partial y_2}{\partial W_{21}} & \cdots & \frac{\partial y_2}{\partial W_{2n}} \\ 0 & 0 & 0 \end{bmatrix}$$

$\in \mathbb{R}^{m \times n}$

Vector Jacobian Product (VJP)

- Jacobian is very sparse and explicit formation of it is too expensive

$x \in \mathbb{R}^n, y \in \mathbb{R}^m, z \in \mathbb{R}, W \in \mathbb{R}^{m \times n}, v \in \mathbb{R}^m$

$y = Wx$

$$\text{reshape}\left(\frac{\partial y}{\partial W}\right) = \begin{bmatrix} \frac{\partial y_1}{\partial W_{11}} & \cdots & \frac{\partial y_1}{\partial W_{1n}} & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & \frac{\partial y_2}{\partial W_{21}} & \cdots & \frac{\partial y_2}{\partial W_{2n}} & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{\partial y_3}{\partial W_{31}} & \cdots & \frac{\partial y_3}{\partial W_{3n}} & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \end{bmatrix}$$

$\in \mathbb{R}^{m \times mn}$

Vector Jacobian Product (VJP)

- Jacobian is very sparse and explicit formation of it is too expensive

$$\frac{\partial z}{\partial y}\,\text{reshape}\left(\frac{\partial y}{\partial W}\right) =$$

$$\begin{bmatrix} \frac{\partial z}{\partial y_1} & \frac{\partial z}{\partial y_2} & \cdots & \frac{\partial z}{\partial y_m} \end{bmatrix} \begin{bmatrix} \frac{\partial y_1}{\partial W_{11}} & \cdots & \frac{\partial y_1}{\partial W_{1n}} & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & \frac{\partial y_2}{\partial W_{21}} & \cdots & \frac{\partial y_2}{\partial W_{2n}} & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{\partial y_3}{\partial W_{31}} & \cdots & \frac{\partial y_3}{\partial W_{3n}} & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial z}{\partial y_1}\frac{\partial y_1}{\partial W_{11}} & \cdots & \frac{\partial z}{\partial y_1}\frac{\partial y_1}{\partial W_{1n}} & \frac{\partial z}{\partial y_2}\frac{\partial y_2}{\partial W_{21}} & \cdots & \frac{\partial z}{\partial y_2}\frac{\partial y_2}{\partial W_{2n}} & \cdots \end{bmatrix}$$

Vector Jacobian Product (VJP)

- Jacobian is very sparse and explicit formation of it is too expensive

$$\frac{\partial z}{\partial W} = \frac{\partial z}{\partial y}\frac{\partial y}{\partial W} = \text{reshape}\left(\frac{\partial z}{\partial y}\,\text{reshape}\left(\frac{\partial y}{\partial W}\right)\right)$$

$$= \text{reshape}\left(\left[\frac{\partial z}{\partial y_1}\frac{\partial y_1}{\partial W_{11}} \quad \cdots \quad \frac{\partial z}{\partial y_1}\frac{\partial y_1}{\partial W_{1n}} \quad \frac{\partial z}{\partial y_2}\frac{\partial y_2}{\partial W_{21}} \quad \cdots \quad \frac{\partial z}{\partial y_2}\frac{\partial y_2}{\partial W_{2n}} \quad \cdots\right]\right)$$

$$= \begin{bmatrix} \frac{\partial z}{\partial y_1}\frac{\partial y_1}{\partial W_{11}} & \cdots & \frac{\partial z}{\partial y_1}\frac{\partial y_1}{\partial W_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial z}{\partial y_m}\frac{\partial y_m}{\partial W_{m1}} & \cdots & \frac{\partial z}{\partial y_m}\frac{\partial y_m}{\partial W_{mn}} \end{bmatrix} = \begin{bmatrix} \frac{\partial z}{\partial y_1}x_1 & \cdots & \frac{\partial z}{\partial y_1}x_n \\ \vdots & \ddots & \vdots \\ \frac{\partial z}{\partial y_m}x_1 & \cdots & \frac{\partial z}{\partial y_m}x_n \end{bmatrix} = \left(\frac{\partial z}{\partial y}\right)^{\top} x^{\top}$$

Vector Jacobian Product (VJP)

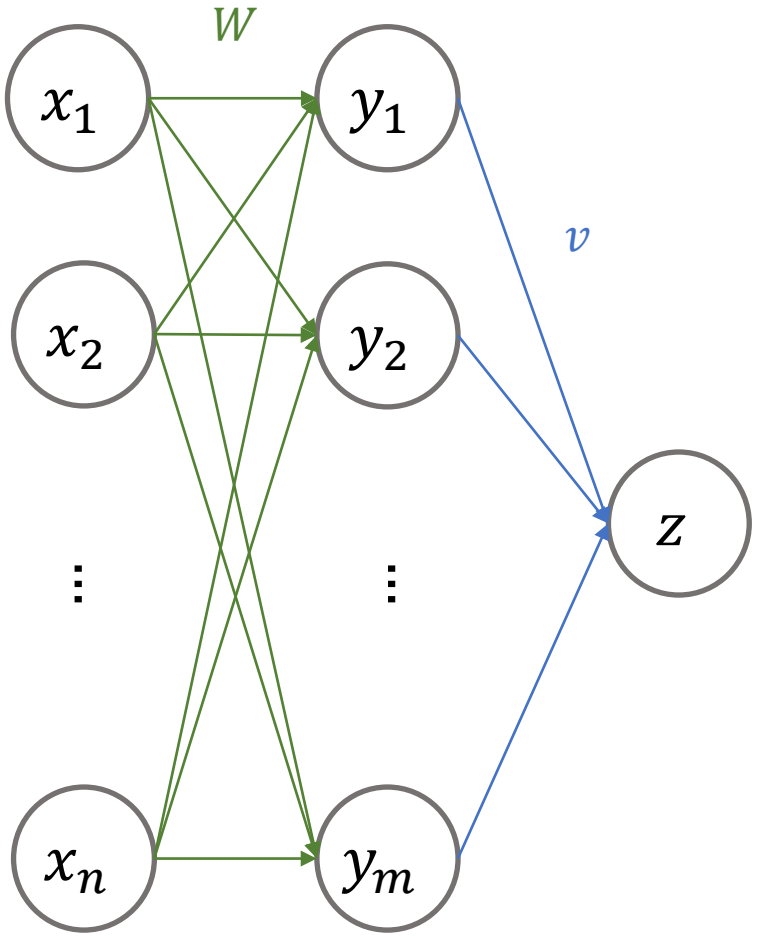
- Explicit formation of Jacobian is too expensive

$$x \in \mathbb{R}^n, y \in \mathbb{R}^m, z \in \mathbb{R}, W \in \mathbb{R}^{m \times n}, v \in \mathbb{R}^m$$

$$y = Wx \qquad z = \sum_{i=1}^{m} v_i y_i = v^\top y$$

$$\underset{\in \mathbb{R}^{m \times n}}{\frac{\partial z}{\partial W}} = \underset{\in \mathbb{R}^{1 \times m}}{\frac{\partial z}{\partial y}} \underset{\in \mathbb{R}^{m \times m \times n}}{\frac{\partial y}{\partial W}}$$

$$\underset{\in \mathbb{R}^{m \times n}}{\frac{\partial z}{\partial W}} = \underset{\in \mathbb{R}^{m \times 1}}{\left(\frac{\partial z}{\partial y}\right)^\top} \underset{\in \mathbb{R}^{1 \times n}}{x^\top}$$



We almost never explicitly construct Jacobians ($\frac{\partial y}{\partial W}$). We instead directly compute vector-Jacobian product (VJP, $\frac{\partial z}{\partial y}\frac{\partial y}{\partial W}$) in more efficient way ($\left(\frac{\partial z}{\partial y}\right)^\top x^\top$)

Vector Jacobian Product (VJP)

- Elementwise activation functions

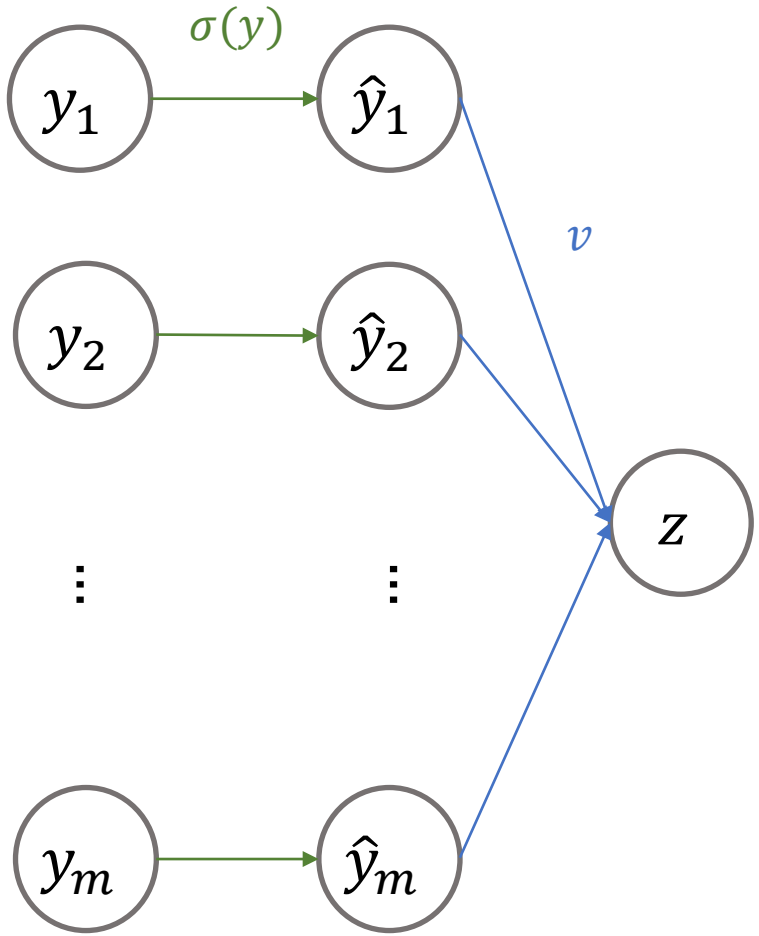
$$y \in \mathbb{R}^m, \hat{y} \in \mathbb{R}^m$$

$$\hat{y} = \sigma(y) \qquad z = \sum_{i=1}^{m} v_i \hat{y}_i = v^\top \hat{y}$$

$$\underset{\in \mathbb{R}^{1\times m}}{\frac{\partial z}{\partial y}} = \underset{\in \mathbb{R}^{1\times m}}{\frac{\partial z}{\partial \hat{y}}} \underset{\in \mathbb{R}^{m\times m}}{\frac{\partial \hat{y}}{\partial y}}$$

$$\frac{\partial \hat{y}}{\partial y} = \begin{bmatrix} \frac{\partial \hat{y}_1}{\partial y_1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \frac{\partial \hat{y}_m}{\partial y_m} \end{bmatrix}$$

$$= \begin{bmatrix} \sigma(y_1)(1-\sigma(y_1)) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma(y_m)(1-\sigma(y_m)) \end{bmatrix}$$



Vector Jacobian Product (VJP)

- Elementwise activation functions

$$\underset{\in \mathbb{R}^{1\times m}}{\frac{\partial z}{\partial y}} = \underset{\in \mathbb{R}^{1\times m}}{\frac{\partial z}{\partial \hat{y}}} \underset{\in \mathbb{R}^{m\times m}}{\frac{\partial \hat{y}}{\partial y}} \qquad \frac{\partial \hat{y}}{\partial y} = \begin{bmatrix} \frac{\partial \hat{y}_1}{\partial y_1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \frac{\partial \hat{y}_m}{\partial y_m} \end{bmatrix} = \begin{bmatrix} \sigma(y_1)(1-\sigma(y_1)) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma(y_m)(1-\sigma(y_m)) \end{bmatrix}$$

Element-wise product

$$\underset{\in \mathbb{R}^{1\times m}}{\frac{\partial z}{\partial y}} = \underset{\in \mathbb{R}^{1\times m}}{\frac{\partial z}{\partial \hat{y}}} \odot \left(\sigma(y)\left(1-\sigma(y)\right)\right)^{\top}$$

Automatic Differentiation

Automatic Differentiation (AD)

- A procedure for automatic evaluation of derivatives of arbitrary algebraic functions
- Backpropagation == reverse-mode AD

Reverse-Mode AD (a.k.a Backpropagation)

$f: \mathbb{R}^{n_1} \to \mathbb{R}^{n_2}$

$g: \mathbb{R}^{n_2} \to \mathbb{R}^{n_3}$

$h: \mathbb{R}^{n_3} \to \mathbb{R}^{n_4}$

$i: \mathbb{R}^{n_4} \to \mathbb{R}$

Loss function:
scalar function

$$b = f(a)$$
$$c = g(b)$$
$$d = h(c)$$
$$e = i(d)$$

$$\frac{\partial e}{\partial a}?$$

Reverse-Mode AD (a.k.a Backpropagation)

$$\frac{\partial e}{\partial d} = \underset{\in \mathbb{R}^{1\times 1}}{\frac{\partial e}{\partial e}} \underset{\in \mathbb{R}^{1\times n_4}}{\frac{\partial e}{\partial d}} = 1\frac{\partial e}{\partial d}$$

$f: \mathbb{R}^{n_1} \to \mathbb{R}^{n_2}$

$g: \mathbb{R}^{n_2} \to \mathbb{R}^{n_3}$

$h: \mathbb{R}^{n_3} \to \mathbb{R}^{n_4}$

$i: \mathbb{R}^{n_4} \to \mathbb{R}$

$b = f(a)$

$c = g(b)$

$d = h(c)$

$e = i(d)$

Loss function:
scalar function

Reverse-Mode AD (a.k.a Backpropagation)

$f: \mathbb{R}^{n_1} \to \mathbb{R}^{n_2}$

$g: \mathbb{R}^{n_2} \to \mathbb{R}^{n_3}$

$h: \mathbb{R}^{n_3} \to \mathbb{R}^{n_4}$

$i: \mathbb{R}^{n_4} \to \mathbb{R}$

Loss function:
scalar function

$$b = f(a)$$
$$c = g(b)$$
$$d = h(c)$$
$$e = i(d)$$

$$\frac{\partial e}{\partial d} = \underbrace{\frac{\partial e}{\partial e}}_{\in \mathbb{R}^{1\times 1}} \underbrace{\frac{\partial e}{\partial d}}_{\in \mathbb{R}^{1\times n_4}} = 1 \frac{\partial e}{\partial d}$$

$$\frac{\partial e}{\partial c} = \underbrace{\frac{\partial e}{\partial e}}_{\in \mathbb{R}^{1\times n_4}} \underbrace{\frac{\partial e}{\partial d}}_{\in \mathbb{R}^{n_4\times n_3}} \frac{\partial d}{\partial c} = \frac{\partial e}{\partial d} \frac{\partial d}{\partial c}$$

Reverse-Mode AD (a.k.a Backpropagation)

$f: \mathbb{R}^{n_1} \to \mathbb{R}^{n_2}$

$g: \mathbb{R}^{n_2} \to \mathbb{R}^{n_3}$

$h: \mathbb{R}^{n_3} \to \mathbb{R}^{n_4}$

$i: \mathbb{R}^{n_4} \to \mathbb{R}$

Loss function:
scalar function

$$b = f(a)$$
$$c = g(b)$$
$$d = h(c)$$
$$e = i(d)$$

$$\frac{\partial e}{\partial d} = \underbrace{\frac{\partial e}{\partial e}}_{\in \mathbb{R}^{1\times 1}} \underbrace{\frac{\partial e}{\partial d}}_{\in \mathbb{R}^{1\times n_4}} = 1\frac{\partial e}{\partial d}$$

$$\frac{\partial e}{\partial c} = \underbrace{\frac{\partial e}{\partial e}\frac{\partial e}{\partial d}}_{\in \mathbb{R}^{1\times n_4}} \underbrace{\frac{\partial d}{\partial c}}_{\in \mathbb{R}^{n_4\times n_3}} = \frac{\partial e}{\partial d}\frac{\partial d}{\partial c}$$

$$\frac{\partial e}{\partial b} = \underbrace{\frac{\partial e}{\partial e}\frac{\partial e}{\partial d}\frac{\partial d}{\partial c}}_{\in \mathbb{R}^{1\times n_3}} \underbrace{\frac{\partial c}{\partial b}}_{\in \mathbb{R}^{n_3\times n_2}} = \frac{\partial e}{\partial c}\frac{\partial c}{\partial b}$$

Reverse-Mode AD (a.k.a Backpropagation)

$f: \mathbb{R}^{n_1} \to \mathbb{R}^{n_2}$

$g: \mathbb{R}^{n_2} \to \mathbb{R}^{n_3}$

$h: \mathbb{R}^{n_3} \to \mathbb{R}^{n_4}$

$i: \mathbb{R}^{n_4} \to \mathbb{R}$

Loss function:
scalar function

$$b = f(a)$$
$$c = g(b)$$
$$d = h(c)$$
$$e = i(d)$$

$$\frac{\partial e}{\partial d} = \underbrace{\frac{\partial e}{\partial e}}_{\in \mathbb{R}^{1\times 1}} \underbrace{\frac{\partial e}{\partial d}}_{\in \mathbb{R}^{1\times n_4}} = 1 \frac{\partial e}{\partial d}$$

Vector-Jacobian Product (VJP)

$$\frac{\partial e}{\partial c} = \frac{\partial e}{\partial e}\frac{\partial e}{\partial d}\underbrace{\frac{\partial d}{\partial c}}_{\in \mathbb{R}^{n_4\times n_3}} = \underbrace{\frac{\partial e}{\partial d}}_{\in \mathbb{R}^{1\times n_4}}\frac{\partial d}{\partial c}$$

$$\frac{\partial e}{\partial b} = \frac{\partial e}{\partial e}\frac{\partial e}{\partial d}\frac{\partial d}{\partial c}\underbrace{\frac{\partial c}{\partial b}}_{\in \mathbb{R}^{n_3\times n_2}} = \underbrace{\frac{\partial e}{\partial c}}_{\in \mathbb{R}^{1\times n_3}}\frac{\partial c}{\partial b}$$

$$\frac{\partial e}{\partial a} = \frac{\partial e}{\partial e}\frac{\partial e}{\partial d}\frac{\partial d}{\partial c}\frac{\partial c}{\partial b}\underbrace{\frac{\partial b}{\partial a}}_{\in \mathbb{R}^{n_2\times n_1}} = \underbrace{\frac{\partial e}{\partial b}}_{\in \mathbb{R}^{1\times n_2}}\frac{\partial b}{\partial a}$$

Reverse-Mode AD (a.k.a Backpropagation)

$f: \mathbb{R}^{n_1} \to \mathbb{R}^{n_2}$

$g: \mathbb{R}^{n_2} \to \mathbb{R}^{n_3}$

$h: \mathbb{R}^{n_3} \to \mathbb{R}^{n_4}$

$i: \mathbb{R}^{n_4} \to \mathbb{R}^{3}$

$b = f(a)$
$c = g(b)$
$d = h(c)$
$e = i(d)$

What if i is vector-valued function?

$$\frac{\partial e}{\partial d} = \underset{\in \mathbb{R}^{3\times 3}}{\frac{\partial e}{\partial e}} \underset{\in \mathbb{R}^{3\times n_4}}{\frac{\partial e}{\partial d}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \frac{\partial e}{\partial d}$$

3 X Computation

$$\frac{\partial e}{\partial c} = \underset{\in \mathbb{R}^{3\times n_4}}{\frac{\partial e}{\partial e}} \underset{\in \mathbb{R}^{n_4\times n_3}}{\frac{\partial d}{\partial d}} \frac{\partial d}{\partial c} = \frac{\partial e}{\partial d} \frac{\partial d}{\partial c}$$

$$\frac{\partial e}{\partial b} = \underset{\in \mathbb{R}^{3\times n_3}}{\frac{\partial e}{\partial e}} \underset{\in \mathbb{R}^{n_3\times n_2}}{\frac{\partial d}{\partial d}} \frac{\partial c}{\partial c} \frac{\partial c}{\partial b} = \frac{\partial e}{\partial c} \frac{\partial c}{\partial b}$$

$$\frac{\partial e}{\partial a} = \underset{\in \mathbb{R}^{3\times n_2}}{\frac{\partial e}{\partial e}} \underset{\in \mathbb{R}^{n_2\times n_1}}{\frac{\partial d}{\partial d}} \frac{\partial c}{\partial c} \frac{\partial b}{\partial b} \frac{\partial b}{\partial a} = \frac{\partial e}{\partial b} \frac{\partial b}{\partial a}$$

Reverse-Mode AD (a.k.a Backpropagation)

$f: \mathbb{R}^{n_1} \rightarrow \mathbb{R}^{n_2}$

$g: \mathbb{R}^{n_2} \rightarrow \mathbb{R}^{n_3}$

$h: \mathbb{R}^{n_3} \rightarrow \mathbb{R}^{n_4}$

$i: \mathbb{R}^{n_4} \rightarrow \mathbb{R}^{3}$

$$b = f(a)$$
$$c = g(b)$$
$$d = h(c)$$
$$e = i(d)$$

$$\frac{\partial e}{\partial d} = \frac{\partial e}{\partial e}\frac{\partial e}{\partial d} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \frac{\partial e}{\partial d}$$

$\in \mathbb{R}^{3\times 3}$ $\in \mathbb{R}^{3\times n_4}$

$$\frac{\partial e_1}{\partial d}$$

Forward-Mode AD

single variable

$f: \mathbb{R} \to \mathbb{R}^{n_1}$

$g: \mathbb{R}^{n_1} \to \mathbb{R}^{n_2}$

$h: \mathbb{R}^{n_2} \to \mathbb{R}^{n_3}$

$L: \mathbb{R}^{n_3} \to \mathbb{R}^{n_4}$

$$b = f(a)$$
$$c = g(b)$$
$$d = h(c)$$
$$e = L(d)$$

$$\frac{\partial b}{\partial a} = \underset{\in \mathbb{R}^{n_1 \times 1}}{\frac{\partial b}{\partial a}} \underset{\in \mathbb{R}^{1 \times 1}}{\frac{\partial a}{\partial a}} = \frac{\partial b}{\partial a} 1$$

Forward-Mode AD

single variable

$f: \mathbb{R} \to \mathbb{R}^{n_1}$

$g: \mathbb{R}^{n_1} \to \mathbb{R}^{n_2}$

$h: \mathbb{R}^{n_2} \to \mathbb{R}^{n_3}$

$L: \mathbb{R}^{n_3} \to \mathbb{R}^{n_4}$

$$b = f(a)$$
$$c = g(b)$$
$$d = h(c)$$
$$e = L(d)$$

$$\frac{\partial b}{\partial a} = \underset{\in \mathbb{R}^{n_1 \times 1}}{\frac{\partial b}{\partial a}} \underset{\in \mathbb{R}^{1 \times 1}}{\frac{\partial a}{\partial a}} = \frac{\partial b}{\partial a} 1$$

$$\frac{\partial c}{\partial a} = \underset{\in \mathbb{R}^{n_2 \times n_1}}{\frac{\partial c}{\partial b}} \underset{\in \mathbb{R}^{n_1 \times 1}}{\frac{\partial b}{\partial a}} \frac{\partial a}{\partial a} = \frac{\partial c}{\partial b} \frac{\partial b}{\partial a}$$

Forward-Mode AD

single variable

$f: \mathbb{R} \to \mathbb{R}^{n_1}$

$g: \mathbb{R}^{n_1} \to \mathbb{R}^{n_2}$

$h: \mathbb{R}^{n_2} \to \mathbb{R}^{n_3}$

$L: \mathbb{R}^{n_3} \to \mathbb{R}^{n_4}$

$$b = f(a)$$
$$c = g(b)$$
$$d = h(c)$$
$$e = L(d)$$

$$\frac{\partial b}{\partial a} = \underset{\in \mathbb{R}^{n_1 \times 1}}{\frac{\partial b}{\partial a}} \underset{\in \mathbb{R}^{1 \times 1}}{\frac{\partial a}{\partial a}} = \frac{\partial b}{\partial a} 1$$

$$\frac{\partial c}{\partial a} = \underset{\in \mathbb{R}^{n_2 \times n_1}}{\frac{\partial c}{\partial b}} \underset{\in \mathbb{R}^{n_1 \times 1}}{\frac{\partial b}{\partial a}} \frac{\partial a}{\partial a} = \frac{\partial c}{\partial b} \frac{\partial b}{\partial a}$$

$$\frac{\partial d}{\partial a} = \underset{\in \mathbb{R}^{n_3 \times n_2}}{\frac{\partial d}{\partial c}} \underset{\in \mathbb{R}^{n_2 \times 1}}{\frac{\partial c}{\partial b}} \frac{\partial b}{\partial a} \frac{\partial a}{\partial a} = \frac{\partial d}{\partial c} \frac{\partial c}{\partial a}$$

Forward-Mode AD

single variable

$f: \mathbb{R} \to \mathbb{R}^{n_1}$

$g: \mathbb{R}^{n_1} \to \mathbb{R}^{n_2}$

$h: \mathbb{R}^{n_2} \to \mathbb{R}^{n_3}$

$L: \mathbb{R}^{n_3} \to \mathbb{R}^{n_4}$

$$b = f(a)$$
$$c = g(b)$$
$$d = h(c)$$
$$e = L(d)$$

$$\frac{\partial b}{\partial a} = \underset{\in \mathbb{R}^{n_1 \times 1}}{\frac{\partial b}{\partial a}} \underset{\in \mathbb{R}^{1 \times 1}}{\frac{\partial a}{\partial a}} = \frac{\partial b}{\partial a} 1$$

Jacobian-Vector Product (JVP)

$$\frac{\partial c}{\partial a} = \underset{\in \mathbb{R}^{n_2 \times n_1}}{\frac{\partial c}{\partial b}} \underset{\in \mathbb{R}^{n_1 \times 1}}{\frac{\partial b}{\partial a}} \frac{\partial a}{\partial a} = \frac{\partial c}{\partial b} \frac{\partial b}{\partial a}$$

$$\frac{\partial d}{\partial a} = \underset{\in \mathbb{R}^{n_3 \times n_2}}{\frac{\partial d}{\partial c}} \underset{\in \mathbb{R}^{n_2 \times 1}}{\frac{\partial c}{\partial b}} \frac{\partial b}{\partial a} \frac{\partial a}{\partial a} = \frac{\partial d}{\partial c} \frac{\partial c}{\partial a}$$

$$\frac{\partial e}{\partial a} = \underset{\in \mathbb{R}^{n_4 \times n_3}}{\frac{\partial e}{\partial d}} \underset{\in \mathbb{R}^{n_3 \times 1}}{\frac{\partial d}{\partial c}} \frac{\partial c}{\partial b} \frac{\partial b}{\partial a} \frac{\partial a}{\partial a} = \frac{\partial e}{\partial d} \frac{\partial d}{\partial a}$$

Forward-Mode AD

$f: \mathbb{R}^3 \to \mathbb{R}^{n_1}$

$g: \mathbb{R}^{n_1} \to \mathbb{R}^{n_2}$

$h: \mathbb{R}^{n_2} \to \mathbb{R}^{n_3}$

$L: \mathbb{R}^{n_3} \to \mathbb{R}^{n_4}$

$b = f(a)$

$c = g(b)$

$d = h(c)$

$e = L(d)$

What if input is Multi-variables?

$$\frac{\partial b}{\partial a} = \underset{\in \mathbb{R}^{n_1 \times 3}}{\frac{\partial b}{\partial a}} \underset{\in \mathbb{R}^{3 \times 3}}{\frac{\partial a}{\partial a}} = \frac{\partial b}{\partial a} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

3 X Computation

$$\frac{\partial c}{\partial a} = \underset{\in \mathbb{R}^{n_2 \times n_1}}{\frac{\partial c}{\partial b}} \underset{\in \mathbb{R}^{n_1 \times 3}}{\frac{\partial b}{\partial a}} \frac{\partial a}{\partial a} = \frac{\partial c}{\partial b} \frac{\partial b}{\partial a}$$

$$\frac{\partial d}{\partial a} = \underset{\in \mathbb{R}^{n_3 \times n_2}}{\frac{\partial d}{\partial c}} \underset{\in \mathbb{R}^{n_2 \times 3}}{\frac{\partial c}{\partial b}} \frac{\partial b}{\partial a} \frac{\partial a}{\partial a} = \frac{\partial d}{\partial c} \frac{\partial c}{\partial a}$$

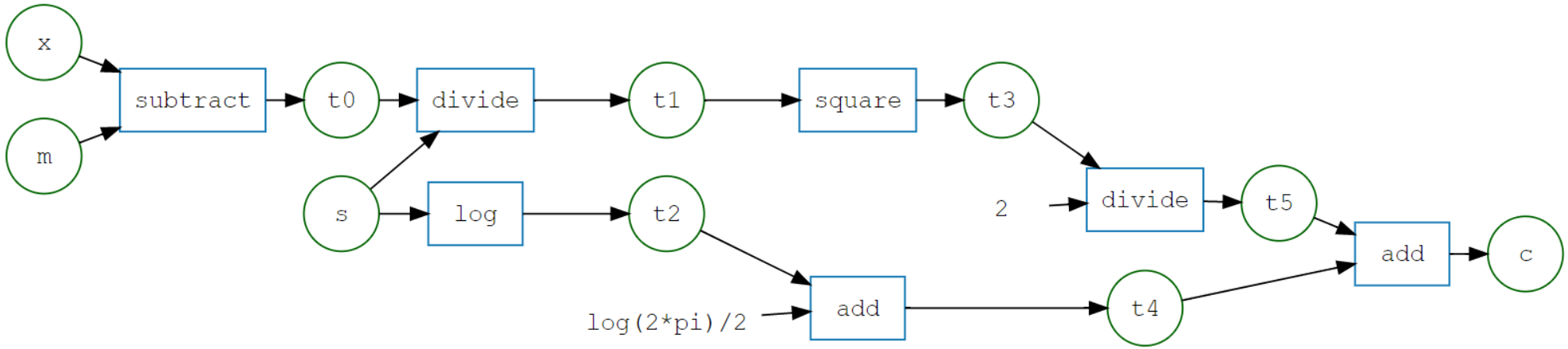
$$\frac{\partial e}{\partial a} = \underset{\in \mathbb{R}^{n_4 \times n_3}}{\frac{\partial e}{\partial d}} \underset{\in \mathbb{R}^{n_3 \times 3}}{\frac{\partial d}{\partial c}} \frac{\partial c}{\partial b} \frac{\partial b}{\partial a} \frac{\partial a}{\partial a} = \frac{\partial e}{\partial d} \frac{\partial d}{\partial a}$$

Automatic Differentiation (AD)

- For low dimensional outputs and high dimensional inputs
  - Objective function w/ deep neural networks
  - reverse-mode AD
- For high dimensional outputs and low dimensional inputs
  - Forward-mode AD

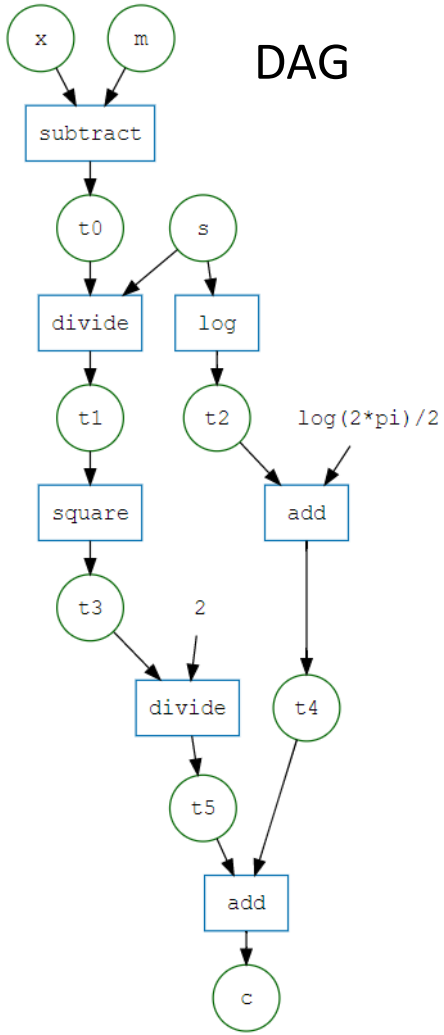
Computational Graph

```
t0 = x - m
t1 = t0 / s
t2 = np.log(s)
t3 = t1**2
t4 = t2 + np.log(2 * np.pi) / 2
t5 = t3 / 2
c = t4 + t5
```

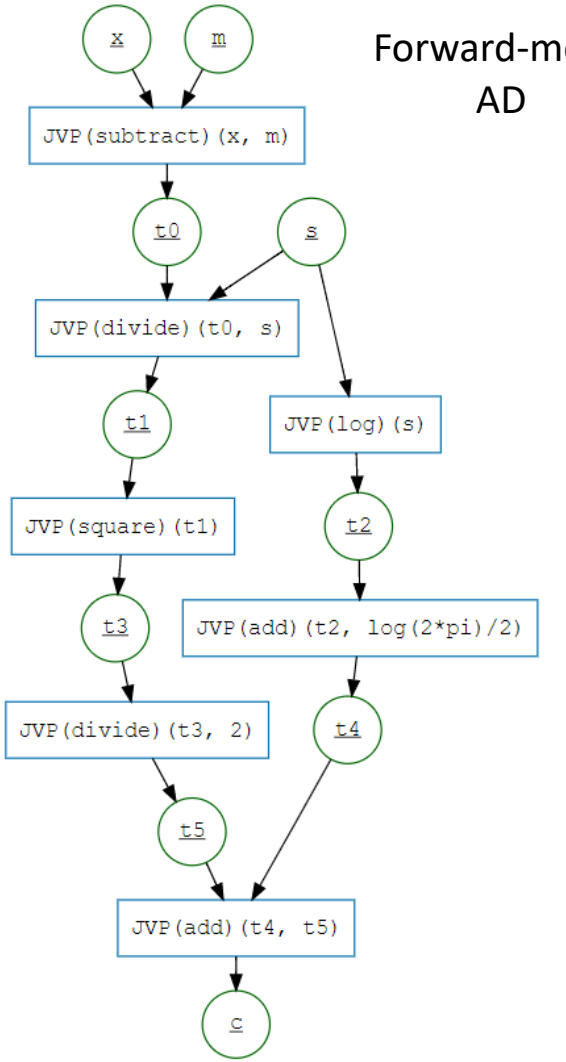


https://matt-graham.github.io/

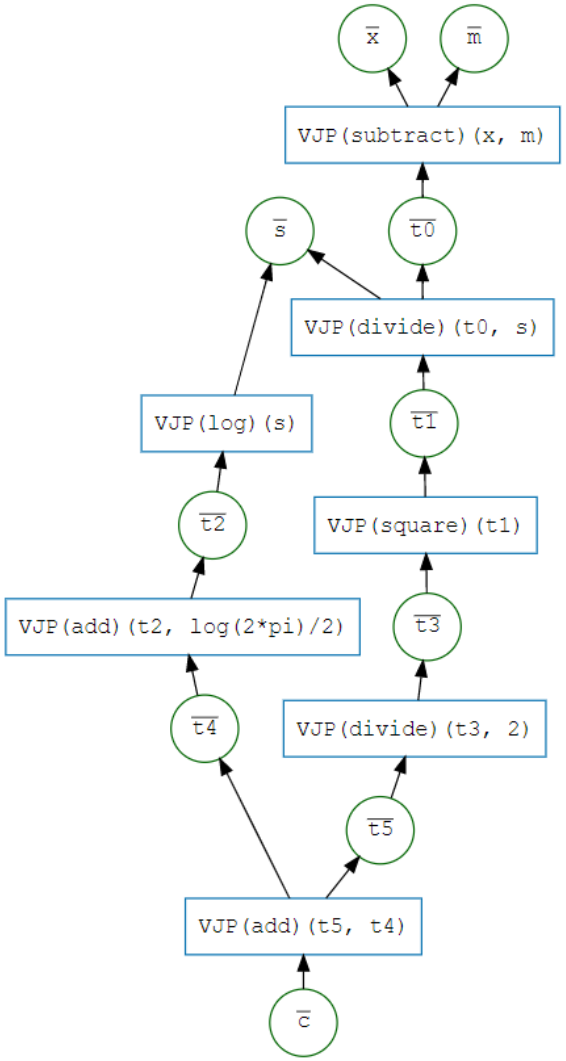
Automatic Differentiation



DAG



Forward-mode AD



Reverse-mode AD

https://matt-graham.github.io/

References

- mattjj/autodidact: A pedagogical implementation of Autograd (github.com)
- [1502.05767] Automatic differentiation in machine learning: a survey (arxiv.org)
- CSC321 Lecture 10: Automatic Differentiation (toronto.edu)