From derivatives to Jacobians

Derivative: scalar input, scalar output

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

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

← how much the "output" changes

← how much the "input" changes

$$f'(x) = \frac{\partial f(x)}{\partial x} = \frac{\partial y}{\partial x} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

$$f(x+h) \cong f'(x)\,h + f(x)$$

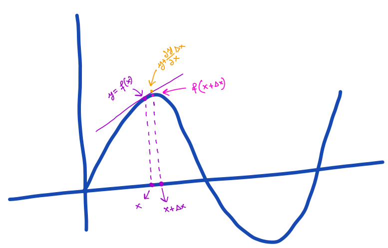
$$f(x+\Delta x) \cong f'(x)\,\Delta x + f(x)$$

$$f(x+\Delta x) \cong \frac{\partial y}{\partial x}\Delta x + f(x)$$

$$y^{NEW} \cong \frac{\partial y}{\partial x}\Delta x + y^{OLD}$$

$$x^{NEW} \to x^{OLD} + \Delta x \Longrightarrow y^{NEW} \overset{\approx}{\to} y^{OLD} + \frac{\partial y}{\partial x}\Delta x$$

"If x is changed by Δx, then y will change approximately by $\frac{\partial y}{\partial x} \cdot \Delta x$"



Chain Rule

$z = f(g(x))$ $x \xrightarrow{g} y \xrightarrow{f} z$

$$x^{NEW} \to x^{OLD} + \Delta x \implies y^{NEW} \xrightarrow{\approx} y^{OLD} + \overbrace{\frac{\partial y}{\partial x}\Delta x}^{\Delta y}$$

$$y^{NEW} \to y^{OLD} + \Delta y \implies z^{NEW} \xrightarrow{\approx} z^{OLD} + \frac{\partial z}{\partial y}\Delta y$$

$$\implies z^{NEW} \xrightarrow{\approx} z^{OLD} + \frac{\partial z}{\partial y} \cdot \frac{\partial y}{\partial x}\Delta x$$

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y} \cdot \frac{\partial y}{\partial x}$$

Gradient : vector input, scalar output

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

$$f\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = y$$

$$\text{gradient} = \left(\frac{\partial y}{\partial x_1}, \frac{\partial y}{\partial x_2}, \ldots\right)$$

$$x^{NEW} \to \underbrace{x^{OLD} + \Delta x}_{\text{vector sum}} \Rightarrow y^{NEW} \overset{\approx}{\to} y^{OLD} + \underbrace{\frac{\partial y}{\partial x} \cdot \Delta x}_{\text{dot product}}$$

$$\frac{\partial y}{\partial x} \cdot \Delta x = \underbrace{\frac{\partial y}{\partial x_1}}_{\text{partial derivative}} \cdot \Delta x_1 + \frac{\partial y}{\partial x_2} \cdot \Delta x_2 + \ldots + \frac{\partial y}{\partial x_N} \cdot \Delta x_N$$

Note: the chain rule applies in the same way as in the scalar case

Jacobian: vector input, vector output

$$f : R^N \rightarrow R^M$$

$$f\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

$$\text{Jacobian} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_N} \\ \vdots & & \vdots \\ \frac{\partial y_M}{\partial x_1} & \cdots & \frac{\partial y_M}{\partial x_N} \end{bmatrix}$$

$$x^{NEW} \rightarrow x^{OLD} + \Delta x \Longrightarrow y^{NEW} \xrightarrow{\approx} y^{OLD} + \underbrace{\frac{\partial y}{\partial x} \Delta x}_{\text{Matrix-vector product}}$$

$$(M \times N) \times (N \times 1) = (M \times 1)$$

Generalized Jacobian: tensor input, tensor output

$$f: \mathbb{R}^{N_1 \times \ldots \times N_{D_x}} \longrightarrow \mathbb{R}^{M_1 \times \ldots M_{D_y}}$$

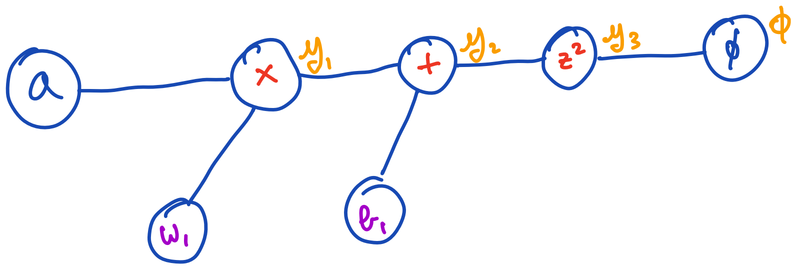
$$f\left(\begin{array}{c}D_x\text{-dimensional} \\ \text{tensor}\end{array}\right) = \begin{array}{c}D_y\text{-dimentional} \\ \text{tensor}\end{array}$$

Generalized Jacobian

$$x^{NEW} \longrightarrow \underbrace{x^{OLD} + \Delta x}_{\text{tensor sum}} \Longrightarrow y^{NEW} \overset{\approx}{\longrightarrow} y^{OLD} + \underbrace{\frac{\partial y}{\partial x}\Delta x}_{\text{Tensor product}}$$

$$(M_1 \times \ldots M_{D_y}) \times (N_1 \times \ldots N_{D_x})$$

Autograd with derivatives



$$\phi = y_3 = (y_2)^2 = (y_1 + b_1)^2 = (a w_1 + b_1)^2$$

$$\frac{\partial \phi}{\partial w_1} = 2(a w_1 + b_1)(a) = 2a(a w_1 + b_1)$$

$$\frac{\partial \phi}{\partial w_1} = \frac{\partial \phi}{\partial y_3} \cdot \frac{\partial y_3}{\partial y_2} \cdot \frac{\partial y_2}{\partial y_1} \cdot \frac{\partial y_1}{\partial w_1} =$$

$$= 1 \cdot 2y_2 \cdot 1 \cdot a =$$

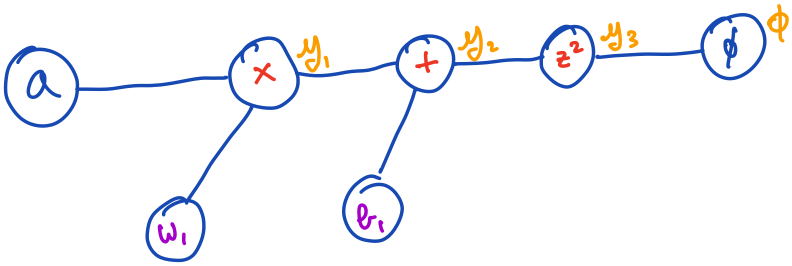
$$= 2a y_2 = 2a(a w_1 + b_1)$$

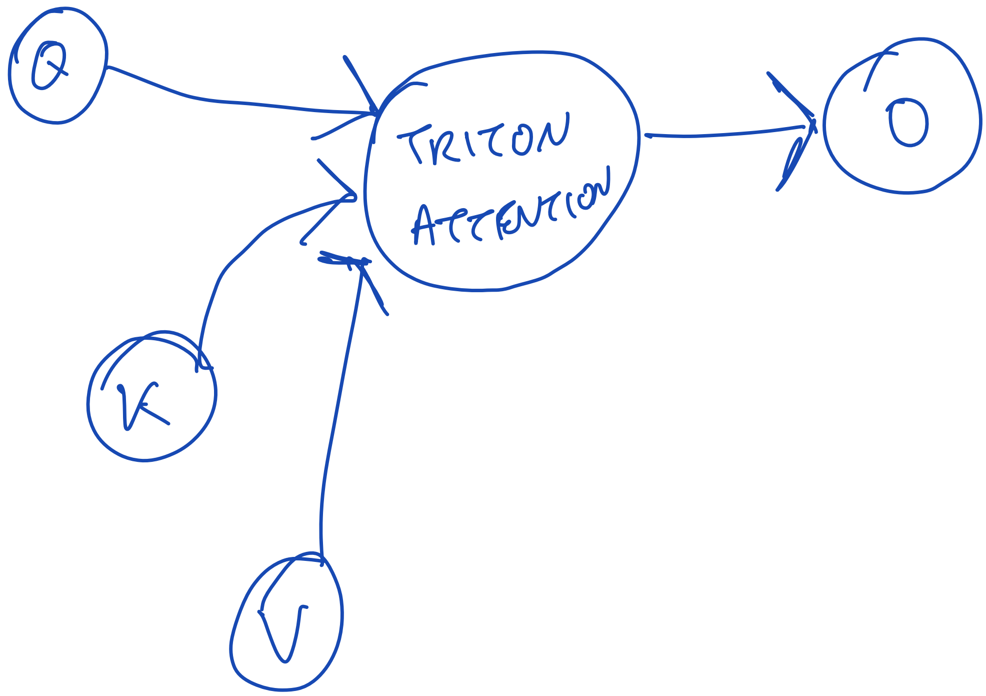
$$\frac{\partial \phi}{\partial w_1} = \overbrace{\underbrace{\frac{\partial \phi}{\partial y_3} \cdot \frac{\partial y_3}{\partial y_2}}_{\frac{\partial \phi}{\partial y_2}} \cdot \frac{\partial y_2}{\partial y_1}}^{\frac{\partial \phi}{\partial y_1}} \cdot \frac{\partial y_1}{\partial w_1}$$

$$\frac{\partial \phi}{y_2} = \frac{\partial \phi}{\partial y_3} \cdot \frac{\partial y_3}{\partial y_2}$$

$$\frac{\partial \phi}{y_1} = \frac{\partial \phi}{y_2} \cdot \frac{\partial y_2}{\partial y_1}$$

$$\frac{\partial \phi}{\partial w_1} = \frac{\partial \phi}{y_1} \cdot \frac{\partial y_1}{\partial w_1}$$





$N = 1024$ $\quad$ 1024

[N, M] $\quad$ [N, D] $\quad$ [D, M] $\quad$ 48

$[N,M]$ $[N,D]$ $[D,M] \rightarrow 20$

$$Y = XW$$

UPSTREAM GRADIENT

$$\frac{\partial \phi}{\partial X} = \frac{\partial \phi}{\partial Y} \cdot \frac{\partial Y}{\partial X}$$

$(N,M) \times (N,D)$

$1024 \times 2048 \times 1024 \times 1024$

LOCAL JACOBIAN

DOWNSTREAM GRADIENT