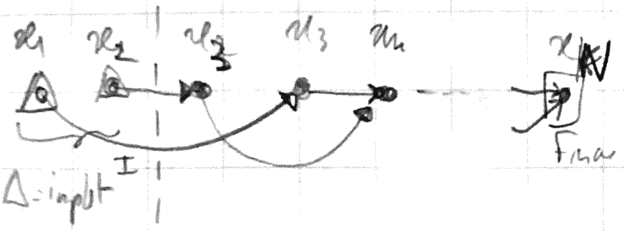
Automatic Differentiation

Math formula ⟷ several implementation ⟷ best impl: (very hard)



Example $f(x) = \log x + \sqrt{\log x}$

$y = \log x$
$z = \sqrt{y}$
$f = y + z$

∂x_1 then ∂x_2

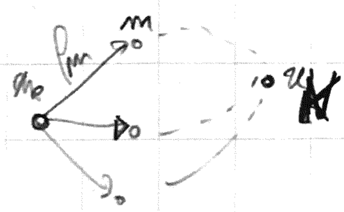
For $k = I+1 \dots N$

$\quad x_k = f_k(x_1 \text{---} x_{k-1})$.

FWD

For $k = I+1 \dots N$

$$\frac{\partial x_k}{\partial x_1} = \sum_{\substack{l \in \mathrm{Father}(k) \\ l < k}} \text{"}\frac{\partial x_k}{\partial x_l}\text{"} \times \frac{\partial x_l}{\partial x_1} = \underbrace{\sum_{l \in \mathrm{Father}(k)} \overbrace{\frac{\partial f_k}{\partial x_l}(x_1 \text{---} x_{k-1})}^{\text{Formula}} \times \frac{\partial x_l}{\partial x_1}}$$



For $k = N-1 \dots I$

$$\frac{\partial x_N}{\partial x_k} = \sum_{\substack{m \in \mathrm{Son}(k) \\ m > k}} \text{"}\frac{\partial x_N}{\partial x_m} \times \left(\frac{\partial x_m}{\partial x_k}\right)\text{"} = \sum_{m \in \mathrm{Son}(k)} \frac{\partial x_N}{\partial x_m} \times \frac{\partial f_m}{\partial x_k}(\underbrace{x_1 \text{---} x_m})$$

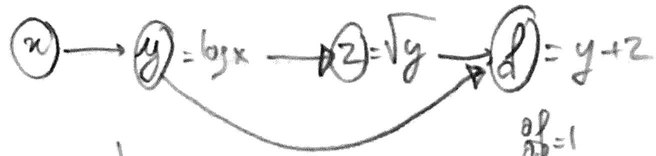
Need to run first the algo once

For gradient, $\overset{f(x_1 \text{---} x_I)}{\text{"}x_N\text{"}} \in \mathbb{R}$ $\qquad \nabla_{x_k} f = \left[\frac{\partial x_N}{\partial x_k}\right]^T$

$$\nabla_{x_k} f = \sum_{m \in \mathrm{Son}(k)} \left[\frac{\partial f}{\partial x_k}(x_1 \text{---} x_m)\right]^T \times \nabla_{x_m} f$$

- → Simple example $\mathbb{R} \to \mathbb{R}$
- → Feed Forward $\mathbb{R}^n \to \mathbb{R}$ (matrix mult)
- → MLP example
- → Res Net

$$f(x) = \log(x) + \sqrt{\log x}$$



$\boxed{\text{FWD}}$

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

$$\frac{\partial y}{\partial x} = \text{"}\frac{\partial y}{\partial x}\text{"} \cdot \frac{\partial x}{\partial x} = \frac{1}{x} \times 1$$

$$\frac{\partial z}{\partial x} = \text{"}\frac{\partial z}{\partial y}\text{"} \times \frac{\partial y}{\partial x} = \frac{1}{2\sqrt{y}} \times \frac{\partial y}{\partial x}$$

$z \to f$, $y \to f$

$$\frac{\partial f}{\partial x} = \text{"}\frac{\partial f}{\partial z}\text{"} \times \frac{\partial z}{\partial x} + \text{"}\frac{\partial f}{\partial y}\text{"} \times \frac{\partial y}{\partial x} = 1 \times \frac{\partial z}{\partial x} + 1 \times \frac{\partial y}{\partial x}$$

$\boxed{\text{BWD}}$

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

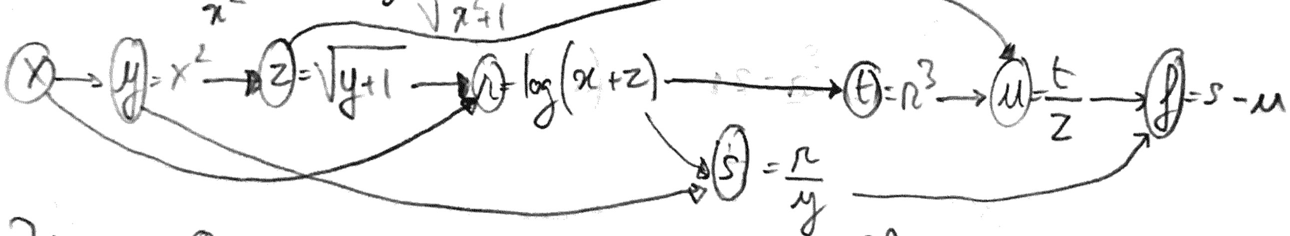
$$\frac{\partial f}{\partial z} = \frac{\partial f}{\partial f} \times \text{"}\frac{\partial f}{\partial z}\text{"} = 1 \times 1$$

$y \to z$, $y \to f$

$$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial z} \times \text{"}\frac{\partial z}{\partial y}\text{"} + \frac{\partial f}{\partial f} \cdot \text{"}\frac{\partial f}{\partial y}\text{"} = \frac{\partial f}{\partial z} \times \frac{1}{2\sqrt{y}} + 1 \times 1$$

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial y} \times \text{"}\frac{\partial y}{\partial x}\text{"} = \frac{\partial f}{\partial y} \times \frac{1}{x}.$$

$$f(x) = \frac{\log(x + \sqrt{x^2+1})}{x^2} - \frac{\log^3(x+\sqrt{x^2+1})}{\sqrt{x^2+1}}$$



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

$$\frac{\partial y}{\partial x} = \text{"}\frac{\partial y}{\partial x}\text{"} \cdot \frac{\partial x}{\partial x} = 2x \cdot \frac{\partial x}{\partial x} \qquad x \to y$$

$$\frac{\partial z}{\partial x} = \text{"}\frac{\partial z}{\partial y}\text{"} \cdot \frac{\partial y}{\partial x} = \frac{1}{2\sqrt{y+1}} \cdot \frac{\partial y}{\partial x} \qquad y \to z$$

$$\frac{\partial r}{\partial x} = \text{"}\frac{\partial r}{\partial x}\text{"} \frac{\partial x}{\partial x} + \text{"}\frac{\partial r}{\partial z}\text{"} \frac{\partial z}{\partial x} = \frac{1}{x+r} \frac{\partial x}{\partial x} + \frac{1}{x+r} \frac{\partial z}{\partial x} \qquad x \to r,\ z \to r$$

$$\frac{\partial s}{\partial x} = \text{"}\frac{\partial s}{\partial r}\text{"} \frac{\partial r}{\partial x} + \text{"}\frac{\partial s}{\partial y}\text{"} \frac{\partial y}{\partial x} = \frac{1}{y} \frac{\partial r}{\partial x} - \frac{r}{y^2} \frac{\partial y}{\partial x} \qquad r \to s,\ y \to s$$

$$\frac{\partial t}{\partial x} = \text{"}\frac{\partial t}{\partial r}\text{"} \frac{\partial r}{\partial x} = 3r^2 \frac{\partial r}{\partial x} \qquad r \to t$$

$$\frac{\partial u}{\partial x} = \text{"}\frac{\partial u}{\partial t}\text{"} \frac{\partial t}{\partial x} + \text{"}\frac{\partial u}{\partial z}\text{"} \frac{\partial z}{\partial x} = \frac{1}{z} \frac{\partial t}{\partial x} - \frac{t}{z^2} \frac{\partial z}{\partial x} \qquad t \to u,\ z \to u$$

$$\frac{\partial f}{\partial x} = \text{"}\frac{\partial f}{\partial s}\text{"} \frac{\partial s}{\partial x} + \text{"}\frac{\partial f}{\partial u}\text{"} \frac{\partial u}{\partial x} = 1 \frac{\partial s}{\partial x} - 1 \frac{\partial u}{\partial x} \qquad s \to f,\ u \to f$$

$\boxed{\text{FWD mode}}$

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

$$\frac{\partial f}{\partial u} = \frac{\partial f}{\partial f} \times \text{"}\frac{\partial f}{\partial u}\text{"} = \frac{\partial f}{\partial f} \cdot (-1) \qquad u \to f$$

$$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial u} \times \text{"}\frac{\partial u}{\partial t}\text{"} = \frac{\partial f}{\partial u} \times \frac{1}{z} \qquad t \to u$$

$$\frac{\partial f}{\partial s} = \frac{\partial f}{\partial f} \times \text{"}\frac{\partial f}{\partial s}\text{"} = \frac{\partial f}{\partial f} (+1) \qquad s \to f$$

$$\frac{\partial f}{\partial r} = \frac{\partial f}{\partial t} \cdot \text{"}\frac{\partial t}{\partial r}\text{"} + \frac{\partial f}{\partial s} \text{"}\frac{\partial s}{\partial r}\text{"} = \frac{\partial f}{\partial t} 3r^2 + \frac{\partial f}{\partial s} \frac{1}{y} \qquad r \to t,\ r \to s$$

$$\frac{\partial f}{\partial z} = \frac{\partial f}{\partial u} \text{"}\frac{\partial u}{\partial z}\text{"} + \frac{\partial f}{\partial r} \text{"}\frac{\partial r}{\partial z}\text{"} = \frac{\partial f}{\partial u} \frac{-t}{z^2} + \frac{\partial f}{\partial r} \frac{1}{x+z} \qquad z \to u,\ z \to r$$

$$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial z} \text{"}\frac{\partial z}{\partial y}\text{"} + \frac{\partial f}{\partial s} \text{"}\frac{\partial s}{\partial y}\text{"} = \frac{\partial f}{\partial z} \frac{1}{2\sqrt{y+1}} + \frac{\partial f}{\partial s} \times \frac{-r}{y^2} \qquad y \to z,\ y \to s$$

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial y} \text{"}\frac{\partial y}{\partial x}\text{"} + \frac{\partial f}{\partial r} \times \text{"}\frac{\partial r}{\partial x}\text{"} = \frac{\partial f}{\partial y} 2x + \frac{\partial f}{\partial r} \times \frac{1}{x+z} \qquad x \to y,\ x \to r$$

$\boxed{\text{BWD mode}}$

Feed Forward: $\overset{\overset{\mathbb{R}^{n_1}}{\Downarrow}}{x_1} \xrightarrow{f_1} \overset{\overset{\mathbb{R}^{n_2}}{\Downarrow}}{x_2} \xrightarrow{f_2} x_3 \to \cdots \xrightarrow{f_{N-1}} \overset{\overset{\mathbb{R}^{n_N}}{\Downarrow}}{x_N} = f(x_1)$

i.e. $f = f_{N-1}\circ f_{N-2}\circ \cdots \circ f_2 \circ f_1$ $\qquad x_{k+1} = f_k(x_k)$

Jacobian: $\partial f(x_1) = \underbrace{\partial f_{N-1}(x_{N-1})}_{\substack{A_{N-1} \\ \mathbb{R}^{n_N\times n_{N-1}}}} \times \underbrace{\partial f_{N-2}(x_{N-2})}_{\substack{A_{N-2} \\ \mathbb{R}^{n_{N-1}\times n_{N-2}}}} \times \cdots \times \underbrace{\partial f_2(x_2)}_{\substack{A_2 \\ \mathbb{R}^{n_3\times n_2}}} \cdot \underbrace{\partial f_1(x_1)}_{\substack{A_1 \\ \mathbb{R}^{n_2\times n_1}}}$

$f_k : \mathbb{R}^{n_k} \to \mathbb{R}^{n_{k+1}}$

FWD mode: $A_{N-1}\times\Big(A_{N-2}\times\Big(\cdots\Big(A_3\times\big(A_2\times A_1\big)\Big)\cdots$

$n_N\times n_{N-1}\times n_1 \qquad n_{N-1}\times n_{N-2}\times n_1 \qquad n_4 n_3 n_1 \qquad n_3 n_2 n_1 \qquad n_3\times n_1$

$$\boxed{\underset{\mathbb{R}^{n\times p}}{A} \times \underset{\mathbb{R}^{p\times q}}{B} \rightsquigarrow \text{cpxity } npq}$$

BWD mode: $\Big(\Big(\big(A_{N-1}\times A_{N-2}\big)\times A_{N-3}\Big)\times\cdots\times A_3\Big)\times A_2\Big)\times A_1$

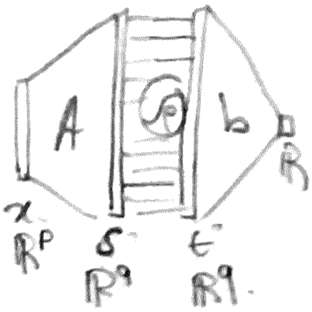
$n_N\times n_{N-1}\times n_{N-2} \qquad n_N n_{N-2} n_{N-3} \qquad n_N n_4 n_3 \qquad n_N n_3 n_2 \qquad n_N n_2 n_1$

Complexity: $\underline{\text{FWD}} = n_1 \sum_{k=2}^{N-1} n_k n_{k+1} \qquad \underline{\text{BWD}} = n_N \sum_{k=1}^{N-2} n_k n_{k+1}$

if $\underset{\gg}{n_1 \ll n_N} \to \underset{\gg}{\text{FWD} \ll \text{BWD}}$

if $n_1 = 1$: FWD optimal (for any graph)
$n_N = 1$ (grad. comput°): BWD optimal (for any graph)

$\boxed{\text{MLP}}$: $f(x,A,b) = b\,\rho(Ax)$. $\quad A\in\mathbb{R}^{q\cdot p}$, $b\in\mathbb{R}^{1\cdot q}$



$$A,\quad b,\quad x \to s = Ax \to t = \rho(s) \to f = bt, \qquad A \to s, \quad b \to f$$

$$\frac{\partial x}{\partial A} = 0 \qquad \frac{\partial b}{\partial A} = 0 \qquad \frac{\partial A}{\partial A} = \mathrm{Id}_{(qp)(qp)}$$

$$\frac{\partial s}{\partial A} = \text{"}\frac{\partial s}{\partial x}\text{"}\underbrace{\frac{\partial x}{\partial A}}_{0} + \text{"}\frac{\partial s}{\partial A}\text{"}\frac{\partial A}{\partial A} = \underbrace{R_x}_{}\cdot\frac{\partial A}{\partial A} \qquad x \to s,\ A \to s$$

$$B\in\mathbb{R}^{q\cdot p} \to Bx\in\mathbb{R}^q$$

$$\frac{\partial t}{\partial A} = \text{"}\frac{\partial t}{\partial s}\text{"}\frac{\partial s}{\partial A} = \mathrm{diag}(\rho'(s))\frac{\partial s}{\partial A} \qquad s \to t$$

$$\frac{\partial f}{\partial A} = \text{"}\frac{\partial f}{\partial b}\text{"}\underbrace{\frac{\partial b}{\partial a}}_{0} + \text{"}\frac{\partial f}{\partial t}\text{"}\cdot\frac{\partial t}{\partial A} = \underbrace{L_b}_{}\frac{\partial t}{\partial A} \qquad b \to f,\ t \to f$$

$$\boxed{t\in\mathbb{R}^q \to bt\in\mathbb{R}}$$

FWD (this is crazy...)

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

$$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial f}\cdot\text{"}\frac{\partial f}{\partial t}\text{"} = \frac{\partial f}{\partial f}\,b \xrightarrow{T} \nabla_t f = b^T \qquad t \to f$$

$$\frac{\partial f}{\partial s} = \frac{\partial f}{\partial t}\,\text{"}\frac{\partial t}{\partial s}\text{"} = \frac{\partial f}{\partial t}\cdot\mathrm{diag}(\rho'(s)) \overset{T}{\rightsquigarrow} \nabla_s f = \mathrm{diag}(\rho'(s))\,\nabla_t f \qquad s \to t$$

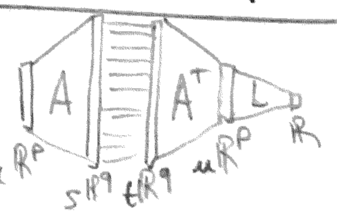
$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial s}\cdot\text{"}\frac{\partial s}{\partial x}\text{"} = \frac{\partial f}{\partial s}\cdot A \to \nabla_x f = A^T\frac{\partial f}{\partial s} \qquad x \to s$$

$$\frac{\partial f}{\partial b} = \frac{\partial f}{\partial f}\,\text{"}\frac{\partial f}{\partial b}\text{"} = 1\cdot t \rightsquigarrow \nabla_b f = t^T \qquad b \to f$$

$\boxed{\text{BWD}}$

$\boxed{\text{Directly}}$:
$$f(x, A+\varepsilon, b) = b\rho(Ax + \varepsilon x) = b\rho(Ax) + b(\rho'(Ax)\odot\varepsilon x) = f(x,A,b) + \langle b^T, \rho'(Ax)\odot\varepsilon x\rangle$$
$$= f(x,A,b) + \langle\varepsilon, [\rho'(Ax)\odot b^T]\,x^T\rangle \Rightarrow \nabla_A f(x,A,b) = [\rho'(Ax)\odot b^T]\,x^T$$
$$f(x,A,b+\varepsilon) = (b+\varepsilon)\rho(Ax) = f(x,A,b) + \langle\varepsilon, \rho(Ax)^T\rangle \Rightarrow \nabla_b f(x,A,b) = \rho(Ax)^T$$

<u>ResNet</u>: $f(x,A) = \underbrace{L}_{\text{Loss, etc}}(A^T\rho(Ax) + x)$



$$A \to s = Ax \to t = \rho(s) \to u = A^T t + x \to f = L(u), \qquad x \to u, \quad A \to u$$

$$\frac{\partial f}{\partial f} = 1;\quad \nabla_u L = \underline{\text{init}};\quad \nabla_t = A\nabla_u;\quad \nabla_s = \mathrm{diag}(\rho'(s))\,\nabla_t$$

$$\nabla_A = \nabla_s x^T + (\nabla_u)^T t$$

$$A \to s,\quad A \to u$$