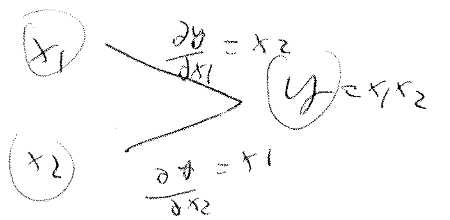
Last lecture Chris R showed how to learn a differential equation meaning to learn the paramaters like learning a neural network.

The derivation was fast & it's worth looking over.

1. Chris's autodiff everything is a vector and a Jacobian

ex1 Graphical as vector - Jacobian

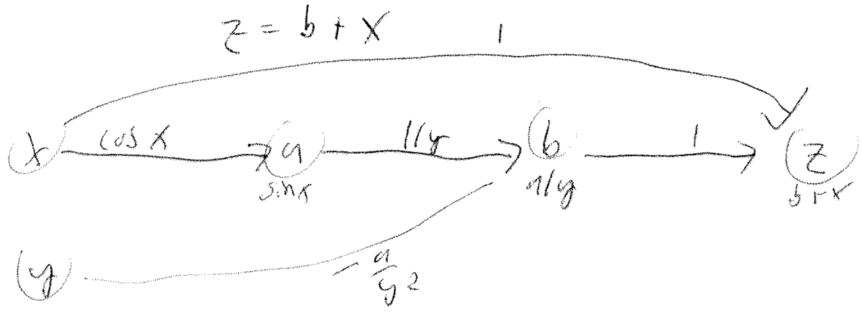


$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \xrightarrow{*} y = x_1 x_2$$

$$J = \begin{pmatrix} x_2 & x_1 \end{pmatrix}$$

ex2) x, y

$a = \sin x$
$b = a/y$
$z = b + x$



Flow written right to left

$$z \longleftarrow \begin{pmatrix} b \\ x \end{pmatrix} \longleftarrow \begin{pmatrix} a \\ x \\ y \end{pmatrix} \longleftarrow \begin{pmatrix} x \\ y \end{pmatrix}$$

a & y can be "garbage collected"

3/20/2023 (2)

$$\begin{pmatrix} da \\ dx \\ dy \end{pmatrix} = \begin{pmatrix} \cos x & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} dx \\ dy \end{pmatrix}$$

$$\begin{pmatrix} db \\ dx \end{pmatrix} = \begin{pmatrix} 1/y & 0 & -a/y^2 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} da \\ dx \\ dy \end{pmatrix}$$

$$(dz) = (1 \quad 1) \begin{pmatrix} db \\ dx \end{pmatrix}$$

$$dz = (1 \quad 1) \begin{pmatrix} 1/y & 0 & -a/y^2 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \cos x & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} dx \\ dy \end{pmatrix}$$

$$1 = \nabla_z z$$

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix} = (1 \quad 1)^T = \nabla_{\binom{b}{x}} z$$

$$\begin{pmatrix} 1/y & 0 \\ 0 & 1 \\ -\frac{a}{y^2} & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \nabla_{\left(\begin{smallmatrix} a \\ x \\ y \end{smallmatrix}\right)} z$$

$$\begin{pmatrix} \cos x & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{y} & 0 \\ 0 & 1 \\ -\frac{a}{y^2} & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \nabla_{\binom{x}{y}} z$$

The pullback function

Let $y = f(x)$ from $\mathbb{R}^n$ to $\mathbb{R}^m$

If $J \in \mathbb{R}^{m,n}$ is the Jacobian matrix at x

$$J = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{pmatrix}$$

then ~~the pullback~~ . $dx \to J\,dx$ is known as the push-forward function. The term comes from differential geometry and applies to any manifold.

If you have a sequence of functions

$f_1, f_2, \ldots, f_K$

then

$J_K \ldots J_2 J_1 v$ is the direction derivative of the composition in the direction v

and $(J_K \ldots J_1)\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$ is a column of the total Jacobian

If $y = f(x)$ (9)

The pullback function

$$\bar{x} = J^T \bar{y} \quad \text{takes } \mathbb{R}^m \text{ to } \mathbb{R}^n$$

It is written $\bar{x} = B_f^x(\bar{y})$

noting that the entries of B_f^x are always functions of x,

e.g. $$f\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} e^{x_1} x_2 \\ \frac{1}{2}\sin(x_1^2 + x_2^2) \\ x_1/x_2 \end{pmatrix}$$

$$J = \begin{pmatrix} e^{x_1} x_2 & e^{x_1} \\ \cos(x_1^2+x_2^2)x_1 & \cos(x_1^2+x_2^2)x_2 \\ 1/x_2 & -x_1/x_2^2 \end{pmatrix}$$

$$B_f^x(\bar{y}) = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \bar{y}_3 \end{pmatrix} \rightarrow \begin{pmatrix} e^{x_1} x_2 & \cos(x_1^2+x_2^2)x_1 & 1/x_2 \\ e^{x_1} & \cos(x_1^2+x_2^2)x_2 & -x_1/x_2^2 \end{pmatrix} \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \bar{y}_3 \end{pmatrix}$$

$$\uparrow$$

$$\begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} & \frac{\partial y_3}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} & \frac{\partial y_3}{\partial x_3} \end{pmatrix}$$

As chris mentioned

$$\begin{pmatrix} J_1^T & \cdots & J_{k-1}^T & J_k^T \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \text{ is}$$

a row in the total Jacobian (written as a column)

& if f_K is a scalar $J_K^T = \nabla_{x_{k-1}} f_K$

Also if f_K is a scalar $f_1 \circ f_2 \circ \ldots \circ f_K$ is a scalar and

$$J_1^T \cdots J_K^T \text{ is } \nabla_{x_1} f_1 \circ \ldots \circ f_K$$

What about matrix to matrix operations like

$Y = X^{-1}$ or matrix to scalar

$$y = \operatorname{tr} M^T X = M \circ X$$

Choice 1:

$$\underset{n^2}{\operatorname{vec}(dY)} = \underset{\underset{J}{n^2 \times n^2}}{-\left(X^{-T} \otimes X^{-1}\right)} \underset{n^2}{\operatorname{vec}(dX)}$$

$$J^T = -\left(X^{-1} \otimes X^{-T}\right)$$ Kronecker Products work that way

$$d\bar{X} = -\left(X^{-1} \otimes X^{-T}\right) \operatorname{vec}(d\bar{Y})$$

$$X \to \operatorname{tr} M^T X = M \circ X$$

$$\mathrm{vec}(X) \to \mathrm{vec}(M) \cdot \mathrm{vec}(X)$$

$$\mathrm{vec}(y) = \mathrm{vec}(M)^T \mathrm{vec}(X)$$

$$\nabla_{\mathrm{vec}(X)} y = \mathrm{vec}(M)$$

Feels ~~Meo~~chanical, which is OK,
I feel adjoints are for grown ups

Recall ~~$\langle y, x \rangle$~~

$\langle u, v \rangle$ = inner product on vector space

$$\langle u, v \rangle = \langle v, u \rangle$$
$$\langle u, u \rangle \geq 0 \quad (= \text{ iff } u = 0)$$
$$\langle u_1 + u_2, v \rangle = \langle u_1, v \rangle + \langle u_2, v \rangle$$
$$\langle \alpha u, v \rangle = \alpha \langle u, v \rangle \quad \text{etc}$$

$\mathcal{L}$ is a linear operator from V_1 to V_2 with inner prods $\langle \rangle_1$ & $\langle \rangle_2$

$\mathcal{L}^T$ is the operator from V_2 to V_1

$\forall$ ~~$\langle \mathcal{L}^T y_1, y_2 \rangle_2 = \ldots$~~

$$\langle \mathcal{L}^T y, x \rangle_1 = \langle y, \mathcal{L} x \rangle_2 \quad \forall x \in V_1, \; y \in V_2$$

7

e.g. $V_1 = \mathbb{R}^{n,n}$ $V_2 = \mathbb{R}$

$\mathcal{L} = \operatorname{tr} M^T X$ is already linear

$\langle y_1, y_2 \rangle_{\mathbb{R}} = y_1 y_2$

$\langle X_1, X_2 \rangle = X_1 \circ X_2 = \operatorname{tr} X_1^T X_2$

$\langle y, \mathcal{L}X \rangle = y \operatorname{tr} M^T X$

$= \operatorname{tr} (yM)^T X$

$= \langle yM, X \rangle$

$\mathcal{L}^T y = My$

Similarly ~~$\langle Y, \mathcal{L}X^{-1}$~~

$\langle Y, -X^{-1}VX \rangle = -\operatorname{tr} Y^T X^{-1} V X$

~~$= -\operatorname{tr} X^T V^T X^{-T} Y$~~

~~$= -\operatorname{tr} X^{-1} V X$~~

~~$= -\operatorname{tr} X^T V^T X^{-T} Y$~~

$= -\operatorname{tr} X^{-T} Y X^T V$

$= -\operatorname{tr}\left((X^{-T} Y X^T)^T V \right)$

so

$(V \to -X^{-1}VX)^T$

$=$

$(Y \to -X^{-T} Y X^T)$

Now $\frac{d}{dx}$ takes functions to functions

$\langle f, g \rangle = \int_I f(x) g(x)\, dx$ is a

common inner product.

Suppose $I = (a,b)$ + $f(a) = f(b) = 0$ for simplicity

$$\int_a^b \left[f(x) \frac{d}{dx} g(x)\right] dx = -\int \left[g(x) \frac{d}{dx} f(x)\right] dx$$

$$\Rightarrow \left(\frac{d}{dx}\right)^T = -\frac{d}{dx}$$