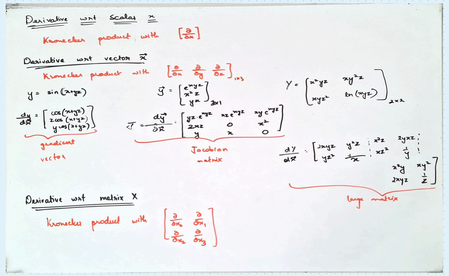


Derivatives of Scalar, Vector, Tensor

	scalar	vector	tensor
scalar	s	$\vec{v}_{size}$	T_{size}
vector	gradient $\vec{v}_{size}$	Jacobian $\vec{t}$	$t_{size_1 size_2}$
tensor	t_{size}	t_{size_1}	$t_{size_1 size_2}$

Derivative of Element-Wise Binary Operator

day we have $f(\vec{x}) \bigcirc g(\vec{x}) = y$, ($\bigcirc$ element-wise binary operator)

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} f_1(\vec{x}) \bigcirc g_1(\vec{x}) \\ f_2(\vec{x}) \bigcirc g_2(\vec{x}) \\ \vdots \\ f_n(\vec{x}) \bigcirc g_n(\vec{x}) \end{bmatrix}$$

$$\text{Jacobian} = \frac{\partial \vec{y}}{\partial \vec{x}} = \begin{bmatrix} \frac{\partial}{\partial x_1}(f_1(\vec{x}) \bigcirc g_1(\vec{x})) & \cdots & \frac{\partial}{\partial x_n}(f_1(\vec{x}) \bigcirc g_n(\vec{x})) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1}(f_n(\vec{x}) \bigcirc g_1(\vec{x})) & \cdots & \frac{\partial}{\partial x_n}(f_n(\vec{x}) \bigcirc g_n(\vec{x})) \end{bmatrix}$$

Here, for all i≠j, derivative becomes 0 for f_i.

$$\Rightarrow \frac{\partial \vec{y}}{\partial \vec{x}} = \text{diag}\left[\frac{\partial}{\partial x_1}(f_1(\vec{x}) \bigcirc g_1(\vec{x})), \cdots, \frac{\partial}{\partial x_n}(f_n(\vec{x}) \bigcirc g_n(\vec{x}))\right]$$

Say you consider vector addition.

$$f_i(\vec{w}) + g_i(\vec{x}) = w_i + x_i$$

$$\frac{\partial \vec{y}}{\partial \vec{w}} = \text{diag}\left[\frac{\partial}{\partial w_1}(w_1 + x_1), \cdots, \frac{\partial}{\partial w_n}(w_n + x_n)\right]$$

$$= I$$

Operator	Derivative
+	I
−	I
⊙ Hadamard multiply	$\text{diag}\left[\frac{\partial}{\partial w_i}\begin{bmatrix} w_1 x_1 \\ \vdots \\ w_n x_n \end{bmatrix}\right] = \vec{x}$
⊘ Hadamard division	$\text{diag}\left[\frac{\partial}{\partial w_i}\begin{bmatrix} w_1/x_1 \\ \vdots \\ w_n/x_n \end{bmatrix}\right] = \text{diag}\left[\frac{1}{x_i} \cdots\right]$
	$\text{diag}\left[\frac{\partial}{\partial x_i}\begin{bmatrix} w_1/x_1 \\ \vdots \\ w_n/x_n \end{bmatrix}\right] = \text{diag}\left[\cdots \frac{-w_i}{x_i^2} \cdots\right]$

Jacobians of scalar-vector multiplication

We can do the same as above by considering the scalar as a vector.

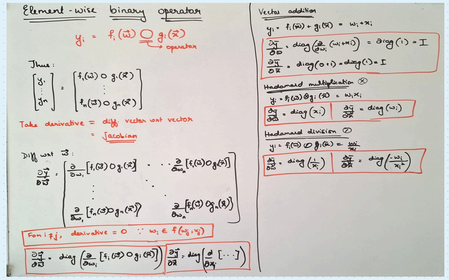
① $\vec{y} = \vec{x} + z = \vec{f}(\vec{x}) + \vec{g}(z)$ ($\vec{f}$: $\vec{x}$, $\vec{g}$: $\vec{1}z$, suitable dimension)

$$J(\vec{y}) = \text{diag}\left[\cdots \frac{\partial}{\partial x_i}[x_i + z] \cdots\right] = \text{diag}[\cdots 1 \cdots] = I$$ same if you diff wrt z

② $\vec{y} = \vec{x}z = \vec{f}(\vec{x}) \otimes \vec{g}(z)$

$$J(\vec{y}) = \text{diag}\left[\cdots \frac{\partial}{\partial x_i}[x_i z] \cdots\right] = \text{diag}[\cdots z \cdots] = Iz$$

$$\frac{\partial}{\partial z}(\vec{x}z) = \text{diag}\left[\cdots \frac{\partial}{\partial z}[x_i z] \cdots\right] = \text{diag}[\cdots x_i \cdots] = \vec{x}$$



Automatic Differentiation

$f(x) = \ln(x^2+1)$

Forward mode:

u	Value	Derivative
x^2+1	x^2+1	$2x$
$\ln(u)$	$\ln(x^2+1)$	$\frac{1}{u} \cdot \frac{1}{x^2+1}$

Total derivative = $\frac{2x}{x^2+1}$

Reverse mode:

$$\frac{\partial \vec{w_i}}{\partial \vec{w_j}} = \sum_{j \in \text{children}(i)} \frac{\partial \vec{w_j}}{\partial \vec{w_i}} \cdot \frac{\partial \vec{w_j}}{\partial \vec{w_i}}$$

$f(xy) = y \cdot \sin(x+y)$. Find $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}$ using reverse mode AD.

$u = x+y$, $v = \sin u$, $f = yv$.

Jacobian = vector of functions

$$\begin{bmatrix} \nabla f(x,y) \\ \nabla g(x,y) \end{bmatrix} = \begin{bmatrix} \frac{\partial f(x,y)}{\partial x} & \frac{\partial f}{\partial y} \\ \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \end{bmatrix}$$ → numerator layout, denominator layout also possible

Single variable chain rule

forward prop: $x \xrightarrow{u} y$ $\quad \frac{dy}{dx} = \frac{du}{dx} \cdot \frac{dy}{du}$

backward prop: $y \xleftarrow{u} x$ $\quad \frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$

Single variable total derivative

Assume $u_1 \ldots u_n$ intermediate variables

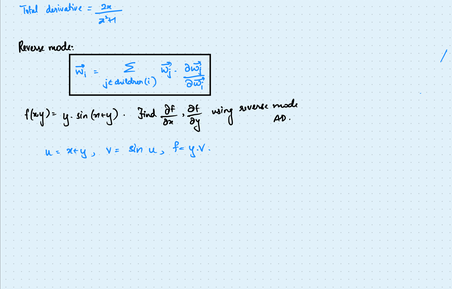
$$\frac{\partial f}{\partial x}(x, u_1, \ldots, u_n) = \frac{\partial f}{\partial x} + \sum_{i=1}^{n} \frac{\partial f}{\partial u_i} \cdot \frac{\partial u_i}{\partial x} = \sum_{i=0}^{n} \frac{\partial f}{\partial u_i} \cdot \frac{\partial u_i}{\partial x}$$

Multiplication of 2 functions

$$\frac{dy}{dx} = \begin{bmatrix} \frac{\partial f(g)}{\partial x} \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial g} & \frac{\partial f}{\partial z} \end{bmatrix} \begin{bmatrix} \frac{\partial g}{\partial x} \\ \frac{\partial z}{\partial x} \end{bmatrix}$$ assuming x is a scalar

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

Chain rule is extended for vectors using Jacobians



Matrix Differentiation Utils

① $y = mx1$, $x = n \times 1$, $A = m \times n$

$$y = Ax = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

$$y_i = \sum_{j=1}^{n} a_{ij} x_j \qquad \frac{\partial y_i}{\partial x_j} = a_{ij}$$

$$\Rightarrow y = \begin{bmatrix} \sum_{j=1}^{n} a_{1j} x_j \\ \vdots \\ \sum_{j=1}^{n} a_{mj} x_j \end{bmatrix}$$

$$\frac{\partial y}{\partial x} = \begin{bmatrix} \frac{\partial}{\partial x_1} & \cdots & \frac{\partial}{\partial x_n} \end{bmatrix} \oplus y = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & & & \vdots \\ a_{m1} & & & a_{mn} \end{bmatrix} = A$$

② $y = Ax$

$y = mx1$, $A = mxn$, $x = nx1$

$x = f(z)$

$$y_i = \sum_{j=1}^{n} a_{ij} x_j \qquad \frac{\partial y_i}{\partial z} = \sum_{j=1}^{n} a_{ij} \frac{\partial x_j}{\partial z}$$ → (1)

$$\frac{\partial y}{\partial z} = A \begin{bmatrix} \frac{\partial x_1}{\partial z} \\ \vdots \\ \frac{\partial x_n}{\partial z} \end{bmatrix} = \begin{bmatrix} a_{11} \frac{\partial x_1}{\partial z} + a_{12} \frac{\partial x_2}{\partial z} \cdots \\ \vdots \end{bmatrix}$$ → (2)

$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$, $\frac{\partial \vec{x}}{\partial z} = \begin{bmatrix} \frac{\partial x_1}{\partial z} \\ \vdots \\ \frac{\partial x_n}{\partial z} \end{bmatrix}$

From ① and ②, we see that

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

③ $\alpha = y^T A x$

$y = mx1$, $x = n \times 1$, $A = mxn$, independent of x and y

Let $w^T = y^T A$

$$\Rightarrow \alpha = w^T x \qquad \frac{\partial \alpha}{\partial x} = w^T = y^T A$$

$\alpha = \alpha^T = x^T A^T y$

Let $w_2^T = x^T A^T$

$$\Rightarrow \alpha = w_2^T y \qquad \frac{\partial \alpha}{\partial y} = w_2^T = x^T A^T$$

④ $\alpha = x^T A x$

$x = nx1$, $A = mxn$, independent of x

$$[x_1 \cdots x_n]_{1\times n} \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}_{m\times n} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}_{n\times 1}$$

$$\begin{bmatrix} \sum_{i=1}^{n} x_i a_{i1} & \cdots \end{bmatrix}_{1\times n} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}_{n\times 1}$$

$$D(U^T V) = U^T (D_v) + V^T (D_u)$$

$$D(x^T A x) = x^T D(Ax) + (Ax)^T (D(x))$$

$$= x^T A + x^T A^T (I)$$

$$\alpha = x^T (A + A^T)$$

⑤ $\alpha = y^T x$

$y = nx1$, $x = nx1$ } functions of z

$$D(U^T V) = U^T (D_v) + V^T (D_u)$$

$$\Rightarrow D_z(y^T x) = y^T (D_z(x)) + x^T (D_z(y))$$

$$\frac{\partial \alpha}{\partial z} = y^T \frac{\partial x}{\partial z} + x^T \frac{\partial y}{\partial z}$$

⑥ $\alpha = y^T A x$ → scalar

$y = mx1$, $A = mxn$, $x = nx1$ → functions of z

$$D(U^T V) = U^T (D_v) + V^T (D_u)$$

$\alpha = y^T A x$

$$D_z(\alpha) = y^T (D_z(Ax)) + (Ax)^T (D_z(y))$$

$$\frac{\partial \alpha}{\partial z} = y^T A \frac{\partial x}{\partial z} + x^T A^T \frac{\partial y}{\partial z}$$

⑦ A: non-singular matrix whose elements = $f(\alpha)$

$$\frac{\partial A^{-1}}{\partial \alpha} = -A^{-1} \frac{\partial A}{\partial \alpha} A^{-1}$$

Proof:

$AA^{-1} = I$ ∵ invertible

$A^{-1} \frac{\partial A}{\partial \alpha} + \frac{\partial A^{-1}}{\partial \alpha} \cdot A = 0$

$\frac{\partial A^{-1}}{\partial \alpha} A = -A^{-1} \frac{\partial A}{\partial \alpha}$

Multiply A^{-1} on both sides

$$\frac{\partial A^{-1}}{\partial \alpha} = -A^{-1} \frac{\partial A}{\partial \alpha} A^{-1}$$

Trace → sum of diagonal elements ∀ y::

(i) $\text{tr}(A+B) = \text{tr}(A) + \text{tr}(B)$

(ii) $\text{tr}(A^T) = \text{tr}(A)$

(iii) $\text{tr}(cA) = c\,\text{tr}(A)$

$$\rightarrow \frac{\partial \text{tr}(Y)}{\partial x} = \frac{\partial \text{tr}(Y')}{\partial x}$$

Let $\vec{x}: nx1$, $X: nxn$; $\vec{y}: nx1$, $Y: nxn$

① Transposition of dependent variable

$\text{tr}(AB) = \text{tr}(BA)$

② Cyclic permutation

$$\frac{\partial \text{tr}(AB)}{\partial x} = \frac{\partial \text{tr}(BA)}{\partial x}$$

③ Trace of scalar

$\alpha = [\alpha]$

$\Rightarrow \text{tr}(\alpha) = \alpha$

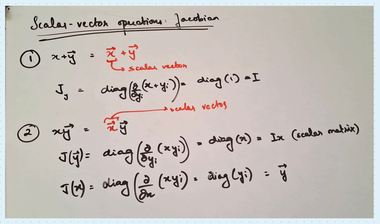
④ Frobenius Inner Product

$\langle x, y \rangle = \text{tr}(x^T y) = \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} y_{ij}$

⑤ Frobenius norm

Add squares of matrix entries and take Main root

$\Rightarrow \|X\|_F = \sqrt{\text{tr}(XX^*)} = \sqrt{\text{tr}(X^2 X)}$



Vector Chain Rule

Let $\vec{x}, \vec{f}, \vec{g}$ such that and $|x|=n$, $|f|=m$, $|g|=k$

$y = f(g(x))$ → product of 2 Jacobians

$$\frac{\partial}{\partial x} f(g(x)) = \begin{bmatrix} \frac{\partial f_1}{\partial g_1} & \cdots & \frac{\partial f_1}{\partial g_k} \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial g_1} & \cdots & \frac{\partial f_m}{\partial g_k} \end{bmatrix}_{m\times k} \begin{bmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial g_k}{\partial x_1} & \cdots & \frac{\partial g_k}{\partial x_n} \end{bmatrix}_{k\times n}$$ → $m \times k$, $k \times n$ ⟹ $m \times n$ Result: $m \times n$

$x \in \mathbb{R}^n$

$g(x): \mathbb{R}^n \to \mathbb{R}^k$

$f(g(x)): \mathbb{R}^k \to \mathbb{R}^m$

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

Data flow = $\vec{x} \xrightarrow{} g(x) \xrightarrow{f} \vec{f}(g(x))$

$\frac{\partial g}{\partial x}$ ← $\frac{\partial f}{\partial g}$