CS 189 MACHINE LEARNING
vector calculus review

§1 fundamentals

$x \in \mathbb{R}^n \xrightarrow{f} y = f(x) \in \mathbb{R}^m$

commonly, we write $f: \underbrace{\mathbb{R}^n}_{\text{domain/ input space}} \longrightarrow \underbrace{\mathbb{R}^m}_{\text{codomain/ output or target space}}$

<u>def.</u> for some function $y = f(x)$, $f: \mathbb{R}^n \to \mathbb{R}^m$ let

$$\boxed{\frac{\partial y}{\partial x} := \left[\frac{\partial y_i}{\partial x_j}\right]_{ij}}$$

matrix who's $(i\,j)$ entry is $\partial y_i / \partial x_j$: the partial derivative of y_i w.r.t. x_j

sometimes, we also write $\frac{\partial f}{\partial x}$

$$= \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix} \in \underbrace{\mathbb{R}^{m \times n}}$$

notice if we think of this matrix as a linear function it has the same "shape" as f: takes input in $\mathbb{R}^n$ and gives output in $\mathbb{R}^m$

<u>def.</u> for $y = f(x)$, we say the <u>jacobian</u> of f at some point x_0 is the derivative above evaluated at $x = x_0$

$$J_f(x_0) := \left.\frac{\partial y}{\partial x}\right|_{x = x_0}$$

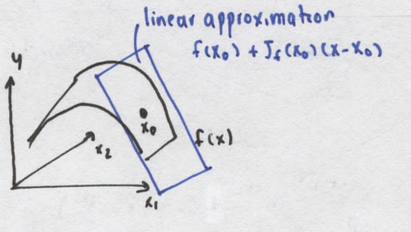
the derivatives change at different points, this notation just allows us specify which point we are looking at

<u>prop.</u> let $y = f(x)$, $f: \mathbb{R}^n \to \mathbb{R}^m$. for some point $x_0 \in \mathbb{R}^n$

$$\boxed{f(x_0 + \overset{\in \mathbb{R}^n}{\Delta x}) \approx f(x_0) + J_f(x_0)\,\Delta x}$$

or

$$f(x) \approx f(x_0) + J_f(x_0)(x - x_0)$$



intuition.

$$f(x_0 + \Delta x) - f(x_0) \approx J_f(x_0)\,\Delta x$$

how much does y_i change when I move x a bit?

$$\begin{bmatrix} \\ \bigcirc \\ \\ \end{bmatrix} = \begin{bmatrix} \frac{\partial y_i}{\partial x_1} & \frac{\partial y_i}{\partial x_2} & \cdots & \frac{\partial y_i}{\partial x_n} \end{bmatrix} \begin{bmatrix} \Delta x_1 \\ \Delta x_2 \\ \vdots \\ \Delta x_n \end{bmatrix}$$

add up all changes

$$\Delta y_i = \underbrace{\frac{\partial y_i}{\partial x_1}\Delta x_1}_{\text{how much } y_i \text{ changes based off } x_1} + \underbrace{\frac{\partial y_i}{\partial x_2}\Delta x_2}_{\text{how much } y_i \text{ changes based off } x_2} + \dots + \frac{\partial y_i}{\partial x_n}\Delta x_n \quad \dots$$

example consider $f: \mathbb{R}^n \longrightarrow \mathbb{R}$ (f is a scalor function) vs a vector map

$$y = f(x) = \sum_{i=1}^{n} x_i^2 = \|x\|_2^2$$

$$J_f(x) = \left[\frac{\partial y}{\partial x_1} \; \frac{\partial y}{\partial x_2} \; \cdots \; \frac{\partial y}{\partial x_n} \right]$$

important: this is a row vector $\mathbb{R}^{1 \times n}$

$$= \left[2x_1 \; 2x_2 \; \cdots \; 2x_n \right]$$

$$= 2x^T$$

since $\frac{\partial y}{\partial x_k} = \frac{\partial}{\partial k} \sum_{i=1}^{n} x_i^2$

$= \frac{\partial}{\partial k} x_k^2$ only 1 term actually depends on x_k

$= 2x_k$

example consider $f: \mathbb{R}^n \longrightarrow \mathbb{R}^n$

$y = f(x) = -\log(x)$ common abuse of notation: write regular functions, to mean componentwise application

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \xmapsto{f} \begin{bmatrix} -\log x_1 \\ -\log x_2 \\ \vdots \\ -\log x_n \end{bmatrix}$$

$$\frac{\partial y}{\partial x} = \begin{bmatrix} -1/x_1 & 0 & \cdots & 0 \\ 0 & -1/x_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & -1/x_n \end{bmatrix}$$

notice only diagonal entries are non-zero!

$$\frac{\partial}{\partial x_j} \underbrace{-\log x_i}_{y_i} = \begin{cases} -1/x_i & : i = j \\ 0 & : i \neq j \end{cases}$$

(if $i \neq j$, it's like we are differentiating a constant)

def. for a scalor function $f: \mathbb{R}^n \longrightarrow \mathbb{R}$, we write the gradient as

$$\boxed{\nabla_x f(x) = \left(\frac{\partial f}{\partial x} \right)^T} \quad \in \mathbb{R}^n$$

always has same dimension/shape as input

recall this was a row vector — transposing makes it a column

$$\underset{\mathbb{R}^n}{x^{(t+1)}} = \underset{\mathbb{R}^n}{x^{(t)}} - \alpha \underset{\mathbb{R}^n}{\nabla_x f(x^{(t)})}$$

(gradient update rule)

→ gradients are convenient because they share the shape of the input

example $f(x) = \|x\|_2^2$

$$\nabla_x f(x) = \left(\frac{\partial f}{\partial x} \right)^T = (2x^T)^T = 2x \; ✓$$

see above

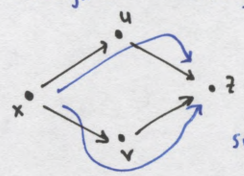
§2 derivative rules

prop (chain rule, multivariate).

let $u = f(x)$
$v = g(x)$
$z = h(u, v)$
$= h(f(x)\ g(x))$

$$\boxed{\frac{\partial z}{\partial x} = \underbrace{\frac{\partial z}{\partial u}\frac{\partial u}{\partial x}}_{\text{how } x \text{ affects } z \text{ through } u} + \underbrace{\frac{\partial z}{\partial v}\frac{\partial v}{\partial x}}_{\text{how } x \text{ affect } z \text{ through } v}}$$

"Math 53 version"



sum effect of x through both paths leading to z

prop (chain rule, multivariate) *

$$x \in \mathbb{R}^n \xrightarrow{f} y \in \mathbb{R}^m \xrightarrow{g} z \in \mathbb{R}^k$$

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

really easy to remember! exactly the same as the regular chain rule...

remember, these are now jacobians

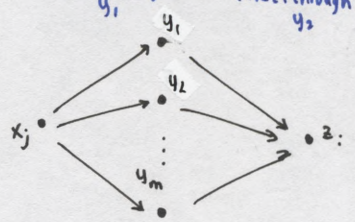
order matters now, this is a matrix multiplication

$$\underset{\frac{\partial z}{\partial x}}{\overset{\mathbb{R}^{k\times n}}{\begin{bmatrix} \frac{\partial z_i}{\partial x_j} \end{bmatrix}}} = \underset{\frac{\partial z}{\partial y}}{\overset{\mathbb{R}^{k\times m}}{\begin{bmatrix} \frac{\partial z_i}{\partial y_1} & \frac{\partial z_i}{\partial y_2} & \cdots & \frac{\partial z_i}{\partial y_m} \end{bmatrix}}} \underset{\frac{\partial y}{\partial x}}{\overset{\mathbb{R}^{m\times n}}{\begin{bmatrix} \frac{\partial y_1}{\partial x_j} \\ \frac{\partial y_2}{\partial x_j} \\ \vdots \\ \frac{\partial y_m}{\partial x_j} \end{bmatrix}}}$$

so this tells us:

$$\frac{\partial z_i}{\partial x_j} = \underbrace{\frac{\partial z_i}{\partial y_1}\frac{\partial y_1}{\partial x_j}}_{\text{effect through } y_1} + \underbrace{\frac{\partial z_i}{\partial y_2}\frac{\partial y_2}{\partial x_j}}_{\text{effect through } y_2} + \ldots + \frac{\partial z_i}{\partial y_m}\frac{\partial y_m}{\partial x_j}$$

...



exact same idea: sum effect through all paths to z_i

↓
transposing both sides, we get:

$$\underbrace{\left(\frac{\partial z}{\partial x}\right)^T}_{} = \left(\frac{\partial y}{\partial x}\right)^T \left(\frac{\partial z}{\partial x}\right)^T$$

remember $(AB)^T = B^T A^T$

useful to write out some rules for gradients!

Corollary. $x \in \mathbb{R}^n \xmapsto{f} y \in \mathbb{R} \xmapsto{g} z \in \mathbb{R}$

$$f: \mathbb{R}^n \to \mathbb{R} \qquad g \circ f : \mathbb{R}^n \longrightarrow \mathbb{R}$$
$$g: \mathbb{R} \to \mathbb{R}$$

$$\nabla_x z = \left(\frac{\partial z}{\partial x}\right)^T$$
$$= \underbrace{\left(\frac{\partial y}{\partial x}\right)^T}_{\text{gradient!}} \left(\frac{\partial z}{\partial y}\right) \in \mathbb{R}$$

both scalars!

$$= \frac{dz}{dy} \nabla_x y \qquad \therefore \boxed{\nabla_x z = \frac{dz}{dy} \nabla_x y}$$

example. compute the gradient of $\|x\|_2^4$ (really messy to do component-wise...)

$$f(x) = \|x\|_2^2$$
$$g(y) = y^2$$
$$\|x\|_2^4 = g(f(x))$$
$$\therefore \nabla_x \|x\|_2^4 = \overbrace{\frac{dy^2}{dy}}^{\frac{dz}{dy}} \cdot \overbrace{\nabla_x \|x\|_2^2}^{\nabla_x y}$$
$$= 2y \cdot 2x$$
$$= 4\|x\|_2^2\, x \quad ✓$$

here $y = f(x) = \|x\|_2^2$

Corollary $x \in \mathbb{R}^n \xmapsto{f} y \in \mathbb{R}^m \xmapsto{g} z \in \mathbb{R}$

$$f: \mathbb{R}^n \to \mathbb{R}^m \qquad g \circ f : \mathbb{R}^n \longrightarrow \mathbb{R}$$
$$g: \mathbb{R}^m \longrightarrow \mathbb{R}$$

$$\nabla_x z = \left(\frac{\partial z}{\partial x}\right)^T$$
$$= \left(\frac{\partial y}{\partial x}\right)^T \left(\frac{\partial z}{\partial y}\right)^T$$
$$= \left(\frac{\partial y}{\partial x}\right)^T \nabla_y z$$

example recall the ℓ_2^2 loss for linear regression, given weights w is defined as:

$$L(w) = \|Xw - y\|_2^2$$

now let $f(w) = Xw - y$
$g(v) = \|v\|_2^2$

we can show $\frac{\partial}{\partial w}(Xw - y) = X$
but this should make sense, since a linear approximation of a linear function should be linear

$$\therefore \nabla_w L(w) = \left(\frac{\partial f}{\partial w}\right)^T \nabla_v g(v)$$
$$= \boxed{X^T} \cdot 2v$$
$$= 2X^T(Xw - b)$$

here $v = f(w) = Xw - b$

term pops out when you compose w/ a linear function

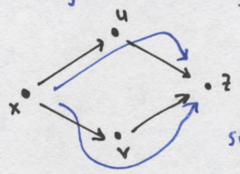
§2 derivative rules

<u>prop</u> (chain rule, multivariate).

let $u = f(x)$
$v = g(x)$
$z = h(u, v)$
$= h(f(x)\ \ g(x))$

$$\boxed{\frac{\partial z}{\partial x} = \underbrace{\frac{\partial z}{\partial u}\frac{\partial u}{\partial x}}_{\text{how } x \text{ affects } z \text{ through } u} + \underbrace{\frac{\partial z}{\partial v}\frac{\partial v}{\partial x}}_{\text{how } x \text{ affect } z \text{ through } v}}$$

"Math 53 version"



sum effect of x through both paths leading to z

<u>prop</u> (chain rule, multivariate) *

$$x \in \mathbb{R}^n \xmapsto{f} y \in \mathbb{R}^m \xmapsto{g} z \in \mathbb{R}^k$$

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

really easy to remember! exactly the same as the regular chain rule...

remember, these are now jacobians

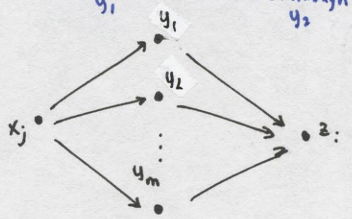
order matters now, this is a matrix multiplication

$$\underset{\frac{\partial z}{\partial x}}{\overset{\mathbb{R}^{k \times n}}{\begin{bmatrix} \frac{\partial z_i}{\partial x_j} \\ \bullet \end{bmatrix}}} = \underset{\frac{\partial z}{\partial y}}{\overset{\mathbb{R}^{k \times m}}{\begin{bmatrix} \frac{\partial z_i}{\partial y_1} & \frac{\partial z_i}{\partial y_2} & \cdots & \frac{\partial z_i}{\partial y_m} \end{bmatrix}}} \underset{\frac{\partial y}{\partial x}}{\overset{\mathbb{R}^{m \times n}}{\begin{bmatrix} \frac{\partial y_1}{\partial x_j} \\ \frac{\partial y_2}{\partial x_j} \\ \vdots \\ \frac{\partial y_m}{\partial x_j} \end{bmatrix}}}$$

so this tells us:

$$\frac{\partial z_i}{\partial x_j} = \underbrace{\frac{\partial z_i}{\partial y_1}\frac{\partial y_1}{\partial x_j}}_{\text{effect through } y_1} + \underbrace{\frac{\partial z_i}{\partial y_2}\frac{\partial y_2}{\partial x_j}}_{\text{effect through } y_2} + \dots + \frac{\partial z_i}{\partial y_m}\frac{\partial y_m}{\partial x_j}$$

...



exact same idea: sum effect through all paths to z_i

↓

transposing both sides, we get:

$$\underbrace{\left(\frac{\partial z}{\partial x}\right)^T = \left(\frac{\partial y}{\partial x}\right)^T \left(\frac{\partial z}{\partial x}\right)^T}$$

remember $(AB)^T = B^T A^T$

useful to write out some rules for gradients!

prop (addition rule).

$$\boxed{\begin{aligned} \frac{\partial}{\partial x}\sum_{i=1}^{m} f_i(x) &= \sum_{i=1}^{m}\frac{\partial}{\partial x} f_i(x) \\ \nabla_x \sum_{i=1}^{m} f_i(x) &= \sum_{i=1}^{m} \nabla_x f_i(x) \end{aligned}}$$

taking transposes

"the derivative is linear"

abuse of notation:
here $\{f_1, f_2, \dots f_m\}$ are different functions,
not the coordinate functions of some map f

proof idea.

let $f_i : \mathbb{R}^n \to \mathbb{R}$

$$\nabla_x \sum_{i=1}^{m} f_i(x) = \begin{array}{c} \\ j \\ \\ \end{array}\begin{bmatrix} \\ \circ \\ \\ \end{bmatrix} \quad \frac{\partial}{\partial x_j}\sum_{i=1}^{m} f_i(x) = \sum_{i=1}^{m}\frac{\partial}{\partial x_j} f_i(x)$$

$$= \begin{array}{c} \\ j \\ \\ \end{array}\underbrace{\begin{bmatrix} \\ \circ \\ \\ \end{bmatrix}}_{\nabla_x f_1(x)} + \dots + \underbrace{\begin{bmatrix} \\ \circ \\ \\ \end{bmatrix}}_{\nabla_x f_m(x)}$$

$\frac{\partial}{\partial x_j} f_m(x)$

$\frac{\partial}{\partial x_j} f_1(x)$

$$\boxed{\sum_{i=1}^{m}\frac{\partial}{\partial x_j} f_i(x)}$$

$$= \sum_{i=1}^{m} \nabla_x f_i(x)$$

prop (scalar multiplication rule)

$$\boxed{\begin{aligned} \frac{\partial}{\partial x} c f(x) &= c\frac{\partial}{\partial x} f(x) \\ \nabla_x c f(x) &= c\,\nabla_x f(x) \end{aligned}}$$

$$\left[\frac{\partial}{\partial x} c f(x)\right]_{ij} = \frac{\partial}{\partial x_j} c f_i(x)$$

refers to i^{th} coordinate of the output $f(x)$

$$= c \cdot \frac{\partial}{\partial x_j} f_i(x)$$

$$= c\left[\frac{\partial}{\partial x} f(x)\right]_{ij}$$

prop (gradient dot product rule).

$$\begin{tikzcd} & u \in \mathbb{R}^m \arrow[dr] & \\ x \in \mathbb{R}^n \arrow[ur, "f"] \arrow[dr, "g"'] & & u^T v \in \mathbb{R} \\ & v \in \mathbb{R}^m \arrow[ur] & \end{tikzcd}$$

$u \in \mathbb{R}^m$
$v \in \mathbb{R}^m$

$$\boxed{\begin{aligned} \nabla_x u^T v &= \nabla_x f(x)^T g(x) \\ &= \underbrace{\left(\frac{\partial u}{\partial x}\right)^T}_{\mathbb{R}^{n\times m}} \underbrace{v}_{\mathbb{R}^{m\times 1}} + \left(\frac{\partial v}{\partial x}\right)^T u \end{aligned}}$$

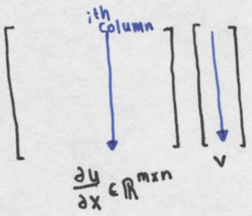
looks similar to regular product rule!

$$\frac{\partial}{\partial x_i} u^T v = \frac{\partial}{\partial x_i} \cdot \sum_{j=1}^{m} u_j v_j$$

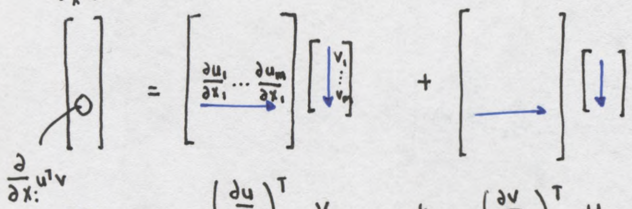
apply scalar product rule & split into two sumations

ith component of gradient

$$= \underbrace{\sum_{j=1}^{m} \frac{\partial u_j}{\partial x_i} v_j}_{\text{looks a bit like matrix multiplication:}} + \sum_{j=1}^{m} u_j \frac{\partial v_j}{\partial x_i}$$



$\nabla_x u^T v$



$$\left(\frac{\partial u}{\partial x}\right)^T v + \left(\frac{\partial v}{\partial x}\right)^T u$$

$\mathbb{R}^n$ $\qquad \mathbb{R}^{n \times m} \cdot \mathbb{R}^{m \times 1} \qquad \mathbb{R}^{n \times m}\ \mathbb{R}^{m \times 1}$

$$\therefore \nabla_x u^T v = \left(\frac{\partial u}{\partial x}\right)^T v + \left(\frac{\partial v}{\partial x}\right)^T u$$

<u>example</u> let $x \in \mathbb{R}^n$
$A \in \mathbb{R}^{n \times n}$ be some matrix, and define

$$f(x) = \underbrace{x^T}_{u} \underbrace{A x}_{v}$$

set $u = x$
$v = Ax$

$$\therefore \nabla_x f(x) = \nabla_x u^T v$$

$$= \left(\frac{\partial u}{\partial x}\right)^T v + \left(\frac{\partial v}{\partial x}\right)^T u$$

$\frac{\partial}{\partial x} Ax = A$, appealing once again to the notion of a derivative as a linear approximation

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

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

$\left[\frac{\partial x}{\partial x}\right]_{ij} = \frac{\partial x_i}{\partial x_j}$, clearly 1 iff $i=j$, 0 otherwise

§3 common derivatives

(linear functions)

$$\nabla_x c^T x = c \qquad \frac{\partial}{\partial x} c^T x = c^T$$

$$\frac{\partial}{\partial x}(Ax+b) = A$$

constant

$$\nabla_x c = 0 \qquad \frac{\partial}{\partial x} c = 0$$

(quadratic functions)

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

$[= 2Ax]$ (when A is symmetric)

$$\nabla_x x^T A x + b^T x + c = (A+A^T)x + b$$

$[= 2Ax+b]$ (when A is symmetric)

$$\nabla_x \|Ax+b\|_2^2 = \nabla_x (Ax+b)^T(Ax+b)$$

$$= \nabla_x \left[x^T \underbrace{A^T A}_{\text{symmetric}} x + 2b^T A x + b^T b \right]$$

$$= 2A^T A x + A^T b$$

$$\nabla_x \|Ax-b\|_2^2 = 2A^T A x - A^T b$$

$$\nabla_x \|x\|_2^2 = 2x$$

$$\nabla_x (Ax+b)^T C (Dx+e) = D^T C^T (Ax+b) + A^T C (Dx+e)$$

$$\nabla_x (Ax-b)^T W (Ax-b) = A^T W^T (Ax-b) + A^T W (Ax-b)$$

$[= 2A^T W(Ax-b)]$ when W is symmetric

$$\nabla_x a^T x x^T b = (ab^T + ba^T)x$$

(other functions)

$$\nabla_x \|x-a\|_2 = \frac{x-a}{\|x-a\|_2}$$