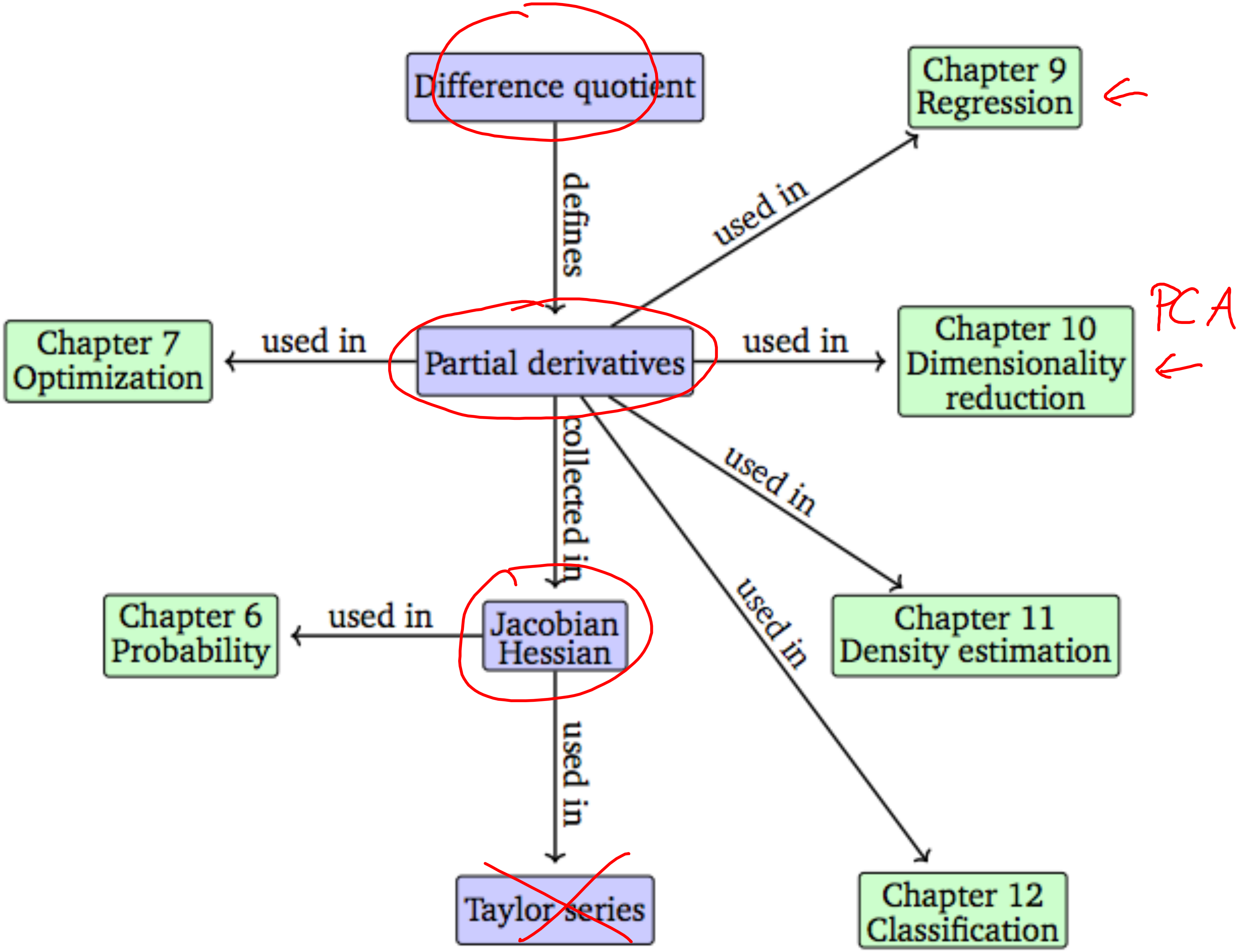
Vector Calculus

Liang Zheng
Australian National University
liang.zheng@anu.edu.au


Difference quotient
defines
Partial derivatives
used in
Chapter 9
Regression
used in
Chapter 10
Dimensionality
reduction
PCA
used in
Chapter 7
Optimization
collected in
Jacobian
Hessian
used in
Chapter 6
Probability
used in
Taylor series
used in
Chapter 11
Density estimation
used in
Chapter 12
Classification

5 Vector Calculus 向量微积分

- We discuss functions

domain codomain norm $\|x\|_2$

$$f : \mathbb{R}^D \to \mathbb{R}$$
$$\boldsymbol{x} \mapsto f(\boldsymbol{x})$$

where $\mathbb{R}^D$ is the domain of f, and the function values $f(\boldsymbol{x})$ are the image/codomain of f.

$f(\boldsymbol{x}) = \boldsymbol{c} \cdot \boldsymbol{x}$ $\mathbb{R}^n \to \mathbb{R}$

- Example (dot product)
- Previously, we write dot product as

$$f(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}, \qquad \boldsymbol{x} \in \mathbb{R}^2$$

$$f(x_1, x_2) = x_1^2 + x_2^2$$

- In this chapter, we write it as

$$f : \mathbb{R}^2 \to \mathbb{R}$$
$$\boldsymbol{x} \mapsto x_1^2 + x_2^2$$

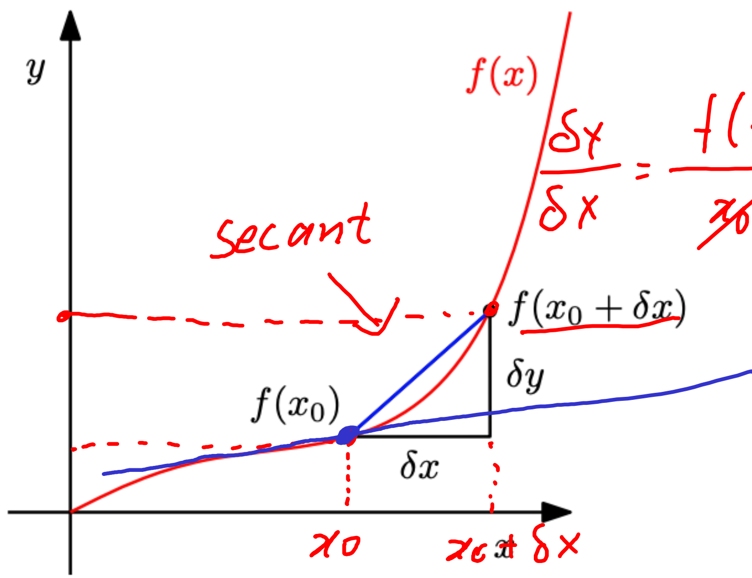
5.1 Differentiation of Univariate Functions

微分

- Given $y = f(x)$, the difference quotient is defined as

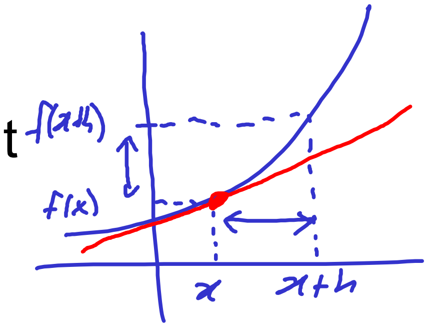
$$\frac{\delta y}{\delta x} := \frac{f(x + \delta x) - f(x)}{\delta x}$$

- It computes the slope of the secant line through two points on the graph of f. In this figure, these are the points with x–coordinates x_0 and $x_0 + \delta x_0$.
- In the limit for $\boxed{\delta x \to 0}$, we obtain the tangent of f at x (if f is differentiable). The tangent is then the derivative of f at x.



$$\frac{\delta y}{\delta x} = \frac{f(x_0 + \delta x) - f(x_0)}{\cancel{x_0} + \delta x - \cancel{x_0}}$$

5.1 Differentiation of Univariate Functions



- For $h > 0$, the derivative of f at x is defined as the limit

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

- The derivative of f points in the direction of steepest ascent of f.

- ~~Example - Derivative of a Polynomial~~

- ~~Compute the derivative of $f(x) = x^n, n \in \mathbb{N}$. (From our high school knowledge, the derivative is nx^{n-1}.)~~

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

$$= \lim_{h \to 0} \frac{(x+h)^n - x^n}{h}$$

$$= \lim_{h \to 0} \frac{\sum_{i=0}^{n} \binom{n}{i} x^{n-i} h^i - x^n}{h}$$

~~we see that $x^n = \binom{n}{0} x^{n-0} h^0$. By starting the sum at 1, the x^n cancels.~~

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

Example. Prove $\boxed{\frac{d}{dx}(x^2) = 2x}$

$$\frac{d}{dx}(x^2) = \lim_{h \to 0} \frac{(x+h)^2 - x^2}{h}$$

$$= \lim_{h \to 0} \frac{\cancel{x^2} + 2xh + h^2 \cancel{- x^2}}{h}$$

$$= \lim_{h \to 0} \frac{2x\cancel{h}}{\cancel{h}} + \frac{h^{\cancel{2}}}{\cancel{h}}\ h$$

$$= \lim_{h \to 0} 2x + h$$

$$= 2x + \cancel{\lim_{h \to 0}} h$$

$$= 2x$$

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

Power Law $\frac{d}{dx}(x^6) = 6x^5$

Example. Prove $\boxed{\frac{d}{dx}(x^n) = nx^{n-1}}$

Binomial Theorem

$$(x+h)^n = \sum_{i=0}^{n} \binom{n}{i} x^{n-i} h^i$$

$$\frac{d}{dx}(x^n) = \lim_{h\to 0} \frac{(x+h)^n - x^n}{h}$$

$$= \lim_{h\to 0} \frac{\sum_{i=0}^{n} \binom{n}{i} x^{n-i} h^i - x^n}{h}$$

$\cancel{\binom{n}{0}} x^n$

$$= \lim_{h\to 0} \frac{\cancel{x^n} + \left(\sum_{i=1}^{n} \binom{n}{i} x^{n-i} h^i\right) \cancel{- x^n}}{h}$$

$$= \lim_{h\to 0} \binom{n}{1} x^{n-1} + \frac{\sum_{i=2}^{n} \binom{n}{i} x^{n-i} h^i}{\cancel{h}}$$

$$\frac{h^i}{h} = h^{i-1}$$

$$= n x^{n-1} + \cancel{\lim_{h\to 0} \sum_{i=2}^{n} \binom{n}{i} x^{n-i} h^{i-1}}$$

$$= n \cdot x^{n-1}$$

Optional Exercise. Prove $\frac{d}{dx} e^x = e^x$

You will need the Taylor Expansion of exp(x) to do this: $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$

and that $e^{x+y} = e^x \cdot e^y$

Why care about derivatives?

$(y_1, x_1), \ldots, (y_n, x_n)$

Model $\hat{y} = f(x \mid \theta)$ ← model, ← param., ↑ guess, ↑ data

Loss

$$\mathcal{L}(X;\theta) = \sum_i \|y_i - \hat{y}_i\|_2^2$$

$\underbrace{\frac{\partial \mathcal{L}}{\partial \theta} = 0}_{\text{easy}}$, $\underbrace{\text{solve for } \theta}_{\text{hard}}$

$f(x) = x^5 - x^2 + x$

$f'(x) = 5x^4 - 2x + 1 = 0$?

Numerically. 1. Guess θ_0

Learning Rate (0.1 to 0.01)

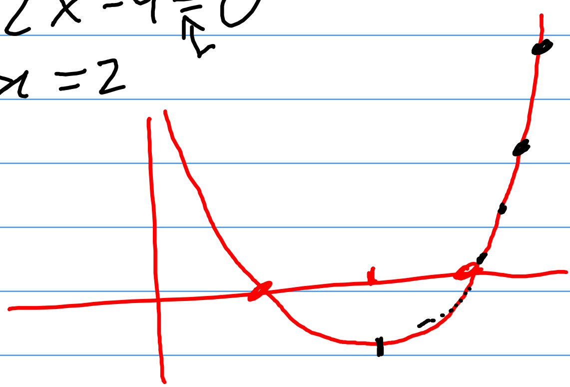
Eta (Hyperparameter)

2. $\theta_{t+1} = \theta_t - \eta \cdot \frac{\partial \mathcal{L}}{\partial \theta}$

Minimise $f(x) = x^2 - 4x + 3$

$f'(x) = 2x - 4 = 0$

$\Rightarrow x = 2$



$\eta = 0.1$

guess $x_0 = 4$

$x_1 = x_0 - \eta \cdot f'(x_0)$

$= 4 - 0.1 \cdot (2 \cdot 4 - 4)$

$= 3.6$

$x_2 = x_1 - \eta \cdot f'(x_1)$

$= 3.6 - 0.1(2 \times 3.6 - 4)$

$= 3.28$ $\qquad x_3 = 3.024, \ldots$

Estimating Intractable Derivatives

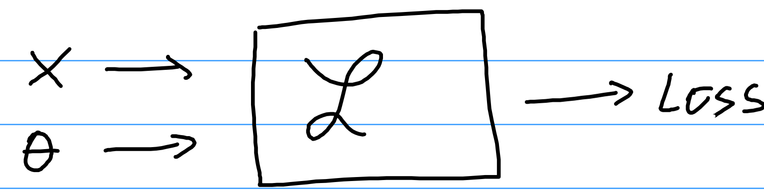
$\frac{\partial \mathcal{L}}{\partial \theta} = ?$ $\quad \mathcal{L} = \text{loss}$

$$\boxed{f'(x) \approx \frac{f(x+\varepsilon) - f(x)}{\varepsilon}} \quad \text{for small } \varepsilon > 0$$

NOT OFTEN

eg estimate derivative of $f(x) = x^2$ at $x = 2$.

True Answer: $f'(x) = 2x \qquad f'(2) = 4$



Estimate: $\frac{f(2+\varepsilon) - f(2)}{\varepsilon} = \frac{(2+\varepsilon)^2 - 2^2}{\varepsilon}$

$\varepsilon = 0.1 \qquad \frac{2.01^2 - 2^2}{0.1} = 4.1$

$\varepsilon = 0.01 \qquad \frac{2.01^2 - 2^2}{0.01} = 4.01$

5.1.2 Differentiation Rules

$$\frac{d}{dx}\left(f(x)g(x)\right) = \frac{df}{dx} \cdot g + f \cdot \frac{dy}{dx} \qquad f'(x) = \frac{d(f(x))}{dx}$$

- Product rule

$$(f(x)g(x))' = f'(x)g(x) + f(x)g'(x)$$

- Quotient rule:

$$\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)g(x) - f(x)g'(x)}{(g(x))^2} \; ?$$

$$(x^2+5x)' = 2x + 5$$

- Sum rule:

Differentiation is linear

$$(f(x) + g(x))' = f'(x) + g'(x)$$

- Chain rule:

$$(g(f(x)))' = (g \circ f)'(x) = g'(f(x))f'(x)$$

Here, $g \circ f$ denotes function composition $g\big(f(x)\big)$

$$\left(\frac{f(x)}{g(x)}\right)' = \left(f(x) \cdot \frac{1}{g(x)}\right)'$$

Prove Sum Rule $\quad (f+g)' = f' + g' \qquad \frac{df}{dx} := \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$

$$\frac{d}{dx}\left(f(x)+g(x)\right) = \lim_{h\to 0} \frac{\left(f(x+h)+g(x+h)\right) - \left(f(x)+g(x)\right)}{h}$$

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

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

$$= f'(x) + g'(x)$$

Prove Product Rule $(fg)' = fg' + f'g$ $\quad \frac{df}{dx} := \lim_{h\to 0} \frac{f(x+h)-f(x)}{h}$

$$\frac{d}{dx}\left(f(x)g(x)\right) = \lim_{h\to 0} \frac{f(x+h)g(x+h) - f(x)g(x)}{h}$$

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

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

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

$$\left(\lim_{h\to 0} f(x+h)\right)\left(\lim_{h\to 0} \frac{g(x+h)-g(x)}{h}\right) + g(x)\boxed{\lim_{h\to 0} \frac{f(x+h)-f(x)}{h}}$$

$f(x)$ $\quad g'(x)$ $\quad + g(x)\cdot f'(x)$

$$= f(x)\cdot g'(x) + g(x)\cdot f'(x)$$

Exercise: Prove Chain Rule.

Exercise: Prove $\frac{d}{dx}(cf(x)) = c\frac{d}{dx}f(x)$

$$(fg)' = f'g + fg'$$
$$(f+g)' = f' + g'$$
$$(f(g(x)))' = f'(g(x)) \cdot g'(x)$$

Examples of Derivative Rules

Derive $\sin(2x+1)$

$$\sin(2x+1) = f(g(x))$$

$$f(x) = \sin x \quad g(x) = 2x+1$$

$$(f(g(x)))'$$
$$= f'(g(x)) \cdot g'(x)$$

$$f'(x) = \cos x$$
$$f'(g(x)) = \cos(2x+1)$$

$$g'(x) = 2$$

$$= 2\cos(2x+1)$$

Derive $(\sin(x^4))^5$

$$h(f(g(x)))$$

$$h(x) = x^5$$
$$f(x) = \sin x$$
$$g(x) = x^4$$

$$(h(f(g(x))))' = h'(f(g(x))) \cdot (f(g(x)))'$$

$$= h'(f(g(x))) \cdot f'(g(x)) \cdot g'(x)$$

$$h'(x) = 5x^4 \quad f'(x) = \cos x \quad g'(x) = 4x^3$$

$$= 5(\sin(x^4))^4 \cdot \cos(x^4) \cdot 4x^3$$

$$= 20(\sin(x^4))^4 \cdot \cos(x^4)$$

$(fg)' = f'g + fg'$

$(f+g)' = f' + g'$

$f(g(x))' = f'(g(x)) \cdot g'(x)$

Product Rule

$$\frac{d}{dx}\left(x^3 \cdot \sin x\right)$$

$$= \left(\frac{d}{dx} x^3\right) \sin x + x^3 \left(\frac{d}{dx} \sin x\right)$$

$$= 3x^2 \cdot \sin x + x^3 \cos x \,.$$

An alternative view of chain rule $\qquad \dfrac{dy}{dx} = \dfrac{dy}{du} \cdot \dfrac{du}{dx}$

Derive $\quad y = (x^2+5x+6)^{10} = u^{10}$

let $\underline{u = x^2 + 5x + 6}$

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$$

$$= 10u^9 \cdot (2x+5)$$

$$= 10(x^2+5x+6)^9(2x+5)$$

5.2 Partial Differentiation and Gradients

$\in \mathbb{R}^2$

- Instead of considering $x \in \mathbb{R}$, we consider $\underbrace{\boldsymbol{x} \in \mathbb{R}^n}$, e.g., $f(\boldsymbol{x}) = f(x_1, x_2)$

nabla "del"

- The generalization of the derivative to functions of several variables is the gradient. $\frac{df}{dx}$ $\underbrace{\nabla_x f \quad \text{grad } f}$
- We find the gradient of the function f with respect to $\boldsymbol{x}$ by
  - varying one variable at a time and keeping the others constant.
  - The gradient is the collection of these partial derivatives.
- For a function $f: \mathbb{R}^n \to \mathbb{R}, \boldsymbol{x} \mapsto f(\boldsymbol{x}), \boldsymbol{x} \in \mathbb{R}^n$ of n variables $x_1, \ldots, x_n$, we define the partial derivatives as

$f(x_1, \ldots, x_n)$

$$\frac{\partial f}{\partial x_1} := \lim_{h \to 0} \frac{f(\underline{x_1 + h}, x_2, \ldots, x_n) - f(\boldsymbol{x})}{h}$$

$$\vdots$$

$$\frac{\partial f}{\partial x_n} := \lim_{h \to 0} \frac{f(x_1, \ldots, x_{n-1}, \boxed{x_n + h}) - f(\boldsymbol{x})}{h}$$

and collect them in the row vector

$$\nabla_x f = \text{grad} f = \frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{x}} = \left[\frac{\partial f(\boldsymbol{x})}{\partial x_1} \quad \frac{\partial f(\boldsymbol{x})}{\partial x_n} \quad \cdots \quad \frac{\partial f(\boldsymbol{x})}{\partial x_n}\right] \in \mathbb{R}^{1 \times n}$$

5.2 Partial Differentiation and Gradients

$f : \mathbb{R}^{n} \to \mathbb{R}^{1}$ $\quad \mathbb{R}^{1\times n}$ $\quad \frac{\partial f(x)}{\partial x_1} : \mathbb{R}^n \to \mathbb{R}$ $\quad \boxed{\mathbb{R}^{\text{codomain} \times \text{domain}}}$

- $\nabla_x f = \text{grad} f = \frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{x}} = \left[\frac{\partial f(\boldsymbol{x})}{\partial x_1} \quad \frac{\partial f(\boldsymbol{x})}{\partial x_n} \quad \dots \quad \frac{\partial f(\boldsymbol{x})}{\partial x_n}\right] \in \boxed{\mathbb{R}^{1\times n}}$ $\quad (\mathbb{R}^n \to \mathbb{R})^{1\times n}$
- n is the number of variables and 1 is the dimension of the image/range/codomain of f
- The row vector $\nabla_x f \in \mathbb{R}^{1\times n}$ is called the gradient of f or the Jacobian.

- Example - Partial Derivatives Using the Chain Rule
- For $f(x, y) = (x + 2y^3)^2$, we obtain the partial derivatives

$u^2 \to 2u \cdot 1 = 2(x + 2y^3)$

$f : \mathbb{R}^2 \to \mathbb{R}$

$$\frac{\partial f(x,y)}{\partial x} = 2(x + 2y^3)\frac{\partial}{\partial x}(x + 2y^3) = 2(x + 2y^3)$$

$$\frac{\partial f(x,y)}{\partial y} = \underline{2(x + 2y^3)}\underbrace{\frac{\partial}{\partial y}(x + 2y^3)}_{6y^2} = \underline{12(x + 2y^3)y^2}$$

$\underline{X = (x, y)}$

$$\nabla_X f = \left[\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right] = \left[2(x+2y^3), 12(x+2y^3)y^2\right]^T \in \mathbb{R}^{1\times 2}$$

5.2 Partial Differentiation and Gradients

$f: \mathbb{R}^2 \to \mathbb{R}$

- For $f(x_1, x_2) = \underline{x_1^2 x_2} + \underline{x_1 x_2^3} \in \mathbb{R}$, the partial derivatives (i.e., the derivatives of f with respect to x_1 and x_2 are

$$\frac{\partial f(x_1, x_2)}{\partial x_1} = 2x_1x_2 + x_2^3$$

$$\frac{\partial f(x_1, x_2)}{\partial x_2} = x_1^2 + 3x_1x_2^2$$

- and the gradient is then

$$\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{x}} = \left[\frac{\partial f(x_1, x_2)}{\partial x_1} \quad \frac{\partial f(x_1, x_2)}{\partial x_2}\right] = [2x_1x_2 + x_2^3, \; x_1^2 + 3x_1x_2^2] \in \mathbb{R}^{1\times 2}$$

5.2.1 Basic Rules of Partial Differentiation

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

$g(f(x))$

- Product rule:

$$\frac{\partial}{\partial \boldsymbol{x}}\big(f(\boldsymbol{x})g(\boldsymbol{x})\big) = \frac{\partial f}{\partial \boldsymbol{x}} g(\boldsymbol{x}) + f(\boldsymbol{x})\frac{\partial g}{\partial \boldsymbol{x}}$$

- Sum rule:

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

$$\frac{\partial}{\partial \boldsymbol{x}}\big(f(\boldsymbol{x}) + g(\boldsymbol{x})\big) = \frac{\partial f}{\partial \boldsymbol{x}} + \frac{\partial g}{\partial \boldsymbol{x}}$$

- Chain rule:

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

$$\frac{\partial}{\partial \boldsymbol{x}}(g \circ f)(x) = \frac{\partial}{\partial \boldsymbol{x}}\Big(g\big(f(\boldsymbol{x})\big)\Big) = \frac{\partial g}{\partial f}\frac{\partial f}{\partial \boldsymbol{x}}$$

$\frac{\partial g}{\partial f} : \mathbb{R} \to \mathbb{R}$ $(\mathbb{R}?)$

$\frac{\partial f}{\partial \boldsymbol{x}} : \mathbb{R}^n \to \mathbb{R}^{1 \times n}$ $(\mathbb{R}^{1 \times n})$

5.2.2 Chain Rule

$f: \mathbb{R}^2 \to \mathbb{R}$ codomain × domain $\mathbb{R}$ $\quad \frac{\partial f}{\partial \boldsymbol{x}} \in \mathbb{R}^{1\times 2}$ $\quad \frac{\partial \boldsymbol{x}}{\partial t} \in \mathbb{R}^{2\times 1}$

$\boldsymbol{x}(t) = [x_1(t), x_2(t)]$

$\boldsymbol{x}: \mathbb{R} \to \mathbb{R}^2$

- Consider a function $f: \mathbb{R}^2 \to \mathbb{R}$ of two variables x_1 and x_2.
- $x_1(t)$ and $x_2(t)$ are themselves functions of t.
- To compute the gradient of f with respect to t, we apply the chain rule:

$$\boxed{\frac{df}{dt}} = \frac{\partial f}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{x}}{\partial t} = \underbrace{\begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix}}_{\frac{\partial f}{\partial x}} \underbrace{\begin{bmatrix} \frac{\partial x_1(t)}{\partial t} \\ \frac{\partial x_2(t)}{\partial t} \end{bmatrix}}_{\frac{\partial x}{\partial t}} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t}$$

Where d denotes the gradient and ∂ partial derivates.

- Example
- ~~Consider~~ $\cancel{f(x_1, x_2) = x_1^2 + 2x_2}$, ~~where~~ $\cancel{x_1 = \sin t}$ ~~and~~ $\cancel{x_2 = \cos t}$, ~~then~~

$$\cancel{\frac{df}{dt} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t}}$$

$$\cancel{= 2\sin t \frac{\partial \sin t}{\partial t} + 2\frac{\partial \cos t}{\partial t}}$$

$$\cancel{= 2\sin t \cos t - 2\sin t = 2\sin t(\cos t - 1)}$$

- ~~The above is the corresponding derivative of f with respect to t~~

$\boxed{f : \mathbb{R}^2 \to \mathbb{R}}$ $\qquad x_1 : \mathbb{R} \to \mathbb{R} \qquad x_2 : \mathbb{R} \to \mathbb{R}$ $\qquad \dfrac{df}{dt} = \dfrac{\partial f}{\partial \mathbf{x}} \dfrac{\partial \mathbf{x}}{\partial t}$

$f(x_1, x_2) = x_1^2 x_2 + 3x_1 + 6x_2$ $\qquad \mathbf{X}(t) = \begin{bmatrix} \sin t \\ e^{2t} \end{bmatrix}$ $\qquad \boxed{\mathbf{X} : \mathbb{R} \to \mathbb{R}^2}$

$\underline{x_1(t) = \sin t} \qquad \underline{x_2(t) = e^{2t}}$ $\qquad$ dom $= 1$ $\qquad$ codom $= 2$

$$\frac{\partial f}{\partial \mathbf{x}} = \begin{bmatrix} \dfrac{\partial f}{\partial x_1}, & \dfrac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2x_1x_2 + 3, & x_1^2 + 6 \end{bmatrix} \in \mathbb{R}^{1 \times 2}$$

$$\boxed{\frac{d}{dx}\left(e^{f(x)}\right) = f'(x)\, e^{f(x)}}$$

$$\boxed{\frac{\partial \mathbf{x}}{\partial t} = \begin{bmatrix} \dfrac{\partial x_1}{\partial t} \\ \dfrac{\partial x_2}{\partial t} \end{bmatrix} = \begin{bmatrix} \cos t \\ 2e^{2t} \end{bmatrix} \in \mathbb{R}^{2 \times 1}}$$

$\mathbb{R}^{\text{codom} \times \text{dom}} = \mathbb{R}^{2 \times 1} \simeq \mathbb{R}^2$

$$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial \mathbf{x}} \frac{\partial \mathbf{x}}{\partial t} = (2x_1x_2 + 3)\cos t + (x_1^2 + 6)\, 2e^{2t}$$

Alternatively:

$$f(t) = \underbrace{(\sin t)^2 e^{2t}}_{\text{product rule}} + 3\sin t + 6e^{2t} \quad : \mathbb{R} \to \mathbb{R}$$

5.2.2 Chain Rule

- If $f(x_1, x_2)$ is a function of x_1 and x_2, where $f: \mathbb{R}^2 \to \mathbb{R}$, $x_1(s,t)$ and $x_2(s,t)$ are themselves functions of two variables s and t, the chain rule yields the partial derivatives

$$\frac{\partial f}{\partial s} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial s} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial s} = \frac{\partial f}{\partial \boldsymbol{x}} \frac{\partial \boldsymbol{x}}{\partial s}$$

$$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t} = \frac{\partial f}{\partial \boldsymbol{x}} \cdot \frac{\partial \boldsymbol{x}}{\partial t}$$

- The gradient can be obtained by the matrix multiplication

$\boldsymbol{x}: \mathbb{R}^2 \longrightarrow \mathbb{R}^2$

$$\nabla_{(s,t)} f = \frac{df}{d(s,t)} = \frac{\partial f}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{x}}{\partial (s,t)} = \underbrace{\begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix}}_{=\frac{\partial f}{\partial \boldsymbol{x}}} \underbrace{\begin{bmatrix} \frac{\partial x_1}{\partial s} & \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial s} & \frac{\partial x_2}{\partial t} \end{bmatrix}}_{=\frac{\partial \boldsymbol{x}}{\partial (s,t)}}$$

$$\frac{\partial \boldsymbol{x}}{\partial (s,t)} \in \mathbb{R}^{2\times 2}$$

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

In general, if $f(x_1 \ldots x_n)$ and each x_i is a function ~~$x_i(u_1, \ldots, u_k)$~~ $x_i(u_1, u_2, \ldots, u_k)$ then $x_i : \mathbb{R}^k \to \mathbb{R}$

$$\frac{\partial f}{\partial u_i} = \sum_{j=1}^{n} \frac{\partial f}{\partial x_j} \boxed{\frac{\partial x_j}{\partial u_i}}$$

$$= \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial u_i} + \ldots + \frac{\partial f}{\partial x_n}\frac{\partial x_n}{\partial u_i}$$

$x_1(u, v) = u + v$

~~$x_2(u) = 2u$~~

$x_2(u, v) = 2u$

$$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \cancel{\frac{\partial f}{\partial x_2}\boxed{\frac{\partial x_2}{\partial t}}}^{\,0}$$

$f(x_1, x_2)$

$x_1(s, t) = s \cdot t$

$x_2(s, p, q) = sp + q$

$x_1(s, t, p, q) = st$

$x_2(s, t, p, q) = sp + q$

$f:\mathbb{R}^3 \to \mathbb{R}$ & $x,y,z:\mathbb{R}^3 \to \mathbb{R}$

Example $f(x,y,z) = 3xy + yz^2$

$x(p,q,r) = pqr$

$y(p,q,r) = 2p^2 + q + 5r$

$z(p,q,r) = p\cos(q+2r)$

$$f = 3\underbrace{pqr}_{x}(\underbrace{2p^2+q+5r}_{y}) + \dots$$

Compute $\dfrac{\partial f}{\partial p} = \dfrac{\partial f}{\partial (x,y,z)} \cdot \dfrac{\partial (x,y,z)}{\partial p}$

$$= \frac{\partial f}{\partial x}\cdot\frac{\partial x}{\partial p} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial p} + \frac{\partial f}{\partial z}\cdot\frac{\partial z}{\partial p}$$

$$= (3y)(qr) + (3x+z^2)(4p) + (2yz)(\cos(q+2r))$$

Exercise: $\dfrac{\partial f}{\partial q}$ & $\dfrac{\partial f}{\partial r}$

$f(X): \mathbb{R}^{n\times m} \to \mathbb{R}$ $\quad$ $F(x): \mathbb{R} \to \mathbb{R}^{n\times m}$

5.3 Gradients of Vector-Valued Functions

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

- We discussed partial derivatives and gradients of function $f: \mathbb{R}^n \to \mathbb{R}$
- We will generalize the concept of the gradient to vector-valued functions (vector fields) $\boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^m$, where $n \geq 1$ and $m > 1$.
- For a function $\boldsymbol{f}: \mathbb{R}^n \to \boxed{\mathbb{R}}^m$ and a vector $\boldsymbol{x} = [x_1, \ldots, x_n]^{\mathrm{T}} \in \mathbb{R}^n$, the corresponding vector of function values is given as

$$\underbrace{\boldsymbol{f}(\boldsymbol{x})}_{} = \begin{bmatrix} f_1(\boldsymbol{x}) \\ \vdots \\ f_m(\boldsymbol{x}) \end{bmatrix} \in \text{"}\mathbb{R}^m\text{"}$$

$\mathbb{R}^n \to \mathbb{R}$ $\quad$ m

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

- Writing the vector-valued function in this way allows us to view a vector valued function $\boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^m$ as a vector of functions $\underbrace{[f_1, \ldots, f_m]^{\mathrm{T}}}_{}$, $\underbrace{f_i: \mathbb{R}^n \to \mathbb{R}}_{}$ that map onto $\mathbb{R}$.
- The differentiation rules for every f_i are exactly the ones we discussed before.

5.3 Gradients of Vector-Valued Functions

- The partial derivative of a vector-valued function $\boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^m$ with respect to $x_i \in \mathbb{R}$, $i = 1, \ldots n$, is given as the vector

$$\frac{\partial \boldsymbol{f}}{\partial x_i} = \begin{bmatrix} \frac{\partial f_1}{\partial x_i} \\ \vdots \\ \frac{\partial f_m}{\partial x_i} \end{bmatrix} = \begin{bmatrix} \lim_{h \to 0} \frac{f_1(x_1, \ldots, x_{i-1}, x_i + h, x_{i+1}, \ldots, x_n) - f_1(\boldsymbol{x})}{h} \\ \vdots \\ \lim_{h \to 0} \frac{f_m(x_1, \ldots, x_{i-1}, x_i + h, x_{i+1}, \ldots, x_n) - f_m(\boldsymbol{x})}{h} \end{bmatrix} \in \mathbb{R}^m$$

- In above, every partial derivative $\frac{\partial f}{\partial x_i}$ is a column vector
- Recall that the gradient of f with respect to a vector is the row vector of the partial derivatives
- Therefore, we obtain the gradient of $\boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^m$ with respect to $\boldsymbol{x} \in \mathbb{R}^n$, by collecting these partial derivatives:

$$\frac{d\boldsymbol{f}(\boldsymbol{x})}{d\boldsymbol{x}} = \begin{bmatrix} \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_n} \end{bmatrix} = \begin{array}{cc} & j \\ i & \begin{bmatrix} \boxed{\frac{\partial f_1(\boldsymbol{x})}{\partial x_1}} & \cdots & \frac{\partial f_1(\boldsymbol{x})}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_n} \end{bmatrix} \end{array} \in \mathbb{R}^{\underset{\text{codomain}}{m} \times \underset{\text{domain}}{n}}$$

Jacobian

$$J_{ij} = \frac{\partial f_i(x)}{\partial x_j}$$

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

$$\frac{\partial f}{\partial x_i} = \lim_{h \to 0} \frac{\overbrace{f(x_1, x_2, \ldots, x_{i-1}, \underbrace{x_i + h}, x_{i+1}, \ldots, x_n)}^{\mathbb{R}} - \overbrace{f(x_1 \ldots x_n)}^{\mathbb{R}}}{h \quad \mathbb{R}}$$

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

$$\boxed{f(x) = x^T x = \|x\|_2^2 \qquad \nabla_x f(x) = 2x^T}$$

$g: \mathbb{R}^n \to \mathbb{R} \quad g(x) = c^T x$

Prove $\boxed{\nabla_x g(x) = c^T}$

Kronecker

~~Kronoker~~ Delta

$$\delta_{ij} = \begin{cases} 1 & i=j \\ 0 & i \neq j \end{cases}$$

$$\nabla_x g(x) = \left[\frac{\partial (c^T x)}{\partial x_1}, \ldots, \frac{\partial (c^T x)}{\partial x_n} \right]^T \in \mathbb{R}^{1 \times n}$$

$$c^T x = \sum_i c_i x_i$$

$$\frac{\partial x_i}{\partial x_j} = \begin{cases} 1 & i=j \\ 0 & i \neq j \end{cases} = \delta_{ij}$$

$$\frac{\partial}{\partial x_j}\left(\sum_{i=1}^n c_i x_i \right) = \sum_{i=1}^n c_i \frac{\partial x_i}{\partial x_j}$$

$$= \sum_{i=1}^n c_i \delta_{ij} = c_j$$

$$\frac{\partial f}{\partial x_1} \qquad \frac{\partial^2 f}{\partial x_1^2}$$

$$\nabla_x g(x) = [c_1, c_2, \ldots, c_n]^{\cancel{T}} = c^T$$

$$= \frac{\partial}{\partial x_1}\left(\frac{\partial f}{\partial x_1} \right)$$

$$\nabla_x (c^T x) = c^T \qquad \frac{d}{dx}(cx) = c.$$

Second Order.

$$\frac{\partial^2 f}{\partial x_1 \partial x_2} = \frac{\partial}{\partial x_1}\left(\frac{\partial f}{\partial x_2} \right)$$

$= \|x\|_2^2$ $\quad x \in \mathbb{R}^N$ $\quad \boxed{f: \mathbb{R}^N \to \mathbb{R}}$

Prove $\nabla_x (x^T x) = 2x^T$

$$\nabla_x (\bar{x}^T \bar{x}) = \left[\frac{\partial x^T x}{\partial x_1}, \dots, \frac{\partial x^T x}{\partial x_n} \right] \in \mathbb{R}^{1 \times n}$$

$$\frac{\partial}{\partial x_j} (x^T x) = \frac{\partial}{\partial x_j} \left(\sum_i x_i^2 \right) = \sum_i \frac{\partial x_i^2}{\partial x_j} \begin{cases} 2x_j & j = i \\ 0 & j \neq i \end{cases}$$

$$= \sum_i 2x_i \delta_{ij} = 2x_j$$

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

$$\nabla_x (x^T x) = [2x_1, \dots, 2x_n] = 2x^T$$

$$\boxed{\nabla_x (c^T x) = c^T}$$

5.3 Gradients of Vector-Valued Functions

- The collection of all first-order partial derivatives of a vector-valued function $\boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^m$ is called the Jacobian. The Jacobian $\boldsymbol{J}$ is an $m \times n$ matrix, which we define and arrange as follows:

$$\boldsymbol{J} = \nabla_{\boldsymbol{x}} \boldsymbol{f} = \frac{d\boldsymbol{f}(\boldsymbol{x})}{d\boldsymbol{x}} = \begin{bmatrix} \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_n} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\boldsymbol{x})}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_n} \end{bmatrix}$$

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \qquad \boldsymbol{J}(i,j) = \frac{\partial f_i}{\partial x_j}$$

- The elements of $\boldsymbol{f}$ define the rows and the elements of $\boldsymbol{x}$ define the columns of the corresponding Jacobian

- Special case: for a function $\boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^1$ which maps a vector $\boldsymbol{x} \in \mathbb{R}^n$ onto a scalar, i.e., $m = 1$, the Jacobian is a row vector of dimension $1 \times n$.

5.3 Gradients of Vector-Valued Functions

$f: \mathbb{R}^n \to \mathbb{R}^m \qquad \nabla_x f: \mathbb{R}^n \to \mathbb{R}^{m \times n}$

codomain domain

- If $f: \mathbb{R} \to \mathbb{R}$, the gradient is a scalar
- If $f: \mathbb{R}^D \to \mathbb{R}$, the gradient is a $1 \times D$ row vector
- If $\boldsymbol{f}: \mathbb{R} \to \mathbb{R}^E$, the gradient is a $E \times 1$ column vector
- If $\boldsymbol{f}: \mathbb{R}^D \to \mathbb{R}^E$, the gradient is an $E \times D$ matrix

$f: \underline{\mathbb{R}^{(n_1 \times \ldots \times n_k)}} \to \underline{\mathbb{R}^{(m_1 \times \ldots \times m_s)}}$

$\nabla_x f: \mathbb{R}^{(n_1 \times \ldots \times n_k)} \to \mathbb{R}^{(m_1 \times \ldots \times m_s) \times (n_1 \times \ldots \times n_k)}$

$\frac{df}{dt} = \underbrace{\frac{df}{d\vec{x}} \cdot \frac{d\vec{x}}{dt}}$

$f = g \circ h$

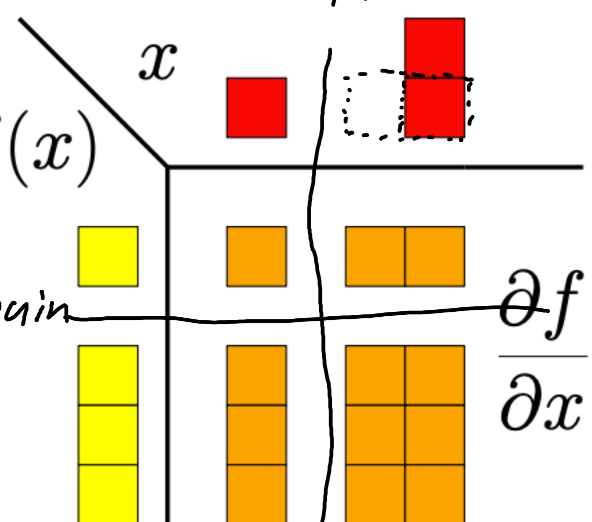
$h: \mathbb{R} \to \mathbb{R}^n$

$g: \mathbb{R}^n \to \mathbb{R}$

$\mathbb{R}^m$

Transposed domain $(\mathbb{R}^{1 \times m})$

$\mathbb{R}^n$



codomain

$\mathbb{R}^{m \times n}$

$f: \underbrace{\mathbb{R}^{a \times b}} \to \mathbb{R}^{c}$

$\nabla_x f(x): \mathbb{R}^{a \times b} \to \mathbb{R}^{c \times (a \times b)}$

$\nabla_x f(x) \in \mathbb{R}^{c \times (a \times b)}$

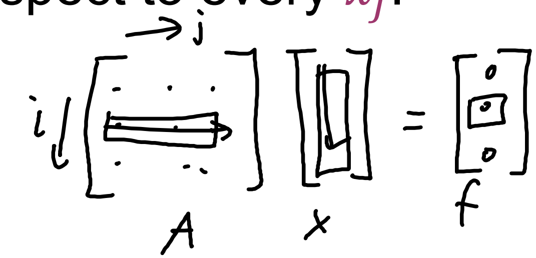
$\underbrace{\mathbb{R}^n} = \underbrace{\mathbb{R}^{n \times 1}}$

Example - Gradient of a Vector-Valued Function

$\vec{f}: \mathbb{R}^{n} \to \mathbb{R}^{m}$

- We are given $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$, $\boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^M$, $\boldsymbol{A} \in \mathbb{R}^{M\times N}$, $\boldsymbol{x} \in \mathbb{R}^N$.
- To compute the gradient $d\boldsymbol{f}/d\boldsymbol{x}$ we first determine the dimension of $d\boldsymbol{f}/d\boldsymbol{x}$: Since $\boldsymbol{f}: \mathbb{R}^N \to \mathbb{R}^M$, it follows that $d\boldsymbol{f}/d\boldsymbol{x} \in \mathbb{R}^{M\times N}$.
- Then, we determine the partial derivatives of $\boldsymbol{f}$ with respect to every x_j:

$$f_i(\boldsymbol{x}) = \sum_{j=1}^{N} A_{ij}x_j \Rightarrow \frac{\partial f_i}{\partial x_j} = A_{ij}$$



$$\boxed{\frac{\partial f_i}{\partial x_j}} = \frac{\partial}{\partial x_j}\left(\sum_k A_{ik}x_k\right) = \sum_k A_{ik}\cancel{\frac{\partial x_k}{\partial x_j}}\,\delta_{jk} = \boxed{A_{ij}}$$

- We collect the partial derivatives in the Jacobian and obtain the gradient

$$\frac{d}{dx}(cx) = c \qquad \frac{d}{d\vec{x}} A\vec{x} = A$$

$$\boxed{\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_N} \\ \vdots & & \vdots \\ \frac{\partial f_M}{\partial x_1} & \cdots & \frac{\partial f_M}{\partial x_N} \end{bmatrix} = \begin{bmatrix} A_{11} & \cdots & A_{1N} \\ \vdots & & \vdots \\ A_{M1} & \cdots & A_{MN} \end{bmatrix} = \boxed{\boldsymbol{A}} \in \mathbb{R}^{M\times N}$$

Example - Chain Rule

$\left(e^{f(x)}\right)' = f'(x)e^{f(x)}$

$\frac{\partial x_1}{\partial t} = \cos t - t\sin t$

$\frac{\partial x_2}{\partial t} = \sin t + t\cos t$

$\frac{\partial f}{\partial x_1} = x_2^2 \exp(x_1 x_2^2)$

$\frac{\partial f}{\partial x_2} = 2x_1x_2 \exp(x_1 x_2^2)$

- Consider the function $h: \mathbb{R} \to \mathbb{R}$, $h(t) = (f \circ g)(t)$ with

$$f: \mathbb{R}^2 \to \mathbb{R}$$
$$g: \mathbb{R} \to \mathbb{R}^2$$
$$f(\boldsymbol{x}) = \exp(x_1 x_2^2)$$
$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = g(t) = \begin{bmatrix} t\cos t \\ t\sin t \end{bmatrix}$$

- We compute the gradient of h with respect to t. Since $f: \mathbb{R}^2 \to \mathbb{R}$ and $g: \mathbb{R} \to \mathbb{R}^2$ we note that

$$\frac{\partial f}{\partial \boldsymbol{x}} \in \mathbb{R}^{1\times 2}, \qquad \frac{\partial g}{\partial t} \in \mathbb{R}^{2\times 1}$$

- The desired gradient is computed by applying the chain rule: *and product rule.*

$$\boxed{\frac{dh}{dt}} = \frac{\partial f}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{x}}{\partial t} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial t} \end{bmatrix}$$

$$= \begin{bmatrix} \exp(x_1x_2^2)x_2^2 & 2\exp(x_1x_2^2)x_1x_2 \end{bmatrix} \begin{bmatrix} \cos t - t\sin t \\ \sin t + t\cos t \end{bmatrix}$$

$$= \exp(x_1x_2^2)\left(x_2^2(\cos t - t\sin t) + 2x_1x_2(\sin t + t\cos t)\right)$$

where $x_1 = t\cos t$ and $x_2 = t\sin t$

$f(x_1, x_2) = \exp(x_1 \cdot x_2^2) \quad g(t) = \begin{bmatrix} t\cos t \\ t\sin t \end{bmatrix}$

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

Alternative: $h(t) = (f \circ g)(t) = f\left(\begin{bmatrix} t\cos t \\ t\sin t \end{bmatrix}\right) = \underline{\exp\left((t\cos t)(t\sin t)^2\right)}$

$h'(t) = \left(\underbrace{t\cos t\,(t\sin t)^2}\right)' \exp\left((t\cos t)(t\sin t)^2\right)$

$\underbrace{\left(t^3 \cos t \sin^2 t\right)'}$

$\left(e^{f(x)}\right)' = f'(x)e^{f(x)}$

$(fgh)'$ Triple Product Rule

$= f'(gh) + f(gh)'$

$= f'(gh) + f(g'h + gh')$

$= f'gh + fg'h + fgh'$

Exercise:

Same answer as before.

Example - Gradient of a Least-Squares Loss in a Linear Model

ϕ features   targets Y

- Let us consider the linear model

$L = \|\hat{y} - y\| = \| y - \phi\theta \|_2^2$

$$\hat{\boldsymbol{y}} = \boldsymbol{\Phi\theta}$$

$L \circ e : \mathbb{R}^D \to \mathbb{R}$

where $\boldsymbol{\theta} \in \mathbb{R}^D$ is a parameter vector, $\boldsymbol{\Phi} \in \mathbb{R}^{N \times D}$ are input features and $\boldsymbol{y} \in \mathbb{R}^N$ are the corresponding observations. We define the functions

$$L(\boldsymbol{e}) := \| \boldsymbol{e} \|_2^2,$$

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

$$\boldsymbol{e}(\boldsymbol{\theta}) := \underbrace{\boldsymbol{y} - \boldsymbol{\Phi\theta}}$$

$e : \mathbb{R}^D \to \mathbb{R}^N$

- We seek $\frac{\partial L}{\partial \boldsymbol{\theta}}$, and we will use the chain rule for this purpose. L is called a least-squares loss function.

- First, we determine the dimensionality of the gradient as

$$\frac{\partial L}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{1 \times D}$$

(as $L \circ e : \mathbb{R}^D \to \mathbb{R}$)

- The chain rule allows us to compute the gradient as

$\nabla_x (Ax) = A$

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = \underbrace{\frac{\partial L}{\partial \boldsymbol{e}}}\,\underbrace{\frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\theta}}}$$

$$\frac{\partial L}{\partial \vec{e}} = \frac{\partial}{\partial \vec{e}} (\vec{e}^T \vec{e}) = 2\vec{e}^T$$

$$\frac{\partial \vec{e}}{\partial \vec{\theta}} = \frac{\partial}{\partial \vec{\theta}} (\cancel{y} - \Phi\vec{\theta}) = \frac{\partial}{\partial \vec{\theta}} (-\Phi\vec{\theta})$$

$$= -\Phi$$

Example - Gradient of a Least-Squares Loss in a Linear Model

$L(\theta): \mathbb{R}^n \longrightarrow \mathbb{R}$

- We know that $||\boldsymbol{e}||^2 = \boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}$ and determine

$$\frac{\partial L}{\partial \boldsymbol{e}} = 2\boldsymbol{e}^{\mathrm{T}} \in \boxed{\mathbb{R}^{1\times N}}$$

$\vec{e} = \vec{y} - \Phi\vec{\theta}$

$e(\theta): \mathbb{R}^D \longrightarrow \mathbb{R}^N$

- Further, we obtain

$$\frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\theta}} = -\boldsymbol{\Phi} \in \boxed{\mathbb{R}^{N\times D}}$$

- Our desired derivative is

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = -2\boldsymbol{e}^{\mathrm{T}}\boldsymbol{\Phi} = -\underbrace{2(\boldsymbol{y}^{\mathrm{T}} - \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}})}_{1\times\mathrm{N}}\underbrace{\boldsymbol{\Phi}}_{\mathrm{N}\times\mathrm{D}} \in \mathbb{R}^{1\times D}$$

Solve $\frac{\partial L}{\partial \theta} = 0$ for θ, to find best θ.

$$-\cancel{2}(Y^T - \theta^T\phi^T)\phi = 0$$

$$Y^T\phi - \theta^T\phi^T\phi = 0$$

$$Y^T\phi = \theta^T\phi^T\phi$$

$$Y^T\phi(\phi^T\phi)^{-1} = \theta^T$$

$$\theta = ((\phi^T\phi)^T)^{-1}\phi^T Y = \underbrace{(\phi^T\phi)^{-1}\phi^T}_{\text{Psued0-Inv.}} Y$$

5.4 Gradients of Matrices

domain of interest — codomain

$$f(A, x) : \mathbb{R}^{M\times N} \times \mathbb{R}^{N} \longrightarrow \mathbb{R}^{M}$$

- Consider the following example

$$\boldsymbol{f} = \boldsymbol{A}\boldsymbol{x}, \qquad \cancel{\boldsymbol{f} \in \mathbb{R}^{M}}, \qquad \boldsymbol{A} \in \mathbb{R}^{M\times N}, \qquad \boldsymbol{x} \in \mathbb{R}^{N}$$

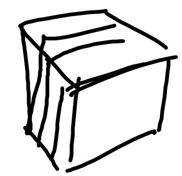
- We seek the gradient $\boxed{\frac{d\boldsymbol{f}}{d\boldsymbol{A}}}$

$$f_x(A) : \boxed{\mathbb{R}^{M\times N}} \rightarrow \boxed{\mathbb{R}^{M}}$$

- First, we determine the dimension of the gradient

$$\frac{d\boldsymbol{f}}{d\boldsymbol{A}} \in \mathbb{R}^{M\times(M\times N)}$$

(3- tensor)

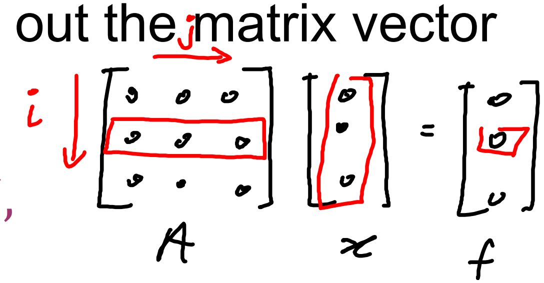


- By definition, the gradient is the collection of the partial derivatives:

$$\frac{d\boldsymbol{f}}{d\boldsymbol{A}} = \begin{bmatrix} \frac{\partial f_1}{\partial \boldsymbol{A}} \\ \vdots \\ \frac{\partial f_M}{\partial \boldsymbol{A}} \end{bmatrix}, \qquad \boxed{\frac{\partial f_i}{\partial \boldsymbol{A}} \in \mathbb{R}^{1\times(M\times N)}} \neq \mathbb{R}^{M\times N}$$

- To compute the partial derivatives, we explicitly write out the matrix vector multiplication

$$f_i = \sum_{j=1}^{N} A_{ij} x_j, \qquad i = 1, \cdots, M,$$



$$\boxed{f_i = \sum_j A_{ij} x_j} \qquad \nabla_A f_i$$

compute $$\frac{\partial f_i}{\partial A_{pq}} = \frac{\partial}{\partial A_{pq}} \left(\sum_j A_{ij} x_j \right)$$

$$= \sum_j x_j \underbrace{\frac{\partial A_{ij}}{\partial A_{pq}}} = \begin{cases} 1 & \boxed{\begin{matrix} i=p \\ j=q \end{matrix}} \\ 0 & \text{else} \end{cases} = \delta_{ip} \delta_{jq}$$

$$= \underbrace{\sum_j x_j \delta_{ip} \underbrace{\delta_{jq}}} = \underline{x_q \delta_{ip}}$$

$$\frac{\partial f_i}{\partial A_{iq}} = x_q \qquad \frac{\partial f_i}{\partial A_{pq}} = 0$$

$$p \neq i$$

$$f_i = \sum_{j=1}^{N} A_{ij} x_j, \qquad i = 1, \cdots, M,$$

$\frac{\partial f_i}{\partial A_{pq}}$

- The partial derivatives are then given as

$$\frac{\partial f_i}{\partial A_{iq}} = x_q$$

$\frac{\partial f_i}{\partial A_{pq}} = 0$

$p \neq i$

- Partial derivatives of f_i with respect to a row of $\boldsymbol{A}$ are given as

$$\frac{\partial f_i}{\partial A_{i,:}} = \boldsymbol{x}^{\mathrm{T}} \in \mathbb{R}^{1\times1\times N}, \qquad \frac{\partial f_i}{\partial A_{k\neq i,:}} = \boldsymbol{0}^{\mathrm{T}} \in \mathbb{R}^{1\times1\times N}$$

- Since f_i maps onto $\mathbb{R}$ and each row of $\boldsymbol{A}$ is of size $1\times N$, we obtain a $1\times1\times N$ sized tensor as the partial derivative of f_i with respect to a row of $\boldsymbol{A}$.
- We stack the partial derivatives and get the desired gradient

$$\boxed{\frac{\partial f_i}{\partial \boldsymbol{A}}} = \begin{array}{c} \begin{bmatrix} \boldsymbol{0}^{\mathrm{T}} \\ \vdots \\ \boldsymbol{0}^{\mathrm{T}} \\ \boldsymbol{x}^{\mathrm{T}} \\ \boldsymbol{0}^{\mathrm{T}} \\ \vdots \\ \boldsymbol{0}^{\mathrm{T}} \end{bmatrix} \end{array} \begin{array}{l} \leftarrow 1 \\ \vdots \\ \leftarrow i-1 \\ \leftarrow i \\ \leftarrow i+1 \\ \vdots \\ \leftarrow M \end{array} \in \mathbb{R}^{1\times(M\times N)} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ x_1 & x_2 & \cdots & x_n \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

So $$\frac{df}{dA} = \left[\begin{bmatrix} x^T \\ 0^T \\ \vdots \\ 0^T \end{bmatrix}, \begin{bmatrix} 0^T \\ x^T \\ \vdots \\ 0^T \end{bmatrix}, \dots, \begin{bmatrix} 0^T \\ 0^T \\ \vdots \\ x^T \end{bmatrix} \right] \in \mathbb{R}^{M \times (M \times N)}$$

$$\frac{df}{dA} = \underbrace{\begin{array}{c} \overset{N}{\longleftrightarrow} \\ M \begin{bmatrix} x_1, x_2, \dots & x_n \\ 0, 0, \dots & 0 \\ \vdots & \\ 0, 0, \dots, & 0 \end{bmatrix} \leftarrow i^{th} \end{array} \quad \begin{bmatrix} 0 & 0 & \dots & 0 \\ x_1 & x_2 & \dots & x_n \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix} \quad \dots \quad \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ x_1 & x_2 & \dots & x_n \end{bmatrix}}_{M}$$

Example - Gradient of Matrices with Respect to Matrices

- Consider a matrix $\boldsymbol{R} \in \mathbb{R}^{M\times N}$ and $\boldsymbol{f}: \mathbb{R}^{M\times N} \to \mathbb{R}^{N\times N}$ with

$$\boldsymbol{f}(\boldsymbol{R}) = \boxed{\boldsymbol{R}^{\mathrm{T}}\boldsymbol{R} =: \boldsymbol{K}} \in \mathbb{R}^{N\times N}$$

- We seek the gradient $\frac{d\boldsymbol{K}}{d\boldsymbol{R}} = \frac{d\, f(R)}{dR}$
- First, the dimension of the gradient is given as

$$\frac{d\boldsymbol{K}}{d\boldsymbol{R}} \in \mathbb{R}^{(N\times N)\times(M\times N)} \qquad (4\text{-tensor})$$

$$\frac{dK_{pq}}{d\boldsymbol{R}} \in \underbrace{\mathbb{R}^{1\times M\times N}} \approx R^{M\times N}$$

$$K = R^T R$$

for $p, q = 1, \ldots, N$, where K_{pq} is the pqth entry of $\boldsymbol{K} = \boldsymbol{f}(\boldsymbol{R})$.

$$K_{pq} = (R^T R)_{pq}$$

- Denoting the ith column of $\boldsymbol{R}$ by $\boldsymbol{r}_i$, every entry of $\boldsymbol{K}$ is given by the dot product of two columns of $\boldsymbol{R}$, i.e.,

$$= \sum_m R^T_{pm} R_{mq}$$

$$\underbrace{K_{pq} = \boldsymbol{r}_p^{\mathrm{T}}\boldsymbol{r}_q = \sum_{m=1}^{M} R_{mp}\, R_{mq}}$$

Example - Gradient of Matrices with Respect to Matrices

- Denoting the ith column of $\boldsymbol{R}$ by $\boldsymbol{r}_i$, every entry of $\boldsymbol{K}$ is given by the dot product of two columns of $\boldsymbol{R}$, i.e.,

$$K_{pq} = \boldsymbol{r}_p^{\mathrm{T}}\boldsymbol{r}_q = \sum_{m=1}^{M} R_{mp}\,R_{mq}$$

$$\frac{\partial K_{pq}}{\partial R_{ij}} = \frac{\partial}{\partial R_{ij}} \sum_{m} R_{mp} R_{mq}$$

- We now compute the partial derivative $\frac{\partial K_{pq}}{\partial R_{ij}}$, we obtain

$$\frac{\partial K_{pq}}{\partial R_{ij}} = \sum_{m=1}^{M} \boxed{\frac{\partial}{\partial R_{ij}} R_{mp} R_{mq}} = \partial_{pqij}$$

δim

$i = m,\ j = p,\ j = q$

Exercise.

$$\boxed{\partial_{pqij}} = \begin{cases} R_{iq} & if\ \ j = p, p \neq q \\ R_{ip} & if\ \ j = q, p \neq q \\ 2R_{iq} & if\ \ j = p, p = q \\ 0 & \text{otherwise} \end{cases}$$

$\leftarrow R_{mq} \quad R_{ij}$

(4-tensor)

- The desired gradient has the dimension $(N\times N)\times(M\times N)$, and every single entry of this tensor is given by ∂_{pqij}, where $p, q, j = 1, \dots, N$ and $i = 1, \dots, M$

5.5 Useful Identities for Computing Gradients

- Some useful gradients that are frequently required in machine learning
- $\mathrm{tr}(\cdot)$: trace $\quad \det(\cdot)$: determinant $\quad \boldsymbol{f}(\boldsymbol{X})^{-1}$: the inverse of $\boldsymbol{f}(\boldsymbol{X})$

$x \cdot y = x^T y$

$\langle x, y \rangle_A = x^T A y$

A is sym and pos. def.

$$\frac{\partial \boldsymbol{x}^{\mathrm{T}}\boldsymbol{a}}{\partial \boldsymbol{x}} = \boldsymbol{a}^{\mathrm{T}}$$

$$\frac{\partial \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{a}^{\mathrm{T}}$$

Already proven.

$$\boxed{\frac{\partial \boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{b}}{\partial \boldsymbol{X}} = \boldsymbol{a}\boldsymbol{b}^{\mathrm{T}}}$$

We will prove this!

$$\boxed{\frac{\partial \boldsymbol{x}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{x}^{\mathrm{T}}(\boldsymbol{B} + \boldsymbol{B}^{\mathrm{T}})}$$

Assignment 2

$$\frac{\partial}{\partial \boldsymbol{s}}\underline{(\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s})^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s})} = -2(\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s})^{\mathrm{T}}\boldsymbol{W}\boldsymbol{A} \quad \text{for symmetric } \boldsymbol{W}$$

You should be able to calculate these gradients

Prove

$$\boxed{\frac{\partial(a^T X b)}{\partial X} = a b^T}$$

$a \in \mathbb{R}^{n\times 1}$ $b \in \mathbb{R}^{m\times 1}$ $X \in \mathbb{R}^{n\times m}$

$a^T X b \in \mathbb{R}$

$1\times n \quad n\times m \quad m\times 1$

$f(X) = \bar{a}^T X \bar{b}$

$: \mathbb{R}^{n\times m} \longrightarrow \mathbb{R}^1$

$\underbrace{\nabla_X}(a^T X b) \in \mathbb{R}^{1\times(n\times m)} \simeq \mathbb{R}^{n\times m}$

$\nabla_X(a^T X b) = \nabla_X(a^T(Xb))$

$$a^T X b = a^T(Xb) = \sum_i a_i (Xb)_i = \sum_i a_i \left(\sum_j x_{ij} b_j\right)$$

(ab^T)

$$\begin{bmatrix} \vdots \end{bmatrix}[\cdots] = \begin{bmatrix} a_1 b_1 & a_2 b_1 & \\ a_1 b_2 & \ddots & \\ \vdots & & \ddots \end{bmatrix}$$

$$= \sum_{i,j} a_i x_{ij} b_j$$

$[\quad][\;] \to \square$

$[\cdot][\cdot] \to [+]$

$$\boxed{\frac{\partial}{\partial X_{pq}}\left(\sum_{i,j} a_i x_{ij} b_j\right)}$$

$$= \sum_{i,j} a_i b_j \frac{\partial x_{ij}}{\partial x_{pq}}$$

$$\begin{cases} 1 & i=p,\ j=q \\ 0 & \text{else} \end{cases}$$

$$\frac{\partial x_{ij}}{\partial x_{pq}} = \delta_{ip}\delta_{jq}$$

$\underbrace{\delta_{jq}} = \underbrace{\delta_{qj}}$

$$= \sum_{i,j} a_i b_j \underbrace{\delta_{ip}}\,\underbrace{\delta_{jq}} \xrightarrow[j=q]{i=p} \underbrace{a_p}\,\underbrace{b_q} = \boxed{(\vec{a}\,\vec{b}^T)_{pq}}$$

h(x)

$$\boxed{\frac{\partial}{\partial X}\left(a^T X b\right)}$$

$f(g) = a^T g$

$g(\cancel{X}) = Xb$

$h(X) = (f \circ g)(X)$

$= f(g(X))$

$g : \mathbb{R}^{n \times m} \to \mathbb{R}^n$

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

$$\frac{\partial h}{\partial X} = \underbrace{\underbrace{\frac{\partial h}{\partial g}}_{\text{vector}} \boxed{\frac{\partial g}{\partial X}}}_{\text{matrix}}$$

3 tensor

K-Means

$$r_{nk} = \begin{cases} 1 & x_n \text{ to cluster } k \\ 0 & \text{else} \end{cases}$$

$$\begin{bmatrix} \vdots & \vdots & & \vdots \\ \mu_1 & \mu_2 & \cdots & \mu_K \\ \vdots & \vdots & & \vdots \end{bmatrix}$$

$$L(M, R) = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|_2^2 \qquad R_{nk} = r_{nk} \quad M = (\mu_1, \ldots, \mu_K)$$

$$\boxed{\mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}}}$$

Compute $\nabla_M L = \left[\frac{\partial L}{\partial \mu_1}, \ldots, \frac{\partial L}{\partial \mu_K} \right] = 0$

$$\frac{\partial L}{\partial \vec{\mu}_j} = \frac{\partial}{\partial \vec{\mu}_j} \left(\sum_n \sum_k r_{nk} \|\bar{x}_n - \bar{\mu}_k\|_2^2 \right)$$

j^{th}

$$r_{n1} \|x_n - \mu_1\|_2^2 + \ldots + r_{nK} \|x_n - \mu_K\|_2^2$$

$$= \sum_n \frac{\partial}{\partial \mu_j} \underbrace{\sum_k r_{nk} \|\bar{x}_n - \bar{\mu}_k\|_2^2}$$

(only non-zero term $k=j$)

$$\frac{\partial}{\partial \bar{x}} (c^T \bar{x}) = c^T$$

$$\frac{\partial}{\partial \bar{x}} (\bar{x}^T \bar{x}) = 2\bar{x}^T$$

$$= \sum_n r_{nj} \frac{\partial}{\partial \mu_j} \|\bar{x}_n - \mu_j\|_2^2$$

$$\|x_n - \mu_j\|_2^2 = (x_n - \mu_j)^T (x_n - \mu_j)$$

$$= \sum_n r_{nj} \frac{\partial}{\partial \mu_j} \left(-2x_n^T \mu_j + \mu_j^T \mu_j \right)$$

$$= \cancel{x_n^T x_n} - \underbrace{\mu_j^T x_n - x_n^T \mu_j} + \mu_j^T \mu_j$$

$$= \sum_n r_{nj} \left(-2x_n^T + 2\mu_j^T \right) = 0$$

$$= -2x_n^T \mu_j + \mu_j^T \mu_j$$

$$\sum_n r_{nj} \left(-2x_n^T + 2\mu_j^T \right) = 0$$

$$= \sum_n r_{nj} \left(-\cancel{2} x_n^T \right) + \sum_n r_{nj} \left(\cancel{2} \mu_j^T \right) = 0$$

$$\sum_n r_{nj} \mu_j = \sum_n r_{nj} x_n$$

$$\mu_j = \frac{\sum_n r_{nj} x_n}{\sum_n r_{nj}}$$