

# Mathematics for Machine Learning

## — Vector Calculus: Gradients of Vector-Valued Functions and Matrices

Joseph Chuang-Chieh Lin

Department of Computer Science & Engineering,  
National Taiwan Ocean University

Fall 2025

# Credits for the resource

- The slides are based on the textbooks:
  - *Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong: Mathematics for Machine Learning. Cambridge University Press. 2020.*
  - *Howard Anton, Chris Rorres, Anton Kaul: Elementary Linear Algebra. Wiley. 2019.*
- We could partially refer to the monograph:  
*Francesco Orabona: A Modern Introduction to Online Learning.*  
<https://arxiv.org/abs/1912.13213>

# Outline

- 1 Gradients of Vector-Valued Functions
- 2 Gradients of Matrices

# Outline

1 Gradients of Vector-Valued Functions

2 Gradients of Matrices

# Our Focus

- Partial derivatives and gradients of functions  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ , for  $n \geq 1, m > 1$ .

# Vector of Functions

Given

- $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ .
- $\mathbf{x} = [x_1, \dots, x_n]^\top \in \mathbb{R}^n$ .

The corresponding *vector of functions*:

$$\mathbf{f}[\mathbf{x}] = \begin{bmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{bmatrix} \in \mathbb{R}^m.$$

# Vector of Functions

Given

- $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ .
- $\mathbf{x} = [x_1, \dots, x_n]^\top \in \mathbb{R}^n$ .

The corresponding *vector of functions*:

$$\mathbf{f}[\mathbf{x}] = \begin{bmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{bmatrix} \in \mathbb{R}^m.$$

We can view  $\mathbf{f}$  as  $[f_1, \dots, f_m]^\top$ , such that  $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$ .

Therefore,

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



So,

$$\begin{aligned}\frac{d\mathbf{f}}{d\mathbf{x}} &= \left[ \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_1} \quad \dots \quad \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_n} \right] \\ &= \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \dots & \frac{\partial f_1(\mathbf{x})}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(\mathbf{x})}{\partial x_1} & \dots & \frac{\partial f_m(\mathbf{x})}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n}.\end{aligned}$$

So,

$$\begin{aligned}\frac{d\mathbf{f}}{d\mathbf{x}} &= \left[ \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_1} \quad \dots \quad \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_n} \right] \\ &= \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \dots & \frac{\partial f_1(\mathbf{x})}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(\mathbf{x})}{\partial x_1} & \dots & \frac{\partial f_m(\mathbf{x})}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n}.\end{aligned}$$

- We call this collection of all first-order partial derivatives of a vector-valued function  $\mathbf{f}$  the **Jacobian**.

So,

$$\begin{aligned}\frac{d\mathbf{f}}{d\mathbf{x}} &= \left[ \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_1} \quad \dots \quad \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_n} \right] \\ &= \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \dots & \frac{\partial f_1(\mathbf{x})}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(\mathbf{x})}{\partial x_1} & \dots & \frac{\partial f_m(\mathbf{x})}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n}.\end{aligned}$$

- We call this collection of all first-order partial derivatives of a vector-valued function  $\mathbf{f}$  the **Jacobian**.
- ★ Denote by  $\mathbf{J} = \nabla_{\mathbf{x}} \mathbf{f} = \frac{d\mathbf{f}(\mathbf{x})}{d\mathbf{x}}$ 
  - $J(i, j) = \frac{\partial f_i}{\partial x_j}$ .

# Example

## Derivative of a Polynomial

Given  $\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ ,  $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^M$ ,  $\mathbf{A} \in \mathbb{R}^{M \times N}$ , and  $\mathbf{x} \in \mathbb{R}^N$ . Compute

$$\frac{d\mathbf{f}}{d\mathbf{x}} = ?$$

Note:

- $\mathbf{f} : \mathbb{R}^N \rightarrow \mathbb{R}^M$ , so  $\frac{d\mathbf{f}}{d\mathbf{x}} \in \mathbb{R}^{M \times N}$ .
- $f_i(\mathbf{x}) =$

# Example

## Derivative of a Polynomial

Given  $\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ ,  $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^M$ ,  $\mathbf{A} \in \mathbb{R}^{M \times N}$ , and  $\mathbf{x} \in \mathbb{R}^N$ . Compute

$$\frac{d\mathbf{f}}{d\mathbf{x}} = ?$$

Note:

- $\mathbf{f} : \mathbb{R}^N \rightarrow \mathbb{R}^M$ , so  $\frac{d\mathbf{f}}{d\mathbf{x}} \in \mathbb{R}^{M \times N}$ .
- $f_i(\mathbf{x}) = \sum_{j=1}^N A_{ij}x_j$

# Example

## Derivative of a Polynomial

Given  $\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ ,  $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^M$ ,  $\mathbf{A} \in \mathbb{R}^{M \times N}$ , and  $\mathbf{x} \in \mathbb{R}^N$ . Compute

$$\frac{d\mathbf{f}}{d\mathbf{x}} = ?$$

Note:

- $\mathbf{f} : \mathbb{R}^N \rightarrow \mathbb{R}^M$ , so  $\frac{d\mathbf{f}}{d\mathbf{x}} \in \mathbb{R}^{M \times N}$ .
- $f_i(\mathbf{x}) = \sum_{j=1}^N A_{ij}x_j \implies \frac{\partial f_i}{\partial x_j} = A_{ij}$ .

# Example

## Derivative of a Polynomial

Given  $\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ ,  $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^M$ ,  $\mathbf{A} \in \mathbb{R}^{M \times N}$ , and  $\mathbf{x} \in \mathbb{R}^N$ . Compute

$$\frac{d\mathbf{f}}{d\mathbf{x}} = \mathbf{A}$$

Note:

- $\mathbf{f} : \mathbb{R}^N \rightarrow \mathbb{R}^M$ , so  $\frac{d\mathbf{f}}{d\mathbf{x}} \in \mathbb{R}^{M \times N}$ .
- $f_i(\mathbf{x}) = \sum_{j=1}^N A_{ij}x_j \implies \frac{\partial f_i}{\partial x_j} = A_{ij}$ .

## Example: Gradient of a Least-Squared Loss in a Linear Model

Consider the linear model

$$\mathbf{y} = \Phi\boldsymbol{\theta},$$

where

- $\boldsymbol{\theta} \in \mathbb{R}^D$ : a parameter vector
- $\Phi \in \mathbb{R}^{N \times D}$ : input features
- $\mathbf{y} \in \mathbb{R}^N$ : the corresponding observations.

We define that

$$L(\mathbf{e}) := \|\mathbf{e}\|^2.$$

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

Compute  $\frac{\partial L}{\partial \boldsymbol{\theta}}$  (using the chain rule).



## Example (2/3)

Note that

- $\frac{\partial L}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{1 \times D}$  ( $\because L : \mathbb{R}^D \rightarrow \mathbb{R}$ ).
- $\frac{\partial \mathbf{e}}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{N \times D}$  ( $\because \mathbf{e} : \mathbb{R}^D \rightarrow \mathbb{R}^N$ ).
- $\frac{\partial L}{\partial \mathbf{e}} \in \mathbb{R}^{1 \times N}$  ( $\because L : \mathbb{R}^N \rightarrow \mathbb{R}$ ).
- $\frac{\partial L}{\partial \boldsymbol{\theta}} = \frac{\partial L}{\partial \mathbf{e}} \frac{\partial \mathbf{e}}{\partial \boldsymbol{\theta}}$  (chain rule).
- The  $d$ th element:  

$$\frac{\partial L}{\partial \boldsymbol{\theta}}[1, d] = \sum_{i=1}^N \frac{\partial L}{\partial \mathbf{e}}[1, i] \frac{\partial \mathbf{e}}{\partial \boldsymbol{\theta}}[i, d].$$
- $L = \|\mathbf{e}\|^2 = \mathbf{e}^\top \mathbf{e}$  and  

$$\frac{\partial L}{\partial \mathbf{e}} = 2\mathbf{e}^\top \in \mathbb{R}^{1 \times N}.$$
- $\frac{\partial \mathbf{e}}{\partial \boldsymbol{\theta}} = -\Phi \in \mathbb{R}^{N \times D}.$

## Example (3/3)

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = -2\mathbf{e}^\top \boldsymbol{\Phi} = -2(\mathbf{y}^\top - \boldsymbol{\theta}^\top \boldsymbol{\Phi}^\top) \boldsymbol{\Phi} \in \mathbb{R}^{1 \times D}$$

## Example (3/3)

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = -2\mathbf{e}^\top \boldsymbol{\Phi} = -2(\mathbf{y}^\top - \boldsymbol{\theta}^\top \boldsymbol{\Phi}^\top) \boldsymbol{\Phi} \in \mathbb{R}^{1 \times D}$$

By the way, we can obtain the same result without using the chain rule:

$$L_2(\boldsymbol{\theta}) := \|\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\theta}\|^2 = (\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\theta})^\top (\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\theta}).$$

## Example (3/3)

$$\frac{\partial L}{\partial \theta} = -2\mathbf{e}^\top \Phi = -2(\mathbf{y}^\top - \theta^\top \Phi^\top) \Phi \in \mathbb{R}^{1 \times D}$$

By the way, we can obtain the same result without using the chain rule:

$$L_2(\theta) := \|\mathbf{y} - \Phi\theta\|^2 = (\mathbf{y} - \Phi\theta)^\top (\mathbf{y} - \Phi\theta).$$

- It becomes impractical for deep function compositions.

# Outline

- 1 Gradients of Vector-Valued Functions
- 2 Gradients of Matrices

# Motivations

- There are scenarios that we need to take gradients of matrices w.r.t. vectors (or other matrices).
  - ⇒ This results in a multidimensional tensor.
    - Multidimensional array.
- Compute the gradient of an  $m \times n$  matrix  $\mathbf{A}$  w.r.t. a  $p \times q$  matrix  $\mathbf{B}$ :
  - The Jacobian  $\mathbf{J}$  would be  $(m \times n) \times (p \times q)$  (4-dimensional tensor).
    - $J_{ijkl} = \frac{\partial A_{ij}}{\partial B_{kl}}$ .
- Matrices  $\Leftrightarrow$  linear mappings, so

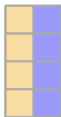
# Motivations

- There are scenarios that we need to take gradients of matrices w.r.t. vectors (or other matrices).
  - ⇒ This results in a multidimensional tensor.
    - Multidimensional array.
- Compute the gradient of an  $m \times n$  matrix  $\mathbf{A}$  w.r.t. a  $p \times q$  matrix  $\mathbf{B}$ :
  - The Jacobian  $\mathbf{J}$  would be  $(m \times n) \times (p \times q)$  (4-dimensional tensor).
    - $J_{ijkl} = \frac{\partial A_{ij}}{\partial B_{kl}}$ .
- Matrices  $\Leftrightarrow$  linear mappings, so

There is a vector-space isomorphism (i.e., linear, invertible mapping) between the space  $\mathbb{R}^{m \times n}$  of  $m \times n$  matrices and the space  $\mathbb{R}^{mn}$  of  $mn$  vectors.

# Visualization of Two Approaches for the Isomorphism

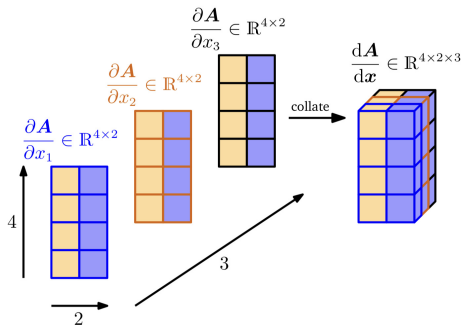
$$A \in \mathbb{R}^{4 \times 2}$$



$$x \in \mathbb{R}^3$$

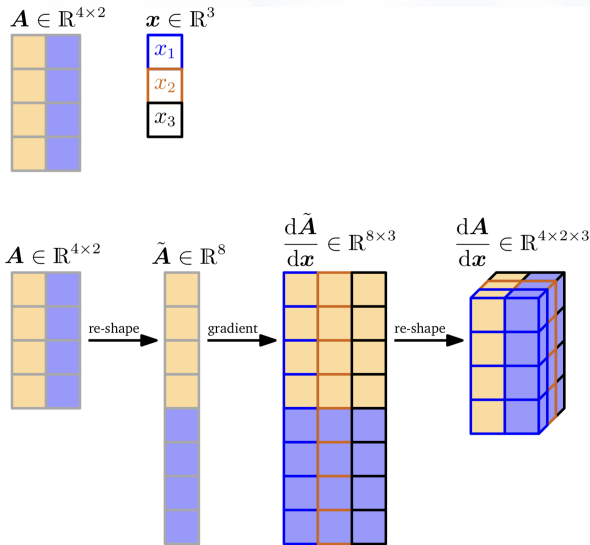


Partial derivatives:





# Visualization of Two Approaches for the Isomorphism



## Example: Gradient of Vectors w.r.t. Matrices

Consider

$$\mathbf{f} = \mathbf{A}\mathbf{x}, \text{ where } \mathbf{f} \in \mathbb{R}^M, \mathbf{A} \in \mathbb{R}^{M \times N}, \mathbf{x} \in \mathbb{R}^N.$$

**Goal:** Compute the gradient  $\frac{d\mathbf{f}}{d\mathbf{A}}$ .

**Note:**

- $\frac{d\mathbf{f}}{d\mathbf{A}} \in$

## Example: Gradient of Vectors w.r.t. Matrices

Consider

$$\mathbf{f} = \mathbf{A}\mathbf{x}, \text{ where } \mathbf{f} \in \mathbb{R}^M, \mathbf{A} \in \mathbb{R}^{M \times N}, \mathbf{x} \in \mathbb{R}^N.$$

**Goal:** Compute the gradient  $\frac{d\mathbf{f}}{d\mathbf{A}}$ .

**Note:**

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

$$\frac{d\mathbf{f}}{d\mathbf{A}} =$$

$$\frac{d\mathbf{f}}{d\mathbf{A}} = \begin{bmatrix} \frac{\partial f_1}{\partial \mathbf{A}} \\ \vdots \\ \frac{\partial f_M}{\partial \mathbf{A}} \end{bmatrix},$$

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

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

- We can explicitly expand  $f_i = \sum_{j=1}^N A_{ij}x_j$ , for  $i = 1, \dots, M$ .

Hence,

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

So we can derive

$$\frac{\partial f_i}{\partial A_{i,:}} = \mathbf{x}^\top$$

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

- We can explicitly expand  $f_i = \sum_{j=1}^N A_{ij}x_j$ , for  $i = 1, \dots, M$ .

Hence,

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

So we can derive

$$\frac{\partial f_i}{\partial A_{i,:}} = \mathbf{x}^\top \in \mathbb{R}^{1 \times (1 \times N)} \quad \text{and} \quad \frac{\partial f_i}{\partial A_{k \neq i,:}} = \mathbf{0}^\top \in \mathbb{R}^{1 \times (1 \times N)}.$$



Stack the partial derivatives:

$$\frac{\partial f_i}{\partial \mathbf{A}} = \begin{bmatrix} \mathbf{0}^\top \\ \vdots \\ \mathbf{0}^\top \\ \mathbf{x}^\top \\ \mathbf{0}^\top \\ \vdots \\ \mathbf{0}^\top \end{bmatrix} \in \mathbb{R}^{1 \times (M \times N)}$$

## Example: Gradient of **Matrices** w.r.t. Matrices

Consider a matrix  $\mathbf{R} \in \mathbb{R}^{M \times N}$  and  $\mathbf{f} : \mathbb{R}^{M \times N} \rightarrow \mathbb{R}^{N \times N}$  with

$$\mathbf{f}(\mathbf{R}) = \mathbf{R}^\top \mathbf{R} := \mathbf{K} \in \mathbb{R}^{N \times N}$$

**Goal:** Compute the gradient  $\frac{d\mathbf{K}}{d\mathbf{R}}$ .

**Note:**

- $\frac{d\mathbf{K}}{d\mathbf{R}} \in$

## Example: Gradient of **Matrices** w.r.t. Matrices

Consider a matrix  $\mathbf{R} \in \mathbb{R}^{M \times N}$  and  $\mathbf{f} : \mathbb{R}^{M \times N} \rightarrow \mathbb{R}^{N \times N}$  with

$$\mathbf{f}(\mathbf{R}) = \mathbf{R}^\top \mathbf{R} := \mathbf{K} \in \mathbb{R}^{N \times N}$$

**Goal:** Compute the gradient  $\frac{d\mathbf{K}}{d\mathbf{R}}$ .

**Note:**

- $\frac{d\mathbf{K}}{d\mathbf{R}} \in \mathbb{R}^{(N \times N) \times (M \times N)}.$

## Example: Gradient of Matrices w.r.t. Matrices

Consider a matrix  $\mathbf{R} \in \mathbb{R}^{M \times N}$  and  $\mathbf{f} : \mathbb{R}^{M \times N} \rightarrow \mathbb{R}^{N \times N}$  with

$$\mathbf{f}(\mathbf{R}) = \mathbf{R}^\top \mathbf{R} := \mathbf{K} \in \mathbb{R}^{N \times N}$$

**Goal:** Compute the gradient  $\frac{d\mathbf{K}}{d\mathbf{R}}$ .

**Note:**

- $\frac{d\mathbf{K}}{d\mathbf{R}} \in \mathbb{R}^{(N \times N) \times (M \times N)}$ .
- $\frac{dK_{pq}}{d\mathbf{R}} \in \mathbb{R}^{1 \times (M \times N)}$ , for  $p, q = 1, \dots, N$ ,  $K_{pq}$ : the  $(p, q)$ th entry of  $\mathbf{K}$ .

$$K_{pq} = \mathbf{r}_p^\top \mathbf{r}_q = \sum_{t=1}^M R_{tp} R_{tq}.$$

$\mathbf{r}_i$ : the  $i$ th column of  $\mathbf{R}$ .

## Example (2/2)

Compute  $\frac{\partial K_{pq}}{\partial R_{ij}}$ : (sum rule)

$$\frac{\partial K_{pq}}{\partial R_{ij}} = \sum_{t=1}^M \frac{\partial}{\partial R_{ij}} R_{tp} R_{tq} = \delta_{pqij},$$

where

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

## Example (2/2)

Compute  $\frac{\partial K_{pq}}{\partial R_{ij}}$ : (sum rule)

$$\frac{\partial K_{pq}}{\partial R_{ij}} = \sum_{t=1}^M \frac{\partial}{\partial R_{ij}} R_{tp} R_{tq} = \delta_{pqij},$$

where

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

Hence, each entry of the desired gradient  $\frac{d\mathbf{K}}{d\mathbf{R}} \in \mathbb{R}^{(N \times N) \times (M \times N)}$  is  $\delta_{pqij}$ , for  $p, q, j = 1, \dots, N$  and  $i = 1, \dots, M$ .

## Useful Identities for Computing Gradients (1/2)

Reference: *The Matrix Cookbook* by Petersen and Pedersen, 2012.

$$\frac{\partial}{\partial \mathbf{X}} f(\mathbf{X})^\top = \left( \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} \right)^\top.$$

$$\frac{\partial}{\partial \mathbf{X}} \text{tr}(f(\mathbf{X})) = \text{tr} \left( \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} \right).$$

$$\frac{\partial}{\partial \mathbf{X}} \det(f(\mathbf{X})) = \det(f(\mathbf{X})) \text{tr} \left( f(\mathbf{X})^{-1} \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} \right)$$

# Useful Identities for Computing Gradients (1/2)

Reference: *The Matrix Cookbook* by Petersen and Pedersen, 2012.

$$\frac{\partial}{\partial \mathbf{X}} f(\mathbf{X})^\top = \left( \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} \right)^\top.$$

$$\frac{\partial}{\partial \mathbf{X}} \text{tr}(f(\mathbf{X})) = \text{tr} \left( \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} \right).$$

$$\frac{\partial}{\partial \mathbf{X}} \det(f(\mathbf{X})) = \det(f(\mathbf{X})) \text{tr} \left( f(\mathbf{X})^{-1} \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} \right) \Rightarrow \text{Jacobi's formula}$$

$$\frac{\partial}{\partial \mathbf{X}} f(\mathbf{X})^{-1} = -f(\mathbf{X})^{-1} \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} f(\mathbf{X})^{-1}$$

$$\frac{\partial \mathbf{a}^\top \mathbf{X}^{-1} \mathbf{b}}{\partial \mathbf{X}} = -(\mathbf{X}^{-1})^\top \mathbf{a} \mathbf{b}^\top (\mathbf{X}^{-1})^\top$$



## Useful Identities for Computing Gradients (2/2)

$$\frac{\partial \mathbf{x}^\top \mathbf{a}}{\partial \mathbf{x}} = \mathbf{a}^\top$$

$$\frac{\partial \mathbf{a}^\top \mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}^\top$$

$$\frac{\partial \mathbf{a}^\top \mathbf{X} \mathbf{b}}{\partial \mathbf{X}} = \mathbf{a} \mathbf{b}^\top$$

$$\frac{\partial \mathbf{x}^\top \mathbf{B} \mathbf{x}}{\partial \mathbf{x}} = \mathbf{x}^\top (\mathbf{B} + \mathbf{B}^\top)$$

$$\frac{\partial}{\partial \mathbf{s}} (\mathbf{x} - \mathbf{A} \mathbf{s})^\top \mathbf{W} (\mathbf{x} - \mathbf{A} \mathbf{s}) = -2(\mathbf{x} - \mathbf{A} \mathbf{s})^\top \mathbf{W} \mathbf{A} \text{ for symmetric } \mathbf{W}.$$

## Clarification of some identities

$$\mathbf{0} = \frac{\partial}{\partial \mathbf{X}} I$$

## Clarification of some identities

$$\mathbf{0} = \frac{\partial}{\partial \mathbf{X}} \mathbf{I} = \frac{\partial}{\partial \mathbf{X}} (f(\mathbf{X})^{-1} f(\mathbf{X}))$$

## Clarification of some identities

$$\mathbf{0} = \frac{\partial}{\partial \mathbf{X}} \mathbf{I} = \frac{\partial}{\partial \mathbf{X}} (f(\mathbf{X})^{-1} f(\mathbf{X})) = \left( \frac{\partial}{\partial \mathbf{X}} f(\mathbf{X})^{-1} \right) f(\mathbf{X}) + f(\mathbf{X})^{-1} \left( \frac{\partial}{\partial \mathbf{X}} f(\mathbf{X}) \right).$$

Hence,

## Clarification of some identities

$$\mathbf{0} = \frac{\partial}{\partial \mathbf{X}} \mathbf{I} = \frac{\partial}{\partial \mathbf{X}} (f(\mathbf{X})^{-1} f(\mathbf{X})) = \left( \frac{\partial}{\partial \mathbf{X}} f(\mathbf{X})^{-1} \right) f(\mathbf{X}) + f(\mathbf{X})^{-1} \left( \frac{\partial}{\partial \mathbf{X}} f(\mathbf{X}) \right).$$

Hence,

$$\frac{\partial}{\partial \mathbf{X}} f(\mathbf{X})^{-1} = -f(\mathbf{X})^{-1} \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} f(\mathbf{X})^{-1}.$$

## About the last formula

For symmetric  $\mathbf{W}$ ,

$$\frac{\partial}{\partial \mathbf{s}} (\mathbf{x} - \mathbf{A}\mathbf{s})^\top \mathbf{W} (\mathbf{x} - \mathbf{A}\mathbf{s}) = -2(\mathbf{x} - \mathbf{A}\mathbf{s})^\top \mathbf{W} \mathbf{A}.$$

## About the last formula

For symmetric  $\mathbf{W}$ ,

$$\frac{\partial}{\partial \mathbf{s}} (\mathbf{x} - \mathbf{A}\mathbf{s})^\top \mathbf{W} (\mathbf{x} - \mathbf{A}\mathbf{s}) = -2(\mathbf{x} - \mathbf{A}\mathbf{s})^\top \mathbf{W} \mathbf{A}.$$

Let  $\mathbf{z} = \mathbf{x} - \mathbf{A}\mathbf{s}$  and use

$$\frac{\partial \mathbf{z}^\top \mathbf{W} \mathbf{z}}{\partial \mathbf{z}} = \mathbf{z}^\top (\mathbf{W} + \mathbf{W}^\top)$$

## About the last formula

For symmetric  $\mathbf{W}$ ,

$$\frac{\partial}{\partial \mathbf{s}} (\mathbf{x} - \mathbf{A}\mathbf{s})^\top \mathbf{W} (\mathbf{x} - \mathbf{A}\mathbf{s}) = -2(\mathbf{x} - \mathbf{A}\mathbf{s})^\top \mathbf{W} \mathbf{A}.$$

Let  $\mathbf{z} = \mathbf{x} - \mathbf{A}\mathbf{s}$  and use

$$\frac{\partial \mathbf{z}^\top \mathbf{W} \mathbf{z}}{\partial \mathbf{z}} = \mathbf{z}^\top (\mathbf{W} + \mathbf{W}^\top)$$

Then compute

$$\frac{\partial \mathbf{z}^\top \mathbf{W} \mathbf{z}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{s}}.$$



## A sketch of Jacobi's formula

Reference: Wikipedia page.

$$\frac{\partial}{\partial \mathbf{X}} \det(f(\mathbf{X})) = \det(f(\mathbf{X})) \operatorname{tr} \left( f(\mathbf{X})^{-1} \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} \right)$$

- To simplify the discussion, let  $\mathbf{M} := f(\mathbf{X})$  and denote the differential of  $\mathbf{M}$  by  $d\mathbf{M}$ . We omit the sizes of matrices if the context is clear.

## A sketch of Jacobi's formula

Reference: Wikipedia page.

$$\frac{\partial}{\partial \mathbf{X}} \det(f(\mathbf{X})) = \det(f(\mathbf{X})) \operatorname{tr} \left( f(\mathbf{X})^{-1} \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} \right)$$

- To simplify the discussion, let  $\mathbf{M} := f(\mathbf{X})$  and denote the differential of  $\mathbf{M}$  by  $d\mathbf{M}$ . We omit the sizes of matrices if the context is clear.

That is,

$$d \det(\mathbf{M}) = \det(\mathbf{M}) \operatorname{tr}(\mathbf{M}^{-1} d\mathbf{M}).$$

Fact

$$\sum_i \sum_j \mathbf{A}_{ij} \mathbf{B}_{ij} = \operatorname{tr}(\mathbf{A}^\top \mathbf{B}) \text{ for any square matrices } \mathbf{A}, \mathbf{B}.$$

## A sketch of Jacobi's formula (2/4)

By the cofactor expansion, we have

$$\det(\mathbf{M}) = \sum_j \mathbf{M}_{ij} \operatorname{adj}^\top(\mathbf{M})_{ij}$$

and recall that

$$\mathbf{M} \operatorname{adj}^\top(\mathbf{M}) = \det(\mathbf{M}) \mathbf{I},$$

## A sketch of Jacobi's formula (2/4)

By the cofactor expansion, we have

$$\det(\mathbf{M}) = \sum_j \mathbf{M}_{ij} \operatorname{adj}^\top(\mathbf{M})_{ij}$$

and recall that

$$\mathbf{M} \operatorname{adj}^\top(\mathbf{M}) = \det(\mathbf{M}) \mathbf{I},$$

which means

$$\mathbf{M}^{-1} = \frac{1}{\det(\mathbf{M})} \operatorname{adj}^\top(\mathbf{M}).$$

## A sketch of Jacobi's formula (2/4)

By the cofactor expansion, we have

$$\det(\mathbf{M}) = \sum_j \mathbf{M}_{ij} \operatorname{adj}^\top(\mathbf{M})_{ij}$$

and recall that

$$\mathbf{M} \operatorname{adj}^\top(\mathbf{M}) = \det(\mathbf{M}) \mathbf{I},$$

which means

$$\mathbf{M}^{-1} = \frac{1}{\det(\mathbf{M})} \operatorname{adj}^\top(\mathbf{M}).$$

Thus, we are actually proving

$$d \det(\mathbf{M}) = \det(\mathbf{M}) \operatorname{tr}(\mathbf{M}^{-1} d\mathbf{M}) = \operatorname{tr}(\operatorname{adj}^\top(\mathbf{M}) d\mathbf{M}).$$

# Note

- Assume  $\det \mathbf{M} = F(\mathbf{M}_{11}, \mathbf{M}_{12}, \dots, \mathbf{M}_{nn})$  is a function of  $\mathbf{M}_{11}, \mathbf{M}_{12}, \dots, \mathbf{M}_{nn}$  and  $\mathbf{M}_{ij} := \mathbf{M}_{ij}(t)$ . Then,

$$\frac{d}{dt} \det(\mathbf{M}) = \sum_i \sum_j \frac{\partial F}{\partial \mathbf{M}_{ij}} \frac{d\mathbf{M}_{ij}}{dt}.$$

That is,

$$d \det(\mathbf{M}) = \sum_i \sum_j \frac{\partial F}{\partial \mathbf{M}_{ij}} d\mathbf{M}_{ij}.$$

## A sketch of Jacobi's formula (3/4)

Differential of the cofactor expansion:

## A sketch of Jacobi's formula (3/4)

Differential of the cofactor expansion:

$$\frac{\partial \det(\mathbf{M})}{\partial \mathbf{M}_{ij}} = \frac{\partial \sum_k \mathbf{M}_{ik} \text{adj}^\top(\mathbf{M})_{ik}}{\partial \mathbf{M}_{ij}} = \sum_k \frac{\partial \mathbf{M}_{ik} \text{adj}^\top(\mathbf{M})_{ik}}{\partial \mathbf{M}_{ij}}$$

Applying the product rule we can derive



## A sketch of Jacobi's formula (3/4)

Differential of the cofactor expansion:

$$\frac{\partial \det(\mathbf{M})}{\partial \mathbf{M}_{ij}} = \frac{\partial \sum_k \mathbf{M}_{ik} \text{adj}^\top(\mathbf{M})_{ik}}{\partial \mathbf{M}_{ij}} = \sum_k \frac{\partial \mathbf{M}_{ik} \text{adj}^\top(\mathbf{M})_{ik}}{\partial \mathbf{M}_{ij}}$$

Applying the product rule we can derive

$$\begin{aligned} \frac{\partial \det(\mathbf{M})}{\partial \mathbf{M}_{ij}} &= \sum_k \frac{\partial \mathbf{M}_{ik}}{\partial \mathbf{M}_{ij}} \text{adj}^\top(\mathbf{M})_{ik} + \sum_k \mathbf{M}_{ik} \frac{\partial \text{adj}^\top(\mathbf{M})_{ik}}{\partial \mathbf{M}_{ij}} \\ &= \sum_k \frac{\partial \mathbf{M}_{ik}}{\partial \mathbf{M}_{ij}} \text{adj}^\top(\mathbf{M})_{ik}. \end{aligned}$$

## A sketch of Jacobi's formula (4/4)

Note that

$$\frac{\partial \mathbf{M}_{ik}}{\partial \mathbf{M}_{ij}} = \delta_{jk}$$

where  $\delta_{jk} = 1$  if  $j = k$  and 0 otherwise. So,

## A sketch of Jacobi's formula (4/4)

Note that

$$\frac{\partial \mathbf{M}_{ik}}{\partial \mathbf{M}_{ij}} = \delta_{jk}$$

where  $\delta_{jk} = 1$  if  $j = k$  and 0 otherwise. So,

$$\frac{\partial \det(\mathbf{M})}{\partial \mathbf{M}_{ij}} = \sum_k \delta_{jk} \operatorname{adj}^\top(\mathbf{M})_{ik} = \operatorname{adj}^\top(\mathbf{M})_{ij}.$$

Thus,

$$\mathrm{d} \det(\mathbf{M}) = \sum_i \sum_j \operatorname{adj}^\top(\mathbf{M})_{ij} \mathrm{d} \mathbf{M}_{ij} = \operatorname{tr}(\operatorname{adj}(\mathbf{M}) \mathrm{d} \mathbf{M}).$$

# Discussions