Mathematics for Machine Learning

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

Joseph Chuang-Chieh Lin

Department of Computer Science & Information Engineering,
Tamkang University

Fall 2023

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

Outline

Our Focus

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

Vector of Functions

Given

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

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

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

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 $[f_1, \ldots, f_m]^\top$, such that $f_i : \mathbb{R}^n \mapsto \mathbb{R}$.

Therefore,

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

So,

$$
\begin{aligned}
\frac{\mathrm{d}\mathbf{f}}{\mathrm{d}\mathbf{x}} &= \begin{bmatrix} \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_n} \end{bmatrix} \\
&= \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\mathbf{x})}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n}.
\end{aligned}
$$

So,

$$\begin{aligned}
\frac{\mathrm{d}\mathbf{f}}{\mathrm{d}\mathbf{x}} &= \begin{bmatrix} \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_n} \end{bmatrix} \\
&= \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\mathbf{x})}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n}.
\end{aligned}$$

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

So,

$$\frac{\mathrm{d}\mathbf{f}}{\mathrm{d}\mathbf{x}} = \begin{bmatrix} \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\mathbf{x})}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n}.$$

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

⋆ Denote by $\boldsymbol{J} = \nabla_{\mathbf{x}} \mathbf{f} = \frac{\mathrm{d}\mathbf{f}(\mathbf{x})}{\mathrm{d}\mathbf{x}}$

- $J(i, j) = \frac{\partial f_i}{\partial x_j}$.

Example

Derivative of a Polynomial

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

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

Note:

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

Example

Derivative of a Polynomial

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

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

Note:

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

Example

Derivative of a Polynomial

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

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

Note:

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

Example

Derivative of a Polynomial

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

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

Note:

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

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

Consider the linear model

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

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

$$\begin{aligned} L(\boldsymbol{e}) &:= \|\boldsymbol{e}\|^2. \\ \boldsymbol{e}(\boldsymbol{\theta}) &:= \mathbf{y} - \boldsymbol{\Phi\theta}. \end{aligned}$$

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

Example (2/3)

Note that

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

Example (2/3)

Note that

- $\frac{\partial L}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{1 \times D} \quad (\because L : \mathbb{R}^D \mapsto \mathbb{R})$.
- $\frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{N \times D}$

Example (2/3)

Note that

- $\frac{\partial L}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{1\times D} \quad (\because L : \mathbb{R}^D \mapsto \mathbb{R})$.
- $\frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{N\times D} \quad (\because \boldsymbol{e} : \mathbb{R}^D \mapsto \mathbb{R}^N)$.

Example (2/3)

Note that

- $\frac{\partial L}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{1 \times D} \quad (\because L : \mathbb{R}^D \mapsto \mathbb{R})$.
- $\frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{N \times D} \quad (\because \boldsymbol{e} : \mathbb{R}^D \mapsto \mathbb{R}^N)$.
- $\frac{\partial L}{\partial \boldsymbol{e}} \in \mathbb{R}^{1 \times N}$

Example (2/3)

Note that

- $\frac{\partial L}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{1\times D} \quad (\because L : \mathbb{R}^D \mapsto \mathbb{R})$.
- $\frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{N\times D} \quad (\because \boldsymbol{e} : \mathbb{R}^D \mapsto \mathbb{R}^N)$.
- $\frac{\partial L}{\partial \boldsymbol{e}} \in \mathbb{R}^{1\times N} \quad (\because L : \mathbb{R}^N \mapsto \mathbb{R})$.

Example (2/3)

Note that

- $\frac{\partial L}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{1 \times D} \quad (\because L : \mathbb{R}^D \mapsto \mathbb{R})$.
- $\frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{N \times D} \quad (\because \boldsymbol{e} : \mathbb{R}^D \mapsto \mathbb{R}^N)$.
- $\frac{\partial L}{\partial \boldsymbol{e}} \in \mathbb{R}^{1 \times N} \quad (\because L : \mathbb{R}^N \mapsto \mathbb{R})$.
- $\frac{\partial L}{\partial \boldsymbol{\theta}} = \frac{\partial L}{\partial \boldsymbol{e}} \frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\theta}}$ (chain rule).

Example (2/3)

Note that

- $\frac{\partial L}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{1\times D} \quad (\because L : \mathbb{R}^D \mapsto \mathbb{R})$.
- $\frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{N\times D} \quad (\because \boldsymbol{e} : \mathbb{R}^D \mapsto \mathbb{R}^N)$.
- $\frac{\partial L}{\partial \boldsymbol{e}} \in \mathbb{R}^{1\times N} \quad (\because L : \mathbb{R}^N \mapsto \mathbb{R})$.
- $\frac{\partial L}{\partial \boldsymbol{\theta}} = \frac{\partial L}{\partial \boldsymbol{e}} \frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\theta}}$ (chain rule).
- The dth element: $\frac{\partial L}{\partial \boldsymbol{\theta}}[1, d] = \sum_{i=1}^{N} \frac{\partial L}{\partial \boldsymbol{e}}[i] \frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\theta}}[i, d]$.

Example (2/3)

Note that

- $\frac{\partial L}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{1\times D} \quad (\because L : \mathbb{R}^D \mapsto \mathbb{R})$.
- $\frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{N\times D} \quad (\because \boldsymbol{e} : \mathbb{R}^D \mapsto \mathbb{R}^N)$.
- $\frac{\partial L}{\partial \boldsymbol{e}} \in \mathbb{R}^{1\times N} \quad (\because L : \mathbb{R}^N \mapsto \mathbb{R})$.
- $\frac{\partial L}{\partial \boldsymbol{\theta}} = \frac{\partial L}{\partial \boldsymbol{e}} \frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\theta}}$ (chain rule).
- The dth element: $\frac{\partial L}{\partial \boldsymbol{\theta}}[1, d] = \sum_{i=1}^{N} \frac{\partial L}{\partial \boldsymbol{e}}[i] \frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\theta}}[i, d]$.
- $L = \|\boldsymbol{e}\|^2 = \boldsymbol{e}^\top \boldsymbol{e}$ and $\frac{\partial L}{\partial \boldsymbol{e}} = 2\boldsymbol{e}^\top \in \mathbb{R}^{1\times N}$.

Example (2/3)

Note that

- $\frac{\partial L}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{1\times D} \quad (\because L : \mathbb{R}^D \mapsto \mathbb{R})$.
- $\frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{N\times D} \quad (\because \boldsymbol{e} : \mathbb{R}^D \mapsto \mathbb{R}^N)$.
- $\frac{\partial L}{\partial \boldsymbol{e}} \in \mathbb{R}^{1\times N} \quad (\because L : \mathbb{R}^N \mapsto \mathbb{R})$.
- $\frac{\partial L}{\partial \boldsymbol{\theta}} = \frac{\partial L}{\partial \boldsymbol{e}} \frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\theta}}$ (chain rule).
- The dth element: $\frac{\partial L}{\partial \boldsymbol{\theta}}[1, d] = \sum_{i=1}^{N} \frac{\partial L}{\partial \boldsymbol{e}}[i] \frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\theta}}[i, d]$.
- $L = \|\boldsymbol{e}\|^2 = \boldsymbol{e}^\top \boldsymbol{e}$ and $\frac{\partial L}{\partial \boldsymbol{e}} = 2\boldsymbol{e}^\top \in \mathbb{R}^{1\times N}$.
- $\frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\theta}} = -\boldsymbol{\Phi} \in \mathbb{R}^{N\times D}$.

Example (3/3)

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

Example (3/3)

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

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

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

Example (3/3)

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

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

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

- It becomes impractical for deep function compositions.

Outline

Motivations

- There are scenarios that we need to take gradients of matrices w.r.t. vectors (or other matrices).

  $\Rightarrow$ This results in a multidimensional tensor.
    - Multidimensional array.

- Compute the gradient of an $m \times n$ matrix $\boldsymbol{A}$ w.r.t. a $p \times q$ matrix $\boldsymbol{B}$:
  - The Jacobian $\boldsymbol{J}$ would be $(m \times n) \times (p \times q)$ (4-dimensional tensor).
    - $J_{ijk\ell} = \frac{\partial A_{ij}}{\partial B_{k\ell}}$.

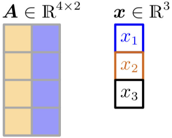
- Matrices $\Leftrightarrow$ linear mappings, so

Motivations

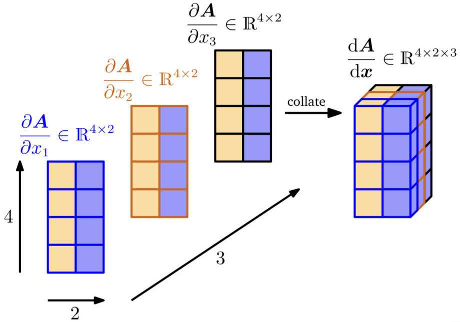
- There are scenarios that we need to take gradients of matrices w.r.t. vectors (or other matrices).
  - $\Rightarrow$ This results in a multidimensional tensor.
    - Multidimensional array.
- Compute the gradient of an $m \times n$ matrix $\boldsymbol{A}$ w.r.t. a $p \times q$ matrix $\boldsymbol{B}$: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

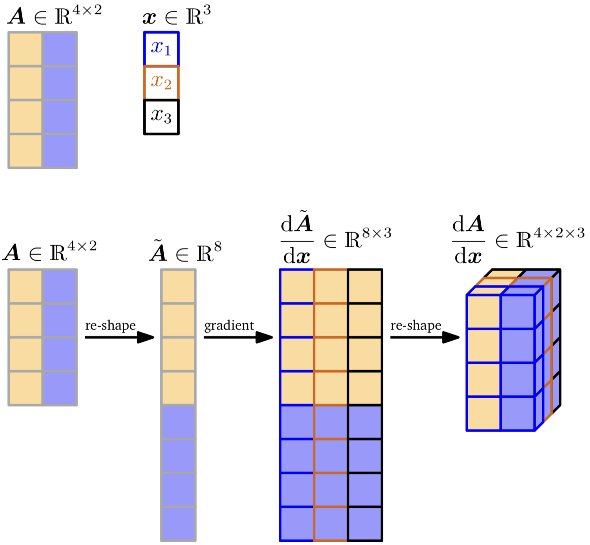
Visualization of Two Approaches for the Isomorphism



Partial derivatives:



Visualization of Two Approaches for the Isomorphism



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

Consider

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

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

Note:

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

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

Consider

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

Goal: Compute the gradient $\dfrac{\mathrm{d}\mathbf{f}}{\mathrm{d}\boldsymbol{A}}$.

Note:

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

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

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

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

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

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

  Hence,

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

  So we can derive

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

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

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

  Hence,

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

  So we can derive

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

Stack the partial derivatives:

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

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

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

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

Goal: Compute the gradient $\dfrac{\mathrm{d}\boldsymbol{K}}{\mathrm{d}\boldsymbol{R}}$.

Note:

- $\dfrac{\mathrm{d}\boldsymbol{K}}{\mathrm{d}\boldsymbol{R}} \in$

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

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

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

Goal: Compute the gradient $\dfrac{\mathrm{d}\boldsymbol{K}}{\mathrm{d}\boldsymbol{R}}$.

Note:

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

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

$\mathbf{r}_i$: the ith column of $\boldsymbol{R}$.

Example (2/2)

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

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

where

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

Example (2/2)

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

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

where

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

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

Useful Identities for Computing Gradients (1/2)

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

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

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

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

Useful Identities for Computing Gradients (1/2)

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

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

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

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

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

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

Useful Identities for Computing Gradients (2/2)

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

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

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

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

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

Discussions