Matrix Differentiation

CS5240 Theoretical Foundations in Multimedia

Leow Wee Kheng

Department of Computer Science
School of Computing
National University of Singapore

Linear Fitting Revisited

Linear fitting solves this problem:

> *Given n data points $\mathbf{p}_i = [x_{i1} \cdots x_{im}]^\top$, $i = 1, \ldots, n$, and their corresponding values v_i, find a linear function f that minimizes the error*

$$E = \sum_{i=1}^{n} (f(\mathbf{p}_i) - v_i)^2 . \tag{1}$$

The linear function $f(\mathbf{p}_i)$ has the form

$$f(\mathbf{p}) = f(x_1, \ldots, x_m) = a_1 x_1 + \cdots + a_m x_m + a_{m+1}. \tag{2}$$

The data points are organized into a matrix equation

$$\mathbf{D}\,\mathbf{a} = \mathbf{v}, \tag{3}$$

where

$$\mathbf{D} = \begin{bmatrix} x_{11} & \cdots & x_{1m} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ x_{n1} & \cdots & x_{nm} & 1 \end{bmatrix}, \quad \mathbf{a} = \begin{bmatrix} a_1 \\ \vdots \\ a_m \\ a_{m+1} \end{bmatrix}, \quad \text{and } \mathbf{v} = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}. \tag{4}$$

The solution of Eq. 3 is

$$\mathbf{a} = (\mathbf{D}^\top \mathbf{D})^{-1} \mathbf{D}^\top \mathbf{v}. \tag{5}$$

Denote each row of $\mathbf{D}$ as $\mathbf{d}_i^\top$. Then,

$$E = \sum_{i=1}^{n} (\mathbf{d}_i^\top \mathbf{a} - v_i)^2 = \|\mathbf{D}\,\mathbf{a} - \mathbf{v}\|^2. \tag{6}$$

So, linear least squares problem can be described very compactly as

$$\min_{\mathbf{a}} \|\mathbf{D}\,\mathbf{a} - \mathbf{v}\|^2. \tag{7}$$

To find the parameters $\mathbf{a}$ that minimizes error E, need to **differentiate** E with respect to $\mathbf{a}$ and set it to zero:

$$\frac{dE}{d\mathbf{a}} = 0. \tag{8}$$

How to do this differentiation?

The obvious (but hard) way:

$$E = \sum_{i=1}^{n} \left(\sum_{j=1}^{m} a_j x_{ij} + a_{m+1} - v_i \right)^2 . \tag{9}$$

Expand equation explicitly giving

$$\frac{\partial E}{\partial a_k} = \begin{cases} 2\sum_{i=1}^{n} \left(\sum_{j=1}^{m} a_j x_{ij} + a_{m+1} - v_i \right) x_{ik}, & \text{for } k \neq m+1 \\ 2\sum_{i=1}^{n} \left(\sum_{j=1}^{m} a_j x_{ij} + a_{m+1} - v_i \right), & \text{for } k = m+1 \end{cases}$$

Then, set $\partial E / \partial a_k = 0$ and solve for a_k.
This is slow, tedious and error prone, especially when E is complex.

Which one do you like to be?

At least like these?

Matrix Derivatives

There are 6 common types of matrix derivatives:

Type	Scalar	Vector	Matrix
Scalar	$\dfrac{\partial y}{\partial x}$	$\dfrac{\partial \mathbf{y}}{\partial x}$	$\dfrac{\partial \mathbf{Y}}{\partial x}$
Vector	$\dfrac{\partial y}{\partial \mathbf{x}}$	$\dfrac{\partial \mathbf{y}}{\partial \mathbf{x}}$	
Matrix	$\dfrac{\partial y}{\partial \mathbf{X}}$		

- Vectors $\mathbf{x}$ and $\mathbf{y}$ are 1-column matrices.
- Regard scalars x, y as 1×1 matrices $[\ x\ ]$, $[\ y\ ]$.
- The 3 remaining cases involve tensors. Omit.

Consider, for example, $\partial \mathbf{y}/\partial \mathbf{x}$.

Suppose $\mathbf{y} = [\ y_1 \ \cdots \ y_m \ ]^\top$ and $\mathbf{x} = [\ x_1 \ \cdots \ x_n \ ]^\top$.
Then, $\partial \mathbf{y}/\partial \mathbf{x}$ will have $m \times n$ components of

$$\frac{\partial y_i}{\partial x_j}, \text{ for } i = 1, \ldots, m, \ j = 1, \ldots, n.$$

How to arrange these $m \times n$ components nicely?

There are two existing layout conventions.

Derivatives by Scalar

Numerator Layout

$$\frac{\partial y}{\partial x}$$

$$\frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} \\ \vdots \\ \frac{\partial y_m}{\partial x} \end{bmatrix}$$

$$\frac{\partial \mathbf{Y}}{\partial x} = \begin{bmatrix} \frac{\partial y_{11}}{\partial x} & \cdots & \frac{\partial y_{1n}}{\partial x} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_{m1}}{\partial x} & \cdots & \frac{\partial y_{mn}}{\partial x} \end{bmatrix}$$

Denominator Layout

$$\frac{\partial y}{\partial x}$$

$$\frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} & \cdots & \frac{\partial y_m}{\partial x} \end{bmatrix} \equiv \frac{\partial \mathbf{y}^\top}{\partial x}$$

Derivatives by Vector

Numerator Layout

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} & \cdots & \frac{\partial y}{\partial x_n} \end{bmatrix} \equiv \frac{\partial y}{\partial \mathbf{x}^\top}$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

$$\equiv \frac{\partial \mathbf{y}}{\partial \mathbf{x}^\top}$$

Denominator Layout

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \vdots \\ \frac{\partial y}{\partial x_n} \end{bmatrix}$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial x_n} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

$$\equiv \frac{\partial \mathbf{y}^\top}{\partial \mathbf{x}}$$

Derivative by Matrix

Numerator Layout

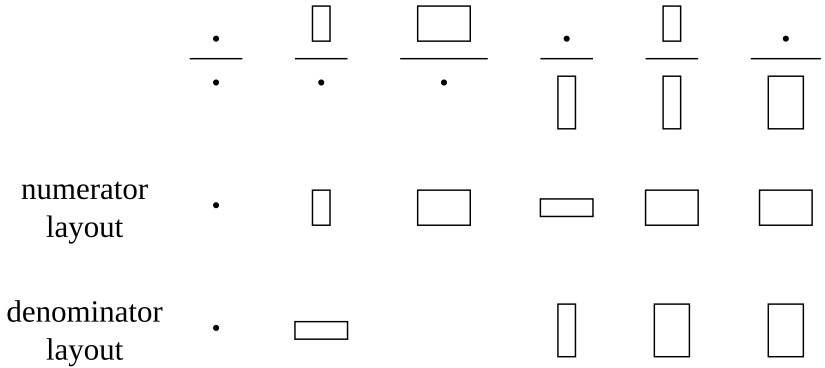
$$\frac{\partial y}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \cdots & \frac{\partial y}{\partial x_{m1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial x_{1n}} & \cdots & \frac{\partial y}{\partial x_{mn}} \end{bmatrix}$$

$$\equiv \frac{\partial y}{\partial \mathbf{X}^\top}$$

Denominator Layout

$$\frac{\partial y}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \cdots & \frac{\partial y}{\partial x_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial x_{m1}} & \cdots & \frac{\partial y}{\partial x_{mn}} \end{bmatrix}$$

Pictorial Representation



Caution!

- Most books and papers don't state which convention they use.
- Gentle [3] uses the explicit notation, e.g., $\partial y / \partial \mathbf{x}^\top$.

$$\frac{\partial y}{\partial \mathbf{x}^\top} = \begin{bmatrix} \dfrac{\partial y}{\partial x_1} & \cdots & \dfrac{\partial y}{\partial x_n} \end{bmatrix} \qquad \frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \dfrac{\partial y}{\partial x_1} \\ \vdots \\ \dfrac{\partial y}{\partial x_n} \end{bmatrix}$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}^\top} = \begin{bmatrix} \dfrac{\partial y_1}{\partial x_1} & \cdots & \dfrac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial y_m}{\partial x_1} & \cdots & \dfrac{\partial y_m}{\partial x_n} \end{bmatrix} \qquad \frac{\partial \mathbf{y}^\top}{\partial \mathbf{x}} = \begin{bmatrix} \dfrac{\partial y_1}{\partial x_1} & \cdots & \dfrac{\partial y_m}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial y_1}{\partial x_n} & \cdots & \dfrac{\partial y_m}{\partial x_n} \end{bmatrix}$$

- It is best not to mix the two conventions in your equations.
- We adopt **numerator layout** convention.

Commonly Used Derivatives

We omit differentiations by scalars which follow the layout definitions.

Differentiation by Vectors

Here, scalar a, vector $\mathbf{a}$ and matrix $\mathbf{A}$ are not functions of $\mathbf{x}$.

(C1) $$\frac{\partial a}{\partial \mathbf{x}} = \mathbf{0}^\top \quad \text{(row matrix)}$$

(C2) $$\frac{\partial \mathbf{a}}{\partial \mathbf{x}} = \mathbf{0} \quad \text{(matrix)}$$

(C3) $$\frac{\partial \mathbf{x}}{\partial \mathbf{x}} = \mathbf{I} \quad \text{(matrix)}$$

(C4) $$\frac{\partial \mathbf{a}^\top \mathbf{x}}{\partial \mathbf{x}} = \frac{\partial \mathbf{x}^\top \mathbf{a}}{\partial \mathbf{x}} = \mathbf{a}^\top$$

(C5) $$\frac{\partial \mathbf{x}^\top \mathbf{x}}{\partial \mathbf{x}} = 2\,\mathbf{x}^\top$$

(C6) $$\frac{\partial (\mathbf{x}^\top \mathbf{a})^2}{\partial \mathbf{x}} = 2\,\mathbf{x}^\top \mathbf{a}\,\mathbf{a}^\top$$

(C7) $$\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}$$

(C8) $$\frac{\partial \mathbf{x}^\top \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{x}^\top (\mathbf{A} + \mathbf{A}^\top)$$

Differentiatiton by Matrix

Here, scalar a and vectors $\mathbf{a}$ and $\mathbf{b}$ are not functions of $\mathbf{X}$.

(C11) $$\frac{\partial a}{\partial \mathbf{X}} = \mathbf{0} \quad \text{(matrix)}$$

(C12) $$\frac{\partial \mathbf{a}^\top \mathbf{X} \mathbf{b}}{\partial \mathbf{X}} = \mathbf{b}\mathbf{a}^\top$$

(C13) $$\frac{\partial \mathbf{a}^\top \mathbf{X}^\top \mathbf{b}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{b}^\top$$

Refer to Appendix for other derivative rules.

Trace

The **trace** of a square matrix $\mathbf{A}$ is the sum of its diagonal elements a_{ii}:

$$\operatorname{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}. \tag{10}$$

Properties

- $\operatorname{tr}(\mathbf{A}+\mathbf{B}) = \operatorname{tr}(\mathbf{A}) + \operatorname{tr}(\mathbf{B})$
- $\operatorname{tr}(c\mathbf{A}) = c\operatorname{tr}(\mathbf{A})$
- $\operatorname{tr}(\mathbf{A}) = \operatorname{tr}(\mathbf{A}^\top)$
- $\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{BA}) = \sum_i \sum_j a_{ij} b_{ji}$
- $\operatorname{tr}(\mathbf{ABCD}) = \operatorname{tr}(\mathbf{BCDA}) = \operatorname{tr}(\mathbf{CDAB}) = \operatorname{tr}(\mathbf{DABC})$

Properties

- $\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i$, where λ_i are the eigenvalues of $\mathbf{A}$.
- $\mathrm{tr}(\mathbf{A}^k) = \sum_{i=1}^{n} \lambda_i^k$.

Matrix trace is useful for expressing quadratic functions.

For example, for matrices $\mathbf{X}$ and $\mathbf{Y}$ (Homework),

$$\sum_i \sum_j x_{ij}^2 = \operatorname{tr}(\mathbf{X}^\top \mathbf{X}). \tag{11}$$

$$\sum_i \sum_j x_{ij} y_{ij} = \operatorname{tr}(\mathbf{X}^\top \mathbf{Y}). \tag{12}$$

So, differentiation of quadratic function of matrix can be performed through matrix trace.

Derivatives of Trace

For variable matrix **X** and constant matrices **A**, **B**, **C**,

(DT1) $$\frac{\partial\,\mathrm{tr}(\mathbf{X})}{\partial\mathbf{X}} = \mathbf{I}$$

(DT2) $$\frac{\partial\,\mathrm{tr}(\mathbf{X}^k)}{\partial\mathbf{X}} = k\mathbf{X}^{k-1}$$

(DT3) $$\frac{\partial\,\mathrm{tr}(\mathbf{A}\mathbf{X})}{\partial\mathbf{X}} = \frac{\partial\,\mathrm{tr}(\mathbf{X}\mathbf{A})}{\partial\mathbf{X}} = \mathbf{A}$$

(DT4) $$\frac{\partial\,\mathrm{tr}(\mathbf{A}\mathbf{X}^\top)}{\partial\mathbf{X}} = \frac{\partial\,\mathrm{tr}(\mathbf{X}^\top\mathbf{A})}{\partial\mathbf{X}} = \mathbf{A}^\top$$

(DT5) $$\frac{\partial\,\mathrm{tr}(\mathbf{X}^\top\mathbf{A}\mathbf{X})}{\partial\mathbf{X}} = \mathbf{X}^\top(\mathbf{A} + \mathbf{A}^\top)$$

(DT6) $$\frac{\partial \operatorname{tr}(\mathbf{X}^{-1}\mathbf{A})}{\partial \mathbf{X}} = -\mathbf{X}^{-1}\mathbf{A}\mathbf{X}^{-1}$$

(DT7) $$\frac{\partial \operatorname{tr}(\mathbf{AXB})}{\partial \mathbf{X}} = \frac{\partial \operatorname{tr}(\mathbf{BAX})}{\partial \mathbf{X}} = \mathbf{BA}$$

(DT8) $$\frac{\partial \operatorname{tr}(\mathbf{AXBX}^\top\mathbf{C})}{\partial \mathbf{X}} = \mathbf{BX}^\top\mathbf{CA} + \mathbf{B}^\top\mathbf{X}^\top\mathbf{A}^\top\mathbf{C}^\top$$

Linear Fitting Revisited

Now, let us show that the solution

$$\mathbf{a} = (\mathbf{D}^\top\mathbf{D})^{-1}\mathbf{D}^\top\mathbf{v}.$$

minimizes error E

$$E = \sum_{i=1}^{n}(\mathbf{d}_i^\top\mathbf{a} - v_i)^2 = \|\mathbf{D}\mathbf{a} - \mathbf{v}\|^2.$$

Proof:

$$\begin{aligned} E = \|\mathbf{D}\mathbf{a} - \mathbf{v}\|^2 \quad &= (\mathbf{D}\mathbf{a} - \mathbf{v})^\top(\mathbf{D}\mathbf{a} - \mathbf{v}) \\ &= (\mathbf{a}^\top\mathbf{D}^\top - \mathbf{v}^\top)(\mathbf{D}\mathbf{a} - \mathbf{v}) \\ &= \mathbf{a}^\top\mathbf{D}^\top\mathbf{D}\mathbf{a} - \mathbf{a}^\top\mathbf{D}^\top\mathbf{v} - \mathbf{v}^\top\mathbf{D}\mathbf{a} + \mathbf{v}^\top\mathbf{v}. \end{aligned}$$

$$E = \mathbf{a}^\top(\mathbf{D}^\top\mathbf{D})\mathbf{a} - \mathbf{a}^\top(\mathbf{D}^\top\mathbf{v}) - (\mathbf{v}^\top\mathbf{D})\mathbf{a} + \mathbf{v}^\top\mathbf{v}.$$

Apply (C8), (C4), (C4) and (C1) to the four terms.

$$\begin{aligned}\frac{dE}{d\mathbf{a}} &= \mathbf{a}^\top(\mathbf{D}^\top\mathbf{D} + \mathbf{D}^\top\mathbf{D}) - (\mathbf{D}^\top\mathbf{v})^\top - \mathbf{v}^\top\mathbf{D} + \mathbf{0}^\top \\ &= 2\,\mathbf{a}^\top\mathbf{D}^\top\mathbf{D} - 2\,\mathbf{v}^\top\mathbf{D}.\end{aligned}$$

Set $dE/d\mathbf{a} = \mathbf{0}^\top$ and obtain

$$\begin{aligned}2\,\mathbf{a}^\top\mathbf{D}^\top\mathbf{D} - 2\,\mathbf{v}^\top\mathbf{D} &= \mathbf{0}^\top \\ \mathbf{a}^\top\mathbf{D}^\top\mathbf{D} &= \mathbf{v}^\top\mathbf{D}\end{aligned}$$

Transpose both sides of the equation and get

$$\begin{aligned}\mathbf{D}^\top\mathbf{D}\,\mathbf{a} &= \mathbf{D}^\top\mathbf{v} \\ \mathbf{a} &= (\mathbf{D}^\top\mathbf{D})^{-1}\mathbf{D}^\top\mathbf{v}. \quad \square\end{aligned}$$

The error $\|\mathbf{D}\,\mathbf{a} - \mathbf{v}\|^2$ is square of **vector norm**:

$$\|\mathbf{u}\|^2 = \sum_i u_i^2. \tag{13}$$

Pseudo-inverse solution $(\mathbf{D}^\top\mathbf{D})^{-1}\mathbf{D}^\top\mathbf{v}$ minimizes this error.

Recall that, in the multivariate case, we have

$$\mathbf{DA} = \mathbf{V},$$

where $\mathbf{A}$ and $\mathbf{V}$ are matrices of known values.

In this case, the error $\|\mathbf{DA} - \mathbf{V}\|_F^2$ is square of **Frobenius norm**:

$$\|\mathbf{U}\|_F^2 = \sum_i \sum_j u_{ij}^2. \tag{14}$$

Pseudo-inverse solution $(\mathbf{D}^\top\mathbf{D})^{-1}\mathbf{D}^\top\mathbf{V}$ minimizes this error.

Summary

- There are 6 common derivatives of matrices (vectors and scalars).
- There are 2 layout conventions:
  numerator layout vs. denominator layout.
- We adopt numerator layout convention.
- Do not mix the two conventions in your equations.
- Use matrix differentiation to prove that pseudo-inverse solution minimizes sum squared error.

Use the right tool
become lightning fast!

Homework

1. Given a constant matrix $\mathbf{A}$ and a variable vector $\mathbf{x}$, what is the result of $d\,\|\mathbf{A}\mathbf{x}\|^2/d\mathbf{x}$?
2. Given constant matrices $\mathbf{A}$ and $\mathbf{B}$ and variable vector $\mathbf{x}$, what is the result of $d\,\|\mathbf{A}\mathbf{B}\mathbf{x}\|^2/d\mathbf{x}$?
3. Given constant vectors $\mathbf{a}$ and $\mathbf{b}$ and variable vector $\mathbf{x}$, what is the result of $d\,\mathbf{a}^\top(\mathbf{x}-\mathbf{b})/d\mathbf{x}$?
4. Given constant matrix $\mathbf{A}$, constant vector $\mathbf{b}$ and variable vector $\mathbf{x}$, what is the result of $d\mathbf{A}(\mathbf{x}+\mathbf{b})/d\mathbf{x}$?
5. Given constant matrix $\mathbf{A}$, constant vector $\mathbf{b}$ and variable vector $\mathbf{x}$, what is the result of $d\,\|\mathbf{A}(\mathbf{x}-\mathbf{b})\|^2/d\mathbf{x}$?
6. Show that, for matrices $\mathbf{X}$ and $\mathbf{Y}$, $\displaystyle\sum_i\sum_j x_{ij}y_{ij} = \operatorname{tr}(\mathbf{X}^\top\mathbf{Y})$.

Appendix

Derivatives of Scalar by Scalar

$$\text{(SS1)} \qquad \frac{\partial(u+v)}{\partial x} = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial x}$$

$$\text{(SS2)} \qquad \frac{\partial uv}{\partial x} = u\frac{\partial v}{\partial x} + v\frac{\partial u}{\partial x} \qquad \text{(product rule)}$$

$$\text{(SS3)} \qquad \frac{\partial g(u)}{\partial x} = \frac{\partial g(u)}{\partial u}\frac{\partial u}{\partial x} \qquad \text{(chain rule)}$$

$$\text{(SS4)} \qquad \frac{\partial f(g(u))}{\partial x} = \frac{\partial f(g)}{\partial g}\frac{\partial g(u)}{\partial u}\frac{\partial u}{\partial x} \qquad \text{(chain rule)}$$

Derivatives of Vector by Scalar

(VS1) $$\frac{\partial a\mathbf{u}}{\partial x} = a\frac{\partial \mathbf{u}}{\partial x}$$

where a is not a function of x.

(VS2) $$\frac{\partial \mathbf{A}\mathbf{u}}{\partial x} = \mathbf{A}\frac{\partial \mathbf{u}}{\partial x}$$

where $\mathbf{A}$ is not a function of x.

(VS3) $$\frac{\partial \mathbf{u}^\top}{\partial x} = \left(\frac{\partial \mathbf{u}}{\partial x}\right)^\top$$

(VS4) $$\frac{\partial (\mathbf{u}+\mathbf{v})}{\partial x} = \frac{\partial \mathbf{u}}{\partial x} + \frac{\partial \mathbf{v}}{\partial x}$$

(VS5) $$\frac{\partial \mathbf{g}(\mathbf{u})}{\partial x} = \frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}} \frac{\partial \mathbf{u}}{\partial x} \qquad \text{(chain rule)}$$

with consistent matrix layout.

(VS6) $$\frac{\partial \mathbf{f}(\mathbf{g}(\mathbf{u}))}{\partial x} = \frac{\partial \mathbf{f}(\mathbf{g})}{\partial \mathbf{g}} \frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}} \frac{\partial \mathbf{u}}{\partial x} \qquad \text{(chain rule)}$$

with consistent matrix layout.

Derivatives of Matrix by Scalar

(MS1) $$\frac{\partial a\mathbf{U}}{\partial x} = a\frac{\partial \mathbf{U}}{\partial x}$$

where a is not a function of x.

(MS2) $$\frac{\partial \mathbf{AUB}}{\partial x} = \mathbf{A}\frac{\partial \mathbf{U}}{\partial x}\mathbf{B}$$

where $\mathbf{A}$ and $\mathbf{B}$ are not functions of x.

(MS3) $$\frac{\partial (\mathbf{U}+\mathbf{V})}{\partial x} = \frac{\partial \mathbf{U}}{\partial x} + \frac{\partial \mathbf{V}}{\partial x}$$

(MS4) $$\frac{\partial \mathbf{UV}}{\partial x} = \mathbf{U}\frac{\partial \mathbf{V}}{\partial x} + \frac{\partial \mathbf{U}}{\partial x}\mathbf{V} \qquad \text{(product rule)}$$

Derivatives of Scalar by Vector

(SV1) $$\frac{\partial a u}{\partial \mathbf{x}} = a \frac{\partial u}{\partial \mathbf{x}}$$

where a is not a function of $\mathbf{x}$.

(SV2) $$\frac{\partial (u + v)}{\partial \mathbf{x}} = \frac{\partial u}{\partial \mathbf{x}} + \frac{\partial v}{\partial \mathbf{x}}$$

(SV3) $$\frac{\partial u v}{\partial \mathbf{x}} = u \frac{\partial v}{\partial \mathbf{x}} + v \frac{\partial u}{\partial \mathbf{x}}$$ (product rule)

(SV4) $$\frac{\partial g(u)}{\partial \mathbf{x}} = \frac{\partial g(u)}{\partial u} \frac{\partial u}{\partial \mathbf{x}}$$ (chain rule)

(SV5) $$\frac{\partial f(g(u))}{\partial \mathbf{x}} = \frac{\partial f(g)}{\partial g} \frac{\partial g(u)}{\partial u} \frac{\partial u}{\partial \mathbf{x}}$$ (chain rule)

(SV6) $$\frac{\partial \mathbf{u}^\top \mathbf{v}}{\partial \mathbf{x}} = \mathbf{u}^\top \frac{\partial \mathbf{v}}{\partial \mathbf{x}} + \mathbf{v}^\top \frac{\partial \mathbf{u}}{\partial \mathbf{x}} \quad \text{(product rule)}$$

where $\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$ and $\frac{\partial \mathbf{v}}{\partial \mathbf{x}}$ are in numerator layout.

(SV7) $$\frac{\partial \mathbf{u}^\top \mathbf{A} \mathbf{v}}{\partial \mathbf{x}} = \mathbf{u}^\top \mathbf{A} \frac{\partial \mathbf{v}}{\partial \mathbf{x}} + \mathbf{v}^\top \mathbf{A}^\top \frac{\partial \mathbf{u}}{\partial \mathbf{x}} \quad \text{(product rule)}$$

where $\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$ and $\frac{\partial \mathbf{v}}{\partial \mathbf{x}}$ are in numerator layout,

and $\mathbf{A}$ is not a function of $\mathbf{x}$.

Derivatives of Scalar by Matrix

(SM1) $$\frac{\partial a u}{\partial \mathbf{X}} = a \frac{\partial u}{\partial \mathbf{X}}$$

where a is not a function of $\mathbf{X}$.

(SM2) $$\frac{\partial (u+v)}{\partial \mathbf{X}} = \frac{\partial u}{\partial \mathbf{X}} + \frac{\partial v}{\partial \mathbf{X}}$$

(SM3) $$\frac{\partial uv}{\partial \mathbf{X}} = u \frac{\partial v}{\partial \mathbf{X}} + v \frac{\partial u}{\partial \mathbf{X}}$$ (product rule)

(SM4) $$\frac{\partial g(u)}{\partial \mathbf{X}} = \frac{\partial g(u)}{\partial u} \frac{\partial u}{\partial \mathbf{X}}$$ (chain rule)

(SM5) $$\frac{\partial f(g(u))}{\partial \mathbf{X}} = \frac{\partial f(g)}{\partial g} \frac{\partial g(u)}{\partial u} \frac{\partial u}{\partial \mathbf{X}}$$ (chain rule)

Derivatives of Vector by Vector

(VV1) $$\frac{\partial a\mathbf{u}}{\partial \mathbf{x}} = a\frac{\partial \mathbf{u}}{\partial \mathbf{x}} + \mathbf{u}\frac{\partial a}{\partial \mathbf{x}}$$ (product rule)

(VV2) $$\frac{\partial \mathbf{A}\mathbf{u}}{\partial \mathbf{x}} = \mathbf{A}\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$$

where $\mathbf{A}$ is not a function of $\mathbf{x}$.

(VV3) $$\frac{\partial (\mathbf{u}+\mathbf{v})}{\partial \mathbf{x}} = \frac{\partial \mathbf{u}}{\partial \mathbf{x}} + \frac{\partial \mathbf{v}}{\partial \mathbf{x}}$$

(VV4) $$\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{x}} = \frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}}\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$$ (chain rule)

(VV5) $$\frac{\partial \mathbf{f}(\mathbf{g}(\mathbf{u}))}{\partial \mathbf{x}} = \frac{\partial \mathbf{f}(\mathbf{g})}{\partial \mathbf{g}}\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}}\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$$ (chain rule)

Notes on Denominator Layout

In some cases, the results of denominator layout are the transpose of those of numerator layout. Moreover, the chain rule for denominator layout goes from right to left instead of left to right.

	Numerator Layout	Denominator Layout
(C7)	$\dfrac{\partial \mathbf{a}^\top \mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}^\top$	$\dfrac{\partial \mathbf{a}^\top \mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}$
(C10)	$\dfrac{\partial \mathbf{x}^\top \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = \mathbf{x}^\top (\mathbf{A} + \mathbf{A}^\top)$	$\dfrac{\partial \mathbf{x}^\top \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}$
(VV5)	$\dfrac{\partial \mathbf{f}(\mathbf{g}(\mathbf{u}))}{\partial \mathbf{x}} = \dfrac{\partial \mathbf{f}(\mathbf{g})}{\partial \mathbf{g}} \dfrac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}} \dfrac{\partial \mathbf{u}}{\partial \mathbf{x}}$	$\dfrac{\partial \mathbf{f}(\mathbf{g}(\mathbf{u}))}{\partial \mathbf{x}} = \dfrac{\partial \mathbf{u}}{\partial \mathbf{x}} \dfrac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}} \dfrac{\partial \mathbf{f}(\mathbf{g})}{\partial \mathbf{g}}$

References

1. Wikipedia, *Matrix Calculus*.
   `en.wikipedia.org/wiki/Matrix_calculus`
2. T. P. Minka. *Old and New Matrix Algebra Useful for Statistics*, MIT Lab Notes, 2000.
3. J. E. Gentle. *Matrix Algebra: Theory, Computations, and Applications in Statistics*, Springer, 2007.
4. K. B. Petersen and M. S. Pedersen. *The Matrix Cookbook*, 2012.
   `www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf`