# Matrix differentiation

Hung-Hsuan Chen

# What is matrix differentiation

- A specialized notation for doing multivariable differentiation, especially over spaces of matrices

# Notations

- $x$ (lower case) ➔ a scalar
- $\boldsymbol{x}$ (lower case, bold face) ➔ a ***column*** vector
  - $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$
- $\boldsymbol{X}$ (upper case, bold face) ➔ a matrix
  - $\boldsymbol{X} = (x_{ij}) = (\boldsymbol{x}_1^T, \boldsymbol{x}_2^T, \ldots, \boldsymbol{x}_n^T)$

# Example

```
4.6,3.2,1.4,0.2,Iris-setosa
5.3,3.7,1.5,0.2,Iris-setosa
5.0,3.3,1.4,0.2,Iris-setosa
7.0,3.2,4.7,1.4,Iris-versicolor
6.4,3.2,4.5,1.5,Iris-versicolor
6.9,3.1,4.9,1.5,Iris-versicolor
5.5,2.3,4.0,1.3,Iris-versicolor
6.5,2.8,4.6,1.5,Iris-versicolor
```

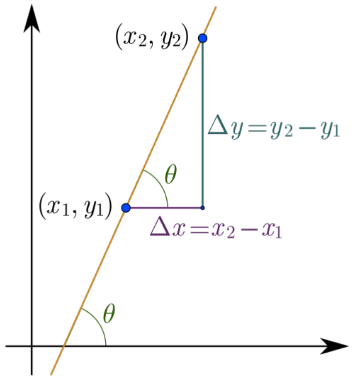
$$\boldsymbol{X} = \begin{bmatrix} 4.6 & \cdots & 0.2 \\ 5.3 & & 0.2 \\ 5.0 & & 0.2 \\ 7.0 & \ddots & 1.4 \\ 6.4 & & 1.5 \\ 6.9 & & 1.5 \\ 5.5 & & 1.3 \\ 6.5 & \cdots & 1.5 \end{bmatrix} \in R^{8\text{x}4}, \boldsymbol{y} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \in R^{8\text{x}1}$$

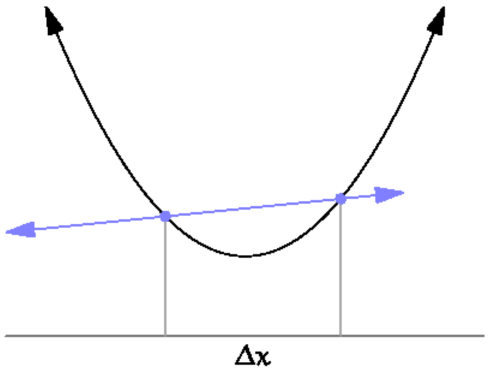
$$x_{1,2} = 3.2, x_{1,3} = 1.4$$

$$\boldsymbol{x}_{1*} = \boldsymbol{x}_1 = \begin{bmatrix} 4.6 \\ 3.2 \\ 1.4 \\ 0.2 \end{bmatrix} \in R^{4\text{x}1}, \quad \boldsymbol{x}_{*1} = \begin{bmatrix} 4.6 \\ 5.3 \\ \vdots \\ 6.5 \end{bmatrix} \in R^{8\text{x}1},$$

# Derivative (scalar)

- An ***infinitesimal*** change in *x* is denoted by *dx*, and the derivative of *y* with respect to *x* is written $dy/dx$ or $f'(x)$
  - If $f(x) = x^k, f'(x) = kx^{k-1}$
  - If $f(x) = e^x, f'(x) = e^x$
  - If $f(x) = a^x, f'(x) = a^x \ln(a)$
  - If $f(x) = \ln x\,, f'(x) = \frac{1}{x}$
  - If $f(x) = \log_a x\,, f'(x) = \frac{1}{x \ln a}$



# Combined function and chain rule

- Given $f$ and $g$ are functions, $a$ and $b$ are real numbers:
  - Sum rule: $(af + bg)' = af' + bg'$
  - Product rule: $(fg)' = f'g + fg'$
  - Quotient rule: $\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}$
- If $f(x) = g(h(x))$
  - Chain rule: $\frac{df(x)}{dx} = \frac{dg(h(x))}{dh(x)} \cdot \frac{dh(x)}{dx}$
    - E.g., $\frac{de^{7x}}{dx} = \frac{de^{7x}}{d7x} \cdot \frac{d7x}{dx} = e^{7x} \cdot 7$

# Partial derivative (scalar)

- A partial derivative of **a function of *several variables*** is its derivative with **respect to one** of those variables, ***with the others held constant***
  - E.g., $f(x_1, x_2) = x_1^2 + 3x_1x_2 + x_2^3$
    - $\frac{\partial f(x_1,x_2)}{\partial x_1} = 2x_1 + 3x_2$
    - $\frac{\partial f(x_1,x_2)}{\partial x_2} = 3x_1 + 3x_2^2$

# Types of matrix derivatives

| Types | Scalar | Vector | Matrix |
|---|---|---|---|
| **Scalar** | $\frac{\partial y}{\partial x}$ | $\frac{\partial \mathbf{y}}{\partial x}$ | $\frac{\partial \mathbf{Y}}{\partial x}$ |
| **Vector** | $\frac{\partial y}{\partial \mathbf{x}}$ | $\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$ | |
| **Matrix** | $\frac{\partial y}{\partial \mathbf{X}}$ | | |

- The three types of derivatives that have not been considered are not as widely considered and a notation is not widely agreed upon

# Layout conventions

- There are mainly two types of layout conventions in matrix calculus
  - Numerator Layout Notation
  - Denominator Layout Notation
- Most books and papers don't state which convention they use
- Even worse, sometimes the two conventions are mixed in the equations
- This confuses the beginners
- We will mostly follow the ***Numerator Layout Notation*** unless otherwise mentioned

# Types of matrix derivatives of different layout conventions

note

| | Scalar *y* | | Vector **y** (size *m*) | | Matrix **Y** (size *m*×*n*) | |
|---|---|---|---|---|---|---|
| | **Notation** | **Type** | **Notation** | **Type** | **Notation** | **Type** |
| **Scalar *x*** | $\frac{\partial y}{\partial x}$ | scalar | $\frac{\partial \mathbf{y}}{\partial x}$ | (numerator layout) size-*m* column vector<br>(denominator layout) size-*m* row vector | $\frac{\partial \mathbf{Y}}{\partial x}$ | (numerator layout) *m*×*n* matrix |
| **Vector x (size *n*)** | $\frac{\partial y}{\partial \mathbf{x}}$ | (numerator layout) size-*n* row vector<br>(denominator layout) size-*n* column vector | $\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$ | (numerator layout) *m*×*n* matrix<br>(denominator layout) *n*×*m* matrix | $\frac{\partial \mathbf{Y}}{\partial \mathbf{x}}$ | |
| **Matrix X (size *p*×*q*)** | $\frac{\partial y}{\partial \mathbf{X}}$ | (numerator layout) *q*×*p* matrix<br>(denominator layout) *p*×*q* matrix | $\frac{\partial \mathbf{y}}{\partial \mathbf{X}}$ | | $\frac{\partial \mathbf{Y}}{\partial \mathbf{X}}$ | |

ex $\vec{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

$\partial y$ ... $= [\partial y$ ...

$\partial \vec{x}$ $\begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \end{bmatrix}$

# Derivative by scalar

- $\dfrac{\partial y}{\partial x}$

- $\dfrac{\partial \boldsymbol{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} \\ \vdots \\ \frac{\partial y_m}{\partial x} \end{bmatrix}$

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

# Derivative by vector

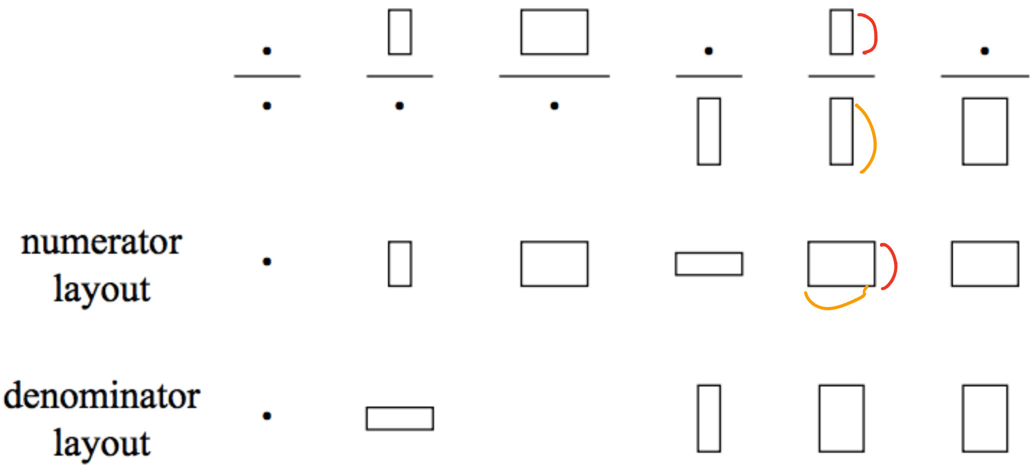
- $\frac{\partial y}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} & \cdots & \frac{\partial y}{\partial x_n} \end{bmatrix}$

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

# Derivative by matrix

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

# Pictorial representation



# Commonly used derivatives

Scalar $a$, vector $\boldsymbol{a}$, and matrix $\boldsymbol{A}$ are not functions of $x$, $\boldsymbol{x}$, and $\boldsymbol{X}$

- $\frac{d\boldsymbol{a}}{dx} = \mathbf{0}$ (column vector)
- $\frac{da}{d\boldsymbol{x}} = \mathbf{0}^T$ (row vector)
- $\frac{da}{d\boldsymbol{X}} = \mathbf{0}^T$ (shape is the same as $\boldsymbol{X}^T$)
- $\frac{d\boldsymbol{a}}{d\boldsymbol{x}} = \mathbf{0}$ (shape is len(**a**) * len(**x**))
- $\frac{d\boldsymbol{x}}{d\boldsymbol{x}} = \boldsymbol{I}$ (shape is len($\boldsymbol{x}$) * len($\boldsymbol{x}$))
- $\frac{d\boldsymbol{a}^T\boldsymbol{x}}{d\boldsymbol{x}} = \frac{d\boldsymbol{x}^T\boldsymbol{a}}{d\boldsymbol{x}} = \boldsymbol{a}^T$
- $\frac{d\boldsymbol{x}^T\boldsymbol{x}}{d\boldsymbol{x}} = 2\boldsymbol{x}^T$ scalar vector
- $\frac{d\boldsymbol{A}\boldsymbol{x}}{d\boldsymbol{x}} = \boldsymbol{A}$

# Exercise

- $\dfrac{d\left(\boldsymbol{x}^T\boldsymbol{a}\right)^2}{d\boldsymbol{x}} = \dfrac{d\left(\boldsymbol{x}^T\boldsymbol{a}\right)^2}{d\left(\boldsymbol{x}^T\boldsymbol{a}\right)} \cdot \dfrac{d\left(\boldsymbol{x}^T\boldsymbol{a}\right)}{d\boldsymbol{x}} = \left(2\boldsymbol{x}^T\boldsymbol{a}\right)\boldsymbol{a}^T$

# Multiple linear regression

$\hat{y} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix}$ $\hat{y}_i = x_{i*}^T \theta$

- $n$: the number of training instances
- $d$: the number of features
- Training instances:
  - $\boldsymbol{X} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,d} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,d} \end{bmatrix}$, $\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$
  - (We assume no coefficient parameter here)

$\sum_{i=1}^{n} (\hat{y}_i - y_i)^2$

- Find $\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_d \end{bmatrix}$ such that $(\widehat{\boldsymbol{y}} - \boldsymbol{y})^T(\widehat{\boldsymbol{y}} - \boldsymbol{y})$ is minimized, where
  - $\widehat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\theta}$

$\hat{y} - y = \begin{bmatrix} \hat{y}_1 - y_1 \\ \vdots \\ \hat{y}_n - y_n \end{bmatrix}$

- The solution is $\boldsymbol{\theta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$

# Solving $\boldsymbol{\theta}$

$$J(\boldsymbol{\theta}) = \cancel{\frac{1}{2}}(\hat{\boldsymbol{y}} - \boldsymbol{y})^T(\hat{\boldsymbol{y}} - \boldsymbol{y}) = \cancel{\frac{1}{2}}(\boldsymbol{X\theta} - \boldsymbol{y})^T(\boldsymbol{X\theta} - \boldsymbol{y})$$

$$\Rightarrow \frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial \cancel{\frac{1}{2}}(\boldsymbol{X\theta}-\boldsymbol{y})^T(\boldsymbol{X\theta}-\boldsymbol{y})}{\partial(\boldsymbol{X\theta}-\boldsymbol{y})} \cdot \frac{\partial(\boldsymbol{X\theta}-\boldsymbol{y})}{\partial \boldsymbol{\theta}} = (\boldsymbol{X\theta} - \boldsymbol{y})^T\boldsymbol{X} := \mathbf{0}$$

$$\Rightarrow (\boldsymbol{\theta}^T\boldsymbol{X}^T - \boldsymbol{y}^T)\boldsymbol{X} = \mathbf{0}$$

$$\Rightarrow \boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{y}^T\boldsymbol{X}$$

$$\Rightarrow (\boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{X})^T = (\boldsymbol{y}^T\boldsymbol{X})^T$$

$$\Rightarrow \boldsymbol{X}^T\boldsymbol{X\theta} = \boldsymbol{X}^T\boldsymbol{y}$$

$$\Rightarrow \boldsymbol{\theta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

- $\frac{d\boldsymbol{x}^T\boldsymbol{x}}{d\boldsymbol{x}} = 2\boldsymbol{x}^T$
- $\frac{d\boldsymbol{Ax}}{d\boldsymbol{x}} = \boldsymbol{A}$

# Summary

- Matrix and vector are compact ways to denote set of variables
- Matrix and vector differentiation may be confusing sometimes, mostly because of inconsistent notations
  - Numerator vs denominator layouts
  - $\frac{d\boldsymbol{a}^T\boldsymbol{x}}{d\boldsymbol{x}} = \boldsymbol{a}^T$ or $\boldsymbol{a}$?
- Sometimes write out the full matrix or vector is helpful

# Quiz

- Given
  - Random variables: $x$ is a scalar, $\boldsymbol{x} \in R^{n\times 1}, \boldsymbol{X} \in R^{n\times m}$
  - Functions of $x, \boldsymbol{x}, \boldsymbol{X}$: $y$ is a scalar, $\boldsymbol{y} \in R^{n\times 1}, \boldsymbol{Y} \in R^{n\times m}$
- What are the shapes of the followings?
  - $\frac{\partial y}{\partial x}; \frac{\partial \boldsymbol{y}}{\partial x}; \frac{\partial \boldsymbol{Y}}{\partial x}$ sca, nx1, nxm
  - $\frac{\partial y}{\partial \boldsymbol{x}}; \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}; \cancel{\frac{\partial \boldsymbol{Y}}{\partial \boldsymbol{x}}}$ 1xn, nxm,
  - $\frac{\partial y}{\partial \boldsymbol{X}}; \cancel{\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{X}}}; \cancel{\frac{\partial \boldsymbol{Y}}{\partial \boldsymbol{X}}}$ mxn,

$$\frac{\partial y}{\partial \vec{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \end{bmatrix}$$

$$\frac{\partial y}{\partial X}$$

$$X = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} \qquad y = f(x_1 \dots x_4)$$

$$\frac{\partial y}{\partial X} = \begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \frac{\partial y}{\partial x_{12}} \\ \frac{\partial y}{\partial x_{21}} & \frac{\partial y}{\partial x_{22}} \end{bmatrix}$$

$$y_1 = x \qquad \vec{y} = \begin{bmatrix} x \\ 2x \end{bmatrix} \qquad \frac{\partial \vec{y}}{\partial x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$y_2 = 2x$$

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

$$y_1 = x_1 + x_2, \quad y_2 = 2x_1 - x_2$$

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

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \qquad \frac{\partial \vec{y}}{\partial \vec{x}} = \begin{bmatrix} 1 & 1 \\ 2 & -1 \end{bmatrix}$$

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\frac{\partial Y}{\partial x} \qquad Y = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} = \begin{bmatrix} x & 2x \\ -x & 3x \end{bmatrix}$$

$$\frac{\partial Y}{\partial x} = \begin{bmatrix} 1 & 2 \\ -1 & 3 \end{bmatrix}$$