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

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

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 show that the solution in Eq. 5 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!

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	$\dfrac{\partial y}{\partial \mathbf{X}}$		

Derivatives by Scalar

Numerator Layout Notation

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

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

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

Denominator Layout Notation

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

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

Derivatives by Vector

Numerator Layout Notation

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

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

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

Denominator Layout Notation

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

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

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

Derivative by Matrix

Numerator Layout Notation

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

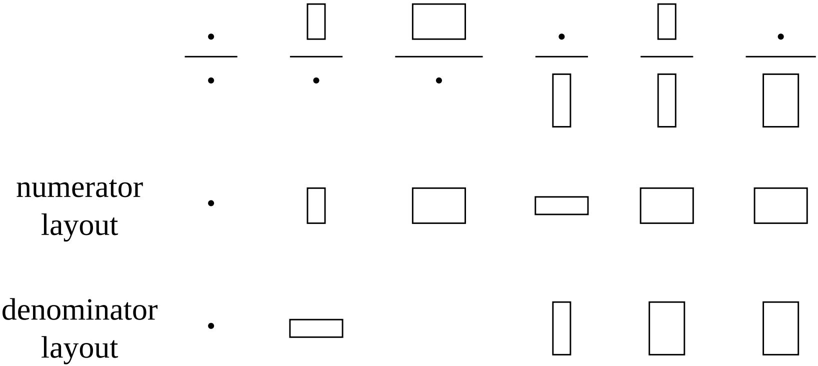
$$\equiv \frac{\partial y}{\partial \mathbf{X}^\top}$$

Denominator Layout Notation

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

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

Pictorial Representation



Caution

- Most books and papers don't state which convention they use.
- Reference [2] uses both conventions but clearly differentiate them.

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

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

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

Commonly Used Derivatives

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

(C1) $$\frac{d\mathbf{a}}{dx} = \mathbf{0} \quad \text{(column matrix)}$$

(C2) $$\frac{da}{d\mathbf{x}} = \mathbf{0}^\top \quad \text{(row matrix)}$$

(C3) $$\frac{da}{d\mathbf{X}} = \mathbf{0}^\top \quad \text{(matrix)}$$

(C4) $$\frac{d\mathbf{a}}{d\mathbf{x}} = \mathbf{0} \quad \text{(matrix)}$$

(C5) $$\frac{d\mathbf{x}}{d\mathbf{x}} = \mathbf{I}$$

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

(C7) $$\frac{d\mathbf{x}^\top\mathbf{x}}{d\mathbf{x}} = 2\,\mathbf{x}^\top$$

(C8) $$\frac{d(\mathbf{x}^\top\mathbf{a})^2}{d\mathbf{x}} = 2\,\mathbf{x}^\top\mathbf{a}\,\mathbf{a}^\top$$

(C9) $$\frac{d\mathbf{A}\mathbf{x}}{d\mathbf{x}} = \mathbf{A}$$

(C10) $$\frac{d\,\mathbf{x}^\top\mathbf{A}}{d\mathbf{x}} = \mathbf{A}^\top$$

(C11) $$\frac{d\,\mathbf{x}^\top\mathbf{A}\mathbf{x}}{d\mathbf{x}} = \mathbf{x}^\top(\mathbf{A} + \mathbf{A}^\top)$$

Math Notation

We represent a vector $\mathbf{x}$ as a column matrix

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}.$$

Its transpose $\mathbf{x}^\top$ is a row matrix

$$\mathbf{x}^\top = \begin{bmatrix} x_1 & x_2 & \cdots & x_m \end{bmatrix}.$$

Consider two vectors $\mathbf{x}$ and $\mathbf{y}$ with the same number of components. Their inner product $\mathbf{x}^\top\mathbf{y}$ is actually a 1×1 matrix:

$$\mathbf{x}^\top\mathbf{y} = [\, s \,]$$

where

$$s = \sum_{i=1}^{m} x_i y_i.$$

For notational inconvenience, we usually drop the matrix and regard the inner product as a scalar, i.e.,

$$\mathbf{x}^\top\mathbf{y} = \sum_{i=1}^{m} x_i y_i.$$

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

Derivatives of Scalar by Vector

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

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) $$\frac{\partial uv}{\partial \mathbf{x}} = u\frac{\partial v}{\partial \mathbf{x}} + v\frac{\partial u}{\partial \mathbf{x}}$$ (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 $\dfrac{\partial \mathbf{u}}{\partial \mathbf{x}}$ and $\dfrac{\partial \mathbf{v}}{\partial \mathbf{x}}$ 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 $\dfrac{\partial \mathbf{u}}{\partial \mathbf{x}}$ and $\dfrac{\partial \mathbf{v}}{\partial \mathbf{x}}$ 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) $$\frac{\partial u v}{\partial \mathbf{X}} = u \frac{\partial v}{\partial \mathbf{X}} + v \frac{\partial u}{\partial \mathbf{X}} \qquad \text{(product rule)}$$

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

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

Derivatives of Vector by Vector

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

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

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

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 Notation	Denominator Layout Notation
(C7)	$\dfrac{d\mathbf{a}^\top\mathbf{x}}{d\mathbf{x}} = \mathbf{a}^\top$	$\dfrac{d\mathbf{a}^\top\mathbf{x}}{d\mathbf{x}} = \mathbf{a}$
(C11)	$\dfrac{d\mathbf{x}^\top\mathbf{A}\mathbf{x}}{d\mathbf{x}} = \mathbf{x}^\top(\mathbf{A}+\mathbf{A}^\top)$	$\dfrac{d\mathbf{x}^\top\mathbf{A}\mathbf{x}}{d\mathbf{x}} = (\mathbf{A}+\mathbf{A}^\top)\mathbf{x}$
(VV5)	$\dfrac{\partial\mathbf{f}(\mathbf{g}(\mathbf{u}))}{\partial\mathbf{x}} = \dfrac{\partial\mathbf{f}(\mathbf{g})}{\partial\mathbf{g}}\dfrac{\partial\mathbf{g}(\mathbf{u})}{\partial\mathbf{u}}\dfrac{\partial\mathbf{u}}{\partial\mathbf{x}}$	$\dfrac{\partial\mathbf{f}(\mathbf{g}(\mathbf{u}))}{\partial\mathbf{x}} = \dfrac{\partial\mathbf{u}}{\partial\mathbf{x}}\dfrac{\partial\mathbf{g}(\mathbf{u})}{\partial\mathbf{u}}\dfrac{\partial\mathbf{f}(\mathbf{g})}{\partial\mathbf{g}}$

Derivations of Derivatives

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

(The not-so-hard way)

Let $s = \mathbf{a}^\top\mathbf{x} = a_1x_1 + \cdots + a_nx_n$. Then, $\dfrac{\partial s}{\partial x_i} = a_i$. So, $\dfrac{ds}{d\mathbf{x}} = \mathbf{a}^\top$.

(The easier way)

Let $s = \mathbf{a}^\top\mathbf{x} = \displaystyle\sum_i a_ix_i$. Then, $\dfrac{\partial s}{\partial x_i} = a_i$. So, $\dfrac{ds}{d\mathbf{x}} = \mathbf{a}^\top$.

(C7) $$\frac{d\mathbf{x}^\top\mathbf{x}}{d\mathbf{x}} = 2\,\mathbf{x}^\top$$

Let $s = \mathbf{x}^\top\mathbf{x} = \displaystyle\sum_i x_i^2$. Then, $\dfrac{\partial s}{\partial x_i} = 2x_i$. So, $\dfrac{ds}{d\mathbf{x}} = 2\,\mathbf{x}^\top$.

(C8) $$\frac{d(\mathbf{x}^\top \mathbf{a})^2}{d\mathbf{x}} = 2\,\mathbf{x}^\top \mathbf{a}\,\mathbf{a}^\top$$

Let $s = \mathbf{x}^\top \mathbf{a}$. Then, $\dfrac{\partial s^2}{\partial x_i} = 2s\dfrac{\partial s}{\partial x_i} = 2\,s\,a_i$. So, $\dfrac{ds^2}{d\mathbf{x}} = 2\,\mathbf{x}^\top \mathbf{a}\,\mathbf{a}^\top$.

(C9) $$\frac{d\mathbf{A}\mathbf{x}}{d\mathbf{x}} = \mathbf{A}$$

(The hard way)

$$\mathbf{A}\mathbf{x} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + \cdots + a_{1n}x_n \\ \vdots \\ a_{n1}x_1 + \cdots + a_{nn}x_n \end{bmatrix}.$$

(The easy way)

Let $\mathbf{s} = \mathbf{A}\mathbf{x}$. Then, $s_i = \displaystyle\sum_j a_{ij}x_j$, and $\dfrac{\partial s_i}{\partial x_j} = a_{ij}$. So, $\dfrac{d\mathbf{s}}{d\mathbf{x}} = \mathbf{A}$.

(C10) $$\frac{d\,\mathbf{x}^\top\mathbf{A}}{d\mathbf{x}} = \mathbf{A}^\top$$

Let $\mathbf{y}^\top = \mathbf{x}^\top\mathbf{A}$, and $\mathbf{a}_j$ denote the j-th column of $\mathbf{A}$. Then, $y_i = \mathbf{x}^\top\mathbf{a}_j$.

Applying (C6) yields $\dfrac{dy_i}{d\mathbf{x}} = \mathbf{a}_j^\top$. So, $\dfrac{d\mathbf{y}^\top}{d\mathbf{x}} = \mathbf{A}^\top$.

(C11) $$\frac{d\,\mathbf{x}^\top\mathbf{A}\mathbf{x}}{d\mathbf{x}} = \mathbf{x}^\top(\mathbf{A} + \mathbf{A}^\top)$$

Apply (SV6) to $\dfrac{d\,\mathbf{x}^\top\mathbf{A}\mathbf{x}}{d\mathbf{x}}$ and obtain $\mathbf{x}^\top\dfrac{d\mathbf{A}\mathbf{x}}{d\mathbf{x}} + (\mathbf{A}\mathbf{x})^\top\dfrac{d\mathbf{x}}{d\mathbf{x}}$,

Next, apply (C9) to the first part of the sum, and obtain

$\mathbf{x}^\top\mathbf{A} + (\mathbf{A}\mathbf{x})^\top$, which is $\mathbf{x}^\top(\mathbf{A} + \mathbf{A}^\top)$.

(Need to prove SV6—Homework.)

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 (C11), (C6), (C9) and (C2) to the four terms.

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

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

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

Transpose both sides of the equation and get

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

Summary

- Matrix calculus studies calculus of matrices.
- There are 6 common derivatives of matrices.
- There are 2 competing notational convention:
  numerator layout notation vs. denominator layout convention.
- We adopt numerator layout notation.
- Do not mix the two conventions in your equations.
- Use matrix differentiation to prove that pseudo-inverse minimizes sum square error.

Use the right tool
become lightning fast!

Probing Questions

- Is there a simple way to double check that the derivative result makes sense?
- Why do we use sum square error for linear fitting? Can we use other forms of errors?
- Six common types of matrix derivatives are discussed. Three other types are left out. Can we work out the other derivatives, e.g., derivatives of vector by matrix or matrix by matrix?

Homework

1. What are the key concepts that you have learned?
2. Prove the product rule SV3 using scalar product rule SS2.

$$\text{(SV3)} \qquad \frac{\partial uv}{\partial \mathbf{x}} = u\frac{\partial v}{\partial \mathbf{x}} + v\frac{\partial u}{\partial \mathbf{x}}$$

3. Prove the product rule SV6 using SV3.

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

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

4. Q2 of AY2015/16 Final Evaluation.

References

1. J. E. Gentle, *Matrix Algebra: Theory, Computations, and Applications in Statistics*, Springer, 2007.
2. H. Lütkepohl, *Handbook of Matrices*, John Wiley & Sons, 1996.
3. K. B. Petersen and M. S. Pedersen, *The Matrix Cookbook*, 2012. `www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf`
4. Wikipedia, *Matrix Calculus*. `en.wikipedia.org/wiki/Matrix_calculus`