L25 Matrices that store all derivatives

1. One function with a matrix of arguments

   (1) Definition
   For $y = f(X) \in R$ where $X = (x_{ij})_{p\times q}$, define matrix $\frac{\partial y}{\partial X} = \left(y'_{x_{ij}}\right)_{p\times q}$. Then this matrix stores all partial derivatives of y to x_{ij}, $i = 1, ..., p$; $j = 1, ..., q$.
   $\left(\frac{\partial y}{\partial X}\right)^T \stackrel{def}{=\!=} \frac{\partial y^T}{\partial X^T} = \frac{\partial y}{\partial X^T}$.

   (2) Formula 1: For $X \in R^{p\times p}$, $\frac{\partial\, \mathrm{tr}(X)}{\partial X} = I_p$.
   Proof $\frac{\partial\, \mathrm{tr}(X)}{\partial X} = \frac{\partial\,(x_{11}+x_{22}+\cdots+x_{pp})}{\partial X} = I_p$.

   (3) Formula 2: For $X \in R^{p\times p}$, $\frac{\partial |X|}{\partial X} = |X| \left(X^T\right)^{-1}$.
   Proof Let $C = (c_{ij})_{p\times p}$ be the cofactor matrix of X. Then $XC^T = |X|\, I_p$.
   So $|X| = x_{i1}c_{i1} + \cdots + x_{ij}c_{ij} + \cdots + x_{ip}c_{ip} \Longrightarrow \frac{\partial |X|}{\partial x_{ij}} = c_{ij} \Longrightarrow \frac{\partial |X|}{\partial X} = C$.
   But $XC^T = |X| I \Longrightarrow C^T = X^{-1}|X| \Longrightarrow C = |X| \left(X^T\right)^{-1}$.
   Thus $\frac{\partial |X|}{\partial X} = |X| \left(X^T\right)^{-1}$.

   (4) Formula 3: For $\alpha \in R^p$, $\beta \in R^q$ and $X = (x_{ij})_{p\times q}$, $\frac{\partial\, \alpha^T X\beta}{\partial X} = \alpha\beta^T$.
   Proof $\alpha^T X\beta = \sum_{i=1}^p \sum_{j=1}^q x_{ij}\alpha_i\beta_j \Longrightarrow \frac{\partial\, \alpha^T X\beta}{\partial x_{ij}} = \alpha_i\beta_j$.
   So $\frac{\partial\, \alpha^T X\beta}{\partial X} = \left(\alpha_i\beta_j\right)_{p\times q} = \alpha\beta^T$.

   (5) Formula 4: For $x \in R^p$, $A \in R^{p\times p}$, $\frac{\partial\, x^T Ax}{\partial x} = (A + A^T)x$.
   Ex1: With $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ and $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$, $x^T Ax = x_1^2 + 5x_1x_2 + 4x_2^2$.
   So $\frac{\partial\, x^T Ax}{\partial x} = \begin{pmatrix} 2x_1 + 5x_2 \\ 5x_1 + 8x_2 \end{pmatrix}$ and

$$\begin{aligned}(A + A^T)x &= \left[\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} + \begin{pmatrix} 1 & 3 \\ 3 & 4 \end{pmatrix}\right] \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 & 5 \\ 5 & 8 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2x_1 + 5x_2 \\ 5x_1 + 8x_2 \end{pmatrix} \\ &= \frac{\partial\, x^T At}{\partial x}.\end{aligned}$$

   Comments: $\frac{\partial\, x^T Ax}{\partial x^T} = x^T(A + A^T)$. If $A^T = A$, then $\frac{\partial\, x^T Ax}{\partial x} = 2Ax$.
   Ex2: For $Y = AXB$ where $A \in R^{m\times p}$ and $B \in R^{q\times n}$, $\frac{\partial\,(AXB)_{st}}{\partial X} = A^T E_{m\times n}(s,t)B^T$.
   Proof Let $I_m = (e_{1m}, ..., e_{mm})$ and $I_n = (e_{1n}, ..., e_{nn})$.
   Then $(AXB)_{st} = e_{sm}^T AXBe_{tn}$. By Formula 3,

$$\frac{\partial\,(AXB)_{st}}{\partial X} = \frac{\partial\, e_{sm}^T AXBe_{tn}}{\partial X} = A^T e_{sm}e_{tn}^T B^T = A^T E_{m\times n}(s,\, t)B^T.$$

2. A matrix of functions with one argument

   (1) Definition
   For $Y = (y_{st}(x))_{m\times n}$ define matrix $\frac{\partial Y}{\partial x} = \left(\frac{\partial y_{st}}{\partial x}\right)_{m\times n}$. Then this matrix stores all derivatives of y_{st} to x, $s = 1, ..., m$; $t = 1, ..., n$.
   $\left(\frac{\partial Y}{\partial x}\right)^T \stackrel{def}{=\!=} \frac{\partial Y^T}{\partial x^T} = \frac{\partial Y^T}{\partial x}$.

(2) Examples

Ex3: For $Y = (2x, x^2-1)$, $\frac{\partial Y}{\partial x} = (2,\, 2x)$ and $\frac{\partial Y^T}{\partial x} = \begin{pmatrix} 2 \\ 2x \end{pmatrix}$.

Ex4: For $Y = AXB$ where $A \in R^{m\times p}$ and $B \in R^{q\times n}$, $\frac{\partial AXB}{\partial x_{ij}} = AE_{p\times q}(i,j)B$.

Proof Write $A = (A_1, .., A_p)$, $A_i = Ae_{ip}$, $B^T = (B_{(1)}, .., B_{(q)})$ and $B_{(j)} = B^T e_{jq}$. Then

$$AXB = (A_1, .., A_p) \begin{pmatrix} x_{11} & \cdots & x_{1q} \\ \vdots & \ddots & \vdots \\ x_{p1} & \cdots & x_{pq} \end{pmatrix} \begin{pmatrix} B_{(1)}^T \\ \vdots \\ B_{(q)}^T \end{pmatrix} = \textstyle\sum_i \sum_j x_{ij} A_i B_{(j)}^T. \text{ So}$$

$$\frac{\partial AXB}{\partial x_{ij}} = A_i B_{(j)}^T = Ae_{ip}\left(B^T e_{jq}\right)^T = Ae_{ip}e_{jq}^T B = AE_{p\times q}(i,\, j)B.$$

3. A vector of functions with a vector of arguments

(1) Definition

For $y = \begin{pmatrix} y_1(x) \\ \vdots \\ y_m(x) \end{pmatrix} \in R^m$ and $x = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} \in R^p$, matrix $\frac{\partial y}{\partial x^T} = \begin{pmatrix} (y_1)'_{x_1} & \cdots & (y_1)'_{x_p} \\ \vdots & \ddots & \vdots \\ (y_m)'_{x_1} & \cdots & (y_m)'_{x_p} \end{pmatrix} \in$ $R^{m\times p}$ stores all partial derivatives of $(y_i)'_{x_j}$, $i = 1, ..., m$; $j = 1, ..., p$.

$\left(\frac{\partial y}{\partial x^T}\right)^T \stackrel{def}{=\!=} \frac{\partial y^T}{\partial x} \in R^{p\times m}$.

Ex5: For $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix}$, $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ and $y = Ax$, $\frac{\partial y}{\partial x^T} = \frac{\partial}{\partial x^T} Ax = \frac{\partial}{\partial x^T} \begin{pmatrix} x_1 + 2x_2 \\ 3x_1 + 4x_2 \\ 5x_1 + 6x_2 \end{pmatrix} =$ $\begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix} = A$.

(2) Formula 5: $\frac{\partial\, Ax}{\partial x^T} = A$. $\qquad \frac{\partial\,(Ax)^T}{\partial x} = A^T$.

(3) A matrix of functions with a matrix of arguments

For $Y = (y_{st}(X))_{m\times n}$ and $X = (x_{ij})_{p\times q}$ there are three ways to store all partial derivatives of $(y_{st})'_{x_{ij}}$.

(i) Matrix $\frac{\partial\, \text{vec}(Y)}{\partial\, [\text{vec}(X)]^T} \in R^{mn\times pq}$ stores all partial derivatives.

(ii) Matrices $\frac{\partial Y}{\partial x_{ij}} \in R^{m\times n}$, $i = 1, ..., p$; $j = 1, ..., q$ store all partial derivatives.

(iii) Matrices $\frac{\partial y_{st}}{\partial X} \in R^{p\times q}$, $s = 1, ..., m$; $t = 1, .., n$ store all partial derivatives.

Ex6: For $Y = AXB \in R^{m\times n}$ and $X \in R^{p\times q}$,

(i) $\frac{\partial\, \text{vec}(AXB)}{\partial\, [\text{vec}(X)]^T} = \frac{\partial\, (B^T \otimes A)\, \text{vec}(X)}{\partial\, [\text{vec}(X)]^T} = B^T \otimes A$.

(ii) By Ex1 $\frac{\partial\, (AXB)_{st}}{\partial X} = A^T E_{m\times n}(s,t) B^T$, $s = 1, .., m$; $t = 1, .., n$.

(iii) By Ex4 $\frac{\partial\, AXB}{\partial x_{ij}} = AE_{p\times q}(i,j)B$, $i = 1, .., p$; $j = 1, .., q$.

L26: Chain rules

1. Derivative matrices in calculus

(1) Gradient vector

With $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ and $y = f(x_1, .., x_n) = f(x)$, $\nabla f(x_1, .., x_n) = \nabla f(x) = \begin{pmatrix} f'_{x_1}(x) \\ \vdots \\ f'_{x_n}(x) \end{pmatrix}$ is called the gradient vector of y at x. $\nabla f(x) = \frac{\partial f(x)}{\partial x} \in R^n$.

Comments: The gradient vector points to the direction along which the directional derivative of y at x is maximized.

x_0 is a stationary point of $y = f(x) \overset{def}{\Longleftrightarrow} \nabla f(x_0) = 0$.

(2) Hessian matrix

$H_f(x) = \begin{pmatrix} f''_{x_1^2} & \cdots & f''_{x_1 x_n} \\ \vdots & \ddots & \vdots \\ f''_{x_n x_1} & \cdots & f''_{x_n^2} \end{pmatrix}$ storing all second order derivatives of $f(x)$ is called the Hessian matrix of y. $H_f(x) = \frac{\partial}{\partial x^T} \nabla(x) = \frac{\partial}{\partial x^T} \frac{\partial f(x)}{\partial x} = \frac{\partial^2 f(x)}{\partial x^T x} = \frac{\partial^2 f(x)}{\partial x x^T} = \frac{\partial}{\partial x} \frac{\partial f(x)}{\partial x^T}$.

Comments: Both $\nabla f(x)$ and $H_{f(x)}$ appear in Taylor expansion at x_0

$$f(x) = f(x_0) + (x - x_0)^T \nabla f(x_0) + \frac{1}{2}(x - x_0)^T H_{f(\xi)}(x - x_0)$$

where $\xi = \alpha x + (1 - \alpha)x_0$ for some $\alpha \in [0,\ 1]$.

If $H_{f(x)} \geq 0$ for all x, then $f(x)$ is minimized at x_0 if and only if $\nabla f(x_0) = 0$, i.e., x_0 is stationary point.

(3) Jacobian matrix

Suppose $y = h_1(x) \in R^n \Longleftrightarrow x = h_2(y) \in R^n$ and $x \in D_x \Longleftrightarrow y \in D_y$. In the definite integral by substitution $y = h_1(x)$,

$$\iint_{D_x} f(x) dx_1, ..dx_n = \iint_{D_y} f(h_2(y)) \text{ abs}|J|\, dy_1, .., dy_n$$

Here J is Jacobian matrix of the transformation.

Comments: $J = \begin{pmatrix} (x_1)'_{y_1} & \cdots & (x_1)'_{y_n} \\ \vdots & \ddots & \vdots \\ (x_n)'_{y_1} & \cdots & (x_n)'_{y_n} \end{pmatrix}$ or $J = \begin{pmatrix} (x_1)'_{y_1} & \cdots & (x_n)'_{y_1} \\ \vdots & \ddots & \vdots \\ (x_1)'_{y_n} & \cdots & (x_n)'_{y_n} \end{pmatrix}$, i.e., $J = \frac{\partial x}{\partial y^T}$ or $J = \frac{\partial x^T}{\partial y}$. Thus notation $J = \frac{\partial\, x_1, .., x_n}{\partial\, y_1, .., y_n}$ is utilized.

2. Chain rules

(1) Three vectors

For $z \in R^m$, $y \in R^n$ and $x \in R^p$, $\quad \frac{\partial z}{\partial x^T} = \frac{\partial z}{\partial y^T} \frac{\partial y}{\partial x^T}$.

For example $\frac{\partial A(x-b)}{\partial x^T} = \frac{\partial Ax - Ab}{\partial x^T} = \frac{\partial Ax}{\partial x^T} = A$ can also be explained as $\frac{\partial A(x-b)}{\partial x^T} = \frac{\partial A(x-b)}{\partial (x-b)^T} \frac{\partial x-b}{\partial x^T} = A \frac{\partial x}{\partial x^T} = AI = A$.

Ex1: For $y = f(x) = \|Ax - b\|^2 = (Ax-b)^T(Ax-b)$, find $\nabla f(x)$ and $H_{f(x)}$.

$\frac{\partial f(x)}{\partial x^T} = \frac{\partial\,(Ax-b)^T(Ax-b)}{\partial\,x^T} = \frac{\partial\,(Ax-b)^T(Ax-b)}{\partial\,(Ax-b)^T}\,\frac{\partial\,Ax-b}{x^T} = (Ax-b)^T 2\,A.$

So $\nabla f(x) = \frac{\partial f(x)}{\partial x} = 2A^T(Ax-b) = 2(A^TAx - A^Tb).$

$H_{f(x)} = \frac{\partial \nabla f(x)}{\partial x^T} = \frac{\partial}{\partial x^T} 2(A^TAx - A^Tb) = 2A^TA.$

(2) One function and two matrices

For $z \in R$, $Y \in R^{m\times n}$ and $X \in R^{p\times q}$,

$$\frac{\partial z}{\partial X} = \sum_s\sum_t \frac{\partial z}{\partial y_{st}}\left(\frac{\partial y_{st}}{\partial X}\right) = \sum_s\sum_t \left(\frac{\partial z}{\partial Y}\right)_{s,t}\left(\frac{\partial y_{st}}{\partial X}\right).$$

Ex2 : With $A \in R^{n\times p}$, $X \in R^{p\times q}$ and $B \in R^{q\times n}$,

$$\begin{aligned}\frac{\partial\,\mathrm{tr}(AXB)}{\partial X} &= \textstyle\sum_{s=1}^n\sum_{t=1}^n \left(\frac{\partial\,\mathrm{tr}(AXB)}{\partial\,(AXB)}\right)_{s,t}\left(\frac{\partial\,(AXB)_{st}}{\partial X}\right)\\ &= \textstyle\sum_{s=1}^n\sum_{t=1}^n (I_n)_{s,t}\left[A^TE_{n\times n}(s,\,t)B^T\right] = A^TI_nB^T = (BA)^T.\end{aligned}$$

3. A example

Find minimizer of $y = f(x) = \|Ax-b\|^2 = (Ax-b)T(Ax-b)$.

(1) Linear algebra approach

$$\|Ax-b\|^2 \text{ is minimized at } \widehat{x} \Longleftrightarrow A\widehat{x} = \pi(b \mid \mathcal{R}(A) \Longleftrightarrow A\widehat{x} = AA^+b.$$

(2) Calculus approach

By Ex1, $\nabla f(x) = 2(A^TAx - A^Tb)$ and $H_{f(x)} = 2A^TA \geq 0$ for all x. Thus

$$\begin{aligned}\|Ax-b\|^2 \text{ is minimized at } \widehat{x} &\Longleftrightarrow \widehat{x} \text{ is a stationary point} \Longleftrightarrow \nabla f(\widehat{x}) = 0\\ &\Longleftrightarrow 2(A^TA\widehat{x} - A^Tb) = 0 \Longleftrightarrow A^TA\widehat{x} = A^Tb.\end{aligned}$$

(3) Equivalency

$$A\widehat{x} = AA^+b \Longleftrightarrow A^TA\widehat{x} = A^Tb.$$

⇒: $A\widehat{x} = AA^+b \Longrightarrow A^TA\widehat{x} = A^T(AA^+)b = A^T(AA^+)^Tb = (AA^+A)^Tb = A^Tb.$

⇐: $A^TA\widehat{x} = A^Tb \Longrightarrow (A^+)^T(A^TA)\widehat{x} = (A^+)^TA^tb \Longrightarrow (AA^+)^TA\widehat{x} = (AA^+)^Tb$
$\Longrightarrow AA^+A\widehat{x} = AA^+b \Longrightarrow A\widehat{x} = AA^+b.$