Matrix derivative on matrix function of matrix variable

How to compute $\nabla_{\mathbf{X}} f(\mathbf{X})$, where $\mathbf{X}$ is a matrix and out of f is matrix?

Andersen Ang

Mathématique et recherche opérationnelle
UMONS, Belgium

manshun.ang@umons.ac.be   Homepage: angms.science

First draft : November 2, 2019
Last update : November 6, 2019

Set up

- All function f in this document are in the form of $f : \Omega \rightarrow \mathbb{R}^{p \times q}$
  - i.e., f maps the elements from the domain set Ω to the set $\mathbb{R}^{p \times q}$
  - i.e., the output of f is a matrix
- We consider in this document : derivative of f with respect to (w.r.t.) matrix
  - where the derivative of f w.r.t. vector is a special case
- Matrix derivative appears in many applications, especially on second order optimization method where Hessian is required. A systematic approach to compute the derivative is important
- To gain understanding of matrix derivative, we first review scalar derivative and vector derivative

Differential and derivatives on function of single variable

Let $\mathbf{y} = f(x)$, where $\mathbf{y}$ is a vector, the *derivative* of $\mathbf{y}$ w.r.t. scalar x is

$$\frac{d\mathbf{y}}{dx} = \frac{df(x)}{dx} = f'(x),$$

where $f'(x)$ is a vector with same size as $\mathbf{y}$

The *differential* of $\mathbf{y}$, a vector, is

$$d\mathbf{y}.$$

The relationship between differential and derivative is

$$d\mathbf{y} = f'(x)dx$$

Recap differential vs derivative

- Differential : the infinitesimal difference in varying variable
- Derivative : the rate of change of the function w.r.t. the variable
- Here $d\mathbf{y}$ is a vector, dx is a scalar and $f'(x)$ is a vector

Differential and derivatives on function of vector variable

Let $\mathbf{y} = f(\mathbf{x})$, denote $d\mathbf{x} = [dx_1 \ dx_2 \ \dots]^\top$ and the derivative

$$\nabla_{\mathbf{x}} f = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial x_n} & \frac{\partial y_2}{\partial x_n} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

we have the (total) *differential* as

$$d\mathbf{y} = \nabla_{\mathbf{x}} f^\top d\mathbf{x}.$$

The equation $d\mathbf{y} = \nabla_{\mathbf{x}} f^\top d\mathbf{x}$. tells that it is possible to compute $\nabla_{\mathbf{x}} f$ by using the information of $d\mathbf{y}$ and $d\mathbf{x}$ (if they are available).

Using differential-derivative relation to compute vector derivative

We can obtain such $\nabla_{\mathbf{x}} f(\mathbf{x})$ by using the formula

$$dy = \langle \nabla_{\mathbf{x}} f, d\mathbf{x} \rangle = \nabla_{\mathbf{x}} f^{\top} d\mathbf{x}$$

To do so we need some basic tools from differential :

$$dc = 0 \tag{1}$$
$$dXY = (dX)Y + XdY \tag{2}$$
$$d\mathbf{x}^{\top} = (d\mathbf{x})^{\top} \tag{3}$$

where c is a constant.

Example

Find the gradient of $\mathbf{y} = f(\mathbf{x}) = \mathbf{A}^\top \mathbf{A}\mathbf{x} - \mathbf{A}^\top \mathbf{b}$ w.r.t. $\mathbf{x}$, where $\mathbf{A}$ is a constant matrix and $\mathbf{b}$ is a constant vector.

We know $\nabla_{\mathbf{x}} f(\mathbf{x}) = \mathbf{A}^\top \mathbf{A}$, but here it is emphasized on how to get this by using the formula $d\mathbf{y} = \langle \nabla_{\mathbf{x}} f, d\mathbf{x} \rangle = \nabla_{\mathbf{x}} f^\top d\mathbf{x}$.

Step-by-step solution :

- Take differential : $d\mathbf{y} = d\mathbf{A}^\top \mathbf{A}\mathbf{x} - d\mathbf{A}^\top \mathbf{b}$
- By (1), $d\mathbf{A}^\top \mathbf{b} = 0$
- By (2),

$$\begin{aligned} d\mathbf{y} &= (d\mathbf{A}^\top \mathbf{A})\mathbf{x} + \mathbf{A}^\top \mathbf{A} d\mathbf{x} \\ &\overset{(1)}{=} \mathbf{A}^\top \mathbf{A} d\mathbf{x} \end{aligned}$$

- By the differential-derivative equation $d\mathbf{y} = \nabla_{\mathbf{x}} f^\top d\mathbf{x}$, we have $\mathbf{A}^\top \mathbf{A} = \nabla_{\mathbf{x}} f^\top$, so $\nabla_{\mathbf{x}} f = \mathbf{A}^\top \mathbf{A}$.

Differential and derivatives on function of matrix variable

On function $\mathbf{Y} = f(\mathbf{X})$, where $\mathbf{X}$ is a m-by-n matrix and $\mathbf{Y}$ is a p-by-q matrix, the gradient of $\mathbf{Y}$ w.r.t. matrix can be defined using the definition of the vector case : by *vectorizing* the matrices, the tools from the vector case can be used.

Definition (Vectorization). Given $\mathbf{X} \in \mathbb{R}^{m \times n}$, $\mathrm{vec}(\mathbf{X})$ is a $mn \times 1$ vector $\mathrm{vec}(\mathbf{X}) = [X_{11},\ X_{21},\ \dots\ X_{m1}, X_{12},\ X_{22},\ \dots\ X_{m2}, \dots, X_{1n},\ X_{2n},\ \dots\ X_{mn}]^\top$

Under vectorization, the gradient of $\mathbf{Y}$ w.r.t. $\mathbf{X}$ is defined as

$$\nabla_{\mathbf{X}} \mathbf{Y} = \frac{\partial \mathbf{Y}}{\partial \mathbf{X}} = \nabla_{\mathrm{vec}(\mathbf{X})} \mathrm{vec}(\mathbf{Y}) = \frac{\partial \mathrm{vec}(\mathbf{Y})}{\partial \mathrm{vec}(\mathbf{X})}$$

The differential-derivative equation is then

$$\mathrm{vec}(d\mathbf{Y}) = (\nabla_{\mathbf{X}} \mathbf{Y})^\top \mathrm{vec}(d\mathbf{X}) = \left(\frac{\partial \mathbf{Y}}{\partial \mathbf{X}}\right)^\top \mathrm{vec}(d\mathbf{X})$$

Note. $\nabla_{\mathbf{X}} \mathbf{Y}$ and $\dfrac{\partial \mathbf{Y}}{\partial \mathbf{X}}$ are just different notations of the same thing.

Differential and derivatives on matrix case

The differential-derivative equation

$$\mathrm{vec}(d\mathbf{Y}) = (\nabla_{\mathbf{X}}\mathbf{Y})^\top \mathrm{vec}(d\mathbf{X}) = \left(\frac{\partial \mathbf{Y}}{\partial \mathbf{X}}\right)^\top \mathrm{vec}(d\mathbf{X})$$

is a compressed expression of the (total) differential of $\mathbf{Y}$.

- In the derivative, $\nabla_{\mathbf{X}}\mathbf{Y} = \nabla_{\mathrm{vec}(\mathbf{X})}\mathrm{vec}(\mathbf{Y})$ is a matrix computed using compressed matrix $\mathrm{vec}(\mathbf{X})$ and $\mathrm{vec}(\mathbf{Y})$ in vector form
- The product $(\nabla_{\mathbf{X}}\mathbf{Y})^\top \mathrm{vec}(d\mathbf{X})$ involves another compressed matrix $\mathrm{vec}(d\mathbf{X})$ in vector form
- The differential itself $\mathrm{vec}(d\mathbf{Y})$ is in compressed form

In fact, the gradient $(\nabla_{\mathbf{X}}\mathbf{Y})$ is a 4-th order tensor with the (i,j,p,q)-th entry equal to $\frac{\partial Y_{ij}}{\partial X_{pq}}$.

Using differential-derivative equation to find derivative

Some basic tools from vectorization

$$\mathrm{vec}(\mathbf{AXB}) = (\mathbf{B}^\top \otimes \mathbf{A})\mathrm{vec}(\mathbf{X}) \quad (4)$$

$$\mathrm{vec}(\mathbf{X}^\top) = \mathbf{C}_{mn}\mathrm{vec}(\mathbf{X}) \quad (5)$$

where $\otimes$ is Kronecker product and $\mathbf{C}_{mn}$ is commutation matrix.

See this wiki page for more on commutation matrix.

See this wiki page for more on Kronecker product.

Seethe Matrix Cookbook section 10.2 for more formula on vectorization.

Example.Find the derivate of $\mathbf{Y} = \mathbf{AXB}$ w.r.t. $\mathbf{X}$

$\mathbf{A}, \mathbf{B}$ are constant matrices

Step-by-step solution.

- Take differential $d\mathbf{Y} = d\mathbf{AXB}$.
- $d\mathbf{AXB} \overset{(2)}{=} (d\mathbf{A})\mathbf{XB} + \mathbf{A}(d\mathbf{X})\mathbf{B} + \mathbf{AX}(d\mathbf{B})$.
  $\mathbf{A}, \mathbf{B}$ are constants so by (1) $d\mathbf{A}, d\mathbf{B}$ are zeros.
  We have $d\mathbf{Y} = \mathbf{A}(d\mathbf{X})\mathbf{B}$.
- Take vectorization $\mathrm{vec}(d\mathbf{Y}) = \mathrm{vec}(A(d\mathbf{X})\mathbf{B}) \overset{(4)}{=} (\mathbf{B}^\top \otimes \mathbf{A})\mathrm{vec}(d\mathbf{X})$
- By differential-derivative equation $\mathrm{vec}(d\mathbf{Y}) = (\nabla_{\mathbf{X}}\mathbf{Y})^\top \mathrm{vec}(d\mathbf{X})$, we have $\nabla_{\mathbf{X}}\mathbf{Y} = (\mathbf{B}^\top \otimes \mathbf{A})^\top = \mathbf{B} \otimes \mathbf{A}^\top$

Example. Find $\nabla_{\mathbf{X}}$ of $\mathbf{Y} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}$

Solution.

- Take differential $d\mathbf{Y} = d\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \overset{(2)}{=} (d\mathbf{X})(\mathbf{X}^\top\mathbf{X})^{-1} + \mathbf{X}d(\mathbf{X}^\top\mathbf{X})^{-1}$
- Consider the second term, let $\mathbf{Y} = \mathbf{X}^\top\mathbf{X}$. By formula 40 of the matrix cookbook, $d\mathbf{Y}^{-1} = -\mathbf{Y}^{-1}(d\mathbf{Y})\mathbf{Y}^{-1}$. Put $\mathbf{Y} = \mathbf{X}^\top\mathbf{X}$ back we get $d(\mathbf{X}^\top\mathbf{X})^{-1} = -(\mathbf{X}^\top\mathbf{X})^{-1}d(\mathbf{X}^\top\mathbf{X})(\mathbf{X}^\top\mathbf{X})^{-1}$.

  As $d(\mathbf{X}^\top\mathbf{X}) \overset{(2)}{=} (d\mathbf{X}^\top)\mathbf{X} + \mathbf{X}^\top d\mathbf{X}$, we have

$$
\begin{aligned}
d(\mathbf{X}^\top\mathbf{X})^{-1} &= -(\mathbf{X}^\top\mathbf{X})^{-1}((d\mathbf{X}^\top)\mathbf{X} + \mathbf{X}^\top d\mathbf{X})(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= -(\mathbf{X}^\top\mathbf{X})^{-1}(d\mathbf{X}^\top)\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top d\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\
\mathbf{X}d(\mathbf{X}^\top\mathbf{X})^{-1} &= -\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}(d\mathbf{X}^\top)\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top d\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}
\end{aligned}
$$

  We have
$$
\begin{aligned}
d\mathbf{Y} =&(d\mathbf{X})(\mathbf{X}^\top\mathbf{X})^{-1} \\
&-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}(d\mathbf{X}^\top)\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\
&-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top d\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}
\end{aligned}
$$

- Take vectorization
$$
\begin{aligned}
\mathrm{vec}(d\mathbf{Y}) =&\mathrm{vec}\Big((d\mathbf{X})(\mathbf{X}^\top\mathbf{X})^{-1}\Big) \\
&-\mathrm{vec}\Big(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}(d\mathbf{X}^\top)\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\Big) \\
&-\mathrm{vec}\Big(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top d\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\Big)
\end{aligned}
$$

Example. Find the gradient of $\mathbf{Y} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}$

- By vec formula (4)

$$\begin{aligned}\mathrm{vec}(d\mathbf{Y}) =& \Big((\mathbf{X}^\top\mathbf{X})^{-1} \otimes \mathbf{I}_m\Big)\mathrm{vec}(d\mathbf{X}) \\ &- \Big((\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1})^\top \otimes \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\Big)\mathrm{vec}(d\mathbf{X}^\top) \\ &- \Big((\mathbf{X}^\top\mathbf{X})^{-1} \otimes \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\Big)\mathrm{vec}(d\mathbf{X})\end{aligned}$$

- By vec formula (5)

$$\begin{aligned}\mathrm{vec}(d\mathbf{Y}) =& \Big((\mathbf{X}^\top\mathbf{X})^{-1} \otimes \mathbf{I}_m\Big)\mathrm{vec}(d\mathbf{X}) \\ &- \Big((\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1})^\top \otimes \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\Big)\mathbf{C}_{mn}\mathrm{vec}(d\mathbf{X}) \\ &- \Big((\mathbf{X}^\top\mathbf{X})^{-1} \otimes \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\Big)\mathrm{vec}(d\mathbf{X})\end{aligned}$$

- Combine terms

$$\begin{aligned}\mathrm{vec}(d\mathbf{Y}) =& \Bigg(\Big((\mathbf{X}^\top\mathbf{X})^{-1} \otimes \mathbf{I}_m\Big) - \Big((\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1})^\top \otimes \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\Big)\mathbf{C}_{mn} \\ &- \Big((\mathbf{X}^\top\mathbf{X})^{-1} \otimes \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\Big)\Bigg)\mathrm{vec}(d\mathbf{X})\end{aligned}$$

Example. Find the gradient of $\mathbf{Y} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}$

- By differential-derivative equation, we have the gradient of $\mathbf{Y}$ as

$$\Bigg(\Big((\mathbf{X}^\top\mathbf{X})^{-1} \otimes \mathbf{I}_m\Big) - \Big((\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1})^\top \otimes \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\Big)\mathbf{C}_{mn}$$

$$-\Big((\mathbf{X}^\top\mathbf{X})^{-1} \otimes \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\Big)\Bigg)^\top$$

which is

$$\Bigg(\Big((\mathbf{X}^\top\mathbf{X})^{-1} \otimes \mathbf{I}_m\Big) - \mathbf{C}_{mn}^\top\Big(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \otimes (\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1})^\top\Big)$$

$$-\Big((\mathbf{X}^\top\mathbf{X})^{-1} \otimes \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\Big)\Bigg)$$

which is very messy as the 4-th order tensor derivative is compressed using Kronecker product and vectorization

End of document