Matrix derivative on scalar function of matrix variable

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

Andersen Ang

Mathématique et recherche opérationnelle
UMONS, Belgium

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

First draft : November 1, 2019
Last update : November 5, 2019

Set up

- All function f in this document are in the form of $f : \Omega \to \mathbb{R}$
  - i.e., f maps the elements from the domain set Ω to real line $\mathbb{R}$
  - i.e., the output of f is a scalar
- 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 has many applications, a systematic approach on computing the derivative is important
- To understand matrix derivative, we first review scalar derivative and vector derivative of f

Differential and derivatives on function of single variable

Let $y = f(x)$, the *derivative* of y w.r.t. x is

$$\frac{dy}{dx} = \frac{df(x)}{dx} = f'(x).$$

The *differential* of y is

$$dy.$$

The relationship between differential and derivative is

$$dy = f'(x)dx$$

Recap differential vs derivative

- Differential : the infinitesimal difference in varying variable
- Derivative : the rate of change of f w.r.t. the (changing) variable

Differential and derivatives on function of vector variable

Let $y = f(x_1, x_2, x_3)$, we have the (total) *differential* of y as

$$dy = \frac{\partial f}{\partial x_1} dx_1 + \frac{\partial f}{\partial x_2} dx_2 + \frac{\partial f}{\partial x_3} dx_3.$$

- Meaning : the (total) amount of change in f (denote as dy) is the sum of amount of changes of its variables.
- The "amount" of change is a scalar quantity for scalar function
- The "sum of amount of changes" is the sum of the amount of change in each variable (dx_i) scaled by how fast the function f is changed w.r.t. that variable ($\frac{\partial f}{\partial x_i}$)

Let $d\mathbf{x} = [dx_1 \ dx_2 \ dx_3]^\top$ and $\nabla_{\mathbf{x}} f = [\frac{\partial f}{\partial x_1} \ \frac{\partial f}{\partial x_2} \ \frac{\partial f}{\partial x_3}]^\top$, we have

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

where $\langle \ . \ \rangle$ denotes inner product. This equation tells that it is possible to compute $\nabla_{\mathbf{x}} f$ by using the information of dy and $d\mathbf{x}$ (if they are available).

Differential and derivatives on function of vector variable

Let $y = f(x_1, \ldots, x_n) = f(\mathbf{x})$.
This is a function of n variables, or a function of vector variable $\mathbf{x}$.

The (total) *differential* of y is

$$dy = \sum_{i}^{n} \frac{\partial f}{\partial x_i} dx_i.$$

The differential-derivative equation

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

still holds, where $d\mathbf{x} = [dx_1 \ \ldots \ dx_n]^{\top}$ and $\nabla_{\mathbf{x}} f = [\frac{\partial f}{\partial x_1} \ \ldots \ \frac{\partial f}{\partial x_n}]^{\top}$.

The differential-derivative equation is the key to compute derivative $\nabla_{\mathbf{x}} f$ for function of vector variable.

Computing derivative by differential-derivative equation

Example. Find the gradient of $y = f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$ w.r.t. $\mathbf{x}$, where $\mathbf{A}$ is a constant matrix and $\mathbf{b}$ is a constant vector.

We know $\nabla_{\mathbf{x}} f(\mathbf{x}) = 2\mathbf{A}^\top(\mathbf{A}\mathbf{x} - \mathbf{b})$, but let's see how to get this by using the formula $dy = \langle \nabla_{\mathbf{x}} f, d\mathbf{x} \rangle = \nabla_{\mathbf{x}} f^\top d\mathbf{x}$.

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, X, Y are two mathematical object (scalar, vector or matrix)

Computing the gradient of $y = \|\mathbf{Ax} - \mathbf{b}\|_2^2$ w.r.t. $\mathbf{x}$

Step-by-step solution :

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

$$
\begin{aligned}
dy &= d\mathbf{x}^\top \mathbf{A}^\top \mathbf{Ax} - 2d\mathbf{b}^\top \mathbf{Ax} \\
&\overset{(2)}{=} d\mathbf{x}^\top \mathbf{A}^\top \mathbf{Ax} - 2(d\mathbf{b}^\top \mathbf{A})\mathbf{x} - 2\mathbf{b}^\top \mathbf{A} d\mathbf{x} \\
&\overset{(1)}{=} d\mathbf{x}^\top \mathbf{A}^\top \mathbf{Ax} - 2\mathbf{b}^\top \mathbf{A} d\mathbf{x} \\
&\overset{(2)}{=} (d\mathbf{x}^\top)\mathbf{A}^\top \mathbf{Ax} + \mathbf{x}^\top (d\mathbf{A}^\top \mathbf{A})\mathbf{x} + \mathbf{x}^\top \mathbf{A}^\top \mathbf{A} d\mathbf{x} - 2\mathbf{b}^\top \mathbf{A} d\mathbf{x} \\
&\overset{(1)}{=} (d\mathbf{x}^\top)\mathbf{A}^\top \mathbf{Ax} + \mathbf{x}^\top \mathbf{A}^\top \mathbf{A} d\mathbf{x} - 2\mathbf{b}^\top \mathbf{A} d\mathbf{x} \\
&\overset{(3)}{=} 2\mathbf{x}^\top \mathbf{A}^\top \mathbf{A} d\mathbf{x} - 2\mathbf{b}^\top \mathbf{A} d\mathbf{x} \\
&= 2(\mathbf{x}^\top \mathbf{A}^\top - \mathbf{b}^\top)\mathbf{A} d\mathbf{x}
\end{aligned}
$$

- By differential-derivative equation $dy = \nabla_{\mathbf{x}} f^\top d\mathbf{x}$, we have $2(\mathbf{x}^\top \mathbf{A}^\top - \mathbf{b}^\top)\mathbf{A} = \nabla_{\mathbf{x}} f^\top$, so $\nabla_{\mathbf{x}} f = 2\mathbf{A}^\top (\mathbf{Ax} - \mathbf{b})$.

Differential and derivatives on function of matrix variable

Let $y = f(\mathbf{X})$, where $\mathbf{X}$ is a m-by-n matrix. The (total) *differential* of y is

$$dy = \sum_{i}^{m} \sum_{j}^{n} \frac{\partial f}{\partial X_{ij}} dX_{ij}.$$

- Let $d\mathbf{X}$ be the matrix of differential dX_{ij} for all i, j
  i.e., $d\mathbf{X} = [dX_{ij}]$ is a m-by-n matrix where the (i, j) entry is dX_{ij}.
- Let $\nabla_{\mathbf{X}} f$ be the derivate (gradient) of f w.r.t. $\mathbf{X}$.
  i.e., $\nabla_{\mathbf{X}} f$ is a m-by-n matrix where the (i, j) entry is $\dfrac{\partial f}{\partial X_{ij}}$.

Follow the same logic $dy = \langle \nabla_{\mathbf{x}} f, d\mathbf{x} \rangle$ for the vector case, we have

$$dy = \langle \nabla_{\mathbf{X}} f, d\mathbf{X} \rangle$$

for the matrix case, where $\langle\ ,\ \rangle$ is the inner product for matrices.

Differential and derivatives on matrix case

The key in the differential-derivative equation

$$dy = \langle \nabla_{\mathbf{X}} f, d\mathbf{X} \rangle$$

is the expression for the inner product : the trace operator

$$dy = \mathrm{Tr}\left(\nabla_{\mathbf{X}} f^{\top} d\mathbf{X} \right).$$

The above can be proved simply by the definition of trace : as

$$\mathrm{Tr}\mathbf{A}^{\top}\mathbf{B} = \sum_{ij} A_{ij} B_{ij},$$

hence

$$dy = \sum_{i}^{m} \sum_{j}^{n} \frac{\partial f}{\partial X_{ij}} dX_{ij} = \mathrm{Tr}\left(\nabla_{\mathbf{X}} f^{\top} d\mathbf{X} \right).$$

Using differential-derivative equation to find derivative

Similar to the vector case, we have some basic tools from differential

$$
\begin{aligned}
d(\mathbf{X}\mathbf{Y}) &= (d\mathbf{X})\mathbf{Y} + \mathbf{X}d\mathbf{Y} \\
d(\mathbf{X}^\top) &= (d\mathbf{X})^\top \\
d\mathrm{Tr}\mathbf{X} &= \mathrm{Tr}(d\mathbf{X}) \qquad (4) \\
d\mathbf{X}^{-1} &= -\mathbf{X}^{-1}(d\mathbf{X})\mathbf{X}^{-1} \qquad (5)
\end{aligned}
$$

We prove (5) here : consider $\mathbf{X}\mathbf{X}^{-1} = \mathbf{I}$.

Take differential we have $d\mathbf{X}\mathbf{X}^{-1} = d\mathbf{I}$ where $d\mathbf{I} \overset{(1)}{=} 0$.
Use (2) on $d\mathbf{X}\mathbf{X}^{-1}$ gives $d\mathbf{X}\mathbf{X}^{-1} = (d\mathbf{X})\mathbf{X}^{-1} + \mathbf{X}d\mathbf{X}^{-1}$ so $(d\mathbf{X})\mathbf{X}^{-1} + \mathbf{X}d\mathbf{X}^{-1} = 0$ which gives (5).

See the Matrix Cookbook for more formula on differential, e.g. (33)-(45).

Example. (eq.101 in the Matrix Cookbook)

Given $\mathbf{A}, \mathbf{B}$ are constant matrices, find $\nabla_{\mathbf{X}} f(\mathbf{X})$ of $y = \mathrm{Tr}(\mathbf{AXB})$

Step-by-step solution.

- Take differential : $dy = d\mathrm{Tr}(\mathbf{AXB}) \overset{(4)}{=} \mathrm{Tr}(d\mathbf{AXB})$.
- $d\mathbf{AXB} \overset{(2)}{=} (d\mathbf{A})\mathbf{XB} + \mathbf{A}(d\mathbf{X})\mathbf{B} + \mathbf{AX}(d\mathbf{B})$.
  $\mathbf{A}, \mathbf{B}$ are constants so by (1) $d\mathbf{A}, d\mathbf{B}$ are zeros.
  We have $dy = \mathrm{Tr}(\mathbf{A}(d\mathbf{X})\mathbf{B})$.
- Trace trick : $\mathrm{Tr}(\mathbf{PQ}) = \mathrm{Tr}(\mathbf{QP})$, let $\mathbf{P} = \mathbf{A}d\mathbf{X}$ and $\mathbf{Q} = \mathbf{B}$ then $dy = \mathrm{Tr}(\mathbf{BA}d\mathbf{X})$.
- By the differential-derivative equation $dy = \mathrm{Tr}\left(\nabla_{\mathbf{X}} f^\top d\mathbf{X}\right)$, we have $\nabla_{\mathbf{X}} f = (\mathbf{BA})^\top = \mathbf{A}^\top \mathbf{B}^\top$.

Special cases : $\nabla_{\mathbf{X}} \mathrm{Tr}(\mathbf{AX}) = \mathbf{A}^\top$ and $\nabla_{\mathbf{X}} \mathrm{Tr}(\mathbf{XB}) = \mathbf{B}^\top$.

Example. (eq. 55 in the Matrix Cookbook)

Find $\nabla_{\mathbf{X}} f(\mathbf{X})$ of $y = \log\det(\mathbf{X}^\top \mathbf{X})$.

Step-by-step solution.

- Let $\mathbf{Y} = \mathbf{X}^\top \mathbf{X}$ so $y = \log\det \mathbf{Y}$, take differential $dy = d\log\det \mathbf{Y}$
- By differential formula (Maxtri Cookbook eq.43), $d\log\det \mathbf{Y} = \mathrm{Tr}(\mathbf{Y}^{-1} d\mathbf{Y})$
- Put back $\mathbf{Y} = \mathbf{X}^\top \mathbf{X}$, we have $dy = \mathrm{Tr}\Big((\mathbf{X}^\top \mathbf{X})^{-1} d\mathbf{X}^\top \mathbf{X}\Big)$
- By (2) and (3), $d\mathbf{X}^\top \mathbf{X} = (d\mathbf{X}^\top)\mathbf{X} + \mathbf{X}^\top d\mathbf{X} = (d\mathbf{X})^\top \mathbf{X} + \mathbf{X}^\top d\mathbf{X}$. We have

$$\begin{aligned} dy &= \mathrm{Tr}\Big((\mathbf{X}^\top \mathbf{X})^{-1}\big((d\mathbf{X})^\top \mathbf{X} + \mathbf{X}^\top d\mathbf{X}\big)\Big) \\ &= \mathrm{Tr}\Big((\mathbf{X}^\top \mathbf{X})^{-1}(d\mathbf{X})^\top \mathbf{X}\Big) + \mathrm{Tr}\Big((\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top d\mathbf{X}\Big) \end{aligned}$$

- Trace trick $\mathrm{Tr}\mathbf{S}\mathbf{Q}^\top \mathbf{P} = \mathrm{Tr}\mathbf{S}\mathbf{P}^\top \mathbf{Q}$, where $\mathbf{S}$ is symmetric, so $dy = 2\mathrm{Tr}\Big((\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top d\mathbf{X}\Big)$
- By the differential-derivative equation $dy = \mathrm{Tr}\left(\nabla_{\mathbf{X}} f^\top d\mathbf{X}\right)$, we have $\nabla_{\mathbf{X}} f = 2\Big((\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}\Big)^\top = 2\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}$

Last page

See here for the proof of $\nabla_{\mathbf{x}} f \log\det(\mathbf{X}^\top \mathbf{X}) = 2\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}$, using different approach.

Comparing the two approaches, the one using $dy = \mathrm{Tr}\left(\nabla_{\mathbf{X}} f^\top dX\right)$ is slightly more systematic and less cumbersome.

Summary

- Matrix derivative $\nabla_{\mathbf{X}} f$ on a function of matrix variable
- Using $dy = \mathrm{Tr}\left(\nabla_{\mathbf{X}} f^\top d\mathbf{X}\right)$ to compute $\nabla_{\mathbf{X}} f$

End of document