Advanced Data Analysis

Whiteboard Lecture

Vector Differential Calculus

Let

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

Then

$$f(\boldsymbol{x}) = \sin(x_1 x_2^2) + 4x_3^7.$$

Example 1

$$f : \mathbb{R}^3 \longrightarrow \mathbb{R}$$

$$f\left(\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}\right) = \sin(x_1 x_2^2) + 4x_3^7.$$

The derivative vector of f with respect to $\boldsymbol{x}$ is defined as:

$$\mathsf{D}\, f(\boldsymbol{x}) = \begin{bmatrix} \dfrac{\partial f(\boldsymbol{x})}{\partial x_1} & \dfrac{\partial f(\boldsymbol{x})}{\partial x_2} & \dfrac{\partial f(\boldsymbol{x})}{\partial x_3} \end{bmatrix}.$$

Note that $\mathsf{D}\, f(\boldsymbol{x})$ is a 1×3 row vector.

$$\mathsf{D}\, f(\boldsymbol{x}) = [x_2^2 \cos(x_1 x_2^2) \quad 2x_1 x_2 \cos(x_1 x_2^2) \quad 28x_3^6].$$

Question

How to efficiently find $\mathrm{D} f(\boldsymbol{x})$ for general $f : \mathbb{R}^p \longrightarrow \mathbb{R}$?

e.g. if

$$f(\boldsymbol{x}) = \cos^2(\boldsymbol{x}^T\boldsymbol{x}) + \mathrm{tr}^3(\boldsymbol{x}\boldsymbol{x}^T)$$

then what is $\mathrm{D} f(\boldsymbol{x})$?

(where $\boldsymbol{x}$ is $p \times 1$).

Start with a simple (familiar)

$$f : \mathbb{R} \longrightarrow \mathbb{R}$$

example.

$$f(x) = x^3 \sin(x).$$

$$\begin{aligned}\frac{d\,f(x)}{dx} &= \left(\frac{dx^3}{dx}\right)\sin(x) + x^3\left(\frac{d\sin(x)}{dx}\right)\\ &= 3x^2\sin(x) + x^3\cos(x)\end{aligned}$$

Alternative approach:

$$\begin{aligned} f(x) &= x^3 \sin x\\ df(x) &= d(x^3 \sin x)\\ &= (dx^3)\sin x + x^3(d\sin x)\\ &= (3x^2 dx)\sin x + x^3(\cos x\, dx)\\ &= (3x^2\sin x + x^3\cos x)\,dx\\ \Longrightarrow \quad \frac{df(x)}{dx} &= 3x^2\sin(x) + x^3\cos(x)\end{aligned}$$

Class Exercise

$$f(x) = (5x^2 + 8x)\ln(x)$$

Find $\frac{df(x)}{dx}$ using the alternative (differential algebra) approach.

Answer

$$\begin{aligned} f(x) &= (5x^2 + 8x)\ln(x) \\ df(x) &= d\{(5x^2 + 8x)\ln(x)\} \\ &= \{d(5x^2 + 8x)\}\ln(x) + (5x^2 + 8x)\{d\ln(x)\} \\ &= \{(10x + 8)dx\}\ln(x) + (5x^2 + 8x)\{(1/x)dx\} \\ &= \{(10x + 8)\ln(x) + (5x + 8)\}dx \end{aligned}$$

Hence

$$\frac{df(x)}{dx} = (10x + 8)\ln(x) + 5x + 8.$$

First Identification Theorem

$$f : \mathbb{R}^p \longrightarrow \mathbb{R}$$

Suppose $\boldsymbol{a}$ $(1 \times p)$ is such that

$$d\,f(\boldsymbol{x}) = \boldsymbol{a}\,d\boldsymbol{x}.$$

Then

$$\mathrm{D}\,f(\boldsymbol{x}) = \boldsymbol{a}.$$

Example 1 (Least squares regression)

Model is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}).$$

Choose $\boldsymbol{\beta}$ to minimise

$$S(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Minimum should occur at stationary point;

i.e. where

$$\mathrm{D}S(\boldsymbol{\beta}) = \boldsymbol{0}.$$

Answer

$$\begin{aligned}
dS(\boldsymbol{\beta}) &= d\{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\} \\
&= \{d(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\
&\quad +(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\{d(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\} \\
&= (-\boldsymbol{X}d\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\
&\quad +(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(-\boldsymbol{X}d\boldsymbol{\beta}) \\
&= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(-\boldsymbol{X}d\boldsymbol{\beta}) \\
&\quad +(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(-\boldsymbol{X}d\boldsymbol{\beta}) \\
&= -2(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{X}d\boldsymbol{\beta}
\end{aligned}$$

i.e. We have just shown

$$dS(\boldsymbol{\beta}) = -2(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{X}d\boldsymbol{\beta}.$$

By the First Identification Theorem

$$\mathrm{D}S(\boldsymbol{\beta}) = -2(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{X}.$$

Setting this to zero gives

$$
\begin{aligned}
\boldsymbol{y}^T\boldsymbol{X} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X} &= \mathbf{0} \\
\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} &= \boldsymbol{X}^T\boldsymbol{y} \\
\boldsymbol{\beta} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}
\end{aligned}
$$

Example 2

$f(\boldsymbol{x}) = \mathbf{1}^T \tan(\boldsymbol{A}\boldsymbol{x})$. What is $\mathsf{D}f(\boldsymbol{x})$?

Here $\boldsymbol{x}$ is $p \times 1$, $\boldsymbol{A}$ is $n \times p$, and

$\mathbf{1} = n \times 1$ vector of ones.

For $\tan(\boldsymbol{A}\boldsymbol{x})$ we're using the convention (as in e.g. Matlab and R)

$$
\tan \begin{bmatrix} 0.2 \\ 7.1 \\ -0.9 \end{bmatrix} = \begin{bmatrix} \tan(0.2) \\ \tan(7.1) \\ \tan(-0.9) \end{bmatrix}
$$

Traditional Answer

$$
\begin{aligned}
f(\boldsymbol{x}) &= \mathbf{1}^T \tan(\boldsymbol{A}\boldsymbol{x}) \\
&= \sum_{i=1}^{n} \tan((\boldsymbol{A}\boldsymbol{x})_i) \\
&= \sum_{i=1}^{n} \tan\left(\sum_{j=1}^{p} A_{ij}x_j\right)
\end{aligned}
$$

For $1 \le k \le p$,

$$
\begin{aligned}
\frac{\partial}{\partial x_k} f(\boldsymbol{x}) &= \sum_{i=1}^{n} \sec^2\left(\sum_{j=1}^{p} A_{ij}x_j\right) A_{ik} \\
&= \sum_{i=1}^{n} \sec^2\left((\boldsymbol{A}\boldsymbol{x})_i\right) A_{ik} \\
&= [\sec^2(\boldsymbol{A}\boldsymbol{x})^T\boldsymbol{A}]_k \\
\Longrightarrow \quad \mathsf{D}\, f(\boldsymbol{x}) &= \sec^2(\boldsymbol{A}\boldsymbol{x})^T\boldsymbol{A}.
\end{aligned}
$$

Drawback of traditional approach:

Need to undo the matrix algebra and then put it back together!

A more efficient approach is...

Vector Calculus Answer

$$\begin{aligned} d\mathbf{1}^T \tan(\boldsymbol{A}\boldsymbol{x}) &= \mathbf{1}^T d\tan(\boldsymbol{A}\boldsymbol{x}) \\ &= \mathbf{1}^T \text{diag}\{\tan'(\boldsymbol{A}\boldsymbol{x})\} d(\boldsymbol{A}\boldsymbol{x}) \\ &= \mathbf{1}^T \text{diag}\{\sec^2(\boldsymbol{A}\boldsymbol{x})\} d(\boldsymbol{A}\boldsymbol{x}) \end{aligned}$$

The 'diag' notation is as follows:

$$\text{diag}\begin{bmatrix} 4 \\ 17 \\ 11 \end{bmatrix} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 17 & 0 \\ 0 & 0 & 11 \end{bmatrix}.$$

Collecting the first two matrices under the transpose operator we get

$$\begin{aligned} d\mathbf{1}^T \tan(\boldsymbol{A}\boldsymbol{x}) &= [\text{diag}\{\sec^2(\boldsymbol{A}\boldsymbol{x})\}\mathbf{1}]^T \boldsymbol{A}\, d\boldsymbol{x} \\ &= \sec^2(\boldsymbol{A}\boldsymbol{x})^T \boldsymbol{A}\, d\boldsymbol{x} \end{aligned}$$

using the rule $\text{diag}(\boldsymbol{a})\mathbf{1} = \boldsymbol{a}$.

Thus

$$\text{D} f(\boldsymbol{x}) = \sec^2(\boldsymbol{A}\boldsymbol{x})^T \boldsymbol{A}.$$

Hessian matrices

If $f : \mathbb{R}^3 \to \mathbb{R}$ then the Hessian matrix of f (with respect to $\boldsymbol{x}$) is

$$\mathsf{H}\, f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1^2} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_2} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_3} \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2^2} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2 \partial x_3} \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_3 \partial x_1} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_3 \partial x_2} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_3^2} \end{bmatrix}$$

Second Identification Theorem

$$f : \mathbb{R}^p \to \mathbb{R}$$

Suppose $\boldsymbol{A}$ ($p \times p$) is such that

$$d^2\, f(\boldsymbol{x}) = (d\boldsymbol{x})^T \boldsymbol{A}\, d\boldsymbol{x}.$$

Then

$$\mathsf{H}\, f(\boldsymbol{x}) = \boldsymbol{A}.$$

Example

Return to example with $f(\boldsymbol{x}) = \mathbf{1}^T \tan(\boldsymbol{A}\boldsymbol{x})$.

We have previously shown that the first differential is:

$$d\, f(\boldsymbol{x}) = \sec^2(\boldsymbol{A}\boldsymbol{x})^T \boldsymbol{A}\, d\boldsymbol{x}.$$

The second differential is

$$\begin{aligned} d^2\, f(\boldsymbol{x}) &= d\{\sec^2(\boldsymbol{A}\boldsymbol{x})^T \boldsymbol{A}\, d\boldsymbol{x}\} \\ &= \{d\, \sec^2(\boldsymbol{A}\boldsymbol{x})\}^T \boldsymbol{A}\, d\boldsymbol{x} \end{aligned}$$

Note that in this last step, both $\boldsymbol{A}$ and $d\boldsymbol{x}$ are treated as constants.

$$
\begin{aligned}
d^2 f(\boldsymbol{x}) &= [\mathsf{diag}\{(\sec^2)'(\boldsymbol{A}\boldsymbol{x})\}\boldsymbol{A}\, d\boldsymbol{x}]^T \boldsymbol{A}\, d\boldsymbol{x} \\
&= (d\boldsymbol{x})^T \boldsymbol{A}^T \mathsf{diag}\{(\sec^2)'(\boldsymbol{A}\boldsymbol{x})\}\boldsymbol{A}\, d\boldsymbol{x}
\end{aligned}
$$

So by the Second Identification Theorem

$$
\mathsf{H}\, f(\boldsymbol{x}) = \boldsymbol{A}^T \mathsf{diag}\{(\sec^2)'(\boldsymbol{A}\boldsymbol{x})\}\boldsymbol{A}.
$$

The last step is to simplify $(\sec^2)'(\boldsymbol{A}\boldsymbol{x})$.

$$
\begin{aligned}
\frac{d}{du}\sec^2(u) &= \frac{d}{du}\cos(u)^{-2} \\
&= \text{mini Class Exercise} \\
&= -2\cos(u)^{-3}(-1)\sin(u) \\
&= 2\cos(u)^{-2}\{\sin(u)/\cos(u)\} \\
&= 2\sec^2(u)\tan(u) \\
\Longrightarrow \quad (\sec^2)'(u) &= 2\sec^2(u)\tan(u) \\
\Longrightarrow \quad \mathsf{H}\, f(\boldsymbol{x}) &= 2\boldsymbol{A}^T \mathsf{diag}\{\sec^2(\boldsymbol{A}\boldsymbol{x}) \odot \tan(\boldsymbol{A}\boldsymbol{x})\}\boldsymbol{A}
\end{aligned}
$$

where $\boldsymbol{a} \odot \boldsymbol{b}$ denotes element-wise product of vectors $\boldsymbol{a}$ and $\boldsymbol{b}$.

♦ ♦ ♦