

Vector and Matrix Calculus

Herman Kamper
kamperh@gmail.com

30 January 2013

1 Introduction

As explained in detail in [1], there unfortunately exists multiple competing notations concerning the layout of matrix derivatives. This can cause a lot of difficulty when consulting several sources, since different sources might use different conventions. Some sources, for example [2] (from which I use a lot of identities), even use a mixed layout (according to [1, Notes]). Identities for both the *numerator layout* (sometimes called the *Jacobian formulation*) and the *denominator layout* (sometimes called the *Hessian formulation*) is given in [1], so this makes it easy to check what layout a particular source uses. I will aim to stick to the denominator layout, which seems to be the most widely used in the field of statistics and pattern recognition (e.g. [3] and [4, pp. 327–332]). Other useful references concerning matrix calculus include [5] and [6]. In this document column vectors are assumed in all cases except where specifically stated otherwise.

Table 1: Derivatives of scalars, vector functions and matrices [1, 6].

	scalar y	column vector $\mathbf{y} \in \mathbb{R}^m$	matrix $\mathbf{Y} \in \mathbb{R}^{m \times n}$
scalar x	scalar $\frac{\partial y}{\partial x}$	row vector $\frac{\partial \mathbf{y}}{\partial x} \in \mathbb{R}^m$	matrix $\frac{\partial \mathbf{Y}}{\partial x}$ (only numerator layout)
column vector $\mathbf{x} \in \mathbb{R}^n$	column vector $\frac{\partial y}{\partial \mathbf{x}} \in \mathbb{R}^n$	matrix $\frac{\partial \mathbf{y}}{\partial \mathbf{x}} \in \mathbb{R}^{n \times m}$	
matrix $\mathbf{X} \in \mathbb{R}^{p \times q}$	matrix $\frac{\partial y}{\partial \mathbf{X}} \in \mathbb{R}^{p \times q}$		

2 Definitions

Table 1 indicates the six possible kinds of derivatives when using the denominator layout. Using this layout notation consistently, we have the following definitions.

The derivative of a scalar function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ with respect to vector $\mathbf{x} \in \mathbb{R}^n$ is

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \stackrel{\text{def}}{=} \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix} \quad (1)$$

This is the transpose of the gradient (some authors simply call this the gradient, irrespective of whether numerator or denominator layout is used).

The derivative of a vector function $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$, where $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}) \ f_2(\mathbf{x}) \ \dots \ f_m(\mathbf{x})]^\top$ and $\mathbf{x} \in \mathbb{R}^n$, with respect to scalar x_i is

$$\frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_i} \stackrel{\text{def}}{=} \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_i} & \frac{\partial f_2(x)}{\partial x_i} & \dots & \frac{\partial f_m(x)}{\partial x_i} \end{bmatrix} \quad (2)$$

The derivative of a vector function $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$, where $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}) \ f_2(\mathbf{x}) \ \dots \ f_m(\mathbf{x})]^\top$, with respect to vector $\mathbf{x} \in \mathbb{R}^n$ is

$$\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} \stackrel{\text{def}}{=} \begin{bmatrix} \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_1} \\ \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \frac{\partial f_2(\mathbf{x})}{\partial x_1} & \dots & \frac{\partial f_m(\mathbf{x})}{\partial x_1} \\ \frac{\partial f_1(\mathbf{x})}{\partial x_2} & \frac{\partial f_2(\mathbf{x})}{\partial x_2} & \dots & \frac{\partial f_m(\mathbf{x})}{\partial x_2} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_1(\mathbf{x})}{\partial x_n} & \frac{\partial f_2(\mathbf{x})}{\partial x_n} & \dots & \frac{\partial f_m(\mathbf{x})}{\partial x_n} \end{bmatrix} \quad (3)$$

This is just the transpose of the Jacobian matrix.

The derivative of a scalar function $f : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ with respect to matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ is

$$\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} \stackrel{\text{def}}{=} \begin{bmatrix} \frac{\partial f(\mathbf{X})}{\partial X_{11}} & \frac{\partial f(\mathbf{X})}{\partial X_{12}} & \dots & \frac{\partial f(\mathbf{X})}{\partial X_{1n}} \\ \frac{\partial f(\mathbf{X})}{\partial X_{21}} & \frac{\partial f(\mathbf{X})}{\partial X_{22}} & \dots & \frac{\partial f(\mathbf{X})}{\partial X_{2n}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f(\mathbf{X})}{\partial X_{m1}} & \frac{\partial f(\mathbf{X})}{\partial X_{m2}} & \dots & \frac{\partial f(\mathbf{X})}{\partial X_{mn}} \end{bmatrix} \quad (4)$$

Observe that the (1) is just a special case of (4) for column vectors. Often (as in [3]) the gradient notation is used as an alternative to the notation used above, for example:

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \quad (5)$$

$$\nabla_{\mathbf{X}} f(\mathbf{X}) = \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} \quad (6)$$

3 Identities

3.1 Scalar-by-vector product rule

If $\mathbf{a} \in \mathbb{R}^m$, $\mathbf{b} \in \mathbb{R}^n$ and $\mathbf{C} \in \mathbb{R}^{m \times n}$ then

$$\mathbf{a}^\top \mathbf{C} \mathbf{b} = \sum_{i=1}^m a_i (\mathbf{C} \mathbf{b})_i = \sum_{i=1}^m a_i \left(\sum_{j=1}^n C_{ij} b_j \right) = \sum_{i=1}^m \sum_{j=1}^n C_{ij} a_i b_j \quad (7)$$

Now assume we have vector functions $\mathbf{u} : \mathbb{R}^m \rightarrow \mathbb{R}^m$, $\mathbf{v} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ and $\mathbf{A} \in \mathbb{R}^{m \times n}$. The vector functions \mathbf{u} and \mathbf{v} are functions of $\mathbf{x} \in \mathbb{R}^q$, but \mathbf{A} is not. We want to find an identity for

$$\frac{\partial \mathbf{u}^\top \mathbf{A} \mathbf{v}}{\partial \mathbf{x}} \quad (8)$$

From (7), we have:

$$\begin{aligned}
\left[\frac{\partial \mathbf{u}^T \mathbf{A} \mathbf{v}}{\partial \mathbf{x}} \right]_l &= \frac{\partial \mathbf{u}^T \mathbf{A} \mathbf{v}}{\partial x_l} = \frac{\partial}{\partial x_l} \sum_{i=1}^m \sum_{j=1}^n A_{ij} u_i v_j \\
&= \sum_{i=1}^m \sum_{j=1}^n A_{ij} \frac{\partial}{\partial x_l} u_i v_j \\
&= \sum_{i=1}^m \sum_{j=1}^n A_{ij} \left[v_j \frac{\partial u_i}{\partial x_l} + u_i \frac{\partial v_j}{\partial x_l} \right] \\
&= \sum_{i=1}^m \sum_{j=1}^n A_{ij} v_j \frac{\partial u_i}{\partial x_l} + \sum_{i=1}^m \sum_{j=1}^n A_{ij} u_i \frac{\partial v_j}{\partial x_l}
\end{aligned} \tag{9}$$

Now we can show (by writing out the elements [Notebook, 2012-05-22]) that:

$$\begin{aligned}
\left[\frac{\partial \mathbf{u}}{\partial \mathbf{x}} \mathbf{A} \mathbf{v} + \frac{\partial \mathbf{v}}{\partial \mathbf{x}} \mathbf{A}^T \mathbf{u} \right]_l &= \sum_{i=1}^m \sum_{j=1}^n A_{ij} v_j \frac{\partial u_i}{\partial x_l} + \sum_{i=1}^m \sum_{j=1}^n (\mathbf{A}^T)_{ji} u_i \frac{\partial v_j}{\partial x_l} \\
&= \sum_{i=1}^m \sum_{j=1}^n A_{ij} v_j \frac{\partial u_i}{\partial x_l} + \sum_{i=1}^m \sum_{j=1}^n A_{ij} u_i \frac{\partial v_j}{\partial x_l}
\end{aligned} \tag{10}$$

A comparison of (9) and (10) completes the proof that

$$\boxed{\frac{\partial \mathbf{u}^T \mathbf{A} \mathbf{v}}{\partial \mathbf{x}} = \frac{\partial \mathbf{u}}{\partial \mathbf{x}} \mathbf{A} \mathbf{v} + \frac{\partial \mathbf{v}}{\partial \mathbf{x}} \mathbf{A}^T \mathbf{u}} \tag{11}$$

3.2 Useful identities from scalar-by-vector product rule

From (11) it follows, with vectors and matrices $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{d} \in \mathbb{R}^q$, $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{B} \in \mathbb{R}^{m \times n}$, $\mathbf{C} \in \mathbb{R}^{m \times q}$, $\mathbf{D} \in \mathbb{R}^{q \times n}$, that

$$\frac{\partial (\mathbf{B} \mathbf{x} + \mathbf{b})^T \mathbf{C} (\mathbf{D} \mathbf{x} + \mathbf{d})}{\partial \mathbf{x}} = \frac{\partial (\mathbf{A} \mathbf{x} + \mathbf{b})}{\partial \mathbf{x}} \mathbf{C} (\mathbf{D} \mathbf{x} + \mathbf{d}) + \frac{\partial (\mathbf{D} \mathbf{x} + \mathbf{d})^T}{\partial \mathbf{x}} \mathbf{C}^T (\mathbf{B} \mathbf{x} + \mathbf{b}) \tag{12}$$

resulting in the identity:

$$\boxed{\frac{\partial (\mathbf{B} \mathbf{x} + \mathbf{b})^T \mathbf{C} (\mathbf{D} \mathbf{x} + \mathbf{d})}{\partial \mathbf{x}} = \mathbf{B}^T \mathbf{C} (\mathbf{D} \mathbf{x} + \mathbf{d}) + \mathbf{D}^T \mathbf{C}^T (\mathbf{B} \mathbf{x} + \mathbf{b})} \tag{13}$$

by using the easily verifiable identities:

$$\boxed{\frac{\partial (\mathbf{u}(\mathbf{x}) + \mathbf{v}(\mathbf{x}))}{\partial \mathbf{x}} = \frac{\partial \mathbf{u}(\mathbf{x})}{\partial \mathbf{x}} + \frac{\partial \mathbf{v}(\mathbf{x})}{\partial \mathbf{x}}} \tag{14}$$

$$\boxed{\frac{\partial \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}^T} \tag{15}$$

$$\boxed{\frac{\partial \mathbf{a}}{\partial \mathbf{x}} = \mathbf{0}} \tag{16}$$

Some other useful special cases of (11):

$$\boxed{\frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{b}}{\partial \mathbf{x}} = \mathbf{A} \mathbf{b}} \tag{17}$$

$$\boxed{\frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^T) \mathbf{x}} \quad (18)$$

$$\boxed{\frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A} \mathbf{x} \text{ if } \mathbf{A} \text{ is symmetric}} \quad (19)$$

3.3 Derivatives of determinant

See [7, p. 374] for definition of cofactors. Also see [Notebook, 2012-05-22].

We can write the determinant of matrix $\mathbf{X} \in \mathbb{R}^{n \times n}$ as

$$|\mathbf{X}| = X_{i1}C_{i1} + X_{i2}C_{i2} + \dots + X_{in}C_{in} = \sum_{j=1}^n X_{ij}C_{ij} \quad (20)$$

Thus the derivative will be

$$\begin{aligned} \left[\frac{\partial |\mathbf{X}|}{\partial \mathbf{X}} \right]_{kl} &= \frac{\partial}{\partial X_{kl}} \{X_{i1}C_{i1} + X_{i2}C_{i2} + \dots + X_{in}C_{in}\} \\ &= \frac{\partial}{\partial X_{kl}} \{X_{k1}C_{k1} + X_{k2}C_{k2} + \dots + X_{kn}C_{kn}\} \\ &\quad \text{(can choose } i \text{ any number, so choose } i = k\text{)} \\ &= C_{kl} \end{aligned} \quad (21)$$

Thus (see [7, p. 386])

$$\frac{\partial |\mathbf{X}|}{\partial \mathbf{X}} = \text{cofactor } \mathbf{X} = (\text{adj } \mathbf{X})^T \quad (22)$$

But we know that the inverse of \mathbf{X} is given by [7, p. 387]

$$\mathbf{X}^{-1} = \frac{1}{|\mathbf{X}|} \text{adj } \mathbf{X} \quad (23)$$

thus

$$\text{adj } \mathbf{X} = |\mathbf{X}| \mathbf{X}^{-1} \quad (24)$$

which, when substituted into (22), results in the identity

$$\boxed{\frac{\partial |\mathbf{X}|}{\partial \mathbf{X}} = |\mathbf{X}| (\mathbf{X}^{-1})^T} \quad (25)$$

From (25) we can also write

$$\left[\frac{\partial \ln |\mathbf{X}|}{\partial \mathbf{X}} \right]_{kl} = \frac{\partial \ln |\mathbf{X}|}{\partial X_{kl}} = \frac{1}{|\mathbf{X}|} \frac{\partial |\mathbf{X}|}{\partial \mathbf{X}} = \frac{1}{|\mathbf{X}|} |\mathbf{X}| (\mathbf{X}^{-1})^T \quad (26)$$

giving the identity

$$\boxed{\frac{\partial \ln |\mathbf{X}|}{\partial \mathbf{X}} = (\mathbf{X}^{-1})^T} \quad (27)$$

References

- [1] Matrix calculus. [Online]. Available: http://en.wikipedia.org/wiki/Matrix_calculus
- [2] K. B. Petersen and M. S. Pedersen, “The matrix cookbook,” 2008.
- [3] A. Ng, *Machine Learning*. Class notes for CS229, Stanford Engineering Everywhere, Stanford University, 2008. [Online]. Available: <http://see.stanford.edu>
- [4] S. R. Searle, *Matrix Algebra Useful for Statistics*. New York, NY: John Wiley & Sons, 1982.
- [5] J. R. Schott, *Matrix Analysis for Statistics*. New York, NY: John Wiley & Sons, 1996.
- [6] T. P. Minka, “Old and new matrix algebra useful for statistics,” 2000. [Online]. Available: <http://research.microsoft.com/en-us/um/people/minka/papers/matrix>
- [7] D. G. Zill and M. R. Cullen, *Advanced Engineering Mathematics*, 3rd ed. Jones and Bartlett, 2006.