

Matrix Differentiation

(and some other stuff)

Randal J. Barnes
Department of Civil Engineering, University of Minnesota
Minneapolis, Minnesota, USA

1 Introduction

Throughout this presentation I have chosen to use a *symbolic matrix notation*. This choice was not made lightly. I am a strong advocate of index notation, when appropriate. For example, index notation greatly simplifies the presentation and manipulation of differential geometry. As a rule-of-thumb, if your work is going to primarily involve differentiation with respect to the spatial coordinates, then index notation is almost surely the appropriate choice.

In the present case, however, I will be manipulating large systems of equations in which the matrix calculus is relatively simply while the matrix algebra and matrix arithmetic is messy and more involved. Thus, I have chosen to use symbolic notation.

2 Notation and Nomenclature 术语

Definition 1 Let $\mathbf{a}_{ij} \in \mathfrak{R}$, $i = 1, 2, \dots, \mathbf{m}$, $j = 1, 2, \dots, \mathbf{n}$. Then the ordered rectangular array

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_{11} & \mathbf{a}_{12} & \cdots & \mathbf{a}_{1\mathbf{n}} \\ \mathbf{a}_{21} & \mathbf{a}_{22} & \cdots & \mathbf{a}_{2\mathbf{n}} \\ \vdots & \vdots & & \vdots \\ \mathbf{a}_{\mathbf{m}1} & \mathbf{a}_{\mathbf{m}2} & \cdots & \mathbf{a}_{\mathbf{m}\mathbf{n}} \end{bmatrix} \quad (1)$$

is said to be a real *matrix* of dimension $\mathbf{m} \times \mathbf{n}$.

When writing a matrix I will occasionally write down its typical element as well as its dimension. Thus,

$$\mathbf{A} = [\mathbf{a}_{ij}], \quad i = 1, 2, \dots, \mathbf{m}; j = 1, 2, \dots, \mathbf{n}, \quad (2)$$

denotes a matrix with \mathbf{m} rows and \mathbf{n} columns, whose typical element is \mathbf{a}_{ij} . Note, the first subscript locates the *row* in which the typical element lies while the second subscript locates the *column*. For example, \mathbf{a}_{jk} denotes the element lying in the j th row and k th column of the matrix \mathbf{A} .

Definition 2 A *vector* is a matrix with only one column. Thus, all vectors are inherently column vectors.

Convention 1

Multi-column matrices are denoted by boldface uppercase letters: for example, **A, B, X**. Vectors (single-column matrices) are denoted by boldfaced lowercase letters: for example, **a, b, x**. I will attempt to use letters from the beginning of the alphabet to designate known matrices, and letters from the end of the alphabet for unknown or variable matrices.

Convention 2

When it is useful to explicitly attach the matrix dimensions to the symbolic notation, I will use an underscript. For example, $\mathbf{A}_{m \times n}$, indicates a known, multi-column matrix with m rows and n columns.

A superscript T denotes the matrix transpose operation; for example, \mathbf{A}^T denotes the transpose of \mathbf{A} . Similarly, if \mathbf{A} has an inverse it will be denoted by \mathbf{A}^{-1} . The determinant of \mathbf{A} will be denoted by either $|\mathbf{A}|$ or $\det(\mathbf{A})$. Similarly, the rank of a matrix \mathbf{A} is denoted by $\text{rank}(\mathbf{A})$. An identity matrix will be denoted by \mathbf{I} , and $\mathbf{0}$ will denote a null matrix.

3 Matrix Multiplication

Definition 3 Let \mathbf{A} be $m \times n$, and \mathbf{B} be $n \times p$, and let the product \mathbf{AB} be

$$\mathbf{C} = \mathbf{AB} \quad (3)$$

then \mathbf{C} is a $m \times p$ matrix, with element (i, j) given by

$$c_{ij} = \sum_{k=1}^n a_{ik} b_{kj} \quad (4)$$

for all $i = 1, 2, \dots, m$, $j = 1, 2, \dots, p$.

Proposition 1 Let \mathbf{A} be $m \times n$, and \mathbf{x} be $n \times 1$, then the typical element of the product

$$\mathbf{z} = \mathbf{Ax} \quad (5)$$

is given by

$$z_i = \sum_{k=1}^n a_{ik} x_k \quad (6)$$

for all $i = 1, 2, \dots, m$. Similarly, let \mathbf{y} be $m \times 1$, then the typical element of the product

$$\mathbf{z}^T = \mathbf{y}^T \mathbf{A} \quad (7)$$

is given by

$$z_i = \sum_{k=1}^n a_{ki} y_k \quad (8)$$

for all $i = 1, 2, \dots, n$. Finally, the scalar resulting from the product

$$\alpha = \mathbf{y}^T \mathbf{Ax} \quad (9)$$

is given by

$$\alpha = \sum_{j=1}^m \sum_{k=1}^n a_{jk} y_j x_k \quad (10)$$

Proof: These are merely direct applications of Definition 3. q.e.d.

Proposition 2 Let \mathbf{A} be $m \times n$, and \mathbf{B} be $n \times p$, and let the product \mathbf{AB} be

$$\mathbf{C} = \mathbf{AB} \quad \text{Dim}(\mathbf{C}) = m \times p \quad (11)$$

then

$$\mathbf{C}^T = \mathbf{B}^T \mathbf{A}^T \quad (12)$$

Proof: The typical element of \mathbf{C} is given by

$$c_{ij} = \sum_{k=1}^n a_{ik} b_{kj} \quad (13)$$

By definition, the typical element of \mathbf{C}^T , say d_{ij} , is given by

$$d_{ij} = c_{ji} = \sum_{k=1}^n a_{jk} b_{ki} \quad (14)$$

Hence,

$$\mathbf{C}^T = \mathbf{B}^T \mathbf{A}^T \quad (15)$$

q.e.d.

Proposition 3 Let \mathbf{A} and \mathbf{B} be $n \times n$ and invertible matrices. Let the product \mathbf{AB} be given by

$$\mathbf{C} = \mathbf{AB} \quad (16)$$

then

$$\mathbf{C}^{-1} = \mathbf{B}^{-1} \mathbf{A}^{-1} \quad (17)$$

Proof:

$$\mathbf{CB}^{-1} \mathbf{A}^{-1} = \mathbf{ABB}^{-1} \mathbf{A}^{-1} = \mathbf{I} \quad (18)$$

q.e.d.

4 Partitioned Matrices

Frequently, I will find it convenient to deal with *partitioned matrices*¹. Such a representation, and the manipulation of this representation, are two of the relative advantages of the symbolic matrix notation.

Definition 4 Let \mathbf{A} be $m \times n$ and write

$$\mathbf{A} = \begin{bmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{D} & \mathbf{E} \end{bmatrix} \quad (19)$$

where \mathbf{B} is $m_1 \times n_1$, \mathbf{E} is $m_2 \times n_2$, \mathbf{C} is $m_1 \times n_2$, \mathbf{D} is $m_2 \times n_1$, $m_1 + m_2 = m$, and $n_1 + n_2 = n$. The above is said to be a *partition* of the matrix \mathbf{A} .

¹Much of the material in this section is extracted directly from Dhrymes (1978, Section 2.7).

Proposition 4 Let \mathbf{A} be a square, nonsingular matrix of order \mathbf{m} . Partition \mathbf{A} as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \quad (20)$$

so that \mathbf{A}_{11} is a nonsingular matrix of order \mathbf{m}_1 , \mathbf{A}_{22} is a nonsingular matrix of order \mathbf{m}_2 , and $\mathbf{m}_1 + \mathbf{m}_2 = \mathbf{m}$. Then

$$\mathbf{A}^{-1} = \begin{bmatrix} (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1} & (\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1} \end{bmatrix} \quad (21)$$

Proof: Direct multiplication of the proposed \mathbf{A}^{-1} and \mathbf{A} yields

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I} \quad (22)$$

q.e.d.

5 Matrix Differentiation

In the following discussion I will differentiate matrix quantities with respect to the elements of the referenced matrices. Although no new concept is required to carry out such operations, the element-by-element calculations involve cumbersome manipulations and, thus, it is useful to derive the necessary results and have them readily available ².

Convention 3

Let

$$\mathbf{y} = \psi(\mathbf{x}), \quad (23)$$

where \mathbf{y} is an \mathbf{m} -element vector, and \mathbf{x} is an \mathbf{n} -element vector. The symbol

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \dots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \dots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \dots & \frac{\partial y_m}{\partial x_n} \end{bmatrix} \quad (24)$$

will denote the $\mathbf{m} \times \mathbf{n}$ matrix of first-order partial derivatives of the transformation from \mathbf{x} to \mathbf{y} . Such a matrix is called the **Jacobian matrix** of the transformation $\psi(\cdot)$.

Notice that if \mathbf{x} is actually a scalar in Convention 3 then the resulting Jacobian matrix is a $\mathbf{m} \times 1$ matrix; that is, a single column (a vector). On the other hand, if \mathbf{y} is actually a scalar in Convention 3 then the resulting Jacobian matrix is a $1 \times \mathbf{n}$ matrix; that is, a single row (the transpose of a vector).

Proposition 5 Let

$$\mathbf{y} = \mathbf{Ax} \quad (25)$$

²Much of the material in this section is extracted directly from Dhrymes (1978, Section 4.3). The interested reader is directed to this worthy reference to find additional results.

where \mathbf{y} is $m \times 1$, \mathbf{x} is $n \times 1$, \mathbf{A} is $m \times n$, and \mathbf{A} does not depend on \mathbf{x} , then

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \mathbf{A} \quad (26)$$

Proof: Since the i th element of \mathbf{y} is given by

$$y_i = \sum_{k=1}^n a_{ik} x_k \quad (27)$$

it follows that

$$\frac{\partial y_i}{\partial x_j} = a_{ij} \quad (28)$$

for all $i = 1, 2, \dots, m$, $j = 1, 2, \dots, n$. Hence

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \mathbf{A} \quad (29)$$

q.e.d.

Proposition 6 Let

$$\mathbf{y} = \mathbf{A}\mathbf{x} \quad (30)$$

where \mathbf{y} is $m \times 1$, \mathbf{x} is $n \times 1$, \mathbf{A} is $m \times n$, and \mathbf{A} does not depend on \mathbf{x} , as in Proposition 5. Suppose that \mathbf{x} is a function of the vector \mathbf{z} , while \mathbf{A} is independent of \mathbf{z} . Then

$$\frac{\partial \mathbf{y}}{\partial \mathbf{z}} = \mathbf{A} \frac{\partial \mathbf{x}}{\partial \mathbf{z}} \quad (31)$$

Proof: Since the i th element of \mathbf{y} is given by

$$y_i = \sum_{k=1}^n a_{ik} x_k \quad (32)$$

for all $i = 1, 2, \dots, m$, it follows that

$$\frac{\partial y_i}{\partial z_j} = \sum_{k=1}^n a_{ik} \frac{\partial x_k}{\partial z_j} \quad (33)$$

but the right hand side of the above is simply element (i, j) of $\mathbf{A} \frac{\partial \mathbf{x}}{\partial \mathbf{z}}$. Hence

$$\frac{\partial \mathbf{y}}{\partial \mathbf{z}} = \frac{\partial \mathbf{y}}{\partial \mathbf{x}} \frac{\partial \mathbf{x}}{\partial \mathbf{z}} = \mathbf{A} \frac{\partial \mathbf{x}}{\partial \mathbf{z}} \quad (34)$$

q.e.d.

Proposition 7 Let the scalar α be defined by

$$\alpha = \mathbf{y}^T \mathbf{A} \mathbf{x} \quad (35)$$

where \mathbf{y} is $m \times 1$, \mathbf{x} is $n \times 1$, \mathbf{A} is $m \times n$, and \mathbf{A} is independent of \mathbf{x} and \mathbf{y} , then

$$\frac{\partial \alpha}{\partial \mathbf{x}} = \mathbf{y}^T \mathbf{A} \quad (36)$$

and

$$\frac{\partial \alpha}{\partial \mathbf{y}} = \mathbf{x}^T \mathbf{A}^T \quad (37)$$

Proof: Define

$$\mathbf{w}^T = \mathbf{y}^T \mathbf{A} \quad (38)$$

and note that

$$\alpha = \mathbf{w}^T \mathbf{x} \quad (39)$$

Hence, by Proposition 5 we have that

$$\frac{\partial \alpha}{\partial \mathbf{x}} = \mathbf{w}^T = \mathbf{y}^T \mathbf{A} \quad (40)$$

which is the first result. Since α is a scalar, we can write

$$\alpha = \alpha^T = \mathbf{x}^T \mathbf{A}^T \mathbf{y} \quad (41)$$

α is a Scalar
 $\alpha = \mathbf{y}^T \cdot \mathbf{A} \cdot \mathbf{x}$
 $\alpha^T = \mathbf{x}^T \mathbf{A}^T \cdot \mathbf{y}$

and applying Proposition 5 as before we obtain

$$\frac{\partial \alpha}{\partial \mathbf{y}} = \mathbf{x}^T \mathbf{A}^T \quad (42)$$

q.e.d.

Proposition 8 For the special case in which the scalar α is given by the quadratic form

$$\alpha = \mathbf{x}^T \mathbf{A} \mathbf{x} \quad (43)$$

where \mathbf{x} is $n \times 1$, \mathbf{A} is $n \times n$, and \mathbf{A} does not depend on \mathbf{x} , then

$$\frac{\partial \alpha}{\partial \mathbf{x}} = \mathbf{x}^T (\mathbf{A} + \mathbf{A}^T) \quad (44)$$

Proof: By definition

$$\alpha = \sum_{j=1}^n \sum_{i=1}^n a_{ij} x_i x_j \quad (45)$$

Differentiating with respect to the k th element of \mathbf{x} we have

$$\frac{\partial \alpha}{\partial x_k} = \sum_{j=1}^n a_{kj} x_j + \sum_{i=1}^n a_{ik} x_i \quad (46)$$

for all $k = 1, 2, \dots, n$, and consequently,

$$\frac{\partial \alpha}{\partial \mathbf{x}} = \mathbf{x}^T \mathbf{A}^T + \mathbf{x}^T \mathbf{A} = \mathbf{x}^T (\mathbf{A}^T + \mathbf{A}) \quad (47)$$

q.e.d.

Proposition 9 For the special case where **A** is a symmetric matrix and

$$\alpha = \mathbf{x}^T \mathbf{A} \mathbf{x} \quad (48)$$

where \mathbf{x} is $n \times 1$, \mathbf{A} is $n \times n$, and \mathbf{A} does not depend on \mathbf{x} , then

$$\frac{\partial \alpha}{\partial \mathbf{x}} = 2\mathbf{x}^T \mathbf{A} \quad (49)$$

Proof: This is an obvious application of Proposition 8. q.e.d.

Proposition 10 Let the scalar α be defined by

$$\alpha = \mathbf{y}^T \mathbf{x} \quad (50)$$

where \mathbf{y} is $n \times 1$, \mathbf{x} is $n \times 1$, and both \mathbf{y} and \mathbf{x} are functions of the vector \mathbf{z} . Then

$$\frac{\partial \alpha}{\partial \mathbf{z}} = \mathbf{x}^T \frac{\partial \mathbf{y}}{\partial \mathbf{z}} + \mathbf{y}^T \frac{\partial \mathbf{x}}{\partial \mathbf{z}} \quad (51)$$

Proof: We have

$$\alpha = \sum_{j=1}^n x_j y_j \quad (52)$$

Differentiating with respect to the k th element of \mathbf{z} we have

$$\frac{\partial \alpha}{\partial z_k} = \sum_{j=1}^n \left(x_j \frac{\partial y_j}{\partial z_k} + y_j \frac{\partial x_j}{\partial z_k} \right) \quad (53)$$

for all $k = 1, 2, \dots, n$, and consequently,

$$\frac{\partial \alpha}{\partial \mathbf{z}} = \frac{\partial \alpha}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial \mathbf{z}} + \frac{\partial \alpha}{\partial \mathbf{x}} \frac{\partial \mathbf{x}}{\partial \mathbf{z}} = \mathbf{x}^T \frac{\partial \mathbf{y}}{\partial \mathbf{z}} + \mathbf{y}^T \frac{\partial \mathbf{x}}{\partial \mathbf{z}} \quad (54)$$

q.e.d.

Proposition 11 Let the scalar α be defined by

$$\alpha = \mathbf{x}^T \mathbf{x} \quad (55)$$

where \mathbf{x} is $n \times 1$, and \mathbf{x} is a function of the vector \mathbf{z} . Then

$$\frac{\partial \alpha}{\partial \mathbf{z}} = 2\mathbf{x}^T \frac{\partial \mathbf{x}}{\partial \mathbf{z}} \quad (56)$$

Proof: This is an obvious application of Proposition 10. q.e.d.

Proposition 12 Let the scalar α be defined by

$$\alpha = \mathbf{y}^T \mathbf{A} \mathbf{x} \quad (57)$$

where \mathbf{y} is $m \times 1$, \mathbf{x} is $n \times 1$, \mathbf{A} is $m \times n$, and both \mathbf{y} and \mathbf{x} are functions of the vector \mathbf{z} , while \mathbf{A} does not depend on \mathbf{z} . Then

$$\frac{\partial \alpha}{\partial \mathbf{z}} = \mathbf{x}^T \mathbf{A}^T \frac{\partial \mathbf{y}}{\partial \mathbf{z}} + \mathbf{y}^T \mathbf{A} \frac{\partial \mathbf{x}}{\partial \mathbf{z}} \quad (58)$$

Proof: *Define*

$$\mathbf{w}^\top = \mathbf{y}^\top \mathbf{A} \quad (59)$$

and note that

$$\alpha = \mathbf{w}^\top \mathbf{x} \quad (60)$$

Applying Proposition 10 we have

$$\frac{\partial \alpha}{\partial \mathbf{z}} = \mathbf{x}^\top \frac{\partial \mathbf{w}}{\partial \mathbf{z}} + \mathbf{w}^\top \frac{\partial \mathbf{x}}{\partial \mathbf{z}} \quad (61)$$

Substituting back in for \mathbf{w} we arrive at

$$\frac{\partial \alpha}{\partial \mathbf{z}} = \frac{\partial \alpha}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial \mathbf{z}} + \frac{\partial \alpha}{\partial \mathbf{x}} \frac{\partial \mathbf{x}}{\partial \mathbf{z}} = \mathbf{x}^\top \mathbf{A}^\top \frac{\partial \mathbf{y}}{\partial \mathbf{z}} + \mathbf{y}^\top \mathbf{A} \frac{\partial \mathbf{x}}{\partial \mathbf{z}} \quad (62)$$

q.e.d.

Proposition 13 *Let the scalar α be defined by the quadratic form*

$$\alpha = \mathbf{x}^\top \mathbf{A} \mathbf{x} \quad (63)$$

where \mathbf{x} is $n \times 1$, \mathbf{A} is $n \times n$, and \mathbf{x} is a function of the vector \mathbf{z} , while \mathbf{A} does not depend on \mathbf{z} . Then

$$\frac{\partial \alpha}{\partial \mathbf{z}} = \mathbf{x}^\top (\mathbf{A} + \mathbf{A}^\top) \frac{\partial \mathbf{x}}{\partial \mathbf{z}} \quad (64)$$

Proof: *This is an obvious application of Proposition 12.* q.e.d.

Proposition 14 *For the special case where \mathbf{A} is a symmetric matrix and*

$$\alpha = \mathbf{x}^\top \mathbf{A} \mathbf{x} \quad (65)$$

where \mathbf{x} is $n \times 1$, \mathbf{A} is $n \times n$, and \mathbf{x} is a function of the vector \mathbf{z} , while \mathbf{A} does not depend on \mathbf{z} . Then

$$\frac{\partial \alpha}{\partial \mathbf{z}} = 2\mathbf{x}^\top \mathbf{A} \frac{\partial \mathbf{x}}{\partial \mathbf{z}} \quad (66)$$

Proof: *This is an obvious application of Proposition 13.* q.e.d.

Definition 5 Let \mathbf{A} be a $m \times n$ matrix whose elements are functions of the scalar parameter α . Then the derivative of the matrix \mathbf{A} with respect to the scalar parameter α is the $m \times n$ matrix of element-by-element derivatives:

$$\frac{\partial \mathbf{A}}{\partial \alpha} = \begin{bmatrix} \frac{\partial a_{11}}{\partial \alpha} & \frac{\partial a_{12}}{\partial \alpha} & \dots & \frac{\partial a_{1n}}{\partial \alpha} \\ \frac{\partial a_{21}}{\partial \alpha} & \frac{\partial a_{22}}{\partial \alpha} & \dots & \frac{\partial a_{2n}}{\partial \alpha} \\ \vdots & \vdots & & \vdots \\ \frac{\partial a_{m1}}{\partial \alpha} & \frac{\partial a_{m2}}{\partial \alpha} & \dots & \frac{\partial a_{mn}}{\partial \alpha} \end{bmatrix} \quad (67)$$

Proposition 15 *Let \mathbf{A} be a nonsingular, $m \times m$ matrix whose elements are functions of the scalar parameter α . Then*

$$\frac{\partial \mathbf{A}^{-1}}{\partial \alpha} = -\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial \alpha} \mathbf{A}^{-1} \quad (68)$$

Proof: *Start with the definition of the inverse*

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I} \quad (69)$$

and differentiate, yielding

$$\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial \alpha} + \frac{\partial \mathbf{A}^{-1}}{\partial \alpha}\mathbf{A} = \mathbf{0} \quad (70)$$

rearranging the terms yields

$$\frac{\partial \mathbf{A}^{-1}}{\partial \alpha} = -\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial \alpha}\mathbf{A}^{-1} \quad (71)$$

q.e.d.

6 References

- Dhrymes, Phoebus J., 1978, **Mathematics for Econometrics**, Springer-Verlag, New York, 136 pp.
- Golub, Gene H., and Charles F. Van Loan, 1983, **Matrix Computations**, Johns Hopkins University Press, Baltimore, Maryland, 476 pp.
- Graybill, Franklin A., 1983, **Matrices with Applications in Statistics**, 2nd Edition, Wadsworth International Group, Belmont, California, 461 pp.