

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/332131671>

# Matrix Calculus

Technical Report · April 2019

DOI: 10.13140/RG.2.2.27251.78888

---

CITATIONS

0

---

READS

80

1 author:



Sourya Dey

University of Southern California

14 PUBLICATIONS 43 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:



Exploring Complexity Reduction for Learning in Deep Neural Networks [View project](#)



A Nonuniform Sampling ADC architecture with reconfigurable digital anti-aliasing filter [View project](#)

# Matrix Calculus

Sourya Dey

## 1 Notation

- Scalars are written as lower case letters.
- Vectors are written as lower case bold letters, such as  $\mathbf{x}$ , and can be either row (dimensions  $1 \times n$ ) or column (dimensions  $n \times 1$ ). Column vectors are the default choice, unless otherwise mentioned. Individual elements are indexed by subscripts, such as  $x_i$  ( $i \in \{1, \dots, n\}$ ).
- Matrices are written as upper case bold letters, such as  $\mathbf{X}$ , and have dimensions  $m \times n$  corresponding to  $m$  rows and  $n$  columns. Individual elements are indexed by double subscripts for row and column, such as  $X_{ij}$  ( $i \in \{1, \dots, m\}, j \in \{1, \dots, n\}$ ).
- Occasionally higher order tensors occur, such as 3rd order with dimensions  $m \times n \times p$ , etc.

Note that a matrix is a 2nd order tensor. A row vector is a matrix with 1 row, and a column vector is a matrix with 1 column. A scalar is a matrix with 1 row and 1 column. Essentially, **scalars and vectors are special cases of matrices**.

The **derivative** of  $f$  with respect to  $x$  is  $\frac{\partial f}{\partial x}$ . Both  $x$  and  $f$  can be a scalar, vector, or matrix, leading to 9 types of derivatives. The **gradient** of  $f$  w.r.t  $x$  is  $\nabla_x f = \left(\frac{\partial f}{\partial x}\right)^T$ , i.e. **gradient is transpose of derivative**. The gradient at any point  $x_0$  in the domain has a physical interpretation, its direction is the direction of maximum increase of the function  $f$  at the point  $x_0$ , and its magnitude is the rate of increase in that direction. We do not generally deal with the gradient when  $x$  is a scalar.

## 2 Basic Rules

**This document follows numerator layout convention.** There is an alternative denominator layout convention, where several results are transposed. *Do not mix different layout conventions.*

We'll first state the most general matrix-matrix derivative type. All other types are simplifications since scalars and vectors are special cases of matrices. Consider a function  $\mathbf{F}(\cdot)$  which maps  $m \times n$  matrices to  $p \times q$  matrices, i.e. domain  $\subset \mathbb{R}^{m \times n}$  and range  $\subset \mathbb{R}^{p \times q}$ . So,  $\mathbf{F}(\cdot) : \underset{m \times n}{\mathbf{X}} \rightarrow \underset{p \times q}{\mathbf{F}(\mathbf{X})}$ . Its derivative  $\frac{\partial \mathbf{F}}{\partial \mathbf{X}}$  is a 4th order tensor of dimensions  $p \times q \times n \times m$ . This is an outer matrix of dimensions  $n \times m$  (transposed dimensions of the denominator  $\mathbf{X}$ ), with

each element being a  $p \times q$  inner matrix (same dimensions as the numerator  $\mathbf{F}$ ). It is given as:

$$\frac{\partial \mathbf{F}}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial \mathbf{F}}{\partial X_{1,1}} & \cdots & \frac{\partial \mathbf{F}}{\partial X_{m,1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mathbf{F}}{\partial X_{1,n}} & \cdots & \frac{\partial \mathbf{F}}{\partial X_{m,n}} \end{bmatrix} \quad (1a)$$

which has  $n$  rows and  $m$  columns, and the  $(i, j)$ th element is given as:

$$\frac{\partial \mathbf{F}}{\partial X_{i,j}} = \begin{bmatrix} \frac{\partial F_{1,1}}{\partial X_{i,j}} & \cdots & \frac{\partial F_{1,q}}{\partial X_{i,j}} \\ \vdots & \ddots & \vdots \\ \frac{\partial F_{p,1}}{\partial X_{i,j}} & \cdots & \frac{\partial F_{p,q}}{\partial X_{i,j}} \end{bmatrix} \quad (1b)$$

which has  $p$  rows and  $q$  columns.

Whew! Now that that's out of the way, let's get to some general rules (for the following,  $x$  and  $y$  can represent scalar, vector or matrix):

- **The derivative  $\frac{\partial y}{\partial x}$  always has outer matrix dimensions = transposed dimensions of denominator  $x$ , and each individual element (inner matrix) has dimensions = same dimensions of numerator  $y$ . If you do a calculation and the dimension doesn't come out right, the answer is not correct.**
- Derivatives usually obey the chain rule, i.e.  $\frac{\partial f(g(x))}{\partial x} = \frac{\partial f(g(x))}{\partial g(x)} \frac{\partial g(x)}{\partial x}$ .
- Derivatives usually obey the product rule, i.e.  $\frac{\partial f(x)g(x)}{\partial x} = f(x) \frac{\partial g(x)}{\partial x} + g(x) \frac{\partial f(x)}{\partial x}$ .

### 3 Types of derivatives

#### 3.1 Scalar by scalar

Nothing special here. The derivative is a scalar, and can also be written as  $f'(x)$ . For example, if  $f(x) = \sin x$ , then  $f'(x) = \cos x$ .

#### 3.2 Scalar by vector

$f(\cdot) : \underset{m \times 1}{\mathbf{x}} \rightarrow \underset{1 \times 1}{f(\mathbf{x})}$ . For this, the derivative is a  $1 \times m$  row vector:

$$\frac{\partial f}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_m} \end{bmatrix} \quad (2)$$

The gradient  $\nabla_{\mathbf{x}} f$  is its transposed column vector.

### 3.3 Vector by scalar

$\underline{f}(\cdot) : \underset{1 \times 1}{\mathbf{x}} \rightarrow \underset{n \times 1}{f(\mathbf{x})}$ . For this, the derivative is a  $n \times 1$  column vector:

$$\frac{\partial \underline{f}}{\partial x} = \begin{bmatrix} \frac{\partial f_1}{\partial x} \\ \frac{\partial f_2}{\partial x} \\ \vdots \\ \frac{\partial f_n}{\partial x} \end{bmatrix} \quad (3)$$

### 3.4 Vector by vector

$\underline{f}(\cdot) : \underset{m \times 1}{\mathbf{x}} \rightarrow \underset{n \times 1}{f(\mathbf{x})}$ . Derivative, also known as the **Jacobian**, is a matrix of dimensions  $n \times m$ . Its  $(i, j)$ th element is the scalar derivative of the  $i$ th output component w.r.t the  $j$ th input component, i.e.:

$$\frac{\partial \underline{f}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_m} \end{bmatrix} \quad (4)$$

#### 3.4.1 Special case – Vectorized scalar function

This is a scalar-scalar function applied element-wise to a vector, and is denoted by  $\underline{f}(\cdot) : \underset{m \times 1}{\mathbf{x}} \rightarrow \underset{m \times 1}{f(\mathbf{x})}$ . For example:

$$\underline{f} \left( \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \right) = \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_m) \end{bmatrix} \quad (5)$$

In this case, both the derivative and gradient are the same  $m \times m$  diagonal matrix, given as:

$$\nabla_{\mathbf{x}} \underline{f} = \frac{\partial \underline{f}}{\partial \mathbf{x}} = \begin{bmatrix} f'(x_1) & & & 0 \\ & f'(x_2) & & \\ & & \ddots & \\ 0 & & & f'(x_m) \end{bmatrix} \quad (6)$$

where  $f'(x_i) = \frac{\partial f(x_i)}{\partial x_i}$ .

**Note:** Some texts take the derivative of a vectorized scalar function by taking element-wise derivatives to get a  $m \times 1$  vector. To avoid confusion with (6), we will refer to this as  $\underline{f}'(\mathbf{x})$ .

$$\underline{f}'(\mathbf{x}) = \begin{bmatrix} f'(x_1) \\ f'(x_2) \\ \vdots \\ f'(x_m) \end{bmatrix} \quad (7)$$

To realize the effect of this, let's say we want to multiply the gradient from (6) with some  $m$ -dimensional vector  $\mathbf{a}$ . This would result in:

$$(\nabla_{\mathbf{x}} f) \mathbf{a} = \begin{bmatrix} f'(x_1) a_1 \\ f'(x_2) a_2 \\ \vdots \\ f'(x_m) a_m \end{bmatrix} \quad (8)$$

Achieving the same result with  $\underline{f}'(\mathbf{x})$  from (7) would require the *Hadamard product*  $\circ$ , defined as element-wise multiplication of 2 vectors:

$$\underline{f}'(\mathbf{x}) \circ \mathbf{a} = \begin{bmatrix} f'(x_1) a_1 \\ f'(x_2) a_2 \\ \vdots \\ f'(x_m) a_m \end{bmatrix} \quad (9)$$

### 3.4.2 Special Case – Hessian

Consider the type of function in Sec. 3.2, i.e.  $f(\cdot) : \underset{m \times 1}{\mathbf{x}} \rightarrow \underset{1 \times 1}{f(\mathbf{x})}$ . Its gradient is a vector-to-vector function given as  $\nabla_{\mathbf{x}} f(\cdot) : \underset{m \times 1}{\mathbf{x}} \rightarrow \underset{m \times 1}{\nabla_{\mathbf{x}} f(\mathbf{x})}$ . The transpose of its derivative is the Hessian:

$$\mathbf{H} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_m \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_m^2} \end{bmatrix} \quad (10)$$

i.e.  $\mathbf{H} = \left( \frac{\partial \nabla_{\mathbf{x}} f}{\partial \mathbf{x}} \right)^T$ . If derivatives are continuous, then  $\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$ , so the Hessian is symmetric.

### 3.5 Scalar by matrix

$f(\cdot) : \underset{m \times n}{\mathbf{X}} \rightarrow \underset{1 \times 1}{f(\mathbf{X})}$ . In this case, the derivative is a  $n \times m$  matrix:

$$\frac{\partial f}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial f}{\partial X_{1,1}} & \cdots & \frac{\partial f}{\partial X_{m,1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial X_{1,n}} & \cdots & \frac{\partial f}{\partial X_{m,n}} \end{bmatrix} \quad (11)$$

The gradient has the same dimensions as the input matrix, i.e.  $m \times n$ .

### 3.6 Matrix by scalar

$f(\cdot) : \underset{1 \times 1}{x} \rightarrow \underset{p \times q}{\mathbf{F}(x)}$ . In this case, the derivative is a  $p \times q$  matrix:

$$\frac{\partial \mathbf{F}}{\partial x} = \begin{bmatrix} \frac{\partial F_{1,1}}{\partial x} & \cdots & \frac{\partial F_{1,q}}{\partial x} \\ \vdots & \ddots & \vdots \\ \frac{\partial F_{p,1}}{\partial x} & \cdots & \frac{\partial F_{p,q}}{\partial x} \end{bmatrix} \quad (12)$$

### 3.7 Vector by matrix

$\mathbf{f}(\cdot) : \underset{m \times n}{\mathbf{X}} \rightarrow \underset{p \times 1}{\mathbf{f}(\mathbf{X})}$ . In this case, the derivative is a 3rd-order tensor with dimensions  $p \times n \times m$ .

This is the same  $n \times m$  matrix in (11), but with  $f$  replaced by the  $p$ -dimensional vector  $\mathbf{f}$ , i.e.:

$$\frac{\partial \mathbf{f}}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial \mathbf{f}}{\partial X_{1,1}} & \cdots & \frac{\partial \mathbf{f}}{\partial X_{m,1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mathbf{f}}{\partial X_{1,n}} & \cdots & \frac{\partial \mathbf{f}}{\partial X_{m,n}} \end{bmatrix} \quad (13)$$

### 3.8 Matrix by vector

$\mathbf{F}(\cdot) : \underset{m \times 1}{\mathbf{x}} \rightarrow \underset{p \times q}{\mathbf{F}(\mathbf{x})}$ . In this case, the derivative is a 3rd-order tensor with dimensions  $p \times q \times m$ .

This is the same  $m \times 1$  row vector in (2), but with  $f$  replaced by the  $p \times q$  matrix  $\mathbf{F}$ , i.e.:

$$\frac{\partial \mathbf{F}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial \mathbf{F}}{\partial x_1} & \frac{\partial \mathbf{F}}{\partial x_2} & \cdots & \frac{\partial \mathbf{F}}{\partial x_m} \end{bmatrix} \quad (14)$$

## 4 Operations and Examples

### 4.1 Commutation

If things normally don't commute (such as for matrices,  $\mathbf{AB} \neq \mathbf{BA}$ ), then order should be maintained when taking derivatives. If things normally commute (such as for vector inner product,  $\mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a}$ ), their order can be switched when taking derivatives. **Output dimensions must always come out right.**

For example, let  $\mathbf{f}(\mathbf{x}) = \underset{n \times 1}{(\mathbf{a}^T)} \underset{1 \times m}{\mathbf{x}} \underset{m \times 1}{\mathbf{b}}$ . The derivative  $\frac{\partial \mathbf{f}}{\partial \mathbf{x}}$  should be a  $n \times m$  matrix. Keeping order fixed, we get  $\frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \mathbf{a}^T \frac{\partial \mathbf{x}}{\partial \mathbf{x}} \mathbf{b} = \mathbf{a}^T \mathbf{I} \mathbf{b} = \mathbf{a}^T \mathbf{b}$ . This is a scalar, which is wrong! The solution? Note that  $(\mathbf{a}^T \mathbf{x})$  is a scalar, which can sit either to the right or the left of vector  $\mathbf{b}$ , i.e. ordering doesn't really matter. Rewriting  $\mathbf{f}(\mathbf{x}) = \mathbf{b} (\mathbf{a}^T \mathbf{x})$ , we get  $\frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \mathbf{b} \mathbf{a}^T \frac{\partial \mathbf{x}}{\partial \mathbf{x}} = \mathbf{b} \mathbf{a}^T \mathbf{I} = \mathbf{b} \mathbf{a}^T$ , which is the correct  $n \times m$  matrix.

If this seems confusing, it might be useful to take a simple example with low values for  $m$  and  $n$ , and write out the full derivative in matrix form as shown in (4). The resulting matrix will be  $\mathbf{b} \mathbf{a}^T$ .

### 4.2 Derivative of a transposed vector

The derivative of a transposed vector w.r.t itself is the identity matrix, but the transpose gets applied to everything *after*. For example, let  $f(\mathbf{w}) = (y - \mathbf{w}^T \mathbf{x})^2 = y^2 - (\mathbf{w}^T \mathbf{x}) y - y (\mathbf{w}^T \mathbf{x}) + (\mathbf{w}^T \mathbf{x}) (\mathbf{w}^T \mathbf{x})$ , where  $y$  and  $\mathbf{x}$  are not a function of  $\mathbf{w}$ . Taking derivative of the terms individually:

- $\frac{\partial y^2}{\partial \mathbf{w}} = \mathbf{0}^T$ , i.e. a row vector of all 0s.
- $\frac{\partial (\mathbf{w}^T \mathbf{x}) y}{\partial \mathbf{w}} = \frac{\partial \mathbf{w}^T}{\partial \mathbf{w}} \mathbf{x} y = (\mathbf{x} y)^T = y^T \mathbf{x}^T$ . Since  $y$  is a scalar, this is simply  $y \mathbf{x}^T$ .
- $\frac{\partial y (\mathbf{w}^T \mathbf{x})}{\partial \mathbf{w}} = y \frac{\partial \mathbf{w}^T}{\partial \mathbf{w}} \mathbf{x} = y \mathbf{x}^T$
- $\frac{\partial (\mathbf{w}^T \mathbf{x}) (\mathbf{w}^T \mathbf{x})}{\partial \mathbf{w}} = \frac{\partial \mathbf{w}^T}{\partial \mathbf{w}} \mathbf{x} (\mathbf{w}^T \mathbf{x}) + (\mathbf{w}^T \mathbf{x}) \frac{\partial \mathbf{w}^T}{\partial \mathbf{w}} \mathbf{x} = (\mathbf{x}^T \mathbf{w}) \mathbf{x}^T + (\mathbf{w}^T \mathbf{x}) \mathbf{x}^T$ . Since vector inner products commute, this is  $2 (\mathbf{w}^T \mathbf{x}) \mathbf{x}^T$ .

So  $\frac{\partial f}{\partial \mathbf{w}} = -2y \mathbf{x}^T + 2 (\mathbf{w}^T \mathbf{x}) \mathbf{x}^T$

### 4.3 Dealing with tensors

A tensor of dimensions  $p \times q \times n \times m$  (such as given in (1)) can be pre- and post-multiplied by vectors just like an ordinary matrix. *These vectors must be compatible with the inner matrices*

of dimensions  $p \times q$ , i.e. for each inner matrix, pre-multiply with a  $1 \times p$  row vector and post-multiply with a  $q \times 1$  column vector to get a scalar. This gives a final matrix of dimensions  $n \times m$ .

Example:  $f(\mathbf{W}) = \mathbf{a}^T \mathbf{W} \mathbf{b}$ . This is a scalar, so  $\frac{\partial f}{\partial \mathbf{W}}$  should be a matrix which has transposed dimensions as  $\mathbf{W}$ , i.e.  $n \times m$ . Now,  $\frac{\partial f}{\partial \mathbf{W}} = \mathbf{a}^T \frac{\partial \mathbf{W}}{\partial \mathbf{W}} \mathbf{b}$ , where  $\frac{\partial \mathbf{W}}{\partial \mathbf{W}}$  has dimensions  $m \times n \times n \times m$ . For example if  $m = 3$ ,  $n = 2$ , then:

$$\frac{\partial \mathbf{W}}{\partial \mathbf{W}} = \begin{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \end{bmatrix} \\ \begin{bmatrix} 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} \end{bmatrix} \quad (15)$$

Note that the  $(i, j)$ th inner matrix has a 1 in its  $(j, i)$ th position. Pre- and post-multiplying the  $(i, j)$ th inner matrix with  $\mathbf{a}^T$  and  $\mathbf{b}$  gives  $a_j b_i$ , where  $i \in \{1, 2\}$  and  $j \in \{1, 2, 3\}$ . So:

$$\mathbf{a}^T \frac{\partial \mathbf{W}}{\partial \mathbf{W}} \mathbf{b} = \begin{bmatrix} a_1 b_1 & a_2 b_1 & a_3 b_1 \\ a_1 b_2 & a_2 b_2 & a_3 b_2 \end{bmatrix} \quad (16)$$

Thus,  $\frac{\partial f}{\partial \mathbf{W}} = \mathbf{b} \mathbf{a}^T$ .

#### 4.4 Gradient Example: L2 Norm

Problem: Given  $f(\mathbf{x}) = \|\mathbf{x} - \mathbf{a}\|_2$ , find  $\nabla_{\mathbf{x}} f$ .

Note that  $\|\mathbf{x} - \mathbf{a}\|_2 = \sqrt{(\mathbf{x} - \mathbf{a})^T (\mathbf{x} - \mathbf{a})}$ , which is a scalar. So the derivative will be a row vector and gradient will be a column vector of the same dimension as  $\mathbf{x}$ . Let's use the chain rule:

$$\frac{\partial f}{\partial \mathbf{x}} = \frac{\partial \sqrt{(\mathbf{x} - \mathbf{a})^T (\mathbf{x} - \mathbf{a})}}{\partial (\mathbf{x} - \mathbf{a})^T (\mathbf{x} - \mathbf{a})} \times \frac{\partial (\mathbf{x} - \mathbf{a})^T (\mathbf{x} - \mathbf{a})}{\partial \mathbf{x}} \quad (17)$$

The first term is a scalar-scalar derivative equal to  $\frac{1}{2\sqrt{(\mathbf{x} - \mathbf{a})^T (\mathbf{x} - \mathbf{a})}}$ . The second term is:

$$\begin{aligned} \frac{\partial (\mathbf{x} - \mathbf{a})^T (\mathbf{x} - \mathbf{a})}{\partial \mathbf{x}} &= \frac{\partial (\mathbf{x}^T \mathbf{x} - \mathbf{a}^T \mathbf{x} - \mathbf{x}^T \mathbf{a} + \mathbf{a}^T \mathbf{a})}{\partial \mathbf{x}} \\ &= (\mathbf{x}^T + \mathbf{x}^T) - \mathbf{a}^T - \mathbf{a}^T + \mathbf{0}^T \\ &= 2(\mathbf{x}^T - \mathbf{a}^T) \end{aligned} \quad (18)$$

$$\text{So } \frac{\partial f}{\partial \mathbf{x}} = \frac{\mathbf{x}^T - \mathbf{a}^T}{\sqrt{(\mathbf{x} - \mathbf{a})^T (\mathbf{x} - \mathbf{a})}}.$$

So  $\nabla_{\mathbf{x}} f = \frac{\mathbf{x} - \mathbf{a}}{\|\mathbf{x} - \mathbf{a}\|_2}$ , which is basically the unit displacement vector from  $\mathbf{a}$  to  $\mathbf{x}$ . This means that to get maximum increase in  $f(\mathbf{x})$ , one should move away from  $\mathbf{a}$  along the straight line joining  $\mathbf{a}$  and  $\mathbf{x}$ . Alternatively, to get maximum decrease in  $f(\mathbf{x})$ , one should move from  $\mathbf{x}$  directly towards  $\mathbf{a}$ , which makes sense geometrically.



## 5 Notes and Further Reading

The chain rule and product rule do not always hold when dealing with matrices. However, some modified forms can hold when using the  $Trace(\cdot)$  function. For a full list of derivatives, the reader should consult a textbook or websites such as Wikipedia's page on Matrix calculus. Keep in mind that some texts may use denominator layout convention, where results will look different.