Matrix Calculus

Sourya Dey

1 Notation

- Scalars are written as lower case letters.
- Vectors are written as lower case bold letters, such as $\boldsymbol{x}$, 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 ($i \in \{1, \cdots, n\}$).
- Matrices are written as upper case bold letters, such as $\boldsymbol{X}$, 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} ($i \in \{1, \cdots, m\}$, $j \in \{1, \cdots, n\}$).
- 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 $\dfrac{\partial f}{\partial x}$. 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 $\nabla_x f = \left(\dfrac{\partial f}{\partial x}\right)^T$, 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 $\boldsymbol{F}(\cdot)$ 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, $\boldsymbol{F}(\cdot) : \underset{m \times n}{\boldsymbol{X}} \rightarrow \underset{p \times q}{\boldsymbol{F}(\boldsymbol{X})}$. Its derivative $\dfrac{\partial \boldsymbol{F}}{\partial \boldsymbol{X}}$ 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 $\boldsymbol{X}$), with each element being a $p \times q$ inner matrix (same dimensions as the numerator $\boldsymbol{F}$). It is given as:

$$\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{X}} = \begin{bmatrix} \frac{\partial \boldsymbol{F}}{\partial X_{1,1}} & \cdots & \frac{\partial \boldsymbol{F}}{\partial X_{m,1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \boldsymbol{F}}{\partial X_{1,n}} & \cdots & \frac{\partial \boldsymbol{F}}{\partial X_{m,n}} \end{bmatrix} \tag{1a}$$

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

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

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. $\frac{\partial f\left(g(x)\right)}{\partial x} = \frac{\partial f\left(g(x)\right)}{\partial g(x)} \frac{\partial g(x)}{\partial x}$.
- 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

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

$$\frac{\partial f}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_m} \end{bmatrix} \tag{2}$$

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

3.3 Vector by scalar

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

$$\frac{\partial \boldsymbol{f}}{\partial x} = \begin{bmatrix} \frac{\partial f_1}{\partial x} \\ \frac{\partial f_2}{\partial x} \\ \vdots \\ \frac{\partial f_n}{\partial x} \end{bmatrix} \tag{3}$$

3.4 Vector by vector

$\boldsymbol{f}(\cdot) : \underset{m\times 1}{\boldsymbol{x}} \to \underset{n\times 1}{\boldsymbol{f}(\boldsymbol{x})}$. 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 ith output component w.r.t the jth input component, i.e.:

$$\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_m} \end{bmatrix} \tag{4}$$

3.4.1 Special case – Vectorized scalar function

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

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

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

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

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

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 $\underline{f}'(\boldsymbol{x})$.

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

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

$$\left(\nabla_{\boldsymbol{x}} \underline{f}\right) \boldsymbol{a} = \begin{bmatrix} f'(x_1)\, a_1 \\ f'(x_2)\, a_2 \\ \vdots \\ f'(x_m)\, a_m \end{bmatrix} \tag{8}$$

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

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

3.4.2 Special Case – Hessian

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

$$\boldsymbol{H} = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \cdots & \dfrac{\partial^2 f}{\partial x_1 \partial x_m} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_m \partial x_1} & \cdots & \dfrac{\partial^2 f}{\partial x_m^2} \end{bmatrix} \tag{10}$$

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

3.5 Scalar by matrix

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

$$\frac{\partial f}{\partial \boldsymbol{X}} = \begin{bmatrix} \frac{\partial f}{\partial X_{1,1}} & \cdots & \frac{\partial f}{\partial X_{m,1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial X_{1,n}} & \cdots & \frac{\partial f}{\partial X_{m,n}} \end{bmatrix} \tag{11}$$

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

3.6 Matrix by scalar

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

$$\frac{\partial \boldsymbol{F}}{\partial x} = \begin{bmatrix} \frac{\partial F_{1,1}}{\partial x} & \cdots & \frac{\partial F_{1,q}}{\partial x} \\ \vdots & \ddots & \vdots \\ \frac{\partial F_{p,1}}{\partial x} & \cdots & \frac{\partial F_{p,q}}{\partial x} \end{bmatrix} \tag{12}$$

3.7 Vector by matrix

$\boldsymbol{f}(\cdot) : \underset{m\times n}{\boldsymbol{X}} \rightarrow \underset{p\times 1}{\boldsymbol{f}(\boldsymbol{X})}$. 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 $\boldsymbol{f}$, i.e.:

$$\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{X}} = \begin{bmatrix} \frac{\partial \boldsymbol{f}}{\partial X_{1,1}} & \cdots & \frac{\partial \boldsymbol{f}}{\partial X_{m,1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \boldsymbol{f}}{\partial X_{1,n}} & \cdots & \frac{\partial \boldsymbol{f}}{\partial X_{m,n}} \end{bmatrix} \tag{13}$$

3.8 Matrix by vector

$\boldsymbol{F}(\cdot) : \underset{m\times 1}{\boldsymbol{x}} \rightarrow \underset{p\times q}{\boldsymbol{F}(\boldsymbol{x})}$. 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 $\boldsymbol{F}$, i.e.:

$$\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial \boldsymbol{F}}{\partial x_1} & \frac{\partial \boldsymbol{F}}{\partial x_2} & \cdots & \frac{\partial \boldsymbol{F}}{\partial x_m} \end{bmatrix} \tag{14}$$

4 Operations and Examples

4.1 Commutation

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

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

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 $f(\boldsymbol{w}) = (y - \boldsymbol{w}^T\boldsymbol{x})^2 = y^2 - \left(\boldsymbol{w}^T\boldsymbol{x}\right) y - y\left(\boldsymbol{w}^T\boldsymbol{x}\right) + \left(\boldsymbol{w}^T\boldsymbol{x}\right)\left(\boldsymbol{w}^T\boldsymbol{x}\right)$, where y and $\boldsymbol{x}$ are not a function of $\boldsymbol{w}$. Taking derivative of the terms individually:

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

So $\dfrac{\partial f}{\partial \boldsymbol{w}} = -2y\boldsymbol{x}^T + 2\left(\boldsymbol{w}^T\boldsymbol{x}\right)\boldsymbol{x}^T$

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: $f(\boldsymbol{W}) = \underset{1\times m}{\boldsymbol{a}^T} \underset{m\times n}{\boldsymbol{W}} \underset{n\times 1}{\boldsymbol{b}}$. This is a scalar, so $\frac{\partial f}{\partial \boldsymbol{W}}$ should be a matrix which has transposed dimensions as $\boldsymbol{W}$, i.e. $n \times m$. Now, $\frac{\partial f}{\partial \boldsymbol{W}} = \boldsymbol{a}^T \frac{\partial \boldsymbol{W}}{\partial \boldsymbol{W}} \boldsymbol{b}$, where $\frac{\partial \boldsymbol{W}}{\partial \boldsymbol{W}}$ has dimensions $m \times n \times n \times m$. For example if $m = 3$, $n = 2$, then:

$$\frac{\partial \boldsymbol{W}}{\partial \boldsymbol{W}} = \begin{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \end{bmatrix} \\ \begin{bmatrix} 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} \end{bmatrix} \tag{15}$$

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 $\boldsymbol{a}^T$ and $\boldsymbol{b}$ gives $a_j b_i$, where $i \in \{1, 2\}$ and $j \in \{1, 2, 3\}$. So:

$$\boldsymbol{a}^T \frac{\partial \boldsymbol{W}}{\partial \boldsymbol{W}} \boldsymbol{b} = \begin{bmatrix} a_1 b_1 & a_2 b_1 & a_3 b_1 \\ a_1 b_2 & a_2 b_2 & a_3 b_2 \end{bmatrix} \tag{16}$$

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

4.4 Gradient Example: L2 Norm

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

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

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

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

$$\begin{aligned} \frac{\partial (\boldsymbol{x} - \boldsymbol{a})^T(\boldsymbol{x} - \boldsymbol{a})}{\partial \boldsymbol{x}} &= \frac{\partial \left(\boldsymbol{x}^T\boldsymbol{x} - \boldsymbol{a}^T\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{a} + \boldsymbol{a}^T\boldsymbol{a}\right)}{\partial \boldsymbol{x}} \\ &= \left(\boldsymbol{x}^T + \boldsymbol{x}^T\right) - \boldsymbol{a}^T - \boldsymbol{a}^T + \boldsymbol{0}^T \\ &= 2\left(\boldsymbol{x}^T - \boldsymbol{a}^T\right) \end{aligned} \tag{18}$$

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

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