

# Notes on differentiation involving vectors

Prabir Barooah

October 26, 2009

We always follow the convention:

A vector, by default, is a column vector. Therefore, a  $n$ -dimensional vector  $\mathbf{x}$  is represented as

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

The derivative of a scalar function with respect to vector is represented as a row vector. Suppose  $\mathbf{x} \in \mathbb{R}^n$  and  $f(\mathbf{x}) \in \mathbb{R}$ . Then

$$\frac{df(\mathbf{x})}{d\mathbf{x}} = \left[ \frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n} \right]. \quad (1)$$

To remember this and the subsequent conventions, think of Taylor series expansion of  $f$  around the “point”  $\mathbf{x}_0$ , which we would like to be able to write in the same manner if all quantities involved were scalars, i.e.,

$$f(\mathbf{x}_0 + \delta\mathbf{x}) = f(\mathbf{x}_0) + \left. \frac{df(\mathbf{x})}{d\mathbf{x}} \right|_{\mathbf{x}_0} \delta\mathbf{x} + \dots$$

Since the left hand side is a scalar, each term on the right hand side must be a scalar as well. Since  $\delta\mathbf{x}$  is a column vector (remember the first convention), this means the derivative  $\left. \frac{df(\mathbf{x})}{d\mathbf{x}} \right|_{\mathbf{x}_0}$  must be a row vector, so that their product is a scalar. This is the justification for the convention that  $\frac{df(\mathbf{x})}{d\mathbf{x}}$  is a row vector, and also serves as a way to “derive” the convention should you forget it.

If you want to expand the Taylor series upto the second derivative, then it becomes

$$f(\mathbf{x}_0 + \delta\mathbf{x}) = f(\mathbf{x}_0) + \left. \frac{df(\mathbf{x})}{d\mathbf{x}} \right|_{\mathbf{x}_0} \delta\mathbf{x} + \frac{1}{2} \delta\mathbf{x}^T \left. \frac{d^2f(\mathbf{x})}{d\mathbf{x}^2} \right|_{\mathbf{x}_0} \delta\mathbf{x} + \dots$$

where  $\frac{d^2 f(\mathbf{x})}{d\mathbf{x}^2}$  is an  $n \times n$  matrix that is called the *Hessian* of the function  $f$  at  $\mathbf{x}$ . It is defined as:

$$\frac{d^2 f(\mathbf{x})}{d\mathbf{x}^2} := \begin{bmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_n^2} \end{bmatrix}$$

It is easy to extrapolate now to the derivative of a vector function with respect to another vector. That can also be “derived” using the Taylor series argument. Let  $g(\mathbf{x}) \in \mathbb{R}^p$  and  $\mathbf{x} \in \mathbb{R}^n$ . Then, the derivative of the vector  $g(\mathbf{x})$  with respect to the vector  $\mathbf{x}$  is defined as the following  $p \times n$  matrix:

$$\frac{dg(\mathbf{x})}{d\mathbf{x}} = \begin{bmatrix} \frac{d}{d\mathbf{x}} g_1(\mathbf{x}) \\ \frac{d}{d\mathbf{x}} g_2(\mathbf{x}) \\ \vdots \\ \frac{d}{d\mathbf{x}} g_p(\mathbf{x}) \end{bmatrix} \in \mathbb{R}^{p \times n}, \quad (2)$$

where  $g_i(\mathbf{x})$  is the  $i^{\text{th}}$  component of the vector  $g(\mathbf{x})$ .

#### Derivatives of linear functions:

You can verify the following:

$$\frac{d}{d\mathbf{x}} (\mathbf{x}^T \mathbf{y}) = \frac{d}{d\mathbf{x}} (\mathbf{y}^T \mathbf{x}) = \mathbf{y}^T. \quad (3)$$

$$\frac{d}{d\mathbf{y}} (\mathbf{x}^T \mathbf{y}) = \mathbf{x}^T. \quad \square \quad (4)$$

hint: you can verify them by writing  $\mathbf{x}^T \mathbf{y}$  as  $\sum_{i=1}^n x_i y_i$  and then applying the definition (1).

Another useful formula that you should verify is the following, where  $A \in \mathbb{R}^{p \times n}$  and  $\mathbf{y} \in \mathbb{R}^n$ :

$$\frac{d}{d\mathbf{y}} (A\mathbf{y}) = A.$$

Note that it has been implicitly assumed that  $\mathbf{y}, A$  etc. are not functions of  $\mathbf{x}$ , otherwise these formulas do not hold.

#### Derivatives of quadratic functions:

Verify the following from the definition:

1. For a vector  $\mathbf{x} \in \mathbb{R}^n$ ,

$$\frac{d}{d\mathbf{x}}(\mathbf{x}^T \mathbf{x}) = 2\mathbf{x}^T$$

2. For a constant matrix  $P \in \mathbb{R}^{n \times n}$ , and a vector  $\mathbf{x} \in \mathbb{R}^n$ ,

$$\frac{d}{d\mathbf{x}}(\mathbf{x}^T P \mathbf{x}) = \mathbf{x}^T P + \mathbf{x}^T P^T.$$

**Chain rule** Let  $\mathbf{x} \in \mathbb{R}^n$  and  $f(\mathbf{x}), g(\mathbf{x}) \in \mathbb{R}^p$ . Then, you can verify that

$$\frac{d}{d\mathbf{x}}(f^T g) = f^T \frac{dg}{d\mathbf{x}} + g^T \frac{df}{d\mathbf{x}}.$$

You can now derive the formula for  $d(x^P x)/dx$  by setting  $f := x$  and  $g = Px$  and applying the chain rule.

Note that:

1. the chain rule involving vectors is no different than the one with scalars, but unlike the scalar case, the order of the two terms matters.
2. Note that some books leave the transpose sign on  $\mathbf{y}$  out on the right hand side of (3). If you are only dealing with a single derivative, that usually does not cause confusion. But such sloppiness cause trouble when you encounter complicated situations, e.g., when you try to apply the chain rule.