Vector Differentiation Rules

Derivatives with respect to a vector come up often in a multitude of areas, such as constrained optimisation, adaptive filtering, and machine learning. We begin by defining the simple case of a scalar differentiated by a vector.

Definition 0.1 (gradient)**.** *Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be a column vector and let $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$ be a function that maps to a scalar. The derivative of f w.r.t $\boldsymbol{x}$, also known as gradient, is defined as:*

$$\nabla_{\boldsymbol{x}} f := \frac{\partial f}{\partial \boldsymbol{x}} := \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \ldots & \frac{\partial f}{\partial x_n} \end{bmatrix}^\top \tag{0.1}$$

Both notations in Eq. (0.1) are acceptable for the gradient

Example 0.1. *Find the gradient of the function $f(x_1, x_2, x_3) = x_1 + 3x_2 + 2x_3$.*

Solution 0.1.

$$\nabla_{\boldsymbol{x}} f = \begin{bmatrix} \frac{\partial(x_1+3x_2+2x_3)}{\partial x_1} & \frac{\partial(x_1+3x_2+2x_3)}{\partial x_2} & \frac{\partial(x_1+3x_2+2x_3)}{\partial x_3} \end{bmatrix}^\top = \begin{bmatrix} 1 & 3 & 2 \end{bmatrix}^\top$$

□

We already notice a property of vector differentiation in the example above; if $f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x}$, where $\mathbf{a}^\top = \begin{bmatrix} 1 & 3 & 2 \end{bmatrix}$ is a coefficient row vector and $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}^\top$ a column vector, then $\nabla_{\mathbf{x}} f = \mathbf{a}$.

For reference, the derivative of vector w.r.t. a vector is also defined just to highlight its difference with the derivative of a scalar w.r.t. a vector.

Definition 0.2 (Jacobian matrix)**.** *Let $\boldsymbol{x} \in \mathbb{R}^m$ and $\boldsymbol{f} : \mathbb{R}^m \to \mathbb{R}^n$ be a function that returns a $n \times 1$ vector, i.e. $\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} f_1(\boldsymbol{x}) & \ldots & f_n(\boldsymbol{x}) \end{bmatrix}^\top$. Then the derivative of vector $\boldsymbol{f}(\boldsymbol{x})$ w.r.t. $\boldsymbol{x}$ is called Jacobian matrix and is defined as [1]*

$$\boldsymbol{J}(\boldsymbol{x}) = \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \frac{\partial f_1(\boldsymbol{x})}{\partial x_2} & \cdots & \frac{\partial f_1(\boldsymbol{x})}{\partial x_m} \\ \frac{\partial f_2(\boldsymbol{x})}{\partial x_1} & \frac{\partial f_2(\boldsymbol{x})}{\partial x_2} & \cdots & \frac{\partial f_2(\boldsymbol{x})}{\partial x_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n(\boldsymbol{x})}{\partial x_1} & \frac{\partial f_n(\boldsymbol{x})}{\partial x_2} & \cdots & \frac{\partial f_n(\boldsymbol{x})}{\partial x_m} \end{bmatrix} \tag{0.2}$$

The determinant of the Jacobian matrix is called Jacobian determinant or Jacobian for short. Therefore each row k contains the derivative of the scalar function $f_k(.)$ with respect to the elements in $\mathbf{x}$.

Example 0.2. *Compute the Jacobian matrix of the transformation $T(u,v) = \begin{bmatrix} u & v & u^v \end{bmatrix}^\top, \quad u > 0$ [2].*

Solution 0.2. *Using the notation from the definition we compute each row at a time.*

$$f_1(u,v) = u \Rightarrow \frac{\partial f_1(u,v)}{\partial u} = 1, \quad \frac{\partial f_1(u,v)}{\partial v} = 0$$

$$f_2(u,v) = v \Rightarrow \frac{\partial f_1(u,v)}{\partial u} = 0, \quad \frac{\partial f_1(u,v)}{\partial v} = 1$$

$$f_3(u,v) = u^v \Rightarrow \frac{\partial f_3(u,v)}{\partial u} = v u^{v-1}, \quad \frac{\partial f_3(u,v)}{\partial v} = u^v \ln u$$

$$\therefore \boldsymbol{J}(u,v) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ v u^{v-1} & u^v \ln u \end{bmatrix}$$

□

Now we can derive and provide the derivatives of some common scalar ((i), (iii), (iv)) and one vector ((ii)) expressions w.r.t. a vector.

Lemma 0.1 (vector differentiation basic properties)**.** *Let $\boldsymbol{x}, \boldsymbol{a} \in \mathbb{R}^n$ be two column vectors where $\boldsymbol{a}$ is not a* *function of $\boldsymbol{x}$ and $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ be a real matrix. Then:*

$$(i) \qquad \frac{\partial(\boldsymbol{a}^\top \boldsymbol{x})}{\partial \boldsymbol{x}} = \frac{\partial(\boldsymbol{x}^\top \boldsymbol{a})}{\partial \boldsymbol{x}} = \boldsymbol{a} \tag{0.3}$$

$$(ii) \qquad \frac{\partial(\boldsymbol{A}\boldsymbol{x})}{\partial \boldsymbol{x}} = \boldsymbol{A} \tag{0.4}$$

$$(iii) \qquad \frac{\partial(\boldsymbol{x}^\top \boldsymbol{A}^\top)}{\partial \boldsymbol{x}} = \boldsymbol{A}^\top, \qquad \text{if} \quad m = n \tag{0.5}$$

$$(iv) \qquad \frac{\partial(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x})}{\partial \boldsymbol{x}} = (\boldsymbol{A} + \boldsymbol{A}^\top)\,\boldsymbol{x}, \qquad \text{if} \quad m = n \tag{0.6}$$

Proof.

(i) From the dot product's definition:

$$\frac{\partial(\mathbf{a}^\top \mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\sum_{i=1}^n a_i x_i}{\partial x_1} \\ \frac{\sum_{i=1}^n a_i x_i}{\partial x_2} \\ \vdots \\ \frac{\sum_{i=1}^n a_i x_i}{\partial x_n} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \mathbf{a}$$

(ii) If we denote $\mathbf{a}_1^\top, \dots, \mathbf{a}_m^\top$ the rows of $\mathbf{A}$ expressed as column vectors, then product $\mathbf{Ax}$ is written as:

$$\mathbf{Ax} = \begin{bmatrix} \mathbf{a}_1^\top \mathbf{x} \\ \mathbf{a}_2^\top \mathbf{x} \\ \vdots \\ \mathbf{a}_m^\top \mathbf{x} \end{bmatrix}$$

We apply definition Eq. (0.1) on each row, since each row is a scalar:

$$\therefore \frac{\partial(\mathbf{Ax})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial(\mathbf{a}_1^\top \mathbf{x})}{\partial \mathbf{x}} \overset{(0.1)}{=} \mathbf{a}_1 \\ \frac{\partial(\mathbf{a}_2^\top \mathbf{x})}{\partial \mathbf{x}} \overset{(0.1)}{=} \mathbf{a}_2 \\ \vdots \\ \frac{\partial(\mathbf{a}_m^\top \mathbf{x})}{\partial \mathbf{x}} \overset{(0.1)}{=} \mathbf{a}_m \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_m \end{bmatrix} = \mathbf{A}$$

(iii) Left as exercise.

(iv) The i-th element of the product $\mathbf{Ax}$, which is a vector, is written with the index notation as follows.

$$(\mathbf{Ax})_i = \sum_{j=1}^n A_{ij} x_j$$

The dot product of $\mathbf{x}$, $\mathbf{Ax}$ is written as:

$$\mathbf{x}^\top \mathbf{Ax} = \sum_i^n x_i \sum_{j=1}^n A_{ij} x_j$$

Applying the definition of the gradient (Eq. (0.1)) to the dot product:

$$\begin{aligned} \frac{\partial(\mathbf{x}^\top \mathbf{Ax})}{\partial \mathbf{x}} &= \frac{\partial(\sum_i^n x_i \sum_{j=1}^n A_{ij} x_j)}{\partial \mathbf{x}} \\ &= \left[\frac{\partial(\sum_{i=1}^n x_i \sum_{j=1}^n A_{ij} x_j)}{\partial x_1} \quad \dots \quad \frac{\partial(\sum_{i=1}^n x_i \sum_{j=1}^n A_{ij} x_j)}{\partial x_n} \right]^\top \end{aligned}$$

Using the product rule on the first element:

$$
\begin{aligned}
\frac{\sum_{i=1}^{n} x_i \sum_{j=1}^{n} A_{ij} x_j}{\partial x_1} &= \cancelto{1,\, i=1 \text{ else } 0}{\frac{\partial \sum_{i=1}^{n} x_i}{\partial x_1}} \sum_{j=1}^{n} A_{ij} x_j \quad + \quad \sum_{i=1}^{n} x_i \cancelto{A_{ij},\, j=1 \text{ else } 0}{\frac{\partial \sum_{j=1}^{n} A_{ij} x_j}{\partial x_1}} \\
&= \sum_{j=1}^{n} A_{1j} x_j + \sum_{i=1}^{n} x_i A_{i1} \\
&= \mathbf{a}_{1:}\mathbf{x} + \mathbf{a}_{:1}\mathbf{x} = (\mathbf{a}_{1:} + \mathbf{a}_{:1})\mathbf{x}
\end{aligned}
$$

, where $\mathbf{a}_{1:}$ denotes the first row of $\mathbf{A}$ and $\mathbf{a}_{:1}$ its first column (Matlab notation). Applying the result to the remaining indexes $2, \ldots, n$, we can rewrite the gradient as:

$$
\frac{\partial(\mathbf{x}^\top \mathbf{A}\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \mathbf{a}_{1:} + \mathbf{a}_{:1} & \ldots & \mathbf{a}_{n:} + \mathbf{a}_{:n} \end{bmatrix} \mathbf{x} = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}
$$

□

Some other properties that can be readily derived from the basic ones are listed below.

Lemma 0.2 (vector differentiation follow-up properties)**.** *If $\mathbf{A} \in \mathbb{R}^{n\times n}$ is square and symmetric ($\mathbf{A} = \mathbf{A}^\top$) and $\boldsymbol{x} \in \mathbb{R}^n$, then*

$$
\frac{\partial(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x})}{\partial \boldsymbol{x}} = 2\boldsymbol{A}\boldsymbol{x} \tag{0.7}
$$

For $\mathbf{A} = \boldsymbol{I}$, we can derive the vector derivative of the squared norm-2:

$$
\frac{\partial(\boldsymbol{x}^\top \boldsymbol{x})}{\partial \boldsymbol{x}} = \frac{\partial \|\boldsymbol{x}\|^2}{\partial \boldsymbol{x}} = 2\boldsymbol{x} \tag{0.8}
$$

Finally, we define the chain rule for vector functions as it's expressed in a slightly different order than the chain rule of scalars.

Lemma 0.3 (chain rule for vector function)**.** *Let $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{y} \in \mathbb{R}^r$, $\boldsymbol{z} \in \mathbb{R}^m$, where $\boldsymbol{z} = \boldsymbol{z}(\boldsymbol{y})$ and $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x})$. Then*

$$
\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{y}} \tag{0.9}
$$

Proof.
Proof is found in [3]. □

References

[1] H. B. Nielsen, *Introduction to vector and matrix differentiation*, 2012. [Online]. Available: https://absalon.instructure.com/files/1853451/download?download_frd=1.

[2] J. Ruan, *Jacobians and their applications*. [Online]. Available: https://www.projectrhea.org/rhea/index.php/Jacobian.

[3] W. Zhu, *Matrix & vector basic linear algebra & calculus*. [Online]. Available: http://www.ams.sunysb.edu/~zhu/ams571/matrixvector.pdf.