On Continuity in Undergraduate Mathematics

Robert Chan

September 5, 2007

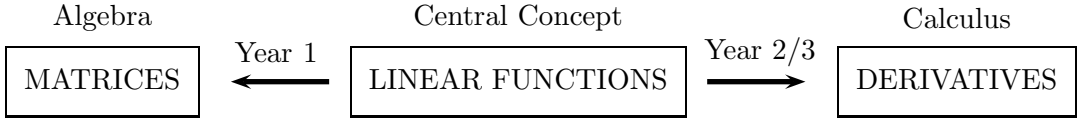
Abstract

The role of linear functions in undergraduate mathematics can be better understood. Their matrix representations are important in algebra and they also appear as derivatives in multivariable calculus. These connections are sometimes apparent to students only at a stage later than desirable. In the undergraduate curriculum some topics make an appearance at all levels and the transition from elementary to advanced levels is often not as smooth as can be. In this article we show some benefits in making linear functions a central concept in first-year mathematics.

1 Introduction

In the undergraduate mathematics curriculum many key concepts and topics are repeated in courses throughout several levels. It is debatable and even controversial as to when students should first meet certain concepts and topics. When topics at first-year level are treated in some detail but with insufficient balance given to the macroscopic view, problems can arise later at more advanced levels. It can be argued that the division and spread of topics and the depth of their treatment is in general appropriate and in a state of flux with a continuous process of review as changes in the teaching personnel occur from time to time. Ocassionally some details escape notice. These are often unimportant and can be ignored. Sometimes, however, these small details have consequences that are not evident until the advanced levels. The cost may well be insignificant for the able student who can self-correct but could be immeasurable for many others. It is of the utmost importance that as mathematics educators we get it right, especially for those who are going to go on to become teachers.

One example is the topic of matrices which pervades our undergraduate curriculum in both pure and applied courses. Matrices are introduced in the first year but without any connection to linear functions on which they are based. In this article we advocate using linear functions as a key concept in providing a foundation, not only for the important topic of matrices in algebra but also for derivatives in calculus as well. Schematically,



Derivatives of real-valued functions of a single variable are introduced in the first year. Derivatives in higher dimensions appear in second and third year courses, but traditionally the generalization of derivatives as linear functions is a hurdle for many students because the definitions used are too dissimilar. The transition from a single real variable to multi-variables can be made smoother if students are given the appropriate definition and preparation at the beginning. We give two examples in first-year mathematics following the definition of linear functions in Section 2. In Section 3 we discuss the representation of linear functions as matrices and in Section 4 establish their composition as matrix multiplication; in Section 5 we propose motivating determinants as a scale factor for area

in planar transformations. In Section 6 we propose an approach to derivatives which makes more natural the generalization to derivatives as linear functions represented by Jacobian matrices. The proposed definition has the advantage of reducing the process of evaluating limits to evaluating zero limits and avoids the restriction of h to nonzero values. An added advantage is that the equation of the tangent line (or hyperplanes in higher dimensions) appears naturally as the linear approximation. A further benefit is that the proof of the chain rule is unchanged in higher dimensions.

2 Linear functions

A function $\boldsymbol{f}\colon \mathbb{R}^n \to \mathbb{R}^m$ is said to be *linear* if

1. $\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{y}) = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{f}(\boldsymbol{y})$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$; and
2. $\boldsymbol{f}(c\boldsymbol{x}) = c\boldsymbol{f}(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^n$ and $c \in \mathbb{R}$.

Equivalently, $\boldsymbol{f}$ is linear if it satisfies $\boldsymbol{f}(c_1\boldsymbol{x}+c_2\boldsymbol{y}) = c_1\boldsymbol{f}(\boldsymbol{x}) + c_2\boldsymbol{f}(\boldsymbol{y})$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and $c_1, c_2 \in \mathbb{R}$. We shall see how matrices are used to represent such linear functions. Linear functions are also called *linear transformations* and we shall use these terms interchangeably.

Before proceeding to the main part of our discussion we remark that the term 'linear' is used in many contexts and can be confusing to students. Here are some examples:

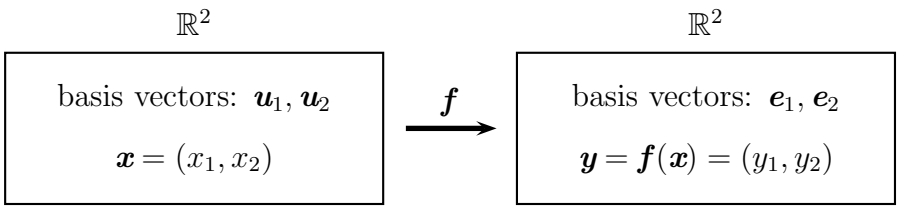
1. The polynomial $p(x) = ax + b$, although referred to as 'linear' is not a linear function in the above sense if $b \neq 0$.
2. A real-valued function of two variables, $f(x, y) = ax + by + c$, is not a linear function if $c \neq 0$. We do say 'f is linear' in x and y. The terms 'homogeneous' and 'inhomogeneous' are also used.

A property of linear functions, $\boldsymbol{f}(\boldsymbol{0}) = \boldsymbol{0}$, becomes important in identifying when a function $\boldsymbol{f}$ is linear. Hence a function $f : \mathbb{R} \to \mathbb{R}$ is linear if its straight-line graph goes through the origin; a function $f : \mathbb{R}^2 \to \mathbb{R}$ is linear if its graph, a plane in $\mathbb{R}^3$, goes through the origin.

3 Planar transformations

In the first year it is sufficient to illustrate the ideas involved with transformations on the plane. Let the basis vectors in the planes of the domain and co-domain be denoted by $\boldsymbol{u}_1, \boldsymbol{u}_2$ and $\boldsymbol{e}_1, \boldsymbol{e}_2$ respectively. If $\boldsymbol{x} = (x_1, x_2)$ is a point in the domain-plane $\mathbb{R}^2$, then we can write

$$\boldsymbol{x} = x_1\boldsymbol{u}_1 + x_2\boldsymbol{u}_2. \tag{3.1}$$



If $\boldsymbol{f}\colon \mathbb{R}^2 \to \mathbb{R}^2$ is linear then, by definition,

$$\boldsymbol{f}(\boldsymbol{x}) = x_1\boldsymbol{f}(\boldsymbol{u}_1) + x_2\boldsymbol{f}(\boldsymbol{u}_2). \tag{3.2}$$

Hence knowledge of $\boldsymbol{f}(\boldsymbol{u}_1)$ and $\boldsymbol{f}(\boldsymbol{u}_2)$ completely determines how $\boldsymbol{x}$ transforms under $\boldsymbol{f}$. Since $\boldsymbol{f}(\boldsymbol{u}_j)$ is in the co-domain $\mathbb{R}^2$, we have the linear combinations

$$\boldsymbol{f}(\boldsymbol{u}_1) = a_{11}\boldsymbol{e}_1 + a_{21}\boldsymbol{e}_2, \quad \boldsymbol{f}(\boldsymbol{u}_2) = a_{12}\boldsymbol{e}_1 + a_{22}\boldsymbol{e}_2, \tag{3.3}$$

where the a_{ij} are real numbers. The a_{ij}'s completely determine the transformation f since, by (3.1), (3.2) and (3.3), we have

$$\begin{aligned}\boldsymbol{f}(\boldsymbol{x}) &= x_1(a_{11}\,\boldsymbol{e}_1 + a_{21}\,\boldsymbol{e}_2) + x_2(a_{12}\,\boldsymbol{e}_1 + a_{22}\,\boldsymbol{e}_2) = (a_{11}x_1 + a_{12}x_2)\boldsymbol{e}_1 + (a_{21}x_1 + a_{22}x_2)\boldsymbol{e}_2 \\ &= y_1\,\boldsymbol{e}_1 + y_2\,\boldsymbol{e}_2, \;\text{ where }\; y_i = a_{i1}x_1 + a_{i2}x_2.\end{aligned} \tag{3.4}$$

How do we represent this transformation?

If the standard basis in each plane are represented by column vectors,

$$\boldsymbol{u}_1 = \boldsymbol{e}_1 = \begin{bmatrix}1\\0\end{bmatrix}, \qquad \boldsymbol{u}_2 = \boldsymbol{e}_2 = \begin{bmatrix}0\\1\end{bmatrix},$$

then points in $\mathbb{R}^2$ are also represented by column vectors; (3.1) becomes

$$\boldsymbol{x} = x_1\begin{bmatrix}1\\0\end{bmatrix} + x_2\begin{bmatrix}0\\1\end{bmatrix} = \begin{bmatrix}x_1\\0\end{bmatrix} + \begin{bmatrix}0\\x_2\end{bmatrix} = \begin{bmatrix}x_1\\x_2\end{bmatrix},$$

and, similarly, (3.3) becomes

$$\boldsymbol{f}(\boldsymbol{u}_1) = \begin{bmatrix}a_{11}\\a_{21}\end{bmatrix}, \qquad \boldsymbol{f}(\boldsymbol{u}_2) = \begin{bmatrix}a_{12}\\a_{22}\end{bmatrix}.$$

Hence, by (3.4),

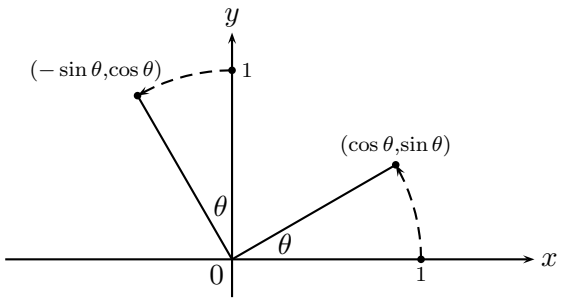
$$\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix}a_{11}x_1 + a_{12}x_2\\a_{21}x_1 + a_{22}x_2\end{bmatrix} = \begin{bmatrix}a_{11} & a_{12}\\a_{21} & a_{22}\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix}.$$

Thus $\boldsymbol{f}$ is represented by the matrix

$$A = \begin{bmatrix}a_{11} & a_{12}\\a_{21} & a_{22}\end{bmatrix},$$

and the value of $\boldsymbol{f}$ at $\boldsymbol{x}$ is represented by a matrix-vector product $A\boldsymbol{x}$. Note that the first column of A is $\boldsymbol{f}(\boldsymbol{u}_1)$ while the second column is $\boldsymbol{f}(\boldsymbol{u}_2)$.

Example 1 *Determine the matrix representing a rotation through angle θ anticlockwise centred at the origin. Now from the diagram (showing both sets of vectors on the same axes)*



we see that

$$\begin{bmatrix}1\\0\end{bmatrix} \longmapsto \begin{bmatrix}\cos\theta\\\sin\theta\end{bmatrix} \quad \textit{and} \quad \begin{bmatrix}0\\1\end{bmatrix} \longmapsto \begin{bmatrix}-\sin\theta\\\cos\theta\end{bmatrix}.$$

Hence the rotation matrix is

$$\begin{bmatrix}\cos\theta & -\sin\theta\\\sin\theta & \cos\theta\end{bmatrix}.$$

□

4 Composition as matrix multiplication

The algebraic rules for matrices and vectors follow from the rules for the corresponding linear functions and linear vector space. We now establish an important link between the composition of linear functions and matrix multiplication of their corresponding matrix representations. The key result is expressed in the following theorem.

Theorem 1 *If $\boldsymbol{f}\colon \mathbb{R}^2 \to \mathbb{R}^2$ and $\boldsymbol{g}\colon \mathbb{R}^2 \to \mathbb{R}^2$ are linear transformations represented by the matrices A and B respectively, then the composition $\boldsymbol{g} \circ \boldsymbol{f}$ is linear and is represented by the matrix product BA.*

Proof: $\boldsymbol{g} \circ \boldsymbol{f}$ is linear since for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^2$ and $c_1, c_2 \in \mathbb{R}$, we have

$$\begin{aligned}
\boldsymbol{g} \circ \boldsymbol{f}(c_1\boldsymbol{x} + c_2\boldsymbol{y}) &= \boldsymbol{g}(\boldsymbol{f}(c_1\boldsymbol{x} + c_2\boldsymbol{y})) && \text{(by definition of composition)} \\
&= \boldsymbol{g}(c_1\boldsymbol{f}(\boldsymbol{x}) + c_2\boldsymbol{f}(\boldsymbol{y})) && \text{(by linearity of } \boldsymbol{f}) \\
&= c_1\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x})) + c_2\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{y})) && \text{(by linearity of } \boldsymbol{g}) \\
&= c_1\boldsymbol{g} \circ \boldsymbol{f}(\boldsymbol{x}) + c_2\boldsymbol{g} \circ \boldsymbol{f}(\boldsymbol{y}).
\end{aligned}$$

Now let $\boldsymbol{x} = [x_1, x_2]^T \in \mathbb{R}^2$ and $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{y} = [y_1, y_2]^T \in \mathbb{R}^2$. From Section 3 we have

$$\begin{aligned}
\boldsymbol{y} &= \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{21}x_1 + a_{22}x_2 \end{bmatrix}, \\
\boldsymbol{g} \circ \boldsymbol{f}(\boldsymbol{x}) &= \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x})) = \boldsymbol{g}(\boldsymbol{y}) = B\boldsymbol{y} \\
&= \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} b_{11}y_1 + b_{12}y_2 \\ b_{21}y_1 + b_{22}y_2 \end{bmatrix} = \begin{bmatrix} b_{11}(a_{11}x_1 + a_{12}x_2) + b_{12}(a_{21}x_1 + a_{22}x_2) \\ b_{21}(a_{11}x_1 + a_{12}x_2) + b_{22}(a_{21}x_1 + a_{22}x_2) \end{bmatrix} \\
&= \begin{bmatrix} (b_{11}a_{11} + b_{12}a_{21})x_1 + (b_{11}a_{12} + b_{12}a_{22})x_2 \\ (b_{21}a_{11} + b_{22}a_{21})x_1 + (b_{21}a_{12} + b_{22}a_{22})x_2 \end{bmatrix} = \begin{bmatrix} b_{11}a_{11} + b_{12}a_{21} & b_{11}a_{12} + b_{12}a_{22} \\ b_{21}a_{11} + b_{22}a_{21} & b_{21}a_{12} + b_{22}a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.
\end{aligned}$$

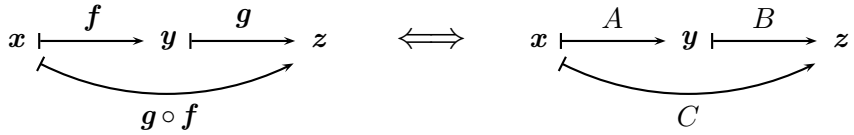
Hence the matrix representing $\boldsymbol{g} \circ \boldsymbol{f}$ is

$$\begin{bmatrix} b_{11}a_{11} + b_{12}a_{21} & b_{11}a_{12} + b_{12}a_{22} \\ b_{21}a_{11} + b_{22}a_{21} & b_{21}a_{12} + b_{22}a_{22} \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = BA. \qquad \square$$

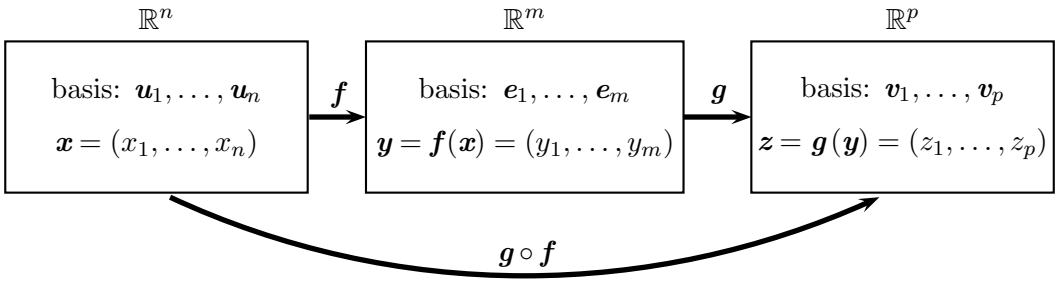
Remarks: In the general case where $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ is represented by an $m \times n$ matrix A, and $\boldsymbol{g}\colon \mathbb{R}^m \to \mathbb{R}^p$ by a $p \times m$ matrix B, then the composition $\boldsymbol{g} \circ \boldsymbol{f}\colon \mathbb{R}^n \to \mathbb{R}^p$ is represented by a $p \times n$ matrix C, and $C = BA$.

1. The proof of the linearity of $\boldsymbol{g} \circ \boldsymbol{f}$ given above is unchanged.

2. The proof that $C = BA$ is easy to generalize, but can it be shortened? Since



   we have $\boldsymbol{g} \circ \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x})) = \boldsymbol{g}(\boldsymbol{y}) = B\boldsymbol{y} = B(\boldsymbol{f}(\boldsymbol{x})) = B(A\boldsymbol{x})$, and since $\boldsymbol{g} \circ \boldsymbol{f}(\boldsymbol{x}) = C\boldsymbol{x}$ we have $B(A\boldsymbol{x}) = C\boldsymbol{x}$ for all $\boldsymbol{x} \in \mathbb{R}^n$. Can we now conclude $C = BA$? Why not? To show $(BA)\boldsymbol{x} = B(A\boldsymbol{x})$, we need to go through the steps exemplified above. But this generalization is now a simple extension of the summation notation.



Since $\boldsymbol{f}$ is linear and $\boldsymbol{f}(\boldsymbol{u}_j) \in \mathbb{R}^m$, we have

$$\boldsymbol{f}(\boldsymbol{x}) = \sum_{j=1}^{n} x_j \boldsymbol{f}(\boldsymbol{u}_j) = \sum_{j=1}^{n} x_j \sum_{i=1}^{m} a_{ij}\, \boldsymbol{e}_i = \sum_{i=1}^{m} \Big(\sum_{j=1}^{n} a_{ij} x_j \Big) \boldsymbol{e}_i = \sum_{i=1}^{m} y_i \boldsymbol{e}_i = \boldsymbol{y},$$

where $y_i = \sum_{j=1}^{n} a_{ij} x_j$, $i = 1, \ldots, m$. Similarly, since $\boldsymbol{g}$ is linear and $\boldsymbol{g}(\boldsymbol{e}_i) \in \mathbb{R}^p$, we have

$$\boldsymbol{g}(\boldsymbol{y}) = \sum_{i=1}^{m} y_i \boldsymbol{g}(\boldsymbol{e}_i) = \sum_{i=1}^{m} y_i \sum_{k=1}^{p} b_{ki} \boldsymbol{v}_k = \sum_{k=1}^{p} \Big(\sum_{i=1}^{m} b_{ki} y_i \Big) \boldsymbol{v}_k = \sum_{k=1}^{p} z_k \boldsymbol{v}_k,$$

where

$$z_k = \sum_{i=1}^{m} b_{ki} y_i = \sum_{i=1}^{m} b_{ki} \sum_{j=1}^{n} a_{ij} x_j = \sum_{j=1}^{n} \Big(\sum_{i=1}^{m} b_{ki} a_{ij} \Big) x_j.$$

Also, since $\boldsymbol{g} \circ \boldsymbol{f}$ is linear and $\boldsymbol{g} \circ \boldsymbol{f}(\boldsymbol{u}_j) \in \mathbb{R}^p$,

$$\boldsymbol{g} \circ \boldsymbol{f}(\boldsymbol{x}) = \sum_{j=1}^{n} x_j (\boldsymbol{g} \circ \boldsymbol{f})(\boldsymbol{u}_j) = \sum_{j=1}^{n} x_j \sum_{k=1}^{p} c_{kj} \boldsymbol{v}_k = \sum_{k=1}^{p} \Big(\sum_{j=1}^{n} c_{kj} x_j \Big) \boldsymbol{v}_k = \sum_{k=1}^{p} z_k \boldsymbol{v}_k,$$

where $z_k = \sum_{j=1}^{n} c_{kj} x_j$. Hence $C = BA$, where

$$c_{kj} = \sum_{i=1}^{m} b_{ki} a_{ij}, \quad k = 1, \ldots, p; \;\; j = 1, \ldots, n.$$

3. The proof gives the rule for matrix multiplication.

Since $\boldsymbol{f} \circ \boldsymbol{g} \neq \boldsymbol{g} \circ \boldsymbol{f}$ in general, it follows as a corollary that $AB \neq BA$. That is, the non-commutativity of matrix multiplication is a direct consequence of the non-commutativity of function composition.

The properties of matrix multiplication follow directly from the properties of function composition, for example, the associative property,

$$\boldsymbol{f} \circ (\boldsymbol{g} \circ \boldsymbol{h}) = (\boldsymbol{f} \circ \boldsymbol{g}) \circ \boldsymbol{h} \quad \Rightarrow \quad A(BC) = (AB)C.$$

5 Determinant as a scale factor of area

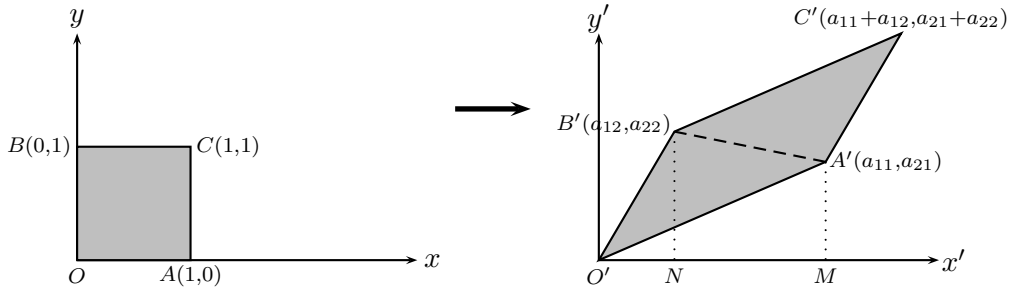
What does the determinant of a square matrix measure? The scale factor of area in linear transformations of the plane provides motivation for the study of determinants which are often introduced devoid of any motivation. Consider the image of the unit square with vertices at $(0, 0)$, $(1, 0)$, $(0, 1)$ and $(1, 1)$. We have

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 0 & a_{11} & a_{12} & a_{11} + a_{12} \\ 0 & a_{21} & a_{22} & a_{21} + a_{22} \end{bmatrix}$$

and so

$$\begin{bmatrix}0\\0\end{bmatrix} \mapsto \begin{bmatrix}0\\0\end{bmatrix}, \quad \begin{bmatrix}1\\0\end{bmatrix} \mapsto \begin{bmatrix}a_{11}\\a_{21}\end{bmatrix}, \quad \begin{bmatrix}0\\1\end{bmatrix} \mapsto \begin{bmatrix}a_{12}\\a_{22}\end{bmatrix}, \quad \begin{bmatrix}1\\1\end{bmatrix} \mapsto \begin{bmatrix}a_{11}+a_{12}\\a_{21}+a_{22}\end{bmatrix}.$$

The image of the unit square is a parallelogram as shown below.



By symmetry, the area of the parallelogram $OA'C'B'$ is twice the area of the triangle $OA'B'$. Now

$$\begin{aligned}\text{area of } \Delta OA'B' &= |\text{area of } \Delta OB'N + \text{area of trapezium } B'NMA' - \text{area of } \Delta OA'M| \\ &= |\tfrac{1}{2}a_{12}a_{22} + \tfrac{1}{2}(a_{21}+a_{22})(a_{11}-a_{12}) - \tfrac{1}{2}a_{11}a_{21}| \\ &= |\tfrac{1}{2}(a_{11}a_{22} - a_{12}a_{21})|.\end{aligned}$$

Hence the area of the parallelogram $OA'C'B'$ is $|a_{11}a_{22} - a_{12}a_{21}|$ which is the absolute value of the determinant of matrix A.

If A is a 3×3 matrix it can be shown that the absolute value of its determinant, $|\det(A)|$, is a measure of the volume of a parallelepiped whose edges are the columns of A. This can be shown equal to $|\boldsymbol{a} \cdot (\boldsymbol{b} \times \boldsymbol{c})|$, where the columns (or rows) of A are $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$.

6 An approach to derivatives

A standard definition given in multivariable calculus states that a function $\boldsymbol{f} : D(\subset \mathbb{R}^n) \to \mathbb{R}^m$ is *differentiable* at $\boldsymbol{x} \in D$ if there exists a linear transformation $\boldsymbol{T_x} : \mathbb{R}^n \to \mathbb{R}^m$ and a function $\boldsymbol{E_x}$ defined in a neighbourhood of $\boldsymbol{0}$ and continuous at $\boldsymbol{0}$, such that

$$\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{h}) = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{T_x}(\boldsymbol{h}) + \|\boldsymbol{h}\|\boldsymbol{E_x}(\boldsymbol{h}). \tag{6.1}$$

If $\boldsymbol{f}$ is differentiable at $\boldsymbol{x}$ its *derivative*, denoted by $\boldsymbol{f}'(\boldsymbol{x})$, is the linear transformation $\boldsymbol{T_x}$. As we have seen in Sections 3 and 4, this linear transformation is represented by a $m \times n$ matrix J. We shall see that J is a matrix of partial derivatives called the Jacobian matrix of $\boldsymbol{f}$ at $\boldsymbol{x}$. The value $\boldsymbol{T_x}(\boldsymbol{h})$ is represented by the matrix-vector product $J\boldsymbol{h}$, where $\boldsymbol{h}$ is represented by a $n \times 1$ column vector. Some applied mathematicians adopt a pragmatic view and define the derivative of $\boldsymbol{f}$ as the Jacobian matrix J. This is sensible if the question of differentiability does not arise.

On the other hand, nearly all textbooks give the definition: A function $f : (a, b) \to \mathbb{R}$ is differentiable at $x \in (a, b)$ if

$$\lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \tag{6.2}$$

exists. If the limit exists it is called the derivative of f at x and is denoted by $f'(x)$. This definition is motivated by rates of change and the slope of a tangent line as a limiting value of the slope of secants. However, it bears little resemblance to the definition in higher dimensions and is a source of difficulty for students. What is needed is a definition in one variable that generalizes more naturally, especially when the concept of a derivative has already been well motivated.

Theorem 2 *Let $K \in \mathbb{R}$ be a constant depending possibly on x but independent of h, and let I_0 denote an open interval centred at 0. The following statements are equivalent.*

(a) $\lim_{h\to 0} \dfrac{f(x+h)-f(x)}{h} = K.$

(b) *There exists a function $E : I_0 \to \mathbb{R}$, continuous at 0, such that $f(x+h) = f(x) + Kh + hE(h)$.*

(c) *There exists a function $F : I_0 \to \mathbb{R}$, continuous at 0, such that $f(x+h) = f(x) + Kh + |h|F(h)$.*

Proof: $(a) \implies (b)$ Define E on an interval I_0 by

$$E(h) = \begin{cases} \dfrac{f(x+h)-f(x)}{h} - K, & \text{if } h \neq 0, \\ 0, & \text{if } h = 0. \end{cases}$$

It then follows that

$$\lim_{h\to 0} E(h) = \lim_{h\to 0} \left(\frac{f(x+h)-f(x)}{h} - K \right) = 0 = E(0),$$

and E is continuous at 0. If $h \neq 0$ the formula for $E(h)$ can be rearranged to give

$$f(x+h) = f(x) + Kh + hE(h).$$

This expression which avoids division is also true when $h = 0$.

$(b) \implies (c)$ Define $F : I_0 \to \mathbb{R}$ by

$$F(h) = \begin{cases} \dfrac{h}{|h|}E(h), & \text{if } h \neq 0, \\ 0, & \text{if } h = 0. \end{cases}$$

Now as $h \to 0^+$, $F(h) = E(h) \to 0$, and as $h \to 0^-$, $F(h) = -E(h) \to 0$. Hence $F(h) \to 0 = F(0)$ as $h \to 0$, and F is continuous at 0. Therefore, $hE(h) = |h|F(h)$ and

$$f(x+h) = f(x) + Kh + hE(h) \;\; \Rightarrow \;\; f(x+h) = f(x) + Kh + |h|F(h).$$

$(c) \implies (a)$ Rearranging the expression yields

$$\lim_{h\to 0} \frac{f(x+h)-f(x)}{h} = \lim_{h\to 0} \left(K + \frac{|h|}{h}F(h) \right) = K,$$

because, since F is continuous at 0, as $h \to 0^+$, $(|h|/h)F(h) = F(h) \to 0$, and as $h \to 0^-$, $(|h|/h)F(h) = -F(h) \to 0$ so that $(|h|/h)F(h) \to 0$ as $h \to 0$. □

Remark: Both the statements (b) and (c) do not involve division by h explicitly and either could be used as an alternative to statement (a) in the definition of differentiability at a point. For example, the statement (b) is used for ease of application, whereas the equivalent statement (c) is a special case of (6.1). We now present some simple examples based on statement (b).

Example 2 *Let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = x^3$. By the binomial theorem,*

$$f(x+h) = (x+h)^3 = x^3 + 3x^2h + 3xh^2 + h^3 = f(x) + Kh + hE(h),$$

where $K = 3x^2$ and $E(h) = h(3x+h)$. Since $E(h) \to 0 = E(0)$ as $h \to 0$ it follows by definition that f is differentiable with derivative $f'(x) = K = 3x^2$. □

Example 3 *Let $f : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ be defined by $f(x) = 1/x$. Here we have*

$$f(x+h) - f(x) = \frac{1}{x+h} - \frac{1}{x} = \frac{-h}{x(x+h)} = \frac{-h}{x^2} + \Big(\frac{h}{x^2} - \frac{h}{x(x+h)}\Big) = Kh + hE(h),$$

where $K = -1/x^2$ and $E(h) = 1/x^2 - 1/x(x+h) \to 0 = E(0)$ as $h \to 0$. Hence f is differentiable at x with derivative $f'(x) = -1/x^2$. □

Example 4 *If $f : [0, \infty) \to \mathbb{R}$ is defined by $f(x) = \sqrt{x}$ we have*

$$\begin{aligned} f(x+h) - f(x) = \sqrt{x+h} - \sqrt{x} &= \frac{(x+h) - x}{\sqrt{x+h} + \sqrt{x}} = \frac{h}{2\sqrt{x}} + \Big(\frac{h}{\sqrt{x+h} + \sqrt{x}} - \frac{h}{2\sqrt{x}}\Big) \\ &= Kh + hE(h), \end{aligned}$$

where $K = \frac{1}{2\sqrt{x}}$ and $E(h) = \frac{1}{\sqrt{x+h}+\sqrt{x}} - \frac{1}{2\sqrt{x}} \to 0 = E(0)$ as $h \to 0$. Hence f is differentiable with derivative $f'(x) = \frac{1}{2\sqrt{x}}$. Note that $x \neq 0$ even though $0 \in dom(f)$ and also that $x + h \geq 0$. This means the interval on which E is defined is restricted in this case. □

Example 5 *Let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = |x|$. Since $\lim_{x\to 0} f(x) = 0 = f(0)$, f is indeed continuous at 0 but its graph shows it is not differentiable there. Let us apply the definition. Now*

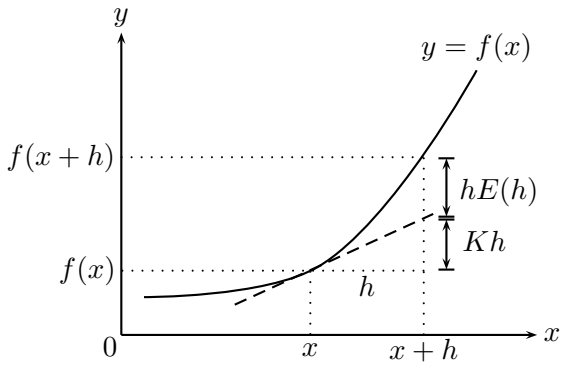
$$f(h) = |h| = \begin{cases} h, & \text{if } h \geq 0, \\ -h, & \text{if } h < 0. \end{cases}$$

If $h > 0$, we have $f(h) = h = f(0) + Kh + hE(h)$, where $K = 1$ and $E(h) = 0 \to 0$ as $h \to 0^+$. If $h < 0$, $f(h) = -h = f(0) + Kh + hE(h)$ with $K = -1$ and $E(h) = 0 \to 0$ as $h \to 0^-$. Since K is not unique, it follows that $f'(0)$ does not exist and f is not differentiable at 0.

Suppose $x > 0$ and $x + h > 0$. Then $f(x+h) = |x+h| = x + h = f(x) + Kh + E(h)$, where $K = 1$ and $E(h) = 0 \to 0$ as $h \to 0$. Hence f is differentiable at x with derivative $f'(x) = 1$. Similarly, $f'(x) = -1$ if $x < 0$. □

As can be seen from the examples, there is little essential difference between using the statement (a) or (b) in the calculation of derivatives. What is different is that the evaluation of limits in (b) is now reduced to finding zero limits (and showing continuity at 0). We remark that a definition based on (b) avoids division by h and therefore the 'sticky and tricky' process of taking limits as $h \to 0$ without allowing $h = 0$ never arises. The statement (b) does not restrict h to nonzero values.

A discussion of the merits of an approach to derivatives based on statement (b) includes its geometrical interpretation and its connection with differentials for a function differentiable at x.



1. The slope or gradient of the tangent line at x (shown dashed) is $K = f'(x)$.

2. The equation of the tangent line is given by the 'linear' part of the expression in (b), that is, $y = f(x) + Kh$ and is obtained for free. Putting $x = a$ and then $h = x - a$ gives the equation of the tangent line at a, $y = f(a) + K(x - a)$.

3. The function E can be called an 'error function' because it is a measure of the difference of the function from its linear approximation. To show differentiability of f at x is to show that E is continuous at 0.

4. In some books the increment h is denoted by δx or Δx with the corresponding change in the function denoted by δy or Δy. Since x is the 'independent' variable we are free to choose the 'differential of x', denoted by dx, to equal h. In this case the differential of y, denoted by dy, is represented by the change to the tangent line, that is,

$$dy = Kh = f'(x)\,dx.$$

The differentials dx and dy are finite quantities, and exist as algebraic elements independently of the notion of 'infinitesimals' used in physics. So it is legitimate to regard the derivative $f'(x)$ as a quotient of two differentials as in

$$\frac{dy}{dx} = f'(x).$$

From the diagram, we then have

$$\Delta y = f(x+h) - f(x) = dy + dxE(dx) \;\Rightarrow\; \frac{\Delta y}{dx} = \frac{dy}{dx} + E(dx) = f'(x) + E(dx),$$

and hence, on recalling that $h = dx = \Delta x$, we have

$$E(h) = \frac{\Delta y}{\Delta x} - \frac{dy}{dx},$$

the difference between the slope of the secant line and the slope of the tangent line. So differentiability is related to

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \frac{dy}{dx}.$$

We now return to consider the Jacobian matrix for a differentiable function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$. For simplicity, we shall consider functions whose domains are the whole of the space. We shall do this in several steps. We have already discussed derivatives of real-valued functions of a single variable. We next consider real and vector-valued functions of several variables. We remark that the terms 'totally differentiable' and 'total derivative' are used widely for 'differentiable' and 'derivative' respectively as a means of distinguishing between various types of partial and directional derivatives.

6.1 Real-valued functions of several variables

First we consider a simple example.

Example 6 *Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by $f(x, y) = x^2 y$. Show f is differentiable at (x, y) and find its derivative $f'(x, y)$. Now*

$$\begin{aligned} f(x+h, y+k) &= (x+h)^2(y+k) = (x^2 + 2xh + h^2)(y+k) \\ &= x^2 y + \Big(2xyh + x^2 k\Big) + yh^2 + 2xhk + h^2 k \\ &= f(x,y) + T_{(x,y)}(h,k) + \|(h,k)\|E(h,k), \end{aligned}$$

where $T_{(x,y)}(h,k) = 2xyh + x^2k$ *and* $E(h,k)$ *is defined by*

$$\|(h,k)\|E(h,k) = yh^2 + 2xhk + h^2k.$$

Now, by the triangle inequality, we have

$$\begin{aligned}\|(h,k)\||E(h,k)| = |yh^2 + 2xhk + h^2k| &\le |y|h^2 + 2|x||h||k| + h^2|k| \\ &\le |y|(h^2+k^2) + 2|x|(h^2+k^2) + (h^2+k^2)^{3/2} \\ &= \|(h,k)\|^2\Big(|y| + 2|x| + \|(h,k)\|\Big), \\ \implies \quad |E(h,k)| &\le \|(h,k)\|\Big(|y| + 2|x| + \|(h,k)\|\Big),\end{aligned}$$

since $|h| \le \sqrt{h^2+k^2}$ *and* $|k| \le \sqrt{h^2+k^2}$. *It follows that as* $(h,k) \to (0,0)$, $|E(h,k)| \to 0$ *and hence* $E(h,k) \to 0$. *It is continuous at* $(0,0)$ *if we define* $E(0,0) = 0$. *Therefore, by definition,* f *is differentiable at* (x,y) *with derivative given by* $f'(x,y) = (2xy, x^2)$. □

Consider a function $f : \mathbb{R}^n \to \mathbb{R}$ with $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and $\boldsymbol{h} = (h_1, \ldots, h_n) \in \mathbb{R}^n$. By definition (6.1), if f is differentiable at $\boldsymbol{x}$ then

$$f(\boldsymbol{x}+\boldsymbol{h}) = f(\boldsymbol{x}) + T_{\boldsymbol{x}}(\boldsymbol{h}) + \|\boldsymbol{h}\|E_{\boldsymbol{x}}(\boldsymbol{h}), \tag{6.3}$$

where $T_{\boldsymbol{x}}$ is a linear function and $E_{\boldsymbol{x}}(\boldsymbol{h}) \to 0 = E_{\boldsymbol{x}}(\boldsymbol{0})$ as $\boldsymbol{h} \to \boldsymbol{0}$. Let $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ be the unit coordinate vectors in $\mathbb{R}^n$. Then, by linearity, we have

$$\boldsymbol{h} = \sum_{j=1}^{n} h_j \boldsymbol{u}_j \quad \implies \quad T_{\boldsymbol{x}}(\boldsymbol{h}) = \sum_{j=1}^{n} h_j T_{\boldsymbol{x}}(\boldsymbol{u}_j). \tag{6.4}$$

Consider $\boldsymbol{h} \to \boldsymbol{0}$ along the j-th axis, that is, $\boldsymbol{h} = h_j\boldsymbol{u}_j$ with $h_j \to 0$. Then $T_{\boldsymbol{x}}(\boldsymbol{h}) = h_j T_{\boldsymbol{x}}(\boldsymbol{u}_j)$ and

$$f(\boldsymbol{x} + h_j\boldsymbol{u}_j) = f(\boldsymbol{x}) + h_j T_{\boldsymbol{x}}(\boldsymbol{u}_j) + |h_j| E_{\boldsymbol{x}}(h_j\boldsymbol{u}_j)$$

with $E_{\boldsymbol{x}}(h_j\boldsymbol{u}_j) \to 0 = E_{\boldsymbol{x}}(\boldsymbol{0})$ as $h_j \to 0$. Hence, by Theorem 2 (c), the *partial derivative* of f along the j-th direction exists and equals $T_{\boldsymbol{x}}(\boldsymbol{u}_j)$ for each $j = 1, \ldots, n$,

$$T_{\boldsymbol{x}}(\boldsymbol{u}_j) = \frac{\partial f(\boldsymbol{x})}{\partial x_j}.$$

Substituting into (6.4) we therefore obtain

$$T_{\boldsymbol{x}}(\boldsymbol{h}) = \sum_{j=1}^{n} h_j \frac{\partial f(\boldsymbol{x})}{\partial x_j} = \nabla f(\boldsymbol{x}) \cdot \boldsymbol{h},$$

where $\nabla f(\boldsymbol{x})$ is the gradient of $f(\boldsymbol{x})$ represented by the $1 \times n$ row-vector,

$$\nabla f(\boldsymbol{x}) = \left[\frac{\partial f(\boldsymbol{x})}{\partial x_1}, \ldots, \frac{\partial f(\boldsymbol{x})}{\partial x_j}, \ldots, \frac{\partial f(\boldsymbol{x})}{\partial x_n}\right].$$

The dot product is represented by matrix multiplication if $\boldsymbol{h}$ is represented by an $n \times 1$ column vector.

Recall that in $\mathbb{R}^2$ the equation of a tangent line to the curve $y = f(x)$ at x_0 is given by

$$y = f(x_0) + f'(x_0)(x - x_0).$$

In R^3 this generalizes to the equation of a tangent plane to the surface $z = f(x,y)$ at (x_0, y_0) given by

$$z = f(x_0,y_0) + \nabla f(x_0,y_0) \cdot (x - x_0, y - y_0) = f(x_0,y_0) + (x - x_0)\frac{\partial f}{\partial x}\bigg|_{(x_0,y_0)} + (y - y_0)\frac{\partial f}{\partial y}\bigg|_{(x_0,y_0)}.$$

6.2 Vector-valued functions of several variables

Finally, we consider $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$. Let $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$, $\boldsymbol{h} = (h_1, \ldots, h_n) \in \mathbb{R}^n$, and let $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ and $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_m$ be the unit coordinate vectors in $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively. Then $\boldsymbol{T_x} : \mathbb{R}^n \to \mathbb{R}^m$ is linear and $\boldsymbol{T_x}(\boldsymbol{h}) \in \mathbb{R}^m$ and so

$$\boldsymbol{T_x}(\boldsymbol{h}) = \sum_{i=1}^{m} T^i_{\boldsymbol{x}}(\boldsymbol{h})\boldsymbol{e}_i.$$

Since $\boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^m$, we can write $\boldsymbol{f}(\boldsymbol{x}) = \sum_{i=1}^m f_i(\boldsymbol{x})\boldsymbol{e}_i$, where $f_i : \mathbb{R}^n \to \mathbb{R}$. Similarly, $\boldsymbol{E_x}(\boldsymbol{h}) = \sum_{i=1}^m E^i_{\boldsymbol{x}}(\boldsymbol{h})\boldsymbol{e}_i$. If $\boldsymbol{f}$ is differentiable at $\boldsymbol{x}$, then

$$\sum_{i=1}^{m} f_i(\boldsymbol{x}+\boldsymbol{h})\boldsymbol{e}_i = \sum_{i=1}^{m} f_i(\boldsymbol{x})\boldsymbol{e}_i + \sum_{i=1}^{m} T^i_{\boldsymbol{x}}(\boldsymbol{h})\boldsymbol{e}_i + \|\boldsymbol{h}\| \sum_{i=1}^{m} E^i_{\boldsymbol{x}}(\boldsymbol{h})\boldsymbol{e}_i,$$

with $E^i_{\boldsymbol{x}}(\boldsymbol{h}) \to 0 = E^i_{\boldsymbol{x}}(\boldsymbol{0})$ as $\boldsymbol{h} \to \boldsymbol{0}$ for each $i = 1, \ldots, m$. Hence

$$f_i(\boldsymbol{x}+\boldsymbol{h}) = f_i(\boldsymbol{x}) + T^i_{\boldsymbol{x}}(\boldsymbol{h}) + \|\boldsymbol{h}\|E^i_{\boldsymbol{x}}(\boldsymbol{h})$$

which has the same form as (6.3). Noting that $T^i_{\boldsymbol{x}} : \mathbb{R}^n \to \mathbb{R}$ is linear for each $i = 1, \ldots, m$ we follow the same procedure as in Section 6.1 and conclude

$$T^i_{\boldsymbol{x}}(\boldsymbol{h}) = \nabla f_i(\boldsymbol{x}) \cdot \boldsymbol{h},$$

and, in terms of matrix representation, we have

$$\boldsymbol{T_x}(\boldsymbol{h}) = \begin{bmatrix} \nabla f_1(\boldsymbol{x}) \cdot \boldsymbol{h} \\ \vdots \\ \nabla f_i(\boldsymbol{x}) \cdot \boldsymbol{h} \\ \vdots \\ \nabla f_m(\boldsymbol{x}) \cdot \boldsymbol{h} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \frac{\partial f_1(\boldsymbol{x})}{\partial x_2} & \cdots & \frac{\partial f_1(\boldsymbol{x})}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_i(\boldsymbol{x})}{\partial x_1} & \frac{\partial f_i(\boldsymbol{x})}{\partial x_2} & \cdots & \frac{\partial f_i(\boldsymbol{x})}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_m(\boldsymbol{x})}{\partial x_1} & \frac{\partial f_m(\boldsymbol{x})}{\partial x_2} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_n} \end{bmatrix} \begin{bmatrix} h_1 \\ \vdots \\ h_j \\ \vdots \\ h_n \end{bmatrix}.$$

Hence the Jacobian matrix is the $m \times n$ matrix of partial derivatives,

$$J(\boldsymbol{f}(\boldsymbol{x})) = \frac{\partial(f_1, \ldots, f_m)}{\partial(x_1, \ldots, x_n)} = \begin{bmatrix} \nabla f_1(\boldsymbol{x}) \\ \vdots \\ \nabla f_i(\boldsymbol{x}) \\ \vdots \\ \nabla f_m(\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \frac{\partial f_1(\boldsymbol{x})}{\partial x_2} & \cdots & \frac{\partial f_1(\boldsymbol{x})}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_i(\boldsymbol{x})}{\partial x_1} & \frac{\partial f_i(\boldsymbol{x})}{\partial x_2} & \cdots & \frac{\partial f_i(\boldsymbol{x})}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_m(\boldsymbol{x})}{\partial x_1} & \frac{\partial f_m(\boldsymbol{x})}{\partial x_2} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_n} \end{bmatrix}.$$

6.3 The chain rule

Theorem 3 *If f is differentiable at x with derivative $f'(x)$ and g is differentiable at $f(x)$ with derivative $g'(f(x))$, then the composition $g \circ f$ is differentiable at x with derivative $(g \circ f)'(x) = g'(f(x))f'(x)$.*

Proof: Suppose f is differentiable at $x \in \mathrm{dom}(f)$ and let $x + h \in \mathrm{dom}(f)$. We then have

$$f(x+h) = f(x) + f'(x)h + hF_x(h),$$

with F_x continuous at 0. If f maps its domain into the domain of g, then let $y = f(x)$ and $y + k = f(x + h)$ so that, by the continuity of f at x,

$$k = f(x+h) - f(x) \to 0 = k(0) \text{ as } h \to 0.$$

If g is also differentiable at y, then

$$g(y+k) = g(y) + g'(y)k + kG_y(k),$$

with G_y continuous at 0. Hence

$$\begin{aligned} g(f(x+h)) &= g(f(x)) + g'(f(x))\Big(f'(x)h + hF_x(h)\Big) + kG_y(k) \\ &= g(f(x)) + g'(f(x))f'(x)h + hE_x(h), \end{aligned}$$

where

$$hE_x(h) = hg'(f(x))F_x(h) + kG_y(k).$$

Now, by the triangular inequality,

$$\begin{aligned} & |k| = |f(x+h) - f(x)| \le |h|(|f'(x)| + |F_x(h)|) \\ \Rightarrow \quad & |h||E_x(h)| \le |h|\Big(|g'(f(x))||F_x(h)| + |G_y(k)|(|f'(x)| + |F_x(h)|)\Big) \\ \Rightarrow \quad & |E_x(h)| \le |g'(f(x))||F_x(h)| + |G_y(k)|(|f'(x)| + |F_x(h)|). \end{aligned}$$

As $h \to 0$ we have $k \to 0 = k(0)$, $F_x(h) \to 0 = F_x(0)$ and $G_y(k) \to 0 = G_y(0)$ and therefore $E_x(h) \to 0 = E_x(0)$. Hence $g \circ f$ is differentiable at x with derivative $(g \circ f)'(x) = g'(f(x))f'(x)$. □

The proof given generalizes in an obvious way to the general case as we illustrate. Let $\boldsymbol{f}\colon \mathrm{dom}(\boldsymbol{f})(\subset \mathbb{R}^n) \to \mathbb{R}^m$ and $\boldsymbol{g}\colon \mathrm{dom}(\boldsymbol{g})(\subset \mathbb{R}^m) \to \mathbb{R}^p$ be functions with the range of $\boldsymbol{f}$, $\boldsymbol{f}(\mathrm{dom}(\boldsymbol{f}))$, a subset of the domain of $\boldsymbol{g}$. The chain rule then reads:

Theorem 4 *If $\boldsymbol{f}$ is differentiable at $\boldsymbol{x}$ with derivative $\boldsymbol{f}'(\boldsymbol{x})$ and $\boldsymbol{g}$ is differentiable at $\boldsymbol{f}(\boldsymbol{x})$ with derivative $\boldsymbol{g}'(\boldsymbol{f}(\boldsymbol{x}))$, then the composition $\boldsymbol{g} \circ \boldsymbol{f}$ is differentiable at $\boldsymbol{x}$ with derivative $(\boldsymbol{g} \circ \boldsymbol{f})'(\boldsymbol{x}) = \boldsymbol{g}'(\boldsymbol{f}(\boldsymbol{x}))\boldsymbol{f}'(\boldsymbol{x})$.*

Proof: Suppose $\boldsymbol{f}$ is differentiable at $\boldsymbol{x}$ and let $\boldsymbol{x}+\boldsymbol{h} \in \mathrm{dom}(\boldsymbol{f})$. We then have

$$\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{h}) = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{f}'(\boldsymbol{x})(\boldsymbol{h}) + \|\boldsymbol{h}\|\boldsymbol{F}_{\boldsymbol{x}}(\boldsymbol{h}),$$

with $\boldsymbol{F}_{\boldsymbol{x}}$ continuous at $\boldsymbol{0}$. If $\boldsymbol{f}$ maps its domain into the domain of $\boldsymbol{g}$, then let $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$ and $\boldsymbol{y}+\boldsymbol{k} = \boldsymbol{f}(\boldsymbol{x}+\boldsymbol{h})$ so that, by the continuity of $\boldsymbol{f}$ at $\boldsymbol{x}$,

$$\boldsymbol{k} = \boldsymbol{f}(\boldsymbol{x}+\boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{x}) \to \boldsymbol{0} = \boldsymbol{k}(\boldsymbol{0}) \ \text{ as } \ \boldsymbol{h} \to \boldsymbol{0}.$$

If $\boldsymbol{g}$ is also differentiable at $\boldsymbol{y}$, then

$$\boldsymbol{g}(\boldsymbol{y}+\boldsymbol{k}) = \boldsymbol{g}(\boldsymbol{y}) + \boldsymbol{g}'(\boldsymbol{y})(\boldsymbol{k}) + \|\boldsymbol{k}\|\boldsymbol{G}_{\boldsymbol{y}}(\boldsymbol{k}),$$

with $\boldsymbol{G}_{\boldsymbol{y}}$ continuous at $\boldsymbol{0}$. Hence

$$\begin{aligned} \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{h})) &= \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x})) + \boldsymbol{g}'(\boldsymbol{f}(\boldsymbol{x}))\Big(\boldsymbol{f}'(\boldsymbol{x})(\boldsymbol{h}) + \|\boldsymbol{h}\|\boldsymbol{F}_{\boldsymbol{x}}(\boldsymbol{h})\Big) + \|\boldsymbol{k}\|\boldsymbol{G}_{\boldsymbol{y}}(\boldsymbol{k}) \\ &= \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x})) + \boldsymbol{g}'(\boldsymbol{f}(\boldsymbol{x}))\boldsymbol{f}'(\boldsymbol{x})(\boldsymbol{h}) + \|\boldsymbol{h}\|\boldsymbol{E}_{\boldsymbol{x}}(\boldsymbol{h}), \end{aligned}$$

where

$$\|\boldsymbol{h}\|\boldsymbol{E}_{\boldsymbol{x}}(\boldsymbol{h}) = \|\boldsymbol{h}\|\boldsymbol{g}'(\boldsymbol{f}(\boldsymbol{x}))\boldsymbol{F}_{\boldsymbol{x}}(\boldsymbol{h}) + \|\boldsymbol{k}\|\boldsymbol{G}_{\boldsymbol{y}}(\boldsymbol{k}).$$

Now, by the triangular and the Cauchy-Schwarz inequality,

$$\begin{aligned} & \|\boldsymbol{k}\| = \|\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{x})\| \le \|\boldsymbol{h}\|(\|\boldsymbol{f}'(\boldsymbol{x})| + \|\boldsymbol{F}_{\boldsymbol{x}}(\boldsymbol{h})\|) \\ \Rightarrow \quad & \|\boldsymbol{h}\|\|\boldsymbol{E}_{\boldsymbol{x}}(\boldsymbol{h})\| \le \|\boldsymbol{h}\|\Big(\|\boldsymbol{g}'(\boldsymbol{f}(\boldsymbol{x}))\|\|\boldsymbol{F}_{\boldsymbol{x}}(\boldsymbol{h})\| + \|\boldsymbol{G}_{\boldsymbol{y}}(\boldsymbol{k})\|(\|\boldsymbol{f}'(\boldsymbol{x})\| + \|\boldsymbol{F}_{\boldsymbol{x}}(\boldsymbol{h})\|)\Big) \\ \Rightarrow \quad & \|\boldsymbol{E}_{\boldsymbol{x}}(\boldsymbol{h})\| \le \|\boldsymbol{g}'(\boldsymbol{f}(\boldsymbol{x}))\|\|\boldsymbol{F}_{\boldsymbol{x}}(\boldsymbol{h})\| + \|\boldsymbol{G}_{\boldsymbol{y}}(\boldsymbol{k})\|(\|\boldsymbol{f}'(\boldsymbol{x})\| + \|\boldsymbol{F}_{\boldsymbol{x}}(\boldsymbol{h})\|). \end{aligned}$$

As $\boldsymbol{h} \to \boldsymbol{0}$ we have $\boldsymbol{k} \to \boldsymbol{0} = \boldsymbol{k}(\boldsymbol{0})$, $\boldsymbol{F}_{\boldsymbol{x}}(\boldsymbol{h}) \to \boldsymbol{0} = \boldsymbol{F}_{\boldsymbol{x}}(\boldsymbol{0})$ and $\boldsymbol{G}_{\boldsymbol{y}}(\boldsymbol{k}) \to \boldsymbol{0} = \boldsymbol{G}_{\boldsymbol{y}}(\boldsymbol{0})$ and therefore $\boldsymbol{E}_{\boldsymbol{x}}(\boldsymbol{h}) \to \boldsymbol{0} = \boldsymbol{E}_{\boldsymbol{x}}(\boldsymbol{0})$. Hence $\boldsymbol{g} \circ \boldsymbol{f}$ is differentiable at $\boldsymbol{x}$ with derivative $(\boldsymbol{g} \circ \boldsymbol{f})'(\boldsymbol{x}) = \boldsymbol{g}'(\boldsymbol{f}(\boldsymbol{x}))\boldsymbol{f}'(\boldsymbol{x})$. □

Remark: As we can see this proof is identical with that given in the one-dimensional case, with the exception that the Cauchy-Schwarz inequality is needed here.

6.4 Matrix form of the chain rule

Let $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$, $\boldsymbol{z} = \boldsymbol{g}(\boldsymbol{y})$ and denote $\boldsymbol{g} \circ \boldsymbol{f} = \boldsymbol{h}$. When the derivatives are represented by Jacobian matrices the chain rule takes the form

$$\begin{aligned}
\boldsymbol{h}'(\boldsymbol{x}) &= \boldsymbol{g}'(\boldsymbol{f}(\boldsymbol{x}))\boldsymbol{f}'(\boldsymbol{x}), \\
J(\boldsymbol{h}(\boldsymbol{x})) &= J(\boldsymbol{g}(\boldsymbol{y}))J(\boldsymbol{f}(\boldsymbol{x})), \\
\frac{\partial(h_1,\ldots,h_p)}{\partial(x_1,\ldots,x_n)} &= \frac{\partial(g_1,\ldots,g_p)}{\partial(y_1,\ldots,y_m)}\frac{\partial(f_1,\ldots,f_m)}{\partial(x_1,\ldots,x_n)}, \\
\begin{bmatrix} \frac{\partial z_1}{\partial x_1} & \cdots & \frac{\partial z_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial z_p}{\partial x_1} & \cdots & \frac{\partial z_p}{\partial x_n} \end{bmatrix} &= \begin{bmatrix} \frac{\partial z_1}{\partial y_1} & \cdots & \frac{\partial z_1}{\partial y_m} \\ \vdots & & \vdots \\ \frac{\partial z_p}{\partial y_1} & \cdots & \frac{\partial z_p}{\partial y_m} \end{bmatrix} \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}.
\end{aligned}$$

6.5 Some miscellaneous results

We consider some immediate results following from the differentiability of a function at a point. We do this for a function $\boldsymbol{f}\colon \mathbb{R}^n \to \mathbb{R}^m$. The first generalizes the one dimensional result that differentiability implies continuity.

Theorem 5 *If $\boldsymbol{f}$ is differentiable at $\boldsymbol{x}$, it is continuous at $\boldsymbol{x}$.*

Proof: If $\boldsymbol{f}$ is differentiable at $\boldsymbol{x}$, then

$$\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{h}) = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{f}'(\boldsymbol{x})(\boldsymbol{h}) + \|\boldsymbol{h}\|\boldsymbol{E}_{\boldsymbol{x}}(\boldsymbol{h})$$

with $\boldsymbol{E}_{\boldsymbol{x}}$ continuous $\boldsymbol{0}$. It follows that $\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{h}) \to \boldsymbol{f}(\boldsymbol{x})$ as $\boldsymbol{h} \to \boldsymbol{0}$. Hence $\boldsymbol{f}$ is continuous at $\boldsymbol{x}$. □

The next result on continuity of a composition follows as a corollary of this result and by the chain rule of Theorem 4.

Theorem 6 *If $\boldsymbol{f}$ is continuous at $\boldsymbol{x}$ and $\boldsymbol{g}$ is continuous at $\boldsymbol{f}(\boldsymbol{x})$, then the composition $\boldsymbol{g} \circ \boldsymbol{f}$ is continuous at $\boldsymbol{x}$.*

Definition 1 *(Directional derivative) The directional derivative of $\boldsymbol{f}$ at $\boldsymbol{x}$ in the direction of $\boldsymbol{s}$ is defined to be the limit*

$$\lim_{t\to 0} \frac{\boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{s}) - \boldsymbol{f}(\boldsymbol{x})}{t}$$

if it exists in which case it is denoted by $\boldsymbol{f}'(\boldsymbol{x};\boldsymbol{s})$.

Theorem 7 *The directional derivative of $\boldsymbol{f}$ at $\boldsymbol{x}$ in the direction of $\boldsymbol{s}$ exists if and only if there exist $\boldsymbol{K}(\boldsymbol{x};\boldsymbol{s}) \in \mathbb{R}^m$ and a function $\boldsymbol{E} : I_0 \to \mathbb{R}^m$, where I_0 is some interval centred at 0, with $\lim_{t\to 0} \boldsymbol{E}(t) = \boldsymbol{0} = \boldsymbol{E}(0)$ such that*

$$\boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{s}) = \boldsymbol{f}(\boldsymbol{x}) + t\boldsymbol{K}(\boldsymbol{x};\boldsymbol{s}) + t\boldsymbol{E}(t) \text{ for all } t \in I_0.$$

Proof: ($\Longrightarrow$) If $\boldsymbol{f}'(\boldsymbol{x};\boldsymbol{s})$ exists, then choose $\boldsymbol{K}(\boldsymbol{x};\boldsymbol{s}) = \boldsymbol{f}'(\boldsymbol{x};\boldsymbol{s})$ and define $\boldsymbol{E}$ by

$$\boldsymbol{E}(t) = \begin{cases} \dfrac{\boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{s}) - \boldsymbol{f}(\boldsymbol{x})}{t} - \boldsymbol{f}'(\boldsymbol{x};\boldsymbol{s}), & \text{if } t \neq 0, \\ \boldsymbol{0}, & \text{if } t = 0. \end{cases}$$

Hence we have

$$\lim_{t\to 0} \boldsymbol{E}(t) = \lim_{t\to 0} \left(\frac{\boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{s}) - \boldsymbol{f}(\boldsymbol{x})}{t} - \boldsymbol{f}'(\boldsymbol{x};\boldsymbol{s}) \right) = \boldsymbol{0} = \boldsymbol{E}(0),$$

and the required expression then follows from a simple rearrangement of the defining formula for $\boldsymbol{E}$.

($\Longleftarrow$) Conversely, given $\boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{s}) = \boldsymbol{f}(\boldsymbol{x}) + t\boldsymbol{K}(\boldsymbol{x};\boldsymbol{s}) + t\boldsymbol{E}(t)$ with $\boldsymbol{E}(t) \to \boldsymbol{0} = \boldsymbol{E}(0)$ as $t \to 0$, we have

$$\lim_{t\to 0} \frac{\boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{s}) - \boldsymbol{f}(\boldsymbol{x})}{t} = \lim_{t\to 0} \Big(\boldsymbol{K}(\boldsymbol{x};\boldsymbol{s}) + \boldsymbol{E}(t) \Big) = \boldsymbol{K}(\boldsymbol{x};\boldsymbol{s}).$$

Hence $\boldsymbol{f}$ has a directional derivative at $\boldsymbol{x}$ in the direction $\boldsymbol{s}$, $\boldsymbol{f}'(\boldsymbol{x};\boldsymbol{s}) = \boldsymbol{K}(\boldsymbol{x};\boldsymbol{s})$. □

Remarks:

1. Theorem 7 may be used instead of Definition 1 to define directional derivatives.

2. The direction vector $\boldsymbol{s}$ is usually taken to be a unit vector.

3. Let $\boldsymbol{s}$ be any vector in $\mathbb{R}^n$ and let $\boldsymbol{g}$ denote the composition $t \mapsto \boldsymbol{x} + t\boldsymbol{s} \mapsto \boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{s})$, so that $\boldsymbol{g}(t) = \boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{s})$. Then $\boldsymbol{f}$ is differentiable at $\boldsymbol{x}$ if and only if $\boldsymbol{g}$ is differentiable at 0, that is,

   $$\boldsymbol{g}(t) = \boldsymbol{g}(0) + \boldsymbol{g}'(0)t + t\boldsymbol{G}(t)$$

   with $\boldsymbol{G}$ continuous at 0. Now $\boldsymbol{g}'(0) = \boldsymbol{f}'(\boldsymbol{x})\boldsymbol{s}$ by the chain rule and, by Theorem 7, it follows that $\boldsymbol{f}'(\boldsymbol{x};\boldsymbol{s}) = \boldsymbol{f}'(\boldsymbol{x})\boldsymbol{s} = J(\boldsymbol{f}(\boldsymbol{x}))\cdot\boldsymbol{s}$. In particular, if $\boldsymbol{f}$ is real-valued we have $f'(\boldsymbol{x};\boldsymbol{s}) = \nabla f(\boldsymbol{x})\cdot\boldsymbol{s}$ and if $\boldsymbol{s} = \boldsymbol{u}_j$, the unit coordinate vector along the j-th axis then the directional derivative is the partial derivative, $f'(\boldsymbol{x};\boldsymbol{u}_j) = \partial f(\boldsymbol{x})/\partial x_j$. Hence directional derivatives at $\boldsymbol{x}$ exist in all directions for a function differentiable at $\boldsymbol{x}$.