CONTINUITY AND DIFFERENTIATION

TSOGTGEREL GANTUMUR

Abstract. After reviewing the notion of continuity for univariate functions, we extend it to multivariate functions. Then we treat differentiability in an analogous way.

Contents

1. Continuity of univariate scalar functions

Let us recall first a definition of continuous functions.

- Intuitively, a continuous function f sends nearby points to nearby points, i.e, if x is close to y then $f(x)$ is close to $f(y)$.
- That is, continuous functions are the ones that send convergent sequences to convergent sequences. This is sometimes called the *sequential criterion of continuity*.

Definition 1.1. Let $K \subset \mathbb{R}$ be a set. A function $f : K \to \mathbb{R}$ is said to be *continuous* at $y \in K$ if and only if $f(x_n) \to f(y)$ as $n \to \infty$ for every sequence $\{x_n\} \subset K$ converging to y.

Example 1.2. (a) Let $f : \mathbb{R} \to \mathbb{R}$ be the function given by $f(x) = x$ for $x \in \mathbb{R}$. Then f is continuous at every point $y \in \mathbb{R}$, because given any sequence $\{x_n\} \subset \mathbb{R}$ converging to y, we have $f(x_n) = x_n \to y = f(y)$ as $n \to \infty$.

(b) Let $g : \mathbb{R} \to \mathbb{R}$ be the function given by $g(x) = |x|$ for $x \in \mathbb{R}$. Then f is continuous at every point $y \in \mathbb{R}$, because given any sequence $\{x_n\} \subset \mathbb{R}$ converging to y, we have $g(x_n) = |x_n| \to |y| = f(y)$ as $n \to \infty$.

(c) We define the *Heaviside step function* $\theta : \mathbb{R} \to \mathbb{R}$ by

$$\theta(x) = \begin{cases} 1 & \text{for } x > 0, \\ 0 & \text{for } x \leq 0. \end{cases} \tag{1}$$

It is clear that θ is continuous at every $x \in \mathbb{R} \setminus \{0\}$. Our intuition tells us that θ is not continuous at $x = 0$. Indeed, let $x_n = \frac{1}{n}$ and $y_n = -\frac{1}{n}$ for $n \in \mathbb{N}$. Then we have $x_n \to 0$ and $y_n \to 0$, but $\theta(x_n) \to 1$ and $\theta(y_n) \to 0$ as $n \to \infty$. Since $1 \neq 0$, the sequential criterion of continuity implies that θ is not continuous at $x = 0$.

Date: 2019/09/14.

(d) The *Dirichlet function* $h : \mathbb{R} \to \mathbb{R}$ is defined by

$$h(x) = \begin{cases} 1 & \text{for } x \in \mathbb{Q}, \\ 0 & \text{for } x \in \mathbb{R} \setminus \mathbb{Q}. \end{cases} \tag{2}$$

For any $x \in \mathbb{R}$, we can find two sequences $\{x_n\} \subset \mathbb{Q}$ and $\{y_n\} \subset \mathbb{R} \setminus \mathbb{Q}$ satisfying $x_n \to x$ and $y_n \to x$ as $n \to \infty$. Since $h(x_n) = 1$ and $h(y_n) = 0$, we have $h(x_n) \to 1$ and $h(y_n) \to 0$, and hence we conclude that h is not continuous at any point $x \in \mathbb{R}$.

Remark 1.3. This remark contains the main practical message of this section.

- In order to show that f is *discontinuous* at y, it suffices to exhibit a sequence $\{x_n\}$ with $x_n \to y$, such that $f(x_n) \not\to f(y)$ as $n \to \infty$.
- If a function is given by a formula, verifying its continuity is usually not hard, because continuous building blocks produce continuous functions. In the rest of this section we shall justify the latter statement.

Lemma 1.4. *Let $K \subset \mathbb{R}$, and let $f, g : K \to \mathbb{R}$ be functions continuous at $x \in K$. Then the sum and difference $f \pm g$, and the product fg are all continuous at x. Moreover, the function $\frac{1}{f}$ is continuous at x, provided that $f(x) \neq 0$.*

Proof. The results are immediate from the definition of continuity. For instance, let us prove that fg is continuous at x. Thus let $\{x_n\} \subset K$ be an arbitrary sequence converging to x. Then $f(x_n) \to f(x)$ and $g(x_n) \to g(x)$ as $n \to \infty$, and hence $f(x_n)g(x_n) \to f(x)g(x)$ as $n \to \infty$. Therefore fg is continuous at x. □

Exercise 1.5. Complete the proof of the preceding lemma.

Example 1.6. (a) We know that the constant function $f(x) = c$ (where $c \in \mathbb{R}$) and the identity function $f(x) = x$ are continuous in $\mathbb{R}$. Then by Lemma 1.4, any monomial $f(x) = ax^n$ with constants $a \in \mathbb{R}$ and $n \in \mathbb{N}_0$, is continuous in $\mathbb{R}$, since we can write $ax^n = a \cdot x \cdots x$. A univariate *polynomial* is a function $p : \mathbb{R} \to \mathbb{R}$ of the form

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + a_0. \tag{3}$$

By Lemma 1.4 again, we conclude that all univariate polynomials are continuous in $\mathbb{R}$.

(b) Let p and q be polynomials, and let $Z = \{x \in \mathbb{R} : q(x) = 0\}$ be the set of zeroes of q. Then by Lemma 1.4, the function $r : \mathbb{R} \setminus Z \to \mathbb{R}$ given by $r(x) = \frac{p(x)}{q(x)}$ is continuous in $\mathbb{R} \setminus Z$. The functions of this form are called *rational functions*. For instance, $f(x) = \frac{x^2+1}{(x-1)(x-2)}$ is continuous at each $x \in \mathbb{R} \setminus \{1, 2\}$.

(c) The function $\tan x = \frac{\sin x}{\cos x}$ is continuous wherever $\cos x \neq 0$.

Lemma 1.7. *Let $K \subset \mathbb{R}$, and let $g : K \to \mathbb{R}$ be a function whose components are all continuous at $x \in K$. Suppose that $U \subset \mathbb{R}$ satisfies $g(K) \subset U$, the latter meaning that $y \in K$ implies $g(y) \in U$. Let $F : U \to \mathbb{R}$ be a function continuous at $g(x)$. Then the composition $F \circ g : K \to \mathbb{R}$, defined by $(F \circ g)(y) = F(g(y))$, is continuous at x.*

Example 1.8. (a) The function $f(x) = \cos(2x + \sin x)$ is continuous in $\mathbb{R}$.

(b) The function $g(x) = \tan(e^x)$ is continuous wherever $e^x \neq \frac{\pi}{2} + \pi n$ for any $n \in \mathbb{Z}$.

2. Continuity of univariate vector functions

The set of all *ordered pairs* of real numbers is denoted by

$$\mathbb{R}^2 = \mathbb{R} \times \mathbb{R} = \{(x, y) : x, y \in \mathbb{R}\}. \tag{4}$$

Ordered means that, for instance, $(1, 3) \neq (3, 1)$. As an example, the position of a point on the surface of the Earth can be described by an element of $\mathbb{R}^2$, by its latitude and longitude.

- For $n \in \mathbb{N}$, we let
$$\mathbb{R}^n = \underbrace{\mathbb{R} \times \ldots \times \mathbb{R}}_{n \text{ times}} = \{(x_1, x_2, \ldots, x_n) : x_1, x_2, \ldots, x_n \in \mathbb{R}\}. \tag{5}$$
An element of $\mathbb{R}^n$ is called an n-tuple, an n-vector, a point, or simply a *vector*.
- In a context where both $\mathbb{R}$ and $\mathbb{R}^n$ (with $n > 1$) are present, an element of $\mathbb{R}$ (i.e., a real number) is called a *scalar*.
- Given a vector $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, the number $x_k \in \mathbb{R}$ is called the *k-th component of x*, for $k \in \{1, \ldots, n\}$.
- At times, it is convenient to denote a vector in $\mathbb{R}^2$ by (x, y) with x and y being real numbers, instead of using subscripts as in (x_1, x_2).
- The advantage of the notation (x_1, x_2), apart from its generalization to n dimensions, is that the vector itself can be denoted by $x = (x_1, x_2)$, whereas for (x, y) we need to invent a letter, such as $P = (x, y)$.

Example 2.1. Consider n foreign currencies, and let x_k be the exchange rate between the k-th currency and Canadian dollar (at a certain moment of time). Then any possible outcome $(x_1, \ldots, x_n)$ can be considered as an element of $\mathbb{R}^n$.

Definition 2.2. For $x, y \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$, we define
$$x \pm y = (x_1 \pm y_1, \ldots, x_n \pm y_n) \qquad \text{and} \qquad \alpha x = x\alpha = (\alpha x_1, \ldots, \alpha x_n). \tag{6}$$

Example 2.3. For $x = (1, -3) \in \mathbb{R}^2$, we have $2x = x \cdot 2 = (2, -6)$, and $x + (-2, 3) = (-1, 0)$.

Let $K \subset \mathbb{R}$, and let $f : K \to \mathbb{R}^n$ be a function.

- Such functions are called *vector valued functions* (of a single variable).
- In contrast, $\mathbb{R}$-valued functions (i.e., $n = 1$) are called *scalar valued functions*.
- For $t \in K$, the value $f(t)$ is an n-vector; Let us denote the k-th component of $f(t) \in \mathbb{R}^n$ by $f_k(t) \in \mathbb{R}$.
- Since t can be any point in K, this defines a function $f_k : K \to \mathbb{R}$, called the *k-th component of f*, for each $k \in \{1, \ldots, n\}$.
- Thus a vector valued function is simply a *collection of scalar valued functions*.

Example 2.4. (a) If $f(t) \in \mathbb{R}^2$ denotes the latitude and longitude of a car at the time moment $t \in \mathbb{R}$, then $f : \mathbb{R} \to \mathbb{R}^2$ is a vector valued function.
(b) Similarly, $f(t) \in \mathbb{R}^n$ could be the list of exchange rates at the time moment $t \in \mathbb{R}$.
(c) An explicit example is $f(t) = (t \cos t, t \sin t)$. This function is of the type $f : \mathbb{R} \to \mathbb{R}^2$.

We define continuity for vector functions component-wise.

Definition 2.5. Let $K \subset \mathbb{R}$ be a set, and let $f : K \to \mathbb{R}^n$ be a vector function. We say that f is *continuous at $y \in K$* if each component of f is continuous at y.

Example 2.6. (a) The function $f : \mathbb{R} \to \mathbb{R}^2$ defined by $f(t) = (t \cos t, t \sin t)$ is obviously continuous at each $t \in \mathbb{R}$.
(b) The function $f(t) = (\theta(t), t^2)$ is discontinuous at 0, and continuous everywhere in $\mathbb{R} \setminus \{0\}$.

Let us briefly discuss vector sequences.

- A *vector sequence* is simply a sequence
$$x^{(1)}, x^{(2)}, \ldots, x^{(i)}, \ldots,$$
consisting of vectors $x^{(i)} \in \mathbb{R}^n$, $i \in \mathbb{N}$.
- If we fix $k \in \{1, \ldots, n\}$, then $\{x_k^{(1)}, x_k^{(2)}, \ldots\}$ is a scalar (i.e., real number) sequence, called the *k-th component* of the vector sequence $\{x^{(i)}\}$.

- Hence a vector sequence is simply a *collection of scalar sequences*.

Example 2.7. An example of an $\mathbb{R}^2$-valued sequence is $(j^2, \arctan j)$, $j = 0, 1, 2, \dots$.

Definition 2.8. We say that a vector sequence $\{x^{(i)}\} \subset \mathbb{R}^n$ *converges to* $y \in \mathbb{R}^n$ if for each $k \in \{1, \dots, n\}$, the k-th component of $\{x^{(i)}\}$ converges to the k-th component of y. That is, we write $x^{(i)} \to y$ as $i \to \infty$ if $x_k^{(i)} \to y_k$ as $i \to \infty$ for each $k \in \{1, \dots, n\}$.

Example 2.9. (a) The sequence $(j^2, \arctan j)$ is *not* convergent as $j \to \infty$, because its first component does not convergent.
(b) However, $(\frac{1}{j+1}, \arctan j)$ is convergent as $j \to \infty$, and the limit is $(0, \frac{\pi}{2})$.

3. Continuity of multivariate functions

In this section, we start our study of functions of several variables.

- A (scalar) function of several variables is simply a function $f : K \to \mathbb{R}$, where $K \subset \mathbb{R}^n$.
- For now, we will focus on *scalar valued* functions of *2 variables* (i.e., $n = 2$).
- Examples of such functions are given by

$$f(x, y) = \log(x + y) \qquad \text{with} \qquad K = \{(x, y) \in \mathbb{R}^2 : x + y > 0\},$$

and

$$f(x, y) = -\max\{x, y\} \qquad \text{with} \qquad K = \mathbb{R}^2.$$

The first question is how we define continuity for functions of several variables. For functions of the sort $g : \mathbb{R} \to \mathbb{R}^n$, we defined continuity as continuity of its components $g_k : \mathbb{R} \to \mathbb{R}$. For a function of the sort $f : \mathbb{R}^2 \to \mathbb{R}$, the component-wise approach to continuity would be to require that the single variable functions $g(x) = f(x, y)$ with y fixed, and $h(y) = f(x, y)$ with x fixed, are both continuous.

Definition 3.1. Let $K \subset \mathbb{R}^2$ be a set, and let $f : K \to \mathbb{R}$ be a function. We say that f is *separately continuous at* $(x, y) \in K$, if the single variable functions $g(t) = f(t, y)$ and $h(t) = f(x, t)$ are both continuous.

Example 3.2. (a) The function $f(x, y) = x^2 + \sin y$ is separately continuous at $(0, \frac{\pi}{2}) \in \mathbb{R}^2$, because $g(x) = x^2 + 1$ and $h(y) = \sin y$ are continuous at $x = 0$ and at $y = \frac{\pi}{2}$, respectively.
(b) Consider

$$f(x, y) = \begin{cases} 1 & \text{for } x < y < 3x \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

This function is separately continuous at $(0, 0)$ with $f(x, y) = 0$, because $f(t, 0) = f(0, t) = 0$ for all $t \in \mathbb{R}$. However, there exist points (x, y) that are arbitrarily close to $(0, 0)$ with $f(x, y) = 1$, such as $(x, 2x)$ for $x > 0$.

We see that separate continuity of $f : \mathbb{R}^2 \to \mathbb{R}$, say, at $(0, 0)$, imposes conditions only on the two axes, and hence it is not dependent on the behaviour of f at points such as (x, x) with $x > 0$ arbitrarily small. The following is a stronger definition, which follows an inherently 2-dimensional approach.

Definition 3.3. Let $K \subset \mathbb{R}^2$ be a set, and let $f : K \to \mathbb{R}$ be a function. We say that f is *jointly continuous* or simply *continuous at* $(x, y) \in K$, if $f(x_i, y_i) \to f(x, y)$ as $i \to \infty$ for every sequence $\{(x_i, y_i)\} \subset K$ converging to (x, y).

Example 3.4. In Example 3.2(b), we have $f(\frac{1}{m}, \frac{2}{m}) = 1$ for $m \in \mathbb{N}$, but $f(\frac{1}{m}, 0) = 0$ for $m \in \mathbb{N}$. This shows that f is *not* jointly continuous at the origin.

Exercise 3.5. Show that the function

$$f(x,y) = \begin{cases} 1 & \text{for } x^2 < y < 3x^2 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

is *not* jointly continuous at the origin, but *is* continuous along any line, that is, the function $g(t) = f(\alpha + at, \beta + bt)$ is continuous in $\mathbb{R}$ for any constants $\alpha, \beta, a, b \in \mathbb{R}$.

Example 3.6. (a) Let $c \in \mathbb{R}$, and let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function given by $f(x,y) = c$. Then f is continuous at every point $(x,y) \in \mathbb{R}^2$, since for any sequence $\{(x_i, y_i)\}$, we have $f(x_i, y_i) = c \to c = f(x,y)$ as $i \to \infty$.

(b) Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function given by $f(x,y) = x$. Then f is continuous at every point $(x,y) \in \mathbb{R}^2$, because given any sequence $\{(x_i, y_i)\} \subset \mathbb{R}^2$ converging to (x,y), we have $f(x_i, y_i) = x_i \to x = f(x,y)$ as $i \to \infty$.

(c) Similarly, $g : \mathbb{R}^2 \to \mathbb{R}$ given by $g(x,y) = y$ is continuous everywhere in $\mathbb{R}^2$.

Remark 3.7. This remark contains the main practical message of this section.

- In order to show that f is *discontinuous* at (x,y), it suffices to exhibit a sequence $\{(x_n, y_n)\}$ with $(x_n, y_n) \to (x,y)$, such that $f(x_n, y_n) \not\to f(x,y)$ as $n \to \infty$.
- If a function is given by a formula, verifying its continuity is usually not hard, because continuous building blocks produce continuous functions. Following the pattern of the single variable theory, in what follows we shall justify the latter statement.

Lemma 3.8. *Let $K \subset \mathbb{R}^2$, and let $f, g : K \to \mathbb{R}$ be functions continuous at $x \in K$. Then the sum and difference $f \pm g$, and the product fg are all continuous at x. Moreover, the function $\frac{1}{f}$ is continuous at x, provided that $f(x) \neq 0$.*

Example 3.9. (a) Recall from Example 3.6 that the constant function $f(x,y) = c$ (where $c \in \mathbb{R}$), and the projection maps $g(x,y) = x$ and $h(x,y) = y$ are continuous in $\mathbb{R}^2$. Then by Lemma 3.8, any monomial $f(x,y) = ax^n y^m$ with constants $a \in \mathbb{R}$ and $n, m \in \mathbb{N}_0^2$, is continuous in $\mathbb{R}^2$. A bivariate polynomial is a function $p : \mathbb{R}^2 \to \mathbb{R}$ of the form

$$p(x) = \sum_{i,k} a_{ik} x^i y^k, \tag{9}$$

where only finitely many of the coefficients $a_{ik} \in \mathbb{R}$ are nonzero. Applying Lemma 3.8 again, we conclude that all bivariate polynomials are continuous in $\mathbb{R}^2$.

(b) Let p and q be polynomials, and let $Z = \{(x,y) \in \mathbb{R}^2 : q(x,y) = 0\}$ be the set of zeroes of q. Then by Lemma 3.8, the function $r : \mathbb{R}^2 \setminus Z \to \mathbb{R}$ given by $r(x,y) = \frac{p(x,y)}{q(x,y)}$ is continuous in $\mathbb{R}^2 \setminus Z$. The functions of this form are called *rational functions*. For instance, $f(x,y) = \frac{x^2+1}{(x-1)^2+y^2}$ is continuous at each $(x,y) \in \mathbb{R}^2 \setminus \{(1,0)\}$.

Lemma 3.10. *Let $K \subset \mathbb{R}^2$, and let $g : K \to \mathbb{R}$ be a function continuous at $(x,y) \in K$. Suppose that $U \subset \mathbb{R}$ satisfies $g(K) \subset U$, the latter meaning that $(x,y) \in K$ implies $g(x,y) \in U$. Let $F : U \to \mathbb{R}$ be a function continuous at $g(x,y)$. Then the composition $F \circ g : K \to \mathbb{R}$, defined by $(F \circ g)(x,y) = F(g(x,y))$, is continuous at x.*

Example 3.11. (a) The function $f(x,y) = \cos(2x + y) - \sin x$ is continuous in $\mathbb{R}^2$.

(b) Let $K \subset \mathbb{R}^2$ and let $f : K \to \mathbb{R}$ be continuous at $(x,y) \in K$. Then

$$g(s,t) = |f(s,t)| \qquad \text{for} \quad (s,t) \in K, \tag{10}$$

is continuous at $(x.y)$.

We now consider *vector functions* of several variables. All that has been said extends to this situation in a straightforward, "componentwise" way.

Definition 3.12. Let $K \subset \mathbb{R}^n$ be a set, and let $f : K \to \mathbb{R}^m$ be a function. We say that f is *continuous at* $y \in K$, if each component of f is continuous at y. If $f : K \to \mathbb{R}^m$ is continuous at each point of K, we say that f is *continuous in* K.

Example 3.13. (a) The function $f : \mathbb{R}^2 \to \mathbb{R}^2$ defined by $f(x, y) = (x \cos y, y \sin x)$ is obviously continuous in $\mathbb{R}^2$.
(b) The function $f(x, y) = (ax + by, cx + dy)$, with a, b, c, d constants, is continuous in $\mathbb{R}^2$.

4. Differentiability of univariable scalar functions

Let us recall the usual definition of differentiability. This is essentially the definition introduced by Augustin-Louis Cauchy in 1821.

Definition 4.1. Let $K \subset \mathbb{R}$ be a set, and let $f : K \to \mathbb{R}$ be a function. We say that f is *differentiable at* $y \in K$, if there exists $\lambda \in \mathbb{R}$ such that

$$\frac{f(x) - f(y)}{x - y} \to \lambda \qquad \text{as} \quad x \to y. \tag{11}$$

We call $f'(y) = \lambda$ the *derivative of* f *at* y. If f is differentiable at each point of K, then f is said to be *differentiable in* K.

Remark 4.2. The following notations are also used:

$$\frac{df}{dx}(x) = \dot{f}(x) = f'(x). \tag{12}$$

Example 4.3. (a) Let us try to differentiate $f(x) = x^2$ at $y = 1$. Taking into account

$$\frac{f(x) - f(y)}{x - y} = \frac{x^2 - 1}{x - 1} = \frac{(x - 1)(x + 1)}{x - 1} = x + 1, \tag{13}$$

we compute

$$\lim_{x \to 1} \frac{f(x) - f(1)}{x - 1} = \lim_{x \to 1} (x + 1) = 2, \tag{14}$$

which means that f is differentiable at 1, with $f'(1) = 2$. Note that (13) can be rewritten as

$$f(x) = f(1) + (x + 1)(x - 1) = f(1) + g(x)(x - 1), \tag{15}$$

with $g(x) = x + 1$, and the derivative $f'(1)$ is simply the value $g(1)$. This is generalized to Carathéodory's criterion (c) in the following lemma.
(b) Let us try to differentiate $f(x) = |x|$ at $x = 0$. With $x_n = \frac{1}{n}$ for $n \in \mathbb{N}$, we have $\{x_n\} \subset \mathbb{R} \setminus \{0\}$ and $x_n \to 0$ as $n \to \infty$. On one hand, we get

$$\lim_{n \to \infty} \frac{f(x_n) - f(0)}{x_n - 0} = \lim_{n \to \infty} \frac{|x_n|}{x_n} = 1, \tag{16}$$

but on the other hand, with $y_n = -x_n$, we infer

$$\lim_{n \to \infty} \frac{f(y_n) - f(0)}{y_n - 0} = \lim_{n \to \infty} \frac{|y_n|}{y_n} = -\lim_{n \to \infty} \frac{x_n}{x_n} = -1. \tag{17}$$

The definition of derivative requires these two limits to be the same, and thus we conclude that $f(x) = |x|$ is *not* differentiable at $x = 0$.
(c) Consider the differentiability of $f(x) = \sqrt[3]{x}$ at $x = 0$. Let $x_n = \frac{1}{n^3}$. It is obvious that $x_n \neq 0$ and $x_n \to 0$. We have

$$\frac{f(x_n) - f(0)}{x_n - 0} = \frac{\sqrt[3]{x_n}}{x_n} = n^2, \tag{18}$$

which diverges as $n \to \infty$. Hence $f(x) = \sqrt[3]{x}$ is *not* differentiable at $x = 0$.

We now state a couple of useful criteria of differentiability. In the following lemma, (b) is called the *sequential criterion*, and (c) is introduced by Constantin Carathéodory in 1950.

Lemma 4.4. *Let $K \subset \mathbb{R}$, let $y \in K$, and let $f : K \to \mathbb{R}$ be a function. Then the following are equivalent.*

(a) f is differentiable at y.
(b) There exists a number $\lambda \in \mathbb{R}$, such that

$$\frac{f(x_n) - f(y)}{x_n - y} \to \lambda \qquad as \quad n \to \infty, \tag{19}$$

for every sequence $\{x_n\} \subset K \setminus \{y\}$ converging to y.
(c) There exists a function $g : K \to \mathbb{R}$, continuous at y, such that

$$f(x) = f(y) + g(x)(x - y) \qquad for \quad x \in K. \tag{20}$$

In (b), the derivative is given by $f'(y) = \lambda$, and in (c), it is given by $f'(y) = g(y)$.

Example 4.5. (a) Let $c \in \mathbb{R}$, and let $f(x) = c$ be a constant function. Then by Caratheodory's criterion, since $f(x) = f(y) + 0 \cdot (x - y)$ for all x, y, we get $f'(y) = 0$ for all y.
(b) Let $a, c \in \mathbb{R}$, and let $f(x) = ax + c$ be a linear (also known as affine) function. Since $f(x) = f(y) + a(x - y)$ for all x, y, by Caratheodory's criterion, we get $f'(y) = a$ for all y.
(c) Let $f(x) = x^3$, and let us try to differentiate it at $y \in \mathbb{R}$. Since

$$f(x) - f(y) = x^3 - y^3 = (x^2 + xy + y^2)(x - y), \tag{21}$$

we identify $g(y) = x^2 + xy + y^2$ in Caratheodory's criterion, to conclude that f is differentiable at y, with $f'(y) = g(y) = 3y^2$ for all y.
(d) Let $f(x) = \frac{1}{x}$, fix $y \in \mathbb{R} \setminus \{0\}$, and for $x \in \mathbb{R} \setminus \{0, y\}$ define

$$g(x) = \frac{\frac{1}{x} - \frac{1}{y}}{x - y} = -\frac{1}{xy}. \tag{22}$$

Upon defining $g(y) = -\frac{1}{y^2}$, the function $g(x) = -\frac{1}{x} \cdot \frac{1}{y}$ becomes continuous at $x = y$, and therefore f is differentiable at y with

$$f'(y) = \big(\frac{1}{y}\big)' = -\frac{1}{y^2} \qquad (y \neq 0). \tag{23}$$

Remark 4.6. Differentiability of f at y is equivalent to the condition that $f(x)$ can be approximated by the linear function $\ell(x) = f(y) + \lambda(x - y)$ with the error going to 0 faster than $|x - y|$. Of course, this linear function is the tangent line to the graph of f through the point $(x, f(x))$. Recall that continuity of f at y is equivalent to saying that $f(x)$ can be approximated by the constant $f(y)$ with the error going to 0 as $x \to y$.

Example 4.7. (a) The linear approximation to $f(x) = x^3$ at $x = 1$ is given by

$$\ell(x) = f(1) + f'(1)(x - 1) = 1 + 3(x - 1) = 3x - 2. \tag{24}$$

So for instance, we can approximate

$$1.1^3 = f(1.1) \approx \ell(1.1) = 3 \cdot 1.1 - 2 = 1.3. \tag{25}$$

Compare this with the true value $1.1^3 = 1.331$.
(b) The tangent line to $f(x) = \frac{1}{x}$ at $x = y$ is given by

$$\ell(x) = f(y) + f'(y)(x - y) = \frac{1}{y} - \frac{x - y}{y^2} = \frac{2}{y} - \frac{x}{y^2}. \tag{26}$$

Remark 4.8. This remark contains the main practical message of this section.

- In order to show that f is *nondifferentiable* at x, it suffices to exhibit a sequence $\{x_n\}$ with $x_n \to x$, such that the quotient $\frac{f(x_n)-f(x)}{x_n-x}$ does *not* have a limit as $n \to \infty$.
- If a function is given by a formula, verifying its differentiability *and* computing its derivative is usually not hard, because differentiable building blocks produce differentiable functions. In the rest of this section, we shall review differentiability of various combinations of differentiable functions.

Theorem 4.9. *Let $f, g : (a,b) \to \mathbb{R}$ be functions differentiable at $x \in (a,b)$. Then the following are true.*

a) The sum and difference $f \pm g$ are differentiable at x, with

$$(f \pm g)'(x) = f'(x) \pm g'(x). \tag{27}$$

These are called the sum and difference rules.

b) The product fg is differentiable at x, with

$$(fg)'(x) = f'(x)g(x) + f(x)g'(x). \tag{28}$$

This is called the product rule.

c) If $F : (c,d) \to \mathbb{R}$ is a function differentiable at $g(x)$, with $g((a,b)) \subset (c,d)$, then the composition $F \circ g : (a,b) \to \mathbb{R}$ is differentiable at x, with

$$(F \circ g)'(x) = F'(g(x))g'(x). \tag{29}$$

This is called the chain rule.

d) If $f : (a,b) \to f((a,b))$ is bijective and $f'(x) \neq 0$, then the inverse $f^{-1} : f((a,b)) \to (a,b)$ is differentiable at $y = f(x)$, with

$$(f^{-1})'(y) = \frac{1}{f'(x)}. \tag{30}$$

Example 4.10. (a) By the product rule, we have

$$\begin{aligned} &(x^2)' = 1 \cdot x + x \cdot 1 = 2x, \\ &(x^3)' = (x^2 \cdot x)' = 2x \cdot x + x^2 \cdot 1 = 3x^2, \dots \\ &(x^n)' = nx^{n-1} \qquad (n \in \mathbb{N}). \end{aligned} \tag{31}$$

(b) By the sum and product rules, all polynomials are differentiable in $\mathbb{R}$, and the derivative of a polynomial is again a polynomial.

(c) Given a function $f : (a,b) \to \mathbb{R}$ that does not vanish anywhere in (a,b), we can write the reciprocal function $\frac{1}{f}$ as $F \circ f$ with $F(z) = \frac{1}{z}$. If f is differentiable at $x \in (a,b)$, then by the chain rule, $\frac{1}{f}$ is differentiable at x and

$$\big(\frac{1}{f}\big)'(x) = (F \circ f)'(x) = F'(f(x))f'(x) = -\frac{f'(x)}{[f(x)]^2}. \tag{32}$$

In particular, we have

$$(x^{-n})' = -\frac{nx^{n-1}}{x^{2n}} = -nx^{-n-1} \qquad (n \in \mathbb{N}). \tag{33}$$

(d) Let $f(x) = x^n$ for $x \in [0,\infty)$, where $n \in \mathbb{N}$. We have $f'(x) = nx^{n-1}$ at $x > 0$, and the inverse function is the arithmetic n-th root $f^{-1}(y) = \sqrt[n]{y}$ $(y \geq 0)$. Since $f'(x) > 0$ for $x > 0$, the inverse f^{-1} is differentiable at each $y > 0$, with

$$(f^{-1})'(y) = \frac{1}{f'(f^{-1}(y))} = \frac{1}{n(\sqrt[n]{y})^{n-1}} = \frac{1}{n} y^{\frac{1-n}{n}}. \tag{34}$$

Moreover, by the chain rule, for $m \in \mathbb{Z}$ and $n \in \mathbb{N}$, we infer

$$(x^{\frac{m}{n}})' = ((\sqrt[n]{x})^m)' = m(\sqrt[n]{x})^{m-1} \cdot \frac{1}{n} x^{\frac{1-n}{n}} = \frac{m}{n} x^{\frac{m-1}{n}+\frac{1-n}{n}} = \frac{m}{n} x^{\frac{m}{n}-1}, \tag{35}$$

that is

$$(x^a)' = ax^{a-1} \qquad \text{at each } x > 0, \quad \text{for } a \in \mathbb{Q}. \tag{36}$$

Exercise 4.11. Let $f, g : (a, b) \to \mathbb{R}$ be functions differentiable at $x \in (a, b)$, with $g(x) \neq 0$. Show that the quotient f/g is differentiable at x, and the following *quotient rule* holds.

$$\left(\frac{f}{g}\right)'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{[g(x)]^2}. \tag{37}$$

Compute the derivative of $q(x) = \frac{3x^3}{x^2+1}$.

5. Differentiability of univariate vector functions

Similarly to continuity, differentiability of vector functions is defined component-wise.

- A function $f : \mathbb{R} \to \mathbb{R}^n$ can be interpreted as a *parametrized curve* in $\mathbb{R}^n$.
- Then its derivative $f'(t)$ is the *tangent vector* to the curve at the point $f(t)$.

Definition 5.1. Let $K \subset \mathbb{R}$ be a set, and let $f : K \to \mathbb{R}^n$ be a vector function. We say that f is *differentiable at* $y \in K$, if each component of f is differentiable at y. We call $f'(y) = (f_1'(y), \ldots, f_n'(y)) \in \mathbb{R}^n$ the *derivative of* f *at* y. If f is differentiable at each point of K, then f is said to be *differentiable in* K.

Example 5.2. (a) Let $f(t) = (t^2, \sin t)$. Then $f'(t) = (2t, \cos t)$.
(b) Let $f(t) = (t \sin t, t \cos t)$. Then $f'(t) = (\sin t + t \cos t, \cos t - t \sin t)$.
(c) Functions $\ell : \mathbb{R} \to \mathbb{R}^n$ of the form

$$\ell(t) = \alpha + t\beta, \tag{38}$$

where $\alpha, \beta \in \mathbb{R}^n$ are fixed vectors, are called *linear (or affine) functions*. For these functions, we have $\ell'(t) = \beta$.

Lemma 5.3. *Let $K \subset \mathbb{R}$, let $y \in K$, and let $f : K \to \mathbb{R}^n$ be a vector function. Then the following are equivalent.*

(a) f is differentiable at y.
(b) There exists a function $g : K \to \mathbb{R}^n$, continuous at y, such that

$$f(x) = f(y) + g(x)(x - y) \qquad \text{for} \quad x \in K. \tag{39}$$

(c) There exists a vector $\lambda \in \mathbb{R}^n$, such that

$$\frac{f(x_i) - f(y)}{x_i - y} \to \lambda \qquad \text{as} \quad i \to \infty, \tag{40}$$

for every sequence $\{x_i\} \subset K \setminus \{y\}$ converging to y.

Exercise 5.4. Let $f : (a, b) \to \mathbb{R}^n$ and $\phi : (a, b) \to \mathbb{R}$ be both differentiable at $y \in (a, b)$. Show that the product $\phi f : (a, b) \to \mathbb{R}^n$ is differentiable at y, with

$$(\phi f)'(y) = \phi'(y) f(y) + \phi(y) f'(y).$$

Exercise 5.5. Let $f : (a, b) \to \mathbb{R}^n$ and $\phi : (c, d) \to (a, b)$, where ϕ is differentiable at $t \in (c, d)$, and f is differentiable at $\phi(t) \in (a, b)$. Show that the composition $f \circ \phi : (c, d) \to \mathbb{R}^n$ is differentiable at t, with

$$(f \circ \phi)'(y) = f'(\phi(t))\phi'(t).$$

6. Partial and directional derivatives

Separate continuity of a function $f : \mathbb{R}^2 \to \mathbb{R}$ at $P = (x, y) \in \mathbb{R}^2$ is defined in terms of continuity of the single variable functions $g(t) = f(x + t, y)$ and $h(t) = f(x, y + t)$, at $t = 0$.

- Observe that g is simply f restricted to the line $\gamma_1(t) = P + (t, 0)$, $t \in \mathbb{R}$, and similarly that h is simply f restricted to the line $\gamma_2(t) = P + (0, t)$, $t \in \mathbb{R}$.
- Apart from continuity, we can talk about differentiability of g and h, leading to the notion of partial derivatives:
$$\frac{\partial f}{\partial x}(x, y) = g'(0), \qquad \frac{\partial f}{\partial y}(x, y) = h'(0).$$
- More generally, given an arbitrary vector $V = (a, b) \in \mathbb{R}^2$, the restriction
$$g(t) = f(P + Vt) = f(x + at, y + bt)$$
to the line $\gamma(t) = P + Vt$ can be considered. This leads us to the notion of *directional derivative* along the direction V:
$$D_V f(x, y) = g'(0).$$
- Similarly to the situation with continuity, partial and directional derivatives turn out to be *not* the correct generalization of the derivative to higher dimensions, but will be a very useful auxiliary tool to get a handle on the ultimate generalization.

Definition 6.1. Let $K \subset \mathbb{R}^2$. The *directional derivative of* $f : K \to \mathbb{R}$ *at* $P = (x, y) \in K$ *along* $V \in \mathbb{R}^2$, is defined to be $D_V f(x) = g'(0)$ if the latter exists, where
$$g(t) = f(P + Vt), \tag{41}$$
is a function of $t \in \mathbb{R}$. The *partial derivatives of* f *at* P are
$$\frac{\partial f}{\partial x}(x, y) = D_{e_1} f(x, y), \qquad \frac{\partial f}{\partial y}(x, y) = D_{e_2} f(x, y), \tag{42}$$
provided that it exists, where $e_1 = (1, 0)$ and $e_2 = (0, 1)$. The row-vector (i.e., 1×2 matrix) consisting of the partial derivatives
$$J_f(x, y) = \begin{pmatrix} \frac{\partial f}{\partial x}(x, y) & \frac{\partial f}{\partial y}(x, y) \end{pmatrix} \tag{43}$$
is called the *Jacobian matrix* of f at x.

Remark 6.2. (a) For partial derivatives, the following notations are also used:
$$f_x(x, y) = f'_x(x, y) = \partial_x f(x, y) = \partial_1 f(x, y) = \frac{\partial f}{\partial x}(x, y). \tag{44}$$
The way to understand the notation $\partial_1 f$ is that we are taking the partial derivative with respect to the *first variable* of f.

(b) For directional derivatives, we have the following (rarely used) alternative notations:
$$\partial_V f(x, y) = \frac{\partial f}{\partial V}(x, y) = D_V f(x, y). \tag{45}$$

Example 6.3. (a) Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by
$$f(x, y) = \begin{cases} 1 & \text{for } x^2 < y < 3x^2 \\ 0 & \text{otherwise.} \end{cases} \tag{46}$$

Given any $(a,b) \in \mathbb{R}^2$, we have $f(at,bt) = 0$ for all $t > 0$ sufficiently small. Hence the directional derivative $D_V f(0,0)$ exists and is equal to 0 for all $V \in \mathbb{R}^2$. In particular, the partial derivatives are

$$\frac{\partial f}{\partial x}(0,0) = D_{(1,0)} f(0,0) = 0, \qquad \frac{\partial f}{\partial y}(0,0) = D_{(0,1)} f(0,0) = 0, \tag{47}$$

and thus the Jacobian matrix of f at the origin is given by $J = (0\ 0) \in \mathbb{R}^{1\times 2}$. However, f is not even continuous at the origin.

(b) Similarly, let

$$f(x,y) = \begin{cases} \frac{x^2 y}{x^4+y^2} & \text{for } |x| + |y| > 0 \\ 0 & \text{for } x = y = 0. \end{cases} \tag{48}$$

For $(a,b) \in \mathbb{R}^2 \setminus \{(0,0)\}$ and $t \neq 0$, we have

$$g(t) = f(at,bt) = \frac{a^2 bt}{a^4 t^2 + b^2} = f(0,0) + \frac{a^2 b}{a^4 t^2 + b^2} \cdot t, \tag{49}$$

which implies that $g'(0)$ exists, with $g'(0) = a^2/b$ for $b \neq 0$ and $g'(0) = 0$ for $b = 0$. Hence, the directional derivative $D_{(a,b)} f(0,0)$ exists, with its value equal to a^2/b for $b \neq 0$ and 0 for $b = 0$. Note that the value a^2/b diverges as $(a,b) \to (1,0)$, even though $D_{(1,0)} f(0,0) = 0$, meaning that the dependence of $D_{(a,b)} f(0,0)$ on (a,b) is *not* continuous. The Jacobian matrix of f at the origin is given by $J = (0\ 0) \in \mathbb{R}^{1\times 2}$.

(c) It is easy to see that the function $f(x,y) = \sqrt{|xy|}$ is differentiable at $(0,0)$ along V if and only if $V = (a,0)$ or $V = (0,a)$ for some $a \in \mathbb{R}$.

Remark 6.4. There is no obvious *a priori* structure on how $D_V f$ depends on V, except to say that $D_V f(x)$ is homogeneous in V, that is, $D_{\alpha V} f(x) = \alpha D_V f(x)$ for $\alpha \in \mathbb{R}$.

Definition 6.5. A vector valued function $f : \mathbb{R}^2 \to \mathbb{R}^m$ is a function with m components:

$$f(x,y) = \big(f_1(x,y), f_2(x,y), \ldots, f_m(x,y)\big), \tag{50}$$

where $f_1, f_2, \ldots$ are scalar functions. We usually think of these components as arranged in a column, and define the *Jacobian of* f to be the $m \times 2$ matrix

$$J_f(x,y) = \begin{pmatrix} \partial_x f_1(x,y) & \partial_y f_1(x,y) \\ \partial_x f_2(x,y) & \partial_y f_2(x,y) \\ \cdots & \\ \partial_x f_m(x,y) & \partial_y f_m(x,y) \end{pmatrix}. \tag{51}$$

Example 6.6. The Jacobian matrix of $f(x,y) = \big(xy, \sin(x+y^2)\big)$ is

$$J_f(x,y) = \begin{pmatrix} y & x \\ \cos(x+y^2) & 2y\cos(x+y^2) \end{pmatrix}. \tag{52}$$

Definition 6.7. The *Jacobian of* $f : \mathbb{R}^n \to \mathbb{R}^m$ is defined to be the $m \times n$ matrix

$$J_f(x) = \begin{pmatrix} \partial_1 f_1(x) & \partial_2 f_1(x) & \ldots & \partial_n f_1(x) \\ \partial_1 f_2(x) & \partial_2 f_2(x) & \ldots & \partial_n f_2(x) \\ \cdots & & & \\ \partial_1 f_m(x) & \partial_2 f_m(x) & \ldots & \partial_n f_m(x) \end{pmatrix}. \tag{53}$$

Example 6.8. The Jacobian matrix of $f(x,y) = \big(xyz, \sin(x+z^2)\big)$ is

$$J_f(x,y,z) = \begin{pmatrix} yz & xz & xy \\ \cos(x+z^2) & 0 & 2z\cos(x+z^2) \end{pmatrix}. \tag{54}$$

7. Gradient in two dimensions

Loosely speaking, the way we defined partial derivatives resembles that of separate continuity. Now we want to introduce a notion of derivative that mirrors joint continuity. To motivate it, recall that $f : \mathbb{R} \to \mathbb{R}$ is differentiable at x if and only if

$$f(s) = f(x) + \lambda(s - x) + e(s), \tag{55}$$

with $e(s)$ tending to 0 faster than $|s - x|$, that is,

$$\frac{e(s)}{s - x} \to 0 \qquad \text{as} \quad s \to x, \tag{56}$$

for some (fixed) number $\lambda \in \mathbb{R}$. In a certain sense, differentiable functions are well approximated locally by linear functions. An equivalent way to characterize this is to say that

$$f(s) = f(x) + g(s)(s - x), \tag{57}$$

with $g(s)$ continuous at $s = x$. The following natural extension of this criterion to functions of several variables was first studied by Karl Weierstrass (1861), Otto Stolz (1893), William H. Young (1910), and Maurice Fréchet (1911).

Definition 7.1. Let $K \subset \mathbb{R}^2$. A function $f : K \to \mathbb{R}$ is called *differentiable* at $(x, y) \in K$ if

$$f(s, t) = f(x, y) + g(s, t)(s - x) + h(s, t)(t - y), \tag{58}$$

for some functions g and h, both continuous at $(s, t) = (x, y)$. We call the row-vector

$$Df(x, y) = \begin{pmatrix} g(x, y) & h(x, y) \end{pmatrix} \tag{59}$$

if it exists, the *derivative* of f at (x, y).

Remark 7.2. For scalar valued functions, the derivative is also called the *gradient*, and the following alternative notations are often used:

$$\operatorname{grad} f(x, y) = \nabla f(x, y) = Df(x, y). \tag{60}$$

Note in particular that if f is differentiable at $(x, y) \in K$ then f is continuous at (x, y). In contrast, recall from Example 6.3 that directional differentiability does not imply continuity.

Example 7.3. (a) Consider $f(x, y) = x^2 + y^3$. We can write

$$f(s, t) - f(x, y) = s^2 - x^2 + t^3 - y^3 = (s - x)(s + x) + (t - y)(t^2 + ty + y^2), \tag{61}$$

and identify $g(s, t) = s + x$ and $h(s, t) = t^2 + ty + y^2$. Note that x and y should be treated as constants, since they are the coordinates of the base point of differentiation. As g and h are obviously both (jointly) continuous at $(s, t) = (x, y)$, we conclude that f differentiable at (x, y), with

$$\nabla f(x, y) = \begin{pmatrix} g(x, y) & h(x, y) \end{pmatrix} = \begin{pmatrix} 2x & 3y^2 \end{pmatrix}. \tag{62}$$

(b) Consider $f(x, y) = xy$, and write

$$f(s, t) - f(x, y) = st - xy = st - xt + xt - xy = (s - x)t + x(t - y). \tag{63}$$

We identify $g(s, t) = t$ and $h(s, t) = x$, which are obviously continuous at $(s, t) = (x, y)$. Hence f differentiable at (x, y), with

$$\nabla f(x, y) = \begin{pmatrix} g(x, y) & h(x, y) \end{pmatrix} = \begin{pmatrix} y & x \end{pmatrix}. \tag{64}$$

Remark 7.4. If f is differentiable at (x, y), then the *linear approximation* at (x, y) to f (or the *tangent plane* to the graph of f) is given by

$$\ell(s, t) = f(x, y) + \nabla f(x, y) \begin{pmatrix} s - x \\ t - y \end{pmatrix}. \tag{65}$$

Note that here x and y are fixed, and s and t are the free parameters of the plane.

Example 7.5. (a) The linear approximation to $f(x,y) = x^2 + y^3$ at $(x,y) = (1,1)$ is

$$\begin{aligned} \ell(s,t) &= f(1,1) + \nabla f(1,1) \begin{pmatrix} s-1 \\ t-1 \end{pmatrix} = 2 + \begin{pmatrix} 2 & 3 \end{pmatrix} \begin{pmatrix} s-1 \\ t-1 \end{pmatrix} \\ &= 2 + 2(s-1) + 3(t-1) = 2s + 3t - 3. \end{aligned} \tag{66}$$

For instance, we can approximate

$$1.1^2 + 0.9^3 = f(1.1, 0.9) \approx \ell(1.1, 0.9) = 1.92 \cdot 1.1 + 3 \cdot 0.9 - 3 = 1.9. \tag{67}$$

Compare this with the true value $f(1.1, 0.9) = 1.939$.

(b) The tangent plane to the graph of $f(x,y) = xy$ at (x,y) is

$$\ell(s,t) = xy + \begin{pmatrix} y & x \end{pmatrix} \begin{pmatrix} s-x \\ t-y \end{pmatrix} = xy + y(s-x) + x(t-y) = ys + xt - xy. \tag{68}$$

Remark 7.6 (Differentiability implies directional derivatives)**.** Suppose that $f : K \to \mathbb{R}$ is differentiable at $(x,y) \in K$. Then for $V = (a,b)$ fixed and $t \in \mathbb{R}$ with $t \to 0$, we have

$$f(x + at, y + bt) = f(x,y) + g(x + at, y + bt)at + h(x + at, y + bt)bt. \tag{69}$$

This leads to

$$\frac{f(x + at, y + bt) - f(x,y)}{t} \to g(x,y)a + h(x,y)b \qquad \text{as} \quad t \to 0, \tag{70}$$

i.e., the directional derivative $D_V f(x,y)$ exists, with

$$D_V f(x,y) = (\nabla f(x,y))V = V \cdot \nabla f(x,y). \tag{71}$$

Thus differentiability of f implies not only directional differentiability, but also a linear dependence of the derivative $D_V f(x,y)$ on the direction V. In particular, by taking $V = e_1$ and $V = e_2$, we see that the gradient is in fact equal to the Jacobian matrix $J_f(x,y)$ of f:

$$\nabla f(x,y) = J_f(x,y). \tag{72}$$

In view of the preceding remark, if we somehow know that f is differentiable, we can compute the derivative as the Jacobian matrix. However, how do we ascertain differentiability of f in the first place? The following result gives a practical way to handle this situation.

Theorem 7.7. *Let $Q = (a,b) \times (c,d)$ be a rectangular domain, and let $f : Q \to \mathbb{R}$. Suppose that all partial derivatives of f exist at each $(x,y) \in Q$, and that the partial derivatives are continuous in Q. Then f is differentiable in Q.*

Example 7.8. (a) Let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by $f(x,y) = x \cos y$. Its Jacobian matrix can be computed as

$$J_f(x,y) = \begin{pmatrix} \partial_x f & \partial_y f \end{pmatrix} = \begin{pmatrix} \cos y & -x \sin y \end{pmatrix}. \tag{73}$$

Since $J_f : \mathbb{R}^2 \to \mathbb{R}^2$ is continuous in $\mathbb{R}^2$, we conclude that f is differentiable in $\mathbb{R}^2$ with $Df(x,y) = J_f(x,y)$.

(b) Let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by $f(x,y) = y \sin x$. Its Jacobian matrix can be computed as

$$J_f(x,y) = \begin{pmatrix} \partial_x f & \partial_y f \end{pmatrix} = \begin{pmatrix} y \cos x & \sin x \end{pmatrix}. \tag{74}$$

Since $J_f : \mathbb{R}^2 \to \mathbb{R}^2$ is continuous in $\mathbb{R}^2$, we conclude that f is differentiable in $\mathbb{R}^2$ with $Df(x,y) = J_f(x,y)$.

Remark 7.9 (Gradient gives direction of fastest growth)**.** Suppose that f is differentiable at $P = (x,y)$ with $\nabla f(P) = \begin{pmatrix} u & v \end{pmatrix}$, and consider the problem of maximizing $D_V f(P)$ over all vectors $V = (a,b)$ with unit length. That is, given a gradient vector $(u,v) \in \mathbb{R}^2$, we want to find $V = (a,b)$ with $|V| = \sqrt{a^2 + b^2} = 1$, such that

$$D_V f(P) = ua + vb \tag{75}$$

takes its maximum value. We claim that V should be parallel to (u,v), that is, $V=(ku,kv)$ with $k=\frac{1}{\sqrt{u^2+v^2}}$. The value of the directional derivative for this particular direction is

$$D_{(ku,kv)}f(P)=\sqrt{u^2+v^2}=|\nabla f(x,y)|, \tag{76}$$

and so what we need to show is

$$ua+vb\le\sqrt{u^2+v^2}, \tag{77}$$

for any (a,b) satisfying $a^2+b^2=1$. To see this, let

$$g(t)=(u+at)^2+(v+bt)^2=u^2+v^2+2(ua+vb)t+(a^2+b^2)t^2. \tag{78}$$

Since $g(t)\ge 0$ for all t, as a quadratic polynomial of t, its discriminant must be *nonpositive*:

$$D=4(ua+vb)^2-4(u^2+v^2)(a^2+b^2)\le 0. \tag{79}$$

Thus we have

$$ua+vb\le\sqrt{u^2+v^2}\sqrt{a^2+b^2}, \tag{80}$$

which is (77) as $a^2+b^2=1$. The argument can be generalized to n dimensions easily, and the inequality (80) is called the *Cauchy-Bunyakowsky-Schwarz inequality*.

Remark 7.10 (Gradient is orthogonal to level surfaces)**.** Let f be differentiable at P, and suppose that V is a vector tangent to the level curve of f at P. Since f does not vary along its level curves, we have $D_Vf(P)=0$. On the other hand, we know that $D_Vf(P)=V\cdot\nabla f(P)$, and hence the gradient $\nabla f(P)$ is orthogonal to the level curve: $V\cdot\nabla f(P)=0$.

Example 7.11. From Example 7.5(a), we know that the tangent plane to the graph of $f(x,y)=x^2+y^3$ at $(x,y)=(1,1)$ is

$$\ell(s,t)=2s+3t-3. \tag{81}$$

This means that the tangent line (lying in the plane $\mathbb{R}^2$) to level curve of f going through $(1,1)$ is given by the equation

$$2s+3t-3=f(1,1)=2. \tag{82}$$

As the gradient of ℓ is $\nabla\ell=(2,3)$, the line normal to the level curve at $(1,1)$ is given by

$$x(t)=1+2t,\qquad y(t)=1+3t. \tag{83}$$

8. Differentiability of multivariate functions

Once we understand the gradient in two dimensions, extensions to more general situations are straightforward. The derivative of $f:\mathbb{R}^2\to\mathbb{R}^m$ is defined component-wise, and if exists, it coincides with the Jacobian matrix of f, cf. Definition 6.5 and Remark 7.6. In the other direction, if the Jacobian matrix of f exists and continuous in an open region, then f is differentiable, cf. Theorem 7.7.

Example 8.1. The Jacobian matrix of $f(x,y)=\big(xy,\sin(x+y^2)\big)$ is

$$J_f(x,y)=\begin{pmatrix} y & x \\ \cos(x+y^2) & 2y\cos(x+y^2)\end{pmatrix}. \tag{84}$$

Since each component of J_f is continuous in $\mathbb{R}^2$, we conclude that f is differentiable in $\mathbb{R}^2$ with $Df(x,y)=J_f(x,y)$.

Differentiability of $f:\mathbb{R}^3\to\mathbb{R}$ at $x=(x_1,x_2,x_3)$ is defined by the condition that

$$f(t_1,t_2,t_3)=f(x_1,x_2,x_3)+g_1(t_1,t_2,t_3)(t_1-x_1)\\ +g_2(t_1,t_2,t_3)(t_2-x_2)+g_3(t_1,t_2,t_3)(t_3-x_3), \quad (85)$$

for some functions g_1, g_2, and g_3, all continuous at $t=x$. Differentiability of functions of the type $f:\mathbb{R}^n\to\mathbb{R}$ is defined analogously. For the n-dimensional gradient

$$\nabla f(x)=\begin{pmatrix}\partial_1 f & \partial_2 f & \dots & \partial_n f\end{pmatrix}, \quad (86)$$

Remark 7.4, Remark 7.9, and Remark 7.10 are true, with obvious modifications.

Example 8.2. The gradient of $f(x,y,z)=xy+\sin z$ is

$$\nabla f(x)=\begin{pmatrix}y & x & \cos z\end{pmatrix}, \quad (87)$$

and so the linear approximation to f at $(x,y,z)=(1,1,0)$ is

$$\begin{aligned}\ell(s,t,u)&=f(1,1,0)+\nabla f(1,1,0)\begin{pmatrix}s-1\\t-1\\u\end{pmatrix}=1+\begin{pmatrix}1&1&1\end{pmatrix}\begin{pmatrix}s-1\\t-1\\u\end{pmatrix}\\&=1+s-1+t-1+u=s+t+u-1.\end{aligned} \quad (88)$$

For instance, we can approximate

$$1+\sin(0.1)=f(1,1,0.1)\approx\ell(1,1,0.1)=1.1. \quad (89)$$

Compare this with the true value $f(1,1,0.1)=1.09983\ldots$.

Finally, differentiability of $f:\mathbb{R}^n\to\mathbb{R}^m$ is defined component-wise. If exists, the derivative of $f:\mathbb{R}^n\to\mathbb{R}^m$ coincides with the Jacobian matrix of f, cf. Definition 6.7 and Remark 7.6. In the other direction, if the Jacobian matrix of f exists and continuous in an open region, then f is differentiable, cf. Theorem 7.7.

Example 8.3. The Jacobian matrix of $f(x,y)=\big(xyz,\sin(x+z^2)\big)$ is

$$J_f(x,y,z)=\begin{pmatrix}yz & xz & xy\\ \cos(x+z^2) & 0 & 2z\cos(x+z^2)\end{pmatrix}. \quad (90)$$

Since each component of J_f is continuous in $\mathbb{R}^3$, we conclude that f is differentiable in $\mathbb{R}^3$ with $Df(x,y,z)=J_f(x,y,z)$.

9. The chain rule

Recall from single variable calculus that the derivative of the composition $(f\circ g)(t)=f(g(t))$ is given by the so-called chain rule

$$\frac{df(g(t))}{dt}=f'(g(t))g'(t). \quad (91)$$

As a preliminary to the chain rule for multivariate functions, let us consider differentiating $f(g_1(t),g_2(t))$ with respect to t, where $f(x,y)$ is a bivariate function, and $g(t)=(g_1(t),g_2(t))$ is a vector valued univariate function. In view of the approximation

$$f(g_1(t+h),g_2(t+h))\approx f(g_1(t)+hg_1'(t),g_2(t)+hg_2'(t)) \quad (92)$$

and of the definition

$$\frac{df(g_1(t),g_2(t))}{dt}=\lim_{h\to0}\frac{f(g_1(t+h),g_2(t+h))-f(g_1(t),g_2(t))}{h}, \quad (93)$$

we formally derive

$$\frac{df(g(t))}{dt}=D_{g'(t)}f(g(t))=\frac{\partial f}{\partial x}(g(t))g_1'(t)+\frac{\partial f}{\partial y}(g(t))g_2'(t). \quad (94)$$

This formal derivation can be made precise, and yields the following theorem.

Theorem 9.1 (Chain rule)**.** *Let $K \subset \mathbb{R}^n$ be a set, and let $f : K \to \mathbb{R}^m$ be a function, differentiable at $y \in K$. Suppose that $U \subset \mathbb{R}^m$, such that $f(K) \subset U$, that is, $f(x) \in U$ for all $x \in K$. Assume that $g : U \to \mathbb{R}^k$ is differentiable at $f(y)$. Then the composition $g \circ f : K \to \mathbb{R}^k$, defined by $(g \circ f)(x) = g(f(x))$ for $x \in K$, is differentiable at y, with*

$$D(g \circ f)(y) = Dg(f(y))Df(y). \tag{95}$$

Example 9.2. Let $f : \mathbb{R}^2 \to \mathbb{R}$ and $g : \mathbb{R}^2 \to \mathbb{R}^2$ be given by

$$f(x,y) = xy, \qquad g(x,y) = \big(xy, \sin(x+y^2)\big). \tag{96}$$

Then we have

$$Df(x,y) = \begin{pmatrix} y & x \end{pmatrix}, \qquad Dg(x,y) = \begin{pmatrix} y & x \\ \cos(x+y^2) & 2y\cos(x+y^2) \end{pmatrix}. \tag{97}$$

The composition $h = f \circ g : \mathbb{R}^2 \to \mathbb{R}$ is $h(x,y) = xy\sin(x+y^2)$, whose derivative can be computed by the chain rule as

$$\begin{aligned} Dh(x,y) = Df(g(x,y))Dg(x,y) &= \begin{pmatrix} \sin(x+y^2) & xy \end{pmatrix} \begin{pmatrix} y & x \\ \cos(x+y^2) & 2y\cos(x+y^2) \end{pmatrix} \\ &= \begin{pmatrix} y\sin(x+y^2) + xy\cos(x+y^2) & x\sin(x+y^2) + 2xy^2\cos(x+y^2) \end{pmatrix}. \end{aligned} \tag{98}$$

Remark 9.3 (Convenient notation)**.** In practice, the following form of the chain rule is often more convenient. Let z be a quantity that depends on $y_1, y_2, \dots, y_n$, which in turn depend on other quantities $x, \dots$, as

$$z = z\big(y_1(x,\dots), y_2(x,\dots), \dots, y_n(x,\dots)\big). \tag{99}$$

Then

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y_1}\frac{\partial y_1}{\partial x} + \frac{\partial z}{\partial y_2}\frac{\partial y_2}{\partial x} + \dots + \frac{\partial z}{\partial y_n}\frac{\partial y_n}{\partial x}. \tag{100}$$

Example 9.4 (Polar coordinates)**.** The formula for transforming polar coordinates to Cartesian coordinates is given by

$$\begin{cases} x = r\cos\theta \\ y = r\sin\theta \end{cases} \tag{101}$$

which can be written as $(x,y) = \Phi(r,\theta)$. We can compute

$$D\Phi(r,\theta) = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}. \tag{102}$$

If we have a curve given in polar coordinates $\gamma(t) = (r(t), \theta(t))$, whose Cartesian representation is $\Gamma(t) = \Phi(\gamma(t))$, then its velocity transforms as

$$\dot\Gamma(t) = D\Phi(\gamma(t))\dot\gamma(t) = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix} \begin{pmatrix} \dot r \\ \dot\theta \end{pmatrix} = \begin{pmatrix} \dot r\cos\theta - r\dot\theta\sin\theta \\ \dot r\sin\theta + r\dot\theta\cos\theta \end{pmatrix}. \tag{103}$$

Following the preceding remark, we could have computed it as

$$\begin{aligned} \dot x = \frac{dx}{dt} &= \frac{\partial x}{\partial r}\frac{dr}{dt} + \frac{\partial x}{\partial\theta}\frac{d\theta}{dt} = (\cos\theta)\dot r - (r\sin\theta)\dot\theta, \\ \dot y = \frac{dy}{dt} &= \frac{\partial y}{\partial r}\frac{dr}{dt} + \frac{\partial y}{\partial\theta}\frac{d\theta}{dt} = (\sin\theta)\dot r + (r\cos\theta)\dot\theta. \end{aligned} \tag{104}$$

On the other hand, given a scalar function $f(x, y)$ of Cartesian coordinates, we can consider it in polar coordinates, as $f = f(r\cos\theta, r\sin\theta)$, and its gradient transforms as

$$\begin{aligned} \partial_r f = \frac{\partial f}{\partial r} &= \frac{\partial f}{\partial x}\frac{\partial x}{\partial r} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial r} = \cos\theta\partial_x f + \sin\theta\partial_y f \\ \partial_\theta f = \frac{\partial f}{\partial \theta} &= \frac{\partial f}{\partial x}\frac{\partial x}{\partial \theta} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial \theta} = -r\sin\theta\partial_x f + r\cos\theta\partial_y f \end{aligned} \tag{105}$$

This can be written as

$$\begin{pmatrix} \partial_r f \\ \partial_\theta f \end{pmatrix} = J^\top \begin{pmatrix} \partial_x f \\ \partial_y f \end{pmatrix}, \qquad \text{where} \qquad J = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}, \tag{106}$$

or

$$\begin{pmatrix} \partial_r f & \partial_\theta f \end{pmatrix} = \begin{pmatrix} \partial_x f & \partial_y f \end{pmatrix} J. \tag{107}$$

Compare this with (103) or (104), that is,

$$\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = J \begin{pmatrix} \dot{r} \\ \dot{\theta} \end{pmatrix}. \tag{108}$$

Quantities, whose components transform like (108), are called *vectors*. Velocity is an example of a vector quantitiy. In contrast, the gradient of a function is *not* a vector, since it transforms according to (107). Such quantities are called *covectors*, or *differential 1-forms*.

Example 9.5 (Linear regression)**.** Suppose that a collection (x_i, y_i), $i = 1, \dots, N$, of points on the plane $\mathbb{R}^2$ is given, which we think of as samples from some unknown functional relation $y = F(x)$. We want to approximate F by a linear function

$$y = f(x) = ax + b, \tag{109}$$

such that the *mean square error*

$$E(a, b) = \sum_{i=1}^{N} (f(x_i) - y_i)^2 = \sum_{i=1}^{N} (ax_i + b - y_i)^2, \tag{110}$$

is as small as possible. We compute the partial derivatives as

$$\begin{aligned} \frac{\partial E}{\partial a} &= \sum_{i=1}^{N} (ax_i + b - y_i)x_i = a\sum_{i=1}^{N} x_i^2 + b\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} x_i y_i, \\ \frac{\partial E}{\partial b} &= \sum_{i=1}^{N} (ax_i + b - y_i) = a\sum_{i=1}^{N} x_i + nb - \sum_{i=1}^{N} y_i. \end{aligned} \tag{111}$$

The parameters a and b are optimal when these partial derivatives vanish. That is, we would need to solve the 2×2 linear system

$$\begin{cases} Ga + Hb = K \\ Ha + nb = L \end{cases} \tag{112}$$

for the unknowns a and b, where

$$G = \sum_{i=1}^{N} x_i^2, \qquad H = \sum_{i=1}^{N} x_i, \qquad K = \sum_{i=1}^{N} x_i y_i, \qquad L = \sum_{i=1}^{N} y_i. \tag{113}$$

Example 9.6 (Backpropagation in a neural net)**.** Suppose that the output s of a neural net depends on the inputs x_1, x_2, x_3, and the weights $a_1, a_2, \dots, a_{12}$, according to the relations

$$\begin{aligned} y_1 &= a_1x_1 + a_2x_2 + a_3x_3, & z_1 &= \sigma(y_1), \\ y_2 &= a_4x_1 + a_5x_2 + a_6x_3, & z_2 &= \sigma(y_2), \\ u_1 &= a_7z_1 + a_8z_2, & v_1 &= \sigma(u_1), \\ u_2 &= a_9z_1 + a_{10}z_2, & v_2 &= \sigma(u_2), \\ w &= a_{11}v_1 + a_{12}v_2, & s &= \sigma(w), \end{aligned} \tag{114}$$

where σ is a given activation function, such as

$$\sigma(t) = \frac{1}{1+e^{-t}}. \tag{115}$$

Then we can compute the derivatives with respect to the weights as, e.g.,

$$\begin{aligned} \frac{\partial s}{\partial a_7} &= \frac{\partial s}{\partial w}\frac{\partial w}{\partial v_1}\frac{\partial v_1}{\partial u_1}\frac{\partial u_1}{\partial a_7} = \sigma'(w)a_{11}\sigma'(u_1)z_1, \\ \frac{\partial s}{\partial a_2} &= \frac{\partial s}{\partial w}\frac{\partial w}{\partial v_1}\frac{\partial v_1}{\partial u_1}\frac{\partial u_1}{\partial z_1}\frac{\partial z_1}{\partial y_1}\frac{\partial y_1}{\partial a_2} + \frac{\partial s}{\partial w}\frac{\partial w}{\partial v_2}\frac{\partial v_2}{\partial u_2}\frac{\partial u_2}{\partial z_1}\frac{\partial z_1}{\partial y_1}\frac{\partial y_1}{\partial a_2} \\ &= \sigma'(w)a_{11}\sigma'(u_1)a_7\sigma'(y_1)x_2 + \sigma'(w)a_{12}\sigma'(u_2)a_9\sigma'(y_1)x_2. \end{aligned} \tag{116}$$