Notes on Calculus on Vector Spaces[1]

Massimo Marinacci and Luigi Montrucchio
Collegio Carlo Alberto, Università di Torino

January 2009

[1]Translated by Gabriella Chiomio. We thank Claudio Mattalia for some very useful comments. Notes for the students of the Collegio Carlo Alberto and of the Università di Torino.

Contents

Chapter 1

Vector Spaces

1.1 Cartesian Products and $\mathbb{R}^n$

Suppose we want to classify a wine according to two characteristics, ageing and alcoholic strength. For example, suppose to read on a label: 2 years of ageing and 12 degrees. We write:

$$(2, 12)\,.$$

Look at another label and read: 1 year of ageing and 10 degrees. In this case we write:

$$(1, 10)\,.$$

The pairs $(2, 12)$ and $(1, 10)$ are called ordered pairs, and in them the first element, ageing, is distinguished from the second one, alcoholic strength. In an ordered pair, therefore, position is fundamental.

Let A_1 be the set of the possible years of ageing and A_2 the set of the possible alcoholic strengths. We can then write:

$$\begin{aligned} (2, 12) &\in A_1 \times A_2, \\ (1, 10) &\in A_1 \times A_2, \end{aligned}$$

A generic element of A_1 is denoted by a_1 and one of A_2 by a_2. For example, in $(2, 12)$ we have: $a_1 = 2$ and $a_2 = 12$.

Definition 1 *Given two sets A_1 and A_2, the Cartesian product $A_1 \times A_2$ is the set of all ordered pairs (a_1, a_2) with $a_1 \in A_1$ and $a_2 \in A_2$.*

In the example, we have $A_1 \subseteq \mathbb{N}$ and $A_2 \subseteq \mathbb{N}$, that is, the elements of A_1 and A_2 are natural numbers. More generally, suppose that $A_1 = A_2 = \mathbb{R}$, so that as elements

of A_1 and A_2 we have any real number, though with a possible different interpretation according to the position.

In this case, $A_1 \times A_2 = \mathbb{R} \times \mathbb{R} = \mathbb{R}^2$ and the pair (a_1, a_2) can be represented in the plane:

Fig. 1

An ordered pair of real numbers $(a_1, a_2) \in \mathbb{R}^2$ is called vector. In some cases, it can be useful to associate a segment to (a_1, a_2):

Fig. 2

Among the subsets of $\mathbb{R}^2$ particularly important are:

(i) $\{(a_1, a_2) \in \mathbb{R}^2 : a_1 = 0\}$, that is, the set of ordered pairs $(0, a_2)$. It is the axis of the ordinates.

(ii) $\{(a_1, a_2) \in \mathbb{R}^2 : a_2 = 0\}$, that is, the set of ordered pairs $(a_1, 0)$. It is the axis of the abscissae.

(iii) $\{(a_1, a_2) \in \mathbb{R}^2 : a_1 > 0 \text{ and } a_2 > 0\}$, that is, the set of ordered pairs (a_1, a_2) with components both positive. It is the first quadrant. In a similar way the other quadrants are defined:

Fig. 3

Before we have classified some wines using two characteristics, ageing and alcoholic strength. Now we consider a more complicated product, for example a portfolio of assets. Suppose there are 4 different assets that can be bought in the market. A portfolio is described by the ordered quadruple:

$$(a_1, a_2, a_3, a_4)\,,$$

where a_1 is the money invested in the first asset, a_2 the one invested in the second asset, and so on. For example:

$$(1000, 1500, 1200, 600)$$

denotes a portfolio in which 1000 euros have been invested in the first asset, 1500 in the second one, and so on. The position is fundamental, the portfolio:

$$(1500, 1200, 1000, 600)$$

is clearly different from the previous one, though the amounts of money invested are still 1500, 1200, 1000 and 600.

As the quantities of money are real numbers, we set $A_1 = A_2 = A_3 = A_4 = \mathbb{R}$, where A_i is the set of the possible amounts of money that can be invested in asset $i = 1, 2, 3, 4$. We have:

$$(a_1, a_2, a_3, a_4) \in A_1 \times A_2 \times A_3 \times A_4 = \mathbb{R}^4.$$

In particular,

$$(1000, 1500, 1200, 600) \in \mathbb{R}^4.$$

In general, we consider n sets $A_1, ..., A_n$.

Definition 2 *Given n sets $A_1, ..., A_n$, the Cartesian product $A_1 \times \cdots \times A_n$ is the set of all the ordered n-tuples $(a_1, ..., a_n)$ with $a_1 \in A_1, ..., a_n \in A_n$.*

If $A_1 = \cdots = A_n = \mathbb{R}$, we write:

$$A_1 \times \cdots \times A_n = \mathbb{R}^n.$$

An element $(a_1, ..., a_n) \in \mathbb{R}$ is called *vector*.

Notation. The Cartesian product $A_1 \times \cdots \times A_n$ is sometimes denoted by $\prod_{i=1}^{n} A_i$.

The vectors in $\mathbb{R}^3$ have a graphical representation:

Fig. 4

1.2 Operations in $\mathbb{R}^n$

Consider two vectors in $\mathbb{R}^n$:

$$\begin{aligned} x &= (x_1, ..., x_n), \\ y &= (y_1, ..., y_n). \end{aligned}$$

We define the *sum* vector $x + y$ as:

$$x + y = (x_1 + y_1, ..., x_n + y_n).$$

For example, consider in $\mathbb{R}^3$ the two vectors $x = (7, 8, 9)$ and $y = (2, 4, 7)$. We have:

$$x + y = (7 + 2, 8 + 4, 9 + 7) = (9, 12, 16).$$

Note that $x + y \in \mathbb{R}^n$, that is, through the sum a new element of $\mathbb{R}^n$ has been constructed.

Let now $\alpha \in \mathbb{R}$ and $x \in \mathbb{R}^n$. We call *product* of the scalar α for the vector x the vector αx defined as:

$$\alpha x = (\alpha x_1, ..., \alpha x_n).$$

For example, given $\alpha = 2$ and $x = (7, 8, 9) \in \mathbb{R}^3$, we have:

$$2x = (2 \cdot 7, 2 \cdot 8, 2 \cdot 9) = (14, 16, 18).$$

Also $\alpha x \in \mathbb{R}^n$, so that also with the scalar multiplication a new element of $\mathbb{R}^n$ has been constructed.

Notation. Let $-x = (-x_1, ..., -x_n)$. We have $-x = (-1)\, x$ and $x - y = x + (-1)\, y$. Moreover, set $\mathbf{0} = (0, ..., 0)$.

We have therefore introduced in $\mathbb{R}^n$ two operations, sum and scalar multiplication. Let us see their properties. We start with the sum.

Proposition 3 *Let x, y and z be three vectors in $\mathbb{R}^n$. We have:*

(i) $x + y \in \mathbb{R}^n$ *($\mathbb{R}^n$ is closed with respect to the sum),*

(ii) $x + y = y + x$ *(commutative property),*

(iii) $(x + y) + z = x + (y + z)$ *(associative property),*

(iv) $x + \mathbf{0} = x$ *(existence of a neutral element for the sum),*

(v) $x + (-x) = \mathbf{0}$ *(existence of the opposite of each vector).*

Proof These properties are easily checked. For example, we prove (ii). We have:

$$x + y = (x_1 + y_1, ..., x_n + y_n) = (y_1 + x_1, ..., y_n + x_n) = y + x,$$

as desired. ■

Consider now the scalar multiplication.

Proposition 4 *Let $x, y \in \mathbb{R}^n$ and $\alpha, \beta \in \mathbb{R}$. We have:*

(i) $\alpha x \in \mathbb{R}^n$ *($\mathbb{R}^n$ is closed with respect to the scalar multiplication),*

(ii) $\alpha\,(x + y) = \alpha x + \alpha y$ *(distributive property),*

(iii) $(\alpha + \beta)\, x = \alpha x + \beta x$ *(distributive property),*

(iv) $1x = x$ *(existence of a neutral element for the scalar multiplication),*

(v) $\alpha(\beta x) = (\alpha\beta) x$ *(associative property).*

Proof Also in this case these properties are easily checked. For example, consider (iii). We have:

$$\begin{aligned}(\alpha+\beta) x &= ((\alpha+\beta) x_1, ..., (\alpha+\beta) x_n) \\ &= (\alpha x_1 + \beta x_1, ..., \alpha x_n + \beta x_n) \\ &= (\alpha x_1, ..., \alpha x_n) + (\beta x_1, ..., \beta x_n) \\ &= \alpha x + \beta x\end{aligned}$$

as desired. ■

1.3 Vector Spaces

We have seen how in $\mathbb{R}^n$ two operations can be defined, sum and scalar multiplication. These operations feature some properties, described in the two last propositions. We now consider another space and show that also in this case two similar operations can be defined.

A polynomial $f(x)$ of degree n has the form $f(x) = a_0 + a_1 x + \cdots + a_n x^n$, with $a_n \neq 0$ and $a_i \in \mathbb{R}$ for each $0 \leq i \leq n-1$. Let $\mathcal{P}_n$ be the set of all the polynomials of degree less than or equal to n, with in addition the degenerated polynomial $\mathbf{0}$, whose coefficients are all zero. Clearly,

$$\mathcal{P}_1 \subseteq \mathcal{P}_2 \subseteq \cdots \subseteq \mathcal{P}_n.$$

Example 5 We have $f(x) = x + x^2 \in \mathcal{P}_2$, while $f(x) = 3x - 10x^4 \in \mathcal{P}_4$. ▲

Let $f(x) = a_0 + a_1 x + \cdots + a_n x^n$ and $g(x) = b_0 + b_1 x + \cdots + b_n x^n$ be two elements of $\mathcal{P}_n$, that is, two polynomials at most of degree n. We define the sum polynomial as:

$$(f+g)(x) = (a_0 + b_0) + (a_1 + b_1) x + \cdots + (a_n + b_n) x^n.$$

For example, consider in $\mathcal{P}_4$ the two polynomials $f(x) = x + x^2$ and $g(x) = 3x - 10x^4$. We have:

$$\begin{aligned}(f+g)(x) &= (0+0) + (1+3) x + (1+0) x^2 + (0+0) x^3 + (0-10) x^4 \\ &= 4x + x^2 - 10x^4.\end{aligned}$$

Note that $f + g \in \mathcal{P}_4$, and so the sum polynomial $f + g$, is also an element of the space $\mathcal{P}_n$.

Now let be $\alpha \in \mathbb{R}$ and $f(x) = a_0 + a_1 x + \cdots + a_n x^n \in \mathcal{P}_n$. The scalar multiplication αf in $\mathcal{P}_n$ is defined as:

$$(\alpha f)(x) = \alpha a_0 + \alpha a_1 x + \cdots + \alpha a_n x^n.$$

For example, consider the polynomial $f(x) = x + 5x^4 \in \mathcal{P}_4$ and $\alpha = 5$. We have:

$$(5f)(x) = 5x + 25x^4.$$

Again, note that $\alpha f \in \mathcal{P}_n$.

Proposition 6 *Let f, g and h be three elements of $\mathcal{P}_n$ and let α and β be two elements of $\mathbb{R}$. We have:*

(i) $\mathcal{P}_n$ *is closed with respect to the sum and to the scalar multiplication.*

(ii) $f + g = g + f$ *(commutative property).*

(iii) $(f + g) + h = f + (g + h)$ *(associative property).*

(iv) $f + \mathbf{0} = f$ *(existence of a neutral element for the sum).*

(v) $f + (-f) = 0$ *(existence of the opposite of each $f \in \mathcal{P}_n$).*

(vi) $\alpha(f + g) = \alpha f + \alpha g$ *(distributive property).*

(vii) $(\alpha + \beta) f = \alpha f + \beta f$ *(distributive property).*

(viii) $1f = f$ *(existence of a neutral element for the scalar multiplication).*

(ix) $\alpha(\beta f) = (\alpha\beta) f$ *(associative property).*

Proof Let prove for example the (vi). Let $f(x) = a_0 + a_1 x + \cdots + a_n x^n$ and $g(x) = b_0 + b_1 x + \cdots + b_n x^n$ be two elements of $\mathcal{P}_n$. We have:

$$\begin{aligned}
\alpha(f + g)(x) &= \alpha\left[(a_0 + b_0) + (a_1 + b_1) x + \cdots + (a_n + b_n) x^n\right] \\
&= \alpha(a_0 + b_0) + \alpha(a_1 + b_1) x + \cdots + \alpha(a_n + b_n) x^n \\
&= \alpha a_0 + \alpha a_1 x + \cdots + \alpha a_n x^n + \alpha b_0 + \alpha b_1 x + \cdots + \alpha b_n x^n \\
&= \alpha f(x) + \beta g(x),
\end{aligned}$$

as desired. ■

We have therefore considered two spaces, $\mathbb{R}^n$ and $\mathcal{P}_n$, in which it is possible to define two operations, sum and scalar multiplication, that share similar properties in the two spaces. This analogy suggests the following fundamental abstraction.

Definition 7 *Let V be a set on which two operations are defined, sum and scalar multiplication. The sum associates to each pair $v, w \in V$ the element $v + w \in V$; the scalar multiplication associates to each $\alpha \in \mathbb{R}$ and $v \in V$ the element $\alpha v \in V$. The set V is said to be a vector space (on $\mathbb{R}$) if, for every $v, w, z \in V$ and every $\alpha, \beta \in \mathbb{R}$, these operations satisfy the following properties:*

(i) $v + w = w + v$ *(commutative property).*

(ii) $(v + w) + z = v + (w + z)$ *(associative property).*

(iii) There exists an element $\mathbf{0} \in V$ *such that* $v + \mathbf{0} = v$ *(existence of a neutral element for the sum).*

(iv) There exists an element $-v \in V$ *such that* $v + (-v) = \mathbf{0}$ *(existence of the opposite of each* $v \in V$*).*

(v) $1v = v$ *(existence of a neutral element for the scalar multiplication).*

(vi) $\alpha(v + w) = \alpha v + \alpha w$ *(distributive property).*

(vii) $(\alpha + \beta) v = \alpha v + \beta v$ *(distributive property).*

(viii) $\alpha(\beta v) = (\alpha\beta) v$ *(associative property).*

>From this definition it follows immediately that $\mathbb{R}^n$ and $\mathcal{P}_n$, endowed with their operations of sum and scalar multiplication, are two examples of vector spaces. In fact, we have already seen in the previous sections as such operations satisfy properties (i)-(viii) of the Definition 7.

To show another example of a vector space, we now introduce matrices. A matrix $m \times n$ is a table of real numbers

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdots & \cdot \\ a_{m1} & a_{m2} & \cdots & a_{mj} & \cdots & a_{mn} \end{bmatrix}$$

For example,

$$\begin{bmatrix} 1 & 5 & 7 & 9 \\ 3 & -2 & -1 & -4 \\ 12 & 15 & 11 & 9 \end{bmatrix}$$

is a matrix 3×4, in which

$$\begin{aligned} a_{11} &= 1, \quad a_{12} = 5, \quad a_{13} = 7, \quad a_{14} = 9, \\ a_{21} &= 3, \quad a_{22} = -2, \quad a_{23} = -1, \quad a_{24} = -4, \\ a_{31} &= 12, \quad a_{32} = 15, \quad a_{33} = 11, \quad a_{34} = 9. \end{aligned}$$

Notation. The matrix of components a_{ij} is sometimes denoted with (a_{ij}).

In a matrix (a_{ij}) we distinguish n columns (said *column vectors*):

$$\begin{bmatrix} a_{11} \\ . \\ . \\ a_{m1} \end{bmatrix}, \begin{bmatrix} a_{12} \\ . \\ . \\ a_{m2} \end{bmatrix}, ..., \begin{bmatrix} a_{1n} \\ . \\ . \\ a_{mn} \end{bmatrix},$$

and m rows (said *row vectors*):

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \end{pmatrix}, \\ \begin{pmatrix} a_{21} & \cdots & a_{2n} \end{pmatrix}, \\ \cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots \\ \begin{pmatrix} a_{m1} & \cdots & a_{mn} \end{pmatrix}.$$

For example, for the matrix

$$\begin{bmatrix} 1 & 5 & 7 & 9 \\ 3 & -2 & -1 & -4 \\ 12 & 15 & 11 & 9 \end{bmatrix}$$

we have 3 row vectors:

$$\begin{pmatrix} 1 & 5 & 7 & 9 \end{pmatrix}, \\ \begin{pmatrix} 3 & -2 & -1 & -4 \end{pmatrix}, \\ \begin{pmatrix} 12 & 15 & 11 & 9 \end{pmatrix},$$

and 4 column vectors:

$$\begin{bmatrix} 1 \\ 3 \\ 12 \end{bmatrix}, \begin{bmatrix} 5 \\ -2 \\ 15 \end{bmatrix}, \begin{bmatrix} 7 \\ -1 \\ 11 \end{bmatrix}, \begin{bmatrix} 9 \\ -4 \\ 9 \end{bmatrix}.$$

When $m = n$, the matrix is said to be *square*, while when $m \neq n$ the matrix is said to be *rectangular*. For example,

$$\begin{bmatrix} 1 & 5 & -1 \\ 3 & 4 & 2 \\ 1 & 7 & 9 \end{bmatrix}$$

is a square matrix 3×3.

Let $M(m,n)$ be the space of all the matrices $m \times n$. In $M(m,n)$ we can define in a natural way the operations of sum and scalar multiplication. As to the sum, let (a_{ij}) and (b_{ij}) be two matrices in $M(m,n)$; the sum matrix $(a_{ij}) + (b_{ij})$ is defined as:

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} + \begin{bmatrix} b_{11} & \cdots & b_{1n} \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ b_{m1} & \cdots & b_{mn} \end{bmatrix} = \begin{bmatrix} a_{11}+b_{11} & \cdots & a_{1n}+b_{1n} \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ a_{m1}+b_{m1} & \cdots & a_{mn}+b_{mn} \end{bmatrix},$$

that is, $(a_{ij}) + (b_{ij}) = (a_{ij} + b_{ij})$. For example:

$$\begin{bmatrix} 1 & 5 & 7 & 9 \\ 3 & -2 & -1 & -4 \\ 12 & 15 & 11 & 9 \end{bmatrix} + \begin{bmatrix} 0 & 2 & 1 & 4 \\ -1 & 3 & 1 & 4 \\ 5 & 8 & 1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 7 & 8 & 13 \\ 2 & 1 & 0 & 0 \\ 17 & 23 & 12 & 11 \end{bmatrix}.$$

Given $\alpha \in \mathbb{R}$ and $(a_{ij}) \in M(m,n)$, the scalar multiplication $\alpha(a_{ij})$ is defined as:

$$\alpha \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} = \begin{bmatrix} \alpha a_{11} & \cdots & \alpha a_{1n} \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ \alpha a_{m1} & \cdots & \alpha a_{mn} \end{bmatrix},$$

that is $\alpha(a_{ij}) = (\alpha a_{ij})$. For example,

$$4 \begin{bmatrix} 1 & 5 & 7 & 9 \\ 3 & -2 & -1 & -4 \\ 12 & 15 & 11 & 9 \end{bmatrix} = \begin{bmatrix} 4 & 20 & 28 & 36 \\ 12 & -8 & -4 & -16 \\ 48 & 60 & 44 & 36 \end{bmatrix}.$$

It is immediate to see that $M(m,n)$ endowed with these two operations is a vector space. We have therefore a third example, besides $\mathbb{R}^n$ and $\mathcal{P}_n$, of a vector space.

1.4 First Properties of Vector Spaces

In this section we show some of the first properties of vector spaces. It is important to observe that the proofs of these properties are based exclusively on the properties (i)-(viii) of Definition 7.

Proposition 8 *In a vector space V the neutral element $\mathbf{0}$ is unique.*

Proof Suppose there exists $\mathbf{0}' \in V$ such that $v + \mathbf{0}' = v$ for every $v \in V$. To prove that $\mathbf{0}$ is unique, we have to show that $\mathbf{0}' = \mathbf{0}$. We have:

$$\begin{aligned} \mathbf{0}' + \mathbf{0} &= \mathbf{0}' \qquad \text{(definition of } \mathbf{0}\text{)}, \\ \mathbf{0} + \mathbf{0}' &= \mathbf{0} \qquad \text{(hypothesis on } \mathbf{0}'\text{)}. \end{aligned}$$

On the other hand, the commutative property implies $\mathbf{0}' + \mathbf{0} = \mathbf{0} + \mathbf{0}'$, and so we conclude that $\mathbf{0}' = \mathbf{0}$, as desired. ■

Proposition 9 *In a vector space V each element has a unique inverse $-v$.*

Proof Suppose there exists $w \in V$ such that $v + w = \mathbf{0}$. To prove that $-v$ is unique, we have to show that $w = -v$. Since $v + w = \mathbf{0}$, we have:

$$\begin{aligned} -v &= (-v) + \mathbf{0} \\ &= (-v) + (v + w) \\ &= (-v + v) + w \qquad \text{(associative property)} \\ &= \mathbf{0} + w \\ &= w. \end{aligned}$$

It follows that $-v = w$, as desired. ■

Proposition 10 *Let v and w be any two vectors of a vector space V. There exists a unique vector $x \in V$ such that $v + x = w$.*

Proof Define $x = w + (-v)$. We have:

$$\begin{aligned} v + x &= v + (w + (-v)) \\ &= (w + (-v)) + v \qquad \text{(commutative property)} \\ &= w + (-v + v) \qquad \text{(associative property)} \\ &= w + (v + (-v)) \qquad \text{(commutative property)} \\ &= w + \mathbf{0} = w. \end{aligned}$$

Therefore, $x = w + (-v)$ is such that $v + x = w$. Let us prove that $w + (-v)$ is also the unique vector of this type. Let x be whatever vector of V such that $v + x = w$. We have to show that $x = w + (-v)$. We have:

$$\begin{aligned} x &= x + \mathbf{0} \\ &= x + (v + (-v)) \\ &= (x + v) + (-v) \qquad \text{(associative property)} \\ &= w + (-v)\,, \end{aligned}$$

as desired. ■

Proposition 11 *Let V be a vector space. For every $v \in V$ we have $0v = \mathbf{0}$, $(-1)\, v = -v$ and $\alpha \mathbf{0} = \mathbf{0}$.*

Proof We start by proving that $0v = \mathbf{0}$. By definition, $\mathbf{0}$ solves $x + v = v$. On the other hand, we have:

$$\begin{aligned} v + 0v &= 1v + 0v \\ &= (1 + 0)\, v \qquad \text{(distributive property)} \\ &= 1v = v. \end{aligned}$$

Therefore, also $0v$ solves $v + x = v$. From Proposition 10 it then follows that $0v = \mathbf{0}$.

To prove that $(-1)\, v = -v$, we observe that for every $v \in V$ we have:

$$\begin{aligned} (-1)\, v + v &= (-1)\, v + 1v \\ &= (-1 + 1)\, v \qquad \text{(distributive property)} \\ &= 0v. \end{aligned}$$

But, according to what has been just proved, we have $0v = \mathbf{0}$ and therefore $(-1)\, v + v = \mathbf{0}$. Hence, both $-v$ and $(-1)\, v$ solve $x + v = \mathbf{0}$. By Proposition 10 we can conclude that $(-1)\, v = -v$.

To prove that $\alpha \mathbf{0} = \mathbf{0}$, observe that $\alpha \mathbf{0} = \alpha\,(v - v) = \alpha v + \alpha\,(-v) = \alpha v - \alpha v = \mathbf{0}$. ■

1.5 Vector Subspaces

Definition 12 *A nonempty subset W of a vector space V is a vector subspace of V if $\alpha v + \beta w \in W$ for every $\alpha, \beta \in \mathbb{R}$ and every $v, w \in W$.*

The following characterization clarifies the nature of vector subspaces.

Proposition 13 *A nonempty subset W of a vector space V is a vector subspace of V if and only if W is itself a vector space with respect to the operations of sum and scalar multiplication inherited by V.*

Proof "If." Let W be a vector space with respect to the operations of sum and scalar multiplication inherited by V. Let $v, w \in W$. Since W is closed with respect to the scalar multiplication, we have that $\alpha v \in W$ and $\beta w \in W$. It follows that $\alpha v + \beta w \in W$ because W is closed with respect to the sum. Therefore, W is a vector subspace.

"Only if." Let W be a vector subspace of V and let $v, w \in W$. Setting $\alpha = \beta = 1$ in Definition 12 we get $v + w \in W$, while setting $\beta = 0$ we get $\alpha v \in W$. Therefore, W is closed with respect to the operations of sum and scalar multiplication inherited by V. We leave to the reader the easy check that these operations satisfy in W the properties (i)-(viii) of Definition 7. ■

Example 14 Let $m \leq n$ and

$$M = \{x \in \mathbb{R}^n : x_1 = \cdots = x_m = 0\}.$$

For example, if $n = 3$ and $m = 2$, we have:

$$M = \left\{x \in \mathbb{R}^3 : x_1 = x_2 = 0\right\}.$$

The set M is a vector subspace of $\mathbb{R}^n$. In fact, let $x, y \in M$ and $\alpha, \beta \in \mathbb{R}$. We have:

$$\begin{aligned} \alpha x + \beta y &= (\alpha x_1 + \beta y_1, ..., \alpha x_n + \beta y_n) \\ &= (0, ..., 0, \alpha x_{m+1} + \beta y_{m+1}, ..., \alpha x_n + \beta y_n) \in M. \end{aligned}$$

In particular, the axis of the ordinates, which corresponds to $M = \{x \in \mathbb{R}^2 : x_1 = 0\}$, is a vector subspace of $\mathbb{R}^2$. ▲

Example 15 Let $r_1, ..., r_m$ be real numbers with $m \leq n + 1$ and let

$$M = \{f \in \mathcal{P}_n : f(r_i) = 0 \quad \text{for } i = 1, ..., m\}.$$

The set M is a vector subspace of $\mathcal{P}_n$. In fact, let $f, g \in \mathcal{P}_n$ and $\alpha, \beta \in \mathbb{R}$. We have:

$$(\alpha f + \beta g)(r_i) = \alpha f(r_i) + \beta g(r_i) \quad \text{for } i = 1, ..., m$$

and therefore $\alpha f + \beta g \in M$. For example, consider $\mathcal{P}_2$ and the numbers 1 and 3. In this case,

$$M = \{f \in \mathcal{P}_2 : f(1) = f(3) = 0\},$$

that is, M is the set of all the polynomials

$$a_0 + a_1 x + a_2 x^2$$

whose coefficients a_0, a_1, a_2 are such that:

$$\begin{cases} a_0 + a_1 + a_2 = 0 \\ a_0 + 3a_1 + 9a_2 = 0 \end{cases}$$

This system of equations is solved by $\left(t, -\frac{4}{3}t, \frac{t}{3}\right)$ for every $t \in \mathbb{R}$. Therefore,

$$M = \left\{ t - \frac{4}{3}tx + \frac{t}{3}x^2 : t \in \mathbb{R} \right\}.$$

If, instead, we consider the three numbers 1, 3 and 5, we have:

$$M = \left\{ f \in \mathcal{P}_2 : f(1) = f(3) = f(5) = 0 \right\},$$

that is, this time M is the set of all the polynomials

$$a_0 + a_1 x + a_2 x^2$$

whose coefficients a_0, a_1, a_2 are such that:

$$\begin{cases} a_0 + a_1 + a_2 = 0 \\ a_0 + 3a_1 + 9a_2 = 0 \\ a_0 + 5a_1 + 25a_2 = 0 \end{cases}$$

This system has the unique solution $(0, 0, 0)$ and therefore M is the trivial vector subspace $M = \{\mathbf{0}\}$ of $\mathcal{P}_2$, where $\mathbf{0}$ is the degenerated polynomial whose coefficients are all null. This result is actually not surprising as the polynomials in $\mathcal{P}_2$ can have at most two roots, while the condition $f(1) = f(3) = f(5) = 0$ requires that 1, 3 and 5 be all roots of the polynomial, which in $\mathcal{P}_2$ can hold only for the degenerated polynomial $\mathbf{0}$. ▲

Example 16 $\mathcal{P}_k$ is a vector subspace of $\mathcal{P}_n$ for every $k \leq n$. In fact, we have:

$$\mathcal{P}_k = \left\{ f \in \mathcal{P}_n : a_{k+1} = \cdots = a_n = 0 \right\}.$$

. ▲

Example 17 Let M be the set of all $x \in \mathbb{R}^4$ such that:

$$\begin{cases} 2x_1 - x_2 + 2x_3 + 2x_4 = 0 \\ x_1 - x_2 - 2x_3 - 4x_4 = 0 \\ x_1 - 2x_2 - 2x_3 - 10x_4 = 0 \end{cases}$$

It it possible to check that the vectors

$$\left(-\frac{10}{3}t, -6t, -\frac{2}{3}t, t \right)$$

solve the system for every $t \in \mathbb{R}$. It follows that:

$$M = \left\{ \left(-\frac{10}{3}t, -6t, -\frac{2}{3}t, t \right) : t \in \mathbb{R} \right\}.$$

For completeness we solve the system. Consider x_4 as a"parameter" and solve the system for x_1, x_2 and x_3; clearly, these solutions will depend on the value of the parameter x_4.

$$\left\{\begin{array}{l} 2x_1 - x_2 + 2x_3 + 2x_4 = 0 \\ x_1 - x_2 - 2x_3 - 4x_4 = 0 \\ x_1 - 2x_2 - 2x_3 - 10x_4 = 0 \end{array}\right. \Rightarrow \left\{\begin{array}{l} 2x_1 - x_2 = -2x_3 - 2x_4 \\ x_1 - x_2 = 2x_3 + 4x_4 \\ x_1 - 2x_2 - 2x_3 - 10x_4 = 0 \end{array}\right.$$

$$\left\{\begin{array}{l} 2\left(x_2 + 2x_3 + 4x_4\right) - x_2 = -2x_3 - 2x_4 \\ x_1 + \left(-2x_3 - 2x_4 - 2x_1\right) = 2x_3 + 4x_4 \\ x_1 - 2x_2 - 2x_3 - 10x_4 = 0 \end{array}\right. \Rightarrow \left\{\begin{array}{l} x_2 = -6x_3 - 10x_4 \\ x_1 = -4x_3 - 6x_4 \\ x_1 - 2x_2 - 2x_3 - 10x_4 = 0 \end{array}\right.$$

$$\left\{\begin{array}{l} x_2 = -6x_3 - 10x_4 \\ x_1 = -4x_3 - 6x_4 \\ \left(-4x_3 - 6x_4\right) - 2\left(-6x_3 - 10x_4\right) - 2x_3 - 10x_4 = 0 \end{array}\right. \Rightarrow \left\{\begin{array}{l} x_2 = -6x_3 - 10x_4 \\ x_1 = -4x_3 - 6x_4 \\ x_3 = -\frac{2}{3}x_4 \end{array}\right. \Rightarrow$$

$$\left\{\begin{array}{l} x_2 = -6\left(-\frac{2}{3}x_4\right) - 10x_4 \\ x_1 = -4\left(-\frac{2}{3}x_4\right) - 6x_4 \\ x_3 = -\frac{2}{3}x_4 \end{array}\right. \Rightarrow \left\{\begin{array}{l} x_2 = -6x_4 \\ x_1 = -\frac{10}{3}x_4 \\ x_3 = -\frac{2}{3}x_4 \end{array}\right.$$

This implies that all and only the vectors of $\mathbb{R}^4$ of the form $\left(-\frac{10}{3}t, -6t, -\frac{2}{3}t, t\right)$ solve the system for every $t \in \mathbb{R}$.. ▲

1.6 Operations on the Subspaces

If W_1 and W_2 are two vector subspaces of V, it is possible to show that also the intersection $W_1 \cap W_2$ is a vector subspace of V. More generally, we have:

Proposition 18 *If $W_1, ..., W_n$ are n vector subspaces of V, then $\bigcap_{i=1}^{n} W_i$ is a vector subspace of V.*

Proof As $\mathbf{0} \in W_i$ for every $1 \leq i \leq n$, we have that $\bigcap_{i=1}^{n} W_i \neq \emptyset$. Let $v, w \in W$ and $\alpha, \beta \in \mathbb{R}$. As $v, w \in \bigcap_{i=1}^{n} W_i$, we have $v, w \in W_i$ for every $i = 1, ..., n$ and therefore $\alpha v + \beta w \in W_i$ for every $i = 1, ..., n$ since each W_i is a vector subspace of V. So, $\alpha v + \beta w \in \bigcap_{i=1}^{n} W_i$ and therefore $\bigcap_{i=1}^{n} W_i$ is a vector subspace of V. ■

The union of vector subspaces is not in general a vector subspace of V, as the next example shows.

Example 19 The sets $W_1 = \{x \in \mathbb{R}^2 : x_1 = 0\}$ and $W_2 = \{x \in \mathbb{R}^2 : x_2 = 0\}$ are both vector subspaces of $\mathbb{R}^2$. We have:

$$W_1 \cup W_2 = \left\{x \in \mathbb{R}^2 : x_1 = 0 \text{or } x_2 = 0\right\},$$

which is not a vector subspace of $\mathbb{R}^2$. In fact, both $(1, 0)$ and $(0, 1)$ belong to $W_1 \cup W_2$, but $(1, 0) + (0, 1) = (1, 1) \notin W_1 \cup W_2$. ▲

1.7 Linear Independence

Definition 20 *A finite set of vectors* $\{v^i\}_{i=1}^n$ *of a vector space* V *is said to be linearly independent if, for each set* $\{\alpha_i\}_{i=1}^n$ *of real numbers, we have:*

$$\alpha_1 v^1 + \alpha_2 v^2 + \cdots + \alpha_n v^n = \mathbf{0} \Longrightarrow \alpha_1 = \alpha_2 = \cdots = \alpha_n = 0.$$

The set $\{v^i\}_{i=1}^n$ is said *linearly dependent* if it is not linearly independent, i.e., if there exists a set $\{\alpha_i\}_{i=1}^n$ of real numbers, not all null, such that:

$$\alpha_1 v^1 + \alpha_2 v^2 + \cdots + \alpha_n v^n = \mathbf{0}.$$

Example 21 In $\mathbb{R}^n$ consider the vectors:

$$\begin{aligned} e^1 &= (1, 0, ..., 0), \\ e^2 &= (0, 1, 0, ..., 0), \\ &\vdots \\ e^n &= (0, 0, ..., 0, 1). \end{aligned}$$

The set $\{e^1, ..., e^n\}$ is linearly independent. In fact, we have

$$\alpha_1 e^1 + \cdots + \alpha_n e^n = (\alpha_1, ..., \alpha_n)$$

and therefore $\alpha_1 e^1 + \cdots + \alpha_n e^n = \mathbf{0}$ implies $\alpha_1 = \cdots = \alpha_n = 0$. ▲

Before continuing with the examples, there is a question of terminology to clarify. Although linear independence and dependence are properties of a set of vectors $\{v^i\}_{i=1}^n$, in the sequel we will often say "sets of linearly independent (dependent) vectors" rather than "linearly independent (dependent) set of vectors".

Example 22 In $\mathbb{R}^3$, the vectors

$$\begin{aligned} x^1 &= (1, 1, 1), \\ x^2 &= (3, 1, 5), \\ x^3 &= (9, 1, 25), \end{aligned}$$

are linearly independent. In fact,

$$\begin{aligned} \alpha_1 x^1 + \alpha_2 x^2 + \alpha_3 x^3 &= \alpha_1 (1, 1, 1) + \alpha_2 (3, 1, 5) + \alpha_3 (9, 1, 25) \\ &= (\alpha_1 + 3\alpha_2 + 9\alpha_3, \alpha_1 + \alpha_2 + \alpha_3, \alpha_1 + 5\alpha_2 + 25\alpha_3) \end{aligned}$$

and therefore $\alpha_1 x^1 + \alpha_2 x^2 + \alpha_3 x^3 = \mathbf{0}$ implies

$$\begin{cases} \alpha_1 + 3\alpha_2 + 9\alpha_3 = 0 \\ \alpha_1 + \alpha_2 + \alpha_3 = 0 \\ \alpha_1 + 5\alpha_2 + 25\alpha_3 = 0 \end{cases},$$

which is a system of equations whose unique solution is $(\alpha_1, \alpha_2, \alpha_3) = (0, 0, 0)$. More generally, to verify if k vectors

$$\begin{aligned} x^1 &= \left(x_1^1, ..., x_n^1\right), \\ x^2 &= \left(x_1^2, ..., x_n^2\right), \\ &\cdot \\ &\cdot \\ &\cdot \\ x^k &= \left(x_1^k, ..., x_n^k\right), \end{aligned}$$

are linearly independent in $\mathbb{R}^n$ it is necessary to solve the following system:

$$\begin{cases} \alpha_1 x_1^1 + \alpha_2 x_1^2 + \cdots + \alpha_k x_1^k = 0 \\ \alpha_1 x_2^1 + \alpha_2 x_2^2 + \cdots + \alpha_k x_2^k = 0 \\ \cdots\cdots \\ \cdots\cdots \\ \alpha_1 x_n^1 + \alpha_2 x_n^2 + \cdots + \alpha_k x_n^k = 0 \end{cases}$$

If $(\alpha_1, ..., \alpha_k) = (0, ..., 0)$ is the unique solution, then these vectors are linearly independent in $\mathbb{R}^n$. For example, consider in $\mathbb{R}^3$ the two vectors $x^1 = (1, 3, 4)$ and $x^2 = (2, 5, 1)$. The system to solve is:

$$\begin{cases} \alpha_1 + 2\alpha_2 = 0 \\ 3\alpha_1 + 5\alpha_2 = 0 \\ 4\alpha_1 + \alpha_2 = 0 \end{cases},$$

whose unique solution is $(\alpha_1, \alpha_2) = (0, 0)$. These two vectors x^1 and x^2 are therefore linearly independent. ▲

Example 23 Consider the vectors:

$$\begin{aligned} x^1 &= (2, 1, 1), \\ x^2 &= (-1, -1, -2), \\ x^3 &= (2, -2, -2), \\ x^4 &= (2, -4, -10). \end{aligned}$$

To verify if these vectors are linearly independent in $\mathbb{R}^3$, we solve the system:

$$\begin{cases} 2\alpha_1 - \alpha_2 + 2\alpha_3 + 2\alpha_4 = 0 \\ \alpha_1 - \alpha_2 - 2\alpha_3 - 4\alpha_4 = 0 \\ \alpha_1 - 2\alpha_2 - 2\alpha_3 - 10\alpha_4 = 0 \end{cases}$$

As we saw before, this system is solved by the vectors

$$\left(-\frac{10}{3}t, -6t, -\frac{2}{3}t, t\right) \tag{1.1}$$

for every $t \in \mathbb{R}$. Therefore, $(0,0,0,0)$ is not the unique solution of the system and so the vectors x^1, x^2, x^3 and x^4 are linearly dependent. In fact, setting for example $t = 1$ in (1.1), we have that the quadruple

$$(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = \left(-\frac{10}{3}, -6, -\frac{2}{3}, 1\right)$$

is an example of set of coefficients not all null such that $\alpha_1 x^1 + \alpha_2 x^2 + \alpha_3 x^3 + \alpha_4 x^4 = \mathbf{0}$.
▲

Example 24 Consider in $\mathcal{P}_2$ the following polynomials:

$$f_1(x) = 1, \quad f_2(x) = x, \quad f_3(x) = x^2.$$

These polynomials are linearly independent. In fact, suppose per contra that this is not true and that f_1, f_2 and f_3 are linearly dependent. By definition, there exists a set $\{\alpha_1^*, \alpha_2^*, \alpha_3^*\}$ of real numbers not all equal to zero such that:

$$\alpha_1^* f_1 + \alpha_2^* f_2 + \alpha_3^* f_3 = \mathbf{0}.$$

This means that for every $x \in \mathbb{R}$ it holds:

$$\alpha_1^* f_1(x) + \alpha_2^* f_2(x) + \alpha_3^* f_3(x) = 0, \tag{1.2}$$

that is,

$$\alpha_1^* + \alpha_2^* x + \alpha_3^* x^2 = 0 \tag{1.3}$$

for every $x \in \mathbb{R}$. If $\alpha_3^* \neq 0$, equation (1.3) would be a second degree equation, which as well known can have at most two solutions in $\mathbb{R}$. Therefore, (1.3) cannot hold for every $x \in \mathbb{R}$, and so $\alpha_3^* = 0$.

On the other hand, if $\alpha_2^* \neq 0$, equation (1.2) would require $\alpha_1^* + \alpha_2^* x = 0$ for every $x \in \mathbb{R}$, which is impossible because it is a first degree equation, with therefore only one solution. It follows that $\alpha_2^* = 0$, and therefore (1.3) implies that also $\alpha_1^* = 0$. We have therefore shown that $\alpha_1^* = \alpha_2^* = \alpha_3^* = 0$, which contradicts the assumption that

these coefficients are not all equal to zero. We conclude that the polynomials f_1, f_2 and f_3 are linearly independent. In a similar way it is possible to show that in $\mathcal{P}_n$ the polynomials

$$1, x, x^2, ..., x^n$$

are linearly independent. ▲

Example 25 Consider in $\mathcal{P}_2$ the following polynomials:

$$f_1(x) = 1 - x, \quad f_2(x) = x(1 - x), \quad f_3(x) = 1 - x^2.$$

These vectors are linearly dependent. In fact, set $\alpha_1 = \alpha_2 = 1$ and $\alpha_3 = -1$. We have:

$$\alpha_1 f_1(x) + \alpha_2 f_2(x) + \alpha_3 f_3(x) = (1 - x) + x(1 - x) - \left(1 - x^2\right) = 0$$

for every $x \in \mathbb{R}$. Therefore, $\alpha_1 f_1 + \alpha_2 f_2 + \alpha_3 f_3 = \mathbf{0}$. ▲

In the next definition we extend the notion of linear independence to sets of vectors of any cardinality, possibly infinite.

Definition 26 *An infinite set S of vectors of a vector space V is said to be linearly independent if each finite subset of vectors of S is linearly independent. Otherwise, S is said to be linearly dependent.*

For example, let $\mathcal{P}$ be the space of all the polynomials of whatever degree. Of course,

$$\mathcal{P}_1 \subseteq \mathcal{P}_2 \subseteq \cdots \subseteq \mathcal{P}_n \subseteq \cdots$$

and $\mathcal{P} = \bigcup_{n \geq 1} \mathcal{P}_n$. The vectors

$$f_1(x) = 1, \ f_2(x) = x, \ ..., \ f_n(x) = x^{n-1}, ...$$

form a linearly independent infinite set. In fact, it is easy to verify that each finite subset of $\{f_n\}_{n \geq 1}$ is linearly independent.

1.8 Linear Combinations

Definition 27 *A vector v of a vector space V is said to be a linear combination of vectors $\{v^i\}_{i=1}^n$ of V if there exist n real coefficients $\{\alpha_i\}_{i=1}^n$ such that $v = \alpha_1 v^1 + \cdots + \alpha_n v^n$.*

Example 28 In $\mathbb{R}^3$ consider the two vectors $e^1 = (1, 0, 0)$ and $e^2 = (0, 1, 0)$. A vector of $\mathbb{R}^3$ is a linear combination of e^1 and e^2 if it has the form $(\alpha_1, \alpha_2, 0)$ for $\alpha_1, \alpha_2 \in \mathbb{R}$. In fact, $(\alpha_1, \alpha_2, 0) = \alpha_1 e^1 + \alpha_2 e^2$. ▲

Example 29 Consider in $\mathcal{P}_2$ the polynomials $f_1(x) = 1$, $f_2(x) = x$, and $f_3(x) = x^2$. Each element of $\mathcal{P}_2$ can be written as $\alpha_1 f_1 + \alpha_2 f_2 + \alpha_3 f_3$ for $\alpha_1, \alpha_2, \alpha_3 \in \mathbb{R}$. Therefore, each element of $\mathcal{P}_2$ is a linear combination of the vectors $\{f_1, f_2, f_3\}$. ▲

The notion of linear combinations allows to establish a fundamental characterization of linear dependence.

Theorem 30 *Assume $n \geq 2$. A set $\{v^i\}_{i=1}^n$ of nonzero vectors of a vector spaceV is linearly dependent if and only if for some $2 \leq k \leq n$ the vector v^k is a linear combination of the vectors $v^1, ..., v^{k-1}$.*

In other words, a set $\{v^i\}_{i=1}^n$ is linearly dependent if and only if there exists at least an element of $\{v^i\}_{i=1}^n$ that is a linear combination of some other elements of $\{v^i\}_{i=1}^n$.

Proof "Only if." Let $\{v^i\}_{i=1}^n$ be a linearly dependent set of vectors of V. Let $2 \leq k \leq n$ be the first natural number between 2 and n such that the set $\left\{v^1, ..., v^k\right\}$ is linearly dependent. In the "worst" case, k will be equal to n since by hypothesis $\{v^i\}_{i=1}^n$ is a linearly dependent set. According to the definition of linear dependence, then there exist k real coefficients $\{\alpha_i\}_{i=1}^k$, not all null, such that:

$$\alpha_1 v^1 + \alpha_2 v^2 + \cdots + \alpha_k v^k = \mathbf{0}.$$

We have $\alpha_k \neq 0$ because otherwise $\left\{v^1, ..., v^{k-1}\right\}$ would be a linearly dependent set, contradicting the fact that k is the smaller natural number between 2 and n such that $\left\{v^1, ..., v^k\right\}$ is a linearly dependent set. Having established that $\alpha_k \neq 0$, we can write:

$$v^k = \frac{-\alpha_1}{\alpha_k} v^1 + \frac{-\alpha_2}{\alpha_k} v^2 + \cdots + \frac{-\alpha_{k-1}}{\alpha_k} v^{k-1}$$

and therefore v^k is a linear combination of the vectors $\left\{v^1, ..., v^{k-1}\right\}$. This proves the "only if ".

"If." Assume that there exists, for some $2 \leq k \leq n$, a set of real coefficients $\{\alpha_i\}_{i=1}^{k-1}$ such that:

$$v^k = \alpha_1 v^1 + \cdots + \alpha_{k-1} v^{k-1}.$$

Consider the real coefficients $\{\beta_i\}_{i=1}^n$ such that:

$$\beta_i = \begin{cases} \alpha_i & 1 \leq i \leq k-1 \\ -1 & i = k \\ 0 & k+1 \leq i \leq n \end{cases}$$

By construction, $\{\beta_i\}_{i=1}^n$ is a set of real coefficients not all null and such that $\sum_{i=1}^n \beta_i v^i = \mathbf{0}$. In fact:

$$\begin{aligned}\sum_{i=1}^{n} \beta_i v^i &= \alpha_1 v^1 + \cdots + \alpha_{k-1} v^{k-1} + (-1)\, v^k + 0 v^{k+1} + \cdots + 0 v^n \\ &= v^k - v^k = \mathbf{0}.\end{aligned}$$

It follows that $\{v^i\}_{i=1}^n$ is a linearly dependent set. ■

Example 31 We saw that the polynomials $f_1(x) = 1 - x$, $f_2(x) = x(1-x)$ and $f_3(x) = 1 - x^2$ are linearly dependent. In this case each of them can be expressed as a linear combination of the other two. In fact:

$$\begin{aligned}f_1(x) &= f_3(x) - f_2(x), \\ f_2(x) &= f_3(x) - f_1(x), \\ f_3(x) &= f_1(x) + f_2(x).\end{aligned}$$

▲

Example 32 Consider in $\mathbb{R}^3$ the vectors $x^1 = (1,3,4)$, $x^2 = (2,5,1)$ and $x^3 = (0,1,7)$. We have that $x^3 = 2x^1 - x^2$, and therefore Theorem 30 can be applied to the set $\{x^1, x^2, x^3\}$ by setting $k = 3$. It follows that $\{x^1, x^2, x^3\}$ is a linearly dependent set. It is immediate to verify that also in this case each of the vectors in the set $\{x^1, x^2, x^3\}$ is a linear combination of the other two. Next example will show that this is not, however, a property of all the sets of linearly dependent vectors. ▲

Example 33 Consider in $\mathbb{R}^3$ the vectors $x^1 = (1,3,4)$, $x^2 = (2,6,8)$, and $x^3 = (2,5,1)$. We have that $x^2 = 2x^1$ and therefore Theorem 30 can be applied to the set $\{x^1, x^2, x^3\}$ by setting $k = 2$. The vectors x^1, x^2, and x^3 are therefore linearly dependent. Note that x^3 is not a linear combination of x^1 and x^2, that is, there do not exist $\alpha_1, \alpha_2 \in \mathbb{R}$ such that $x^3 = \alpha_1 x^1 + \alpha_2 x^2$. Therefore, though Theorem 30 guarantees that in a set of linearly dependent vectors some of them are a linear combination of other vectors of the set, this property does not necessarily hold for all the vectors of the set. For example, this property held for all vectors in the two previous examples, but not in this last one. ▲

Next result is an immediate, but fundamental, consequence of Theorem 30.

Corollary 34 *A set S of cardinality greater than 1, finite or infinite, of vectors of a vector space V is linearly independent if and only if none of the vectors in the set S is a linear combination of other vectors in S.*

Note that the case $n = 1$, i.e., $S = \{v\}$, has to be treated separately. On the other hand, by definition $\{v\}$ is linearly independent if and only if $v \neq \mathbf{0}$.

1.9 Generated Subspaces

Definition 35 *Let S be a subset of V. The subspace generated by S, denoted by $span\,(S)$, is the smallest vector subspace of V containing S.*

Therefore, $span\,(S)$ is the smallest vector subspace in which the set S "lives."

Proposition 36 *Let $\{W_\alpha\}$ be the collection of all vector subspaces of V containing the set S. We have that $span\,(S) = \bigcap_\alpha W_\alpha$.*

Proof It is easy to see that the proof of Proposition 18 holds in reality for any collection, finite or infinite, of vector subspaces. Therefore, $\bigcap_\alpha W_\alpha$ is itself a vector subspace of V. As $S \subseteq W_\alpha$ for each W_α, we have $span\,(S) \subseteq \bigcap_\alpha W_\alpha$ because, by definition, $span\,(S)$ is the smallest vector subspace of V containing S.

On the other hand, $span\,(S)$ belongs to the collection $\{W_\alpha\}$ because is a vector subspace of V containing S. It follows that $\bigcap_\alpha W_\alpha \subseteq span\,(S)$, and we can therefore conclude that $\bigcap_\alpha W_\alpha = span\,(S)$. ■

The intersection $\bigcap_\alpha W_\alpha$ is always nonempty since it contains at least the null vector $\mathbf{0}$. Therefore, Proposition 36 guarantees that $span\,(S)$ exists for every subset S. Moreover, $span\,(S)$ is unique as it coincides with $\bigcap_\alpha W_\alpha$.

The next important result shows that $span\,(S)$ has a "concrete" representation in terms of linear combinations of S.

Theorem 37 *Let S be a subset of V. A vector $v \in V$ belongs to $span\,(S)$ if and only if it is a linear combination of vectors of S, that is, if and only if there exists a finite set $\{v^i\}_{i\in I}$ of S and a set $\{\alpha_i\}_{i\in I}$ of real coefficients such that $v = \sum_{i\in I} \alpha_i v^i$.*

Proof "If." Let $v \in V$ be a linear combination of a finite set $\{v^i\}_{i\in I}$ of vectors of S. For simplicity, set $\{v^i\}_{i\in I} = \{v^1, ..., v^n\}$. There exists therefore a set $\{\alpha_i\}_{i=1}^n$ of real coefficients such that $v = \sum_{i=1}^n \alpha_i v^i$. By definition of vector subspace, we have $\alpha_1 v^1 + \alpha_2 v^2 \in span\,(S)$ because $v^1, v^2 \in span\,(S)$. Moreover, $(\alpha_1 v^1 + \alpha_2 v^2) \in span\,(S)$ implies $(\alpha_1 v^1 + \alpha_2 v^2) + \alpha_3 v^3 \in span\,(S)$, and proceeding in this way we get that $v = \sum_{i=1}^n \alpha_i v^i \in span\,(S)$, as desired.

"Only if". Let W be the set of all vectors v of V that can be expressed as linear combinations of vectors of S; that is, $v \in W$ if there exist finite sets $\{v^i\}_{i\in I} \subseteq S$ and $\{\alpha^i\}_{i\in I} \subseteq \mathbb{R}$ such that $v = \sum_{i=1}^n \alpha_i v^i$. It is easy to see that W is a vector subspace of V containing S. It follows that $span\,(S) \subseteq W$ and therefore every $v \in span\,(S)$ is a linear combination of vectors of S. ■

Before illustrating Theorem 37 with some examples, we state a simple consequence of this theorem.

Corollary 38 *Let S be a subset of V. The vector $v \in V$ is a linear combination of vectors of S if and only if $span(S) = span(S \cup \{v\})$.*

Example 39 Let $S = \{v^1, ..., v^k\} \subseteq V$. For Theorem 37 we have:

$$\begin{aligned} span(S) &= \left\{ v \in V : v = \sum_{i=1}^{k} \alpha_i v^i \quad \text{with } \alpha_i \in \mathbb{R} \text{ for every } i = 1, ..., k \right\} \\ &= \left\{ \sum_{i=1}^{k} \alpha_i v^i : \alpha_i \in \mathbb{R} \text{ for every } i = 1, ..., k \right\}. \end{aligned}$$

For example, consider k vectors $\{f_i\}_{i=1}^{k}$ of the vector space $\mathcal{P}_n$, with $k \leq n+1$. We have:

$$span(S) = \left\{ \sum_{i=1}^{k} \alpha_i f_i : \alpha_i \in \mathbb{R} \text{ forevery } i = 1, ..., k \right\}.$$

▲

Example 40 Let $S = \{(1,0,0), (0,1,0), (0,0,1)\} \subseteq \mathbb{R}^3$. We have:

$$\begin{aligned} & span(S) \\ = \; & \{x \in \mathbb{R}^3 : x = \alpha_1 (1,0,0) + \alpha_2 (0,1,0) + \alpha_3 (0,0,1) \\ & \text{with } \alpha_i \in \mathbb{R} \text{ for every } i = 1,2,3\} \\ = \; & \{(\alpha_1, \alpha_2, \alpha_3) : \alpha_i \in \mathbb{R} \text{ for every } i = 1,2,3\} = \mathbb{R}^3. \end{aligned}$$

More generally, let $S = \{e^1, ..., e^n\} \subseteq \mathbb{R}^n$. We have:

$$\begin{aligned} span(S) &= \left\{ x \in \mathbb{R}^n : x = \sum_{i=1}^{n} \alpha_i e^i \quad \text{with } \alpha_i \in \mathbb{R} \text{ for every } i = 1, ..., n \right\} \\ &= \{(\alpha_1, ..., \alpha_n) : \alpha_i \in \mathbb{R} \text{ forevery } i = 1, ..., n\} = \mathbb{R}^n. \end{aligned}$$

▲

Example 41 If $S = \{v\}$, we have $span(S) = \{\alpha v : \alpha \in \mathbb{R}\}$. For example, let $v = (2,3) \in \mathbb{R}^2$. We have:

$$span(S) = \{(2\alpha, 3\alpha) : \alpha \in \mathbb{R}\},$$

that is, in this case $span(S)$ it is nothing but the straight line that passes through the point v.

Fig.

▲

Example 42 Consider a financial market in which the returns of the assets depend on the state of the economy, which can be of three types:

$$s_1 = \text{“recession”} \quad s_2 = \text{“stasis”} \quad s_3 = \text{“growth”}.$$

Each asset can be described as a vector (x_1, x_2, x_3) of $\mathbb{R}^3$, in which x_i is the return of the asset in case state s_i obtains, for $i = 1, 2, 3$. Suppose that there exist only three assets on the market: x^1, x^2, and x^3. Let α_i be the quantity of asset x^i held, so that the vector of coefficients $(\alpha_1, \alpha_2, \alpha_3)$ represents a portfolio formed by these assets. The quantities α_i can be both positive and negative. In the first case we are “long” in the asset and we have the return x^i in case state s_i obtains; when α_i is negative we are instead “short” on the asset and we have to pay the return x^i when s_i obtains. The return of a portfolio $(\alpha_1, \alpha_2, \alpha_3)$ in the different states is therefore given by the linear combination

$$\alpha_1 x^1 + \alpha_2 x^2 + \alpha_3 x^3.$$

The set $span\,(x^1, x^2, x^3)$ is thus the collection of all the returns that can be earned with portfolios constituted by three assets available on the market. It is therefore a vector subspace of $\mathbb{R}^3$.

More generally, consider a financial market in which are traded n assets $\{x^i\}_{i=1}^n$,whose returns depend on k states of nature $\{s_i\}_{i=1}^k$. Each asset is then described by a vector $(x_1, ..., x_k) \in \mathbb{R}^k$, in which x_i is the return of the asset if state s_i obtains, for $i = 1, ..., k$. A portfolio in this market is represented by a vector $(\alpha_1, ..., \alpha_n)$ of real coefficients and the return of the portfolio in the different states of nature is given by the linear combination $\sum_{i=1}^n \alpha_i x^i$. It follows that $span\,(x^1, ..., x^n)$ is the set of all the returns that can be earned with portfolios of assets of this market. These returns form, therefore, a vector subspace of $\mathbb{R}^k$. ▲

1.10 Bases

Suppose S is a subset of a vector space V. By Theorem 37, $span\,(S)$ consists of the linear combinations of vectors in S. Suppose that S is a linearly dependent set. By Theorem 30 and by Corollary 34, some vectors in S can in turn be expressed as linear combinations of other elements of S. By Corollary 38, such vectors are therefore redundant for the generation of $span\,(S)$. In fact, if a vector v in $span\,(S)$ is a linear combination of vectors of S, by Corollary 38 we have $span\,(S) = span\,(S - \{v\})$, where $S - \{v\}$ is the set S without the vector v.

A linearly dependent set S thus contains some redundant elements with respect to the generation of $span\,(S)$. This does not happen if, instead, S is a linearly independent set, a case in which by Corollary 34 none of the vectors of S can be expressed as linear

combination of other elements of S. In other words, when S is linearly independent, all its vectors are essential for the generation of $span\,(S)$.

These considerations lead us to introduce the notion of basis of a vector space.

Definition 43 *A set of vectors S of V is said to be a basis of V if S is a linearly independent set such that $span\,(S) = V$.*

If S is a basis of V we have therefore:

(i) each $v \in V$ is representable as a linear combination of vectors in S;

(ii) all the vectors of S are essential for this representation, none of them is redundant.

The following result makes clear this "essentiality" of a basis for the representation as linear combinations of the elements of V.

Theorem 44 *A subset S of a vector space V is a basis of V if and only if each vector $v \in V$ can be written in a unique way as a linear combination of vectors in S.*

In other words, there is a unique finite set of coefficients $\{\alpha_i\}_{i\in I}$ and of vectors $\{v^i\}_{i\in I} \subseteq S$ such that $v = \sum_{i\in I} \alpha_i v^i$.

Proof We prove the theorem only for finite S; in the infinite case the proof is similar, though notationally tedious. "Only if." Let $S = \{v^1, ..., v^n\}$ be a basis of V. Suppose there exist two sets of real coefficients, $\{\alpha_i\}_{i=1}^n$ and $\{\beta_i\}_{i=1}^n$, such that:

$$v = \sum_{i=1}^{n} \alpha_i v^i = \sum_{i=1}^{n} \beta_i v^i.$$

We therefore have:

$$\sum_{i=1}^{n} (\alpha_i - \beta_i)\, v^i = \mathbf{0},$$

and, being the vectors in S linearly independent, this implies that $\alpha_i - \beta_i = 0$ for every $i = 1, ..., n$, that is, $\alpha_i = \beta_i$ for every $i = 1, ..., n$.

"If." Let $S = \{v^1, ..., v^n\}$ and suppose that every $v \in V$ can be written in a unique way as a linear combination of vectors in S. Clearly, by Theorem 37 we have $V = span\,(S)$. It remains to prove that S is a linearly independent set. Suppose that the real coefficients $\{\alpha_i\}_{i=1}^n$ are such that:

$$\sum_{i=1}^{n} \alpha_i v^i = \mathbf{0}.$$

Since we also have that $\sum_{i=1}^{n} 0 v^i = \mathbf{0}$, we conclude that $\alpha_i = 0$ for every $i = 1, ..., n$ since by hypothesis the vector $\mathbf{0}$ can be written in a unique way as a linear combination of vectors in S. ■

Example 45 The standard basis of $\mathbb{R}^n$ is given by the vectors $\{e^1, ..., e^n\}$. Every $x \in \mathbb{R}^n$ can be written in a unique way as linear combination of these vectors. In particular:

$$x = x_1 e^1 + \cdots + x_n e^n = \sum_{i=1}^{n} x_i e^i,$$

that is, the coefficients of the linear combination are the components of the vector x.▲

Example 46 Since the standard basis of $\mathcal{P}_n$ is given by the polynomials

$$f_1(x) = 1, f_2(x) = x, ..., f_{n+1}(x) = x^n,$$

each $f \in \mathcal{P}_n$ can be uniquely written as

$$f = \sum_{i=1}^{n+1} \alpha_i f_i$$

with $\alpha_i \in \mathbb{R}$ for $i = 1, ..., n+1$. ▲

Example 47 Since the standard basis of $\mathcal{P}$ is given by the polynomials

$$f_1(x) = 1, f_2(x) = x, ..., f_{n+1}(x) = x^n, ...$$

each polynomial $f \in \mathcal{P}$ can be uniquely written as a linear combination of vectors taken from the infinite set $\{f_i\}_{i \geq 1}$. ▲

1.11 Dimension

Each vector of a vector space V can be "reconstructed" as a linear combination of the vectors of a basis of V. In a sense, a basis is therefore a "genetic code" for a vector space, which contains all the information that is necessary to identify its elements. Since there are in general multiple bases of a given vector space, these information can thus be "summarized" through different sets of vectors.

In view of these observations, it is therefore natural to think that a vector space is the "bigger" the more elements its bases have, that is, the bigger is the quantity of information needed to identify the elements of the vector space. In this section we will to formalize this simple and natural intuition.

Let start with a definition.

Definition 48 *A vector space V is said to have finite dimension if it has a basis with a finite number of elements.*

For example, the spaces $\mathbb{R}^n$ and $\mathcal{P}_n$ have both finite dimension because for example $\{e^1, ..., e^n\}$ and $\{1, x, ..., x^n\}$ are, respectively, bases of these spaces featuring a finite number of elements.

Next theorem has an importance that is inversely proportional to the tediousness of its proof.

Theorem 49 *Let V be a finite dimensional vector space with a basis of n elements. For each linearly independent set of vectors $\left\{v^1, ..., v^k\right\}$, with $k \leq n$, there exist $n-k$ vectors $\left\{v^{k+1}, ..., v^n\right\}$ such that the set $\{v^i\}_{i=1}^n$ is a basis of V.*

Proof We prove the theorem by induction. We start therefore with $k = 1$, that is, from a singleton $\{v^1\}$. We want to show that there exist $n-1$ vectors that, when added to v^1, form a bases of V. Let $\{w^1, ..., w^n\}$ be a basis of n elements of V. There exist coefficients $\{\alpha_i^*\}_{i=1}^n \subseteq \mathbb{R}$ such that

$$v^1 = \sum_{i=1}^{n} \alpha_i^* w^i. \tag{1.4}$$

As $v^1 \neq \mathbf{0}$, not all these coefficients are zero (why is $v^1 \neq \mathbf{0}$?). Suppose, for example, that $\alpha_1^* \neq 0$. We have that:

$$w^1 = \frac{1}{\alpha_1^*} v^1 - \frac{1}{\alpha_1^*} \sum_{i=2}^{n} \alpha_1^* w^i,$$

and therefore for each set of coefficients $\{\alpha_i^*\}_{i=1}^n \subseteq \mathbb{R}$ we have:

$$\begin{aligned} \sum_{i=1}^{n} \alpha_i w^i &= \sum_{i=2}^{n} \alpha_i w^i + \alpha_1 \left[\frac{1}{\alpha_1^*} v^1 - \frac{1}{\alpha_1^*} \sum_{i=2}^{n} \alpha_1^* w^i \right] \\ &= \left(\frac{\alpha_1}{\alpha_1^*} \right) v^1 + \sum_{i=2}^{n} \left(\alpha_i - \frac{\alpha_i^*}{\alpha_1^*} \right) w^i. \end{aligned}$$

It follows that $span\,(v^1, w^2, ..., w^n) = span\,(w^1, w^2, ..., w^n)$, and so $span\,(v^1, w^2, ..., w^n) = V$.

We now show that the vectors $\{v^1, w^2, ..., w^n\}$ are linearly independent, so that we can conclude that $\{v^1, w^2, ..., w^n\}$ is a basis of V. Let $\{\beta_i\}_{i=1}^n \subseteq \mathbb{R}$ be coefficients such that

$$\beta_1 v^1 + \sum_{i=2}^{n} \beta_i w^i = \mathbf{0}. \tag{1.5}$$

We want to show that $\beta_1 = \cdots = \beta_n = 0$. Suppose $\beta_1 \neq 0$. We have:

$$v^1 = \sum_{i=2}^{n} \left(\frac{-\beta_i}{\beta_1} \right) w^i.$$

On the other hand, by (1.4) we have $v^1 = \sum_{i=1}^{n} \alpha_i^* w^i$, and therefore:

$$\alpha_1^* = 0 \quad \text{and} \quad \alpha_i^* = \frac{-\beta_i}{\beta_1} \quad \text{for every } i = 2, ..., n$$

since by Theorem 44 the vector v^1 can be uniquely written as linear combination of the basis $\{w^i\}_{i=1}^n$.

But, $\alpha_1^* = 0$ contradicts the assumption $\alpha_1^* \neq 0$, and we are therefore arrived at a contradiction. We conclude that $\beta_1 = 0$. At this point, setting $\beta_1 = 0$, (1.5) reduces to

$$\sum_{i=2}^{n} \beta_i w^i = \mathbf{0}.$$

As $\{w^2, ..., w^n\}$ is a linearly independent set (see Exercise 13.0.8), we thus have that $\beta_2 = \cdots = \beta_n = 0$, and this completes the proof that $\{v^1, w^2, ..., w^n\}$ is a linearly independent set of V and therefore a basis of it. The case $k = 1$ is therefore proved.

Suppose now that the theorem is true for every set of $k-1$ vectors; we want to show that the theorem is true for every set of k vectors. Let therefore $\left\{v^1, ..., v^k\right\}$ be a set of k linearly independent vectors. The subset $\left\{v^1, ..., v^{k-1}\right\}$ is linearly independent and it has $k-1$ elements. There exist therefore $n-(k-1)$ vectors $\left\{\widetilde{w}^k, ..., \widetilde{w}^n\right\}$ such that $\left\{v^1, ..., v^{k-1}, \widetilde{w}^k, ..., \widetilde{w}^n\right\}$ is a basis of V. Then, there exist coefficients $\{\alpha_i^*\}_{i=1}^n \subseteq \mathbb{R}$ such that

$$v^k = \sum_{i=1}^{k-1} \alpha_i^* v^i + \sum_{i=k}^{n} \alpha_i^* \widetilde{w}^i. \tag{1.6}$$

As the vectors $\left\{v^1, ..., v^{k-1}\right\}$ are linearly independent, at least one of the coefficients $\{\alpha_i^*\}_{i=1}^n$ is not zero. Otherwise, we would have $v^k = \sum_{i=1}^{k-1} \alpha_i^* v^i$, and the vector v^k would be therefore a linear combination of the vectors $\left\{v^1, ..., v^{k-1}\right\}$, something that by Corollary 34 cannot happen. Let, for example, $\alpha_k^* \neq 0$. We have:

$$\widetilde{w}^k = \frac{1}{\alpha_k^*} v^k + \sum_{i=1}^{k-1} -\frac{\alpha_i^*}{\alpha_k^*} v^i + \sum_{i=k}^{n} -\frac{\alpha_i^*}{\alpha_k^*} \widetilde{w}^i.$$

For each set of coefficients $\{\alpha_i\}_{i=1}^n \subseteq \mathbb{R}$ we have:

$$\begin{aligned}
& \sum_{i=1}^{k-1} \alpha_i v^i + \sum_{i=k}^{n} \alpha_i \widetilde{w}^i \\
= & \sum_{i=1}^{k-1} \alpha_i v^i + \alpha_k \left[\frac{1}{\alpha_k^*} v^k + \sum_{i=1}^{k-1} -\frac{\alpha_i^*}{\alpha_k^*} v^i + \sum_{i=k}^{n} -\frac{\alpha_i^*}{\alpha_k^*} \widetilde{w}^i \right] + \sum_{i=k+1}^{n} \alpha_i \widetilde{w}^i \\
= & \sum_{i=1}^{k-1} \left(\alpha_i - \frac{\alpha_k \alpha_i^*}{\alpha_k^*} \right) v^i + \frac{\alpha_k}{\alpha_k^*} v^k + \sum_{i=k+1}^{n} \left(\alpha_i - \frac{\alpha_k \alpha_i^*}{\alpha_k^*} \right) \widetilde{w}^i
\end{aligned}$$

and therefore:

$$span\left(v^1,...,v^k,\widetilde{w}^{k+1},...,\widetilde{w}^n\right)=span\left(v^1,...,v^{k-1},\widetilde{w}^k,...,\widetilde{w}^n\right)=V.$$

It remains to show that the vectors $\left\{v^1,...,v^k,\widetilde{w}^{k+1},...,\widetilde{w}^n\right\}$ are linearly independent. Let $\{\beta_i\}_{i=1}^n\subseteq\mathbb{R}$ be coefficients such that:

$$\sum_{i=1}^{k}\beta_i v^i+\sum_{i=k+1}^{n}\beta_i\widetilde{w}^i=\mathbf{0}. \tag{1.7}$$

We want to show that $\beta_1=\cdots=\beta_n=0$. Suppose $\beta_k\neq 0$. We have:

$$v^k=\sum_{i=1}^{k-1}\left(-\frac{\beta_i}{\beta_k}\right)v^i+\sum_{i=k+1}^{n}\left(-\frac{\beta_i}{\beta_k}\right)\widetilde{w}^i.$$

Being $\left\{v^1,...,v^{k-1},\widetilde{w}^k,...,\widetilde{w}^n\right\}$ a basis of V, the vector v^k can be written in a unique way as their linear combination. Therefore, (1.6) implies that

$$\alpha_i^*=-\frac{\beta_i}{\beta_k}\qquad\text{for } i=1,...,k-1 \text{ and } i=k+1,...,n,$$

while $\alpha_k^*=0$. This contradicts the previous assumption $\alpha_k^*\neq 0$, and we thus conclude that $\beta_k=0$. Equality (1.7) reduces to:

$$\sum_{i=1}^{k-1}\beta_i v^i+\sum_{i=k+1}^{n}\beta_i\widetilde{w}^i=\mathbf{0}.$$

But, the vectors $\left\{v^1,...,v^{k-1},\widetilde{w}^{k+1},...,\widetilde{w}^n\right\}$ are linearly independent (see again Exercise 13.0.8), and therefore we conclude that

$$\beta_1=\cdots=\beta_{k-1}=\beta_{k+1}=\cdots=\beta_n=0.$$

This shows that the vectors $\left\{v^1,...,v^k,\widetilde{w}^k,...,\widetilde{w}^n\right\}$ are linearly independent and form therefore a basis of V. ■

Next result is a simple, but important, consequence of Theorem 49.

Corollary 50 *Let V be a finite dimensional vector space with a basis of n elements. We have:*

(i) Each linearly independent set of V that has n elements is a basis of V.

(ii) Each linearly independent set of V has at most n elements.

Proof (i) It is sufficient to set $k = n$ in Theorem 49.

(ii) Let S be a linearly independent set in V. We consider first the case of S finite, say $S = \left\{v^1, ..., v^k\right\}$. We want to show that $k \leq n$. By contradiction, suppose $k > n$. Then, $\{v^1, ..., v^n\}$ is itself a linearly independent set in V (see Exercise 13.0.8) and from part (i) is a basis of V. Therefore, the vectors $\left\{v^{n+1}, ..., v^k\right\}$ are a linear combination of the vectors $\{v^1, ..., v^n\}$, which by Corollary 34 contradicts the fact that the vectors $\left\{v^1, ..., v^k\right\}$ are linearly independent. Therefore, $k \leq n$, and this completes the proof for S finite.

Suppose now that S is infinite. By Definition 26, each of its finite subset is linearly independent. Therefore, for what it has been just proved it can have at most n elements. But, a set whose finite subsets can have at most n elements must have itself at most n elements. It follows that S has at most n elements and the proof is complete. ■

We finally arrive at the main result of the section.

Theorem 51 *In a finite dimensional vector space V, each basis has the same number of elements.*

In other words, though the "genetic" information of a vector space can be coded through different sets of vectors, that is, through different bases, all these sets have the same number of elements, the same "magnitude."

Proof Suppose that V has a basis of n elements. By part (ii) of Corollary 50, each other basis of V can have at most n elements. Let $\left\{v^1, ..., v^k\right\}$ be any another basis of V. We show that it cannot hold $k < n$, so that $k = n$. Suppose $k < n$ holds. By Theorem 49, there would exist $n - k$ vectors $\left\{v^{k+1}, ..., v^n\right\}$ such that the set $\left\{v^1, ..., v^k, v^{k+1}, ..., v^n\right\}$ would be a basis of V. This, however, contradicts the assumption that $\left\{v^1, ..., v^k\right\}$ is a basis of V since the vectors $\left\{v^{k+1}, ..., v^n\right\}$ are not linear combination of the vectors $\left\{v^1, ..., v^k\right\}$, being $\{v^1, ..., v^n\}$ a linearly independent set. In conclusion, it cannot hold $k < n$, and so $k = n$. ■

Theorem 51 motivates the following fundamental definition.

Definition 52 *The dimension of a finite dimensional vector space V is the number of elements of a basis of V.*

By Theorem 51 this number is unique. We denote it by $\dim(V)$.

Example 53 We have $\dim(\mathbb{R}^n) = n$ and $\dim(\mathcal{P}_n) = n + 1$. ▲

Example 54 The space $\mathcal{P}$ is not finite dimensional. In fact, suppose on the contrary that $\dim(\mathcal{P}) = n$ for some $n \in \mathbb{N}$. By Theorem 51, each basis of $\mathcal{P}$ has n elements, which is not possible because we saw that the infinite set $\{1, x, ..., x^n, ...\}$ is a basis of $\mathcal{P}$. A space that, like $\mathcal{P}$, has not finite dimension is said to be infinite dimensional. Therefore, $\mathcal{P}$ is a first example of an infinite dimensional vector space. ▲

Example 55 If $V = \{\mathbf{0}\}$, that is, if V is the trivial vector space constituted only by the neutral element $\mathbf{0}$, we set $\dim(V) = 0$. Observe that V does not contain linearly independent vectors (why ?) and therefore has as basis the empty set $\{\emptyset\}$. ▲

Chapter 2

Linear Functionals

In the previous sections we studied in detail the structure of a vector space. In this section we move to the study linear functionals, a first important family of "inhabitants" of vector spaces.

Definition 56 *A function $L : V \to \mathbb{R}$ with real values defined on a vector space V is called functional. A functional $L : V \to \mathbb{R}$ is linear if*

$$L(\alpha v + \beta w) = \alpha L(v) + \beta L(w) \tag{2.1}$$

for every $v, w \in V$ and every $\alpha, \beta \in \mathbb{R}$.

Example 57 Consider $\mathbb{R}^n$. Given two vectors $x, y \in \mathbb{R}^n$, their scalar (or inner) product $x \cdot y$ is defined as $x \cdot y = \sum_{i=1}^{n} x_i y_i$. Using inner products it is easy to define linear functionals. In fact, given a vector $\chi \in \mathbb{R}^n$, define $L : \mathbb{R}^n \to \mathbb{R}$ by $L(x) = \chi \cdot x$ for each $x \in \mathbb{R}^n$. The functional L is linear:

$$\begin{aligned} L(\alpha x + \beta y) &= \chi \cdot (\alpha x + \beta y) = \sum_{i=1}^{n} \chi_i (\alpha x_i + \beta y_i) = \alpha \sum_{i=1}^{n} \chi_i x_i + \beta \sum_{i=1}^{n} \chi_i y_i \\ &= \alpha (\chi \cdot x) + \beta (\chi \cdot y) = \alpha L(x) + \beta L(y) \end{aligned}$$

for every $x, y \in \mathbb{R}^n$ and every $\alpha, \beta \in \mathbb{R}$. ▲

Example 58 Consider $\mathcal{P}$, the vector space of all polynomials defined on $\mathbb{R}$. Fix $r \in \mathbb{R}$ and define $L : \mathcal{P} \to \mathbb{R}$ as follows:

$$L(f) = f(r) \qquad \text{for every } f \in \mathcal{P}.$$

For example, if $f(x) = x^2$ and $r = 3$, we have $L(f) = r^2 = 9$. The functional L is linear:

$$L(\alpha f + \beta g) = (\alpha f + \beta g)(r) = \alpha f(r) + \beta g(r) = \alpha L(f) + \beta L(g)$$

for every $f, g \in \mathcal{P}$ and every $\alpha, \beta \in \mathbb{R}$. ▲

We now give a fundamental characterization of linear functionals, which shows that a functional is linear if and only if it preserves the operations of sum and scalar multiplication.

Proposition 59 *A functional $L : V \to \mathbb{R}$ is linear if and only if*

$$L(v+w) = L(v) + L(w), \tag{2.2}$$
$$L(\alpha v) = \alpha L(v) \tag{2.3}$$

for every $v, w \in V$ and every $\alpha \in \mathbb{R}$.

Proof "If." Suppose that (2.2) and (2.3) hold. Then,

$$L(\alpha v + \beta w) = L(\alpha v) + L(\beta w) = \alpha L(v) + \beta L(w),$$

and therefore L is a linear functional.

"Only if." Let L be a linear functional. If in (2.1) we set $\alpha = \beta = 1$, we get (2.2). If, instead, in (2.1) we set $\beta = 0$ we get (2.3). ■

Before considering other examples, we give an important property of linear functionals.

Proposition 60 *Let $L : V \to \mathbb{R}$ be a linear functional. We have $L(\mathbf{0}) = 0$ and*

$$L\left(\sum_{i=1}^{n} \alpha_i v^i\right) = \sum_{i=1}^{n} \alpha_i L\left(v^i\right) \tag{2.4}$$

for each set of vectors $\{v^i\}_{i=1}^n$ in V and each set of real numbers $\{\alpha_i\}_{i=1}^n$.

Proof We show that $L(\mathbf{0}) = 0$. By (2.3), we have $L(\alpha\mathbf{0}) = \alpha L(\mathbf{0})$ for every $\alpha \in \mathbb{R}$. Since $\alpha\mathbf{0} = \mathbf{0}$, we therefore have $L(\mathbf{0}) = \alpha L(\mathbf{0})$ for every $\alpha \in \mathbb{R}$, and this can happen only if $L(\mathbf{0}) = 0$. We leave the proof of (2.4) as an exercise. ■

Property (2.4) has a simple, but important consequence: once we know which values a linear functional takes on the elements of a basis, we can determine the values that the functional assumes in correspondence of all vectors of the vector space. In fact, let S be a basis of V, finite or infinite. Each vector $v \in V$ can be written as linear combination of elements of S, that is, there exist a finite set of vectors $\{v^i\}_{i\in I}$ in S and of real coefficients $\{\alpha_i\}_{i\in I}$ such that:

$$v = \sum_{i\in I} \alpha_i v^i.$$

By property (2.4) of Proposition 60, we have

$$L(v) = \sum_{i \in I} \alpha_i L\left(v^i\right),$$

and this means that, once we know the values $\{L(v) : v \in S\}$, we can determine all the values $L(v)$ of the vectors $v \in V$ by exploiting the linearity of the functional L.

We continue with the examples.

Example 61 We saw in Example 42 a market in which there are n assets $\{x^i\}_{i=1}^n$, whose returns depend on k states of nature $\{s_i\}_{i=1}^k$. Each asset can be represented as a vector of $\mathbb{R}^k$ and the set of the returns of the portfolios that can be formed in this market is given by $span(x^1, ..., x^n)$, the vector subspace of $\mathbb{R}^k$ generated by the set $\{x^i\}_{i=1}^n$. Each portfolio, and in particular each asset, has a price at which is traded on the market. Therefore, to each vector $v \in span(x^1, ..., x^n)$ is associated a real number that represents its market price. In other words, there exists a functional $L : span(x^1, ..., x^n) \to \mathbb{R}$, called price functional, in which $L(v)$ is the market price of the portfolio $v \in span(x^1, ..., x^n)$. We often assume that in a financial market arbitrages cannot exist, and this implies that the price functional is linear.[1] By (2.4), the linearity of the price functional implies that the portfolios $\sum_{i=1}^n \alpha_i x^i$ have a price $\sum_{i=1}^n \alpha_i L(x^i)$. Hence, it is sufficient to know the price of the assets $\{x^i\}_{i=1}^n$ in order to determine the price of all the portfolios that can be formed with them. This is a simple, but important, consequence of the hypothesis of no-arbitrages. ▲

Example 62 Consider again the vector space $\mathcal{P}$. Let $\{r_i\}_{i=0}^\infty$ be an arbitrary set of infinite real numbers (e.g., $r_i = 2^i$ for $i \geq 0$). Since each element f of $\mathcal{P}$ has the form $\sum_{i=0}^n a_i x^i$ with $n \in \mathbb{N}$, we can define $L : \mathcal{P} \to \mathbb{R}$ as follows:

$$L(f) = \sum_{i=0}^n a_i r_i \quad \text{for each} \quad f \in \mathcal{P}.$$

This functional is linear. In fact, let $f(x) = \sum_{i=0}^n a_i x^i$ and $g(x) = \sum_{i=0}^m b_i x^i$ be two elements of $\mathcal{P}$. Without loss of generality, suppose that $m \leq n$, so that

$$(f+g)(x) = \sum_{i=0}^m (a_i + b_i) x^i + \sum_{i=m+1}^n a_i x^i,$$

(of course the second sum is superfluous if $m = n$). We therefore have that:

$$\begin{aligned} L(\alpha f + \beta g) &= \sum_{i=0}^m (\alpha a_i + \beta b_i) r_i + \sum_{i=m+1}^n \alpha a_i r_i \\ &= \alpha \sum_{i=0}^n a_i r_i + \beta \sum_{i=0}^m b_i r_i = \alpha L(f) + \beta L(g). \end{aligned}$$

[1] See, for instance, H. Varian, "The arbitrage principle in financial economics", *Journal of Economic Perspectives* 1, pp. 55-72, 1987.

▲

2.1 Dual Spaces

Definition 63 *The set of all linear functionals $L : V \to \mathbb{R}$ defined on a vector space V is called dual space of V and is denoted by V'.*

The space V' is therefore the set of all linear functionals defined on the vector space V. In V' it is possible to define in a natural way sum and scalar multiplication. In fact:

(i) If $L_1, L_2 \in V'$, the sum $L_1 + L_2$ is the element of V' defined as:

$$(L_1 + L_2)(v) = L_1(v) + L_2(v) \tag{2.5}$$

for every $v \in V$.

(ii) If $L \in V'$ and $\alpha \in \mathbb{R}$, the scalar multiplication αL is the element of V' defined as:

$$(\alpha L)(v) = \alpha L(v) \tag{2.6}$$

for every $v \in V$ and every $\alpha \in \mathbb{R}$.

Endowed with these two operations, V' becomes a vector space. We state now this important property, whose simple proof is left as an exercise.

Proposition 64 *The dual space V' of a vector space V is itself a vector space with respect to the operations of sum and scalar multiplication defined in (2.5) and (2.6), and with neutral element the linear functional $\mathbf{0} : V \to \mathbb{R}$ such that $\mathbf{0}(v) = 0$ for every $v \in V$.*

Given a vector space V, it is not always easy to describe its dual V', that is, to say which form have the elements of V'. Fortunately, this is possible for some important vector spaces. For example, consider $\mathbb{R}^n$. We have seen how each vector $\chi \in \mathbb{R}^n$ induces a linear functional $L : V \to \mathbb{R}$ defined by $L(x) = \chi \cdot x$ for every $x \in \mathbb{R}^n$. Next we show that actually all linear functionals defined on $\mathbb{R}^n$ have this form; that is, the dual space $(\mathbb{R}^n)'$ is constituted by the linear functionals of the form $L(x) = \chi \cdot x$ for some $\chi \in \mathbb{R}^n$.

Theorem 65 *A functional $L : \mathbb{R}^n \to \mathbb{R}$ is linear if and only if there exists a vector $\chi \in \mathbb{R}^n$ such that $L(x) = \chi \cdot x$ for every $x \in \mathbb{R}^n$. In particular, such vector is unique.*

Proof We have already seen the "if" part in Example 1. It remains to show the "only if" part. Let $L : \mathbb{R}^n \to \mathbb{R}$ be a linear functional and consider the standard basis $\{e^1, ..., e^n\}$. Set $\chi = (L(e^1), ..., L(e^n))$. For each vector $x \in \mathbb{R}^n$ we have $x = \sum_{i=1}^n x_i e^i$, and so:

$$L(x) = L\left(\sum_{i=1}^n x_i e^i\right) = \sum_{i=1}^n x_i L\left(e^i\right) = \sum_{i=1}^n \chi_i x_i = \chi \cdot x$$

for every $x \in \mathbb{R}^n$.

As to uniqueness, let $\chi' \in \mathbb{R}^n$ be a vector such that $L(x) = \chi' \cdot x$ for every $x \in \mathbb{R}^n$. Then, for each $i = 1, ..., n$ we have:

$$\chi'_i = \chi' \cdot e^i = L(x) = \chi \cdot e^i = \chi_i,$$

and therefore $\chi' = \chi$. This completes the proof. ▲

Example 66 Consider again a financial market in which the assets can be represented as vectors of $\mathbb{R}^k$. Suppose that on the market the assets $\{e^i\}_{i=1}^k$ are available (the asset e^i pays 1 euro if the state of nature s_i obtains, and 0 otherwise). These assets are called "Arrow securities" and of course they are nothing else that the standard basis of $\mathbb{R}^k$, which here represents the set of the returns of the portfolios that can be formed on this financial market.

Let $\pi = \left(L(e^1), ..., L\left(e^k\right)\right)$ be the vector of prices of Arrow securities assigned by a given linear functional of price $L : \mathbb{R}^k \to \mathbb{R}$. Assume that on this financial market there is no arbitrage and that, therefore, the price functional L is linear. From the proof of Theorem 65 we know that:

$$L(x) = \sum_{i=1}^k \pi_i x_i = \pi \cdot x \qquad \text{for every } x \in \mathbb{R}^k.$$

The price of every portfolio is therefore determined by the price of Arrow securities. In other words, it is enough to know the price of these k "fundamental" assets in order to determine the price of all infinite portfolios that can be formed on the market, whose set is given by the whole $\mathbb{R}^k$. This is another important consequence of the hypothesis of no-arbitrage. ▲

We saw that V' is itself a vector space. It is natural to ask if, for example, it has finite dimension when V has finite dimension and, in this case, which is the relation between the dimensions of the vector spaces V and V'. To satisfy these curiosities we first have to prove the following result:

Proposition 67 *Let V be a finite dimensional vector space with basis $\{v^1, ..., v^n\}$, and let $\{r_1, ..., r_n\}$ be a set of n real numbers. There exists one and only one linear functional $L : V \to \mathbb{R}$ such that:*

$$L\left(v^i\right) = r_i \qquad \text{for every } i = 1, ..., n.$$

Proof Since $\{v^1, ..., v^n\}$ is a basis, for each $v \in V$ there exists a unique set of real coefficients $\{\alpha_i\}_{i=1}^n$ such that $v = \sum_{i=1}^n \alpha_i v^i$. Define $L : V \to \mathbb{R}$ as follows:

$$L(v) = \sum_{i=1}^{n} \alpha_i r_i \qquad \text{for every } v \in V.$$

It is easy to verify that the functional $L : V \to \mathbb{R}$ defined in this way is linear. In fact, let $v, w \in V$ be such that $v = \sum_{i=1}^n \alpha_i v^i$ and $w = \sum_{i=1}^n \beta_i v^i$. For every $\alpha, \beta \in \mathbb{R}$ we have:

$$\begin{aligned} L(\alpha v + \beta w) &= L\left(\alpha \sum_{i=1}^{n} \alpha_i v^i + \beta \sum_{i=1}^{n} \beta_i v^i\right) = L\left(\sum_{i=1}^{n} (\alpha\alpha_i) v^i + \sum_{i=1}^{n} (\beta\beta_i) v^i\right) \\ &= L\left(\sum_{i=1}^{n} (\alpha\alpha_i + \beta\beta_i) v^i\right) = \sum_{i=1}^{n} (\alpha\alpha_i + \beta\beta_i) r_i \\ &= \alpha \sum_{i=1}^{n} \alpha_i r_i + \beta \sum_{i=1}^{n} \beta_i r_i = \alpha L(v) + \beta L(w) \end{aligned}$$

and therefore $L : V \to \mathbb{R}$ is a linear functional. Since the functional L is such that $L(v^i) = r_i$ for each $i = 1, ..., n$, to complete the proof we still have to prove its uniqueness. Let $L' : V \to \mathbb{R}$ be another linear functional such that $L'(v^i) = r_i$ for each $i = 1, ..., n$. For every $v \in V$ we have:

$$L'(v) = L'\left(\sum_{i=1}^{n} \alpha_i v^i\right) = \sum_{i=1}^{n} \alpha_i L'\left(v^i\right) = \sum_{i=1}^{n} \alpha_i r_i = \sum_{i=1}^{n} \alpha_i L\left(v^i\right) = L\left(\sum_{i=1}^{n} \alpha_i v^i\right) = L(v)$$

and so $L(v) = L'(v)$ for every $v \in V$. There exists therefore one and only one linear functional $L : V \to \mathbb{R}$ such that $L(v^i) = r_i$ for every $i = 1, ..., n$. ■

Using Proposition 67, we can now study the relations between the dimensions of V and V'. To ease notation, we will use the "delta of Kronecker":

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

Theorem 68 *Let V be a finite dimensional vector space with basis $\{v^1, ..., v^n\}$. Then,*

$$\dim(V) = \dim(V') = n$$

and a basis of V' is given by the set of linear functionals $\{L_1, ..., L_n\}$ defined by:

$$L_i\left(v^j\right) = \delta_{ij}$$

for every $i = 1, ..., n$ and every $j = 1, ..., n$.

In other words, the functional $L_i : V \to \mathbb{R}$ assumes the following values on the basis $\{v^1, ..., v^n\}$:

$$\begin{aligned} L_i\left(v^i\right) &= 1, \\ L_i\left(v^j\right) &= 0 \qquad \text{if } i \neq j. \end{aligned}$$

By Proposition 67, such functional $L_i : V \to \mathbb{R}$ exists and is unique for every $i = 1, ..., n$.

Proof As already observed, the linear functionals $\{L_1, ..., L_n\}$ exist and are unique. It remains to show that $\{L_1, ..., L_n\}$ is indeed a basis of V'. Begin by show that it is a linearly independent set. Let $\{\alpha_i\}_{i=1}^n$ be a set of real coefficients such that $\sum_{i=1}^n \alpha_i L_i = \mathbf{0}$, where $\mathbf{0}$ is the linear functional on V such that $\mathbf{0}(v) = 0$ for every $v \in V$ (it is the neutral element according to Proposition 64). We therefore have that $\sum_{i=1}^n \alpha_i L_i(v) = 0$ for every $v \in V$. In particular, for the vectors of the basis $\{v^1, ..., v^n\}$ we have:

$$\begin{gathered} \alpha_1 L_1(v^1) + \alpha_2 L_2(v^1) + \cdots + \alpha_n L_n(v^1) = 0 \\ \alpha_1 L_1(v^2) + \alpha_2 L_2(v^2) + \cdots + \alpha_n L_n(v^2) = 0 \\ \cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots \\ \alpha_1 L_1(v^n) + \alpha_2 L_2(v^n) + \cdots + \alpha_n L_n(v^n) = 0 \end{gathered}$$

and therefore:

$$\begin{gathered} \alpha_1 \cdot 1 + \alpha_2 \cdot 0 + \cdots + \alpha_n \cdot 0 = 0 \\ \alpha_1 \cdot 0 + \alpha_2 \cdot 1 + \cdots + \alpha_n \cdot 0 = 0 \\ \cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots \\ \alpha_1 \cdot 0 + \alpha_2 \cdot 0 + \cdots + \alpha_n \cdot 1 = 0 \end{gathered}$$

that implies $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$. The functionals $\{L_1, ..., L_n\}$ are therefore linearly independent. It remains to show that $V' = span(L_1, ..., L_n)$. To do this, we need two observations:

(i) For each set $\{\alpha_i\}_{i=1}^n$ of real coefficients we have:

$$L_i\left(\alpha_1 v^1 + \cdots + \alpha_n v^n\right) = \alpha_1 L_i\left(v^1\right) + \cdots + \alpha_n L_i\left(v^n\right) = \alpha_i L_i\left(v^i\right) = L_i\left(\alpha_i v^i\right)$$

for every $i = 1, ..., n$.

(ii) For each $L \in V'$ we have:

$$L\left(v^i\right) = L\left(v^i\right) L_i\left(v^i\right)$$

for every $i = 1, ..., n$.

We can now prove that $V' = span\left(L_1, ..., L_n\right)$. Let $L \in V'$ and set $\beta_i = L\left(v^i\right)$ for every $i = 1, ..., n$. Since $\{v^1, ..., v^n\}$ is a basis, there exist real coefficients $\{\alpha_i\}_{i=1}^n$ such that $v = \sum_{i=1}^n \alpha_i v^i$ for a given $v \in V$. Thanks to observations (i) and (ii), for every $v \in V$ we have:

$$\begin{aligned} L(v) &= L\left(\sum_{i=1}^n \alpha_i v^i\right) = \sum_{i=1}^n \alpha_i L\left(v^i\right) = \sum_{i=1}^n \alpha_i \beta_i L\left(v^i\right) \\ &= \sum_{i=1}^n \beta_i L_i\left(\alpha_i v^i\right) = \sum_{i=1}^n \beta_i L_i\left(\sum_{i=1}^n \alpha_i v^i\right) = \sum_{i=1}^n \beta_i L_i(v). \end{aligned}$$

Therefore, L is a linear combination of the linear functionals $\{L_1, ..., L_n\}$. Hence, we conclude that $L \in span\left(L_1, ..., L_n\right)$. Since L was an arbitrary element of V', this proves that $V' = span\left(L_1, ..., L_n\right)$, as desired. ■

2.2 Extension of Linear Functionals

Let W be a vector subspace of V and let $L_W : W \to \mathbb{R}$ be a linear functional defined on W. The question we address in this section is whether it is in general possible to extend L_W from W to the whole space V or if, instead, L_W can remain "trapped" in the subspace W without having any extension on V. In other words, does there exist a linear functional $L : V \to \mathbb{R}$ defined on the whole space V and such that $L(v) = L_W(v)$ for every $v \in W$?

This is a problem of great importance, not only theoretical, but also for the applications. For example, suppose that W is the vector subspace of $\mathbb{R}^k$ generated by the assets traded on a financial market and suppose that the price of the portfolios of this market is given by the linear price functional $L_W : W \to \mathbb{R}$. Suppose that there exists the possibility of introducing on the market some new assets that, when added to the existing ones, would generate the whole space $\mathbb{R}^k$. If the functional L_W were not extendible to $\mathbb{R}^k$, this would mean that a priori the introduction of new assets is not compatible with the current market prices, which therefore should necessarily change with the appearance on the market of the new assets.[2]

[2] Assuming, of course, that absence of arbitrage keeps to be true in the enlarged market, so that also the new functional of price is linear.

If, instead, L_W is extendible to $\mathbb{R}^k$, the introduction of new assets does not necessarily lead to a modification of the current prices because there would exist linear price functionals $L : \mathbb{R}^k \to \mathbb{R}$ compatible with L_W, that is, such that $L(v) = L_W(v)$ for each $v \in W$. A positive answer to the question we study in this section is therefore important in this economic application because, a priori, there is no reason to think that the introduction of new assets necessarily leads to a variation in the prices of the portfolios already existing on the market.

Let us begin by considering finite dimensional spaces. In this case the extension is always possible.

Theorem 69 *Let W be a vector subspace of a finite dimensional vector space V. Each linear functional $L_W : W \to \mathbb{R}$ can be extended to V.*

Proof Let $\dim(V) = n$ and $\dim(W) = k$, with $k \leq n$. By Theorem 49, there exist $n - k$ vectors $\{v^{k+1}, ..., v^n\}$ such that the whole set $\{v^1, ..., v^n\}$ is a basis of V. Let $\{r^{k+1}, ..., r^n\}$ be a set of $n-k$ real numbers and let $L : V \to \mathbb{R}$ be the linear functional on V defined by:

$$L(v^i) = \begin{cases} L_W(v^i) & \text{for } i = 1, ..., k \\ r_i & \text{for } i = k+1, ..., n. \end{cases}$$

By Proposition 67, this linear functional $L : V \to \mathbb{R}$ exists and is unique. Furthermore, being $\{v^1, ..., v^k\}$ a basis of the subspace W, for each $v \in W$ there exist k real coefficients $\{\alpha_i\}_{i=1}^k$ such that $v = \sum_{i=1}^k \alpha_i v^i$. It follows that for each $v \in W$ we have:

$$\begin{aligned} L(v) &= L\left(\sum_{i=1}^k \alpha_i v^i\right) = \sum_{i=1}^k \alpha_i L(v^i) = \sum_{i=1}^k \alpha_i L_W(v^i) \\ &= L_W\left(\sum_{i=1}^k \alpha_i v^i\right) = L_W(v). \end{aligned}$$

Therefore, $L : V \to \mathbb{R}$ extends to V the linear functional $L_W : W \to \mathbb{R}$. ■

Example 70 We consider $\mathbb{R}^3$ and the vector subspace

$$W = \{(x_1, x_2, 0) : x_1, x_2 \in \mathbb{R}\}$$

generated by the vectors e^1 and e^2. By Theorem 69, each linear functional $L_W : W \to \mathbb{R}$ can be extended to a linear functional $L : \mathbb{R}^3 \to \mathbb{R}$, that is, there exists a linear functional $L : \mathbb{R}^3 \to \mathbb{R}$ such that $L(x) = L_W(x)$ for each $x \in W$. For example, let $L_W : W \to \mathbb{R}$ be defined by: $L_W(x) = x_1 + x_2$ for each $x \in W$. A possible extension of L_W on $\mathbb{R}^3$ is given by the functional $L : \mathbb{R}^3 \to \mathbb{R}$ defined by $L(x) = x_1 + x_2 + x_3$ for each $x \in \mathbb{R}^3$. ▲

As it is clear from the proof of Theorem 69, the extension is far from being unique. For example, a different extension is associated to each set $\{r_i\}_{i=k+1}^n$. This lack of uniqueness of the extension can be described in a more precise way by using the following result.

Proposition 71 *Let W be a vector subspace of a finite dimensional vector space V. There exists a vector subspace $\widetilde{W}$ of V, called complement space of W, such that:*

(i) $W \cap \widetilde{W} = \{\mathbf{0}\}$,

(ii) $W + \widetilde{W} = V$,

(iii) $dim(W) + \dim\left(\widetilde{W}\right) = \dim(V)$.

Proof Let $\dim(V) = n$ and $\dim(W) = k$, with $k \leq n$. Let $\left\{v^1, ..., v^k\right\}$ be a basis of W. By Theorem 49, there exist $n-k$ vectors $\left\{v^{k+1}, ..., v^n\right\}$ such that the whole set $\{v^1, ..., v^n\}$ is a basis of V. These $n-k$ vectors are not in W because, being $\{v^1, ..., v^n\}$ a linearly independent set, the vectors $\left\{v^{k+1}, ..., v^n\right\}$ are not linear combination of the vectors $\left\{v^1, ..., v^k\right\}$. Let $\widetilde{W} = span\left(v^{k+1}, ..., v^n\right)$. If $v \in \widetilde{W}$ and $v \neq \mathbf{0}$, then $v \notin W$. In fact, there exist $n-k$ real coefficients $\{\alpha_i^*\}_{i=k+1}^n$, not all zero, such that $v = \sum_{i=k+1}^n \alpha_i^* v^i$. Since $\{v^1, ..., v^n\}$ is a basis of V, $\sum_{i=k+1}^n \alpha_i^* v^i$ is also the only way in which the vector v can be written as linear combination of the vectors $\{v^1, ..., v^n\}$. Therefore, $v = \sum_{i=1}^n \alpha_i v^i$ if and only if $\alpha_1 = \cdots = \alpha_k = 0$ and $\alpha_i = \alpha_i^*$ for $i = k+1, ..., n$. It follows that $v \notin W$ and therefore we can conclude that $\mathbf{0} \neq v \in \widetilde{W}$ implies $v \notin W$, which implies $W \cap \widetilde{W} = \{\mathbf{0}\}$. The easy proof of points (ii) and (iii) is left to the reader. ■

Thanks to Proposition 71, we can state a more complete version of Theorem 69, in which the lack of uniqueness of the extension is clear.

Corollary 72 *Let W be a vector subspace of a finite dimensional vector space V, and let $L_W : W \to \mathbb{R}$ be a linear functional defined on W. For each basis $\{\widetilde{w}_i\}_{i \in I}$ of the complement space $\widetilde{W}$ of W and for each set of real numbers $\{r_i\}_{i \in I}$, there exists one and only one linear functional $L : V \to \mathbb{R}$ such that:*

(i) $L(v) = L_W(v)$ *for every* $v \in W$,

(ii) $L(\widetilde{w}_i) = r_i$ *for every* $i \in I$.

Proof It is enough to observe that the vectors $\left\{v^{k+1}, ..., v^n\right\}$ used in the proof of Theorem 69 are a basis of $\widetilde{W}$, as it should be clear from the proof of Lemma 71. ■

Example 73 Consider again $\mathbb{R}^3$ and its subspace

$$W = \{(x_1, x_2, 0) : x_1, x_2 \in \mathbb{R}\}.$$

The complement space $\widetilde{W}$ is $\{(0, 0, x_3) : x_3 \in \mathbb{R}\}$ and a basis is given by the singleton $\{e^3\}$. By Corollary 72, given a real number r, there exists a linear functional $L : \mathbb{R}^3 \to \mathbb{R}$ that extends L_W on V and such that $L(e^3) = r$. In particular, for each $x \in \mathbb{R}^3$ we have:

$$L(x) = x_1 L_W(e^1) + x_2 L_W(e^2) + x_3 r.$$

▲

Example 74 Let us go back to the example of the financial market. Let $\{\widetilde{w}_i\}_{i \in I}$ be the assets that will be introduced on the market and assume that together with the existing assets they form a basis of $\mathbb{R}^k$. As $\widetilde{W} = span(\{\widetilde{w}_i\}_{i \in I})$, Corollary 72 guarantees that, given a set of possible future prices $\{p_i\}_{i \in I}$ of these new assets, there exists a linear price functional $L : \mathbb{R}^k \to \mathbb{R}$ such that $L(\widetilde{w}_i) = p_i$ for each $i \in I$ and $L(v) = L_W(v)$ for each $v \in W$.

The functional of price L is therefore compatible with the current prices of the portfolios in W, given by the linear price functional $L_W : W \to \mathbb{R}$. Therefore, introduction of new assets does not necessarily modify the current prices. ▲

We consider now the general case, in which V has not necessarily finite dimension. To treat this case we must introduce sublinear functionals.

Definition 75 *A functional $L : V \to \mathbb{R}$ defined on a vector space V is sublinear if:*

(i) $L(\alpha v) = \alpha L(v)$ *for every $\alpha \geq 0$ and every $v \in V$,*

(ii) $L(v + w) \leq L(v) + L(w)$ *for every $v, w \in V$.*

Example 76 Consider in $\mathbb{R}^n$ the functional $L : \mathbb{R}^n \to \mathbb{R}$ defined by $L(x) = \|x\| = \sqrt{\sum_{i=1}^n x_i^2}$ for every $x \in \mathbb{R}^n$. It is easy to verify that L is sublinear. ▲

We can now state the celebrated Hahn-Banach Theorem, whose proof is omitted.

Theorem 77 (Hahn-Banach) *Let W be a vector subspace of a vector space V and let $L_W : W \to \mathbb{R}$ be a linear functional defined on W. There exists an extension $L : V \to \mathbb{R}$ of L_W on the whole space V if and only if there exists a sublinear functional $L^* : V \to \mathbb{R}$ such that $L_W(v) \leq L^*(v)$ for every $v \in W$. Moreover, for such an extension $L : V \to \mathbb{R}$ we have $L(v) \leq L^*(v)$ for every $v \in V$.*

The Hahn-Banach Theorem therefore guarantees the existence of an extension, provided a condition is met: in the vector space V there must exist a sublinear functional $L^* : V \to \mathbb{R}$ that on W "dominates" the functional L_W.

Example 78 Consider the vector subspace $\mathcal{P}_n$ of $\mathcal{P}$. Let $L^* : \mathcal{P} \to \mathbb{R}$ be the functional defined by $L^*(f) = \max_{r\in[0,1]} f(r)$ for every $f \in \mathcal{P}$. By the Weierstrass Theorem, the functional L^* is well defined and it is easy to verify that it is sublinear. Therefore, by Hahn-Banach Theorem each linear functional $L_n : \mathcal{P}_n \to \mathbb{R}$ such that $L_n(f) \leq L^*(f)$ for every $f \in \mathcal{P}_n$ can be extended on the whole space $\mathcal{P}$, that is, there exists a linear functional $L : \mathcal{P} \to \mathbb{R}$ such that $L_n(f) = L(f)$ for every $f \in \mathcal{P}_n$.

Moreover, we have $L(f) \leq L^*(f)$ for every $f \in \mathcal{P}$. For example, it is easy to verify that all this holds for the linear functionals $L_n : \mathcal{P}_n \to \mathbb{R}$ defined by $L_n(f) = f(r)$ for every $f \in \mathcal{P}_n$, where r is a given real number belonging to $[0, 1]$. ▲

Chapter 3

Linear Applications

3.1 Definition and First Properties

Definition 79 *A function $T : V_1 \to V_2$ defined on a vector space V_1 and with values in a vector space V_2 is called application. An application $T : V_1 \to V_2$ is linear if*

$$T(\alpha v + \beta w) = \alpha T(v) + \beta T(w) \tag{3.1}$$

for every $v, w \in V_1$ and every $\alpha, \beta \in \mathbb{R}$.

The notion of linear application generalizes that of linear functional, which is the special case where V_2 is the real line $\mathbb{R}$. Before considering some examples, we show that an application is linear if and only if it preserves the operations of sum and scalar multiplication between the two spaces. We omit the proof, which is similar to that of Proposition 59.

Proposition 80 *An application $T : V_1 \to V_2$ is linear if and only if*

$$T(v + w) = T(v) + T(w) \text{ and} \tag{3.2}$$

$$T(\alpha v) = \alpha T(v), \tag{3.3}$$

for every $v, w \in V_1$ and every $\alpha \in \mathbb{R}$.

We give few examples of linear applications..

Example 81 Let $A = (a_{ij})$ be a matrix $m \times n$. Given a vector $x \in \mathbb{R}^n$, set

$$Ax = \left(\sum_{k=1}^{n} a_{1k} x_k, \sum_{k=1}^{n} a_{2k} x_k, ..., \sum_{k=1}^{n} a_{mk} x_k \right) \in \mathbb{R}^m. \tag{3.4}$$

For example, if $x = (1, 2, 6)$ and

$$A = \begin{bmatrix} 0 & 2 & -1 \\ 2 & 1 & 5 \\ 1 & -2 & 3 \end{bmatrix},$$

we have

$$\begin{aligned} \sum_{k=1}^{3} a_{1k}x_k &= 0 \cdot 1 + 2 \cdot 2 + (-1) \cdot 6 = -2, \\ \sum_{k=1}^{3} a_{2k}x_k &= 2 \cdot 1 + 1 \cdot 2 + 5 \cdot 6 = 34, \\ \sum_{k=1}^{3} a_{3k}x_k &= 1 \cdot 1 + (-2) \cdot 2 + 3 \cdot 6 = 15. \end{aligned}$$

Therefore,

$$Ax = (-2, 34, 15) \in \mathbb{R}^3.$$

Define the application $T : \mathbb{R}^n \to \mathbb{R}^m$ as

$$T(x) = Ax \tag{3.5}$$

for every $x \in \mathbb{R}^n$. It is easy to see that T is linear. Theorem 93 will show that all linear applications $T : \mathbb{R}^n \to \mathbb{R}^m$ have actually this form. ▲

Example 82 Consider the application $D : \mathcal{P} \to \mathcal{P}$ defined by

$$D(f) = f'$$

for every $f \in \mathcal{P}$, where f' is the derivative of f. This important application is linear because, by the properties of the derivatives, we have

$$D(\alpha f + \beta g) = (\alpha f + \beta g)' = \alpha f' + \beta g' = \alpha D(f) + \beta D(g)$$

for every $f, g \in \mathcal{P}$ and every $\alpha, \beta \in \mathbb{R}$. ▲

Example 83 Consider the application $T : \mathbb{R}^{n+1} \to \mathcal{P}_n$ defined by

$$T(\alpha) = \alpha_0 + \alpha_1 x + \cdots + \alpha_n x^n$$

for every $\alpha = (\alpha_0, \alpha_1, ..., \alpha_n) \in \mathbb{R}^{n+1}$. Also in this case it is easy to verify that T is linear. ▲

Example 84 As a final example, consider the application $0 : V_1 \to V_2$ defined as

$$0(v) = \mathbf{0}$$

for every $v \in V$. This linear application is called null application. ▲

An important special case is when $V_1 = V_2$, so that we can write $T : V \to V$. The application $T : \mathcal{P} \to \mathcal{P}$ of Example 2 has this form; let us see another example:

Example 85 Let $A = (a_{ij})$ be a square $n \times n$ matrix and, similarly to Example 81, define the application $T : \mathbb{R}^n \to \mathbb{R}^n$ by

$$T(x) = Ax$$

for every $x \in \mathbb{R}^n$. Therefore, if in Example 81 we use square $n \times n$ matrices, we have $V_1 = V_2 = \mathbb{R}^n$. ▲

Example 86 Among the linear applications $T : V \to V$, an important role is played by the identity applications $I : V \to V$, defined by

$$I(v) = v$$

for every $v \in V$. Clearly, I is a linear application. ▲

We conclude this first section with some other simple properties of linear applications, analogous to those stated in Proposition 60 for linear functionals. The simple proof is left to the reader.

Proposition 87 *Let $T : V_1 \to V_2$ be a linear application. We have $T(\mathbf{0}) = \mathbf{0}$ and*

$$T\left(\sum_{i=1}^{n} \alpha_i v^i\right) = \sum_{i=1}^{n} \alpha_i T\left(v^i\right) \tag{3.6}$$

for every set of vectors $\{v^i\}_{i=1}^n$ in V_1 and every set of real numbers $\{\alpha_i\}_{i=1}^n$.

As we have already seen for linear functionals, property (3.6) has an important consequence: once we know the values of a linear application T on the elements of a basis of V_1, we can determine the values of T on all the vectors of the vector space V_1.

3.2 Algebra of the Applications

We can define in a natural way the sum and the scalar multiplication for applications.

Definition 88 *Given two applications $S, T : V_1 \to V_2$ and a real number $\alpha \in \mathbb{R}$, define:*

$$\begin{aligned} (S+T)(v) &= S(v) + T(v) \quad \text{for every } v \in V, \\ (\alpha T)(v) &= \alpha T(v) \quad \text{for every } v \in V. \end{aligned}$$

Let $L(V_1, V_2)$ be the space of all linear applications $T : V_1 \to V_2$. In the case of linear functionals, that is, $V_2 = \mathbb{R}$, the space $L(V_1, V_2)$ is nothing but the dual space V_1', which we studied in detail in the previous chapter. Next we shows that, like dual spaces, also the space $L(V_1, V_2)$ of linear applications forms a vector space. The simple proof is left to the reader.

Proposition 89 *The space $L(V_1, V_2)$ is a vector space with respect to the operations of sum and scalar multiplication introduced in Definition 88. In particular, the neutral element is given by the null application $\mathbf{0} : V_1 \to V_2$.*

We now introduce the fundamental notion of product of applications.

Definition 90 *Given two applications $T : V_1 \to V_2$ and $S : V_2 \to V_3$, their product is the transformation $ST : V_1 \to V_3$ defined by*

$$(ST)(v) = S(T(v))$$

for every $v \in V_1$.

In other words, the product application ST is the composite function $S \circ T$. If the applications S and T are linear, then also the product ST is. In fact:

$$\begin{aligned} (ST)(\alpha v + \beta w) &= S(T(\alpha v + \beta w)) = S(\alpha T(v) + \beta T(w)) \\ &= \alpha S(T(v)) + \beta S(T(w)) = \alpha (ST)(v) + \beta (ST)(w) \end{aligned}$$

for every $v, w \in V_1$ and every $\alpha, \beta \in \mathbb{R}$. Therefore, the product of two linear applications is a new linear application.

Example 91 Consider the differential application $D : \mathcal{P} \to \mathcal{P}$ of Example 82, together with the linear application $T : \mathcal{P} \to \mathcal{P}$ defined by

$$T(f)(x) = xf(x)$$

for every $f \in \mathcal{P}$ and every $x \in \mathbb{R}$. Since in this case we have $V_1 = V_2 = V_3 = \mathcal{P}$, both products DT and TD are well defined. In particular:

$$\begin{aligned}(DT)(f)(x) &= \frac{d}{dx}(xf(x)) = f(x) + xf'(x),\\ (TD)(f)(x) &= xf'(x).\end{aligned}$$

We can observe as DT and TD are two different linear applications, i.e., $DT \neq TD$. Therefore, this simple example shows that the product of applications is not, in general, commutative. ▲

As we just saw, even when both products ST and TS are well defined, the product of applications is not in general a commutative operation and, consequently, in the notation ST it is important the order in which the applications S and T appear.

Apart from this, many of the properties of the multiplication among real numbers remain true for multiplication among applications. In particular, consider the linear applications $T : V_1 \to V_2$, $S : V_2 \to V_3$, $R : V_2 \to V_3$, and $Q : V_4 \to V_2$. The following properties can be immediately verified:

$$(S+R)T = ST + RT, \tag{3.7}$$

$$Q(S+R) = QS + QR, \tag{3.8}$$

$$(QS)T = Q(ST). \tag{3.9}$$

If we consider the identity application $I : V_2 \to V_2$, we also have $SI = S$ and $IT = T$. Finally, it is easy to see how the null application plays in the product of applications a role analogous to that of zero.

These properties take a particularly simple and compact form in the case of the linear applications $T : V \to V$. For brevity, we denote by $L(V)$ in place of $L(V, V)$ the space of such applications. Let $S, T, Q, I, 0 \in L(V)$; we have:

$$\begin{aligned}T0 &= 0T = 0,\\ TI &= IT = T,\\ (Q+S)T &= QT + ST,\\ T(Q+S) &= TQ + TS,\\ (QS)T &= Q(ST).\end{aligned}$$

These properties allow us to introduce in $L(V)$ polynomials of applications. In fact, given an application $T : V \to V$, the associative property of the product allows us to write:

$$T^m = \overbrace{TT\cdots T}^{m \text{ times}}.$$

If we set $T^0 = I$, it follows that given a polynomial $p(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_n x^n$, it is possible to construct the linear application

$$p(T) = \alpha_0 I + \alpha_1 T + \alpha_2 T^2 + \cdots + \alpha_n T^n,$$

which can be seen as a polynomial in T.

Example 92 Let $D : \mathcal{P} \to \mathcal{P}$ be the differential application. In this case $D^2 : \mathcal{P} \to \mathcal{P}$ is defined by $D^2(f) = D(D(f))$ and is therefore the second derivative; in general, $D^k : \mathcal{P} \to \mathcal{P}$ is the derivative of order k. Therefore, if we take the polynomial $p(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3$, we have

$$\begin{aligned} p(D)(f) &= \left(\alpha_0 I + \alpha_1 D + \alpha_2 D^2 + \alpha_3 D^3\right)(f) \\ &= \alpha_0 f + \alpha_1 f' + \alpha_2 f'' + \alpha_3 f'''. \end{aligned}$$

▲

Besides the lack of commutativity, another curious property of the product of applications is the existence of the so-called divisors of the zero. For example, consider the differential applications D^1 and D^2 on the space $\mathcal{P}_2$. It is easy to see that $D^2 D^1 = 0$. We therefore have a non null application, D^2, for which there exists another non null application, D^1, whose product is the null application 0. Such applications are called divisors of the zero. It is easy to see as all differential applications are actually divisors of the zero on the spaces $\mathcal{P}_n$.

3.3 Applications among Euclidean Spaces

In this section we study in more in detail applications among Euclidean spaces, that is, applications of the form $T : \mathbb{R}^n \to \mathbb{R}^m$. We start by giving a representation. In Theorem 65 we saw that a functional $L : \mathbb{R}^n \to \mathbb{R}$ is linear if and only if there exists a vector $\chi \in \mathbb{R}^n$ such that $L(x) = \chi \cdot x$ for every $x \in \mathbb{R}^n$. Next we generalize that result to linear applications.

Theorem 93 *An application $T : \mathbb{R}^n \to \mathbb{R}^m$ is linear if and only if there exists a matrix $\underset{m\times n}{A}$ such that*

$$T(x) = Ax \tag{3.10}$$

for every $x \in \mathbb{R}^n$. In particular, such matrix A is unique.

The matrix A is called the matrix associated to the application T.

Proof "If." This direction is essentially proved in Example 81. "Only if." Let T be a linear application. Set

$$A = \left[T\left(e^1\right), T\left(e^2\right), ..., T\left(e^n\right)\right],$$

i.e., A is the $m \times n$ matrix whose n columns are given by the column vectors $T\left(e^i\right)$ for $i = 1, ..., n$. The set $\{e^i\}_{i=1}^n$ is a basis of $\mathbb{R}^n$ and for every $x \in \mathbb{R}^n$ we have $x = \sum_{i=1}^n x_i e^i$. Therefore:

$$T(x) = T\left(\sum_{i=1}^n x_i e^i\right) = \sum_{i=1}^n x_i T\left(e^i\right) = Ax$$

for every $x \in \mathbb{R}^n$.

As to uniqueness, let B be a matrix $m \times n$ for which (3.10) holds. We have

$$\begin{aligned}
(a_{11}, a_{21}, ..., a_{m1}) &= T\left(e^1\right) = Be^1 = (b_{11}, b_{21}, ..., b_{m1}), \\
(a_{12}, a_{22}, ..., a_{m2}) &= T\left(e^2\right) = Be^2 = (b_{12}, b_{22}, ..., b_{m2}), \\
&\cdots\cdots\cdots\cdots \\
(a_{1n}, a_{2n}, ..., a_{mn}) &= T\left(e^n\right) = Be^n = (b_{1n}, b_{2n}, ..., b_{mn}).
\end{aligned}$$

Therefore, $A = B$. ■

Example 94 Let $T : \mathbb{R}^3 \to \mathbb{R}^3$ be defined by

$$T(x) = (0, x_2, x_3)$$

for every $x \in \mathbb{R}^3$. In other words, T is the projection of each vector in $\mathbb{R}^3$ in the plane $\{x \in \mathbb{R}^3 : x_1 = 0\}$. For example, $T(2, 3, 5) = (0, 3, 5)$. We have

$$\begin{aligned}
T\left(e^1\right) &= (0, 0, 0), \\
T\left(e^2\right) &= (0, 1, 0), \\
T\left(e^3\right) &= (0, 0, 1),
\end{aligned}$$

and therefore

$$A = \left[T\left(e^1\right), T\left(e^2\right), T\left(e^3\right)\right] = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

In conclusion, $T(x) = Ax$ for every $x \in \mathbb{R}^3$. ▲

Example 95 Let $T : \mathbb{R}^3 \to \mathbb{R}^2$ be defined by

$$T(x) = (x_1 - x_3, x_1 + x_2 + x_3)$$

for every $x \in \mathbb{R}^3$. For example, $T(2,3,5) = (-3,10)$. We have

$$\begin{aligned} T\left(e^1\right) &= (1,1), \\ T\left(e^2\right) &= (0,1), \\ T\left(e^3\right) &= (-1,1), \end{aligned}$$

and therefore

$$A = \left[T\left(e^1\right), T\left(e^2\right), T\left(e^3\right)\right] = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 1 & 1 \end{bmatrix}.$$

It is thus possible to write $T(x) = Ax$ for every $x \in \mathbb{R}^3$. ▲

3.3.1 Matrix Representation of Operations

A natural question that arises at this point is what are the representations in terms of matrices of the operations just introduced, when defined among applications in $L(\mathbb{R}^n, \mathbb{R}^m)$.

For sum and scalar multiplication we have the following simple result, whose obvious proof is omitted.

Proposition 96 *Let $S, T : \mathbb{R}^n \to \mathbb{R}^m$ be two linear applications and let $\alpha \in \mathbb{R}$. Let A and B be the two $m \times n$ matrices associated to S and T, respectively. Then, $A + B$ is the matrix associated to the application $S + T$, while αA is the matrix associated to the application αS.*

Example 97 Let $S, T : \mathbb{R}^3 \to \mathbb{R}^3$ be two linear applications defined as:

$$\begin{aligned} S(x) &= (0, x_2, x_3), \\ T(x) &= (2x_1 - x_3, x_1 + x_2 + 3x_3, 2x_1 - x_2), \end{aligned}$$

for every $x \in \mathbb{R}^3$. In Example 94 we saw that

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

is the matrix associated to the application S. By proceeding in the same way, we can see that

$$B = \begin{bmatrix} 2 & 0 & -1 \\ 1 & 1 & 3 \\ 2 & -1 & 0 \end{bmatrix}$$

is the matrix associated to the application T. By Proposition 96,

$$A+B=\begin{bmatrix} 2 & 0 & -1 \\ 1 & 2 & 3 \\ 2 & -1 & 1 \end{bmatrix}$$

is the matrix associated to the application $S+T$. Moreover, if for example we set $\alpha=10$, by Proposition 96

$$\alpha A=\begin{bmatrix} 0 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \end{bmatrix}$$

is the matrix associated to the application αS. ▲

After having considered sum and scalar multiplication, we move to the more interesting case of product of applications.

Proposition 98 *Consider two linear applications $S:\mathbb{R}^m\to\mathbb{R}^q$ and $T:\mathbb{R}^n\to\mathbb{R}^m$, whose associated matrices are, respectively,*

$$\underset{(q\times m)}{A}=(a_{ij}) \quad \text{and} \quad \underset{(m\times n)}{B}=(b_{ij}).$$

Then, the matrix associated to the product application $ST:\mathbb{R}^n\to\mathbb{R}^q$ is given by the matrix $\underset{(q\times n)}{AB}=(ab_{ij})$, whose components are defined by

$$ab_{ij}=\sum_{k=1}^{m}a_{ik}b_{kj} \tag{3.11}$$

for $i=1,...,q$ and $j=1,...,n$.

The matrix AB defined through the rule (3.11) is called *product matrix* of A and B. To understand this rule, denote by $a_{i\cdot}=(a_{i1},...,a_{im})$ the row vector i of the matrix A and by $b_{\cdot j}=(b_{1j},...,b_{mj})$ the column vector j of the matrix B. By (3.11), the component ab_{ij} of AB is nothing but the inner product of the vectors $a_{i\cdot}$ and $b_{\cdot j}$, that is, $ab_{ij}=a_{i\cdot}\cdot b_{\cdot j}$.

Proof Let $\{e^i\}_{i=1}^n$ and $\{\overline{e}^i\}_{i=1}^m$ be respectively the standard bases of $\mathbb{R}^n$ and $\mathbb{R}^m$. We have

$$\begin{aligned} T\left(e^j\right) &= Be^j=(b_{1j},b_{2j},...,b_{mj}) \\ &= b_{1j}(1,0,...,0)+b_{2j}(0,1,0,...,0)\cdots+b_{mj}(0,0,...,1) \\ &= \sum_{k=1}^{m}b_{kj}\overline{e}^k. \end{aligned}$$

Similarly, we have

$$S\left(\overline{e}^k\right) = A\overline{e}^k = (a_{1k}, ..., a_{qk}) = \sum_{i=1}^{q} a_{ik}e^i.$$

We can therefore write:

$$\begin{aligned}(ST)\left(e^j\right) &= S\left(T\left(e^j\right)\right) = S\left(\sum_{k=1}^{m} b_{kj}\overline{e}^k\right) = \sum_{k=1}^{m} b_{kj}S\left(\overline{e}^k\right) \\ &= \sum_{k=1}^{m} b_{kj}\left(\sum_{i=1}^{q} a_{ik}e^i\right) = \sum_{i=1}^{q}\left(\sum_{k=1}^{m} a_{ik}b_{kj}\right)e^i.\end{aligned}$$

On the other hand, let C be the matrix associated to the application ST. We have:

$$(ST)\left(e^j\right) = Ce^j = (c_{1j}, ..., c_{qj}) = \sum_{i=1}^{q} c_{ij}e^i.$$

Therefore, $c_{ij} = \sum_{k=1}^{m} a_{ik}b_{kj}$ and we conclude that $C = AB$. ■

Notice that the product of matrices can be only applied to two matrixes $\underset{(q\times m)}{A}$ and $\underset{(m\times n)}{B}$ such that the number of columns of A is equal to the number of rows of B.

In general the product is not commutative, which naturally reflects the non commutativity of the product of applications that we saw in the previous section. Example 100 will show a simple case in which both products AB and BA are well defined, but $AB \neq BA$.

We now illustrate with few examples this new operation among matrices.

Example 99 Let A and B be defined as:

$$\underset{(2\times 3)}{A} = \begin{bmatrix} 1 & 3 & 1 \\ 0 & 1 & 4 \end{bmatrix} \quad \text{and} \quad \underset{(3\times 4)}{B} = \begin{bmatrix} 1 & 2 & 1 & 0 \\ 2 & 5 & 2 & 2 \\ 0 & 1 & 3 & 2 \end{bmatrix}.$$

The product matrix AB is 2×4.[1] Using rule (3.11), we have

$$\begin{aligned}AB &= \begin{bmatrix} 1 & 3 & 1 \\ 0 & 1 & 4 \end{bmatrix}\begin{bmatrix} 1 & 2 & 1 & 0 \\ 2 & 5 & 2 & 2 \\ 0 & 1 & 3 & 2 \end{bmatrix} \\ &= \begin{bmatrix} 1\cdot 1+3\cdot 2+1\cdot 0 & 1\cdot 2+3\cdot 5+1\cdot 1 & 1\cdot 1+3\cdot 2+1\cdot 3 & 1\cdot 0+3\cdot 2+1\cdot 2 \\ 0\cdot 1+1\cdot 2+4\cdot 0 & 0\cdot 2+1\cdot 5+4\cdot 1 & 0\cdot 1+1\cdot 2+4\cdot 3 & 0\cdot 0+1\cdot 2+4\cdot 2 \end{bmatrix} \\ &= \begin{bmatrix} 7 & 18 & 10 & 8 \\ 2 & 9 & 14 & 10 \end{bmatrix}.\end{aligned}$$

[1] *To determine the number of rows and colums of AB, a useful trick to remember is $(2 \times 4) = (2 \times 3)(3 \times 4)$.*

▲

Example 100 Let A and B be defined as

$$A = \begin{bmatrix} 1 & 0 & 3 \\ 2 & 1 & 0 \\ 1 & 4 & 6 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 2 & 1 & 4 \\ 0 & 3 & 1 \\ 4 & 2 & 4 \end{bmatrix}.$$

Since A and B are square matrices, both BA and AB exist and they 3×3 matrices . Using rule (3.11), we have

$$\begin{aligned} BA &= \begin{bmatrix} 2 & 1 & 4 \\ 0 & 3 & 1 \\ 4 & 2 & 4 \end{bmatrix} \begin{bmatrix} 1 & 0 & 3 \\ 2 & 1 & 0 \\ 1 & 4 & 6 \end{bmatrix} \\ &= \begin{bmatrix} 2\cdot 1+1\cdot 2+4\cdot 1 & 2\cdot 0+1\cdot 1+4\cdot 4 & 2\cdot 3+1\cdot 0+4\cdot 6 \\ 0\cdot 1+3\cdot 2+1\cdot 1 & 0\cdot 0+3\cdot 1+1\cdot 4 & 0\cdot 3+3\cdot 0+1\cdot 6 \\ 4\cdot 1+2\cdot 2+4\cdot 1 & 4\cdot 0+2\cdot 1+4\cdot 4 & 4\cdot 3+2\cdot 0+4\cdot 6 \end{bmatrix} \\ &= \begin{bmatrix} 8 & 17 & 30 \\ 7 & 7 & 6 \\ 12 & 18 & 36 \end{bmatrix}, \end{aligned}$$

while,

$$\begin{aligned} AB &= \begin{bmatrix} 1 & 0 & 3 \\ 2 & 1 & 0 \\ 1 & 4 & 6 \end{bmatrix} \begin{bmatrix} 2 & 1 & 4 \\ 0 & 3 & 1 \\ 4 & 2 & 4 \end{bmatrix} \\ &= \begin{bmatrix} 1\cdot 2+0\cdot 0+3\cdot 4 & 1\cdot 1+0\cdot 3+3\cdot 2 & 1\cdot 4+0\cdot 1+3\cdot 4 \\ 2\cdot 2+1\cdot 0+0\cdot 4 & 2\cdot 1+1\cdot 3+0\cdot 2 & 2\cdot 4+1\cdot 1+0\cdot 4 \\ 1\cdot 2+4\cdot 0+6\cdot 4 & 1\cdot 1+4\cdot 3+6\cdot 2 & 1\cdot 4+4\cdot 1+6\cdot 4 \end{bmatrix} \\ &= \begin{bmatrix} 14 & 7 & 16 \\ 4 & 5 & 9 \\ 26 & 25 & 32 \end{bmatrix}. \end{aligned}$$

Notice that $AB \neq BA$. Therefore, this is an example where the product is not commutative. ▲

Example 101 Go back to Example 81, in which we associated to a $m \times n$ matrix A and to a vector $x \in \mathbb{R}^n$ the vector Ax in $\mathbb{R}^m$ defined by

$$Ax = \left(\sum_{k=1}^{n} a_{1k}x_k, \sum_{k=1}^{n} a_{2k}x_k, ..., \sum_{k=1}^{n} a_{mk}x_k \right).$$

If we consider x as a column vector $n \times 1$, it is easy to see that the vector $\underset{(m\times 1)}{Ax}$ so defined is exactly the product of the matrices

$$\underset{(m\times n)}{A} = (a_{ij}) \quad \text{and} \quad \underset{(n\times 1)}{x} = (x_{i1}) .$$

In Example 81 we had $x = (1, 2, 6)$ and

$$A = \begin{bmatrix} 0 & 2 & -1 \\ 2 & 1 & 5 \\ 1 & -2 & 3 \end{bmatrix} .$$

In view of what we just said, we can therefore write:

$$\begin{aligned} Ax &= \begin{bmatrix} 0 & 2 & -1 \\ 2 & 1 & 5 \\ 1 & -2 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 6 \end{bmatrix} \\ &= \begin{bmatrix} 0 \cdot 1 + 2 \cdot 2 + (-1) \cdot 6 \\ 2 \cdot 1 + 1 \cdot 2 + 5 \cdot 6 \\ 1 \cdot 1 + 2 \cdot (-2) + 3 \cdot 6 \end{bmatrix} = \begin{bmatrix} -2 \\ 34 \\ 15 \end{bmatrix} . \end{aligned}$$

▲

3.4 Isomorphisms

As we have seen, a fundamental characteristic of a linear application $T : V_1 \to V_2$ is to preserve the operations of sum and scalar multiplication from the space V_1 to the space V_2. Suppose now that the application T is injective and surjective, that is, suppose that T is a bijection between the two spaces V_1 and V_2. Define its *inverse* $T^{-1} : V_2 \to V_1$ as $T^{-1}(w) = v$ if and only if $T(v) = w$ for every $w \in V_2$. Since T is injective, the function T^{-1} is well defined; moreover, being T surjective, the domain of T^{-1} is the entire space V_2.

Lemma 102 *Given two vector spaces V_1 and V_2, we have $T \in L(V_1, V_2)$ if and only if $T^{-1} \in L(V_2, V_1)$.*

Proof The simple proof is left to the reader. ■

Therefore, the inverse T^{-1} is itself a linear application and, as such, it preserves the operations of sum and scalar multiplication from the space V_2 to the space V_1. We can thus say that with respect to a linear bijective application $T : V_1 \to V_2$, the operations of sum and scalar multiplication in the two spaces V_1 and V_2 are completely

interchangeable. For example, if we have to compute the sum $v, w \in V_1$, we can make the operation directly in V_1. But, we can also transfer the problem in V_2 through the images $T(v)$ and $T(w)$, and then make the operation in V_2 computing $T(v) + T(w)$, and go back in V_1 through the inverse T^{-1}, by considering $T^{-1}(T(v) + T(w))$. Since $v + w = T^{-1}(T(v) + T(w))$, also in this way we get the sum $v + w$. In a similar way, the operations in V_2 can be done in V_1, by transferring them in this space through the inverse T^{-1}.

Sometimes, this "tour" is useful also operationally because it may happen that to carry out the operations in one of the two spaces is significantly simpler than in the other one. But, what is most interesting is to observe that all this shows that two spaces V_1 and V_2 among which there exists a linear bijective application $T : V_1 \to V_2$ are mutually interchangeable with respect to the operations of sum and scalar multiplication.

All this leads us to the following definition.

Definition 103 *Two vector spaces V_1 and V_2 are called isomorphic if there exists a linear application $T : V_1 \to V_2$ that is both injective and surjective. Such application is called isomorphism.*

In light of what just observed, two isomorphic vector spaces behave in a similar way with respect to the operations of sum and scalar multiplication. That is, they are similar from the point of view of their vector structure.

We have therefore a criterion to bring some order among all different examples of vector spaces: isomorphic spaces can be viewed as belonging to the same "category." The next remarkable result shows that, for finite dimensional spaces, this classification is equivalent to the one based on dimension.

Theorem 104 *Two finite dimensional vector spaces are isomorphic if and only if they have the same dimension.*

Proof "If." Let V_1 and V_2 be two vector spaces such that $\dim(V_1) = \dim(V_2)$. Let $\{v^i\}_{i=1}^n$ and $\{\overline{v}^i\}_{i=1}^n$ be bases of V_1 and V_2, respectively. For each $v \in V_1$, there exists $\{\alpha_i\}_{i=1}^n \subseteq \mathbb{R}$ such that $v = \sum_{i=1}^n \alpha_i v^i$. We can therefore define an application $T : V_1 \to V_2$ as follows:

$$T(v) = \sum_{i=1}^{n} \alpha_i \overline{v}^i$$

for every $v \in V$. We first verify that T is linear. Let $v, w \in V$ be such that $v = \sum_{i=1}^{n} \alpha_i v^i$ and $w = \sum_{i=1}^{n} \beta_i v^i$. For every $\alpha, \beta \in \mathbb{R}$, we have:

$$\begin{aligned} T(\alpha v + \beta w) &= T\left(\alpha \sum_{i=1}^{n} \alpha_i v^i + \beta \sum_{i=1}^{n} \beta_i v^i\right) = T\left(\sum_{i=1}^{n} (\alpha\alpha_i + \beta\beta_i) v^i\right) \\ &= \sum_{i=1}^{n} (\alpha\alpha_i + \beta\beta_i) \overline{v}^i = \alpha \sum_{i=1}^{n} \alpha_i \overline{v}^i + \beta \sum_{i=1}^{n} \beta_i \overline{v}^i = \alpha T(v) + \beta T(w). \end{aligned}$$

Therefore, T is linear. It is also injective In fact, let v and w be two vectors in V with $v \neq w$ and with $v = \sum_{i=1}^{n} \alpha_i v^i$ and $w = \sum_{i=1}^{n} \beta_i v^i$. Suppose *per contra* that $T(v) = T(w)$. By definition, this implies that $\sum_{i=1}^{n} \alpha_i \overline{v}^i = \sum_{i=1}^{n} \beta_i \overline{v}^i$, and so $\sum_{i=1}^{n} (\alpha_i - \beta_i) \overline{v}^i = 0$. As the vectors in $\{\overline{v}^i\}_{i=1}^{n}$ are linearly independent, we have $\alpha_i = \beta_i$ for every $i = 1, ..., n$, which contradicts $v \neq w$. Therefore, $T(v) \neq T(w)$ and we conclude that T is injective.

To conclude the proof of the "If," it remains to prove that T is surjective, that is, for each $v \in V_2$ there exists $v^* \in V_1$ such that $T(v^*) = v$. Fist notice that for every $i = 1, ..., n$ we have $T(v^i) = \overline{v}^i$. Let $v \in V_2$. As $\{\overline{v}^i\}_{i=1}^{n}$ is a basis of V_2, there exists $\{\alpha_i\}_{i=1}^{n} \subseteq \mathbb{R}$ such that $v = \sum_{i=1}^{n} \alpha_i \overline{v}^i$. Set $v^* = \sum_{i=1}^{n} \alpha_i v^i$. Clearly, $v^* \in V_1$; moreover, by definition we have

$$T(v^*) = \sum_{i=1}^{n} \alpha_i \overline{v}^i = v,$$

and therefore T is surjective.

"Only if." Let V_1 and V_2 be two isomorphic spaces, that is, there exists a linear application $T : V_1 \to V_2$ that is both injective and surjective. Assume that $\dim(V_1) = n$ and let $\{v^i\}_{i=1}^{n}$ be a basis of V_1. To prove that $\dim(V_2) = n$, it is sufficient to prove that $\{T(v^i)\}_{i=1}^{n}$ is a basis of V_2. We leave to the reader the easy proof. ■

As an immediate consequence of the previous theorem, we have:

Corollary 105 *A vector space has dimension n if and only if it is isomorphic to $\mathbb{R}^n$.*

Before stating the next result, we introduce some important notions. Given an application $T : V_1 \to V_2$, its kernel $\ker(T)$ is the set

$$\ker(T) = \{v \in V_1 : T(v) = \mathbf{0}\}. \tag{3.12}$$

That is, $\ker(T) = T^{-1}(\mathbf{0})$. In other words, the kernel is the set of the points in which the application is null, that is, it takes on as value the null vector $\mathbf{0}$ of V_2.

Another important set is the imagine of T, which is defined in the usual way as:

$$\mathrm{Im}(T) = \{v \in V_2 : v = T(w) \text{ for some } w \in V_1\}. \tag{3.13}$$

The image is therefore the set of the vectors of V_2 that are "reached" by V_1 through the application T.

It is easy to see that, if T is linear, then both $\ker(T)$ and $\mathrm{Im}(T)$ are vector subspaces. These two subspaces are important to study the properties of injectivity and surjectivity of the applications. In particular, by definition T is surjective when $\mathrm{Im}(T) = V_2$, while by exploiting the linearity of T we have the following simple characterization of injectivity.

Proposition 106 *A linear application T is injective if and only if* $\ker(T) = \{\mathbf{0}\}$.

Proof "If." Let $T : V_1 \to V_2$ be a linear application such that $\ker(T) = \{\mathbf{0}\}$. Let $v, w \in V_1$ with $v \neq w$. Being $v - w \neq \mathbf{0}$, the hypothesis $\ker(T) = \{\mathbf{0}\}$ implies $T(v - w) \neq \mathbf{0}$, and so $T(v) \neq T(w)$.

"Only if." Let $T : V_1 \to V_2$ be a linear injective application and let $v \in \ker(T)$. If $v \neq \mathbf{0}$, we have $T(v) \neq T(\mathbf{0}) = \mathbf{0}$, a contradiction. Therefore $v = \mathbf{0}$, which implies $\ker(T) = \{\mathbf{0}\}$. ■

We can now state an important result, which shows that the dimension of V_2 is the sum of the dimensions of the two subspaces $\ker(T)$ and $\mathrm{Im}(T)$. To this end, we first give a name to these two dimensions.

Definition 107 *Let V_1 and V_2 be two finite dimensional vector spaces. The rank $\rho(T)$ of a linear application $T : V_1 \to V_2$ is the dimension of* $\mathrm{Im}(T)$*, while the nullity $\nu(T)$ is the dimension of* $\ker(T)$.

Using this terminology, we can now state and prove the result.

Theorem 108 *Let V_1 and V_2 be two finite dimensional vector spaces. Given a linear application $T : V_1 \to V_2$, we have*

$$\rho(T) + \nu(T) = \dim(V_1). \tag{3.14}$$

Proof Setting $\rho(T) = k$ and $\nu(T) = l$, let $\{\overline{v}_i\}_{i=1}^k$ be a basis of the vector subspace $\mathrm{Im}(T)$ of V_2 and $\{v_i\}_{i=1}^l$ a basis of the vector subspace $\ker(T)$ of V_1. Being $\{\overline{v}_i\}_{i=1}^k \subseteq \mathrm{Im}(T)$, by definition there exist k vectors $\{w_i\}_{i=1}^k$ in V_1 such that $T(w_i) = \overline{v}_i$ for every $i = 1, ..., k$. Set

$$S = \{w_1, ..., w_k, v_1, ..., v_l\}.$$

To prove the theorem it is sufficient to prove that S is a basis of V_1. We first show that S is a linearly independent set. Let $\{\alpha_1, ..., \alpha_k, \beta_1, ..., \beta_l\} \subseteq \mathbb{R}$ such that

$$\sum_{i=1}^{k} \alpha_i w_i + \sum_{i=1}^{l} \beta_i v_i = \mathbf{0}. \tag{3.15}$$

As $\{v_i\}_{i=1}^{l}$ is a basis of $\ker(T)$, we have $\sum_{i=1}^{l}\beta_i T(v_i) = T\left(\sum_{i=1}^{l}\beta_i v_i\right) = \mathbf{0}$. Therefore, (3.15) implies

$$\sum_{i=1}^{k}\alpha_i T(w_i) + \sum_{i=1}^{l}\beta_i T(v_i) = \sum_{i=1}^{k}\alpha_i T(w_i) = \sum_{i=1}^{k}\alpha_i \overline{v}_i = \mathbf{0}. \tag{3.16}$$

Being a basis, $\{\overline{v}_i\}_{i=1}^{k}$ is a linearly independent set and so (3.16) implies $\alpha_i = 0$ for every $i = 1, ..., k$. Therefore, (3.15) reduces to $\sum_{i=1}^{l}\beta_i v_i = \mathbf{0}$, which implies $\beta_i = 0$ for every $i = 1, ..., l$ because also $\{v_i\}_{i=1}^{l}$, being a basis, is a linearly independent set. In conclusion, the set S is linearly independent.

It remains to prove that $V_1 = span(S)$. Let $v \in V_1$ and consider its image $T(v)$. By definition, $T(v) \in \mathrm{Im}(T)$ and therefore, since $\{\overline{v}_i\}_{i=1}^{k}$ is a basis of $\mathrm{Im}(T)$, there exists a set $\{\alpha_i\}_{i=1}^{k} \subseteq \mathbb{R}$ such that $T(v) = \sum_{i=1}^{k}\alpha_i\overline{v}_i$. Having set $\overline{v}_i = T(w_i)$ for every $i = 1, ..., k$, this implies

$$T(v) = \sum_{i=1}^{k}\alpha_i T(w_i) = T\left(\sum_{i=1}^{k}\alpha_i w_i\right).$$

Consequently, $T\left(v - \sum_{i=1}^{k}\alpha_i w_i\right) = 0$ and therefore $\left(v - \sum_{i=1}^{k}\alpha_i w_i\right) \in \ker(T)$. On the other hand, $\{v_i\}_{i=1}^{l}$ is a basis of $\ker(T)$, and so there exists a set $\{\beta_i\}_{i=1}^{l} \subseteq \mathbb{R}$ such that $v - \sum_{i=1}^{k}\alpha_i w_i = \sum_{i=1}^{l}\beta_i v_i$. In conclusion, $v = \sum_{i=1}^{k}\alpha_i w_i + \sum_{i=1}^{l}\beta_i v_i$, which proves that $v \in span(S)$, as desired. ■

Theorem 108 has some important consequences. We begin by analyzing the relationships with Theorem 104. One of the implications of this theorem is that if the application $T : V_1 \to V_2$ is an isomorphism, then $\dim(V_1) = \dim(V_2)$. Next corollary refines this conclusion by considering separately injectivity and surjectivity.

Corollary 109 *Let V_1 and V_2 be two finite dimensional vector spaces. A linear application $T : V_1 \to V_2$ is injective only if $\dim(V_1) \leq \dim(V_2)$, while it is surjective only if $\dim(V_1) \geq \dim(V_2)$.*

Proof Let T be injective, so that $\ker(T) = \{\mathbf{0}\}$. Since $\mathrm{Im}(T)$ is a vector subspace of V_2, we have $\rho(T) = \dim(\mathrm{Im}(T)) \leq \dim(V_2)$. Therefore, (3.14) reduces to

$$\dim(V_1) = \rho(T) + \dim(\mathbf{0}) = \rho(T) \leq \dim(V_2).$$

Assume now that T is surjective, that is, $\mathrm{Im}(T) = V_2$. Since $\nu(T) \geq 0$, (3.14) implies:

$$\dim(V_1) = \rho(T) + \nu(T) = \dim(V_2) + \nu(T) \geq \dim(V_2),$$

as desired. ■

We can now see an important consequence of Theorem 108. Usually, the properties of injectivity and surjectivity are very different and altogether independent features of a function. It is very easy to construct examples of functions that are injective, but not surjective, and viceversa. Next we shows how that linear applications among spaces of the same dimension, these two properties are actually equivalent.

Corollary 110 *Let V_1 and V_2 be two finite dimensional vector spaces with $\dim(V_1) = \dim(V_2)$. A linear application $T : V_1 \to V_2$ is injective if and only if it is surjective. In particular, the following conditions are equivalent:*

(i) T is an isomorphism,

(ii) $\ker(T) = \{\mathbf{0}\}$,

(iii) $\mathrm{Im}(T) = V_2$.

Proof We prove that conditions (i)-(iii) are equivalent. By Proposition 106, (i) implies (ii). Assume (ii), that is, $\ker(T) = \{\mathbf{0}\}$. Using (3.14) and the hypothesis $\dim(V_1) = \dim(V_2)$, we have:

$$\dim(V_2) = \dim(V_1) = \rho(T)$$

Being $\mathrm{Im}(T)$ a subspace of V_2, the equality $\dim(V_2) = \rho(T)$ implies $\mathrm{Im}(T) = V_2$. Therefore, (ii) implies (iii).

It remains to prove that (iii) implies (i). Assume therefore (iii), that is, $\mathrm{Im}(T) = V_2$. To prove that T is an isomorphism it is sufficient to prove that it is injective. Using (3.14) and the hypothesis $\dim(V_1) = \dim(V_2)$, we have

$$\rho(T) + \nu(T) = \dim(V_1) = \dim(V_2) = \rho(T).$$

Therefore, $\nu(T) = 0$, which implies $\ker(T) = \{\mathbf{0}\}$. By Proposition 106, T is then injective, as desired. ■

Notice that an equivalent way to state the second part of Corollary 110 is to say that, setting $n = \dim(V_1) = \dim(V_2)$, the following conditions are equivalent:

(i) T is an isomorphism,

(ii) $\nu(T) = 0$,

(iii) $\rho(T) = n$.

3.5 Invertible Applications

3.5.1 Definitions and Properties

In the previous section we introduced isomorphisms and we saw some important properties in the finite dimensional case. In particular, by Theorem 104 we have $\dim(V_1) = \dim(V_2)$ and, therefore, to study isomorphisms among finite dimensional vector spaces it is necessary to consider the case $\dim(V_1) = \dim(V_2)$. At this point, we assume directly that $V \equiv V_1 = V_2$, though without assuming a priori that V is necessarily finite dimensional.

In this case, an application $T \in L(V)$ that is an isomorphism is usually called *invertible*. In other words, a linear application $T \in L(V)$ is invertible if it is both injective and surjective.

Given an invertible application $T \in L(V)$, consider its inverse $T^{-1} : V \to V$. It is easy to verify that

$$T^{-1}T = TT^{-1} = I, \tag{3.17}$$

and that the application T^{-1} is itself linear, that is, $T^{-1} \in L(V)$.

Example 111 The identity $I : V \to V$ is invertible and we have $I^{-1} = I$. ▲

Example 112 Let $T : \mathbb{R}^2 \to \mathbb{R}^2$ be defined as $T(x) = Ax$ for every $x \in \mathbb{R}^2$, where

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix}.$$

The application T is invertible, with $T^{-1}(x) = Bx$ for every $x \in \mathbb{R}^2$, where

$$B = \begin{bmatrix} 1 & 0 \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

▲

Thanks to Corollary 110, we have a first characterization of invertibility on finite dimensional spaces. In fact, by this corollary the following properties are equivalent for $T \in L(V)$, when V is finite dimensional:

(i) T is invertible,

(ii) $\ker(T) = \{\mathbf{0}\}$,

(iii) $\mathrm{Im}(T) = V$.

We can however give another characterization of the invertibility that, unlike the previous one, holds for any space V, not necessarily finite dimensional.

Theorem 113 *An application $T \in L(V)$ is invertible if and only if there exist $S, R \in L(V)$ such that*

$$TS = RT = I. \tag{3.18}$$

In this case, S and R are unique and we have $S = R = T^{-1}$.

Proof "Only if." Let T be invertible; (3.17) implies that (3.18) holds with $S = R = T^{-1}$.

"If." Assume there exist $S, R \in L(V)$ such that (3.18) holds. Let $v, w \in V$ be such that $v \neq w$. We have $T(v) \neq T(w)$ and therefore T is injective. In fact, if it were $T(v) = T(w)$, from (3.18) we would have

$$v = R(T(v)) = R(T(w)) = w,$$

which contradicts $v \neq w$. It remains to prove that T is surjective. Let $v \in V$ and set $w = S(v)$. From (3.18), we have

$$T(w) = T(S(v)) = v,$$

and therefore $v \in \mathrm{Im}(T)$. This implies $V = \mathrm{Im}(T)$, as desired. In conclusion, T is invertible.

Using (3.17) and (3.18), we have:

$$\begin{aligned} S(v) &= \left(T^{-1} \circ T\right)(S(v)) = T^{-1}((T \circ S)(v)) = T^{-1}(v), \\ R(v) &= R\left(\left(T \circ T^{-1}\right)(v)\right) = (R \circ T)\left(T^{-1}(v)\right) = T^{-1}(v), \end{aligned}$$

for every $v \in V$, and therefore $S = R = T^{-1}$. ■

In (3.18) we need both $TS = I$ and $RT = I$. Otherwise, T might not be invertible, as the following example shows.

Example 114 Let $S : \mathcal{P} \to \mathcal{P}$ be the integral application defined by

$$S(f)(t) = \int_0^t f(s)\, ds$$

for every $f \in \mathcal{P}$, and let $D : \mathcal{P} \to \mathcal{P}$ be the usual differential operator defined as $D(f)(t) = f'(t)$ for every $f \in \mathcal{P}$. As well known, we have $DS = I$. On the other hand, neither D nor S are invertible. For example, D is clearly not injective. ▲

3.5.2 Inverse matrices and Determinants

Inverse

Consider now the case $V = \mathbb{R}^n$, and let $T \in L(\mathbb{R}^n)$ be a linear application on $\mathbb{R}^n$ to which is associated the square matrix A. If T is invertible, the matrix A is called *invertible*; the matrix associated to the inverse application T^{-1} is called *inverse* matrix of A and is denoted by A^{-1}. Going back to Example 112, we have

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix} \quad \text{and} \quad A^{-1} = \begin{bmatrix} 1 & 0 \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

>From (3.17) we have:

$$A^{-1}A = AA^{-1} = I.$$

More generally, by Theorem 113 we have the following immediate characterization of the invertibility of matrices.

Corollary 115 *A square matrix* $\underset{n\times n}{A}$ *is invertible if and only if there exist two square matrices* $\underset{n\times n}{B}$ *and* $\underset{n\times n}{C}$ *such that*

$$AB = CA = I.$$

In this case, these matrices are unique, with $B = C = A^{-1}$.

So far so good, but now the problem is to compute the inverse of an invertible matrix, that is, given an invertible matrix A, to find what are the components of its inverse A^{-1}. To do this, we must stop and introduce determinants.

Determinants

Given a $m \times n$ matrix A, the submatrix A_{ij} is the matrix $(m-1) \times (n-1)$ obtained from A by cancelling the row i and the column j.

Example 116 Let

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} 2 & 1 & 4 \\ 3 & 1 & 0 \\ 1 & 6 & 3 \end{bmatrix}$$

We have, for example,

$$A_{12} = \begin{bmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 1 & 3 \end{bmatrix}, \quad A_{32} = \begin{bmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{bmatrix} = \begin{bmatrix} 2 & 4 \\ 3 & 0 \end{bmatrix},$$

$$A_{22} = \begin{bmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{bmatrix} = \begin{bmatrix} 2 & 4 \\ 1 & 3 \end{bmatrix}, \quad A_{31} = \begin{bmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{bmatrix} = \begin{bmatrix} 1 & 4 \\ 1 & 0 \end{bmatrix}.$$

▲

Using submatrices, we can define recursively the determinants of square matrices.

Definition 117 *The determinant is a function* $\det : M(n) \to \mathbb{R}$ *such that, for every* $A \in M(n)$*, we have:*

(i) if $n = 1$*, i.e.* $A = [a_{11}]$*, we set* $\det A = a_{11}$*,*

(ii) if $n > 1$*, i.e.* $A = (a_{ij})$*, we set* $\det A = \sum_{j=1}^{n} (-1)^{1+j} a_{1j} \det A_{1j}$*.*

We now illustrate the computation of determinants with some examples.

Example 118 If $n = 2$, the determinant of the matrix

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

is

$$\begin{aligned} \det A &= (-1)^{1+1} a_{11} \det([a_{22}]) + (-1)^{1+2} a_{12} \det([a_{21}]) \\ &= a_{11}a_{22} - a_{12}a_{21.} \end{aligned}$$

For example, if

$$A = \begin{bmatrix} 2 & 4 \\ 1 & 3 \end{bmatrix}$$

we have $\det A = 2 \cdot 3 - 4 \cdot 1 = 2$. ▲

Example 119 If $n = 3$, the determinant of the matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

is given by

$$\begin{aligned} \det A &= (-1)^{1+1} a_{11} \det A_{11} + (-1)^{1+2} a_{12} \det A_{12} + (-1)^{1+3} a_{13} \det A_{13} \\ &= a_{11} \det A_{11} - a_{12} \det A_{12} + a_{13} \det A_{13} \\ &= a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{23}a_{31}) + a_{13}(a_{21}a_{32} - a_{22}a_{31}) \\ &= a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31}. \end{aligned}$$

For example, suppose we want to compute the determinant of the following matrix:

$$A = \begin{bmatrix} 2 & 1 & 4 \\ 3 & 1 & 0 \\ 1 & 6 & 3 \end{bmatrix}.$$

First of all we compute the determinants of the three submatrices A_{11}, A_{12}, and A_{13}. We have

$$\begin{aligned} \det A_{11} &= 1 \cdot 3 - 0 \cdot 6 = 3, \\ \det A_{12} &= 3 \cdot 3 - 0 \cdot 1 = 9, \\ \det A_{13} &= 3 \cdot 6 - 1 \cdot 1 = 17, \end{aligned}$$

and therefore

$$\det A = 2 \det A_{11} - 1 \det A_{12} + 4 \det A_{13} = 2 \cdot 3 - 1 \cdot 9 + 4 \cdot 17 = 65.$$

▲

Example 120 A matrix of the form

$$\begin{bmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ 0 & a_{22} & 0 & \cdots & 0 \\ 0 & 0 & a_{33} & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & a_{nn} \end{bmatrix}$$

is called diagonal. It is easy to see that in this case $\det A = a_{11}a_{22}a_{33} \cdots a_{nn}$. ▲

Inverse and Determinants

We saw that the determinant of any square matrix can be computed through a well specified procedure – an algorithm – based on submatrices. There exist different techniques to simplify the computation of determinants, but, for our purposes, it is sufficient to know that they are quantities that can be computed through algorithms.

Next result, whose proof if omitted, shows the importance of the determinants in the computation of the inverses.

Theorem 121 *A square matrix A is invertible if and only if $\det A \neq 0$. In this case, the components a_{ij}^{-1} of the inverse matrix A^{-1} are given by:*

$$a_{ij}^{-1} = (-1)^{i+j} \frac{\det A_{ji}}{\det A}. \tag{3.19}$$

A matrix A for which $\det A = 0$ is called *singular*. Using this terminology, Theorem 121 states that a matrix is invertible if and only if it is non-singular.

This theorem is important because, through determinants, it gives us an algorithm that allows both to verify the invertibility of A and to compute the components of the inverse A^{-1}. Note that in formula (3.19) the subscript of A_{ji} is precisely ji and not ij.

Example 122 Let A be a square matrix 2×2. We use formula (3.19) to compute the inverse A^{-1}. For concreteness, let

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 5 \end{bmatrix}$$

We have:

$$\begin{aligned} a_{11}^{-1} &= (-1)^{1+1} \frac{\det A_{11}}{\det A} = \frac{a_{22}}{a_{11}a_{22} - a_{12}a_{21}} = \frac{5}{-1} = -5, \\ a_{12}^{-1} &= (-1)^{1+2} \frac{\det A_{21}}{\det A} = -\frac{a_{12}}{a_{11}a_{22} - a_{12}a_{21}} = -\frac{2}{-1} = 2, \\ a_{21}^{-1} &= (-1)^{2+1} \frac{\det A_{12}}{\det A} = -\frac{a_{21}}{a_{11}a_{22} - a_{12}a_{21}} = -\frac{3}{-1} = 3, \\ a_{22}^{-1} &= (-1)^{2+2} \frac{\det A_{22}}{\det A} = \frac{a_{11}}{a_{11}a_{22} - a_{12}a_{21}} = \frac{1}{-1} = -1, \end{aligned}$$

and therefore

$$A^{-1} = \begin{bmatrix} \frac{a_{22}}{\det A} & -\frac{a_{12}}{\det A} \\ -\frac{a_{21}}{\det A} & \frac{a_{11}}{\det A} \end{bmatrix} = \begin{bmatrix} -5 & 2 \\ 3 & -1 \end{bmatrix}.$$

▲

Example 123 Let A be a diagonal matrix. Using formula (3.19), it is possible to show that

$$a_{ij}^{-1} = \begin{cases} \frac{1}{a_{ij}} & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

▲

Linear Systems

Using inverse matrices it is possible to give a procedure to solve the linear systems of n equations in n unknowns:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2, \\ \qquad\qquad \cdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n, \end{cases}$$

Consider the system in matrix form

$$\underset{(n\times n)}{A} \underset{(n\times 1)}{x} = \underset{n\times 1}{b}, \tag{3.20}$$

where A is a square matrix $n \times n$, while x and b are vectors in $\mathbb{R}^n$. We consider two problems with reference to the system (3.20):

- Existence: under what conditions the system has solution for each vector $b \in \mathbb{R}^n$, that is, when, for each given $b \in \mathbb{R}^n$, there exists $x \in \mathbb{R}^n$ such that $Ax = b$.

- Uniqueness: under what conditions such solution is unique, that is, when, for each given $b \in \mathbb{R}^n$, there exists at most a unique $x \in \mathbb{R}^n$ such that $Ax = b$.

To set this problem in what we have studied till now, consider the linear application $T : \mathbb{R}^n \to \mathbb{R}^n$ associated to A, defined as $T(x) = Ax$ for every $x \in \mathbb{R}^n$. It is immediate to see that:

- the system admits solutions for a given $b \in \mathbb{R}^n$ if and only if $b \in \operatorname{Im} T$; in particular, the system has a solution for each $b \in \mathbb{R}^n$ if and only if T is surjective, that is, $\operatorname{Im} T = \mathbb{R}^n$;

- the system admits a unique solution for a given $b \in \mathbb{R}^n$ if and only if $T^{-1}(b)$ is a singleton; in particular, the system admits a unique solution for each $b \in \mathbb{R}^n$ if and only if T is injective.[2]

Since injectivity and surjectivity are, by Corollary 110, equivalent properties, it follows from this that the two problems of existence and uniqueness are equivalent: there exists a solution for the system (3.20) for each $b \in \mathbb{R}^n$ if and only if this solution is unique.

In particular, a necessary and sufficient condition for this unique solution to exist for each $b \in \mathbb{R}^n$ is that the application T is invertible; equivalently, that the matrix A is invertible.

The desired condition is, therefore, the invertibility of the matrix A. Formally, we have the following result, often called "Cramer's Rule," which therefore easily follows from what we have seen till now.

Proposition 124 *The system (3.20) has one and only one solution for each $b \in \mathbb{R}^n$ if and only if the matrix A is invertible. In this case, the solution is given by*

$$x = A^{-1}b.$$

Proof "If." Let A be invertible. The associated linear application $T : \mathbb{R}^n \to \mathbb{R}^n$ is invertible, and so both surjective and injective. Since T is surjective, the system has solution. Since T is injective, such solution is unique. In particular, the solution

[2]Remember that a function $f : A \to B$ among two generic sets A and B is injective if and only if all the counterimages $f^{-1}(y)$ are singletons.

corresponding to a certain $b \in \mathbb{R}^n$ is given by $T^{-1}(b)$. Since $T^{-1}(y) = A^{-1}y$ for each $y \in \mathbb{R}^n$, it follows that the solution is given by $T^{-1}(b) = A^{-1}b$.[3]

"Only if." Assume that the system (3.20) admits one and only one solution for each $b \in \mathbb{R}^n$. This means that for each vector $b \in \mathbb{R}^n$ there exists one and only one vector $x \in \mathbb{R}^n$ such that $T(x) = b$. Therefore, the application T is bijective, and therefore it is invertible. It follows that also A is invertible. ■

Therefore, the system (3.20) admits solution if and only if the matrix A is invertible and, more importantly, the unique solution can be expressed in terms of A^{-1}. Since thanks to Theorem 121 we know how to compute A^{-1} through the determinants, we have thus derived a solution procedure for linear systems of n equations in n unknowns. Though we omit the details, this procedure can be easily extended to general systems of m equations in n unknowns.

Example 125 A special case of the system (3.20) is when $b = \mathbf{0}$. In this case, the system is called homogeneous and, by Proposition 124, the unique possible solution is $x = \mathbf{0}$. ▲

Example 126 Consider the following system of 2 equations in 2 unknowns:

$$\begin{cases} x_1 + 2x_2 = b_1 \\ 3x_1 + 5x_2 = b_2 \end{cases}$$

In this case, we have

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 5 \end{bmatrix}$$

From Example 122 we know that A is invertible. By Proposition 124, the unique solution of the system is given by

$$x = A^{-1}b = \begin{bmatrix} -5 & 2 \\ 3 & -1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} -5b_1 + 2b_2 \\ 3b_1 - b_2 \end{bmatrix}.$$

▲

[3] Alternatively, it is possible to prove the "If" also in the following way, quite mechanical. Set $x = A^{-1}b$; we have

$$Ax = A\left(A^{-1}b\right) = \left(AA^{-1}\right)b = Ib = b,$$

and therefore $x = A^{-1}b$ solves the system. It is also the unique solution. In fact, let $\widetilde{x} \in \mathbb{R}^n$ be another solution. We have

$$\widetilde{x} = I\widetilde{x} = \left(A^{-1}A\right)\widetilde{x} = A^{-1}\left(A\widetilde{x}\right) = A^{-1}b,$$

as desired.

Example 127 Consider the following system of 3 equations in 3 unknowns:

$$\begin{cases} x_1 - 2x_2 + 2x_3 = b_1 \\ 2x_2 - x_3 = b_2 \\ x_2 - x_3 = b_3 \end{cases}$$

We have

$$A = \begin{bmatrix} 1 & -2 & 2 \\ 0 & 2 & -1 \\ 0 & 1 & -1 \end{bmatrix}$$

Using the sub-matrices, it is easy (but tedious) to verify that $\det A = -1 \neq 0$. Therefore, A is invertible and, using formula (3.19), it is possible to verify that

$$A^{-1} = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & -1 \\ 0 & 1 & -2 \end{bmatrix}$$

By Proposition 124, the unique solution of the system is therefore given by

$$x = A^{-1}b = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & -1 \\ 0 & 1 & -2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} b_1 + 2b_3 \\ b_2 - b_3 \\ b_2 - 2b_3 \end{bmatrix}.$$

For example, if $b = (1, -1, 2)$, we have

$$x = (1 + 2 \cdot 2, -1 - 2, -1 - 2 \cdot 2) = (5, -3, -5).$$

▲

Chapter 4

Differential Calculus in Several Variables

4.1 Gateaux Differential

4.1.1 Directional Derivatives

We know from Calculus that for a scalar functions $f : A \subseteq \mathbb{R} \to \mathbb{R}$ defined on an open set A, the derivative $f'(x)$ in the point $x \in A$ is given by:

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h},$$

when such limit exists and is finite.[1] To give a first extension of this notion to the case of functions of several variables, it is useful to see the derivative from a "directional" point of view. In order to do this, we remind a basic result concerning bilateral and unilateral limits: given a scalar function $f : A \subseteq \mathbb{R} \to \mathbb{R}$ and a point $x_0 \in A$, we have

$$\lim_{x \to x_0} f(x) = L \quad \text{if and only if} \quad \lim_{x \to x_0+} f(x) = \lim_{x \to x_0-} f(x) = L. \tag{4.1}$$

In the case of limits of incremental ratios, (4.1) becomes:

$$\lim_{h \to 0} \frac{f(x+h) - f(x)}{h} = f'(x) \Longleftrightarrow \lim_{h \to 0+} \frac{f(x+h) - f(x)}{h} = \lim_{h \to 0-} \frac{f(x+h) - f(x)}{h} = f'(x).$$

On the other hand, it is immediate to see that

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

and therefore we have

$$\lim_{h \to 0} \frac{f(x+h) - f(x)}{h} = f'(x)$$

[1] Throughout all the chapter, A will always denote an open set.

if and only if

$$\lim_{h\to 0+} \frac{f(x+h)-f(x)}{h} = -\lim_{h\to 0+} \frac{f(x-h)-f(x)}{h} = f'(x).$$

It is useful to rewrite this equivalence in the following way, which for simplicity we write as a proposition.

Proposition 128 *Given a scalar function $f : A \subseteq \mathbb{R} \to \mathbb{R}$, the derivative $f'(x)$ exists if and only if the limits:*

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

exist finite for $y = \pm 1$. In particular,we have:

$$f'(x) = \lim_{t\to 0+} \frac{f(x+ty)-f(x)}{t} \quad \text{for } y = 1, \tag{4.2}$$

and

$$f'(x) = -\lim_{t\to 0+} \frac{f(x+ty)-f(x)}{t} \quad \text{for } y = -1. \tag{4.3}$$

On the real line $\mathbb{R}$ there are two fundamental directions, the positive direction "+" and the negative one "−." Given a point $x \in \mathbb{R}$, when $y = 1$ the limit

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

tells us which is the infinitesimal increment of the function f at the point x when we move in the direction "+;" in the same way, when $y = -1$ the limit

$$\lim_{t\to 0+} \frac{f(x+ty)-f(x)}{t} \tag{4.4}$$

tells us which is the infinitesimal increment of the function f at the point x when we move in the direction "−." By Proposition 128, the derivative $f'(x)$ exists when the increments considered in both directions coincide, except for the sign.

While on the real line there exist only two directions, this is no longer true in $\mathbb{R}^n$, where from each point we can move along infinite directions. It becomes therefore natural to consider the increments along all possible directions. Using the limit (4.4), we have therefore the following definition for functions of several variables.

Definition 129 *Given a function $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$, the derivative of f at $x \in A$ along the direction $y \in \mathbb{R}^n$ is given by*

$$f'(x;y) = \lim_{t\to 0+} \frac{f(x+ty)-f(x)}{t}, \tag{4.5}$$

when such limit exists finite.

In other words, the derivative $f'(x;y)$ represents the infinitesimal increment of the function f at the point x when we move along the direction determined by the vector y. Fixed $x \in \mathbb{R}^n$, the function $f'(x;\cdot) : D \subseteq \mathbb{R}^n \to \mathbb{R}$ is called the *directional derivative* of f at x. Its domain D is the set of all directions along which the limit (4.5) exists finite.

To better understand this notion, observe that, given any two vectors $x, y \in \mathbb{R}^n$, the segment $[x, y]$ that joins them is given by:

$$\{(1-t)x + ty : t \in [0,1]\}.$$

On the other hand, going back to (4.5), we have

$$f(x+ty) = f((1-t)x + t(x+y)),$$

and therefore the ratio

$$\frac{f(x+ty) - f(x)}{t}$$

tells us which is the "incremental" behavior of the function when we move along the segment $[x, x+y]$. Each $y \in \mathbb{R}^n$ identifies a segment and therefore gives us a direction along which we can study the increments of the function.

Not all segments $[x, x+y]$ identify different directions. In fact, for a fixed vector $y \in \mathbb{R}^n$, all vectors αy, with $\alpha > 0$, identify the same direction. To see why this is the case, notice that two different segments $[x, x+y]$ and $[x, x+y']$ identify the same direction when one of the two is the extension of the other, that is, when

$$[x, x+y] \subseteq [x, x+y'] \quad \text{or} \quad [x, x+y'] \subseteq [x, x+y]. \tag{4.6}$$

Proposition 130 *Given a point $x \in \mathbb{R}^n$, for every $y, y' \in \mathbb{R}^n$ we have*

$$[x, x+y] \subseteq [x, x+y'] \quad \textit{or} \quad [x, x+y'] \subseteq [x, x+y]$$

if and only if there exists $\alpha > 0$ such that $y' = \alpha y$.

Proof "If." Suppose that $y' = \alpha y$ with $\alpha > 0$. We assume $\alpha \leq 1$. We prove that in this case we have $[x, x+y'] \subseteq [x, x+y]$. We have

$$x + y' = x + \alpha y = \alpha x + (1-\alpha)x + \alpha y = (1-\alpha)x + \alpha(x+y),$$

and therefore, being $\alpha \leq 1$, we have $x + y' \in [x, x+y]$. This implies $[x, x+y'] \subseteq [x, x+y]$, as desired.

Proceeding in a similar way, we prove that if $\alpha > 1$, we have on the contrary $[x, x+y] \subseteq [x, x+y']$. We conclude therefore that if $y' = \alpha y$ with $\alpha > 0$, (4.6) holds.

"Only if". Suppose that $[x, x+y'] \subseteq [x, x+y]$. Since $x+y' \in [x, x+y]$, there exists $t \in (0,1)$ such that $x+y' = (1-t)\,x + t\,(x+y)$. This implies that $y' = ty$ and therefore, setting $\alpha = t$, we have the result desired. ■

As the next corollary shows, this redundancy of the directions is reflected in a simple and elegant way by the positive homogeneity of the directional derivative, a property that allows to determine immediately the value of $f'(x;\alpha y)$ for every $\alpha \geq 0$ once we know the value of $f'(x;y)$.

Corollary 131 *Given a point $x \in A$, for every $y \in D$ and every $\alpha \geq 0$, we have*

$$f'(x;\alpha y) = \alpha f'(x;y), \tag{4.7}$$

that is, the directional derivative $f'(x;\cdot) : D \subseteq \mathbb{R}^n \to \mathbb{R}$ is a positively homogeneous function.

Proof Let $\alpha > 0$. Since $t \to 0+$ if and only if $(\alpha t) \to 0+$, we have:

$$\lim_{t\to 0+} \frac{f(x+(\alpha t)\,y) - f(x)}{\alpha t} = \lim_{(\alpha t)\to 0+} \frac{f(x+(\alpha t)\,y) - f(x)}{\alpha t} = f'(x;y).$$

Dividing and multiplying by α, we therefore have:

$$\lim_{t\to 0+} \frac{f(x+t\,(\alpha y)) - f(x)}{t} = \alpha \lim_{t\to 0+} \frac{f(x+(\alpha t)\,y) - f(x)}{\alpha t} = \alpha f'(x;y).$$

It follows that the limit

$$f'(x;\alpha y) = \lim_{t\to 0+} \frac{f(x+t\,(\alpha y)) - f(x)}{t}$$

exists finite and is equal to $\alpha f'(x;y)$, as desired.

On the other hand, if $\alpha = 0$, we have

$$f'(x;\alpha y) = f'(x;\mathbf{0}) = \lim_{t\to 0+} \frac{f(x+\mathbf{0}) - f(x)}{t} = 0,$$

and therefore $f'(x;\alpha y) = 0 = \alpha f'(x;y)$, which completes the proof. ■

4.1.2 Calculus and Algebra of Directional Derivatives

Thanks to a simple observation, the calculus of directional derivatives can be reduced to the calculus of ordinary derivatives of scalar functions, which we know very well from Calculus. In fact, given a point $x \in \mathbb{R}^n$ and a direction $y \in \mathbb{R}^n$, it is possible to define an auxiliary scalar function ϕ as $\phi(t) = f(x+ty)$ for every $t \in \mathbb{R}$. The domain

of ϕ is the set $\{t \in \mathbb{R} : x + ty \in A\}$, which is an open set in $\mathbb{R}$ containing the point 0. By definition of right-side derivative, we have

$$\phi'_+(0) = \lim_{t \to 0+} \frac{\phi(t) - \phi(0)}{t} = \lim_{t \to 0+} \frac{f(x+ty) - f(x)}{t},$$

and therefore

$$f'(x;y) = \phi'_+(0). \tag{4.8}$$

The derivative $f'(x;y)$ can therefore be seen as the right-side ordinary derivative of the scalar function ϕ computed in the point 0. Naturally, when ϕ can be derived at 0, (4.8) reduces to $f'(x;y) = \phi'(0)$.

Example 132 Let $f : \mathbb{R}^3 \to \mathbb{R}$ be defined as $f(x_1, x_2, x_3) = x_1^2 + x_2^2 + x_3^2$. Compute the derivative of f *at* $x = (1, -1, 2)$ in the direction $y = (2, 3, 5)$. We have:

$$x + ty = (1 + 2t, -1 + 3t, 2 + 5t),$$

and therefore

$$\phi(t) = f(x + ty) = (1 + 2t)^2 + (-1 + 3t)^2 + (2 + 5t)^2.$$

It follows that $\phi'(t) = 76t + 18$ and, by (4.8), we can conclude that

$$f'(x;y) = \phi'(0) = 18.$$

▲

Example 133 Generalize the previous example and consider the function $f : \mathbb{R}^n \to \mathbb{R}$ defined by $f(x) = \|x\|^2$ for every $x \in \mathbb{R}^n$. We have

$$\phi'(t) = \frac{d}{dt} \sum_{i=1}^{n} (x_i + ty_i)^2 = 2 \sum_{i=1}^{n} y_i (x_i + ty_i) = 2y(x + ty),$$

and therefore

$$f'(x;y) = \phi'(0) = 2x \cdot y.$$

The directional derivative of $f(x) = \|x\|^2$ thus exists at all the points and along all possible directions. Its general form is $f'(x;y) = 2x \cdot y$. In the special case of the previous example, we have

$$f'(x;y) = 2(1, -1, 2)(2, 3, 5) = 2(2 - 3 + 10) = 18.$$

▲

Example 134 Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined as

$$f(x_1, x_2) = \begin{cases} \frac{x_1 x_2^2}{x_1^2 + x_2^2} & \text{if } (x_1, x_2) \neq (0,0), \\ 0 & \text{if } (x_1, x_2) = (0,0). \end{cases}$$

Set $x = (0,0)$. For every $y \in \mathbb{R}^2$ we have

$$\phi(t) = f(ty) = t\frac{y_1 y_2^2}{y_1^2 + y_2^2},$$

and therefore

$$f'(\mathbf{0}; y) = \phi'(0) = \frac{y_1 y_2^2}{y_1^2 + y_2^2}.$$

In conclusion, $f'(\mathbf{0}; y) = f(y)$ for every $y \in \mathbb{R}^2$. ▲

Using the auxiliary functions ϕ it is easy to prove the next result, which shows that for directional derivatives the usual algebraic rules hold:

Proposition 135 *Let $f, g : A \subseteq \mathbb{R}^n \to \mathbb{R}$ be functions that admit directional derivative at x along the direction y. Then:*

(i) $\alpha f + \beta g$ admits directional derivative at x along the direction y for every $\alpha, \beta \in \mathbb{R}$, and we have

$$(\alpha f + \beta g)'(x; y) = \alpha f'(x; y) + \beta g'(x; y),$$

(ii) fg admits directional derivative at x along the direction y, and we have

$$(fg)'(x; y) = f'(x; y)\, g(x) + f(x)\, g'(x; y),$$

(iii) f/g admits directional derivative at x along the direction y, and we have

$$\left(\frac{f}{g}\right)'(x; y) = \frac{f'(x; y)\, g(x) - f(x)\, g'(x; y)}{g^2(x)},$$

provided $g(x) \neq 0$.

Proof Denote by ϕ_f the auxiliary function of a function f. It is immediate to verify that:

$$\phi_{\alpha f + \beta g} = \alpha \phi_f + \beta \phi_g, \quad \phi_{fg} = \phi_f \phi_g, \quad \text{and} \quad \phi_{f/g} = \phi_f / \phi_g, \tag{4.9}$$

where $\phi_{\alpha f + \beta g}$ denotes the auxiliary function associated to the function $\alpha f + \beta g$, and so on. For example,

$$\phi_{fg}(t) = (fg)(x + ty) = f(x + ty)\, g(x + ty) = \phi_f(t)\, \phi_g(t).$$

As a consequence of (4.9), the rules (i)-(iii) follow directly from similar rules that hold for ordinary right-side derivatives of functions of one variable. For example, let us verify (ii). By (4.8) and (4.9) we have:

$$\begin{aligned}(fg)'(x;y) &= D_+\phi_{fg}(0) = D_+\left(\phi_f\phi_g\right)(0)\\ &= D_+\left(\phi_f\right)(0)\,\phi_g(0) + \phi_f(0)\,D_+\left(\phi_f\right)(0)\\ &= f'(x;y)\,g(x) + f(x)\,g'(x;y),\end{aligned}$$

as desired.[2] ■

4.1.3 Partial Derivatives

The vectors e^1, ..., e^n represent the fundamental directions in $\mathbb{R}^n$. The directional derivatives computed along these directions are called *partial derivatives* and have great importance. We give now their definition, in which it is required that $f'(x;-e^i) = -f'(x;e^i)$ for every $i = 1,...,n$. In other words, it is required that the incremental behavior of f along the opposite directions e^i and $-e^i$ is equal, apart from the sign.

Definition 136 *Given a function $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$, the directional derivatives*

$$f'\left(x;e^1\right),...,f'(x;e^n)$$

of f at $x \in A$ along the directions $e^1,...,e^n$ are called partial derivatives when

$$f'\left(x;-e^i\right) = -f'\left(x;e^i\right) \quad \text{for every } i = 1,...,n. \tag{4.10}$$

In this case, $f'(x;e^i)$ is called partial derivative of f with respect to x_i.

The partial derivative of f with respect to x_i tells us therefore which is the incremental behavior of f when we increment only the variable x_i, keeping fixed the other variables.

Moreover, the existence of the partial derivatives is the counterpart of differentiability for scalar functions. In fact, in the case $n = 1$ we have $e^1 = 1$, and it is therefore immediate to see how (4.10) becomes:

$$\lim_{t\to 0+}\frac{f(x+t)-f(x)}{t} = -\lim_{t\to 0+}\frac{f(x+t(-1))-f(x)}{t}$$

By Proposition 128, this equality holds if and only if there exists the ordinary derivative $f'(x)$, and in this case:

$$f'(x) = \lim_{t\to 0+}\frac{f(x+t)-f(x)}{t} = -\lim_{t\to 0+}\frac{f(x+t(-1))-f(x)}{t}.$$

[2] To simplify notation, we used $D_+\phi(0)$ to denote $\phi'_+(0)$.

Therefore, in the case $n = 1$ the partial derivative is nothing but the usual ordinary derivative.

These observations suggest, *inter alia,* also a simple method to compute partial derivatives. First of all, we observe that thanks to condition (4.10) the partial derivative at x with respect to x_i is given by the bilateral limit:

$$f'\left(x;e^i\right)=\lim_{t\to 0}\frac{f\left(x+te^i\right)-f\left(x\right)}{t}. \tag{4.11}$$

In fact,

$$\begin{aligned}-f'\left(x;-e^i\right) &= -\lim_{t\to 0+}\frac{f\left(x+t\left(-e^i\right)\right)-f\left(x\right)}{t}=\lim_{t\to 0+}\frac{f\left(x+t\left(-e^i\right)\right)-f\left(x\right)}{-t}\\ &= \lim_{t\to 0-}\frac{f\left(x+te^i\right)-f\left(x\right)}{t},\end{aligned}$$

so that (4.10) implies:

$$\lim_{t\to 0+}\frac{f\left(x+te^i\right)-f\left(x\right)}{t}=\lim_{t\to 0-}\frac{f\left(x+te^i\right)-f\left(x\right)}{t}.$$

Being equal, the two unilateral limits are in turn equal to the bilateral limit

$$\lim_{t\to 0}\frac{f\left(x+te^i\right)-f\left(x\right)}{t},$$

and therefore (4.11) holds.

Consider the scalar auxiliary function ϕ_i defined by

$$\phi_i\left(x_i\right)=f\left(x_1,...,x_{i-1},x_i,x_{i+1},...x_n\right).$$

Notice that ϕ_i is a function of only the i-th variable x_i, while the other variables are considered as constants. Using the function ϕ_i, (4.11) becomes:

$$f'\left(x;e^i\right)=\lim_{t\to 0}\frac{\phi_i\left(x_i+t\right)-\phi_i\left(x_i\right)}{t}=\phi_i'\left(x_i\right).$$

Therefore, the partial derivative $f'\left(x;e^i\right)$ is nothing but the ordinary derivative ϕ_i' of the function ϕ_i computed in the point x_i, that is, in the i-th coordinate of the vector x.

Notation. To denote the partial derivative different notations are used, among which $\frac{\partial f}{\partial x_i}$ and $D_i f$. The vector

$$\left(\frac{\partial f}{\partial x_1}\left(x\right),...,\frac{\partial f}{\partial x_n}\left(x\right)\right)$$

of the partial derivatives at x is called *gradient* of f at x and is denoted by $\nabla f\left(x\right)$.

Example 137 Let $f : \mathbb{R}^3 \to \mathbb{R}$ be defined by $f(x_1, x_2, x_3) = x_1 x_2 x_3$. We compute the partial derivatives of f *at* $x = (1, -1, 2)$. We have:

$$\begin{aligned} \phi_1(x_1) &= f(x_1, -1, 2) = -2x_1, \\ \phi_2(x_2) &= f(1, x_2, 2) = 2x_2, \\ \phi_3(x_3) &= f(1, -1, x_3) = -x_3, \end{aligned}$$

and therefore

$$\phi_1'(x_1) = -2, \quad \phi_2'(x_2) = 2, \quad \phi_3'(x_3) = -1.$$

More generally, we have:

$$\phi_1'(x_1) = x_2 x_3, \quad \phi_2'(x_2) = x_1 x_3, \quad \phi_3'(x_3) = x_1 x_2,$$

and therefore

$$\frac{\partial f}{\partial x_1}(x) = x_2 x_3, \quad \frac{\partial f}{\partial x_2}(x) = x_1 x_3, \quad \frac{\partial f}{\partial x_3}(x) = x_1 x_2,$$

that is,

$$\nabla f(x) = (x_2 x_3, x_1 x_3, x_1 x_2).$$

▲

Example 138 Let $f : \mathbb{R}^4 \to \mathbb{R}$ be defined as $f(x_1, x_2, x_3, x_4) = x_1 + e^{x_2 x_3} + 2x_4^2$. It is immediate to verify that:

$$\frac{\partial f}{\partial x_1}(x) = 1, \quad \frac{\partial f}{\partial x_2}(x) = x_3 e^{x_2 x_3}, \quad \frac{\partial f}{\partial x_3}(x) = x_2 e^{x_2 x_3}, \quad \frac{\partial f}{\partial x_4}(x) = 4x_4,$$

and therefore $\nabla f(x) = (1, x_3 e^{x_2 x_3}, x_2 e^{x_2 x_3}, 4x_4)$. ▲

4.1.4 Gateaux Differential

Corollary 131 shows that the directional derivative $f'(x; \cdot) : \mathbb{R}^n \to \mathbb{R}$ is a positively homogeneous function, that is, $f'(x; \alpha y) = \alpha f'(x; y)$ for every $\alpha \geq 0$. We observed how this reflected a redundancy in the directions identified by the vectors y.

We now consider two other properties that are desirable for the directional derivative. The first property is that it be "symmetric" with respect to the opposite directions y and $-y$, in the sense that:

$$f'(x; -y) = -f'(x; y). \tag{4.12}$$

In other terms, the incremental behavior of the function is the same in the two opposite directions, changing only the sign.

This property was already assumed along the fundamental directions in defining partial derivatives. More in general, it is a property that is desirable for the directional derivatives along each direction. Together with the positive homogeneity, (4.12) implies:

$$f'(x;\alpha y) = \alpha f'(x;y) \quad \text{for every } \alpha \in \mathbb{R}, \tag{4.13}$$

as it is immediate to verify by observing that with $\alpha = -1$ we get exactly (4.12).

A second desirable property of the directional derivative is that it be additive along the directions, that is:

$$f'(x;y_1+y_2) = f'(x;y_1) + f'(x;y_2) \tag{4.14}$$

for every $y_1, y_2 \in \mathbb{R}^n$. In this case, the incremental behavior of the function along the direction y_1+y_2 can be decomposed in the sum of the behaviors along the two directions y_1 and y_2. The utility of this property consists in the possibility of reconstructing the behavior along "compound" directions, such as $y_1 + y_2$, starting from the elementary directions y_1 and y_2.

When the directional derivative $f'(x;y)$ "behaves well" and satisfies both (4.13) and (4.14), it becomes a linear functional. In this case we have the following definition, in which together with the linearity we also assume $D = \mathbb{R}^n$, i.e., we assume that the directional derivative $f'(x;y)$ exists along all possible directions $y \in \mathbb{R}^n$.

Definition 139 *A function $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ is called differentiable according to Gateaux at $x \in A$ if $D = \mathbb{R}^n$ and if the directional derivative $f'(x;\cdot) : \mathbb{R}^n \to \mathbb{R}$ is a linear functional.*

When $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ is differentiable according to Gateaux at $x \in A$, the linear functional $f'(x;\cdot) : \mathbb{R}^n \to \mathbb{R}$ is called the *Gateaux differential*.

By Theorem 65, each linear functional on $\mathbb{R}^n$ admits a representation as scalar product. The next result, which follows immediately from Theorem 65, shows that this representation assumes a particularly interesting form in our case.

Theorem 140 *A function $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ is differentiable according to Gateaux at $x \in A$ if and only if there exists $\chi \in \mathbb{R}^n$ such that*

$$f'(x;y) = \chi \cdot y = \sum_{i=1}^{n} \chi_i y_i, \quad \forall y \in \mathbb{R}^n.$$

In this case, $\chi = \nabla f(x)$.

Proof The "If" is obvious. As to the "only if", go back to the proof of Theorem 65. In that case we showed that for a linear functional $L : \mathbb{R}^n \to \mathbb{R}$, if we put $\chi = (L(e^1), ..., L(e^n))$ we have $L(x) = \chi \cdot x$ for every $x \in \mathbb{R}^n$. In our case, we have therefore

$$\chi = \left(f'\left(x; e^1\right), ..., f'\left(x; e^n\right)\right) = \nabla f(x),$$

as desired. ■

Therefore, the vector $\chi \in \mathbb{R}^n$ in the representation is nothing but the gradient $\nabla f(x)$; that is,

$$f'(x; y) = \nabla f(x) \cdot y = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x) y_i, \quad \forall y \in \mathbb{R}^n.$$

The differentiability according to Gateaux guarantees that, once we know the value of the gradient $\nabla f(x)$, we can reconstruct the incremental behavior of the function along all directions, that is, the value of the directional derivative $f'(x; y)$ along each $y \in \mathbb{R}^n$.

Example 141 Consider again the function $f(x) = \|x\|^2$ of Example 133. We showed that $f'(x; y) = 2x \cdot y$ for every $x, y \in \mathbb{R}^n$, and therefore the function is Gateaux differentiable in each $x \in \mathbb{R}^n$. As to the gradient, we have:

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x) = 2x_1, ..., \frac{\partial f}{\partial x_n}(x) = 2x_n\right) = 2x,$$

and therefore

$$f'(x; y) = \nabla f(x) \cdot y,$$

as in Theorem 140. ▲

Example 142 Go back to the function $f : \mathbb{R}^2 \to \mathbb{R}$ of Example 134 defined as

$$f(x_1, x_2) = \begin{cases} \frac{x_1 x_2^2}{x_1^2 + x_2^2} & \text{if } (x_1, x_2) \neq (0, 0), \\ 0 & \text{if } (x_1, x_2) = (0, 0). \end{cases}$$

Let $x = (0, 0)$. We showed that $f'(\mathbf{0}; y) = f(y)$ for every $y \in \mathbb{R}^2$. Therefore, being f non-linear, the directional derivative $f'(\mathbf{0}; y)$ is not a linear functional and the function f is not Gateaux differentiable in $(0, 0)$. In particular, notice that $f'(x, y)$ satisfies property (4.13), but not (4.14). As to the last one, take $y_1 = (1, 0)$ and $y_2 = (0, 1)$. We have $f'(x, y_1) = f'(x, y_2) = 0$, while $f'(x, y_1 + y_2) = 1/2$. ▲

The last example has shown that the existence of all partial derivatives (indeed of all directional derivatives) at a given point does not imply in general that the function is Gateaux differentiable at this point. It is important to observe how, on the other hand, this is true in the special case $n = 1$.

Proposition 143 *A scalar function $f : A \subseteq \mathbb{R} \to \mathbb{R}$ is Gateaux differentiable at $x \in A$ if and only if it has a derivative at this point.*

Proof "Only if." By Proposition 128, $f'(x)$ exists if the directional derivatives $f'(x;1)$ and $f'(x;-1)$ exist, and therefore if f is Gateaux differentiable at x.

"If." Let f has a derivative at x. Setting $h = ty$ we have:

$$\begin{aligned} f'(x;y) &= \lim_{t\to 0+} \frac{f(x+ty)-f(x)}{t} = y \lim_{t\to 0+} \frac{f(x+ty)-f(x)}{ty} \\ &= y \lim_{h\to 0} \frac{f(x+h)-f(x)}{h} = yf'(x), \end{aligned}$$

and therefore $f'(x;y) = f'(x)\, y$ for every $y \in \mathbb{R}$. According to Theorem 140, f is Gateaux differentiable at x. ■

We conclude the section by observing that, as an immediate consequence of Proposition 135, for Gateaux differentiability the usual algebraic rules hold:

Corollary 144 *Let $f, g : A \subseteq \mathbb{R}^n \to \mathbb{R}$ be differentiable according to Gateaux at $x \in A$. Then:*

(i) $\alpha f + \beta g$ is differentiable according to Gateaux at x for every $\alpha, \beta \in \mathbb{R}$, and we have

$$\nabla(\alpha f + \beta g)(x) = \alpha \nabla f(x) + \beta \nabla g(x),$$

(ii) fg is differentiable according to Gateaux at x, and we have

$$\nabla(fg)(x) = g(x)\nabla f(x) + f(x)\nabla g(x),$$

(iii) f/g is differentiable according to Gateaux at x, and we have

$$\nabla\left(\frac{f}{g}\right)(x) = \frac{g(x)\nabla f(x) - f(x)\nabla g(x)}{g^2(x)},$$

provided $g(x) \neq 0$.

4.2 Frechet Differential

Two are the fundamental aspects of the derivative $f'(x)$ of a scalar function $f : A \subseteq \mathbb{R} \to \mathbb{R}$ at a point $x \in A$. On the one hand, the derivative $f'(x)$ represents the incremental behavior of the function at x; on the other hand, through the differential $df(x)(h) = f'(x)\, h$ the derivative gives a linear approximation of the function at x. Although strictly linked, these two aspects of the notion of derivative are conceptually different.

The first aspect motivated the notion of differential according to Gateaux, as we detailed in the previous section. The second aspect, of linear approximation, leads to the notion of differential according to Frechet, which we will treat in this section.

To introduce the Frechet differential it is necessary to go back to the differential $df(x)$ of a scalar function $f : A \subseteq \mathbb{R} \to \mathbb{R}$. As known, such a function is called differentiable at x if there exists a function $df(x) : \mathbb{R} \to \mathbb{R}$ of the form $df(x)(h) = \chi \cdot h$ for every $h \in \mathbb{R}$, such that

$$f(x+h) = f(x) + df(x)(h) + o(h). \tag{4.15}$$

In other words, f is differentiable at x if it can be approximated in a linear way at x. The term $o(h)$ indicates the degree accuracy of the approximation, which is the better the smaller is h.

Expression (4.15) can be rewritten as

$$\lim_{h\to 0} \frac{f(x+h) - f(x) - df(x)(h)}{h} = 0,$$

or, equivalently, as

$$\lim_{h\to 0} \frac{|f(x+h) - f(x) - df(x)(h)|}{|h|} = 0. \tag{4.16}$$

On the other hand, the differential $df(x) = \chi \cdot h$ is a linear functional on $\mathbb{R}$. Therefore, we can equivalently say that a function $f : A \subseteq \mathbb{R} \to \mathbb{R}$ is differentiable at $x \in A$ when there exists a linear functional $df(x) : \mathbb{R} \to \mathbb{R}$ such that

$$\lim_{h\to 0} \frac{|f(x+h) - f(x) - df(x)(h)|}{|h|} = 0.$$

This formulation is the most useful to generalize the notion of differential to functions of several variables. In fact, at this point it is natural to give the following definition:

Definition 145 *A function $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ is called differentiable according to Frechet at $x \in A$ if there exists a linear functional $df(x) : \mathbb{R}^n \to \mathbb{R}$ such that*

$$\lim_{h\to 0} \frac{|f(x+h) - f(x) - df(x)(h)|}{\|h\|} = 0. \tag{4.17}$$

The functional $df(x)$ is called differential according to Frechet of f at x.

Expression (4.17) is the version for functions of several variables of (4.16), in which the absolute value in the denominator is replaced by the Euclidean norm. Apart from this, the idea is the same: a function is differentiable according to Frechet at x if there exists a linear functional that approximates the function, with accuracy given here by $o(\|h\|)$. Expression (4.15) becomes:

$$f(x+h) = f(x) + df(x)(h) + o(\|h\|). \tag{4.18}$$

We now give a first important property of Frechet differentials.

Proposition 146 *When it exists, the differential $df(x): \mathbb{R}^n \to \mathbb{R}$ is unique.*

Proof To prove the uniqueness it is enough to prove that $df(x)(y) = f'(x;y)$ for every $y \in \mathbb{R}^n$. Given $y \in \mathbb{R}^n$, in (4.17) set $h = ty$. In this case, (4.17) implies:

$$\lim_{t\to 0} \frac{|f(x+ty) - f(x) - df(x)(ty)|}{\|ty\|} = 0,$$

and therefore,

$$\lim_{t\to 0} \left| \frac{f(x+ty) - f(x)}{t} - df(x)(y) \right| = 0,$$

which implies

$$\lim_{t\to 0} \frac{f(x+ty) - f(x)}{t} = df(x)(y).$$

It follows that $df(x)(y) = f'(x;y)$. ■

Before going on, we remind the fundamental result for differentials of functions of one variable: a function $f : A \subseteq \mathbb{R} \to \mathbb{R}$ is differentiable at x if and only if it has a derivative at this point; in this case, $\chi = f'(x)$. Therefore, derivability and differentiability according to Frechet are equivalent properties for functions of one variable and the differential, when it exists, is given by the function $df(x)(h) = f'(x)h$ for each $h \in \mathbb{R}$.

We saw in the previous section that in the case of scalar functions also the differentiability according to Gateaux is equivalent to the ordinary derivability. It follows from this that in the case $n = 1$ the notions of differentiability according to Gateaux and Frechet are equivalent notions. In the case of functions of several variables things become more complicated. We begin with a first result.

Theorem 147 *Let $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ be differentiable according to Frechet at x. Then, it is Gateaux differentiable at this point and the two notions of differential coincide. That is, we have*

$$df(x)(h) = f'(x;h) = \nabla f(x) \cdot h = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x) h_i \tag{4.19}$$

for every $h \in \mathbb{R}^n$.

Proof In the proof of Proposition 146 we saw that, when it exists, we have $df(x)(h) = f'(x;h)$ for each $h \in \mathbb{R}^n$ (here we use h instead of y, but in any case they are mute variables). Since $df(x)$ is by hypothesis a linear functional, it follows that the directional derivative $f'(x;\cdot) : \mathbb{R}^n \to \mathbb{R}$ is itself a linear functional. The function is therefore differentiable according to Gateaux in x and, thanks to Theorem 140, we have

$$df(x)(h) = f'(x;h) = \nabla f(x) \cdot h$$

for every $h \in \mathbb{R}^n$. ■

Therefore, the differentiability according to Frechet implies the one according to Gateaux. The gradient $\nabla f(x)$ is the vector that gives the representation in terms of scalar product of the differential $df(x)$. In an imprecise, but expressive, way (4.19) is often denoted by:

$$df = \frac{\partial f}{\partial x_1} dx_1 + \cdots + \frac{\partial f}{\partial x_n} dx_n, \tag{4.20}$$

which is often called the formula of the *total differential*. This formula shows how the total effect df on f can be decomposed in the sum of the effects that the infinitesimal variations dx_i of the single variables have on f.

For example, if $f : \mathbb{R}^n \to \mathbb{R}$ is a production function with n inputs, (4.20) tells us that the total variation df of the output is the result of the sum of the effects $(\partial f/\partial x_i)\, dx_i$ that the infinitesimal variations dx_i of the single inputs have on the production function. In a more economic language, the total variation of the output df is given by the sum of the infinitesimal variations of the factors dx_i, multiplied by their respective marginal productivites $\partial f/\partial x_i$.

By Theorem 147, expression (4.18) becomes:

$$f(x+h) = f(x) + \nabla f(x) \cdot h + o(\|h\|), \tag{4.21}$$

which in the scalar case $n = 1$ reduces to the well known formula:

$$f(x+h) = f(x) + f'(x) \cdot h + o(h).$$

Together with Corollary 144, Theorem 147 also implies that for the Frechet differentiability the usual algebraic rules, that for brevity we do not state explicitly, hold.

The converse of Theorem 147 is false when $n \geq 2$: next example shows that there exist functions of several variables that are differentiable at a point according to Gateaux, but not according to Frechet. This is an important observation because it indicates a first fundamental difference of the case of several variables with respect to the scalar case.

Example 148 Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined as

$$f(x_1, x_2) = \begin{cases} \frac{x_1^4 x_2^2}{x_1^8 + x_2^4} & \text{if } (x_1, x_2) \neq (0,0), \\ 0 & \text{if } (x_1, x_2) = (0,0). \end{cases}$$

If we set $x = \mathbf{0} = (0,0)$, for every $y \in \mathbb{R}^2$ we have:

$$\begin{aligned} f'(\mathbf{0}; y) &= \lim_{t \to 0+} \frac{f(ty)}{t} = \lim_{t \to 0+} \frac{(ty_1)^4 (ty_2)^2}{t\left[(ty_1)^8 + (ty_2)^4\right]} \\ &= \lim_{t \to 0+} \frac{t^6 y_1^4 y_2^2}{t^5 (t^4 y_1^8 + y_2^4)} = \lim_{t \to 0+} \frac{t y_1^4 y_2^2}{t^4 y_1^8 + y_2^4} = 0. \end{aligned}$$

Therefore, $f'(\mathbf{0};y) = 0$ for every $y \in \mathbb{R}^2$ and the directional derivative in $(0,0)$ is consequently the null linear functional. It follows that f is Gateaux differentiable at $(0,0)$. However, it is not Frechet differentiable at $(0,0)$. In fact, f is not continuous at $(0,0)$, and this implies that f cannot be differentiable at $(0,0)$ because, as Theorem 149 will show, continuity is implied by Frechet differentiability. We show therefore that f is not continuous at $(0,0)$. Consider the points $(t,t^2) \in \mathbb{R}^2$ that lie on the graphic of the parabola $x_2 = x_1^2$. We have

$$f\left(t,t^2\right) = \frac{t^4\left(t^2\right)^2}{t^8 + \left(t^2\right)^4} = \frac{t^4 t^4}{t^8 + t^8} = \frac{1}{2},$$

and therefore along these points the function is constant and takes the value $1/2$. It follows that $\lim_{t\to 0} f(t,t^2) = 1/2$ and, since $f(0,0) = 0$, the function is discontinuous at $(0,0)$. ▲

One of the classic results in the case of functions of one variable is that differentiability implies continuity. In the case of functions of several variables, it is necessary to distinguish between the two notions of differentiability that we have seen. Example 148 has just exhibited a function that is Gateaux differentiable at a point, even if it is discontinuous at this point. Therefore, differentiability according to Gateaux does not imply continuity, a second fundamental difference with respect to the scalar case, in which differentiability implied continuity.

On the contrary, it is true that Frechet differentiability implies continuity, as the following result shows.

Theorem 149 *If a function $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ is Frechet differentiable at a point of A, then it is continuos at this point.*

Proof Let f be Frechet differentiable at $x_0 \in A$. Set $x = x_0 + h$ and therefore $h = x - x_0$. We have $x \to x_0$ if and only if $x - x_0 \to 0$; therefore, using (4.21) we can write:

$$\begin{aligned}\lim_{x\to x_0}\left(f(x) - f(x_0)\right) &= \lim_{x\to x_0} \nabla f(x_0)(x - x_0) + \lim_{x\to x_0} o(\|x - x_0\|) \\ &= \lim_{x - x_0\to 0} \nabla f(x_0)(x - x_0) + \lim_{x - x_0\to 0} o(\|x - x_0\|) = 0.\end{aligned}$$

Since

$$\lim_{x\to x_0}\left(f(x) - f(x_0)\right) = \lim_{x\to x_0} f(x) - \lim_{x\to x_0} f(x_0) = \lim_{x\to x_0} f(x) - f(x_0),$$

it follows that $\lim_{x\to x_0} f(x) = f(x_0)$, as desired. ■

4.3 Classes $\mathcal{C}^1$

In the previous sections we introduced two notions of differentiability. The first one, a la Gateaux, is linked to the incremental behavior of the function at a point, while the second one, a la Frechet, is motivated by the desire to approximate in a linear way the function at a point. The most significant characteristics that we found are:

- Frechet differentiability implies Gateaux differentiability (Theorem 147), but the converse is false when $n \geq 2$ (Example 148).
- Frechet differentiability implies continuity (Theorem 149), while when $n \geq 2$ this is not true for Gateaux differentiability (Example 148).

It is now important to give simple sufficient conditions that imply that a function is differentiable according to these notions, something otherwise difficult to verify using directly the definitions (especially in the case of Frechet).

The most important result of this type gives a simple sufficient condition for Frechet differentiability, and therefore also according to Gateaux. To state it, observe that partial derivatives $\partial f/\partial x_i$ can be seen as functions $\partial f/\partial x_i : A_i \subseteq \mathbb{R}^n \to \mathbb{R}$, where A_i is the set in which these partial derivatives exist. Being functions, it is meaningful to talk of continuity of the partial derivatives at a point $x \in A_i$. In Example 137 we saw that the partial derivatives of $f(x_1, x_2, x_3) = x_1 x_2 x_3$ are the functions $x_2 x_3$, $x_1 x_3$ and $x_1 x_2$. In this case, each of the partial derivatives is continuous on all $\mathbb{R}^3$.

Theorem 150 *Let $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ be a function that has partial derivatives in a neighborhood of the point $x \in A$. If these derivatives are continuos at x, then f is Frechet differentiable at x.*

Proof For simplicity of notation, we consider the case in which $n = 2$, f is defined on all $\mathbb{R}^2$, and the partial derivatives $\partial f/\partial x_1$ and $\partial f/\partial x_2$ exist on all $\mathbb{R}^2$. Apart from the complication of notation, the general case can be proved in a similar way.

Let therefore $f : \mathbb{R}^2 \to \mathbb{R}$ and $x \in \mathbb{R}^2$. Assume that $\partial f/\partial x_1$ and $\partial f/\partial x_2$ are both continuous at x. Adding and subtracting $f(x_1 + h_1, x_2)$, for each $h \in \mathbb{R}^2$ we have:

$$\begin{aligned} & f(x+h) - f(x) \qquad (4.22) \\ = \; & f(x_1 + h_1, x_2) - f(x_1, x_2) + f(x_1 + h_1, x_2 + h_2) - f(x_1 + h_1, x_2). \end{aligned}$$

The partial derivative $\partial f/\partial x_1(x)$ is the derivative of the function $\phi_1 : \mathbb{R} \to \mathbb{R}$ defined as $\phi_1(x_1) = f(x_1, x_2)$, in which x_2 is considered as a constant. By the Mean Value

Theorem, there exists $z_1 \in (x_1, x_1 + h_1) \subseteq \mathbb{R}$ such that

$$\begin{aligned} \phi_1'(z_1) &= \frac{\phi_1(x_1 + h_1) - \phi_1(x_1)}{x_1 + h_1 - x_1} = \frac{\phi_1(x_1 + h_1) - \phi_1(x_1)}{h_1} \\ &= \frac{f(x_1 + h_1, x_2) - f(x_1, x_2)}{h_1}. \end{aligned}$$

Similarly, the partial derivative $\partial f/\partial x_2(x + h)$ is the derivative of the function $\phi_2 : \mathbb{R} \to \mathbb{R}$ defined as $\phi_2(x_2) = f(x_1 + h_1, x_2)$, in which $x_1 + h_1$ is considered as a constant. Again by the Mean Value Theorem, there exists $z_2 \in (x_2, x_2 + h_2) \subseteq \mathbb{R}$ such that

$$\begin{aligned} \phi_2'(z_2) &= \frac{\phi_2(x_2 + h_2) - \phi_2(x_2)}{x_2 + h_2 - x_2} = \frac{\phi_2(x_2 + h_2) - \phi_2(x_2)}{h_2} \\ &= \frac{f(x_1 + h_1, x_2 + h_2) - f(x_1 + h_1, x_2)}{h_2}. \end{aligned}$$

Since by construction $\partial f/\partial x_1(z_1, x_2) = \phi_1'(z_1)$ and $\partial f/\partial x_2(x_1 + h_1, z_2) = \phi_2'(z_2)$, we can rewrite (4.22) as:

$$f(x + h) - f(x) = \frac{\partial f}{\partial x_1}(z_1, x_2) h_1 + \frac{\partial f}{\partial x_2}(x_1 + h_1, z_2) h_2.$$

On the other hand, by definition $\nabla f(x) \cdot h = \partial f/\partial x_1(x_1, x_2) h_1 + \partial f/\partial x_2(x_1, x_2) h_2$. Thus:

$$\begin{aligned} & \lim_{h\to 0} \frac{|f(x + h) - f(x) - \nabla f(x) \cdot h|}{\|h\|} \\ = & \lim_{h\to 0} \frac{\left| \frac{\partial f}{\partial x_1}(z_1, x_2) h_1 + \frac{\partial f}{\partial x_2}(x_1 + h_1, z_2) h_2 - \left(\frac{\partial f}{\partial x_1}(x_1, x_2) h_1 + \frac{\partial f}{\partial x_2}(x_1, x_2) h_2 \right) \right|}{\|h\|} \\ = & \lim_{h\to 0} \frac{\left| \left(\frac{\partial f}{\partial x_1}(z_1, x_2) - \frac{\partial f}{\partial x_1}(x_1, x_2) \right) h_1 + \left(\frac{\partial f}{\partial x_2}(x_1 + h_1, z_2) - \frac{\partial f}{\partial x_2}(x_1, x_2) \right) h_2 \right|}{\|h\|} \\ \leq & \lim_{h\to 0} \frac{\left| \left(\frac{\partial f}{\partial x_1}(z_1, x_2) - \frac{\partial f}{\partial x_1}(x_1, x_2) \right) h_1 \right|}{\|h\|} + \lim_{h\to 0} \frac{\left| \left(\frac{\partial f}{\partial x_2}(x_1 + h_1, z_2) - \frac{\partial f}{\partial x_2}(x_1, x_2) \right) h_2 \right|}{\|h\|} \\ = & \lim_{h\to 0} \left| \left(\frac{\partial f}{\partial x_1}(z_1, x_2) - \frac{\partial f}{\partial x_1}(x_1, x_2) \right) \right| \frac{|h_1|}{\|h\|} + \lim_{h\to 0} \left| \left(\frac{\partial f}{\partial x_2}(x_1 + h_1, z_2) - \frac{\partial f}{\partial x_2}(x_1, x_2) \right) \right| \frac{|h} \\ \leq & \lim_{h\to 0} \left| \left(\frac{\partial f}{\partial x_1}(z_1, x_2) - \frac{\partial f}{\partial x_1}(x_1, x_2) \right) \right| + \lim_{h\to 0} \left| \left(\frac{\partial f}{\partial x_2}(x_1 + h_1, z_2) - \frac{\partial f}{\partial x_2}(x_1, x_2) \right) \right|, \end{aligned}$$

where the last inequality holds because

$$0 \leq \frac{|h_1|}{\|h\|} \leq 1 \text{ and } 0 \leq \frac{|h_2|}{\|h\|} \leq 1.$$

On the other hand, since $z_1 \in (x_1, x_1 + h_1)$ and $z_2 \in (x_2, x_2 + h_2)$, we have $z_1 \to x_1$ for $h_1 \to 0$ and $z_2 \to x_2$ for $h_2 \to 0$. Therefore, being $\partial f/\partial x_1$ and $\partial f/\partial x_2$ both continuous at x, we have

$$\lim_{h\to 0} \frac{\partial f}{\partial x_1}(z_1, x_2) = \frac{\partial f}{\partial x_1}(x_1, x_2) \text{ and } \lim_{h\to 0} \frac{\partial f}{\partial x_2}(x_1 + h_1, z_2) = \frac{\partial f}{\partial x_2}(x_1, x_2),$$

which implies

$$\begin{aligned}\lim_{h\to 0}\left|\left(\frac{\partial f}{\partial x_1}(z_1,x_2)-\frac{\partial f}{\partial x_1}(x_1,x_2)\right)\right| &= 0,\\ \lim_{h\to 0}\left|\left(\frac{\partial f}{\partial x_2}(x_1+h_1,z_2)-\frac{\partial f}{\partial x_2}(x_1,x_2)\right)\right| &= 0.\end{aligned}$$

In conclusion, we have proved that

$$\lim_{h\to 0}\frac{|f(x+h)-f(x)-\nabla f(x)\cdot h|}{\|h\|}=0,$$

and the function f is thus Frechet differentiable at x. ■

The importance of the sufficient condition given by Theorem 150 lies in its easy verifiability. In fact, it is sufficient to compute the partial derivatives and to verify their continuity at the given point, something much easier than to verify directly the definition of Frechet differentiability, as shown by the next examples.

Example 151 Going back again to the function $f(x_1,x_2,x_3)=x_1x_2x_3$ of Example 137, we already saw that the partial derivatives

$$\frac{\partial f}{\partial x_1}(x)=x_2x_3,\quad \frac{\partial f}{\partial x_2}(x)=x_1x_3,\quad \frac{\partial f}{\partial x_3}(x)=x_1x_2,$$

are continuous functions on all $\mathbb{R}^3$. By Theorem 150, f is therefore Frechet differentiable at each point of $\mathbb{R}^3$. ▲

Example 152 Consider the function $f(x_1,x_2,x_3,x_4)=x_1+e^{x_2x_3}+2x_4^2$ of Example 138. We saw that

$$\frac{\partial f}{\partial x_1}(x)=1,\quad \frac{\partial f}{\partial x_2}(x)=x_3e^{x_2x_3},\quad \frac{\partial f}{\partial x_3}(x)=x_2e^{x_2x_3},\quad \frac{\partial f}{\partial x_4}(x)=4x_4,$$

and therefore the partial derivatives are continuous on all $\mathbb{R}^4$. By Theorem 150, f is Frechet differentiable at each point of $\mathbb{R}^4$. ▲

Example 153 Consider the function $f(x_1,x_2)=\lg(x_1-x_2)$, whose domain is the open set $A=\{x\in\mathbb{R}^2 : x_1>x_2\}$. For each $x\in A$, we have:

$$\frac{\partial f}{\partial x_1}(x)=\frac{1}{x_1-x_2}\quad\text{and}\quad\frac{\partial f}{\partial x_2}(x)=\frac{1}{x_2-x_1}.$$

The partial derivatives are therefore continuous on the entire domain A; by Theorem 150, f is Frechet differentiable at each point of A. ▲

Before going on, we give an example that shows how the condition contained in Theorem 150 is indeed only sufficient.

Example 154 Consider the function $f : \mathbb{R} \to \mathbb{R}$ defined as:

$$f(x) = \begin{cases} x^2 \sin \frac{1}{x} & x \neq 0, \\ 0 & x = 0. \end{cases}$$

As the reader can verify, we have

$$f'(x) = \begin{cases} 2x \sin \frac{1}{x} - \cos \frac{1}{x} & x \neq 0, \\ 0 & x = 0. \end{cases}$$

Consider the point $x = 0$. The first derivative f' is discontinuous at 0, and so the condition of Theorem 150 is violated. Nevertheless, the function has a derivative at 0 and it is therefore Frechet differentiable at this point. ▲

Many functions of interest have continuous derivative on their entire domain and, by Theorem 150, are therefore Frechet differentiable at each point of their domain. They are therefore functions that behave very well with respect to differentiability and, for this reason, it is useful to give a name to this class of functions.

Definition 155 *A function $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ is called of class $\mathcal{C}^1$ if it has partial derivatives that are continuous on its domain A. The set of all these functions is denoted by $\mathcal{C}^1(A)$.*

By Theorem 150, each function of class $\mathcal{C}^1$ is Frechet differentiable at each point of A. For example, all the functions seen in Examples 9, 10 and 11 are of class $\mathcal{C}^1$.

The way we stated Theorem 150 is the one operationally useful when the Frechet differentiability of a function has to be checked. But, we can restate Theorem 150 in a different form, which is conceptually important. To see it, observe that an application $f : A \subseteq \mathbb{R}^n \to \mathbb{R}^m$ can be regarded as a m-tuple $(f_1, ..., f_m)$ of functions defined on A and with values in $\mathbb{R}$:

$$\begin{aligned} y_1 &= f_1(x_1, ..., x_n), \\ y_2 &= f_2(x_1, ..., x_n), \\ &\dots \\ y_m &= f_m(x_1, ..., x_n). \end{aligned}$$

In particular, an application $f = (f_1, ..., f_m) : A \subseteq \mathbb{R}^n \to \mathbb{R}^m$ is continuous at $x \in A$ if and only if each $f_i : A \subseteq \mathbb{R}^n \to \mathbb{R}$ is continuous at $x \in A$.

Example 156 Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be defined by $f(x_1, x_2) = (x_1, x_1 x_2)$ for each vector $(x_1, x_2) \in \mathbb{R}^2$. For example, if $(x_1, x_2) = (2, 5)$, then $f(x_1, x_2) = (2, 2 \cdot 5) = (2, 10) \in \mathbb{R}^2$. In this case we have:

$$\begin{aligned} f_1(x_1, x_2) &= x_1, \\ f_2(x_1, x_2) &= x_1 x_2. \end{aligned}$$

Since both f_1 and f_2 are continuous on $\mathbb{R}^2$, the function f is also continuous on $\mathbb{R}^2$. ▲

Suppose $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ has partial derivatives in a neighborhood $B_\varepsilon(x)$ of the point $x \in A$. Then, we can write

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, ..., \frac{\partial f}{\partial x_n} \right) : B_\varepsilon(x) \subseteq \mathbb{R}^n \to \mathbb{R}^n,$$

that is, the gradient can be regarded as a function from a subset of $\mathbb{R}^n$ to $\mathbb{R}^n$ that has the partial derivatives as its components. In particular, ∇f is continuous at x if and only if each partial derivative $\partial f / \partial x_i : B_\varepsilon(x) \subseteq \mathbb{R}^n \to \mathbb{R}$ is continuous.

In view of all this we can restate Theorem 150 as follows, where the sufficient condition for Frechet differentiability is viewed as a continuity condition of the gradient mapping.

Theorem 157 *Let $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ be a function such that its gradient ∇f is well defined in a neighborhood of the point $x \in A$. If ∇f is continuous at x, then f is Frechet differentiable at x.*

In a similar vein, we can say that a function $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ is of class $\mathcal{C}^1$ if its gradient mapping ∇f is well defined and continuous on its domain A.

4.4 Differential of Applications

The notions of Gateaux and Frechet differentials can be easily extended to functions $f : A \subseteq \mathbb{R}^n \to \mathbb{R}^m$, that is, to applications. For brevity, we limit here to study the Frechet differential, as the Gateaux case can be studied along similar lines.

4.4.1 Definition and Representation

We start by giving the extension of the definition of Frechet differential to the case of applications.

Definition 158 *An application $f : A \subseteq \mathbb{R}^n \to \mathbb{R}^m$ is said to be differentiable according to Frechet at $x \in A$ if there exists a linear application $df(x) : \mathbb{R}^n \to \mathbb{R}^m$ such that*

$$\lim_{h \to 0} \frac{\|f(x+h) - f(x) - df(x)(h)\|}{\|h\|} = 0. \tag{4.23}$$

The application $df(x)$ is said to be Frechet differential of f at x.

This definition generalizes Definition 145, that is, the special case $m = 1$. The linear approximation is now given by a linear application with values in $\mathbb{R}^m$, while at the numerator of the incremental ratio in (4.23) we find the Euclidean norm instead of the absolute value because we now have to deal with vectors in $\mathbb{R}^m$.

The Frechet differential for applications satisfies properties that are similar to those seen in the case $m = 1$ in Proposition 146 and in Theorems 149 and 150. Naturally, instead of the vector representation of Theorem 147 we have now a more general matrix representation. To see its form, we introduce the Jacobian matrix. Recall how we just observed that an application $f : A \subseteq \mathbb{R}^n \to \mathbb{R}^m$ can be regarded as a m-tuple $(f_1, ..., f_m)$ of functions defined on A and with values in $\mathbb{R}$. The Jacobian matrix $Df(x)$ of an application $f : A \subseteq \mathbb{R}^n \to \mathbb{R}^m$ at $x \in A$ is then a matrix $m \times n$ given by:

$$Df(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x) & \frac{\partial f_1}{\partial x_2}(x) & \cdots & \frac{\partial f_1}{\partial x_n}(x) \\ \frac{\partial f_2}{\partial x_1}(x) & \frac{\partial f_2}{\partial x_2}(x) & \cdots & \frac{\partial f_2}{\partial x_n}(x) \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f_m}{\partial x_1}(x) & \frac{\partial f_m}{\partial x_2}(x) & \cdots & \frac{\partial f_m}{\partial x_n}(x) \end{bmatrix}$$

that is,

$$Df(x) = \begin{bmatrix} \nabla f_1(x) \\ \nabla f_2(x) \\ \cdots \\ \nabla f_m(x) \end{bmatrix}. \tag{4.24}$$

In Example 156 we have

$$Df(x) = \begin{bmatrix} 1 & 0 \\ x_2 & x_1 \end{bmatrix}.$$

We can now give the matrix representation of Frechet differentials, which shows that the Jacobian matrix $Df(x)$ is the matrix associated to the linear application $df(x)$. This representation generalizes the vector representation given in Theorem 147 because it is immediate to see from (4.24) that the Jacobian matrix $Df(x)$ reduces to the gradient $\nabla f(x)$ in the special case $m = 1$.

Theorem 159 *Let $f : A \subseteq \mathbb{R}^n \to \mathbb{R}^m$ be Frechet differentiable at $x \in A$. Then,*

$$df(x)(h) = Df(x)\, h, \qquad \forall h \in \mathbb{R}^n.$$

Proof We begin by observing a property of the Euclidean norm. Let $x = (x_1, ..., x_n) \in \mathbb{R}^n$. For every $j = 1, .., n$ we have:

$$|x_j| = \sqrt{x_j^2} \leq \sqrt{\sum_{j=1}^{n} x_j^2} = \|x\|. \tag{4.25}$$

Assume that f is differentiable at $x \in A$. Set $h = te^j$ with $j = 1, .., n$. By definition,

$$\lim_{t \to 0} \frac{\|f(x + te^j) - f(x) - df(x)(te^j)\|}{\|te^j\|} = 0,$$

and therefore, being $\|te^j\| = |t|$, we have

$$\lim_{t \to 0} \left\| \frac{f(x + te^j) - f(x)}{|t|} - df(x)\left(e^j\right) \right\| = 0. \tag{4.26}$$

>From inequality (4.25), for each $i = 1, ..., m$ we have

$$\left| \frac{f_i(x + te^j) - f_i(x)}{|t|} - df_i(x)\left(e^j\right) \right| \leq \left\| \frac{f(x + te^j) - f(x)}{|t|} - df(x)\left(e^j\right) \right\|.$$

Together with (4.26), this implies

$$\lim_{t \to 0} \left| \frac{f_i(x + te^j) - f_i(x)}{|t|} - df_i(x)\left(e^j\right) \right| = 0$$

for each $i = 1, ...m$. We can therefore conclude that for every $i = 1, ..., m$ and every $j = 1, ..., n$ we have:

$$\frac{\partial f_i}{\partial x_j}(x) = \lim_{t \to 0} \frac{f_i(x + te^j) - f_i(x)}{t} = df_i(x)\left(e^j\right). \tag{4.27}$$

In the proof of Theorem 93 of the previous chapter we showed that the matrix associated to a linear application $f : \mathbb{R}^n \to \mathbb{R}^m$ was

$$A = \left[f\left(e^1\right), f\left(e^2\right), ..., f\left(e^n\right)\right].$$

In our case, thanks to (4.27) we therefore have

$$\begin{aligned} A &= \left[df(x)\left(e^1\right), ..., df(x)\left(e^n\right)\right] \\ &= \begin{bmatrix} df_1(x)(e^1) & df_1(x)(e^2) & \cdots & df_1(x)(e^n) \\ df_2(x)(e^1) & df_2(x)(e^2) & \cdots & df_2(x)(e^n) \\ \cdots & \cdots & \cdots & \cdots \\ df_m(x)(e^1) & df_m(x)(e^2) & \cdots & df_m(x)(e^n) \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x) & \frac{\partial f_1}{\partial x_2}(x) & \cdots & \frac{\partial f_1}{\partial x_n}(x) \\ \frac{\partial f_2}{\partial x_1}(x) & \frac{\partial f_2}{\partial x_2}(x) & \cdots & \frac{\partial f_2}{\partial x_n}(x) \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f_m}{\partial x_1}(x) & \frac{\partial f_m}{\partial x_2}(x) & \cdots & \frac{\partial f_m}{\partial x_n}(x) \end{bmatrix} = Df(x), \end{aligned}$$

as desired. ■

We have thus shown how the Jacobian matrix $Df(x)$ is the matrix associated to the linear application $df(x)$. We have observed that when $m = 1$ we have $Df(x) = \nabla f(x)$, and so Theorem 159 generalizes Theorem 147 to the case of applications. We illustrate what we have done til now with some examples.

Example 160 Let $f : \mathbb{R}^3 \to \mathbb{R}^2$ be defined by $f(x_1, x_2, x_3) = (2x_1^2 + x_2 + x_3, x_1 - x_2^4)$ for each vector $x \in \mathbb{R}^3$. For example, if $x = (2, 5, -3)$, then $f(x_1, x_2, x_3) = (2 \cdot 4 + 5 - 3, 2 - (10, -623) \in \mathbb{R}^2$. We have:

$$\begin{aligned} f_1(x_1, x_2, x_3) &= 2x_1^2 + x_2 + x_3, \\ f_2(x_1, x_2, x_3) &= x_1 - x_2^4. \end{aligned}$$

and therefore

$$Df(x) = \begin{bmatrix} 4x_1 & 1 & 1 \\ 1 & -4x_2^3 & 0 \end{bmatrix}.$$

By Theorem 159, the Frechet differential at x is given by the linear application $df(x) : \mathbb{R}^3 \to \mathbb{R}^2$ defined by

$$df(x)(h) = Df(x)h = \left(4x_1h_1 + h_2 + h_3, h_1 - 4x_2^3h_2\right)$$

for each $h \in \mathbb{R}^3$. For example, at $x = (2, 5, -3)$ we have:

$$df(x)(h) = (8h_1 + h_2 + h_3, h_1 - 500h_2).$$

▲

Example 161 Let $f : \mathbb{R} \to \mathbb{R}^3$ be defined by $f(x) = (x, \sin x, \cos x)$ for each $x \in \mathbb{R}$. For example, if $x = \pi$, then $f(x) = (\pi, 0, -1) \in \mathbb{R}^3$. We have:

$$\begin{aligned} f_1(x) &= x, \\ f_2(x) &= \sin x, \\ f_3(x) &= \cos x, \end{aligned}$$

and so

$$Df(x) = \begin{bmatrix} 1 \\ \cos x \\ -\sin x \end{bmatrix}$$

By Theorem 159, the Frechet differential at x is given by the linear application $df(x) : \mathbb{R} \to \mathbb{R}^3$ defined by

$$df(x)(h) = Df(x)h = (h, h\cos x, -h\sin x)$$

for each $h \in \mathbb{R}$. For example, at $x = \pi$ we have:

$$df(x)(h) = (h, -h, 0).$$

▲

Example 162 Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be the linear application defined as $f(x) = Ax$ for each $x \in \mathbb{R}^n$, with

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

Let $a_1, ..., a_m$ be the row vectors, that is $a_1 = (a_{11}, a_{12}, ..., a_{1n}), ..., a_m = (a_{m1}, a_{m2}, ..., a_{mn})$. We have:

$$\begin{aligned} f_1(x_1, ..., x_n) &= a_1 \cdot x = a_{11}x_1 + \cdots + a_{1n}x_n, \\ f_2(x_1, ..., x_n) &= a_2 \cdot x = a_{21}x_1 + \cdots + a_{2n}x_n, \\ &\cdots \\ f_m(x_1, ..., x_n) &= a_m \cdot x = a_{m1}x_1 + \cdots + a_{mn}x_n, \end{aligned}$$

which implies $Df(x) = A$. Hence, the Jacobian matrix of a linear application coincides with the associated matrix A. By Theorem 159, the Frechet differential at x is therefore given by the linear application Ah itself. This naturally generalizes the well known result that for scalar functions of the form $f(x) = ax$, with $a \in \mathbb{R}$, the differential is $df(x)(h) = ah$. ▲

4.4.2 Chain Rule

One of the most useful rules of derivation in Calculus is that about compound scalar functions $f \circ g$, which says that $(f \circ g)'(x) = f'(g(x)) g'(x)$. We now to generalize this rule to the case of composition of applications. In this more general context, it is known as the *chain rule*.

Theorem 163 *Let $g : A \subseteq \mathbb{R}^n \to \mathbb{R}^m$ and $f : B \subseteq \mathbb{R}^m \to \mathbb{R}^q$ with $g(A) \subseteq B$. If g is Frechet differentiable at $x \in A$ and if f is Frechet differentiable at $g(x)$, then the composition $f \circ g : A \subseteq \mathbb{R}^n \to \mathbb{R}^q$ is Frechet differentiable at x, with:*

$$D(f \circ g)(x) = Df(g(x)) Dg(x). \tag{4.28}$$

The proof is based on this lemma.

Lemma 164 *Given a linear application $T : \mathbb{R}^n \to \mathbb{R}^m$, there exists a constant $k > 0$ such that $\|T(x)\| \leq k\|x\|$ for every $x \in \mathbb{R}^n$.*

Proof Set $k = \sum_{i=1}^{n} \|T(e^i)\|$. We have:

$$\|T(x)\| = \left\| T\left(\sum_{i=1}^{n} x_i e^i\right) \right\| = \left\| \sum_{i=1}^{n} x_i T\left(e^i\right) \right\| \leq \sum_{i=1}^{n} |x_i| \cdot \left\| T\left(e^i\right) \right\|.$$

By (4.25), we have $|x_i| \leq \|x\|$ for each $i = 1, ..., n$. Therefore,

$$\sum_{i=1}^{n} |x_i| \cdot \left\| T\left(e^i\right) \right\| \leq \sum_{i=1}^{n} \|x\| \cdot \left\| T\left(e^i\right) \right\| = \|x\| \sum_{i=1}^{n} \left\| T\left(e^i\right) \right\| = k\|x\|,$$

which implies $\|T(x)\| \leq k\|x\|$, as desired. ■

We can now prove Theorem 163.

Proof It is sufficient to prove that

$$d(f \circ g)(x) = df(g(x))\, dg(x), \tag{4.29}$$

where the right-hand side is the product of the linear applications $df(g(x))$ and $dg(x)$. In fact, by Theorem 98 of the previous chapter, the matrix representation of the product linear application $df(g(x))\, dg(x)$ is given by the product matrix $Df(g(x))\, Dg(x)$. Therefore, (4.29) implies (4.28).

We show therefore that (4.29) holds. In other words, we must show that

$$\lim_{h \to 0} \frac{\|(f \circ g)(x+h) - (f \circ g)(x) - (df(g(x))\, dg(x))(h)\|}{\|h\|} = 0. \tag{4.30}$$

Set

$$\begin{aligned} \phi(h) &= g(x+h) - g(x) - dg(x)(h), \\ \psi(k) &= f(g(x)+k) - f(g(x)) - df(g(x))(k). \end{aligned}$$

We have

$$\begin{aligned} & (f \circ g)(x+h) - (f \circ g)(x) - (df(g(x))\, dg(x))(h) \\ = \; & f(g(x+h)) - f(g(x)) - df(g(x))(dg(x)(h)) \\ = \; & f(g(x+h)) - f(g(x)) - df(g(x))(g(x+h) - g(x) - \phi(h)) \\ = \; & f(g(x+h)) - f(g(x)) - df(g(x))(g(x+h) - g(x)) + df(g(x))(\phi(h)) \\ = \; & \psi(g(x+h) - g(x)) + df(g(x))(\phi(h)). \end{aligned}$$

To prove (4.30) thus amounts to proving that

$$\lim_{h\to 0}\frac{\|\psi\left(g\left(x+h\right)-g\left(x\right)\right)+df\left(g\left(x\right)\right)\left(\phi\left(h\right)\right)\|}{\|h\|}=0. \tag{4.31}$$

Consider the linear application $df\left(g\left(x\right)\right)$. By Lemma 164, there exists $k>0$ such that $\|df\left(g\left(x\right)\right)\left(h\right)\|\leq k\|h\|$ for each $h\in\mathbb{R}^m$. Since $\phi\left(h\right)\in\mathbb{R}^m$ for each $h\in\mathbb{R}^n$, we therefore have $\|df\left(g\left(x\right)\right)\left(\phi\left(h\right)\right)\|\leq k\|\phi\left(h\right)\|$. On the other hand, g is differentiable at x, and so

$$\lim_{h\to 0}\frac{\|\phi\left(h\right)\|}{\|h\|}=0.$$

It follows that

$$\lim_{h\to 0}\frac{\|df\left(g\left(x\right)\right)\left(\phi\left(h\right)\right)\|}{\|h\|}\leq k\lim_{h\to 0}\frac{\|\phi\left(h\right)\|}{\|h\|}=0. \tag{4.32}$$

Since f is differentiable at $g\left(x\right)$, we have

$$\lim_{k\to 0}\frac{\|\psi\left(k\right)\|}{\|k\|}=0. \tag{4.33}$$

Fix $\varepsilon>0$. By (4.33), there exists $\delta_\varepsilon>0$ such that $\|k\|\leq\delta_\varepsilon$ implies $\|\psi\left(k\right)\|/\|k\|\leq\varepsilon$. In other words, there exists $\delta_\varepsilon>0$ such that $\|g\left(x+h\right)-g\left(x\right)\|\leq\delta_\varepsilon$ implies

$$\frac{\|\psi\left(g\left(x+h\right)-g\left(x\right)\right)\|}{\|g\left(x+h\right)-g\left(x\right)\|}\leq\varepsilon.$$

On the other hand, since g is continuous at x, there exists $\delta^1>0$ such that $\|h\|\leq\delta^1$ implies $\|g\left(x+h\right)-g\left(x\right)\|\leq\delta_\varepsilon$. Therefore, for $\|h\|$ sufficiently small we have

$$\|\psi\left(g\left(x+h\right)-g\left(x\right)\right)\|\leq\varepsilon\|g\left(x+h\right)-g\left(x\right)\|.$$

By applying Lemma 164 to the linear application $dg\left(x\right)$, there exists $k>0$ such that

$$\begin{aligned}\|\psi\left(g\left(x+h\right)-g\left(x\right)\right)\| &\leq \varepsilon\|g\left(x+h\right)-g\left(x\right)\| \\ &\leq \varepsilon\|\phi\left(h\right)+dg\left(x\right)\left(h\right)\| \\ &\leq \varepsilon\|\phi\left(h\right)\|+\varepsilon\|dg\left(x\right)\left(h\right)\|\leq\varepsilon\|\phi\left(h\right)\|+\varepsilon k\|h\|.\end{aligned} \tag{4.34}$$

Go back to (4.31). Using (4.32) and (4.34), we have:

$$\begin{aligned}&\lim_{h\to 0}\frac{\|\psi\left(g\left(x+h\right)-g\left(x\right)\right)+df\left(g\left(x\right)\right)\left(\phi\left(h\right)\right)\|}{\|h\|} \\ \leq\ &\lim_{h\to 0}\frac{\|\psi\left(g\left(x+h\right)-g\left(x\right)\right)\|}{\|h\|}+\lim_{h\to 0}\frac{\|df\left(g\left(x\right)\right)\left(\phi\left(h\right)\right)\|}{\|h\|} \\ \leq\ &\varepsilon\lim_{h\to 0}\frac{\|\phi\left(h\right)\|}{\|h\|}+\varepsilon k\lim_{h\to 0}\frac{\|h\|}{\|h\|}=\varepsilon k.\end{aligned}$$

Since ε was fixed arbitrarily, it can be taken as small as we like. Therefore:

$$\lim_{h\to 0}\frac{\|\psi\left(g\left(x+h\right)-g\left(x\right)\right)+df\left(g\left(x\right)\right)\left(\phi\left(h\right)\right)\|}{\|h\|}\leq k\lim_{\varepsilon\to 0}\varepsilon=0,$$

as desired. ■

In the scalar case $n=m=q=1$, (4.28) reduces to the classic rule

$$\left(f\circ g\right)'\left(x\right)=f'\left(g\left(x\right)\right)g'\left(x\right).$$

Another important special case is when $q=1$. In this case we have $f:B\subseteq\mathbb{R}^m\to\mathbb{R}$ and $g=(g_1,...,g_m):A\subseteq\mathbb{R}^n\to\mathbb{R}^m$, with $g\left(A\right)\subseteq B$. For the compound function $f\circ g:A\subseteq\mathbb{R}^n\to\mathbb{R}$ the chain rule (4.28) takes the form:

$$\begin{aligned}
&\nabla\left(f\circ g\right)\left(x\right)\\
=\;&\nabla f\left(g\left(x\right)\right)Dg\left(x\right)\\
=\;&\left(\frac{\partial f}{\partial x_1}\left(g\left(x\right)\right),...,\frac{\partial f}{\partial x_m}\left(g\left(x\right)\right)\right)\begin{bmatrix}\frac{\partial g_1}{\partial x_1}\left(x\right) & \frac{\partial g_1}{\partial x_2}\left(x\right) & \cdots & \frac{\partial g_1}{\partial x_n}\left(x\right)\\ \frac{\partial g_2}{\partial x_1}\left(x\right) & \frac{\partial g_2}{\partial x_2}\left(x\right) & \cdots & \frac{\partial g_2}{\partial x_n}\left(x\right)\\ \cdots & \cdots & \cdots & \cdots\\ \frac{\partial g_m}{\partial x_1}\left(x\right) & \frac{\partial g_m}{\partial x_2}\left(x\right) & \cdots & \frac{\partial g_m}{\partial x_n}\left(x\right)\end{bmatrix}\\
=\;&\left(\sum_{i=1}^{m}\frac{\partial f}{\partial x_i}\left(g\left(x\right)\right)\frac{\partial g_i}{\partial x_1}\left(x\right),...,\sum_{i=1}^{m}\frac{\partial f}{\partial x_i}\left(g\left(x\right)\right)\frac{\partial g_i}{\partial x_n}\left(x\right)\right).
\end{aligned}$$

As to the differential, for each $h\in\mathbb{R}^n$ we have

$$\begin{aligned}
d\left(f\circ g\right)\left(x\right)\left(h\right)&=\nabla\left(f\circ g\right)\left(x\right)\cdot h\\
&=\sum_{i=1}^{m}\frac{\partial f}{\partial x_i}\left(g\left(x\right)\right)\frac{\partial g_i}{\partial x_1}\left(x\right)h_1+\cdots+\sum_{i=1}^{m}\frac{\partial f}{\partial x_i}\left(g\left(x\right)\right)\frac{\partial g_i}{\partial x_n}\left(x\right)h_n.
\end{aligned}$$

Grouping the terms for $\partial f/\partial x_i$, we get the following equivalent form:

$$d\left(f\circ g\right)\left(x\right)\left(h\right)=\sum_{i=1}^{n}\frac{\partial f}{\partial x_1}\left(g\left(x\right)\right)\frac{\partial g_1}{\partial x_i}\left(x\right)h_i+\cdots+\sum_{i=1}^{n}\frac{\partial f}{\partial x_m}\left(g\left(x\right)\right)\frac{\partial g_m}{\partial x_i}\left(x\right)h_i,$$

which can be reformulated in the following imprecise, but expressive way:

$$d\left(f\circ g\right)=\sum_{i=1}^{n}\left(\frac{\partial f}{\partial g_1}\frac{\partial g_1}{\partial x_i}dx_i+\cdots+\frac{\partial f}{\partial g_m}\frac{\partial g_m}{\partial x_i}dx_i\right).\tag{4.35}$$

This is the formula of the total differential for the compound function $f\circ g$. The total variation $d\left(f\circ g\right)$ of $f\circ g$ is the result of the sum of the effects on the function f of the variations of the single functions g_i determined by infinitesimal variations dx_i of the different variables.

In the next two examples we consider two important subcases of the case $q=1$.

Example 165 Suppose that, besides $q = 1$, we have $n = 1$. Let $f : B \subseteq \mathbb{R}^m \to \mathbb{R}$ and $g : A \subseteq \mathbb{R} \to \mathbb{R}^m$, with $g(A) \subseteq B$. The compound function $f \circ g : A \subseteq \mathbb{R} \to \mathbb{R}$ is scalar and for this function we have:

$$\begin{aligned}(f \circ g)'(x) &= \nabla f(g(x)) Dg(x) = \left(\frac{\partial f}{\partial x_1}(g(x)), ..., \frac{\partial f}{\partial x_m}(g(x))\right) \begin{bmatrix} \frac{dg_1}{dx}(x) \\ \cdots \\ \frac{dg_m}{dx}(x) \end{bmatrix} \\ &= \sum_{i=1}^{m} \frac{\partial f}{\partial x_i}(g(x)) \frac{dg_i}{dx}(x).\end{aligned}$$

The differential is

$$d(f \circ g)(x)(h) = \sum_{i=1}^{m} \frac{\partial f}{\partial x_i}(g(x)) \frac{dg_i}{dx}(x) h$$

for each $h \in \mathbb{R}$, and the total differential (4.35) becomes:

$$d(f \circ g) = \frac{\partial f}{\partial g_1}\frac{dg_1}{dx}dx + \cdots + \frac{\partial f}{\partial g_m}\frac{dg_m}{dx}dx.$$

For example, let $f : \mathbb{R}^m \to \mathbb{R}$ be a production function whose m inputs depend on a common parameter, the time t, which indicates the availability of the different inputs at t.

Inputs are therefore represented by the function $g = (g_1, ..., g_m) : \mathbb{R} \to \mathbb{R}^m$, where $g_i(t)$ denotes what is the quantity of input i at time t. The composition $f \circ g : \mathbb{R} \to \mathbb{R}$ is a scalar function that tells us how the output varies according to the parameter t. We have

$$d(f \circ g) = \frac{\partial f}{\partial g_1}\frac{dg_1}{dt}dt + \cdots + \frac{\partial f}{\partial g_m}\frac{dg_m}{dt}dt, \tag{4.36}$$

that is, the total variation $d(f \circ g)$ of the output is the result of the sum of the effects that the variations of the availability of the different inputs due to infinitesimal variations dt of time have on the production function. In this example, (4.36) has therefore a clear economic interpretation. More concretely, let $g : \mathbb{R} \to \mathbb{R}^3$ be defined as $g(t) = (1/t, 3/t, e^{-t})$ for $t \neq 0$, and let $f : \mathbb{R}^3 \to \mathbb{R}$ be defined as $f(x_1, x_2, x_3) = 3x_1^2 - x_1x_2 + 6x_1x_3$. We have:

$$\begin{aligned}(f \circ g)'(t) &= \frac{\partial f}{\partial x_1}(g(t)) \frac{dg_1}{dt}(t) + \frac{\partial f}{\partial x_2}(g(t)) \frac{dg_2}{dt}(t) + \frac{\partial f}{\partial x_3}(g(t)) \frac{dg_3}{dt}(t) \\ &= 6e^{-t}\left(-\frac{1}{t^2} - \frac{1}{t}\right).\end{aligned}$$

Therefore,

$$d(f \circ g)(t)(h) = \left(6e^{-t}\left(-\frac{1}{t^2} - \frac{1}{t}\right)\right) h \qquad \textit{for every } h \in \mathbb{R},$$

and the total differential (4.36) is:

$$d(f \circ g) = \left(6e^{-t}\left(-\frac{1}{t^2} - \frac{1}{t}\right)\right) dt.$$

▲

Example 166 Here assume that besides $q = 1$ we have $m = 1$. Let $f : B \subseteq \mathbb{R} \to \mathbb{R}$ and $g : A \subseteq \mathbb{R}^n \to \mathbb{R}$, with the usual condition $g(A) \subseteq B$. For the compound function $f \circ g : A \subseteq \mathbb{R}^n \to \mathbb{R}$ we have

$$\nabla (f \circ g)(x) = f'(g(x)) \nabla g(x) = \left(f'(g(x)) \frac{\partial g}{\partial x_1}, ..., f'(g(x)) \frac{\partial g}{\partial x_n}\right),$$

to which it corresponds the differential

$$d(f \circ g)(x)(h) = \sum_{i=1}^{n} f'(g(x)) \frac{\partial g}{\partial x_i} h_i$$

for each $h \in \mathbb{R}^n$. In this case the total differential (4.35) becomes:

$$d(f \circ g) = \frac{df}{dg}\frac{\partial g}{\partial x_1} dx_1 + \cdots + \frac{df}{dg}\frac{\partial g}{\partial x_n} dx_n. \tag{4.37}$$

Let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = e^{2x}$ for each $x \in \mathbb{R}$ and let $g : \mathbb{R}^2 \to \mathbb{R}$ be defined by $g(x_1, x_2) = x_1 x_2^2$ for each $x \in \mathbb{R}^2$. We have $\nabla (f \circ g)(x) = \left(2x_2^2 e^{2x_1 x_2^2}, 4x_1 x_2 e^{2x_1 x_2^2}\right)$ and therefore

$$d(f \circ g)(x)(h) = 2e^{2x_1 x_2^2}\left(x_2^2 h_1 + 2x_1 x_2 h_2\right)$$

for each $h \in \mathbb{R}^2$, while (4.37) is:

$$d(f \circ g) = 2e^{2x_1 x_2^2}\left(x_2^2 dx_1 + 2x_1 x_2 dx_2\right).$$

▲

We conclude this section with a chain rule example with $q \neq 1$.

Example 167 Consider the applications seen in Examples 156 and 160. Therefore, let $g : \mathbb{R}^3 \to \mathbb{R}^2$ be defined by $g(x_1, x_2, x_3) = (2x_1^2 + x_2 + x_3, x_1 - x_2^4)$ for each $x \in \mathbb{R}^3$, while $f : \mathbb{R}^2 \to \mathbb{R}^2$ is defined by $f(x_1, x_2) = (x_1, x_1 x_2)$ for each $x \in \mathbb{R}^2$. Since both f and g are Frechet differentiable at each point of their domain, by Theorem 163 the composition $f \circ g : \mathbb{R}^3 \to \mathbb{R}^2$ is itself Frechet differentiable at each point of its domain $\mathbb{R}^3$. By the chain rule (4.28), the Jacobian matrix of $f \circ g : \mathbb{R}^3 \to \mathbb{R}^2$ is given by:

$$D(f \circ g)(x) = Df(g(x)) Dg(x).$$

In Example 160 we saw that

$$Dg(x) = \begin{bmatrix} 4x_1 & 1 & 1 \\ 1 & -4x_2^3 & 0 \end{bmatrix}.$$

On the other hand, we also know that:

$$Df(x) = \begin{bmatrix} 1 & 0 \\ x_2 & x_1 \end{bmatrix},$$

and therefore

$$Df(g(x)) = \begin{bmatrix} 1 & 0 \\ x_1 - x_2^4 & 2x_1^2 + x_2 + x_3 \end{bmatrix}.$$

It follows that:

$$\begin{aligned} & Df(g(x))\, Dg(x) \\ = & \begin{bmatrix} 1 & 0 \\ x_1 - x_2^4 & 2x_1^2 + x_2 + x_3 \end{bmatrix} \begin{bmatrix} 4x_1 & 1 & 1 \\ 1 & -4x_2^3 & 0 \end{bmatrix} \\ = & \begin{bmatrix} 4x_1 & 1 & 1 \\ 6x_1^2 - 4x_1x_2^4 + x_2 + x_3 & x_1 - 8x_1^2x_2^3 - 5x_2^4 - 4x_2^3x_3 & x_1 - x_2^4 \end{bmatrix}, \end{aligned}$$

which implies that the Frechet differential at x of $f \circ g$ is given by the linear application $df(x) : \mathbb{R}^3 \to \mathbb{R}^2$ defined as

$$\begin{aligned} & d(f \circ g)(x)(h) \\ = & \begin{bmatrix} 4x_1 & 1 & 1 \\ 6x_1^2 - 4x_1x_2^4 + x_2 + x_3 & x_1 - 8x_1^2x_2^3 - 5x_2^4 - 4x_2^3x_3 & x_1 - x_2^4 \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix}. \end{aligned}$$

For example, at $x = (2, 1, -1)$ we have:

$$d(f \circ g)(x)(h) = (8h_1 + h_2 + h_3, 16h_1 - 31h_2 + h_3).$$

Naturally, though it is in general more complicated, the Jacobian matrix of the composition $f \circ g$ can be computed directly, without using the chain rule, by writing explicitly the form of $f \circ g$ and by computing its partial derivatives. In this example, $f \circ g : \mathbb{R}^3 \to \mathbb{R}^2$ is given by

$$\begin{aligned} (f \circ g)(x_1, x_2, x_3) &= \left(2x_1^2 + x_2 + x_3, \left(x_1 - x_2^4\right)\left(2x_1^2 + x_2 + x_3\right)\right) \\ &= \left(2x_1^2 + x_2 + x_3, 2x_1^3 + x_1x_2 + x_1x_3 - 2x_1^2x_2^4 - x_2^5 - x_2^4x_3\right). \end{aligned}$$

Therefore,

$$\begin{aligned} (f \circ g)_1(x) &= 2x_1^2 + x_2 + x_3, \\ (f \circ g)_2(x) &= 2x_1^3 + x_1x_2 + x_1x_3 - 2x_1^2x_2^4 - x_2^5 - x_2^4x_3, \end{aligned}$$

and we have:

$$\begin{aligned}
\frac{\partial (f\circ g)_1}{\partial x_1} &= 4x_1, \quad \frac{\partial (f\circ g)_1}{\partial x_2} = 1, \quad \frac{\partial (f\circ g)_1}{\partial x_3} = 1,\\
\frac{\partial (f\circ g)_2}{\partial x_1} &= 6x_1^2 - 4x_1x_2^4 + x_2 + x_3,\\
\frac{\partial (f\circ g)_2}{\partial x_2} &= x_1 - 8x_1^2x_2^3 - 5x_2^4 - 4x_2^3x_3,\\
\frac{\partial (f\circ g)_2}{\partial x_3} &= x_1 - x_2^4,
\end{aligned}$$

The Jacobian matrix

$$\begin{bmatrix} \frac{\partial (f\circ g)_1}{\partial x_1} & \frac{\partial (f\circ g)_1}{\partial x_2} & \frac{\partial (f\circ g)_1}{\partial x_3} \\ \frac{\partial (f\circ g)_2}{\partial x_1} & \frac{\partial (f\circ g)_2}{\partial x_2} & \frac{\partial (f\circ g)_2}{\partial x_3} \end{bmatrix}$$

coincides with the one found through the chain rule. ▲

4.5 Subsequent Differentials

4.5.1 Derivatives

Till now we always talked of differentials, never of derivatives. It is time to see which form takes this important notion in our general case.

Definition 168 *Given a function $f : A \subseteq \mathbb{R}^n \to \mathbb{R}^m$, let $\Omega \subseteq A$ be the set of the points at which f is Frechet differentiable. The function $f' : \Omega \subseteq \mathbb{R}^n \to L(\mathbb{R}^n, \mathbb{R}^m)$ defined by*

$$f'(x) = df(x), \qquad \forall x \in \Omega,$$

is called (Frechet) derivative of f. In particular, the value $f'(x)$ is called derivative of f at x.

In other words, the derivative is a function that associates to each point x of Ω the Frechet differential $df(x)$ at x. This differential is an element of the space $L(\mathbb{R}^n, \mathbb{R}^m)$ since by definition it is a linear application defined on $\mathbb{R}^n$ and with values in $\mathbb{R}^m$.

Example 169 Let $f : \mathbb{R}^3 \to \mathbb{R}^2$ be defined by $f(x_1, x_2, x_3) = (2x_1^2 + x_2 + x_3, x_1 - x_2^4)$ for each $x \in \mathbb{R}^3$. In Example 160 we saw that the Frechet differential at each $x \in \mathbb{R}^3$ is given by the linear application $df(x) : \mathbb{R}^3 \to \mathbb{R}^2$ defined by

$$df(x)(h) = Df(x)\,h = \left(4x_1h_1 + h_2 + h_3, h_1 - 4x_2^3h_2\right)$$

for each $h \in \mathbb{R}^3$. Therefore, $\Omega = \mathbb{R}^3$ and the derivative $f' : \mathbb{R}^3 \to L(\mathbb{R}^3, \mathbb{R}^2)$ of f is the function that associates to each $x \in \mathbb{R}^3$ the linear application $df(x)(h)$, which is

$$x \longmapsto df(x)(h) = \left(4x_1h_1 + h_2 + h_3, h_1 - 4x_2^3h_2\right).$$

For example, the derivative of f at $x = (2, 5, -3)$ is given by

$$f'(x) = df(x)(h) = (8h_1 + h_2 + h_3, h_1 - 500h_2).$$

▲

Example 170 Let $f : \mathbb{R} \to \mathbb{R}^3$ be defined by $f(x) = (x, \sin x, \cos x)$ for each $x \in \mathbb{R}$. In Example 161 we saw that in this case the Frechet differential at each $x \in \mathbb{R}$ is given by the linear application $df(x) : \mathbb{R} \to \mathbb{R}^3$ defined by

$$df(x)(h) = (h, h\cos x, -h\sin x)$$

for each $h \in \mathbb{R}$. Therefore, $\Omega = \mathbb{R}$ and the derivative $f' : \mathbb{R} \to L(\mathbb{R}, \mathbb{R}^3)$ of f is the function that associates to each $x \in \mathbb{R}$ the linear application $df(x)(h)$, which is

$$x \longmapsto df(x)(h) = (h, h\cos x, -h\sin x).$$

For example, the derivative of f at $x = \pi$ is given by

$$f'(x) = df(x)(h) = (h, -h, 0).$$

▲

By Theorem 159, each differential $df(x)(h)$ admits a matrix representation $df(x)(h) = Df(x)h$ through the Jacobian matrix $Df(x)$. Thanks to this representation we can identify the derivative $f' : \Omega \subseteq \mathbb{R}^n \to L(\mathbb{R}^n, \mathbb{R}^m)$ with the function that associates to each $x \in \Omega$ the corresponding Jacobian matrix $Df(x)$, which is:

$$x \longmapsto Df(x), \qquad \forall x \in \Omega.$$

For this reason, from now on for (Frechet) derivative of f we will mean the function, always denoted by f', defined by $f'(x) = Df(x)$ for each $x \in \Omega$. It is therefore a function defined on $\mathbb{R}^n$ and with values in the space of the matrices $m \times n$; that is, $f' : \Omega \subseteq \mathbb{R}^n \to M(m, n)$.

In the special case $m = 1$, we simply have $f' : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^n$ since in this case $f'(x) = \nabla f(x) \in \mathbb{R}^n$. In the even more special case $m = n = 1$ the derivative at x is given by the ordinary derivative $f'(x)$.

Example 171 Let $f : \mathbb{R}^4 \to \mathbb{R}$ be defined by $f(x_1, x_2, x_3, x_4) = 4x_1x_4 + 3x_2^2x_3 + 2x_4^2$. In each $x \in \mathbb{R}^4$ we have:

$$\frac{\partial f}{\partial x_1}(x) = 4x_4, \quad \frac{\partial f}{\partial x_2}(x) = 6x_2x_3, \quad \frac{\partial f}{\partial x_3}(x) = 3x_2^2, \quad \frac{\partial f}{\partial x_4}(x) = 4x_1 + 4x_4.$$

Therefore, $\Omega = \mathbb{R}^4$ and the derivative $f' : \mathbb{R}^4 \to \mathbb{R}^4$ is defined by

$$f'(x) = \nabla f(x) = \left(4x_4, 6x_2x_3, 3x_2^2, 4x_1 + 4x_4\right), \qquad \forall x \in \mathbb{R}^4.$$

At the point $x = (0, 1, 3, 2) \in \mathbb{R}^4$, the derivative is given by $f'(x) = (8, 18, 3, 8)$. ▲

Example 172 Let $f:\mathbb{R}^3\to\mathbb{R}^2$ be defined by $f(x_1,x_2,x_3)=(2x_1^2+x_2+x_3,x_1-x_2^4)$ for each $x\in\mathbb{R}^3$. In Example 160 we showed that

$$Df(x)=\begin{bmatrix}4x_1 & 1 & 1\\ 1 & -4x_2^3 & 0\end{bmatrix}$$

at each $x\in\mathbb{R}^3$. Therefore, $\Omega=\mathbb{R}^3$ and the derivative $f':\mathbb{R}^3\to M(2,3)$ of f is the function that associates to each $x\in\mathbb{R}^3$ the Jacobian matrix $Df(x)$, that is ,

$$x\longmapsto Df(x)=\begin{bmatrix}4x_1 & 1 & 1\\ 1 & -4x_2^3 & 0\end{bmatrix}.$$

The derivative of f at $x=(2,5,-3)$ is given by

$$f'(x)=Df(x)=\begin{bmatrix}8 & 1 & 1\\ 1 & -500 & 0\end{bmatrix}.$$

▲

4.5.2 Second-Order Differentials

We can now introduce the second-order differential of a function $f:A\subseteq\mathbb{R}^n\to\mathbb{R}$.

Definition 173 *A function $f:A\subseteq\mathbb{R}^n\to\mathbb{R}$ is said to be twice Frechet differentiable at $x\in\Omega$ if the derivative $f':\Omega\subseteq\mathbb{R}^n\to\mathbb{R}^n$ is Frechet differentiable at x. The second differential of f at x is given by*

$$d^2f(x)(h)=df'(x)(h),\qquad\forall h\in\mathbb{R}^n.$$

We give right away an example.

Example 174 Let $f:\mathbb{R}^4\to\mathbb{R}$ be defined by $f(x_1,x_2,x_3,x_4)=4x_1x_4+3x_2^2x_3+2x_4^2$. In Example 171 we saw that in this case the derivative $f':\mathbb{R}^4\to\mathbb{R}^4$ is defined by

$$f'(x)=\left(4x_4,6x_2x_3,3x_2^2,4x_1+4x_4\right),\qquad\forall x\in\mathbb{R}^4.$$

To find out the differential $df'(x)$ it is necessary to compute the Jacobian matrix $Df'(x)$ of f'. We have:

$$f_1'(x)=4x_4,\quad f_2'(x)=6x_2x_3,\quad f_3'(x)=3x_2^2,\quad f_4'(x)=4x_1+4x_4,$$

and therefore

$$\begin{array}{rcllll}
\dfrac{\partial f_1'}{\partial x_1}(x) &=& 0, & \dfrac{\partial f_1'}{\partial x_2}(x)=0, & \dfrac{\partial f_1'}{\partial x_3}(x)=0, & \dfrac{\partial f_1'}{\partial x_4}(x)=4,\\
\dfrac{\partial f_2'}{\partial x_1}(x) &=& 0, & \dfrac{\partial f_2'}{\partial x_2}(x)=6x_3, & \dfrac{\partial f_2'}{\partial x_3}(x)=6x_2, & \dfrac{\partial f_2'}{\partial x_4}(x)=0,\\
\dfrac{\partial f_3'}{\partial x_1}(x) &=& 0, & \dfrac{\partial f_3'}{\partial x_2}(x)=6x_2, & \dfrac{\partial f_3'}{\partial x_3}(x)=0, & \dfrac{\partial f_3'}{\partial x_4}(x)=0,\\
\dfrac{\partial f_4'}{\partial x_1}(x) &=& 4, & \dfrac{\partial f_4'}{\partial x_2}(x)=0, & \dfrac{\partial f_4'}{\partial x_3}(x)=0, & \dfrac{\partial f_4'}{\partial x_4}(x)=4.
\end{array}$$

Consequently, the Jacobian matrix is

$$Df'(x) = \begin{bmatrix} 0 & 0 & 0 & 4 \\ 0 & 6x_3 & 6x_2 & 0 \\ 0 & 6x_2 & 0 & 0 \\ 4 & 0 & 0 & 4 \end{bmatrix},$$

which implies:

$$df'(x)(h) = Df'(x)\,h = \begin{bmatrix} 0 & 0 & 0 & 4 \\ 0 & 6x_3 & 6x_2 & 0 \\ 0 & 6x_2 & 0 & 0 \\ 4 & 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \end{bmatrix}.$$

Making the computations, by Definition 173 we therefore get that the second differential $d^2 f(x) : \mathbb{R}^4 \to \mathbb{R}^4$ is defined by:

$$d^2 f(x)(h) = (4h_4, 6x_3h_2 + 6x_2h_3, 6x_2h_2, 4h_1 + 4h_4) \quad \textit{for each } h \in \mathbb{R}^4.$$

▲

In the example just seen, the matrix $Df'(x)$ has been computed taking the derivatives of the functions f_i' that form the application $f' : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^n$. Since $f'(x) = \nabla f(x)$, these functions are nothing but the partial derivatives of f, that is,

$$f_i'(x) = \frac{\partial f}{\partial x_i}(x) \quad \text{for } i = 1, ..., n.$$

Therefore, for each $i, j = 1, ..., n$ we have

$$\frac{\partial f_i'}{\partial x_j}(x) = \frac{\partial\left(\frac{\partial f}{\partial x_i}\right)}{\partial x_j}(x).$$

In other words, $\partial f_i'/\partial x_j$ is the partial derivative with respect to x_j of the partial derivative $\partial f/\partial x_i$. The usual notation for such partial derivatives of the second order is

$$\frac{\partial^2 f}{\partial x_i \partial x_j}. \tag{4.38}$$

In particular, when $i = j$ we write $\partial^2 f/\partial x_i^2$ instead of $\partial^2 f/\partial x_i \partial x_i$.[3] Using this notation, the general form of the Jacobian $Df'(x)$ becomes:

$$\begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_2^2}(x) & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n}(x) \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(x) & \frac{\partial^2 f}{\partial x_n \partial x_2}(x) & \cdots & \frac{\partial^2 f}{\partial x_n^2}(x) \end{bmatrix}. \tag{4.39}$$

[3]Note that with this notation the order of i and j is inverted. For this reason, instead of (4.38) sometimes the notation $\partial^2 f/\partial x_j \partial x_i$ is used. On the other hand, thanks to Theorem 180 this choice of notation is irrelevant in most of the cases of interest.

This matrix of second-order partial derivatives is called the *Hessian matrix* of f and is denoted by $\nabla^2 f(x)$. The Hessian Matrix is therefore the matrix that gives us the matrix representation of the linear application $d^2 f(x)$, that is,

$$d^2 f(x)(h) = \nabla^2 f(x) h, \qquad \forall h \in \mathbb{R}^n.$$

Example 175 Let $f : \mathbb{R}^3 \to \mathbb{R}$ be defined by $f(x) = e^{x_1 x_2} + 3x_2 x_3$ for each $x \in \mathbb{R}^3$. Compute the Hessian matrix. We have:

$$\frac{\partial f}{\partial x_1}(x) = x_2 e^{x_1 x_2}, \quad \frac{\partial f}{\partial x_2}(x) = x_1 e^{x_1 x_2} + 3x_3, \quad \frac{\partial f}{\partial x_3}(x) = 3x_2,$$

and therefore

$$\begin{aligned}
\frac{\partial^2 f}{\partial x_1^2}(x) &= x_2^2 e^{x_1 x_2}, \quad \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) = (1 + x_1 x_2) e^{x_1 x_2}, \quad \frac{\partial^2 f}{\partial x_1 \partial x_3}(x) = 0, \\
\frac{\partial^2 f}{\partial x_2 \partial x_1}(x) &= (1 + x_1 x_2) e^{x_1 x_2}, \quad \frac{\partial^2 f}{\partial x_2^2}(x) = x_1^2 e^{x_1 x_2}, \quad \frac{\partial^2 f}{\partial x_2 \partial x_3}(x) = 3, \\
\frac{\partial^2 f}{\partial x_3 \partial x_1}(x) &= 0, \quad \frac{\partial^2 f}{\partial x_3 \partial x_2}(x) = 3, \quad \frac{\partial^2 f}{\partial x_3^2}(x) = 0.
\end{aligned}$$

We can conclude that the Hessian matrix of f is given by:

$$\nabla^2 f(x) = \begin{bmatrix} x_2^2 e^{x_1 x_2} & (1 + x_1 x_2) e^{x_1 x_2} & 0 \\ (1 + x_1 x_2) e^{x_1 x_2} & x_1^2 e^{x_1 x_2} & 3 \\ 0 & 3 & 0 \end{bmatrix}. \tag{4.40}$$

Consequently, the matrix representation of $d^2 f(x) : \mathbb{R}^3 \to \mathbb{R}^3$ is given by:

$$d^2 f(x)(h) = \begin{bmatrix} x_2^2 e^{x_1 x_2} & (1 + x_1 x_2) e^{x_1 x_2} & 0 \\ (1 + x_1 x_2) e^{x_1 x_2} & x_1^2 e^{x_1 x_2} & 3 \\ 0 & 3 & 0 \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix}.$$

▲

As to the differentials of order higher than two, it is possible to proceed in a similar way. Let $\Omega_2 \subseteq \mathbb{R}^n$ be the set on which f is twice Frechet differentiable. Using the identification seen above for first-order derivatives, we can define the second derivative $f'' : \Omega_2 \subseteq \mathbb{R}^n \to M(n, n)$ as the function that associates to each point $x \in \mathbb{R}^n$ the matrix $n \times n$ associated to the linear application $d^2 f(x)$. At this point, the third-order differential $d^3 f(x)(h)$ is defined as $d^3 f(x)(h) = df''(x)(h)$ for each $h \in \mathbb{R}^n$. And so on for the differentials of order n generic.

The problem of all this is that we still do not know what it means to differentiate an application of the form $f : A \subseteq \mathbb{R}^n \to M(n, n)$, what is the second derivative. Til now, in fact, we only studied differentials of applications $f : A \subseteq \mathbb{R}^n \to \mathbb{R}^m$ among

Euclidean spaces. However, conceptually the differentials of applications of the type $f : A \subseteq \mathbb{R}^n \to M(n,n)$ can be defined similarly to how we did for applications among Euclidean spaces. Because of the limited conceptual novelty involved, for brevity we do not enter into details and we do not go beyond differentials of order two.

For the same reason in Definition 173 we limited ourselves to the case $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$. In fact, to define second differentials of applications $f : A \subseteq \mathbb{R}^n \to \mathbb{R}^m$ it is necessary to talk of differentials of functions of the type $f : A \subseteq \mathbb{R}^n \to M(n,n)$, something we do not pursue in these lecture notes.

4.5.3 Symmetry of Hessian Matrices

Though we do not go beyond second-order Frechet differentials, it is however possible to consider partial derivatives of whatever order. In fact, second-order partial derivatives can be regarded as functions of their variables and we can therefore look for their partial derivatives, which (if they exist) become the partial derivatives of third order. On the other hand, also third-order partial derivatives can be seen as functions of their variables, whose partial derivatives (if they exist) become the derivatives of fourth order, and so on.

For example going back to Example 175 consider the partial derivative $(\partial^2 f/\partial x_1 \partial x_2)(x) = (1 + x_1 x_2) e^{x_1 x_2}$. We have the following third-order derivatives:

$$\begin{aligned} \frac{\partial^3 f}{\partial x_1 \partial x_2 \partial x_1}(x) &= \frac{\partial\left(\frac{\partial^2 f}{\partial x_1 \partial x_2}\right)}{\partial x_1}(x) = \left(2x_2 + x_1 x_2^2\right) e^{x_1 x_2}, \\ \frac{\partial^3 f}{\partial x_1 \partial x_2^2}(x) &= \frac{\partial\left(\frac{\partial^2 f}{\partial x_1 \partial x_2}\right)}{\partial x_2}(x) = \left(2x_1 + x_1^2 x_2\right) e^{x_1 x_2}, \end{aligned}$$

and so on for the fourth-order partial derivatives, etc.

Using the successive partial derivatives we can extend in a natural way Definition 155.

Definition 176 *A function $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ is said of class $\mathcal{C}^k$, with $k \geq 1$, if it has partial derivatives up to the order k and if such partial derivatives are continuos on their domain A. The set of all these functions is denoted by $\mathcal{C}^k(A)$.*

Clearly,

$$\mathcal{C}^k(A) \subseteq \mathcal{C}^{k-1}(A) \subseteq \cdots \subseteq \mathcal{C}^1(A),$$

that is, each function of class $\mathcal{C}^k$ is also of class $\mathcal{C}^{k-1}$, and so on. Of particular interest is the set $\bigcap_{k=1}^{\infty} \mathcal{C}^k(A)$, denoted by $\mathcal{C}^\infty(A)$. It is the class of the functions that have continuous partial derivatives of each order $k \geq 1$.

Example 177 Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined as $f(x_1, x_2) = x_1 x_2$. It is immediate to see that f has continuous partial derivatives of any order and therefore belongs to $\mathcal{C}^k(\mathbb{R}^2)$ for each $k \geq 1$. Consequently, $f \in \mathcal{C}^\infty(\mathbb{R}^2)$. More generally, the polynomials in several variables are all functions of class $\mathcal{C}^\infty$. ▲

Example 178 Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by:

$$f(x_1, x_2) = \begin{cases} x_1 x_2 \frac{x_1^2 - x_2^2}{x_1^2 + x_2^2} & \text{if } (x_1, x_2) \neq (0,0), \\ 0 & \text{if } (x_1, x_2) = (0,0). \end{cases}$$

Making the computations, the reader can verify that: (i) f has partial derivatives $\partial f/\partial x_1$ and $\partial f/\partial x_2$ continuous on $\mathbb{R}^2$; (ii) f has second-order partial derivatives $\partial^2 f/\partial x_1 \partial x_2$ and $\partial^2 f/\partial x_2 \partial x_1$ defined on all $\mathbb{R}^2$, but discontinuous at $(0,0)$. We can therefore conclude that $f \in \mathcal{C}^1(\mathbb{R}^2)$, but $f \notin \mathcal{C}^2(\mathbb{R}^2)$. ▲

The introduction of the functions of class $\mathcal{C}^1$ was motivated by Theorem 150, which showed that such functions are well behaved with respect to differentiability. Similar results hold for the classes $\mathcal{C}^k$ and for higher order differentials. For our purposes it is sufficient to consider the case $k = 2$.

Theorem 179 *Let $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ be a function that has second-order partial derivatives on a neighborhood of the point $x \in A$. If these derivatives are continuous at x, then f is twice Frechet differentiable at x.*

Proof The Hessian matrix of f at x was the Jacobian matrix $Df'(x)$ associated to the derivative $f' : \mathbb{R}^n \to \mathbb{R}^n$, whose components we saw in (4.39) to be the second-order partial derivatives of f. As we observed after Definition 158, Theorem 150 holds also for applications. In our case this means that the derivative f' is Frechet differentiable at x if the components of $Df'(x)$ are continuos at x, that is, if the second-order partial derivatives of f are continuous at x. ■

A straightforward consequence of this result is that the functions of class $\mathcal{C}^2$ are twice Frechet differentiable at each point of their domain. For these functions the following fundamental theorem holds at each point of their domain.

Theorem 180 *Let $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ be a function that has second-order partial derivatives on a neighborhood of the point $x \in A$. If these derivatives are continuous at x, then*

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i} \tag{4.41}$$

for each $i, j = 1, ..., n$.

For the functions of class $\mathcal{C}^2$, like for example polynomials in several variables, it therefore does not matter the order in which partial derivatives are considered: we can indifferently compute first the partial derivative with respect to x_i and then the one with respect to x_j, or do the contrary. The result does not change, and we can therefore choose the way that seems easier to compute, thus obtaining "for free" also the other second partial derivative. All this simplifies considerably the computation of derivatives and, moreover, gives an elegant property of symmetry to the Hessian matrix, as we will see after the proof.

Proof For simplicity we consider the case $n = 2$. In this case, (4.41) reduces to:

$$\frac{\partial^2 f}{\partial x_1 \partial x_2} = \frac{\partial^2 f}{\partial x_2 \partial x_1}. \tag{4.42}$$

Always for simplicity, we also assume that the domain A is the whole space $\mathbb{R}^2$, so that we can consider a function $f : \mathbb{R}^2 \to \mathbb{R}$. By definition,

$$\frac{\partial f}{\partial x_1}(x) = \lim_{h_1 \to 0} \frac{f(x_1 + h_1, x_2) - f(x_1, x_2)}{h_1}$$

and therefore:

$$\begin{aligned} \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) &= \lim_{h_2 \to 0} \frac{\frac{\partial f}{\partial x_1}(x_1, x_2 + h_2) - \frac{\partial f}{\partial x_1}(x_1, x_2)}{h_2} \\ &= \lim_{h_2 \to 0} \frac{1}{h_2} \left(\lim_{h_1 \to 0} \frac{f(x_1 + h_1, x_2 + h_2) - f(x_1, x_2 + h_2)}{h_1} \right. \\ &\quad \left. - \lim_{h_1 \to 0} \frac{f(x_1 + h_1, x_2) - f(x_1, x_2)}{h_1}. \right) \end{aligned}$$

Let $\psi : \mathbb{R}^2 \to \mathbb{R}$ be an auxiliary function defined by:

$$\psi(h_1, h_2) = f(x_1 + h_1, x_2 + h_2) - f(x_1, x_2 + h_2) - f(x_1 + h_1, x_2) + f(x_1, x_2),$$

for each $(h_1, h_2) \in \mathbb{R}^2$. Using the function ψ, we can write:

$$\frac{\partial^2 f}{\partial x_1 \partial x_2}(x) = \lim_{h_2 \to 0} \lim_{h_1 \to 0} \frac{\psi(h_1, h_2)}{h_2 h_1}. \tag{4.43}$$

Consider the scalar auxiliary function $\phi_1 : \mathbb{R} \to \mathbb{R}$ defined by $\phi_1(x) = f(x, x_2 + h_2) - f(x, x_2)$ for each $x \in \mathbb{R}$. We have:

$$\phi_1'(x) = \frac{\partial f}{\partial x_1}(x, x_2 + h_2) - \frac{\partial f}{\partial x_1}(x, x_2). \tag{4.44}$$

Moreover, by the Mean Value Theorem there exists $z_1 \in (x_1, x_1 + h_1)$ such that

$$\phi_1'(z_1) = \frac{\phi_1(x_1 + h_1) - \phi_1(x_1)}{h_1} = \frac{\psi(h_1, h_2)}{h_1},$$

and therefore, by (4.44), such that

$$\frac{\partial f}{\partial x_1}(z_1, x_2 + h_2) - \frac{\partial f}{\partial x_1}(z_1, x_2) = \frac{\psi(h_1, h_2)}{h_1}. \tag{4.45}$$

Let $\phi_2 : \mathbb{R} \to \mathbb{R}$ be a new auxiliary scalar function defined by $\phi_2(x) = \frac{\partial f}{\partial x_1}(z_1, x)$ for each $x \in \mathbb{R}$. We have:

$$\phi_2'(x) = \frac{\partial^2 f}{\partial x_2 \partial x_1}(z_1, x). \tag{4.46}$$

By the Mean Value Theorem there exists $z_2 \in (x_2, x_2 + h_2)$ such that

$$\phi_2'(z_2) = \frac{\phi_2(x_2 + h_2) - \phi_2(x_2)}{h_2} = \frac{\frac{\partial f}{\partial x_1}(z_1, x_2 + h_2) - \frac{\partial f}{\partial x_1}(z_1, x_2)}{h_2},$$

and therefore, by (4.46), such that

$$\frac{\partial^2 f}{\partial x_2 \partial x_1}(z_1, z_2) = \frac{\frac{\partial f}{\partial x_1}(z_1, x_2 + h_2) - \frac{\partial f}{\partial x_1}(z_1, x_2)}{h_2}.$$

Together with (4.45), this implies that

$$\frac{\partial^2 f}{\partial x_2 \partial x_1}(z_1, z_2) = \frac{\psi(h_1, h_2)}{h_2 h_1}. \tag{4.47}$$

Go back now to (4.43). Thanks to (4.47), expression (4.43) becomes:

$$\frac{\partial^2 f}{\partial x_1 \partial x_2}(x) = \lim_{h_2 \to 0} \lim_{h_1 \to 0} \frac{\partial^2 f}{\partial x_2 \partial x_1}(z_1, z_2). \tag{4.48}$$

On the other hand, since $z_i \in (x_i, x_i + h_i)$ for $i = 1, 2$, we have $z_i \to x_i$ when $h_i \to 0$. Being $\partial^2 f / \partial x_1 \partial x_2$ continuous by hypothesis at $x = (x_1, x_2)$, we therefore have

$$\lim_{h_2 \to 0} \lim_{h_1 \to 0} \frac{\partial^2 f}{\partial x_2 \partial x_1}(z_1, z_2) = \frac{\partial^2 f}{\partial x_2 \partial x_1}(x_1, x_2). \tag{4.49}$$

Putting together (4.48) and (4.49) we get (4.42), as desired. ■

Example 181 Let $f : \mathbb{R}^3 \to \mathbb{R}$ be defined by $f(x_1, x_2, x_3) = x_1^2 x_2 x_3 + e^{x_2 x_3}$. We have $f \in \mathcal{C}^\infty(\mathbb{R}^3)$ and, with some simple algebra, it is possible to verify that:

$$\frac{\partial^2 f}{\partial x_1 \partial x_2}(x) = 2x_1 x_3 \quad \text{and} \quad \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) = 2x_1 x_3,$$

as granted by Theorem 180. ▲

A square matrix $n \times n$ A is called *symmetric* if $a_{ij} = a_{ji}$ for each $i, j = 1, ..., n$. For example,

$$A = \begin{bmatrix} 1 & 3 & -2 & 0 \\ 3 & 5 & 9 & 2 \\ -2 & 9 & -4 & -3 \\ 0 & 2 & -3 & 3 \end{bmatrix}$$

is a symmetric matrix 4×4.

An important consequence of Theorem 180 is that the Hessian matrix $\nabla^2 f(x)$ of a function $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ of class $\mathcal{C}^2$ is symmetric at each point of the domain. For instance, in Example 175 we considered a function of class $\mathcal{C}^\infty$ and in fact its Hessian matrix (4.40) is symmetric. Consider now another example.

Example 182 Let $f : \mathbb{R}^3 \to \mathbb{R}$ be defined by $f(x_1, x_2, x_3) = \cos(x_1 x_2) + e^{-x_3}$. It is a function of class $\mathcal{C}^\infty$, whose Hessian matrix is

$$\nabla^2 f(x) = \begin{bmatrix} -x_2^2 \cos(x_1 x_2) & -\sin(x_1 x_2) - x_1 x_2 \cos(x_1 x_2) & 0 \\ -\sin(x_1 x_2) - x_1 x_2 \cos(x_1 x_2) & -x_1^2 \cos(x_1 x_2) & 0 \\ 0 & 0 & e^{-x_3} \end{bmatrix}.$$

As Theorem 180 guarantees, this matrix is symmetric. ▲

To conclude, we show a case not covered by Theorem 180.

Example 183 Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ seen in Example 178. We saw how the second-order partial derivatives $\partial^2 f / \partial x_1 \partial x_2$ and $\partial^2 f / \partial x_2 \partial x_1$ are defined on all $\mathbb{R}^2$, but are discontinuous at $(0, 0)$. Therefore, the hypothesis of continuity of the second-order partial derivatives, required in Theorem 180, does not hold at the point $(0, 0)$. Theorem 180 can therefore tell nothing about the behavior of such derivatives at $(0, 0)$. If we compute them, we discover that:

$$\frac{\partial^2 f}{\partial x_1 \partial x_2}(0, 0) = 1 \quad \text{and} \quad \frac{\partial^2 f}{\partial x_2 \partial x_1}(0, 0) = -1,$$

and therefore

$$\frac{\partial^2 f}{\partial x_1 \partial x_2}(0, 0) \neq \frac{\partial^2 f}{\partial x_2 \partial x_1}(0, 0).$$

The hypothesis of continuity of the second-order partial derivatives is therefore essential for the validity of equality (4.41). ▲

4.6 Taylor's Formula

Using successive differentials, we can give a version of the fundamental Taylor's Formula for functions of several variables. Before doing this, it is necessary to introduce quadratic forms.

4.6.1 Quadratic Forms

A function $f : \mathbb{R}^n \to \mathbb{R}$ of the form

$$f(x_1, ..., x_n) = k\left(x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}\right)$$

with $k \in \mathbb{R}$ and $\alpha_i \in \mathbb{N}$, is called a *monomial* of degree $m \in \mathbb{N}$ if $\sum_{i=1}^{n} \alpha_i = m$. For example, $f(x_1, x_2) = 2x_1x_2$ is a second-degree monomial, while $f(x_1, x_2, x_3) = 5x_1x_2^3x_3^4$ is an eight-degree monomial.

Definition 184 *A function $f : \mathbb{R}^n \to \mathbb{R}$ is a quadratic form if it is the sum of second-degree monomials.*

For example, $f(x_1, x_2, x_3) = 3x_1x_3 - x_2x_3$ is a quadratic form because it is the sum of the second-degree monomials $3x_1x_3$ and $-x_2x_3$. It is easy to see that the following functions are quadratic forms:

$$\begin{aligned} f(x) &= x^2, \\ f(x_1, x_2) &= x_1^2 + x_2^2 - 4x_1x_2, \\ f(x_1, x_2, x_3) &= x_1x_3 + 5x_2x_3 + x_3^2, \\ f(x_1, x_2, x_3, x_4) &= x_1x_4 - 2x_1^2 + 3x_2x_3. \end{aligned}$$

There is a one-to-one correspondence between quadratic forms and symmetric matrices, as the following result, whose proof we omit, shows.

Proposition 185 *There exists a one-to-one correspondence between quadratic forms $f : \mathbb{R}^n \to \mathbb{R}$ and symmetric matrices $\underset{n\times n}{A}$, determined by:*

$$f(x) = x \cdot Ax = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_ix_j \quad \text{for every } x \in \mathbb{R}^n. \tag{4.50}$$

Therefore, given a symmetric matrix $\underset{n\times n}{A}$ there exists a unique quadratic form $f : \mathbb{R}^n \to \mathbb{R}$ for which (4.50) holds; viceversa, given a quadratic form $f : \mathbb{R}^n \to \mathbb{R}$ there exists a unique symmetric matrix $\underset{n\times n}{A}$ for which (4.50) holds.

The matrix A is called matrix associated to the quadratic form f. We see some examples.

Example 186 The matrix associated to the quadratic form $f(x_1, x_2, x_3) = 3x_1x_3 - x_2x_3$ is given by:

$$A = \begin{bmatrix} 0 & 0 & \frac{3}{2} \\ 0 & 0 & -\frac{1}{2} \\ \frac{3}{2} & -\frac{1}{2} & 0 \end{bmatrix}.$$

In fact, for each $x \in \mathbb{R}^3$ we have:

$$\begin{aligned} x \cdot Ax &= (x_1, x_2, x_3) \cdot \begin{bmatrix} 0 & 0 & \frac{3}{2} \\ 0 & 0 & -\frac{1}{2} \\ \frac{3}{2} & -\frac{1}{2} & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \\ &= (x_1, x_2, x_3) \cdot \left(\frac{3}{2} x_3, -\frac{1}{2} x_3, \frac{3}{2} x_1 - \frac{1}{2} x_2 \right) \\ &= \frac{3}{2} x_1 x_3 - \frac{1}{2} x_2 x_3 + \frac{3}{2} x_1 x_3 - \frac{1}{2} x_2 x_3 = 3x_1 x_3 - x_2 x_3. \end{aligned}$$

Notice that also the matrices

$$A = \begin{bmatrix} 0 & 0 & 3 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 3 & -1 & 0 \end{bmatrix} \tag{4.51}$$

are such that $f(x) = x \cdot Ax$, though they are not symmetric. What we loose without symmetry is the one-to-one correspondence between quadratic forms and matrices. In fact, while given the quadratic form $f(x_1, x_2, x_3) = 3x_1 x_3 - x_2 x_3$ there exists a unique symmetric matrix such that (4.50) holds, this is no longer true if we do not require the symmetry of the matrix, as shown by the two matrices in (4.51), for which (4.50) holds. ▲

Example 187 Concerning the quadratic form $f(x_1, x_2) = x_1^2 + x_2^2 - 4x_1 x_2$, we have:

$$A = \begin{bmatrix} 1 & -2 \\ -2 & 1 \end{bmatrix}.$$

In fact, for each $x \in \mathbb{R}^2$ we have:

$$\begin{aligned} x \cdot Ax &= (x_1, x_2) \cdot \begin{bmatrix} 1 & -2 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= (x_1, x_2) \cdot (x_1 - 2x_2, -2x_1 + x_2) \\ &= x_1^2 - 2x_1 x_2 - 2x_1 x_2 + x_2^2 = x_1^2 + x_2^2 - 4x_1 x_2. \end{aligned}$$

▲

Example 188 Let $f : \mathbb{R}^n \to \mathbb{R}$ be defined by $f(x) = \|x\|^2 = \sum_{i=1}^n x_i^2$ for each $x \in \mathbb{R}^n$. The symmetric matrix associated to this quadratic form is the identity matrix I. In fact,

$$x \cdot Ix = x \cdot x = \sum_{i=1}^n x_i^2.$$

More generally, let $f(x) = \sum_{i=1}^{n} \alpha_i x_i^2$ with $\alpha_i \in \mathbb{R}$ for every $i = 1, ..., n$. It is easy to see that the matrix associated to f is the diagonal matrix

$$\begin{bmatrix} \alpha_1 & 0 & 0 & \cdots & 0 \\ 0 & \alpha_2 & 0 & \cdots & 0 \\ 0 & 0 & \alpha_3 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & \alpha_n \end{bmatrix}.$$

▲

For our purposes it is important to classify quadratic forms according to their sign.

Definition 189 *A quadratic form $f : \mathbb{R}^n \to \mathbb{R}$ is called:*

(i) positive (negative) semidefinite if $f(x) \geq 0$ (≤ 0) for every $x \in \mathbb{R}^n$,

(ii) positive (negative) definite if $f(x) > 0$ (< 0) for every $x \in \mathbb{R}^n$ with $x \neq \mathbf{0}$,

(iii) indefinite if there exist $x, x' \in \mathbb{R}^n$ such that $f(x) < 0$ and $f(x') > 0$.

By Proposition 185, we have a similar classification for symmetric matrices, where the matrix is called positive semidefinite if the corresponding quadratic form is, and so on.

In some cases it is easy to verify the sign of a quadratic form. For example, it is immediate to see that the quadratic form $f(x) = \sum_{i=1}^{n} \alpha_i x_i^2$ of Example 35 is positive (negative) semidefinite if and only if $\alpha_i \geq 0$ ($\alpha_i \leq 0$) for every $i = 1, ..., n$, while it is positive (negative) definite if and only if $\alpha_i > 0$ ($\alpha_i < 0$) for every $i = 1, ..., n$.

In general, however, it is not simple to establish directly what is the sign of a quadratic form and therefore some methods have been developed in order to facilitate this task. Among them, we present, as an example, the Sylvester-Jacobi criterion.

Given a symmetric matrix A, construct the following sequence of square submatrices A_1, A_2, ..., A_n:

$$A_1 = [a_{11}], \; A_2 = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \; A_3 = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}, ..., \; A_n = A.$$

Using this sequence, we have the following criterion of Sylvester-Jacobi.

Proposition 190 *A symmetric matrix A is:*

(i) positive definite if and only if $\det A_i > 0$ for every $i = 1, ..., n$;

(ii) negative definite if and only if $\det A_i$ *alternate in sign starting with the first negative (i.e.* $\det A_1 < 0, \det A_2 > 0, \det A_3 < 0$ *and so on).*

Example 191 Let $f(x_1, x_2, x_3) = x_1^2 + 2x_2^2 + x_3^2 + (x_1 + x_3) x_2$. The matrix associated to f is:

$$A = \begin{bmatrix} 1 & \frac{1}{2} & 0 \\ \frac{1}{2} & 2 & \frac{1}{2} \\ 0 & \frac{1}{2} & 1 \end{bmatrix}.$$

In fact,

$$\begin{aligned} x \cdot Ax &= (x_1, x_2, x_3) \cdot \begin{bmatrix} 1 & \frac{1}{2} & 0 \\ \frac{1}{2} & 2 & \frac{1}{2} \\ 0 & \frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \\ &= (x_1, x_2, x_3) \cdot \left(x_1 + \frac{1}{2}x_2, \frac{1}{2}x_1 + 2x_2 + \frac{1}{2}x_3, \frac{1}{2}x_2 + x_3 \right) \\ &= x_1^2 + 2x_2^2 + x_3^2 + (x_1 + x_3) x_2. \end{aligned}$$

Let us study the sign of this quadratic form with Sylvester-Jacobi criterion. We have:

$$\begin{aligned} \det A_1 &= 1, \\ \det A_2 &= \det \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 2 \end{bmatrix} = \frac{7}{4} > 0, \\ \det A_3 &= \det A = \frac{3}{2} > 0. \end{aligned}$$

By the Sylvester-Jacobi criterion, we can conclude that this quadratic form is positive definite. ▲

There exist versions of the Sylvester-Jacobi criterion able to determine whether a symmetric matrix is positive semidefinite, negative semidefinite, or if it is instead indefinite. For brevity, we omit the details of these versions and we move, instead, to Taylor's Formula.

4.6.2 Taylor's Formula

As we already know, when a function $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ is Frechet differentiable at a point $x \in A$, then it can be linearly approximated at this point. In particular, we have

$$\begin{aligned} f(x + h) &= f(x) + df(x)(h) + o(\|h\|) \\ &= f(x) + \nabla f(x) h + o(\|h\|) \end{aligned} \tag{4.52}$$

for each $h \in \mathbb{R}^n$. With a change in notation, denote by x_0 the point at which f is Frechet differentiable and set $h = x - x_0$. With this notation, (4.52) assumes the following equivalent, but more expressive, form:

$$\begin{aligned} f(x) &= f(x_0) + df(x_0)(x - x_0) + o(\|x - x_0\|) \\ &= f(x_0) + \nabla f(x_0)(x - x_0) + o(\|x - x_0\|) \end{aligned} \tag{4.53}$$

for each $x \in \mathbb{R}^n$.

We can now present Taylor's formula for functions of several variables; as in the scalar case, also in this more general case Taylor's formula refines the approximation (4.53). We limit ourselves to an approximation up to the second order both because we have seen Frechet differentials only up to such order (and because this is enough for our purposes). We also assume the standard hypothesis that the function is of class $\mathcal{C}^2$, which thanks to Theorem 179 guarantees that the function is twice Frechet differentiable on its domain.

Theorem 192 *Let $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ be a function of class $\mathcal{C}^2$. Then, at each $x_0 \in A$ we have:*

$$\begin{aligned} f(x) &= f(x_0) + df(x_0)(x - x_0) + \frac{1}{2}(x - x_0) \cdot d^2 f(x_0)(x - x_0) + o\left(\|x - x_0\|^2\right) \\ &= f(x_0) + \nabla f(x_0)(x - x_0) + \frac{1}{2}(x - x_0) \cdot \nabla^2 f(x_0)(x - x_0) + o\left(\|x - x_0\|^2\right) \end{aligned}$$

for every $x \in \mathbb{R}^n$.

The expression

$$f(x_0) + \nabla f(x_0)(x - x_0) + \frac{1}{2}(x - x_0) \cdot \nabla^2 f(x_0)(x - x_0)$$

is called the Taylor polynomial of second degree at x_0. The term of second degree is a quadratic form, whose associated matrix – the Hessian $\nabla^2 f(x)$ – is symmetric since f is of class $\mathcal{C}^2$. Naturally, if arrested to the first order, Taylor's formula reduces to (4.52). Moreover, observe that in the scalar case the Taylor polynomial assumes the well-known form:

$$f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2.$$

For, in this case we have $\nabla^2 f(x_0) = [f''(x_0)]$ and therefore

$$(x - x_0) \cdot \nabla^2 f(x_0)(x - x_0) = f''(x_0)(x - x_0)^2. \tag{4.54}$$

Like in the scalar case, also here we have a trade-off between the simplicity of the approximation and its accuracy. In fact, the approximation arrested to the first order

(4.52) has the advantage of simplicity with respect to that arrested to the second order - we approximate with a linear function rather than with a second-degree polynomial - but with a loss in the degree of accuracy of the approximation, which is given by $o(\|x-x_0\|)$ rather than by the better $o\left(\|x-x_0\|^2\right)$.

The choice of the order to which arrest Taylor's formula thus depends on the particular use which we are interested in, which determines what aspect of the approximation is more important in it, simplicity or accuracy.

Proof of Theorem 192 For simplicity, assume that the domain of f is all $\mathbb{R}^n$. Fixed a point $y \in \mathbb{R}^n$, introduce the auxiliary scalar functions $\phi, \psi : \mathbb{R} \to \mathbb{R}$ defined as $\phi(t) = f(x_0 + ty)$ and $\psi(t) = x_0 + ty$ for each $t \in \mathbb{R}$. We have $\phi(t) = f(\psi(t))$ for every $t \in \mathbb{R}$, i.e., $\phi = f \circ \psi$. In particular,

$$\phi(0) = f(\psi(0)) = f(x_0). \tag{4.55}$$

Since f is of class $\mathcal{C}^2$, it is easy to see that by Theorem 163 the function ϕ is twice differentiable on $\mathbb{R}$. In particular, Taylor's formula for scalar functions gives us:

$$\phi(t) = \phi(0) + \phi'(0)\, t + \frac{1}{2}\phi''(0)\, t^2 + o\left(t^2\right) \tag{4.56}$$

for each $t \in \mathbb{R}$. Since $\phi = f \circ \psi$, by the chain rule (Theorem 163) and recalling what we saw in Example 165, we have:

$$\phi'(t) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\psi(t))\, \psi_i'(t) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x_0 + ty)\, y_i \tag{4.57}$$

for each $t \in \mathbb{R}$. In particular,

$$\phi'(0) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x_0)\, y_i = \nabla f(x_0)\, y. \tag{4.58}$$

Consider now the auxiliary scalar function $\varphi_i : \mathbb{R} \to \mathbb{R}$ defined as $\varphi_i(t) = (\partial f/\partial x_i)(x_0 + ty)$ for each $t \in \mathbb{R}$ and each $i = 1, .., n$. We have $\varphi_i = (\partial f/\partial x_i) \circ \psi$, and so by the chain rule we have:

$$\varphi_i'(t) = \sum_{j=1}^{n} \frac{\partial\left(\frac{\partial f}{\partial x_i}\right)}{\partial x_j}(\psi(t))\, \psi_j'(t) = \sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(x_0 + ty)\, y_j.$$

Together with (4.57), this implies:

$$\phi''(t) = \sum_{i=1}^{n} \varphi_i'(t)\, y_i = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(x_0 + ty)\, y_j y_i,$$

and therefore

$$\phi''(0) = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(x_0)\, y_j y_i = y \cdot \nabla^2 f(x_0)\, y. \tag{4.59}$$

Till now y was an arbitrary point of $\mathbb{R}^n$. Set now $y = (x - x_0) / \|x - x_0\|$ in (4.55), (4.58) and (4.59). When computed at $t = \|x - x_0\|$, the expansion (4.56) becomes:

$$\begin{aligned} \phi(\|x - x_0\|) &= f(x_0) + \nabla f(x_0) \frac{x - x_0}{\|x - x_0\|} \|x - x_0\| \\ &\quad + \frac{1}{2} \frac{x - x_0}{\|x - x_0\|} \cdot \nabla^2 f(x_0) \frac{x - x_0}{\|x - x_0\|} \|x - x_0\|^2 + o\left(\|x - x_0\|^2\right) \end{aligned}$$

By definition,

$$\phi(\|x - x_0\|) = f\left(x_0 + \|x - x_0\| \frac{x - x_0}{\|x - x_0\|}\right) = f(x),$$

and therefore we can conclude that:

$$f(x) = f(x_0) + \nabla f(x_0)(x - x_0) + \frac{1}{2}(x - x_0) \cdot \nabla^2 f(x_0)(x - x_0) + o\left(\|x - x_0\|^2\right),$$

as desired. ■

Example 193 Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined as $f(x_1, x_2) = 3x_1^2 e^{x_2^2}$. Making the computations, we get:

$$\begin{aligned} \nabla f(x) &= \left(6x_1 e^{x_2^2}, 6x_1^2 x_2 e^{x_2^2}\right), \\ \nabla^2 f(x) &= \begin{bmatrix} 6e^{x_2^2} & 12x_1 x_2 e^{x_2^2} \\ 12x_1 x_2 e^{x_2^2} & 6x_1^2 e^{x_2^2}(1 + 2x_2^2) \end{bmatrix}. \end{aligned}$$

By Theorem 192, Taylor's formula at $x_0 = (1, 1)$ is

$$\begin{aligned} f(x) &= f(1,1) + \nabla f(1,1)(x_1 - 1, x_2 - 1) \\ &\quad + \frac{1}{2}(x_1 - 1, x_2 - 1) \cdot \nabla^2 f(1,1)(x_1 - 1, x_2 - 1) + o\left(\|(x_1 - 1, x_2 - 1)\|^2\right) \\ &= 3e + (6e, 6e)(x_1 - 1, x_2 - 1) + \\ &\quad \frac{1}{2}(x_1 - 1, x_2 - 1) \cdot \begin{bmatrix} 6e & 12e \\ 12e & 18e \end{bmatrix} \begin{bmatrix} x_1 - 1 \\ x_2 - 1 \end{bmatrix} + o\left((x_1 - 1)^2 + (x_2 - 1)^2\right) \\ &= 3e\left(x_1^2 - 4x_1 + 5 - 8x_2 + 4x_1 x_2 + 3x_2^2\right) + o\left((x_1 - 1)^2 + (x_2 - 1)^2\right). \end{aligned}$$

Therefore, the function $f(x_1, x_2) = 3x_1^2 e^{x_2^2}$ is approximated at the point $(1, 1)$ by the second-degree Taylor's polynomial

$$3e\left(x_1^2 - 4x_1 + 5 - 8x_2 + 4x_1 x_2 + 3x_2^2\right),$$

with a level of accuracy given by $o\left((x_1 - 1)^2 + (x_2 - 1)^2\right)$. ▲

Chapter 5

Free Classic Optimization

>From Calculus we know that, given a function $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$, a point $x_0 \in A$ is called point of *local* (or *relative*) *maximum* if there exists a neighborhood $B_{x_0}(\varepsilon)$ of x_0 such that $f(x_0) \geq f(x)$ for each $x \in B_{x_0}(\varepsilon) \cap A$. In particular, when we have $f(x_0) > f(x)$ for each $x \in B_{x_0}(\varepsilon) \cap A$ with $x \neq x_0$, the point x_0 is called of *strong local maximum*. Finally, the point x_0 is called of *global* (or *absolute*) *maximum* if $f(x_0) \geq f(x)$ for each $x \in A$. In a similar way it is possible to define the points of local and global minimum.

Like in the scalar case, also for functions of several variables a fundamental application of the differential calculus is the research of the points of local maximum and minimum. Conceptually, there are not many novelties relative to the scalar case, though the analysis is more complicated because of the greater sophistication of differential calculus in several variables with respect to the scalar one. In any case, also in this more general case we will divide the analysis between first-order conditions and second-order conditions.

5.1 First-Order Conditions

For functions that have an open set as their domain, the first-order condition is based on the next result.[1]

Theorem 194 *Let $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ and let $x_0 \in A$ be a point of local maximum or minimum. If f is Gateaux differentiable at x_0, then $\nabla f(x_0) = \mathbf{0}$.*

Proof Assume that x_0 is a point of local maximum (a similar argument holds if is a local minimum). By definition, there exists a neighborhood $B_{x_0}(\varepsilon)$ such that $f(x_0) \geq f(x)$ for each $x \in B_{x_0}(\varepsilon) \cap A$. Since x_0 is an interior point, there exists a neighborhood

[1] Throughout the chapter, A will always denote an open set.

$B_{x_0}(\varepsilon')$ such that $B_{x_0}(\varepsilon') \subseteq A$. Set $B = B_{x_0}(\varepsilon) \cap B_{x_0}(\varepsilon')$. Clearly, we have $f(x_0) \geq f(x)$ for each $x \in B$.

For each fundamental versor e^i, there exists $k_i > 0$ sufficiently small such that $x_0 + te^i \in B$ if $t \in (-k_i, k_i)$. Being x_0 a point of local maximum, we thus have $f(x_0) \geq f(x_0 + te^i)$ for each $t \in (-k_i, k_i)$. Consequently,

$$\lim_{t \to 0+} \frac{f(x_0 + te^i) - f(x_0)}{t} \leq 0 \leq \lim_{t \to 0-} \frac{f(x_0 + te^i) - f(x_0)}{t}.$$

On the other hand, by definition of partial derivative the bilateral limit

$$\frac{\partial f}{\partial x_i}(x_0) = \lim_{t \to 0} \frac{f(x_0 + te^i) - f(x_0)}{t},$$

holds, and therefore we conclude that $(\partial f/\partial x_i)(x_0) = 0$, as desired. ■

By Theorem 194, a necessary condition for a point to be a local maximum or minimum is that in this point the gradient vanishes. This is the so-called *first order condition* (often abbreviated as FOC) and the points that satisfy it are called *stationary* (or *critical*) points.

Observe that the condition $\nabla f(x_0) = \mathbf{0}$ is only necessary, but not sufficient, as the next simple example shows.

Example 195 Consider the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^3$. At the point $x = 0$ we have $f'(x) = 0$, but this point is neither a local minimum nor a local maximum. ▲

By Theorem 194, the search of points of relative maximum or minimum can be restricted only to stationary points, that is, to points $x \in A$ such that $\nabla f(x) = \mathbf{0}$. This amounts to solving the system

$$\frac{\partial f}{\partial x_1}(x) = 0, ..., \frac{\partial f}{\partial x_n}(x) = 0. \tag{5.1}$$

We illustrate the first order condition with some examples.

Example 196 Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined as $f(x_1, x_2) = x_1^2 - x_2^2$. We have:

$$\nabla f(x) = (2x_1, -2x_2)$$

and the system (5.1) has the form

$$\begin{cases} 2x_1 = 0 \\ -2x_2 = 0 \end{cases}$$

The only solution of this system is $x = (0, 0)$, which is therefore the only stationary point of f. It is easy to see that this point is neither a maximum nor a minimum. In fact, consider a generic point $(0, x_2)$, different from the origin, on the vertical axis and a generic point $(x_1, 0)$, also different from the origin, on the horizontal axis. By the definition of f, we have:

$$f(0, x_2) = -x_2^2 < 0 \quad \text{and} \quad f(x_1, 0) = x_1^2 > 0,$$

and therefore in each neighborhood of the point $(0, 0)$ there are both points in which the function is strictly positive and points in which it is strictly negative.

As $f(0, 0) = 0$, this implies that $x = (0, 0)$ cannot be neither a point of maximum nor a point of minimum. Therefore, this example shows how, like in the scalar case, also for functions of several variables there can be stationary points that are neither maxima nor minima. ▲

Example 197 Let $f : \mathbb{R}^n \to \mathbb{R}$ be defined as $f(x) = \sum_{i=1}^n x_i^2$ for each $x \in \mathbb{R}^n$. We have:

$$\nabla f(x) = 2x,$$

and therefore the system (5.1) has the form:

$$2x_1 = 0, ..., 2x_n = 0,$$

whose only solution is clearly $x = \mathbf{0}$. Therefore, the function f has the only stationary point $x = \mathbf{0}$. Since we have $f(x) \geq 0$ for each $x \in \mathbb{R}^n$, it is a point of global minimum. ▲

Example 198 More generally, let $f : \mathbb{R}^n \to \mathbb{R}$ be defined as $f(x) = \sum_{i=1}^n \alpha_i x_i^2$ for each $x \in \mathbb{R}^n$, with $0 \neq \alpha_i \in \mathbb{R}$ for $i = 1, ..., n$. We have:

$$\nabla f(x) = (2\alpha_1 x_1, ..., 2\alpha_n x_n),$$

and therefore the system (5.1) has the form:

$$2\alpha_1 x_1 = 0, ..., 2\alpha_n x_n = 0,$$

whose only solution is given by $x = \mathbf{0}$. Also in this more general case, $x = \mathbf{0}$ is therefore the only stationary point of f.

When for each $i = 1, ..., n$ we have $\alpha_i > 0$, then it is immediate to see that $x = \mathbf{0}$ would be a point of global minimum. Similarly, if instead we have $\alpha_i < 0$ for each $i = 1, ..., n$, then $x = \mathbf{0}$ would be a point of global maximum. When some α_i' are positive, while others are negative, we cannot conclude anything for the moment: we have to wait for the second order conditions in order to be able to say something more. ▲

Example 199 Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined as $f(x_1, x_2) = 2x_1^2 + x_2^2 - 3(x_1 + x_2) + x_1x_2 - 3$. We have:

$$\nabla f(x) = (4x_1 - 3 + x_2, 2x_2 - 3 + x_1).$$

To find the stationary points it is necessary to solve the system (5.1), which here takes the form:

$$\begin{cases} 4x_1 - 3 + x_2 = 0 \\ 2x_2 - 3 + x_1 = 0 \end{cases}$$

It is easy to see that $x = (3/7, 9/7)$ is the only solution of this system and so is the only stationary point of f. By Theorem 194, $x = (3/7, 9/7)$ is therefore the only point that could be of local minimum or maximum. Also in this case, in order to be able to say something more about the nature of this point we have to wait for the second-order conditions. ▲

Example 200 Let $f : \mathbb{R}^3 \to \mathbb{R}$ be defined as $f(x_1, x_2, x_3) = x_1^3 + x_2^3 + 3x_3^2 - 2x_3 + x_1^2x_2^2$. We have:

$$\nabla f(x) = \left(3x_1^2 + 2x_1x_2^2, 3x_2^2 + 2x_1^2x_2, 6x_3 - 2\right).$$

In this case, the system (5.1) becomes:

$$\begin{cases} 3x_1^2 + 2x_1x_2^2 = 0 \\ 3x_2^2 + 2x_1^2x_2 = 0 \\ 6x_3 - 2 = 0 \end{cases}$$

It is a nonlinear system and therefore to solve it we cannot use the methods seen in Section 3.5.2. Let us try to solve it directly. We start by observing that x_3 is alone in the third equation; therefore in each solution we have $x_3 = 1/3$. To find the values of x_1 and x_2 it remains to consider the subsystem

$$\begin{cases} 3x_1^2 + 2x_1x_2^2 = 0 \\ 3x_2^2 + 2x_1^2x_2 = 0 \end{cases}$$

A solution is given by $x_1 = x_2 = 0$. This is also the unique solution in which either variable, x_1 and x_2, vanishes. In fact, if $x_1 = 0$, then the second equation implies $x_2 = 0$, and, similarly, $x_2 = 0$ implies $x_1 = 0$.

Therefore, if besides $(0, 0)$ there exists another solution (x_1, x_2) of the subsystem, in this solution we must have $x_1 \neq 0$ and $x_2 \neq 0$. In view of all this, we can rewrite the system as

$$\begin{cases} x_1(3x_1 + 2x_2^2) = 0 \\ x_2(3x_2 + 2x_1^2) = 0 \end{cases} \tag{5.2}$$

Since we saw that, except for $(0,0)$, the solutions (x_1, x_2) of the subsystem are such that $x_1 \neq 0$ and $x_2 \neq 0$, (5.2) implies that for these solutions it holds:

$$\begin{cases} 3x_1 + 2x_2^2 = 0 \\ 3x_2 + 2x_1^2 = 0 \end{cases}$$

From the first equation we get $x_1 = (-2/3)\, x_2^2$ and, substituting it in the second equation, we find $x_2^3 = -27/8$. Consequently, $x_2 = -3/2$, and so $x_1 = -3/2$. Hence we conclude that $(-3/2, -3/2)$ is the other solution of the system besides $(0,0)$.

In conclusion, the two stationary points of this function are

$$x = (0, 0, 1/3) \quad \text{and} \quad x = (-3/2, -3/2, 1/3)\,.$$

Using the second-order conditions later we will try to determine if they are indeed points of local maximum or minimum. ▲

5.2 Second-Order Conditions

Theorem 192 showed that a function $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ of class $\mathcal{C}^2$ can be locally approximated at a point $x_0 \in A$ with a second-degree polynomial, as follows:

$$f(x) = f(x_0) + \nabla f(x_0)(x - x_0) + \frac{1}{2}(x - x_0) \cdot \nabla^2 f(x_0)(x - x_0) + o\left(\|x - x_0\|^2\right).$$

If x_0 is a point of local maximum or minimum, from Theorem 194 we have $\nabla f(x_0) = \mathbf{0}$. Therefore, the approximation becomes:

$$f(x) = f(x_0) + \frac{1}{2}(x - x_0) \cdot \nabla^2 f(x_0)(x - x_0) + o\left(\|x - x_0\|^2\right). \tag{5.3}$$

By working on this simple observation we obtain the second order conditions, which are based on the sign of the quadratic form $x \cdot \nabla^2 f(x_0)\, x$.

Theorem 201 *Let $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ be a function of class $\mathcal{C}^2$ and let $x_0 \in A$ be a stationary point of this function. We have:*

(i) If x_0 is a point of local maximum (minimum), then the quadratic form $x \cdot \nabla^2 f(x_0)\, x$ is negative (positive) semidefinite.

(ii) If the quadratic form $x \cdot \nabla^2 f(x_0)\, x$ is negative (positive) definite, then x_0 is a point of strong local maximum (minimum).

In the scalar case we get back to the usual second order conditions, based on the sign of the second derivative $f''(x_0)$. In fact, we already observed in (4.54) that in

the scalar case we have $x \cdot \nabla^2 f(x_0) x = f''(x_0) x^2$, so that in this case the sign of the quadratic form depends only on the sign of $f''(x_0)$; that is, it is negative (positive) definite if and only if $f''(x_0) < 0$ (> 0) and it is negative (positive) semidefinite if and only if $f''(x_0) \leq 0$ (≥ 0).

Naturally, like in the scalar case, also in the present more general case condition (i) is only necessary for x_0 to be a local maximum or minimum. In fact, consider the scalar function $f(x) = x^3$. We have $\nabla^2 f(x) = [f''(x)]$ and therefore at $x_0 = 0$ we have $\nabla^2 f(x_0) = [0]$. The corresponding quadratic form $x \cdot \nabla^2 f(x_0) x$ is identically null and it is therefore both negative semidefinite and positive semidefinite. Nevertheless, $x_0 = 0$ is neither a point of local maximum nor a point of local minimum.

Similarly, condition (ii) is only sufficient for x_0 to be a point of local maximum or minimum. Consider the scalar function $f(x) = -x^4$. The point $x_0 = 0$ is clearly a point of maximum (even absolute) for the function f. But, $\nabla^2 f(x_0) = [0]$ and therefore the corresponding quadratic form $x \cdot \nabla^2 f(x_0) x$ is not negative definite.

Proof of Theorem 201. We first show (ii). Let $x \cdot \nabla^2 f(x_0) x$ be negative definite (the positive definite case is similarly handled). We want to prove that x_0 is a point of strong local maximum.

Let $U \equiv \{x \in \mathbb{R}^n : \|x\| = 1\}$ be the unit ball of $\mathbb{R}^n$, that is, the set of vectors that have unit norm. It is a compact set because it is easy to see that it is both closed and bounded. Let $Q(x) = x \cdot \nabla^2 f(x_0) x$ for each $x \in U$. The function Q is therefore the restriction of the quadratic form $x \cdot \nabla^2 f(x_0) x$ on the unit ball U. As U is compact, by the Weierstrass Theorem the function Q has a point of absolute maximum in U and we can therefore set $M = \max_{x \in U} Q(x)$. In particular, $M < 0$ because $Q(x) < 0$ for each $x \in U$, since by hypothesis $x \cdot \nabla^2 f(x_0) x$ is definite negative.

Let now x be a generic point of $\mathbb{R}^n$. Clearly,

$$\left\| \frac{x - x_0}{\|x - x_0\|} \right\| = \frac{\|x - x_0\|}{\|x - x_0\|} = 1,$$

and therefore the vector $(x - x_0) / \|x - x_0\|$ belongs to the unit ball U. It follows that, using (4.50), we can write:

$$\begin{aligned} M &\geq \frac{(x - x_0)}{\|x - x_0\|} \cdot \nabla^2 f(x_0) \frac{(x - x_0)}{\|x - x_0\|} = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{x_i - x_{0,i}}{\|x - x_0\|} \frac{\partial^2 f}{\partial x_i \partial x_j} \frac{x_j - x_{0,j}}{\|x - x_0\|} \\ &= \frac{1}{\|x - x_0\|^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - x_{0,i}) \frac{\partial^2 f}{\partial x_i \partial x_j} (x_j - x_{0,j}) \\ &= \frac{1}{\|x - x_0\|^2} (x - x_0) \cdot \nabla^2 f(x_0) (x - x_0), \end{aligned}$$

so that

$$\frac{1}{2}(x-x_0)\cdot\nabla^2 f(x_0)(x-x_0)\leq\frac{M}{2}\|x-x_0\|^2.$$

By (5.3), it then follows that:

$$\begin{aligned}f(x)-f(x_0) &= \frac{1}{2}(x-x_0)\cdot\nabla^2 f(x_0)(x-x_0)+o\left(\|x-x_0\|^2\right)\\ &\leq \frac{M}{2}\|x-x_0\|^2+o\left(\|x-x_0\|^2\right)\\ &= \|x-x_0\|^2\left(\frac{M}{2}+\frac{o\left(\|x-x_0\|^2\right)}{\|x-x_0\|^2}\right).\end{aligned}$$

Set $\phi(x)=M/2+o\left(\|x-x_0\|^2\right)/\|x-x_0\|^2$ for each $x\in\mathbb{R}^n$. Clearly, $\lim_{x\to x_0}\phi(x)=M/2<0$. Therefore, by the Theorem of the Permanence of the Sign there exists a neighborhood $B_{x_0}(\varepsilon)$ of x_0 such that $\phi(x)<0$ for each $x\in B_{x_0}(\varepsilon)$. It follows that

$$f(x)-f(x_0)=\|x-x_0\|^2\phi(x)<0$$

for each $x\in B_{x_0}(\varepsilon)$ and we conclude that x_0 is a point of strong local maximum.

To complete the proof it remains to show (i). Also here we limit ourselves to the case of x_0 local maximum (the argument for the local minimum is analogous). Let x_0 be a point of local maximum and let $B_{x_0}(\varepsilon)$ be a neighborhood of x_0 such that $f(x_0)\geq f(x)$ for each $x\in B_{x_0}(\varepsilon)$.

Fixed $x\in\mathbb{R}^n$, set $\phi(t)=f(x_0+tx)$. Let $A_\phi=\{t:\phi(t)\in B_{x_0}(\varepsilon)\}$. It is easy to see that A_ϕ is an open set containing 0. As x_0 is a point of local maximum, we have $\phi(t)=f(x_0+tx)\leq f(x_0)=\phi(0)$ for each $t\in A_\phi$. Consequently, 0 is a point of local maximum for ϕ and therefore $\phi''(0)\leq 0$. On the other hand, (4.59) showed that $\phi''(0)=x\cdot\nabla^2 f(x_0)\,x$.[2] Therefore, the quadratic form $x\cdot\nabla^2 f(x_0)\,x\leq 0$ is negative semidefinite, as desired. ■

Since f is of class $\mathcal{C}^2$, the Hessian matrix $\nabla^2 f(x_0)$ is the symmetric matrix associated to the quadratic form $x\cdot\nabla^2 f(x_0)\,x$; we can therefore equivalently state Theorem 201 in the following way:

- a necessary condition for x_0 to be a point of maximum (minimum) is that the Hessian matrix $\nabla^2 f(x_0)$ be negative (positive) semidefinite,
- a sufficient condition for x_0 to be a point of strong maximum (minimum) is that such matrix be negative (positive) definite.

[2]Notice that in (4.59) the generic point of $\mathbb{R}^n$ was denoted by y instead of x.

Operationally, this is an important observation because there exist some criteria, like Sylvester-Jacobi, able to determine whether a symmetric matrix is positive/negative definite or semidefinite.

Example 202 As in Example 199, let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined as $f(x_1, x_2) = 2x_1^2 + x_2^2 - 3(x_1 + x_2) + x_1x_2 - 3$. We have:

$$\nabla f(x) = (4x_1 - 3 + x_2, 2x_2 - 3 + x_1),$$

and therefore

$$\nabla^2 f(x) = \begin{bmatrix} 4 & 1 \\ 1 & 2 \end{bmatrix}$$

The only stationary point is $x = (3/7, 9/7)$. By the Sylvester-Jacobi criterion, the Hessian matrix $\nabla^2 f(x)$ is positive definite. By Theorem 201, we can conclude that $(3/7, 9/7)$ is a point of strong local minimum. ▲

Example 203 Going back to Example 200, let $f : \mathbb{R}^3 \to \mathbb{R}$ be defined as $f(x_1, x_2, x_3) = x_1^3 + x_2^3 + 3x_3^2 - 2x_3 + x_1^2x_2^2$. We have:

$$\nabla f(x) = \left(3x_1^2 + 2x_1x_2^2, 3x_2^2 + 2x_1^2x_2, 6x_3 - 2\right),$$

and therefore

$$\nabla^2 f(x) = \begin{bmatrix} 6x_1 + 2x_2^2 & 4x_1x_2 & 0 \\ 4x_1x_2 & 6x_2 + 2x_1^2 & 0 \\ 0 & 0 & 6 \end{bmatrix}.$$

The stationary points are $x = (-3/2, -3/2, 1/3)$ and $x = (0, 0, 1/3)$. In $x = (-3/2, -3/2, 1/$ we have

$$\nabla^2 f(x) = \begin{bmatrix} -\frac{9}{2} & 9 & 0 \\ 9 & -\frac{9}{2} & 0 \\ 0 & 0 & 6 \end{bmatrix},$$

and therefore

$$\det\left[-\frac{9}{2}\right] < 0, \quad \det\begin{bmatrix} -\frac{9}{2} & 9 \\ 9 & -\frac{9}{2} \end{bmatrix} < 0, \quad \det \nabla^2 f(x) < 0.$$

Consequently, by the Sylvester-Jacobi criterion (Proposition 190) this Hessian matrix is neither positive definite nor negative definite (and it is also neither positive semidefinite nor negative semidefinite). By Theorem 201, we can conclude that the point $x = (-3/2, -3/2, 1/3)$ is neither a local minimum nor a local maximum. At the point $x = (0, 0, 1/3)$ we have

$$\nabla^2 f(x) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 6 \end{bmatrix},$$

and so all the determinants of the matrices used in the Sylvester-Jacobi criterion are null. By Proposition 190, the Hessian matrix is therefore neither positive definite nor negative definite and the second-order sufficient condition does not tell us anything in this case. ▲

Chapter 6

Metric Spaces

6.1 Definition

In Calculus the distance $d(a,b)$ between two points a and b of the real line $\mathbb{R}$ is given by $|a-b|$, while for two vectors x and y of $\mathbb{R}^n$ their distance $d(x,y)$ is given by $\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$.[1]

Our aim in this chapter is to extend the notion of distance to abstract spaces. To this end, the first thing to observe is that the distance among vectors just mentioned is certainly not the only possible one. For example, consider two vectors $x=(x_1,x_2)$ and $y=(y_1,y_2)$ in the plane $\mathbb{R}^2$. Suppose that these two vectors give us the coordinates of two places of Torino, for example x corresponds to Piazza Carlina while y corresponds to Piazza Carlo Felice. As the historic center of Torino is essentially at square plan, the length of the shortest way for a pedestrian to move between these two squares is certainly not given by $\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$, that is, by the length of the segment that joins the points x and y (a segment that could be covered only by an hypothetical subway joining the two squares). Looking at the map, it is easy to see that the effective distance is given by $|x_1-y_1|+|x_2-y_2|$. Formally, given two vectors $x,y\in\mathbb{R}^n$, we define the distance $d(x,y)$ as $\sum_{i=1}^{n}|x_i-y_i|$. In the case $n=2$ we find again the "pedestrian" distance $d(x,y)=|x_1-y_1|+|x_2-y_2|$ just discussed.

To see another example of distance, suppose that two vectors x and y in $\mathbb{R}^n$ denote the allocations of income in a society composed by n individuals. Therefore, x_i is the income that individual i has under allocation x, while y_i is his income under allocation y. How we measure the "distance" between the allocations x and y? A possibility is to evaluate the individual differences of income $|x_i-y_i|$ among the two allocations, and to take the quantity $\max_{1\leq i\leq n}|x_i-y_i|$ as the distance between the two allocations x and y.

[1] For this basic notions, see for instance Ambrosetti and Musu (1988) chapters II and III.

In other words, we evaluate the distance between the two allocations by considering the individual whose income is subject to the greatest variation (in absolute value). Given two vectors $x, y \in \mathbb{R}^n$, we define therefore the distance $d(x,y) = \max_{1\leq i\leq n} |x_i - y_i|$.

We try now to abstract from the particular examples, in order to arrive at a general definition of distance. We observe that the distances just discussed have the following properties:

- The distance between two vectors is always non-negative: $d(x,y) \geq 0$.
- Two vectors have zero distance if and only if they coincide: $d(x,y) = 0$ if and only if $x = y$.
- The distance between two vectors is symmetric: $d(x,y) = d(y,x)$.
- Given three vectors, the triangular inequality holds: $d(x,y) \leq d(x,z) + d(z,y)$.

All this leads us to the following definition, in which X is a general set.

Definition 204 *A space X is called metric if there exists a function $d : X \times X \to \mathbb{R}_+$, called distance (or metric), such that, for each $x, y, z \in X$,*

(i) $d(x,y) = 0$ *if and only if* $x = y$,

(ii) $d(x,y) = d(y,x)$

(iii) $d(x,y) \leq d(x,z) + d(z,y)$.

Naturally, on the same set X different metrics can be defined, each corresponding to a different concept of distance, relevant according to the different problems considered.

We illustrate this definition with few examples.

Example 205 We already saw different distances that make $\mathbb{R}^n$ a metric space. We will denote them as follows:

$$\begin{aligned} d_1(x,y) &= \sum_{i=1}^{n} |x_i - y_i|, \\ d_2(x,y) &= \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}, \\ d_\infty(x,y) &= \max_{1\leq i\leq n} |x_i - y_i|. \end{aligned}$$

In the case $n = 1$, all these different distances reduce to the standard distance $|x - y|$ among real numbers $x, y \in \mathbb{R}$. ▲

Example 206 Let $X = \mathcal{C}([0,1])$ be the space of continuous functions $f : [0,1] \to \mathbb{R}$ defined on the interval $[0,1]$. Define $d_\infty : \mathcal{C}([0,1]) \times \mathcal{C}([0,1]) \to \mathbb{R}_+$ by

$$d_\infty(f,g) = \max_{t\in[0,1]} |f(t) - g(t)|, \qquad \forall f, g \in \mathcal{C}([0,1]). \tag{6.1}$$

In other words, we consider for each $t \in [0,1]$ the distance $|f(t) - g(t)|$ between the two functions at that point, and we then take the maximum among all these distances. The idea is similar to that used in defining d_∞ on $\mathbb{R}^n$ and for this reason we denote also this metric by d_∞.

Observe that the max in (6.1) exists. In fact, set $h(t) = |f(t) - g(t)|$ for each $t \in [0,1]$. The function h is continuous, and so by the Weierstrass Theorem it assume a maximum value. We now verify that d_∞ is indeed a distance. By the definition of absolute value, we have $|f(t) - g(t)| = 0$ if and only if $f(t) = g(t)$. On the other hand, $\max_{t\in[0,1]} |f(t) - g(t)| = 0$ if and only if $|f(t) - g(t)| = 0$ for each $t \in [0,1]$, and so $d_\infty(f,g) = 0$ if and only if $f(t) = g(t)$ for each $t \in [0,1]$. It follows that $d_\infty(f,g) = 0$ if and only if $f = g$, which verifies property (i) of Definition 204.

Property (ii) is obvious. As to (iii), observe that, given a generic $h \in \mathcal{C}([0,1])$, for each $t \in [0,1]$ we have:

$$\begin{aligned} |f(t) - g(t)| &= |f(t) - h(t) + h(t) - g(t)| \\ &\leq |f(t) - h(t)| + |h(t) - g(t)|. \end{aligned} \tag{6.2}$$

Therefore,

$$\begin{aligned} d_\infty(f,g) &= \max_{t\in[0,1]} |f(t) - g(t)| \\ &\leq \max_{t\in[0,1]} (|f(t) - h(t)| + |h(t) - g(t)|) \\ &\leq \max_{t\in[0,1]} |f(t) - h(t)| + \max_{t\in[0,1]} |h(t) - g(t)| \\ &= d_\infty(f,h) + d_\infty(h,g), \end{aligned}$$

and (iii) is consequently satisfied. The function d_∞ is therefore a distance and the pair $(\mathcal{C}([0,1]), d_\infty)$ is a metric space. ▲

Example 207 Let again $X = \mathcal{C}([0,1])$ and define $d_1 : \mathcal{C}([0,1]) \times \mathcal{C}([0,1]) \to \mathbb{R}_+$ by

$$d_1(f,g) = \int_0^1 |f(t) - g(t)|\, dt, \qquad \forall f, g \in \mathcal{C}([0,1]). \tag{6.3}$$

We used the notation d_1 because, mutatis mutandis, this definition is similar to that of the distance d_1 in $\mathbb{R}^n$. We now verify property (i) of Definition 204. It is obvious that $f = g$ implies $d_1(f,g) = 0$. To show the converse, suppose that $f \neq g$, that is, that there exists $t^* \in [0,1]$ such that $f(t^*) \neq g(t^*)$. For simplicity, suppose that

$t^* \in (0,1)$.[2] Setting again $h(t) = |f(t) - g(t)|$, this is equivalent to $h(t^*) > 0$. Since h is continuous, by the Theorem of the Permanence of Sign there exists a neighborhood $B_\varepsilon(t^*) = (t^* - \varepsilon, t^* + \varepsilon) \subseteq [0,1]$ such that $h(t) > 0$ for each $t \in B_\varepsilon(t^*)$. It follows that $\int_{t^*-\varepsilon}^{t^*+\varepsilon} h(t)\,dt > 0$, and therefore:

$$d_1(f,g) = \int_0^1 |f(t) - g(t)|\,dt \geq \int_{t^*-\varepsilon}^{t^*+\varepsilon} h(t)\,dt > 0,$$

and this verifies property (i).

Property (ii) is obvious. As to (iii), using (6.2) we have:

$$\begin{aligned} d_1(f,g) &= \int_0^1 |f(t) - g(t)|\,dt \\ &\leq \int_0^1 |f(t) - h(t)|\,dt + \int_0^1 |h(t) - g(t)|\,dt \\ &= d_1(f,h) + d_1(h,g). \end{aligned}$$

The function d_1 is therefore a distance and also the pair $(\mathcal{C}([0,1]), d_1)$ is a metric space. ▲

Example 208 Let X be any set and define $d : X \times X \to \mathbb{R}_+$ by

$$d(x,y) = \begin{cases} 0 & \text{if } x = y, \\ 1 & \text{if } x \neq y. \end{cases} \tag{6.4}$$

It is easy to see that this function is a metric, usually called the discrete metric. Clearly, it a very "coarse" metric, which is however well defined on any set X. ▲

6.2 Topology

Having defined the notion of distance for general spaces, we can now topologize such spaces, that is, we can introduce the notions of closed and open sets.

We start by introducing neighborhoods. Let (X,d) be a metric space.

Definition 209 *Given a point $x \in X$, the neighborhood $B_\varepsilon(x)$ of x of width ε is given by:*

$$B_\varepsilon(x) = \{y \in X : d(x,y) < \varepsilon\}.$$

In other words, $B_\varepsilon(x)$ is the set of all points of X whose distance from the given point x is lower than ε .

[2] It is easy to see that the case in which t^* is a boundary point of $[0,1]$, that is, $t^* = 0$ or $t^* = 1$, can be similarly studied.

Example 210 When $X = \mathbb{R}^n$ is endowed with the Euclidean distance d_2, we have

$$B_\varepsilon(x) = \left\{ y \in \mathbb{R}^n : \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} < \varepsilon \right\},$$

and therefore we find again the notion of spheric neighborhood seen in the basic courses. With respect to the distance d_1 we have:

$$B_\varepsilon(x) = \left\{ y \in \mathbb{R}^n : \sum_{i=1}^{n} |x_i - y_i| < \varepsilon \right\},$$

while with respect to d_∞ we have:

$$B_\varepsilon(x) = \left\{ y \in \mathbb{R}^n : \max_{1 \le i \le n} |x_i - y_i| < \varepsilon \right\}.$$

▲

Example 211 Consider $X = \mathcal{C}([0,1])$ and the distance d_∞ defined by (6.1). Given $f \in \mathcal{C}([0,1])$, we have:

$$B_\varepsilon(f) = \left\{ g \in \mathcal{C}([0,1]) : \max_{t \in [0,1]} |f(t) - g(t)| < \varepsilon \right\}.$$

Consider the two functions $f + \varepsilon$ and $f - \varepsilon$. Both belong to $C([0,1])$, and we have:

$$B_\varepsilon(f) = \{ g \in \mathcal{C}([0,1]) : f - \varepsilon < g < f + \varepsilon \}.$$

In fact, $\max_{t \in [0,1]} |f(t) - g(t)| < \varepsilon$ if and only if $|f(t) - g(t)| < \varepsilon$ for each $t \in [0,1]$, that is, if and only if $f(t) - \varepsilon < g(t) < f(t) + \varepsilon$ for each $t \in [0,1]$. ▲

Example 212 For $X = \mathcal{C}([0,1])$ with the distance defined by (6.3), we have:

$$B_\varepsilon(f) = \left\{ g \in \mathcal{C}([0,1]) : \int_0^1 |f(t) - g(t)| \, dt < \varepsilon \right\}.$$

▲

The notion of neighborhood for general metric spaces is conceptually similar to that studied in Calculus courses; the only novelty is given by the general notion of distance that is now available. The same similarity holds also for the other notions that we will now introduce, that is, interior points, accumulation points, open sets, etc. They are in fact defined in a completely analogous way to that seen in Calculus courses for $\mathbb{R}^n$. For this reason, we will now introduce them quite briefly.

- Given a set $A \subseteq X$, a point $x \in A$ is an *interior point* of A if there exists a neighborhood $B_\varepsilon(x)$ included in A, that is, $B_\varepsilon(x) \subseteq A$. A point $x \in A$ is called an *isolated point* of A if it is not of accumulation point, that is, if there exists a neighborhood $B_\varepsilon(x)$ such that $B_\varepsilon(x) \cap A = \{x\}$. The set of all interior points of A is denoted by $\overset{\circ}{A}$.

- Given a set $A \subseteq X$, a point $x \in X$ is called a *frontier* (or *boundary*) *point* of A if it is neither an interior point of A nor an interior point of A^c. A point $x \in X$ is called an *accumulation point* of A if each neighborhood $B_\varepsilon(x)$ contains a point $y \in A$ different from x. The set of the frontier points of A is denoted by ∂A, while A' denotes the set of the accumulation points of A (A' is usually called the derived set of A).

- A set whose points are all interior is called *open*.

- A set A is called *closed* if its complement A^c is open.

- A set A is called *bounded* if there exists $\varepsilon > 0$ and $x \in X$ such that $A \subseteq B_\varepsilon(x)$.

We now illustrate these notions with some examples.

Example 213 The sets $\emptyset$ and X are both open and closed. ▲

Example 214 Let A be a singleton in a metric space. We have $\overset{\circ}{A} = A' = \emptyset$ and $\partial A = A$. More generally, for each finite set we have $\overset{\circ}{A} = A' = \emptyset$ and $\partial A = A$. ▲

Example 215 Let $X = \mathbb{R}$ and $A = (0,1)$. We have $A = \overset{\circ}{A}$, i.e., A is an open set. Furthermore, $\partial A = \{0,1\}$ and $A' = [0,1]$. ▲

Example 216 Let $X = \mathbb{R}$ and $A = (0,1]$. We have $\overset{\circ}{A} = (0,1)$, $\partial A = \{0,1\}$, and $A' = [0,1]$. In this case A is neither open nor closed. ▲

Example 217 Let $A = [0,1]$. We have $\overset{\circ}{A} = (0,1)$, $\partial A = \{0,1\}$ and $A' = [0,1]$. Since $A^c = (-\infty, 0) \cup (1, +\infty)$ is open, it follows that A is closed. ▲

Example 218 Let $X = \mathbb{R}$ and $A = [0,1) \cup \{3\}$. We have $\overset{\circ}{A} = (0,1)$, $A' = [0,1]$, and $\partial A = \{0,1,3\}$. Note that the point $x = 3$ is isolated. The set A is neither open nor closed. ▲

Example 219 Let $X = \mathbb{R}^2$ and $A = \left\{x \in \mathbb{R}^2_+ : x_1 + x_2 = 1\right\}$. We have $\overset{\circ}{A} = \emptyset$ and $\partial A = A' = A$. ▲

We now present some basic properties of metric spaces. First of all we observe that the neighborhoods are actually open sets.

Proposition 220 *Neighborhoods are open sets.*

Proof. Let $B_\varepsilon(x)$ be a neighborhood of a given point $x \in X$. Let $y \in B_\varepsilon(x)$. We want to show that y is an interior point of $B_\varepsilon(x)$. By definition, $d(x,y) < \varepsilon$. Let $\varepsilon' > 0$ such that $d(x,y) = \varepsilon - \varepsilon'$, and consider the neighborhood $B_{\varepsilon'}(y)$ of y. We want to show that $B_{\varepsilon'}(y) \subseteq B_\varepsilon(x)$. Let $z \in B_{\varepsilon'}(y)$. We have:

$$d(x,z) \leq d(x,y) + d(y,z) < \varepsilon - \varepsilon' + \varepsilon' = \varepsilon,$$

and therefore $z \in B_\varepsilon(x)$, as desired. ■

The next important property of accumulation points shows that each of their neighborhoods contains a great number of elements of the reference set.

Proposition 221 *Let x be an accumulation point of a set A. Then, each neighborhood of x contains infinite points of A.*

Proof Suppose *per contra* that there exists a neighborhood $B_\varepsilon(x)$ of x that contains a finite number of points $\{x_1, ..., x_n\}$ of A, all distinct from x. As $\{x_1, ..., x_n\}$ is a finite set, $\min_{i=1,...,n} d(x, x_i)$ exists and it is strictly positive, i.e., $\min_{i=1,...,n} d(x, x_i) > 0$. Set $\varepsilon' = \min_{i=1,...,n} d(x, x_i)$ and consider the neighborhood $B_{\varepsilon'}(x)$. By construction, we have $B_{\varepsilon'}(x) \cap A \subseteq \{x\}$. Therefore, the only point of A that $B_\varepsilon(x)$ can contain is, at most, x itself. But this contradicts the hypothesis that x is an accumulation point of A. ■

The next result describes the behavior of the open sets with respect to the basic set operations.

Proposition 222 *If $\{G_i\}_{i \in I}$ is a collection of open sets, then $\bigcup_{i \in I} G_i$ is an open set. If I is a finite set $\{1, ..., n\}$, then $\bigcap_{i=1}^{n} G_i$ is an open set.*

In other words, the union of any number of open sets is an open set, while only the intersection of a finite number of open sets is still an open set.

Proof We start by proving that the set $\bigcup_{i \in I} G_i$ is open. Let $x \in \bigcup_{i \in I} G_i$. By definition, there exists $i \in I$ such that $x \in G_i$. As G_i is open, there exists a neighborhood $B_\varepsilon(x)$ of x such that $B_\varepsilon(x) \subseteq G_i$. *A fortiori*, $B_\varepsilon(x) \subseteq \bigcup_{i \in I} G_i$ and therefore x is an interior point of $\bigcup_{i \in I} G_i$. It follows that $\bigcup_{i \in I} G_i$ is an open set.

Consider now $\bigcap_{i=1}^{n} G_i$. Let $x \in \bigcap_{i=1}^{n} G_i$. By definition, for each $i \in I$ there exists a neighborhood $B_{\varepsilon_i}(x)$ of x such that $B_{\varepsilon_i}(x) \subseteq G_i$. Set $\varepsilon = \min_{i=1,...,n} \varepsilon_i$. Since the collection $\{\varepsilon_i\}_{i=1}^{n}$ is finite, such minimum exists and $\varepsilon > 0$. Consider now $B_\varepsilon(x)$. By definition, $B_\varepsilon(x) \subseteq \bigcap_{i=1}^{n} B_{\varepsilon_i}(x) \subseteq \bigcap_{i=1}^{n} G_i$. Therefore, x is an interior point of $\bigcap_{i=1}^{n} G_i$, which is therefore an open set. ■

To derive the corresponding properties of the closed sets we need the following result.

Lemma 223 *Given any collection $\{A_i\}_{i\in I}$ of sets, we have*

$$\left(\bigcup_{i\in I} A_i\right)^c = \bigcap_{i\in I} A_i^c.$$

Proof We show that $\left(\bigcup_{i\in I} A_i\right)^c \subseteq \bigcap_{i\in I} A_i^c$. Let $x \in \left(\bigcup_{i\in I} A_i\right)^c$. Therefore, $x \notin A_i$ for each $i \in I$, that is, $x \in A_i^c$ for each $i \in I$. It follows that $x \in \bigcap_{i\in I} A_i^c$, as desired.

To show that $\bigcap_{i\in I} A_i^c \subseteq \left(\bigcup_{i\in I} A_i\right)^c$, consider $x \in \bigcap_{i\in I} A_i^c$. Therefore, $x \in A_i^c$ for each $i \in I$, that is, $x \notin A_i$ for each $i \in I$. It follows that $x \notin \bigcup_{i\in I} A_i$, that is, $x \in \left(\bigcup_{i\in I} A_i\right)^c$. This completes the proof. ■

The next result is, at this point, an obvious consequence of Proposition 222 and of Lemma 223.

Corollary 224 *If $\{F_i\}_{i\in I}$ is a collection of closed sets, then $\bigcap_{i\in I} F_i$ is a closed set. If I is a finite set $\{1, ..., n\}$, then $\bigcup_{i=1}^{n} F_i$ is a closed set.*

The behavior of closed sets is therefore specular relative to that of open sets: the intersection of a any number of closed sets is a closed set, while only the union of a finite number of closed sets is still a closed set.

Example 225 Let $G_n = (-1/n, 1/n)$. We have $\bigcap_{i\geq 1} G_n = \{0\}$, that is, the intersection of the infinite collection of open sets G_n is the singleton $\{0\}$. Since this singleton is not open, this shows that the hypothesis that I is finite is essential in Proposition 222. Similarly, consider for example $F_n = [1/n, 1 - (1/n)]$. We have $\bigcup_{n\geq 1} F_n = (0, 1)$, and therefore an infinite union of closed sets may not be a closed set. ▲

In general, the collection of open sets (and so of closed sets) of a given metric space can be complicated to describe. For this reason, the metric spaces where this collection can be described explicitly are of special interest. For instance, this is the case for the discrete metric, as Exercise13.0.30 shows. A much more important example is given by the open sets of the real line, whose structure is described in the following result (whose proof we omit).

Theorem 226 *Each open set of the real line is the union, finite or infinite, of open intervals that are pairwise disjoint.*

In other words, given an open set G of $\mathbb{R}$, there exists a collection of open intervals $\{(a_i, b_i)\}_{i\in I}$ such that $G = \bigcup_{i\in I} (a_i, b_i)$, with $(a_i, b_i) \cap (a_{i'}, b_{i'}) = \emptyset$ for each $i, i' \in I$.[3]

6.2.1 A Closer Look at Closed Sets

The definition we gave of closed set is a little bit "mechanical": a set is closed if its complement is open. We now have a more close look at closed sets, in order to better understand their nature.

Theorem 227 *A set is closed if and only if it contains all its accumulation points.*

This theorem gives us a characterizing property of closed sets: a set is closed if it contains all its accumulation points, and, among all the subsets of a metric space, only the closed ones satisfy this property. In other words, the property of containing all its own accumulation points is a property that distinguishes closed sets among all subsets of a metric space.

Proof Let A be closed. We prove that $A' \subseteq A$. By contradiction, assume that this is not true, i.e., that there exists $x \in A'$ such that $x \notin A$. As A is closed, A^c is open. The point $x \in A^c$ is therefore an interior point of A^c, and so there exists a neighborhood $B_\varepsilon(x)$ of x such that $B_\varepsilon(x) \subseteq A^c$, that is, such that $B_\varepsilon(x) \cap A = \emptyset$. But this contradicts the definition of accumulation point. We conclude that $A' \subseteq A$.

Viceversa, assume that $A' \subseteq A$. We want to show that A is closed, i.e., that A^c is open. Let $x \in A^c$. As $A' \subseteq A$, the point x is not an accumulation point of A. There exists therefore a neighborhood $B_\varepsilon(x)$ of x such that $B_\varepsilon(x) \cap A \subseteq \{x\}$. As $x \notin A$, it follows that $B_\varepsilon(x) \cap A = \emptyset$, that is, $B_\varepsilon(x) \subseteq A^c$. The point x is therefore an interior point of A^c and, since x was an arbitrary point of A^c, the set A^c is open. ■

Example 228 The inclusion $A' \subseteq A$ in Theorem 227 can be strict, and the set $A - A'$ is formed by the isolated points of A. For example, let $A = [0, 1] \cup \{-1, 4\}$. The set A is closed and we have $A' = [0, 1]$. Therefore, A' is strictly included in A and the set $A - A' = \{-1, 4\}$ is formed by the isolated points of A. ▲

Theorem 229 *A set is closed if and only if it contains all its frontier points.*

[3] In Theorem 226 we can have $a = -\infty$ and $b = +\infty$. The collection $\{(a_i, b_i)\}_{i\in I}$ can therefore contain the open intervals $(-\infty, b)$, $(a, +\infty)$ and $(-\infty, +\infty)$.

The property of containing all its frontier points is therefore a further property that characterizes closed sets among all the subsets of a metric space.

Proof Let A be closed. We prove that $\partial A \subseteq A$. Let $x \in \partial A$. By definition of frontier point, x is not an interior point neither of A nor of A^c. As A^c is open, all its points are interior and therefore x cannot belong to such set. It follows that $x \in A$, and so $\partial A \subseteq A$.

Viceversa, assume that $\partial A \subseteq A$. We want now to show that A is closed, i.e., that A^c is open. Let $x \in A^c$. As $\partial A \subseteq A$, $x \notin \partial A$. Therefore, either x is an interior point of A or it is an interior point of A^c. Since $x \in A^c$, the point x is an interior point of A^c, which is therefore an open set. ■

Example 230 Also in this case the inclusion $\partial A \subseteq A$ in Theorem 229 can be strict. The set $A - \partial A$ consists of all interior points of A, that is, $A - \partial A = \overset{\circ}{A}$. Consider again the closed set $A = [0,1] \cup \{-1,4\}$. We have $\partial A = \{-1,0,1,4\}$; the set $A - \partial A = (0,1)$ is formed by the interior points of A. ▲

Until now, to prove that a set A is closed we had to consider its complement A^c, and to prove that it is an open set. Thanks to the characterizations given in Theorems 227 and 229, we now have two criteria that can be used to establish directly whether a set A is closed (the choice among the two criteria is only a matter of convenience: in some cases it may be easier to use one of the two criteria).

Example 231 Let $A = [0,1) \cup [2,3]$. We have $A' = [0,1] \cup [2,3]$ and therefore $A' \not\subseteq A$. By Theorem 227, A is not closed. We can get the same result using Theorem 229. In fact, $\partial A = \{0,1,2,3\} \not\subseteq A$. ▲

Example 232 Let A be a finite set of a metric space. In Example 214 we saw that $A' = \emptyset$. Therefore, by Theorem 227 the set A is closed. On the other hand, $\partial A = A$ and therefore we can arrive at the same conclusion via Theorem 229. ▲

6.2.2 Closure

Given a set A, it is easy to see that the set of the interior points $\overset{\circ}{A}$ is an open set and that $\overset{\circ}{A}$ is actually the largest open set contained in A. That is, if G is an open set such that $G \subseteq A$, we have $G \subseteq \overset{\circ}{A}$ (see Exercise 13.0.32). In a similar way, it is possible to ask which is the smallest closed set that contains A. The next definition is motivated by this question.

Definition 233 *Given a set A, its closure $\overline{A}$ is given by the set $A \cup \partial A$.*

The closure $\overline{A}$ of A is therefore the union of the set itself with all its frontier points.

Example 234 Let $A = (0,1) \cup [2,3] \cup \{-10,10\}$. As $\partial A = \{-10,0,1,2,3,10\}$, we have

$$\begin{aligned} \overline{A} &= A \cup \partial A = (0,1) \cup [2,3] \cup \{-10,0,1,2,3,10\} \\ &= [0,1] \cup [2,3] \cup \{-10,10\}. \end{aligned}$$

▲

The next result collects the most important properties of the closure. In particular, by (i) the set $\overline{A}$ is closed and, by (iii), is the smallest closed set that contains A.

Theorem 235 *Given a set A, we have:*

(i) $\overline{A}$ *is closed,*

(ii) $A = \overline{A}$ *if and only if* A *is closed,*

(iii) $\overline{A} \subseteq F$ *for each closed set* F *such that* $A \subseteq F$*,*

(iv) $\overline{A} = A \cup A'$*.*

Proof (i) We prove that $\overline{A}$ is closed, i.e., that $\overline{A}^c$ is open. To this end we prove that $\overline{A}^c = \overset{\circ}{A}{}^c$, where $\overset{\circ}{A}{}^c$ is the set of the interior points of A^c. Let $x \in \overline{A}^c$. As $\overline{A}^c = (A \cup \partial A)^c = A^c \cap \partial(A^c)$, we have $x \in A^c$ and $x \notin \partial A$. From $x \notin \partial A$ it follows that either x is an interior point of A or it is an interior point of A^c. Since $x \in A^c$, x is therefore an interior point of A^c, that is, $x \in \overset{\circ}{A}{}^c$. This proves that $\overline{A}^c \subseteq \overset{\circ}{A}{}^c$.

Viceversa, let $x \in \overset{\circ}{A}{}^c$. We have $x \in A^c$. Furthermore, $x \notin \partial A$ because x is an interior point of A^c. It follows that $x \in A^c \cap \partial(A^c)$ and therefore $x \in \overline{A}^c$. In sum, $\overline{A}^c = \overset{\circ}{A}{}^c$ and therefore $\overline{A}^c$ is an open set.

(ii) The "only if" is an immediate consequence of (i). Consider the "if." Let A be a closed set. We want to prove that $A = \overline{A}$. By Theorem 229, $\partial A \subseteq A$ and therefore $\overline{A} = A \cup \partial A = A$, as desired.

(iii) Let F be a closed set such that $A \subseteq F$. Since $\overline{A} = A \cup \partial A$, to prove that $\overline{A} \subseteq F$ we have to show that $\partial A \subseteq F$. Suppose *per contra* that there exists $x \in \partial A$ such that $x \notin F$. Since F is closed, its complement F^c is open and therefore x is an interior point of F^c. There exists therefore a neighborhood $B_\varepsilon(x)$ of x such that $B_\varepsilon(x) \subseteq F^c$. From $A \subseteq F$ it follows that $F^c \subseteq A^c$, and so $B_\varepsilon(x) \subseteq F^c \subseteq A^c$. The point x is thus an interior point also of A^c, and this implies $\partial A \subseteq \overset{\circ}{A}{}^c$. But, this is a contradiction

because by definition frontier points are neither interior points of A nor interior points of A^c. Therefore, $\partial A \subseteq F$.

(iv) We begin by proving that $A \cup A' \subseteq \overline{A}$. Since $A \subseteq \overline{A}$, what we have to prove is that $A' \subseteq \overline{A}$. Let $x \in A'$. If $x \in A$, then we trivially have $x \in \overline{A}$. Suppose $x \notin A$. We prove that $x \in \partial A$, and so $x \in \overline{A}$. As $x \in A'$, for each neighborhood $B_\varepsilon(x)$ of x there exists $y \in A$ such that $y \in B_\varepsilon(x)$. Therefore, x cannot be an interior point of A^c. On the other hand, being $x \notin A$, x is not an interior point of A. We conclude that $x \in \partial A$, as desired.

It remains to prove that $\overline{A} \subseteq A \cup A'$. In view of point (iii), it is enough to prove that $A \cup A'$ is closed. Let $x \in (A \cup A')^c = A^c \cap (A')^c$. As $x \in (A')^c$, x is not an accumulation point of A and there exists therefore a neighborhood $B_\varepsilon(x)$ of x such that $B_\varepsilon(x) \cap A = \emptyset$. On the other hand, we have also $B_\varepsilon(x) \cap A' = \emptyset$. In fact, suppose that this is not true and let $y \in B_\varepsilon(x) \cap A'$. By Proposition 220, $B_\varepsilon(x)$ is an open set and so there exists a neighborhood $B_\varepsilon(y)$ of y such that $B_\varepsilon(y) \subseteq B_\varepsilon(x)$. As $y \in A'$, we have

$$\emptyset \neq B_\varepsilon(y) \cap A \subseteq B_\varepsilon(x) \cap A,$$

which contradicts $B_\varepsilon(x) \cap A = \emptyset$. Therefore, $B_\varepsilon(x) \cap A' = \emptyset$, and we conclude that $B_\varepsilon(x) \subseteq A^c \cap (A')^c$. The set $(A \cup A')^c$ is open, and so $A \cup A'$ is closed. ■

Point (iv) shows that a possible alternative definition of closure, sometimes actually adopted in the literature, is as union of A with its derived set A'.

The closure of a set A thus contains three types of points:

- the points of accumulation of A that belong to A (i.e., the set $A' \cap A$);
- the points of accumulation of A that do not belong to A (i.e., the set $A' \cap A^c$);
- the isolated points of A (i.e., the set $A - A'$).

In Example 234, these three types of points are respectively given by:

$$A' \cap A = (0,1) \cup [2,3], \quad A' \cap A^c = \{0,1\}, \quad \text{and} \quad A - A' = \{-10, 10\}.$$

6.3 Sequences

6.3.1 Definition

>From basic Calculus courses we know that sequences of real numbers are defined as functions $f : \mathbb{N} - \{0\} \to \mathbb{R}$ that associate to each natural number $n \geq 1$ a real number $f(n)$. For brevity, the image $f(n)$ is typically denoted by x_n.

Example 236 The sequence of odd numbers

$$\{1, 3, 5, 7, ...\} \tag{6.5}$$

has as generic element $x_n = 2(n-1)+1$ for $n \geq 2$, while the sequence

$$\left\{1, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{4}}, \frac{1}{\sqrt{8}}, ...\right\} \tag{6.6}$$

has as generic element $x_n = 1/\sqrt{2^{n-1}}$ for $n \geq 2$.

Formally, the sequence (6.5) corresponds to the function $f : \mathbb{N} - \{0\} \to \mathbb{R}$ given by $f(n) = 2(n-1)+1$ for each $n \geq 1$, while the sequence (6.6) corresponds to the function $f : \mathbb{N} - \{0\} \to \mathbb{R}$ given by $f(n) = 1/\sqrt{2^{n-1}}$ for each $n \geq 1$. ▲

Example 237 The sequence with generic element $x_n = (-1)^n$ is given by

$$\{-1, 1, -1, 1, ...\}.$$

Therefore, in a sequence the same value can appear several times. For example, the constant sequence

$$\{2, 2, 2, ...\}$$

is formed only by numbers 2 and its generic element is $x_n = 2$ for each $n \geq 1$ (the corresponding function f is therefore the constant $f(n) = 2$ for each $n \geq 1$). ▲

Still from basic calculus we know that a sequence of real numbers $\{x_n\}_{n\geq 1}$ *converges* to a real number x if, for each $\varepsilon > 0$, there exists $\overline{n} \geq 1$ such that

$$|x_n - x| < \varepsilon \tag{6.7}$$

for each $n \geq \overline{n}$. In this case, we write $\lim_{n\to\infty} x_n = x$ or $x_n \to x$. In other words, we have $\lim_{n\to\infty} x_n = x$ if the elements x_n become, when n grows, closer and closer to x.

Example 238 The sequence $\left\{1, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{5}}, \frac{1}{\sqrt{7}}, ...\right\}$ converges to $x = 0$, while the sequence $\{2, 2, 2, ...\}$ converges to $x = 2$. ▲

Sequences may well not converge to any real number. For example, neither the sequence $\{1, 3, 5, 7, ...\}$ nor the sequence $\{-1, 1, -1, 1, ...\}$ converge; the first diverges to $+\infty$, while the second continues to alternate between the values -1 and 1.

These elementary notions can be extended in a natural way from $\mathbb{R}$ to a generic metric space X. Formally, a sequence in a metric space X is a function $f : \mathbb{N} - \{0\} \to X$ that associates to each natural number $n \geq 1$ a point $f(n)$ of X. In this more general case as well, we still denote the image $f(n)$ by x_n.

Example 239 Let $X = \mathbb{R}^n$. In this case $\{x_n\}_{n\geq 1}$ is a sequence of vectors, usually denoted by $\{x^n\}_{n\geq 1}$ to distinguish vectors and their components.[4] For example, $x^n = (x_1^n, x_2^n) = (1/n, 1/n^2)$ is the generic element of the sequence

$$\left\{(1,1), \left(\frac{1}{2}, \frac{1}{4}\right), \left(\frac{1}{3}, \frac{1}{9}\right), ..., \left(\frac{1}{n}, \frac{1}{n^2}\right), ...\right\} \tag{6.8}$$

of vectors of $\mathbb{R}^2$, while $x^n = (0, 1/n, n^3)$ the generic element of the sequence

$$\left\{(0,1,1), \left(0, \frac{1}{2}, 8\right), \left(0, \frac{1}{3}, 81\right), ..., \left(0, \frac{1}{n}, n^3\right), ...\right\} \tag{6.9}$$

of vectors of $\mathbb{R}^3$. ▲

Example 240 Let $X = \mathcal{C}([0,1])$. In this case $\{x_n\}_{n\geq 1}$ is a sequence of continuous functions, usually denoted by $\{f_n\}_{n\geq 1}$. For example, $f_n(t) = t^n$ is the generic element of the sequence

$$\left\{t, t^2, t^3, ..., t^n, ...\right\} \tag{6.10}$$

of continuous functions on $[0,1]$, while $f_n(t) = t^n/n$ is the generic element of the sequence

$$\left\{t, \frac{t^2}{2}, \frac{t^3}{3}, ..., \frac{t^n}{n}, ...\right\} \tag{6.11}$$

▲

As to convergence, we have the following natural definition.

Definition 241 *A sequence $\{x_n\}_{n\geq 1}$ of points of a metric space X converges to $x \in X$ if for each neighborhood $B_\varepsilon(x)$ of x there exists $\overline{n} \geq 1$ such that $x_n \in B_\varepsilon(x)$ for each $n \geq \overline{n}$. In this case, we write $\lim_{n\to\infty} x_n = x$ or $x_n \to x$.*

The condition $x_n \in B_\varepsilon(x)$ generalizes in an obvious way (6.7). The point x is called *limit* of the sequence $\{x_n\}_{n\geq 1}$; a sequence that has a limit point is called *convergent*.

Lemma 242 *A sequence $\{x_n\}_{n\geq 1}$ of points of a metric space X converges to $x \in X$ if and only if $\lim_{n\to\infty} d(x_n, x) = 0$.*

Proof Let $\lim_{n\to\infty} x_n = x$. Set $\varepsilon = 1/m$, with $m \geq 1$. By Definition 241, there exists $\overline{n} \geq 1$ such that $x_n \in B_{\frac{1}{m}}(x)$ for each $n \geq \overline{n}$, that is, such that $d(x, x_n) < 1/m$ for each $n \geq \overline{n}$. Since this is true for every $m \geq 1$, we have

$$0 \leq \lim_{n\to+\infty} d(x, x_n) \leq \lim_{m\to+\infty} \frac{1}{m} = 0,$$

[4] As in the first chapters, we use pedices and apices to distinguish the vectors of $\mathbb{R}^n$ and their components. In particular, the i-th component x_i^n of the vector x^n is denoted by the pedix i, while the apix n denotes the vector itself. Naturally, there is no relation between the pedix n in x_n and the apix n in $\mathbb{R}^n$. To avoid any possible confusion, sometimes we use the notation $\mathbb{R}^m$ instead of $\mathbb{R}^n$.

and so $\lim_{n\to+\infty} d(x, x_n) = 0$.

Viceversa, suppose that $\lim_{n\to+\infty} d(x, x_n) = 0$. Let $\varepsilon > 0$. There exists $\overline{n} \geq 1$ such that $d(x, x_n) < \varepsilon$ for each $n \geq \overline{n}$. Therefore, $x_n \in B_\varepsilon(x)$ for each $n \geq \overline{n}$, as desired. ■

Thanks to this simple result we can reduce the study of the convergence of sequences in general metric spaces to that, much simpler, of the convergence to 0 of the sequence of real numbers $\{d(x_n, x)\}_{n\geq 1}$. In other words, to verify if $x_n \to x$ it is enough to verify that $d(x_n, x) \to 0$.

Example 243 Let $X = \mathbb{R}^2$, endowed with the Euclidean metric, and consider the sequence (6.8). We have

$$d_2\left((0,0), \left(\frac{1}{n}, \frac{1}{n^2}\right)\right) = \sqrt{\left(\frac{1}{n}\right)^2 + \left(\frac{1}{n^2}\right)^2} = \frac{\sqrt{n^2+1}}{n^2} \to 0$$

and therefore the sequence converges to $x = (0, 0)$. The sequence (6.9), instead, does not converges to any vector of $\mathbb{R}^3$. ▲

Example 244 Let $X = \mathcal{C}([0, 1])$, endowed with the metric d_∞, and let $\mathbf{0}$ be the identically null function. For the sequence $\{f_n\}_{n\geq 1}$ given by (6.11) we have

$$d_\infty(f_n, \mathbf{0}) = \max_{t\in[0,1]} |f_n(t)| = \max_{t\in[0,1]} \frac{t^n}{n} \leq \frac{1}{n} \to 0,$$

which implies $d_\infty(f_n, \mathbf{0}) \to 0$ since $d_\infty(f_n, \mathbf{0}) \geq 0$ for each $n \geq 1$. The sequence converges therefore to the function $\mathbf{0} \in C([0, 1])$.

On the contrary, the sequence (6.10) does not converge to any function of $C([0, 1])$. In fact, suppose *per contra* that this is the case, and let $f \in C([0, 1])$ be such that $t^n \overset{d_\infty}{\to} f$. Therefore, $\max_{t\in[0,1]} |t^n - f(t)| \to 0$, so that $|t^n - f(t)| \to 0$ for each $t \in [0, 1]$, that is, $t^n \to f(t)$ for each $t \in [0, 1]$. Since $t^n \to 0$ for each $t \in [0, 1)$, it follows that $f(t) = 0$ for each $t \in [0, 1)$. Since f is continuous, this implies $f(1) = 0$, and therefore $f = \mathbf{0}$. But,

$$d_\infty(f_n, \mathbf{0}) = \max_{t\in[0,1]} |f_n(t)| = \max_{t\in[0,1]} t^n = 1, \qquad \forall n \geq 1$$

and the sequence thus does not converge to $\mathbf{0}$. This contradiction shows that the sequence (6.10) is not convergent in $C([0, 1])$. ▲

Example 245 Let again $X = C([0, 1])$, this time endowed with the metric d_1. For the sequence $\{f_n\}_{n\geq 1}$ given by (6.11) we have

$$d_1(f_n, \mathbf{0}) = \int_0^1 |f_n(t)| = \int_0^1 \frac{t^n}{n} dt = \frac{1}{n(n+1)} \to 0.$$

The sequence therefore converges to the function $\mathbf{0} \in C([0,1])$. As to the sequence (6.10), we have

$$d_1(f_n, \mathbf{0}) = \int_0^1 |f_n(t)| = \int_0^1 t^n dt = \frac{1}{n+1} \to 0,$$

and so also this sequence converges to $\mathbf{0}$. It is important to note that this is an example of a sequence that is convergent according to a metric (the d_1), but divergent according to another metric (the d_∞). ▲

We conclude this subsection by considering convergence in the metric space $\mathbb{R}^n$. The next result shows that a sequence of vectors converges if and only if the sequences in $\mathbb{R}$ formed by their components converge.

Proposition 246 *Let $\{x^n\}_{n\geq 1}$ be a sequence of vectors $x^n = (x_1^n, ..., x_m^n)$ of $\mathbb{R}^m$. We have $x^n \xrightarrow{d_1} x \in \mathbb{R}^m$ if and only if $x_i^n \to x_i \in \mathbb{R}$ for each $i = 1, ..., m$.*

In other words, x^n converges to x under the metric d_1 if and only if for every $i = 1, .., n$ the sequence $\{x_i^n\}_{n\geq 1}$ of the i-th components of the vectors x^n converges to the i-th component x_i of the vector x.

Proof Suppose that $x^n \xrightarrow{d_1} x \in \mathbb{R}^m$. By Lemma 242,

$$\sum_{i=1}^n |x_i^n - x_i| = d_1(x_n, x) \to 0.$$

Therefore,

$$0 \leq |x_i^n - x_i| \leq \sum_{i=1}^n |x_i^n - x_i| \to 0, \qquad \text{for each } i = 1, ..., m,$$

and consequently $x_i^n \to x_i \in \mathbb{R}$ for each $i = 1, ..., m$.

Viceversa, suppose that $x_i^n \to x_i \in \mathbb{R}$ for each $i = 1, ..., m$. Let $\varepsilon > 0$. For each $i = 1, ..., m$ there exists $\overline{n}_i \geq 1$ such that $|x_i^n - x_i| < \varepsilon/m$ for each $n \geq \overline{n}_i$. Therefore, we have $|x_i^n - x_i| < \varepsilon/m$ for each $n \geq \max_{i=1,...,m} \overline{n}_i$ and each $i = 1, ..., m$. It follows that for each $n \geq \max_{i=1,...,m} \overline{n}_i$ we have

$$d_1(x^n, x) = \sum_{i=1}^m |x_i^n - x_i| < \frac{\varepsilon}{m} + \cdots + \frac{\varepsilon}{m} = \varepsilon,$$

and therefore $x^n \in B_\varepsilon(x)$. In conclusion, $x^n \xrightarrow{d_1} x$. ■

In Exercise 13.0.33 we will see how Proposition 246 holds also for the metrics d_2 and d_∞.

6.3.2 First Properties

The next uniqueness result shows that sequences of points in metric spaces can converge to at most a unique point.

Theorem 247 *A convergent sequence of points of a metric space has a unique limit.*

Proof Let $\{x_n\}_{n\geq 1}$ be a convergent sequence of a metric space X. Suppose there exist $x, y \in X$ such that $x_n \to x$ and $x_n \to y$. We want to prove that $x = y$. Suppose that this is not true, i.e., that $x \neq y$. For ε sufficiently small, we have two neighborhoods $B_\varepsilon(x)$ and $B_\varepsilon(y)$ such that $B_\varepsilon(x) \cap B_\varepsilon(y) = \emptyset$. By Definition 241, there exist $\overline{n}_1, \overline{n}_2 \geq 1$ such that $x_n \in B_\varepsilon(x)$ for each $n \geq \overline{n}_1$ and $x_n \in B_\varepsilon(y)$ for each $n \geq \overline{n}_2$. Therefore, $x_n \in B_\varepsilon(x) \cap B_\varepsilon(y)$ for each $n \geq \overline{n}_1 \vee \overline{n}_2$, which contradicts $B_\varepsilon(x) \cap B_\varepsilon(y) = \emptyset$.[5] It follows that $x = y$, and the limit is therefore unique. ■

The next result shows that when a sequence converges to a point x, in each neighborhood of this point lies the majority of the points of the sequence.

Proposition 248 *A sequence $\{x_n\}_{n\geq 1}$ of points of a metric space X converges to $x \in X$ if and only if each neighborhood $B_\varepsilon(x)$ of x contains all the points of the sequence, except at most a finite number of them.*

Proof Suppose that $x_n \to x$. For each $\varepsilon > 0$, there exists $\overline{n} \geq 1$ such that $x_n \in B_\varepsilon(x)$ for each $n \geq \overline{n}$. Therefore, except for the points x_n with $1 \leq n < \overline{n}$, all other points of the sequence belong to $B_\varepsilon(x)$.

Viceversa, given a neighborhood $B_\varepsilon(x)$ of x, suppose that all points of the sequence belong to it, except at most a finite number of them. Denote by $\{x_{n_k}\}_{k=1}^{m}$ the set of the points of the sequence that do not belong to $B_\varepsilon(x)$. Setting $\overline{n} = n_m + 1$, we therefore have that $x_n \in B_\varepsilon(x)$ for each $n \geq \overline{n}$. Since this is true for each neighborhood $B_\varepsilon(x)$ of x, it follows that $x_n \to x$. ■

Given a sequence $f : \mathbb{N} - \{0\} \to X$ of points of a metric space, $f(\mathbb{N})$ is called *image* of the sequence. Naturally, the image does not take into account the repetitions that may occur in the sequence. For example, the constant sequence $\{2, 2, 2, ...\}$ in $\mathbb{R}$ has as image the singleton $\{2\}$, while the sequence $\{-1, 1, -1, 1, ...\}$ has as image the set of the two elements $\{-1, 1\}$.

A sequence $\{x_n\}_{n\geq 1}$ is said to be *bounded* if its image is a bounded set of X.

Proposition 249 *A convergent sequence in a metric space is bounded.*

[5] $\overline{n}_1 \vee \overline{n}_2$ denotes $\max\{\overline{n}_1, \overline{n}_2\}$.

Proof Suppose that $x_n \to x$. Setting $\varepsilon = 1$, there exists $\overline{n} \geq 1$ such that $x_n \in B_1(x)$ for each $n \geq \overline{n}$. Let $M > 0$ be such that $M > \max\{1, d(x_1, x), ..., d(x_{\overline{n}-1}, x)\}$. We have $d(x_n, x) < M$ for each $n \geq 1$ and, therefore, the image of the sequence is contained in the neighborhood $B_{M'}(x)$. The sequence is therefore bounded. ■

Proposition 249 gives us a simple sufficient condition for a sequence to diverge: if the image of the sequence is not bounded, then the sequence diverges. For example, the sequence that has as generic element $x_n = 2n$ diverges because it has as image the unbounded set $\{1, 2, 3, ..., n, ...\}$.

A sequence $\{x_n\}_{n\geq 1}$ in $\mathbb{R}$ is *increasing* if $x_n \leq x_{n+1}$ for each $n \geq 1$, while it is *decreasing* if $x_n \geq x_{n+1}$ for each $n \geq 1$. In general, a sequence in $\mathbb{R}$ is *monotonic* if it is increasing or decreasing (it is both increasing and decreasing if and only if it is constant). For this class of sequences, boundedness is a necessary and sufficient condition for convergence.[6]

Proposition 250 *A monotonic sequence in* $\mathbb{R}$ *is convergent if and only if it is bounded.*

Proof Let $\{x_n\}_{n\geq 1}$ be an increasing sequence in $\mathbb{R}$ (the proof of the decreasing case is similar). If it is convergent, Proposition 249 guarantees that it is bounded.

Viceversa, suppose that this sequence is bounded. We want to prove that it is convergent. Let E be the image of the sequence. By hypothesis, it is a bounded subset of $\mathbb{R}$. By the completeness of $\mathbb{R}$, $\sup E$ exists. Set $x = \sup E$. We now show that $x_n \to x$. Let $\varepsilon > 0$. Since x is the supremum of E, we have: (i) $x \geq x_n$ for each $n \geq 1$, (ii) there exists an element of E, denoted by $x_{\overline{n}}$, such that $x_{\overline{n}} > x - \varepsilon$.[7] Since $\{x_n\}_{n\geq 1}$ is an increasing sequence, it follows that

$$x \geq x_n \geq x_{\overline{n}} > x - \varepsilon, \qquad \forall n \geq \overline{n}$$

and therefore $x_n \in B_\varepsilon(x)$ for each $n \geq \overline{n}$, as desired. ■

Notation. The limit x of an increasing sequence $\{x_n\}_{n\geq 1}$ is easily seen to be such that $x_n \leq x$ for each $n \geq 1$. For this reason, often this convergence is denoted by $x_n \uparrow x$. Analogously, often it is denoted by $x_n \downarrow x$ the convergence of a decreasing sequence $\{x_n\}_{n\geq 1}$ to a limit point x.

[6] For simplicity, Proposition 250 considers sequences of real numbers. The general case of sequences in $\mathbb{R}^n$ follows from Proposition 246 and from Exercise 13.0.33.

[7] For these properties, see for example Ambrosetti and Musu (1988) p. 37.

There exists a partial converse of Proposition 249. To state it, we introduce subsequences. Given a sequence of distinct natural numbers $\{n_k\}_{k\geq 1}$, i.e., such that $n_1 < n_2 < n_3 < \cdots < n_k < \cdots$, the sequence $\{x_{n_k}\}_{k\geq 1}$ is said to be a *subsequence* of $\{x_n\}_{n\geq 1}$.

Naturally, the image of a subsequence is included in that of the sequence.

Example 251 Consider the sequence in $\mathbb{R}$

$$\left\{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, ..., \frac{1}{n}, ...\right\} \tag{6.12}$$

with generic point $x_n = 1/n$. A subsequence is given by

$$\left\{1, \frac{1}{3}, \frac{1}{5}, \frac{1}{7}, ..., \frac{1}{2k+1}, ...\right\},$$

where the sequence $\{n_k\}_{k\geq 1}$ of natural numbers considered is that of the odd numbers $\{1, 3, 5, ...\}$. Another subsequence of (6.12) is given by

$$\left\{\frac{1}{2}, \frac{1}{8}, \frac{1}{16}, ..., \frac{1}{2^n}, ...\right\},$$

where the sequence $\{n_k\}_{k\geq 1}$ of natural numbers considered is that of the powers of 2, that is, $\{2, 2^2, 2^3, ...\}$. ▲

Example 252 As to the sequence in $\mathbb{R}$ with generic point $x_n = (-1)^n$, a subsequence is given by

$$\{1, 1, 1, ..., 1, ...\},$$

where the sequence $\{n_k\}_{k\geq 1}$ of natural numbers considered is that of the even numbers. ▲

As the last example shows, it can happen that even though the original sequence diverges, there exist subsequences that are convergent. In other words, it is sometimes possible that from a divergent behavior we can "extract" a convergent one, by selecting in a proper way among the points of the sequence. In Example 252 we have an oscillating sequence, from which we have selected a constant subsequence considering only the points with even index.

We conclude with some simple "algebraic" properties of sequences of vectors, whose proof we omit. In the statement the space $\mathbb{R}^n$ is endowed with any one of the metrics d_1, d_2 and d_∞.

Proposition 253 *Let $\{x^n\}_{n\geq 1}$ and $\{y^n\}_{n\geq 1}$ be two sequences of vectors of $\mathbb{R}^n$, with $x^n \to x$ and $y^n \to y$. Then:*

(i) $\alpha x^n + \beta y^n \to \alpha x + \beta y$ *for each* $\alpha, \beta \in \mathbb{R}$,

(ii) $x^n \cdot y^n \to x \cdot y$,

(iii) $x \geq y$ *if* $x^n \geq y^n$ *for each* $n \geq 1$.

6.3.3 Sequences and Topology

Using sequences we can give a characterization of the closure of a set.

Theorem 254 *Let A be a set of a metric space X. We have $x \in \overline{A}$ if and only if there exists a sequence $\{x_n\}_{n\geq 1} \subseteq A$ such that $x_n \to x$.*

Proof Let $x \in \overline{A}$. If x is an isolated point of A, set $x_n = x$ for each $n \geq 1$. If $x \in A'$, consider the neighborhoods $B_{\frac{1}{n}}(x)$ of x. Each of these neighborhoods contains a point x_n of A distinct from x. The sequence $\{x_n\}_{n\geq 1}$ clearly converges to x, as desired.

Let now $x \in X$ be such that there exists a sequence $\{x_n\}_{n\geq 1} \subseteq A$ with $x_n \to x$. Each neighborhood $B_\varepsilon(x)$ of x contains points of the sequence, that is, $B_\varepsilon(x) \cap A \neq \emptyset$. The point x therefore is not an interior point of A^c, that is, $x \notin \overset{\circ}{A^c}$. It follows that $x \in \overline{A}$ since in the first part of the proof of point (i) of Theorem 235 we proved that $\overset{\circ}{A^c} = \overline{A}^c$. ■

In this result is therefore crucial that in a sequence it is possible that some points can appear repeatedly. In fact, this allows to consider an isolated point x of A as a limit of the constant sequence $\{x, x, x, ..., x, ...\}$.

Next result is an immediate consequence of Theorem 254 and is probably the most useful criterion to determine whether a set is closed.

Corollary 255 *A set A is closed if and only if $x \in A$ whenever there exists a sequence $\{x_n\}_{n\geq 1} \subseteq A$ such that $x_n \to x$.*

Proof "Only if." Let A be closed. By Theorem 254, $x \in A$ if and only if there exists a sequence $\{x_n\}_{n\geq 1} \subseteq A$ such that $x_n \to x$.

"If." Let $x \in \overline{A}$. By Theorem 254 there exists $\{x_n\}_n \subseteq A$ such that $x_n \to x$. By hypothesis, this implies $x \in A$, that is, $\overline{A} \subseteq A$. Therefore, $\overline{A} = A$, which implies that A is closed. ■

Thanks to this corollary, to establish whether a set A is closed is sufficient to consider a generic sequence $\{x_n\}_{n\geq 1} \subseteq A$ such that $x_n \to x$. If we prove that also the limit point x belongs to A, we can then conclude that A is closed.

Example 256 Consider the subset $A = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_m, b_m]$ of R^m. Let $\{x^n\}_{n\geq 1} \subseteq A$ such that $x^n \xrightarrow{d_1} x$. By Proposition 246, $x_i^n \to x_i$. Since $x_i^n \in [a_i, b_i]$, the convergence $x_i^n \to x_i$ implies $x_i \in [a_i, b_i]$. Therefore, $x \in A$, and thanks to Corollary 255 we conclude that A is closed. ▲

Example 257 In the metric space $(C([0,1]), d_\infty)$ consider the set

$$A = \{f \in \mathcal{C}([0,1]) : -1 \leq f(t) \leq 1, \qquad \forall t \in [0,1]\}.$$

Consider a sequence $\{f_n\}_{n\geq 1} \subseteq A$ such that $f_n \xrightarrow{d_\infty} f$. Therefore, $\max_{t\in[0,1]} |f_n(t) - f(t)| \to 0$, which implies $|f_n(t) - f(t)| \to 0$ for each $t \in [0,1]$. Since for each $t \in [0,1]$ we have $f_n(t) \in [-1,1]$, it follows that $f(t) \in [-1,1]$. Therefore, $f \in C([0,1])$ and by Corollary 255 we conclude that A is closed. ▲

6.3.4 Completeness

Definition 258 *A sequence $\{x_n\}_{n\geq 1}$ of a metric space (X,d) satisfies the Cauchy criterion if, for each $\varepsilon > 0$, there exists $\overline{n} \geq 1$ such that $d(x_n, x_m) < \varepsilon$ for each $n, m \geq \overline{n}$.*

A sequence that satisfies Cauchy criterion is called a *Cauchy sequence*. When n increases, the points x_n of these sequences become therefore closer and closer among them.

Proposition 259 *Each convergent sequence in a metric space is a Cauchy sequence.*

Proof Let $x_n \to x$. Given $\varepsilon > 0$, there exists $\overline{n} \geq 1$ such that $d(x_n, x) < \varepsilon/2$ for each $n \geq \overline{n}$. Hence, for each $n, m \geq \overline{n}$, we have

$$d(x_n, x_m) \leq d(x_n, x) + d(x, x_m) \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

and the sequence $\{x_n\}_{n\geq 1}$ is Cauchy. ■

As this proof shows, if a sequence converges to a limit point, the points of the sequence become closer and closer to this point, and, consequently, they also become closer and closer among them. On the other hand, there are metric spaces in which there exist Cauchy sequences that are not convergent.

Example 260 Let $X = (0,1)$, endowed with the standard metric of the real line. The sequence $\{1/n\}_{n\geq 1}$ converges to the point 0, which does not belong to X. This sequence is therefore Cauchy in X, but it is not convergent in such space. ▲

A metric space where Cauchy sequences may not converge is as if it were lacking some points, it were "incomplete," with respect to the convergence. In fact, in such spaces there are sequences whose points behave among themselves as if they were converging to some limit point, but at the end they do not converge to any point of the space. These considerations lead us to the following definition.

Definition 261 *A metric space in which Cauchy sequences are convergent is called complete.*

The space $X = (0, 1)$ seen in the last example is not complete. The set of rational points $\mathbb{Q}$ endowed with the usual distance $|q' - q''|$ is not compete.

Theorem 262 *The space $\mathbb{R}^n$, endowed with any of the metrics d_1, d_2 and d_∞, is complete.*

Proof It is sufficient to consider the case $n = 1$. In fact, the general case follows from Proposition 246 and from Exercise 13.0.33. Let $\{x_n\}_{n\geq 1}$ be a sequence in $\mathbb{R}$ that satisfies Cauchy criterion. We want to prove that this sequence is convergent. We start by proving that it is bounded. Setting $\varepsilon = 1$, there exists $\overline{n} \geq 1$ such that $|x_n - x_m| < 1$ for each $n, m \geq \overline{n}$. Hence, for each $n \geq \overline{n}$ we have:

$$|x_n| = |x_n - x_{\overline{n}} + x_{\overline{n}}| \leq |x_n - x_{\overline{n}}| + |x_{\overline{n}}| < 1 + |x_{\overline{n}}|,$$

which implies that for the image E of the sequence $\{x_n\}_{n\geq 1}$ we have $E \subseteq (-1 - |x_{\overline{n}}|, 1 + |x_{\overline{n}}|$ The sequence is therefore bounded.

Let $\varepsilon > 0$. There exists $\overline{n} \geq 1$ such that $|x_n - x_m| < \varepsilon/2$ for each $n, m \geq \overline{n}$. Consider the subsequence $\{x_n\}_{n\geq\overline{n}}$. It is clearly bounded and therefore the sup exists. That is $\lambda = \sup_{n\geq\overline{n}} x_n$ with $\lambda \in \mathbb{R}$. In particular, we can pick an element x_m, with $m \geq \overline{n}$, for which $x_m > \lambda - \varepsilon/2$. Therefore,

$$d(x_n, \lambda) = |x_n - \lambda| \leq |x_n - x_m + x_m - \lambda| \leq |x_n - x_m| + |x_m - \lambda| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

holds for all $n \geq \overline{n}$

This implies $x_n \in B_\varepsilon(x)$ for each $n \geq \overline{n}$, and therefore $x_n \to \lambda$. ■

To know that a metric space is complete simplifies the study of the convergence of sequences. In fact, in complete metric spaces a sequence is convergent if and only if it satisfies Cauchy criterion, and this can checked by only considering the values x_n of the sequence, without any need to specify a limit point x to which the sequence can tend.

This is a key feature of this criterion because for many sequences it may not obvious, a priori, which may be a limit point. This makes difficult to check convergence through Definition 241 or Lemma 242.

We conclude with another important example of a complete metric space (we omit the proof), and with an example of non-complete metric space.

Theorem 263 *The space* $(\mathcal{C}([0,1]), d_\infty)$ *is complete.*

Example 264 The metric space $(C([0,1]), d_1)$ is not complete, that is, there exists a Cauchy sequence in this space that is not convergent (we omit the check). ▲

6.4 Compactness

Compact sets are a very important class of closed sets, crucial in the formulation of many classical results. In Calculus compact sets of $\mathbb{R}$ are defined as the closed and bounded sets. The most important example of such sets was given by the closed and bounded intervals $[a, b]$.

We now extend the notion of compactness to general metric spaces. To this end, we start by considering the compactness in $\mathbb{R}$ from a different perspective. Given a set A of a metric space, an *open cover* of A is any collection of open sets $\{G_i\}_{i\in I}$ such that $A \subseteq \bigcup_{i\in I} G_i$.

Proposition 265 *Each open cover of a closed and bounded interval* $[a, b]$ *of* $\mathbb{R}$ *has a finite subcover.*

In other words, each open cover $\{G_i\}_{i\in I}$ of a closed and bounded interval $[a, b]$ of $\mathbb{R}$ has a finite subcover $\{G_i\}_{i=1}^n \subseteq \{G_i\}_{i\in I}$ such that $[a, b] \subseteq \bigcup_{i=1}^n G_i$. The proof is based on the next lemma.

Lemma 266 *Let* $\{[a_n, b_n]\}_{n\geq 1}$ *be a collection of closed and bounded intervals of* $\mathbb{R}$ *with* $[a_{n+1}, b_{n+1}] \subseteq [a_n, b_n]$ *for each* $n \geq 1$. *We have* $\bigcap_{n\geq 1} [a_n, b_n] \neq \emptyset$.

Proof Given the collection $\{[a_n, b_n]\}_{n\geq 1}$, let $A = \{a_1, a_2, ..., a_n, ...\}$. We have $A \subseteq [a_1, b_1]$ and therefore A is a bounded set. By the completeness of $\mathbb{R}$, we can set $x = \sup A$. Since each b_n is an upper bound for A, we have $b_n \geq x$ for each $n \geq 1$. Hence, $a_n \leq x \leq b_n$ for each $n \geq 1$, and therefore $x \in \bigcap_{n\geq 1} [a_n, b_n]$. This implies $\bigcap_{n\geq 1} [a_n, b_n] \neq \emptyset$, as desired. ■

Proof of Proposition 265. Suppose *per contra* that there exists an open cover $\{G_i\}_{i\in I}$ of $[a, b]$ that does not contain any finite subcover of $[a, b]$. Let $\delta = b - a$ and $c_1 = (a + b)/2$. The collection $\{G_i\}_{i\in I}$ is an open cover also of the intervals $[a, c_1]$ and $[c_1, b]$. Therefore, at least one of these two intervals has no finite subcover of $\{G_i\}_{i\in I}$. Otherwise, from $[a, b] = [a, c_1] \cup [c_1, b]$ it would follow that $[a, b]$ itself would have

such a subcover. Without loss of generality, suppose therefore that $[a, c_1]$ has no finite subcover of $\{G_i\}_{i\in I}$. Set $c_2 = (a + c_1)/2$. By repeating the argument just seen, we can assume that also $[a, c_2]$ does not have not such a finite subcover. By proceeding in this way we can construct a collection of intervals $\{[a, c_n]\}_{n\geq 1}$ such that $[a, c_{n+1}] \subseteq [a, c_n]$ and $c_n - a = \delta/2^n$ for each $n \geq 1$. Moreover, none of these closed intervals has a finite subcover of $\{G_i\}_{i\in I}$.

By Lemma 242, $\bigcap_{n\geq 1} [a, c_n] \neq \emptyset$. Let $x \in \bigcap_{n\geq 1} [a, c_n]$. Since $[a, b] \subseteq \bigcup_{i=1}^{n} G_i$, there exists G_i such that $x \in G_i$. As x is an interior point of G_i, there exists a neighborhood $(x - \varepsilon, x + \varepsilon)$ such that $(x - \varepsilon, x + \varepsilon) \subseteq G_i$. For n sufficiently large, we have $\delta/2^n < \varepsilon$, and therefore

$$[a, c_n] \subseteq \left(x - \frac{\delta}{2^n}, x + \frac{\delta}{2^n}\right) \subseteq (x - \varepsilon, x + \varepsilon) \subseteq G_i.$$

Consequently, the singleton $\{G_i\}$ is a finite subcover of $\{G_i\}_{i\in I}$ that covers $[a, c_n]$, which contradicts the fact that all the intervals $[a, c_n]$ do not have such subcovers. From this contradiction it follows that $[a, b]$ has a finite subcover of $\{G_i\}_{i\in I}$. ■

Proposition 265 motivates the next definition.

Definition 267 *A subset A of a metric space is compact if each open cover of A has a finite subcover.*

Example 268 By Proposition 265, the closed and bounded intervals $[a, b]$ are compact sets of $\mathbb{R}$. ▲

Example 269 In any metric space, the finite sets are compact. In fact, let $A = \{x_i\}_{i=1}^{n}$ be a finite set and let $\{G_i\}_{i\in I}$ be an open cover of A. For each point $x_i \in A$ there exists an open set G_i in this cover such that $x_i \in G_i$. Therefore, $\{G_i\}_{i=1}^{n}$ is a finite subcover of A.

Example 270 Let $\{x_n\}_{n\geq 1}$ be a convergent sequence in a metric space X, with image E and limit point x. Then, the set $E \cup \{x\}$ is compact. In fact, let $\{G_i\}_{i\in I}$ be an open cover of $E \cup \{x\}$. Let G_x be an open set in such cover that contains the limit point x. There exists $\overline{n} \geq 1$ such that $x_n \in G_x$ for each $n \geq \overline{n}$. Therefore, the set $E \cap G_x^c$ is at most finite. Since $E \cap G_x^c \subseteq \bigcup_{i\in I} G_i$, for each $y \in E \cap G_x^c$ there exists an open set G_y of the cover $\{G_i\}_{i\in I}$ such that $y \in G_y$. It follows that

$$E = (E \cap G_x^c) \cup (E \cap G_x) \subseteq \left(\bigcup_{y\in E\cap G_x^c} G_y\right) \cup G_x,$$

and therefore $\{G_y\}_{y\in E\cap G_x^c} \cup \{G_x\}$ is a finite subcover of $E \cup \{x\}$. ▲

We start by proving that compact sets are closed and bounded sets.

Theorem 271 *A compact subset of a metric space X is closed and bounded.*

Proof Let K be a compact set of a metric space X. We prove that K^c is open. Let $x \in K^c$. For each $y \in K$, let G_y and V_y be, respectively, neighborhoods of y and x with radius lower than $d(x,y)/2$. This implies that $G_y \cap V_y = \emptyset$ for each $y \in K$. Since the collection $\{G_y\}_{y\in K}$ is an open cover of K, there exists a finite subcover $\{G_{y_i}\}_{i=1}^n$ of K. Setting $V = \bigcap_{i=1}^n V_{y_i}$ and $G = \bigcup_{i=1}^n G_i$, we have $K \subseteq G$ and $V \cap G = \emptyset$. Hence, V is a neighborhood of x such that $V \subseteq K^c$, which implies that x is an interior point of K^c. Thus, K^c is open.

It remains to show that K is bounded. Given $\varepsilon > 0$, for each $x \in K$ let $B_\varepsilon(x)$ be a neighborhood of radius ε. Since the collection $\{B_\varepsilon(x)\}_{x\in K}$ is an open cover of K, there exists a finite set $E \subseteq K$ such that $\{B_\varepsilon(x)\}_{x\in E}$ is a subcover of K, that is, $K \subseteq \bigcup_{x\in E} B_\varepsilon(x)$. Set $M = \max_{x',x''\in E} d(x',x'')$. Let $y', y'' \in K$. There exist $x', x'' \in E$ such that $y' \in B_\varepsilon(x')$ and $y'' \in B_\varepsilon(x'')$. Therefore,

$$d(y',y'') \leq d(y',x') + d(x',x'') + d(x'',y'') < 2\varepsilon + M.$$

It follows that, taking any $y \in K$, we have $K \subseteq B_{2\varepsilon+M}(y)$, which implies that K is a bounded set. ■

The next result, often called the Heine-Borel Theorem, generalizes Proposition 265 and shows that the converse of Theorem 271 holds in Euclidean spaces. Therefore, Euclidean spaces the new definition of compactness reduces to the one seen in Calculus.

Theorem 272 (Heine-Borel) *Let $X = \mathbb{R}^n$, endowed with any of the metrics d_1, d_2 and d_∞. A subset of $\mathbb{R}^n$ is compact if and only if it is closed and bounded.*

Proof In view of Proposition 271, we only have to prove that a closed and bounded subset of $\mathbb{R}^n$ is compact. The special case $X = \mathbb{R}$ is given by Proposition 265. We leave to the reader the proof of the general case $X = \mathbb{R}^n$. ■

At this point it is important to see an example of a space where there is a closed and bounded set that is not compact, thus showing that Theorem 272 is in general false in non-Euclidean spaces.

Example 273 Let X be the space of rational numbers $\mathbb{Q}$, with the usual metric $d(q_1,q_2) = |q_1 - q_2|$ for each $q_1, q_2 \in \mathbb{Q}$. Consider the set

$$A = \left\{ q \in \mathbb{Q} : \sqrt{2} < q < \sqrt{3} \right\}.$$

Clearly, A is bounded. It is also closed. In fact, let $\{q_n\}_{n\geq 1} \subseteq A$ be such that $q_n \to q \in \mathbb{Q}$. Since for each $n \geq 1$ we have $\sqrt{2} < q_n < \sqrt{3}$, it follows that $\sqrt{2} \leq q \leq \sqrt{3}$. Being q a rational number, this implies $\sqrt{2} < q < \sqrt{3}$, that is $q \in A$. By Corollary 255, we conclude that A is closed.

We leave to the reader to check that A is not compact (i.e., there exists at least an open cover of A that does not have any finite subcover). We close by showing that A is also open. To this end, take $q \in A$. By the density of $\mathbb{Q}$ in $\mathbb{R}$,[8] there exist $q_1, q_2 \in \mathbb{Q}$ such that $\sqrt{2} < q_1 < q < q_2 < \sqrt{3}$. Since (q_1, q_2) is an open neighborhood of q entirely contained in A, the point q is an interior point of A, which is therefore open. In conclusion, A is an example of closed and bounded set that is not compact. It is also an example of a set that is both closed and open. ▲

The next result shows that compactness is inherited by closed subsets of compact sets.

Proposition 274 *In a metric space the closed subsets of a compact set are themselves compact.*

Proof Let F be a closed subset of a compact set K. Let $\{G_i\}_{i\in I}$ be an open cover of F. We want to prove that there exists a finite subcover of F. Since F is closed, the complement F^c is open. Therefore, the collection $\{G_i\}_{i\in I} \cup \{F^c\}$ is an open cover of K. Therefore, there exists a finite subcover $\{G_i\}_{i=1}^n \cup \{F^c\}$ of K, i.e., $K \subseteq F^c \cup \bigcup_{i=1}^n G_i$. This implies $F \subseteq \bigcup_{i=1}^n G_i$, and therefore $\{G_i\}_{i=1}^n$ is a finite subcover of F. ■

We now give a fundamental characterization of compact sets, based sequences and accumulation points. Condition (ii) is often called the Bolzano-Weierstrass property of compact sets.

Theorem 275 *For a subset A of a metric space, the following properties are equivalent:*

(i) A is compact,

(ii) each infinite subset of A has at least one accumulation point,

(iii) each sequence of points of A has at least a convergent subsequence.

[8] See for instance Theorem 1 p. 27 of Ambrosetti and Musu (1988).

Proof (i) implies (ii). Let A be a compact set and let E be an infinite subset of A. If E does not have accumulation points, then all its points are isolated. Therefore, each $x \in E$ has a neighborhood G_x such that $G_x \cap E = \{x\}$. Consequently, $\{G_x\}_{x\in E}$ is an open cover of E that does not contain any finite subcover. For $x \in A - E$, let V_x be a generic neighborhood of x. The collection $\{G_x\}_{x\in E} \cup \{V_x\}_{x\in A-E}$ is therefore an open cover of the compact set A that does not have a finite subcover. This contradiction proves that E has at least one accumulation point.

(ii) implies (iii). Let $\{x_n\}_{n\geq 1}$ be a sequence contained in A. Let E be its image. If E is finite, there exist at least one $x \in E$ and a sequence of natural numbers $n_1 < n_2 < \cdots < n_k < \cdots$ such that $x_{n_k} = x$ for each $k \geq 1$. The constant subsequence $\{x_{n_k}\}_{k\geq 1}$ is obviously convergent. Let now E be an infinite set. There exists therefore an accumulation point x of E. Fixing $k \geq 1$, let $x_{n_k} \in B_{\frac{1}{k}}(x) \cap E$. The subsequence $\{x_{n_k}\}_{k\geq 1}$ constructed in this way converges to x.

(iii) implies (i). We omit the proof of this last point. ■

The (iii) is a fundamental property of compact sets and guarantees that from any sequence, however irregular it may be, of a compact set it is always possible to extract at least a convergent subsequence.

By Theorem 275, this property characterizes compact sets and therefore it is not in general true for sets that are only closed and bounded. For instance, in Example 273 consider a sequence $\{q_n\}_n \subseteq A$ of rational numbers with $q_n \to \sqrt{2}$. It is easy to see that this sequence does not have any convergent subsequence.

A metric space X is called *compact* if X itself is a compact set. In other words, if each collection $\{G_i\}_{i\in I}$ of open sets such that $X = \bigcup_{i\in I} G_i$ has a subcollection $\{G_i\}_{i=1}^n$ such that $X = \bigcup_{i=1}^n G_i$.

Example 276 Each finite metric space is compact. The space $X = [a, b]$, with $a, b \in R$, is compact. ▲

By Proposition 274, closed subsets of a compact space are themselves compact. Moreover, thanks to Theorem 275, we have the following characterization of compact metric spaces.

Corollary 277 *A metric space is compact if and only if every sequence has at least a convergent subsequence.*

Metric spaces have the following fundamental property.

Theorem 278 *A compact metric space is complete.*

To prove this result we prove couple of useful properties. The first one generalizes Lemma 266.

Lemma 279 *Let $\{K_i\}_{i\in I}$ be a collection of compact sets such that, for each finite subcollection $\{K_i\}_{i\in J}\subseteq\{K_i\}_{i\in I}$, we have $\bigcap_{i\in J}K_i\neq\emptyset$. Then, $\bigcap_{i\in I}K_i\neq\emptyset$.*

Notice that the set I has any cardinality. In Lemma 266 we have $I=\{1,...,n,...\}$ and $K_n=[a_n,b_n]$. Since $[a_{n+1},b_{n+1}]\subseteq[a_n,b_n]$, we have $K_{n+1}\subseteq K_n$ and therefore for each finite subcollection $\{K_n\}_{n\in J}$ we obviously have $\bigcap_{i\in J}K_i\neq\emptyset$. Lemma 279 implies $\bigcap_{n\geq1}K_n\neq\emptyset$, which was exactly what stated in Lemma 266.

Proof Take K_1 and set $I_1=\{i\in I:i\neq1\}$. Suppose that $K_1\cap\bigcap_{i\in I_1}K_i=\emptyset$. Then, $K_1\subseteq\bigcup_{i\in I_1}K_i^c$ and therefore $\{K_i^c\}_{i\in I_1}$ is an open cover of the compact set K_1. Therefore, there exists a finite subcover $\left\{K_{i_j}^c\right\}_{j=1}^n$ of K_1. This implies that for the finite subcollection $\left\{K_{i_j}\right\}_{j=1}^n\cup\{K_1\}$ we have $K_1\cap K_{j_1}\cap\cdots\cap K_{j_n}=\emptyset$, which contradicts the hypothesis. Therefore, $\bigcap_{i\in I}K_i=K_1\cap\bigcap_{i\in I_1}K_i\neq\emptyset$, as desired. ■

The *diameter* of a set E of a metric space, denoted by $diam\,(E)$, is defined as:

$$diam\,(E)=\sup\{d\,(x,y):x,y\in E\}.$$

Lemma 280 *We have $diam\,(E)=diam\left(\overline{E}\right)$.*

Proof Since $E\subseteq\overline{E}$, clearly $diam\,(E)\leq diam\left(\overline{E}\right)$. We prove that it also holds $diam\,(E)\geq diam\left(\overline{E}\right)$. Let $\varepsilon>0$ and let $x,y\in\overline{E}$. By Theorem 235, $\overline{E}=E\cup E'$. Therefore, there exist $x',y'\in E$ such that $d\,(x,x')<\varepsilon$ and $d\,(y,y')<\varepsilon$. It follows that:

$$d\,(x,y)\leq d\,(x,x')+d\,(x',y')+d\,(y',y)\leq\varepsilon+d\,(x',y')+\varepsilon\leq diam\,(E)+2\varepsilon,$$

which implies:

$$diam\left(\overline{E}\right)=\sup\left\{d\,(x,y):x,y\in\overline{E}\right\}\leq diam\,(E)+2\varepsilon.$$

Since ε is arbitrary, we conclude that $diam\left(\overline{E}\right)\leq diam\,(E)$, as desired. ■

Proof of Theorem 278. Let $\{x_n\}_{n\geq1}$ be a Cauchy sequence of points of X. We want to prove that it is convergent. For simplicity, assume that all the terms of the sequence are distinct. For each $n\geq1$, set

$$E_n=\{x_n,x_{n+1},...\}.$$

Since $E_{n+1} \subseteq E_n$, we have $\overline{E}_{n+1} \subseteq \overline{E}_n$. Moreover, being X compact, also the sets $\overline{E}_n$ are compact. By Lemma 279, $\bigcap_{n\geq 1} \overline{E}_n \neq \emptyset$. Let $x \in \bigcap_{n\geq 1} \overline{E}_n$. We prove that $x_n \to x$. Let $\varepsilon > 0$ and let $\varepsilon' \in (0, \varepsilon)$. Since the sequence $\{x_n\}_{n\geq 1}$ is Cauchy, there exists $n_{\varepsilon'} \geq 1$ such that $d(x_n, x_m) < \varepsilon'$ for each $n, m \geq n_{\varepsilon'}$.[9] Therefore, $diam\left(E_{n_{\varepsilon'}}\right) \leq \varepsilon'$. As $x \in \overline{E}_{n_{\varepsilon'}}$, by Lemma 280 we have:

$$d(x, x_n) \leq diam\left(\overline{E}_{n_{\varepsilon'}}\right) = diam\left(E_{n_{\varepsilon'}}\right) \leq \varepsilon' < \varepsilon,$$

for each $n \geq n_{\varepsilon'}$. Therefore, $x_n \in B_\varepsilon(x)$ for each $n \geq n_{\varepsilon'}$, and so $x_n \to x$, as desired. ■

The converse of Theorem 278 is clearly false. The real line $\mathbb{R}$ is a simple example of a complete metric space that is not compact.

There exists, however, a condition that combined with completeness makes a metric space compact. A metric space X is called *totally bounded* if, for each $\varepsilon > 0$, there exists a collection of points $\{x_i\}_{i=1}^n$ of X such that $X = \bigcup_{i=1}^n B_\varepsilon(x_i)$. In other words, for each $\varepsilon > 0$ we can find a finite "net" of points from which each point of the space has distance lower than ε. Naturally, what gives bite to the property is that such net is finite. For example, the real line does not have this property. This is not by chance: next result completes Theorem 278 by showing that total boundedness is exactly the property that, combined with completeness, leads to compactness (we omit the proof).

Theorem 281 *A metric space is compact if and only if it is complete and totally bounded.*

6.5 Limits and Continuity of Functions

6.5.1 Limits

We now move to the study of functions defined on metric spaces. We start with the notion of limit of a function. In Calculus, given a function $F : A \subseteq \mathbb{R} \to \mathbb{R}$ and given an accumulation point x_0 of A, we write $\lim_{x\to x_0} F(x) = L \in \mathbb{R}$ if for each $\varepsilon > 0$ there exists $\delta_\varepsilon > 0$ such that, for each $x \in A$ with $0 < |x - x_0| < \delta_\varepsilon$, we have $|F(x) - L| < \varepsilon$. The point x_0 is not required to belong to A, but it is enough that it is an accumulation point of A.

This notion can be naturally extended to general metric spaces.

[9] To underline its dependence on ε', here we denote by $n_{\varepsilon'}$ the number $\overline{n}$ of Definition 258.

Definition 282 *Let (X, d_X) and (Y, d_Y) be two metric spaces. Given a function $f : A \subseteq X \to Y$ and an accumulation point x_0 of A, we write $\lim_{x\to x_0} f(x) = y \in Y$ if, for each $\varepsilon > 0$, there exists $\delta_\varepsilon > 0$ such that*

$$d_Y(f(x), y) < \varepsilon$$

for all $x \in A$ with $0 < d_X(x, x_0) < \delta_\varepsilon$.

The special case seen in Calculus corresponds to $(X, d_X) = (\mathbb{R}^n, d_2)$ and $(Y, d_Y) = (\mathbb{R}, d)$, where d is the standard metric of the real line. Also here the point x_0 is only required to be an accumulation point of A.

The limit value $\lim_{x\to x_0} F(x)$ is, therefore, the value to which the function tends as x becomes closer and closer to x_0. Note that we do not require any relation between such limit value and the value $f(x_0)$ that the function actually takes at the point x_0. Indeed, the point x_0 may even not belong to the domain A of the function (in which case $f(x_0)$ does not exist), but be only an accumulation point of A. A standard example is the function $f : \mathbb{R} - \{0\} \to \mathbb{R}$ given by $f(x) = (\sin x)/x$ for each $x \in \mathbb{R} - \{0\}$. In this case $A = \mathbb{R} - \{0\}$, and the point $x_0 = 0$ is of accumulation of A, but it does not belong to A. As well known, we have $\lim_{x\to 0} f(x) = 1$.

Before showing some examples, we characterize limits through sequences.

Proposition 283 *Let (X, d_X) and (Y, d_Y) be two metric spaces. Given a function $f : A \subseteq X \to Y$ and an accumulation point x_0 of A, we have $\lim_{x\to x_0} f(x) = y \in Y$ if and only if $f(x_n) \to y$ for each sequence $\{x_n\}_{n\geq 1}$ of points of A, with $x_n \neq x_0$ for each n, such that $x_n \to x_0$.*

To be precise, we should have written $x_n \overset{d_X}{\to} x_0$ and $F(x_n) \overset{d_Y}{\to} y$, but for simplicity we only write $x_n \to x_0$ and $F(x_n) \to y$.

Proof "If": suppose $F(x_n) \to y$ for each sequence $\{x_n\}_{n\geq 1}$ of points of A, with $x_n \neq x_0$ for each n, such that $x_n \to x_0$. Suppose it is false that $\lim_{x\to x_0} F(x) = y$. Then there exists $\varepsilon > 0$ such that for each $\delta > 0$ there exists $x_\delta \in A$ such that $0 < d_X(x_\delta, x_0) < \delta$ and $d_Y(F(x_\delta), y) \geq \varepsilon$. For each n, set $\delta = 1/n$ and let x_n be the corresponding point of A, just denoted by x_δ. For the sequence $\{x_n\}_{n\geq 1}$of points of A constructed in this way we have $d_X(x_0, x_n) < 1/n$ for each n, and therefore $\lim_{n\to\infty} d_X(x_0, x_n) = 0$. By Lemma 242, $x_n \to x$. But, by construction we have $d_Y(F(x_n), y) \geq \varepsilon$ for each n, and therefore the sequence $F(x_n)$ does not converge to y. This contradicts the hypothesis, and we conclude that $\lim_{x\to x_0} F(x) = y$.

"Only if": suppose $\lim_{x\to x_0} F(x) = y \in Y$. Let $\{x_n\}_{n\geq 1}$ be a sequence of points of A, with $x_n \neq x_0$ for each n, such that $x_n \to x_0$. Let $\varepsilon > 0$. There exists $\delta_\varepsilon > 0$

such that for each $x \in A$ with $0 < d_X(x, x_0) < \delta_\varepsilon$ we have $d_Y(F(x), y) < \varepsilon$. Since $x_n \to x_0$ and $x_n \neq x_0$, there exists $\overline{n} \geq 1$ such that $0 < d_X(x_n, x_0) < \delta_\varepsilon$ for each $n \geq \overline{n}$. Therefore, for each $n \geq \overline{n}$ we have $d_Y(F(x_n), y) < \varepsilon$, which implies $F(x_n) \to y$. ■

Note that, in view of Proposition 283, Theorem 247 also ensures the uniqueness of the limit value y in Definition 282. In fact, each sequence $\{F(x_n)\}_{n\geq 1}$ converges to a unique limit.

Example 284 Let $(X, d_X) = (\mathbb{R}, d)$ and $(Y, d_Y) = (\mathbb{R}^2, d_2)$, and let $F : \mathbb{R} \to \mathbb{R}^2$ be defined by $F(x) = (\sin(x), \cos(x))$ for each $x \in \mathbb{R}$. Using Proposition 283 it is easy to verify that $\lim_{x\to 0} F(x) = (0, 1) \in \mathbb{R}^2$. ▲

Example 285 Let $(X, d_X) = (Y, d_Y) = (\mathbb{R}^2, d_2)$, and let $F = (F_1, F_2) : \mathbb{R}^2 \to \mathbb{R}^2$ be defined by

$$F_1(x) = \begin{cases} x_1 + x_2 + 1 & x \neq (0,0) \\ 0 & x = (0,0) \end{cases} \quad \text{and} \quad F_2(x) = 1 + x_1^2 x_2, \qquad \forall x \in \mathbb{R}^2.$$

For example, for $x = (2, 1)$ we have $F(x) = (F_1(x), F_2(x)) = (4, 5)$, while for $x = (0, 0)$ we have$F(x) = (F_1(x), F_2(x)) = (0, 1)$. Using Proposition 283, it is easy to verify that

$$\lim_{x\to(2,1)} F(x) = (4, 5) \quad \text{and} \quad \lim_{x\to(0,0)} F(x) = (1, 1).$$

Note that $\lim_{x\to(2,1)} F(x) = F(2, 1)$, while $\lim_{x\to(0,0)} F(x) \neq F(0, 0)$. ▲

Example 286 Let $(X, d_X) = (C([0, 1]), d_\infty)$ and $(Y, d_Y) = (\mathbb{R}, d)$, and let $F : C([0, 1]) \to \mathbb{R}$ be defined by $F(f) = \int_0^1 f(t)\, dt$ for each $f \in C([0, 1])$. Let $\mathbf{0} \in C([0, 1])$ be the identically null function. We have $\lim_{f\to\mathbf{0}} F(f) = 0$, that is, $\lim_{f\to\mathbf{0}} F(f) = F(\mathbf{0})$. In fact, let $\{f_n\}_{n\geq 1}$ be a sequence in $C([0, 1])$, with $f_n \neq 0$, such that $f_n \to \mathbf{0}$. By Proposition 283, in order to prove that $\lim_{f\to\mathbf{0}} F(f) = 0$ we have to prove that $F(f_n) \to 0$. We have:

$$d_\infty(f_n, \mathbf{0}) = \max_{t\in[0,1]} |f_n(t) - \mathbf{0}(t)| = \max_{t\in[0,1]} |f_n(t)| \to 0. \tag{6.13}$$

Let $g_n \in C([0, 1])$ be the constant function such that $g_n(t) = \max_{t\in[0,1]} |f_n(t)|$ for each $t \in [0, 1]$. Expression (6.13) implies:

$$\begin{aligned} d(F(f_n), 0) &= |F(f_n) - 0| = \left| \int_0^1 f_n(t)\, dt \right| \leq \int_0^1 |f_n(t)|\, dt \\ &\leq \int_0^1 g_n(t)\, dt = \max_{t\in[0,1]} |f_n(t)| \to 0. \end{aligned}$$

Therefore, by Lemma 242 we have $F(f_n) \to 0$, as desired. ▲

Example 287 Let $(X, d_X) = (Y, d_Y) = (C([0,1]), d_\infty)$, and let $F : C([0,1]) \to C([0,1])$ be defined by $F(f) = f^2$ for each $f \in C([0,1])$. For example, if $f(t) = \sin t$, we have $F(f)(t) = \sin^2 t$. We prove that $\lim_{f \to \mathbf{0}} F(f) = 0$, that is, $\lim_{f \to \mathbf{0}} F(f) = F(\mathbf{0})$. Let $\{f_n\}_{n \geq 1}$ be a sequence in $C([0,1])$, with $f_n \neq 0$, such that $f_n \to 0$. By Proposition 283, in order to prove that $\lim_{f \to \mathbf{0}} F(f) = 0$ we have to prove that $F(f_n) \to 0$. We have:

$$d_\infty(f_n, \mathbf{0}) = \max_{t \in [0,1]} |f_n(t) - \mathbf{0}(t)| = \max_{t \in [0,1]} |f_n(t)| \to 0,$$

which implies:

$$\begin{aligned} d_\infty(F(f_n), \mathbf{0}) &= \max_{t \in [0,1]} |F(f_n)(t) - \mathbf{0}(t)| = \max_{t \in [0,1]} |F(f_n)(t)| = \max_{t \in [0,1]} \left| f_n^2(t) \right| \\ &= \max_{t \in [0,1]} (|f_n(t)| |f_n(t)|) \leq \left(\max_{t \in [0,1]} |f_n(t)| \right) \left(\max_{t \in [0,1]} |f_n(t)| \right) \to 0. \end{aligned}$$

By Lemma 242, we have $F(f_n) \to \mathbf{0}$, as desired. ▲

For real valued functions, we have the following algebraic properties, whose simple proof we omit.

Proposition 288 *Let $f : A \subseteq X \to \mathbb{R}$ and $g : A \subseteq X \to \mathbb{R}$ be real valued functions defined on a subset A of a metric space X. Given $x_0 \in A'$, we have:*

(i) $\lim_{x \to x_0} (\alpha f + \beta g)(x) = \alpha \lim_{x \to x_0} f(x) + \beta \lim_{x \to x_0} g(x)$ *for each* $\alpha, \beta \in \mathbb{R}$;

(ii) $\lim_{x \to x_0} (fg)(x) = \lim_{x \to x_0} f(x) \lim_{x \to x_0} g(x)$;

(iii) $\lim_{x \to x_0} \left(\frac{f}{g} \right)(x) = \frac{\lim_{x \to x_0} f(x)}{\lim_{x \to x_0} g(x)}$ *when* $\lim_{x \to x_0} g(x) \neq 0$.

6.5.2 Continuity

Definition 289 *Let (X, d_X) and (Y, d_Y) be two metric spaces. A function $f : A \subseteq X \to Y$ is said to be continuous at the point $x_0 \in A$ if, for each $\varepsilon > 0$, there exists $\delta_\varepsilon > 0$ such that*

$$d_Y(f(x), f(x_0)) < \varepsilon$$

for each $x \in A$ with $d_X(x, x_0) < \delta_\varepsilon$.

In other words, f is continuous at $x_0 \in A$ if, for each neighborhood $B_\varepsilon(f(x_0))$ of $f(x_0)$, there exists a neighborhood $B_{\delta_\varepsilon}(x_0)$ of x_0 such that $f(x) \in B_\varepsilon(f(x_0))$ for each $x \in B_{\delta_\varepsilon}(x_0)$. Equivalently, for each open set V containing $f(x_0)$, there exists an open set G containing x_0 such that $f(x) \in V$ for each $x \in G$.

Observe that the point x_0 must belong to the domain of the function. A function continuous at each point of a set $E \subseteq A$ is called *continuous* on E. The function is called *continuous* if it is continuous at all the points of its domain.

Lemma 290 *Let (X, d_X) and (Y, d_Y) be two metric spaces. A function $f : A \subseteq X \to Y$ is continuous at each isolated point of A.*

Proof Let x_0 be an isolated point of A. There exists a neighborhood $B_\delta(x_0)$ of radius $\delta > 0$ such that $B_\delta(x_0) \cap A = \emptyset$. Therefore, we have $x \in A$ and $d_X(x, x_0) < \delta$ if and only if $x = x_0$. It follows that, for each $\varepsilon > 0$, there exists $\delta > 0$ such that $x \in A$ and $d_X(x, x_0) < \delta$ implies $x = x_0$. Therefore, $|F(x) - F(x_0)| = 0 < \varepsilon$. ■

All functions are thus always continuous at the isolated points of their domains. Such points are therefore of no interest for the notion of continuity. For accumulation points we have instead the following characterization of continuity, based on the notion of limit. Recall that $A \cap A'$ is the set of the points of A that are accumulation points of A.

Proposition 291 *Let (X, d_X) and (Y, d_Y) be two metric spaces. A function $f : A \subseteq X \to Y$ is continuous at a point $x_0 \in A \cap A'$ if and only if we have $\lim_{x \to x_0} f(x) = f(x_0)$.*

This results follows immediately from Definitions 282 and 289. The notion of continuity seen in Calculus, usually presented as $\lim_{x \to x_0} f(x) = f(x_0)$, can therefore be recovered by setting $(X, d_X) = (\mathbb{R}^n, d_2)$ and $(Y, d_Y) = (\mathbb{R}, d)$.

Thanks to the characterization given by Proposition 291, we can fully understand the meaning of continuity: a function is continuous at an accumulation point $x_0 \in A \cap A'$ when the value that the function takes at x_0, that is $f(x_0)$, is consistent with the value to which the function tends when x gets closer and closer to x_0, that is, $\lim_{x \to x_0} f(x)$.

Proposition 283 leads us to a fundamental criterion to establish the continuity of a function.

Corollary 292 *Let (X, d_X) and (Y, d_Y) be two metric spaces. A function $f : A \subseteq X \to Y$ is continuous at a point $x_0 \in A$ if and only if we have $f(x_n) \to f(x_0)$ for each sequence $\{x_n\}_{n \geq 1}$ of points of A such that $x_n \to x_0$.*

Proof The result follows immediately by Proposition 291, once we observe that when x_0 is an isolated point of A we can consider the constant sequence $\{x_0, x_0, ...\}$. ■

Example 293 The function of Example 284 is continuous at each point of its domain. The function of Example 285 is not continuous at the point $(0, 0)$, but it is continuous at all the other points of its domain. ▲

In the next three examples we consider the metric space $(C([0, 1]), d_\infty)$.

Example 294 The function of Example 286 is continuous at each point of its domain. To see this, given a function $f \in C([0,1])$, let $\{f_n\}_{n\geq 1}$ be a sequence in $C([0,1])$, with $f_n \neq f$, such that $f_n \to f$. By Corollary 292, to prove that F is continuous in f we have to prove that $F(f_n) \to F(f)$. Since $f_n \to f$, we have:

$$d_\infty(f_n, f) = \max_{t\in[0,1]} |f_n(t) - f(t)| \to 0. \tag{6.14}$$

Let $g_n \in C([0,1])$ be the constant function such that $g_n(t) = \max_{t\in[0,1]} |f_n(t) - f(t)|$ for each $t \in [0,1]$. Expression (6.14) implies:

$$\begin{aligned} d(F(f_n), F(f)) &= |F(f_n) - F(f)| = \left| \int_0^1 f_n(t)\,dt - \int_0^1 f(t)\,dt \right| = \left| \int_0^1 (f_n(t) - f(t))\,d \\ &\leq \int_0^1 |f_n(t) - f(t)|\,dt \leq \int_0^1 g_n(t)\,dt = \max_{t\in[0,1]} |f_n(t) - f(t)| \to 0. \end{aligned}$$

Therefore, by Lemma 242 we have $F(f_n) \to F(f)$, as desired. ▲

Example 295 The function of Example 287 is continuous at each point of its domain. In fact, given $f \in C([0,1])$, let $\{f_n\}_{n\geq 1}$ be a sequence in $C([0,1])$ such that $f_n \to f$. By Corollary 292, to prove that F is continuous in f we have to prove that $F(f_n) \to F(f)$. We have:

$$d_\infty(f_n, f) = \max_{t\in[0,1]} |f_n(t) - f(t)| \to 0, \tag{6.15}$$

which implies:

$$\begin{aligned} d_\infty(F(f_n), F(f)) &= \max_{t\in[0,1]} |F(f_n)(t) - F(f)(t)| = \max_{t\in[0,1]} \left| f_n^2(t) - f^2(t) \right| = \max_{t\in[0,1]} |(f_n(t) - \\ &\leq \max_{t\in[0,1]} (|f_n(t) - f(t)|\,|f_n(t) + f(t)|) \leq \max_{t\in[0,1]} |f_n(t) - f(t)| \max_{t\in[0,1]} |f_n(\end{aligned}$$

Let $m = \min_{t\in[0,1]} f(t)$ and $M = \max_{t\in[0,1]} f(t)$. By the Weierstrass Theorem, m and M are well defined. Without loss of generality, assume that $|m| \leq |M|$, so that $-|M| \leq m \leq |M|$.[10] Let $\varepsilon > 0$. By (6.15), there exists $\overline{n} \geq 1$ such that

$$\max_{t\in[0,1]} |f_n(t) - f(t)| < \varepsilon, \quad \forall n \geq \overline{n}. \tag{6.16}$$

Therefore, for each $n \geq \overline{n}$ we have $f(t) - \varepsilon < f_n(t) < f(t) + \varepsilon$, and consequently

$$\begin{aligned} f_n(t) + f(t) &< f(t) + \varepsilon + f(t) \leq 2|M| + \varepsilon, \quad \forall t \in [0,1], \\ f_n(t) + f(t) &> f(t) - \varepsilon + f(t) \geq 2m - \varepsilon \geq -2|M| - \varepsilon, \quad \forall t \in [0,1] \end{aligned}$$

[10] Remember that, given $x \in \mathbb{R}$ and $c > 0$, we have $|x| < c$ if and only if $-c < x < c$ (see Ambrosetti and Musu, 1998, p. 32).

that is

$$|f_n(t) + f(t)| < 2|M| + \varepsilon, \quad \forall t \in [0,1].$$

Together with (6.16), this implies that

$$\max_{t\in[0,1]} |f_n(t) + f(t)| < 2|M| + \varepsilon \quad \text{and} \quad \max_{t\in[0,1]} |f_n(t) - f(t)| < \varepsilon, \qquad \forall n \geq \overline{n}.$$

It follows that, for each $n \geq \overline{n}$,

$$d_\infty(F(f_n), F(f)) \leq \max_{t\in[0,1]} |f_n(t) - f(t)| \max_{t\in[0,1]} |f_n(t) + f(t)| \leq \varepsilon(2|M| + \varepsilon) = 2\varepsilon|M| + \varepsilon^2.$$

Therefore, $\lim_{n\to\infty} d_\infty(F(f_n), F(f)) \leq \varepsilon(2|M| + \varepsilon)$. Since this holds for each $\varepsilon > 0$, we conclude that $\lim_{n\to\infty} d_\infty(F(f_n), F(f)) = 0$. By Lemma 242, we have $F(f_n) \to F(f)$, as desired. ▲

Example 296 Given $t_0 \in (0,1)$, let $F : A \subseteq C([0,1]) \to \mathbb{R}$ be defined by

$$F(f) = f'(t_0), \quad \forall f \in C([0,1]).$$

The domain A consists of the functions in $C([0,1])$ that are differentiable at x_0. We prove that the function F is discontinuous at each point of its domain. For each $n \geq 1$, let $f_n \in C([0,1])$ be such that $f_n'(t_0) = 1$ and $\max_{t\in[0,1]} f(t) < 1/n$. For example, let

$$f_n(t) = \begin{cases} 0 & \text{if} \quad 0 \leq t \leq t_0 - \frac{1}{4n} \\ t - t_0 + \frac{\varepsilon}{4n} & \text{if} \quad t_0 - \frac{1}{4n} < t \leq t_0 + \frac{1}{4n} \\ \frac{1}{2n} & \text{if} \quad t_0 + \frac{1}{4n} < t \leq 1 \end{cases}$$

Given $f \in A$, for each $n \geq 1$ set $g_n = f + f_n$. We have $g_n \in C([0,1])$ and

$$d_\infty(g_n, f) = \max_{t\in[0,1]} |(f + f_n - f)(t)| = \max_{t\in[0,1]} |f_n(t)| < 1/n \longrightarrow 0, \tag{6.17}$$

and

$$F(g_n) = (f + f_n)'(t_0) = f'(t_0) + f_n'(t_0) = F(f) + 1, \quad \forall n \geq 1. \tag{6.18}$$

Therefore, (6.18) shows that the sequence $\{F(g_n)\}_{n\geq 1}$ does not converge to $F(f)$, though by (6.17) we have $g_n \to f$. It follows that, by Corollary 292, F is not continuous at f. Since f was an arbitrary function in A, we conclude that F is not continuous at any point of A. ▲

We conclude this subsection by showing that in compact metric spaces continuity takes a stronger form. Given two metric spaces X and Y, a function $f : X \to Y$ is called *uniformly continuous* if for each $\varepsilon > 0$ there exists $\delta_\varepsilon > 0$ such that

$$d_Y(f(x_1), f(x_2)) < \varepsilon$$

for each $x_1, x_2 \in X$ with $d_X(x_1, x_2) < \delta_\varepsilon$.

With respect to simple continuity, uniform continuity is not linked to a point x_0 but it holds on the whole space. In particular, for a fixed ε, the quantity δ_ε must work for all points of the space, and not only for a given x_0. It is therefore a substantially stronger notion of continuity, which the next result (whose proof we omit) shows to be automatically satisfied in compact metric spaces.

Theorem 297 *Let X and Y be two metric spaces, with X compact. Each function $f : X \to Y$ continuous on X is uniformly continuous.*

6.5.3 Intermezzo: Images and Counterimages

In the sequel we will need some properties of images and counterimages. Recall that, given a function $f : X \to Y$, the *counterimage* of $E \subseteq Y$ is given by

$$f^{-1}(E) = \{x \in X : f(x) \in E\},$$

while the *image* of $A \subseteq X$ is

$$f(A) = \{y \in Y : f(x) = y \text{ for some } x \in A\}.$$

Recall also that f is *surjective* if $f(X) = Y$ and is *injective* if $f(x_1) \neq f(x_2)$ for each $x_1, x_2 \in X$ with $x_1 \neq x_2$. A function that is both injective and surjective is called *bijective*.

It is immediate to see that

$$f(A_1) \subseteq f(A_2) \text{ if } A_1 \subseteq A_2 \subseteq X \quad \text{and} \quad f^{-1}(E_1) \subseteq f^{-1}(E_2) \text{ if } E_1 \subseteq E_2 \subseteq Y.$$

The next result, whose proof is left to the reader, collects other important properties of images and counterimages. Properties (i) and (v) are especially important as they show that counterimages are well behaved with respect to the set operations, that is, unions, intersections, and complements.

Lemma 298 *Let $f : X \to Y$ be a function between two sets X and Y. We have:*

(i) $f^{-1}\left(\bigcup_{i\in I} E_i\right) = \bigcup_{i\in I} f^{-1}(E_i)$ *and* $f^{-1}\left(\bigcap_{i\in I} E_i\right) = \bigcap_{i\in I} f^{-1}(E_i)$*, where $E_i \subseteq Y$ for each $i \in I$.*

(ii) $f\left(\bigcap_{i\in I} A_i\right) \subseteq \bigcap_{i\in I} f(A_i)$ *and* $f\left(\bigcup_{i\in I} A_i\right) = \bigcup_{i\in I} f(A_i)$*, where $A_i \subseteq X$ for each $i \in I$.*

(iii) $f(f^{-1}(E)) \subseteq E$ *for each $E \subseteq Y$; the equality holds if f is surjective.*

(iv) $f^{-1}(f(A)) \supseteq A$ *for each $A \subseteq X$; the equality holds if f is injective.*

(v) $f^{-1}(E^c) = (f^{-1}(E))^c$ *for each $E \subseteq Y$.*

6.5.4 Continuity and Topology

We now give a fundamental characterization of continuous functions. For simplicity we consider the case of functions having as domain the whole space.

Theorem 299 *Let (X, d_X) and (Y, d_Y) be two metric spaces. A function $f : X \to Y$ is continuous if and only if the counterimage $f^{-1}(G)$ of each open set G of Y is itself an open set of X.*

Proof Suppose that f is continuous. Let G be an open set of Y, and let $x_0 \in f^{-1}(G)$. We prove that x_0 is an interior point of $f^{-1}(G)$. Since $f(x_0) \in G$ and the set G is open, there exists a neighborhood $B_\varepsilon(f(x_0))$ of $f(x_0)$ such that $B_\varepsilon(f(x_0)) \subseteq G$. Since f is continuous at x_0, there exists δ_ε such that $f(x) \in B_\varepsilon(f(x_0))$ for each $x \in B_{\delta_\varepsilon}(x_0)$. Being $B_\varepsilon(f(x_0)) \subseteq G$, we have therefore $f(x) \in G$ for each $x \in B_{\delta_\varepsilon}(x_0)$, which implies $B_{\delta_\varepsilon}(x_0) \subseteq f^{-1}(G)$. This proves that x_0 is an interior point of $f^{-1}(G)$, as desired.

Suppose that, for each open set G of Y, the set $f^{-1}(G)$ is itself an open set of X. We prove that, taken any $x_0 \in X$, f is continuous at x_0. Let V be an open set containing $f(x_0)$. Since V is open, $f^{-1}(V)$ is itself an open set. Since $x_0 \in f^{-1}(V)$, there exists an open set G (for instance, a neighborhood) of x_0 such that $G \subseteq f^{-1}(V)$, and therefore such that $f(x) \in V$ for each $x \in G$. We conclude that f is continuous at x_0. ■

Thanks to Lemma 298-(v), we have the following version of Theorem 299 for closed sets: a function $f : X \to Y$ is continuous if and only if the counterimage $f^{-1}(F)$ of each closed set F of Y is itself a closed set of X.

Example 300 Consider a real valued function $f : X \to \mathbb{R}$ defined on a metric space X. Denote by $(f < t)$ the set $\{x \in X : f(x) < t\}$, that is, $(f < t) = f^{-1}((-\infty, t))$. If f is continuous, then by Theorem 299 we have that $(f < t)$ is an open set in X because it is the counterimage of the open set $(-\infty, t)$ of $\mathbb{R}$. In a similar way, we prove that also $(f > t)$ is an open set in X, while $(f \leq t)$ and $(f \geq t)$ are closed sets in X. Finally, $(f = t)$ is a closed set in X because we have $(f = t) = f^{-1}(\{t\})$ and the singleton $\{t\}$ is a closed set of $\mathbb{R}$. ▲

There is no, instead, counterpart of Theorem 299 for images: given a continuous function f, in general the image $f(G)$ of an open set is not open and the image $f(F)$ of a closed set is not closed.

Example 301 Let $f : \mathbb{R} \to \mathbb{R}$ be defined as $f(x) = x^2$. For $G = (-1, 1)$ we have $f(G) = [0, 1)$, which is not open. ▲

Example 302 Let $f : \mathbb{R} \to \mathbb{R}$ be defined as $f(x) = e^x$. The real line $\mathbb{R}$ is a closed set (it is also an open set, but here this is not of interest), and we have $f(\mathbb{R}) = (0, +\infty)$, which is not closed. ▲

In the last example the closed set considered, i.e. $\mathbb{R}$, is not bounded and hence it is not a compact set. This does not happen by chance: the next important result shows that when a closed set is compact, its image is then compact.

Theorem 303 *Let (X, d_X) and (Y, d_Y) be two metric spaces. If a function $f : A \subseteq X \to Y$ is continuous on a compact subset $K \subseteq A$ of X, then the image $f(K)$ of K is a compact set of Y.*

Proof Given any sequence $\{y_n\}_n \subseteq f(K)$, by Theorem 275 to show that $f(K)$ is compact is enough to show that there is subsequence $\{y_{n_k}\}_k$ that converges to some $y \in f(K)$. Since $\{y_n\}_n \subseteq f(K)$, by definition there is a sequence $\{x_n\}_n \subseteq K$ such that $f(x_n) = y_n$ for each n. Since K is compact, by Theorem 275 there is a subsequence $\{x_{n_k}\}_k$ that converges to some $x \in K$. Since f is continuous at x, we have $\lim_k f(x_{n_k}) = f(x)$. Hence, by setting $y = f(x)$, we have that $\{y_{n_k}\}_k$ converges to $y \in f(K)$, as desired. ■

The next section will show that the classic Weierstrass Theorem is a consequence of this theorem (Exercise 13.0.45 gives another proof of this important result).

We conclude this section with some simple but useful results on continuity, whose simple proof we omit. For real valued functions, by Proposition 249 we have:

Proposition 304 *Let $f : A \subseteq X \to \mathbb{R}$ and $g : A \subseteq X \to \mathbb{R}$ be functions defined on a subset A of a metric space X. If both f and g are continuous at a point $x_0 \in A$, we have:*

(i) $\alpha f + \beta g$ is continuous at x_0 for each $\alpha, \beta \in \mathbb{R}$;

(ii) fg is continuous at x_0;

(iii) f/g is continuous at x_0, when $g(x_0) \neq 0$.

This shows, *inter alia*, that the set of the functions that are continuous at a point $x_0 \in X$ is a vector space.

Consider now the functions $f : A \subseteq \mathbb{R}^n \to \mathbb{R}^m$. Recall from Subsection 4.4.1 that these functions can be written as $f = (f_1, ...f_m) : A \subseteq \mathbb{R}^n \to \mathbb{R}^m$, where each $f_i : A \subseteq \mathbb{R}^n \to \mathbb{R}$, for $i = 1, ..., m$, is a real valued function on A.

Proposition 305 *A function $f = (f_1, ...f_n) : A \subseteq \mathbb{R}^n \to \mathbb{R}^m$ is continuous at $x \in A$ if and only if each function $f_i : A \subseteq \mathbb{R}^n \to \mathbb{R}$, for $i = 1, ..., m$, is itself continuous at x.*

Example 306 Let $F = (F_1, F_2) : \mathbb{R}^2 \to \mathbb{R}^2$ be defined as in Example 285, that is

$$F_1(x) = \begin{cases} x_1 + x_2 + 1 & x \neq (0,0) \\ 0 & x = (0,0) \end{cases} \quad \text{and} \quad F_2(x) = 1 + x_1^2 x_2, \qquad \forall x \in \mathbb{R}^2.$$

The function $F_2 : \mathbb{R}^2 \to \mathbb{R}$ is continuous on $\mathbb{R}^2$, while the function $F_1 : \mathbb{R}^2 \to \mathbb{R}$ is continuous at all the points of $\mathbb{R}^2$, except the origin. By Proposition 305, F is therefore continuous at all the points of $\mathbb{R}^2$ except the origin. This is not by chance: in Example 285 we verified that $\lim_{x \to (0,0)} F(x) \neq F(0,0)$. ▲

Given two functions, $f : A \subseteq \mathbb{R}^n \to \mathbb{R}^m$ and $g : A \subseteq \mathbb{R}^n \to \mathbb{R}^m$, the function $fg : A \subseteq \mathbb{R}^n \to \mathbb{R}$ is given by

$$(fg)(x) = f(x) \cdot g(x), \qquad \forall x \in A,$$

that is, it is given by the inner product of the two vectors $f(x)$ and $g(x)$ of $\mathbb{R}^m$. It is easy to see that if f and g are continuous at a point $x \in A$, then also the functions fg and $f + g$ are continuous at x. In other words, *mutatis mutandis*, points (i) and (ii) of Proposition 304 hold also for functions with values in $\mathbb{R}^m$.

6.6 Weierstrass Theorem

In this section we study the famous Weierstrass Theorem, and some of its variants. First of all we state and prove the classic version of this result, whose proof is a consequence of the Theorem 303 just seen.

Theorem 307 (Weierstrass) *Let $f : A \subseteq X \to \mathbb{R}$ be a function continuous on a compact subset $K \subseteq A$ of X. Then, both $\arg\min_{x \in K} f(x)$ and $\arg\max_{x \in K} f(x)$ are nonempty and compact.*

In other words, f has both global minima and maxima in K, that is, there exist $x_1, x_2 \in K$ such that

$$f(x_1) = \max_{x \in K} f(x) \qquad \text{and} \qquad f(x_2) = \min_{x \in K} f(x).$$

The version of Weierstrass theorem for functions $f : [a, b] \to \mathbb{R}$ studied in basic Calculus is therefore a very special case of this general result.

Proof By Theorem 303, $f(K)$ is a compact set in $\mathbb{R}$. By the Heine-Borel Theorem 272, $f(K)$ is therefore a closed and bounded set of $\mathbb{R}$. Since $f(K)$ is bounded, by the

completeness of $\mathbb{R}$ there exist $\sup f(K)$ and $\inf f(K)$. On the other hand, $\sup f(K)$ and $\inf f(K)$ belong to the closure of $f(K)$. Consider for instance $\sup f(K)$. For each neighborhood $B_{\frac{1}{n}}(\sup f(K))$ there exists a point of $f(K)$, denoted x_n, such that $x_n > \sup f(K) - 1/n$. The sequence $\{x_n\}_n$ thus constructed converges to $\sup f(K)$ and so, by Theorem 254, $\sup f(K) \in \overline{f(K)}$. A similar argument shows that also $\inf f(K) \in \overline{f(K)}$. Since $f(K)$ is closed, by Theorem 235 we have $f(K) = \overline{f(K)}$. Hence, both $\sup f(K)$ and $\inf f(K)$ belong to $f(K)$, that is, $\sup f(K) = \max f(K)$ and $\inf f(K) = \min f(K)$.

It remains to show that $\arg\max_{x \in K} f(x)$ is a closed (and so compact) subset of K. Let $\{x_n\}_n \subseteq \arg\max_{x \in K} f(x)$ be such that $x_n \to x_0 \in X$. Since K is compact (and so closed), $x_0 \in K$. By Corollary 255, we need to show that $x_0 \in \arg\max_{x \in K} f(x)$. Let $x \in K$. We have $f(x_n) \geq f(x)$ for each $n \geq 1$, and so, f being continuous, $f(x_0) = \lim_n f(x_n) \geq f(x)$. Since x is an arbitrary element of K, we conclude that $x_0 \in \arg\max_{x \in K} f(x)$, as desired. A similar argument shows that also $\arg\min_{x \in K} f(x)$ is compact. ■

The Weierstrass Theorem ensures the existence of both maxima and minima. In some cases, however, we are only interested in the existence of either maxima or minima. For example, in many economic applications it is crucial the existence of maxima, while the possible existence also of minima is of little or no interest at all.

For these reasons we now introduce a weakened version of continuity, with the objective of establishing a version of the Weierstrass Theorem that, under weaker hypothesis, guarantees the existence of maxima, without considering the possible existence of minima.

First of all, recall that a function $f : A \subseteq X \to \mathbb{R}$ is continuous at a point $x_0 \in A$ when, for each $\varepsilon > 0$, there exists $\delta_\varepsilon > 0$ such that $f(x_0) - \varepsilon < f(x) < f(x_0) + \varepsilon$ for each $x \in A$ with $d_X(x, x_0) < \delta_\varepsilon$.

If in this definition we keep the second inequality, we have the following weakening of continuity.

Definition 308 *A function $f : A \subseteq X \to \mathbb{R}$ is said to be upper semicontinuous at $x_0 \in A$ if for each $\varepsilon > 0$ there exists $\delta_\varepsilon > 0$ such that*

$$f(x) < f(x_0) + \varepsilon$$

for each $x \in A$ with $d_X(x, x_0) < \delta_\varepsilon$.

A function that is upper semicontinuous at each point of a set E is called *upper semicontinuous* on E. The function is called *upper semicontinuous* when it is upper semicontinuous at all the points of its domain.

The next result is the version of Corollary 292 for semicontinuous functions, and it helps to understand the importance of this weakening of the notion of continuity.

Proposition 309 *A function $f : A \subseteq X \to \mathbb{R}$ is upper semicontinuous at the point $x_0 \in X$ if and only if $\limsup_n f(x_n) \leq f(x_0)$ for each sequence $\{x_n\}_n \subseteq A$ such that $x_n \to x_0$.*

Proof Let f be upper semicontinuous at the point x_0. Let $\{x_n\}_n$ be such that $x_n \to x_0$. By Definition 308, fixed $\varepsilon > 0$ we have $f(x_n) < f(x_0) + \varepsilon$ for each n. Therefore, $\limsup_n f(x_n) \leq f(x_0) + \varepsilon$. Since this is true for each $\varepsilon > 0$, we conclude that $\limsup_n f(x_n) \leq f(x_0)$.

Suppose now that $\limsup_n f(x_n) \leq f(x_0)$ for each sequence $\{x_n\}_n$ such that $x_n \to x_0$. Let $\varepsilon > 0$ and suppose, by contradiction, that f is not upper semicontinuous at x_0. Therefore, for each $\delta > 0$ there exists x_δ such that $d_X(x_\delta, x_0) < \delta$ and $f(x_\delta) \geq f(x_0) + \varepsilon$. Setting $\delta = 1/n$, it follows that for each n there exists x_n such that $d_X(x_n, x_0) < 1/n$ and $f(x_n) \geq f(x_0) + \varepsilon$. In this way we can construct a sequence $\{x_n\}_n$ such that $x_n \to x_0$ and $f(x_n) \geq f(x_0) + \varepsilon$ for each n. Therefore, $\liminf_n f(x_n) \geq f(x_0) + \varepsilon$, which contradicts $\limsup_n f(x_n) \leq f(x_0)$ and thus proves that f is upper semicontinuous at x_0. ■

Example 310 The function $f : [0,1] \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} 1 & \text{if} \quad x = 0 \\ x & \text{if} \quad x \in (0,1] \end{cases}$$

is upper semicontinuous. In fact, it is continuous at each $x \in (0,1]$. As to $x = 0$, consider $\{x_n\}_{n \geq 1} \subseteq [0,1]$ with $x_n \to 0$. For each such x_n we have $f(x_n) \leq 1$ and therefore $\limsup_n f(x_n) \leq 1 = f(0)$. By Proposition 309, f is upper semicontinuous also at 0. ▲

>From Example 300 we know that the upper level sets $(f \geq t)$ of continuous functions are closed. Next result shows how this property is still true for upper semicontinuous functions and actually characterizes this weakened notion of continuity.

Proposition 311 *A function $f : A \subseteq X \to \mathbb{R}$ is upper semicontinuous on a closed subset $B \subseteq A$ of X only if the sets $(f \geq t) \cap B$ are closed for each $t \in \mathbb{R}$. The converse is true when $A = B$.*

Proof For simplicity, assume $A = X$ (see Exercise 13.0.46 for the general case). Let f be upper semicontinuous on B. Fixed $t \in \mathbb{R}$, we want to show that $(f \geq t) \cap B$ is closed. Let $\{x_n\}_n \subseteq (f \geq t) \cap B$ with $x_n \to x \in X$. By Corollary 255, we have to

show that $x \in (f \geq t) \cap B$. First observe that $x \in B$ since B is closed. Moreover, $f(x_n) \geq t$ for each $n \geq 1$. Since f is upper semicontinuous, by Proposition 309 we have $\limsup_n f(x_n) \leq f(x)$. Therefore $t \leq f(x)$, i.e, $x \in (f \geq t)$. We conclude that $x \in (f \geq t) \cap B$, as desired.

Viceversa, suppose $A = B$ and that the sets $(f \geq t) \cap A$ are closed for each $t \in \mathbb{R}$. Fix $x \in A$ and let $\{x_n\}_n \subseteq A$ be such that $x_n \to x$. We want to show that $\limsup_n f(x_n) \leq f(x)$. Assume *per contra* that $\limsup_n f(x_n) > f(x)$. Let $\alpha \in \mathbb{R}$ be such that $\limsup_n f(x_n) > \alpha > f(x)$. There exists a subsequence $\{x_{n_k}\}_k$ such that $f(x_{n_k}) \geq \alpha$ for each k. On the other hand $x_n \to x$ implies $x_{n_k} \to x$, and therefore by Corollary 255 we have $x \in (f \geq \alpha) \cap A$ since $(f \geq \alpha) \cap A$ is closed. But, this implies $f(x) \geq \alpha > f(x)$, and this contradiction allows us to conclude that $\limsup_n f(x_n) \leq f(x)$. ■

Example 312 Given a closed subset F of a metric space X, let $1_F : X \to \mathbb{R}$ be defined by

$$1_F(x) = \begin{cases} 1 & \text{if } x \in F \\ 0 & \text{if } x \notin F \end{cases}.$$

The function 1_F is upper semicontinuous. In fact,

$$(1_F \geq t) = \begin{cases} X & \text{if } \quad t \leq 0 \\ F & \text{if } \ t \in (0,1] \\ \emptyset & \text{if } \quad t > 1 \end{cases}$$

and therefore the sets $(1_F \geq t)$ are closed for each $t \in \mathbb{R}$. ▲

We now introduce a last notion.

Definition 313 *A function $f : A \subseteq X \to \mathbb{R}$ is said to be coercive on a subset B if there exists a scalar $t \in \mathbb{R}$ such that $(f \geq t) \cap B$ is compact and nonempty.*

A function $f : A \subseteq X \to \mathbb{R}$ is called *coercive* when $B = A$, i.e., when B is the domain of the function.[11]

Example 314 Let $f : \mathbb{R} \to \mathbb{R}$ be defined as $f(x) = 1 - x^2$. This function is coercive; in fact:

$$(f \geq t) = \begin{cases} \left[-\sqrt{1-t}, \sqrt{1-t}\right] & \text{if } t \leq 1 \\ \emptyset & \text{if } t > 1 \end{cases}$$

and therefore $(f \geq t)$ is compact and nonempty for each $t \leq 1$. ▲

[11] Notice that in this definition B is not required to be a subset of the domain A. This is without any loss of generality since $(f \geq t) \cap B \subseteq A$, and so only the points in $A \cap B$ matter for the definition.

Example 315 Let $f : \mathbb{R} \to \mathbb{R}$ be defined as $f(x) = e^{-|x|}$. This function is coercive; in fact:

$$(f \geq t) = \begin{cases} \mathbb{R} & \text{if } t \leq 0 \\ [\lg t, -\lg t] & \text{if } t \in (0,1] \\ \emptyset & \text{if } t > 1 \end{cases}$$

and therefore $(f \geq t)$ is compact and nonempty for each $t \in (0,1]$. ▲

Example 316 Let $f : \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) = \begin{cases} \lg(|x|) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

Set $B = [-1,1]$. We have

$$(f \geq t) = \begin{cases} (-\infty, -e^t] \cup [e^t, +\infty) \cup \{0\} & \text{if } t \leq 0 \\ (-\infty, -e^t] \cup [e^t, +\infty) & \text{if } t > 0 \end{cases},$$

and therefore

$$(f \geq t) \cap B = \begin{cases} \emptyset & \text{if } t > 0 \\ [-1, -e^t] \cup [e^t, 1] \cup \{0\} & \text{if } t \leq 0 \end{cases}.$$

This function is therefore coercive on B (observe that B is not a compact). ▲

We can now state the following fundamental version of the Weierstrass Theorem, in which only the existence of maxima is guaranteed.

Theorem 317 *Let $f : A \subseteq X \to \mathbb{R}$ be a function that is both upper semicontinuous and coercive on a subset $B \subseteq A$ of X. Then, $\arg\max_{x \in B} f(x)$ is nonempty and compact.*

In other words, f has at least a point of maximum in B, that is, there exists $\widehat{x} \in B$ such that

$$f(\widehat{x}) = \max_{x \in B} f(x).$$

Proof For simplicity, assume $A = X$. Since f is coercive on B, there exists $\overline{t} \in \mathbb{R}$ such that $\left(f \geq \overline{t}\right) \cap B$ is compact and nonempty. Taking $t \geq \overline{t}$, we have $(f \geq t) \subseteq \left(f \geq \overline{t}\right)$, and therefore $(f \geq t) \cap B \subseteq \left(f \geq \overline{t}\right) \cap B$. We prove that also the set $(f \geq t) \cap B$ is compact. By Proposition 274, it is enough to prove that $(f \geq t) \cap B$ is closed. Let $\{x_n\}_n \subseteq (f \geq t) \cap B$ with $x_n \to x \in X$. By Corollary 255, we have to prove that $x \in (f \geq t') \cap B$. Since $\{x_n\}_n \subseteq \left(f \geq \overline{t}\right) \cap B$ and this set is closed, we have $x \in \left(f \geq \overline{t}\right) \cap B$ thanks to Corollary 255. Therefore, $x \in B$. On the other hand, $\{x_n\}_n \subseteq (f \geq t)$ implies $F(x_n) \geq t$ for each n. Since f is upper semicontinuous on B,

it follows that $F(x) \geq \limsup_n F(x_n) \geq t$, and therefore $x \in (f \geq t)$. This proves that $x \in (f \geq t) \cap B$, which allows us to conclude that $(f \geq t) \cap B$ is compact.

Set $K_t = (f \geq t) \cap B$ for each $t \geq \overline{t}$, so that $\{K_t\}_{t \geq \overline{t}}$ is a collection of compact sets. We have $K_t \subseteq K_{t'}$ if $t \geq t'$, and therefore given any finite collection $\{K_{t_i}\}_{i=1}^n$ of $\{K_t\}_{t \in \mathbb{R}}$, with $t_1 < t_2 < \cdots < t_n$, we have

$$\bigcap_{i=1}^{n} K_{t_i} = K_{t_n}. \tag{6.19}$$

We begin by observing that $\sup_{x \in B} f(x)$ exists. In fact, suppose that this is not true, so that $K_t \neq \emptyset$ for each $t \geq \overline{t}$. Expression (6.19) implies that for each finite subcollection $\{K_{t_i}\}_{i \in J}$ of $\{K_t\}_{t \geq \overline{t}}$ we have $\bigcap_{i \in J} K_i \neq \emptyset$. By Lemma 279, we have $\bigcap_{t \geq \overline{t}} K_t \neq \emptyset$. Let $x \in \bigcap_{t \geq \overline{t}} K_t$. We have $f(x) \geq t$ for each $t \in \mathbb{R}$, and therefore $f(x) = +\infty$, which contradicts the fact that f is real valued. We conclude that $\sup_{x \in B} f(x)$ exists.

Set now $\alpha = \sup_{x \in B} f(x)$. Clearly, $\alpha \geq \overline{t}$ because by hypothesis $K_{\overline{t}} \neq \emptyset$. If $\alpha = \overline{t}$, then there exists $x \in K_{\overline{t}}$ such that $f(x) = \alpha$, which proves the statement.

Suppose on the contrary that $\alpha > \overline{t}$, so that $K_t \neq \emptyset$ for each t if $\overline{t} \leq t < \alpha$. Also here (6.19) implies that for each finite subcollection $\{K_{t_i}\}_{i \in J}$ of $\{K_t\}_{\overline{t} \leq t < \alpha}$ we have $\bigcap_{i \in J} K_i \neq \emptyset$. By Lemma 279, we therefore have $\bigcap_{\overline{t} \leq t < \alpha} K_t \neq \emptyset$. Let $x \in \bigcap_{\overline{t} \leq t < \alpha} K_t$. We have $f(x) \geq t$ for each $t < \alpha$, and therefore $f(x) \geq \alpha$. On the other hand, by the definition of supremum we cannot have $f(x) > \alpha$. It follows that $f(x) = \alpha$, and so x is a point of maximum.

It remains to show that $\arg\max_{x \in B} f(x)$ is compact. It is enough to show that $\arg\max_{x \in B} f(x)$ is closed since it is a subset of the compact set $\left(f \geq \overline{t}\right) \cap B$. Let $\{x_n\}_n \subseteq \arg\max_{x \in B} f(x)$ be such that $x_n \to x_0 \in X$. By Corollary 255, we need to show that $x_0 \in \arg\max_{x \in B} f(x)$. Since $\left(f \geq \overline{t}\right) \cap B$ is compact, we have $x_0 \in \left(f \geq \overline{t}\right) \cap B$, i.e., $x_0 \in B$. Now, let $x \in B$. We have $f(x_n) \geq f(x)$ for each $n \geq 1$, and so, f being upper continuous, $f(x_0) \geq \limsup_n f(x_n) \geq f(x)$. Since x is an arbitrary element of B, we conclude that $x_0 \in \arg\max_{x \in B} f(x)$, as desired. ■

Example 318 By Theorem 317, the function of Example 315 has at least one point of maximum in $\mathbb{R}$. Notice that, instead, the function does not have points of minimum; on the other hand, here the Weierstrass Theorem does not hold because $\mathbb{R}$ is not a compact set. ▲

Example 319 Again by Theorem 317, also the function of Example 316 has a point of maximum on the set B. It is easy to see that the function does not have points

of minimum on B (here Weierstrass Theorem cannot be applied because B is not compact). ▲

In conclusion, Theorem 317 only ensures the existence of maxima, but it has the advantage of using hypotheses that, as shown by Examples 316 and 322, are substantially weaker than those of the Weierstrass Theorem. As a result, Theorem 317 has a much greater scope than Weierstrass Theorem and it plays a key role in optimization problems that are only interested in the existence of maxima.

That said, even a very general Theorem 317 gives only a sufficient condition of optimality. It is easy to give examples where there exist points of global maximum even though the hypotheses of Theorem 317 do not hold.

Example 320 Let $f : \mathbb{R}^n \to \mathbb{R}$ be a constant function, say $f(x) = 1$ for each $x \in \mathbb{R}^n$. All points of $\mathbb{R}^n$ are trivially points of global maximum (and of minimum as well). This function is continuous, but it is not coercive, and therefore the hypotheses of Theorem 317 are not satisfied. ▲

The next corollary is a simple consequence of Theorem 317 and shows that coercitivity is implied by the upper semicontinuity when the set in which we search the point of maximum is compact. This version of Theorem 317 is especially suited for a comparison with the Weierstrass Theorem, since it holds under the same hypotheses, except upper semicontinuity in place of continuity.

Corollary 321 *Let $f : A \subseteq X \to \mathbb{R}$ be a function that is upper semicontinuous on a compact subset $K \subseteq A$ of X. Then,* $\arg\max_{x \in K} f(x)$ *is nonempty and compact.*

That is, f has at least a maximum in K, that is, there exists $\widehat{x} \in K$ such that

$$f(\widehat{x}) = \max_{x \in K} f(x).$$

Proof By Proposition 311, $(f \geq t) \cap K$ is closed for each $t \in \mathbb{R}$, and therefore is compact by Proposition 274. Hence, f is coercive on K. ■

Example 322 Consider the upper semicontinuous function $f : [0,1] \to \mathbb{R}$ seen in Example 310, given by

$$f(x) = \begin{cases} 1 & \text{if} \quad x = 0 \\ x & \text{if} \quad x \in (0,1] \end{cases}.$$

By Corollary 321, this function has at least a point of maximum in its domain $[0,1]$. We can verify that this function does not have points of minimum in its domain (here the Weierstrass Theorem cannot be applied because the function is not continuous).▲

For functions defined on $\mathbb{R}^n$ we can give a condition that, together with the upper continuity, guarantees coercitivity (and that therefore allows to apply Theorem 317).

Proposition 323 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be an upper semicontinuous function. The sets $(f \geq t)$ are compact for each $t \in \mathbb{R}$ (and so f is coercive) if $f(x) \to -\infty$ when $\|x\| \to +\infty$.*

Proof Suppose to have $f(x) \to -\infty$ when $\|x\| \to +\infty$. We want to prove that the sets $(f \geq t)$ are compact for each $t \in \mathbb{R}$. Given $t \in \mathbb{R}$, the set $(f \geq t)$ is closed since f is upper semicontinuous. It remains to verify its boundedness. By hypothesis, there exists $k > 0$ such that $\|x\| \geq k$ implies $f(x) \leq t$. Therefore, we have $x \in (f \geq t)$ only if $\|x\| < k$ for a suitable k. Hence, $(f \geq t) \subseteq B_k(\mathbf{0})$, and the set $(f \geq t)$ is therefore bounded. Since it is also closed, it follows that it is compact. ■

The condition $f(x) \to -\infty$ when $\|x\| \to +\infty$ can sometimes be not easy to verify. For this reason next we give a stronger condition, which is however easier to verify.

Corollary 324 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be an upper semicontinuous function. The sets $(f \geq t)$ are compact for each $t \in \mathbb{R}$ (and f is therefore coercive) if there exist $\alpha > 0$ and $\beta \in \mathbb{R}$ such that $f(x) \leq \beta - \alpha \|x\|$.*

Proof Clearly $f(x) \leq \beta - \alpha \|x\|$ implies $f(x) \to -\infty$ when $\|x\| \to +\infty$. ■

Example 325 Let $f : \mathbb{R}^n \to \mathbb{R}$ be defined as $f(x) = 1 - \|x\|$ for each $x \in \mathbb{R}^n$. The function f is continuous and, by Corollary 324, is also coercive. By Theorem 317 the function has at least one point of maximum in $\mathbb{R}^n$. On the other hand, it is easy to see how this function does not have points of minimum in $\mathbb{R}^n$ (the Weierstrass Theorem here does not hold since $\mathbb{R}^n$ is not compact). ▲

We conclude by observing that for the points of minimum hold results that are specular relative to those that we have established here. In this case the lower level sets $(f \leq t)$ become relevant; in particular, it is easy to see that f admits at least a point of minimum in B provided the sets $(f \leq t) \cap B$ are compact for each $t \in \mathbb{R}$.

Chapter 7

Normed Vector Spaces

7.1 Norms and Metrics

In this chapter we go back to vector spaces and we show how it is possible to introduce in a natural way a topological structure on them.

In Calculus is sometimes studied the so-called Euclidean norm $\|\cdot\|_2$ of $\mathbb{R}^n$, defined as $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$ for each $x \in \mathbb{R}^n$. This norm satisfies the following properties:[1]

- $\|x\|_2 \geq 0$ for each $x \in \mathbb{R}^n$, and $\|x\|_2 = 0$ if and only if $x = \mathbf{0}$;
- $\|x + y\|_2 \leq \|x\|_2 + \|y\|_2$ for each $x, y \in \mathbb{R}^n$;
- $\|\alpha x\|_2 = |\alpha| \|x\|_2$ for each $\alpha \in \mathbb{R}$ and each $x \in \mathbb{R}^n$.

For each $x \in \mathbb{R}^n$ we have $\|x\|_2 = d_2(x, \mathbf{0})$ and, more generally, for each $x, y \in \mathbb{R}^n$ we have $d_2(x, y) = \|x - y\|_2$. We can therefore see the Euclidean norm as the primitive notion, in whose terms it is then possible to define the Euclidean distance. This is actually the approach sometimes followed in Calculus.[2]

Since $\mathbb{R}^n$ is a vector space, all this suggests a possible way to introduce distances in vector spaces. This motivates the next definition.

Definition 326 *Given a vector space V, a functional $\|\cdot\| : V \to \mathbb{R}$ is said to be a* ***norm*** *if:*

(i) $\|v\| \geq 0$ *for each* $v \in V$, *and* $\|v\| = 0$ *if and only if* $v = \mathbf{0}$;

(ii) $\|v + w\| \leq \|v\| + \|w\|$ *for each* $v, w \in V$;

(iii) $\|\alpha v\| = |\alpha| \|v\|$ *for each* $\alpha \in \mathbb{R}$ *and each* $v \in V$.

[1] See Ambrosetti and Musu (1988) pp. 56 and 57.

[2] See Chapter III of Ambrosetti and Musu (1988).

The Euclidean norm is an example of norm in the vector space $\mathbb{R}^n$. We see another example.

Example 327 In the vector space $C([0,1])$, the function $\|\cdot\|_\infty : C([0,1]) \to \mathbb{R}$ defined by

$$\|f\|_\infty = \max_{t\in[0,1]} |f(t)|, \qquad \forall f \in C([0,1]),$$

is a norm. This can be easily checked by recalling what we did in Example 206. ▲

Definition 328 *A vector space V endowed with a norm $\|\cdot\|$ is called a* ***normed vector space****.*

The spaces $(\mathbb{R}^n, \|\cdot\|_2)$ and $(C([0,1]), \|\cdot\|_\infty)$ are examples of normed vector spaces.

The next result follows immediately from the properties of the norm and shows how each normed space has a natural metric structure.

Lemma 329 *Let $(V, \|\cdot\|)$ be a normed vector space. The function $d : V \times V \to \mathbb{R}$ defined by*

$$d(v,w) = \|v - w\|, \qquad \forall v, w \in V$$

is a distance.

Each normed space thus becomes a metric space and this generalizes the procedure seen in earlier courses for the Euclidean norm and distance.

Example 330 In the vector space $\mathbb{R}^n$, the function $\|\cdot\|_1 : \mathbb{R}^n \to \mathbb{R}$ defined by:

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|, \qquad \forall x \in \mathbb{R}^n,$$

is a norm. The normed vector space $(\mathbb{R}^n, \|\cdot\|_1)$ induces the metric space $(\mathbb{R}^n, d_1)$. ▲

Example 331 In the vector space $\mathbb{R}^n$, the function $\|\cdot\|_\infty : \mathbb{R}^n \to \mathbb{R}$ defined by:

$$\|x\|_\infty = \max_{i=1,\ldots,n} |x_i|, \qquad \forall x \in \mathbb{R}^n,$$

is a norm. The normed vector space $(\mathbb{R}^n, \|\cdot\|_\infty)$ induces the metric space $(\mathbb{R}^n, d_\infty)$. ▲

Example 332 In the vector space $C([0,1])$, the function $\|\cdot\|_1 : C([0,1]) \to \mathbb{R}$ defined by:

$$\|f\|_1 = \int_0^1 |f(t)|\, dt, \qquad \forall f \in C([0,1]),$$

is a norm. The normed vector space $(C([0,1]), \|\cdot\|_1)$ induces the metric space $(C([0,1]), d_1)$.
▲

Since a normed vector space have a natural metric structure, we can define in them all the topological notions we studied in the previous chapter. For example, given a normed vector space $(V, \|\cdot\|)$, a neighborhood $B_\varepsilon(v)$ of a vector v of V is given by:

$$B_\varepsilon(v) = \{w \in V : \|v - w\| < \varepsilon\}.$$

Vectors can thus viewed as interior points or boundary points of a certain set A of V, and so on. In particular, we have sets of V that are open, sets that are closed, and sets that are compact.

The neighborhoods $B_\varepsilon(\mathbf{0})$ of $\mathbf{0}$ are particularly important. Among them the one with radius 1, i.e.,

$$B_1(\mathbf{0}) = \{v \in V : \|v\| < 1\},$$

is called the *open unit ball*, while its closure:

$$\overline{B}_1(\mathbf{0}) = \{v \in V : \|v\| \leq 1\},$$

is called the *closed unit ball*. In view of its importance, in what follows we will often denote the closed unit ball $\overline{B}_1(\mathbf{0})$ by B_V. Also important is the unit sphere $\overline{B}_1(\mathbf{0}) \setminus B_1(\mathbf{0}) = \{v \in V : \|v\| = 1\}$, which will be denoted by S_V.

Finally, notice that in this context a set A is bounded when there exists $\varepsilon > 0$ such that $A \subseteq B_\varepsilon(\mathbf{0})$.

As to convergence, a sequence of vectors $\{v_n\}_{n\geq 1}$ converges to a vector v if for each $\varepsilon > 0$ there exists $\overline{n} \geq 1$ such that:

$$\|v_n - v\| < \varepsilon, \qquad \forall n \geq \overline{n}.$$

In particular, by Lemma 242 we have $v_n \to v$ if and only if $\|v_n - v\| \to 0$, and this is a useful criterion to check the convergence of a sequence of vectors.

Notice that $v_n \to v$ if and only if $v_n - v \to \mathbf{0}$. More generally, it is easy to see that Proposition 253-(i) can be extended to normed vector spaces: given two sequences $\{v_n\}_{n\geq 1}$ and $\{w_n\}_{n\geq 1}$ of vectors of a normed vector space V, with $v_n \to v$ and $w_n \to w$, we have:

$$\alpha v_n + \beta w_n \to \alpha v + \beta w, \qquad \forall \alpha, \beta \in \mathbb{R}. \tag{7.1}$$

A sequence of vectors $\{v_n\}_{n\geq 1}$ is Cauchy if, for each $\varepsilon > 0$, there exists $\overline{n} \geq 1$ such that:

$$\|v_m - v_n\| < \varepsilon, \qquad \forall m, n \geq \overline{n}.$$

Completeness characterizes the following fundamental class of normed vector spaces.

Definition 333 *A normed vector space whose metric is complete is called* ***Banach space.***

By Theorems 262 and 263, the normed vector spaces $(\mathbb{R}^n, \|\cdot\|_1)$, $(\mathbb{R}^n, \|\cdot\|_2)$, $(\mathbb{R}^n, \|\cdot\|_\infty)$ and $(C([0,1]), \|\cdot\|_\infty)$ are all examples of Banach spaces.

Example 334 In light of Example 264, $(C([0,1]), d_1)$ is an example of a normed vector space that is not Banach. ▲

7.2 Functionals and Operators

Now that we have a topological structure in vector spaces, we can talk of continuity of the functionals $L : V \to \mathbb{R}$ and, more generally, of the applications $T : V_1 \to V_2$.

Let $(V_1, \|\cdot\|_1)$ and $(V_2, \|\cdot\|_2)$ be two normed vector spaces. An application $T : V_1 \to V_2$ is continuous at $v \in V_1$ if, for each $\varepsilon > 0$, there exists $\delta_\varepsilon > 0$ such that:

$$\|w - v\|_1 < \delta_\varepsilon \Longrightarrow \|T(w) - T(v)\|_2 < \varepsilon, \qquad \forall w \in V_1.$$

Equivalently, by Corollary 292 T is continuous at v if and only if:

$$v_n \to v \Longrightarrow T(v_n) \to T(v).$$

Example 335 Let $(V_1, \|\cdot\|_1) = (V_2, \|\cdot\|_2) = (C([0,1]), \|\cdot\|_\infty)$, with the application $T : C([0,1]) \to C([0,1])$ defined as $T(f) = f^2$ for each $f \in C([0,1])$. As shown in Example 295, the application T is continuous. Observe that this application is not linear. ▲

For the special case of functionals $L : V \to \mathbb{R}$, where $(V_2, \|\cdot\|_2) = (\mathbb{R}, |\cdot|)$, we have continuity at $v \in V$ when, for each $\varepsilon > 0$, there exists $\delta_\varepsilon > 0$ such that:

$$\|v - w\| < \delta_\varepsilon \Longrightarrow |L(w) - L(v)| < \varepsilon, \qquad \forall w \in V.$$

Example 336 Let $(V, \|\cdot\|) = (C([0,1]), \|\cdot\|_\infty)$, with $L : C([0,1]) \to \mathbb{R}$ defined as $L(f) = \int_0^1 f(t)\,dt$ for each $f \in C([0,1])$. As shown in Example 294, the linear functional L is continuous. ▲

Example 337 In every normed space $(V, \|\cdot\|)$, the norm itself $\|\cdot\| : V \to \mathbb{R}$ is a continuous (nonlinear) functional. In fact, the reader can verify that for each $v, w \in V$ we have:

$$|\|v\| - \|w\|| \leq \|v - w\|, \qquad \forall v, w \in V \tag{7.2}$$

If $\{v_n\}_{n\geq 1}$ is a sequence in V such that $v_n \to v$, then:

$$|\|v_n\| - \|v\|| \leq \|v_n - v\| \to 0,$$

and therefore $\|v_n\| \to \|v\|$. In conclusion:

$$v_n \to v \Longrightarrow \|v_n\| \to \|v\|, \tag{7.3}$$

and $\|\cdot\| : V \to \mathbb{R}$ is therefore a continuous functional. ▲

Linear applications play a central role in the theory of normed spaces. We start therefore by studying their continuity properties.

Proposition 338 *A linear application $T : V_1 \to V_2$ between normed vector spaces is continuous at a point $v \in V_1$ if and only if it is continuous on V_1.*

We therefore have a first remarkable property: for linear applications, the continuity at a point guarantees automatically the continuity on the entire space. To verify whether a linear application is continuous it is therefore sufficient to verify that it is continuous at some point of V_1 (for example at the neutral element $\mathbf{0}$).[3]

Proof. We prove the "only if," the "if" being obvious. Let T be continuous at a vector $v_0 \in V$. We prove that it is continuous at any other vector $v \in V$. Let $\{v_n\}_{n\geq 1}$ be a sequence in V such that $v_n \to v$. Therefore, $v_n - v \to \mathbf{0}$, and, by (7.1), $v_n - v + v_0 \to v_0$. Being T continuous at v_0, we have $T(v_n - v + v_0) \to T(v_0)$. For the linearity of T, this implies:

$$T(v_n) - T(v) + T(v_0) \to T(v_0),$$

and hence $T(v_n) - T(v) \to \mathbf{0}$. By Corollary 292, T is continuous at v. ■

We now give a characterization of the continuity of linear applications. To do this, we need the following definition.

Definition 339 *A linear application $T : V_1 \to V_2$ between normed vector spaces is said to be bounded if there exists a scalar $K > 0$ such that:*

$$\|T(v)\|_2 \leq K \|v\|_1, \qquad \forall v \in V_1. \tag{7.4}$$

In particular, a linear functional $L : V \to \mathbb{R}$ is bounded if there exists a scalar $K > 0$ such that $|T(v)| \leq K \|v\|$ for each $v \in V$.

This definition of boundedness is a bit different from the usual one for functions. Recall from Calculus that a function $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ is bounded when there exists a scalar $M > 0$ such that $|f(x)| \leq M$ for each $x \in A$; i.e. if the image $f(A)$ is a bounded set in $\mathbb{R}$.[4] In the case of applications, this definition is generalized by saying that an application $T : V_1 \to V_2$ is bounded if there exists a scalar $M > 0$ such that $\|T(v)\|_2 \leq M$ for each $v \in V_1$; i.e., recalling (3.13), if the image $\mathrm{Im}(T)$ of T is a bounded set in V_2.

[3] For simplicity, in the following we denote by $\mathbf{0}$ both the neutral element of V_1 and that of V_2. The context should clarify to which neutral element we refer. In the same way, the symbol of convergence $\to$ is used both for the convergence in V_1 and in V_2.

[4] See Ambrosetti and Musu (1988) p. 84.

The next result clarifies the relationships between the notion of boundedness for linear applications introduced in Definition 339 and the usual notion of boundedness for functions just recalled.

Lemma 340 *A linear application $T : V_1 \to V_2$ between normed vector spaces is bounded if and only if the image $T(B_{V_1})$ of the closed unit ball of V_1 is a bounded set in V_2.*

A linear application is therefore bounded if and only if it is bounded in the traditional sense when restricted to the closed unit ball.

Proof. Let $T : V_1 \to V_2$ be bounded. By (7.4) we have $\|T(v)\|_2 \leq K$ for each $v \in B_{V_1}$, and therefore:

$$T(B_{V_1}) = \{T(v) : v \in B_{V_1}\} \subseteq \overline{B}_K(\mathbf{0}),$$

where $\overline{B}_K(\mathbf{0}) = \{w \in V_2 : \|w\|_2 \leq K\}$. Hence, $T(B_{V_1}) \subseteq B_{K'}(\mathbf{0})$ for any $K' > K$.

Viceversa, suppose that $T(B_{V_1})$ is a bounded set in V_2. There exists therefore a neighborhood $B_K(\mathbf{0})$ of $\mathbf{0} \in V_2$ such that $T(B_{V_1}) \subseteq B_K(\mathbf{0})$. Hence, given any $K' > K$, we have $\|T(v)\|_2 \leq K'$ for each $v \in B_{V_1}$. Given any vector v of V_1, we have:

$$\left\| \frac{v}{\|v\|_1} \right\|_1 = \frac{\|v\|_1}{\|v\|_1} = 1,$$

and so:

$$\left\| T\left(\frac{v}{\|v\|_1} \right) \right\|_2 \leq K', \qquad \forall v \in V_1,$$

which in turn implies (7.4) for $v \neq \mathbf{0}$. On the other hand, (7.4) clearly holds for $v = \mathbf{0}$, and we have therefore completed the proof. ■

Example 341 When $V_1 = V_2 = \mathbb{R}$, the only applications $T : \mathbb{R} \to \mathbb{R}$ have the form $T(x) = \alpha x$ with $\alpha \in \mathbb{R}$. The function T is obviously unbounded on the entire real line. But, its restriction on the closed unit ball $[-1, 1]$ is bounded because $|T(x)| \leq |\alpha|$ for each $x \in [-1, 1]$. In particular, $T([-1, 1]) \subseteq [-\alpha, \alpha]$ and therefore by Lemma 340 T is bounded as a linear application in the sense of Definition 339. ▲

The next result shows why we devoted all this attention to boundedness.

Theorem 342 *A linear application $T : V_1 \to V_2$ between normed vector spaces is continuous if and only if it is bounded.*

Hence, continuity and boundedness are equivalent properties for linear applications. This is another remarkable property of linear applications, which is in general altogether false for nonlinear applications, as the example that follows the proof will show.

Proof. We suppose that T is bounded and we prove that in this case T is continuous. Let $v_n \to v$, that is, $\|v_n - v\|_1 \to 0$. Since T is bounded, by (7.4) we have:

$$\|T(v_n) - T(v)\|_2 = \|T(v_n - v)\|_2 \leq K \|v_n - v\|_1 \to 0,$$

and therefore $T(v_n) \to T(v)$. By Corollary 292, T is continuous at v.

Let now T be continuous, and we show that this implies the boundedness of T. Suppose that this is not true, and that there exists therefore at least a sequence $\{v_n\}_{n\geq 1}$ in V_1 such that:

$$\frac{\|T(v_n)\|_2}{\|v_n\|_1} \to +\infty.$$

Set $\alpha_n = \|T(v_n)\|_2 / \|v_n\|_1$ and:

$$w_n = \frac{v_n}{\alpha_n \|v_n\|_1}, \qquad \forall n \geq 1.$$

Since $\|w_n\|_1 = 1/\alpha_n \to 0$, we have $w_n \to \mathbf{0}$. Being T continuous, it follows that $T(w_n) \to T(\mathbf{0}) = \mathbf{0}$, which implies $\|T(w_n)\|_2 \to \|\mathbf{0}\|_2 = 0$ by (7.3). But,

$$\|T(w_n)\|_2 = \left\| T\left(\frac{v_n}{\alpha_n \|v_n\|_1} \right) \right\|_2 = \frac{\|T(v_n)\|_2}{\alpha_n \|v_n\|_1} = 1,$$

which contradicts $\|T(w_n)\|_2 \to 0$. This contradiction proves that T is bounded. ■

Example 343 The application $T : C([0,1]) \to C([0,1])$ of Example 335 is continuous and is not linear. We prove that T is not bounded. If it were so, there would exist $K > 0$ such that:

$$\left\|f^2\right\|_\infty = \|T(f)\|_\infty \leq K \|f\|_\infty, \qquad \forall f \in C([0,1]).$$

Let $f_n(t) = e^{nt}$ for each $t \in [0,1]$. We have $\|f^2\|_\infty = e^{2n}$ and $\|f\|_\infty = e^n$, so that:

$$\frac{\|f_n^2\|_\infty}{\|f_n\|_\infty} = \frac{e^{2n}}{e^n} = e^n \to +\infty,$$

which proves that such a K cannot exist. It follows that T is not bounded. ▲

Example 344 Let $(\mathcal{P}([-1,1]), \|\cdot\|_\infty)$ be the normed vector space of the polynomials defined on the interval $[-1,1]$, with the norm $\|f\|_\infty = \max_{t\in[-1,1]} f(t)$. Let $L : \mathcal{P}([-1,1]) \to \mathbb{R}$ be the linear functional defined as $L(f) = f'(0)$ for each $f \in P([-1,1])$. Set $f_n(t) = 1 - n + nt$ for each $t \in [-1,1]$ and for each $n \geq 1$. We have $\|f_n\|_\infty = 1$ and $|L(f_n)| = |f'(0)| = n$, and therefore:

$$\frac{|L(f_n)|}{\|f_n\|_\infty} = n \to +\infty,$$

which shows that L is not bounded. By Theorem 342, the functional L is not continuous. Observe that the space $\mathcal{P}([-1,1])$ is infinite dimensional, and Theorem 359 will show that this is a crucial feature of this example. ▲

We conclude the study of the continuity with a characterization that holds for linear functionals. Recall from (3.12) that the kernel $\ker L$ of a functional $L : V \to \mathbb{R}$ is given by $\ker L = \{v \in V : L(v) = 0\}$, that is, $\ker L = L^{-1}(0)$.

Theorem 345 *A linear functional $L : V \to \mathbb{R}$ defined on a normed vector space is continuous if and only if $\ker L$ is a closed set.*

>From Chapter 3 we know that $\ker L$ is a vector subspace of V that plays an important role in the study of the finite dimensional vector spaces. Theorem 345 shows how the continuity of a linear functional is reflected in a simple topological property of such a subspace, i.e., in being a closed subset of V.

This result is false without linearity. For example, the function $f : \mathbb{R} \to \mathbb{R}$ given by:

$$f(x) = \begin{cases} x & \text{if } x \geq 0, \\ x - 1 & \text{if } x < 0, \end{cases}$$

is discontinuous at 0, though $\ker f = f^{-1}(0) = \{0\}$ is a singleton, and it is therefore a closed set.

Proof. We omit the proof of the "If." We prove the "Only if." Let $L : V \to \mathbb{R}$ be a continuous linear functional. Let $\{v_n\}_{n\geq 1}$ be a sequence of vectors of $\ker L$ such that $v_n \to v$. By Corollary 255, to prove that $\ker L$ is closed is sufficient to prove that $v \in \ker L$. Being L continuous, $L(v_n) \to L(v)$. On the other hand, $\{v_n\}_{n\geq 1} \subseteq \ker L$ implies $L(v_n) = 0$ for each n, and we conclude that $L(v) = 0$. It follows that $v \in \ker L$, as desired. ■

7.3 Topological Duals

In Definition 63 we defined dual space V' of a vector space V as the set of linear functionals defined on V. In particular, we saw that V' were themselves vector spaces. When V is normed, it becomes natural to consider the following subset of V'.

Definition 346 *The set of all linear and continuous functionals $L : V \to \mathbb{R}$ defined on a normed vector space V is called the **topological dual space** of V and is denoted by V^*.*

To distinguish it from the topological dual V^*, the space V' is often called the *algebraic dual* of V. By Proposition 304-(i), given $L_1, L_2 \in V^*$ we have $\alpha L_1 + \beta L_2 \in V^*$ for each $\alpha, \beta \in \mathbb{R}$. Therefore, the topological dual V^* is a vector subspace of the algebraic dual V'.

Similar considerations can be done for the vector space $L(V_1, V_2)$ of all linear applications $T : V_1 \to V_2$, introduced in Subsection 3.2. In this case we denote by $B(V_1, V_2)$ the vector subspace of $L(V_1, V_2)$ formed by all linear applications $T : V_1 \to V_2$ that are continuous. Naturally, when $V_2 = \mathbb{R}$ we have $B(V, \mathbb{R}) = V^*$.

For simplicity, forget for a moment that $B(V_1, V_2)$ is a vector subspace of $L(V_1, V_2)$, and think of it as a vector space on its own. The natural question to ask is whether $B(V_1, V_2)$ is a normed vector space. To this end, we first need to introduce a norm on $B(V_1, V_2)$. By Theorem 342, we know that for each continuous application $T : V_1 \to V_2$ there exists a scalar $K > 0$ such that $\|T(v)\|_2 \leq K \|v\|_1$ for each $v \in V_1$. Naturally, there exist many scalars that satisfy this inequality; in particular, if a scalar K satisfies it, this will be true also for all scalars K' such that $K' \geq K$. Among all these scalars, however, the one that really tells us how much the application is bounded is the smallest among them, i.e., the smallest scalar K for which the inequality $\|T(v)\|_2 \leq K \|v\|_1$ holds for each $v \in V_1$. We start by seeing that such a minimum actually exists.

Lemma 347 *The set:*

$$\{K \in \mathbb{R}_+ : \|T(v)\|_2 \leq K \|v\|_1, \qquad \forall v \in V_1\}$$

has a minimum.

Proof. Set $A = \{K \in \mathbb{R}_+ : \|T(v)\|_2 \leq K \|v\|_1, \forall v \in V_1\}$. Since A consists of positive numbers, is it obviously lower bounded and by the completeness of $\mathbb{R}$ it has infimum $\inf A$. On the other hand, it is easy to see that A is a closed subset of $\mathbb{R}$, which implies that $\inf A \in A$ (proceed, for instance, as in the proof of Theorem 307). Hence $\inf A$ is the minimum of A. ■

At this point we can introduce the norm of a linear application.

Definition 348 *Given a continuous linear application $T : V_1 \to V_2$ between normed vector spaces, its norm, denoted by $\|T\|$, is given by the quantity:*

$$\|T\| = \min \{K \in \mathbb{R}_+ : \|T(v)\|_2 \leq K \|v\|_1, \forall v \in V_1\}. \tag{7.5}$$

Notice that, by definition, we have:

$$\|T(v)\|_2 \leq \|T\| \|v\|_1, \qquad \forall v \in V_1, \tag{7.6}$$

an inequality that will turn out to be very useful.

Proposition 349 *Given a continuous linear application $T : V_1 \to V_2$ between normed vector spaces, the function $\|\cdot\| : B(V_1, V_2) \to \mathbb{R}$ defined in (7.5) is a norm. Moreover, we have:*[5]

$$\begin{aligned} \|T\| &= \sup\left\{\frac{\|T(v)\|_2}{\|v\|_1} : v \in V_1 \text{ with } v \neq \mathbf{0}\right\} \\ &= \sup\{\|T(v)\|_2 : v \in B_{V_1}\} \\ &= \sup\{\|T(v)\|_2 : v \in S_{V_1}\}. \end{aligned}$$

Proof. We start by proving that $\|\cdot\|$ is a norm. Obviously, $\|T\| \geq 0$ for each $T \in B(V_1, V_2)$. The neutral element of the vector space $B(V_1, V_2)$ is the null application $\mathbf{0} : V_1 \to V_2$. Clearly, $\|\mathbf{0}\| = 0$. Moreover, if $\|T\| = 0$, by (7.6) we have $\|T(v)\|_2 = 0$ for each $v \in V_1$. Therefore, $T(v) = \mathbf{0}$ for each $v \in V_1$, that is, T is the null application. Condition (i) of Definition 326 is therefore verified.

As to condition (ii), given any two applications T_1 and T_2 in $B(V_1, V_2)$, for each $v \in V_1$ we have:[6]

$$\begin{aligned} \|(T_1 + T_2)(v)\|_2 &= \|T_1(v) + T_2(v)\|_2 \leq \|T_1(v)\|_2 + \|T_2(v)\|_2 \\ &\leq \|T_1\| \|v\|_1 + \|T_2\| \|v\|_1 = (\|T_1\| + \|T_2\|) \|v\|_1, \end{aligned}$$

which implies $\|T_1 + T_2\| \leq \|T_1\| + \|T_2\|$.

Finally, for each $\alpha \in \mathbb{R}$ we have:[7]

$$\|\alpha T(v)\|_2 = |\alpha| \|T(v)\|_2 \leq |\alpha| \|T\|_2 \|v\|_1, \qquad \forall v \in V_1,$$

and hence $\|\alpha T\| \leq |\alpha| \|T\|$. In conclusion, $\|\cdot\|$ is a norm.

For each $v \neq \mathbf{0}$, (7.6) implies:

$$\frac{\|T(v)\|_2}{\|v\|_1} \leq \|T\|.$$

The following inequalities are therefore clear:

$$\begin{aligned} \|T\| &\geq \sup\left\{\frac{\|T(v)\|_2}{\|v\|_1} : v \in V_1 \text{ with } v \neq \mathbf{0}\right\} \geq \sup\{\|T(v)\|_2 : v \in B_{V_1}\} \\ &\geq \sup\{\|T(v)\|_2 : v \in S_{V_1}\}. \end{aligned}$$

To complete the proof it remains to prove that:

$$\|T\| \leq \sup\{\|T(v)\|_2 : v \in S_{V_1}\}. \tag{7.7}$$

[5] S_V denotes the unit sphere $\{v \in V : \|v\| = 1\}$.

[6] The second inequality follows from (7.6).

[7] The inequality follows from (7.6).

Set $\alpha = \sup\{\|T(v)\|_2 : v \in S_{V_1}\}$. For each $v \neq \mathbf{0}$, we have:

$$\left\| \frac{v}{\|v\|_1} \right\|_1 = \frac{1}{\|v\|_1} \|v\|_1 = 1.$$

Hence, for each $v \in V_1$ with $v \neq \mathbf{0}$ we have:

$$\left\| T\left(\frac{v}{\|v\|_1} \right) \right\|_2 \leq \alpha. \tag{7.8}$$

Since $T(\mathbf{0}) = \mathbf{0}$, (7.8) implies $\|T(v)\|_2 \leq \alpha \|v\|_1$ for each $v \in V_1$. Hence:

$$\alpha \in \{K \in \mathbb{R}_+ : \|T(v)\|_2 \leq K \|v\|_1, \forall v \in V_1\},$$

from which $\|T\| \leq \alpha$, as desired. ■

By this proposition, $(B(V_1, V_2), \|\cdot\|)$ is therefore a normed vector space.

Example 350 Consider $F : C([0,1]) \to \mathbb{R}$ defined as $F(f) = \int_0^1 f(t)\, dt$ for each $f \in C([0,1])$. It is a linear functional defined on the normed vector space $(C([0,1]), \|\cdot\|_\infty)$, and it is continuous for what seen in Example 294. Let us compute its norm $\|F\|$. Since $|f(t)| \leq \|f\|_\infty$ for each $t \in [0,1]$, we have:

$$|F(f)| = \left| \int_0^1 f(t)\, dt \right| \leq \int_0^1 |f(t)|\, dt \leq \|f\|_\infty,$$

which implies:

$$\|F\| = \sup\{|F(f)| : \|f\|_\infty = 1\} \leq \|f\|_\infty = 1.$$

On the other hand, consider the constant function $1_{[0,1]} : [0,1] \to \mathbb{R}$ such that $1_{[0,1]}(t) = 1$ for each $t \in [0,1]$. We have $\left\|1_{[0,1]}\right\|_\infty = 1$ and:

$$|F(g)| = \left| \int_0^1 1_{[0,1]}(t)\, dt \right| = 1.$$

Hence, $1 = |F(g)| \leq \|F\| \leq 1$, and we conclude that $\|F\| = 1$. ▲

Example 351 Let $(V_1, \|\cdot\|_1) = (V_2, \|\cdot\|_2) = (C([0,1]), \|\cdot\|_\infty)$, and define the application $T : C([0,1]) \to C([0,1])$ as:

$$T(f)(s) = \int_0^1 (h_1(t) + h_2(s)) f(t)\, dt, \qquad \forall s \in [0,1],$$

for each $f \in C([0,1])$, where h_1 and h_2 are any two positive functions in $C([0,1])$. For example, if $f(t) = t$, $h_1(t) = e^t$ and $h_2(s) = s^2$, then:[8]

$$T(f)(s) = \int_0^1 te^t dt + \int_0^1 s^2 e^t dt = (e-1)s^2, \qquad \forall s \in [0,1].$$

[8]See Ambrosetti and Musu (1988) p. 349.

We have:

$$\begin{aligned}
\|T(f)\|_\infty &= \max_{s\in[0,1]}\left|\int_0^1 (h_1(t)+h_2(s))\,f(t)\,dt\right| \leq \max_{s\in[0,1]}\int_0^1 |(h_1(t)+h_2(s))\,f(t)|\,dt \\
&= \max_{s\in[0,1]}\int_0^1 (h_1(t)+h_2(s))\,|f(t)|\,dt \\
&= \int_0^1 h_1(t)\,|f(t)|\,dt + \max_{s\in[0,1]}\int_0^1 h_2(s)\,|f(t)|\,dt \\
&\leq \|f\|_\infty\left(\int_0^1 h_1(t)\,dt + \max_{s\in[0,1]} h_2(s)\right),
\end{aligned}$$

which implies:

$$\|T\| = \sup\{\|T(f)\|_\infty : \|f\|_\infty = 1\} \leq \int_0^1 h_1(t)\,dt + \max_{s\in[0,1]} h_2(s).$$

On the other hand, consider the constant function $1_{[0,1]}$. We have:

$$\begin{aligned}
\left\|T\left(1_{[0,1]}\right)\right\|_\infty &= \max_{s\in[0,1]}\left|\int_0^1 (h_1(t)+h_2(s))\,dt\right| = \max_{s\in[0,1]}\int_0^1 (h_1(t)+h_2(s))\,dt \\
&= \int_0^1 h_1(t)\,dt + \max_{s\in[0,1]} h_2(s),
\end{aligned}$$

and so:

$$\|T\| = \int_0^1 h_1(t)\,dt + \max_{s\in[0,1]} h_2(s).$$

▲

We saw that $(B(V_1,V_2),\|\cdot\|)$ is a normed vector space. Since completeness is a central property for normed vector spaces, it is therefore natural to ask under what conditions the space $(B(V_1,V_2),\|\cdot\|)$ is complete, i.e., when it is a Banach space. A natural conjecture is that this can happen when both spaces V_1 and V_2 are themselves Banach. The next result (whose proof we omit) shows that, surprisingly, it is sufficient that V_2 be a Banach space in order for $(B(V_1,V_2),\|\cdot\|)$ to be also a Banach space, independently of whether or not V_1 is a Banach space.

Theorem 352 *If V_2 is a Banach space, also the normed vector space $(B(V_1,V_2),\|\cdot\|)$ is a Banach space.*

Since $V_2 = \mathbb{R}$ is obviously a Banach space, we have the following important corollary.

Corollary 353 *The topological dual space $(V^*,\|\cdot\|)$ is a Banach space.*

7.4 Intermezzo: Homeomorphisms and Isometries

In Theorem 299 we saw that, given a continuous function $f : X \to Y$ between two metric spaces, the counterimages $f^{-1}(G)$ of each open set G of Y are themselves open sets of X, while the images $f(G)$ of open sets G of X are in general not open sets of Y. Next definition introduces a class of functions for which this is always true.

Definition 354 *A continuous and surjective function $f : X \to Y$ between two metric spaces is called homeomorphism if it is injective and if its inverse function $f^{-1} : Y \to X$ is continuous. When such a function exists, the spaces X and Y are said to be homeomorphic.*

Since f is in turn the inverse function of f^{-1}, i.e., $f = (f^{-1})^{-1}$, when f is an homeomorphism we have that $f(G)$ is an open set. Specifically, let τ_X and τ_Y be the collections of all the open sets of the metric spaces X and Y. Such collections are called topologies of X and Y, respectively.

Proposition 355 *Let $f : X \to Y$ be an homeomorphism between the two metric spaces X and Y. Then,*

$$\tau_Y = \{f(G) : G \in \tau_X\} \qquad \text{and} \qquad \tau_X = \left\{f^{-1}(V) : V \in \tau_Y\right\}.$$

Proof. It is sufficient to prove that $\tau_Y = \{f(G) : G \in \tau_X\}$, the other equality being specular, with f^{-1} in place of f. Clearly, $\{f(G) : G \in \tau_X\} \subseteq \tau_Y$. Let $V \in \tau_Y$. We have $f^{-1}(V) \in \tau_X$, and therefore:

$$V = f\left(f^{-1}(V)\right) \in \{f(G) : G \in \tau_X\},$$

as desired. ■

The homeomorphism f preserves therefore the open sets between the two spaces X and Y. Once we know the topology τ_X, through the function f we can determine also the topology τ_Y, and viceversa. In this sense, the topologies of two homeomorphic metric spaces can be seen as equivalent.

As seen in Chapter 3, a linear and bijective application $T : V_1 \to V_2$ between two normed vector spaces is called isomorphism. It becomes an homeomorphism when both T and T^{-1} are continuous. Next lemma gives a necessary and sufficient condition for an isomorphism to be also an homeomorphism.

Lemma 356 *A surjective linear application $T : V_1 \to V_2$ between two vector spaces V_1 and V_2 is both an isomorphism and an homeomorphism if and only if there exist $c_1, c_2 > 0$ such that:*

$$c_1 \|v\|_1 \leq \|T(v)\|_2 \leq c_2 \|v\|_1, \qquad \forall v \in V_1. \tag{7.9}$$

Proof. If $V_1 = \{\mathbf{0}\}$, the result is trivially true since $\|T(v)\|_2 = \|v\| = 0$ for each $v \in V_1$.

Assume therefore that $V_1 \neq \{\mathbf{0}\}$. Suppose that T is both an isomorphism and an homeomorphism. It follows that $V_2 \neq \{\mathbf{0}\}$, and therefore $\|T\| > 0$ and $\|T^{-1}\| > 0$ by Proposition 349. Moreover, being T linear and continuous, by (7.6) we have:

$$\|T(v)\|_2 \leq \|T\| \|v\|_1, \qquad \forall v \in V_1. \tag{7.10}$$

Moreover, being also T^{-1} linear and continuous, we also have:

$$\left\|T^{-1}(T(v))\right\|_1 \leq \left\|T^{-1}\right\| \|T(v)\|_2, \qquad \forall v \in V_1,$$

that is,

$$\frac{\|v\|_1}{\|T^{-1}\|} \leq \|T(v)\|_2, \qquad \forall v \in V_1. \tag{7.11}$$

Thanks to (7.10) and (7.11), there exist therefore $c_1, c_2 > 0$ such that (7.9) holds.

As to the converse, assume that (7.9) holds. We first prove that T is an isomorphism. Since by hypothesis T is surjective, it is necessary to prove that it is also injective. By Proposition 106, this is equivalent to prove that $\ker T = \{\mathbf{0}\}$. Let $v \neq \mathbf{0}$. We have $\|v\|_1 > 0$ and hence (7.9) implies $\|T(v)\|_2 \geq c_1 \|v\|_1 > 0$. It follows that $T(v) \neq \mathbf{0}$, and so $v \notin \ker T$. We therefore conclude that $\ker T = \{\mathbf{0}\}$, as desired.

To complete the proof it is necessary to show that T and T^{-1} are both continuous. Let $\{v^n\}_{n\geq 1}$ be a sequence in V_1 such that $v^n \to v \in V_1$. By (7.9) we have:

$$\|T(v^n) - T(v)\|_2 = \|T(v^n - v)\|_2 \leq c_2 \|v^n - v\|_1 \to 0,$$

and therefore $T(v^n) \to T(v)$, which proves that T is continuous. Let now $\{w^n\}_{n\geq 1}$ be a sequence in V_2 such that $w^n \to w \in V_2$. By (7.9) we have:

$$\begin{aligned} \left\|T^{-1}(w^n) - T^{-1}(w)\right\|_1 &= \left\|T^{-1}(w^n - w)\right\|_1 \leq \frac{1}{c_1} \left\|T\left(T^{-1}(w^n - w)\right)\right\|_2 \\ &= \frac{1}{c_1} \|w^n - w\|_2 \to 0, \end{aligned}$$

and therefore $T^{-1}(w^n) \to T^{-1}(w)$. This implies that also T^{-1} is continuous. ■

Consider the case $V \equiv V_1 = V_2$, and suppose that $\|\cdot\|_1$ and $\|\cdot\|_2$ are two norms defined on V. Let τ_1 and τ_2 be the topologies induced on V by the norms $\|\cdot\|_1$ and $\|\cdot\|_2$, respectively.

If the isomorphism $I : V \to V$ given by the identity application is an homeomorphism between $(V, \|\cdot\|_1)$ and $(V, \|\cdot\|_2)$, by Proposition 355 we have:

$$\tau_2 = \{I(G) : G \in \tau_1\} = \{G : G \in \tau_1\},$$

that is, $\tau_1 = \tau_2$. In other words, the two norms $\|\cdot\|_1$ and $\|\cdot\|_2$, though different, induce the same topology on the vector space V.

On the other hand, by Lemma 356 the isomorphism I is an homeomorphism if and only if there exist $c_1, c_2 > 0$ such that:

$$c_1 \|v\|_1 \leq \|v\|_2 \leq c_2 \|v\|_1, \qquad \forall v \in V.$$

This motivates the following definition.

Definition 357 *Two norms $\|\cdot\|_1$ and $\|\cdot\|_2$ on a vector space V are called **equivalent** if there exist $c_1, c_2 > 0$ such that:*

$$c_1 \|v\|_1 \leq \|v\|_2 \leq c_2 \|v\|_1, \qquad \forall v \in V.$$

Two equivalent norms induce therefore the same topology on the vector space V. This means, for example, that a function is continuous with respect to the norm $\|\cdot\|_1$ if and only if it is so with respect to $\|\cdot\|_2$, and a sequence $\{v^n\}_{n\geq 1}$ converges to v according to the norm $\|\cdot\|_1$ if and only if it converges also according to $\|\cdot\|_2$. Hence, once verified if a function or a sequence satisfies these properties with respect to one norm, this will be true with respect to all its equivalent norms.

We close this Intermezzo, with a last notion. An isomorphism $T : V_1 \to V_2$ between two normed vector spaces V_1 and V_2 is called *isometry* if $\|T(v)\|_2 = \|v\|_1$ for each $v \in V_1$. An isometry therefore preserves the norms between two metric spaces. As immediate consequence of Lemma 356 we have that an isometry between two normed vector spaces V_1 and V_2 is also an homeomorphism between such spaces.

7.5 Finite Dimensional Spaces

Finite dimensional vector spaces are a fundamental class of vector spaces, which we studied in detail in Chapters 1-3. Here we will see that finite dimensional normed vector spaces enjoy important properties.

We begin with a fundamental lemma, whose proof we omit.

Lemma 358 *Let $\{v^i\}_{i=1}^n$ be a set of linearly independent vectors of a normed vector space V. Then, there exists a constant $c > 0$ such that, for each collection of scalars $\{\alpha_i\}_{i=1}^n$, we have:*

$$\sum_{i=1}^{n} |\alpha_i| \leq c \left\| \sum_{i=1}^{n} \alpha_i v^i \right\|.$$

The first important consequence of this lemma is the continuity of linear applications defined on finite dimensional spaces.

Theorem 359 *Let V_1 and V_2 be two normed vector spaces such that V_1 is finite dimensional. Then, each linear application $T : V_1 \to V_2$ is continuous.*

In other words, when V_1 is finite dimensional we have $B(V_1, V_2) = L(V_1, V_2)$. As Example 344 shows, the finite dimensionality of V_1 is crucial for the validity of Theorem 359, which is in general false when V_1 is infinite dimensional.

Proof Let $\{v^i\}_{i=1}^n$ be a basis of the space V_1. For each $v \in V_1$ there exists a collection of scalars $\{\alpha_i\}_{i=1}^n$ such that $v = \sum_{i=1}^n \alpha_i v^i$. Hence, by Lemma 358 we have:

$$\begin{aligned} \|T(v)\|_2 &= \left\| T\left(\sum_{i=1}^n \alpha_i v^i\right) \right\|_2 = \left\| \sum_{i=1}^n \alpha_i T\left(v^i\right) \right\|_2 \leq \sum_{i=1}^n |\alpha_i| \left\| T\left(v^i\right) \right\|_2 \\ &\leq c \left\| \sum_{i=1}^n \alpha_i v^i \right\|_1 \left(\max_{i=1,\dots n} \left\| T\left(v^i\right) \right\|_2 \right) = c \left(\max_{i=1,\dots n} \left\| T\left(v^i\right) \right\|_2 \right) \|v\|_1 . \end{aligned}$$

Setting $K = c\left(\max_{i=1,\dots n} \|T(v^i)\|_2\right)$, we conclude that for each $v \in V$ we have $\|T(v)\|_2 \leq K \|v\|_1$. Hence, T is a bounded application and, by Theorem 342, it is continuous. ■

Corollary 360 *Each linear functional $L : V \to \mathbb{R}$ defined on a finite dimensional normed vector space V is continuous.*

Hence, the duals V' and V^* coincide when V is finite dimensional. For example, by Theorem 65 we have that $\mathbb{R}^n$ is both the algebraic and the topological dual of $\mathbb{R}^n$.

In Theorem 104 we saw that two finite dimensional spaces are isomorphic if and only if they have the same dimension. Next we show that, in the case of normed vector spaces, such isomorphism is also an homeomorphism. Hence, finite dimensional spaces that have the same dimension have also similar topological structures.

Theorem 361 *Two finite dimensional normed vector spaces that have the same dimension are homeomorphic.*

Proof Let V_1 and V_2 be two finite dimensional normed vector spaces, with $\dim V_1 = \dim V_2$. By Theorem 104, there exists an isomorphism $T : V_1 \to V_2$. By Theorem 359, T is continuous. Similarly, also the inverse application $T^{-1} : V_2 \to V_1$ is continuous. It follows that T is an homeomorphism, as desired. ■

A trivial, but legitimate, case of two normed vector spaces that have the same dimension is given by $(V, \|\cdot\|_1)$ and $(V, \|\cdot\|_2)$, where $\|\cdot\|_1$ and $\|\cdot\|_2$ are two norms defined on the same vector space V. This simple observation leads to the following important consequence of Theorem 361.

Corollary 362 *In a finite dimensional vector space all norms are equivalent.*

Proof. Let $\|\cdot\|_1$ and $\|\cdot\|_2$ be two norms defined on a finite dimensional space V. The identity application $I : V \to V$ is an isomorphism between the normed vector spaces $(V, \|\cdot\|_1)$ and $(V, \|\cdot\|_2)$. By Theorem 361, I is an homeomorphism. By Lemma 356, there exist $c_1, c_2 > 0$ such that:

$$c_1 \|v\|_1 \leq \|v\|_2 \leq c_2 \|v\|_1 , \qquad \forall v \in V.$$

Hence, the two norms are equivalent. ■

Corollary 362 is an important principle of order. Though in a finite dimensional space it is possible to introduce several norms, very different among them, Corollary 362 guarantees that they are topologically equivalent. For example, the norms $\|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|_\infty$ defined on $\mathbb{R}^n$ are equivalent and so they induce the same topology on $\mathbb{R}^n$, i.e., the same collection of open sets.

>From Corollary 105 we know that all vector spaces of dimension n are isomorphic to $\mathbb{R}^n$. Due to Corollary 362, they are also homeomorphic to $\mathbb{R}^n$. Since the latter space is Banach, we have the following result.

Corollary 363 *Each finite dimensional normed vector space is a Banach space.*

Proof Let $(V, \|\cdot\|)$ be a finite dimensional normed vector space, and let $\mathbb{R}^n$ be endowed with its Euclidean norm $\|\cdot\|_2$. By Corollary 105 there exists an homeomorphism $T : V \to \mathbb{R}^n$, which Corollary 362 guarantees to be also an homeomorphism between $(V, \|\cdot\|)$ and $(\mathbb{R}^n, \|\cdot\|_2)$. By Lemma 356 there exists therefore a constant $c > 0$ such that $\|T(v)\|_2 \leq c\|v\|$ for each $v \in V$.

Let $\{v^n\}_{n\geq 1}$ be a Cauchy sequence in V. To show that V is a Banach space we have to prove that $\{v^n\}_{n\geq 1}$ is a convergent sequence. For each $m, n \geq 1$ we have:

$$\|T(v^n) - T(v^m)\|_2 = \|T(v^n - v^m)\|_2 \leq c\|v^n - v^m\|$$

and hence the sequence $\{\geq\}_{n\geq 1}$ is a Cauchy sequence in $(\mathbb{R}^n, \|\cdot\|_2)$. Since this space is a Banach space, there exists $x \in \mathbb{R}^n$ such that $T(v^n) \overset{\|\cdot\|_2}{\to} x$. Since T^{-1} is a continuous application, we have:

$$v^n = T^{-1}(T(v^n)) \overset{\|\cdot\|}{\to} T^{-1}(x),$$

and so $\{v^n\}_{n\geq 1}$ converges to $T^{-1}(x)$. ■

We close this study of finite dimensional normed vector spaces with a surprising characterization through the notion of compactness. In Section 6.4 we saw that compact sets are closed and bounded, but how the converse is in general false, with the remarkable exception of $\mathbb{R}^n$ thanks to the Heine-Borel Theorem 272. This motivates

the following terminology: we say that a normed vector space enjoys the *Heine-Borel property* when all closed and bounded sets are compact. Clearly, $\mathbb{R}^n$ enjoys such property.

The next deep result, proved in 1918 by Frederic Riesz, shows that the Heine-Borel property actually characterizes finite dimensional normed vector spaces. Moreover, the result shows that the compactness of an "elementary" closed and bounded set, i.e., the closed unit ball, is per se equivalent to the finite dimensionality of the space.

Theorem 364 *Given a normed vector space V, the following properties are equivalent:*

(i) V is finite dimensional;

(ii) V enjoys the Heine-Borel property;

(iii) the closed unit ball $B_V = \{v : \|v\| \leq 1\}$ is compact.

In other words, in infinite dimensional normed vector spaces the closed and bounded sets are not in general compact. In particular, it is not compact the closed unit ball, nor, more generally, any set having interior points.

Corollary 365 *In an infinite dimensional vector space, the sets with interior points are not compact.*

Proof Let A be a subset with $\overset{\circ}{A} \neq \emptyset$ of an infinite dimensional normed vector space V. Suppose that A is compact, and let $x \in \overset{\circ}{A}$. There exists therefore a neighborhood $B_\varepsilon(x)$ of x such that $B_\varepsilon(x) \subseteq A$. Since A is compact, it is also closed by Theorem 271; hence, $\overline{B}_\varepsilon(x) \subseteq A$. It follows that $\overline{B}_\varepsilon(\mathbf{0}) = \overline{B}_\varepsilon(x) - x \subseteq A - x$. It is easy to see that also $A - x$ is compact, and therefore its closed subset $\overline{B}_\varepsilon(\mathbf{0})$ is also compact by Proposition 274. In turn, this implies that the closed unit ball B_V is compact (why?), and hence the space V is finite dimensional by Theorem 364. This contradiction shows that A cannot be compact. ■

All this shows that in infinite dimensional spaces compactness with respect to the metric induced by the norms is a very strong property, enjoyed by relatively few sets. Hence, such compactness is a not very useful notion in studying sets in infinite dimensional space, while it plays a fundamental role in finite dimensional spaces.

7.6 Some Classical Spaces

We introduce some classic normed vector spaces and study some of their properties.

7.6.1 Bounded Functions

Given any set X, consider the vector space $B(X)$ of all bounded functions $f : X \to \mathbb{R}$, endowed with the supnorm $\|f\|_\infty = \sup_{x\in X} |f(x)|$.

Proposition 366 *$(B(X), \|\cdot\|_\infty)$ is a Banach space.*

Proof Consider a Cauchy sequence $\{f_n\}_n \subseteq B(X)$. Fix $x \in X$. From the obvious inequality

$$|f_n(x) - f_m(x)| \leq \|f_n - f_m\|_\infty,$$

it follows that the sequence $\{f_n(x)\}_n \subseteq \mathbb{R}$ is Cauchy. Hence, this sequence has a limit point, denoted $f(x)$. In this way we define a function $f : X \to \mathbb{R}$, with $\lim_n f_n(x) = f(x)$ for all $x \in X$. Given any $\varepsilon > 0$, there is $n_0 \geq 1$ such that

$$|f_n(x) - f_m(x)| \leq \|f_n - f_m\|_\infty \leq \varepsilon, \qquad \forall n, m \geq n_0.$$

Letting $m \to \infty$, we get $|f_n(x) - f(x)| \leq \varepsilon$. This implies that $\|f_n - f\|_\infty \leq \varepsilon$ and thus f is the uniform limit of the sequence. Hence, $f \in B(X)$. For, take $\varepsilon = 1$ and let n be such that $\|f_n - f\|_\infty \leq 1$. Then

$$|f(x)| \leq |f(x) - f_n(x)| + |f_n(x)| \leq 1 + \|f_n\|_\infty$$

and f is bounded, i.e., $f \in B(X)$. ■

7.6.2 Continuous Functions

Given any metric space X, consider the vector space $C_b(X)$ of all bounded functions $f : X \to \mathbb{R}$, endowed with the supnorm $\|f\|_\infty = \sup_{x\in X} |f(x)|$.

Proposition 367 *$(C_b(X), \|\cdot\|_\infty)$ is a Banach space.*

The proof relies on the following very important property, which shows that continuity is preserved by uniform convergence (i.e., convergence under the supnorm).

Lemma 368 *Let $\{f_n\}_n \subseteq C(X)$ be such that $f_n \xrightarrow{\|\cdot\|_\infty} f$ for some $f : X \to \mathbb{R}$. Then, $f \in C(X)$.*

Proof Let $\{x_m\}_m \subseteq X$ be such that $x_m \to x \in X$. Fix $\varepsilon > 0$. Let n_ε be such that $\|f_{n_\varepsilon} - f\|_\infty \leq \varepsilon/3$, and, being f_{n_ε} continuous, let m_ε be such that $|f_{n_\varepsilon}(x_m) - f_{n_\varepsilon}(x)| \leq \varepsilon/3$ for all $m \geq m_\varepsilon$. Then,

$$\begin{aligned} |f(x_m) - f(x)| &\leq |f(x_m) - f_{n_\varepsilon}(x_m)| + |f_{n_\varepsilon}(x_m) - f_{n_\varepsilon}(x)| + |f_{n_\varepsilon}(x) - f_{n_\varepsilon}(x)| \\ &\leq \|f_{n_\varepsilon} - f\|_\infty + |f_{n_\varepsilon}(x_m) - f(x)| + \|f_{n_\varepsilon} - f\|_\infty \leq \varepsilon \end{aligned}$$

for all $m \geq m_\varepsilon$. This shows that $f \in C(X)$. ■

Proof of Proposition 367. In view of Proposition 366, it is enough to show that $C(X)$ is a closed subset of $B(X)$. Let $\{f_n\}_n \subseteq C(X)$ be such that $\|f_n - f\|_\infty \to 0$ for some $f \in B(X)$. By Lemma 368, $f \in C(X)$. Hence, by Corollary 255 $C(X)$ is a closed subset of $B(X)$. ■

Notice that by the Weierstrass Theorem we have $C(X) = C_b(X)$ when the metric space X is compact. We thus have the following simple but important corollary of Proposition 367.

Corollary 369 $(C(X), \|\cdot\|_\infty)$ *is a Banach space if* X *is a compact metric space.*

We now characterize the compact subsets of the Banach space $(C_b(X), \|\cdot\|_\infty)$. To this end, say that a subset $A \subseteq C(X)$ is *equicontinuous* if, given any $\varepsilon > 0$, there is $\delta_\varepsilon > 0$ such that

$$d(x,y) < \delta_\varepsilon \Longrightarrow |f(x) - f(y)| < \varepsilon, \qquad \forall x, y \in X, \forall f \in A.$$

In other words, the collection A of continuous functions if equicontinuous if, given any $\varepsilon > 0$, they all share the same δ_ε.

We can now state and prove the classic Ascoli-Arzelà Theorem, due to Arzelà (1882-1883) and (1895) and Ascoli (1883-1884), which characterizes compact subsets of the Banach space $(C(X), \|\cdot\|_\infty)$, when X is compact.

Theorem 370 (Ascoli-Arzelà) *Let* X *be a compact metric space. A subset* H *of* $C(X)$ *is relatively compact (i.e., its closure is a compact set) in the Banach space* $(C(X), \|\cdot\|_\infty)$ *if and only it is uniformly bounded*[9] *and equicontinuous.*

Proof. *Necessity.* If H is relatively compact then is totally bounded. This implies in turn that it is norm bounded. Let us show that the family of functions H is equicontinuous. Fix $\varepsilon > 0$. There is a finite number of functions $\{f_i\}_{i=1}^n \subseteq H$ such that $f \in H$ implies $\|f - f_{i_0}\|_\infty \leq \varepsilon/3$ for some index $i_0 \in \{1, .., n\}$. Fix $x \in X$. There is a neighborhood $B_\eta(x)$ such that $|f_i(x) - f_i(y)| \leq \varepsilon/3$ for all $y \in B_\eta(x)$ and every $i \in \{1, 2, .., n\}$. Consequently, for any $f \in H$ and any $y \in B_\eta(x)$ we have

$$\begin{aligned} |f(y) - f(x)| &= |f(y) - f_{i_0}(y) + f_{i_0}(y) - f_{i_0}(x) + f_{i_0}(x) - f(x)| \\ &\leq |f(y) - f_{i_0}(y)| + |f_{i_0}(y) - f_{i_0}(x)| + |f_{i_0}(x) - f(x)| \\ &\leq \varepsilon/3 + \varepsilon/3 + \varepsilon/3 = \varepsilon, \end{aligned}$$

[9] That is, bounded with respect to the supnorm: there is $K > 0$ such that $H \subseteq \{f \in C(X) : \|f\|_\infty \leq K\}$.

and thus H is equicontinuous.

Sufficiency. Assume that H is equicontinuous. For any $x \in X$, let $B_\eta(x)$ be a neighborhood of x such that $|f(x) - f(y)| \leq \varepsilon/4$ for all $y \in B_\eta(x)$ and $f \in H$. As X is compact, there is a finite number of points $\{x_i\}_{i=1}^n \subseteq X$ such that the neighborhoods $B_{\eta_i}(x_i)$ cover X. Since H is norm bounded, say $\|f\| \leq K$ for all $f \in H$, we have $|f(x)| \leq K$, i.e., $f(x) \in [-K, K]$ for all $f \in H$ and $x \in X$. We can hence construct a finite sequence of points $\{c_j\}_{j=1}^m \subseteq [-K, K]$ so that if $f \in H$ and $x \in X$, $|f(x) - c_{j_0}| \leq \varepsilon/4$ for some c_{j_0}. Consider any map $\varphi : \{1, 2, ..., n\} \to \{1, 2, ..., m\}$ and, associated with φ, the subset

$$L_\varphi = \left\{ f \in H : \left| f(x_i) - c_{\varphi(i)} \right| \leq \varepsilon/4 \right\}.$$

Clearly, some L_φ may be empty but, by construction, $H = \cup_\varphi L_\varphi$, where the union is made over all the finite number of maps $\varphi : \{1, 2, ..., n\} \to \{1, 2, ..., m\}$. We claim that the diameter of each nonempty set L_φ is less than ε. Actually, if $f, g \in L_\varphi$, for all i we have

$$\begin{aligned} |f(x_i) - g(x_i)| &= \left| f(x_i) - c_{\varphi(i)} + c_{\varphi(i)} - g(x_i) \right| \\ &\leq \left| f(x_i) - c_{\varphi(i)} \right| + \left| c_{\varphi(i)} - g(x_i) \right| \\ &\leq \varepsilon/4 + \varepsilon/4 = \varepsilon/2. \end{aligned}$$

Consequently, for any $x \in X$, if $x \in B_{\eta_i}(x_i)$, it follows

$$\begin{aligned} |f(x) - g(x)| &= |f(x) - f(x_i) + f(x_i) - g(x_i) + g(x_i) - g(x)| \\ &\leq |f(x) - f(x_i)| + |f(x_i) - g(x_i)| + |g(x_i) - g(x)| \\ &\leq \varepsilon/4 + \varepsilon/2 + \varepsilon/4 = \varepsilon. \end{aligned}$$

Hence, $f, g \in L_\varphi$ implies $\|f - g\| \leq \varepsilon$ and $diam(L_\varphi)$. Summarizing, for every $\varepsilon > 0$ the set H can be covered by a finite number of sets with diameter less than ε. Hence H is totally bounded. As $C(X)$ is complete, H is relatively compact, as desired. ■

An immediate consequence of this classic result is that a subset A of $C(X)$ is compact in $(C(X), \|\cdot\|_\infty)$ if and only it is bounded, closed, and equicontinuous.

Example 371 Given $K > 0$, let A be a subset of $C([a, b])$ such that

$$|f(x) - f(y)| \leq K|x - y|, \qquad \forall x, y \in [a, b],$$

for all $f \in A$. The set A is easily seen to be uniformly bounded and equicontinuous (why?). By the Ascoli-Arzelà Theorem, the set A is relatively compact in the Banach space $(C(X), \|\cdot\|_\infty)$. ▲

7.6.3 Differentiable Functions

Given a closed and bounded interval $[a,b]$, consider the vector space $C^1([a,b])$ of the continuously differentiable functions $f:[a,b]\to\mathbb{R}$. Endow $C^1([a,b])$ with the norm $\|\cdot\|_1$ given by

$$\|f\|_1 = \max\{\|f\|_\infty, \|f'\|_\infty\}.$$

In other words, $\|f\|_1$ is the largest value among the supnorms of f and of its derivative f'.

Proposition 372 *$(C^1([a,b]), \|\cdot\|_1)$ is a Banach space.*

Proof. Let us first observe that if (g_n) is a sequence of continuous functions uniformly converging on $[a,b]$ to a function g, then the sequence of their primitives

$$\varphi_n(t) = \int_a^t g_n(s)\,ds,\ t\in[a,b]$$

converges uniformly to $\varphi(t) = \int_a^t g(s)\,ds$. Actually,

$$\begin{aligned}|\varphi(t)-\varphi_n(t)| &= \left|\int_a^t [g(s)-g_n(s)]\,ds\right| \le \int_a^t |g(s)-g_n(s)|\,ds\\ &\le \int_a^t \|g-g_n\|_\infty\,ds = \|g-g_n\|_\infty \int_a^t ds \le (b-a)\,\|g-g_n\|_\infty.\end{aligned}$$

Hence

$$\|\varphi-\varphi_n\|_\infty \le (b-a)\,\|g-g_n\|_\infty,$$

which is the desired result.

Consider now the space $C^1([a,b])$. It is clearly a vector space and $\|f\|_1$ is a norm on this space. It remains to prove that $C^1([a,b])$ is complete when endowed with the norm $\|f\|_1$.

Let $\{f_n\}\subseteq C^1([a,b])$ be a Cauchy sequence. We have $\|f_n-f_m\|_1\le\varepsilon$ for $n,m\ge n_0$. This implies $\|f_n-f_m\|_\infty\le\varepsilon$ as well as $\|f_n'-f_m'\|_\infty\le\varepsilon$. Hence, the two sequences $\{f_n\}$ and $\{f_n'\}$ are Cauchy in $C([a,b],\|\cdot\|_\infty)$. Completeness of this space implies that $f_n\to f$ and $f_n'\to g$, uniformly over $[a,b]$ and with $f,g\in C([a,b],\|\cdot\|_\infty)$. We must show that $f'=g$. On the other hand, we have

$$f_n(t) = f_n(a) + \int_a^t f_n'(s)\,ds. \tag{7.12}$$

We have seen that $\int_a^t f_n'(s)\to\int_a^t g(s)$ uniformly. In particular it converges point-wise. Likewise, $f_n(a)\to f(a)$, and $f_n(t)\to f(t)$. Therefore, taking limit in (7.12), we get

$$f(t) = f(a) + \int_a^t g(s)\,ds$$

for all t. As g is continuous, this leads to $f'(t) = g(t)$ for all $t \in [a,b]$. We conclude that $f_n \to f$ and $f_n' \to f'$ uniformly. Namely, $\|f_n - f\|_\infty \to 0$ and $\|f_n' - f'\|_\infty \to 0$. As the function $(r,s) \to Max\{r,s\}$ is continuous, it follows that $\|f_n - f\|_1 = \max\{\|f_n - f\|_\infty, \|f_n' - f'\|_\infty\} \to 0$. ■

It should be noticed that there are several norms equivalent to $\|f\|_1$ in $C^1([a,b])$.. For instance:

$$\begin{aligned} \|f\| &= \|f\|_\infty + \|f'\|_\infty \\ \|f\| &= |f(a)| + \|f'\|_\infty \\ \|f\| &= Max\{|f(a)|, \|f'\|_\infty\}. \end{aligned}$$

Let us check the last one. Clearly,

$$Max\{|f(a)|, \|f'\|_\infty\} \leq Max\{\|f\|_\infty, \|f'\|_\infty\}.$$

On the other hand, from

$$f(t) = f(a) + \int_a^t f'(s)\, ds$$

we obtain easily

$$\|f\|_\infty \leq |f(a)| + (b-a)\|f'\|_\infty.$$

Hence,

$$\begin{aligned} Max\{\|f\|_\infty, \|f'\|_\infty\} &\leq Max\{|f(a)| + (b-a)\|f'\|_\infty, \|f'\|_\infty\} \\ &\leq (1+b-a)\, Max\{|f(a)|, \|f'\|_\infty\} \end{aligned}$$

that proves the claim.

The set $C^1([a,b])$ is a subset of $C([a,b])$. The next result shows that its unit ball is actually relatively compact (i.e., its closure is compact) when viewed as a subset of the Banach space $(C([a,b]), \|\cdot\|_\infty)$. We thus have a "concrete" example of a relatively compact subset of $(C([a,b]), \|\cdot\|_\infty)$.

Proposition 373 *The unit ball* $\{f \in C^1([a,b]) : \|f\|_1 \leq 1\}$ *is relatively compact in* $(C([a,b]), \|\cdot\|_\infty)$.

Proof Set $B = \{f \in C([a,b]) : \|f\|_1 \leq 1\}$. If $f \in B$, by Example 419 f is Lipschitz with

$$|f(x) - f(y)| \leq \|f\|_1 |x-y| \leq |x-y|, \qquad \forall x, y \in [a,b].$$

Hence, B is equicontinuous. It is also bounded since $B \subseteq \{f \in C([a,b]) : \|f\|_\infty \leq 1\}$. By the Ascoli-Arzelà Theorem 370, B is relatively compact in $(C([a,b]), \|\cdot\|_\infty)$. ■

7.7 Differentiability

The notions of directional derivative and Gateaux differentiability studied in Chapter 4 can be naturally extended to applications $f : V \to W$ among normed vector spaces. To ease notation, we focus on the important special case $W = \mathbb{R}$, that is, on functionals $f : V \to \mathbb{R}$. This is also the case that is more relevant for the rest of these lecture notes, and we leave to the reader the more general case $f : V \to W$, which can be studied along the same lines.

We first give the versions for functionals defined on normed vector spaces of Definitions 129 and 139.

Let A be an open subset of a normed vector space V. Given a functional $f : A \to \mathbb{R}$, the *derivative of f at $v \in A$ in the direction $w \in V$* is given by

$$f'(v; w) = \lim_{t \to 0+} \frac{f(v + tw) - f(v)}{t}, \tag{7.13}$$

when such limit exists finite. Fixed $v \in V$, the function $f'(v; \cdot) : D \subseteq V \to \mathbb{R}$ is the *directional derivative* of f at v. Its domain D is the set of the directions along which the limit (7.13) exists finite.

The functional $f : A \to \mathbb{R}$ is called the *Gateaux differentiable* at $v \in A$ if $D = V$ and if the directional derivative $f'(v; \cdot) : V \to \mathbb{R}$ is a linear and continuous functional, i.e., if it belongs to the topological dual V^*. The linear functional $f'(v; \cdot) : V \to \mathbb{R}$ is called *Gateaux differential*.

Notice that Gateaux differentiability requires that the functional is linear and continuous. In the finite dimensional case seen in Chapter 4, linear functionals are automatically continuous thanks to Theorem 359 and for this reason it was not necessary to require explicitly the continuity.

Frechet differentiability can be also naturally extended to general normed vector spaces: say that a function $f : A \to \mathbb{R}$ is Frechet differentiable at $v \in A$ if there exists a continuous linear functional $df(v) : V \to \mathbb{R}$ such that

$$\lim_{h \to 0} \frac{|f(v + h) - f(v) - df(v)(h)|}{\|h\|} = 0. \tag{7.14}$$

The functional $df(v)$ is the Frechet differential of f at $v \in A$.

This extends Definition 145 to general normed vector spaces. It is easy to check that the properties of Frechet differentials established in Chapter 4 for functionals defined on $\mathbb{R}^n$ still hold in normed vector spaces. In particular, if $f : A \to \mathbb{R}$ is Frechet differentiable at $v \in A$, then:

(i) the Frechet differential $df(v): V \to \mathbb{R}$ is unique;

(ii) f is Gateaux differentiable at v, with $df(v)(w) = f'(v; w)$ for all $w \in V$;

(iii) f is continuous at v.

Frechet differentiability is a stronger notion than Gateaux differentiability, and in fact the converse in (ii) is false: a function $f : A \to \mathbb{R}$ that is Gateaux differentiable at v may not be Frechet differentiable at v (and may also be not continuous at v). The next result clarifies the extent to which Frechet differentiability is a stronger notion than Gateaux differentiability. In particular, it shows that Frechet differentiability can be regarded as a "uniform" version of Gateaux differentiability, where the limit (7.13) is required to converge to zero uniformly across all w that belong to unit sphere $S_V = \{w \in V : \|w\| = 1\}$ of V. In other words, all such limits must converge to zero at a similar pace. This is something that Gateaux differentiability per se does not require, as it only require that each individual limit (7.13) exists, without any assumption on their relative behavior.[10]

Proposition 374 *A function $f : A \to \mathbb{R}$ is Frechet differentiable at $v \in A$ if and only if is Gateaux differentiable at v and the limit (7.13) exists uniformly on the unit sphere S_V of V.*[11]

Proof. Suppose f is Gateaux differentiable at v. If $w \in S_V$, then by setting $h = tw$ we have

$$\left|\frac{f(v+tw) - f(v)}{t} - f'(v; w)\right| = \frac{|f(v+tw) - f(v) - f'(v; tw)|}{t} = \frac{|f(v+h) - f(v) - f'(v; h)|}{\|h\|}, \tag{7.15}$$

for all $t > 0$. Suppose the limit (7.13) exists uniformly on S_V, that is, for all $\varepsilon > 0$ there exists $t_\varepsilon > 0$ such that

$$\left|\frac{f(v+tw) - f(v)}{t} - f'(v; w)\right| < \varepsilon \tag{7.16}$$

for all $0 < t \leq t_\varepsilon$ and all $w \in S_V$. Wlog, we can assume that $t_\varepsilon \leq 1$. By (7.15),

$$\frac{|f(v+h) - f(v) - f'(v; h)|}{\|h\|} < \varepsilon \tag{7.17}$$

[10] Though we could have already stated this insightful result in Chapter 4 for $\mathbb{R}^n$, we state it here because its nature is best understood in general normed vector spaces.

[11] That is, for all $\varepsilon > 0$ there exists $t_\varepsilon > 0$ such that

$$\left|\frac{f(v+tw) - f(v)}{t} - f'(v; w)\right| < \varepsilon$$

for all $0 < t \leq t_\varepsilon$ and all $w \in S_V$.

for all $h \in V$ such that $\|h\| = t\|w\| = t \leq t_\varepsilon$. Hence, (7.14) holds and so f is Frechet differentiable at $v \in A$.

Conversely, suppose f is Frechet differentiable at $v \in A$. Hence, it is Gateaux differentiable at v with $f'(v;w) = df(v)(w)$ for all $w \in V$. Moreover, for each $\varepsilon > 0$ there exists $\delta_\varepsilon > 0$ such that (7.17) holds for all $h \in V$ such that $0 < \|h\| \leq \delta_\varepsilon$. Let $w \in S_V$. For each $0 < t \leq \delta_\varepsilon$, by setting $h = tw$ from (7.15) and (7.17) we have:

$$\left| \frac{f(v+tw) - f(v)}{t} - f'(v;w) \right| < \varepsilon.$$

Since this holds for any $w \in S_V$, we conclude that the limit (7.13) exists uniformly on S_V. ■

Next we extend Theorem 157 to normed vector spaces. To this end, consider the topological dual V^* with the norm $\|\cdot\|$ given by (7.5). The Gateaux differential $f'(v;\cdot) : V \to \mathbb{R}$ is continuous at v if $\|f'(v_n;\cdot) - f'(v;\cdot)\| \to 0$ whenever $v_n \to v$.

Proposition 375 *A function $f : A \to \mathbb{R}$ is Frechet differentiable at $v \in A$ if is Gateaux differentiable on a neighborhood of v and $f'(v;\cdot) : V \to \mathbb{R}$ is continuous at v.*

Proof. Let $B_\varepsilon(v)$ be the neighborhood of v where f is Gateaux differentiable. Let $h \in V$ be such that $\|h\| < \varepsilon$. By Exercise 13.0.59, there is an open interval (a,b), with $[0,1] \subseteq (a,b)$, such that $v + th \in B_\varepsilon(v)$ for all $t \in (a,b)$. Let $\varphi : (a,b) \to \mathbb{R}$ be given by $\varphi(t) = f(v+th)$. Then, φ is differentiable on (a,b), with $\varphi'(t) = f'(v+th;h)$. By the Mean Value Theorem there exists $\xi \in (0,1)$ such that $\varphi(1) - \varphi(0) = \varphi'(\xi)$. We can thus write:

$$\begin{aligned} |f(v+h) - f(v) - f'(v;h)| &= |\varphi(1) - \varphi(0) - f'(v;h)| = |\varphi'(\xi) - f'(v;h)| \\ &= |f'(v+\xi h;h) - f'(v;h)| \leq \|f'(v+\xi h;\cdot) - f'(v;\cdot)\| \, \|h\| \end{aligned}$$

where the last inequality follows from (7.6). By the continuity of $f'(v;\cdot) : V \to \mathbb{R}$ at v, $\|f'(v+\xi h;\cdot) - f'(v;\cdot)\| \to 0$ as $h \to \mathbf{0}$. Hence, $|f(v+h) - f(v) - f'(v;h)| = o(\|h\|)$, and we conclude that f is Frechet differentiable at v. ■

Thus, the continuity of the Gateaux differential of f, which in Theorem 157 amounted to the continuity of the gradient mapping ∇f, ensures the Frechet differentiability of f. This is an important observation since it is in general easier to compute the Gateaux differential of a function: by Proposition 375 it is then enough to check whether the Gateaux differential is continuous to conclude that f is also Frechet differentiable.

We close by showing that in finite dimensional spaces Frechet and Gateaux differentiability are equivalent notions for the important class of locally Lipschitz functions (see Definition 422 below).

Proposition 376 *Suppose $f : A \to \mathbb{R}$ is a locally Lipschitz function defined on an open set A of a finite dimensional normed vector space. Then, f is Gateaux differentiable at $v \in A$ if and only if f is Frechet differentiable at v.*

Proof. We prove the "only if" part, the converse being trivial by now. Suppose that f is Gateaux differentiable at $v \in A$. Fix $\varepsilon > 0$. By Proposition 374, to show that f is also Frechet differentiable at v we must show that there exists $t_\varepsilon > 0$ such that

$$\left| \frac{f(v+tw) - f(v)}{t} - f'(v;w) \right| < \varepsilon \tag{7.18}$$

for all $0 < t \leq t_\varepsilon$ and all $w \in S_V$.

Fix $w \in S_V$. Since f is Gateaux differentiable at v, there is $t_w > 0$ such that

$$\left| \frac{f(v+tw) - f(v)}{t} - f'(v;w) \right| < \varepsilon$$

for all $0 < t \leq t_w$. Moreover, since f is locally Lipschitz at v, there exists $B_\varepsilon(v)$ and $K > 0$ such that

$$\left| \frac{f(v+tw') - f(v)}{t} - \frac{f(v+tw'') - f(v)}{t} \right| \leq K \|w' - w''\|$$

for all $0 < t < \varepsilon$ and all $w', w'' \in S_V$ (by Exercise 13.0.59, if $w \in S_V$, then $v+tw \in B_\varepsilon(v)$ if and only if $0 < t < \varepsilon$). Hence, given any $z \in S_V$, for all $0 < t \leq \min\{t_w, \varepsilon\}$ we have:

$$\begin{aligned} \left| \frac{f(v+tz) - f(v)}{t} - f'(v;z) \right| &= \left| \frac{f(v+tz) - f(v)}{t} - \frac{f(v+tw) - f(v)}{t} \right. \\ &\quad \left. + \frac{f(v+tw) - f(v)}{t} - f'(v;w) + f'(v;w) - f'(v;z) \right| \\ &\leq K\|w-z\| + \varepsilon + \|f'(v;\cdot)\| \|w-z\|. \end{aligned}$$

Hence, there exists $\delta_w > 0$ such that the neighborhood $B_{\delta_w}(w)$ of w is such that

$$\left| \frac{f(v+tz) - f(v)}{t} - f'(v;z) \right| \leq 3\varepsilon \tag{7.19}$$

for all $0 < t \leq \max\{t_w, \varepsilon\}$ and all $z \in B_{\delta_w}(w)$.

By considering for each $w \in S_V$ such a neighborhood $B_{\delta_w}(w)$, we have an open cover $\{B_{\delta_w}(w)\}_{w \in S_V}$ of S_V . The unit sphere S_V is compact because it is a closed subset of the closed unit ball $\overline{B}_1(0)$, which is compact by Theorem 364. Hence, there is a finite subcover $\{B_\delta(w_i)\}_{i=1}^n$ of S_V, with $w_i \in S_V$ for all $i = 1, ..., n$, and we can thus write

$$\left| \frac{f(v+tz) - f(v)}{t} - f'(v;z) \right| \leq 3\varepsilon$$

for all $0 < t \leq \min_{i=1,...n}\{t_{w_1}, ..., t_{w_n}\varepsilon\}$ and all $z \in S_V$. By setting $t_\varepsilon = \min_{i=1,...n}\{t_{w_1}, ..., t_{w_n}\varepsilon\}$, this implies (7.18), as desired. ■

7.8 Convex Sets

We devote this last section to the study of the convex sets, a fundamental class of sets of a vector space.

Definition 377 *A set C of a vector space V is said to be **convex** if, for each $v, w \in V$, we have $tv + (1-t)\, w \in C$ for each $t \in [0,1]$.*

In other words, a set is convex if it contains the segment:

$$[v, w] = \{(1-t)\, v + tw : t \in [0,1]\}$$

that joins two any points v and w of the set.

Notice that the points of the segment $[v, w]$ can be seen as linear combinations of the vectors v and w in which the coefficients are required to be positive and to add up to one. In general, given a collection $\{v^i\}_{i=1}^n$ of vectors, a linear combination $\sum_{i=1}^n \alpha_i v^i$ is called a *convex combination* of the vectors $\{v^i\}_{i=1}^n$ if $\alpha_i \in [0,1]$ for each $i = 1, ..., n$ and if $\sum_{i=1}^n \alpha_i = 1$. In the case $n = 2$, $\alpha_1 + \alpha_2 = 1$ implies $\alpha_2 = 1 - \alpha_1$, and hence convex combinations of two vectors have the form $\alpha v + (1-\alpha)\, w$ with $\alpha \in [0,1]$.

Lemma 378 *A set C of a vector space V is convex if and only if it is closed with respect to all convex combinations of its own elements.*

Hence, C is convex if and only if $\sum_{i=1}^n \alpha_i v^i \in C$ for each collection $\{v^i\}_{i=1}^n$ of vectors of C and each collection $\{\alpha_{i=1}^n\}$ of scalars such that $\alpha_i \in [0,1]$ for each $i = 1, ..., n$ and $\sum_{i=1}^n \alpha_i = 1$.

Proof The "If" is obvious because by considering the convex combinations with $n = 2$ we get Definition 377. We prove the "Only if." Let C be convex and let $\{v^i\}_{i=1}^n$ be a collection of vectors of C and $\{\alpha_i\}_{i=1}^n$ a collection of scalars such that $\alpha_i \in [0,1]$ for each $i = 1, ..., n$ and $\sum_{i=1}^n \alpha_i = 1$. We want to prove that $\sum_{i=1}^n \alpha_i v^i \in C$. By Definition 377, this is true for $n = 2$. We proceed by induction on n: we assume that it is true for $n-1$ and we show that this implies that the property holds also for n. We have:

$$\sum_{i=1}^{n} \alpha_i v^i = \sum_{i=1}^{n-1} \alpha_i v^i + \alpha_n v^n = (1-\alpha_n) \sum_{i=1}^{n-1} \frac{\alpha_i}{1-\alpha_n} v^i + \alpha_n v^n.$$

Since we have assumed that C is closed with respect to the convex combinations of $n-1$ elements, we have:

$$\sum_{i=1}^{n-1} \frac{\alpha_i}{1-\alpha_n} v^i \in C.$$

Hence, the convexity of C implies:

$$(1-\alpha_n)\sum_{i=1}^{n-1}\frac{\alpha_i}{1-\alpha_n}v^i+\alpha_n v^n\in C,$$

from which it follows that C is closed with respect to the convex combinations of n elements, as desired. ■

As to set operations, the next result shows that set intersection preserves convexity. On the contrary, the union of convex sets is not in general a convex set. For example, in $\mathbb{R}^2$ the horizontal and vertical axes are convex sets, while their union is not.

Lemma 379 *The intersection of any collection of convex subsets of a vector space is a convex set.*

Proof Let $\{C_\alpha\}$ be any collection of convex subsets of a vector space V. Let $C=\bigcap_\alpha C_\alpha$. The empty set is trivially convex, and hence if $C=\emptyset$ the result holds. Suppose therefore that $C\neq\emptyset$. Let $v,w\in C$ and let $t\in[0,1]$. We want to prove that $tv+(1-t)\,w\in C$. Since $v,w\in C_\alpha$ for each α, we have that $tv+(1-t)\,w\in C_\alpha$ for each α because each set C_α is convex. Hence, $tv+(1-t)\,w\in\bigcap_\alpha C_\alpha$, as desired. ■

Definition 380 *Given any subset A of a vector space V, its convex envelope $co\,(A)$ is the smallest convex set that contains A. If V is normed, its closed convex envelope $\overline{co}\,(A)$ is the smallest closed and convex set that contains A.*

Next results show that convex envelopes are the counterpart for convex combinations of what generated subspaces were for linear combinations (remember Section 1.9). We begin with a useful lemma.

Lemma 381 *Let C be a convex subset of a normed vector space V. Then, its closure $\overline{C}$ is a convex subset of V.*

Proof Let $v,w\in\overline{C}$ and let $t\in[0,1]$. By Theorem 254, there exist two sequences $\{v^n\}_{n\geq1}$ and $\{w^n\}_{n\geq1}$ in C such that $v^n\to v$ and $w^n\to w$. Since C is convex, $tv^n+(1-t)\,w^n\in C$ for each n. Since $tv^n+(1-t)\,w^n\to tv+(1-t)\,w$ and $\overline{C}$ is closed, it follows that $tv+(1-t)\,w\in\overline{C}$, as desired. ■

Proposition 382 *Given a subset A of a vector space V, let $\{C_\alpha\}$ be the collection of all convex subsets of V containing A. We have $co\,(A)=\bigcap_\alpha C_\alpha$. If, moreover, V is normed, we have:*

$$\overline{co}\,(A)=\overline{co\,(A)}=\bigcap_\alpha\overline{C}_\alpha. \tag{7.20}$$

Expression (7.20) shows that to obtain $\overline{co}(A)$ it is necessary first to construct $co(A)$ and then, in order to pass to the closure $\overline{co(A)}$, it is necessary to consider all the limits of sequences in $co(A)$ (remember Theorem 254).

Proof By Lemma 379, $\bigcap_\alpha C_\alpha$ is a convex set of V. Since $A \subseteq C_\alpha$ for each α, we have $co(A) \subseteq \bigcap_\alpha C_\alpha$ since, by definition, $co(A)$ is the smallest convex subset of V containing A. On the other hand, $co(A)$ belongs to the collection $\{C_\alpha\}$, being a convex subset of V containing A. It follows that $\bigcap_\alpha C_\alpha \subseteq co(A)$ and we can therefore conclude that $\bigcap_\alpha C_\alpha = co(A)$.

Let $\{F_\alpha\}$ be the collection of all closed and convex subsets of V containing the set A. By proceeding as above we get $\overline{co}(A) = \bigcap_\alpha F_\alpha$. By Lemma 381, $\overline{co(A)} \in \{F_\alpha\}$. On the other hand, $\overline{co(A)} = \bigcap_\alpha F_\alpha$ since $\overline{co(A)}$ is by Theorem 235-(iii) the smallest closed set containing $co(A)$.

It remains to show that $\{F_\alpha\} = \left\{\overline{C}_\alpha\right\}$. By Lemma 381, $\left\{\overline{C}_\alpha\right\} \subseteq \{F_\alpha\}$, while from $F_\alpha = \overline{F}_\alpha$ it follows that $\{F_\alpha\} \subseteq \left\{\overline{C}_\alpha\right\}$. This completes the proof. ■

The next result shows that $co(A)$ can be represented through convex combinations of vectors of A.

Theorem 383 *Let A be a subset of a vector space V. A vector $v \in V$ belongs to $co(A)$ if and only if it is a convex combination of vectors of A, i.e., if and only if there exists a finite set $\{v^i\}_{i\in I}$ of A and a finite set $\{\alpha_i\}_{i\in I}$ of scalars, with $\alpha_i \in [0,1]$ for each $i \in I$ and $\sum_{i\in I} \alpha_i = 1$, such that $v = \sum_{i\in I} \alpha_i v^i$.*

Proof "If." Let $v \in V$ be convex combination of a finite set $\{v^i\}_{i\in I}$ of vectors of A. The set $co(A)$ is convex and, since $\{v^i\}_{i\in I} \subseteq co(A)$, Lemma 378 implies $v \in co(A)$, as desired.

"Only if." Let C be the set of all the vectors v of V that can be expressed as convex combinations of vectors of A, i.e., $v \in C$ if there exist finite sets $\{v^i\}_{i\in I} \subseteq A$ and $\{\alpha^i\}_{i\in I} \subseteq \mathbb{R}$, with $\alpha_i \in [0,1]$ for each $i \in I$ and $\sum_{i\in I} \alpha_i = 1$, such that $v = \sum_{i=1}^n \alpha_i v^i$. It is easy to see that C is a convex subset of V containing A. It follows that $co(A) \subseteq C$ and hence each $v \in co(A)$ is a linear combination of vectors of A. ■

Example 384 Let $A = \left\{v^1, ..., v^k\right\} \subseteq V$. By Theorem 383 we have:

$$co(A) = \left\{ \sum_{i=1}^k \alpha_i v^i : \alpha_i \in [0,1] \ \ \forall i = 1, ..., k \quad \text{and} \quad \sum_{i=1}^k \alpha_i = 1 \right\}.$$

The convex sets that are a convex envelope of a finite collection of vectors are called polytopes. In $\mathbb{R}^2$, the polytopes are nothing else that the polygons studied in high

school. For example, if $A = \{(0,1),(1,0),(-1,0),(0,-1)\}$, then $co(A)$ is the rhomb that has as vertices the four points of the set A. Note that the set

$$A' = \{(0,1),(1,0),(-1,0),(0,-1),(1/2,1/2)\}$$

is such that $co(A) = co(A')$. Hence, it may well happen that the finite collection of vectors that generates a polygon is made only by the vertices, even though they must necessary belong to it. ▲

Example 385 Let $A = \{(1,0,0),(0,1,0),(0,0,1)\} \subseteq \mathbb{R}^3$. We have:

$$\begin{aligned} & co(A) \\ = \; & \{x \in \mathbb{R}^3 : x = \alpha_1(1,0,0) + \alpha_2(0,1,0) + (1-\alpha_1-\alpha_2)(0,0,1) \\ & \text{with } \alpha_i \in [0,1] \;\; \forall i = 1,2,3 \quad \text{and} \quad \alpha_1 + \alpha_2 \leq 1\} \\ = \; & \{(\alpha_1, \alpha_2, 1-\alpha_1-\alpha_2) : \alpha_i \in [0,1] \;\; \forall i = 1,2,3 \quad \text{and} \quad \alpha_1 + \alpha_2 \leq 1\}. \end{aligned}$$

More generally, let $A = \{e^1, ..., e^n\} \subseteq \mathbb{R}^n$. We have:

$$\begin{aligned} co(A) &= \left\{ \sum_{i=1}^n \alpha_i e^i : \alpha_i \in [0,1] \;\; \forall i = 1,...,n \quad \text{and} \quad \sum_{i=1}^n \alpha_i = 1 \right\} \\ &= \left\{ (\alpha_1, ..., \alpha_n) : \alpha_i \in [0,1] \;\; \forall i = 1,...,n \quad \text{and} \quad \sum_{i=1}^n \alpha_i = 1 \right\}. \end{aligned}$$

This polytope $co(A)$ is called the simplex of $\mathbb{R}^n$, often denoted by Δ_{n-1}. ▲

By Theorem 383, each vector of the convex envelope $co(A)$ can be obtained as convex combination of a finite set $\{v^i\}_{i \in I}$ of vectors of A. The next important result, often called the Caratheodory Theorem, shows that in a vector space of finite dimension n it is actually sufficient to consider convex combinations of at most $n+1$ vectors of A, i.e., with $|I| \leq n+1$.

Theorem 386 *Let V be a vector space of finite dimension n. Each vector that belongs to $co(A)$ can be written as convex combination of at most $n+1$ elements of A.*

Proof. Wlog, let $A \subseteq \mathbb{R}^n$. We prove that if $v \in co(A)$ and $v = \sum_{i=0}^n t_i v^i$ with $v^i \in A$, $t_i > 0$, $\sum_{i=0}^n t_i = 1$ and $N > d$, then v can be written as convex combination of at most $N-1$ points of A. Wlog, set $v^0 = 0$. But, $N > d$ implies that $\{v^i\}_{i=1}^n$ are linearly dependent. Therefore, there exist $\lambda_1, \lambda_2, ..., \lambda_N \in \mathbb{R}$ not all zero such that $\sum_{i=1}^n \lambda_i v^i = 0$. Wlog, assume $\sum_{i=0}^n \lambda_i \geq 0$ (otherwise we change all their signs) and $\lambda_j > 0$ for some j. Observe that if $\lambda_i \leq 0$, then $t_i - t\lambda_i \geq 0$, while if $\lambda_i > 0$, then $t_i - t\lambda_i \geq 0$ if and only if $t \leq \frac{t_i}{\lambda_i}$. Set $\bar{t} = \min\left\{\frac{t_n}{\lambda_n} : \lambda_n > 0\right\} = \frac{t_j}{\lambda_j}$. Then

$$\begin{aligned} v &= \sum_{i=0}^{n} t_i v^i = \sum_{i=1}^{n} t_i v^i = \sum_{i=1}^{n} t_i v^i - \bar{t} \sum_{i=1}^{n} t_i v^i \\ &= \sum_{i=1}^{n} (t_i - \bar{t}\lambda_i) v^n = \sum_{j \neq i=1}^{n} (t_i - \bar{t}\lambda_i) v^i \end{aligned}$$

and $\sum_{i=1}^{n} (t_i - \bar{t}\lambda_i) = \sum_{i=1}^{n} t_i - \bar{t} \sum_{i=1}^{n} \lambda_i \leq 1$ (since $\sum_{i=1}^{n} \lambda_i \geq 0$). Finally:

$$v = \sum_{j \neq i=1}^{n} (t_i - \bar{t}\lambda_i) a_i + \left(1 - \sum_{i=1}^{n} (t_i - \bar{t}\lambda_i) \right) 0.$$

■

Next result is a consequence of the Caratheodory Theorem and shows that convex envelopes preserve compactness.

Corollary 387 *Let V be a finite dimensional normed vector space. The convex envelope of a compact subset of V is compact.*

When A is a compact subset of a finite dimensional normed vector space, we therefore have that $co(A)$ is compact, and so $co(A) = \overline{co}(A)$.

Proof Wlog, let $A \subseteq \mathbb{R}^l$. Let $x_k = \sum_{j=1}^{N+1} \lambda_j^{(k)} a_j^{(k)}$ for $k \in \mathbb{N}$ be a sequence in $co(A)$. But,

$$\left\{ \lambda_1^{(k)} \right\}_{k \in \mathbb{N}}, \left\{ \lambda_2^{(k)} \right\}_{k \in \mathbb{N}}, ..., \left\{ \lambda_{N+1}^{(k)} \right\}_{k \in \mathbb{N}} \subseteq [0,1]$$

compact and $\left\{ a_1^{(k)} \right\}_{k \in \mathbb{N}}, \left\{ a_2^{(k)} \right\}_{k \in \mathbb{N}}, ..., \left\{ a_{N+1}^{(k)} \right\}_{k \in \mathbb{N}} \subseteq A$ compact, hence in at most $2N+2$ steps we can obtain a subsequence $\{k_m\}$ of $\{k\}$ such that for $m \to \infty$ we have $\lambda_j^{(k_m)} \to \bar{\lambda}_j \in [0,1]$ and $a_j^{(k_m)} \to \bar{a}_j \in A$, then for the continuity of the operations: $\sum_{j=1}^{N+1} \lambda_j^{(k_m)} a_j^{(k_m)} \to \sum_{j=1}^{N+1} \bar{\lambda}_j \bar{a}_j$ for $m \to \infty$, moreover $1 = \sum_{j=1}^{N+1} \lambda_j^{(k_m)} \to \sum_{j=1}^{N+1} \bar{\lambda}_j$ implies $\sum_{j=1}^{N+1} \bar{\lambda}_j = 1$. The subsequence x_{k_m} of x_k converges to $\sum_{j=1}^{N+1} \bar{\lambda}_j \bar{a}_j \in co(A)$. ■

7.8.1 Affine spaces

Let V be a vector space and $W \subseteq V$ be a linear subspace of V. The translation $A = W + u$ is called an affine subspace of V. We also say that A is parallel to W. The dimension of A is the dimension of W, whenever it is finite.

For instance, if $\dim(W) = 1$, A is called a straight line. The straight line can be written in the parametric form

$$A = \{u + \lambda v : \lambda \in \mathbb{R}\}.$$

The opposite extreme to the straight lines is represented by the (affine) hyperplanes. Consider any non-zero linear functional $L : V \to \mathbb{R}$. If α is a fixed element of $\mathbb{R}$ the set

$$H = \{v \in V : L(v) = \alpha\}$$

is called an hyperplane of V. Clearly H is an affine subspace because it is parallel to $\ker(L)$. Actually, $H = \ker(L) + \overline{v}$, where $\overline{v}$ is any element for which $L(\overline{v}) = \alpha$.

Observe that if the affine subspace A has the representation $A = W + u$, where W is a linear subspace, then $u \in A$, as $\mathbf{0} \in W$. Moreover, if u_1 is another point $u_1 \in A$, then $A = W + u = W + u_1$. Actually, from $A = W + u$, it follows $u_1 = w + u$ for some $w \in W$. Hence, $W + u_1 = W + w + u = W + u$.

This fact implies another important property. If A is an affine space and u is any point $u \in A$, then $A - u$ is the linear subspace parallel to A and we have $A = W + u$, where $W = A - u$.

A linear combination

$$v = \lambda_1 v_1 + \lambda_2 v_2 + ... + \lambda_m v_m$$

in a vector space V is called an affine combination if $\lambda_1 + \lambda_2 + ... + \lambda_m = 1$. The vectors $\{v_i\}_{i=1}^m$ are said to be affinely independent, provided

$$\lambda_1 v_1 + \lambda_2 v_2 + ... + \lambda_m v_m = 0$$

and $\lambda_1 + \lambda_2 + ... + \lambda_m = 0$ implies necessarily that $\lambda_1 = \lambda_2 = ... = \lambda_m = 0$.

The affine combinations play for affine spaces the role played by linear combinations for vector spaces and the convex combinations for convex sets.

Proposition 388 *$A \subseteq V$ is affine if and only if it contains all the affine combinations of elements in A.*

Proof Assume that A is affine. That is, $A = a + W$ where W is a linear subspace of V. Let $(v_i)_{i=1}^m \subseteq A$ and $(\lambda_i)_{i=1}^m$ be with $\sum_{i=1}^m \lambda_i = 1$. Since $v_i - a \in W$, it follows

$$\sum_{i=1}^m \lambda_i v_i = a + \sum_{i=1}^m \lambda_i (v_i - a) \in a + W.$$

Conversely, assume that A is closed under affine combinations. Pick $a \in A$. We must prove that $A - a$ is a vector space. Given $v_1, v_2 \in A$ and $\lambda_1, \lambda_2 \in \mathbb{R}$, we have

$$\lambda_1 (v_1 - a) + \lambda_2 (v_2 - a) + a = \lambda_1 v_1 + \lambda_2 v_2 + (1 - \lambda_1 + \lambda_2) a \in A.$$

Hence, $\lambda_1 (v_1 - a) + \lambda_2 (v_2 - a) \in A - a$. Consequently $A - a$ is a vector space. ■

Given a nonempty subset $C \subseteq V$, we define the affine hull of C, denoted by $\text{aff}(C)$, the smallest affine subspace containing C. The next proposition characterizes this set $\text{aff}(C)$.

Proposition 389 *The following definitions are equivalent:*
i)

$$\text{aff}\,(C) = \bigcap_{\alpha} A_\alpha$$

where the intersection is made over the class of all affine subspaces $A_\alpha \supseteq C$;
ii)

$$\text{aff}\,(C) = \left\{ \sum_{i=1}^{m} \lambda_i v_i : \{v_i\}_{i=1}^{m} \subseteq C,\ \sum_{i=1}^{m} \lambda_i = 1 \right\}. \tag{7.21}$$

Proof We must first prove that the definition as intersection of affine subspaces is meaningful. Note that the class is nonempty since $V \supseteq C$ is an affine space. Let us show that intersection is still an affine space. Since C is nonempty, let $c \in C$. Hence, $c \in A_\alpha$ for all α. Hence, $A_\alpha - c = W_\alpha$ is a linear subspace, and $A_\alpha = W_\alpha + c$. Therefore,

$$\bigcap_{\alpha} A_\alpha = \bigcap_{\alpha} (W_\alpha + c) = \left(\bigcap_{\alpha} W_\alpha \right) + c$$

that implies $\bigcap_\alpha A_\alpha$ to be an affine space.

ii) We now show that definition 7.21 turns out to be equivalent to the first one. The set defined in 7.21 is an affine space. More specifically, if c is a fixed element of C, then

$$\begin{aligned} & \left\{ \sum_{i=1}^{m} \lambda_i v_i : \{v_i\}_{i=1}^{m} \subseteq C,\ \sum_{i=1}^{m} \lambda_i = 1 \right\} \\ = \ & \text{span}\,\{v - c : v \in C\} + c. \end{aligned}$$

Actually, if $v = \sum_{i=1}^{m} \lambda_i v_i$, with $v_i \in C$ and $\sum_{i=1}^{m} \lambda_i = 1$, then

$$\begin{aligned} v &= \sum_{i=1}^{m} \lambda_i v_i = \sum_{i=1}^{m} \lambda_i v_i - \sum_{i=1}^{m} \lambda_i c + c \\ &= \sum_{i=1}^{m} \lambda_i (v_i - c) + c \end{aligned}$$

and so $v \in \text{span}\,\{v - c : v \in C\} + c$. Conversely, if $v \in \text{span}\,\{v - c : v \in C\} + c$, then

$$v = \sum_{i=1}^{m} \lambda_i (v_i - c) + c$$

for scalars λ_i. Consequently,

$$v = \sum_{i=1}^{m} \lambda_i v_i + \left(1 - \sum_{i=1}^{m} \lambda_i \right) c$$

and v is the affine combination of elements in C.

It remains to prove that if $A_\alpha \supseteq C$ then

$$A_\alpha \supseteq \left\{ \sum_{i=1}^{m} \lambda_i v_i : \{v_i\}_{i=1}^{m} \subseteq C, \ \sum_{i=1}^{m} \lambda_i = 1 \right\}.$$

But this is obvious since the operation of affine combinations is closed for affine spaces. The claim is hence proved. ■

Proposition 390 *If H is an hyperplane of V, and $v_0 \notin H$, then*

$$\text{aff}\,(H, v_0) = \text{span}\,(H, v_0) = V.$$

Consequently, if $A \supseteq H$ is affine, then either $A = H$ or $A = V$.

Proof Let $H = \{v : V : L(v) = \alpha\}$, where L is a linear functional. Let $L(v_0) = \beta \neq \alpha$, as $v_0 \notin H$. Take any point $v \in V$ with $L(v) \neq \beta$. Consider the point

$$v_1 = \frac{\alpha - \beta}{L(v) - \beta} v + \left(1 - \frac{\alpha - \beta}{L(v) - \beta}\right) v_0.$$

It is easy to see that $L(v_1) = \alpha$. Hence, $v_1 \in H$ and consequently

$$v = \frac{L(v) - \beta}{\alpha - \beta} v_1 + \left(1 - \frac{L(v) - \beta}{\alpha - \beta}\right) v_0.$$

We deduce that all the points outside the hyperplane $L(v) = \beta$ lie in aff (H, v_0). This is enough as an affine space is connected and therefore also the set $L(v) = \beta$ lies in aff (H, v_0). Note further that from the obvious relation aff $(H, v_0) \subseteq$ span $(H, v_0) \subseteq V$ it follows that aff $(H, v_0) =$ span (H, v_0).

To finish, if $A \supset H$, there is a point $v_0 \in A$ and $v_0 \notin H$. Hence, aff $(H, v_0) \subseteq A$, that implies $A = V$. ■

Corollary 391 *Let H be an hyperplane of a vector space V.*
i) if $\dim(V) = n$, *then* $\dim(H) = n - 1$.
ii) if V is a normed space, then either H is closed or $\overline{H} = V$.

Proof (i) From the relation span $(H, v_0) = V$, it follows that $\dim(H) \geq n - 1$. Clearly $\dim(H)$ cannot be n. Otherwise the parallel linear space W would have dimension n. Hence $H = a + W = a + V = V$ which is not an hyperplane.

(ii) Suppose that H is not closed, i.e., $H \subset \overline{H}$. It is easy to see that $\overline{H}$ is still affine. Let $v_0 \in \overline{H}$ and $v_0 \notin H$. Then

$$\overline{H} \supseteq \text{aff}\,(H, v_0) = V$$

and therefore H is dense in V. ■

We end this subsection by giving a useful criterion for the determination of the dimension of affine spaces, related to the notion of affinely independent vectors.

Proposition 392 *Let $A \subseteq V$ be an affine space. $\dim(A) = m$ if and only if $m+1$ is the maximum number of elements in A which are affinely independent.*

Proof Let $A = W + a$ where W is the parallel subspace. Assume that $\dim(A) = m = \dim W$. This implies the existence of m linearly independent vectors $(v_i)_{i=1}^m \subset W$. Consider the $m+1$ vectors $a, v_1 + a, ..., v_m + a$ in A. We show that they are affinely independent. Let $\lambda_0, \lambda_1, ..., \lambda_m$ be scalars with $\lambda_0 + \sum_{i=1}^m \lambda_i = 0$. We have

$$\begin{aligned} \lambda_0 a + \sum_{i=1}^m \lambda_i (v_i + a) &= \left(\lambda_0 + \sum_{i=1}^m \lambda_i\right) a + \sum_{i=1}^m \lambda_i v_i \\ &= \sum_{i=1}^m \lambda_i v_i. \end{aligned}$$

Because $(v_i)_{i=1}^m$ are linearly independent, $\lambda_i = 0$ for $i = 1, ..., m \Rightarrow \lambda_0 = 0$. Hence, the vectors $a, v_1 + a, ..., v_m + a$ are affinely independent. Thus the maximum number is greater or equal to $m+1$. To prove that it is just $m+1$, suppose by contradiction the existence of $p > m+1$ elements $(v_i)_{i=1}^p \subset A$ which are affinely independent. Consider the $p-1 > m$ points $v_2 - v_1, ..., v_p - v_1 \in W$. They are linearly independent. Actually, set

$$\sum_{i=2}^p \lambda_i (v_i - v_1) = 0.$$

It follows

$$0 = \sum_{i=2}^p \lambda_i (v_i - v_1) = \sum_{i=2}^p \lambda_i v_i - \left(\sum_{i=2}^p \lambda_i\right) v_1.$$

Hence, $\lambda_i = 0$ for $i = 2, ..., p$ and the $p-1$ vectors are linearly independent. This contradicts the fact that $\dim W = m < p - 1$. ■

Definition 393 *Let $v_1, v_2, ..., v_{n+1} \in \mathbb{R}^n$ be affinely independent. The set $\Delta = co(v_1, v_2, ..., v$ is called a n-dimensional simplex.*

Proposition 394 *Let $v_1, v_2, ..., v_{n+1} \in \mathbb{R}^n$ be affinely independent. Any point $v \in \mathbb{R}^n$ admits a unique representation as $v = \lambda_1 v_1 + \lambda_2 v_2 + ... + \lambda_{n+1} v_{n+1}$ with $\lambda_1 + \lambda_2 + ... + \lambda_{n+1} = 1$. In particular, if $\lambda_i \geq 0$ we get the points of the simplex Δ.*

Proof Clearly any point is representable as an affine combination because $aff(v_1, v_2, ...,$ $\mathbb{R}^n$. Let us shows that the coefficient λ_i are uniquely determined. Suppose that there is a point $v \in \mathbb{R}^n$ for which

$$\begin{aligned} v &= \lambda_1 v_1 + \lambda_2 v_2 + ... + \lambda_{n+1} v_{n+1} \\ v &= \lambda'_1 v_1 + \lambda'_2 v_2 + ... + \lambda'_{n+1} v_{n+1} \end{aligned}$$

with $\lambda_1 + \lambda_2 + ... + \lambda_{n+1} = 1 = \lambda'_1 + \lambda'_2 + ... + \lambda'_{n+1}$. This implies

$$(\lambda_1 - \lambda'_1) v_1 + (\lambda_2 - \lambda'_2) v_2 + ... + \left(\lambda_{n+1} - \lambda'_{n+1}\right) v_{n+1} = 0$$

and

$$(\lambda_1 - \lambda'_1) + (\lambda_2 - \lambda'_2) + ... + \left(\lambda_{n+1} - \lambda'_{n+1}\right) = 0.$$

Since the vectors are affinely independent, $\lambda_i = \lambda'_i$ for $i = 1, ..., n+1$ and hence the representation is unique. ■

Remark. When $v_1, v_2, ..., v_{n+1} \in \mathbb{R}^n$ are affinely independent, the map $v \leftrightarrow (\lambda_1, \lambda_2, ..., \lambda_{n+1})$, as described in Proposition 394, is called the barycenter coordinates of v.

7.8.2 Separation properties

In this subsection we treat only convex sets of some finite-dimensional space. With no loss of generality we can consider convex sets C of $\mathbb{R}^n$. By using the operation of affine closure, $\text{aff}(C)$, we show that a non-empty convex set in Euclidean spaces with empty interior, has a smaller ambient space, for which it acquires an interior. This important property makes the finite-dimensional situation radically different from the infinite-dimensional case.

Proposition 395 *Let $C \subseteq \mathbb{R}^n$ be a convex set. If $intC = \varnothing$, then $\dim \text{aff}(C) < n$.*

Proof It suffices to prove there is an affine space $A \supseteq C$ and with $\dim(A) < n$. First we claim there are no $n+1$ affinely independent points $v_1, v_2, ..., v_{n+1}$ in C. For if there were such points, then $\Delta = co(v_1, v_2, ..., v_{n+1}) \subseteq C$. In view of Proposition 394, consider the barycenter coordinates $v \leftrightarrow (\lambda_1, \lambda_2, ..., \lambda_{n+1})$. To the $n+1$-ple $\left(\frac{1}{n+1}, ..., \frac{1}{n+1}\right)$ corresponds the point $\overline{v} = \frac{1}{n+1}(v_1 + ... + v_{n+1}) \in \Delta \subset C$. Consequently, there is a small neighborhood U of $\overline{v}$ in $\mathbb{R}^n$ such that for $w \in U$ its barycenter coordinates are all positive. Hence $U \subseteq \Delta \subset C$ and C would have nonempty interior. Now we can deduce that $\dim \text{aff}(C) < n$. Actually, let $k < n+1$ be the maximum number of affinely independent points in C and let $v_1, v_2, ..., v_k$ be such points. If v is any point of C, then the linear system

$$\begin{aligned} &\lambda_1 v_1 + \lambda_2 v_2 + ... + \lambda_k v_k + \lambda v = 0 \\ &\lambda_1 + \lambda_2 + ... + \lambda_k + \lambda = 0 \end{aligned}$$

has a non trivial solution. Clearly, as $v_1, v_2, ..., v_k$ are affinely independent, λ must be different from 0. Hence,

$$v = \sum \left(-\frac{\lambda_i}{\lambda}\right) v_i$$

and v is an affine combination of the points $v_1, v_2, ..., v_k$. Hence, $C \subseteq \text{aff}\{v_1, v_2, ..., v_k\} \Longrightarrow \text{aff}(C) \subseteq \text{aff}\{v_1, v_2, ..., v_k\}$, and thus $\dim \text{aff}(C) < n$. ■

We have seen above how, given any set, the notion of convex envelope allows to construct through the convex combinations of its elements the smallest convex set that contains it. Now we consider, in a sense, the opposite problem: given a convex set C, we ask what is the smallest set of its points from which C can be reconstructed as their convex combinations. In other words, we ask what is the minimal set $A \subseteq C$ such that $co(A) = C$.

If it exists, such set A gives us the essence of the set C, its "skeleton." From the point of view of convexity, the knowledge of A would be equivalent to the knowledge of the entire set C, since C could be reconstructed from A in a "mechanical" way through convex combinations of its elements.

To understand how to address this problem, we go back to the rhomb described in Example 384. There we saw how this polygon is the convex envelope of its vertices $A = \{(0,1),(1,0),(-1,0),(0,-1)\}$. In general, it is immediate to see how any polygon can be seen as the convex envelope of its own vertices.

On the other hand, we observed how the same rhomb can be seen as the convex envelope of the set:

$$A' = \{(0,1),(1,0),(-1,0),(0,-1),(1/2,1/2)\}.$$

In this set, besides the vertices there is also the vector $(1/2, 1/2)$, which is however completely useless for the representation of the polygon because is itself a convex combination of the vertices.[12] We therefore have a redundancy in the set A', while this does not happen in the set A of the vertices, whose elements are all essential for the representation of the rhomb.

Hence, for a polygon the set of the vertices is the natural candidate to be the minimal set that allows to represent each point of the polygon as a convex combination of its elements.

Motivated by all this, we introduce the notion of extreme point, which generalizes that of vertex to any convex sets.

Definition 396 *Let C be a convex subset of a vector space V. A point $v_0 \in C$ is said to be an extreme point for C if $v_0 = tv + (1-t)w$ with $t \in (0,1)$ and $v, w \in C$ implies $w = v = v_0$.*

[12] In fact:

$$\left(\frac{1}{2}, \frac{1}{2}\right) = \frac{1}{2}(1,0) + \frac{1}{2}(0,1).$$

A point $v_0 \in C$ is therefore extreme if it is not convex combination of other two vectors of C. The set of the extreme points of C is denoted by $\text{ext}C$, and in the case of polytopes the extreme points are called *vertices*.

Next result gives a simple characterization of extreme points, by showing that they are the points that can be eliminated without altering the convex nature of the set considered.

Lemma 397 *A point v_0 of a convex set C is extreme if and only if the set $C \setminus \{v_0\}$ is convex.*

Proof Let $v_0 \in \text{ext}C$ and let $v, w \in C \setminus \{v_0\}$. Since C is convex, $tv + (1-t)\, w \in C$ for each $t \in [0,1]$. To prove that $tv + (1-t)\, w \in C \setminus \{v_0\}$, it is therefore sufficient to prove that $v_0 \neq tv + (1-t)\, w$. This is obvious if $t \in \{0,1\}$. On the other hand, if it held $v_0 = tv + (1-t)\, w$ for some $t \in (0,1)$, then Definition 396 implies $u = v = v_0$, which contradicts $v, w \in C \setminus \{v_0\}$. In conclusion, $tv + (1-t)\, w \in C \setminus \{v_0\}$, and the set $C \setminus \{v_0\}$ is therefore convex.

Viceversa, assume that $v_0 \in C$ is such that the set $C \setminus \{v_0\}$ is convex. We prove that $v_0 \in \text{ext}C$. Let $v, w \in C$ be such that $v_0 = tv + (1-t)\, w$ with $t \in (0,1)$. Since $C \setminus \{v_0\}$ is convex, if v, w belong to $C \setminus \{v_0\}$, then $tv + (1-t)\, w \in C \setminus \{v_0\}$ for each $t \in [0,1]$. Hence, $v_0 \neq tv + (1-t)\, w$ for each $t \in [0,1]$. It follows that v, w do not belong to $C \setminus \{v_0\}$, which is equivalent to say that $w = v = v_0$. In conclusion, $v_0 \in \text{ext}C$. ■

The next result shows that the extreme points belong necessarily to the frontier of the set, that is, no interior point of a convex set can be an extreme point.

Proposition 398 *Given a convex set C of a normed vector space, we have* $\text{ext}C \subseteq \partial C$.

Proof Let v be an interior point of C. We prove that $v \notin \text{ext}C$. Since v is an interior point, there exists a neighborhood $B_\varepsilon(v)$ such that $B_\varepsilon(v) \subseteq C$. Consider the points $(1-\varepsilon/n)\, v$ and $(1+\varepsilon/n)\, v$. We have:

$$\|(1-\varepsilon/n)\, v - v\| = \frac{\varepsilon}{n}\|v\| \qquad \text{and} \qquad \|(1+\varepsilon/n)\, v - v\| = \frac{\varepsilon}{n}\|v\|,$$

and hence $(1-\varepsilon/n)\, v, (1+\varepsilon/n)\, v \in B_\varepsilon(v)$ for n sufficiently large. On the other hand,

$$v = \frac{1}{2}(1-\varepsilon/n)\, v + \frac{1}{2}(1+\varepsilon/n)\, v,$$

and so $v \notin \text{ext}C$. ■

An immediate consequence of Proposition 398 is that open convex sets (like, for example, open unit balls) do not have extreme points.

We now see other examples in which we get the set of the extreme points of some convex sets.

Example 399 Consider the polytope $co(A)$ generated by a finite collection $A = \{v^1, ..., v^k\}$ of vectors of a vector space V. It is easy to see that $\text{ext}co(A)$ is not empty and that $\text{ext}co(A) \subseteq A$, i.e., the vertices of the polytope necessarily belong to the finite collection that generates the polytope. ▲

Example 400 Consider the closed unit ball $B_V = \{v \in V : \|v\| \leq 1\}$ of a normed vector space V. Since $\partial B_V = \{v \in V : \|v\| = 1\}$, by Proposition 398 we have:

$$\text{ext}B_V \subseteq \{v \in V : \|v\| = 1\}. \tag{7.22}$$

In the next examples we will see cases in which this inclusion is strict, and others in which it is instead an equality. ▲

Example 401 Consider the closed unit ball $B_{\mathbb{R}^n} = \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}$ of $\mathbb{R}^n$, endowed with its Euclidean norm $\|\cdot\|_2$. In this case, we have:

$$\text{ext}B_{\mathbb{R}^n} = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}.$$

The set of the extreme points is therefore given by the "circumference" of the ball and the inclusion (7.22) in this case is an equality. To prove this statement, we take a point $x \in \mathbb{R}^n$ such that $\|x\|_2 = 1$, and we show that $x \in \text{ext}B_{\mathbb{R}^n}$. To this end, we use Exercise 13.0.50. Let therefore $y \in B_{\mathbb{R}^n}$ be such that $x + y \in B_{\mathbb{R}^n}$ and $x - y \in B_{\mathbb{R}^n}$. Therefore, $\|x + y\| \leq 1$ and $\|x - y\| \leq 1$, from which:[13]

$$\begin{aligned} 1 &\geq \|x + y\|^2 = \|x\|^2 + \|y\|^2 + 2x \cdot y, \\ 1 &\geq \|x - y\|^2 = \|x\|^2 + \|y\|^2 - 2x \cdot y. \end{aligned}$$

Being $\|x\| = 1$, it follows that:

$$\|y\|^2 + 2x \cdot y \leq 0 \qquad \text{and} \qquad \|y\|^2 - 2x \cdot y \leq 0,$$

which implies:

$$\|y\|^2 + 2x \cdot y \leq 2x \cdot y - \|y\|^2,$$

and hence $\|y\|^2 = 0$, that is $y = \mathbf{0}$. By Exercise 13.0.50 we can conclude that $x \in \text{ext}B_{\mathbb{R}^n}$, as desired. ▲

Example 402 Consider the closed unit ball $B_{\mathbb{R}^n} = \{x \in \mathbb{R}^n : \|x\|_1 \leq 1\}$ of $\mathbb{R}^n$, endowed this time with the norm $\|\cdot\|_1$. In this case we have:

$$\text{ext}B_{\mathbb{R}^n} = \{\pm e^i : i = 1, ..., n\}, \tag{7.23}$$

[13] See Ambrosetti and Musu (1988) p. 56.

that is, $\text{ext}B_{\mathbb{R}^n}$ is the set of the fundamental versors e^i of $\mathbb{R}^n$ and of their negatives $-e^i$. The set $\text{ext}B_{\mathbb{R}^n}$ has therefore cardinality $2n$ and the inclusion (7.22) is in this case strict. For simplicity, consider the case $n = 2$. Let therefore $x \in \mathbb{R}^2$ be such that $\|x\|_1 = 1$, and show that if $x \neq \pm e^i$ for $i = 1, 2$, then $x \notin \text{ext}B_{\mathbb{R}^n}$. Since $x \neq \pm e^i$ and $\|x\|_1 = |x_1| + |x_2| = 1$, we have $0 < |x_i| < 1$ for $i = 1, 2$. Set $y = (x_1, 0)$ and $z = (0, x_2)$, so that $1 = \|x\|_1 = \|y\|_1 + \|z\|_1$. Moreover, set:

$$\widetilde{y} = \frac{y}{\|y\|_1} \qquad \text{and} \qquad \widetilde{z} = \frac{z}{\|z\|_1}.$$

We have $\|\widetilde{y}\|_1 = \|\widetilde{z}\|_1 = 1$, and $x = \|y\|_1 \widetilde{y} + (1 - \|y\|_1) \widetilde{z}$. Since $\|y\|_1 \in (0, 1)$, it follows that x is a convex combination of two other vectors of $\overline{B}_1(\mathbf{0})$, and so it is not an extreme point. ▲

Example 403 Consider the vector space $(C([0,1]), \|\cdot\|_\infty)$. In this case, $B_{C([0,1])} = \{f \in C([0,1]) : \|f\|_\infty \leq 1\}$. Denote by $1_{[0,1]}$ the constant function equal to 1, that is, $1_{[0,1]}(t) = 1$ for each $t \in [0,1]$. We have $\text{ext}B_{C([0,1])} = \{\pm 1_{[0,1]}\}$, that is, the closed unit ball has only two extreme points, the function $1_{[0,1]}$ and its negative $-1_{[0,1]}$. By (7.22), it is enough to consider $f \in C([0,1])$ such that $\|f\|_\infty = 1$, and to show that if $f \neq \pm 1_{[0,1]}$, then $f \notin \text{ext}B_{C([0,1])}$. Let therefore $f \in C([0,1])$ with $\|f\|_\infty = 1$ and $f \neq \pm 1_{[0,1]}$. Hence there exists $t_0 \in [0,1]$ such that $|f(t_0)| < 1$. Set $g = 1 - |f|$. It is easy to see that $\|f + g\|_\infty \leq 1$ and $\|f - g\|_\infty \leq 1$, and hence $f + g$ and $f - g$ belong to $\overline{B}_1(\mathbf{0})$. On the other hand, $g(t_0) \neq 0$, and so $g \neq \mathbf{0}$. By Exercise 13.0.50, we have $f \notin \text{ext}B_{C([0,1])}$, as desired. ▲

Next fundamental result shows that a convex and compact set can be reconstructed from its extreme points, by taking all their convex combinations.

Theorem 404 (Minkowski) *Let K be a convex and compact subset of a finite dimensional normed vector space V. Then:*

$$K = \text{co}(\text{ext}K). \tag{7.24}$$

Proof Luigi

Note that $\text{ext}(K)$ is the minimal set in K for which (7.24) holds: if $A \subseteq K$ is another set for which $K = co(A)$, then $\text{ext}(K) \subseteq A$. In fact, by definition the extreme points of K cannot be expressed as convex combinations of other vectors of the set K, and must therefore necessarily belong to such set A.

Hence:

- all the points of a compact and convex set K can be expressed as convex combinations of the extreme points;

- the set of the extreme points of K is the minimal set in K for which this is true.

We can conclude that Minkowski Theorem answers the question with which we started this Subsection, that is, the characterization of the minimal set of points of a convex set, whose convex combinations makes it is possible to reconstruct the convex set considered.

The Minkowski Theorem can be extended to infinite dimensional spaces, where it takes the name of Krein-Milman Theorem. We see a version of this result for normed vector spaces, though it is not the most interesting form in which this theorem appears.

Theorem 405 (Krein-Milman) *Let K be a convex and compact subset of a normed vector space V. Then:*

$$K = \overline{\text{co}}\left(\text{ext}K\right). \tag{7.25}$$

In Minkowski Theorem we only used the convex envelope and, therefore, the result obtained in (7.24) is stronger than what we have now in (7.25). In fact, convex envelopes are a simpler set to construct than closed and convex envelopes, as it has been observed in Proposition 382. On the other hand, Krein-Milman Theorem applies to any normed vector space, not necessarily finite dimensional.[14]

As to the minimality of $\text{ext}K$, in the more general context of Krein-Milman Theorem we have the following result, based on closures.

Theorem 406 *Let K be a convex and compact subset of a normed vector space V. Then, given a set $A \subseteq K$, we have:*

$$K = \overline{\text{co}}A \Longleftrightarrow \overline{A} \supseteq \text{ext}K.$$

Finally, observe that Example 403 shows that in the Krein-Milman Theorem the hypothesis that K is compact is crucial. In fact, in this example we saw that the set of the extreme points of the closed unit ball of the space $\left(C\left([0,1]\right), \|\cdot\|_\infty\right)$ is given by $\left\{\pm 1_{[0,1]}\right\}$. Hence,

$$\overline{\text{co}}\left(\text{ext}K\right) = \left\{\alpha 1_{[0,1]} : \alpha \in [-1,1]\right\} \neq \overline{B}_1\left(\mathbf{0}\right),$$

which does not contradict the Krein-Milman Theorem since $\left(C\left([0,1]\right), \|\cdot\|_\infty\right)$ is an infinite dimensional space and therefore, by Theorem 364, the convex set $B_{C([0,1])}$ is not compact.

[14] In this regard it is, however, necessary to remember Theorem 364, according to which in infinite dimensional spaces there are "few" sets that are compact in the metric induced by the norms. This is why before we said that Theorem 405 is not the most interesting version of Krein-Milman Theorem.

Chapter 8

Concavity

8.1 Definitions

8.1.1 Concavity

This chapter is devoted to concave functionals, a fundamental class of non linear functionals that have as their natural domain the convex sets studied in the previous chapter.

Definition 407 *A functional $f : C \to \mathbb{R}$ defined on a convex set C of a vector space is called concave if*

$$f(\lambda v + (1-\lambda) w) \geq \lambda f(v) + (1-\lambda) f(w), \tag{8.1}$$

for each $v, w \in C$ and each $\lambda \in [0,1]$, while it is called convex if

$$f(\lambda v + (1-\lambda) w) \leq \lambda f(v) + (1-\lambda) f(w), \tag{8.2}$$

for each $v, w \in C$ and each $\lambda \in [0,1]$.

Graphically, a functional is concave if the chord that joins any two points $(v, f(v))$ and $(w, f(w))$ of its graph lies below the graph of the functional itself, while it is convex if the opposite happens, i.e., if such chord lies above the graph of the functional.

Example 408 Each norm $\|\cdot\| : V \to \mathbb{R}$ is a convex functional.[1] In fact:

$$\|\lambda v + (1-\lambda) w\| \leq \|\lambda v\| + \|(1-\lambda) w\| = \lambda \|v\| + (1-\lambda) \|w\|, \tag{8.3}$$

for each $v, w \in V$ and each $\lambda \in [0,1]$. ▲

[1] Throughout the chapter, V denotes a vector space and C a convex subset of V.

Notice that a functional f is convex if and only if $-f$ is concave. This simple duality between the convexity and concavity of functionals implies that the properties of convex functionals can be immediately derived from those of concave functionals. For this reason we will consider only the properties of concave functionals.

An important subclass of concave functionals is given by *strictly concave* functionals, which are the functionals $f : C \to \mathbb{R}$ such that

$$f(\lambda v + (1-\lambda) w) > \lambda f(v) + (1-\lambda) f(w),$$

for each $v \neq w \in C$ and each $\lambda \in (0,1)$. In other words, the inequality (8.1) here is required to be strict, which implies that the graph of a strictly concave functional has no straight lines.

Similarly, a functional $f : C \to \mathbb{R}$ is *strictly convex* if

$$f(\lambda v + (1-\lambda) w) < \lambda f(v) + (1-\lambda) f(w),$$

for each $v \neq w \in C$ and each $\lambda \in (0,1)$. In particular, a functional is strictly convex if and only if $-f$ is strictly concave.

Since the inequalities (8.1) and (8.2) are weak, it is possible for a functional to be at the same time concave and convex. In such case, the functional is called affine. That is, a functional $f : C \to \mathbb{R}$ is *affine* if

$$f(\lambda v + (1-\lambda) w) = \lambda f(v) + (1-\lambda) f(w),$$

for each $v, w \in C$ and each $\lambda \in [0,1]$. Next result shows that affine functionals defined on vector spaces are nothing but translations of linear functionals.

Proposition 409 *A functional $f : V \to \mathbb{R}$ defined on a vector space V is affine if and only if there exist a linear functional $L : V \to \mathbb{R}$ and a scalar $\alpha \in \mathbb{R}$ such that $f(v) = L(v) + \alpha$ for each $v \in V$.*

Notice that $f = L + \alpha$ implies $f(\mathbf{0}) = \alpha$, and hence by Proposition 409 it follows that an affine functional f is linear if and only if $f(\mathbf{0}) = 0$. In other words, linear functionals can be viewed as affine functionals that become equal to zero at the neutral element $\mathbf{0}$.

Proof Let $L \in V'$. For each $v, w \in C$ and each $\lambda \in [0,1]$, we have:

$$\begin{aligned} f(\lambda v + (1-\lambda) w) &= L(\lambda v + (1-\lambda) w) + \alpha = \lambda L(v) + (1-\lambda) L(w) + \alpha \\ &= \lambda f(v) + (1-\lambda) f(w), \end{aligned}$$

and hence f is affine.

Viceversa, let $f : V \to \mathbb{R}$ be affine and set $L(v) = f(v) - f(\mathbf{0})$ for each $v \in V$. We prove that $L \in V'$. We start by proving that $L(\alpha v) = \alpha L(v)$ for each $v \in V$ and each $\alpha \in \mathbb{R}$. For each $\alpha \in [0,1]$ we have

$$\begin{aligned} L(\alpha v) &= f(\alpha v) - f(\mathbf{0}) = f(\alpha v + (1-\alpha)\mathbf{0}) - (1-\alpha) f(\mathbf{0}) - f(\mathbf{0}) \\ &= \alpha f(v) + (1-\alpha) f(\mathbf{0}) - (1-\alpha) f(\mathbf{0}) - f(\mathbf{0}) = \alpha f(v) - f(\mathbf{0}) \\ &= \alpha L(v). \end{aligned}$$

Now let $\alpha > 1$. Setting $w = \alpha v$, from what we just proved we have

$$L(v) = L\left(\frac{w}{\alpha}\right) = \frac{1}{\alpha} L(w),$$

and so $L(\alpha v) = \alpha L(v)$. On the other hand,

$$\begin{aligned} 0 &= L(\mathbf{0}) = L\left(\frac{1}{2}v - \frac{1}{2}v\right) = f\left(\frac{1}{2}v - \frac{1}{2}v\right) - f(\mathbf{0}) \\ &= \frac{1}{2} f(v) + \frac{1}{2} f(-v) - \frac{1}{2} f(\mathbf{0}) - \frac{1}{2} f(\mathbf{0}) \\ &= \frac{1}{2} L(v) + \frac{1}{2} L(-v), \end{aligned}$$

from which $L(v) = -L(-v)$. If $\alpha < 0$, we therefore have:

$$L(\alpha v) = L((-\alpha)(-v)) = (-\alpha) L(-v) = (-\alpha)(-L(v)) = \alpha L(v).$$

In conclusion, $L(\alpha v) = \alpha L(v)$ for each $v \in V$ and each $\alpha \in \mathbb{R}$. By Proposition 59, to complete the proof that $L \in V'$ we have to prove that $L(v+w) = L(v) + L(w)$ for each $v, w \in V$. We have:

$$\begin{aligned} L(v+w) &= 2L\left(\frac{v+w}{2}\right) = 2L\left(\frac{v}{2} + \frac{w}{2}\right) = 2\left(f\left(\frac{v}{2} + \frac{w}{2}\right) - f(\mathbf{0})\right) \\ &= 2\left(\frac{1}{2} f(v) + \frac{1}{2} f(w) - \frac{1}{2} f(\mathbf{0}) - \frac{1}{2} f(\mathbf{0})\right) = L(v) + L(w), \end{aligned}$$

as desired. ■

Example 410 By Theorem 65, a functional $f : \mathbb{R}^n \to \mathbb{R}$ is affine if and only if there exist $\chi \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$ such that $f(x) = \chi \cdot x + \alpha$ for each $x \in \mathbb{R}$. When $n = 1$, we have $f(x) = \beta x + \alpha$, with $\alpha, \beta \in \mathbb{R}$, that is, the affine functionals on $\mathbb{R}$ are the straight lines. ▲

Superlinear Functionals

A very important class of concave functionals are the *superlinear* functionals, which are the functionals $f : V \to \mathbb{R}$ defined on the whole space V and such that:

(i) $f(\alpha v) = \alpha f(v)$ for each $\alpha \geq 0$ and each $v \in V$,

(ii) $f(v + w) \geq f(v) + f(w)$ for each $v, w \in V$.

Property (i) is called positive homogeneity, while property (ii) is called superadditivity. Hence, a functional is superlinear if it is positively homogeneous and superadditive. Similarly, a functional $f : V \to \mathbb{R}$ is *sublinear* if it is positively homogeneous and subadditive, i.e., if $f(v + w) \leq f(v) + f(w)$ for each $v, w \in V$. We already introduced this notion in Definition 75, and it is immediate to see that f is sublinear if and only if $-f$ is superlinear.

Example 411 Norms $\|\cdot\| : V \to \mathbb{R}$ are sublinear functionals. ▲

Example 412 Consider the functional $f : V \to \mathbb{R}$ defined by

$$f(v) = \inf_{i \in I} L_i(v), \qquad \forall v \in V,$$

where $\{L_i\}_{i \in I}$ be a collection, finite or infinite, of linear functionals defined on a vector space V. This functional f is easily seen to be superlinear.

Superlinear functionals are concave (and so sublinear functionals are convex). In fact:

$$f(\lambda v + (1 - \lambda) w) \geq f(\lambda v) + f((1 - \lambda) w) = \lambda f(v) + (1 - \lambda) f(w),$$

for each $v, w \in V$ and each $\lambda \in [0, 1]$ (note the analogy with (8.3), obviously due to the sublinearity of the norm).

In the sequel superlinear functionals will play an important role. For this reason next we give some useful properties of superlinear functionals and we then provide a key characterization.

Lemma 413 *Let $f : V \to \mathbb{R}$ be a superlinear functional. Then, $f(\mathbf{0}) = 0$ and*

$$-f(-v) \geq f(v), \qquad \forall v \in V. \tag{8.4}$$

Furthermore, f is linear if and only if $f(-v) = -f(v)$ for each $v \in V$.

Proof Since f is positively homogeneous, we have $f(\alpha\mathbf{0}) = \alpha f(\mathbf{0})$ for each $\alpha \geq 0$. Since $\alpha\mathbf{0} = \mathbf{0}$, we therefore have $f(\mathbf{0}) = \alpha f(\mathbf{0})$ for each $\alpha \geq 0$, which can happen only if $f(\mathbf{0}) = 0$ (note that the argument is analogous to the one used in the proof of Proposition 60). Hence, for each $v \in V$, we have:

$$0 = f(\mathbf{0}) = f(v - v) \geq f(v) + f(-v),$$

which implies $-f(-v) \geq f(v)$, i.e., (8.4).

Obviously, if f is linear we have $f(-v) = -f(v)$ for each $v \in V$. We conclude the proof by showing that also the viceversa is true. Let $f(-v) = -f(v)$ for each $v \in V$. In Exercise 13.0.55 we considered the sublinear functional $\overline{f} : V \to \mathbb{R}$ defined as $\overline{f}(v) = -f(-v)$ for each $v \in V$. From $f(-v) = -f(v)$ it follows that $f(v) = \overline{f}(v)$ for each $v \in V$, and hence f is affine. By Proposition 409, there exist $L \in V'$ and $\alpha \in \mathbb{R}$ such that $f = L + \alpha$. On the other hand, $\alpha = f(\mathbf{0}) = 0$, and hence $f = L$, i.e., f is a linear functional. ■

Note that in the final part of the previous proof we implicitly proved also the following result.

Lemma 414 *A functional $f : V \to \mathbb{R}$ is both superlinear and sublinear if and only if is linear.*

The next result is a key characterization of superlinear functionals. It shows that they can be viewed as the lower envelopes of the linear functionals that pointwise dominate them.

Theorem 415 *A functional $f : V \to \mathbb{R}$ is superlinear if and only if*

$$f(v) = \min_{\{L \in V' : L \geq f\}} L(v), \qquad \forall v \in V. \tag{8.5}$$

If, in addition, f is continuous, then in (8.5) we can replace the algebraic dual V' with the topological dual V^*.

Proof. We prove the "only if" part, as the "if" follows from Example 412. Suppose f is superlinear and set $\Gamma = \{L \in V^* : L \geq f\}$. Let $v \in V$. Consider the vector subspace $M_v = \{\alpha v : \alpha \in \mathbb{R}\}$ generated by v (see Example 41). Define $L_v : M_v \to \mathbb{R}$ by $L_v(\alpha v) = \alpha f(v)$ for all $\alpha \in \mathbb{R}$. The functional L_v is linear on the vector subspace M_v. By the Hahn-Banach Theorem 77, there is $\widetilde{L} \in V'$ such that $\widetilde{L} \geq f$ on V and $\widetilde{L} = L_v$ on M_v. Hence, $\widetilde{L} \in \Gamma$ and $f(v) = \widetilde{L}(v)$. Consequently, $f(v) = \widetilde{L}(v) = \min_{\{L \in V' : L \geq f\}} L(v)$.

Finally, suppose f is continuous. By Theorem 432, f is lower bounded. Hence, given $v \in V$, there is a neighborhood $B_\varepsilon(v)$ and a constant $M_v \in \mathbb{R}$ such that $f(w) \geq M$

for all $w \in B_\varepsilon(v)$. Then, for all $L \in \Gamma$ it holds $L(w) \geq M$ for all $w \in B_\varepsilon(v)$. Again by Theorem 432, this implies that all $L \in \Gamma$ are continuous. Hence, $\Gamma \subseteq V^*$. ■

We close with a further characterization of the linearity of superlinear functionals.

Corollary 416 *A superlinear functional $f : V \to \mathbb{R}$ is linear (resp., continuous linear) if and only if there exists a unique $L \in V'$ (resp., $L \in V^*$) such that $L \geq f$.*

Proof. Suppose f is linear. Let $L \in V'$ be such that $L \geq f$. Then,

$$f(v) = -f(-v) \geq -L(-v) = L(v) \geq f(v), \qquad \forall v \in V,$$

and so $f = L$. Conversely, suppose there is a unique $L \in V'$ such that $L \geq f$. Then (8.5) implies $f = L$.

A similar argument proves the continuous case. ■

8.1.2 Lipschitzianity

We introduce now Lipschitzian functionals, another fundamental class of nonlinear functionals. Differently from convexity, which relies on the vector structure, Lipschitzianity can be introduced in any metric space.

Definition 417 *A function $f : A \subseteq X \to Y$ between two metric spaces X and Y is said to be **Lipschitz** on a subset B of A if there exists a positive scalar $M > 0$ such that*

$$d_Y(f(x_1), f(x_2)) \leq M d_X(x_1, x_2), \qquad \forall x_1, x_2 \in B. \tag{8.6}$$

A function is called Lipschitz, without further qualifications, when $A = B$; i.e. when (8.6) holds on all the domain of the function.

In Lipschitz functions the distance $d_Y(f(x_1), f(x_2))$ between the images of two points x_1 and x_2 is therefore controlled, through a positive coefficient M, by the distance $d_X(x_1, x_2)$ between the same two points x_1 and x_2.

In the case which we are interested in, we have that a functional $f : A \subseteq V \to \mathbb{R}$ defined on a subset A of a normed vector space V is Lipschitz on $B \subseteq A$ if there exists a positive scalar $M > 0$ such that

$$|f(v) - f(w)| \leq M \|v - w\|, \qquad \forall v, w \in B. \tag{8.7}$$

When $A = B$, the functional f is called Lipschitz (or Lipschitzian). Note that B is not required to be convex, exactly because Lipschitzianity is not based on the vector structure of V, but only on the metric induced by the norm.

Example 418 By (7.4), bounded linear functionals are Lipschitz. By (7.2), also the norms are Lipschitz functionals. ▲

Example 419 Let $C^1([a,b])$ be the vector space of the continuously differentiable functions $f : [a,b] \to \mathbb{R}$ (remember Definition 155). All these functions are Lipschitz. In fact, let $f \in C^1([a,b])$ and set $M = \max_{x\in[a,b]} |f'(x)|$. Since the derivative f' is continuous on $[a,b]$, by the Weierstrass Theorem the constant M is well defined.

On the other hand, given $x, y \in [a,b]$, by the Mean Value Theorem there exists $c \in [x,y]$ such that

$$\frac{f(x) - f(y)}{x - y} = f'(c).$$

Hence,

$$\frac{|f(x) - f(y)|}{|x - y|} = |f'(c)| \leq M,$$

and f is therefore Lipschitz. ▲

Example 420 Consider $f : \mathbb{R}_+ \to \mathbb{R}$ given by $f(x) = \sqrt{x}$ for each $x \geq 0$. This function is not Lipschitz. In fact,

$$\lim_{x\to 0+} \frac{f(x) - f(0)}{x - 0} = \lim_{x\to 0+} \frac{\sqrt{x}}{x} = \lim_{x\to 0+} \frac{1}{\sqrt{x}} = +\infty,$$

and, setting $y = 0$, it cannot exist $M > 0$ such that $|f(x) - f(y)| \leq M|x - y|$ for each $x, y \in \mathbb{R}_+$.

On the other hand, the previous Example shows that f is Lipschitz on each interval $[a,b]$ with $a, b > 0$. Hence, f is a function that is not of Lipschitz, i.e., (8.7) does not hold for any two points of its domain, but it is Lipschitz on appropriate subsets of the domain. ▲

Lipschitz functionals are obviously continuous. In fact, if $v^n \to v$, we have:

$$|f(v^n) - f(v)| \leq M\|v^n - v\| \to 0,$$

and hence $f(v^n) \to f(v)$. Next lemma shows that such functionals are actually uniformly continuous.[2]

Lemma 421 *A Lipschitz functional $f : A \subseteq V \to \mathbb{R}$ defined on a normed vector space V is uniformly continuous.*

[2]Even if this property, like other ones that we will see, holds for Lipschitz functions defined on metric spaces, we only consider the case that is here relevant of functionals defined on normed vector spaces.

Proof For each $\varepsilon > 0$, take $\delta_\varepsilon \in (0, \varepsilon/M)$. We have:

$$|f(v) - f(w)| \leq M \|v - w\| < \varepsilon$$

for each $v, w \in V$ such that $\|v - w\| < \delta_\varepsilon$. ■

Lipschitzianity is a global property since the constant M in (8.7) is required to be the same for each pair of vectors v and w in B. It is, however, possible to give a local version of Lipschitzianity.

Definition 422 *A functional $f : A \subseteq V \to \mathbb{R}$ is locally Lipschitz at a point $v_0 \in A$ if there exist a neighborhood $B_\varepsilon(v_0)$ and a positive scalar $M_{v_0} > 0$ such that*

$$|f(v) - f(w)| \leq M_{v_0} \|v - w\|, \qquad \forall v, w \in B_\varepsilon(v_0).$$

Notice the local character of this definition: the constant M_{v_0} depends on the particular point v_0 considered and, moreover, (8.7) is required to hold only among the points of a neighborhood of the point v_0 considered (and not between any two points of the domain).

When f is locally Lipschitz at each point of a set B we say that it is *locally Lipschitz on B*. Clearly, a functional that is locally Lipschitz on B is also continuous on B. It is also clear that a functional that is Lipschitz on B is also locally Lipschitz on B. On the contrary, since the constant M_{v_0} depends on the point v_0, a functional that is locally Lipschitz on B might well not be Lipschitz on B.

Example 423 Consider $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^2$ for each $x \in \mathbb{R}$. By proceeding as in Example 419 it is easy to see that f is locally Lipschitz at each $x \in \mathbb{R}$. But, f is not Lipschitz. In fact, if it were so there would exist M such that

$$\left|x^2 - y^2\right| \leq M |x - y|, \qquad \forall x, y \in \mathbb{R},$$

i.e., such that $|x + y| \leq M$ for each $x, y \in \mathbb{R}$ with $x \neq y$, which is impossible. ▲

There is, however, an important case where local Lipschitzianity implies the global one.

Proposition 424 *Let $f : A \subseteq V \to \mathbb{R}$ be a functional defined on a subset A of a normed vector space V, and let $K \subseteq A$ be a compact subset. Then, f is Lipschitz on K if and only if it is locally Lipschitz on K.*

Proof Since the "Only if " is obvious, we prove the "If." Let f be locally Lipschitz at each point of K. For each $v_0 \in K$ there exist a neighborhood $B_\varepsilon(v_0)$ and a constant $M_{v_0} > 0$ such that

$$|f(v) - f(w)| \leq M_{v_0} \|v - w\|, \qquad \forall v, w \in B_\varepsilon(v_0). \tag{8.8}$$

The collection $\{B_\varepsilon(v_0)\}_{v_0 \in K}$ is an open cover of K, and therefore there exists a finite subcover $\{B_\varepsilon(v_0^i)\}_{i=1}^n$. Let $M = \max_{i=1,\ldots,n} M_{v_0^i}$. Since $K \subseteq \bigcup_{i=1}^n B_\varepsilon(v_0^i)$, (8.8) implies that

$$|f(v) - f(w)| \leq M_{v_0} \|v - w\| \leq M \|v - w\| \qquad \forall v, w \in K,$$

as desired. ■

8.2 First Properties

Consider in the set

$$\operatorname{ipo}(f) = \{(v, t) \in C \times \mathbb{R} : f(v) \geq t\}, \tag{8.9}$$

called the *ipograph* of f, which consists of the points (v, t) that lie below the graph $G(f)$ of the function f, which we remind is given by $G(f) = \{(v, t) \in C \times \mathbb{R} : f(v) = t\}$.

The next result shows that the concavity of f is equivalent to the convexity of its ipograph. This provides a simple characterization of concave functionals through convex sets.

Proposition 425 *A functional $f : C \to \mathbb{R}$ defined on a convex set C of a vector space is concave if and only if its ipograph* $\operatorname{ipo}(f)$ *is a convex set in $C \times \mathbb{R}$.*

Proof Let f be concave, and let $(v, t), (w, s) \in \operatorname{ipo}(f)$. By definition, $t \leq f(v)$ and $s \leq f(w)$, from which:

$$\lambda t + (1 - \lambda) s \leq \lambda f(v) + (1 - \lambda) f(w) \leq f(\lambda v + (1 - \lambda) w),$$

for each $\lambda \in [0, 1]$. Hence, $(\lambda v + (1 - \lambda) w, \lambda t + (1 - \lambda) s) \in \operatorname{ipo}(f)$, which proves that $\operatorname{ipo}(f)$ is convex.

Viceversa, suppose that $\operatorname{ipo}(f)$ is convex. Hence, for each $v, w \in C$ and $\lambda \in [0, 1]$,

$$(\lambda v + (1 - \lambda) w, \lambda f(v) + (1 - \lambda) f(w)) \in \operatorname{ipo} f$$

that is,

$$\lambda f(v) + (1 - \lambda) f(w) \leq f(\lambda v + (1 - \lambda) w),$$

as desired. ■

Though concavity is defined through convex combinations of only two elements, next we show that it actually holds for all convex combinations.

Proposition 426 *A functional $f : C \to \mathbb{R}$ defined on a convex set C of a vector space is concave if and only if, for each finite collection $\{v^1, ..., v^n\}$ of elements of C, we have*

$$f\left(\sum_{i=1}^{n} \lambda_i v^i\right) \geq \sum_{i=1}^{n} \lambda_i f\left(v^i\right) \tag{8.10}$$

for each $\lambda_i \geq 0$ and $\sum_{i=1}^{n} \lambda_i = 1$.

Expression (8.10) is known as *Jensen inequality*.

Proof The "If" is obvious. As to the "Only if" we proceed, as for Lemma 378, by induction on n. Let f be concave. Expression (8.10) obviously holds for $n = 2$. Suppose that it holds for $n-1$, i.e., that $f\left(\sum_{i=1}^{n-1} \lambda_i v^i\right) \geq \sum_{i=1}^{n-1} \lambda_i f\left(v^i\right)$ for each convex combination of $n-1$ elements of C. If $\lambda_n = 1$, (8.10) trivially holds. Let $\lambda_n < 1$. We have:

$$\begin{aligned} f\left(\sum_{i=1}^{n} \lambda_i v^i\right) &= f\left(\sum_{i=1}^{n-1} \lambda_i v^i + \lambda_n v^n\right) = f\left((1-\lambda_n)\sum_{i=1}^{n-1} \frac{\lambda_i}{1-\lambda_n} v^i + \lambda_n v^n\right) \\ &\geq (1-\lambda_n) f\left(\sum_{i=1}^{n-1} \frac{\lambda_i}{1-\lambda_n} v^i\right) + \lambda_n f\left(v^n\right) \\ &\geq (1-\lambda_n) \sum_{i=1}^{n-1} \frac{\lambda_i}{1-\lambda_n} f\left(v^i\right) + \lambda_n f\left(v^n\right) = \sum_{i=1}^{n} \lambda_i f\left(v^i\right), \end{aligned}$$

as desired. ■

Turn now to the properties of the space of concave functionals. Given two functionals $f, g : C \to \mathbb{R}$, the minimum functional $f \wedge g : C \to \mathbb{R}$ is defined by $(f \wedge g)(v) = \min\{f(v), g(v)\}$ for each $v \in V$.

Proposition 427 *Let $f, g : C \to \mathbb{R}$ be two concave functionals defined on a convex set C of a vector space. The functionals $f + g$ and $f \wedge g$ are concave, while αf is concave provided $\alpha \geq 0$.*

The proof of this result is left to the reader. Note that the space of concave functions is a convex cone (see Section 11.2) but not a vector space. In fact, it is not closed with respect to scalar multiplication, except when the scalar is non-negative.

Given a function $f : C \to \mathbb{R}$ defined on a convex set, its concavity on C is intuitively closely related to its concavity on all line segments $[v, w] = \{tv + (1-t)w : t \in [0, 1]\}$ determined by vectors v and w that belong to C. Proposition 429 will make precise this intuition that is important both conceptually, to better understand the scope of

concavity, and operationally since the restrictions on line segments of f are scalar functions, in general much easier to study than the original function f.

Given a convex set C and $v, w \in C$, set $C_{v,w} = \{t \in \mathbb{R} : (1-t)\, v + tw \in C\}$. That is, $C_{v,w}$ is the set of all t values such that $(1-t)\, v + tw \in C$. Clearly, $[0,1] \subseteq C_{v,w}$.

Lemma 428 *$C_{v,w}$ is an open interval when C is open convex.*

Proof. Suppose C is open convex. Let $t \in [v,w]_C$, i.e., $(1-t)\, v + tw \in C$. We want to show that t an interior point. Since C is open, there exists $B_\varepsilon((1-t)\, v + tw)$ such that $B_\varepsilon((1-t)\, v + tw) \subseteq C$. Then, given any $0 < \delta < \varepsilon / \|v - w\|$, by Exercise 13.0.59,

$$\begin{aligned}(1-t-\delta)\, v + (t+\delta)\, w &= (1-t)\, v + tw + \delta(w - v) \in B_\varepsilon((1-t)\, v + tw),\\ (1-t+\delta)\, v + (t-\delta)\, w &= (1-t)\, v + tw + \delta(v - w) \in B_\varepsilon((1-t)\, v + tw)\end{aligned}$$

Hence, $(t - \varepsilon/\|v-w\|, t + \varepsilon/\|v-w\|) \subseteq C_{v,w}$, as desired. ■

Define $\phi_{v,w} : C_{v,w} \to \mathbb{R}$ by

$$\phi_{v,w}(t) = f((1-t)\, v + tw). \tag{8.11}$$

Proposition 429 *For a function $f : C \to \mathbb{R}$ defined on a convex set C of a vector space, the following properties are equivalent:*

(i) f is concave (resp., strictly concave);

(ii) $\phi_{v,w}$ is concave (resp., strictly concave) for all $v, w \in C$;

(iii) $\phi_{v,w}$ is concave (resp., strictly concave) on $[0,1]$ for all $v, w \in C$.

Proof. We consider the concave case, and leave to the reader the strictly concave one.

(i) implies (ii). Suppose f is concave. Let $v, w \in C$ and $t_1, t_2 \in C_{v,w}$. Then, for each $\alpha \in [0,1]$,

$$\begin{aligned}\phi_{v,w}(\alpha t_1 + (1-\alpha)\, t_2) &= f((1 - (\alpha t_1 + (1-\alpha)\, t_2))\, v + (\alpha t_1 + (1-\alpha)\, t_2)\, w)\\ &= f(\alpha((1-t_1)\, v + t_1 w) + (1-\alpha)((1-t_2)\, v + t_2 w))\\ &\geq \alpha f((1-t_1)\, v + t_1 w) + (1-\alpha)\, f((1-t_2)\, v + t_2 w)\\ &= \alpha\phi_{v,w}(t_1) + (1-\alpha)\, \phi_{v,w}(t_2)\end{aligned}$$

and so $\phi_{v,w}$ is concave.

Since (ii) trivially implies (iii), it remains to prove that (iii) implies (i). Let $v, w \in C$. Since $\phi_{v,w}$ is concave on $[0,1]$, we have

$$f((1-t)\, v + tw) = \phi_{v,w}(t) \geq t\phi_{v,w}(1) + (1-t)\, \phi_{v,w}(0) = (1-t)\, f(v) + tf(w),$$

for all $t \in [0,1]$, as desired. ■

We close with a property of convex sets of normed vector spaces.

Lemma 430 *Let C be a convex set of a normed vector space. Then, its interior $\overset{\circ}{C}$ is convex.*

Proof. Let $U = \{v \in V : \|v\| < 1\}$. By Exercise 13.0.59,

$$B_\varepsilon(v) = \{v + w : \|w\| < \varepsilon\} = v + \varepsilon U, \qquad \forall v \in V. \tag{8.12}$$

Hence, given any $v \in \overset{\circ}{C}$, there is $\varepsilon > 0$ small enough so that $v + \varepsilon U \subseteq C$. Therefore, given any $v_1, v_2 \in \overset{\circ}{C}$, there is $\varepsilon > 0$ small enough so that both $v_1 + \varepsilon U \subseteq C$ and $v_2 + \varepsilon U \subseteq C$. Then, for all $t \in [0,1]$, we have

$$tv_1 + (1-t)v_2 + \varepsilon U = t(v_1 + \varepsilon U) + (1-t)(v_2 + \varepsilon U) \subseteq C,$$

Hence, by (8.12), $B_\varepsilon(tv_1 + (1-t)v_2) \subseteq C$, and so $tv_1 + (1-t)v_2 \in \overset{\circ}{C}$. ■

In view of this lemma, when in the sequel we will state results for concave functions defined on open convex sets, they actually hold on the interior points of any convex set, possibly not open, on which a concave function is defined. This should be kept in mind when reading the results of the rest of the chapter, which for convenience will be often stated in terms of concave functions defined on open convex sets (see for example the discussion after Proposition 454).

8.3 Continuity

Concave functionals have remarkable continuity properties, similar to those enjoyed by linear functionals. Since the latter are a very particular class of concave functionals, these properties then turn out to hold for a much larger class of functionals.

We begin by introducing a property that plays a key role in the study of continuity of concave functionals. A functional $f : C \to \mathbb{R}$ is said to be *lower bounded at* $v \in C$ if there exist a neighborhood $B_\varepsilon(v)$ of v and a constant $M \in \mathbb{R}$ such that $f(w) \geq M$ for each $w \in B_\varepsilon(v)$. The functional is *lower bounded on* C if it is lower bounded at each $v \in C$.

Example 431 Consider a linear functional $L : V \to \mathbb{R}$ defined on a normed vector space V. If L is bounded in the sense of Definition 339, then it is lower bounded on V. In fact, let $v \in V$ and let $B_\varepsilon(v)$ be a neighborhood of v. By (7.6), we have:

$$|L(w) - L(v)| \leq \|L\| \|w - v\| \leq \|L\| \varepsilon, \qquad \forall w \in B_\varepsilon(v),$$

and so $L(w) \geq L(v) - \|L\| \varepsilon$ for each $w \in B_\varepsilon(v)$. ▲

The next result is truly remarkable: a concave function defined on an open convex set C that is just lower bounded at some point of C turns out to be automatically locally Lipschitz (and so continuous) on the entire C. Thus, a property that in general is much weaker than continuity, let alone than locally Lipschitzianity, becomes equivalent to it when f is concave.

Theorem 432 *Let $f : C \to \mathbb{R}$ be a concave functional defined on an open convex set C of a normed vector space. Then, the following properties are equivalent:*

(i) f is lower bounded at some point of C;

(ii) f is continuous at some point of C;

(iii) f is locally Lipschitz on C.

Proof.[3](ii) implies (i). If f is continuous at $v_0 \in C$, then f is locally bounded at v_0. Actually there exists a neighborhood $B_\varepsilon(v_0) \subseteq C$ such that $|f(v) - f(v_0)| \leq 1$ for all $v \in B_\varepsilon(v_0)$. In particular, $f(v) = f(v) - f(v_0) + f(v_0) \geq f(v_0) - 1$.

As (iii) trivially implies (ii), it remains to prove that (i) implies (ii). Suppose f is lower bounded at $v_0 \in C$, i.e., there exists $M \in \mathbb{R}$ and a neighborhood $B_\varepsilon(v_0)$ such that $f(w) \geq M$ for all $w \in B_\varepsilon(v_0)$. We divide the proof in three steps.

Step 1: f is bounded above on $B_\varepsilon(v_0)$. For, let $w \in B_\varepsilon(v_0)$. Consider the point $z = 2v_0 - w = v_0 - (w - z_0)$. Clearly, $z \in B_\varepsilon(v_0)$ and v_0 is the mid-point between z and w. By concavity,

$$f(v_0) = f\left(\frac{1}{2}z + \frac{1}{2}w\right) \geq \frac{1}{2}f(z) + \frac{1}{2}f(w).$$

We conclude that

$$f(w) \leq 2f(v_0) - f(z) \leq 2f(v_0) - M, \qquad \forall w \in B_\varepsilon(v_0),$$

and this completes the proof of Step 1.

Step 2: f is is locally bounded on C, i.e., at all $v \in C$. By Step 1, f is bounded at v_0, i.e., there exists $M \in \mathbb{R}$ and a neighborhood $B_\varepsilon(v_0)$ such that $|f(w)| \leq M$ for all $w \in B_\varepsilon(v_0)$. We now prove that f is is locally bounded at all $v \in C$. Since C is open, by Lemma 428 there is $z \in C_{v_0,v}$ such that $z = (1-t)v_0 + tv$ with $t > 1$. For each $w \in B_\varepsilon(v_0)$, we have

$$\begin{aligned}\frac{1}{t}z + \left(1 - \frac{1}{t}\right)w &= \frac{1}{t}((1-t)v_0 + tv) + \left(1 - \frac{1}{t}\right)w = \\ &= v + \left(1 - \frac{1}{t}\right)(w - v_0)\end{aligned}$$

[3]The proof is based on Roberts and Varberg (1974).

and so, by Exercise 13.0.59,

$$B_{\left(1-\frac{1}{t}\right)\varepsilon} = \left\{ \frac{1}{t}z + \left(1-\frac{1}{t}\right) w : w \in B_\varepsilon (v_0) \right\}.$$

By concavity,

$$f\left(\frac{1}{t}z + \left(1-\frac{1}{t}\right) w\right) \geq \frac{1}{t} f(z) + \left(1-\frac{1}{t}\right) f(w) \geq \frac{1}{t} f(z) + \left(1-\frac{1}{t}\right) M,$$

and so f is bounded below on the neighborhood $B_{\left(1-\frac{1}{t}\right)\varepsilon}$.

Step 3: We want to show that f is locally Lipschitz at any $v \in C$. By Step 2, f is locally bounded at v, i.e., there exists $M \in \mathbb{R}$ and a neighborhood $B_{2\varepsilon}(v)$, wlog of radius 2ε, such that $|f(w)| \leq M$ for all $w \in B_{2\varepsilon}(v)$. Given $w_1, w_2 \in B_{2\varepsilon}(v)$, set

$$w_3 = w_2 + \frac{\varepsilon}{\|w_2 - w_1\|} (w_2 - w_1).$$

Then, $w_3 \in B_{2\varepsilon}(v)$ since

$$\|w_3 - v\| = \left\| w_3 - w_2 + \frac{\varepsilon}{\|w_2 - w_1\|} (w_2 - w_1) \right\| \leq 2\varepsilon.$$

Since

$$w_2 = \frac{\varepsilon}{\|w_2 - w_1\| + \varepsilon} w_1 + \frac{\|w_2 - w_1\|}{\|w_2 - w_1\| + \varepsilon} w_3,$$

concavity implies

$$f(w_2) \geq \frac{\varepsilon}{\|w_2 - w_1\| + \varepsilon} f(w_1) + \frac{\|w_2 - w_1\|}{\|w_2 - w_1\| + \varepsilon} f(w_3),$$

so that

$$f(w_1) - f(w_2) \leq \frac{\|w_2 - w_1\|}{\|w_2 - w_1\| + \varepsilon} (f(w_1) - f(w_3)) \leq \frac{\|w_2 - w_1\|}{\varepsilon} 2M. \tag{8.13}$$

Interchanging the roles of w_1 and w_2, we get

$$f(w_2) - f(w_1) \leq \frac{\|w_1 - w_2\|}{\|w_1 - w_2\| + \varepsilon} (f(w_2) - f(w_3)) \leq \frac{\|w_1 - w_2\|}{\varepsilon} 2M.$$

Along with (8.13), this implies

$$|f(w_1) - f(w_2)| \leq \frac{2M}{\varepsilon} \|w_1 - w_2\|,$$

and so f is locally Lipschitz at v. ■

Thanks to Theorem 432, the following properties are thus equivalent for a concave functional $f : C \to \mathbb{R}$ defined on an open and convex set C:

(i) f is lower bounded at a point v of C;

(ii) f is continuous at a point $v \in C$;

(iii) f is locally Lipschitz at a point v of C

(iv) f is continuous on C;

(v) f is locally Lipschitz on C.

The equivalence between (i)-(v) is an impressive property of concavity, which generalizes to concave functionals what established in Proposition 338 and Theorem 342 for linear functionals.

Turn now to the finite dimensional case. Here Theorem 432 takes an even more remarkable form since the next lemma shows that lower boundedness is always satisfied in the finite dimensional case.

Lemma 433 *A concave functional $f : C \to \mathbb{R}$ defined on an open convex set C of a finite dimensional normed vector space is lower bounded at a point v of C.*

Proof. By Corollary 105, wlog consider $\mathbb{R}^n$. Let $v \in C$. Since C is open, there is a neighborhood $B_\varepsilon(v)$ of v such that $B_\varepsilon(v) \subseteq C$. By Exercise 13.0.59, for $0 < \delta < \varepsilon$ $v + \delta e^i \in C$ for all $i = 1, ..., n$. Consider the convex hull $co(v, v + \delta e^1, ..., v + \delta e^n)$. Since C is convex, we have $co(v, v + \delta e^1, ..., v + \delta e^n) \subseteq C$. To ease notation, denote by G the interior of $co(v, v + \delta e^1, ..., v + \delta e^n)$. By Lemma 430, G is convex; by Exercise 13.0.54, G is nonempty.

Let $w \in G$. There is $\{t_i\}_{i=0}^n$, with each $t_i \geq 0$ and $\sum_{i=0}^n t_i = 1$, such that $w = t_0 v + \sum_{i=1}^n t_i (v + \delta e^i)$. By concavity,

$$f(w) \geq t_0 f(v) + \sum_{i=1}^n t_i f\left(v + \delta e^i\right) \geq \min_{i=1,...,n} \left\{f(v), f\left(v + \delta e^i\right)\right\}, \qquad \forall w \in G,$$

and so f is bounded on G. In particular, f is locally bounded at any point of G. ■

Thanks to Lemma 433, Theorem 432 implies that in finite dimensional vector spaces concave functions are always locally Lipschitz on open convex sets. This is a far reaching generalization of Corollary 360, which showed a similar properties for the very special case of linear functionals.

Corollary 434 *Let $f : C \to \mathbb{R}$ be a concave functional defined on an open convex set C of a finite dimensional normed vector space. Then, f is locally Lipschitz on C and is Lipschitz on each compact $K \subseteq C$.*

The next simple example shows that Corollary 434 does not hold for closed and convex sets.

Example 435 Let $f : [-1, 1] \to R$ be defined by:

$$f(x) = \begin{cases} 2 - x^2 & \text{if } x \in (-1, 1) \\ 0 & \text{if } x \in \{-1, 1\} \end{cases}$$

This function is concave on the entire domain $[-1, 1]$, and it is discontinuous at -1 and 1, i.e., at the boundary points of the domain (actually it is lower semicontinuous at $x \in \{-1, 1\}$). According to Corollary 434, f is instead continuous when considered on the open interval $(-1, 1)$. ▲

Next examples further illustrate Corollary 434.

Example 436 Consider $f : \mathbb{R}_{++} \to \mathbb{R}$ given by $f(x) = \sqrt{x}$ for each $x > 0$. It is the restriction on $(0, +\infty)$ of the function $\sqrt{x}$ studied in Example 420. By Corollary 434, f is locally Lipschitz at each point of the domain. On the other hand, by proceeding as in Example 420 it is easy to see that f is not Lipschitz on the entire domain. By Corollary 434, f is Lipschitz on each compact contained in the domain. In fact, in Example 420 we observed how f is Lipschitz on each closed and bounded interval $[a, b]$, with $a, b > 0$. ▲

Example 437 Consider $f : \mathbb{R}_{++} \to \mathbb{R}$ given by $f(x) = \lg x$ for each $x > 0$. By Corollary 434, f is locally Lipschitz at each point of the domain, and it is Lipschitz on each compact subset of the domain, for example on the closed and bounded intervals $[a, b]$ with $a, b > 0$. On the other hand, also this function is not Lipschitz on the entire domain. In fact, we have:

$$\lim_{x \to 0+} \frac{|f(x) - f(y)|}{|x - y|} = \lim_{x \to 0+} \frac{|\lg x - \lg y|}{|x - y|} = +\infty,$$

and therefore there does not exist $M > 0$ such that $|f(x) - f(y)| \leq M |x - y|$ for each $x, y \in \mathbb{R}_{++}$. ▲

8.4 Differentiability

8.4.1 Directional Derivatives

Concave functionals have important differential properties. We begin by studying the directional derivatives. Corollary 131 showed how directional derivatives are positively homogeneous functionals. The next fundamental result shows that for concave

functionals such result can be improved in two important aspects: the directional derivative in the concave case exists along all directions and is a superlinear functional, i.e., positively homogeneous and superadditive.

Theorem 438 *Let $f : C \to \mathbb{R}$ be a concave functional defined on an open and convex set C of a normed vector space V. Given any $v \in C$, the directional derivative*

$$f'(v; w) = \lim_{t \to 0+} \frac{f(v + tw) - f(v)}{t} \tag{8.14}$$

exists at each direction $w \in V$. Moreover:

(i) $f'(v; \cdot) : V \to \mathbb{R}$ is a superlinear functional;

(ii) $f'(v; \cdot)$ is continuous on V provided f is continuous at v.

The proof relies on the following important lemma, which shows that the difference quotient is decreasing.

Lemma 439 *Let $f : C \to \mathbb{R}$ be a concave functional defined on a convex set C of a vector space V. Given any $v \in C$ and $w \in V$, then the function*

$$0 < t \longmapsto \frac{f(v + tw) - f(v)}{t} \tag{8.15}$$

is decreasing for all $t > 0$ values such that $v + tw \in C$.

Proof. Let $v \in C$. Assume first that $v = \mathbf{0}$ and $f(\mathbf{0}) = 0$. Fix $w \in V$ and let $0 < t_1 < t_2$. By concavity,

$$f(t_1 w) = f\left(\frac{t_1}{t_2} t_2 w\right) \geq \frac{t_1}{t_2} f(t_2 w) + \left(1 - \frac{t_1}{t_2}\right) f(\mathbf{0}) = \frac{t_1}{t_2} f(t_2 w),$$

and so $f(t_1 w)/t_1 \geq f(t_2 w)/t_2$. To complete the proof, define $g : C - \{v\} \to \mathbb{R}$ by $g(z) = f(z + v) - f(v)$ for all $z \in C - \{v\}$. Then, $g(\mathbf{0}) = 0$ and $g(tw)/t = (f(v + tw) - f(v))/t$. We can conclude that the difference quotient (8.15) has the desired properties. ■

Proof of Theorem 438. By Lemma 439, the difference quotient is decreasing, and so the limit (8.14) exists and

$$\lim_{t \to 0^+} \frac{f(v + tw) - f(v)}{t} = \sup_{t > 0} \frac{f(v + tw) - f(v)}{t}.$$

It remains to show that such a limit is finite. Note that the scalar function $t \to f(v + tw)$ is concave and, as C is open and $v \in C$, it is defined at least in a complete

neighborhood of the origin. Therefore $t \to f(v+tw)$ is locally Lipschitz at 0. Hence, $|f(v+tw) - f(v)| \leq Lt$ for some L and for small values of t. Therefore,

$$\frac{f(v+tw) - f(v)}{t} \leq L$$

and the right-hand side limit is finite and, in turn, $f'(v;w)$ exists for all $w \in V$.

To prove (i), observe that the positive homogeneity of $f'(v;\cdot) : V \to \mathbb{R}$ follows from Corollary 131. Observe that, for each $\alpha \in [0,1]$,

$$\begin{aligned} & \frac{f(v + t(\alpha w_1 + (1-\alpha) w_2)) - f(v)}{t} \\ \geq & \frac{\alpha(f(v+tw_1) - f(v)) + (1-\alpha)(f(v+tw_2) - f(v))}{t} \end{aligned}$$

Taking limits as $t \to 0+$, this implies that $f'(v;\cdot) : V \to \mathbb{R}$ is concave. Hence,

$$\begin{aligned} f'(v; w_1 + w_2) &= f'\left(v; 2\frac{w_1+w_2}{2}\right) = 2f'\left(v; \frac{w_1+w_2}{2}\right) \\ &\geq 2\left(\frac{f'(v;w_1)}{2} + \frac{f'(v;w_2)}{2}\right) = f'(v;w_1) + f'(v;w_2). \end{aligned}$$

This shows that $f'(v;\cdot) : V \to \mathbb{R}$ is superadditive, and so superlinear.

It remains to show (ii). Let f be continuous at $v \in C$. By Theorem 432, f is locally Lipschitz. Hence, there exists $B_\varepsilon(v)$ and M_v such that, if $w \in V$,

$$|f(v+tw) - f(v)| \leq M_v t \|w\|$$

provided t is small enough so that $v + tw \in B_\varepsilon(v)$. Thus, $|f'(v;w)| \leq M_v \|w\|$ for all $w \in V$, and so $f'(v;\cdot) : V \to \mathbb{R}$ is continuous. In fact, let $w_n \to w$. Then,

$$-M_v \|w_n - w\| \leq f'(v; w_n - w) \leq M_v \|w_n - w\|,$$

which implies $f'(v; w_n - w) \to 0$. ■

By Lemma 413, we therefore have that

$$-f'(v;-w) \geq f'(v;w), \qquad \forall w \in V,$$

and that $f'(v;\cdot) : V \to \mathbb{R}$ is linear if and only if $f'(v;-w) = -f'(v;w)$ for each $w \in V$.

Recalling what we saw in Chapter 4 about (4.11), it is easy to see that

$$-f'(v;-w) = \lim_{t \to 0-} \frac{f(v+tw) - f(v)}{t}.$$

Hence, condition $f'(v;-w) = -f'(v;w)$ is nothing but the equality between the right and left limit of the difference quotient

$$\frac{f(v+tw) - f(v)}{t},$$

i.e., its bilateral limit. Thus:

Corollary 440 *Let $f : C \to \mathbb{R}$ be a concave and continuous functional defined on an open and convex set C of a normed vector space V. Then, f is Gateaux differentiable at a point $v \in C$ if and only if the bilateral limit*

$$f'(v; w) = \lim_{t \to 0} \frac{f(v + tw) - f(v)}{t} \tag{8.16}$$

exists finite for each $w \in V$.

Proof Fix $v \in V$. For each $w \in V$ we have:

$$\begin{aligned} -f'(v; -w) &= -\lim_{t \to 0+} \frac{f(v + t(-w)) - f(v)}{t} = \lim_{t \to 0+} \frac{f(v + t(-w)) - f(v)}{-t} \\ &= \lim_{t \to 0-} \frac{f(v + tw) - f(v)}{t}. \end{aligned}$$

Hence, $f'(v; \cdot) : V \to \mathbb{R}$ is linear if and only if

$$\lim_{t \to 0+} \frac{f(v + tw) - f(v)}{t} = f'(v; w) = -f'(v; -w) = \lim_{t \to 0-} \frac{f(v + tw) - f(v)}{t},$$

i.e., if and only if the bilateral limit

$$f'(v; w) = \lim_{t \to 0} \frac{f(v + tw) - f(v)}{t}$$

exists finite for each $w \in V$.

If f is lower bounded at v, by Theorem 438 $f'(v; \cdot) : V \to \mathbb{R}$ is continuous on V. Hence, if the limit (8.16) exists, we can conclude that $f'(v; \cdot) \in V'$, i.e., f is Gateaux differentiable at v. ■

The next result gives an important characterization of Gateaux differentiability of concave functionals.

Theorem 441 *Let $f : C \to \mathbb{R}$ be a concave and continuous functional defined on an open and convex subset C of a normed vector space V. Then, f is Gateaux differentiable at a point $v \in C$ if and only if there exists a unique functional $L \in V^*$ such that*

$$f(w) \leq f(v) + L(w - v), \qquad \forall w \in C, \tag{8.17}$$

or, equivalently,

$$f'(v; z) \leq L(z) \qquad \forall z \in V, \tag{8.18}$$

In this case, $f'(v; w) = L(w)$ for each $w \in V$.

We first prove as a separate lemma the equivalence of conditions (8.17) and (8.18), an important fact that will be used later in the chapter.

Lemma 442 *Let $f : C \to \mathbb{R}$ be a concave functional defined on an open convex subset C of a normed vector space V. Then, $L \in V^*$ satisfies (8.17) if and only if satisfies (8.18).*

Proof. Suppose $L \in V^*$ satisfies (8.17). Let $z \in V$. For $t > 0$ small enough we have $v + tz \in C$, and so

$$tL(z) = L((v + tz) - v) \geq f(v + tz) - f(v).$$

This implies that L satisfies (8.17). Conversely, suppose $L \in V^*$ satisfies (8.18). Let $w \in C$ and consider $t > 0$ small enough so that $v + t(w - v) \in C$. Then, by Lemma 439,

$$L(w - v) \geq f'(v; w - v) \geq \frac{f(v + t(w - v)) - f(v)}{t}, \tag{8.19}$$

which is (8.17) when $t = 1$. This completes the proof. ■

Proof of Theorem 441. Let $v \in C$. By Theorem 438, $f'(v; \cdot) : V \to \mathbb{R}$ is superlinear. Hence, by Corollary 416 $f'(v; \cdot) : V \to \mathbb{R}$ is a continuous linear functional (and so f is Gateaux differentiable at v) if and only if there exists a unique $L \in V^*$ such that $L(w) \geq f'(v; w)$ for all $w \in V$, i.e., a a unique $L \in V^*$ that satisfies condition (8.18). To complete the proof it remains to observe that, by Lemma 442, conditions (8.17) and (8.18) are equivalent. ■

Therefore, a necessary and sufficient condition for f to be Gateaux differentiable at v is that there exists a unique linear and continuous functional $L : V \to \mathbb{R}$ that satisfies either of the equivalent conditions (8.17) and (8.18). When exists, this unique linear functional is exactly the Gateaux differential $f'(v; \cdot)$ of f at v, for which the following inequality thus holds:

$$f(w) \leq f(v) + f'(v; w - v), \qquad \forall w \in C. \tag{8.20}$$

Inequality (8.20) is a noteworthy property of Gateaux differentials of concave functions.

The Euclidean Case

In view of Theorem 65, in the special case $V = \mathbb{R}^n$ Theorem 441 takes the following form, where the continuity of f is no longer required since, by Corollary 434, it is already implied by concavity.

Corollary 443 *Let $f : C \to \mathbb{R}$ be a concave functional defined on an open convex set C of $\mathbb{R}^n$. Then, f is Gateaux differentiable at $x \in C$ if and only if there exists a unique vector $\chi \in \mathbb{R}^n$ such that*

$$f(y) \leq f(x) + \chi \cdot (y - x), \qquad \forall y \in C,$$

or, equivalently,

$$f'(x;z) \leq \chi \cdot z, \qquad \forall z \in \mathbb{R}^n.$$

In this case, $\chi = \nabla f(x)$.

By Theorem 140, in $\mathbb{R}^n$ we have $f'(x;y) = \nabla f(x) \cdot y$. Therefore, in this case (8.20) becomes:

$$f(y) \leq f(x) + \nabla f(x) \cdot (y - x), \qquad \forall y \in C,$$

which in this version can be regarded as a remarkable property of the gradient of concave functions.

We close by showing that in $\mathbb{R}^n$ the Gateaux and Frechet differentiability of concave functions are equivalent notions, which are in turn equivalent to the mere existence of the partial derivatives. Thus, for concave functions a substantially stronger version of Theorem 150 holds.

Proposition 444 *Let* $f : C \to \mathbb{R}$ *be a concave functional defined on an open convex subset* C *of* $\mathbb{R}^n$. *Given* $x \in C$, *the following properties are equivalent:*

(i) f *is Gateaux differentiable at* x;

(ii) f *is Frechet differentiable at* x;

(iii) the partial derivatives $\partial f(x)/\partial x_i$ *exist for* $i = 1, ..., n$.

Proof. Since by Theorem 432 f is locally Lipschitz, by Proposition 376 (i) implies (ii). Clearly, (ii) implies (iii). It remains to show that (iii) implies (i). Suppose the partial derivatives $\partial f(x)/\partial x_i$ exist for $i = 1, ..., n$. Let $T_x : \mathbb{R}^n \to \mathbb{R}$ be the linear functional given by

$$T_x(y) = \sum_{i=1}^{n} y_i \frac{\partial f(x)}{\partial x_i} = y \cdot \nabla f(x), \qquad \forall y \in \mathbb{R}^n.$$

By (4.13), $f'(x;\alpha e^i) = \alpha f'(x;e^i)$ for every $\alpha \in \mathbb{R}$ and $i = 1, ..., n$, and so

$$f'\left(x; y_i e^i\right) = y_i f'\left(x; e^i\right) = y_i \frac{\partial f(x)}{\partial x_i} = T_x\left(y_i e^i\right), \qquad \forall i = 1, ..., n.$$

By Theorem 438, $f'(x;\cdot) : \mathbb{R}^n \to \mathbb{R}$ is superlinear. Hence,

$$T_x(y) = \sum_{i=1}^{n} T_x\left(y_i e^i\right) = \sum_{i=1}^{n} f'\left(x; y_i e^i\right) \leq f'\left(x; \sum_{i=1}^{n} y_i e^i\right) = f'(x;y), \qquad \forall y \in \mathbb{R}^n,$$

and so

$$f'(x;y) \geq T_x(y) \geq -f'(x;-y) \geq f'(x;y), \qquad \forall y \in \mathbb{R}^n.$$

This implies $f'(x;y) = T_x(y)$ for all $y \in \mathbb{R}^n$. We conclude that $f'(x;\cdot) : \mathbb{R}^n \to \mathbb{R}$ is linear, and so f is Gateaux differentiable at x. ■

The next example shows that without concavity the mere existence of the partial derivatives in general does not guarantee Gateaux differentiability.

Example 445 Consider $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x_1, x_2) = \begin{cases} \frac{x_1 x_2}{|x_1|+|x_2|} & \text{if } (x_1, x_2) \neq (0,0), \\ 0 & \text{if } (x_1, x_2) = (0,0). \end{cases}$$

It is easy to check that the partial derivatives exist at $(0,0)$, with

$$\frac{\partial f(0,0)}{\partial x_1} = \frac{\partial f(0,0)}{\partial x_2} = 0.$$

But, for all $\alpha \neq 0$,

$$\frac{f((0,0) + t(\alpha,\alpha)) - f(0,0)}{t} = \frac{\alpha^2}{2|\alpha|} = \frac{1}{2}|\alpha|, \qquad \forall t > 0,$$

and so the Gateaux derivative at $(0,0)$ does not exist. ▲

8.4.2 Superdifferentials

Theorem 441 shows that for a concave functional Gateaux differentiability is equivalent to the existence of a unique linear functional for which (8.17) holds. Hence, non differentiability is equivalent to either the existence of more than one linear functional for which (8.17) holds or to the non existence of any such linear functional. This observation motivates the next definition.

Definition 446 *Let $f : A \to \mathbb{R}$ be a functional defined on a subset A of a normed vector space V. Given $v \in A$, we call superdifferential of f at v the set $\partial f(v)$ formed by the functionals $L \in V^*$ such that*

$$f(w) \leq f(v) + L(w - v), \qquad \forall w \in A. \tag{8.21}$$

The functional $f : A \to \mathbb{R}$ is called superdifferentiable at v if $\partial f(v) \neq \emptyset$.

The superdifferential consists therefore of all linear and continuous functionals for which (8.17) holds. Clearly, it may not exist any such linear functional; in this case the superdifferential is empty and the function is not superdifferentiable at the point considered.[4]

Next we give few basic properties of the superdifferential.

[4]Even if the superdifferential is a notion motivated by the study of concave functionals, notice that in Definition 446 we did not assume that f is concave and A convex (this observation will be useful in the sequel).

Proposition 447 *Let $f : A \to \mathbb{R}$ be a functional defined on a subset A of a normed vector space V. Then, $\partial f(v)$ is closed and convex. If, in addition, A is open and f is locally Lipschitz on A, then $\partial f(v)$ is also bounded.*

In the important special case when $f : C \to \mathbb{R}$ is a concave and continuous function defined on an open convex subset of $\mathbb{R}^n$, it follows that $\partial f(v)$ is a nonempty convex compact set at all $v \in C$.

Proof. It is immediate to check that $\partial f(v)$ is closed and convex. Suppose f is locally Lipschitz on the open set A. Wlog, suppose $\mathbf{0} \in A$ and $f(\mathbf{0}) = 0$. Consider $v = \mathbf{0}$. There exists a neighborhood $B_\varepsilon(\mathbf{0}) \subseteq A$ and a constant $M > 0$ such that $|f(w)| \leq M \|w\|$ for all $w \in B_\varepsilon(\mathbf{0})$. Let $L \in \partial f(v)$. Since $w \in B_\varepsilon(\mathbf{0})$ if and only if $-w \in B_\varepsilon(\mathbf{0})$, by (8.21) we have:

$$M \|w\| \geq -f(-w) \geq -L(-w) = L(w) \geq f(w) \geq -M \|w\|, \qquad \forall w \in B_\varepsilon(\mathbf{0}),$$

Hence,

$$\frac{|L(w)|}{\|w\|} \leq M, \qquad \forall w \in B_\varepsilon(\mathbf{0}).$$

It is easy to check that this implies $\|L\| \leq M$. Since L is any element of $\partial f(v)$, we conclude that $\|L\| \leq M$ for all $L \in \partial f(v)$, i.e., $\partial f(v)$ is a bounded set. ■

By Theorem 65, in the special case $V = \mathbb{R}^n$ the superdifferential $\partial f(x)$ of a function $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ can be equivalently defined as the set of the vectors $\chi \in \mathbb{R}^n$ such that

$$f(y) \leq f(x) + \chi \cdot (y - x), \qquad \forall y \in A. \tag{8.22}$$

Clearly, (8.22) is altogether equivalent to (8.21), and it simply uses the fact that, thanks to Theorem 65, we are able to fully describe the dual space of $\mathbb{R}^n$ and so to give a more concrete form to (8.21).

To visualize graphically the superdifferential, consider the affine functional $f_v : A \subseteq V \to \mathbb{R}$ defined by:

$$f_v(w) = f(v) + L(w - v), \qquad \forall w \in V.$$

with $L \in \partial f(v)$. The affine functional f_v is therefore such that $f_v(v) = f(v)$ and $f_v(w) \geq f(w)$ for each $w \in V$. In other words, f_v is equal to f at the point v and dominates f at all the other points of A.

It follows that $\partial f(v)$ characterizes the set of all affine functionals that touch the graph of f at the point v and that lie above it at all other points of the domain.

If we draw the graph of a concave function defined on $\mathbb{R}$, it is easy to see that at the points in which the function is differentiable the only affine functional that enjoys this

property is the tangent straight line $f(x) + f'(x)(y - x)$. Where the function is not differentiable, we can have several straight lines that touch the graphic of the function at the point considered and that lie above such graph in the other points. The set of these straight lines can be viewed as a surrogate of the tangent straight line, i.e., of the differential. This is the idea that lies behind the superdifferential: a surrogate of the differential when this does not exist.

There is a tight and very important relationship between the superdifferential and the directional derivative. This is established in the next result, which easily follows from Theorem 415 and Lemma 442. In particular, Lemma 442 implies (8.23), while Theorem 415 implies (8.24).

Theorem 448 *Let $f : C \to \mathbb{R}$ be a concave and continuous functional defined on an open and convex subset C of a normed vector space V. Then,*

$$\partial f(v) = \{L \in V^* : L(w) \geq f'(v; w) \text{ for all } w \in V\} \tag{8.23}$$

and

$$f'(v; w) = \min_{L \in \partial f(v)} L(w), \qquad \forall w \in V. \tag{8.24}$$

Thus, by (8.23) the superdifferential $\partial f(v)$ can be viewed as the collection of all continuous linear functionals that pointwise dominate the directional derivative $f'(v; \cdot) : V \to \mathbb{R}$. In turn, by (8.24) the directional derivative can be regarded as the the lower envelope of the superdifferential. This "dual" relationship between superdifferentials and directional derivatives is a main feature of concavity.

A first important consequence of this duality is the next result, which shows that for concave functions superdifferentiability is indeed a generalization of differentiability.

Corollary 449 *Let $f : C \to \mathbb{R}$ be a concave and continuous functional defined on an open and convex subset C of a normed vector space V. Then, f is Gateaux differentiable at a point $v \in C$ if and only if $\partial f(v)$ is a singleton. In such case, $\partial f(v) = \{f'(v; \cdot)\}$.*

We omit the proof of this result since, in view of (8.23), it is a straightforward consequence of Corollary 416. When C is a subset of $\mathbb{R}^n$, the condition $\partial f(v) = \{f'(v; \cdot)\}$ becomes $\partial f(x) = \{\nabla f(x)\}$.

We are now ready to see an example.

Example 450 Consider $f : \mathbb{R} \to \mathbb{R}$ defined as $f(x) = 1 - |x|$ for each $x \in \mathbb{R}$. The only point in which the function has not a derivative is $x = 0$. Hence, by Theorem 449

we have $\partial f(x) = \{f'(x)\}$ for each $x \neq 0$. It remains to determine $\partial f(0)$. By (8.22), this is equivalent to determine for which $\alpha \in \mathbb{R}$ the inequality

$$1 - |y| \leq 1 - |0| + \alpha(y - 0), \qquad \forall y \in \mathbb{R},$$

holds, i.e., $-|y| \leq \alpha y$ for each $y \in \mathbb{R}$. This inequality holds always for any α if $y = 0$. If $y \neq 0$, we have

$$\alpha \frac{y}{|y|} \geq -1. \tag{8.25}$$

Since

$$\frac{y}{|y|} = \begin{cases} 1 & \text{if } y \geq 0 \\ -1 & \text{if } y < 0 \end{cases},$$

(8.25) implies $\alpha \geq -1$ and $\alpha(-1) \geq -1$, that is, $\alpha \in [-1, 1]$. It follows that $\partial f(0) = [-1, 1]$, and hence:

$$\partial f(x) = \begin{cases} -1 & \text{if } x > 0 \\ [-1, 1] & \text{if } x = 0 \\ 1 & \text{if } x < 0 \end{cases}.$$

▲

Example 450 can be nicely generalized to any concave function $f : (a, b) \to \mathbb{R}$ defined on a, possibly unbounded, interval (a, b) of the real line. To this end, we first report some properties of its one-sided derivatives, which partly anticipate some more general results that we will see later in the chapter. We leave the proof to the reader.

Lemma 451 *Let $f : (a, b) \to \mathbb{R}$ be concave. Then,*

(i) the left $f'_+(x)$ and right $f'_-(x)$ derivatives exist at all $x \in (a, b)$;

(ii) the left $f'_+(x)$ and right $f'_-(x)$ derivatives are both decreasing on (a, b);

(iii) $f'_+(x) \leq f'_-(x)$ for all $x \in (a, b)$.

Using this lemma we can characterize the superdifferential of scalar concave functions.

Proposition 452 *Let $f : (a, b) \to \mathbb{R}$ be a concave function defined on a, possibly unbounded, interval of the real line. Then,*

$$\partial f(x) = \left[f'_+(x), f'_-(x)\right], \qquad \forall x \in (a, b). \tag{8.26}$$

In other words, the superdifferential consists of all coefficients that lie between the right and left derivatives. This makes precise the geometric intuition we gave above through the graph of a scalar function.

Proof. By Exercise 13.0.61, for the scalar function f we have

$$f'_+(x) = f'(x;1) \quad \text{and} \quad f'_-(x) = -f'(x;-1).$$

Hence, (8.26) amounts to $\partial f(x) = [f'(x;1), -f'(x;-1)]$. By (8.23), we have $\partial f(x) \subseteq [f'(x;1), -f'(x;-1)]$. To prove the converse inclusion, let $\chi \in [f'(x;1), -f'(x;-1)]$. Then, by the positive homogeneity of the directional derivative,

$$f'(x;t) = f'(x;1)\, t \leq \chi \cdot t \leq -f'(x;-1)\, t = -f'(x;-t), \qquad \forall t \geq 0.$$

which implies $\chi \cdot t \geq f'(x;t)$ for all $t \in \mathbb{R}$. By (8.23), $\chi \in \partial f(x)$. ■

Next we give an example where the superdifferential is empty.

Example 453 Consider $f : \mathbb{R}_+ \to \mathbb{R}$ defined by $f(x) = \sqrt{x}$ for each $x \in \mathbb{R}$. Also here the only point of the domain in which the function has not a derivative is $x = 0$, so that by Theorem 449 we have $\partial f(x) = \{f'(x)\}$ for each $x \neq 0$. As to $\partial f(0)$, it is given by the $\alpha \in \mathbb{R}$ such that

$$\sqrt{y} \leq \sqrt{0} + \alpha(y - 0), \qquad \forall y \in \mathbb{R}_+, \tag{8.27}$$

i.e., such that $\sqrt{y} \leq \alpha y$ for each $y \geq 0$. This inequality holds always for any α if $y = 0$, while for each $y > 0$ it is equivalent to:

$$\alpha \geq \frac{\sqrt{y}}{y} = \frac{1}{\sqrt{y}}.$$

But, letting y tend to 0, this implies:

$$\alpha \geq \lim_{y \to 0+} \frac{1}{\sqrt{y}} = +\infty,$$

and therefore there does not exist any $\alpha \in \mathbb{R}$ for which (8.27) holds. It follows that $\partial f(0) = \emptyset$, and the function is not superdifferentiable at 0. ▲

Before we argued that the superdifferential is a surrogate of the differential when this does not exist. In order to be a useful surrogate, however, it is necessary that it often exists, otherwise it would be of very little help. The previous example showed an instance where the superdifferential is indeed empty.

Fortunately, the next important result guarantees that concave functionals defined on open convex sets are everywhere superdifferentiable and that, moreover, this is exactly a property that characterizes concave functionals (another proof of the tight link between superdifferentiability and concavity).

Proposition 454 *Let $f : C \to \mathbb{R}$ be a concave and continuous functional defined on an open and convex subset C of a normed vector space V. Then, f is concave if and only if $\partial f(v)$ is nonempty for all $v \in C$.*

Proof. Suppose f is concave. Let $v \in C$. By proceeding as in the proof of Theorem 415, it is easy to check that the Hahn-Banach Theorem implies that there exists $L \in V^*$ such that $L(w) \geq f'(v; w)$ for all $w \in V$. Hence, by (8.23), $\partial f(v)$ is nonempty.

Conversely, suppose $\partial f(v) \neq \emptyset$ at all $v \in C$. Let $v_1, v_2 \in C$ and $t \in [0, 1]$. Let $L \in \partial f(tv_1 + (1-t)v_2)$. By (8.21),

$$\begin{aligned} f(v_1) &\leq f(tv_1 + (1-t)v_2) + L(v_1 - (tv_1 + (1-t)v_2)), \\ f(v_2) &\leq f(tv_1 + (1-t)v_2) + L(v_2 - (tv_1 + (1-t)v_2)), \end{aligned}$$

that is,

$$\begin{aligned} f(v_1) - (1-t)L(v_1 - v_2) &\leq f(tv_1 + (1-t)v_2), \\ f(v_2) - tL(v_2 - v_1) &\leq f(tv_1 + (1-t)v_2). \end{aligned}$$

Hence,

$$\begin{aligned} & f(tv_1 + (1-t)v_2) \\ \geq\ & tf(v_1) - t(1-t)L(v_1 - v_2) + (1-t)f(v_2) - (1-t)tL(v_2 - v_1) \\ =\ & tf(v_1) + (1-t)f(v_2), \end{aligned}$$

as desired. ■

The fact that C is open is key for Proposition 454: in fact, the function with an empty superdifferential in Example 453 is defined on the closed convex set $\mathbb{R}_+$. It is noteworthy that in this example the superdifferential is empty at 0, a boundary point of the domain. In fact, by Lemma 430, we could equivalently state Proposition 454 by saying that a concave and continuous functional $f : C \to \mathbb{R}$, defined on a convex subset C of a normed vector space V, is concave on $\operatorname{int} C$ if and only if $\partial f(v)$ is nonempty at all $v \in \operatorname{int} C$, i.e., at all interior points v of C. Getting back to Example 453, the concave function $f(x) = \sqrt{x}$ is indeed differentiable (and so superdifferentiable) at all $x \in (0, \infty)$, that is, at all interior points of the function's domain $\mathbb{R}_+$.

We close with couple of examples.

Example 455 Let $f : V \to \mathbb{R}$ be a superlinear continuous functional. Then, for all $v \in V$,

$$\partial f(v) = \{L \in V^* : L(v) = f(v) \text{ and } L(w) \geq f(w) \text{ for all } w \in V\}. \tag{8.28}$$

For, suppose $L \in \partial f(v)$. If we consider $w = 0$ in (8.21), we have $L(v) \leq f(v)$. On the other hand, if we consider $w = 2v$ in (8.21), we have $L(v) \geq f(v)$. We conclude that $L(v) = f(v)$. In turn, by (8.21) this implies $L(w) \geq f(w)$ for all $w \in V$. Conversely, suppose $L \in V^*$ is such that $L(v) = f(v)$ and $L(w) \geq f(w)$ for all $w \in V$. Then, (8.21) trivially holds and so $L \in \partial f(v)$.

Observe that $\partial f(\mathbf{0}) = \{L \in V^* : L(w) \geq f(w) \text{ for all } w \in V\}$. We can thus write (8.28) as

$$\partial f(v) = \{L \in \partial f(\mathbf{0}) : L(v) = f(v)\}, \qquad \forall v \in V. \tag{8.29}$$

In other words, to find $\partial f(v)$ is enough to determine the superdifferential $\partial f(\mathbf{0})$ at $\mathbf{0}$. ▲

Example 456 To make more concrete the previous example, consider the function $f : \mathbb{R}^n \to \mathbb{R}$ given by

$$f(x) = \min_{i=1,...,n} x_i, \qquad \forall x \in \mathbb{R}^n.$$

The function f is superlinear (and so continuous, by Corollary 434). By (8.29), to find $\partial f(x)$ is enough to determine $\partial f(\mathbf{0})$, i.e., $\{\chi \in \mathbb{R}^n : \chi \cdot x \geq f(x) \text{ for all } x \in \mathbb{R}^n\}$. Let $\chi \in \partial f(\mathbf{0})$. From:

$$\begin{aligned} \chi_i &= \chi \cdot e^i \geq f\left(e^i\right) = 0, \qquad \forall i = 1, ..., n, \\ \sum_{i=1}^{n} \chi_i &= \chi \cdot (1, ..., 1) \geq f(1, ..., 1) = 1, \\ -\sum_{i=1}^{n} \chi_i &= \chi \cdot (-1, ..., -1) \geq f(-1, ..., -1) = -1, \end{aligned}$$

we conclude that $\sum_{i=1}^{n} \chi_i = 1$ and $\chi_i \geq 0$ for each $i = 1, ..., n$. That is, χ belongs to the simplex Δ_{n-1} (see Exercise 385), so that $\partial f(\mathbf{0}) \subseteq \Delta_{n-1}$. On the other hand, if $\chi \in \Delta_{n-1}$, then

$$\chi \cdot x \geq \chi \cdot \left(\min_{i=1,...,n} x_i, ..., \min_{i=1,...,n} x_i \right) = \min_{i=1,...,n} x_i, \qquad \forall x \in \mathbb{R}^n,$$

and so $\chi \in \partial f(\mathbf{0})$. We conclude that $\partial f(\mathbf{0}) = \Delta_{n-1}$. By (8.29),

$$\partial f(x) = \{\chi \in \Delta_{n+1} : \chi \cdot x = f(x)\},$$

i.e., $\partial f(x)$ consists of the vectors x of the simplex such that $\chi \cdot x = f(x)$. ▲

8.4.3 Concavity and Differentiability

Up to now we have considered the properties of differentiability of concave functionals. We change now perspective and ask if, given a suitably differentiable function, there

exist some criteria based on this differentiability that allow us to determine whether the given function is concave.

The first thing to observe is that Theorem 441 has shown that for Gateaux differentials of a concave functional $f : C \to \mathbb{R}$ we have

$$f(w) \leq f(v) + f'(v; w - v), \qquad \forall w \in C.$$

A possible conjecture is that this is an inequality satisfied only by Gateaux differentials of concave functions, and hence a property that is typical of concave functions. The next result shows that this is true, thus establishing a first differential characterization of concavity.

Theorem 457 *Let $f : C \to \mathbb{R}$ be Gateaux differentiable at each point of an open and convex subset C of a normed vector space V. Then, f is concave if and only if*

$$f(w) \leq f(v) + f'(v; w - v), \qquad \forall v, w \in C, \tag{8.30}$$

while f is strictly concave if and only if inequality (8.30) is strict for each $v, w \in C$ with $v \neq w$.

Proof. "If." Let f be concave. Fix $v, w \in C$. Let $\phi_{v,w} : C_{v,w} \to \mathbb{R}$ be given by (8.11). By Lemma 428, $C_{v,w}$ is an open interval, and by Proposition 429 $\phi_{v,w}$ is concave on $C_{v,w}$. Hence,[5]

$$\begin{aligned} \phi'_+(t) &= \lim_{\varepsilon \to 0+} \frac{\phi(t+\varepsilon) - \phi(t)}{\varepsilon} \\ &= \lim_{\varepsilon \to 0+} \frac{f((1-t)v + tw + \varepsilon(v-w)) - f((1-t)v + tw)}{\varepsilon} \\ &= f'((1-t)v + tw; v - w). \end{aligned}$$

for each $t \in C_{v,w}$. Since $[0,1] \subseteq C_{v,w}$, by Corollary 443 we have

$$\phi(1) \leq \phi(0) + \phi'_+(0) = \phi(0) + f'(v; v - w),$$

i.e., $f(w) \leq f(v) + f'(v; w - v)$.

Conversely, suppose (8.30) holds. For each $v \in C$, consider the affine function $F_v : C \to \mathbb{R}$ given by $F_v(w) = f(v) + f'(v; w - v)$. By (8.30), $f(w) \leq F_v(w)$ for all $v, w \in C$. Since $F_v(v) = f(v)$, we conclude that $f(w) = \min_{v \in C} F_v(w)$ for each $w \in C$. Since each F_v is affine (why?), by Exercise 13.0.64 we conclude that f is concave since is a minimum of a family of concave functions. ■

[5] To ease notation, in the rest of the proof we use ϕ in place of $\phi_{v,w}$.

The Euclidean Case

By Theorem 457, a functional $f : C \to \mathbb{R}$ Gateaux differentiable at each point of an open convex subset C of $\mathbb{R}^n$ is concave if and only if

$$f(y) \leq f(x) + \nabla f(x) \cdot (y - x), \qquad \forall x, y \in C. \tag{8.31}$$

We now continue to study the special case $\mathbb{R}^n$. A functional $f : C \subseteq \mathbb{R}^n \to \mathbb{R}^n$ is *monotone (decreasing)* if

$$(f(x) - f(y)) \cdot (x - y) \leq 0, \qquad \forall x, y \in C, \tag{8.32}$$

and *strictly monotone (decreasing)* if the inequality (8.32) is strict for each $x, y \in C$ with $x \neq y$.

When $n = 1$ we go back to the usual notion of monotonicity, that is, $f : C \subseteq \mathbb{R} \to \mathbb{R}$ is monotone decreasing if $x \leq y$ implies $f(x) \geq f(y)$, and strictly monotone if $x < y$ implies $f(x) > f(y)$.[6]

Example 458 Consider an affine function $f : R^n \to R^n$ given by $f(x) = Ax + b$, where A is a symmetric $n \times n$ matrix and $b \in R^n$. Then, f is monotone if and only if A is negative semidefinite, and f is strictly monotone if and only if A is negative definite (why?).

In Section 4.5.1 we called derivative of f the application $f' : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^n$ defined by $f(x) = \nabla f(x)$, where Ω is the set of the points where f is Frechet differentiable. We now give a slightly more general meaning of this notion of derivative by assuming that Ω is the set of the points where f is only Gateaux differentiable.

Theorem 459 *Let $f : C \to \mathbb{R}$ be Gateaux differentiable at each point of an open and convex subset C of $\mathbb{R}^n$. Then,*

(i) f is concave if and only if its derivative $f' : C \to \mathbb{R}^n$ is continuous and monotone.

(ii) f is strictly concave if and only if $f' : C \to \mathbb{R}^n$ is continuous and strictly monotone.

Proof (i) Suppose f is concave. Let $x, y \in C$. By (8.31),

$$\begin{aligned} f(y) &\leq f(x) + \nabla f(x) \cdot (y - x), \\ f(x) &\leq f(y) + \nabla f(y) \cdot (x - y), \end{aligned}$$

[6] Notice that (8.32) cannot be defined on any vector space because it is based on internal products, which we have defined only in $\mathbb{R}^n$. This is why here we are only considering $\mathbb{R}^n$.

and so $\nabla f(x) \cdot (x-y) \leq f(x) - f(y) \leq \nabla f(y) \cdot (x-y)$. Hence,

$$(f'(x) - f'(y)) \cdot (x-y) \leq 0.$$

and we conclude that $f' : C \to \mathbb{R}^n$ is monotone decreasing.

Conversely, suppose $f' : C \to \mathbb{R}^n$ is monotone decreasing, i.e., (8.32) holds. Suppose first that $n = 1$. Let $x \in C$, and define $\psi_x : C \to \mathbb{R}$ by $\psi_x(y) = f(y) - f(x) - f'(x)(y-x)$. Then, $\psi_x'(y) = f'(y) - f'(x)$, and so $\psi_x'(y) \geq 0$ if $y < x$ and $\psi_x'(y) \leq 0$ if $y > x$. Hence, ψ_x has a minimum at x, i.e.,

$$0 = \psi_x(x) \leq \psi_x(y) = f(y) - f(x) - f'(x)(y-x), \qquad \forall y \in C.$$

Since x was arbitrary, we conclude that $f(y) \leq f(x) + f'(x)(y-x)$ for all $x, y \in C$. By Theorem 457, f is concave. This completes the proof for $n = 1$.

Suppose now that $n > 1$. Let $x, y \in C$ and let $\phi_{x,y} : C_{x,y} \to \mathbb{R}$ be given by (8.11). By Lemma 428, $C_{x,y}$ is an open interval, with $[0,1] \subseteq C_{x,y}$. Then, $\phi_{x,y}$ is concave and differentiable on (a,b), with

$$\phi_{x,y}'(t) = \nabla f((1-t)x + ty) \cdot (x-y), \qquad \forall t \in C_{x,y}. \tag{8.33}$$

Let $t_2 \geq t_1 \in C_{x,y}$. Since f' is monotone, then

$$\begin{aligned} & (\nabla f((1-t_1)x + t_1 y) - \nabla f((1-t_2)x + t_2 y)) \cdot ((1-t_1)x + t_1 y - ((1-t_2)x + t_2 y)) \\ = \; & (t_2 - t_1)(\nabla f((1-t_1)x + t_1 y) - \nabla f((1-t_2)x + t_2 y)) \cdot (x-y) \leq 0 \end{aligned}$$

and so, by (8.33),

$$0 \geq (\nabla f((1-t_1)x + t_1 y) - \nabla f((1-t_2)x + t_2 y)) \cdot (x-y) = \phi_{x,y}'(t_1) - \phi_{x,y}'(t_2),$$

and we conclude that $\phi_{x,y}'(t_1) \leq \phi_{x,y}'(t_2)$, i.e., $\phi_{x,y}'$ is monotone on $C_{x,y}$. By what already proved, $\phi_{x,y}$ is then concave, and so:

$$f((1-t)x + ty) = \phi_{x,y}(t) \geq (1-t)\phi_{x,y}(0) + t\phi_{x,y}(1) = (1-t)f(x) + tf(y),$$

which shows that f is concave.

(ii) For simplicity, we consider the case $n = 1$. We leave to the reader the extension to $n \geq 1$, with the help of Proposition 429.

Suppose f is strictly concave. Since f is concave, f' is decreasing and continuous by (i). Let $x_1, x_2 \in U$ with $x_1 < x_2$. Suppose, *per contra*, that $f'(x_1) = f'(x_2) \equiv \alpha$. We have:

$$\begin{aligned} f(x) &\leq f(x_1) + \alpha(x - x_1), \\ f(x) &\leq f(x_2) + \alpha(x - x_2), \end{aligned} \tag{8.34}$$

for all $x \in U$. In particular, $f(x_2) \leq f(x_1) + \alpha(x_2 - x_1)$ and $f(x_1) \leq f(x_2) + \alpha(x_1 - x_2)$, so that

$$f(x_2) - f(x_1) \leq \alpha(x_2 - x_1) \leq f(x_2) - f(x_1),$$

which implies $f(x_2) - f(x_1) = \alpha(x_2 - x_1)$. Given $\beta \in (0,1)$, by (8.34),

$$f(\beta x_1 + (1-\beta)x_2) \leq f(x_1) + \alpha(1-\beta)(x_2 - x_1) = \beta f(x_1) + (1-\beta) f(x_2),$$

which contradicts strict concavity.

Conversely, suppose f' is strictly decreasing. Then, the function f is concave. It remains to show that it is strictly concave. Suppose, per contra, that there exist $x_1, x_2 \in U$, with $x_1 < x_2$, and $\lambda \in (0,1)$ such that $f((1-\lambda)x_1 + \lambda x_2) = (1-\lambda) f(x_1) + \lambda f(x_2)$. Define $\phi : [0,1] \to \mathbb{R}$ by $\phi(\alpha) = f((1-\alpha)x_1 + \alpha x_2)$. Then, ϕ is concave and continuous, with $\phi(\lambda) = (1-\lambda)\phi(0) + \lambda\phi(1)$. This implies $\phi(\alpha) = (1-\alpha)\phi(0) + \alpha\phi(1)$ for all $\alpha \in [0,1]$. Then,

$$f'(x_1) = \lim_{\alpha \downarrow 0} \frac{f((1-\alpha)x_1 + \alpha x_2) - f(x_1)}{(1-\alpha)x_1 + \alpha x_2 - x_1} = \frac{f(x_2) - f(x_1)}{x_2 - x_1},$$
$$f'(x_2) = \lim_{\alpha \uparrow 1} \frac{f((1-\alpha)x_1 + \alpha x_2) - f(x_2)}{(1-\alpha)x_1 + \alpha x_2 - x_2} = \frac{f(x_1) - f(x_2)}{x_1 - x_2},$$

so that $f'(x_1) = f'(x_2)$, a contradiction. ■

In the case $n = 1$ this means that a differentiable function on an interval (a,b) is concave if and only if its derivative f' is monotone decreasing (and f is strictly concave if f' is strictly monotonic decreasing).

Recall from Calculus that a differentiable function $f : (a,b) \to \mathbb{R}$ is monotone if and only if $f'(x) \leq 0$ for each $x \in (a,b)$, that is, if and only if $f''(x) \leq 0$ for each $x \in (a,b)$ when f is twice differentiable. On the other hand, $f : (a,b) \to \mathbb{R}$ is strictly monotone if $f'(x) < 0$ for each $x \in (a,b)$, that is, if $f''(x) \leq 0$ for each $x \in (a,b)$ when f is twice differentiable. But, the "only if" is false for strict monotonicity, as $f(x) = x^3$ shows.

This simple observation leads to the following result, in which the role of the second derivative is played in the general case by the Hessian matrix.

Theorem 460 *Let $f : C \to \mathbb{R}$ be a functional defined on an open and convex subset C of $\mathbb{R}^n$. If $f \in \mathcal{C}^2(C)$, then*

(i) f is concave if and only if the Hessian matrix $\nabla^2 f(x)$ is negative semidefinite for each $x \in C$.

(ii) f is strictly concave if $\nabla^2 f(x)$ is negative definite for each $x \in C$.

Proof. We only prove (i) and leave (ii) to the reader. By Theorem 459, we have to prove that f' is monotone on C if and only if $\nabla^2 f(x)$ is negative semidefinite for each $x \in C$. Suppose first that f' is monotone. Let $x \in C$ and $y \in \mathbb{R}^n$. Then, for $t > 0$ small enough we have $(f'(x+ty) - f'(x)) \cdot ((x+ty) - x) \leq 0$. Hence,

$$\begin{aligned} 0 &\geq \lim_{t \to 0+} \frac{(f'(x+ty) - f'(x)) \cdot ((x+ty) - x)}{t} \\ &= \lim_{t \to 0+} \frac{(f'(x+ty) - f'(x))}{t} \cdot y = \nabla^2 f(x)\, y \cdot y \end{aligned}$$

and since this holds for any $y \in \mathbb{R}^n$ we conclude that $\nabla^2 f(x)$ is negative semidefinite.

Conversely, suppose that $\nabla^2 f(x)$ is negative semidefinite at all $x \in C$. Let $x_1, x_2 \in C$ and consider the function $\phi : [0,1] \to \mathbb{R}$ given by

$$\phi(t) = (x_1 - x_2) \cdot (f'(tx_1 + (1-t)x_2) - f'(x_2)).$$

To prove that f' is monotone we must show that $\phi(1) \geq 0$. But, $\phi(0) = 0$ and ϕ is monotone since, for all $t \in (0,1)$,

$$\phi'(t) = (x_1 - x_2) \cdot \nabla^2 f(tx_1 + (1-t)x_2)(x_1 - x_2) \geq 0.$$

Hence, $\phi(1) \geq \phi(0) = 0$. ■

This is the most useful criterion to determine if a function is concave. Naturally, specular results hold for convex functions, which are characterized by having positive semidefinite Hessian matrices.

Example 461 Let $f : \mathbb{R}_{++} \to \mathbb{R}$ be defined as $f(x) = \sqrt{x}$ for each $x > 0$. We have $f''(x) = (-1/4)\, x^{-\frac{3}{2}}$ for each $x > 0$, and hence f is strictly concave since $f''(x) < 0$ for each $x > 0$. ▲

Example 462 Let $f : \mathbb{R} \to \mathbb{R}$ be defined as $f(x) = e^x$ for each $x \in \mathbb{R}$. We have $f''(x) = e^x$, and so f is strictly convex since $f''(x) > 0$ for each $x \in \mathbb{R}$. ▲

Example 463 In Example 191 we considered the function $f : \mathbb{R}^3 \to \mathbb{R}$ defined by

$$f(x) = x_1^2 + 2x_2^2 + x_3^2 + (x_1 + x_3)\, x_2, \qquad \forall x \in \mathbb{R}^3,$$

and we saw how its Hessian matrix was positive definite. By Theorem 460, this function is strictly convex. ▲

8.5 Optimization

Concave functions have their most classical application in the study of optimization problems, in which they enjoy remarkable properties. The first one is that concave functions can only have global maxima.

Theorem 464 *Let $f : C \to \mathbb{R}$ be a concave functional defined on a convex subset C of a normed vector space V. If the point $\widehat{v} \in C$ is a local maximum, then it is also a global maximum.*

Proof Let $\widehat{v} \in C$ be a local maximum. By definition, there exists a neighborhood $B_\varepsilon(\widehat{v})$ such that

$$f(\widehat{v}) \geq f(v), \qquad \forall v \in B_\varepsilon(\widehat{v}) \cap C. \tag{8.35}$$

Suppose that $\widehat{v}$ is not a global maximum. There exists therefore $w \in C$ such that $f(w) > f(\widehat{v})$. Since f is concave, for each $t \in (0,1)$ we have:

$$f(t\widehat{v} + (1-t)w) \geq tf(\widehat{v}) + (1-t)f(w) > tf(\widehat{v}) + (1-t)f(\widehat{v}) = f(\widehat{v}). \tag{8.36}$$

Since C is convex, we have $t\widehat{v} + (1-t)w \in C$ for each $t \in (0,1)$. On the other hand,

$$\lim_{t\to 1-} \|t\widehat{v} + (1-t)w - \widehat{v}\| = \|w - \widehat{v}\| \lim_{t\to 1-} (1-t) = 0,$$

and therefore there exists $\overline{t} \in (0,1)$ such that $t\widehat{v} + (1-t)w \in B_\varepsilon(\widehat{v})$ for each $t \in \left(\overline{t}, 1\right)$. Expression (8.36) implies that for such t we have $f(t\widehat{v} + (1-t)w) > f(\widehat{v})$, which contradicts (8.35). This contradiction proves that $\widehat{v}$ is a point of global maximum. ■

By Theorem 464, the points of maximum of concave functions are necessarily points of global maximum. We denote by $\arg\max_C f(x)$ the set of such points, i.e.,

$$\arg\max_C f = \left\{ v \in C : f(v) = \max_{v\in C} f(v) \right\}.$$

When f is concave, this set is convex. In fact, let $v_1, v_2 \in \arg\max_C f$ and let $t \in [0,1]$. By concavity we have:

$$f(tv_1 + (1-t)v_2) \geq tf(v_1) + (1-t)f(v_2) = \max_{v\in C} f(v),$$

and therefore

$$f(tv_1 + (1-t)v_2) = \max_{v\in C} f(v),$$

that is, $tv_1 + (1-t)v_2 \in \arg\max_C f$.

Being convex, for this set we have three possibilities:

(i) $\arg\max_C f$ is empty, i.e., there are no global maxima;

(ii) $\arg\max_C f$ is a singleton, i.e., there exists a unique global maximum;

(iii) $\arg\max_C f$ consists of infinite points, i.e., there exist infinite global maxima.

We illustrate these different possibilities with some examples.

Example 465 Let $f : \mathbb{R}_{++} \to \mathbb{R}$ be defined by $f(x) = \lg x$ for each $x > 0$. Since $f''(x) = -(1/x^2) < 0$, this function is strictly concave by Theorem 460. On the other hand, it is easy to see that this function does not have points of global maximum, i.e., $\arg\max_{\mathbb{R}_{++}} f = \emptyset$. ▲

Example 466 Let $f : \mathbb{R} \to \mathbb{R}$ be defined as $f(x) = 1 - x^2$ for each $x \in \mathbb{R}$. This function is strictly concave and the only point of global maximum is $\widehat{x} = 0$. Hence, $\arg\max_{\mathbb{R}_{++}} f = \{0\}$. ▲

Example 467 Let $f : \mathbb{R} \to \mathbb{R}$ be defined by:

$$f(x) = \begin{cases} x & \text{if} \quad x \leq 1 \\ 1 & \text{if} \quad x \in (1,2) \\ 3 - x & \text{if} \quad x > 2 \end{cases}$$

It is a concave function with $\arg\max_{\mathbb{R}} f = [1, 2]$. ▲

In the last example, in which we have infinite global maxima, the function is concave but not strictly concave. This is not by chance: the next result shows that, remarkably, strict concavity implies that the maximum, if it exists, is unique. In other words, for strictly concave functions $\arg\max_C f$ is at most a singleton.

Theorem 468 *A functional $f : C \to \mathbb{R}$ strictly concave defined on a convex subset C of a normed vector space V has at most a unique point of maximum.*

Proof Suppose that $\widehat{v}_1, \widehat{v}_2 \in C$ are two points of global maximum for f. We want to prove that $\widehat{v}_1 = \widehat{v}_2$. Suppose that this is not the case, i.e., $\widehat{v}_1 \neq \widehat{v}_2$. Since $\widehat{v}_1$ and $\widehat{v}_2$ are global maxima, we have $f(\widehat{v}_1) = f(\widehat{v}_2) = \max_{v \in C} f(v)$. Set $\overline{v} = \frac{1}{2}\widehat{v}_1 + \frac{1}{2}\widehat{v}_2 \in C$. By strict concavity we have:

$$f(\overline{v}) = f\left(\frac{1}{2}\widehat{v}_1 + \frac{1}{2}\widehat{v}_2\right) > \frac{1}{2}f(\widehat{v}_1) + \frac{1}{2}f(\widehat{v}_2) = \max_{v \in C} f(v),$$

a contradiction. It follows that $\widehat{v}_1 = \widehat{v}_2$, as desired. ■

Now that we have seen what are the remarkable properties that the points of maximum of a concave function enjoy, we face the problem of how to find them. The next result gives an interesting characterization of the points of global maximum of any functional, not necessarily concave. Notice that here $\mathbf{0}$ denotes the identically null linear functional.

Theorem 469 *Let $f : A \to \mathbb{R}$ be a functional defined on a subset A of a normed vector space V. Then, $\widehat{v} \in A$ is a global maximum if and only if f is superdifferentiable at $\widehat{v}$ and $\mathbf{0} \in \partial f(\widehat{v})$.*

Proof Let $\widehat{v} \in A$ be a maximum. We have:

$$f(v) \leq f(\widehat{v}) + \mathbf{0}(v - \widehat{v}), \qquad \forall v \in C,$$

and hence $\mathbf{0} \in \partial f(\widehat{v})$.

Viceversa, let $\mathbf{0} \in \partial f(\widehat{v})$. We have

$$f(v) \leq f(\widehat{v}) + \mathbf{0}(v - \widehat{v}), \qquad \forall v \in A,$$

that is, $f(v) \leq f(\widehat{v})$ for each $v \in A$, which implies that $\widehat{v}$ is a maximum. ■

By Proposition 454, the superdifferential of a concave and continuous functional defined on an open convex set is nonempty. We therefore have the following fundamental consequence of Theorem 469, which gives us the more general version of the so-called first order condition for concave functionals.

Corollary 470 *Let $f : C \to \mathbb{R}$ be a concave and continuous functional defined on an open and convex subset C of a normed vector space V. Then, $\widehat{v} \in A$ is a global maximum if and only if $\mathbf{0} \in \partial f(\widehat{v})$.*

The next example shows how this corollary makes it possible to find the global maxima of a function for which the classical Fermat Theorem 194 does not apply since there are points where it is not differentiable.

Example 471 We go back to Example 450, where we considered the function $f : \mathbb{R} \to \mathbb{R}$ defined as $f(x) = 1 - |x|$ for each $x \in \mathbb{R}$. We have

$$\partial f(x) = \begin{cases} 1 & \text{if } x > 0 \\ [-1, 1] & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}$$

Hence, by Corollary 470 we have that $\widehat{x} = 0$ is a global maximum since $\mathbf{0} \in \partial f(0)$. ▲

The most interesting aspect of Theorem 469 emerges when it is applied to concave functions. In fact, by Corollary 449, Gateaux differentiability implies that the superdifferential is a singleton, which consists exactly of the Gateaux differential. This simple observation implies that by Theorem 469 we have the following version of Fermat Theorem 194 for concave functions that are Gateaux differentiable on normed spaces.

Corollary 472 *Let $f : C \to \mathbb{R}$ be a concave functional defined on an open and convex subset C of a normed vector space V. If f is Gateaux differentiable at $\widehat{v} \in C$, then $\widehat{v} \in C$ is a global maximum if and only if $f'(\widehat{v}; \cdot) = \mathbf{0}$.*

Proof Since f is Gateaux differentiable at $\widehat{v} \in C$, by Corollary 449 we have $\partial f(\widehat{v}) = \{f'(\widehat{v}; \cdot)\}$. By Corollary 470, $\widehat{v}$ is a global maximum if and only if $\mathbf{0} \in \partial f(\widehat{v})$, i.e., if and only if $f'(\widehat{v}; \cdot) = \mathbf{0}$. ■

In the special case $V = \mathbb{R}^n$, which was exactly the one considered in Theorem 194, Corollary 472 takes the following form.

Corollary 473 *Let $f : C \subseteq \mathbb{R}^n \to \mathbb{R}$ be a concave functional defined on an open and convex subset C of $\mathbb{R}^n$. If f is Gateaux differentiable at a point $\widehat{x}$ of C, then $\widehat{x} \in C$ is a global maximum if and only if $\nabla f(\widehat{x}) = \mathbf{0}$.*

Consider for a moment this corollary. Theorem 194 gave us the first order condition for local maxima and minima. Corollary 473 has two very important consequences in the concave case:

- the first order condition characterizes global maxima;
- the first order condition is necessary and sufficient for a point to be a maximum.

All this considerably simplifies the study of maxima of concave functions. Naturally, specular considerations hold for convex functions, in which it is the study of minima to be equally facilitated.

Example 474 Let $f : \mathbb{R} \to \mathbb{R}$ be defined as $f(x) = 1 - x^2$. It is a strictly concave function, and by Theorem 468 it has therefore at most a unique global maximum. To find it, observe that $f'(x) = -2x = 0$ if and only if $x = 0$. By Corollary 473, we conclude that $\widehat{x} = 0$ is a global maximum for this function. ▲

Example 475 Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x) = 2x_1^2 + x_2^2 - 3(x_1 + x_2) + x_1 x_2 - 3, \qquad \forall x \in \mathbb{R}^2.$$

In Example 202 we saw how its Hessian matrix was negative definite at each point of the domain. By Theorem 460, f is strictly convex, and hence by Theorem 468 it has at most a unique global minimum. We have

$$\nabla f(x) = (4x_1 - 3 + x_2, 2x_2 - 3 + x_1),$$

and so $\nabla f(x) = 0$ implies $x = (3/7, 9/7)$. By Corollary 473, $\widehat{x} = (3/7, 9/7)$ is a global minimum. We reached this conclusion also in Example 202, but using the second order condition, which we now know is superfluous for this function. ▲

8.5.1 Minima

We now consider global minima of concave functions. For these points we have the following classical result.

Proposition 476 *Let $f : C \to \mathbb{R}$ be a concave functional defined on a convex subset C of a normed vector space V. If f is not constant, then its global minima (if they exist) belong to the frontier of C.*

Proof Let $\widehat{v}$ be a minimum of f. Since f is not constant, there exists $w \in C$ such that $f(w) > f(\widehat{v})$. Suppose that $\widehat{v}$ is an interior point of C. Set $z_\alpha = \alpha\widehat{v} + (1-\alpha)w$ with $\alpha \in \mathbb{R}$. The points z_α are the points of the straight line that pass through $\widehat{v}$ and w. Since $\widehat{v}$ is an interior point of C, there exists $\alpha > 1$ such that $z_\alpha \in C$. On the other hand,

$$\widehat{v} = \frac{1}{\alpha} z_\alpha + \left(1 - \frac{1}{\alpha}\right) w,$$

and therefore we get the contradiction

$$\begin{aligned} f(\widehat{v}) &= f\left(\frac{1}{\alpha} z_\alpha + \left(1 - \frac{1}{\alpha}\right) w\right) \geq \frac{1}{\alpha} f(z_\alpha) + \left(1 - \frac{1}{\alpha}\right) f(w) \\ &> \frac{1}{\alpha} f(\widehat{v}) + \left(1 - \frac{1}{\alpha}\right) f(\widehat{v}) = f(\widehat{v}). \end{aligned}$$

It follows that $\widehat{v} \in \partial C$, as desired. ■

Hence,

$$\min_{v \in C} f(v) = \min_{v \in \partial C} f(v)$$

and the search of global minima can be restricted to the frontier ∂C of C.

Example 477 Consider the concave function $f : [-1, 1] \to \mathbb{R}$ defined by:

$$f(x) = \begin{cases} 2 - x^2 & \text{if } x \in (0, 1) \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x = 1 \end{cases},$$

a slight modification of the function seen in Example 435. Since the frontier of $[0, 1]$ is given by $\{0, 1\}$, by Theorem 476 global minima belong to the set $\{0, 1\}$. In particular, it is immediate to see that $x = 1$ is the global minimum. ▲

In the finite dimensional case we can refine the previous result by showing that at least some of the points of global minimum are extreme points of C. This result is interesting because the set of the extreme points can be a very small subset of the frontier (think of Example 402).

Theorem 478 *Let $f : C \to \mathbb{R}$ be a concave functional defined on a convex and compact subset C of a finite dimensional normed vector space. If f is not constant, then*

$$\text{ext}C \cap \arg\min_{C} f \neq \emptyset,$$

i.e., at least one of the points of global minimum of f is an extreme point of C.

Proof Suppose that v^* is a global minimum. By the Minkowski Theorem 404, we have $C = \text{co}(\text{ext}C)$, and therefore there exist a finite collection $\{v^i\}_{i\in I} \subseteq \text{ext}C$ and a finite collection $\{\lambda_i\}_{i\in I}$, with $\lambda_i \in [0,1]$ and $\sum_{i\in I}\lambda_i = 1$, such that $v^* = \sum_{i\in I}\lambda_i v^i$. Since v^* is a global minimum, we have $f(v^i) \geq f(v^*)$ for each $i \in I$. Together with concavity, this implies that:

$$f(v^*) = f\left(\sum_{i\in I}\lambda_i v^i\right) \geq \sum_{i\in I}\lambda_i f\left(v^i\right) \geq \sum_{i\in I}\lambda_i f(v^*) = f(v^*). \tag{8.37}$$

Hence, $\sum_{i\in I}\lambda_i f(v^i) = f(v^*)$, which implies $f(v^i) = f(v^*)$ for at least one $i \in I$. In fact, if it were $f(v^i) > f(v^*)$ for each $i \in I$, we would have $\sum_{i\in I}\lambda_i f(v^i) > f(v^*)$, which contradicts (8.37). It follows that for at least one $i \in I$ we have $v^i \in \arg\min_C f$, and so $\text{ext}C \cap \arg\min_C f \neq \emptyset$. ■

By this theorem, if C is a compact and convex set of a finite dimensional space we have:

$$\min_{v\in C} f(v) = \min_{v\in \text{ext}C} f(v), \tag{8.38}$$

and at least some minima belong to ext$C$.

Example 479 The function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x) = -2x_1^2 - x_2^2 + 3(x_1 + x_2) - x_1x_2 + 3, \qquad \forall x \in \mathbb{R}^2,$$

is concave since $-f$ is convex (see Example 475). We look for its points of global minimum on the closed unit ball $\overline{B}_1(0) = \{x \in \mathbb{R}^2 : \|x\|_1 \leq 1\}$. By (7.23) and (8.38) we have:

$$\min_{x\in\overline{B}_1(0)} f(x) = \min_{i=1,2}\left\{\pm f\left(e^i\right)\right\}.$$

Since

$$f\left(e^1\right) = -2,\ f\left(e^2\right) = 5,\ f\left(-e^1\right) = -2,\ f\left(-e^2\right) = -1,$$

the points e^1 and $-e^1$ are therefore global minima. ▲

Specular properties hold for convex functionals, whose global maxima enjoy the properties that we have just seen hold for the global minima of concave functionals. If we consider affine functionals, i.e., functionals that are both concave and convex, we therefore have the following corollary of Theorem 478, which reinforces the conclusions of the Weierstrass Theorem in the affine case.

Corollary 480 *Let $f : C \to \mathbb{R}$ be an affine functional defined on a convex and compact subset C of a finite dimensional normed vector space. If $f \neq \mathbf{0}$, then there exist $x_1, x_2 \in \text{ext}C$ such that*

$$f(x_1) = \max_{x \in K} f(x) \qquad \text{and} \qquad f(x_2) = \min_{x \in K} f(x).$$

For affine functionals we therefore have a particularly effective version of the Weierstrass Theorem: not only global maxima and minima exist, but at least some of them are necessarily extreme points. This result and its variations play a fundamental role in linear programming, whose object of study are problems of optimum in which objective functions are affine.

Example 481 Consider the affine functional $L : \mathbb{R}^2 \to \mathbb{R}$

$$L(x) = 1 + x_1 - 2x_2 \qquad \forall x \in \mathbb{R}^2,$$

and study its points of global minimum and maximum on $\overline{B}_1(0) = \{x \in \mathbb{R}^2 : \|x\|_\infty \leq 1\}$. Since

$$\text{ext}C = \{(1,1), (1,-1), (-1,1), (-1,-1)\},$$

and

$$L(1,1) = 0,\ L(1,-1) = 4,\ L(-1,1) = -2,\ L(-1,-1) = 2$$

by Corollary 480 it follows that $(-1,1)$ is a global minimum, while $(1,-1)$ is a global maximum.

Notice that to find the points of global maximum and minimum on $\overline{B}_1(0)$ it has been enough to examine only four points. ▲

8.5.2 Noncoercive Optimality

For concave functions we can establish existence results for global maxima that do not rely on any form of compactness, unlike the Weierstrass-type theorems of Section 6.6. This is a noteworthy feature of concavity, which we investigate in this section.

Though the main existence result of this section, Theorem 497, requires that the space be finite dimensional, the key concept of recession cone and some of its properties can be introduced in general normed vector spaces.

Definition 482 *The recession cone R_C of a set C of V is defined by*

$$R_C = \{w \in V : v + tw \in C \text{ for all } v \in C \text{ and all } t \geq 0\},$$

with the convention $R_\emptyset = V$.

The vectors in R_C are called *directions of recession*. Intuitively, along these directions the set C is unbounded.

Lemma 483 *Let C be a subset of V. Then, R_C is a convex cone, which is closed if C is.*

Proof. The set R_C is a cone since, given any $\lambda \geq 0$,

$$v + t(\lambda w) = v + (t\lambda) w \in C, \qquad \forall v \in C, \forall t \geq 0.$$

To show that it is convex, let $w', w'' \in R_C$ and $\alpha \in (0,1)$. For all $v \in C$, we have $v + \alpha t w' \in C$, and so $v + t(\alpha w' + (1-\alpha) w'') = v + \alpha t w' + (1-\alpha) t w'' \in C$.

We now show that it is closed when C is. Let $\{w_n\}_n \subseteq R_C$ with $w_n \to w$. Then, $v + \lambda w_n \to v + \lambda w$ for all $v \in C$ and all $t \geq 0$, and so $v + \lambda w \in C$ since C is closed. ■

The next lemma gives some basic properties of recession cones of closed convex sets. Observe that by point (i) to see if a vector w is a direction of recession is actually enough to check a single $v \in C$.

Lemma 484 *Let C be a closed convex subset of V. Then,*

(i) A vector $w \in V$ belongs to R_C if and only if there is some $v \in C$ such that $v + tw \in C$ for all $t \geq 0$.

(ii) A vector $w \in V$ belongs to R_C if and only if there exist $\{v_n\}_n \subseteq C$ and $\{\lambda_n\}_n \subseteq \mathbb{R}_+$, with $\lambda_n \uparrow \infty$, such that $\lim_n (v_n/\lambda_n) = w$.

(iii) If D is a closed convex subset of V such that $C \cap D \neq \emptyset$, then $R_{C \cap D} = R_C \cap R_D$.

Proof. (i) We prove the "if" part, the converse being trivial. Suppose there is $v \in C$ such that $v + \lambda w \in C$ for all $t \geq 0$. Consider any $z \in C$ and $\eta \geq 0$. We want to show that $z + \eta w \in C$. By convexity $(1 - \eta/t)z + (\eta/t)(v + \lambda w) \in C$ for all $t \geq \eta$. Setting $\lambda = t$, we have

$$\lim_{t \to \infty} (1 - \frac{\eta}{t}) z + \frac{\eta}{t}(v + \lambda w) = \lim_{t \to \infty} \left[(1 - \frac{\eta}{t}) z + \frac{\eta}{t} v + \eta w \right] = z + \eta w \in C$$

by closedness of C.

(ii) Let $w \in R_C$. Given $v \in C$ and $\{\lambda_n\}_n \subseteq \mathbb{R}_+$, with $\lambda_n \uparrow \infty$, set $v_n = v + \lambda_n w$. We have $v_n \in C$ for all $n \geq 1$ since $w \in R_C$, and $\lim_n (v_n/\lambda_n) = w$. Conversely, suppose $w \in V$ is such that $\lim_n (v_n/\lambda_n) = w$ for some $\{v_n\}_n \subseteq C$ and $\{\lambda_n\}_n \subseteq \mathbb{R}_+$, with $\lambda_n \uparrow \infty$. Given any $v \in C$ and $t \geq 0$, set

$$z_n = \left(1 - \frac{t}{\lambda_n}\right) v + \frac{t}{\lambda_n} v_n, \qquad \forall n \geq 1.$$

We have $z_n \in C$ for all n large enough (i.e., such that $t/\lambda_n \leq 1$), and $\lim_n z_n = v + tw$. Since C is closed, we have $v + tw \in C$, and so $w \in R_C$.

(iii) We only prove that $R_{C\cap D} \subseteq R_C \cap R_D$, the converse being trivial. Let $w \in R_{C\cap D}$ and $v \in C \cap D$. Then, $v + tw \in C$ for all $t \geq 0$. By point (i), w belongs to both R_C and R_D. ■

It is easy to see that (iii) can be generalized as follows: given a collection $\{C_i\}_{i\in I}$ of closed convex sets with nonempty intersection, it holds $\bigcap_{i\in I} R_{C_i} = R_{\bigcap_{i\in I} C_i}$.

Next we define the recession cone of a function as the intersection of all recession cones of its upper level sets $(f \geq \lambda)$. We will see momentarily that for concave functions all these cones actually coincide.

Definition 485 *The recession cone R_f of a function $f : C \to \mathbb{R}$ is defined by*

$$R_f = \bigcap_{\lambda\in\mathbb{R}} R_{(f\geq\lambda)},$$

The following result clarifies the nature of R_f for concave functions by showing that $R_f = R_{(f\geq\lambda)}$ for all nonempty $(f \geq \lambda)$.

Lemma 486 *Let $f : C \to \mathbb{R}$ be an upper semicontinuous concave function defined on a closed convex subset C of V. Then, all its nonempty upper level sets $(f \geq \lambda)$ have the same recession cone. In particular,*

$$R_{(f\geq\lambda)} = \left\{w \in V : (w, 0) \in R_{\mathrm{ipo}(f)}\right\}$$

for all $\lambda \in \mathbb{R}$ such that $(f \geq \lambda) \neq \emptyset$.[7]

Proof. Fix $\lambda \in \mathbb{R}$ such that $(f \geq \lambda) \neq \emptyset$. We have $w \in R_{(f\geq\lambda)}$ if and only $v + tw \in (f \geq \lambda)$ for all $v \in (f \geq \lambda)$ and all $t \geq 0$, i.e., if and only if $f(v + tw) \geq \lambda$ for all $v \in (f \geq \lambda)$ and all $t \geq 0$. Set $A = \{(v, \lambda) \in V \times \mathbb{R} : v \in (f \geq \lambda)\}$. Then,

$$\begin{aligned}
(w, s) \quad \in \quad & R_A \iff (v, \lambda) + t(w, s) \in A, \qquad \forall (v, \lambda) \in A, \forall t \geq 0, \\
\iff \quad & (v + tw, \lambda + ts) \in A, \qquad \forall (v, \lambda) \in A, \forall t \geq 0, \\
\iff \quad & f(v + tw) \geq \lambda \text{ and } s = 0, \qquad \forall v \in (f \geq \lambda), \forall t \geq 0.
\end{aligned}$$

Hence, $R_A = \left\{(w, 0) : w \in R_{(f\geq\lambda)}\right\}$. On the other hand, $A = \mathrm{ipo}(f) \cap \{(v, \lambda) : v \in V\}$. Since $R_{\{(v,\lambda):v\in V\}} = \{(w, 0) : w \in V\}$, by Lemma 484-(iii) we have $R_A = R_{\mathrm{ipo}(f)} \cap R_{\{(v,\lambda):v\in V\}} = \left\{(w, 0) : (w, 0) \in R_{\mathrm{ipo}(f)}\right\}$. We conclude that

$$\left\{(w, 0) \in V \times \mathbb{R} : w \in R_{(f\geq\lambda)}\right\} = \left\{(w, 0) \in V \times \mathbb{R} : (w, 0) \in R_{\mathrm{ipo}(f)}\right\},$$

[7] Recall that the ipograph $\mathrm{ipo}(f)$ was defined in (8.9).

i.e., $R_{(f\geq\lambda)} = \left\{w \in V : (w,0) \in R_{\text{ipo}(f)}\right\}$. ■

Next we characterize R_f for concave functions. In particular, point (iii) shows that R_f is the set of all directions of increase of f, that is, the directions along which f increases.

Proposition 487 *Let $f : C \to \mathbb{R}$ be an upper semicontinuous concave function defined on a closed convex subset C of V. Then, the following conditions are equivalent:*

(i) $w \in R_f$;

(ii) $f(v+tw) \geq f(v)$ *for all* $t \geq 0$ *and all* $v \in C$;

(iii) $f(v+tw)$ *is, as a function of* t, *nondecreasing on* $[0,\infty)$ *for all* $v \in C$;

(iv) $\lim_{t\to\infty} f(v+tw) > -\infty$ *for all* $v \in C$;

(v) $\lim_{t\to\infty} f(v+tw)/t \geq 0$ *for all* $v \in C$.

Remark. It is easy to see that properties (i)-(iii) are equivalent even if f is not upper semicontinuous.

Proof. (i) implies (ii). Let $v \in C$ and let $t \geq 0$. Fix $\varepsilon > 0$. We have $v \in (f \geq f(v) - \varepsilon)$, and so $w \in \bigcap_{\lambda\in\mathbb{R}} R_{(f\geq\lambda)}$ implies $v + tw \in (f \geq f(v) - \varepsilon)$ for all $t \geq 0$, i.e., $f(v+tw) \geq f(v) - \varepsilon$. Since ε is arbitrary, we conclude that $f(v+tw) \geq f(v)$.

(ii) implies (iii). Let $v \in C$. Let $t' > t''$. As $f(v+tw) \geq f(v)$ for all $t \geq 0$, we have $v + tw \in C$ for all $t \geq 0$. Hence, $f(v+t'w) = f(v + (t'-t'')w + t''w) \geq f(v+t''w)$ since $v + t''w \in C$.

(iii) trivially implies (iv).

(iv) implies (v). By Exercise 13.0.66, it is enough consider the function $\phi : [0,\infty) \to \mathbb{R}$ defined by $\phi(t) = f(v+tw)$.

(v) implies (i). Consider again the function ϕ. By Exercise 13.0.66, ϕ is nondecreasing. Hence, $f(v+tw) = \phi(t) \geq \phi(0) = f(v)$. That is, $(v, f(v)) \in \text{ipo}(f)$ is such that $(v, f(v)) + t(w,0) \in \text{ipo}(f)$ for all $t \geq 0$. By Lemma 484-(i), this implies $(w,0) \in R_{\text{ipo}(f)}$, and so, by Lemma 486, $w \in R_f$. ■

Though conceptually illuminating, the properties established in Properties 487 are less useful to actually find the elements of R_f. The next result shows that points (iv) and (v) still characterize the elements of R_f if they just hold for some $v \in V$. This greatly simplifies the identification of the elements of R_f.

Proposition 488 *Let $f : C \to \mathbb{R}$ be an upper semicontinuous and concave function defined on a closed convex subset C of V. Then, the following conditions are equivalent:*

(i) $w \in R_f$;

(ii) there is $v \in C$ *such that* $\lim_{t\to\infty} f(v+tw) > -\infty$;

(iii) there is $v \in C$ *such that* $\lim_{t\to\infty} f(v+tw)/t \geq 0$.

Proof. (i) implies (iii) by Theorem 487.

(iii) implies (ii). Let $v_0 \in C$ be such that $\lim_{t\to\infty} f(v_0+tw)/t \geq 0$. Define $\phi : [0,\infty) \to \mathbb{R}$ by $\phi(t) = f(v_0+tw)$. By Exercise 13.0.66, $\lim_{t\to\infty} f(v_0+tw) = \lim_{t\to\infty} \phi(t) > -\infty$.

(ii) implies (i). Let $v_0 \in C$ be such that $\lim_{t\to\infty} f(v_0+tw) > -\infty$. Define $\phi : [0,\infty) \to [-\infty,\infty)$ by $\phi(t) = f(v_0+tw)$. The function ϕ is concave. Since $\lim_{t\to\infty} f(v_0+tw) > -\infty$, by Exercise 13.0.66 ϕ is nondecreasing. Hence, $f(v_0+tw) = \phi(t) \geq \phi(0) = f(v_0)$. That is, $(v_0, f(v_0)) \in \text{ipo}(f)$ is such that $(v_0, f(v_0)) + t(w,0) \in \text{ipo}(f)$ for all $t \geq 0$. By Lemma 484-(i), this implies $(w,0) \in R_{\text{hyp}(f)}$, and so, by Lemma 486, $w \in R_f$. ■

Example 489 Let $f : V \to \mathbb{R}$ be an upper semicontinuous superlinear functional. Set $v = \mathbf{0}$ in Proposition 488. Then, $w \in R_f$ if and only if $f(w) = \lim_{t\to\infty} f(tw)/t \geq 0$. Hence, $R_f = \{w \in V : f(w) \geq 0\}$. △

A vector $w \in V$ is a direction of recession if, given any $v \in C$, we remain in C by moving forward along the direction w, i.e., $v + tw$ for all $t \geq 0$. The next stronger definition requires that this happens by moving both backward and forward, i.e., $v + tw \in C$ for all $t \in \mathbb{R}$.

Definition 490 *The lineality space L_C of a set C of V is defined by*

$$L_C = \{w \in V : v + tw \in C \text{ for all } v \in C \text{ and all } t \in \mathbb{R}\},$$

with the convention $L_\emptyset = V$.

Under this stronger condition we get a vector space and not only a cone.

Lemma 491 *The lineality space L_C is a vector space, with*

$$L_C = R_C \cap R_{-C} = R_C \cap -R_C. \tag{8.39}$$

Proof. It is easy to check that L_C is a vector space. Let $w \in R_C \cap R_{-C}$. Given any $t < 0$ and $v \in C$, consider $v + tw$. Then, $-v \in -C$ and so $-v + (-t)\, w \in -C$, i.e., $v + tw \in C$. This shows that $w \in L_C$. Conversely, let $w \in L_C$. Clearly, $L_C \subseteq R_C$. Moreover, given any $t < 0$ and $v \in C$, we have $v + tw \in C$, i.e., $-v + (-t)\, w \in -C$. This implies $w \in R_{-C}$.

It remains to show that $R_{-C} = -R_C$. We have

$$\begin{aligned} w \quad \in \quad & R_{-C} \Longleftrightarrow v + tw \in -C \quad \forall v \in -C, \forall t \geq 0 \Longleftrightarrow z + t\,(-w) \in C \quad \forall z \in C, \forall t \geq 0 \\ \Longleftrightarrow \quad & -w \in R_C \Longleftrightarrow w \in -R_C, \end{aligned}$$

and so $R_{-C} = -R_C$. ■

Definition 492 *The lineality space of a function $f : C \to \mathbb{R}$ is defined by*

$$L_f = \bigcap_{\lambda \in \mathbb{R}} L_{(f \geq \lambda)},$$

Next we show that also the vector spaces $L_{(f \geq \lambda)}$ coincide for the nonempty upper level sets $(f \geq \lambda)$ of concave functions.

Lemma 493 *Let $f : C \to \mathbb{R}$ be an upper semicontinuous concave function. Then,*

$$L_f = L_{(f \geq \lambda)}$$

for all nonempty $(f \geq \lambda)$.

Proof. Define the auxiliary function $g : -C \to \mathbb{R}$ by $g(v) = f(-v)$ for all $v \in -C$. The function g is upper semicontinuous, proper, and concave if f is. We show that

$$R_{-(f \geq \lambda)} = \left\{ w \in V : (w, 0) \in R_{\text{hyp}(g)} \right\} \tag{8.40}$$

for all $\lambda \in \mathbb{R}$ such that $-(f \geq \lambda) \neq \emptyset$, i.e., such that $(f \geq \lambda) \neq \emptyset$. Fix $\lambda \in \mathbb{R}$ such that $-(f \geq \lambda) \neq \emptyset$. We have $R_{-(f \geq \lambda)} = R_{(g \geq \lambda)}$. In fact, $w \in R_{-(f \geq \lambda)}$ if and only $v + tw \in -(f \geq \lambda)$ for all $v \in -(f \geq \lambda)$ and all $t \geq 0$, i.e., if and only if $g(v + tw) \geq \lambda$ for all $v \in (g \geq \lambda)$ and all $\lambda \geq 0$. By Lemma 486, (8.40) holds.

We conclude that $R_{-(f \geq \lambda)}$ are all equal provided $(f \geq \lambda) \neq \emptyset$. In turn, this implies $L_f = L_{(f \geq \lambda)}$ for all $(f \geq \lambda) \neq \emptyset$. ■

Propositions 487 and 488 take the following form for lineality spaces.

Proposition 494 *Let $f : C \to \mathbb{R}$ be an upper semicontinuous concave function defined on a closed convex subset C of V. Then, the following conditions are equivalent:*

(i) $w \in L_f$;

(ii) $f(v+tw) = f(v)$ *for all* $t \in \mathbb{R}$ *and all* $v \in C$*;*

(iii) $f(v+tw)$ *is, as a function of* t*, constant on* $\mathbb{R}$ *for all* $v \in C$*;*

(iv) $\lim_{t\to\pm\infty} f(v+tw) > -\infty$ *for all* $v \in C$*;*

(v) $\lim_{t\to\pm\infty} f(v+tw)/t = 0$ *for all* $v \in C$*;*

(vi) there is $v \in C$ *such that* $\lim_{t\to\pm\infty} f(v+tw) > -\infty$*;*

(vii) there is $v \in C$ *such that* $\lim_{t\to\pm\infty} f(v+tw)/t = 0$*.*

Proof. (i) implies (ii). Let $w \in L_f$. Fix $v \in C$. By Proposition 487, $f(v+tw) \geq f(v)$ for all $t \geq 0$. Let $t < 0$ and consider $v + tw$. Since $-v \in -(f \geq \lambda)$ and $w \in R_{-(f\geq\lambda)}$, we have $-(v+tw) = -v + (-t)w \in -(f \geq \lambda)$, i.e., $v + tw \in (f \geq \lambda)$. We thus have $f(v+tw) \geq f(v)$ for all $t \in \mathbb{R}$ and all $v \in C$.

The other implications follow by applying Exercise 13.0.67 to the function $\phi : \mathbb{R} \to \mathbb{R}$ given by $\phi(t) = f(v+tw)$. ■

Example 495 Let $f : V \to \mathbb{R}$ be an upper semicontinuous superlinear functional. Set $v = \mathbf{0}$ in Proposition 494. Then, $w \in L_f$ if and only if

$$f(w) = \lim_{t\to\infty} \frac{f(tw)}{t} = 0 \quad \text{and} \quad -f(-w) = \lim_{t\to-\infty} \frac{f(-t(-w))}{t} = \lim_{t\to-\infty} \frac{f(tw)}{t} = 0,$$

that is, $L_f = \{w \in V : f(w) = -f(-w) = 0\}$. △

Example 496 Given a collection $\{F_i\}_{i\in I}$ of lower semicontinuous convex functionals $F_i : V \to \mathbb{R}$ and a collection $\{b_i\}_{i\in I} \subseteq \mathbb{R}$, consider the set

$$C = \{v \in V : F_i(v) \leq b_i \text{ for all } i \in I\}.$$

If the closed and convex set C is nonempty, then

$$L_C = \bigcap_{i\in I} L_{F_i}. \tag{8.41}$$

For, suppose $w \in \bigcap_{i\in I} L_{F_i}$. Fix $i \in I$. By Proposition 494-(ii), $F_i(v+tw) = F_i(v)$ for all $v \in V$ and all $t \in \mathbb{R}$. Hence, if $v \in C$ and $t \in \mathbb{R}$, we have $F_i(v+tw) = F_i(v) \leq b_i$, i.e., $v + tw \in C$. Since this holds for any $i \in I$, we conclude that $w \in L_C$. Conversely, suppose $w \in L_C$. Let $v \in C$, i.e., $F_i(v) \leq b_i$ for all $i \in I$. Then, for all $t \in \mathbb{R}$ we have $v + tw \in C$, i.e., $F_i(v+tw) \leq b_i$ for all $i \in I$. Define $\phi_i : \mathbb{R} \to \mathbb{R}$ by $\phi_i(t) = F_i(v+tw)$. Each function ϕ_i is convex and $\phi_i(t) \leq b_i$ for all $t \in \mathbb{R}$. By Exercise 13.0.67, $\phi_i(t) = \phi_i(0)$ for all $t \in \mathbb{R}$, and so $F_i(v+tw) = F_i(v) \leq b_i$ for all $i \in I$. Hence, $w \in \bigcap_{i\in I} L_{F_i}$. ▲

We can finally state and prove the main result of this section, an existence result for global maxima of concave functions that does not rely on any compactness assumption. Recall that we already observed that $\arg\max_{v\in B} f(v)$ is a convex set.

Theorem 497 *Let B and C be nonempty convex closed subsets of a finite dimensional normed vector space V, with $B \subseteq C$. If a concave function $f : C \to \mathbb{R}$ is upper semicontinuous on B, then,*

(i) $\arg\max_{v\in B} f(v)$ *is nonempty if* $R_B \cap R_f = L_B \cap L_f$*;*

(ii) $\arg\max_{v\in B} f(v)$ *is nonempty and compact if and only if* $R_B \cap R_f = \{0\}$*.*

By Lemmas 486 and 484-(iii), $R_B \cap R_f = R_B \cap R_{(f\geq t)} = R_{B\cap(f\geq t)}$ for each $t \in \mathbb{R}$ such that $(f \geq t)\cap B$ is nonempty. Similarly, $L_B \cap L_f = L_{B\cap(f\geq t)}$. Hence, the condition $R_B \cap R_f = L_B \cap L_f$ requires that each nonempty set $(f \geq t) \cap B$ be "symmetrically unbounded:" w is a direction of recession of $(f \geq t) \cap B$ if and only if also $-w$ is.

It is important to observe that in this finite dimensional convex setup, Theorem 497 improves Theorem 317, the fundamental Weierstrass-type existence result proved in Section 6.6. For, let B and C as in Theorem 497 and suppose that the concave function $f : C \to \mathbb{R}$ is upper semicontinuous and coercive on B. Then, there is $t \in \mathbb{R}$ such that $(f \geq t) \cap B$ is compact and nonempty and so, by Lemmas 484-(iii) and 499,

$$R_B \cap R_f = R_B \cap R_{(f\geq t)} = R_{B\cap(f\geq t)} = \{0\}.$$

Hence, $\arg\max_{v\in B} f(v)$ is nonempty and compact by Theorem 497-(ii). This shows that Theorem 317 is a special case of Theorem 497 in this finite dimensional convex setup. A constant function $f : \mathbb{R}^n \to \mathbb{R}$ is a trivial instance where Theorem 497, but not Theorem 317, applies (see Example 320).

A final remark: an important example when the convex set $\arg\max_{v\in B} f(v)$ is nonempty and compact is when there is a unique a solution to the optimal problem $\max_{v\in B} f(v)$. Hence, $R_B \cap R_f = \{0\}$ is also a necessary condition for such uniqueness.

Before proving it, we illustrate Theorem 497 with some examples.

Example 498 Given a superlinear (and so continuous) functional $f : \mathbb{R}^n \to \mathbb{R}$, consider $\max_{x\in B} f(x)$. In view of Examples 489 and 495, by Theorem 497-(i) we have that $\arg\max_{x\in B} f(x)$ is nonempty provided

$$\{x \in L_B : f(x) = -f(-x) = 0\} = \{x \in R_B : f(x) \geq 0\},$$

which is equivalent to

$$f(x) \geq 0 \Longrightarrow x \in -R_B \text{ and } -f(-x) = 0, \qquad \forall x \in R_B. \tag{8.42}$$

For example, consider the superlinear functional

$$f(x) = \min_{i=1,\ldots,n} x_i, \qquad \forall x \in \mathbb{R}^n.$$

Intuitively, $\arg\max_{x \in B} f(x)$ is nonempty if the set B has no positive directions of recession, i.e., if $R_B \cap \mathbb{R}^n_+ = \{\mathbf{0}\}$. In fact, these are directions along which f can keep growing. Condition (8.42) makes precise this simple insight. For, let $x \in R_B$ be such that $f(x) \geq 0$. From $f(x) \geq 0$ it follows that $x \in \mathbb{R}^n_+$, and so $R_B \cap \mathbb{R}^n_+ = \{\mathbf{0}\}$ implies $x = 0$. Hence, condition (8.42) holds and we conclude that $\arg\max_{x \in B} f(x)$ is nonempty. ▲

The proof of Theorem 497 relies on two important lemmas, of independent interest. The first one formalizes the intuition that a set without directions of recession is bounded, an intuition that turns out to be correct as long as we consider finite dimensional spaces.

Lemma 499 *Let C be a closed convex subset of a finite dimensional normed vector space V. Then, C is bounded (and so compact) if and only if $R_C = \{0\}$.*

Proof. We only prove the "if" part, the converse being trivial. Suppose $R_C = \{0\}$ and suppose, *per contra*, that C is unbounded. Then, there is a sequence $\{v_n\}_n \subseteq C$ with $\|v_n\| \uparrow \infty$. Then, each $v_n / \|v_n\|$ belongs to the unit ball B_V of V and so, being B_V compact, there is $z \in B_V$ and a subsequence $\{v_{n_k}\}_k$ such that $v_{n_k} / \|v_{n_k}\| \to z$. Let $v \in C$ and $t \geq 0$. Set

$$z_k = \left(1 - \frac{t}{\|v_{n_k}\|}\right) v + \frac{t}{\|v_{n_k}\|} v_{n_k}, \qquad \forall k \geq 1.$$

We have $z_k \in C$ for all k large enough so that $t / \|v_{n_k}\| \leq 1$. Moreover, $z_k \to v + tz$. Since C is closed, this implies $v + tz \in C$. We conclude that $z \in R_C$, which contradicts $R_C = \{0\}$ since z belongs to B_V and so is nonzero. ■

The next lemma gives a condition under which a monotone sequence of closed convex sets has nonempty intersection. It generalizes Lemma 266 and is close in spirit to Lemma 279. Here as well the finite dimensionality of V is required.

Lemma 500 *Let $\{C_n\}_n$ be a monotone sequence of closed convex sets, with $C_1 \supset \cdots \supset C_n \supset \cdots$, of a finite dimensional normed vector space V. Then $\bigcap_n C_n \neq \emptyset$ provided $\bigcap_n R_{C_n} = \bigcap_n L_{C_n}$.*

Proof. Given any $\overline{v}_n \in C_n$, by Theorem 364 the set $\{v \in C : \|v\| \leq \|\overline{v}_n\|\}$ is compact since V is finite dimensional. We can then take $v_n^* \in \arg\min_{v \in C_n} \|v_n\|$ for all $n \geq 1$. The monotone sequence $\{\|v_n^*\|\}_n$ is bounded. Suppose, *per contra*, that $\|v_n^*\| \uparrow \infty$. The sequence $v_n^*/\|v_n^*\|$ belongs to the unit ball of V, which is compact by Theorem 364, and so there exist a subsequence $\left\{v_{n_k}^*\right\}_k$ and $w \in \{v \in C : \|v\| \leq 1\}$ such that $\lim_k \left(v_{n_k}^*/\left\|v_{n_k}^*\right\|\right) = w$.

Fix $m \geq 1$. Then, $v_{n_k}^* \in C_m$ for all k large enough. By Lemma 484-(ii), $w \in R_{C_m}$. Since m is arbitrary, $w \in \bigcap_n R_{C_n}$, and so, by hypothesis, $w \in \bigcap_n L_{C_n}$.

Since $\lim_k \left(v_{n_k}^*/\left\|v_{n_k}^*\right\|\right) = w$, we have

$$\lim_k \left\| \frac{v_{n_k}^* - \left\|v_{n_k}^*\right\| w}{\left\|v_{n_k}^*\right\|} \right\| = 0. \tag{8.43}$$

Moreover, $w \in \bigcap_n L_{C_n}$ implies $v_{n_k}^* + t_k w \in C_{n_k}$ for all $t_k \in \mathbb{R}$ and all $k \geq 1$. Then, $\left\|v_{n_k}^* + t_k w\right\| \geq \left\|v_{n_k}^*\right\|$ for all $k \geq 1$ since by construction each $v_{n_k}^*$ is a minimum norm vector. Setting $t_k = -\left\|v_{n_k}^*\right\|$, we then have

$$\left\| \frac{v_{n_k}^* - \left\|v_{n_k}^*\right\| w}{\left\|v_{n_k}^*\right\|} \right\| \geq 1, \qquad \forall k \geq 1, \tag{8.44}$$

which contradicts (8.43). We conclude that the sequence $\{\|v_n^*\|\}_n$ is bounded. Since V is finite dimensional, there is a subsequence $\left\{v_{n_k}^*\right\}_k$ and $z \in V$ such that $\lim_k v_{n_k}^* = z$. It is easy to check that $z \in \bigcap_n C_n$, and so $\bigcap_n C_n \neq \emptyset$. ■

Proof of Theorem 497. (i) Set $\alpha = \sup_{v \in B} f(v)$ and consider an increasing sequence $\{\alpha_n\}_n \subseteq \mathbb{R}$ with $\alpha_n \uparrow \alpha$. Set $B_n = (f \geq \alpha_n) \cap B$ for all $n \geq 1$. Since f is upper semicontinuous and concave, each B_n is closed and convex. Moreover, $C \cap B \neq \emptyset$ implies that each B_n is nonempty. Then, by Lemma 484-(iii),

$$R_{B_n} = R_{(f \geq \alpha_n) \cap B} = R_{(f \geq \alpha_n)} \cap R_B = R_f \cap R_B.$$

Similarly, $L_{B_n} = L_f \cap L_B$, and so $R_{B_n} = L_{B_n}$ for all $n \geq 1$. By Theorem 500, $\bigcap_n B_n \neq \emptyset$. Let $v^* \in \bigcap_n B_n$. Since f is upper semicontinuous, we have

$$f(v^*) \geq \limsup_n f(v_n^*) \geq \lim_n \alpha_n = \alpha$$

and so $v^* \in \arg\max_{v \in B} f(v)$.

(ii) Suppose $R_B \cap R_f = \{0\}$. Then, the condition $R_B \cap R_f \subseteq L_B \cap L_f$ is trivially satisfied, and so $\arg\max_{v \in B} f(v) \neq \emptyset$. Moreover, in this case we have $R_{B_n} = R_f \cap R_B = \{0\}$ for each $n \geq 1$, and so, by Theorem 499, each B_n is compact. This implies that $\arg\max_{v \in B} f(v)$ is compact.

Conversely, suppose $\arg\max_{v\in B} f(v)$ is nonempty and compact. Set $\alpha = \max_{v\in B} f(v)$. Then, $\arg\max_{v\in B} f(v) = (f \geq \alpha) \cap B$ and so, by Theorem 499,

$$\{0\} = R_{\arg\max_{v\in B} f(v)} = R_{(f\geq\alpha)\cap B} = R_{(f\geq\alpha)} \cap R_B = R_f \cap R_B,$$

as desired. ■

Chapter 9

Classical Constrained Optimization

9.1 Introduction

In Chapter 5 we considered the problem of unconstrained optimization for functions $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$, in which the search of maxima and minima was done on all the domain A of the function, which was assumed to be an open set.

Nevertheless, in many economic problems the results of Chapter 5 are not very satisfactory. To see this, consider the classical problem of the consumer. In this problem a consumer has to choose the best bundle of goods for him, under the constraint of a given wealth and given prices of the goods.

Formally, suppose that the preferences of the consumer are represented by a utility function $u : A \subseteq \mathbb{R}^n \to \mathbb{R}$, where A is an (open) set of bundles of goods, and that his wealth is $b \in \mathbb{R}$. Moreover, we denote by $p \in \mathbb{R}^n_+$ the vector of the prices of the goods.

If we assume that the consumer must spend all his income, his budget constraint is given by the set

$$C(p, b) = \{x \in A : p \cdot x = b\}$$

and his problem of optimum is given by:

$$\max_{x \in C(p,b)} u(x), \tag{9.1}$$

whose solutions have necessarily to be searched in the subset of A given by $C(p, b)$.

Two are the aspects of this example to which pay attention:

- the search of the points of maximum does not take place on the entire domain of the function, but on the subset $C(p, b)$ of $\mathbb{R}^n$ of the points that satisfy an equality constraint defined by the function $g : \mathbb{R}^n \to \mathbb{R}$ given by $g(x) = p \cdot x$ for each $x \in \mathbb{R}^n$;

- the points of interest are those of global maximum, i.e., the bundles of goods that are the best possible ones that the consumer can get given the constrains he faces. Instead, the points that are only local maxima are of little interest.

In this chapter we will therefore try to solve problems of optimum in which the search of the solutions is restricted to a subset of the domain of the function determined by equality constraints (domain that we will continue to assume open) and in which the solutions of interest are global. We will see that the resolution of these problems is an elegant combination of local methods based on differential calculus and of global existence results à la Weierstrass.

9.2 Formalization of the Problem

The general form of a problem of optimum with equality constraints is given by

$$\begin{array}{rcl} & & \max_{x \in A} f(x) \\ \text{sub } g_1(x) & = & b_1,\, g_2(x) = b_2,\, ...,\, g_m(x) = b_m, \end{array} \tag{9.2}$$

where $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ is our objective function, while the functions $g_i : A \subseteq \mathbb{R}^n \to \mathbb{R}$ and the scalars $b_i \in \mathbb{R}$ induce m equality constraints.

In an analogous way we can define a minimization problem:

$$\begin{array}{rcl} & & \min_{x \in A} f(x) \\ \text{sub } g_1(x) & = & b_1,\, g_2(x) = b_2,\, ...,\, g_m(x) = b_m. \end{array} \tag{9.3}$$

Even if these two problems are specular, we will give more emphasis to the problem (9.2) due to its greater relevance in economics. Furthermore, a problem of minimum can be always transformed in a problem of maximum by considering $-f$ as objective function. In other words, (9.3) is equivalent to:

$$\begin{array}{rcl} & & \max_{x \in A} -f(x) \\ \text{sub } g_1(x) & = & b_1,\, g_2(x) = b_2,\, ...,\, g_m(x) = b_m. \end{array}$$

The set

$$C = \{x \in A : g_i(x) = b_i \text{ for each } i = 1, ..., m\}, \tag{9.4}$$

is the subset of A identified by the constraints, and hence the problem of optimum (9.2) can be equivalently formulated as:

$$\max_{x \in C} f(x).$$

A point $\widehat{x} \in B$ is a (global) *solution* of the problem of optimum (9.2) if $f(\widehat{x}) \geq f(x)$ for each $x \in C$, while $\widehat{x} \in B$ is called a *local solution* of such problem if there exists a neighborhood $B_{x_0}(\varepsilon)$ of $\widehat{x}$ such that $f(\widehat{x}) \geq f(x)$ for each $x \in B_{x_0}(\varepsilon) \cap C$. Obviously, a solution of the problem of optimum is also a local solution.

As we already observed, in the resolution of the problem of optimum a fundamental role is played by the results of existence à la Weierstrass seen in Section 6.6, and in particular Theorem 317, which was the most general among them. Though very interesting, these results of existence have however the strong limit of telling nothing on how to find these solutions. It is therefore necessary to develop some techniques that allow us to find a set, hopefully small, of points that are the possible candidates to be solutions of the problem of optimum, and on which to concentrate the search of the solutions.

Since the set C is not in general open,[1] the first order and second order conditions seen in Chapter 5 cannot be applied to this problem. Fortunately, there exists a classical method that allows to give first order and second order conditions also in presence of constraints. In next section we begin to illustrate it for the case of a unique equality constraint.

9.3 One Constraint

In this case the problem of optimum (9.2) has the form:

$$\begin{aligned} &\max_{x \in A} f(x) \\ &\text{sub } g(x) = b \end{aligned} \tag{9.5}$$

where $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ is our objective function, A is an open set, while the function $g : A \subseteq \mathbb{R}^n \to \mathbb{R}$ and the scalar $b \in \mathbb{R}$ define the equality constraint.

The next fundamental lemma gives us the key to find the solutions of problem 9.5.

Lemma 501 *Let $\widehat{x}$ be a local solution of the problem of optimum (9.5). If f and g are of class $\mathcal{C}^1$ and if $\nabla g(\widehat{x}) \neq \mathbf{0}$, then there exists a scalar $\widehat{\lambda} \in \mathbb{R}$ such that*

$$\nabla f(\widehat{x}) = \widehat{\lambda} \nabla g(\widehat{x}). \tag{9.6}$$

It should be noticed that we have assumed A to be an open set of $\mathbb{R}^n$. Clearly we can drop this assumption by assuming that the local solution $\widehat{x} \in int(A)$.

[1] Recall that the empty set and the set $\mathbb{R}^n$ itself are the only sets that are both open and closed in $\mathbb{R}^n$.

Proof The proof of this lemma is a special case of that of Lemma 509 that we will see below. ■

Expression (9.6) tells us that a necessary condition for $\widehat{x}$ to be local solution of the problem of optimum (9.5) is that the gradients of the functions f and g are between them proportional. The "hat" over λ reminds us that such scalar depends on the point $\widehat{x}$ considered.

Next example shows how condition (9.6) is necessary, but not sufficient.

Example 502 Consider the problem of optimum:

$$\begin{array}{l} \max\limits_{x\in\mathbb{R}} \dfrac{x_1^3+x_2^3}{2} \\ \text{sub } x_1 - x_2 = 0 \end{array} \tag{9.7}$$

It is of the form (9.5), where $f : \mathbb{R}^2 \to \mathbb{R}$ and $g : \mathbb{R}^2 \to \mathbb{R}$ are given by $f(x) = 2^{-1}(x_1^3 + x_2^3)$ and $g(x) = x_1 - x_2$, while $b = 0$. We have:

$$\nabla f(0,0) = (0,0) \quad \text{and} \quad \nabla g(0,0) = (1,-1)$$

and hence $\lambda = 0$ is such that $\nabla f(0,0) = \lambda \nabla g(0,0)$. The point $(0,0)$ satisfies therefore with $\lambda = 0$ condition (9.6), but this point is not solution of the problem of optimum (9.7). In fact,

$$f(t,t) = t^3 > 0 = f(0,0), \qquad \forall t > 0. \tag{9.8}$$

Notice that $(0,0)$ is not even a constrained (global) minimum since $f(t,t) = t^3 < 0$ for each $t < 0$. ▲

To understand at an intuitive level condition (9.6), assume that f and g are defined on $\mathbb{R}^2$, so that (9.6) has the form:

$$\left(\frac{\partial f}{\partial x_1}(\widehat{x}), \frac{\partial f}{\partial x_2}(\widehat{x})\right) = \widehat{\lambda}\left(\frac{\partial g}{\partial x_1}(\widehat{x}), \frac{\partial g}{\partial x_2}(\widehat{x})\right),$$

that is,

$$\frac{\partial f}{\partial x_1}(\widehat{x}) = \widehat{\lambda}\frac{\partial g}{\partial x_1}(\widehat{x}) \qquad \text{and} \qquad \frac{\partial f}{\partial x_2}(\widehat{x}) = \widehat{\lambda}\frac{\partial g}{\partial x_2}(\widehat{x}). \tag{9.9}$$

The condition $\nabla g(\widehat{x}) \neq \mathbf{0}$ requires that at least one of the two partial derivatives $(\partial g/\partial x_i)(\widehat{x})$ is different from zero. If, for simplicity, we suppose that both of them are so and that $\widehat{\lambda} \neq 0$, then (9.9) is equivalent to

$$\frac{\frac{\partial f}{\partial x_1}(\widehat{x})}{\frac{\partial g}{\partial x_1}(\widehat{x})} = \frac{\frac{\partial f}{\partial x_2}(\widehat{x})}{\frac{\partial g}{\partial x_2}(\widehat{x})} \tag{9.10}$$

We try now to understand, through an heuristic argument, what is the intuition for (9.10) to be a necessary condition for $\widehat{x}$ to be a solution of the optimum problem (9.5). The differential of f and g at the point $\widehat{x}$ is given by

$$\begin{aligned} df\left(\widehat{x}\right)(h) &= \nabla f\left(\widehat{x}\right)\cdot h = \frac{\partial f}{\partial x_1}\left(\widehat{x}\right)h_1 + \frac{\partial f}{\partial x_2}\left(\widehat{x}\right)h_2, \qquad \forall h \in \mathbb{R}^2, \\ dg\left(\widehat{x}\right)(h) &= \nabla g\left(\widehat{x}\right)\cdot h = \frac{\partial g}{\partial x_1}\left(\widehat{x}\right)h_1 + \frac{\partial g}{\partial x_2}\left(\widehat{x}\right)h_2, \qquad \forall h \in \mathbb{R}^2, \end{aligned}$$

and provides a linear approximation of the difference $f\left(\widehat{x}+h\right)-f\left(\widehat{x}\right)$ and $g\left(\widehat{x}+h\right)-g\left(\widehat{x}\right)$, respectively, i.e., of the effect on f and g that results from moving from $\widehat{x}$ to $\widehat{x}+h$. As well know, this approximation is the better the smaller is h. Suppose, heuristically, that h is infinitesimal and that the approximation is exact, so that $f\left(\widehat{x}+h\right)-f\left(\widehat{x}\right) = df\left(\widehat{x}\right)(h)$ and $g\left(\widehat{x}+h\right)-g\left(\widehat{x}\right) = dg\left(\widehat{x}\right)(h)$. Formally this is clearly incorrect, but here we are proceeding heuristically, trying to understand at an intuitive level what lies behind (9.10).

Proceeding heuristically, we start now from our point $\widehat{x}$ and consider changes $\widehat{x}+h$ with h infinitesimal. The first problem to face is to make sure that they are legitimate, i.e., that they respect the equality constraint $g\left(\widehat{x}+h\right)=b$. This means that $g\left(\widehat{x}+h\right)=g\left(\widehat{x}\right)$, and hence h must be such that $dg\left(\widehat{x}\right)(h)=0$. It follows that:

$$\frac{\partial g}{\partial x_1}\left(\widehat{x}\right)h_1 + \frac{\partial g}{\partial x_2}\left(\widehat{x}\right)h_2 = 0,$$

and so

$$h_1 = -\frac{\frac{\partial g}{\partial x_2}\left(\widehat{x}\right)}{\frac{\partial g}{\partial x_1}\left(\widehat{x}\right)}h_2. \tag{9.11}$$

The effect of moving from $\widehat{x}$ to $\widehat{x}+h$ on our objective function f is given by $df\left(\widehat{x}\right)(h)$. When h is legitimate, by (9.11) such effect is given by:

$$df\left(\widehat{x}\right)(h) = \frac{\partial f}{\partial x_1}\left(\widehat{x}\right)\left(-\frac{\frac{\partial g}{\partial x_2}\left(\widehat{x}\right)}{\frac{\partial g}{\partial x_1}\left(\widehat{x}\right)}\right)h_2 + \frac{\partial f}{\partial x_2}\left(\widehat{x}\right)h_2. \tag{9.12}$$

If $\widehat{x}$ is solution of our optimum problem, we must necessarily have $df\left(\widehat{x}\right)(h)=0$ for each legitimate change h. In fact, if it were $df\left(\widehat{x}\right)(h)>0$ such change would give us a point $\widehat{x}+h$ that satisfies the equality constraint and such that $f\left(\widehat{x}+h\right)>f\left(\widehat{x}\right)$. On the other hand, if it were $df\left(\widehat{x}\right)(h)<0$ a similar argument, this time for $-h$ (which is obviously a legitimate change), would lead us to the point $\widehat{x}-h$ with $f\left(\widehat{x}-h\right)>f\left(\widehat{x}\right)$.

The necessary condition of optimum $df\left(\widehat{x}\right)(h)=0$ together with (9.12) gives us:

$$\frac{\partial f}{\partial x_1}\left(\widehat{x}\right)\left(-\frac{\frac{\partial g}{\partial x_2}\left(\widehat{x}\right)}{\frac{\partial g}{\partial x_1}\left(\widehat{x}\right)}\right)h_2 + \frac{\partial f}{\partial x_2}\left(\widehat{x}\right)h_2 = 0.$$

If, as natural, we assume $h_2 > 0$, we have

$$\frac{\partial f}{\partial x_1}(\widehat{x})\left(-\frac{\frac{\partial g}{\partial x_2}(\widehat{x})}{\frac{\partial g}{\partial x_1}(\widehat{x})}\right)+\frac{\partial f}{\partial x_2}(\widehat{x})=0,$$

which is exactly (9.10). Intuitively, all this explains why (9.6) is a necessary condition for $\widehat{x}$ to be solution of the optimum problem.

Lemma 501 therefore established a necessary condition for optimality, with a clear intuitive meaning. This condition can be equivalently written as

$$\nabla f(\widehat{x})-\widehat{\lambda}\nabla g(\widehat{x})=\mathbf{0}.$$

Remembering the algebra of gradients, the expression $\nabla f(x)-\lambda\nabla g(x)$ leads naturally to think of the function $L: A\times\mathbb{R}\subseteq\mathbb{R}^{n+1}\to\mathbb{R}$ defined by:

$$L(x,\lambda)=f(x)+\lambda(b-g(x)),\qquad\forall(x,\lambda)\in A\times\mathbb{R}. \tag{9.13}$$

This function is called the *Lagrangian function* and plays a fundamental role in optimization problems. Its gradient is

$$\nabla L(x,\lambda)=\left(\frac{\partial L}{\partial x_1}(x,\lambda),...,\frac{\partial L}{\partial x_n}(x,\lambda),\frac{\partial L}{\partial\lambda}(x,\lambda)\right)\in\mathbb{R}^{n+1}$$

and it is important to distinguish in it the two parts $\nabla_x L$ and $\nabla_\lambda L$ given by:

$$\begin{aligned}\nabla_x L(x,\lambda) &= \left(\frac{\partial L}{\partial x_1}(x,\lambda),...,\frac{\partial L}{\partial x_n}(x,\lambda)\right)\in\mathbb{R}^n,\\ \nabla_\lambda L(x,\lambda) &= \frac{\partial L}{\partial\lambda}(x,\lambda)\in\mathbb{R}.\end{aligned}$$

Using this notation, we have

$$\nabla_x L(x,\lambda)=\nabla f(x)-\lambda\nabla g(x) \tag{9.14}$$

and

$$\nabla_\lambda L(x,\lambda)=b-g(x), \tag{9.15}$$

which leads us to the following fundamental formulation in terms of the Lagrangian function of the necessary optimality condition of Lemma 501.

Theorem 503 *Let $\widehat{x}$ be a local solution of the optimum problem (9.5). If f and g are of class $\mathcal{C}^1$ and if $\nabla g(\widehat{x})\neq\mathbf{0}$, then there exists a scalar $\widehat{\lambda}\in\mathbb{R}$, called Lagrange multiplier, such that the pair $\left(\widehat{x},\widehat{\lambda}\right)\in\mathbb{R}^{n+1}$ is a stationary point of the Lagrangian function.*

Proof Let $\widehat{x}$ be a solution of the problem of optimum (9.5). By Lemma 501 there exists $\widehat{\lambda} \in \mathbb{R}$ such that

$$\nabla f(\widehat{x}) - \widehat{\lambda} \nabla g(\widehat{x}) = \mathbf{0}.$$

By (9.14), this condition is equivalent to

$$\nabla_x L\left(\widehat{x}, \widehat{\lambda}\right) = \mathbf{0}.$$

On the other hand, by (9.14) we have $\nabla_\lambda L(x, \lambda) = b - g(x)$. Hence, we will also have $\nabla_\lambda L\left(\widehat{x}, \widehat{\lambda}\right) = 0$ since $b - g(\widehat{x}) = 0$. It follows that $\left(\widehat{x}, \widehat{\lambda}\right)$ is a stationary point of L. ■

Thanks to Theorem 503, the constrained optimum problem (9.5) is reduced to the search of stationary points of a suitable function in several variables, the Lagrangian function. It is a more complicated function than the original function f because we have the new variable λ. But, thanks to the Lagrangian the solutions of the optimum problem can be found by solving a standard first order condition, of the type seen for the unconstrained problems.

Naturally, we are talking of conditions that are only necessary, and so there is no guarantee that the stationary points are actually solutions of the problem. But, to have a first order condition that selects possible candidates to be solutions of the constrained problem is a substantial advance.

Before going on, we have to make two important observations. First, notice that in general the pair $\left(\widehat{x}, \widehat{\lambda}\right)$ is not a point of maximum of the Lagrangian, even when $\widehat{x}$ is actually a solution of the optimum problem. The pair $\left(\widehat{x}, \widehat{\lambda}\right)$ is a stationary point for the Lagrangian, but nothing more. Therefore, to say that the the solution of the constrained optimum problem is reduced to the search of the maximum points of the Lagrangian is a big mistake, as Chapter 10 will further clarify.

The second observation to make is that the problem (9.5) has a specular version

$$\begin{array}{l} \min\limits_{x \in A} f(x) \\ \text{sub } g(x) = b \end{array} \tag{9.16}$$

where instead of constrained maxima we look for constrained minima. Condition (9.6) is necessary also for this version of the problem (9.5) and therefore the stationary points of the Lagrangian could be constrained minima instead of maxima, but they could also be neither maxima nor minima. This is the usual ambiguity of the first order conditions, already met in the case of unconstrained optimization, which reflects the fact that first order conditions are only necessary.

9.3.1 The Method of Elimination

Thanks to Theorems 503 and 317, we can establish a procedure for solving the optimum problem (9.5) when both functions f and g are of class $\mathcal{C}^1$. We will call *method of elimination* this procedure, and in order to describe it we set

$$\begin{aligned} D_0 &= \{x \in A : \nabla g(x) = \mathbf{0}\}, \\ D_1 &= \{x \in A : \nabla g(x) \neq \mathbf{0}\}. \end{aligned}$$

The elements of D_0 and D_1 are called *singular points* and *regular points* of g, respectively. The method of elimination is based on four steps:

(i) We check if Theorem 317 can be applied, i.e., if f is upper semicontinuous and coercive on $C = \{x \in A : g(x) = b\}$.

(ii) We find the set $D_0 \cap C$ of the singular points that satisfy the constraint.

(iii) We find the set S of the points $x \in D_1$ for which there exists λ such that the pair (x, λ) is a stationary point of the Lagrangian.[2]

(iv) We construct the set $\{f(x) : x \in S \cup (D_0 \cap C)\}$. If $\widehat{x} \in S \cup (D_0 \cap C)$ is such that $f(\widehat{x}) \geq f(x)$ for each $x \in S \cup (D_0 \cap C)$, then such $\widehat{x}$ is solution of the optimum problem (9.5). In other words, we construct the set that consists of the stationary points and of the singular points; the points of this set in which f has maximum value are the solutions of the optimum problem.

To understand why the method of elimination works, note that by Theorem 503 the set S consists of the points of D_1 that are candidates to be local solutions of the optimum problem (9.5).

On the other hand, if f is upper semicontinuous and coercive on C, by Theorem 317 there exists at least one solution of the optimum problem. Since $D_0 \cup D_1 = A$, we have $C = (D_0 \cap C) \cup (D_1 \cap C)$ and hence such a solution must belong to either $D_0 \cap C$ or $D_1 \cap C$. But, if it belongs to $D_1 \cap C$, for what we just observed it must be in S. It follows that the solutions belong to $S \cup (D_0 \cap C)$, and hence the points $\widehat{x} \in S \cup (D_0 \cap C)$ such that $f(\widehat{x}) \geq f(x)$ for each $x \in S \cup (D_0 \cap C)$ are the solutions of the problem of optimum (9.5).

The method of elimination is an elegant combination of global existence results, like Theorem 317, and of local results, like Theorem 503.

[2]Notice that these points x automatically satisfy the constraint and hence we always have $S \subseteq D_1 \cap C$; it is therefore not necessary to verify if for a point $x \in S$ we also have $x \in C$.

Before illustrating this method with some examples, note that when in the first step of the method of elimination we use Weierstrass Theorem, that is, a result stronger than Theorem 317, as a "by-product" of this method we also find the points of global minimum, that is, the points $x \in C$ that solve problem (9.16). In fact, it is easy to see that they are the points $x^* \in S \cup D_0$ such that $f(x^*) \leq f(x)$ for each $x \in S \cup D_0$. Naturally, this is no longer true if we use Theorem 317, or one of its variants seen in Section 6.6, because these results guarantee only the existence of maximum points, and they do not say anything on the existence of possible minimum points.

Example 504 Consider the optimum problem:

$$\begin{aligned} &\max_{x \in \mathbb{R}^2} \left(2x_1^2 - 5x_2^2\right) \\ &\text{sub } x_1^2 + x_2^2 = 1 \end{aligned} \tag{9.17}$$

It is of the form (9.5), where $f : \mathbb{R}^2 \to \mathbb{R}$ and $g : \mathbb{R}^2 \to \mathbb{R}$ are given by $f(x_1, x_2) = 2x_1^2 - 5x_2^2$ and $g(x_1, x_2) = x_1^2 + x_2^2$, while $b = 1$. The set C is compact, and this completes the first step of the method of elimination (here the Weierstrass Theorem holds). We have

$$\nabla g(x) = (2x_1, 2x_2)$$

and so $x = \mathbf{0}$ is the only singular point, i.e., $D_0 = \{(0,0)\}$. Since it does not satisfy the constraint, we have $D_0 \cap C = \emptyset$. The Lagrangian $L : \mathbb{R}^3 \to \mathbb{R}$ is given by

$$L(x_1, x_2, \lambda) = 2x_1^2 - 5x_2^2 + \lambda\left(1 - x_1^2 - x_2^2\right), \qquad \forall (x_1, x_2, \lambda) \in \mathbb{R}^3,$$

and to find the set S of its stationary points it is necessary to solve the first order condition given by

$$\begin{cases} \frac{\partial L}{\partial x_1} = 0 \\ \frac{\partial L}{\partial x_2} = 0 \\ \frac{\partial L}{\partial \lambda} = 0 \end{cases}$$

It is therefore necessary to solve the following (nonlinear) system of 3 equations

$$\begin{cases} 2x_1(2 - \lambda) = 0 \\ -2x_2(5 + \lambda) = 0 \\ 1 - x_1^2 - x_2^2 = 0 \end{cases}$$

in the 3 unknowns x_1, x_2 and λ. Clearly $x_1 = 0$ and $x_2 = 0$ do not satisfy the third equation. While $x_1 = 0$ and $\lambda = -5$ implies $x_2 = \pm 1$. So $\lambda = 2$ and $x_2 = 0$ lead to $x_1 = \pm 1$. In conclusion, we have

$$S = \{(0,1), (0,-1), (1,0), (-1,0)\},$$

which coincides with $S \cup D_0$ since $D_0 = \emptyset$. We therefore have:

$$f(0,1) = f(0,-1) = 5,\ f(1,0) = f(-1,0) = 2,$$

and this implies that the points $(0,1)$ and $(0,-1)$ are solutions of the optimum problem (9.17). Instead, $(1,0)$ and $(-1,0)$ are constrained (global) minima. ▲

Example 505 Consider the optimum problem:

$$\begin{array}{l} \max\limits_{x \in \mathbb{R}^n} e^{-\|x\|^2} \\ \text{sub} \sum\limits_{i=1}^{n} x_i = 1 \end{array} \tag{9.18}$$

It is of the form (9.5), where $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ are given by $f(x) = e^{-\|x\|^2}$ and $g(x) = \sum_{i=1}^{n} x_i$, while $b = 1$. The set C is not compact, and hence the Weierstrass Theorem cannot be applied. But, f is coercive. In fact,

$$(f \geq t) = \begin{cases} \mathbb{R}^n & \text{if } t \leq 0 \\ \left\{x \in \mathbb{R}^n : \|x\| \leq \sqrt{-\lg t}\right\} & \text{if } t \in (0,1] \\ \emptyset & \text{if } t > 1 \end{cases}$$

and hence $(f \geq t)$ is compact and nonempty for each $t \in (0,1]$. By Theorem 317, f has then at least one maximum point on C, and this completes the first step of the method of elimination. We have

$$\nabla g(x) = (1, ..., 1)$$

and therefore there exist no singular points, that is, $D_0 = \emptyset$. The Lagrangian $L : \mathbb{R}^n \to \mathbb{R}$ is given by

$$L(x_1, x_2, \lambda) = e^{-\|x\|^2} + \lambda \left(1 - \sum_{i=1}^{n} x_i\right), \qquad \forall (x, \lambda) \in \mathbb{R}^{n+1},$$

and to find the set S of its stationary points it is necessary to solve the first order condition given by the following (nonlinear)system of $n+1$ equations

$$\begin{cases} \frac{\partial L}{\partial x_i} = -2x_i e^{-\|x\|^2} - \lambda = 0 & \forall i = 1, ..., n \\ \frac{\partial L}{\partial \lambda} = 1 - \sum_{i=1}^{n} x_i = 0 \end{cases}$$

The first n equations imply

$$x_i = -\frac{\lambda}{2} e^{\|x\|^2}$$

and by substituting these values in the last equation we find

$$1 - n\frac{\lambda}{2}e^{\|x\|^2} = 0,$$

that is,

$$\lambda = -\frac{2}{n}e^{-\|x\|^2}.$$

Substituting this value of λ in each of the first n equations we find $x_i = 1/n$, and so S is the singleton given by:

$$S = \left\{\left(\frac{1}{n}, ..., \frac{1}{n}\right)\right\},$$

which coincides with $S \cup (D_0 \cap C)$ since $D_0 = \emptyset$. As $S \cup (D_0 \cap C)$ is a singleton, the method of elimination is completed and we conclude that the vector $x = (1/n, ..., 1/n)$ is the only solution of the optimum problem (9.17). ▲

The next example shows the importance of the set $D_0 \cap C$.

Example 506 Consider the optimum problem:

$$\begin{array}{l} \max\limits_{x \in \mathbb{R}^2} e^{-x_1} \\ \text{sub } x_1^3 - x_2^2 = 0 \end{array} \tag{9.19}$$

It is of the form (9.5), where $f : \mathbb{R}^2 \to \mathbb{R}$ and $g : \mathbb{R}^2 \to \mathbb{R}$ are given by $f(x) = e^{-x_1}$ and $g(x) = x_1^3 - x_2^2$, while $b = 0$. The set C is closed but not compact, and hence the Weierstrass Theorem cannot be applied. The function f is continuous on $\mathbb{R}^2$ and we have:

$$(f \geq t) = \begin{cases} \mathbb{R}^2 & \text{if } t \leq 0 \\ (-\infty, -\lg t] \times \mathbb{R} & \text{if } t \in (0, 1] \\ \emptyset & \text{if } t > 1 \end{cases}$$

The function f is therefore not coercive. On the other hand, the constraint $x_1^3 = x_2^2$ is such that x_1 can satisfy such constraint only if $x_1 \geq 0$. Hence, $C \subseteq \mathbb{R}_+ \times \mathbb{R}$ and

$$(f \geq t) \cap C \subseteq ((-\infty, -\lg t] \times \mathbb{R}) \cap (\mathbb{R}_+ \times \mathbb{R}) = [0, -\lg t] \times \mathbb{R}, \qquad \forall t \in (0, 1].$$

If $x_1 \in [0, -\lg t]$, the constraint is such that $x_2^2 \in \left[0, (-\lg t)^3\right]$, that is,

$$x_2^2 \in \left[-\sqrt{(-\lg t)^3}, \sqrt{(-\lg t)^3}\right].$$

It follows that:

$$(f \geq t) \cap C \subseteq [0, -\lg t] \times \left[-\sqrt{(-\lg t)^3}, \sqrt{(-\lg t)^3}\right], \qquad \forall t \in (0, 1],$$

and hence $(f \geq t) \cap C$ is compact since it is a closed subset of a compact. The function f is therefore coercive and continuous on C, and by Theorem 317 it has at least one point of maximum on C. This completes the first step of the method of elimination.

As to the second step, we have

$$\nabla g(x) = \left(3x_1^2, 2x_2\right)$$

and hence $D_0 \cap C = \{(0,0)\}$.

The Lagrangian $L : \mathbb{R}^2 \to \mathbb{R}$ is given by

$$L(x_1, x_2, \lambda) = e^{-x_1} + \lambda\left(x_2^2 - x_1^3\right), \qquad \forall\, (x, \lambda) \in \mathbb{R}^3,$$

and to find the set S of its stationary points it is necessary to solve the first order condition given by the following (nonlinear) system of 3 equations

$$\begin{cases} \frac{\partial L}{\partial x_1} = -e^{-x_1} - 3\lambda x_1^2 = 0 \\ \frac{\partial L}{\partial x_2} = 2\lambda x_2 = 0 \\ \frac{\partial L}{\partial \lambda} = x_2^2 - x_1^3 = 0 \end{cases}$$

We observe that in no solution we can have $\lambda = 0$. In fact, if it were $\lambda = 0$ the first equation would become $e^{-x_1} = 0$, which does not have solution. Suppose therefore that $\lambda \neq 0$. The second equation implies $x_2 = 0$, and from the third one it follows that $x_1 = 0$. The first equation becomes $-1 = 0$, and this contradiction shows that the system does not have solution. Hence, $S = \emptyset$.

In conclusion, $S \cup (D_0 \cap C) = \{(0,0)\}$ and the method of elimination allows us to conclude that $(0,0)$ is the only solution of the optimum problem (9.19). ▲

Though the method of elimination is a very general procedure of resolution of constrained optimum problems, it can happen that there are solutions of the problem of optimum that are not found through this procedure. In fact, the first step of the method of elimination is based on results of existence à la Weierstrass that are sufficient, but not necessary, conditions for the existence of optimum points. In this regard, recall Example 320, where the hypotheses of these existence results did not hold and, nevertheless, there were points of global maximum.

It can therefore happen that the first step of the method of elimination does not hold, though there exist solutions of the optimum problem. The only thing that we can say when the first step does not hold is that the solutions, if they exist, belong to the set $S \cup D_0$. But, it can well happen that none of such points is solution of the system.

Next examples shows these features of the method of elimination.

Example 507 Consider the optimum problem:

$$\begin{array}{l} \max\limits_{x\in\mathbb{R}^n} x_1^3 - x_2^3 \\ \text{sub } x_1 - x_2 = 0 \end{array} \tag{9.20}$$

It is of the form (9.5), where $f : \mathbb{R}^2 \to \mathbb{R}$ and $g : \mathbb{R}^2 \to \mathbb{R}$ are given by $f(x) = x_1^3 - x_2^3$ and $g(x) = x_1 - x_2$, while $b = 0$. The set C is not compact and the function f is not coercive on C. In fact,

$$(f \geq t) \cap C = \begin{cases} \emptyset & \text{if } t > 0 \\ C & \text{if } t \leq 0 \end{cases}$$

It follows that the hypotheses of Theorem 317 are not satisfied, and therefore the first step of the method of elimination does not hold. As a result, this method cannot be applied to the optimum problem (9.20).

On the other hand, it is immediate to verify that $f(x) = 0$ for each $x \in C$, and hence every point of C is solution of the problem (9.20). Therefore, the method of elimination has not been able here to solve a quite simple constrained optimum problem. ▲

Example 508 We slightly change the previous example, and we consider the optimum problem:

$$\begin{array}{l} \max\limits_{x\in\mathbb{R}^n} x_1^3 - x_2^3 \\ \text{sub } x_1 + x_2 = 0 \end{array} \tag{9.21}$$

The only difference with respect to the previous example is that $g(x) = x_1 + x_2$. Also here the set C is not compact and the function f is not coercive on C. In fact,

$$(f \geq t) \cap C = \left[\sqrt[3]{\frac{t}{2}}, +\infty \right), \qquad \forall t \in \mathbb{R}.$$

The hypotheses of Theorem 317 do not hold and therefore the method of elimination cannot be applied even to the optimum problem (9.21).

Unlike the previous example, this problem does not have solution. In fact, consider the sequence $\{x_n\}_n$ given by $x_n = (n, -n)$ for each $n \geq 1$. We have $\{x_n\}_n \subseteq C$ and $f(x_n) = 2n^3$, so that $\lim_n f(x_n) = +\infty$, what proves that the problem (9.21) does not have solutions.

If we study the Lagrangian, we get $S \cup (D_0 \cap C) = \{(0,0)\}$, and therefore the point $(0,0)$ is the only candidate to be solution of the problem. But, $f(n,-n) > f(0,0)$ and hence also by this way we conclude that problem (9.21) does not have solutions.[3] ▲

[3]This second way that uses the Lagrangian is in this case convolute with respect to the direct study of the sequence $x_n = (n, -n)$. But, there can be cases in which it is useful first to reduce to $S \cup (D_0 \cap C)$, and then to establish that none of these candidates is actually solution of the problem.

Finally, observe that we did not mention any second order conditions to use in solving constrained optimum problems, while in Chapter 5 we talked in detail of these conditions in the problems without constraints. This omission is not by chance and is due to the change of perspective, from local to global, that we did in this chapter with respect to Chapter 5. The second order conditions are in fact of little interest in the search of global maximum points, and their marginal contribution to such search is usually more than compensated by the heaviness of the computations that they require. For this reason we do not talk about them and we refer the interested reader to Section 5.4 of Montrucchio (1998).

9.4 Several Constraints

Consider now the general optimum problem (9.2), in which there can be several equality constraints. Lemma 501 and Theorem 503 generalize in a natural way to the case of several constraints. We write problem (9.2) as

$$\begin{array}{l} \max\limits_{x \in A} f(x) \\ \text{sub } g(x) = b \end{array} \tag{9.22}$$

where $g = (g_1, ..., g_m) : A \subseteq \mathbb{R}^n \to \mathbb{R}^m$ and $b = (b_1, ..., b_m) \in \mathbb{R}^m$. The Jacobian matrix $Dg(x)$ is given by

$$Dg(x) = \begin{bmatrix} \nabla g_1(x) \\ \nabla g_2(x) \\ \cdots \\ \nabla g_m(x) \end{bmatrix}$$

and the points x where $Dg(\widehat{x})$ has maximum rank (i.e., rank m) are called *regular*, while the points where this is not true are called *singular*.

The Jacobian $Dg(\widehat{x})$ has maximum rank if and only if the gradients $\nabla g_1(\widehat{x})$, $\nabla g_2(\widehat{x})$, ..., $\nabla g_m(\widehat{x})$ are linearly independent. In this case the vector subspace of $\mathbb{R}^n$ generated by these gradients has dimension m and the condition of maximum rank is such that $m \leq n$, i.e., the condition of regularity can hold only if the number of constraints does not exceed the dimensionality of the space.

When $m = 1$, we have $Dg(x) = \nabla g(x)$ and the condition of maximum rank is equivalent to require $\nabla g(x) \neq \mathbf{0}$. In this case we thus go back to the notions of regular points and singular points previously seen.[4]

On the other hand, when $m > 1$ the Jacobian $Dg(\widehat{x})$ has maximum rank if and only if the gradients $\nabla g_1(\widehat{x})$, $\nabla g_2(\widehat{x})$, ..., $\nabla g_m(\widehat{x})$ are linearly independent. In this

[4]Note that in a vector space V, a singleton $\{v\}$ is linearly independent when $\alpha v = \mathbf{0}$ implies $\alpha = 0$, which is equivalent to require $v \neq \mathbf{0}$.

case the vector subspace of $\mathbb{R}^n$ generated by these gradients has dimension m and the condition of maximum rank is such that $m \leq n$, i.e., the condition of regularity can hold only if the number of constraints does not exceed the dimensionality of the space.

Naturally, when $m = n$ the Jacobian has maximum rank if and only if it is a non singular matrix, i.e., if and only if $\det Dg(\widehat{x}) \neq 0$.

The next result extends Lemma 501 to the case of several constraints, and shows that the condition of regularity $\nabla g(\widehat{x}) \neq \mathbf{0}$ of Lemma 501 is generalized by requiring that the Jacobian $Dg(\widehat{x})$ has maximum rank. In other words, also here $\widehat{x}$ must be a regular point.

Lemma 509 *Let $\widehat{x}$ be solution of the optimum problem (9.22). If the functions $f, g_1, ..., g_m$ are of class $\mathcal{C}^1$ and if $Dg(\widehat{x})$ has rank m, then there exists a vector $\widehat{\lambda} \in \mathbb{R}^m$ such that*

$$\nabla f(\widehat{x}) = \sum_{i=1}^{m} \widehat{\lambda}_i \nabla g_i(\widehat{x}). \tag{9.23}$$

Proof Let $\|\cdot\|$ be the Euclidean norm. Since A is an open, there exists $\widehat{\varepsilon} > 0$ sufficiently small such that $\overline{B}_{\widehat{\varepsilon}}(\widehat{x}) = \{x \in A : \|x - \widehat{x}\| \leq \widehat{\varepsilon}\} \subseteq A$. Given $\varepsilon \in (0, \widehat{\varepsilon}]$, set $S_\varepsilon(\widehat{x}) = \{x \in A : \|x - \widehat{x}\| = \varepsilon\}$. In view of Exercise 13.0.56, the set $S_\varepsilon(\widehat{x})$ is compact.

We first prove a property that we will use in the following.

Fact 1. For each $\varepsilon \in (0, \widehat{\varepsilon}]$, there exists $N > 0$ such that

$$f(x) - f(\widehat{x}) - \|x - \widehat{x}\|^2 - N \sum_{i=1}^{m} (g_i(x) - g_i(\widehat{x}))^2 < 0, \tag{9.24}$$

for each $x \in S_\varepsilon(\widehat{x})$.

Proof of Fact 1. We proceed by contradiction, and we assume therefore that there exists $\varepsilon \in (0, \widehat{\varepsilon}]$ for which we do not have any $N > 0$ such that (9.24) holds. Take an increasing sequence $\{N_n\}_n$ with $N_n \uparrow +\infty$, and for each of these N_n take $x_n \in S_\varepsilon(\widehat{x})$ for which (9.24) does not hold, i.e., x_n such that:

$$f(x_n) - f(\widehat{x}) - \|x_n - \widehat{x}\|^2 - N_n \sum_{i=1}^{m} (g_i(x_n) - g_i(\widehat{x}))^2 \geq 0.$$

Hence, for each $n \geq 1$ we have:

$$\frac{f(x_n) - f(\widehat{x}) - \|x_n - \widehat{x}\|^2}{N_n} \geq \sum_{i=1}^{m} (g_i(x_n) - g_i(\widehat{x}))^2. \tag{9.25}$$

Since the sequence $\{x_n\}_n$ just constructed is contained in the compact set $S_\varepsilon(\widehat{x})$, by Theorem 275 there exists a subsequence $\{x_{n_k}\}_k$ convergent in $S_\varepsilon(\widehat{x})$, i.e., there exists $x^* \in S_\varepsilon(\widehat{x})$ such that $x_{n_k} \to x^*$. Expression (9.25) implies that, for each $k \geq 1$, we have:

$$\frac{f(x_{n_k}) - f(\widehat{x}) - \|x_{n_k} - \widehat{x}\|^2}{N_{n_k}} \geq \sum_{i=1}^{m} (g_i(x_{n_k}) - g_i(\widehat{x}))^2. \tag{9.26}$$

Since f is continuous, we have $\lim_k f(x_{n_k}) = f(x^*)$. Moreover, $\lim_k \|x_{n_k} - \widehat{x}\| = \|x^* - \widehat{x}\|$. Since $\lim_k N_{n_k} = +\infty$, we have

$$\lim_k \frac{f(x_{n_k}) - f(\widehat{x}) - \|x_{n_k} - \widehat{x}\|^2}{N_{n_k}} = 0,$$

and hence (9.26) implies, thanks to the continuity of the functions g_i,

$$\sum_{i=1}^{m} (g_i(x^*) - g_i(\widehat{x}))^2 = \lim_k \sum_{i=1}^{m} (g_i(x_{n_k}) - g_i(\widehat{x}))^2 = 0.$$

It follows that $(g_i(x^*) - g_i(\widehat{x}))^2 = 0$ for each $i = 1, ..., m$, from which $g_i(x^*) = g_i(\widehat{x}) = b_i$ for each $i = 1, ..., m$. Therefore, x^* satisfies the equality constraints, and we therefore have $f(\widehat{x}) \geq f(x^*)$ since $\widehat{x}$ is a solution of the optimum problem (9.22).

On the other hand, since $x_{n_k} \in S_\varepsilon(\widehat{x})$ for each $k \geq 1$, (9.26) implies

$$f(x_{n_k}) - f(\widehat{x}) \geq \|x_{n_k} - \widehat{x}\|^2 + N_{n_k} \sum_{i=1}^{m} (g_i(x_{n_k}) - g_i(\widehat{x}))^2 \geq \varepsilon^2, \qquad \forall k \geq 1,$$

and therefore $f(x_{n_k}) \geq f(\widehat{x}) + \varepsilon^2$ for each $k \geq 1$. By the continuity of f, this leads to

$$f(x^*) = \lim_k f(x_{n_k}) \geq f(\widehat{x}) + \varepsilon^2 > f(\widehat{x}),$$

what contradicts $f(\widehat{x}) \geq f(x^*)$. This contradiction proves Fact 1. △

Using Fact 1, we prove now a second property that we will need. Here we set $S_1^{m+1}(\mathbf{0}) = \{x \in \mathbb{R}^{m+1} : \|x\| = 1\}$.

Fact 2. For each $\varepsilon \in (0, \widehat{\varepsilon}]$, there exist $x^\varepsilon \in B_\varepsilon(\widehat{x})$ and a vector $(\lambda_0^\varepsilon, \lambda_1^\varepsilon, ..., \lambda_m^\varepsilon) \in S_1^{m+1}(\mathbf{0})$ such that

$$\lambda_0^\varepsilon \left(\frac{\partial f}{\partial x_j}(x^\varepsilon) - 2\left(x_j^\varepsilon - \widehat{x}_j\right) \right) - \sum_{i=1}^{m} \lambda_i^\varepsilon \frac{\partial g_i}{\partial x_j}(x^\varepsilon) = 0, \qquad \forall j = 1, ..., n. \tag{9.27}$$

Proof of Fact 2. Given $\varepsilon \in (0, \widehat{\varepsilon}]$, let $N_\varepsilon > 0$ be the positive constant whose existence is guaranteed by Fact 1. Define the function $h_\varepsilon : A \subseteq \mathbb{R}^n \to \mathbb{R}$ by:

$$h_\varepsilon(x) = f(x) - f(\widehat{x}) - \|x - \widehat{x}\|^2 - N_\varepsilon \sum_{i=1}^{m} (g_i(x) - g_i(\widehat{x}))^2, \qquad \forall x \in A.$$

We have $h_\varepsilon(\widehat{x}) = 0$ and, given how we chose N_ε,

$$h_\varepsilon(x) < 0, \qquad \forall x \in S_\varepsilon(\widehat{x}). \tag{9.28}$$

The function h_ε is continuous on the compact $\overline{B}_\varepsilon(\widehat{x}) = \{x \in A : \|x - \widehat{x}\| \leq \varepsilon\}$, and by the Weierstrass Theorem there exists $x^\varepsilon \in \overline{B}_\varepsilon(\widehat{x})$ such that $h_\varepsilon(x^\varepsilon) \geq h_\varepsilon(x)$ for each $x \in \overline{B}_\varepsilon(\widehat{x})$. In particular, $h_\varepsilon(x^\varepsilon) \geq h_\varepsilon(\widehat{x}) = 0$, and so (9.28) implies that $\|x^\varepsilon\| < \varepsilon$, i.e., $x^\varepsilon \in B_\varepsilon(\widehat{x})$. The point x^ε is therefore a maximum point on the open $B_\varepsilon(\widehat{x})$ and by Theorem 194 we have $\nabla h_\varepsilon(x^\varepsilon) = \mathbf{0}$. Hence,

$$\frac{\partial f}{\partial x_j}(x^\varepsilon) - 2\left(x_j^\varepsilon - \widehat{x}_j\right) - 2N_\varepsilon \sum_{i=1}^{m} g_i(x^\varepsilon)\frac{\partial g_i}{\partial x_j}(x^\varepsilon) = 0, \qquad \forall j = 1, ..., n. \tag{9.29}$$

Set:

$$\begin{aligned} c_\varepsilon &= 1 + \sum_{i=1}^{m} \left(2N_\varepsilon g_i(x^\varepsilon)\right)^2, \\ \lambda_0^\varepsilon &= \frac{1}{c_\varepsilon}, \\ \lambda_i^\varepsilon &= \frac{2N_\varepsilon g_i(x^\varepsilon)}{c_\varepsilon}, \qquad \forall i = 1, ..., m, \end{aligned}$$

so that (9.27) is obtained dividing (9.29) by c_ε. Note that $\sqrt{\sum_{i=0}^{m}(\lambda_i^\varepsilon)^2} = 1$, i.e., $(\lambda_0^\varepsilon, \lambda_1^\varepsilon, ..., \lambda_m^\varepsilon)$ belongs to a $S_1^{m+1}(\mathbf{0})$. △

Using Fact 2, we can now complete the proof. Take a decreasing sequence $\{\varepsilon_n\}_n \subseteq (0, \widehat{\varepsilon}]$ with $\varepsilon_n \downarrow 0$,[5] and consider the relative sequence $\{(\lambda_0^n, \lambda_1^n, ..., \lambda_m^n)\}_n \subseteq \mathbb{R}^{m+1}$, whose existence is guaranteed by Fact 2.

Since the sequence $\{(\lambda_0^n, \lambda_1^n, ..., \lambda_m^n)\}_n$ is contained in the compact set $S_1^{m+1}(\mathbf{0})$, by Theorem 275 there exists a subsequence $\{(\lambda_0^{n_k}, \lambda_1^{n_k}, ..., \lambda_m^{n_k})\}_k$ convergent in $S_1^{m+1}(\mathbf{0})$, i.e., there exists $(\lambda_0, \lambda_1, ..., \lambda_m) \in S_1^{m+1}(\mathbf{0})$ such that

$$(\lambda_0^{n_k}, \lambda_1^{n_k}, ..., \lambda_m^{n_k}) \to (\lambda_0, \lambda_1, ..., \lambda_m).$$

Thanks to Fact 2, for each ε_{n_k} there exists $x^{n_k} \in B_{\varepsilon_{n_k}}(\widehat{x})$ for which (9.27) holds, that is,

$$\lambda_0^{n_k}\left(\frac{\partial f}{\partial x_j}(x^{n_k}) - 2\left(x^{n_k} - \widehat{x}_j\right)\right) - \sum_{i=1}^{m} \lambda_i^{n_k}\frac{\partial g_i}{\partial x_j}(x^{n_k}) = 0, \qquad \forall j = 1, ..., n.$$

[5] Naturally, this use of the index n must not be confused with the use of n as index of dimensionality of the space $\mathbb{R}^n$ on which the functions f and g_i are defined.

Consider the sequence $\{x^{n_k}\}_k$ so constructed. From $x^{n_k} \in B_{\varepsilon_{n_k}}(\widehat{x})$ it follows that

$$\|x^{n_k} - \widehat{x}\| < \varepsilon_{n_k} \to 0,$$

and hence, for each $j = 1, ..., n$,

$$\begin{aligned} & \lambda_0 \frac{\partial f}{\partial x_j}(\widehat{x}) - \sum_{i=1}^{m} \lambda_i \frac{\partial g_i}{\partial x_j}(\widehat{x}) \qquad (9.30) \\ = & \lim_k \left(\lambda_0^{n_k} \left(\frac{\partial f}{\partial x_j}\left(x^k\right) - 2\left(x^{n_k} - \widehat{x}_j\right) \right) - \sum_{i=1}^{m} \lambda_i^{n_k} \frac{\partial g_i}{\partial x_j}(x^{n_k}) \right) = 0. \end{aligned}$$

On the other hand, $\lambda_0 \neq 0$. In fact, if it were $\lambda_0 = 0$, then from (9.30) it follows that

$$\sum_{i=1}^{m} \lambda_i \frac{\partial g_i}{\partial x_j}(\widehat{x}) = 0, \qquad \forall j = 1, ..., n.$$

Since the gradients $\nabla g_1(\widehat{x})$, $\nabla g_2(\widehat{x})$, ..., $\nabla g_m(\widehat{x})$ are linearly independent, this implies $\lambda_i = 0$ for each $i = 1, ..., m$, which contradicts $(\lambda_0, \lambda_1, ..., \lambda_m) \in S_1^{m+1}(\mathbf{0})$.

In conclusion, if we set $\widehat{\lambda}_i = \lambda_i/\lambda_0$ for each $i = 1, ..., m$, (9.30) implies (9.23). ■

The Lagrangian is now the function $L : A \times \mathbb{R} \subseteq \mathbb{R}^{n+m} \to \mathbb{R}$ defined as:

$$L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i (b_i - g_i(x)) = f(x) + \lambda \cdot (b - g(x)), \qquad (9.31)$$

for each $(x, \lambda) \in A \times \mathbb{R}^m$. Theorem 503 generalizes as follows (we omit the proof, which is completely analogous to that of Theorem 503).

Theorem 510 *Let $\widehat{x}$ be solution of the optimum problem (9.22). If the functions $f, g_1, ..., g_m$ are of class $\mathcal{C}^1$ and if $Dg(\widehat{x})$ has rank m, then there exists a vector $\widehat{\lambda} \in \mathbb{R}^m$ such that the pair $\left(\widehat{x}, \widehat{\lambda}\right) \in \mathbb{R}^{n+m}$ is a stationary point of the Lagrangian function.*

The components $\widehat{\lambda}_i$ of the vector $\widehat{\lambda} \in \mathbb{R}^m$ are called *Lagrange multipliers*. It is worthwhile remarking that the vector $\widehat{\lambda}$ of the multipliers associated with a solution to (9.23) is necessarily unique: by assumption the vector $\{\nabla g_i(\widehat{x})\}_{i=1}^m$ are linearly independent and thus the representation $\nabla f(\widehat{x}) = \sum_{i=1}^m \widehat{\lambda}_i \nabla g_i(\widehat{x})$ is unique.

Theorem 510 allows to extend to the case of several constraints the method of elimination that we introduced in the previous section for the search of the solutions of the optimum problem with only one constraint. Next examples illustrate the procedure in this more general form. It is however important to observe that the determination of the singular points, satisfying the constraints, when $m > 1$, requires some job and

the study of the rank of the Jacobian matrix is rather complicated and not the best way to follow.

Observe that we must find points $\overline{x}$ such that $g_i(\overline{x}) = b_i$ and the gradients $\nabla g_i(\overline{x})$ be linearly dependent. Therefore, we must verify if the system

$$\begin{cases} \sum_{i=1}^{m} \lambda_i \nabla g_i(x) = 0 \\ g_i(\overline{x}) = b_i \quad i = 1, ..., m \end{cases}$$

admits solutions (λ_i) which are not all null. Clearly this system can be write as

$$\begin{cases} \sum_{i=1}^{m} \lambda_i \partial g_i \partial x_1 = 0 \\ ..... \\ \sum_{i=1}^{m} \lambda_i \partial g_i \partial x_n = 0 \\ g_i(\overline{x}) = b_i \quad i = 1, ..., m \end{cases} . \tag{9.32}$$

Example 511 Consider the optimum problem:

$$\begin{aligned} &\max_{x \in \mathbb{R}^3} (7x_1 - 3x_3) \\ &\text{sub } x_1^2 + x_2^2 = 1 \text{ and } x_1 + x_2 - x_3 = 1 \end{aligned} \tag{9.33}$$

This problem is of the form (9.22), where $f : \mathbb{R}^3 \to \mathbb{R}$ and $g = (g_1, g_2) : \mathbb{R}^3 \to \mathbb{R}^2$ are given by $f(x_1, x_2) = 3x_1 - 4x_3$, $g_1(x_1, x_2, x_3) = x_1^2 + x_2^2$ and $g_2(x_1, x_2, x_3) = x_1 + x_2 - x_3$, while $b = (1, 1) \in \mathbb{R}^2$. Since C is obviously closed, to prove that it is compact it is sufficient to prove that it is also bounded. For the x_1 and x_2 such that $x_1^2 + x_2^2 = 1$ we have $x_1, x_2 \in [-1, 1]$ and hence for the x_3 such that $x_3 = x_1 + x_2 - 1$ we have $x_3 \in [-3, 1]$. It follows that $C \subseteq [0, 1] \times [0, 1] \times [-3, 1]$ and the set C is therefore bounded, and hence compact. This completes the first step of the method of elimination (because the Weierstrass Theorem holds). Let us find the singular point. System (9.32) becomes

$$\begin{cases} 2\lambda x_1 + \mu = 0 \\ 2\lambda x_2 + \mu = 0 \\ -\mu = 0 \\ x_1^2 + x_2^2 = 1 \\ x_1 + x_2 - x_3 = 1 \end{cases} .$$

Since $\mu = 0$, λ must be different from 0. This implies $x_1 = x_2 = 0$ that contradicts the forth equation. Hence no singular points satisfies the constraints.

The Lagrangian $L : \mathbb{R}^3 \to \mathbb{R}$ is given by

$$L(x_1, x_2, \lambda) = 7x_1 - 3x_3 + \lambda_1 \left(1 - x_1^2 - x_2^2\right) + \lambda_2 (1 - x_1 - x_2 + x_3)$$

for each $(x_1, x_2, x_3, \lambda_1, \lambda_2) \in \mathbb{R}^5$ and to find the set S of its stationary points it is necessary to solve the first order condition given by the following (nonlinear) system

of 5 equations

$$\begin{cases} \frac{\partial L}{\partial x_1} = 7 - 2\lambda_1 x_1 - \lambda_2 = 0 \\ \frac{\partial L}{\partial x_2} = -2\lambda_1 x_2 - \lambda_2 = 0 \\ \frac{\partial L}{\partial x_3} = -3 + \lambda_2 = 0 \\ \frac{\partial L}{\partial \lambda_1} = 1 - x_1^2 - x_2^2 = 0 \\ \frac{\partial L}{\partial \lambda_2} = 1 - x_1 - x_2 + x_3 = 0 \end{cases}$$

in the 5 unknowns x_1, x_2, x_3, λ_1 and λ_2. The third equation implies $\lambda_2 = 3$ and therefore the system reduces to:

$$\begin{cases} -2\lambda_1 x_1 + 4 = 0 \\ -\lambda_1 x_2 - 3 = 0 \\ 1 - x_1^2 - x_2^2 = 0 \\ 1 - x_1 - x_2 + x_3 = 0 \end{cases}$$

The first equation implies $\lambda_1 \neq 0$. Hence, from the first two equations it follows that:

$$\frac{2}{\lambda_1} = x_1 \qquad \textit{and} \qquad -\frac{3}{2\lambda_1} = x_2$$

Substituting in the third equation we have:

$$\lambda_1 = \pm\frac{5}{2}.$$

If $\lambda_1 = 5/2$, we obtain $x_1 = 4/5$, $x_2 = -3/5$, $x_3 = -4/5$. If $\lambda_1 = -5/2$, we have $x_1 = -4/5$, $x_2 = 3/5$, $x_3 = -7/5$. We have therefore found two stationary points

$$\left\{ \left(\frac{4}{5}, -\frac{3}{5}, -\frac{4}{5}, \frac{5}{2}, 3 \right), \left(-\frac{4}{5}, \frac{3}{5}, -\frac{7}{5}, -\frac{5}{2}, 3 \right) \right\},$$

and hence

$$S = \left\{ \left(\frac{4}{5}, -\frac{3}{5}, -\frac{4}{5} \right), \left(-\frac{4}{5}, \frac{3}{5}, -\frac{7}{5} \right) \right\}.$$

We have:

$$\begin{aligned} f\left(\frac{4}{5}, -\frac{3}{5}, -\frac{4}{5} \right) &= 8 \\ f\left(-\frac{4}{5}, \frac{3}{5}, -\frac{7}{5} \right) &= -\frac{7}{5} \end{aligned}$$

and this implies that

$$\left(\frac{4}{5}, -\frac{3}{5}, -\frac{4}{5} \right)$$

is the solution of the optimum problem (9.33), while

$$\left(-\frac{4}{5}, \frac{3}{5}, -\frac{7}{5} \right)$$

is a constrained (global) minimum. ▲

Example 512 Consider the optimum problem:

$$\begin{array}{l} \max\limits_{x\in\mathbb{R}^3} -x_1 \\ \text{sub } -x_1^2+x_2^3=0 \quad \text{and} \quad x_3^2+x_2^2-2x_2=0 \end{array} \tag{9.34}$$

This problem is of the form (9.22), where $f:\mathbb{R}^3\to\mathbb{R}$ and $g=(g_1,g_2):\mathbb{R}^3\to\mathbb{R}^2$ are given by $f(x_1,x_2)=-x_1$, $g_1(x_1,x_2,x_3)=-x_1^2+x_2^3$, $g_2(x_1,x_2,x_3)=x_3^2+x_2^2-2x_2$, while $b=(0,-1)\in\mathbb{R}^2$. Since C is obviously closed, to prove that is compact it is sufficient to prove that it is also bounded. The second constraint can be written as $x_3^2+(x_2-1)^2=1$, and so the x_2 and x_3 that satisfy it are such that $x_2\in[0,2]$ and $x_3\in[-1,1]$. At this point the constraint $x_1^2=x_2^3$ implies that $x_1^2\in[0,8]$, from which $x_1\in\left[-\sqrt{8},\sqrt{8}\right]$. In conclusion,

$$C\subseteq\left[-\sqrt{8},\sqrt{8}\right]\times[0,2]\times[-1,1],$$

which shows that C is bounded, and so compact. As in the previous example, also here the hypotheses of the Weierstrass Theorem are satisfied and this completes the first step of the method of elimination.

To check the existence of singular points we solve the system

$$\begin{cases} -2\lambda x_1=0 \\ 3\lambda x_2^2+\mu(2x_2-2)=0 \\ 2\mu x_3=0 \\ -x_1^2+x_2^3=0 \\ x_3^2+x_2^2-2x_2=0 \end{cases}$$

In view of the first and the third equation, we discuss three cases:

$$\begin{array}{rrcl} i) & \lambda & = & 0, x_3=0 \\ ii) & \mu & = & 0, x_1=0 \\ iii) & x_3 & = & 0, x_1=0 \end{array}$$

Case (i) implies that $\mu\neq 0$. We have $x_2=1$ which contradicts the last equation.
Case (ii) implies $\lambda\neq 0$ and we get the solution $x_1=x_2=x_3=0$.
Case (iii) clearly leads to the same solution $x_1=x_2=x_3=0$. W can conclude that:

$$C\cap D_0=\{(0,0,0)\}. \tag{9.35}$$

We now move to the set S of the stationary points of the Lagrangian, which is here given by

$$L(x_1,x_2,\lambda)=-x_1+\lambda_1\left(x_1^2-x_2^3\right)+\lambda_2\left(-x_3^2-x_2^2+2x_2\right)$$

for each $(x_1, x_2, x_3, \lambda_1, \lambda_2) \in \mathbb{R}^5$. To find S it is necessary to solve the first order condition given by the following (nonlinear) system of 5 equations

$$\begin{cases} \frac{\partial L}{\partial x_1} = -1 + 2\lambda_1 x_1 = 0 \\ \frac{\partial L}{\partial x_2} = -3\lambda_1 x_2^2 - 2\lambda_2 (x_2 - 1) = 0 \\ \frac{\partial L}{\partial x_3} = -2\lambda_2 x_3 = 0 \\ \frac{\partial L}{\partial \lambda_1} = x_1^2 - x_2^3 = 0 \\ \frac{\partial L}{\partial \lambda_2} = -x_3^2 + x_2^2 - 2x_2 = 0 \end{cases}$$

in the 5 unknowns x_1, x_2, x_3, λ_1 and λ_2. The first equation implies $\lambda_1 \neq 0$ and $x_1 \neq 0$. From the fourth equation it follows that $x_2 \neq 0$ and, hence, from the second equation it follows that $\lambda_2 \neq 0$.

Since $\lambda_2 \neq 0$, from the third equation it follows $x_3 = 0$, and hence the fifth equation implies that $x_2 = 0$ or $x_2 = 2$. Since $x_2 = 0$ contradicts what established above, we consider $x_2 = 2$. The fourth equation implies $x_1 = \pm\sqrt{8}$, and hence the first equation implies

$$\lambda_1 = \pm\frac{1}{4\sqrt{2}},$$

so that from the second equation it follows that

$$\lambda_2 = \mp\frac{3}{2\sqrt{2}}.$$

In conclusion, the stationary points are

$$\left\{ \left(\sqrt{8}, 2, 0, \frac{1}{4\sqrt{2}}, -\frac{3}{2\sqrt{2}} \right), \left(\sqrt{8}, 2, 0, -\frac{1}{4\sqrt{2}}, \frac{3}{2\sqrt{2}} \right) \right\}$$

and therefore $S = \left\{ \left(\sqrt{8}, 2, 0 \right), \left(\sqrt{8}, 2, 0 \right) \right\}$, from which

$$S \cup D_0 = \left\{ \left(\sqrt{8}, 2, 0 \right), \left(\sqrt{8}, 2, 0 \right), (0, 0, 0) \right\}.$$

We have:

$$f\left(\sqrt{8}, 2, 0 \right) = -\sqrt{8}, \quad f\left(-\sqrt{8}, 2, 0 \right) = \sqrt{8}, \quad f(0, 0, 0) = 0.$$

and this implies that $\left(-\sqrt{8}, 2, 0 \right)$ is the solution of the optimum problem (9.34). Instead, $\left(\sqrt{8}, 2, 0 \right)$ is the constrained (global) minimum. ▲

Example 513 Consider the optimum problem:

$$\begin{aligned} &\max_{x \in \mathbb{R}^3} -\left(x_1^2 + x_2^2 + x_3^2 \right) \\ &\text{sub } x_1^2 - x_2 = 1 \quad \text{and} \quad x_1 + x_3 = 0 \end{aligned} \tag{9.36}$$

This problem is of the form (9.22), where $f : \mathbb{R}^3 \to \mathbb{R}$ and $g = (g_1, g_2) : \mathbb{R}^3 \to \mathbb{R}^2$ are given by $f(x_1, x_2, x_3) = -(x_1^2 + x_2^2 + x_3^2)$, $g_1(x_1, x_2, x_3) = x_1^2 - x_2$ and $g_2(x_1, x_2, x_3) = x_1 + x_3$, while $b = (1, 1) \in \mathbb{R}^2$. The set C is closed, but is not bounded (and hence it is not compact). To see this, consider the sequence $\{x_n\}_n$ given by $x_n = \left(\sqrt{1+n}, n, 1-n\right)$. This sequence belongs to C, but $\|x_n\| \to +\infty$ and hence there does not exist any neighborhood in $\mathbb{R}^3$ that can contain it.

On the other hand, thanks to Proposition 323 f is coercive, and hence Theorem 317 holds. The first step of the method of elimination is therefore satisfied.

Let us in this case study directly the rank of the Jacobian:

$$Dg(x) = \begin{bmatrix} -2x_1 & -1 & 0 \\ 1 & 0 & 1 \end{bmatrix}.$$

It is easy to see how for no value of x_1 the two row vectors, i.e., the two gradients $\nabla g_1(x)$ and $\nabla g_2(x)$, are linearly dependent (at a "mechanical" level it is easily verified that there does not exist any value of x_1 for which the matrix $Dg(x)$ does not have full rank). Therefore, there do not exist singular points, i.e., we have $D_0 = \emptyset$.

We now determine the set S of the stationary points of the Lagrangian, here given by

$$L(x_1, x_2, \lambda) = -\left(x_1^2 + x_2^2 + x_3^2\right) + \lambda_1\left(1 - x_1^2 + x_2\right) + \lambda_2(1 - x_1 - x_3)$$

for each $(x_1, x_2, x_3, \lambda_1, \lambda_2) \in \mathbb{R}^5$. To find S it is necessary to solve the following (nonlinear) system of 5 equations

$$\begin{cases} \frac{\partial L}{\partial x_1} = -2x_1 - 2\lambda_1 x_1 - \lambda_2 = 0 \\ \frac{\partial L}{\partial x_2} = -2x_2 + \lambda_1 = 0 \\ \frac{\partial L}{\partial x_3} = -2x_3 - \lambda_2 = 0 \\ \frac{\partial L}{\partial \lambda_1} = 1 - x_1^2 + x_2 = 0 \\ \frac{\partial L}{\partial \lambda_2} = 1 - x_1 - x_3 = 0 \end{cases}$$

We have $\lambda_1 = 2x_2$ and $\lambda_2 = -2x_3$, which substituted in the first equation leads to the following nonlinear system of 3 equations:

$$\begin{cases} x_1 + 2x_1x_2 - x_3 = 0 \\ 1 - x_1^2 + x_2 = 0 \\ 1 - x_1 - x_3 = 0 \end{cases}$$

From the last two equations we have $x_2 = x_1^2 - 1$ and $x_3 = 1 - x_1$, which substituted in the first one lead to $2x_1^3 - 1 = 0$, from which

$$x_1 = \frac{1}{\sqrt[3]{2}}.$$

In turn this implies:

$$x_2 = \frac{1}{\sqrt[3]{4}} - 1 \quad \text{and} \quad x_3 = 1 - \frac{1}{\sqrt[3]{2}}.$$

Being $D_0 = \emptyset$, we therefore have:

$$S \cup D_0 = S = \left\{ \left(\frac{1}{\sqrt[3]{2}}, \frac{1}{\sqrt[3]{4}} - 1, 1 - \frac{1}{\sqrt[3]{2}} \right) \right\}.$$

We can therefore conclude that the point

$$\left(\frac{1}{\sqrt[3]{2}}, \frac{1}{\sqrt[3]{4}} - 1, 1 - \frac{1}{\sqrt[3]{2}} \right)$$

is the solution of the optimum problem (9.36). Notice that since we found just one critical point of the Lagrangian, we can infer that there is no global (or local) minimum.
▲

Chapter 10

Differential Non Linear Programming

10.1 Introduction

Let us go back to the problem of the consumer seen at the beginning of the previous chapter, in which we considered a consumer with utility function $u : A \subseteq \mathbb{R}^n \to \mathbb{R}$ and with an income $b \in \mathbb{R}$. Given the vector $p \in \mathbb{R}^n_+$ of prices of the goods, we wrote his budget constraint as

$$C(p, b) = \{x \in A : p \cdot x = b\},$$

and his optimum problem as:

$$\max_{x \in C(p,b)} u(x).$$

In this formulation we assumed that the consumer has to spend all his income, from which the sign of equality in the budget constraint, and we did not impose other constraints on the bundle x except that of satisfying the budget constraint. As to the income, the hypothesis that all the income is spent can be too strong. Think for example of intertemporal problems, where it can be crucial to leave to the consumer the possibility of saving in some periods, something that is impossible if we require that the budget constraint is satisfied with equality at each period. It becomes therefore natural to ask what happens to the optimum problem of the consumer if we weaken the constraint to $p \cdot x \leq b$, that is, if the constraint is given by an inequality and not anymore by an equality.

As to the bundles of goods $x \in A$, in many cases it can have no sense to talk of negative quantities. Think for example of the purchase of physical goods, maybe fruit or vegetables to the market, in which the quantity purchased of goods has to be non-negative. This suggests to impose the constraint $x \in \mathbb{R}^n_+$ in the optimum problem.

Keeping in mind these observations, the consumer problem becomes:

$$\begin{aligned} &\max_{x \in A} u(x) \\ &\text{sub} \quad p \cdot x \leq b \text{ and } x \in \mathbb{R}^n_+ \end{aligned} \tag{10.1}$$

with constraints now given by inequalities. If we write the budget constraint as

$$C(p,b) = \left\{ x \in A : x \in \mathbb{R}^n_+ \text{ and } p \cdot x \leq b \right\}, \tag{10.2}$$

the problem of optimum becomes:

$$\max_{x \in C(p,b)} u(x), \tag{10.3}$$

which has a form similar to the problem (9.1), though the set $C(p,b)$ has now a different definition.

The general form of an optimum problem, in which there can be both equality and inequality constraints, is given by

$$\begin{aligned} &\max_{x \in A} f(x) \\ &\text{sub} \quad g_i(x) = b_i, \qquad \forall i \in I, \\ &h_j(x) \leq c_j, \qquad \forall j \in J, \end{aligned} \tag{10.4}$$

where I and J are finite sets of indices (possibly empty), $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ is the objective function, the functions $g_i : A \subseteq \mathbb{R}^n \to \mathbb{R}$ and the associated scalars $b_i \in \mathbb{R}$ characterize $|I|$ equality constraints, while the functions $h_j : A \subseteq \mathbb{R}^n \to \mathbb{R}$ with the associated scalars $c_j \in \mathbb{R}$ induce $|J|$ inequality constraints.

The set

$$C = \{ x \in A : g_i(x) = b_i \text{ and } h_j(x) \leq c_j, \quad \forall i \in I, \forall j \in J \} \tag{10.5}$$

identified by the constraints is called *admissible region* for the optimum problem. The optimum problem (10.4) can be equivalently formulated as:

$$\max_{x \in C} f(x).$$

A point $\widehat{x} \in B$ is called (global) *solution* of the optimum problem (10.4) if $f(\widehat{x}) \geq f(x)$ for each $x \in C$, while $\widehat{x} \in B$ is called *local solution* of such problem if there exists a neighborhood $B_{x_0}(\varepsilon)$ of $\widehat{x}$ such that $f(\widehat{x}) \geq f(x)$ for each $x \in B_{x_0}(\varepsilon) \cap C$.

The formulation (10.4) is extremely versatile. First observe that it encompasses as special cases the optimum problems seen up to now. In fact:

(i) If $I = J = \emptyset$, we go back to the unconstrained optimum problem of Chapter 5, without constraints.

(ii) If $I \neq \emptyset$ and $J = \emptyset$, we go back to the optimum problem with only equality constraints of the previous chapter.

Moreover, observe that:

(iii) A constraint of the form $h(x) \geq c$ can be included in the formulation (10.4) by considering $-h(x) \leq -c$. In particular, a constraint of the form $x \geq 0$ can be included in the formulation (10.4) by considering $-x \leq 0$.

(iv) A constrained minimization problem for f can be written in the formulation (10.4) by considering $-f$.

(v) In the light of (iii) it should be clear that also the choice of the sign $\leq$ in expressing the inequality constraints is simply a convention.

Observations (i)-(iv) show the scope and flexibility of formulation (10.4). Before illustrating all this with some examples, we give a minimum of discipline to this formulation.

Definition 514 *The problem (10.4) is said to be* well posed *(or* superconsistent*) if for each $j \in J$ there exists $x \in C$ such that $h_j(x) < c$.*

To understand this definition observe that an equality constraint $g(x) = b$ can be written in form of inequality constraint as $g(x) \leq b$ and $-g(x) \leq -b$. This makes uncertain the distinction between equality constraints and inequality constraints in (10.4). To avoid this, and so to have a clear distinction between the two types of constraints, in what follows we will consider always problems (10.4) that are well posed, so that it is not possible to express possible equality constraints in the form of inequality constraints. In fact, there cannot exist any $x \in C$ for which we can have both $g(x) \leq b$ and $-g(x) < -b$ (naturally, if $J = \emptyset$, Definition 514 is automatically satisfied and there is nothing to worry about).[1]

Example 515 The optimum problem:

$$\begin{aligned} &\max_{x \in \mathbb{R}^3} \left(x_1^2 + x_2^2 + x_3^3\right) \\ &\text{sub} \quad x_1 + x_2 - x_3 = 1 \quad \text{and} \quad x_1^2 + x_2^2 \leq 1 \end{aligned}$$

[1] For simplicity, in the statements of the results concerning problem (10.4) we will assume implicitely that the problem is well posed, even without explicitely mention it.

is of the form (10.4) with $|I| = |J| = 1$, $f(x) = x_1^2 + x_2^2 + x_3^3$, $g(x) = x_1 + x_2 - x_3$, $h(x) = x_1^2 + x_2^2$ and $b = c = 1$.[2] ▲

Example 516 The optimum problem:

$$\begin{array}{l} \max\limits_{x \in \mathbb{R}^3} -x_1 \\ \text{sub } -x_1^2 + x_2^2 = 0 \quad \text{and} \quad x_3^2 + x_2^2 - 2x_2 = 0 \end{array}$$

is of the form (10.4) with $I = \{1, 2\}$, $J = \emptyset$, $f(x) = -x_1$, $g_1(x) = -x_1^2 + x_2^2$, $g_2(x) = x_3^2 + x_2^2 - 2x_2$ and $b_1 = b_2 = 0$. ▲

Example 517 The optimum problem:

$$\begin{array}{l} \max\limits_{x \in \mathbb{R}^3} e^{x_1 + x_2 + x_3} \\ \text{sub } x_1 + x_2 + x_3 = 1,\ x_1^2 + x_2^2 + x_3^2 = \dfrac{1}{2},\ x_1 \geq 0 \quad \text{and} \quad x_2 \geq \dfrac{1}{10} \end{array}$$

is of the form (10.4) with $I = J = \{1, 2\}$, $f(x) = e^{x_1 + x_2 + x_3}$, $g_1(x) = x_1 + x_2 + x_3$, $g_2(x) = x_1^2 + x_2^2 + x_3^2$, $h_1(x) = -x_1$, $h_2(x) = -x_2$, $b_1 = 1$, $b_2 = 2^{-1}$, $c_1 = 0$ and $c_2 = -10^{-1}$. ▲

Example 518 The optimum problem:

$$\begin{array}{l} \max\limits_{x \in \mathbb{R}^3} x_1^3 - x_2^3 \\ \text{sub } x_1 + x_2 \leq 1 \quad \text{and} \quad -x_1 + x_2 \leq 1 \end{array}$$

is of the form (10.4) with $I = \emptyset$, $J = \{1, 2\}$, $f(x) = x_1^3 - x_2^3$, $h_1(x) = x_1 + x_2$, $h_2(x) = -x_2 + x_1$ and $c_1 = c_2 = 1$. ▲

Example 519 The minimum problem:

$$\begin{array}{l} \min\limits_{x \in \mathbb{R}^3} x_1 + x_2 + x_3 \\ \text{sub } x_1 + x_2 = 1 \quad \text{and} \quad x_2^2 + x_3^2 \leq \dfrac{1}{2} \end{array}$$

can be written in the form (10.4) as

$$\begin{array}{l} \max\limits_{x \in \mathbb{R}^3} -(x_1 + x_2 + x_3) \\ \text{sub } x_1 + x_2 = 1 \quad \text{and} \quad x_2^2 + x_3^2 \leq \dfrac{1}{2} \end{array}$$

▲

[2] To be pedantic, here we should have set $I = J = \{1\}$, $g_1(x) = x_1 + x_2 - x_3$, $h_1(x) = x_1^2 + x_2^2$ and $b_1 = c_1 = 1$. But, in this case in which we have only one equality constraint and only one inequality constraint, pedices make the notation heavy without utility.

10.1.1 An Alternative Formulation

An optimum problem with inequality constraints is often written as

$$\begin{array}{rcl} & \max\limits_{x\in A} f(x) & \\ \text{sub} \quad g_1(x) & \leq & b_1, g_2(x) \leq b_2, ..., g_m(x) \leq b_m \end{array} \tag{10.6}$$

where $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ is our objective function, while the functions $g_i : A \subseteq \mathbb{R}^n \to \mathbb{R}$ and the scalars $b_i \in \mathbb{R}$ induce m inequality constraints.

In this formulation also possible equality constraints are included, by resorting to the usual trick of writing the equality constraint $g(x) = b$ as two inequality constraints $g(x) \leq b$ and $-g(x) \leq -b$. Note, however, that this formulation requires the presence of at least one constraint (it is the case $m = 1$) and hence it is less general than (10.4). Moreover, the indirect way in which (10.6) encompasses the equality constraints makes sometimes less transparent the formulation of the results and this is a further reason for which we chose formulation (10.4) with the equality constraints fully specified.

10.2 Resolution of the Problem

In this section we extend to the optimum problem (10.4) the solution methods studied in the previous chapter for the problem (9.2), which we saw to be a special case of (10.4).

The first thing to do is to find the general version of Lemma 509 that also holds for problem (10.4). To this end set, for a given point $x \in A$,

$$A(x) = I \cup \{j \in J : h_j(x) = c_j\}.$$

In other words, $A(x)$ is the set of the so called *binding constraints* at x, that is, of the constraints that hold as equalities at the given point x. The constraints that are not binding are called *non binding*.

For example, in the problem

$$\begin{array}{l} \max\limits_{x\in\mathbb{R}^3} f(x) \\ \text{sub} \quad x_1 + x_2 - x_3 = 1 \ \text{ and } x_1^2 + x_2^2 \leq 1 \end{array}$$

the first constraint is binding at all the points of C, while the second constraint is for example binding at the point $\left(1/\sqrt{2}, 1/\sqrt{2}, \sqrt{2}-1\right)$ and is not binding at the point $(1/2, 1/2, 0)$.

Definition 520 *The problem (10.4) has regular constraints at a point $x \in \mathbb{R}^n$ if the gradients $\nabla g_i(x)$ and the gradients $\nabla h_j(x)$ with $j \in A(x)$ are linearly independent.*

In other words, the constraints are regular at a point x if the gradients of the functions that induce constraints binding at such point are linearly independent. This condition is the generalization to the problem (10.4) of the condition of linear independence on which Lemma 509 was based, and in fact it implies that x is a regular point for the function $g : A \subseteq \mathbb{R}^{|I|} \to \mathbb{R}$.

In particular, if we form the matrix whose rows consist of the gradients of the functions that induce binding constraints at the point considered, the regularity condition of the constraints is equivalent to require that such matrix has maximum rank.

Finally, observe that in view of Corollary 50-(ii) the regularity condition of the constraints can be satisfied at a point x only if $|A(x)| \leq n$, that is, only if the number of the binding constraints at x does not exceed the dimension of the space on which the optimum problem is defined.

We can now give the generalization of Lemma 509 for problem (10.4).

Lemma 521 *Let $\widehat{x}$ be solution of the optimum problem (10.4). If the functions $f, \{g_i\}_{i\in I}$ and $\{h_j\}_{j\in J}$ are of class $\mathcal{C}^1$ and if the constraints are regular in $\widehat{x}$, then there exist a vector $\widehat{\lambda} \in \mathbb{R}^{|I|}$ and a vector $\widehat{\mu} \in \mathbb{R}_+^{|J|}$ such that*

$$\nabla f(\widehat{x}) = \sum_{i\in I} \widehat{\lambda}_i \nabla g_i(\widehat{x}) + \sum_{j\in J} \widehat{\mu}_j \nabla h_j(\widehat{x}), \tag{10.7}$$

$$\widehat{\mu} \cdot (c - h(\widehat{x})) = 0. \tag{10.8}$$

Note how the vector $\widehat{\mu}$ associated to the inequality constraints has positive sign, while there is no restriction on the sign of the vector $\widehat{\lambda}$ associated to the equality constraints.

Lemma 521 generalizes Theorem 194 and Lemma 509. In fact, if $I = J = \emptyset$ (optimization without constraints), (10.7) reduces to the condition $\nabla f(\widehat{x}) = \mathbf{0}$ of Theorem 194, while if $I \neq \emptyset$ and $J = \emptyset$ (optimization with only equality constraints), (10.7) reduces to the condition $\nabla f(\widehat{x}) = \sum_{i\in I} \widehat{\lambda}_i \nabla g_i(\widehat{x})$ of Lemma 509.

With respect to Theorem 194 and to Lemma 509, the novelty of Lemma 521 is the condition (10.8). Since $\widehat{\mu}$ has positive sign, this condition is equivalent to require

$$\widehat{\mu}_j (c - h_j(\widehat{x})) = 0, \qquad \forall j \in J,$$

and often (10.8) is written in this form. To understand the role of this condition it is useful the following characterization.

Lemma 522 *Condition (10.8) holds if and only if $\widehat{\mu}_j = 0$ for each j such that $h_j(\widehat{x}) < c_j$, that is, for each $j \notin A(\widehat{x})$.*

Proof Assume (10.8). Since for each $j \in J$ we have $h_j(\widehat{x}) \leq c_j$, from the positive sign of $\widehat{\mu}$ it follows that (10.8) implies $c_j - h_j(\widehat{x}) = 0$ for each $j \in J$, and therefore $\widehat{\mu}_j = 0$ for each j such that $h_j(\widehat{x}) < c_j$.

On the other hand, if this last property holds we have

$$\widehat{\mu}_j \cdot (c_j - h_j(\widehat{x})) = 0, \qquad \forall j \in J. \tag{10.9}$$

because, being $h_j(\widehat{x}) \leq c_j$ for each $j \in J$, we have $h_j(\widehat{x}) < c_j$ or $h_j(\widehat{x}) = c_j$. Expression (10.9) immediately implies (10.8). ■

In other words, (10.8) is equivalent to require the nullity of each $\widehat{\mu}_j$ associated to a non-binding constraint. Hence, we can have $\widehat{\mu}_j > 0$ only if the constraint j is binding in correspondence of the solution $\widehat{x}$.

For example, if $\widehat{x}$ is such that $g_j(\widehat{x}) < c_j$ for each $j \in J$, i.e., if in correspondence of $\widehat{x}$ all the inequality constraints are non-binding, then we have $\widehat{\mu}_j = 0$ for each $j \in J$ and the vector $\widehat{\mu}$ does not play any role in the determination of $\widehat{x}$. Naturally, this reflects the fact that for the solution $\widehat{x}$ the inequality constraints do not play any role.

The next example shows that conditions (10.7) and (10.8) are necessary but not sufficient, similarly to what we saw for Theorem 194 and Lemma 509.

Example 523 Consider the optimum problem:

$$\begin{array}{l} \max\limits_{x \in \mathbb{R}} \dfrac{x_1^3 + x_2^3}{2} \\ \text{sub } x_1 - x_2 \leq 0 \end{array} \tag{10.10}$$

It is a simple modification of Example 502, and it is of the form (10.4) with $f : \mathbb{R}^2 \to \mathbb{R}$ and $h : \mathbb{R}^2 \to \mathbb{R}$ given by $f(x) = 2^{-1}(x_1^3 + x_2^3)$ and $h(x) = x_1 - x_2$, while $c = 0$. We have:

$$\nabla f(0,0) = (0,0) \quad \text{and} \quad \nabla g(0,0) = (1,-1)$$

and hence $\lambda = 0$ is such that

$$\begin{array}{rcl} \nabla f(0,0) & = & \mu \nabla g(0,0), \\ \mu \cdot (0-0) & = & 0. \end{array}$$

The point $(0,0)$ satisfies with $\mu = 0$ the conditions (10.7) and (10.8), but $(0,0)$ is not solution of the optimum problem (10.10), as (9.8) shows. ■

We now move to the proof of Lemma 521. It is possible to give a partial proof of this lemma by reducing problem (10.4) to a problem with only equality constraints,

and then by exploiting the results seen in the previous chapter. For simplicity, we give this argument for the special case

$$\begin{aligned} &\max_{x\in A} f(x) \\ &\text{sub} \quad g(x) = b \quad \text{and} \quad h(x) \leq c \end{aligned} \tag{10.11}$$

where $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ is the objective function, and $g : A \subseteq \mathbb{R}^n \to \mathbb{R}$ and $h : A \subseteq \mathbb{R}^n \to \mathbb{R}$ induce one equality and one inequality constraint.

Define $H : A \times \mathbb{R} \subseteq \mathbb{R}^{n+1} \to \mathbb{R}$ as $H(x,z) = h(x) + z^2$ for each $x \in A$ and each $z \in \mathbb{R}$. Given $x \in A$, we have $h(x) \leq c$ if and only if there exists $z \in \mathbb{R}$ such that $h(x) + z^2 = c$, i.e., if and only if $H(x,z) = c$.[3]

Define $F : A \times \mathbb{R} \subseteq \mathbb{R}^{n+1} \to \mathbb{R}$ and $G : A \times \mathbb{R} \subseteq \mathbb{R}^{n+1} \to \mathbb{R}$ by $F(x,z) = f(x)$ and $G(x,z) = g(x)$ for each $x \in A$ and each $z \in \mathbb{R}$. The dependence of F and G on z is only fictitious, but it allows to formulate the following classical optimum problem:

$$\begin{aligned} &\max_{(x,z)\in A\times\mathbb{R}} F(x,z) \\ &\text{sub} \quad G(x,z) = b \quad \text{and} \quad H(x,z) = c \end{aligned} \tag{10.12}$$

Problems (10.11) and (10.12) are equivalent: $\widehat{x}$ is solution of problem (10.11) if and only if there exists $\widehat{z} \in \mathbb{R}$ such that $(\widehat{x}, \widehat{z})$ is solution of problem (10.12).

We have therefore reduced problem (10.11) to a problem with only equality constraints. By Lemma 509, $(\widehat{x}, \widehat{z})$ is solution of such problem only if there exists a vector $\left(\widehat{\lambda}, \widehat{\mu}\right) \in \mathbb{R}^2$ such that:

$$\nabla F(\widehat{x}, \widehat{z}) = \widehat{\lambda} \nabla G(\widehat{x}, \widehat{z}) + \widehat{\mu} \nabla H(\widehat{x}, \widehat{z})$$

that is, only if

$$\begin{aligned} \frac{\partial F}{\partial x_i}(\widehat{x}, \widehat{z}) &= \widehat{\lambda} \frac{\partial G}{\partial x_i}(\widehat{x}, \widehat{z}) + \widehat{\mu} \frac{\partial H}{\partial x_i}(\widehat{x}, \widehat{z}), \qquad \forall i = 1, ..., n \\ \frac{\partial F}{\partial z}(\widehat{x}, \widehat{z}) &= \widehat{\lambda} \frac{\partial G}{\partial z}(\widehat{x}, \widehat{z}) + \widehat{\mu} \frac{\partial H}{\partial z}(\widehat{x}, \widehat{z}), \end{aligned}$$

which is equivalent to:

$$\begin{aligned} \nabla f(\widehat{x}) &= \widehat{\lambda} \nabla g(\widehat{x}) + \widehat{\mu} \nabla h(\widehat{x}) \\ 2\widehat{\mu} z &= 0 \end{aligned}$$

On the other hand, we have $2\widehat{\mu} z = 0$ if and only if $\widehat{\mu} z^2 = 0$. Recalling the equivalence between problems (10.11) and (10.12), we can therefore conclude that $\widehat{x}$ is solution of

[3] Note that the positivity of the square z^2 preserves the inequality $g(x) \leq b$. The auxiliary variable z is often called *slack variable*.

problem (10.11) only if there exists a vector $(\lambda, \mu) \in \mathbb{R}^2$ such that:

$$\nabla f(\widehat{x}) = \widehat{\lambda} \nabla g(\widehat{x}) + \widehat{\mu} \nabla h(\widehat{x})$$
$$\widehat{\mu}(c - h(x)) = 0$$

We therefore have conditions (10.7) and (10.8) of Lemma 521. What we have not been able to prove is the positivity of the multiplier μ, and for this reason the proof just seen is incomplete.[4]

To have a complete proof of Lemma 521 it is necessary to generalize the argument used to prove Lemma 509; for the sake of completeness, the proof will be given in full detail, at the cost of some repetitions relative to the proof of Lemma 509.

Proof of Lemma 521. Let $\|\cdot\|$ be the Euclidean norm. We have $h_j(\widehat{x}) < c_j$ for each $j \notin A(\widehat{x})$. Since A is an open, there exists $\widetilde{\varepsilon} > 0$ sufficiently small such that $\overline{B}_{\widetilde{\varepsilon}}(\widehat{x}) = \{x \in A : \|x - \widehat{x}\| \leq \widetilde{\varepsilon}\} \subseteq A$. Moreover, since each h_j is continuous, for each $j \notin A(\widehat{x})$ there exists ε_j sufficiently small such that $h_j(x) < c_j$ for each $x \in \overline{B}_{\varepsilon_j}(\widehat{x}) = \{x \in A : \|x - \widehat{x}\| \leq \varepsilon_j\}$. Let $\varepsilon' = \min_{j \notin A(\widehat{x})} \varepsilon_j$ and $\widehat{\varepsilon} = \min\{\widetilde{\varepsilon}, \varepsilon'\}$; in other words, $\widehat{\varepsilon}$ is the minimum between $\widetilde{\varepsilon}$ and the ε_j. In this way we have $\overline{B}_{\widehat{\varepsilon}}(\widehat{x}) = \{x \in A : \|x - \widehat{x}\| \leq \widehat{\varepsilon}\} \subseteq A$ and $h_j(x) < c_j$ for each $x \in \overline{B}_{\widehat{\varepsilon}}(\widehat{x})$ and each $j \notin A(\widehat{x})$.

Given $\varepsilon \in (0, \widehat{\varepsilon}]$, set $S_\varepsilon(\widehat{x}) = \{x \in A : \|x - \widehat{x}\| = \varepsilon\}$. In the light of Exercise 13.0.56, the set $S_\varepsilon(\widehat{x})$ is compact. Moreover, by what just seen $h_j(x) < c_j$ for each $x \in S_\varepsilon(\widehat{x})$ and each $j \notin A(\widehat{x})$, that is, in $S_\varepsilon(\widehat{x})$ all the non binding constraints are always satisfied.

For each $j \in J$, let $h_j^+ : A \subseteq \mathbb{R}^n \to \mathbb{R}$ be defined as $h_j^+(x) = \max\{h_j(x) - c_j, 0\}$ for each $x \in A$. Thanks to Exercise 13.0.57, we have $\left(h_j^+\right)^2 \in \mathcal{C}^1(A)$ and

$$\frac{\partial \left(h_j^+\right)^2(x)}{\partial x_p} = 2h_j^+(x) \frac{\partial h_j(x)}{\partial x_p}, \qquad \forall p = 1, ..., n \tag{10.13}$$

We first prove a property that we will use after.

Fact 1. For each $\varepsilon \in (0, \widehat{\varepsilon}]$, there exists $N > 0$ such that

$$f(x) - f(\widehat{x}) - \|x - \widehat{x}\|^2 \tag{10.14}$$
$$-N\left(\sum_{i \in I} (g_i(x) - g_i(\widehat{x}))^2 + \sum_{i \in J \cap A(\widehat{x})} \left(h_j^+(x) - h_j^+(\widehat{x})\right)^2\right) < 0,$$

for each $x \in S_\varepsilon(\widehat{x})$.

[4] Since it is, in any case, an incomplete argument, for simplicity we did not check the rank condition required by Lemma 509.

Proof of Fact 1. We proceed by contradiction, and we assume therefore that there exists $\varepsilon \in (0, \widehat{\varepsilon}]$ for which there is no $N > 0$ such that (10.14) holds. Take an increasing sequence $\{N_n\}_n$ with $N_n \uparrow +\infty$, and for each of these N_n take $x_n \in S_\varepsilon(\widehat{x})$ for which (10.14) does not hold, that is, x_n such that:

$$f(x_n) - f(\widehat{x}) - \|x_n - \widehat{x}\|^2 - N_n \left(\sum_{i \in I} (g_i(x_n) - g_i(\widehat{x}))^2 + \sum_{i \in J \cap A(\widehat{x})} \left(h_j^+(x_n) - h_j^+(\widehat{x})\right)^2 \right) \geq 0.$$

Hence, for each $n \geq 1$ we have:

$$\frac{f(x_n) - f(\widehat{x}) - \|x_n - \widehat{x}\|^2}{N_n} \geq \sum_{i \in I} (g_i(x_n) - g_i(\widehat{x}))^2 + \sum_{j \in J \cap A(\widehat{x})} \left(h_j^+(x_n) - h_j^+(\widehat{x})\right)^2 \tag{10.15}$$

Since the sequence $\{x_n\}_n$ just constructed is contained in the compact set $S_\varepsilon(\widehat{x})$, by Theorem 275 there exists a subsequence $\{x_{n_k}\}_k$ convergent in $S_\varepsilon(\widehat{x})$, i.e., there exists $x^* \in S_\varepsilon(\widehat{x})$ such that $x_{n_k} \to x^*$. Expression (10.15) implies that, for each $k \geq 1$, we have:

$$\frac{f(x_{n_k}) - f(\widehat{x}) - \|x_{n_k} - \widehat{x}\|^2}{N_{n_k}} \geq \sum_{i \in I} (g_i(x_{n_k}) - g_i(\widehat{x}))^2 + \sum_{j \in J \cap A(\widehat{x})} \left(h_j^+(x_{n_k}) - h_j^+(\widehat{x})\right)^2. \tag{10.16}$$

Since f is continuous, we have $\lim_k f(x_{n_k}) = f(x^*)$. Moreover, $\lim_k \|x_{n_k} - \widehat{x}\| = \|x^* - \widehat{x}\|$. Since $\lim_k N_{n_k} = +\infty$, we have

$$\lim_k \frac{f(x_{n_k}) - f(\widehat{x}) - \|x_{n_k} - \widehat{x}\|^2}{N_{n_k}} = 0,$$

and hence (10.16) implies, thanks to the continuity of the functions g_i and h_j^+,

$$\sum_{i \in I} (g_i(x^*) - g_i(\widehat{x}))^2 + \sum_{i \in J \cap A(\widehat{x})} \left(h_j^+(x^*) - h_j^+(\widehat{x})\right)^2 = \lim_k \left(\sum_{i \in I} (g_i(x_{n_k}) - g_i(\widehat{x}))^2 + \sum_{i \in J \cap A(\widehat{x})} \left(h_j^+(x_{n_k}) - h_j^+(\widehat{x})\right)^2 \right) = 0.$$

It follows that $(g_i(x^*) - g_i(\widehat{x}))^2 = \left(h_j^+(x^*) - h_j^+(\widehat{x})\right)^2 = 0$ for each $i \in I$ and for each $j \in J \cap A(\widehat{x})$, from which $g_i(x^*) = g_i(\widehat{x}) = b_i$ for each $i \in I$ and $h_j^+(x^*) = h_j^+(\widehat{x}) = c_j$ for each $j \in J \cap A(\widehat{x})$.

Since in $S_\varepsilon(\widehat{x})$ the non binding constraints are always satisfied, i.e., $h_j(x) < c_j$ for each $x \in S_\varepsilon(\widehat{x})$ and each $j \notin A(\widehat{x})$, we can conclude that x^* satisfies all the constraints. We therefore have $f(\widehat{x}) \geq f(x^*)$ given that $\widehat{x}$ is a solution of the optimum problem (9.22).

On the other hand, since $x_{n_k} \in S_\varepsilon(\widehat{x})$ for each $k \geq 1$, (10.16) implies

$$f(x_{n_k}) - f(\widehat{x}) \geq \|x_{n_k} - \widehat{x}\|^2 + N_{n_k}\left(\sum_{i\in I}\left(g_i(x_{n_k}) - g_i(\widehat{x})\right)^2 + \sum_{j\in J\cap A(\widehat{x})}\left(h_j^+(x_{n_k}) - h_j^+(\widehat{x})\right)^2\right) \geq \varepsilon^2,$$

for each $k \geq 1$, and hence $f(x_{n_k}) \geq f(\widehat{x}) + \varepsilon^2$ for each $k \geq 1$. Thanks to the continuity of f, this leads to

$$f(x^*) = \lim_k f(x_{n_k}) \geq f(\widehat{x}) + \varepsilon^2 > f(\widehat{x}),$$

which contradicts $f(\widehat{x}) \geq f(x^*)$. This contradiction proves Fact 1. △

Using Fact 1, we prove now a second property that we will need. Here we set $U_1(\mathbf{0}) = \left\{x \in \mathbb{R}^{|I|+|J|+1} : \|x\| = 1\right\}$.

Fact 2. For each $\varepsilon \in (0, \widehat{\varepsilon}]$, there exist $x^\varepsilon \in B_\varepsilon(\widehat{x})$ and a vector

$$\left(\lambda_0^\varepsilon, \lambda_1^\varepsilon, ..., \lambda_{|I|}^\varepsilon, \mu_1^\varepsilon, ..., \mu_{|J|}^\varepsilon\right) \in U_1(\mathbf{0}),$$

with $\mu_j^\varepsilon \geq 0$ for each $j \in J$, such that

$$\lambda_0^\varepsilon\left(\frac{\partial f}{\partial x_z}(x^\varepsilon) - 2\left(x_j^\varepsilon - \widehat{x}_j\right)\right) - \sum_{i\in I}\lambda_i^\varepsilon\frac{\partial g_i}{\partial x_z}(x^\varepsilon) - \sum_{j\in J\cap A(\widehat{x})}\mu_j^\varepsilon\frac{\partial h_j}{\partial x_z}(x^\varepsilon) = 0, \tag{10.17}$$

for each $z = 1, ..., n$.

Proof of Fact 2. Given $\varepsilon \in (0, \widehat{\varepsilon}]$, let $N_\varepsilon > 0$ be the positive constant whose existence is guaranteed by Fact 1. Define the function $\rho_\varepsilon : A \subseteq \mathbb{R}^n \to \mathbb{R}$ as:

$$\rho_\varepsilon(x) = f(x) - f(\widehat{x}) - \|x - \widehat{x}\|^2 - N_\varepsilon\left(\sum_{i\in I}\left(g_i(x) - g_i(\widehat{x})\right)^2 + \sum_{j\in J\cap A(\widehat{x})}\left(h_j^+(x) - h_j^+(\widehat{x})\right)^2\right)$$

for each $x \in A$. We have $\rho_\varepsilon(\widehat{x}) = 0$ and, given how N_ε has been chosen,

$$\rho_\varepsilon(x) > 0, \qquad \forall x \in S_\varepsilon(\widehat{x}). \tag{10.18}$$

The function ρ_ε is continuous on the compact set $\overline{B}_\varepsilon(\widehat{x}) = \{x \in A : \|x - \widehat{x}\| \leq \varepsilon\}$ and, by the Weierstrass Theorem, there exists $x^\varepsilon \in \overline{B}_\varepsilon(\widehat{x})$ such that $\rho_\varepsilon(x^\varepsilon) \geq \rho_\varepsilon(x)$ for

each $x \in \overline{B}_{\varepsilon}(\widehat{x})$. In particular, $\rho_{\varepsilon}(x^{\varepsilon}) \geq \rho_{\varepsilon}(\widehat{x}) = 0$, and hence (10.18) implies that $\|x^{\varepsilon}\| < \varepsilon$, that is, $x^{\varepsilon} \in B_{\varepsilon}(\widehat{x})$. Point x^{ε} is therefore a maximum on the open set $B_{\varepsilon}(\widehat{x})$ and by Theorem 194 we have $\nabla\rho_{\varepsilon}(x^{\varepsilon}) = \mathbf{0}$. Therefore, by (10.13), we have:

$$\frac{\partial f}{\partial x_z}(x^{\varepsilon}) - 2(x_z^{\varepsilon} - \widehat{x}_z) - 2N_{\varepsilon}\left(\sum_{i=1}^{m} g_i(x^{\varepsilon})\frac{\partial g_i}{\partial x_z}(x^{\varepsilon}) + \sum_{j\in J\cap A(\widehat{x})} h_j^{+}(x^{\varepsilon})\frac{\partial h_j}{\partial x_z}(x^{\varepsilon})\right) = 0, \tag{10.19}$$

for each $z = 1, ..., n$. Set:

$$\begin{aligned}
c_{\varepsilon} &= 1 + \sum_{i=1}^{m}\left(2N_{\varepsilon}g_i(x^{\varepsilon})\right)^2 + \sum_{j\in J\cap A(\widehat{x})}\left(2N_{\varepsilon}h_j^{+}(x^{\varepsilon})\right)^2,\\
\lambda_0^{\varepsilon} &= \frac{1}{c_{\varepsilon}},\\
\lambda_i^{\varepsilon} &= \frac{2N_{\varepsilon}g_i(x^{\varepsilon})}{c_{\varepsilon}}, \qquad \forall i \in I,\\
\mu_j^{\varepsilon} &= \frac{2N_{\varepsilon}h_j^{+}(x^{\varepsilon})}{c_{\varepsilon}}, \qquad \forall j \in J\cap A(\widehat{x}),\\
\mu_j^{\varepsilon} &= 0, \qquad \forall j \notin A(\widehat{x}),
\end{aligned}$$

so that (10.17) is obtained by dividing (10.19) by c_{ε}. Observe that $\mu_i^{\varepsilon} \geq 0$ for each $j \in J$ and that $\sqrt{\sum_{i\in I}(\lambda_i^{\varepsilon})^2 + \sum_{j\in J}\left(\mu_j^{\varepsilon}\right)^2} = 1$, i.e.,

$$\left(\lambda_0^{\varepsilon}, \lambda_1^{\varepsilon}, ..., \lambda_{|I|}^{\varepsilon}, \mu_1^{\varepsilon}, ..., \mu_{|J|}^{\varepsilon}\right) \in U_1(\mathbf{0}).$$

△

Using Fact 2, we can now complete the proof. Take a decreasing sequence $\{\varepsilon_n\}_n \subseteq (0, \widehat{\varepsilon}]$ with $\varepsilon_n \downarrow 0$, and consider the associated sequence $\left\{\left(\lambda_0^n, \lambda_1^n, ..., \lambda_{|I|}^n, \mu_1^n, ..., \mu_{|J|}^n\right)\right\}_n \subseteq U_1(\mathbf{0})$ whose existence is guaranteed by Fact 2.

Since the sequence $\left\{\left(\lambda_0^n, \lambda_1^n, ..., \lambda_{|I|}^n, \mu_1^n, ..., \mu_{|J|}^n\right)\right\}_n$ is contained in the compact set $U_1(\mathbf{0})$, by Theorem 275 there exists a subsequence

$$\left\{\left(\lambda_0^{n_k}, \lambda_1^{n_k}, ..., \lambda_{|I|}^{n_k}, \mu_1^{n_k}, ..., \mu_{|J|}^{n_k}\right)\right\}_k$$

convergent in $U_1(\mathbf{0})$, that is, there exists $\left(\lambda_0, \lambda_1, ..., \lambda_{|I|}, \mu_1, ..., \mu_{|J|}\right) \in U_1(\mathbf{0})$ such that

$$\left(\lambda_0^{n_k}, \lambda_1^{n_k}, ..., \lambda_{|I|}^{n_k}, \mu_1^{n_k}, ..., \mu_{|J|}^{n_k}\right) \to \left(\lambda_0, \lambda_1, ..., \lambda_{|I|}, \mu_1, ..., \mu_{|J|}\right).$$

By Fact 2, for each ε_{n_k} there exists $x^{n_k} \in B_{\varepsilon_{n_k}}(\widehat{x})$ for which (10.17) holds, i.e.,

$$\lambda_0^{n_k}\left(\frac{\partial f}{\partial x_z}(x^{n_k}) - 2(x^{n_k} - \widehat{x}_z)\right) - \sum_{i\in I}\lambda_i^{n_k}\frac{\partial g_i}{\partial x_z}(x^{n_k}) - \sum_{j\in J\cap A(\widehat{x})}\mu_j^{n_k}\frac{\partial h_j}{\partial x_z}(x^{n_k}) = 0,$$

for each $z = 1, ..., n$. Consider the sequence $\{x^{n_k}\}_k$ so constructed. From $x^{n_k} \in B_{\varepsilon_{n_k}}(\widehat{x})$ it follows that

$$\|x^{n_k} - \widehat{x}\| < \varepsilon_{n_k} \to 0,$$

and hence, for each $z = 1, ..., n$,

$$\begin{aligned} & \lambda_0 \frac{\partial f}{\partial x_z}(\widehat{x}) - \sum_{i \in I} \lambda_i \frac{\partial g_i}{\partial x_z}(\widehat{x}) - \sum_{j \in J \cap A(\widehat{x})} \mu_j \frac{\partial h_j}{\partial x_z}(x) \qquad (10.20) \\ = & \lim_k \left(\lambda_0^{n_k} \left(\frac{\partial f}{\partial x_z}(x^k) - 2(x^{n_k} - \widehat{x}_z) \right) - \sum_{i \in I} \lambda_i^{n_k} \frac{\partial g_i}{\partial x_z}(x^{n_k}) - \sum_{j \in J \cap A(\widehat{x})} \mu_j^{n_k} \frac{\partial h_j}{\partial x_z}(x^{n_k}) \right) \\ = & 0. \end{aligned}$$

On the other hand, $\lambda_0 \neq 0$. In fact, if it were $\lambda_0 = 0$, then by (10.20) it follows that

$$\sum_{i \in I} \lambda_i \frac{\partial g_i}{\partial x_z}(\widehat{x}) + \sum_{j \in J \cap A(\widehat{x})} \mu_j \frac{\partial h_j}{\partial x_z}(\widehat{x}) = 0, \qquad \forall z = 1, ..., n.$$

The linear independence of the gradients associated to the constraints that holds for the hypothesis of regularity of the constraints implies $\lambda_i = 0$ for each $i \in I$, which contradicts $(\lambda_0, \lambda_1, ..., \lambda_{|I|}, \mu_1, ..., \mu_{|J|}) \in U_1(\mathbf{0})$.

In conclusion, if we set $\widehat{\lambda}_i = \lambda_i / \lambda_0$ for each $i \in I$ and $\widehat{\mu}_j = \mu_j / \lambda_0$ for each $j \in J$, (10.20) implies (10.7). ■

10.2.1 Kuhn-Tucker Conditions

In view of Lemma 521, the Lagrangian associated to the problem of optimum (10.4) is the function

$$L : A \times \mathbb{R}^{|I|} \times \mathbb{R}_+^{|J|} \subseteq \mathbb{R}^{n+|I|+|J|} \to \mathbb{R}$$

defined by:[5]

$$\begin{aligned} L(x; \lambda, \mu) &= f(x) + \sum_{i \in I} \lambda_i (b_i - g_i(x)) + \sum_{j \in J} \mu_j (c_j - h_j(x)) \qquad (10.21) \\ &= f(x) + \lambda \cdot (b - g(x)) + \mu \cdot (c - h(x)), \end{aligned}$$

for each $(x; \lambda, \mu) \in A \times \mathbb{R}^{|I|} \times \mathbb{R}_+^{|J|}$. Note that in this case μ is required to be a non-negative vector.

We can now generalize Theorem 510 to the optimum problem (10.4). As we did for Theorem 510, also here we omit the proof because it is analogous to the one of Theorem 503.

[5] Observe the use of the notation $(x; \lambda, \mu)$ to underline the different status of x with respect to λ and μ.

Theorem 524 *Let $\widehat{x}$ be solution of the optimum problem (10.4). If the functions $f, \{g_i\}_{i\in I}$ and $\{h_j\}_{j\in J}$ are of class $\mathcal{C}^1$ and if the constraints are regular in $\widehat{x}$, then there exists a pair of vectors $\left(\widehat{\lambda}, \widehat{\mu}\right) \in \mathbb{R}^{|I|} \times \mathbb{R}^{|J|}_+$ such that the triple $\left(\widehat{x}; \widehat{\lambda}, \widehat{\mu}\right)$ satisfies the conditions:*

$$\nabla L_x \left(\widehat{x}; \widehat{\lambda}, \widehat{\mu}\right) = \mathbf{0}, \tag{10.22}$$

$$\widehat{\mu} \cdot \nabla L_\mu \left(\widehat{x}; \widehat{\lambda}, \widehat{\mu}\right) = 0, \tag{10.23}$$

$$\nabla L_\lambda \left(\widehat{x}; \widehat{\lambda}, \widehat{\mu}\right) = \mathbf{0}, \tag{10.24}$$

$$\nabla L_\mu \left(\widehat{x}; \widehat{\lambda}, \widehat{\mu}\right) \in \mathbb{R}^{|J|}_+. \tag{10.25}$$

The components $\widehat{\lambda}_i$ and $\widehat{\mu}_j$ of the vectors $\widehat{\lambda}$ and $\widehat{\mu}$ are called *Lagrange multipliers*, while the conditions (10.22)-(10.25) are called *Kuhn-Tucker conditions*.

The points $x \in A$ for which there exists a pair $(\lambda, \mu) \in \mathbb{R}^{|I|} \times \mathbb{R}^{|J|}_+$ such that the triple (x, λ, μ) satisfies the conditions (10.22)-(10.25) are called *points of Kuhn-Tucker*. The points of Kuhn-Tucker are therefore the solutions of the system (typically nonlinear) of equations and inequalities given by conditions (10.22)-(10.25). By Theorem 524 we can say that a necessary condition for a point x in which the constrains are regular to be solution of the optimum problem (10.4) is that it is a point of Kuhn-Tucker.[6]

Observe that a Kuhn-Tucker point $(x; \lambda, \mu)$ is not necessarily a stationary point for the Lagrangian, since the condition (10.25) requires only $\nabla L_\mu (x; \lambda, \mu) \in \mathbb{R}^{|J|}_+$, and not the stronger property $\nabla L_\mu (x; \lambda, \mu) = \mathbf{0}$.

Let (x, λ, μ) be a Kuhn-Tucker point, that is, a triple that satisfies conditions (10.22)-(10.25). By Lemma 522, expression (10.23) is equivalent to require $\mu_j = 0$ for each j such that $h_j(x) < c_j$. Hence, $\mu_j > 0$ implies that the correspondent constraint is binding at the point x, that is, $h_j(x) = c_j$.

Given its importance, we state formally this observation.

Proposition 525 *At a Kuhn-Tucker point (x, λ, μ) we have $\mu_j > 0$ only if $h_j(x) = c_j$.*

10.2.2 The Method of Elimination

Like Theorems 503 and 510, also Theorem 524 allows to solve the optimum problem (10.4) through a suitable generalization of the method of elimination seen in the previous chapter.

[6]Note the caveat "in which the constraints are regular". In fact, a point of Kuhn-Tucker in which the constraints are not regular does not fall in the ambit of Theorem 524.

In this case, let D_0 be the set of the points $x \in A$ where the regularity condition of the constraints does not hold, and let D_1 be instead the set of the points $x \in A$ where this condition holds.

The method of elimination consists of four steps:

1. We determine whether Theorem 317 can be applied, that is, if f is upper semi-continuous and coercive on C.

2. We find the set $D_0 \cap C$.

3. We find the set S of the points of Kuhn-Tucker that belong to D_1, i.e., the set of the points $x \in D_1$ for which there exists $(\lambda, \mu) \in \mathbb{R}^{|I|} \times \mathbb{R}_+^{|J|}$ such that the triple $(x; \lambda, \mu)$ satisfies the Kuhn-Tucker conditions (10.22)-(10.25).[7]

4. We build the set $\{f(x) : S \cup (D_0 \cap C)\}$. If $\widehat{x} \in S \cup (D_0 \cap C)$ is such that $f(\widehat{x}) \geq f(x)$ for each $x \in S \cup (D_0 \cap C)$, then such $\widehat{x}$ is solution of the optimum problem (10.4).

The first step of the method of elimination is the same of the previous chapter, while the other steps are the obvious extension of the method to the case of the problem (10.4).

Example 526 Consider the optimum problem:

$$\begin{array}{l} \max\limits_{x \in \mathbb{R}^2} \left(x_1 - 2x_2^2\right) \\ \text{sub } x_1^2 + x_2^2 \leq 1 \end{array} \tag{10.26}$$

This problem is of the form (10.4), where $f : \mathbb{R}^2 \to \mathbb{R}$ and $h : \mathbb{R}^2 \to \mathbb{R}$ are given by $f(x_1, x_2) = x_1 - 2x_2^2$, $h(x_1, x_2) = x_1^2 + x_2^2$, while $b = 1$. Since C is compact, the first step is completed by observing that here a solution exists by the Weierstrass Theorem.

We have

$$\nabla h(x) = (2x_1, 2x_2)$$

and hence the constraint is regular at each point $x \in C$, that is, $D_0 \cap C = \emptyset$.

The Lagrangian $L : \mathbb{R}^3 \to \mathbb{R}$ is given by

$$L(x_1, x_2, \mu) = x_1 - 2x_2^2 + \mu\left(1 - x_1^2 - x_2^2\right), \qquad \forall (x_1, x_2, \mu) \in \mathbb{R}^3,$$

[7] Observe that these points x satisfy for sure the constraints and hence we always have $S \subseteq D_1 \cap C$; it is therefore not necessary to check if for a point $x \in S$ we have also $x \in C$. A similar observation was made in the previous chapter.

and to find the set S of its Kuhn-Tucker points it is necessary to solve the system

$$\begin{cases} \frac{\partial L}{\partial x_1} = 1 - 2\mu x_1 = 0 \\ \frac{\partial L}{\partial x_2} = -4x_2 - 2\mu x_2 = 0 \\ \mu \frac{\partial L}{\partial \mu} = \mu\left(1 - x_1^2 - x_2^2\right) = 0 \\ \frac{\partial L}{\partial \mu} = 1 - x_1^2 - x_2^2 \geq 0 \\ \mu \geq 0 \end{cases}$$

We start by observing that $\mu \neq 0$, that is, $\mu > 0$. In fact, if $\mu = 0$ the first equation becomes $1 = 0$, a contradiction. We therefore assume that $\mu > 0$. The second equation implies $x_2 = 0$, and in turn the third equation implies $x_1 = \pm 1$. From the first equation it follows $\mu = \mp (1/2)$, and hence the only solution of the system is $(-1, 0, (1/2))$. The only Kuhn-Tucker point is therefore $(-1, 0)$, i.e., $S = \{(-1, 0)\}$.

In conclusion, $S \cup (D_0 \cap C) = \{(-1, 0)\}$ and the method of elimination allows to conclude that $(-1, 0)$ is the only solution of the optimum problem 10.26. Note that in this solution the constrain is binding (i.e., it is satisfied with equality), and in fact $\mu = (1/2) > 0$, as required by Proposition 525. ▲

Example 527 Consider the optimum problem:

$$\begin{aligned} &\max_{x \in \mathbb{R}^n} -\sum_{i=1}^{n} x_i^2 \\ &\text{sub } \sum_{i=1}^{n} x_i = 1,\ x_1 \geq 0,\ ...,\ x_n \geq 0 \end{aligned} \tag{10.27}$$

This problem is of the form (10.4), where $f : \mathbb{R}^n \to \mathbb{R}$ is given by $f(x) = -\sum_{i=1}^{n} x_i^2$, $g : \mathbb{R}^n \to \mathbb{R}$ is given by $g(x) = \sum_{i=1}^{n} x_i$ and $h_j(x) : \mathbb{R}^n \to \mathbb{R}$ are given by $h_j(x) = -x_j$ for $j = 1, ..., n$, while $b = 1$ and $c_j = 0$ for $j = 1, ..., n$. Observe that $C = \left\{x \in \mathbb{R}^n_+ : \sum_{i=1}^{n} x_i\right\}$ is the simplex Δ_{n+1} that we considered in Example 385. It is a compact set and hence also in this case the first step is completed thanks to the Weierstrass Theorem.

For each $x \in \mathbb{R}^n$ we have

$$\nabla g(x) = (1, ..., 1) \quad \text{and} \quad \nabla h_j(x) = -e^j,$$

and therefore the value of these gradients does not depend on the point x considered. To verify the regularity of the constraints, we consider the collection $\{(1, ..., 1), e^1, ..., e^n\}$ of these gradients. This collection has $n + 1$ elements and it is obviously linearly dependent (the fundamental versors e^1,..., e^n are the most classic basis of $\mathbb{R}^n$).

On the other hand, it is immediate to see that any subcollection with at most n elements is instead linearly independent. Hence, the only way to violate the regularity

of the constraints is that they are all binding, so that all the collection of $n+1$ elements have to be considered. Fortunately, however, there does not exist any point $x \in \mathbb{R}^n$ where all constraints are binding. In fact, the only point that satisfies with equality all the constraints $-x_j \leq 0$ is the origin 0, which nevertheless does not satisfy the equality constraint $\sum_{i=1}^{n} x_i = 1$.

We can conclude that the constraints are regular at all the points $x \in \mathbb{R}^n$, i.e., $D_0 = \emptyset$. Hence, $D_0 \cap C = \emptyset$ and also the second step of the method of elimination is complete.

The Lagrangian $L : \mathbb{R}^{2n+1} \to \mathbb{R}$ is given by

$$L(x_1, x_2, \mu) = -\sum_{i=1}^{n} x_i^2 + \lambda \left(1 - \sum_{i=1}^{n} x_i\right) + \sum_{i=1}^{n} \mu_i x_i, \qquad \forall (x, \lambda, \mu) \in \mathbb{R}^{2n+1},$$

and to find the set S of its Kuhn-Tucker points it is necessary to solve the system

$$\begin{cases} \frac{\partial L}{\partial x_i} = -2x_i - \lambda + \mu_i = 0, & \forall i = 1, ..., n \\ \lambda \frac{\partial L}{\partial \lambda} = \lambda \left(1 - \sum_{i=1}^{n} x_i\right) = 0 \\ \frac{\partial L}{\partial \lambda} = 1 - \sum_{i=1}^{n} x_i = 0 \\ \mu_i \frac{\partial L}{\partial \mu_i} = \mu_i x_i = 0, & \forall i = 1, ..., n \\ \frac{\partial L}{\partial \mu_i} = x_i \geq 0, & \forall i = 1, ..., n \\ \mu_i \geq 0, & \forall i = 1, ..., n \end{cases}$$

If we multiply by x_i the first n equations, we get

$$-2x_i^2 - \lambda x_i + \mu_i x_i = 0, \qquad \forall i = 1, ..., n$$

Adding up these new equations, we have

$$-2\sum_{i=1}^{n} x_i^2 - \lambda \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \mu_i x_i = 0,$$

and therefore

$$-2\sum_{i=1}^{n} x_i^2 - \lambda = 0,$$

that is, $\lambda = -2\sum_{i=1}^{n} x_i^2$. We conclude that $\lambda \leq 0$.

If $x_i = 0$, from the condition $\partial L / \partial x_i = 0$ it follows that $\lambda = \mu_i$. Since $\mu_i \geq 0$ and $\lambda \leq 0$, it follows that $\mu_i = 0$. In turn, this implies $\lambda = 0$ and hence using again the condition $\partial L / \partial x_i = 0$ we conclude that $x_i = \lambda = 0$ for each $i = 1, ..., n$. But this contradicts the condition $\lambda \left(1 - \sum_{i=1}^{n} x_i\right) = 0$, and we can therefore conclude that $x_i \neq 0$, that is, $x_i > 0$.

Since this holds for each $i = 1, ..., n$, it follows that $x_i > 0$ for each $i = 1, ..., n$. From the condition $\mu_i x_i = 0$ it follows that $\mu_i = 0$ for each $i = 1, ..., n$, and the first n equations become:

$$-2x_i - \lambda = 0, \qquad \forall i = 1, ..., n$$

that is,

$$x_i = -\frac{\lambda}{2}, \qquad \forall i = 1, ..., n.$$

The x_i are therefore all equal, and from $\sum_{i=1}^{n} x_i = 1$ it follows that

$$x_i = \frac{1}{n}, \qquad \forall i = 1, ..., n.$$

In conclusion,

$$S = \left\{ \left(\frac{1}{n}, ..., \frac{1}{n} \right) \right\}.$$

Since $D_0 = \emptyset$, we have $S \cup (D_0 \cap C) = \{(1/n, ..., 1/n)\}$, and the method of elimination allows to conclude that the point $(1/n, ..., 1/n)$ is the solution of the optimum problem 10.27. ▲

Example 528 Let $h : \mathbb{R} \to \mathbb{R}$ be a continuous function with $h'(t) < 0$ and $h''(t) < 0$ for each $t > 0$. In view of Exercise 13.0.58, both h and h' are strictly decreasing on $\mathbb{R}_+$. Therefore, h is also strictly concave on $\mathbb{R}_+$. Instead, we do not ask anything on the behavior of h on $(-\infty, 0)$. Still in view of Exercise 13.0.58, the inverse $(h')^{-1}$ is strictly decreasing on $h'(\mathbb{R}_+)$.[8]

For example, the function $h(t) = -t^{2m}$ with $m \geq 1$ satisfies these conditions, and so does the function

$$h(t) = \begin{cases} t \lg t & \text{if } t > 0 \\ 0 & \text{if } t \leq 0 \end{cases}$$

Consider the optimum problem:

$$\max_{x \in \mathbb{R}^n} \sum_{i=1}^{n} h(x_i) \tag{10.28}$$
$$\text{sub} \sum_{i=1}^{n} x_i = 1,\ x_1 \geq 0,\ ...,\ x_n \geq 0$$

This problem is of the form (10.4), where $f : \mathbb{R}^n \to \mathbb{R}$ is given by $f(x) = \sum_{i=1}^{n} h(x_i)$, $g : \mathbb{R}^n \to \mathbb{R}$ is given by $g(x) = \sum_{i=1}^{n} x_i$ and $h_j(x) : \mathbb{R}^n \to \mathbb{R}$ are given by $h_j(x) = -x_j$ for $j = 1, ..., n$, while $b = 1$ and $c_j = 0$ for $j = 1, ..., n$. Observe that this optimum problem generalizes Example 527, in which we had $h(x_i) = -x_i^2$. Now we will see how a simple (but not trivial) modification of what we did in Example 527 allows to solve also this much more general problem.

[8] To require $h'(t) < 0$ and $h''(t) < 0$ only for $t > 0$ and not more generally for $t \geq 0$ is not pedantic. In fact, there are important functions for which the differentiability at 0 is problematic (think for example of roots and logarithms) and therefore it is better not to assume it unless necessary. This explains also the interest of Exercise 13.0.58.

The first two steps of the method of elimination are identical to those seen in Example 527. In particular, we have $D_0 = \emptyset$.

The Lagrangian $L : \mathbb{R}^{2n+1} \to \mathbb{R}$ is given by

$$L(x_1, x_2, \mu) = \sum_{i=1}^{n} h(x_i) + \lambda \left(1 - \sum_{i=1}^{n} x_i\right) + \sum_{i=1}^{n} \mu_i x_i, \qquad \forall (x, \lambda, \mu) \in \mathbb{R}^{2n+1},$$

and to find the set S of its Kuhn-Tucker points it is necessary to solve the system

$$\begin{cases} \frac{\partial L}{\partial x_i} = h'(x_i) - \lambda + \mu_i = 0, & \forall i = 1, ..., n \\ \lambda \frac{\partial L}{\partial \lambda} = \lambda \left(1 - \sum_{i=1}^{n} x_i\right) = 0 \\ \frac{\partial L}{\partial \lambda} = 1 - \sum_{i=1}^{n} x_i = 0 \\ \mu_i \frac{\partial L}{\partial \mu_i} = \mu_i x_i = 0, & \forall i = 1, ..., n \\ \frac{\partial L}{\partial \mu_i} = x_i \geq 0, & \forall i = 1, ..., n \\ \mu_i \geq 0, & \forall i = 1, ..., n \end{cases}$$

If we multiply by x_i the first n equations, we get

$$h'(x_i) x_i - \lambda x_i + \mu_i x_i = 0, \qquad \forall i = 1, ..., n$$

Adding up these new equations, we have

$$\sum_{i=1}^{n} h'(x_i) x_i - \lambda \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \mu_i x_i = 0,$$

and therefore $\lambda = \sum_{i=1}^{n} h'(x_i) x_i$. Since $h'(x_i) \leq 0$ when $x_i \geq 0$, the condition $x_i \geq 0$ allows to conclude that $\lambda \leq 0$.

The condition $\sum_{i=1}^{n} x_i = 1$ is such that there exists $i = 1, ..., n$ such that $x_i > 0$. From the condition $\mu_i x_i = 0$ it follows that $\mu_i = 0$ for such i. In turn the condition $\partial L / \partial x_i = 0$ implies that $h'(x_i) = \lambda$. Therefore, $(h')^{-1}(\lambda) = x_i > 0$.

If $\mu_i > 0$, the condition $\mu_i x_i = 0$ implies $x_i = 0$. From the condition $\partial L / \partial x_i = 0$ it follows that $h'(0) = \lambda - \mu_i < \lambda$, and so

$$0 = (h')^{-1}(h'(0)) > (h')^{-1}(\lambda) > 0 \tag{10.29}$$

because the inverse $(h')^{-1}$ is a strictly decreasing function on R_+. The contradiction (10.29) allows to conclude that $\mu_i = 0$ for each $i = 1, ..., n$.

At this point the condition $\partial L / \partial x_i = 0$ implies $h'(x_i) = \lambda$ for each $i = 1, ..., n$, and therefore,

$$x_i = (h')^{-1}(\lambda), \qquad \forall i = 1, ..., n.$$

The x_i are therefore all equal, and from $\sum_{i=1}^{n} x_i = 1$ it follows that

$$x_i = \frac{1}{n}, \qquad \forall i = 1, ..., n.$$

In conclusion,

$$S = \left\{ \left(\frac{1}{n}, ..., \frac{1}{n} \right) \right\}.$$

Since $D_0 = \emptyset$, we have $S \cup (D_0 \cap C) = \{(1/n, ..., 1/n)\}$, and the method of elimination allows us to conclude that the point $(1/n, ..., 1/n)$ is the solution also of the optimum problem 10.28. ▲

10.3 Concave Programming

In Section 8.5 we saw how concave functions enjoy remarkable properties from the point of view of optimization. In this section we will see how such remarkable properties make the concave functions of particular interest in nonlinear programming.

We start with a simple but important result.

Proposition 529 *Consider the optimum problem (10.4). If the functions g_i are convex for each $i \in I$ and if the functions h_j are affine for each $j \in J$, then the admissible region C defined in (10.5) is convex. If $A = \mathbb{R}^n$, then C is also closed.*

It is very easy to give examples where C is no longer convex when the conditions of convexity and affinity used in this result are not satisfied.

Notice that the convexity condition of the g_i is much weaker than that of affinity on the h_j. This shows that the convexity of the admissible region is more natural for inequality constraints than for equality ones. This is a crucial "structural" difference between the two types of constraints, which differentiate them in a much stronger way than it may appear *prima facie*.

Proof (i) Suppose that the functions g_i are convex for each $i \in I$ and that the functions h_j are affine for each $j \in J$. Set:

$$\begin{aligned} C_i &= \{x \in A : g_i(x) \leq b_i\}, \qquad \forall i \in I, \\ C_j &= \{x \in A : h_j(x) = c_j\}, \qquad \forall j \in J. \end{aligned}$$

Let $x_1, x_2 \in C_i$ and $t \in [0,1]$. By the convexity of g_i we have:

$$g_i(tx_1 + (1-t)x_2) \leq tg_i(x_1) + (1-t)g_i(x_2) \leq tb_i + (1-t)b_i = b_i$$

and hence $tx_1 + (1-t)x_2 \in C_i$. Each set C_i is therefore convex. A similar argument shows that also each C_j is convex, and this implies the convexity of the set C defined in (10.5) since $C = \left(\bigcap_{i \in I} C_i\right) \cap \left(\bigcap_{j \in J} C_j\right)$.

Let $A = \mathbb{R}^n$. Since each g_i is convex on $\mathbb{R}^n$, by Corollary 434 the g_i are also continuous. This is true *a fortiori* for the functions h_j. We have $C_i = g_i^{-1}((-\infty, b_i])$

and $C_j = h_j^{-1}(c_j)$ and therefore thanks to Theorem 299 the sets C_i and C_j are closed because $(-\infty, b_i]$ and $\{c_j\}$ are closed sets of $\mathbb{R}$. Hence, also C is closed. ■

Motivated by Proposition 529, we give the following definition.

Definition 530 *An optimum problem (10.4) is called concave if the function f is concave and if the functions g_i are convex.*[9]

A concave problem of optimum has therefore the form

$$\max_{x \in A} f(x) \tag{10.30}$$
$$\text{sub} \quad g_i(x) = b_i, \qquad \forall i \in I,$$
$$h_j(x) \leq c_j, \qquad \forall j \in J, \tag{10.31}$$

where I and J are finite sets of indices (possibly empty), $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ is a concave objective function defined on an open and convex A, the affine functions $g_i : A \subseteq \mathbb{R}^n \to \mathbb{R}$ and the associated scalars $b_i \in \mathbb{R}$ characterize $|I|$ equality constraints, while the convex functions $h_j : A \subseteq \mathbb{R}^n \to \mathbb{R}$ with the associated scalars $c_j \in \mathbb{R}$ induce $|J|$ inequality constraints.

Concave Programming studies concave optimum problems, and it has important theoretical aspects that we will consider in the next chapter. Here we will study instead the solution techniques of concave problems.

Recall from Corollary 473 of Section 8.5 that the search of the solutions of an unconstrained optimum problem for concave functions was based on a very remarkable property: the first order necessary condition for the existence of a local maximum becomes sufficient for the existence of a global maximum in the case of concave functions.

The next fundamental result is the "constrained" version of this property. Notice that the regularity of the constraints does not play any role in this result.

Theorem 531 *In a concave optimum problem in which the functions $f, \{g_i\}_{i \in I}$ and $\{h_j\}_{j \in J}$ are Gateaux differentiable on A, the Kuhn-Tucker points are solutions of the problem.*

Proof Let $(x^*; \lambda^*, \mu^*)$ be a Kuhn-Tucker point for the optimum problem (10.4), that is, $(x; \lambda, \mu)$ satisfies the conditions (10.22)-(10.25). In particular, this means that

$$\nabla f(x^*) = \sum_{i \in I} \lambda_i^* \nabla g_i(x^*) + \sum_{j \in A(x^*)} \mu_j^* \nabla h_j(x^*) \tag{10.32}$$

[9] Implicitely here we are assuming that A is convex.

Since each g_i is affine and each h_j is convex, by (8.31) it follows that:

$$\begin{aligned} h_j(x) &\geq h_j(x^*) + \nabla h_j(x^*)(x - x^*), \qquad \forall j \in J, \forall x \in A, \quad (10.33) \\ g_i(x) &= g_i(x^*) + \nabla g_i(x^*)(x - x^*), \qquad \forall i \in I, \forall x \in A, \quad (10.34) \end{aligned}$$

For each $j \in A(x^*)$ we have $h_j(x^*) = c_j$, and hence $h_j(x) \leq h_j(x^*)$ for each $x \in C$ and each $j \in A(x^*)$. Moreover, $g_i(x^*) = g_i(x)$ for each $i \in I$ and each $x \in C$. By (10.33) and (10.34) it follows

$$\begin{aligned} \nabla h_j(x^*)(x - x^*) &\leq 0, \qquad \forall j \in A(x^*), \forall x \in C, \\ \nabla g_i(x^*)(x - x^*) &= 0, \qquad \forall i \in I, \forall x \in C \end{aligned}$$

Together with (10.32), we therefore have:

$$\nabla f(x^*)(x - x^*) = \sum_{i \in I} \widehat{\lambda}_i \nabla g_i(x^*)(x - x^*) + \sum_{j \in A(x^*)} \widehat{\mu}_j \nabla h_j(x^*)(x - x^*) \leq 0,$$

for each $x \in C$. On the other hand, by (8.31) we have:

$$f(x) \leq f(x^*) + \nabla f(x^*)(x - x^*), \qquad \forall x \in A,$$

and we conclude that $f(x) \leq f(x^*)$ for each $x \in C$, as desired. ■

Theorem 531 gives us a sufficient condition of optimum: if a point is of Kuhn-Tucker, then it is solution of the optimum problem. The condition is, however, not necessary: there can be solutions of a concave optimum problem that are not Kuhn-Tucker points. In view of Theorem 524, this can happen only if the solution is a point in which the constraints are not regular. Next example illustrates this situation.

Example 532 Consider the optimum problem:

$$\begin{aligned} &\max_{x \in \mathbb{R}^3} \left(-x_1 - x_2 - x_3^2\right) \qquad (10.35) \\ &\text{sub } x_1^2 + x_2^2 - 2x_1 \leq 0 \quad \text{and} \quad x_1^2 + x_2^2 + 2x_1 \leq 0 \end{aligned}$$

This problem is of the form (10.4), where $f : \mathbb{R}^3 \to \mathbb{R}$, $h_1 : \mathbb{R}^3 \to \mathbb{R}$ and $h_2 : \mathbb{R}^3 \to \mathbb{R}$ are given by $f(x_1, x_2, x_3) = -x_1 - x_2 - x_3^2$, $h_1(x_1, x_2, x_3) = x_1^2 + x_2^2 - 2x_1$, $h_2(x_1, x_2, x_3) = x_1^2 + x_2^2 + 2x_1$, while $c_1 = c_2 = 0$. Using Theorem 460 it is easy to verify that f is concave and that h_1 and h_2 are convex, so that (10.35) is a concave optimum problem.

The system of inequalities

$$\begin{aligned} x_1^2 + x_2^2 - 2x_1 \leq 0 \\ x_1^2 + x_2^2 + 2x_1 \leq 0 \end{aligned}$$

has the point $(0,0)$ as its unique solution. Hence, $C = \{x \in \mathbb{R}^3 : x_1 = x_2 = 0\}$ and the unique solution of the problem (10.35) is the point $(0,0,0)$. On the other hand,

$$\nabla h_1(0,0,0) = (-2,0,0) \quad \text{and} \quad \nabla h_2(0,0,0) = (2,0,0),$$

and hence the constraints are not regular at $(0,0,0)$. Since

$$\nabla f(0,0,0) = (-1,-1,0)$$

there does not exist any pair $(\mu_1, \mu_2) \in \mathbb{R}^2_+$ such that:

$$\nabla f(0,0,0) = \mu_1 \nabla h_1(0,0,0) + \mu_2 \nabla h_2(0,0,0)$$

and therefore the solution $(0,0,0)$ is not a Kuhn-Tucker point. ▲

By combining Theorems 524 and 531 we get the following necessary and sufficient optimality condition.

Theorem 533 *Consider a concave optimum problem in which the functions $f, \{g_i\}_{i\in I}$ and $\{h_j\}_{j\in J}$ are of class $\mathcal{C}^1$ on A. A point $x \in A$ where the constraints are regular is solution of such problem if and only if it is a Kuhn-Tucker point.*

Theorem 533 is a refinement of Theorem 524, and as such it allows to refine the method of elimination, which we will call *convex method of elimination* (convex method, for brevity). Such method is based on the following steps:

1. We determine if the problem is concave, that is, if the function f is concave, if the functions g_i are affine and if the functions h_j are convex.

2. We find the set $D_0 \cap C$.

3. We find the set T of the Kuhn-Tucker points,[10] i.e., the set of the points $x \in A$ for which there exists $(\lambda, \mu) \in \mathbb{R}^{|I|} \times \mathbb{R}^{|J|}_+$ such that the triple $(x; \lambda, \mu)$ satisfies the Kuhn-Tucker conditions (10.22)-(10.25).[11]

4. If $T \neq \emptyset$, then taken any $x^* \in T$, we construct the set

$$\{f(x) : \{x^*\} \cup (D_0 \cap C)\}.$$

All the points of T are solutions of the problem,[12] and a point $x \in D_0 \cap C$ is itself solution if and only if $f(x) = f(x^*)$.

[10] The set T considered here is therefore slightly different from the set T seen in the previous versions of the method of elimination.

[11] These points x satisfy surely the constraints and hence we have always $T \subseteq D_1 \cap C$; it is therefore not necessary to verify if for a point $x \in T$ we have also $x \in C$. A similar observation was done in Chapter 9.

[12] The set T is at most a singleton when f is strictly concave because in such a case there is at most a solution of the problem (Theorem 468).

5. If $T = \emptyset$, we determine if Theorem 317 can be applied, i.e., if f is upper semi-continuous and coercive on C. If this is the case, we then consider the optimum problem $\max_{x \in D_0 \cap C} f(x)$ and the solutions of this problem are also solutions of problem (10.4).

Since either step 4 or 5 applies, depending on whether or not T is empty, the actual steps of the convex method are four.

The convex method works thanks to Theorems 531 and 533. In fact, if $T \neq \emptyset$, then by Theorem 531 all points of T are solutions of the problem. In this case, a point $x \in D_0 \cap C$ that does not belong to T can in turn be a solution only if its value $f(x)$ is equal to that of any point in T.

When, instead, we have $T = \emptyset$, then Theorem 533 guarantees that no point in D_1 is solution of the problem. At this stage, if Theorem 317 ensures the existence of at least a solution, we can restrict the search to the set $D_0 \cap C$. In other words, it is sufficient to solve the optimum problem $\max_{x \in D_0 \cap C} f(x)$: the solutions of this problem are also solutions of problem (10.4), and viceversa.[13]

As it is easy to understand, the convex method becomes very powerful when $T \neq \emptyset$ because in such a case there is no need to verify the validity of existence theorems à la Weierstrass, but it is sufficient to find the Kuhn-Tucker points.

If we are satisfied with the solutions that are points of Kuhn-Tucker, without worrying about the possible existence of solutions that are not so, we can give a shortened version of the convex method, based uniquely on Theorem 531, which we call the *short convex method*.

This method is based only on two steps:

1. We determine whether the optimum problem (10.4) is concave, i.e., if the function f is concave, if the functions g_i are affine, and if the functions h_j are convex.

2. We find the set T of the Kuhn-Tucker points.

By Theorem 531, all the points of T are solutions of the problem. The short convex method is simpler than the convex method, and it does not require neither the use of existence theorems à la Weierstrass nor the study of the regularity of the constraints. The price of this simplification is in the possible inaccuracy of this method, which, being based on sufficient conditions, is not able to find the solutions where these conditions are not satisfied (by Theorem 533, these possible solutions are points where

[13] Observe that the problem $\max_{x \in D_0 \cap C} f(x)$ of the step 5 has for sure solution. In fact, if the problem $\max_{x \in C} f(x)$ has solutions and if none of them belong to $D_0 \cap C$, it follows that $\arg\max_{x \in C} f(x) = \arg\max_{x \in D_0 \cap C} f(x)$.

the constraints are not regular). Furthermore, the short method cannot be applied when $T = \emptyset$, and in such a case it is necessary to apply the complete convex method.

The convex short method is particularly powerful when the objective function f is strictly concave. In fact, in such a case a solution found with the short method is necessarily also the unique solution of the concave optimum problem. Therefore, in this case the short method is as effective as the complete convex method.

Example 534 Consider the optimum problem:

$$\begin{array}{l} \max\limits_{x\in\mathbb{R}^3} -\left(x_1^2+x_2^2+x_3^2\right) \\ \text{sub } 3x_1+x_2+2x_3\geq 1 \quad \text{and} \quad x_1\geq 0 \end{array} \tag{10.36}$$

This problem is of the form (10.4), where $f:\mathbb{R}^3\to\mathbb{R}$ is given by $f(x) = -(x_1^2+x_2^2+x_3^2)$, $h_1:\mathbb{R}^3\to\mathbb{R}$ is given by $h_1(x) = -(3x_1+x_2+2x_3)$ and $h_2(x):\mathbb{R}^3\to\mathbb{R}$ is given by $h_2(x) = -x_1$, while $c_1 = -1$ and $c_2 = 0$.

Using Theorem 460 it is easy to verify that f is strictly concave, while it is immediate to verify that h_1 and h_2 are convex. Therefore, (10.36) is a concave optimum problem. Since f is strictly concave, we apply without doubts the short convex method. To do this we have to find the set T of the Kuhn-Tucker points.

The Lagrangian $L:\mathbb{R}^5\to\mathbb{R}$ is given by

$$L(x_1,x_2,x_3,\mu_1,\mu_2) = -\left(x_1^2+x_2^2+x_3^2\right)+\mu_1(-1+3x_1+x_2+2x_3)+\mu_2x_1,$$

for each $(x_1,x_2,x_3,\mu_1,\mu_2)\in\mathbb{R}^3$, and to find the set T of its Kuhn-Tucker points it is necessary to solve the system of equalities and inequalities:

$$\begin{cases} \frac{\partial L}{\partial x_1} = -2x_1+3\mu_1+\mu_2 = 0 \\ \frac{\partial L}{\partial x_2} = -2x_2+\mu_1 = 0 \\ \frac{\partial L}{\partial x_3} = -2x_3+2\mu_1 = 0 \\ \mu_1\frac{\partial L}{\partial \mu_1} = \mu_1(-1+3x_1+x_2+2x_3) = 0 \\ \mu_2\frac{\partial L}{\partial \mu_2} = \mu_2x_1 = 0 \\ \frac{\partial L}{\partial \mu_1} = -1+3x_1+x_2+2x_3\geq 0 \\ \frac{\partial L}{\partial \mu_2} = x_1\geq 0 \\ \mu_1\geq 0, \mu_2\geq 0 \end{cases} \tag{10.37}$$

We consider four cases, depending on the fact that the multipliers μ_1 and μ_2 are null or not.

Case 1: $\mu_1 > 0$ and $\mu_2 > 0$. The conditions $\mu_2\partial L/\partial\mu_2 = \partial L/\partial x_1 = 0$ imply $x_1 = 0$ and $3\mu_1+\mu_2 = 0$. This last equation does not have strictly positive solutions μ_1 and μ_2, and hence we conclude that we cannot have $\mu_1 > 0$ and $\mu_2 > 0$.

Case 2: $\mu_1 = 0$ and $\mu_2 > 0$. The conditions $\mu_2 \partial L/\partial \mu_2 = \partial L/\partial x_1 = 0$ imply $x_1 = 0$ and $3\mu_1 = 0$, that is $\mu_1 = 0$. This contradiction shows that we cannot have $\mu_1 = 0$ and $\mu_2 > 0$.

Case 3: $\mu_1 > 0$ and $\mu_2 = 0$. The conditions $\mu_1 \partial L/\partial \mu_1 = \partial L/\partial x_1 = \partial L/\partial x_2 = \partial L/\partial x_3 = 0$ imply:

$$\begin{cases} -2x_1 + 3\mu_1 = 0 \\ -2x_2 + \mu_1 = 0 \\ -2x_3 + 2\mu_1 = 0 \\ 3x_1 + x_2 + 2x_3 = 1 \end{cases}$$

Solving for μ_1, we get $\mu_1 = 1/7$, and hence $x_1 = 3/14$, $x_2 = 1/14$ and $x_3 = 1/7$. The quintuple $(3/14, 1/14, 1/7, 1/7, 0)$ solves the system (10.37), and hence $(3/14, 1/14, 1/7)$ is a Kuhn-Tucker point.

Case 4: $\mu_1 = \mu_2 = 0$. The condition $\partial L/\partial x_1 = 0$ implies $x_1 = 0$, while the conditions $\partial L/\partial x_2 = \partial L/\partial x_3 = 0$ imply $x_2 = x_3 = 0$. It follows that the condition $\partial L/\partial \mu_1 \geq 0$ implies $-1 \geq 0$, and this contradiction shows that we cannot have $\mu_1 = \mu_2 = 0$.

In conclusion,

$$T = \{((3/14, 1/14, 1/7))\}$$

and since f is strictly concave the short convex method allows to conclude that $(3/14, 1/14, 1/7)$ is the unique solution of the optimum problem (10.36). ▲

We conclude with a last observation. The methods of solution seen in this chapter are based on the search of the Kuhn-Tucker points, and therefore they require the resolution of systems of nonlinear equations. In general these systems are not easy to solve and this limits the computational utility of these methods, whose importance is mostly theoretical. At a numerical level, nonlinear programming problems are solved with other methods, which the interested reader can find in books of numerical analysis.

Chapter 11

Explicit Constraints

The admissible region C of the optimum problem (10.4) is identified by a finite number of equality and inequality constraints. This type of constraints are sometimes called *implicit constraints* since they are defined through suitable functions g and h.

There are problems, especially in dynamic contexts, where it is required that the solution belongs to a region X of $\mathbb{R}^n$ that cannot be identified through a finite number of implicit constraints. In this case the constraint is of the general form $x \in X$ and this type of constraints are called *explicit* since they are not defined through functions.

Of course, explicit constraints are more general than the implicit ones: while there are explicit constraints that cannot be written as implicit constraints, the converse is always true: each implicit constraint can be always written in explicit form. In fact, the implicit constraints $g_i(x) = b_i$ and $h_j(x) = c_j$ for each $i \in I$ and $j \in J$ are equivalent to the explicit constraint $x \in C$, where C is given by (10.5).

When possible, the advantage of the implicit formulation of the constraints lies in the possibility of using Theorems 510 and 524, while the study of the explicit constraints $x \in X$ is much less easy, though something interesting can still be said at least when X is a closed and convex set, as it will be seen later. Note however that explicit constraints are often necessary and in principle unavoidable.

We illustrate this point by an example. Suppose to have the following familiar problem

$$\begin{aligned} &\max f(x) \\ &\text{sub } h_j(x) \leq c_j, \qquad \forall j \in J, \\ &x \geq 0 \end{aligned}$$

which has been widely discussed in the previous section. We have seen that the positivity constraint $x \geq 0$ may be incorporated as constraints $-x \leq 0$. Note that this can be incorrect. The reason is that the implicit assumption is that objective function $f(x)$ has to be defined over an open set containing the positive orthant $x \geq 0$. If f is

just only defined over $x \geq 0$ (think for instance of functions containing square roots or logaritms) this reduction method fails.[1]

The possible existence of explicit constraints leads us to the following generalization of the optimum problem (10.4):

$$\begin{aligned} &\max_{x \in A} f(x) \\ &\text{sub } g_i(x) = b_i, \qquad \forall i \in I, \\ &h_j(x) \leq c_j, \qquad \forall j \in J, \\ &x \in X \end{aligned} \tag{11.1}$$

where X is a subset of A. In the general optimum problem (11.1) there are both explicit and implicit constraints; in particular, we get back to the optimum problem (10.4) when $X = A$.

Formulation (11.1) is actually more general than (10.4) when the explicit constraint $x \in X$ is *irreducible*, i.e., when it cannot be expressed through implicit constraints. Formulation (11.1) is, however, also useful when there are conditions on the sign or on the value of the x_i. The classic example is the non-negativity condition of the x_i, which is very useful to express as an explicit constraint $x \in \mathbb{R}^n_+$, rather than implicitly through n inequalities $-x_i \leq 0$ (on the other hand, in (10.1) we already wrote $x \in \mathbb{R}^n_+$). In this case we write

$$\begin{aligned} &\max_{x \in A} f(x) \\ &\text{sub} \quad g_i(x) = b_i, \qquad \forall i \in I, \\ &h_j(x) \leq c_j, \qquad \forall j \in J, \\ &x \in \mathbb{R}^n_+ \end{aligned}$$

Here the use of the explicit constraint is done to simplify the exposition.

11.1 Variational Inequalities

Explicit constraints can either a necessity, when they are irreducible, or they can just simplify the exposition when they provide a simpler way to express implicit constraints. In any case, we want to generalize to the optimum problem (11.1) the solution methods previously seen for problem (10.4).

[1] Of course we may extend the objective function outside of $x \geq 0$. Unfortunately in several cases these extensions may lose some nice properties such as differentiability and concavity properties.

To this end, we will assume that X is a closed and convex subset of $\mathbb{R}^n$ as, otherwise, little can be said on explicit constraints. We start by considering directly the optimum problem:

$$\max_{x \in X} f(x) \tag{11.2}$$

where X is a generic closed and convex set of $\mathbb{R}^n$, without asking whether it is irreducible or if, instead, it can be written through equality/inequality constraints. Needless to say, a point $\widehat{x} \in X$ is solution of the optimum problem (11.2) if $f(\widehat{x}) \geq f(x)$ for each $x \in X$.

Consider the simplest case, i.e., $X = [a, b]$ with $a, b \in \mathbb{R}$. The optimum problem (11.2) becomes:

$$\max_{x \in [a,b]} f(x) \tag{11.3}$$

Suppose that $\widehat{x} \in [a, b]$ is solution of problem (11.3). It is easy to see that we can have two cases:

(i) $\widehat{x} \in (a, b)$, i.e., $\widehat{x}$ is an interior point; in this case, $f'(\widehat{x}) = 0$.

(ii) $\widehat{x} \in \{a, b\}$, i.e., $\widehat{x}$ is a boundary point; in this case, we have $f'(\widehat{x}) \leq 0$ if $\widehat{x} = a$, while we have $f'(\widehat{x}) \geq 0$ if $\widehat{x} = b$.

The next lemma gives a simple and elegant way to unify these two cases.

Proposition 535 *Let $f : A \subseteq \mathbb{R} \to \mathbb{R}$ be a function differentiable on an open set A in $\mathbb{R}$ and let $[a, b]$ be a closed and bounded interval contained in A. If $\widehat{x} \in [a, b]$ is solution of the optimum problem (11.3), then*

$$f'(\widehat{x})(x - \widehat{x}) \leq 0, \qquad \forall x \in [a, b]. \tag{11.4}$$

The viceversa holds if f is concave.

The proof of this result rests on the following lemma.

Lemma 536 *Under the hypotheses of Proposition 535, expression (11.4) is equivalent to $f'(\widehat{x}) = 0$ if $\widehat{x} \in (a, b)$, to $f'(\widehat{x}) \leq 0$ if $\widehat{x} = a$, and to $f'(\widehat{x}) \geq 0$ if $\widehat{x} = b$.*

Proof We divide the proof in three parts, one for each of the equivalences to prove.

(i) Let $\widehat{x} \in (a, b)$. We prove that (11.4) is equivalent to $f'(\widehat{x}) = 0$. If it holds $f'(\widehat{x}) = 0$, then $f'(\widehat{x})(x - \widehat{x}) = 0$ for each $x \in [a, b]$, and hence (11.4) holds. Viceversa, suppose that (11.4) holds. Setting $x = a$, we have $(a - \widehat{x}) < 0$ and so (11.4) implies $f'(\widehat{x}) \geq 0$. On the other hand, setting $x = b$, we have $(b - \widehat{x}) > 0$ and so (11.4) implies $f'(\widehat{x}) \leq 0$. In conclusion, $\widehat{x} \in (a, b)$ implies $f'(\widehat{x}) = 0$.

(ii) Let $\widehat{x} = a$. We prove that (11.4) is equivalent to $f'(a) \leq 0$. Let $f'(a) \leq 0$. Since $(x - a) > 0$ for each $x \in (a, b]$, it follows that $f'(a)(x - a) \leq 0$ for each $x \in [a, b]$, and hence (11.4) holds. Viceversa, suppose that (11.4) holds. Taking $x \in (a, b]$, we have $(x - a) > 0$ and so (11.4) implies $f'(a) \leq 0$.

(iii) Let $\widehat{x} = b$. We prove that (11.4) is equivalent to $f'(b) \geq 0$. Let $f'(b) \geq 0$. Since $(x - b) < 0$ for each $x \in [a, b)$, we have $f'(b)(x - b) \leq 0$ for each $x \in [a, b]$ and (11.4) is valid. Viceversa, suppose that (11.4) holds. Taking $x \in [a, b)$, we have $(x - b) < 0$ and so (11.4) implies $f'(b) \geq 0$. ■

Proof of Proposition 535. In view of Lemma 536, it only remains to prove that (11.4) becomes a sufficient condition when f is concave. Suppose therefore that f is concave and that $\widehat{x} \in [a, b]$ is such that (11.4) holds. We prove that this implies that $\widehat{x}$ is solution of problem (11.3). In fact, by Corollary 443 we have:

$$f(x) \leq f(\widehat{x}) + f'(\widehat{x})(x - \widehat{x}), \qquad \forall x \in [a, b],$$

and (11.4) therefore implies that $f(x) \leq f(\widehat{x})$ for each $x \in [a, b]$. Hence, $\widehat{x}$ is solution of the optimum problem (11.3). ■

Inequality (11.4) is an example of variational inequality. Beyond unifying the two cases, this variational inequality is interesting because when f is concave it gives us a necessary and sufficient condition for a point to be solution of the optimum problem considered.

Even more interesting is the fact that this characterization can be naturally extended to the case of functions of several variables.

Theorem 537 *Let $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ be a function Gateaux differentiable on the open and convex set A and let X be a closed and convex set contained in A. If $\widehat{x} \in X$ is solution of the optimum problem (11.2), then*

$$\nabla f(\widehat{x}) \cdot (x - \widehat{x}) \leq 0, \qquad \forall x \in X. \tag{11.5}$$

The viceversa holds if f is concave.

Proof Let $\widehat{x} \in X$ be solution of the optimum problem (11.2), i.e., $f(\widehat{x}) \geq f(x)$ for each $x \in X$. Given $x \in X$, set $z_t = (x + t(x - \widehat{x}))$ for $t \in [0, 1]$. Since X is convex, $z_t \in X$ for each $t \in [0, 1]$. Moreover, by (4.8) we have $f'(\widehat{x}; x - \widehat{x}) = \phi'_+(0)$, where $\phi(t) = f(z_t)$ for each $t \in [0, 1]$. For each $t \in [0, 1]$ we have $\phi(0) = f(\widehat{x}) \geq f(z_t) = \phi(t)$, and so $\phi : [0, 1] \to \mathbb{R}$ has a point of (global) maximum at $t = 0$. It follows that $\phi'_+(0) \leq 0$, which implies $f'(\widehat{x}; x - \widehat{x}) \leq 0$. By Proposition 140, $f'(\widehat{x}; x - \widehat{x}) = \nabla f(\widehat{x}) \cdot (x - \widehat{x})$ and we can therefore conclude that (11.5) holds.

To prove the converse, assume that f is concave. By (8.31) we have $f(x) \leq f(\widehat{x}) + \nabla f(\widehat{x}) \cdot (x - \widehat{x})$ for each $x \in A$, and therefore (11.5) implies $f(x) \leq f(\widehat{x})$ for each $x \in X$. ■

Expression (11.5) is therefore a necessary condition for $\widehat{x}$ to be solution of the optimum problem (11.2), and it becomes also sufficient when f is concave.

It is easy to check that (11.5) is equivalent to $\nabla f(\widehat{x}) = \mathbf{0}$ when $\widehat{x}$ is an interior point of X. For such points we therefore find the classic condition $\nabla f(\widehat{x}) = 0$ of Fermat Theorem 194. But, the importance of (11.5) is that it is a necessary condition also when $\widehat{x}$ is a boundary point. Note further that it is not "selfdual" in the sense that for the minimum problems it has changed into $\nabla f(\widehat{x}) \cdot (x - \widehat{x}) \geq 0$. Not so for interior maximum and minimum for which the condition $\nabla f(\widehat{x}) = 0$ is the same in both cases.

Theorem 537 provides interesting optimality conditions when the maximization is restricted to a generic closed and convex set X. Even if elegant, this condition is, however, of limited operational interest, i.e., in the actual resolution of the optimum problem (11.2). In fact, to check condition (11.5) it is necessary to consider all the points x belonging to X, something that can be very difficult.

Fortunately, there exists an important class of convex sets, the convex cones, in which condition (11.5) becomes easier to verify. We pause therefore for a while to introduce this class of convex sets.

11.2 Intermezzo: Convex Cones

11.2.1 Basic Properties

Cones can be defined in any vector space.

Definition 538 *A subset C of a vector space V is a cone if $v \in C$ implies $\alpha v \in C$ for each $\alpha \geq 0$.*

A cone is therefore a set closed with respect to non-negative scalar multiplication.

A first property of the cones to observe is that they always contain the neutral element. In fact, given any vector v belonging to a cone C, we have $0v = \mathbf{0}$ and therefore $\mathbf{0} \in C$.

In the sequel we will always consider convex cones and next simple result characterizes them.

Lemma 539 *A set C of a vector space is a convex cone if and only if $\alpha v + \beta w \in C$ for each $v, w \in C$ and each $\alpha, \beta \in \mathbb{R}_+$.*

Proof Let C be a convex cone and let $v, w \in C$ and $\alpha, \beta \in \mathbb{R}_+$. Since C is a cone, αv and βw belong to C. Since C is convex, we have

$$\alpha v + \beta w = (\alpha + \beta)\left(\frac{\alpha}{\alpha+\beta}v + \frac{\beta}{\alpha+\beta}w\right) \in C,$$

as desired.

Suppose now that $\alpha v + \beta w \in C$ for each $v, w \in C$ and each $\alpha, \beta \in \mathbb{R}_+$. Taken any two $v, w \in C$, for each $\alpha \geq 0$ we have $\alpha v = \alpha v + 0w \in C$. Therefore, C is a cone. Convexity of C is obvious, and so C is a convex cone. ■

A convex cone is therefore closed with respect to any positive linear combination, without the requirement that the coefficients must add to one. The basic properties of convex cones reflect this strong form of convexity that they feature. We now briefly state these properties, without proofs as they are obvious counterparts of similar results that we saw for convex sets in Section 7.8.

We first introduce a natural generalization of convex combinations: a linear combination $\sum_{i=1}^n \alpha_i v^i$ is a *positive combination* of the vectors $\{v^i\}_{i=1}^n$ if $\alpha_i \geq 0$ for each $i = 1, ..., n$. In the case $n = 2$, positive combinations can be written in the form $\alpha v + \beta w$ with $\alpha, b \in \mathbb{R}_+$.

Lemma 540 *A set C of a vector space V is a convex cone if and only if it is closed with respect to all the positive combinations of its own elements.*

We now see some examples of convex cones.

Example 541 *Both* the positive orthant $\mathbb{R}^n_+$ and the whole space $\mathbb{R}^n$ are convex cones. ▲

Example 542 Taken a vector $x \in \mathbb{R}^n$, the half-line $\{\alpha x : \alpha \geq 0\}$ that passes through the point x is a convex cone. More generally, given a finite set of vectors $\{v^i\}_{i=1}^n$ of a vector space, the set

$$\left\{\sum_{i=1}^n \alpha_i v^i : \alpha_i \geq 0 \ \forall i = 1, ..., n\right\}$$

of all their positive combinations is a convex cone. ▲

Example 543 Given a family of linear functionals $\{L_i\}_{i=1}^m$ of the form $L_i : V \to \mathbb{R}$, the set:

$$\{v \in V : L_i(v) \leq 0 \ \forall i = 1, ..., m\}$$

is a convex cone in V. In particular, given a matrix $\underset{m\times n}{A}$, the set

$$\{x \in \mathbb{R}^n : Ax \leq \mathbf{0}\} \tag{11.6}$$

is a convex cone in $\mathbb{R}^n$. This cone is determined by the m linear inequality constraints $Ax \leq \mathbf{0}$. ▲

Example 544 Given two linear functionals $L_1 : V \to \mathbb{R}$ and $L_2 : V \to \mathbb{R}$, the set $H = \{v \in V : L_1(v) \leq L_2(v)\}$ is a convex (possibly empty) cone. In fact, for each $v \in H$ we have

$$L_1(\alpha v) = \alpha L_1(v) \leq \alpha L_2(v) = L_2(\alpha v), \qquad \forall \alpha \geq 0,$$

and therefore $\alpha v \in H$. Similarly, it is possible to prove the convexity. ▲

Next lemma shows that the intersection preserves the structure of convex cone.

Lemma 545 *The intersection of any collection of convex subsets of a vector space is a convex set.*

This lemma leads us to the next fundamental notion of generated cone.

Definition 546 *Given any set S of vectors in a vector space V, we denote by $cone(S)$ the smallest convex cone in V that contains S. If V is normed, we denote by $\overline{cone}(S)$ the smallest closed and convex cone in V that contains S.*

The parallel with the notions of convex envelope seen in Section 7.8 should be clear.

Proposition 547 *Given a subset C of a vector space V, let $\{C_\alpha\}$ be the collection of all convex cones of V that contain C. We have $cone(C) = \bigcap_\alpha C_\alpha$. If, moreover, V is normed, we have $\overline{cone}(C) = \overline{cone(C)}$.*

The next result is the analog of Theorem 383 for convex cones and shows that $cone(S)$ is the set of all possible positive combinations of vectors of S.

Proposition 548 *Let A be a subset of a vector space V. A vector $v \in V$ belongs to $cone(S)$ if and only if it is a positive combination of vectors of S, i.e., if and only if there exist a finite set $\{v^i\}_{i \in I}$ of S and a finite set $\{\alpha_i\}_{i \in I}$ of scalars, with $\alpha_i \geq 0$ for each $i \in I$, such that $v = \sum_{i \in I} \alpha_i v^i$.*

Example 542 is a first illustration of this result. In fact, if we set $S = \{v^i\}_{i=1}^n$, we have

$$cone(S) = \left\{ \sum_{i=1}^{n} \alpha_i v^i : \alpha_i \geq 0 \ \forall i = 1, ..., n \right\}. \tag{11.7}$$

In the special case where S is the singleton $\{x\}$, with $x \in \mathbb{R}^n$, $cone(S)$ is the half-line passing through that point. If, instead, $S = \{e^1, ..., e^n\} \subseteq \mathbb{R}^n$, (11.7) becomes:

$$cone(S) = \{(\alpha_1, ..., \alpha_n) : \alpha_i \geq 0 \ \forall i = 1, ..., n\} = \mathbb{R}^n_+. \tag{11.8}$$

Hence, the positive orthant $\mathbb{R}^n_+$ is the convex cone generated by the fundamental versors. Instead, the space $\mathbb{R}^n$ can be viewed as the convex cone generated by $S = \{\pm e^i : i = 1, ..., n\}$.

Example 549 Let $S = \{1, x, ..., x^n, ...\}$ be the collection of the powers in the space of the polynomials $\mathcal{P}$. Here $cone\,(S)$ is the set of the polynomials with positive coefficients. ▲

Note that in all these examples the cardinality of S is much smaller than that of the generated cone $cone\,(S)$. When S is a finite set, the generated cone $cone\,(S)$ is called *polytope*. For example, (11.8) shows that $\mathbb{R}^n_+$ is a polytope.

The following result, due to Weyl (1935), shows that polytopes are precisely the cones determined by a finite set of linear inequality constraints (we omit the proof of this classic theorem).

Theorem 550 (Weyl) *A cone C in $\mathbb{R}^n$ is a polytope if and only if there exists a matrix $\underset{m\times n}{A}$ such that $C = \{x \in \mathbb{R}^n : Ax \leq \mathbf{0}\}$.*

This result is important because linear inequality constraints are a classic class of constraints in optimization problems and Weyl's Theorem characterizes them in terms of polytopes.

Given a norm $\|\cdot\|$ on $\mathbb{R}^n$, set $U = \{x \in \mathbb{R}^n : \|x\| = 1\}$; that is, U is the unit sphere in $\mathbb{R}^n$ according to this norm.

Theorem 551 *Given a closed and convex cone C in $\mathbb{R}^n$, we have*

$$C = cone\,(\text{ext}\,(C \cap U))\,. \tag{11.9}$$

Proof The set $C \cap U$ is compact and convex, and hence by Theorem 404 we have $C \cap U = co\,(\text{ext}\,(C \cap U))$. On the other hand, given $x \in C$, we have $x/\|x\| \in C \cap U$. Hence, $C = cone\,(C \cap U)$, from which it follows (11.9). ■

Theorem 551 is a version of the Minkowski Theorem for cones. Its importance lies in reducing the study of the cone to a set of extreme points. Two observations: (i) thanks to the special structure of the cones, relative to Theorem 404 here compactness is not required; (ii) the choice of the norm $\|\cdot\|$ in $\mathbb{R}^n$ is arbitrary and, depending on the set C considered, a certain norm can turn out to be particularly useful in the construction of $\text{ext}\,(C \cap U)$.

Notice that when $C \subseteq \mathbb{R}^n_+$, if we consider the norm $\|\cdot\|_1$ expression (11.9) becomes

$$C = cone\,(\text{ext}\,(C \cap \Delta_{n-1}))\,, \tag{11.10}$$

and so C is determined by the extreme points of its intersection with the simplex. For example, if $C = \mathbb{R}^n_+$, we have $\text{ext}\left(\mathbb{R}^n_+ \cap \Delta_{n-1}\right) = \{e^i\}_{i=1}^n$ and so (11.10) becomes $\mathbb{R}^n_+ = cone\left(\{e^i\}_{i=1}^n\right)$. We thus find again expression (11.8).

11.2.2 The Normal Cone and Equation (11.5)

Given a convex set C of $\mathbb{R}^n$ and a point $\overline{x} \in C$, the *normal cone* $N_C(\overline{x})$ of C with respect to $\overline{x}$ is given by

$$N_C(\overline{x}) = \{y \in \mathbb{R}^n : y \cdot (x - \overline{x}) \leq 0 \quad \forall x \in C\}.$$

Next we give couple of important properties of $N_C(\overline{x})$. In particular, (ii) shows that $N_C(\overline{x})$ is non trivial only if $\overline{x}$ is a boundary point.

Lemma 552 *Let C be a convex set of $\mathbb{R}^n$ and let $\overline{x} \in C$. We have:*

(i) $N_C(\overline{x})$ is a closed and convex cone;

(ii) $N_C(\overline{x}) = \{\mathbf{0}\}$ if $\overline{x}$ is an interior point of C.

Proof (i) The set $N_C(\overline{x})$ is clearly closed. Moreover, given $x_1, x_2 \in N_C(x)$ and $\alpha, \beta \geq 0$, we have

$$(\alpha x_1 + \beta x_2) \cdot (x - \overline{x}) = \alpha x_1 \cdot (x - \overline{x}) + \beta x_2 \cdot (x - \overline{x}) \leq 0 \quad \forall x \in C$$

and so $\alpha x_1 + \beta x_2 \in N_C(\overline{x})$. By Lemma 539, $N_C(\overline{x})$ is a convex cone.

(ii) Let $\overline{x}$ be an interior point of C. Then, there exists a neighborhood $B_\varepsilon(\overline{x})$ included in C, so $\overline{x} + \varepsilon' e^i$ and $\overline{x} - \varepsilon' e^i$ belong to C for each $i = 1, ..., n$ and $0 < \varepsilon' \leq \varepsilon$. Hence, for each $y \in N_C(\overline{x})$ and each $i = 1, ..., n$, we have:

$$y \cdot \left(\overline{x} + \varepsilon' e^i - \overline{x}\right) = \varepsilon' y_i \leq 0 \quad \text{and} \quad y \cdot \left(\overline{x} - \varepsilon' e^i - \overline{x}\right) = -\varepsilon' y_i \leq 0,$$

which implies $y_i = 0$, that is, $y = \mathbf{0}$. It follows that $N_C(\overline{x}) = \{\mathbf{0}\}$. ■

To see the importance of normal cones, observe that the fundamental equation (11.5) can be written as:

$$\nabla f(\widehat{x}) \in N_X(\widehat{x}). \tag{11.11}$$

Therefore, $\widehat{x}$ is solution of the optimum problem (11.2) only if the gradient $\nabla f(\widehat{x})$ belongs to the normal cone of X with respect to $\widehat{x}$. This way of writing condition (11.5) is useful because, given a set X, if we can describe the form that the normal cone has – something that does not require any knowledge of the objective function f – we can then have an idea of the form that take the "first order condition" for the optimum problems that have X as an explicit constraint.

In other words, (11.11) allows to distinguish two parts in the first order condition: the part $N_X(\widehat{x})$, determined by the explicit constraint X, and the part $\nabla f(\widehat{x})$, determined by the objective function. This distinction between the roles of the objective

function and of the constraint allows to face the optimum problem (11.2) with greater effectiveness.

The next result characterizes the normal cone for a fundamental class of convex sets.

Proposition 553 *Let $C = \overline{cone}\,(S)$ be the closed and convex cone generated by a set S in $\mathbb{R}^n$. If $\widehat{x} \in C$, then*

$$N_C(\widehat{x}) = \{y \in \mathbb{R}^n : y \cdot \widehat{x} = 0 \text{ and } y \cdot x \leq 0 \ \forall x \in S\}. \tag{11.12}$$

Proof Let $y \in \mathbb{R}^n$ be such that $y \cdot \widehat{x} = 0$ and $y \cdot x \leq 0$ for each $x \in S$. We prove that $y \in N_C(\widehat{x})$. Given $x \in cone\,(S)$, by Theorem 551 there exist a finite set $\{x_i\}_{i \in I}$ in S and a finite set $\{\alpha_i\}_{i \in I}$ of scalars with $\alpha_i \geq 0$ for each $i \in I$, such that $x = \sum_{i \in I} \alpha_i x_i$. Hence, we have:

$$y \cdot x = y \cdot \left(\sum_{i \in I} \alpha_i x_i \right) = \sum_{i \in I} \alpha_i (y \cdot x_i) \leq 0. \tag{11.13}$$

Let $x \in C$. Since $C = \overline{cone}\,(S)$, there exists a sequence $\{x_n\}_n \subseteq cone\,(S)$ such that $x_n \to x$. Expression (11.13) implies $y \cdot x_n \leq 0$ for each $n \geq 1$, and hence $y \cdot x = \lim_n y \cdot x_n \leq 0$. Since $y \cdot \widehat{x} = 0$, we therefore have $y \cdot (x - \widehat{x}) = 0$ for each $x \in C$, which implies $y \in N_C(\widehat{x})$, as desired.

Viceversa, let $y \in N_C(\widehat{x})$. Since $\mathbf{0} \in N_C(\widehat{x})$, we have $y \cdot \widehat{x} \geq 0$. On the other hand, $\widehat{x} \in C$ implies $2\widehat{x} \in C$ since C is a cone. Therefore, $y \cdot \widehat{x} = y \cdot (2\widehat{x} - \widehat{x}) \leq 0$, and we can conclude that $y \cdot \widehat{x} = 0$. In turn, this implies $y \cdot x = y \cdot (x - \widehat{x}) \leq 0$ for each $x \in C$, and in particular $y \cdot x \leq 0$ for each $x \in S$. This completes the proof. ■

By Theorem 551, for each closed and convex cone C in $\mathbb{R}^n$ we have

$$C = cone\,(\text{ext}\,(C \cap U)).$$

Therefore, (11.12) implies

$$N_C(\widehat{x}) = \{y \in \mathbb{R}^n : y \cdot \widehat{x} = 0 \text{ and } y \cdot x \leq 0 \ \forall x \in \text{ext}\,(C \cap U)\},$$

which is an especially useful representation of the cone $N_C(\widehat{x})$.

Example 554 If $C = \mathbb{R}^n_+$, (11.12) becomes:

$$N_C(\widehat{x}) = \{y \in \mathbb{R}^n : y_i \widehat{x}_i = 0 \text{ and } y_i \leq 0 \quad \forall i = 1, ..., n\}. \tag{11.14}$$

The condition $y_i \leq 0$ for each $i = 1, ..., n$ follows directly from (11.8), that is, from $cone\left(\{e^i\}_{i=1}^n\right) = \mathbb{R}^n_+$. On the other hand, from $y \cdot \widehat{x} = 0$ it follows that $y_i \leq 0$ for each $i = 1, ..., n$ implies $y_i x_i = 0$ for each $i = 1, ..., n$. ▲

11.3 Variational Inequalities on Cones

After having introduced cones, we get back to the analysis of the optimum problem (11.2). Thanks to (11.11) and to Proposition 553, for convex cones Theorem 537 takes the following form.

Proposition 555 *Let $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ be Gateaux differentiable on the open set A and let $X = \overline{cone}\,(S)$ be the closed and convex cone generated by a set $S \subseteq A$. If $\widehat{x} \in X$ is solution of the optimum problem (11.2), then*

$$\nabla f(\widehat{x}) \cdot \widehat{x} = 0, \tag{11.15}$$

$$\nabla f(\widehat{x}) \cdot x \leq 0, \qquad \forall x \in S. \tag{11.16}$$

The viceversa holds if f is concave.

Relative to condition (11.5), conditions (11.15) and (11.16) are easier to verify since they only involve $\widehat{x}$ and the elements of S, which are in general much fewer than those of $\overline{cone}\,(S)$.

Proof By (11.11) and (11.12), expression (11.5) is equivalent to

$$\nabla f(\widehat{x}) \in N_X(\widehat{x}) = \{y \in \mathbb{R}^n : y \cdot \widehat{x} = 0 \text{ and } y \cdot x \leq 0 \ \forall x \in S\},$$

which proves that conditions (11.15) and (11.16) are in this case equivalent to (11.5). The result is therefore a consequence of Theorem 537. ■

In the important special case $X = \mathbb{R}^n_+$, thanks to (11.14) Proposition 555 takes a particularly interesting form. In fact, conditions (11.15) and (11.16) reduce to check n equalities – i.e., (11.17) – and n inequalities – i.e., (11.18).

Corollary 556 *Let $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ be Gateaux differentiable on the open set A containing $\mathbb{R}^n_+$ and let $X = \mathbb{R}^n_+$. If $\widehat{x} \in X$ is solution of the optimum problem (11.2), then, for each $i = 1, ..., n$, we have:*

$$\widehat{x}_i \frac{\partial f}{\partial x_i}(\widehat{x}) = 0, \tag{11.17}$$

$$\frac{\partial f}{\partial x_i}(\widehat{x}) \leq 0. \tag{11.18}$$

The viceversa holds if f is concave.

11.4 Resolution of the General Optimum Problem (sketch)

We now have all we need to extend Theorem 524 to the general optimum problem (11.1). We first extend to the general case Definition 520 of regularity of the constraints. As usual, $A(x)$ denotes the set of inequality constraints non binding at x.

Definition 557 *Problem (11.1) has regular constraints at a point $x \in X$ if there does not exist a vector $\alpha \in \mathbb{R}^{|I|+|A(x)|}$ such that $\alpha \neq \mathbf{0}$ and*

$$\sum_{i \in I} \alpha_i \nabla g_i(x) + \sum_{j \in A(x)} \alpha_j \nabla h_j(x) \in N_X(x). \tag{11.19}$$

Definition 557 thus generalizes Definition 520 to optimum problems with explicit constraints. In fact, the optimum problem (10.4) is the special case of problem (11.1) in which $X = A$. Since A is an open, Lemma 552 implies $N_A(x) = \{\mathbf{0}\}$ for each $x \in A$, and (11.19) becomes

$$\sum_{i \in I} \alpha_i \nabla g_i(x) + \sum_{j \in A(x)} \alpha_j \nabla h_j(x) = \mathbf{0}.$$

In the case $X = A$ the regularity of the constraints at x is precisely the requirement that the gradients $\nabla g_i(x)$ and the gradients $\nabla h_j(x)$ with $j \in A(x)$ are linearly independent, and therefore we find again Definition 520.

The next result, proved as Theorem 4.2 p. 198 in Rockafellar (1993), generalizes Theorem 524 to the optimum problem (11.1).

Theorem 558 *Let $\widehat{x}$ be local solution of the optimum problem (11.1), where X is a closed and convex subset of A. If the functions $f, \{g_i\}_{i \in I}$ and $\{h_j\}_{j \in J}$ are of class $\mathcal{C}^1$ and if the constraints are regular at $\widehat{x}$, then there exists a pair of vectors $\left(\widehat{\lambda}, \widehat{\mu}\right) \in \mathbb{R}^{|I|} \times \mathbb{R}_+^{|J|}$ such that the triple $\left(\widehat{x}; \widehat{\lambda}, \widehat{\mu}\right)$ satisfies the conditions:*

$$\nabla L_x\left(\widehat{x}; \widehat{\lambda}, \widehat{\mu}\right) \in N_X(\widehat{x}), \tag{11.20}$$

$$\widehat{\mu} \cdot \nabla L_\mu\left(\widehat{x}; \widehat{\lambda}, \widehat{\mu}\right) = 0, \tag{11.21}$$

$$\nabla L_\lambda\left(\widehat{x}; \widehat{\lambda}, \widehat{\mu}\right) = \mathbf{0}, \tag{11.22}$$

$$\nabla L_\mu\left(\widehat{x}; \widehat{\lambda}, \widehat{\mu}\right) \in \mathbb{R}_+^{|J|}. \tag{11.23}$$

Conditions (11.20)-(11.23) are the *Kuhn-Tucker conditions* of the optimum problem (11.1), while the points $x \in A$ for which there exists a pair $(\lambda, \mu) \in \mathbb{R}^{|I|} \times \mathbb{R}_+^{|J|}$ such

that the triple (x, λ, μ) satisfies conditions (11.20)-(11.23) are the *Kuhn-Tucker points* of problem (11.1). For them still apply the considerations made after Theorem 524 about the Kuhn-Tucker points of the problem without explicit constraints (10.4).

Thanks to Proposition 553, Theorem 558 takes a simpler form when $X = \overline{cone}\,(S)$.

Corollary 559 *Let $\widehat{x}$ be local solution of the optimum problem (11.1), where $X = \overline{cone}\,(S)$ is the closed and convex cone generated by a set S contained in A. If the functions $f, \{g_i\}_{i\in I}$ and $\{h_j\}_{j\in J}$ are of class $\mathcal{C}^1$ and if the constraints are regular at $\widehat{x}$, then there exists a pair of vectors $\left(\widehat{\lambda}, \widehat{\mu}\right) \in \mathbb{R}^{|I|} \times \mathbb{R}_+^{|J|}$ such that the triple $\left(\widehat{x}; \widehat{\lambda}, \widehat{\mu}\right)$ satisfies the conditions:*

$$\nabla L_x\left(\widehat{x}; \widehat{\lambda}, \widehat{\mu}\right) \cdot x \leq 0, \qquad \forall x \in S, \tag{11.24}$$

$$\nabla L_x\left(\widehat{x}; \widehat{\lambda}, \widehat{\mu}\right) \cdot \widehat{x} = 0, \tag{11.25}$$

$$\widehat{\mu} \cdot \nabla L_\mu\left(\widehat{x}; \widehat{\lambda}, \widehat{\mu}\right) = 0, \tag{11.26}$$

$$\nabla L_\lambda\left(\widehat{x}; \widehat{\lambda}, \widehat{\mu}\right) = \mathbf{0}, \tag{11.27}$$

$$\nabla L_\mu\left(\widehat{x}; \widehat{\lambda}, \widehat{\mu}\right) \in \mathbb{R}_+^{|J|}. \tag{11.28}$$

When $X = \mathbb{R}_+^n$, thanks to Corollary 556 conditions (**??**) and (11.23) are equivalent to:

$$\widehat{x}_i \frac{\partial f}{\partial x_i}(\widehat{x}) = 0, \qquad \forall i = 1, ..., n. \tag{11.29}$$

$$\frac{\partial f}{\partial x_i}(\widehat{x}) \leq 0, \qquad \forall i = 1, ..., n. \tag{11.30}$$

Theorem 558 allows to solve the problem (11.1) with the Method of Elimination, which is based on the following steps:

(i) We determine if Theorem 317 can be applied, i.e., if f is upper semicontinuous and coercive on $C \cap X$.

(ii) We find the set $D_0 \cap C \cap X$.

(iii) We find the set T of the Kuhn-Tucker points that belong to D_1, that is, the set of the points $x \in D_1$ for which there exists $(\lambda, \mu) \in \mathbb{R}^{|I|} \times \mathbb{R}_+^{|J|}$ such that the triple $(x; \lambda, \mu)$ satisfies the Kuhn-Tucker conditions (11.20)-(11.23).

(iv) We construct the set $\{f(x) : T \cup (D_0 \cap C \cap X)\}$. If $\widehat{x} \in T \cup (D_0 \cap C \cap X)$ is such that $f(\widehat{x}) \geq f(x)$ for each $x \in T \cup (D_0 \cap C \cap X)$, then such $\widehat{x}$ is solution of the optimum problem (10.4).

Chapter 12

Abstract Equations

Given any two spaces X and Y, in its most general form an equation has the form

$$f(x) = y_0, \tag{12.1}$$

where f is a function $f : X \to Y$ and y_0 is a given element of Y.[1] The solutions of equation (12.1) are all $x \in X$ such that $f(x) = y_0$.

For example, the second order equation

$$\alpha_0 + \alpha_1 x + \alpha_2 x^2 = 0.$$

can be written as

$$f(x) = 0 \tag{12.2}$$

where $f : \mathbb{R} \to \mathbb{R}$ is the polynomial $f(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2$ and $y_0 = 0$. Its solutions are all $x \in \mathbb{R}$ that satisfy (12.2). Also the linear equation system

$$Ax = b, \tag{12.3}$$

where A is a $m \times n$ matrix and x and b are vectors in $\mathbb{R}^n$, can be written as

$$T(x) = b,$$

where $T : \mathbb{R}^n \to \mathbb{R}^m$ is the linear application $T(x) = Ax$ (see Section 3.5.2). Here the solutions are all $x \in \mathbb{R}^n$ such that $T(x) = b$.

Two main questions can be asked on the solutions of equation (12.1):

(i) can the equation be solved globally: given any $y_0 \in Y$ is there $x \in X$ that satisfies (12.1)? if so, is the solution unique?

[1] We write y_0 in place of y to make clear that y_0 should be regarded as a fixed element of Y and not as a variable.

(ii) can the equation be solved locally: given a $y_0 \in Y$ is there $x \in X$ that satisfies (12.1)? if so, is the solution unique?

To discuss these questions, consider the weak inverse (correspondence) $f^{-1} : Y \to 2^X$ given by $f^{-1}(y) = \{x \in X : f(x) = y\}$. We say that f is *weakly invertible at* $y \in Y$ if $f^{-1}(y)$ is nonempty, while we say that f is *invertible at* y if $f^{-1}(y)$ is a singleton. If f is weakly invertible (resp., invertible) at all $y \in Y$, we say that f is *weakly invertible* (resp., *invertible*).

Using this terminology, the above two questions can be rephrased in more precise terms as follows:

(i) is f weakly invertible? if so, is it invertible?

(ii) is f weakly invertible at $y_0 \in Y$? if so, is it invertible at y_0?

Question (i) is clearly much more demanding than (ii). In fact, (i) requires the global invertibility of f, while (ii) only requires f to be locally invertible. Nevertheless, in Section 3.5.2 we were able to answer question (i) for linear equation systems (12.3). In fact, by Proposition 124 the linear application $T : \mathbb{R}^n \to \mathbb{R}^n$ is invertible if and only if A is invertible, that is, if $\det A \neq 0$. Condition $\det A \neq 0$ thus ensures that in correspondence of each $b \in \mathbb{R}^n$ there is a unique solution $x \in \mathbb{R}^n$ given by $T^{-1}(b) = A^{-1}b$.

12.1 Operator Equations

At this point, to make further progress we consider the special case of (12.1) where both X and Y are a vector space V. That is, we consider the operator equation

$$T(v) = w_0, \tag{12.4}$$

where $T : V \to V$ is an operator and $w_0 \in V$. Though special, (12.4) includes many equations of interest. For example, observe that equations (12.2) and (12.3), when $m = n$, are special cases of (12.4) with $V = \mathbb{R}$ and $V = \mathbb{R}^n$, respectively.

A first taxonomy: the operator equation (12.4) is

(i) *linear* if the operator $T : V \to V$ is linear and is *nonlinear* otherwise;

(ii) *homogeneous* if $w_0 = \mathbf{0}$ and *nonhomogeneous* otherwise.

For example, equation (12.2) is nonlinear and homogeneous, while equation (12.3) is linear and nonhomogeneous, unless $b = \mathbf{0}$.[2] Next we present few other classic operator equations.

Volterra Integral Equations

Given a continuous function $\psi : [a,b] \times [a,b] \times \mathbb{R} \to \mathbb{R}$, consider the operator $F : C([a,b]) \to C([a,b])$ given by

$$F(f)(s) = \int_a^s \psi(s,t,f(t))\,dt, \qquad \forall s \in [a,b], \tag{12.5}$$

for each $f \in C([a,b])$. The *Volterra operator equation* is

$$f(s) = \int_a^s \psi(s,t,f(t))\,dt + g(s), \qquad \forall s \in [a,b],$$

for a given $g \in C([a,b])$. The unknown function $f \in C([a,b])$ has to be determined.

In operator notation, the Volterra equation can be written as

$$f - F(f) = g. \tag{12.6}$$

In terms of equation (12.4) we have $w_0 = g$ and $T = I - F$. It is a nonlinear equation, which is homogeneous when g is zero.

In (12.5) there is a variable limit s of integration. When this limit is constant, we need to consider the operator $G : C([a,b]) \to C([a,b])$ given by

$$G(f)(s) = \int_a^b \psi(s,t,f(t))\,dt, \qquad \forall s \in [a,b], \tag{12.7}$$

for each $f \in C([a,b])$. Given a $g \in C([a,b])$, the equation

$$f(s) = \int_a^b \psi(s,t,f(t))\,dt + g(s), \qquad \forall s \in [a,b],$$

is called a *Fredholm operator equation*. In operator notation we have

$$f - G(f) = g. \tag{12.8}$$

[2]Notice that a nonhomogeneous equation $T(v) = w_0$ can be made homogeneous by considering the transformation $T_{w_0}(v) = T(v) - w_0$. However, as the above discussion on equation (12.1) showed, it is important to consider explicitly the term w_0. This is why we study equation (12.4) rather than a general homogeneous equation $T(v) = \mathbf{0}$.

Hammerstein Equation We now present an important class of Volterra and Fredholm equations. Given a continuous function $\phi : [a,b] \times \mathbb{R} \to \mathbb{R}$, the *Nemitski operator* $N : C([a,b]) \to C([a,b])$ is given by:

$$N(f)(t) = \phi(t, f(t)), \qquad \forall t \in [a,b],$$

for each $f \in C([a,b])$. Moreover, given any continuous kernel $k : [a,b] \times [a,b] \to \mathbb{R}$, let $K : C([a,b]) \to C([a,b])$ be the linear operator given by

$$K(f)(s) = \int_a^s k(s,t) f(t)\, dt, \qquad \forall s \in [a,b], \tag{12.9}$$

for each $f \in C([a,b])$. For instance, the operator of Example 351 features an additively separable kernel $k(s,t) = h_1(t) + h_2(s)$.

The *Volterra-Hammerstein operator equation* is

$$f(s) = \int_a^s k(s,t)\, \phi(t, f(t))\, dt + g(s), \qquad \forall s \in [a,b], \tag{12.10}$$

for a given $g \in C([a,b])$. The unknown function $f \in C([a,b])$ has to be determined.

In operator notation, the Volterra-Hammerstein equation can be written as[3]

$$f - KN(f) = g.$$

Hence, in terms of equation (12.4) here we have $w_0 = g$ and $T = I - KN$. A Volterra-Hammerstein equation is thus a Volterra equation with $F = KN$ in (12.6).

If in (12.9) we make constant the limit s of integration we get a *Fredholm-Hammerstein equation*

$$f(s) = \int_a^b k(s,t)\, \phi(t, f(t))\, dt + g(s), \qquad \forall s \in [a,b],$$

for a given $g \in C([a,b])$.

12.1.1 Fixed Points

Thanks to the vector structure of V, the solution of vector equations can be reduced to the search of fixed points of suitable operators. For, given any function $\lambda : V \to \mathbb{R}$ with $\lambda(v) \neq 0$ for all $v \in V$, define $T_{w_0} : V \to V$ by

$$T_{w_0}(v) = \lambda(v)(T(v) - w_0) + v.$$

Then, $\overline{v} \in V$ solves the operator equation (12.4) if and only if $T_{w_0}(\overline{v}) = \overline{v}$, that is, if and only if $\overline{v}$ is a fixed point of T_{w_0}. To solve the operator equation (12.4) thus amounts to find the fixed points of the application T_{w_0}.

[3] Recall from Definition 90 that the product KN of operators is the composition $K \circ N$.

For this reason, the results that provide conditions for the existence of fixed points, the so-called fixed point theorems, play a key role in the solution of the vector equation (12.4). Motivated by this observation, we now study these theorems in some detail. In particular, we present two fundamental classes of fixed point theorems, originated respectively by the Banach Contraction Theorems of Banach (1922) and by the Brouwer Fixed Point Theorem of Brouwer (1912).

12.2 Banach Contraction Theorem

The first fundamental fixed point theorem we present is the Banach Contraction Mapping Theorem, due to Banach (1922).[4] It can be stated in any metric space (X, d) and is based on successive approximations. The notion of contraction is key in its derivation.

Definition 560 *A selfmap $T : X \to X$ defined on a metric space X is a contraction if there exists a scalar $0 < \alpha < 1$ such that*

$$d(Tx, Ty) \leq \alpha d(x, y), \qquad \forall x, y \in X. \tag{12.11}$$

That is, contractions are Lipschitz (and so continuous) functions with constant $\alpha \in (0, 1)$.[5]

To see the basic properties of contractions we need some notation. Given any selfmap $T : X \to X$, its second iterate $T \circ T : X \to X$ is denoted by T^2. More generally, T^n denotes the n-th iterate $T^n = T^{n-1} \circ T$, i.e., $T^n(x) = T(T^{n-1}(x))$ for all $x \in X$. We adopt the convention $T^0 = I$, that is, T^0 is the identity map.

Using the iterates T^n of a selfmap $T : X \to X$ we can construct a sequence $\{T^n(x_0)\}_n$ of points in X by starting from any initial point x_0 of X. These sequences will play a key role.

Lemma 561 *If exists, the fixed point $\overline{x}$ of a α-contraction $T : X \to X$ is unique and globally attracting, that is,*

$$\lim_{n\to\infty} d(T^n(x_0), \overline{x}) = 0, \qquad \forall x_0 \in X. \tag{12.12}$$

[4] It is sometimes called the Banach-Caccioppoli Contraction Mapping Theorem since it was independently discovered by Caccioppoli (1930).

[5] To ease notation, we call α-contractions the contractions with constant $\alpha \in (0, 1)$. Moreover, throughout the book, the terms functions and maps are synonyms (a selfmap is a map with the same domain and codomain).

By (12.12), any sequence $\{T^n(x_0)\}_n$ of iterates converges to the fixed point (if exists). Therefore, the iterates can be regarded as successive, better and better, approximations of the fixed point, which can thus be actually computed, when it exists. Uniqueness and global attractivity are thus truly remarkable properties of contractions' fixed points.

Proof We first prove uniqueness. Suppose x_1 and x_2 are fixed points. Then, for some $\alpha \in (0,1)$,

$$0 \leq d(x_1, x_2) = d(T(x_1), T(x_2)) \leq \alpha d(x_1, x_2),$$

and so $d(x_1, x_2) = 0$. This implies $x_1 = x_2$, as desired.

As to attractivity, let $x_0 \in X$. We have, $d(T(\overline{x}), T(x_0)) \leq \alpha d(\overline{x}, x_0)$, i.e., $d(\overline{x}, x_1) \leq \alpha d(\overline{x}, x_0)$. By iterating we get

$$d(T^n(x_0), \overline{x}) \leq \alpha^n d(\overline{x}, x_0), \qquad \forall n \geq 1,$$

and so $\lim_{n\to\infty} d(T^n(x_0), \overline{x}) = 0$. ■

We can now state and prove Banach's fundamental theorem.

Theorem 562 (Banach) *Let $T : X \to X$ be a α-contraction defined on a complete metric space X. Then, T has a unique and globally attracting fixed point $\overline{x}$. Moreover, the following error estimate holds:*

$$d(\overline{x}, T^n(x_0)) \leq \frac{\alpha^n}{1-\alpha} d(T(x_0), x_0). \tag{12.13}$$

Under the only mild assumption that X is complete, this theorem ensures the existence of fixed points of contractions, which, by Lemma 561, have the remarkable properties of being unique and globally attracting. The latter property, through the successive approximations $T^n(x_0)$ from any initial point $x_0 \in X$, gives a procedure to actually compute the unique fixed point, with an error upper bound given by (12.13).

As the use of the exponential supnorm in the proof of Proposition 581 illustrates, the true power of this result lies in the careful choice of the metric. In fact, the contractive nature of a selfmap $T : X \to X$ is not an intrinsic property of T but it entirely depends on the metric d. A clever choice of a metric may thus make contractive even selfmaps $T : X \to X$ that, *prima facie*, do not appear to be contractive.

Proof Let $\{x_n\}_n$ be the sequence of iterates defined by $x_n = T^n(x_0)$, given any initial $x_0 \in X$. We want to prove that this sequence is Cauchy. Fix n and take any $m \geq n$. We have $d(x_m, x_n) \leq \alpha d(x_{m-1}, x_{n-1})$. By iterating, we get

$$d(x_m, x_n) \leq \alpha^n d(x_{m-n}, x_0).$$

By the triangular inequality,

$$d(x_m, x_n) \leq \alpha^n \left[d(x_{m-n}, x_{m-n-1}) + d(x_{m-n-1}, x_{m-n-2}) + ... + d(x_1, x_0) \right].$$

Hence,

$$\begin{aligned} d(x_m, x_n) &\leq \alpha^n \left[\alpha^{m-n-1} + \alpha^{m-n-2} + ... + 1 \right] d(x_1, x_0) \qquad (12.14) \\ d(x_m, x_n) &\leq \frac{\alpha^n}{1-\alpha} d(x_1, x_0). \end{aligned}$$

Therefore, if $m, p \geq n$, we have

$$d(x_m, x_p) \leq d(x_m, x_n) + d(x_n, x_p) \leq 2\alpha^n (1-\alpha)^{-1} d(x_1, x_0).$$

As $\alpha^n \to 0$, this implies that the sequence is Cauchy. By the completeness of the space, $x_n \to \overline{x}$. Clearly, $\overline{x}$ is a fixed point. Since contractions are continuous, we have $T(x_n) \to T(\overline{x})$, i.e., $x_{n+1} \to T(\overline{x}) = \overline{x}$. By (12.14), letting $m \to \infty$, we get the estimate (12.13) and this completes the proof as the rest follows from Lemma 561. ■

The uniqueness and global attractivity of contractions' fixed points gives Banach's Theorem a central role in the solution of vector equations. In this regard it is useful to give the vector space version of this theorem. Because of the completeness hypothesis, the result holds in Banach spaces, i.e., in complete normed vector spaces (see Definition 333).

Corollary 563 *Let $T : V \to V$ be a α-contraction defined on a Banach space V. Then, T has a unique and globally attracting fixed point $\overline{v}$, that is,*

$$\lim_n \|T^n(v_0) - \overline{v}\| = 0, \qquad \forall v_0 \in V.$$

Moreover, the following estimate holds:

$$\|T^n(v_0) - \overline{v}\| \leq \frac{\alpha^n}{1-\alpha} \|T(v_0) - v_0\|.$$

A simple but useful principle of localization easily follows from Banach's Theorem 562.

Corollary 564 *Under the assumption of Theorem 562, if D is a closed subset of X that is invariant under T,[6] then the fixed point of T lies in D.*

Proof Given any point $x_0 \in D$, by assumption the sequence $T^n(x_0)$ belongs to D. As D is closed, $T^n(x_0) \to \overline{x} \in D$ by Theorem 254. ■

This corollary can be equivalently stated in the following version, which we report for later reference.

[6] That is, $x \in D$ implies $T(x) \in D$ (equivalently, $T(D) \subseteq D$).

Corollary 565 *Let $T : D \to D$ be a α-contraction defined on a closed subset D of a complete metric space X.*[7] *Then, T has a unique and globally attracting fixed point $\overline{x} \in D$.*

12.2.1 Variations on the theme

Consider the following application of Corollary 564. Let $x_0 \in X$ and consider the closed ball $\overline{B}_r(x_0)$ with center x_0 and radius

$$r = \frac{d(x_0, T(x_0))}{1-\alpha}.$$

Then, $\overline{B}_r(x_0)$ is invariant under T and the fixed point lies in $\overline{B}_r(x_0)$. For, given any $x \in \overline{B}_r(x_0)$,

$$\begin{aligned} d(T(x), x_0) &\leq d(T(x), T(x_0)) + d(T(x_0), x_0) \leq \alpha d(x, x_0) + d(T(x_0), x_0) \\ &\leq \left(\frac{\alpha}{1-\alpha} + 1\right) d(T(x_0), x_0) = \frac{1}{1-\alpha} d(T(x_0), x_0) = r, \end{aligned}$$

and so $T(x) \in \overline{B}_r(x_0)$. By Corollary 564, T has a fixed point in $\overline{B}_r(x_0)$.

This observation leads to the next powerful "local" formulation of the contraction theorem, for operators defined on open balls (an open domain, and so outside the scope of Corollary 564). It plays a fundamental role in many existence theorems.

Proposition 566 *Let X be a complete metric space. Given $x_0 \in X$, suppose:*

(i) $T : B_r(x_0) \to X$ is a α-contraction

(ii) $d(x_0, T(x_0)) < (1-\alpha) r$,

Then, T has a unique fixed point $\overline{x} \in B_r(x_0)$ that attracts all points of $B_r(x_0)$.[8]

Proof. By the triangular inequality, the open ball $B_r(x_0)$ is invariant under T. In fact, if $x \in B_r(x_0)$, then $d(x, x_0) < r$. Hence,

$$\begin{aligned} d(T(x), x_0) &\leq d(T(x), T(x_0)) + d(T(x_0), x_0) \leq \alpha d(x, x_0) + d(T(x_0), x_0) \\ &< \alpha r + (1-\alpha) r = r. \end{aligned}$$

It follows that in $B_r(x_0)$ there is at most a fixed point. Notice that $B_r(x_0)$ is not complete and therefore we cannot deduce the existence of the fixed point.

[7] That is, there is $\alpha \in (0,1)$ such that $d(Tx, Ty) \leq \alpha d(x, y)$ for each $x, y \in D$.

[8] That is, $\lim_n d(\overline{x}, T^n(x_0)) = 0$ for all $x_0 \in \overline{B}_r(x_0)$.

Consider any closed ball $\overline{B}_{r_1}(x_0)$ with

$$\frac{d(x_0, T(x_0))}{1-\alpha} \leq r_1 < r.$$

It is invariant, i.e., $T\left(\overline{B}_{r_1}(x_0)\right) \subseteq \overline{B}_{r_1}(x_0)$. In fact, from $x \in \overline{B}_{r_1}(x_0)$, it follows

$$\begin{aligned} d(T(x), x_0) &\leq d(T(x), T(x_0)) + d(T(x_0), x_0) \leq \alpha d(x, x_0) + d(T(x_0), x_0) \\ &\leq \alpha r_1 + (1-\alpha) r_1 = r_1. \end{aligned}$$

By Corollary 564, there exists a fixed point in the closed ball $\overline{B}_{r_1}(x_0)$, and so in $B_r(x_0)$, as desired. ■

The next theorem weakens the hypotheses of the Banach Theorem 562 by requiring that only some iterate T^n be a contraction, and not necessarily T itself as in Theorem 562.

Proposition 567 *Let $T : X \to X$ be a selfmap defined on a complete metric space X. If T^n is an α-contraction for some $n \geq 1$, then T has a unique and globally attracting fixed point.*

Proof. Set $Q = T^n$. By Theorem 562, Q has a unique fixed point $\overline{x}$. Notice that if x_1 is a fixed point of T, then x_1 is a fixed point of Q as well. We deduce that T has at most the fixed point $\overline{x}$. On the other hand,

$$T(\overline{x}) = (T \circ Q)(\overline{x}) = T^{n+1}(\overline{x}) = Q(T(\overline{x})).$$

Hence, $T(\overline{x})$ is a fixed point of Q, and so $T(\overline{x}) = \overline{x}$. It remains to prove that it is globally attracting. Set $\beta = \sqrt[n]{\alpha}$ and define

$$\rho(x, y) = d(x, y) + \frac{1}{\beta} d(T(x), T(y)) + \cdots + \frac{1}{\beta^{n-1}} d\left(T^{n-1}(x), T^{n-1}(y)\right). \quad (12.15)$$

It is immediately seen that ρ is a new metric on X. Moreover,

$$\begin{aligned} \rho(T(x), T(y)) &= d(T(x), T(y)) + \frac{1}{\beta} d\left(T^2(x), T^2(y)\right) + \cdots + \frac{1}{\beta^{n-1}} d(T^n(x), T^n(y)) \\ &\leq d(T(x), T(y)) + \frac{1}{\beta} d\left(T^2(x), T^2(y)\right) + \cdots + \beta d(x, y) \\ &= \beta \left[d(x, y) + \frac{1}{\beta} d(T(x), T(y)) + \cdots + \frac{1}{\beta^{n-1}} d\left(T^{n-1}(x), T^{n-1}(y)\right) \right] \\ &= \beta \rho(x, y). \end{aligned}$$

Hence, T is a contraction for the metric ρ. It follows that $\rho(T^n x_0, \overline{x}) \to 0$. On the other hand, $d(x, y) \leq \rho(x, y)$. Therefore, $d(T^n x_0, \overline{x}) \to 0$ for all $x_0 \in X$. ■

Even the weaker hypothesis of Proposition 567 that a selfmap $T : X \to X$ has a contracting iterate can be difficult to check. A useful criterion is condition (12.16) below, which controls the growth of Lipschitz constants of the iterates. Another well known unpleasant fact is that a selfmap may be contracting with respect to a metric and not with respect to another, equivalent, metric. These two facts are closely related and will be discussed in the next result. We first define a notion of equivalence for metrics, related to that for norms given in Definition 357.

Definition 568 *Two metrics d_1 and d_2 on X are (Lipschitz) equivalent if there exist $c_1, c_2 > 0$ such that $c_1 d_1(x,y) \leq d_2(x,y) \leq c_2 d_1(x,y)$ for all $x, y \in X$.*

Clearly, d_1 is complete if and only if d_2 is.

Proposition 569 *Let $T : X \to X$ be a selfmap defined on a metric space X. The following conditions are equivalent.*

(i) T is Lipschitz and T^n is an contraction for some $n \geq 1$;

(ii) there is some $k > 0$ and $\alpha \in (0,1)$ such that

$$d(T^n x, T^n y) \leq k\alpha^n d(x,y) \qquad \forall n \geq 0, \forall x, y \in X; \tag{12.16}$$

(iii) T is a contraction for a metric ρ that is equivalent to d;

(iv) there is a sequence k_n of Lipschitz constants of the iterates T^n such that[9]

$$\liminf_{n\to\infty} \sqrt[n]{k_n} < 1;$$

(v) there exists a sequence k_n of Lipschitz constants of the iterates such that

$$\lim_{n\to\infty} \sqrt[n]{k_n} < 1.$$

Proof. (ii) implies (i). From (12.16) it follows that $d(T(x), T(y)) \leq kd(x,y)$ and T is Lipschitz. Moreover, there is some n such that $k\alpha^n \in (0,1)$. Therefore, T^n is a contraction.

(i) implies (iii). As in the proof of Proposition 567, we can construct the metric ρ given by (12.15) and T is a contraction with respect to ρ. If T is Lipschitz with constant $k > 0$, from (12.15) it follows

$$d(x,y) \leq \rho(x,y) \leq \left(1 + \left(\frac{k}{\beta}\right) + ... + \left(\frac{k}{\beta}\right)^{n-1}\right) d(x,y)$$

[9] That is, $d(T^n x, T^n y) \leq k_n d(x,y)$ for each n.

and ρ is Lipschitz equivalent to d.

(iii) implies (ii). Assume that $\rho(T(x),T(y)) \leq \alpha\rho(x,y)$ with $\alpha \in (0,1)$ and where ρ is equivalent to d. We have $\rho(T^n(x),T^n(y)) \leq \alpha^n\rho(x,y)$. Therefore, if $Ad(x,y) \leq \rho(x,y) \leq Bd(x,y)$, we have

$$Ad(T^n(x),T^n(y)) \leq \rho(T^n(x),T^n(y)) \leq \alpha^n\rho(x,y) \leq \alpha^n Bd(x,y).$$

Hence, $d(T^n(x),T^n(y)) \leq \alpha^n(B/A)\,d(x,y)$.

(v) clearly implies (iv).

(iv) implies (i). There exists some $n \geq 1$ such that $\sqrt[n]{k_n} \leq \beta < 1$, so that $k_n \leq \beta^n$. Therefore, T^n is a contraction.

(ii) implies (v). From $k_n = k\alpha^n$, it follows $\sqrt[n]{k_n} = \alpha\sqrt[n]{k}$. Consequently, $\alpha\sqrt[n]{k} \to \alpha$, as $n \to \infty$. ■

Remarks. (i) By Propositions 567 and 569, if the sequence $\{k_n\}_n$ of the Lipschitz constants of the iterates T^n satisfies

$$\liminf_{n\to\infty} \sqrt[n]{k_n} = \alpha,$$

then for each $\varepsilon > 0$ there is an equivalent metric ρ such that $\rho(T(x),T(y)) \leq (\alpha+\varepsilon)\,\rho(x,y)$.

(ii) Even if X is a Banach space, and consequently $d(x,y) = \|x-y\|$, the new distance $\rho(x,y)$ given by (12.15) may not be induced by some norm. This is case, however, if T is linear. For,

$$\|x\|_\rho = \|x\| + \frac{1}{\beta}\|T(x)\| + ... + \frac{1}{\beta^{n-1}}\left\|T^{n-1}(x)\right\|$$

is the new norm associated with $\rho(x,y)$.

(iii) The condition that T is Lipschitz in (i) of Proposition 569 is essential. In general, if T^n is a contraction with $n > 1$ it does not follow not even that T is continuous. For instance, consider the discontinuous selfmap $T : \mathbb{R} \to \mathbb{R}$ given by $T(x) = (1/2)|x|$ if $|x| \leq 1$, and $T(x) = -1/2$ else. The second iterate is $T^2(x) = (1/4)|x|$ for $|x| \leq 1$ and $T^2(x) = 1/4$ else. Clearly, T^2 is a contraction.

The contraction condition $d(T(x),T(y)) \leq \alpha d(x,y)$ in (12.11) is stronger than the condition

$$d(T(x),T(y)) < d(x,y), \qquad \forall x \neq y, \tag{12.17}$$

which in general does not ensure the existence of fixed points. For instance, the function

$$f(x) = x + \frac{1}{1+x}$$

satisfies (12.17) on $\mathbb{R}_+$ and it does not have fixed points.

The next result, due to Edelstein (1962), shows that under a stronger condition on X – that is, compactness in place of completeness – selfmaps that satisfy the weaker contraction condition (12.17) have a fixed point.

Proposition 570 *If the selfmap $T : X \to X$ satisfies (12.17) on a compact metric space X, then it has a unique and globally attracting fixed point.*

Proof As T is continuous, the function $x \to d(x, T(x))$ is continuous on X. Consider $\inf_{x \in X} d(x, T(x))$. As X is compact, the infimum is attained. Hence, $d(x, T(x)) \geq d(\overline{x}, T(\overline{x}))$ for some $\overline{x} \in X$. Clearly $\overline{x}$ is a fixed point. Otherwise, $T(\overline{x}) \neq \overline{x}$ implies $d(T^2(\overline{x}), T(\overline{x})) < d(\overline{x}, T(\overline{x}))$, a contradiction.

We must prove that $\overline{x}$ attracts every point. Consider a trajectory $x_n = T^n(x_0)$. From $d(x_{n+1}, \overline{x}) \leq d(x_n, \overline{x})$ it follows that $d(x_n, \overline{x}) \to \inf_n d(x_n, \overline{x}) = \lambda$. If $\lambda = 0$ the claim is proved. Assume, by contradiction, $\lambda > 0$. By compactness (Corollary 277), there is a convergent subsequence $x_{n_k} \to y \in X$ and $d(y, \overline{x}) = \lambda > 0$. Hence, $d(T(y), \overline{x}) < \lambda$. On the other hand, from $x_{n_k} \to y$, it follows that $x_{n_k+1} \to Ty$ which implies $d(T(y), \overline{x}) = \lambda$, a contradiction. ■

Example 571 Consider an increasing and concave function $f : \mathbb{R}_+ \to \mathbb{R}$ with $f(0) > 0$ and $\lim_{x \to \infty} f'_+(x) < 1$. If $f'_+(0) < 1$, then f is a contraction. If $1 \leq f'_+(0) < \infty$, there is an iterate f^n that is a contraction. If $f'_+(0) = \infty$ (e.g., $f(x) = 1 + \sqrt{x}$), no iterate is a contraction though there is a globally attracting fixed point.

An additional remark is useful to understand the Definition 568 on Lipschitz equivalent metrics. In a general ground, two metrics d_1 and d_2 on X are said to be equivalent provided they induce the same topology on X. This is a much weaker condition than Lipschitz equivalence. This fact will be illustrated by an example.

Example 572 Consider in R the standard metric $d_1(x, y) = |x - y|$ induced by the norm $t \to |t|$. Now we introduce the new metric

$$d_2(x, y) = |\arctan(x) - \arctan(y)|.$$

We show the following properties:
i) $d_2(x, y)$ is a metric;
ii) d_1 and d_2 are equivalent;

iii) d_2 is incomplete and consequently the two metrics are not Lipschitz equivalent.
Point (i) is trivial to check. Observe that the property $d_2(x,y) = 0 \Longleftrightarrow x = y$ is a direct consequence of the fact that $\arctan(x)$ is strictly increasing. To prove that d_1 and d_2 are equivalent, it suffices to show that for any sequence x_n we have $d_1(x_n, \overline{x}) \to 0 \Longleftrightarrow d_2(x_n, \overline{x}) \to 0$. Clearly this follows easily from the continuity of $\arctan(x)$.
Consider any sequence $x_n \to +\infty$. It is Cauchy for the metric d_2. Actually, $\lim_{n\to\infty} \arctan(x_n) = \pi/2$. Given any $\varepsilon > 0$, there is an index n_0 such that $|\arctan(x_n) - \pi/2| < \varepsilon/2$. Hence, if $n, m \geq n_0$ we have

$$\begin{aligned} & |\arctan(x_n) - \arctan(x_m)| = \left|\arctan(x_n) - \frac{\pi}{2} + \frac{\pi}{2} - \arctan(x_m)\right| \\ \leq & \left|\arctan(x_n) - \frac{\pi}{2}\right| + \left|\frac{\pi}{2} - \arctan(x_m)\right| < \varepsilon, \end{aligned}$$

and (x_n) is Cauchy for d_2, though (x_n) does not convergent to any point.

12.2.2 Parametric Versions

We close by considering a parametric version of Banach Theorem, in which there is a family of contractions that depend on a parameter $\theta \in \Theta$.

Theorem 573 *Let $T : X \times \Theta \to X$ be a continuous map, where X is a complete metric space and Θ is a metric space. Assume that, for some $\alpha \in (0,1)$,*

$$d(T(x_1,\theta), T(x_2,\theta)) \leq \alpha d(x_1,x_2), \qquad \forall x_1, x_2 \in X, \forall \theta \in \Theta.$$

Then, there exists a unique function $x : \Theta \to X$ such that

$$T(x(\theta),\theta) = x(\theta), \qquad \forall \theta \in \Theta.$$

Moreover, $x : \Theta \to X$ is a continuous function and can be obtained by successive iterations.

Proof. By Theorem 562, there is a unique fixed point $x(\theta)$ for each $\theta \in \Theta$. Therefore, there is a unique function $x : \Theta \to X$ such that $T(x(\theta),\theta) = x(\theta)$ for all θ. It remains to prove that $x(\theta)$ is continuous. If the starting point $x_0(\theta)$ depends continuously on θ, then the first iterate $x_1(\theta) = T(x_0(\theta),\theta)$ is continuous. By iterating we have that $x(\theta)$ is the limit of continuous functions $x_n(\theta)$. It suffices to prove that this limit is uniform. From (12.13) we have

$$d(x(\theta), x_n(\theta)) \leq \frac{\alpha^n}{1-\alpha} d(x_1(\theta), x_0(\theta)) = \frac{\alpha^n}{1-\alpha} d(T(x_0(\theta),\theta), x_0(\theta)).$$

Taking $x_0(\theta) \equiv x_0$,

$$d(T(x_0(\theta),\theta),x_0(\theta)) = d(T(x_0,\theta),x_0)$$

which is bounded on a neighborhood of $\theta = \theta_0$. We conclude that the limit is uniform over a neighborhood of θ_0. Hence, $x(\theta)$ is continuous over that neighborhood. This argument is valid for any point θ_0 and this concludes the proof. ■

Also the local fixed point result formulated in Proposition 566 has a parametric counterpart.

Proposition 574 *Let $T : B_r(x_0) \times \Theta \to X$ a continuous map, where $B_r(x_0)$ is an open ball of a complete metric space X and Θ is a metric space. Assume that, for some $\alpha \in (0,1)$,*

$$d(T(x_1,\theta),T(x_2,\theta)) \leq \alpha d(x_1,x_2), \qquad \forall x_1,x_2 \in B_r(x_0), \forall \theta \in \Theta$$

If $d(x_0,T(x_0,\theta_0)) < (1-\alpha)r$, then there exists a continuous function $x : \Theta \to X$, defined over a neighborhood of θ_0, such that $x(\theta) \in B_r(x_0)$ and $T(x(\theta),\theta) = x(\theta)$.

Proof From the condition $d(x_0,T(x_0,\theta_0)) < (1-\alpha)r$ and the continuity of T, we have $d(x_0,T(x_0,\theta)) < (1-\alpha)r$ for all θ sufficiently close to θ_0. Therefore we can apply Proposition 566. ■

12.2.3 Contractions on Functions Spaces

A very useful contraction theorem holds in the Banach space $(B(X),\|\cdot\|_\infty)$, due to Blackwell (1965). Observe that in $B(X)$ there is a natural pointwise order $\leq$, where $f \leq g$ if and only if $f(x) \leq g(x)$ for all $x \in X$.[10]

Theorem 575 (Blackwell) *Suppose the selfmap $T : B(X) \to B(X)$ satisfies the following conditions:*

(i) T is monotone, i.e., $f \leq g \Longrightarrow Tf \leq Tg$;

(ii) T has the discounting property, i.e., there is $\beta \in (0,1)$ such that

$$T(f+c) \leq Tf + \beta c, \qquad \forall c \in \mathbb{R}_+.$$

Then, T is a β-contraction.

[10] In the statement, a scalar number λ denotes also the constant function $f \equiv \lambda$. For instance, the notation $f \leq \lambda$ means $f(x) \leq \lambda$ for all x.

Proof. Let $f, g \in B(X)$. Clearly, $f - g \leq \|f - g\|$. Hence, $f \leq g + \|f - g\|$ and so, by conditions (i) and (ii),

$$Tf \leq T(g + \|f - g\|) \leq Tg + \beta \|f - g\|.$$

which implies $Tf - Tg \leq \beta \|f - g\|$. Changing the role between f and g, we also have $Tg - Tf \leq \beta \|f - g\|$. Hence, $|Tf - Tg| \leq \beta \|f - g\|$, and so T is a β-contraction. ■

When X is a metric space, we may consider the subspace $C(X)$ of all bounded and continuous functions. It can be shown that $(C(X), \|\cdot\|_\infty)$ is also a Banach space (as a special case we have the Banach space $(C([0,1]), \|\cdot\|_\infty)$ of Chapter 7). It is easy to see that the Blackwell Contraction Theorem holds also for selfmaps on this Banach space.

12.3 Brouwer Fixed Point Theorem

The Brouwer Fixed Point Theorem, due to Brouwer (1912), and its generalizations are a second fundamental class of fixed points. In its simplest form it is an immediate consequence of the Intermediate Value Theorem.

Lemma 576 *A continuous selfmap $f : [0,1] \to [0,1]$ has a fixed point.*

Proof The result is obviously true if either $f(0) = 0$ or $f(1) = 1$. Suppose $f(0) > 0$ and $f(1) < 1$. Define the auxiliary function $g : [0,1] \to \mathbb{R}$ by $g(x) = x - f(x)$. Then, $g(0) < 0$ and $g(1) > 0$. Since g is continuous, by the Intermediate Value Theorem there exists $\overline{x} \in (0,1)$ such that $g(\overline{x}) = 0$. Hence, $f(\overline{x}) = \overline{x}$, and so $\overline{x}$ is a fixed point. ■

The generalization to $\mathbb{R}^n$ (and so to any finite dimensional vector space) of this lemma is Brouwer's result. It is an highly nontrivial generalization and, in fact, its proof is surprisingly complicated. For this reason we omit it and refer the interested reader to Border (1985) for a combinatorial proof and to Rogers (1980) for an analytic one.

Theorem 577 (Brouwer) *A continuous selfmap $f : K \to K$ defined on a convex compact subset K of a finite dimensional vector space has a fixed point.*

Let us compare the fixed point results of Banach and Brouwer. In particular, compare Theorem 577 with Corollary 565, which is the version of Banach's result that is best compared with Brouwer's result. Relative to Corollary 565, Brouwer's Theorem is less demanding on the selfmap, which is only required to be continuous. On the other hand, Brouwer's Theorem is more demanding on the domain, which is no longer any

closed subset of a complete metric space, but has to be a convex compact subset of a finite dimensional vector space. More importantly, Brouwer's Theorem is dramatically less informative on the fixed point since it only guarantees its existence, without any information about its uniqueness and attractivity, two key features of Banach's result. For this reason, Brouwer's Theorem can be viewed as a "surrogate" of Banach's one, which at least ensures the existence of fixed points for selfmaps that do not have the Lipschitzianity required by Banach's result.

Brouwer's Theorem has been extended to infinite dimensional spaces by Schauder (1930).

Theorem 578 (Schauder) *A continuous selfmap $f : K \to K$ defined on a convex compact subset K of a Banach space has a fixed point.*

Proof By Theorem 303, $f(K)$ is compact. Hence, it is totally bounded (see Theorem 281), i.e., given any n there exists a finite collection $\{w_i\}_{i=1}^{N_n} \subseteq f(K)$ of vectors such that

$$\min_{i=1,...,N_n} \|f(v) - w_i\| < \frac{1}{n}, \qquad \forall v \in f(K). \tag{12.18}$$

Let $M_n = span(w_1, ..., w_{N_n})$ be the vector subspace generated by the finite collection $\{w_i\}_{i=1}^{m}$, and let $\Delta_n = co(w_1, ..., w_{N_n})$ be its convex hull. Observe that M_n is finite dimensional and so, by Corollary 387, $\Delta_n \subseteq f(K)$ is a compact subset of M_n. Later in the proof this observation will make it possible to invoke the Brouwer Theorem.

For each w_i define the real valued function $\phi_i : f(K) \to \mathbb{R}$ by

$$\phi_i(v) = \max\left\{\frac{1}{n} - \|f(v) - w_i\|, 0\right\}, \qquad \forall v \in f(K).$$

The function ϕ_i is continuous (why?) and, by (12.18), is nonzero. Define the selfmap $g_n : \Delta_n \to \Delta_n$ by

$$g_n(v) = \frac{\sum_{i=1}^{N_n} \phi_i(v) w_i}{\sum_{i=1}^{N_n} \phi_i(v)}, \qquad \forall v \in \Delta_n.$$

The selfmap g_n is continuous. Moreover,

$$\begin{aligned} \|g_n(v) - f(v)\| &= \left\| \frac{\sum_{i=1}^{N_n} \phi_i(v) w_i}{\sum_{i=1}^{N_n} \phi_i(v)} - f(v) \right\| = \left\| \frac{\sum_{i=1}^{N_n} \phi_i(v) (w_i - f(v))}{\sum_{i=1}^{N_n} \phi_i(v)} \right\| \\ &= \frac{\left\| \sum_{i=1}^{N_n} \phi_i(v) (w_i - f(v)) \right\|}{\sum_{i=1}^{N_n} \phi_i(v)} \leq \frac{\sum_{i=1}^{N_n} \phi_i(v) \|w_i - f(v)\|}{\sum_{i=1}^{N_n} \phi_i(v)} \leq \frac{1}{n}. \end{aligned}$$

By the Brouwer Theorem 577, there exists $\overline{v}_n \in \Delta_n$ such that $g_n(\overline{v}_n) = \overline{v}_n$. Since $\Delta_n \subseteq f(K)$ for each n, by Theorem 275 there exists a subsequence $\{\overline{v}_{n_k}\}_k$ and $\overline{v} \in$

$f(K) \subseteq K$ such that $\lim_k \|\overline{v}_{n_k} - \overline{v}\| = 0$. Since

$$\begin{aligned} \|\overline{v}_{n_k} - f(\overline{v})\| &= \|g_{n_k}(\overline{v}_{n_k}) - f(\overline{v})\| \leq \|g_{n_k}(\overline{v}_{n_k}) - f(\overline{v}_{n_k})\| + \|f(\overline{v}_{n_k}) - f(\overline{v})\| \\ &\leq \frac{1}{n_k} + \|f(\overline{v}_{n_k}) - f(\overline{v})\|, \end{aligned}$$

we then conclude that $\lim_k \|\overline{v}_{n_k} - f(\overline{v})\| = 0$. In turn this implies $f(\overline{v}) = \overline{v}$, i.e., $\overline{v} \in K$ is a fixed point of f. ■

As we discussed in Section 7.5, compactness is a strong property in infinite dimensional spaces. This limits the scope of this version of Schauder's Theorem. The following generalization, due to Leray and Schauder (1934), has a wider applicability.

Theorem 579 (Leray-Schauder) *A continuous selfmap $f : C \to C$ defined on a convex closed and bounded subset C of a Banach space has a fixed point provided $f(C)$ is relatively compact (i.e., its closure $\overline{f(C)}$ is compact).*

Proof Since C is closed, we have $\overline{f(C)} \subseteq C$. Moreover, $f(C)$ is totally bounded, i.e., given any n there exists a finite collection $\{w_i\}_{i=1}^{N_n} \subseteq f(C)$ of vectors such that $\min_{i=1,\ldots,N_n} \|f(v) - w_i\| < 1/n$ for each $v \in f(C)$. A simple modification of the previous proof is now enough to prove the theorem. ■

12.4 Application I: Volterra Integral Equations

12.4.1 Existence

Consider the Volterra integral equation (12.5), namely

$$f(s) = \int_a^s \psi(s, t, f(t))\, dt + g(s), \qquad \forall s \in [a, b],$$

for a given $g \in C([a, b])$. Here the unknown function $f \in C([a, b])$ has to be determined.

We first give a general existence result, based on the Leray-Schauder Fixed Point Theorem 579, which shows that under very mild conditions on ψ the Volterra equations has a solution.

Proposition 580 *Suppose the function $\psi : [a, b] \times [a, b] \times \mathbb{R} \to \mathbb{R}$ is bounded and continuous. Then, given any $g \in C([a, b])$, the Volterra integral equation (12.5) has a solution $f \in C([a, b])$.*

Proof To ease the derivation, consider the homogeneous equation

$$f(s) = \int_a^s \phi(t, f(t))\, dt, \qquad \forall s \in [a,b]. \tag{12.19}$$

Let $G : C([a,b]) \to C([a,b])$ be the operator given by

$$G(f)(t) = \phi(t, f(t)), \qquad \forall (s,t) \in [a,b] \times [a,b].$$

Since ϕ is continuous and bounded,

$$\|G(f)\|_\infty \leq \max_{t \in [a,b]} \phi(t, f(t)), \qquad \forall f \in C([a,b]).$$

Hence, the image $G(C([a,b]))$ is bounded in $(C([a,b]), \|\cdot\|_\infty)$.

Let $S : C([a,b]) \to C^1([a,b])$ be the integral operator given by

$$S(f)(s) = \int_a^s f(x)\, dx, \qquad \forall s \in [a,b], \forall f \in C([a,b]).$$

Since $S(f)'(s) = f(s)$ for all $s \in [a,b]$ (see, e.g., Rudin (1976) p. 133), we have $\|S(f)'\|_\infty = \|f\|_\infty$. Moreover, $\|S(f)\|_\infty \leq \|f\|_\infty (b-a)$, so that

$$\|S(f)\|_1 \leq \|f\|_\infty \max\{b-a, 1\}.$$

Hence, $SG(C([a,b]))$ is bounded in $(C^1([a,b]), \|\cdot\|_1)$.

Finally, let $J : C^1([a,b]) \to C([a,b])$ be the natural embedding $J(f) = f$ for all $f \in C^1([a,b])$. Clearly, $JSG(C([a,b]))$ is also bounded in $(C^1([a,b]), \|\cdot\|_1)$. Hence, by Proposition 373, $JSG(C([a,b]))$ is relatively compact in $(C([a,b]), \|\cdot\|_\infty)$.

Consider the operator $T : C([a,b]) \to C([a,b])$ given by

$$T(f)(s) = \int_a^s \phi(t, f(t))\, dt, \qquad \forall f \in C([a,b]).$$

A function $f \in C([a,b])$ is a solution of the Volterra equation (12.19) if and only if is a fixed point of this operator. We have $T = JSG$, and so $T(C([a,b]))$ is relatively compact in $(C([a,b]), \|\cdot\|_\infty)$. By the Leray-Schauder Fixed Point Theorem 579, T has a fixed point. ■

12.4.2 Uniqueness

Though very general, Proposition 580 does not say anything about the uniqueness, let alone the attractivity, of the solutions. The next result, based on the Banach Contraction Theorem 562, addresses this issue by showing that, under a Lipschitz condition, the solution is indeed both unique and attractive.

Proposition 581 *Suppose the continuous function $\psi : [a,b] \times [a,b] \times \mathbb{R} \to \mathbb{R}$ is such that, for all $(s,t) \in [a,b] \times [a,b]$,*

$$|\psi(s,t,z_1) - \psi(s,t,z_2)| \leq K|z_1 - z_2|, \qquad \forall z_1, z_2 \in \mathbb{R}.$$

Then, given any $g \in C([a,b])$, the Volterra integral equation (12.5) has a unique solution $f \in C([a,b])$. In particular, the sequence $\{f_n\}_n \subseteq C([a,b])$ defined inductively by choosing $f_0 \in C([a,b])$ and setting

$$f_{n+1}(s) = \int_a^s \psi(s,t,f_n(t))\,dt + g(s), \qquad \forall s \in [a,b],$$

is such that $\|f_n - f\|_\infty \to 0$.

Proof Endow $C([a,b])$ with the exponential supnorm

$$\|f\|_{e^\infty} = \max_{t\in[a,b]} e^{-Kt}|f(t)|.$$

The norm $\|\cdot\|_{e^\infty}$ is complete and equivalent (in the sense of Definition 357) to the supnorm $\|\cdot\|_\infty$ (why?). The vector space $(C([a,b]), \|\cdot\|_{e^\infty})$ is thus Banach.

Consider the operator $T : C([a,b]) \to C([a,b])$ given by:

$$T(f)(s) = \int_a^s \psi(s,t,f(t))\,dt + g(s), \qquad \forall s \in [a,b].$$

A function $f \in C([a,b])$ is a solution of the operator equation (12.5) if and only if it is a fixed point of T. Given any $h_1, h_2 \in C([a,b])$, we have:

$$\begin{aligned}
& \|T(h_1) - T(h_2)\|_{e^\infty} \\
= & \left\| \left(\int_a^s \psi(s,t,h_1(t))\,dt + g(s) \right) - \left(\int_a^s \psi(s,t,h_2(t))\,dt + g(s) \right) \right\|_{e^\infty} \\
= & \left\| \int_a^s \psi(s,t,h_1(t))\,dt - \int_a^s \psi(s,t,h_2(t))\,dt \right\|_{e^\infty} \\
= & \max_{s\in[a,b]} e^{-Ks} \left| \int_a^s \psi(s,t,h_1(t))\,dt - \int_a^s \psi(s,t,h_2(t))\,dt \right| \\
\leq & \max_{s\in[a,b]} e^{-Ks} \int_a^s |\psi(s,t,h_1(t))\,dt - \psi(s,t,h_2(t))|\,dt
\end{aligned}$$

$$\begin{aligned}
&\leq K \max_{s\in[a,b]} e^{-Ks} \int_a^s |h_1(t) - h_2(t)|\, dt \\
&= K \max_{s\in[a,b]} e^{-Ks} \int_a^s e^{Kt} e^{-Kt} |h_1(t) - h_2(t)|\, dt \\
&\leq K \|h_1 - h_2\|_{e^\infty} \max_{s\in[a,b]} e^{-Ks} \int_a^s e^{Kt} dt \\
&= K \|h_1 - h_2\|_{e^\infty} \max_{s\in[a,b]} e^{-Ks} \frac{e^{Ks} - e^{Ka}}{K} \\
&= K \|h_1 - h_2\|_{e^\infty} \max_{s\in[a,b]} \frac{1 - e^{-K(s-a)}}{K} \\
&\leq \left(1 - e^{-K(b-a)}\right) \|h_1 - h_2\|_{e^\infty}.
\end{aligned}$$

Since $1 - e^{-K(b-a)} < 1$, we conclude that T is a contraction with respect to the exponential supnorm $\|\cdot\|_{e^\infty}$. The result then follows from Corollary 563, which also implies that the sequence $\{f_n\}_n \subseteq C([a,b])$ in the statement is such that $\|f_n - f\|_{e^\infty} \to 0$. In turn this implies $\|f_n - f\|_\infty \to 0$ since $\|\cdot\|_{e^\infty}$ and $\|\cdot\|_\infty$ are equivalent norms. ■

In this proof it clearly emerges the importance of the choice of the norm (and so of the metric) that makes a selfmap on a Banach space a contraction. If in the proof we consider the supnorm $\|\cdot\|_\infty$ instead of the exponential supnorm $\|\cdot\|_{e^\infty}$, we get

$$\|T(h_1) - T(h_2)\|_\infty \leq K(b-a) \|h_1 - h_2\|_\infty,$$

instead of

$$\|T(h_1) - T(h_2)\|_{e^\infty} \leq \left(1 - e^{-K(b-a)}\right) \|h_1 - h_2\|_{e^\infty}.$$

Hence, T is a contraction with respect to the supnorm $\|\cdot\|_\infty$ if $K(b-a) < 1$, and so here Corollary 563 ensures existence of a solution only on small enough domains $[a,b]$. In contrast, the use of the exponential supnorm $\|\cdot\|_{e^\infty}$ eliminates any such restriction and allowed us to prove the existence of a solution on any interval $[a,b]$.

In sum, the choice of suitable norms (and so metrics) is a key issue in solving operator equations through contractions.

12.4.3 Systems of Volterra Integral Equations

Systems of Volterra integral equations play an important role in applications and they can be solved with a natural generalization of the arguments we just used for the scalar case.

Let $C([a,b], \mathbb{R}^n)$ be the vector space of all continuous functions $f = (f_1, ..., f_n) : [a,b] \to \mathbb{R}^n$, often called *curves*. The vector space $C([a,b], \mathbb{R}^n)$ becomes a Banach space (why?) once endowed with the supnorm

$$\|f\|_\infty = \sup_{i=1,...,n} \|f_i\|_\infty = \sup_{(i,t)\in\{1,...,n\}\times[a,b]} |f_i(t)|.$$

Given a continuous function $\psi = (\psi_1, ..., \psi_n) : [a,b] \times [a,b] \times \mathbb{R}^n \to \mathbb{R}^n$, consider the Volterra integral equation (12.5), namely for each $i = 1, ..., n$,

$$f_i(s) = \int_a^s \psi(s, t, f_1(t), ..., f_n(t))\, dt + g_i(s), \qquad \forall s \in [a,b],$$

for a given $g \in C([a,b], \mathbb{R}^n)$. Here the unknown function $f \in C([a,b], \mathbb{R}^n)$ has to be determined.

We can now state and prove the extension of Proposition 581 to systems of integral equations. Though the proof follows the same lines of that of Proposition 581, for the sake of completeness we report it.

Proposition 582 *Suppose the continuous function $\phi : [a,b] \times [a,b] \times \mathbb{R}^n \to \mathbb{R}^n$ is such that, for all $(s,t) \in [a,b] \times [a,b]$,*

$$\|\phi(s, t, z_1) - \phi(s, t, z_2)\|_\infty \leq K \|z_1 - z_2\|_\infty, \qquad \forall z_1, z_2 \in \mathbb{R}^n.$$

Then, given any $g \in C([a,b])$, the Volterra integral equation (12.5) has a unique solution $f \in C([a,b], \mathbb{R}^n)$. In particular, the sequence $\{f^k\}_k \subseteq C([a,b], \mathbb{R}^n)$ defined inductively by choosing $f^0 \in C([a,b], \mathbb{R}^n)$ and setting, for each $i = 1, ..., n$,

$$f_i^{k+1}(s) = \int_a^s \psi\left(s, t, f_1^k(t), ..., f_n^k(t)\right) dt + g_i(s), \qquad \forall s \in [a,b],$$

is such that $\left\|f^k - f\right\|_\infty \to 0$.[11]

Proof Endow $C([a,b], \mathbb{R}^n)$ with the exponential supnorm

$$\|f\|_{e^\infty} = \sup_{i=1,...,n} \|f_i\|_{e^\infty} = \sup_{(i,t) \in \{1,...,n\} \times [a,b]} e^{-Kt} |f_i(t)|.$$

Here as well the norm $\|\cdot\|_{e^\infty}$ is complete and equivalent (in the sense of Definition 357) to the supnorm $\|\cdot\|_\infty$ (why?). The vector space $(C([a,b]), \|\cdot\|_{e^\infty})$ is Banach.

Consider the operator $T : C([a,b], \mathbb{R}^n) \to C([a,b], \mathbb{R}^n)$ given by:

$$T_i(f)(s) = \int_a^s \psi_i(s, t, f_1(t), ..., f_n(t))\, dt + g_i(s), \qquad \forall s \in [a,b],$$

for each $i = 1, ..., n$. A function $f \in C([a,b], \mathbb{R}^n)$ is a solution of the operator equation (12.5) if and only if it is a fixed point of T. Given any $h^1, h^2 \in C([a,b], \mathbb{R}^n)$, we have:

$$h^1(t) = \left(h_1^1(t), ..., h_n^1(t)\right) \in \mathbb{R}^n \qquad \text{and} \qquad h^2(t) = \left(h_1^2(t), ..., h_n^2(t)\right) \in \mathbb{R}^n.$$

[11] We write f^k in place of f_n to ease notation since here n denotes the dimensionality of the system and f_i^n is the i-th component of the function $f^n = (f_1^n, ..., f_n^n) : [a,b] \to \mathbb{R}^n$.

Hence,

$$\begin{aligned}
&\left\|T\left(h_1\right)-T\left(h_2\right)\right\|_{e^\infty}\\
=&\sup_{i=1,\ldots,n}\left\|\left(\int_a^s\psi_i\left(s,t,h^1\left(t\right)\right)dt+g_i\left(s\right)\right)-\left(\int_a^s\psi_i\left(s,t,h^2\left(t\right)\right)dt+g_i\left(s\right)\right)\right\|_{e^\infty}\\
=&\sup_{i=1,\ldots,n}\left\|\int_a^s\psi_i\left(s,t,h^1\left(t\right)\right)dt-\int_a^s\psi_i\left(s,t,h^2\left(t\right)\right)dt\right\|_{e^\infty}\\
=&\sup_{i=1,\ldots,n}\left(\max_{s\in[a,b]}e^{-Ks}\left|\int_a^s\psi_i\left(s,t,h^1\left(t\right)\right)dt-\int_a^s\psi_i\left(s,t,h^2\left(t\right)\right)dt\right|\right)\\
\leq&\sup_{i=1,\ldots,n}\left(\max_{s\in[a,b]}e^{-Ks}\int_a^s\left|\psi_i\left(s,t,h^1\left(t\right)\right)dt-\psi_i\left(s,t,h^2\left(t\right)\right)\right|dt\right)\\
\leq&K\left(\max_{s\in[a,b]}e^{-Ks}\int_a^s\left\|h^1\left(t\right)-h^2\left(t\right)\right\|_{e^\infty}dt\right)\\
=&K\max_{s\in[a,b]}e^{-Ks}\int_a^s e^{Kt}e^{-Kt}\left\|h^1\left(t\right)-h^2\left(t\right)\right\|_{e^\infty}dt\\
\leq&K\left\|h^1\left(t\right)-h^2\left(t\right)\right\|_{e^\infty}\max_{s\in[a,b]}e^{-Ks}\int_a^s e^{Kt}dt\\
=&K\left\|h^1\left(t\right)-h^2\left(t\right)\right\|_{e^\infty}\max_{s\in[a,b]}e^{-Ks}\frac{e^{Ks}-e^{Ka}}{K}\\
=&K\left\|h^1\left(t\right)-h^2\left(t\right)\right\|_{e^\infty}\max_{s\in[a,b]}\frac{1-e^{-K(s-a)}}{K}\\
\leq&\left(1-e^{-K(b-a)}\right)\left\|h^1\left(t\right)-h^2\left(t\right)\right\|_{e^\infty}.
\end{aligned}$$

Since $1-e^{-K(b-a)}<1$, we conclude that T is a contraction with respect to the exponential supnorm $\left\|\cdot\right\|_{e^\infty}$. The result then follows from Corollary 563, which also implies that the sequence $\left\{f^n\right\}_n\subseteq C\left(\left[a,b\right],\mathbb{R}^n\right)$ in the statement is such that $\left\|f^n-f\right\|_{e^\infty}\to 0$. In turn this implies $\left\|f^n-f\right\|_\infty\to 0$ since $\left\|\cdot\right\|_{e^\infty}$ and $\left\|\cdot\right\|_\infty$ are equivalent norms. ■

12.4.4 A Volterra-Hammerstein Equation

We now give an illustration of the importance of Proposition 567 by analyzing a Volterra-Hammerstein integral equation (12.10). Specifically, consider the simple Nemitski operator $N:C\left(\left[a,b\right]\right)\to C\left(\left[a,b\right]\right)$ given by:

$$N\left(f\right)\left(t\right)=\lambda f\left(t\right),\qquad\forall t\in\left[a,b\right],\forall f\in C\left(\left[a,b\right]\right),$$

where $\lambda\neq 0$. In this case equation (12.10) becomes

$$f\left(s\right)=\lambda\int_a^s K\left(s,t\right)f\left(t\right)dt+g\left(s\right),\qquad\forall s\in\left[a,b\right]. \tag{12.20}$$

The unknown continuous function $f\in C\left(\left[a,b\right]\right)$ has to be determined. It is not restrictive to consider the homogeneous case

$$f\left(s\right)=\lambda\int_a^s K\left(s,t\right)f\left(t\right)dt,\qquad\forall s\in\left[a,b\right], \tag{12.21}$$

where we set $g = \mathbf{0}$. Let $T : C([a,b]) \to C([a,b])$ be given by

$$T(f)(s) = \lambda \int_a^s K(s,t) f(t)\, dt, \qquad \forall s \in [a,b].$$

The operator T is linear and allows to write (12.21) as:

$$f = T(f).$$

That is, f is a solution of equation (12.21) if and only if it is a fixed point of T.

Since T is linear, $f = \mathbf{0}$ is a fixed point (and so a solution). To find nonzero fixed points, endow $C([a,b])$ with its supnorm $\|\cdot\|_\infty$; that is, consider the Banach space $(C([a,b]), \|\cdot\|_\infty)$.

Lemma 583 *For n large enough the iterate T^n is a contraction with respect to the supnorm $\|\cdot\|_\infty$.*

Proof The iterate T^n is linear and so, by (7.6), is enough to prove that, for n large enough, we have $\|T^n\| < 1$. Denote by M the maximum value

$$M = \max_{(s,t)\in[a,b]\times[a,b]} |K(s,t)|.$$

We have, for all $s \in [a,b]$ and all $f \in C([a,b])$,

$$\begin{aligned} |Tf(s)| &= |\lambda| \left| \int_a^s K(s,t) f(t)\, dt \right| \leq |\lambda| M \int_a^s |f(t)|\, dt \leq |\lambda| M \|f\|_\infty \int_a^s dt \\ &= |\lambda| M \|f\|_\infty (s-a) \leq |\lambda| M \|f\|_\infty (b-a). \end{aligned}$$

Hence, $\|T(f)\|_\infty \leq |\lambda| M \|f\|_\infty (b-a)$ and so

$$\|T\| = \sup\left\{ \frac{\|T(f)\|_\infty}{\|f\|_\infty} : f \neq \mathbf{0} \right\} \leq |\lambda| M (b-a).$$

Consider the second iterate T^2. We have

$$\begin{aligned} \left|T^2 f(s)\right| &\leq |\lambda| M \int_a^s |Tf(t)|\, dt \leq \lambda^2 M^2 \|f\|_\infty \int_a^s (t-a)\, dt \\ &= \frac{\lambda^2 M^2 \|f\|_\infty}{2} (s-a)^2 \leq \frac{\lambda^2 M^2 \|f\|_\infty}{2} (b-a)^2. \end{aligned}$$

Therefore, $\|T^2(f)\|_\infty \leq 2^{-1}\lambda^2 M^2 \|f\|_\infty (b-a)^2$, and so

$$\left\|T^2\right\| = \sup\left\{ \frac{\|T^2(f)\|_\infty}{\|f\|_\infty} : f \neq \mathbf{0} \right\} \leq \frac{|\lambda|^2 M^2 (b-a)^2}{2}.$$

Iterating this procedure, we get

$$\|T^n\| \leq \frac{|\lambda|^n M^n (b-a)^n}{n!}.$$

On the other hand,

$$\left[\frac{|\lambda|^n M^n (b-a)^n}{n!}\right]^{1/n} = \frac{|\lambda| M (b-a)}{\sqrt[n]{n!}} \to 0$$

as $n \to \infty$. Hence, $\|T^n\| < 1$ for n large enough, as desired. ■

By Proposition 567, we thus have the following result.

Proposition 584 *The Volterra-Hammerstein integral equation (12.20) has a unique solution $f \in C([a,b])$ for all $g \in C([a,b])$. In particular, the sequence $\{f_n\}_n \subseteq C([a,b])$ defined inductively by choosing $f_0 \in C([a,b])$ and setting*

$$f_{n+1}(s) = \lambda \int_a^s K(s,t) f_n(t)\, dt + g(s), \qquad \forall s \in [a,b],$$

is such that $\|f_n - f\|_\infty \to 0$.

12.5 Application II: Differential Equations

Let J be any closed, bounded or unbounded, interval of $\mathbb{R}$. That is, either $J = [a,b]$, with $a, b \in \mathbb{R}$, or $J = (-\infty, \infty)$. Given a function $f : J \times \mathbb{R} \to \mathbb{R}$ and a pair $(t_0, x_0) \in J \times \mathbb{R}$, a solution of the initial value problem[12]

$$\begin{aligned} x'(t) &= f(t,x) \\ x(t_0) &= x_0 \end{aligned} \tag{12.22}$$

is a differentiable function $x : J \to \mathbb{R}$ such that

$$\begin{aligned} x'(t) &= f(t, x(t)), \qquad \forall t \in J, \\ x(t_0) &= x_0. \end{aligned}$$

In this section we present two classic results, originally due to Giuseppe Peano and Charles Picard, about the solution of this initial value problem. These results rely on a simple but crucial "duality" between initial value problems and suitable Volterra integral equations. This allows to reduce the solution of initial value problems to that of the dual Volterra equations, thus making possible to solve the initial value problems by using the results on Volterra equations established in the previous section.

[12] Initial value problems are also called Cauchy problems.

To see in a nutshell the duality, suppose $J = [a, b]$ and consider the initial value problem

$$\begin{aligned} x'(t) &= f(t, x) \\ x(a) &= x_0 \end{aligned} \tag{12.23}$$

and the Volterra equation

$$x(t) = \int_a^t f(z, x(z))\, dz + x_0, \qquad \forall t \in [a, b], \tag{12.24}$$

where we set $\psi(s, t, z) = f(t, z)$ in (12.5). A function $x : [a, b] \to \mathbb{R}$ is a solution of this Volterra integral equation (12.24) if and only if is also a solution of the initial value problem (12.23). In fact, if $x : [a, b] \to \mathbb{R}$ is a solution of this integral equation, then

$$x'(t) = f(t, x(t)), \qquad \forall t \in [a, b]$$

by a basic Calculus result (see, e.g., Rudin (1976) p. 133). Conversely, if $x : [a, b] \to \mathbb{R}$ is a solution of the initial value problem (12.23), then by integrating we get (12.24).

12.5.1 Peano's Theorem

We begin with Peano's Theorem, originally due to Peano (1886) and (1890), a classic existence result that establishes the existence of solutions of initial value problems under very mild conditions. It is based on Proposition 580, which in turn was based on the Leray-Schauder Fixed Point Theorem 579.

Theorem 585 (Peano) *Suppose $f : J \times \mathbb{R} \to \mathbb{R}$ is continuous and bounded. Then, the initial value problem (12.22) has a solution $x : J \to \mathbb{R}$ for all $(t_0, x_0) \in J \times \mathbb{R}$.*

Proof It is easy to see that a function $x : [t_0, b] \to \mathbb{R}$ the Volterra equation

$$x(t) = \int_{t_0}^t f(z, x(z))\, dz + x_0, \qquad \forall t \in [t_0, b],$$

if and only if is also a solution of the initial value problem

$$\begin{aligned} x'(t) &= f(t, x), \qquad \forall t \in [t_0, b], \\ x(t_0) &= x_0 \end{aligned}$$

By Theorem 580, such a solution $x : [t_0, b] \to \mathbb{R}$ exists. A similar argument, based on Exercise 13.0.79, shows that a solution $x : [a, t_0] \to \mathbb{R}$ exists (cf. the proof of Theorem 586). To find a solution $x : [a, b] \to \mathbb{R}$ is enough to paste together these solutions as in the proof of Theorem 586. ■

12.5.2 Picard's Theorem

We turn now to Picard's Theorem, originally due to Picard (1890), which shows that under a Lipschitz condition the initial value problem has a unique and attractive solution. The proof shows that the solution of the initial value problem 12.22 can be reduced to the solution of suitable Volterra integral equations, which in turn can be solved via Proposition 581.

Theorem 586 (Picard) *Suppose $f : J \times \mathbb{R} \to \mathbb{R}$ is continuous and satisfies the Lipschitz condition*

$$|f(t, s_1) - f(t, s_2)| \leq K |s_1 - s_2|, \qquad \forall s_1, s_2 \in \mathbb{R}, \forall t \in J.$$

Then, the initial value problem (12.22) has a unique solution $x : J \to \mathbb{R}$ for all $(t_0, x_0) \in J \times \mathbb{R}$.

Proof Suppose first that $J = [a, b]$. Let $(t_0, x_0) \in [a, b] \times \mathbb{R}$. First consider the Volterra equation

$$x(t) = \int_{t_0}^{t} f(z, x(z))\, dz + x_0, \qquad \forall t \in [t_0, b], \tag{12.25}$$

where we set $\psi(s, t, z) = f(t, z)$ in (12.5). It is easy to see that a function $x : [t_0, b] \to \mathbb{R}$ is a solution of this integral equation if and only if is also a solution of the initial value problem

$$\begin{aligned} x'(t) &= f(t, x), \qquad \forall t \in [t_0, b], \\ x(t_0) &= x_0 \end{aligned} \tag{12.26}$$

Since, by Proposition 581, the Volterra equation (12.25) has a unique solution, we conclude that there is a unique $x^* : [t_0, b] \to \mathbb{R}$ that solves the initial value problem (12.26).

Now consider the backward Volterra integral equation

$$x(t) = \int_{t}^{t_0} f(z, x(z))\, dz + x_0, \qquad \forall t \in [a, t_0], \tag{12.27}$$

Here as well it is easy to check that a function $x : [a, t_0] \to \mathbb{R}$ is a solution of this integral equation if and only if is also a solution of the initial value problem

$$\begin{aligned} x'(t) &= f(t, x), \qquad \forall t \in [a, t_0], \\ x(t_0) &= x_0 \end{aligned} \tag{12.28}$$

Hence, there is a unique $x_* : [a, t_0] \to \mathbb{R}$ that solves the initial value problem (12.28) since, by Exercise 13.0.78, the backward Volterra equation (12.27) has a unique solution.

To complete the proof for the case $J = [a, b]$, it is enough to paste together the "partial" solutions x^* and x_* by defining $x : [a, b] \to \mathbb{R}$ as follows:

$$x(t) = \begin{cases} x^*(t) & \text{if } t \in [t_0, b] \\ x_*(t) & \text{if } t \in [a, t_0] \end{cases}$$

The function x is the unique solution of the initial value problem (12.22).

It remains to consider the case $J = (-\infty, \infty)$. Clearly,

$$\bigcup_{n \geq 1} \left[t_0 - \frac{1}{n}, t_0 + \frac{1}{n} \right] = \mathbb{R}.$$

On each interval $[t_0 - 1/n, t_0 + 1/n]$ there is a unique solution $x_n : [t_0 - 1/n, t_0 + 1/n] \to \mathbb{R}$ such that $x_n(t_0) = x_0$ and $x_n'(t) = f(t, x_n(t))$ for all $t \in [t_0 - 1/n, t_0 + 1/n]$. Because of uniqueness, we have

$$x_n(t) = x_{n+1}(t), \qquad \forall t \in \left[t_0 - \frac{1}{n}, t_0 + \frac{1}{n} \right].$$

Hence, we can define $x : \mathbb{R} \to \mathbb{R}$ by $x(t) = x_n(t)$ for each $t \in [t_0 - 1/n, t_0 + 1/n]$. Clearly, $x(t_0) = x_0$ and $x'(t) = f(t, x(t))$ for all $t \in \mathbb{R}$, and so $x : \mathbb{R} \to \mathbb{R}$ is a solution of the initial value problem (12.22). It is also the unique solution since any other solution has to agree with each x_n on $[t_0 - 1/n, t_0 + 1/n]$. ■

In this proof the solution $x : J \to \mathbb{R}$ of the initial value problem has been reduced to that of a suitable Volterra integral equation. As a result, the solution can be determined via successive approximations. For, by Proposition 581 and Exercise 13.0.78, the sequence $\{\varphi_n\}_n \subseteq C(J)$ defined inductively by choosing $\varphi_0 \in C(J)$ and setting

$$\varphi_{n+1}(s) = \begin{cases} \int_{t_0}^{s} f(t, \varphi_n(t))\, dt + x_0 & \text{if } s \in J \cap [t_0, \infty), \\ \int_{s}^{t_0} f(t, \varphi_n(t))\, dt + x_0 & \text{if } s \in J \cap (\infty, t_0]. \end{cases}$$

is such that $\|\varphi_n - x\|_\infty \to 0$. Hence, the sequence $\{\varphi_n\}_n$ uniformly converges to the solution $x : J \to \mathbb{R}$ of the initial value problem (12.22). The function φ_n are often called *successive Picard approximations*.

Example 587 Consider the initial value problem

$$\begin{aligned} x'(t) &= x(t), \qquad \forall t \geq 0, \\ x(0) &= 1. \end{aligned} \tag{12.29}$$

Let $\varphi_0 : [0,\infty) \to \mathbb{R}$ be the constant $\varphi_0(t) = 1$ for all $t \geq 0$. Then, the successive Picard approximations are:

$$\begin{aligned}
\varphi_1(s) &= \int_0^s \varphi_0(t)\,dt + 1 = 1 + s, \qquad \forall s \geq 0, \\
\varphi_2(s) &= \int_0^s \varphi_1(t)\,dt + 1 = \int_0^s (1+t)\,dt + 1 = 1 + s + \frac{s^2}{2}, \qquad \forall s \geq 0 \\
&\cdots \\
\varphi_n(s) &= \int_0^s \varphi_{n-1}(t)\,dt + 1 = 1 + s + \frac{s^2}{2} + \cdots + \frac{s^n}{n!}, \qquad \forall s \geq 0
\end{aligned}$$

Hence,

$$x(t) = \lim_n \varphi_n(t) = \lim_n \left(1 + t + \frac{t^2}{2} + \cdots + \frac{t^n}{n!}\right) = e^t, \qquad \forall t \geq 0$$

and we conclude that e^t is the unique solution of the initial value problem (12.29). ▲

Example 588 The initial value problem

$$\begin{aligned}
x'(t) &= \sqrt{|x(t)|}, \qquad \forall t \in \mathbb{R}, \\
x(0) &= 0,
\end{aligned}$$

has a continuum of solutions, besides $x = 0$. In fact, all functions

$$x(t) = \begin{cases} -\frac{1}{2}(t+b)^2 & t \leq -b \\ 0 & -b \leq t \leq a \\ \frac{1}{2}(t-a)^2 & t \geq a \end{cases}$$

with $a, b \geq 0$, are easily seen to be solutions.

The continuous function $x \longmapsto \sqrt{|x|}$ is not Lipschitz at $x = 0$. This proves that the Lipschitz condition in Picard's Theorem cannot be omitted.

Lavrentieff (1925) (see also Hartman, 1963) gave an example of a continuous function $f(x,t)$ defined on the rectangle $[0,1] \times [0,1]$ such that, for any $(x_0, t_0) \in (0,1) \times (0,1)$, the initial value problem

$$\begin{aligned}
x'(t) &= f(x,t), \\
x(t_0) &= x_0
\end{aligned}$$

has more than one solution on every interval $[t_0, t_0 + \varepsilon]$ and $[t_0 - \varepsilon, t_0]$ for small enough ε. ▲

12.5.3 Systems of Differential Equations

The proof of Picard's Theorem was based on the solution of a suitable Volterra integral equation, as ensured by Proposition 581. In section 12.4 we also solved systems of Volterra integral equations thanks to Proposition 582. By using this generalization of Proposition 581 we can solve systems of differential equations.

Specifically, given a function $f = (f_1, ..., f_n) : J \times \mathbb{R}^n \to \mathbb{R}^n$ and a pair $(t_0, x_0) \in J \times \mathbb{R}^n$, a solution of the initial value problem[13]

$$\begin{aligned} x_i'(t) &= f_i(t, x_1, ..., x_n), \qquad \forall i = 1, ..., n \\ x_i(t_0) &= x_{0i}, \qquad \forall i = 1, ..., n \end{aligned} \tag{12.30}$$

is a differentiable function $x : J \to \mathbb{R}^n$ such that, for each $i = 1, ..., n$,

$$\begin{aligned} x_i'(t) &= f(t, x(t)), \qquad \forall t \in J, \\ x_i(t_0) &= x_{0i}. \end{aligned}$$

Proposition 589 *Suppose $f : J \times \mathbb{R}^n \to \mathbb{R}^n$ is continuous and satisfies the Lipschitz condition*

$$\|f(t, s_1) - f(t, s_2)\|_\infty \leq K \|s_1 - s_2\|_\infty, \qquad \forall s_1, s_2 \in \mathbb{R}^n, \forall t \in J.$$

Then, the initial value problem (12.30) has a unique solution $x : J \to \mathbb{R}^n$ for all $(t_0, x_0) \in J \times \mathbb{R}^n$.

Proof We sketch the proof, which is basically that of Theorem 586 with Proposition 582 in place of Proposition 581.

Suppose first that $J = [a, b]$. Let $(t_0, x_0) \in [a, b] \times \mathbb{R}^n$. Consider the system of Volterra equations

$$x_i(t) = \int_{t_0}^{t} f_i(z, x_1(z), ..., x_n(z))\, dz + x_{0i}, \qquad \forall t \in [t_0, b],$$

with $i = 1, ..., n$. A function $x : [t_0, b] \to \mathbb{R}^n$ is a solution of this system of integral equations if and only if, for each $i = 1, ..., n$,

$$\begin{aligned} x_i'(t) &= f(t, x(t)), \qquad \forall t \in [t_0, b], \\ x_i(t_0) &= x_{0i}. \end{aligned}$$

We can now proceed as in the proof of the previous theorem, with Proposition 582 in place of Proposition 581, to prove the theorem. ■

Finally, it is easy to see that also the solutions of systems of differential equations can be uniformly approximated by successive Picard approximations. We omit the details for brevity.

[13] Initial value problems are also called Cauchy problems.

Chapter 13

Exercises

Exercise 13.0.1 *Two vector subspaces W_1 and W_2 of V are said to be disjoint if $W_1 \cap W_2 = \{\vec{0}\}$, i.e., if their intersection is the trivial vector subspace $\left\{\vec{0}\right\}$. For example, the horizontal and vertical axes are two disjoint subspaces in $\mathbb{R}^2$. A polynomial g is even if $g(-x) = g(x)$ for each $x \in \mathbb{R}$, while g is odd if $g(-x) = -g(x)$ for each $x \in \mathbb{R}$. Let W_1 be the set of all even polynomials $g \in \mathcal{P}$ and W_2 the set of all odd polynomials. Show that:*

(a) W_1 and W_2 are two vector subspaces of $\mathcal{P}$;

(b) W_1 and W_2 are disjoint;

(c) $W_1 + W_2 = \mathcal{P}$, where $W_1 + W_2 = \{w_1 + w_2 : \ w_1 \in W_1 \text{ and } w_2 \in W_2\}$.

Sol.: (a) To verify whether W_1 is a vector subspace we must check that any linear combination of two even polynomials is still an even polynomial. In fact, for $\alpha, \beta \in \mathbb{R}$ and $f, g \in W_1$ we have

$$(\alpha f + \beta g)(-x) = \alpha f(-x) + \beta g(-x) = \alpha f(x) + \beta g(x) = (\alpha f + \beta g)(x)$$

for each $x \in \mathbb{R}$, i.e., $\alpha f + \beta g \in W_1$. Similarly, if $\alpha, \beta \in \mathbb{R}$ and $f, g \in W_2$ we have

$$(\alpha f + \beta g)(-x) = \alpha f(-x) + \beta g(-x) = -\alpha f(x) - \beta g(x) = -(\alpha f + \beta g)(x)$$

for each $x \in \mathbb{R}$, i.e., $\alpha f + \beta g \in W_2$.

(b) We need to check that the null polynomial is the only polynomial that is both even and odd. Let $g \in W_1 \cap W_2$. For each $x \in \mathbb{R}$ we have $g(-x) = g(x)$ since $g \in W_1$ and $g(-x) = -g(x)$ since $g \in W_2$. Hence, $g(x) = -g(x)$ for each $x \in \mathbb{R}$, which implies $g(x) = 0$ for each $x \in \mathbb{R}$.

(c) The equality $W_1 + W_2 = \mathcal{P}$ means that every polynomial can be written as a sum of an even polynomial and an odd polynomial. For, if $g \in \mathcal{P}$ we can write

$$g(x) = A(x) + B(x) \text{ con } \begin{cases} A(x) = \frac{1}{2}(g(x) + g(-x)), \\ B(x) = \frac{1}{2}(g(x) - g(-x)). \end{cases}$$

Let us check that $A \in W_1$ and $B \in W_2$:

$$\begin{cases} A(-x) = \frac{1}{2}(g(-x) + g(x)) = A(x), \\ B(-x) = \frac{1}{2}(g(-x) - g(x)). \end{cases}$$

(It is easy to see that A is given by the sum of all terms of even degree in g, and B is given by the sum of the odd degree terms). □

Exercise 13.0.2 *Let $W = \{x \in \mathbb{R}^3 : x_1 + x_2 + x_3 = 0\}$. Verify if W is a vector subspace of $\mathbb{R}^3$.*

Sol.: Siano $\alpha, \beta \in \mathbb{R}$, $x = (x_1, x_2, x_3), y = (y_1, y_2, y_3) \in \mathbb{R}^3$. Dobbiamo verificare che $\alpha x + \beta y \in W$. Abbiamo

$$\alpha x + \beta y = (\alpha x_1 + \beta y_1, \alpha x_2 + \beta y_2, \alpha x_3 + \beta y_3),$$

e la somma delle tre componenti di $\alpha x + \beta y$ è

$$\begin{aligned} & (\alpha x_1 + \beta y_1) + (\alpha x_2 + \beta y_2) + (\alpha x_3 + \beta y_3) \\ = \; & \alpha(\underbrace{x_1 + x_2 + x_3}_{=0,\ siccome\ x \in W}) + \beta(\underbrace{y_1 + y_2 + y_3}_{=0,\ siccome\ y \in W}) = 0. \end{aligned}$$

Quindi W è un sottospazio vettoriale di $\mathbb{R}^3$. □

Exercise 13.0.3 *Sia*

$$W = \{x \in \mathbb{R}^3 : x_1^2 + x_2^2 + x_3^2 \leq 1\}.$$

Verificare se W è un sottospazio vettoriale di $\mathbb{R}^3$.

Sol.: W non è un sottospazio vettoriale di $\mathbb{R}^3$ dal momento che $(1, 0, 0) \in W$ e $(0, 1, 0) \in W$ ma la loro somma $(1, 1, 0)$ non è in W: il quadrato della sua distanza dall'origine vale 2 (> 1). □

Exercise 13.0.4 *Sia W il sottoinsieme di $\mathcal{P}$ composto dai polinomi che hanno come coefficienti numeri interi. Verificare se W è un sottospazio vettoriale di $\mathcal{P}$.*

Sol.: W non è un sottospazio vettoriale di $\mathcal{P}$, dal momento che, ad esempio, il polinomio x è in W ma il suo prodotto con il numero reale $\frac{1}{2}$ (non intero!) non sta in W. □

Exercise 13.0.5 *Mostrare che i vettori* $(1,0,0),\ (0,1,0),\ (0,0,1)$*, e* $(1,1,1)$ *sono linearmente dipendenti, ma ogni loro sottoinsieme di tre elementi è linearmente indipendente.*

Sol.: La dipendenza lineare dei quattro vettori segue osservando che $(1,1,1)$ è uguale alla somma degli altri tre vettori. Alternativamente si può studiare l'equazione

$$\alpha(1,0,0)+\beta(0,1,0)+\gamma(0,0,1)+\delta(1,1,1)=(0,0,0),$$

equivalente al sistema

$$\begin{cases} \alpha+\delta=0, \\ \beta+\delta=0, \\ \gamma+\delta=0. \end{cases}$$

il quale, considerato δ come paramentro, ha la soluzione $\alpha=-\delta,\beta=-\delta,\gamma=-\delta$. Questo dimostra che i quattro vettori dati sono linearmente dipendenti.

Per quanto riguarda la seconda affermazione nel testo dell'esercizio, dovremmo ragionare in maniera analoga a quanto precede considerando però soltanto insiemi di tre vettori. Lo facciamo considerando i vettori $(1,0,0),\ (0,1,0),\ (1,1,1)$, lasciando al lettore gli altri tre casi. L'equazione e il relativo sistema sono ora dati rispettivamente da

$$\alpha(1,0,0)+\beta(0,1,0)+\gamma(1,1,1)=(0,0,0),$$

$$\begin{cases} \alpha+\gamma=0, \\ \beta+\gamma=0, \\ \gamma=0, \end{cases}$$

da cui risulta l'unica soluzione banale $\alpha=\beta=\gamma=0$. Questo dimostra l'indipendenza lineare dei tre vettori. □

Exercise 13.0.6 *Mostrare che i vettori* $f_1(x)=1,\ f_2(x)=x,\ f_3(x)=x^2,\ f_4(x)=1+x+x^2$ *sono linearmente dipendenti, ma ogni loro sottoinsieme di tre elementi è linearmente indipendente.*

Sol.: La dipendenza lineare di f_1,f_2,f_3,f_4 segue dal fatto che $f_4=f_1+f_2+f_3$. Alternativamente si può studiare l'equazione

$$\alpha+\beta x+\gamma x^2+\delta(1+x+x^2)=0 \quad \text{per ogni } x\in\mathbb{R},$$

che implica

$$\begin{cases} \alpha+\delta=0, \\ \beta+\delta=0, \\ \gamma+\delta=0, \end{cases}$$

che è il primo sistema incontrato nella risoluzione dell'Esercizio 5. Pertanto tutte le conclusioni di tale esercizio possono ripetersi senza alcun cambiamento. Anche per il caso di tre vettori, è immediato ricondursi all'Esercizio 5. □

Exercise 13.0.7 *Sia M l'insieme di tutti gli $x \in \mathbb{R}^4$ tali che*

$$\begin{cases} 2x_1 + x_2 - 2x_3 + 2x_4 = 0, \\ x_1 - x_2 - 2x_3 - 4x_4 = 0, \\ x_1 - 2x_2 - 2x_3 - 10x_4 = 0. \end{cases}$$

Verificare che M è un sottospazio vettoriale di $\mathbb{R}^4$ e dare una descrizione esplicita di M.

Sol.: Ciascuna delle tre equazioni che compongono il sistema definisce un sottospazio vettoriale di $\mathbb{R}^4$ (verifica identica a quella dell'esercizio 2), pertanto M sarà un sottospazio di $\mathbb{R}^4$ essendo l'intersezione dei tre sottospazi suddetti. Per darne una descrizione esplicita risolviamo il sistema. Ricavando x_1 dalla terza equazione e sostituendo nelle altre due otteniamo

$$\begin{cases} 5x_2 + 2x_3 + 22x_4 = 0, \\ x_2 + 6x_4 = 0, \\ x_1 = 2x_2 + 2x_3 + 10x_4, \end{cases}$$

ricavando x_2 dalla seconda equazione e sostituendo nelle altre due otteniamo

$$\begin{cases} 2x_3 - 8x_4 = 0, \\ x_2 = -6x_4, \\ x_1 = 2x_3 - 2x_4, \end{cases}$$

ricavando x_3 dalla prima e sostituendo nell'ultima, si ha

$$\begin{cases} x_3 = 4x_4, \\ x_2 = -6x_4, \\ x_1 = 6x_4. \end{cases}$$

Le soluzioni del sistema (ovvero i vettori di M) sono tutte e sole le quaterne del tipo $(6x_4, -6x_4, 4x_4, x_4)$ al variare del parametro reale x_4. Pertanto

$$M = \{x_4(6, -6, 4, 1), x_4 \in \mathbb{R}\}$$

risulta il sottospazio vettoriale di $\mathbb{R}^4$ generato dal vettore $(6, -6, 4, 1)$. □

Exercise 13.0.8 *Mostrare che ogni sottoinsieme di un insieme S di vettori linearmente indipendenti è a sua volta linearmente indipendente.*

Sol.: Per assurdo, supponiamo che esista un sottoinsieme $S' \subseteq S$ linearmente dipendente. Questo significa che esistono alcuni vettori di S', diciamo $v_1, \ldots, v_r$, e dei numeri reali $\alpha_1, \ldots, \alpha_r$ non tutti nulli tali che

$$\alpha_1 \mathrm{v}_1 + \cdots + \alpha_r \mathrm{v}_r = \vec{0}.$$

Ma, poiché $S' \subseteq S$, questa stessa uguaglianza vale in S, cioè considerando i vettori $v_1, \ldots, v_r$ come vettori di S, cosa che contraddice l'indipendenza lineare di S. Questo conclude la dimostrazione. □

Exercise 13.0.9 *(a)Per quali valori di* α *i vettori* $(1,1,1)$ *e* $(1,\alpha,\alpha^2)$ *sono linearmente indipendenti?*

(b) Per quali valori di α *i vettori* $(0,1,\alpha)$, $(\alpha,0,1)$, $(\alpha,1,1+\alpha)$ *sono una base di* $\mathbb{R}^3$*?*

Sol.: (a) Occorre determinare gli $\alpha \in \mathbb{R}$ per i quali l'equazione

$$x(1,1,1) + y(1,\alpha,\alpha^2) = (0,0,0),$$

o equivalentemente il sistema

$$\begin{cases} x + y = 0, \\ x + \alpha y = 0, \\ x + \alpha^2 y = 0, \end{cases}$$

ha come unica soluzione quella banale: $x = 0,\ y = 0$. Ricavando y dalla prima equazione e sostituendo nelle altre otteniamo

$$\begin{cases} y = -x, \\ (1-\alpha)\, x = 0, \\ (1-\alpha^2)\, x = 0. \end{cases}$$

Pertanto, per $\alpha \neq 1$ abbiamo la sola soluzione $x = 0,\ y = 0$, mentre per $\alpha = 1$ si hanno come soluzioni tutte le coppie del tipo $(x, -x)$, ossia $y = -x$. Quindi i due vettori $(1,1,1)$ e $(1,\alpha,\alpha^2)$ sono linearmente indipendenti per $\alpha \neq 1$, mentre sono linearmente dipendenti per $\alpha = 1$ (in realtà, per $\alpha = 1$ essi sono proprio uguali).

(b) I tre vettori assegnati non sono una base di $\mathbb{R}^3$ per alcun valore di α, dal momento che non sono linearmente indipendenti: risulta

$$(\alpha, 1, 1+\alpha) = (0,1,\alpha) + (\alpha,0,1).$$

Alternativamente è possibile controllare la dipendenza lineare dei tre vettori risolvendo l'equazione

$$x(0,1,\alpha) + y(\alpha,0,1) + z(\alpha,1,1+\alpha) = (0,0,0),$$

vale a dire il sistema

$$\begin{cases} \alpha y + \alpha z = 0, \\ x + z = 0, \\ \alpha x + y + (1+\alpha) z = 0. \end{cases}$$

Esso è equivalente al sistema

$$\begin{cases} \alpha (y - x) = 0, \\ z = -x, \\ y - x = 0, \end{cases}$$

che ammette le infinite soluzioni $y = x,\ z = -x$ qualunque sia il numero reale α. □

Exercise 13.0.10 *Siano W_1 e W_2 due sottospazi vettoriali di V con $\dim(W_1) = \dim(W_2)$. Mostrare che se W_1 è incluso in W_2, allora $W_1 = W_2$.*

Sol.: Sia $n = \dim(W_1) = \dim(W_2)$ e $S = \{e_1, \ldots, e_n\}$ una base di W_1. Allora S costituisce un sistema di n vettori linearmente indipendenti dello spazio vettoriale W_2 (perché $W_1 \subseteq W_2$) il quale ha dimensione n. Pertanto S sarà una base anche per W_2 e

$$W_1 = span(S) = W_2.$$

□

Exercise 13.0.11 *Siano x, y e z tre vettori tali che $x + y + z = \vec{0}$. Mostrare che $span(x, y) = span(y, z)$.*

Sol.: Poiché $z = -x - y$, risulta $z \in span(x, y)$, potendosi esprimere z come una combinazione lineare di x e y. Di qui segue che $span(x, y) = span(x, y, z)$. Analogamente, da $x = -y - z$ risulta $x \in span(y, z)$, e quindi $span(y, z) = span(x, y, z)$. Allora la tesi è immediata:

$$span(x, y) = span(x, y, z) = span(y, z).$$

□

Exercise 13.0.12 *Define a linear functional $L : \mathbb{R}^3 \to \mathbb{R}$ such that $L(x) = L(y) = 0$, where $x = (1, 1, 1)$ and $y = (1, 1, -1)$.*

Sol.: Si può prendere come L il funzionale identicamente nullo. Se si vuole determinarne uno non identicamente nullo, basta osservare che ogni funzionale $L \in (\mathbb{R}^3)'$ è della forma $L(x) = ax_1 + bx_2 + cx_3$ per qualche $a, b, c \in \mathbb{R}$ (ossia $L(x) = x \cdot x^*$ con $x^* = (a, b, c)$). Imponendo la condizione $L(x) = L(y) = 0$, otteniamo

$$\begin{cases} a + b + c = 0 \\ a + b - c = 0 \end{cases} \Longrightarrow \begin{cases} a + b = -c \\ -2c = 0 \end{cases}$$

da cui $b = -a,\ c = 0$. Pertanto i funzionali che soddisfano le condizioni richieste sono quelli del tipo $L(x) = ax_1 - ax_2$, con a un qualunque numero reale. Si può ad esempio prendere $L(x) = x_1 - x_2$(che corrisponde ad $a = 1$). □

Exercise 13.0.13 *Provare che i vettori* $x = (1,1,1), y = (1,1,-1)$ *e* $z = (1,-1,-1)$ *formano una base di* $\mathbb{R}^3$. *Se* $\{L_1, L_2, L_3\}$ *è la corrispondente base duale di* $(\mathbb{R}^3)'$, *trovare* $L_1(\tilde{x})$, $L_2(\tilde{x})$ *e* $L_3(\tilde{x})$, *dove* $\tilde{x} = (0,1,0)$.

Sol.: I tre vettori dati sono una base di $\mathbb{R}^3$ perché sono linearmente indipendenti. Per verificarlo ci si riconduce a risolvere il sistema

$$\begin{cases} \alpha+\beta+\gamma=0 \\ \alpha+\beta-\gamma=0 \\ \alpha-\beta-\gamma=0 \end{cases} \Longrightarrow \begin{cases} 2\alpha=0 \\ 2\beta=0 \\ \gamma=\alpha-\beta \end{cases} \Longrightarrow \begin{cases} \alpha=0 \\ \beta=0 \\ \gamma=0 \end{cases}$$

che ammette appunto la sola soluzione banale.

La base duale di $\{x, y, z\}$, indicata con $\{L_1, L_2, L_3\}$, è definita da

$$\begin{cases} L_1(x)=1, & L_1(y)=L_1(z)=0, \\ L_2(y)=1, & L_2(x)=L_2(z)=0, \\ L_3(z)=1, & L_3(x)=L_3(y)=0. \end{cases}$$

Sfruttando queste condizioni e la linearità di L_1, L_2, L_3 è facile calcolare i valori che questi tre funzionali assumono in $\tilde{x} = (0,1,0)$: esprimiamo $\tilde{x}$ come combinazione lineare della base x, y, z:

$$(0,1,0) = \alpha(1,1,1) + \beta(1,1,-1) + \gamma(1,-1,-1),$$

ovvero

$$\begin{cases} \alpha+\beta+\gamma=0 \\ \alpha+\beta-\gamma=1 \\ \alpha-\beta-\gamma=0 \end{cases} \Longrightarrow \begin{cases} 2\alpha=0 \\ 2\beta=1 \\ \gamma=\alpha-\beta \end{cases} \Longrightarrow \begin{cases} \alpha=0 \\ \beta=\frac{1}{2} \\ \gamma=-\frac{1}{2} \end{cases}$$

Pertanto $\tilde{x} = \frac{1}{2}y - \frac{1}{2}z$, e quindi, siccome L_1, L_2, L_3 sono lineari, abbiamo

$$\begin{cases} L_1(\widetilde{x}) = L_1\left(\frac{1}{2}y - \frac{1}{2}z\right) = \frac{1}{2}L_1(y) - \frac{1}{2}L_1(z) = 0 - 0 = 0, \\ L_2(\widetilde{x}) = L_2\left(\frac{1}{2}y - \frac{1}{2}z\right) = \frac{1}{2}L_2(y) - \frac{1}{2}L_2(z) = \frac{1}{2} - 0 = \frac{1}{2}, \\ L_3(\widetilde{x}) = L_3\left(\frac{1}{2}y - \frac{1}{2}z\right) = \frac{1}{2}L_3(y) - \frac{1}{2}L_3(z) = 0 - \frac{1}{2} = -\frac{1}{2}. \end{cases}$$

□

Exercise 13.0.14 *Sia* $L : V \to \mathbb{R}$ *un funzionale lineare definito su uno spazio vettoriale* V. *Mostrare che*

$$W = \{v \in V : L(v) = 0\}$$

è un sottospazio vettoriale di V. *Qual è la sua dimensione?*

Sol.: W è un sottospazio vettoriale perché, presi arbitrariamente $\alpha, \beta \in \mathbb{R}$ e $w_1, w_2 \in W$, risulta $\alpha w_1 + \beta w_2 \in W$ dato che

$$L(\alpha w_1 + \beta w_2) = \alpha L(w_1) + \beta L(w_2) = 0 + 0 = 0.$$

Quanto alla sua dimensione, possono presentarsi due casi.

(a) Se Lè il funzionale identicamente nullo ($L = \vec{0}_{V'}$), allora $W = V$ e quindi $\dim W = \dim V$.

(b) Se L non è identicamente nullo (ovvero $L \neq \vec{0}_{V'}$), verifichiamo che $\dim W = \dim V - 1$. Infatti, in tal caso esiste un vettore z tale che $L(z) \neq 0$. Indichiamo con Z lo spazio vettoriale generato da z, ossia

$$Z = span(z) = \{\alpha z; \alpha \in \mathbb{R}\}.$$

Lo spazio Z ha dimensione 1. Ora notiamo che $W \cap Z = \{\vec{0}\}$. Infatti, se imponiamo che un vettore di Z, e quindi del tipo αz, appartenga anche a W, cioé $L(\alpha z) = 0$, otteniamo $\alpha L(z) = 0$ (perché Lè lineare) e quindi $\alpha = 0$ (perché $L(z) \neq 0$)). Pertanto, se dimostriamo che $W + Z = V$ abbiamo concluso, perché allora avremmo che W e Z sono complementari e quindi $\dim W = \dim V - \dim Z = \dim V - 1$. Basta quindi dimostrare che $W + Z = V$, in altri termini, che ogni $v \in V$ si può scrivere come $v = \alpha z + w$ per opportuni $\alpha \in \mathbb{R}$ e $w \in W$. Ricordiamo che $w \in W$ significa $L(w) = 0$, ovvero $L(v - \alpha z) = 0$ e quindi $L(v) - \alpha L(z) = 0$; risolvendo rispetto ad α otteniamo $\alpha = L(v)/L(z)$ (ricordare che $L(z) \neq 0$). Quindi, comunque preso $v \in V$, scegliamo α in questo modo e poi $w = v - \alpha z$. Allora abbiamo che $w \in W$e $v = \alpha z + w$. □

Exercise 13.0.15 *Nell'Esercizio 14, sia* $V = \mathbb{R}^3$ *e sia* $L(x) = x_1 + x_2 + x_3$ *per ogni* $x \in \mathbb{R}^3$. *Trovare una base del sottospazio* $W = \{x \in \mathbb{R}^3 : L(x) = 0\}$.

Sol.: Risolvendo rispetto a x_1 l'equazione

$$x_1 + x_2 + x_3 = 0,$$

che definisce W, si vede come W sia formato da tutti i vettori $x = (x_1, x_2, x_3)$ tali che $x_1 = -x_2 - x_3$. Quindi,

$$\begin{aligned} W &= \{(-x_2 - x_3, x_2, x_3) : x_2, x_3 \in \mathbb{R}\} \\ &= \{x_2(-1, 1, 0) + x_3(-1, 0, 1) : x_2, x_3 \in \mathbb{R}\}. \end{aligned}$$

Pertanto, i due vettori $(-1, 1, 0)$ e $(-1, 0, 1)$ generano lo spazio W. Siccome sono linearmente indipendenti (verificarlo!), essi costituiscono una base per W.[1] □

[1] Si noti che abbiamo trovato, in particolare, che $\dim(W) = 2$, in accordo con quanto dimostrato nell'Esercizio 14.

Exercise 13.0.16 *I vettori $x = (1, 1)$ e $y = (0, 1)$ formano una base di $\mathbb{R}^2$. Trovare il funzionale lineare $L : \mathbb{R}^2 \to \mathbb{R}$ tale che $L(x) = 3$ e $L(y) = -2$.*

Sol.: Come nell'esercizio 12, bisogna determinare una coppia (a, b) di numeri reali tale che il funzionale definito da $L(x) = ax_1 + bx_2$ soddisfi alle condizioni $L(x) = 3$ e $L(y) = -2$:

$$\begin{cases} L(x) = a + b = 3, \\ L(y) = b = -2 \end{cases} \Longrightarrow \begin{cases} a = 5 \\ b = -2 \end{cases}$$

Il funzionale lineare cercato è dunque dato da $L(x) = 5x_1 - 2x_2$. □

Exercise 13.0.17 *Show that the functional $L : \mathbb{R}^3 \to \mathbb{R}$ given by $L(x) = 2x_1 - 3x_2 + 4x_3$ is linear.*

Sol.: Può scriversi $L(x) = x \cdot x^*$, con $x^* = (2, -3, 4)$ e quindi la linearità di L segue dalla linearità del prodotto scalare rispetto al primo argomento. Alternativamente, si può fare una verifica diretta: per ogni $\alpha, \beta \in \mathbb{R}$ e $x = (x_1, x_2, x_3)$, $y = (y_1, y_2, y_3) \in \mathbb{R}^3$ risulta

$$\begin{aligned} L(\alpha x + \beta y) &= L((\alpha x_1 + \beta y_1, \alpha x_2 + \beta y_2, \alpha x_3 + \beta y_3)) \\ &= 2(\alpha x_1 + \beta y_1) - 3(\alpha x_2 + \beta y_2) + 4(\alpha x_3 + \beta y_3) \\ &= \alpha(2x_1 - 3x_2 + 4x_3) + \beta(2y_1 - 3y_2 + 4y_3) \\ &= \alpha L(x) + \beta L(y). \end{aligned}$$

□

Exercise 13.0.18 *Verificare se il funzionale $L : \mathbb{R}^2 \to \mathbb{R}$ definito da $L(x) = x_1 \cdot x_2$ è lineare.*

Sol.: L non è lineare dal momento che, ad esempio, prendendo $x = (1, 1)$risulta

$$L(2x) = L((2, 2)) = 2 \cdot 2 = 4,$$

mentre se L fosse lineare dovrebbe risultare

$$L(2x) = 2L(x) = 2L((1, 1)) = 2(1 \cdot 1) = 2.$$

□

Exercise 13.0.19 *Sia V uno spazio vettoriale. Mostrare che due vettori x e y di V sono linearmente dipendenti se e solo se si verifica uno dei seguenti fatti: $x = \vec{0}$ oppure esiste un $\alpha \in \mathbb{R}$ tale che $y = \alpha x$.*

Sol.: "Se". Se $x = \vec{0}$ risulta $1 \cdot x + 0 \cdot y = \vec{0}$ per cui x e y sono linearmente dipendenti. Allo stesso modo, da $y = \alpha x$ segue $\alpha \cdot x - 1 \cdot y = \vec{0}$, ovvero la dipendenza lineare di x e y.

"Solo se". Se x e y sono linearmente dipendenti allora sarà $\alpha x + \beta y = \vec{0}$ con $(\alpha, \beta) \neq (0, 0)$. Se è $\beta \neq 0$ sarà $y = -\frac{\alpha}{\beta} x$. Altrimenti, se $\beta = 0$, dovrà essere $\alpha \neq 0$ e allora (poiché $\alpha x + \beta y = \vec{0}$) segue $x = \vec{0}$. □

Exercise 13.0.20 *Si determini per quali valori dell'intero $n \geq 0$ l'insieme $M_n = \{g \in \mathcal{P} : g \equiv 0$ oppure $\deg(g) = n\}$ è un sottospazio vettoriale di $\mathcal{P}$.*

Sol.: Per $n = 0$ abbiamo M_0 che è un sottospazio vettoriale di $\mathcal{P}$, essendo l'insieme dei polinomi costanti, ossia $span(1)$.

Per $n > 0$ M_n non è un sottospazio vettoriale di $\mathcal{P}$ dal momento che i due polinomi $x^n + x^{n-1}$ e $-x^n$ sono in M_n ma la loro somma, ossia x^{n-1}, non è in M_n. □

Exercise 13.0.21 *Dire quali dei seguenti sottoinsiemi di $\mathcal{P}$ sono dei sottospazi vettoriali. Per ciascuno di quelli che non risultano dei sottospazi si determini il più piccolo sottospazio di $\mathcal{P}$ che li contiene.*

(a) $A = \{g \in \mathcal{P} : g(0) = 0\}$;

(b) $B = \{g \in \mathcal{P} : g(0) = 2\}$;

(c) $C = \{g \in \mathcal{P} : g(0) = g(1)\}$;

(d) $D = \{g \in \mathcal{P} : g$ *è divisibile per* $x^2\}$;

(e) $E = \{g \in \mathcal{P} : g(x) = \alpha x + \beta x^2;\ \alpha, \beta \in \mathbb{R}\}$;

(f) $F = \{g \in \mathcal{P} : g \equiv 0$ *oppure* $\deg(g) \geq 2\}$.

Sol.: (a) A è un sottospazio dato che se due polinomi si annullano in $x = 0$ ogni loro combinazione lineare si annullerà anch'essa in $x = 0$.

(b) B non è un sottospazio perché, ad esempio, non contiene il polinomio identicamente nullo. Il più piccolo sottospazio che contiene B è $span(B)$. È facile vedere che $span(B) = \mathcal{P}$. Infatti, preso arbitrariamente $g \in \mathcal{P}$, se $g(0) = 0$ scrivendo $g(x) = (g(x) + 2) - (2)$ si vede che g può esprimersi come differenza di due polinomi in B, mentre se $g(0) \neq 0$ si può scrivere

$$g(x) = \frac{g(0)}{2} \left(\frac{2}{g(0)} g(x) \right)$$

e quindi g è dato dal prodotto di un numero reale per un polinomio in B. In ogni caso g risulta una combinazione lineare di polinomi in B, e come tale sta in $span(B)$.

(c) C è un sottospazio. Infatti se $f, g \in C$ e $\alpha, \beta \in \mathbb{R}$ risulta $\alpha f + \beta g \in C$ dal momento che

$$(\alpha f + \beta g)(0) = \alpha f(0) + \beta g(0) = \alpha f(1) + \beta g(1) = (\alpha f + \beta g)(1).$$

(d) D è un sottospazio: esso è costituito dai polinomi che mancano del termine noto e del termine di primo grado, ed è evidente che una qualunque combinazione lineare di due polinomi in D è ancora in D.

(e) E è il sottospazio vettoriale generato dai polinomi x e x^2, ossia $E = span(x, x^2)$.

(f) F non è un sottospazio dal momento che i polinomi $x^2 + x$ e $-x^2$sono in F, ma $(x^2 + x) + (-x^2) = x \notin F$. Ovviamente, come in (b), il più piccolo sottospazio che contiene F è $span(F)$. Verifichiamo che $span(F) = \mathcal{P}$. $span(F)$ è un sottospazio vettoriale che contiene tutti i polinomi di F, in particolare i polinomi $x^2 + x$ e $-x^2$, e con loro la loro somma x. Allo stesso modo, contenendo i polinomi x^2+x+1 e $-x^2-x$, dovrà contenere anche il polinomio costante 1. Così, oltre ai polinomi x^n con $n \geq 2$, contiene anche i polinomi 1 e x, per cui contiene qualunque polinomio. □

Exercise 13.0.22 *Si consideri il sottospazio vettoriale di* $\mathbb{R}^3$ *dato da*

$$M = span((1, 1, 0), (1, 2, 3)).$$

Determinare un'equazione lineare omogenea (ossia del tipo $ax + by + cz = 0$*) che sia soddisfatta da tutti e soli i vettori di* M*.*

Sol.: I vettori di M sono tutti e solo gli (x, y, z) che si possono scrivere nella forma

$$(x, y, z) = \alpha\,(1, 1, 0) + \beta\,(1, 2, 3)$$

per qualche $\alpha, \beta \in \mathbb{R}$, ossia

$$\begin{cases} x = \alpha + \beta, \\ y = \alpha + 2\beta, \\ z = 3\beta. \end{cases}$$

Sottraendo la prima equazione dalla seconda, si trova $\beta = y - x$. Sostituendo questa espressione di β nella terza equazione, troviamo $z = 3y - 3x$, ossia $3x - 3y + z = 0$.

Viceversa, se (x, y, z) soddisfa l'equazione $3x - 3y + z = 0$, si verifica subito che risulta

$$(x, y, z) = (2x - y)(1, 1, 0) + (y - x)(1, 2, 3)$$

e quindi $(x, y, z) \in M$.[2] □

[2]Usando la notazione precedente, si è preso $\beta = y - x$ e $\alpha = x - \beta = 2x - y$.

Exercise 13.0.23 *Si consideri lo spazio $\mathcal{P}_2$. Mostrare che gli insiemi*

$$B_1 = \left\{1, x, x^2\right\} \text{ e } B_2 = \left\{\frac{1}{2}x\left(x-1\right), 1-x^2, \frac{1}{2}x\left(x+1\right)\right\}$$

sono due basi di $\mathcal{P}_2$. In particolare, come si può esprimere il polinomio

$$7 - 3x + 2x^2$$

nei termini di queste due basi B_1 e B_2?

Sol.: Poichè $\dim \mathcal{P}_2 = 3$, per la prima parte dell'esercizio è sufficiente dimostrare che i tre polinomi di B_1 sono linearmente indipendenti e poi lo stesso per B_2; lasciamo la verifica al lettore. Per quanto riguarda l'altro quesito, osserviamo che il polinomio $7 - 3x + 2x^2$ è già espresso in termini della base B_1 e quindi ha componenti $(7, -3, 2)$ rispetto a B_1. Per esprimerlo nei termini della base B_2, incominciamo a scrivere i polinomi $1, x, x^2$ come combinazioni lineari dei polinomi della base B_2, che indichiamo, nell'ordine, con A, B, C, ovvero poniamo $A = \frac{1}{2}x(x-1),\ B = 1 - x^2, C = \frac{1}{2}x(x+1)$. Partento dal polinomio 1 andiamo quindi a cercare dei coefficienti α, β, γ tali che

$$1 = \alpha A + \beta B + \gamma C,$$

ossia

$$1 = \alpha\left(\frac{1}{2}x(x-1)\right) + \beta\left(1-x^2\right) + \gamma\left(\frac{1}{2}x(x+1)\right).$$

Svolgendo i conti a secondo membro e raccogliendo i termini dello stesso grado, otteniamo

$$1 = (\alpha/2 + \gamma/2 - \beta)x^2 + (-\alpha/2 + \gamma/2)x + \beta.$$

Uguagliando i coefficienti dei termini dello stesso grado a primo e secondo membro, otteniamo il sistema

$$\begin{cases} \frac{\alpha}{2} + \frac{\gamma}{2} - \beta = 0, \\ -\frac{\alpha}{2} + \frac{\gamma}{2} = 0, \\ \beta = 1, \end{cases}$$

che risolto produce $\alpha = 1, \beta = 1, \gamma = 1$. Pertanto possiamo scrivere $1 = A + B + C$. Ragionando in maniera analoga per i polinomi x e x^2 otteniamo (verificarlo!)

$$\begin{cases} 1 = A + B + C, \\ x = C - A, \\ x^2 = A + C \end{cases}$$

(In realtà, per questo esercizio, dove le espressioni dei polinomi sono semplici, si sarebbe potuto vedere subito "a occhio" che valgono queste tre espressioni per $1, x, x^2$, senza quindi andare a risolvere i tre sistemi).

Sostituendo queste espressioni per $1, x, x^2$ nel polinomio $7 - 3x + 2x^2$ troviamo

$$7 - 3x + 2x^2 = 7(A + B + C) - 3(C - A) + 2(A + C) = 12A + 7B + 6C.$$

Pertanto le componenti del polinomio $7 - 3x + 2x^2$ nella base B_2 sono $(12, 7, 6)$. □

Exercise 13.0.24 *Dimostrare il "Solo se" del Teorema 104.*

Exercise 13.0.25 *Mostrare che* $\ker(T)$ *che* $\mathrm{Im}(T)$ *sono sottospazi vettoriali.*

Exercise 13.0.26 *Si consideri un'applicazione lineare* $T : \mathbb{R}^n \to \mathbb{R}^m$*, con matrice associata* A*. Si definisce rango di* A*, indicato con* $\rho(A)$*, il numero di vettori colonna di* A *linearmente indipendenti. Mostrare che* $\rho(A) = \rho(T)$*.*

Exercise 13.0.27 *Si mostri questa versione più completa della Proposizione 130: Dato un punto* $x \in \mathbb{R}^n$*, per ogni* $y, y' \in \mathbb{R}^n$ *si ha*

$$[x, x + y] \subseteq [x, x + y']$$

se e solo se esiste $\alpha \geq 1$ *tale che* $y' = \alpha y$*, mentre si ha*

$$[x, x + y'] \subseteq [x, x + y]$$

se e solo se esiste $0 < \alpha \leq 1$ *tale che* $y' = \alpha y$*.*

Exercise 13.0.28 *Consideriamo le applicazioni viste negli Esempi 160 e 161. Sia quindi* $g : \mathbb{R} \to \mathbb{R}^3$ *definita da* $g(x) = (x, \sin x, \cos x)$ *per ogni vettore* $x \in \mathbb{R}$*, mentre* $f : \mathbb{R}^3 \to \mathbb{R}^2$ *è definita come* $f(x_1, x_2, x_3) = (2x_1^2 + x_2 + x_3, x_1 - x_2^4)$ *per ogni vettore* $x \in \mathbb{R}^3$*. Siccome sia* f *e* g *sono Frechet differenziabili in ogni punto del loro dominio, per il Teorema 163 la composizione* $f \circ g : \mathbb{R} \to \mathbb{R}^2$ *è anch'essa Frechet differenziabile in ogni punto del suo dominio* $\mathbb{R}$*. Usando la regola della catena, si calcoli la matrice Jacobiana di* $f \circ g : \mathbb{R} \to \mathbb{R}^2$*.*

Sol.: Per la regola della catena (4.28), la matrice Jacobiana di $f \circ g : \mathbb{R} \to \mathbb{R}^2$ è data da:

$$D(f \circ g)(x) = Df(g(x))\, Dg(x).$$

Dall'Esempio 161 sappiamo che:

$$Dg(x) = \begin{bmatrix} 1 \\ \cos x \\ -\sin x \end{bmatrix},$$

mentre dall'Esempio 160 si ha:

$$Df(x) = \begin{bmatrix} 4x_1 & 1 & 1 \\ 1 & -4x_2^3 & 0 \end{bmatrix},$$

e quindi

$$Df(g(x)) = \begin{bmatrix} 4x & 1 & 1 \\ 1 & -4\sin^3 x & 0 \end{bmatrix}.$$

Pertanto,

$$\begin{aligned} Df(g(x))\, Dg(x) &= \begin{bmatrix} 4x & 1 & 1 \\ 1 & -4\sin^3 x & 0 \end{bmatrix} \begin{bmatrix} 1 \\ \cos x \\ -\sin x \end{bmatrix} \\ &= \begin{bmatrix} 4x + \cos x - \sin x \\ 1 - 4\sin^3 x \cos x \end{bmatrix}. \end{aligned}$$

Il differenziale di Frechet in x di $f \circ g$ è dunque dato dall'applicazione lineare $df(x) : \mathbb{R} \to \mathbb{R}^2$ definita come

$$d(f \circ g)(x)(h) = \begin{bmatrix} 4x + \cos x - \sin x \\ 1 - 4\sin^3 x \cos x \end{bmatrix} h.$$

Ad esempio, in $x = \pi$ si ha:

$$d(f \circ g)(x) = (4\pi - 1)h + h = 4\pi h.$$

Anche in questo caso il lettore può verificare l'esattezza di quanto abbiamo fatto usando la regola della catena, scrivendo esplicitamente la forma di $f \circ g$ e calcolandone la matrice Jacobiana. □

Exercise 13.0.29 *Calcolare la derivata della funzione $f : \mathbb{R} \to \mathbb{R}^3$ definita da $f(x) = (x, \sin x, \cos x)$ per ogni vettore $x \in \mathbb{R}$.*

Sol.: Nell'Esempio 161 avevamo visto come

$$Df(x) = \begin{bmatrix} 1 \\ \cos x \\ -\sin x \end{bmatrix}.$$

Quindi, la derivata $f' : \mathbb{R}^3 \to M(3,1)$ di f è la funzione che associa ad ogni $x \in \mathbb{R}^3$ la matrice Jacobiana $Df(x)$, ossia

$$x \longmapsto Df(x) = \begin{bmatrix} 1 \\ \cos x \\ -\sin x \end{bmatrix}.$$

La derivata di f in $x = \pi$ è data da

$$f'(x) = Df(x) = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}.$$

□

Exercise 13.0.30 Sia X uno spazio metrico rispetto alla metrica discreta (6.4). Quali sono gli insiemi aperti e chiusi in questo spazio metrico? Quali sono i compatti?

Exercise 13.0.31 Mostrare che in ogni spazio metrico i singoletti sono insiemi chiusi.

Exercise 13.0.32 Dato un insieme A di uno spazio metrico, mostrare che $\overset{\circ}{A}$ è un aperto e che, per ogni aperto $G \subseteq A$, si ha $G \subseteq \overset{\circ}{A}$.

Sol. Sia G un aperto tale che $G \subseteq A$. Sia $x \in G$. Poichè x è punto interno di G, esiste un intorno $B_\varepsilon(x)$ tale che $B_\varepsilon(x) \subseteq G$, e quindi tale $B_\varepsilon(x) \subseteq A$. Ne segue che $x \in \overset{\circ}{A}$, e quindi $G \subseteq \overset{\circ}{A}$.

Rimane da mostrare che $\overset{\circ}{A}$ è aperto. Sia $x \in \overset{\circ}{A}$. Per definizione di punto interno, esiste $B_\varepsilon(x)$ di x tale che $B_\varepsilon(x) \subseteq A$. Per la Proposizione 220, $B_\varepsilon(x)$ è aperto e quindi, per quanto appena dimostrato, $B_\varepsilon(x) \subseteq \overset{\circ}{A}$. Il punto x è perciò interno di $\overset{\circ}{A}$, che è dunque un aperto.

Exercise 13.0.33 Dati $x, y \in \mathbb{R}^n$ e un numero naturale $p \geq 1$, si definisca

$$d_p(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}.$$

Si mostri che d_p è una metrica su $\mathbb{R}^n$, che generalizza le metriche d_1 e d_2. Si mostri inoltre che per ogni metrica d_p vale la Proposizione 246.

Exercise 13.0.34 (Teorema del Confronto per successioni) Siano $\{x_n\}_{n\geq 1}$, $\{y_n\}_{n\geq 1}$ e $\{z_n\}_{n\geq 1}$ tre successioni di reali tali che $x_n \geq y_n \geq z_n$ per ogni n. Mostrare che, se $x_n \to x$ e $z_n \to x$, allora $y_n \to x$.

Exercise 13.0.35 Sia $\{x_n\}_{n\geq 1}$ una successione di reali. Si mostri che $\{|x_n|\}_{n\geq 1}$ è convergente se $\{x_n\}_{n\geq 1}$ è convergente.[3] Vale anche il viceversa?

Exercise 13.0.36 *(i)* Siano $\{x_n\}_n$ e $\{y_n\}_n$ due successioni in uno spazio metrico (X, d) tali che $x_n \to x$ e $y_n \to y$. Si mostri che $d(x_n, y_n) \to d(x, y)$.

[3]Si ricordi la disuguagliaza $||x| - |y|| \leq |x - y|$ per ogni $x, y \in \mathbb{R}$.

(ii) Siano $\{x_n\}_n$ e $\{y_n\}_n$ due successioni di Cauchy in uno spazio metrico (X, d). Si mostri che la successione $\{d(x_n, y_n)\}_n$ è convergente.

Exercise 13.0.37 Sia A un sottoinsieme di uno spazio metrico (X, d) e si definisca la funzione $f : X \to \mathbb{R}$ definita come $f(x) = \inf_{y \in A} d(x, y)$. Si verifichi se f è una funzione continua.

Exercise 13.0.38 Sia $(B([0,1]), d_\infty)$ lo spazio delle funzioni $f : [0,1] \to \mathbb{R}$ che sono limitate, ossia esiste $K > 0$ tale che $|f(x)| \leq K$ per ogni $x \in [0,1]$. Per il Teorema di Weierstrass, si ha $C([0,1]) \subseteq B([0,1])$.

(i) Sia $\{f_n\}_n \subseteq C([0,1])$ una successione di funzioni continue tali che $f_n \overset{d_\infty}{\to} f \in B([0,1])$. Si mostri che $f \in C([0,1])$.

(ii) Si mostri che $C([0,1])$ è un sottoinsieme chiuso di $B([0,1])$.

Exercise 13.0.39 Si mostri che $(C([0,1]), d_\infty)$ è uno spazio metrico completo.

Exercise 13.0.40 Si mostri che l'insieme $A_1 \times \cdots \times A_m$ è un chiuso di R^m se ogni A_i è un chiuso di R.

Exercise 13.0.41 Dimostrare le Proposizioni **??** e 304.

Exercise 13.0.42 Verificare la disuglianza (7.2).

Exercise 13.0.43 Dimostrare il Lemma 7.1.

Exercise 13.0.44 Dimostrare il seguente risultato sulla composizione di funzioni: Siano (X, d_X), (Y, d_Y) e (Z, d_Z) tre spazi metrici, e siano $f : A \subseteq X \to Y$ e $g : B \subseteq Y \to Z$ funzioni tali che $f(A) \subseteq B$. Se f è continua in $x \in A$ e g è continua in $f(x)$, allora la funzione composta $g \circ f : A \subseteq X \to Z$ è continua in x.[4]

Exercise 13.0.45 *Assuming $A = X$, give a proof of Theorem 303 only based on Lemma 298 and Theorem 299.*

[4]Si ricordi che la funzione composta $g \circ f : A \subseteq X \to Z$ si definisce come

$$(g \circ f)(x) = g(f(x)), \qquad \forall x \in A.$$

Sol. Let K be a compact set in X. We want to prove that $f(K)$ is compact. Let $\{G_i\}_{i\in I}$ be an open cover of $f(K)$. As $f(K) \subseteq \bigcup_{i\in I} G_i$, by Lemma 298 we have:

$$K \subseteq f^{-1}(f(K)) \subseteq f^{-1}\left(\bigcup_{i\in I} G_i\right) = \bigcup_{i\in I} f^{-1}(G_i).$$

By Theorem 299, each $f^{-1}(G_i)$ is open, and therefore $\{f^{-1}(G_i)\}_{i\in I}$ is a an open cover of K. Since K is compact, there exists a finite subcover $\{f^{-1}(G_i)\}_{i=1}^{n}$of K. Therefore, again by Lemma 298, we have

$$f(K) \subseteq f\left(\bigcup_{i=1}^{n} f^{-1}(G_i)\right) = f\left(f^{-1}\left(\bigcup_{i=1}^{n} G_i\right)\right) \subseteq \bigcup_{i=1}^{n} G_i.$$

This implies that $\{G_i\}_{i=1}^{n}$ is a finite subcover of $f(K)$, which is therefore compact. □

Exercise 13.0.46 Dimostrare la seguente generalizzazione della Proposizione 311: una funzione $f : X \to \mathbb{R}$ è superiormente semicontinua su un sottoinsieme chiuso F di X solo se gli insiemi $(f \geq t) \cap F$ sono chiusi per ogni $t \in \mathbb{R}$. Il vicecersa vale se $F = X$.

Sol. Sia f superiormente semicontinua su F. Fissato $t \in \mathbb{R}$, vogliamo mostrare che $(f \geq t) \cap F$ è chiuso. Sia $\{x_n\}_n \subseteq (f \geq t) \cap F$ con $x_n \to x \in X$. Alla luce del Corollario 255, occorre mostrare che $x \in (f \geq t) \cap F$.

Per ogni n si ha $x_n \in F$ e $f(x_n) \geq t$. Siccome F è chiuso, si ha $x \in F$ per il Corollario 255. Siccome f è superiormente semicontinua su F, per la Proposizione 309 si ha $\limsup_n f(x_n) \leq f(x)$, il che implica $t \leq f(x)$, ossia $x \in (f \geq t)$. In conclusione, $x \in (f \geq t) \cap F$, come desiderato.

Viceversa, supponiamo $F = X$ e che gli insiemi $(f \geq t)$ siano chiusi per ogni $t \in \mathbb{R}$. Fissato $x \in F$, sia $\{x_n\}_n$ tale che $x_n \to x$. Vogliamo mostrare che $\limsup_n f(x_n) \leq f(x)$. Per contraddizione, assumiamo che $\limsup_n f(x_n) > f(x)$. Sia $\alpha \in \mathbb{R}$ tale che $\limsup_n f(x_n) > \alpha > f(x)$. Esiste una sottosuccessione $\{x_{n_k}\}_k$ tale che $f(x_{n_k}) \geq \alpha$ per ogni k. D'altra parte $x_n \to x$ implica $x_{n_k} \to x$, e quindi per il Corollario 255 si ha $x \in \{f \geq \alpha\}$ poichè $\{f \geq \alpha\}$ è chiuso. Ma ciò implica $f(x) \geq \alpha > f(x)$, e questa contraddizione ci permette di concludere che $\limsup_n f(x_n) \leq f(x)$. □

Exercise 13.0.47 Fissata una funzione $g \in C([a,b])$, si consideri il funzionale lineare $F : C([a,b]) \to \mathbb{R}$ dato da

$$F(f) = \int_a^b f(t)\, g(t)\, dt, \qquad \forall f \in C([a,b]).$$

Si mostri che $\|F\| = \int_a^b |g(t)|\, dt$.

Exercise 13.0.48 Si consideri lo spazio $C^1([0,1])$ delle funzioni $f:[0,1]\to R$ differenziabili con continuità. Si verifichi se

$$\|f\| = \sup_{t\in[0,1]} |f(t)| + |f'(t)|$$

è una norma e se il funzionale lineare $F: C^1([0,1]) \to \mathbb{R}$ dato da

$$F(f) = \int_0^1 |f'(t)|\, dt, \qquad \forall f \in C^1([0,1]).$$

è continuo su $\left(C^1([0,1]), \|\cdot\|\right)$

Exercise 13.0.49 Siano $(V_1, \|\cdot\|_1)$ e $(V_2, \|\cdot\|_2)$ due spazi vettoriali normati. Si mostri che $V_1 \times V_2$ diventa anch'esso uno spazio vettoriale normato una volta dotato di una delle seguenti norme:

(i) $\|(v_1, v_2)\| = \|v_1\| + \|v_2\|$,

(ii) $\|(v_1, v_2)\| = \max\{\|v_1\|, \|v_2\|\}$,

(iii) $\|(v_1, v_2)\| = \sqrt{\|v_1\|_1^2 + \|v_2\|_2^2}$

Sia $B: V_1 \times V_2 \longrightarrow \mathbb{R}$ bilineare, ossia

$$\begin{aligned} B(v_1, \alpha v_2' + \beta v_2'') &= \alpha B(v_1, v_2') + \beta B(v_1, v_2'') \\ B(\alpha v_1' + \beta v_1'', v_2) &= \alpha B(v_1', v_2) + \beta B(v_1'', v_2) \end{aligned}$$

per ogni $\alpha, \beta \in R$. Si mostri che B è continua se e solo se esiste $c > 0$ tale che:

$$|B(v_1, v_2)| \le c\|v_1\|\|v_2\|$$

Exercise 13.0.50 Dimostrare il seguente risultato: sia C un insieme convesso di uno spazio vettoriale, e sia $x \in C$; allora, $x \in \text{ext}C$ se e solo se le condizioni $x + y \in C$ e $x - y \in C$ implicano $y = \mathbf{0}$.

Exercise 13.0.51 Mostrare che nell'Esempio 399 si ha $\text{ext}co(A) \neq \emptyset$ e $\text{ext}co(A) \subseteq A$.

Exercise 13.0.52 Dimostrare la Proposizione 427.

Exercise 13.0.53 I coni sono insiemi convessi C tali per cui $\alpha v \in C$ per ogni $\alpha \ge 0$ se $v \in C$. In altre parole, i coni sono insiemi convessi che sono chiusi riseptto alla moltiplicazione scalare non-negativa. Si mostri che i coni contengono sempre l'elemento neutro, e che, data una collezione finita $\{v^i\}_{i\in I}$ di vettori, l'insieme

$$\left\{ \sum_{i\in I} \alpha_i v^i : \alpha_i \ge 0 \text{ per ogni } i \in I \right\}$$

è un cono contenente tutti i vettori dati.

Exercise 13.0.54 Let C be a convex set of $\mathbb{R}^n$ that contains n linearly independent vectors. Show that its interior is nonempty.

Exercise 13.0.55 Dato un funzionale $f : V \to \mathbb{R}$, si definisca il funzionale $\overline{f} : V \to \mathbb{R}$ come $\overline{f}(v) = -f(-v)$ per ogni $v \in V$. Si mostri che $\overline{f}$ è sublineare se f è superlineare.

Exercise 13.0.56 Sia $(V, \|\cdot\|)$ uno spazio vettoriale normato. Dato $v_0 \in V$, si mostri che l'insieme $U_1(v_0) = \{v \in V : \|v - v_0\| = 1\}$ è chiuso. Si mostri anche che $U_1(v_0)$ è compatto se V è finito dimensionale,

Sol. Sia $\{v^n\}_n$ una successione contenuta in $U_1(v_0)$ e tale che $v^n \to v$. Per il Corollario 255, occorre mostrare che $v \in U_1(v_0)$.

Da $v^n \to v$ segue che $\|v^n - v_0\| \to \|v - v_0\|$. Infatti, $\|(v^n - v_0) - (v - v_0)\| = \|v^n - v\| \to 0$. D'altra parte, $\|v^n - v_0\| = 1$ per ogni n, ed insieme a $\|v^n - v_0\| \to \|v - v_0\|$ questo implica $\|v - v_0\| = 1$, ossia $v \in U_\varepsilon(v_0)$.

Si supponga che V sia finito dimensionale. Per il Teorema 364, la palla unitaria chiusa $\overline{B}_1(v_0) = \{v \in V : \|v - v_0\| \leq 1\}$ è compatta. Dunque, $U_1(v_0)$ è un sottoinsieme chiuso di un compatto, ed è quindi a sua volta un insieme compatto grazie alla Proposizione 274.

Exercise 13.0.57 Given a function $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$, set $f^+(x) = \max\{f(x), 0\}$ for all $x \in A$. If $f \in \mathcal{C}^1(A)$, then $(f^+)^2 \in \mathcal{C}^1(A)$, with

$$\frac{\partial (f^+)^2(x)}{\partial x_i} = 2f^+(x) \frac{\partial f(x)}{\partial x_i}, \qquad \forall i = 1, ..., n \tag{13.1}$$

Sol. Consider first the scalar function $\varphi(t) = t^+ = \max\{t, 0\}$. We show that $[\varphi(t)]^2$ is of class $\mathcal{C}^1(A)$ and

$$D[\varphi(t)]^2 = 2t^+.$$

This is just a matter of computation. If $t_0 < 0$, $[\varphi(t)]^2$ is locally equal to 0. Hence, $D[\varphi(t_0)]^2 = 0 = 2(t_0^+)$. If $t_0 > 0$, $[\varphi(t)]^2$ agrees locally with the function t^2. It follows that $D[\varphi(t_0)]^2 = 2t = 2(t_0^+)$. It remains to check the derivative at $t = 0$. The incremental ratio is

$$\begin{aligned} \frac{[\varphi(h)]^2 - [\varphi(0)]^2}{h} &= \frac{h^2}{h} \text{ for } h > 0 \\ \frac{[\varphi(h)]^2 - [\varphi(0)]^2}{h} &= \frac{0}{h} \text{ for } h < 0. \end{aligned}$$

Hence $D[\varphi(0)]^2 = 0 = 2(0^+)$. We conclude that $[\varphi(t)]^2$ is everywhere differentiable and with derivative $2t^+$. As $t \to 2t^+$ is continuous, $[\varphi(t)]^2$ is of class C^1.

Now the function $[f^+(x)]^2 = [\varphi(f(x))]^2$ is C^1 with derivatives (13.1) by the chain rule. □

Exercise 13.0.58 Dato $a \in \mathbb{R}$, si consideri una funzione continua $f : [a, +\infty) \to \mathbb{R}$ con $f'(x) > 0$ per ogni $x > a$. Si mostri che f è strettamente crescente su $[a, +\infty)$ e che la sua inversa f^{-1} è strettamente crescente su $f([a, +\infty))$.

Sol. Per risultati di base si ha che f è strettamente crescente su (a, ∞). Per completare la prova, mostriamo che $f(x) > f(a)$ per ogni $x > a$. Se ciò non fosse il caso, esiste $\overline{x} > a$ tale che $f(\overline{x}) \leq f(a)$. Preso un qualsiasi $\varepsilon > 0$ con $\overline{x} - \varepsilon > a$, si ha

$$f(a) \geq f(\overline{x}) > f(\overline{x} - \varepsilon) > f(x), \qquad \forall x \in (a, \overline{x} - \varepsilon),$$

e quindi la continuità di f porta alla contraddizione:

$$f(a) = \lim_{x \to a+} f(x) \leq f(\overline{x} - \varepsilon) < f(a).$$

Concludiamo che $f(x) > f(a)$, come desiderato.

Poichè f è continua e strettamente crescente, si ha $f([a, +\infty)) = [f(a), +\infty)$. Da risultati di base si ha:

$$\left(f^{-1}\right)'(y) = \frac{1}{f'\left(f^{-1}(y)\right)} > 0$$

per ogni $y \in (f(a), +\infty)$. Quindi $\left(f^{-1}\right)'(y) > 0$ per ogni $y \in (f(a), +\infty)$, e per quanto mostrato nella prima parte possiamo concludere che f^{-1} è strettamente crescente su $[f(a), +\infty)$. □

Exercise 13.0.59 *Show that for any neighborhhod $B_\varepsilon(v)$ in a normed vector space we have*

$$B_\varepsilon(v) = \{v + w : \|w\| < \varepsilon\}.$$

Sol. Let $v + w$ be such that $\|w\| < \varepsilon$. Then, $\|v - (v + w)\| = \|w\| < \varepsilon$, and so $v + w \in B_\varepsilon(v)$. Hence, $\{v + w : \|w\| < \varepsilon\} \subseteq B_\varepsilon(v)$. Conversely, suppose $z \in B_\varepsilon(v)$. Set $w = z - v$. Then, $\|w\| < \varepsilon$ and $v + w = z$. Hence, $B_\varepsilon(v) \subseteq \{v + w : \|w\| < \varepsilon\}$. □

Exercise 13.0.60 Let $\phi : \mathbb{R} \to \mathbb{R}$ be concave. Show that ϕ is unbounded above if $\lim_{x \to +\infty} \phi'(x) > 0$.

Sol. By hypothesis, there is a sequence $\{x_n\}_n$, with $x_n \uparrow +\infty$, such that $\lim_{n \to +\infty} \phi'(x_n) > 0$. Given n, for all $y \in \mathbb{R}$ we have $\phi(y) \leq \phi(x_n) + \phi'(x_n)(y - x_n)$. In particular, $y = 0$ implies $\phi'(x_n) x_n \leq \phi(x_n)$. Then, setting $\alpha = \lim_{n \to +\infty} \phi'(x_n)$, we have $\alpha x_n \leq \phi(x_n)$ for all n, which implies $\lim_{n \to +\infty} \phi(x_n) = +\infty$. □

Exercise 13.0.61 Let $f : (a, b) \to \mathbb{R}$ be a concave function defined on the, possibly unbounded, interval (a, b). Show that

$$f'_+(x) = f'(x; 1) \quad \text{and} \quad f'_-(x) = -f'(x; -1).$$

Exercise 13.0.62 Prove what stated in Example 458.

Exercise 13.0.63 Prove (ii) of Theorem 460.

Exercise 13.0.64 Let $\{f_i\}_{i\in I}$ a family of concave functions $f_i : C \to \mathbb{R}$ defined on a convex set C. Extend Proposition 427 by showing that the function $f : C \to \mathbb{R}$ given by $f(x) = \inf_{i\in I} f_i(x)$ for all $x \in C$ is concave. Moreover, show that f is strictly concave provided each f_i is strictly concave and I is finite (what happens when I is not finite?).

Exercise 13.0.65 Prove Lemma 451.

Exercise 13.0.66 For a concave function $\phi : [a, \infty) \to \mathbb{R}$, with $a \geq 0$, the following conditions are equivalent:

(i) ϕ is nondecreasing;

(ii) $\limsup_{t\to\infty} \phi(t) > -\infty$;

(iii) $\liminf_{t\to\infty} \frac{\phi(t)}{t} \geq 0$.

In particular, the limits in (ii) and (iii) exist.

Sol. (i) implies both (ii) and (iii). Suppose ϕ that is nondecreasing. Since ϕ is proper, there is $t_0 \geq 0$ such that $\phi(t_0) \in \mathbb{R}$. Hence, $\limsup_{t\to\infty} \phi(t) \geq \phi(t_0) > -\infty$. Moreover, $\liminf_{t\to\infty} \phi(t)/t = \liminf_{t\to\infty} (\phi(t) - \phi(t_0))/t \geq 0$.

(ii) implies (i). Suppose $\limsup_{t\to\infty} \phi(t) > -\infty$ and suppose, per contra, that there is $0 \leq t_1 < t_2$ such that $\phi(t_1) > \phi(t_2)$. Given any $t > t_2$, set $\alpha = (t_2 - t_1)/(t - t_1)$, so that $t_2 = \alpha t + (1 - \alpha) t_1$. By concavity,

$$\phi(t) \leq \frac{t}{t_2 - t_1} (\phi(t_2) - \phi(t_1)) + t_2\phi(t_1) - t_1\phi(t_2),$$

and so $\limsup_{t\to\infty} \phi(t) = -\infty$, a contradiction.

(iii) implies (i). Set $\operatorname{int}\operatorname{dom}(\phi) = (a, b)$. We have $b = \infty$. For, if $b < \infty$, then by concavity $\phi(t) = -\infty$ for all $t > b$, which contradicts (iii). We thus have $\operatorname{dom}(\phi) = [a, \infty)$. Let $t_0 > a$. We have $\phi(t) \leq \phi(t_0) + \phi'_+(t_0)(t - t_0)$ for all $t > a$, and so

$$0 \leq \liminf_{t\to\infty} \frac{\phi(t)}{t} \leq \liminf_{t\to\infty} \left(\frac{\phi(t_0)}{t} + \phi'_+(t_0) - \frac{t_0}{t} \right) = \phi'_+(t_0).$$

In turn, this implies that ϕ is nondecreasing on (a, ∞), and so on $[a, \infty)$ by concavity. Since, $\operatorname{dom}(\phi) = [a, \infty)$, we have $\phi(t) = -\infty$ for all $t \in [0, a)$ and this shows that ϕ is nondecreasing on $[0, \infty)$.

Since (ii) implies that ϕ is nondecreasing, $\lim_{t\to\infty}\phi(t)$ exists. As to the limit in (iii), let $t_0 \in \operatorname{dom}(\phi)$ and set $\psi(t) = \phi(t) - \phi(t_0)$. Then ψ is concave and subadditive, so that $\psi(t)/t$ is decreasing. Hence, $\lim_{t\to\infty}\psi(t)/t$ exists and, being $\lim_{t\to\infty}\psi(t)/t = \lim_{t\to\infty}\phi(t)/t$, we conclude that $\lim_{t\to\infty}\phi(t)/t$ exists. ■

Exercise 13.0.67 For a concave function $\phi : \mathbb{R} \to \mathbb{R}$, the following conditions are equivalent:

(i) ϕ is constant;

(ii) $\inf_{t\in\mathbb{R}}\phi > -\infty$;

(iii) $\limsup_{t\to\pm\infty}\phi(t) > -\infty$;

(iv) $\liminf_{t\to\pm\infty}\frac{\phi(t)}{t} = 0$.

In particular, the limits in (iii) and (iv) exist.

Proof. (i) trivially implies (ii) and (ii) trivially implies (iii). To prove the other implications, wlog assume $\phi(0) = 0$, and define $\psi', \psi'' : [0,\infty) \to [-\infty,\infty)$ by $\psi'(t) = \phi(t)$ and $\psi''(t) = \phi(-t)$ for all $t \geq 0$. Both ψ' and ψ'' are proper concave functions on $[0,\infty)$. If we apply Exercise 13.0.66 to ψ' and ψ'' we easily see that the limits in (iii) and (iv) exist.

(iii) implies (iv). By Exercise 13.0.66,

$$\lim_{t\to\infty}\phi(t) = \lim_{t\to\infty}\psi'(t) > -\infty \Longrightarrow \lim_{t\to\infty}\frac{\phi(t)}{t} = \lim_{t\to\infty}\frac{\psi'(t)}{t} \geq 0$$

Moreover, $\lim_{t\to-\infty}\phi(t) > -\infty$ if and only if $\lim_{t\to\infty}\phi(-t) > -\infty$. Hence, Exercise 13.0.66 implies:

$$\lim_{t\to\infty}\phi(-t) = \lim_{t\to\infty}\psi''(t) > -\infty \Longrightarrow \lim_{t\to\infty}\frac{\phi(t)}{t} \leq \lim_{t\to\infty} -\frac{\phi(-t)}{t} = -\lim_{t\to\infty}\frac{\psi''(t)}{t} \leq 0.$$

We conclude that $\lim_{t\to\infty}\phi(t)/t = 0$. A similar argument shows that $\lim_{t\to-\infty}\phi(t)/t = 0$.

(iv) implies (i). Suppose $\lim_{t\to\pm\infty}\phi(t)/t = 0$. This implies $\lim_{t\to\infty}\psi'(t)/t = \lim_{t\to\infty}\psi''(t)/t = 0$. By Exercise 13.0.66, both ψ' and ψ'' are nondecreasing. Since $\psi'(0) = \psi''(0) = 0$ and $-\psi'' \geq \psi'$, we conclude that $-\psi'' = \psi'$, i.e., ϕ is constant. ■

Exercise 13.0.68 Sia $h : \mathbb{R} \to \mathbb{R}$ una funzione con $h'(t) < 0$ per ogni $t > 0$ e con $h''(t) > 0$ per ogni $t > 0$. Si assuma inoltre che $h'(0) = 0$. Si risolva il problema di ottimo:

$$\begin{aligned} &\max_{x \in \mathbb{R}^n} \sum_{i=1}^{n} h(x_i) \\ &\text{sub } \sum_{i=1}^{n} x_i = 1,\ x_1 \geq 0,\ ...,\ x_n \geq 0 \end{aligned} \tag{13.2}$$

Sol. Alla luce dell'Esercizio 13.0.58, h è una funzione strettamente decrescente su $\mathbb{R}_+$, mentre h' è strettamente crescente. Quindi, h è strettamente convessa su $\mathbb{R}_+$. Inoltre, sempre alla luce dell'Esercizio 13.0.58, si ha che l'inversa $(h')^{-1}$ è ben definita poichè h è iniettiva.

Il problema (13.2) è uguale a quello risolto nell'Esempio 528. Le prime due fasi del metodo di eliminazione sono identiche a quelle viste nell'Esempio 527. In particolare, si ha $D_0 = \emptyset$.

Il Lagrangiano $L : \mathbb{R}^{2n+1} \to \mathbb{R}$ è dato da

$$L(x_1, x_2, \mu) = \sum_{i=1}^{n} h(x_i) + \lambda \left(1 - \sum_{i=1}^{n} x_i\right) + \sum_{i=1}^{n} \mu_i x_i, \qquad \forall (x, \lambda, \mu) \in \mathbb{R}^{2n+1},$$

e per trovare l'insieme S dei suoi punti di Kuhn-Tucker occorre risolvere il sistema

$$\begin{cases} \frac{\partial L}{\partial x_i} = h'(x_i) - \lambda + \mu_i = 0, \qquad \forall i = 1, ..., n \\ \lambda \frac{\partial L}{\partial \lambda} = \lambda \left(1 - \sum_{i=1}^{n} x_i\right) = 0 \\ \frac{\partial L}{\partial \lambda} = 1 - \sum_{i=1}^{n} x_i = 0 \\ \mu_i \frac{\partial L}{\partial \mu_i} = \mu_i x_i = 0, \qquad \forall i = 1, ..., n \\ \frac{\partial L}{\partial \mu_i} = x_i \geq 0, \qquad \forall i = 1, ..., n \\ \mu_i \geq 0, \qquad \forall i = 1, ..., n \end{cases}$$

Se moltiplichiamo per x_i le prime n equazioni, si ottiene

$$h'(x_i) x_i - \lambda x_i + \mu_i x_i = 0, \qquad \forall i = 1, ..., n$$

Sommando queste nuove equazioni, si ha

$$\sum_{i=1}^{n} h'(x_i) x_i - \lambda \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \mu_i x_i = 0,$$

e quindi $\lambda = \sum_{i=1}^{n} h'(x_i) x_i$. Poichè $h'(x_i) \leq 0$ quando $x_i \geq 0$, la condizione $x_i \geq 0$ ci permette di concludere che $\lambda \leq 0$.

Se $x_i = 0$, si ha $h'(x_i) = 0$ e dalla condizione $\partial L / \partial x_i = 0$ segue che $\lambda = \mu_i$. Siccome $\mu_i \geq 0$ e $\lambda \leq 0$, ne segue che $\mu_i = 0$. A sua volta, questo implica $\lambda = 0$ e quindi usando

di nuovo la condizione $\partial L/\partial x_i = 0$ si conclude that $x_i = \lambda = 0$ per ogni $i = 1, ..., n$. Ma questo contraddice la condizione $\lambda\left(1 - \sum_{i=1}^{n} x_i\right) = 0$, e possiamo perciò concludere che $x_i \neq 0$, ossia $x_i > 0$.

Poichè questo vale per ogni $i = 1, ..., n$, ne segue che $x_i > 0$ per ogni $i = 1, ..., n$. Dalla condizione $\mu_i x_i = 0$ segue che $\mu_i = 0$ per ogni $i = 1, ..., n$, e le prime n equazione diventano:

$$h'(x_i) - \lambda = 0, \qquad \forall i = 1, ..., n$$

ossia, visto che l'inversa $(h')^{-1}$ è ben definita,

$$x_i = (h')^{-1}\left(\frac{\lambda}{2}\right), \qquad \forall i = 1, ..., n.$$

Le x_i sono quindi tutte uguali tra loro, e da $\sum_{i=1}^{n} x_i = 1$ segue che

$$x_i = \frac{1}{n}, \qquad \forall i = 1, ..., n.$$

In conclusione,

$$S = \left\{\left(\frac{1}{n}, ..., \frac{1}{n}\right)\right\}.$$

Poichè $D_0 = \emptyset$, si ha $S \cup (D_0 \cap C) = \{(1/n, ..., 1/n)\}$, e il metodo di eliminazione ci permette di concludere che il punto $(1/n, ..., 1/n)$ è la soluzione anche del problema di ottimo 10.28. ▲

Exercise 13.0.69 Si dimostri il seguente risultato: Sia $f : A_1 \times A_2 \subseteq \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ una funzione Gateaux differenziabile su $A_1 \times A_2$, dove A_1 e A_2 sono insiemi aperti, e siano K_1 e K_2 due sottoinsiemi chiusi e convessi di A_1 e A_2, rispettivamente. Se $(\widehat{x}, \widehat{y}) \in K_1 \times K_2$ è punto di sella di f su $K_1 \times K_2$, ossia se

$$f(\widehat{x}, y) \geq f(\widehat{x}, \widehat{y}) \geq f(x, \widehat{y}), \qquad \forall x \in K_1, \forall y \in K_2,$$

allora

$$\nabla_x f(\widehat{x}, \widehat{y}) \cdot (x - \widehat{x}) \leq 0, \qquad \forall x \in K_1, \tag{13.3}$$

$$\nabla_y f(\widehat{x}, \widehat{y}) \cdot (y - \widehat{y}) \geq 0, \qquad \forall y \in K_2. \tag{13.4}$$

Il viceversa vale se f è una funzione di sella su $K_1 \times K_2$.

Sol.: Si osservi che $\nabla_x f^y(x) = \nabla_x f(x, y)$ per ogni sezione f^y e $\nabla_y f^x(y) = \nabla_y f(x, y)$ per ogni sezione f^x. Poichè $\widehat{x}$ è punto di massimo globale per la sezione $f^{\widehat{y}}$, il Teorema 537 implica

$$\nabla_x f(\widehat{x}, \widehat{y}) \cdot (x - \widehat{x}) = \nabla_x f^y(x) \cdot (x - \widehat{x}) \leq 0, \qquad \forall x \in A_1.$$

Analogamente, essendo $\widehat{y}$ punto di minimo globale per la sezione $f^{\widehat{x}}$, per la (**??**) si ha:

$$\nabla_y f(\widehat{x},\widehat{y})\cdot(y-\widehat{y}) = \nabla_y f^x(y)\cdot(y-\widehat{y}) \geq 0, \qquad \forall y \in K_2.$$

Infine, se le sezioni $f^{\widehat{y}}$ e $f^{\widehat{x}}$ sono concave e convesse, rispettivamente, allora per il Teorema 537 le condizioni (13.3) e (13.4) sono anche sufficienti affinchè $(\widehat{x},\widehat{y})$ sia punto di sella. □

Exercise 13.0.70 Si mostri che nell'Esercizio 13.0.69 la condizione (13.3) assume la forma $\nabla_x f(\widehat{x},\widehat{y}) = 0$ se $\widehat{x}$ è punto interno di K_1, mentre la (13.4) assume la forma $\nabla_y f(\widehat{x},\widehat{y}) = 0$ se $\widehat{y}$ è punto interno di K_2.

Exercise 13.0.71 Si mostri che gli insiemi:

$$C_1 = \{f \in C([0,1]) : |f(x)| \leq 1 \text{for all } x \in [0,1]\}$$

e

$$C_2 = \{f \in C([0,1]) : f(x) \leq 0 \text{for all } x \in [0,1]\}$$

sono insiemi convessi dello spazio vettoriale $C([0,1])$.

Exercise 13.0.72 Siano A e B sottoinsiemi di uno spazio vettoriale V. Si mostri che:

(i) $co(co(A)) = co(A)$;

(ii) $co(A) \subseteq co(B)$ se $A \subseteq B$;

(iii) $co(A) \cup co(B) \subseteq co(A \cup B)$.

Infine, si dia un esempio in cui $co(A) \cup co(B) \neq co(A \cup B)$.

Exercise 13.0.73 Let $f : \mathbb{R} \to \mathbb{R}$ be a differentiable function. Show that f is a contraction provided

$$\sup_{x\in\mathbb{R}} |f'(x)| < 1.$$

Exercise 13.0.74 Let $f \in C^1([a,b])$. Show that f is a contraction provided $|f'(x)| < 1$ for all $x \in [a,b]$. Notice that the compactness of the domain is key. For instance, the function

$$f(x) = x + \frac{1}{1+x}$$

is not a contraction on $\mathbb{R}_+$ though it satisfies $f'(x) < 1$ over $\mathbb{R}_+$.

Exercise 13.0.75 Let $T_1, T_2 : B(X) \to B(X)$ be two contractions.

(i) Show that it is not restrictive to assume that both operators have the same contraction modulus β.

(ii) Show that the operator $T = \max\{T_1, T_2\}$ is a β-contraction.

(iii) What can be said on the minimum $Q = \min\{T_1, T_2\}$?

Exercise 13.0.76 Let $T : B(X) \to B(X)$ be defined as $Tf = h + \beta f$, where $0 \leq \beta < 1$ and h is a fixed element of $B(X)$. Prove that T is a Blackwell operator. What is the fixed point?

Exercise 13.0.77 Let $T : B(\mathbb{R}) \to B(\mathbb{R})$ be defined as $Tf(x) = h(x) + \beta f(x + c)$, where $h \in B(\mathbb{R})$ and $c > 0$. Prove that it is a contraction. Find the fixed point.

Sol Consider

$$\overline{f} = h(x) + \beta h(x + c) + \beta^2 h(x + 2c) + \beta^3 h(x + 3c) + ...$$

. □

Exercise 13.0.78 Prove the following version of Proposition 581: Suppose the continuous function $\phi : [a, b] \times [a, b] \times \mathbb{R} \to \mathbb{R}$ is such that, for all $(s, t) \in [a, b] \times [a, b]$,

$$|\phi(s, t, z_1) - \phi(s, t, z_2)| \leq K |z_1 - z_2|, \qquad \forall z_1, z_2 \in \mathbb{R}.$$

Then, given any $g \in C([a, b])$, the Volterra backward integral equation (12.5)

$$f(s) = \int_s^b \psi(s, t, f(t))\, dt + g(s), \qquad \forall s \in [a, b],$$

has a unique solution $f \in C([a, b])$. In particular, the sequence $\{f_n\}_n \subseteq C([a, b])$ defined inductively by choosing $f_0 \in C([a, b])$ and setting

$$f_{n+1}(s) = \int_a^s \psi(s, t, f_n(t))\, dt + g(s), \qquad \forall s \in [a, b],$$

is such that $\|f_n - f\|_\infty \to 0$.

Exercise 13.0.79 Prove the following version of Proposition 580: Suppose the function $\psi : [a, b] \times [a, b] \times R \to R$ is bounded and continuous. Then, given any $g \in C([a, b])$, the Volterra backward integral equation

$$f(s) = \int_s^b \psi(s, t, f(t))\, dt + g(s), \qquad \forall s \in [a, b],$$

has a solution $f \in C([a, b])$.

Bibliography

[1] C. Arzelà, Un'osservazione intorno alle serie di funzioni, *Rend. dell' Accad. R. delle Sci. dell'Istituto di Bologna*, 142–159, 1882-1883.

[2] C. Arzelà, Sulle funzioni di linee, *Mem. Accad. Sci. Ist. Bologna Cl. Sci. Fis. Mat.*, 5, 55-74, 1895.

[3] G. Ascoli, Le curve limiti di una varietà data di curve, *Atti della R. Accad. dei Lincei Memorie della Cl. Sci. Fis. Mat. Nat.*, 18, 521-586, 1883-1884.

[4] S. Banach, Sur les opérations dans les ensembles abstracts et leur application aux équations intégrales, *Fundamenta Mathematicae*, 3, 133-181, 1922.

[5] D. Blackwell, Discounted dynamic programming, *Annals of Mathematical Statistics*, 36, 226-235, 1965.

[6] K. Border, *Fixed point theorems with applications to economics and game theory*, Cambridge University Press, 1985.

[7] L.E.J. Brouwer, Uber abbildung von mannigfaltigkeiten, *Math. Ann.*, 71, 97-115, 1912.

[8] R. Caccioppoli, Un teorema generale sull'esistenza di elementi uniti in una transformazione funzionale, *Rendiconti Accademia Nazionale Lincei*, 11, 794-799, 1930.

[9] M. Edelstein, An extension of Banach's contraction principle, *Proc. Amer. Math. Soc.* 12 (1961), 7-10.

[10] M. Edelstein, On fixed and periodic points under contractive mappings, *J. London Math. Soc.* 37 (1962), 74-79.

[11] P. Hartman, A differential equation with non-unique solutions, *Amer. Math. Monthly*, 70, 255-259, 1963.

[12] J. Leray and J. Schauder, Topologie et équations fonctionnelles, Ann. Sci. Ecole Norm. Sup., 51, 45-78, 1934.

[13] M. Lavrentieff, Sur une équation différentielle du premier ordre, *Mathematische Zeitschrift*, 23, 197-209, 1925.

[14] L. Montrucchio, Introduzione alla teoria delle scelte, Carocci Editore, Roma, 1998.

[15] G. Peano, Sull'integrabilità delle equazioni differenziali del primo ordine, *Atti Accad. Sci. Torino*, 21, 677-685, 1886.

[16] G. Peano, Demonstration de l'intégrabilité des équations différentielles ordinaires, *Mathematische Annalen*, 37, 182-228, 1890.

[17] C. E. Picard, Mémoire sure théorie des equations aux derives partielles et la méthode des approximations successives, *J. Math. Pures Appl.*, 5, 423-441, 1890.

[18] A. W. Roberts and D. E. Varberg, Another proof that convex functions are locally Lipschitz, *The American Mathematical Monthly*, 81, 1014-1016, 1974.

[19] C. A. Rogers, A less strange version of Milnor's proof of Brouwer's fixed-point theorem, *Amer. Math. Monthly*, 87, 525-527, 1980.

[20] W. Rudin, *Principles of mathematical analysis*, McGraw-Hill, 1976.

[21] J. Schauder, Der fixpunktsatz in functionalraumen, *Studia Math.*, 2, 171-180, 1930.

[22] H. Weyl, Elementare theorie der konvexen polyeder, *Comentarii Mathematici Helvetici,* 7, 290-306, 1935.