Fundamentals of Linear Algebra and Optimization

Jean Gallier and Jocelyn Quaintance
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104, USA
e-mail: jean@cis.upenn.edu

December 31, 2018

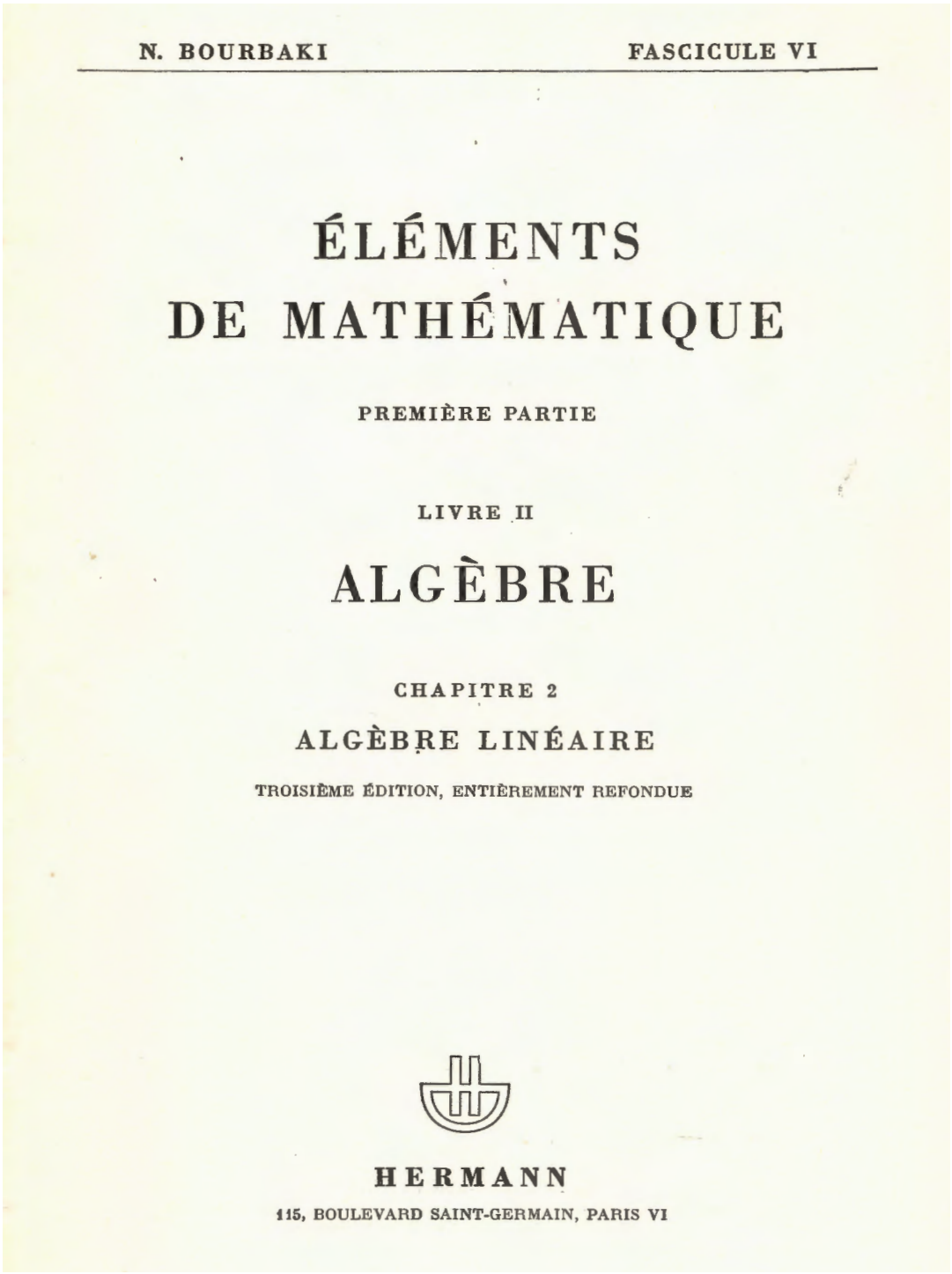


Figure 1: Cover page from Bourbaki, Fascicule VI, Livre II, Algèbre, 1962

a) *u est bijectif ;*
b) *u est injectif ;*
c) *u est surjectif ;*
d) *u est inversible à droite ;*
e) *u est inversible à gauche ;*
f) *u est de rang n.*

Si E est un espace vectoriel de dimension infinie, il y a des endomorphismes injectifs (resp. surjectifs) de E qui ne sont pas bijectifs (exerc. 9).

Soient K, K′ deux corps, $\sigma : K \to K'$ un *isomorphisme* de K sur K′, E un K-espace vectoriel, E′ un K′-espace vectoriel, $u : E \to E'$ une application *semi-linéaire* relative à σ (§ 1, n° 13) ; on appelle encore *rang* de u la dimension du sous-espace $u(E)$ de E′. C'est aussi le rang de u considéré comme application linéaire de E dans $\sigma_*(E')$, car toute base de $u(E)$ est aussi une base de $\sigma_*(u(E))$.

5. *Dual d'un espace vectoriel.*

THÉORÈME 4. — *La dimension du dual* E* *d'un espace vectoriel* E *est au moins égale à la dimension de* E. *Pour que* E* *soit de dimension finie, il faut et il suffit que* E *le soit, et on a alors* $\dim E^* = \dim E$.

Si K est le corps des scalaires de E, E est isomorphe à un espace $K_s^{(I)}$ et par suite E* est isomorphe à K_d^I (§ 2, n° 6, prop. 10). Comme $K_d^{(I)}$ est un sous-espace de K_d^I, on a $\dim E = \mathrm{Card}(I) \leqslant \dim E^*$ (n° 2, cor. 4 du th. 3) ; en outre, si I est fini, on a $K_d^I = K_d^{(I)}$ (cf. exerc. 3 *d*)).

COROLLAIRE. — *Pour un espace vectoriel* E, *les relations* $E = \{0\}$ *et* $E^* = \{0\}$ *sont équivalentes.*

THÉORÈME 5. — *Étant données deux suites exactes d'espaces vectoriels (sur un même corps* K*) et d'applications linéaires*

$$0 \to E' \to E \to E'' \to 0$$
$$0 \to F' \to F \to F'' \to 0$$

Figure 2: Page 156 from Bourbaki, Fascicule VI, Livre II, Algèbre, 1962

Contents

Part I

Linear Algebra

Chapter 1

Vector Spaces, Bases, Linear Maps

1.1 Motivations: Linear Combinations, Linear Independence and Rank

In linear optimization problems, we encounter systems of linear equations. For example, consider the problem of solving the following system of three linear equations in the three variables $x_1, x_2, x_3 \in \mathbb{R}$:

$$\begin{aligned} x_1 + 2x_2 - x_3 &= 1 \\ 2x_1 + x_2 + x_3 &= 2 \\ x_1 - 2x_2 - 2x_3 &= 3. \end{aligned}$$

One way to approach this problem is introduce the "vectors" u, v, w, and b, given by

$$u = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} \qquad v = \begin{pmatrix} 2 \\ 1 \\ -2 \end{pmatrix} \qquad w = \begin{pmatrix} -1 \\ 1 \\ -2 \end{pmatrix} \qquad b = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

and to write our linear system as

$$x_1 u + x_2 v + x_3 w = b.$$

In the above equation, we used implicitly the fact that a vector z can be multiplied by a scalar $\lambda \in \mathbb{R}$, where

$$\lambda z = \lambda \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} \lambda z_1 \\ \lambda z_2 \\ \lambda z_3 \end{pmatrix},$$

and two vectors y and and z can be added, where

$$y + z = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} + \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} y_1 + z_1 \\ y_2 + z_2 \\ y_3 + z_3 \end{pmatrix}.$$

The set of all vectors with three components is denoted by $\mathbb{R}^{3\times 1}$. The reason for using the notation $\mathbb{R}^{3\times 1}$ rather than the more conventional notation $\mathbb{R}^3$ is that the elements of $\mathbb{R}^{3\times 1}$ are *column vectors*; they consist of three rows and a single column, which explains the superscript 3×1. On the other hand, $\mathbb{R}^3 = \mathbb{R} \times \mathbb{R} \times \mathbb{R}$ consists of all triples of the form (x_1, x_2, x_3), with $x_1, x_2, x_3 \in \mathbb{R}$, and these are *row vectors*. However, there is an obvious bijection between $\mathbb{R}^{3\times 1}$ and $\mathbb{R}^3$ and they are usually identifed. For the sake of clarity, in this introduction, we will denote the set of column vectors with n components by $\mathbb{R}^{n\times 1}$.

An expression such as

$$x_1u + x_2v + x_3w$$

where u, v, w are vectors and the x_is are scalars (in $\mathbb{R}$) is called a *linear combination*. Using this notion, the problem of solving our linear system

$$x_1u + x_2v + x_3w = b.$$

is equivalent to *determining whether b can be expressed as a linear combination of* u, v, w.

Now, if the vectors u, v, w are *linearly independent*, which means that there is *no* triple $(x_1, x_2, x_3) \neq (0, 0, 0)$ such that

$$x_1u + x_2v + x_3w = 0_3,$$

it can be shown that *every* vector in $\mathbb{R}^{3\times 1}$ can be written as a linear combination of u, v, w. Here, 0_3 is the *zero vector*

$$0_3 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

It is customary to abuse notation and to write 0 instead of 0_3. This rarely causes a problem because in most cases, whether 0 denotes the scalar zero or the zero vector can be inferred from the context.

In fact, every vector $z \in \mathbb{R}^{3\times 1}$ can be written *in a unique way* as a linear combination

$$z = x_1u + x_2v + x_3w.$$

This is because if

$$z = x_1u + x_2v + x_3w = y_1u + y_2v + y_3w,$$

then by using our (linear!) operations on vectors, we get

$$(y_1 - x_1)u + (y_2 - x_2)v + (y_3 - x_3)w = 0,$$

which implies that

$$y_1 - x_1 = y_2 - x_2 = y_3 - x_3 = 0,$$

by linear independence. Thus,

$$y_1 = x_1, \quad y_2 = x_2, \quad y_3 = x_3,$$

which shows that z has a unique expression as a linear combination, as claimed. Then, our equation

$$x_1u + x_2v + x_3w = b$$

has a *unique solution*, and indeed, we can check that

$$\begin{aligned} x_1 &= 1.4 \\ x_2 &= -0.4 \\ x_3 &= -0.4 \end{aligned}$$

is the solution.

But then, *how do we determine that some vectors are linearly independent?*

One answer is to compute the *determinant* $\det(u, v, w)$, and to check that it is nonzero. In our case,

$$\det(u, v, w) = \begin{vmatrix} 1 & 2 & -1 \\ 2 & 1 & 1 \\ 1 & -2 & -2 \end{vmatrix} = 15,$$

which confirms that u, v, w are linearly independent.

Other methods consist of computing an LU-decomposition or a QR-decomposition, or an SVD of the *matrix* consisting of the three columns u, v, w,

$$A = \begin{pmatrix} u & v & w \end{pmatrix} = \begin{pmatrix} 1 & 2 & -1 \\ 2 & 1 & 1 \\ 1 & -2 & -2 \end{pmatrix}.$$

If we form the vector of unknowns

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix},$$

then our linear combination $x_1u + x_2v + x_3w$ can be written in matrix form as

$$x_1u + x_2v + x_3w = \begin{pmatrix} 1 & 2 & -1 \\ 2 & 1 & 1 \\ 1 & -2 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix},$$

so our linear system is expressed by

$$\begin{pmatrix} 1 & 2 & -1 \\ 2 & 1 & 1 \\ 1 & -2 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix},$$

or more concisely as

$$Ax = b.$$

Now, what if the vectors u, v, w are *linearly dependent*? For example, if we consider the vectors

$$u = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} \qquad v = \begin{pmatrix} 2 \\ 1 \\ -1 \end{pmatrix} \qquad w = \begin{pmatrix} -1 \\ 1 \\ 2 \end{pmatrix},$$

we see that

$$u - v = w,$$

a nontrivial *linear dependence*. It can be verified that u and v are still linearly independent. Now, for our problem

$$x_1 u + x_2 v + x_3 w = b$$

to have a solution, it must be the case that b can be expressed as linear combination of u and v. However, it turns out that u, v, b are linearly independent (because $\det(u, v, b) = -6$), so b cannot be expressed as a linear combination of u and v and thus, our system has *no* solution.

If we change the vector b to

$$b = \begin{pmatrix} 3 \\ 3 \\ 0 \end{pmatrix},$$

then

$$b = u + v,$$

and so the system

$$x_1 u + x_2 v + x_3 w = b$$

has the solution

$$x_1 = 1, \quad x_2 = 1, \quad x_3 = 0.$$

Actually, since $w = u - v$, the above system is equivalent to

$$(x_1 + x_3)u + (x_2 - x_3)v = b,$$

and because u and v are linearly independent, the unique solution in $x_1 + x_3$ and $x_2 - x_3$ is

$$\begin{aligned} x_1 + x_3 &= 1 \\ x_2 - x_3 &= 1, \end{aligned}$$

which yields an infinite number of solutions parameterized by x_3, namely

$$\begin{aligned} x_1 &= 1 - x_3 \\ x_2 &= 1 + x_3. \end{aligned}$$

In summary, a 3×3 linear system may have a unique solution, no solution, or an infinite number of solutions, depending on the linear independence (and dependence) or the vectors

u, v, w, b. This situation can be generalized to any $n \times n$ system, and even to any $n \times m$ system (n equations in m variables), as we will see later.

The point of view where our linear system is expressed in matrix form as $Ax = b$ stresses the fact that the map $x \mapsto Ax$ is a *linear transformation*. This means that

$$A(\lambda x) = \lambda(Ax)$$

for all $x \in \mathbb{R}^{3\times 1}$ and all $\lambda \in \mathbb{R}$ and that

$$A(u + v) = Au + Av,$$

for all $u, v \in \mathbb{R}^{3\times 1}$. We can view the matrix A as a way of expressing a linear map from $\mathbb{R}^{3\times 1}$ to $\mathbb{R}^{3\times 1}$ and solving the system $Ax = b$ amounts to determining whether b belongs to the image of this linear map.

Yet another fruitful way of interpreting the resolution of the system $Ax = b$ is to view this problem as an intersection problem. Indeed, each of the equations

$$\begin{aligned} x_1 + 2x_2 - x_3 &= 1 \\ 2x_1 + x_2 + x_3 &= 2 \\ x_1 - 2x_2 - 2x_3 &= 3 \end{aligned}$$

defines a subset of $\mathbb{R}^3$ which is actually a *plane*. The first equation

$$x_1 + 2x_2 - x_3 = 1$$

defines the plane H_1 passing through the three points $(1, 0, 0), (0, 1/2, 0), (0, 0, -1)$, on the coordinate axes, the second equation

$$2x_1 + x_2 + x_3 = 2$$

defines the plane H_2 passing through the three points $(1, 0, 0), (0, 2, 0), (0, 0, 2)$, on the coordinate axes, and the third equation

$$x_1 - 2x_2 - 2x_3 = 3$$

defines the plane H_3 passing through the three points $(3, 0, 0), (0, -3/2, 0), (0, 0, -3/2)$, on the coordinate axes. The intersection $H_i \cap H_j$ of any two distinct planes H_i and H_j is a line, and the intersection $H_1 \cap H_2 \cap H_3$ of the three planes consists of the single point $(1.4, -0.4, -0.4)$. Under this interpretation, observe that we are focusing on the *rows* of the matrix A, rather than on its *columns*, as in the previous interpretations.

Another great example of a real-world problem where linear algebra proves to be very effective is the problem of *data compression*, that is, of representing a very large data set using a much smaller amount of storage.

Typically the data set is represented as an $m \times n$ matrix A where each row corresponds to an n-dimensional data point and typically, $m \geq n$. In most applications, the data are not independent so the rank of A is a lot smaller than $\min\{m, n\}$, and the the goal of *low-rank decomposition* is to factor A as the product of two matrices B and C, where B is a $m \times k$ matrix and C is a $k \times n$ matrix, with $k \ll \min\{m, n\}$ (here, $\ll$ means "much smaller than"):

$$\begin{pmatrix} \\ A \\ m \times n \\ \\ \end{pmatrix} = \begin{pmatrix} \\ B \\ m \times k \\ \\ \end{pmatrix} \begin{pmatrix} \\ C \\ k \times n \\ \\ \end{pmatrix}$$

Now, it is generally too costly to find an exact factorization as above, so we look for a low-rank matrix A' which is a "good" *approximation* of A. In order to make this statement precise, we need to define a mechanism to determine how close two matrices are. This can be done using *matrix norms*, a notion discussed in Chapter 6. The norm of a matrix A is a nonnegative real number $\|A\|$ which behaves a lot like the absolute value $|x|$ of a real number x. Then, our goal is to find some low-rank matrix A' that minimizes the norm

$$\|A - A'\|^2,$$

over all matrices A' of rank at most k, for some given $k \ll \min\{m, n\}$.

Some advantages of a low-rank approximation are:

1. Fewer elements are required to represent A; namely, $k(m + n)$ instead of mn. Thus less storage and fewer operations are needed to reconstruct A.

2. Often, the process for obtaining the decomposition exposes the underlying structure of the data. Thus, it may turn out that "most" of the significant data are concentrated along some directions called *principal directions*.

Low-rank decompositions of a set of data have a multitude of applications in engineering, including computer science (especially computer vision), statistics, and machine learning. As we will see later in Chapter 16, the *singular value decomposition* (SVD) provides a very satisfactory solution to the low-rank approximation problem. Still, in many cases, the data sets are so large that another ingredient is needed: *randomization*. However, as a first step, linear algebra often yields a good initial solution.

We will now be more precise as to what kinds of operations are allowed on vectors. In the early 1900, the notion of a *vector space* emerged as a convenient and unifying framework for working with "linear" objects and we will discuss this notion in the next few sections.

1.2 Vector Spaces

A (real) vector space is a set E together with two operations, $+\colon E \times E \to E$ and $\cdot\colon \mathbb{R} \times E \to E$, called *addition* and *scalar multiplication*, that satisfy some simple properties. First of all, E under addition has to be a commutative (or abelian) group, a notion that we review next.

However, keep in mind that vector spaces are not just algebraic objects; they are also geometric objects.

Definition 1.1. A *group* is a set G equipped with a binary operation $\cdot\colon G \times G \to G$ that associates an element $a \cdot b \in G$ to every pair of elements $a, b \in G$, and having the following properties: $\cdot$ is associative, has an identity element $e \in G$, and every element in G is invertible (w.r.t. $\cdot$). More explicitly, this means that the following equations hold for all $a, b, c \in G$:

(G1) $a \cdot (b \cdot c) = (a \cdot b) \cdot c$. (associativity);

(G2) $a \cdot e = e \cdot a = a$. (identity);

(G3) For every $a \in G$, there is some $a^{-1} \in G$ such that $a \cdot a^{-1} = a^{-1} \cdot a = e$. (inverse).

A group G is *abelian* (or *commutative*) if

$$a \cdot b = b \cdot a \quad \text{for all } a, b \in G.$$

A set M together with an operation $\cdot\colon M \times M \to M$ and an element e satisfying only conditions (G1) and (G2) is called a *monoid*. For example, the set $\mathbb{N} = \{0, 1, \ldots, n, \ldots\}$ of natural numbers is a (commutative) monoid under addition. However, it is not a group.

Some examples of groups are given below.

Example 1.1.

1. The set $\mathbb{Z} = \{\ldots, -n, \ldots, -1, 0, 1, \ldots, n, \ldots\}$ of integers is a group under addition, with identity element 0. However, $\mathbb{Z}^* = \mathbb{Z} - \{0\}$ is not a group under multiplication.

2. The set $\mathbb{Q}$ of rational numbers (fractions p/q with $p, q \in \mathbb{Z}$ and $q \neq 0$) is a group under addition, with identity element 0. The set $\mathbb{Q}^* = \mathbb{Q} - \{0\}$ is also a group under multiplication, with identity element 1.

3. Similarly, the sets $\mathbb{R}$ of real numbers and $\mathbb{C}$ of complex numbers are groups under addition (with identity element 0), and $\mathbb{R}^* = \mathbb{R} - \{0\}$ and $\mathbb{C}^* = \mathbb{C} - \{0\}$ are groups under multiplication (with identity element 1).

4. The sets $\mathbb{R}^n$ and $\mathbb{C}^n$ of n-tuples of real or complex numbers are groups under componentwise addition:

$$(x_1, \ldots, x_n) + (y_1, \ldots, y_n) = (x_1 + y_1, \ldots, x_n + y_n),$$

with identity element $(0, \ldots, 0)$. All these groups are abelian.

5. Given any nonempty set S, the set of bijections $f\colon S \to S$, also called *permutations of* S, is a group under function composition (i.e., the multiplication of f and g is the composition $g \circ f$), with identity element the identity function id_S. This group is not abelian as soon as S has more than two elements.

6. The set of $n \times n$ matrices with real (or complex) coefficients is a group under addition of matrices, with identity element the null matrix. It is denoted by $\mathrm{M}_n(\mathbb{R})$ (or $\mathrm{M}_n(\mathbb{C})$).

7. The set $\mathbb{R}[X]$ of all polynomials in one variable X with real coefficients,
$$P(X) = a_n X^n + a_{n-1}X^{n-1} + \cdots + a_1 X + a_0,$$
(with $a_i \in \mathbb{R}$), is a group under addition of polynomials.

8. The set of $n \times n$ invertible matrices with real (or complex) coefficients is a group under matrix multiplication, with identity element the identity matrix I_n. This group is called the *general linear group* and is usually denoted by $\mathbf{GL}(n, \mathbb{R})$ (or $\mathbf{GL}(n, \mathbb{C})$).

9. The set of $n \times n$ invertible matrices with real (or complex) coefficients and determinant $+1$ is a group under matrix multiplication, with identity element the identity matrix I_n. This group is called the *special linear group* and is usually denoted by $\mathbf{SL}(n, \mathbb{R})$ (or $\mathbf{SL}(n, \mathbb{C})$).

10. The set of $n \times n$ invertible matrices with real coefficients such that $RR^\top = R^\top R = I_n$ and of determinant $+1$ is a group called the *special orthogonal group* and is usually denoted by $\mathbf{SO}(n)$ (where $R^\top$ is the *transpose* of the matrix R, i.e., the rows of $R^\top$ are the columns of R). It corresponds to the rotations in $\mathbb{R}^n$.

11. Given an open interval (a, b), the set $\mathcal{C}(a, b)$ of continuous functions $f\colon (a, b) \to \mathbb{R}$ is a group under the operation $f + g$ defined such that
$$(f + g)(x) = f(x) + g(x)$$
for all $x \in (a, b)$.

It is customary to denote the operation of an abelian group G by $+$, in which case the inverse a^{-1} of an element $a \in G$ is denoted by $-a$.

The identity element of a group is *unique*. In fact, we can prove a more general fact:

Fact 1. If a binary operation $\cdot\colon M \times M \to M$ is associative and if $e' \in M$ is a left identity and $e'' \in M$ is a right identity, which means that
$$e' \cdot a = a \quad \text{for all} \quad a \in M \tag{G2l}$$
and
$$a \cdot e'' = a \quad \text{for all} \quad a \in M, \tag{G2r}$$
then $e' = e''$.

Proof. If we let $a = e''$ in equation (G2l), we get

$$e' \cdot e'' = e'',$$

and if we let $a = e'$ in equation (G2r), we get

$$e' \cdot e'' = e',$$

and thus

$$e' = e' \cdot e'' = e'',$$

as claimed. □

Fact 1 implies that the identity element of a monoid is unique, and since every group is a monoid, the identity element of a group is unique. Furthermore, every element in a group has a *unique inverse*. This is a consequence of a slightly more general fact:

Fact 2. In a monoid M with identity element e, if some element $a \in M$ has some left inverse $a' \in M$ and some right inverse $a'' \in M$, which means that

$$a' \cdot a = e \tag{G3l}$$

and

$$a \cdot a'' = e, \tag{G3r}$$

then $a' = a''$.

Proof. Using (G3l) and the fact that e is an identity element, we have

$$(a' \cdot a) \cdot a'' = e \cdot a'' = a''.$$

Similarly, Using (G3r) and the fact that e is an identity element, we have

$$a' \cdot (a \cdot a'') = a' \cdot e = a'.$$

However, since M is monoid, the operation $\cdot$ is associative, so

$$a' = a' \cdot (a \cdot a'') = (a' \cdot a) \cdot a'' = a'',$$

as claimed. □

Remark: Axioms (G2) and (G3) can be weakened a bit by requiring only (G2r) (the existence of a right identity) and (G3r) (the existence of a right inverse for every element) (or (G2l) and (G3l)). It is a good exercise to prove that the group axioms (G2) and (G3) follow from (G2r) and (G3r).

Vector spaces are defined as follows.

Definition 1.2. A *real vector space* is a set E (of vectors) together with two operations $+\colon E \times E \to E$ (called *vector addition*)[1] and $\cdot\colon \mathbb{R} \times E \to E$ (called *scalar multiplication*) satisfying the following conditions for all $\alpha, \beta \in \mathbb{R}$ and all $u, v \in E$;

(V0) E is an abelian group w.r.t. $+$, with identity element 0;[2]

(V1) $\alpha \cdot (u + v) = (\alpha \cdot u) + (\alpha \cdot v)$;

(V2) $(\alpha + \beta) \cdot u = (\alpha \cdot u) + (\beta \cdot u)$;

(V3) $(\alpha * \beta) \cdot u = \alpha \cdot (\beta \cdot u)$;

(V4) $1 \cdot u = u$.

In (V3), $*$ denotes multiplication in $\mathbb{R}$.

Given $\alpha \in \mathbb{R}$ and $v \in E$, the element $\alpha \cdot v$ is also denoted by αv. The field $\mathbb{R}$ is often called the field of scalars.

In Definition 1.2, the field $\mathbb{R}$ may be replaced by the field of complex numbers $\mathbb{C}$, in which case we have a *complex* vector space. It is even possible to replace $\mathbb{R}$ by the field of rational numbers $\mathbb{Q}$ or by any other field K (for example $\mathbb{Z}/p\mathbb{Z}$, where p is a prime number), in which case we have a *K-vector space* (in (V3), $*$ denotes multiplication in the field K). In most cases, the field K will be the field $\mathbb{R}$ of reals.

From (V0), a vector space always contains the null vector 0, and thus is nonempty. From (V1), we get $\alpha \cdot 0 = 0$, and $\alpha \cdot (-v) = -(\alpha \cdot v)$. From (V2), we get $0 \cdot v = 0$, and $(-\alpha) \cdot v = -(\alpha \cdot v)$.

Another important consequence of the axioms is the following fact: For any $u \in E$ and any $\lambda \in \mathbb{R}$, if $\lambda \neq 0$ and $\lambda \cdot u = 0$, then $u = 0$.

Indeed, since $\lambda \neq 0$, it has a multiplicative inverse λ^{-1}, so from $\lambda \cdot u = 0$, we get

$$\lambda^{-1} \cdot (\lambda \cdot u) = \lambda^{-1} \cdot 0.$$

However, we just observed that $\lambda^{-1} \cdot 0 = 0$, and from (V3) and (V4), we have

$$\lambda^{-1} \cdot (\lambda \cdot u) = (\lambda^{-1}\lambda) \cdot u = 1 \cdot u = u,$$

and we deduce that $u = 0$.

Remark: One may wonder whether axiom (V4) is really needed. Could it be derived from the other axioms? The answer is **no**. For example, one can take $E = \mathbb{R}^n$ and define $\cdot\colon \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ by

$$\lambda \cdot (x_1, \ldots, x_n) = (0, \ldots, 0)$$

[1] The symbol $+$ is overloaded, since it denotes both addition in the field $\mathbb{R}$ and addition of vectors in E. It is usually clear from the context which $+$ is intended.

[2] The symbol 0 is also overloaded, since it represents both the zero in $\mathbb{R}$ (a scalar) and the identity element of E (the zero vector). Confusion rarely arises, but one may prefer using $\mathbf{0}$ for the zero vector.

for all $(x_1, \ldots, x_n) \in \mathbb{R}^n$ and all $\lambda \in \mathbb{R}$. Axioms (V0)–(V3) are all satisfied, but (V4) fails. Less trivial examples can be given using the notion of a basis, which has not been defined yet.

The field $\mathbb{R}$ itself can be viewed as a vector space over itself, addition of vectors being addition in the field, and multiplication by a scalar being multiplication in the field.

Example 1.2.

1. The fields $\mathbb{R}$ and $\mathbb{C}$ are vector spaces over $\mathbb{R}$.

2. The groups $\mathbb{R}^n$ and $\mathbb{C}^n$ are vector spaces over $\mathbb{R}$, with salar multiplication given by

$$\lambda(x_1, \ldots, x_n) = (\lambda x_1, \ldots, \lambda x_n),$$

   for any $\lambda \in \mathbb{R}$ and with $(x_1, \ldots, x_n) \in \mathbb{R}^n$ or $(x_1, \ldots, x_n) \in \mathbb{C}^n$, and $\mathbb{C}^n$ is a vector space over $\mathbb{C}$ with scalar multiplication as above, but with $\lambda \in \mathbb{C}$.

3. The ring $\mathbb{R}[X]_n$ of polynomials of degree at most n with real coefficients is a vector space over $\mathbb{R}$, and the ring $\mathbb{C}[X]_n$ of polynomials of degree at most n with complex coefficients is a vector space over $\mathbb{C}$, with scalar multiplication $\lambda \cdot P(X)$ of a polynomial

$$P(X) = a_m X^m + a_{m-1} X^{m-1} + \cdots + a_1 X + a_0$$

   (with $a_i \in \mathbb{R}$ or $a_i \in \mathbb{C}$) by the scalar λ (in $\mathbb{R}$ or $\mathbb{C}$), with $m \leq n$, given by

$$\lambda \cdot P(X) = \lambda a_m X^m + \lambda a_{m-1} X^{m-1} + \cdots + \lambda a_1 X + \lambda a_0.$$

4. The ring $\mathbb{R}[X]$ of all polynomials with real coefficients is a vector space over $\mathbb{R}$, and the ring $\mathbb{C}[X]$ of all polynomials with complex coefficients is a vector space over $\mathbb{C}$, with the same scalar multiplication as above.

5. The ring of $n \times n$ matrices $\mathrm{M}_n(\mathbb{R})$ is a vector space over $\mathbb{R}$.

6. The ring of $m \times n$ matrices $\mathrm{M}_{m,n}(\mathbb{R})$ is a vector space over $\mathbb{R}$.

7. The ring $\mathcal{C}(a, b)$ of continuous functions $f\colon (a, b) \to \mathbb{R}$ is a vector space over $\mathbb{R}$, with the scalar multiplication λf of a function $f\colon (a, b) \to \mathbb{R}$ by a scalar $\lambda \in \mathbb{R}$ given by

$$(\lambda f)(x) = \lambda f(x), \qquad \text{for all } x \in (a, b).$$

Let E be a vector space. We would like to define the important notions of linear combination and linear independence.

Before defining these notions, we need to discuss a strategic choice which, depending how it is settled, may reduce or increase headackes in dealing with notions such as linear combinations and linear dependence (or independence). The issue has to do with using sets of vectors versus sequences of vectors.

1.3 Indexed Families; the Sum Notation $\sum_{i \in I} a_i$

Our experience tells us that *it is preferable to use sequences of vectors*; even better, indexed families of vectors. (We are not alone in having opted for sequences over sets, and we are in good company; for example, Artin [6], Axler [8], and Lang [62] use sequences. Nevertheless, some prominent authors such as Lax [66] use sets. We leave it to the reader to conduct a survey on this issue.)

Given a set A, recall that a *sequence* is an ordered n-tuple $(a_1, \ldots, a_n) \in A^n$ of elements from A, for some natural number n. The elements of a sequence need not be distinct and the order is important. For example, (a_1, a_2, a_1) and (a_2, a_1, a_1) are two distinct sequences in A^3. Their underlying set is $\{a_1, a_2\}$.

What we just defined are *finite* sequences, which can also be viewed as functions from $\{1, 2, \ldots, n\}$ to the set A; the ith element of the sequence $(a_1, \ldots, a_n)$ is the image of i under the function. This viewpoint is fruitful, because it allows us to define (countably) infinite sequences as functions $s \colon \mathbb{N} \to A$. But then, why limit ourselves to ordered sets such as $\{1, \ldots, n\}$ or $\mathbb{N}$ as index sets?

The main role of the index set is to tag each element uniquely, and the order of the tags is not crucial, although convenient. Thus, it is natural to define an *I-indexed family* of elements of A, for short a *family*, as a function $a \colon I \to A$ where I is any set viewed as an index set. Since the function a is determined by its graph

$$\{(i, a(i)) \mid i \in I\},$$

the family a can be viewed as the set of pairs $a = \{(i, a(i)) \mid i \in I\}$. For notational simplicity, we write a_i instead of $a(i)$, and denote the family $a = \{(i, a(i)) \mid i \in I\}$ by $(a_i)_{i \in I}$. For example, if $I = \{r, g, b, y\}$ and $A = \mathbb{N}$, the set of pairs

$$a = \{(r, 2), (g, 3), (b, 2), (y, 11)\}$$

is an indexed family. The element 2 appears twice in the family with the two distinct tags r and b.

When the indexed set I is totally ordered, a family $(a_i)_{i \in I}$ often called an I-sequence. Interestingly, sets can be viewed as special cases of families. Indeed, a set A can be viewed as the A-indexed family $\{(a, a) \mid a \in I\}$ corresponding to the identity function.

Remark: An indexed family should not be confused with a multiset. Given any set A, a *multiset* is a similar to a set, except that elements of A may occur more than once. For example, if $A = \{a, b, c, d\}$, then $\{a, a, a, b, c, c, d, d\}$ is a multiset. Each element appears with a certain multiplicity, but the order of the elements does not matter. For example, a has multiplicity 3. Formally, a multiset is a function $s \colon A \to \mathbb{N}$, or equivalently a set of pairs $\{(a, i) \mid a \in A\}$. Thus, a multiset is an A-indexed family of elements from $\mathbb{N}$, but not a $\mathbb{N}$-indexed family, since distinct elements may have the same multiplicity (such as c an d in

the example above). An indexed family is a generalization of a sequence, but a multiset is a generalization of a set.

We also need to take care of an annoying technicality, which is to define sums of the form $\sum_{i\in I} a_i$, where I is any finite index set and $(a_i)_{i\in I}$ is a family of elements in some set A equiped with a binary operation $+\colon A \times A \to A$ which is associative (axiom (G1)) and commutative. This will come up when we define linear combinations.

The issue is that the binary operation $+$ only tells us how to compute $a_1 + a_2$ for two elements of A, but it does not tell us what is the sum of three of more elements. For example, how should $a_1 + a_2 + a_3$ be defined?

What we have to do is to define $a_1+a_2+a_3$ by using a sequence of steps each involving two elements, and there are two possible ways to do this: $a_1 + (a_2 + a_3)$ and $(a_1 + a_2) + a_3$. If our operation $+$ is not associative, these are different values. If it associative, then $a_1+(a_2+a_3) = (a_1 + a_2) + a_3$, but then there are still six possible permutations of the indices $1, 2, 3$, and if $+$ is not commutative, these values are generally different. If our operation is commutative, then all six permutations have the same value. Thus, if $+$ is associative and commutative, it seems intuitively clear that a sum of the form $\sum_{i\in I} a_i$ does not depend on the order of the operations used to compute it.

This is indeed the case, but a rigorous proof requires induction, and such a proof is surprisingly involved. Readers may accept without proof the fact that sums of the form $\sum_{i\in I} a_i$ are indeed well defined, and jump directly to Definition 1.3. For those who want to see the gory details, here we go.

First, we define sums $\sum_{i\in I} a_i$, where I is a finite sequence of distinct natural numbers, say $I = (i_1, \ldots, i_m)$. If $I = (i_1, \ldots, i_m)$ with $m \geq 2$, we denote the sequence $(i_2, \ldots, i_m)$ by $I - \{i_1\}$. We proceed by induction on the size m of I. Let

$$\sum_{i\in I} a_i = a_{i_1}, \quad \text{if } m = 1,$$
$$\sum_{i\in I} a_i = a_{i_1} + \Big(\sum_{i\in I-\{i_1\}} a_i\Big), \quad \text{if } m > 1.$$

For example, if $I = (1, 2, 3, 4)$, we have

$$\sum_{i\in I} a_i = a_1 + (a_2 + (a_3 + a_4)).$$

If the operation $+$ is not associative, the grouping of the terms matters. For instance, in general

$$a_1 + (a_2 + (a_3 + a_4)) \neq (a_1 + a_2) + (a_3 + a_4).$$

However, if the operation $+$ is associative, the sum $\sum_{i\in I} a_i$ should not depend on the grouping of the elements in I, as long as their order is preserved. For example, if $I = (1, 2, 3, 4, 5)$,

$J_1 = (1, 2)$, and $J_2 = (3, 4, 5)$, we expect that

$$\sum_{i\in I} a_i = \Big(\sum_{j\in J_1} a_j\Big) + \Big(\sum_{j\in J_2} a_j\Big).$$

This indeed the case, as we have the following proposition.

Proposition 1.1. *Given any nonempty set A equipped with an associative binary operation $+\colon A \times A \to A$, for any nonempty finite sequence I of distinct natural numbers and for any partition of I into p nonempty sequences $I_{k_1}, \ldots, I_{k_p}$, for some nonempty sequence $K = (k_1, \ldots, k_p)$ of distinct natural numbers such that $k_i < k_j$ implies that $\alpha < \beta$ for all $\alpha \in I_{k_i}$ and all $\beta \in I_{k_j}$, for every sequence $(a_i)_{i\in I}$ of elements in A, we have*

$$\sum_{\alpha\in I} a_\alpha = \sum_{k\in K}\Big(\sum_{\alpha\in I_k} a_\alpha\Big).$$

Proof. We proceed by induction on the size n of I.

If $n = 1$, then we must have $p = 1$ and $I_{k_1} = I$, so the proposition holds trivially.

Next, assume $n > 1$. If $p = 1$, then $I_{k_1} = I$ and the formula is trivial, so assume that $p \geq 2$ and write $J = (k_2, \ldots, k_p)$. There are two cases.

Case 1. The sequence I_{k_1} has a single element, say β, which is the first element of I. In this case, write C for the sequence obtained from I by deleting its first element β. By definition,

$$\sum_{\alpha\in I} a_\alpha = a_\beta + \Big(\sum_{\alpha\in C} a_\alpha\Big),$$

and

$$\sum_{k\in K}\Big(\sum_{\alpha\in I_k} a_\alpha\Big) = a_\beta + \Big(\sum_{j\in J}\Big(\sum_{\alpha\in I_j} a_\alpha\Big)\Big).$$

Since $|C| = n - 1$, by the induction hypothesis, we have

$$\Big(\sum_{\alpha\in C} a_\alpha\Big) = \sum_{j\in J}\Big(\sum_{\alpha\in I_j} a_\alpha\Big),$$

which yields our identity.

Case 2. The sequence I_{k_1} has at least two elements. In this case, let β be the first element of I (and thus of I_{k_1}), let I' be the sequence obtained from I by deleting its first element β, let I'_{k_1} be the sequence obtained from I_{k_1} by deleting its first element β, and let $I'_{k_i} = I_{k_i}$ for $i = 2, \ldots, p$. Recall that $J = (k_2, \ldots, k_p)$ and $K = (k_1, \ldots, k_p)$. The sequence I' has $n - 1$ elements, so by the induction hypothesis applied to I' and the I'_{k_i}, we get

$$\sum_{\alpha\in I'} a_\alpha = \sum_{k\in K}\Big(\sum_{\alpha\in I'_k} a_\alpha\Big) = \Big(\sum_{\alpha\in I'_{k_1}} a_\alpha\Big) + \Big(\sum_{j\in J}\Big(\sum_{\alpha\in I_j} a_\alpha\Big)\Big).$$

If we add the lefthand side to a_β, by definition we get

$$\sum_{\alpha\in I} a_\alpha.$$

If we add the righthand side to a_β, using associativity and the definition of an indexed sum, we get

$$\begin{aligned} a_\beta + \Big(\Big(\sum_{\alpha\in I'_{k_1}} a_\alpha\Big) + \Big(\sum_{j\in J}\Big(\sum_{\alpha\in I_j} a_\alpha\Big)\Big)\Big) &= \Big(a_\beta + \Big(\sum_{\alpha\in I'_{k_1}} a_\alpha\Big)\Big) + \Big(\sum_{j\in J}\Big(\sum_{\alpha\in I_j} a_\alpha\Big)\Big) \\ &= \Big(\sum_{\alpha\in I_{k_1}} a_\alpha\Big) + \Big(\sum_{j\in J}\Big(\sum_{\alpha\in I_j} a_\alpha\Big)\Big) \\ &= \sum_{k\in K}\Big(\sum_{\alpha\in I_k} a_\alpha\Big), \end{aligned}$$

as claimed. □

If $I = (1, \ldots, n)$, we also write $\sum_{i=1}^n a_i$ instead of $\sum_{i\in I} a_i$. Since $+$ is associative, Proposition 1.1 shows that the sum $\sum_{i=1}^n a_i$ is independent of the grouping of its elements, which justifies the use the notation $a_1 + \cdots + a_n$ (without any parentheses).

If we also assume that our associative binary operation on A is commutative, then we can show that the sum $\sum_{i\in I} a_i$ does not depend on the ordering of the index set I.

Proposition 1.2. *Given any nonempty set A equipped with an associative and commutative binary operation $+\colon A \times A \to A$, for any two nonempty finite sequences I and J of distinct natural numbers such that J is a permutation of I (in other words, the unlerlying sets of I and J are identical), for every sequence $(a_i)_{i\in I}$ of elements in A, we have*

$$\sum_{\alpha\in I} a_\alpha = \sum_{\alpha\in J} a_\alpha.$$

Proof. We proceed by induction on the number p of elements in I. If $p = 1$, we have $I = J$ and the proposition holds trivially.

If $p > 1$, to simplify notation, assume that $I = (1, \ldots, p)$ and that J is a permutation $(i_1, \ldots, i_p)$ of I. First, assume that $2 \leq i_1 \leq p-1$, let J' be the sequence obtained from J by deleting i_1, I' be the sequence obtained from I by deleting i_1, and let $P = (1, 2, \ldots, i_1-1)$ and $Q = (i_1+1, \ldots, p-1, p)$. Observe that the sequence I' is the concatenation of the sequences P and Q. By the induction hypothesis applied to J' and I', and then by Proposition 1.1 applied to I' and its partition (P, Q), we have

$$\sum_{\alpha\in J'} a_\alpha = \sum_{\alpha\in I'} a_\alpha = \Big(\sum_{i=1}^{i_1-1} a_i\Big) + \Big(\sum_{i=i_1+1}^{p} a_i\Big).$$

If we add the lefthand side to a_{i_1}, by definition we get

$$\sum_{\alpha \in J} a_\alpha.$$

If we add the righthand side to a_{i_1}, we get

$$a_{i_1} + \left(\left(\sum_{i=1}^{i_1 - 1} a_i \right) + \left(\sum_{i=i_1+1}^{p} a_i \right) \right).$$

Using associativity, we get

$$a_{i_1} + \left(\left(\sum_{i=1}^{i_1 - 1} a_i \right) + \left(\sum_{i=i_1+1}^{p} a_i \right) \right) = \left(a_{i_1} + \left(\sum_{i=1}^{i_1 - 1} a_i \right) \right) + \left(\sum_{i=i_1+1}^{p} a_i \right),$$

then using associativity and commutativity several times (more rigorously, using induction on $i_1 - 1$), we get

$$\begin{aligned} \left(a_{i_1} + \left(\sum_{i=1}^{i_1 - 1} a_i \right) \right) + \left(\sum_{i=i_1+1}^{p} a_i \right) &= \left(\sum_{i=1}^{i_1 - 1} a_i \right) + a_{i_1} + \left(\sum_{i=i_1+1}^{p} a_i \right) \\ &= \sum_{i=1}^{p} a_i, \end{aligned}$$

as claimed.

The cases where $i_1 = 1$ or $i_1 = p$ are treated similarly, but in a simpler manner since either $P = ()$ or $Q = ()$ (where $()$ denotes the empty sequence). □

Having done all this, we can now make sense of sums of the form $\sum_{i \in I} a_i$, for any finite indexed set I and any family $a = (a_i)_{i \in I}$ of elements in A, where A is a set equipped with a binary operation $+$ which is associative and commutative.

Indeed, since I is finite, it is in bijection with the set $\{1, \ldots, n\}$ for some $n \in \mathbb{N}$, and any total ordering $\preceq$ on I corresponds to a permutation $I_\preceq$ of $\{1, \ldots, n\}$ (where we identify a permutation with its image). For any total ordering $\preceq$ on I, we define $\sum_{i \in I, \preceq} a_i$ as

$$\sum_{i \in I, \preceq} a_i = \sum_{j \in I_\preceq} a_j.$$

Then, for any other total ordering $\preceq'$ on I, we have

$$\sum_{i \in I, \preceq'} a_i = \sum_{j \in I_{\preceq'}} a_j,$$

and since $I_{\preceq}$ and $I_{\preceq'}$ are different permutations of $\{1, \ldots, n\}$, by Proposition 1.2, we have

$$\sum_{j \in I_{\preceq}} a_j = \sum_{j \in I_{\preceq'}} a_j.$$

Therefore, the sum $\sum_{i \in I, \preceq} a_i$ does not depend on the total ordering on I. We define *the* sum $\sum_{i \in I} a_i$ as the common value $\sum_{i \in I, \preceq} a_i$ for all total orderings $\preceq$ of I.

1.4 Linear Independence, Subspaces

One of the most useful properties of vector spaces is that there possess bases. What this means is that in every vector space, E, there is some set of vectors, $\{e_1, \ldots, e_n\}$, such that *every* vector $v \in E$ can be written as a linear combination,

$$v = \lambda_1 e_1 + \cdots + \lambda_n e_n,$$

of the e_i, for some scalars, $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$. Furthermore, the n-tuple, $(\lambda_1, \ldots, \lambda_n)$, as above is unique.

This description is fine when E has a finite basis, $\{e_1, \ldots, e_n\}$, but this is not always the case! For example, the vector space of real polynomials, $\mathbb{R}[X]$, does not have a finite basis but instead it has an infinite basis, namely

$$1,\ X,\ X^2,\ \ldots, X^n,\ \ldots$$

For simplicity, in this chapter, we will restrict our attention to vector spaces that have a finite basis (we say that they are *finite-dimensional*).

Given a set A, recall that an *I-indexed family* $(a_i)_{i \in I}$ of elements of A (for short, a *family*) is a function $a \colon I \to A$, or equivalently a set of pairs $\{(i, a_i) \mid i \in I\}$. We agree that when $I = \emptyset$, $(a_i)_{i \in I} = \emptyset$. A family $(a_i)_{i \in I}$ is finite if I is finite.

Remark: When considering a family $(a_i)_{i \in I}$, there is no reason to assume that I is ordered. The crucial point is that every element of the family is uniquely indexed by an element of I. Thus, unless specified otherwise, we do not assume that the elements of an index set are ordered.

Given two disjoint sets I and J, the union of two families $(u_i)_{i \in I}$ and $(v_j)_{j \in J}$, denoted as $(u_i)_{i \in I} \cup (v_j)_{j \in J}$, is the family $(w_k)_{k \in (I \cup J)}$ defined such that $w_k = u_k$ if $k \in I$, and $w_k = v_k$ if $k \in J$. Given a family $(u_i)_{i \in I}$ and any element v, we denote by $(u_i)_{i \in I} \cup_k (v)$ the family $(w_i)_{i \in I \cup \{k\}}$ defined such that, $w_i = u_i$ if $i \in I$, and $w_k = v$, where k is any index such that $k \notin I$. Given a family $(u_i)_{i \in I}$, a subfamily of $(u_i)_{i \in I}$ is a family $(u_j)_{j \in J}$ where J is any subset of I.

In this chapter, unless specified otherwise, it is assumed that all families of scalars are finite (i.e., their index set is finite).

Definition 1.3. Let E be a vector space. A vector $v \in E$ is a *linear combination of a family* $(u_i)_{i\in I}$ *of elements of* E iff there is a family $(\lambda_i)_{i\in I}$ of scalars in $\mathbb{R}$ such that

$$v = \sum_{i\in I} \lambda_i u_i.$$

When $I = \emptyset$, we stipulate that $v = 0$. (By Proposition 1.2, sums of the form $\sum_{i\in I} \lambda_i u_i$ are well defined.) We say that a family $(u_i)_{i\in I}$ is *linearly independent* iff for every family $(\lambda_i)_{i\in I}$ of scalars in $\mathbb{R}$,

$$\sum_{i\in I} \lambda_i u_i = 0 \quad \text{implies that} \quad \lambda_i = 0 \text{ for all } i \in I.$$

Equivalently, a family $(u_i)_{i\in I}$ is *linearly dependent* iff there is some family $(\lambda_i)_{i\in I}$ of scalars in $\mathbb{R}$ such that

$$\sum_{i\in I} \lambda_i u_i = 0 \quad \text{and} \quad \lambda_j \neq 0 \text{ for some } j \in I.$$

We agree that when $I = \emptyset$, the family $\emptyset$ is linearly independent.

Observe that defining linear combinations for families of vectors rather than for sets of vectors has the advantage that the vectors being combined need not be distinct. For example, for $I = \{1, 2, 3\}$ and the families (u, v, u) and $(\lambda_1, \lambda_2, \lambda_1)$, the linear combination

$$\sum_{i\in I} \lambda_i u_i = \lambda_1 u + \lambda_2 v + \lambda_1 u$$

makes sense. Using sets of vectors in the definition of a linear combination does not allow such linear combinations; this is too restrictive.

Unravelling Definition 1.3, a family $(u_i)_{i\in I}$ is linearly dependent iff either I consists of a single element, say i, and $u_i = 0$, or $|I| \geq 2$ and some u_j in the family can be expressed as a linear combination of the other vectors in the family. Indeed, in the second case, there is some family $(\lambda_i)_{i\in I}$ of scalars in $\mathbb{R}$ such that

$$\sum_{i\in I} \lambda_i u_i = 0 \quad \text{and} \quad \lambda_j \neq 0 \text{ for some } j \in I,$$

and since $|I| \geq 2$, the set $I - \{j\}$ is nonempty and we get

$$u_j = \sum_{i\in(I-\{j\})} -\lambda_j^{-1} \lambda_i u_i.$$

Observe that one of the reasons for defining linear dependence for families of vectors rather than for sets of vectors is that our definition allows multiple occurrences of a vector. This is important because a matrix may contain identical columns, and we would like to say

that these columns are linearly dependent. The definition of linear dependence for sets does not allow us to do that.

The above also shows that a family $(u_i)_{i\in I}$ is linearly independent iff either $I = \emptyset$, or I consists of a single element i and $u_i \neq 0$, or $|I| \geq 2$ and no vector u_j in the family can be expressed as a linear combination of the other vectors in the family.

When I is nonempty, if the family $(u_i)_{i\in I}$ is linearly independent, note that $u_i \neq 0$ for all $i \in I$. Otherwise, if $u_i = 0$ for some $i \in I$, then we get a nontrivial linear dependence $\sum_{i\in I} \lambda_i u_i = 0$ by picking any nonzero λ_i and letting $\lambda_k = 0$ for all $k \in I$ with $k \neq i$, since $\lambda_i 0 = 0$. If $|I| \geq 2$, we must also have $u_i \neq u_j$ for all $i, j \in I$ with $i \neq j$, since otherwise we get a nontrivial linear dependence by picking $\lambda_i = \lambda$ and $\lambda_j = -\lambda$ for any nonzero λ, and letting $\lambda_k = 0$ for all $k \in I$ with $k \neq i, j$.

Thus, the definition of linear independence implies that a nontrivial linearly independent family is actually a set. This explains why certain authors choose to define linear independence for sets of vectors. The problem with this approach is that linear dependence, which is the logical negation of linear independence, is then only defined for sets of vectors. However, as we pointed out earlier, it is really desirable to define linear dependence for families allowing multiple occurrences of the same vector.

Example 1.3.

1. Any two distinct scalars $\lambda, \mu \neq 0$ in $\mathbb{R}$ are linearly dependent.
2. In $\mathbb{R}^3$, the vectors $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$ are linearly independent.
3. In $\mathbb{R}^4$, the vectors $(1, 1, 1, 1)$, $(0, 1, 1, 1)$, $(0, 0, 1, 1)$, and $(0, 0, 0, 1)$ are linearly independent.
4. In $\mathbb{R}^2$, the vectors $u = (1, 1)$, $v = (0, 1)$ and $w = (2, 3)$ are linearly dependent, since

$$w = 2u + v.$$

When I is finite, we often assume that it is the set $I = \{1, 2, \ldots, n\}$. In this case, we denote the family $(u_i)_{i\in I}$ as $(u_1, \ldots, u_n)$.

The notion of a subspace of a vector space is defined as follows.

Definition 1.4. Given a vector space E, a subset F of E is a *linear subspace* (or *subspace*) of E iff F is nonempty and $\lambda u + \mu v \in F$ for all $u, v \in F$, and all $\lambda, \mu \in \mathbb{R}$.

It is easy to see that a subspace F of E is indeed a vector space, since the restriction of $+\colon E \times E \to E$ to $F \times F$ is indeed a function $+\colon F \times F \to F$, and the restriction of $\cdot\colon \mathbb{R} \times E \to E$ to $\mathbb{R} \times F$ is indeed a function $\cdot\colon \mathbb{R} \times F \to F$.

It is also easy to see that any intersection of subspaces is a subspace.

Since F is nonempty, if we pick any vector $u \in F$ and if we let $\lambda = \mu = 0$, then $\lambda u + \mu u = 0u + 0u = 0$, so every subspace contains the vector 0. For any nonempty finite index set I, one can show by induction on the cardinality of I that if $(u_i)_{i \in I}$ is any family of vectors $u_i \in F$ and $(\lambda_i)_{i \in I}$ is any family of scalars, then $\sum_{i \in I} \lambda_i u_i \in F$.

The subspace $\{0\}$ will be denoted by (0), or even 0 (with a mild abuse of notation).

Example 1.4.

1. In $\mathbb{R}^2$, the set of vectors $u = (x, y)$ such that

$$x + y = 0$$

is a subspace.

2. In $\mathbb{R}^3$, the set of vectors $u = (x, y, z)$ such that

$$x + y + z = 0$$

is a subspace.

3. For any $n \geq 0$, the set of polynomials $f(X) \in \mathbb{R}[X]$ of degree at most n is a subspace of $\mathbb{R}[X]$.

4. The set of upper triangular $n \times n$ matrices is a subspace of the space of $n \times n$ matrices.

Proposition 1.3. *Given any vector space E, if S is any nonempty subset of E, then the smallest subspace $\langle S \rangle$ (or $\mathrm{Span}(S)$) of E containing S is the set of all (finite) linear combinations of elements from S.*

Proof. We prove that the set $\mathrm{Span}(S)$ of all linear combinations of elements of S is a subspace of E, leaving as an exercise the verification that every subspace containing S also contains $\mathrm{Span}(S)$.

First, $\mathrm{Span}(S)$ is nonempty since it contains S (which is nonempty). If $u = \sum_{i \in I} \lambda_i u_i$ and $v = \sum_{j \in J} \mu_j v_j$ are any two linear combinations in $\mathrm{Span}(S)$, for any two scalars $\lambda, \mu \in \mathbb{R}$,

$$\begin{aligned}
\lambda u + \mu v &= \lambda \sum_{i \in I} \lambda_i u_i + \mu \sum_{j \in J} \mu_j v_j \\
&= \sum_{i \in I} \lambda \lambda_i u_i + \sum_{j \in J} \mu \mu_j v_j \\
&= \sum_{i \in I - J} \lambda \lambda_i u_i + \sum_{i \in I \cap J} (\lambda \lambda_i + \mu \mu_i) u_i + \sum_{j \in J - I} \mu \mu_j v_j,
\end{aligned}$$

which is a linear combination with index set $I \cup J$, and thus $\lambda u + \mu v \in \mathrm{Span}(S)$, which proves that $\mathrm{Span}(S)$ is a subspace. □

One might wonder what happens if we add extra conditions to the coefficients involved in forming linear combinations. Here are three natural restrictions which turn out to be important (as usual, we assume that our index sets are finite):

(1) Consider combinations $\sum_{i\in I} \lambda_i u_i$ for which

$$\sum_{i\in I} \lambda_i = 1.$$

These are called *affine combinations*. One should realize that every linear combination $\sum_{i\in I} \lambda_i u_i$ can be viewed as an affine combination. For example, if k is an index not in I, if we let $J = I \cup \{k\}$, $u_k = 0$, and $\lambda_k = 1 - \sum_{i\in I} \lambda_i$, then $\sum_{j\in J} \lambda_j u_j$ is an affine combination and

$$\sum_{i\in I} \lambda_i u_i = \sum_{j\in J} \lambda_j u_j.$$

However, we get new spaces. For example, in $\mathbb{R}^3$, the set of all affine combinations of the three vectors $e_1 = (1, 0, 0), e_2 = (0, 1, 0)$, and $e_3 = (0, 0, 1)$, is the plane passing through these three points. Since it does not contain $0 = (0, 0, 0)$, it is not a linear subspace.

(2) Consider combinations $\sum_{i\in I} \lambda_i u_i$ for which

$$\lambda_i \geq 0, \quad \text{for all } i \in I.$$

These are called *positive* (or *conic*) *combinations*. It turns out that positive combinations of families of vectors are *cones*. They show up naturally in convex optimization.

(3) Consider combinations $\sum_{i\in I} \lambda_i u_i$ for which we require (1) *and* (2), that is

$$\sum_{i\in I} \lambda_i = 1, \quad \text{and} \quad \lambda_i \geq 0 \quad \text{for all } i \in I.$$

These are called *convex combinations*. Given any finite family of vectors, the set of all convex combinations of these vectors is a *convex polyhedron*. Convex polyhedra play a very important role in convex optimization.

1.5 Bases of a Vector Space

Given a vector space E, given a family $(v_i)_{i\in I}$, the subset V of E consisting of the null vector 0 and of all linear combinations of $(v_i)_{i\in I}$ is easily seen to be a subspace of E. The family $(v_i)_{i\in I}$ is an economical way of representing the entire subspace V, but such a family would be even nicer if it was not redundant. Subspaces having such an "efficient" generating family (called a basis) play an important role, and motivate the following definition.

Definition 1.5. Given a vector space E and a subspace V of E, a family $(v_i)_{i\in I}$ of vectors $v_i \in V$ *spans* V or *generates* V iff for every $v \in V$, there is some family $(\lambda_i)_{i\in I}$ of scalars in $\mathbb{R}$ such that

$$v = \sum_{i\in I} \lambda_i v_i.$$

We also say that the elements of $(v_i)_{i\in I}$ are *generators* of V and that V is *spanned by* $(v_i)_{i\in I}$, or *generated by* $(v_i)_{i\in I}$. If a subspace V of E is generated by a finite family $(v_i)_{i\in I}$, we say that V is *finitely generated*. A family $(u_i)_{i\in I}$ that spans V and is linearly independent is called a *basis* of V.

Example 1.5.

1. In $\mathbb{R}^3$, the vectors $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$ form a basis.
2. The vectors $(1, 1, 1, 1), (1, 1, -1, -1), (1, -1, 0, 0), (0, 0, 1, -1)$ form a basis of $\mathbb{R}^4$ known as the *Haar basis*. This basis and its generalization to dimension 2^n are crucial in wavelet theory.
3. In the subspace of polynomials in $\mathbb{R}[X]$ of degree at most n, the polynomials $1, X, X^2, \ldots, X^n$ form a basis.
4. The *Bernstein polynomials* $\binom{n}{k}(1-X)^{n-k}X^k$ for $k = 0, \ldots, n$, also form a basis of that space. These polynomials play a major role in the theory of *spline curves*.

The first key result of linear algebra that every vector space E has a basis. We begin with a crucial lemma which formalizes the mechanism for building a basis incrementally.

Lemma 1.4. *Given a linearly independent family $(u_i)_{i\in I}$ of elements of a vector space E, if $v \in E$ is not a linear combination of $(u_i)_{i\in I}$, then the family $(u_i)_{i\in I} \cup_k (v)$ obtained by adding v to the family $(u_i)_{i\in I}$ is linearly independent (where $k \notin I$).*

Proof. Assume that $\mu v + \sum_{i\in I} \lambda_i u_i = 0$, for any family $(\lambda_i)_{i\in I}$ of scalars in $\mathbb{R}$. If $\mu \neq 0$, then μ has an inverse (because $\mathbb{R}$ is a field), and thus we have $v = -\sum_{i\in I}(\mu^{-1}\lambda_i)u_i$, showing that v is a linear combination of $(u_i)_{i\in I}$ and contradicting the hypothesis. Thus, $\mu = 0$. But then, we have $\sum_{i\in I} \lambda_i u_i = 0$, and since the family $(u_i)_{i\in I}$ is linearly independent, we have $\lambda_i = 0$ for all $i \in I$. □

The next theorem holds in general, but the proof is more sophisticated for vector spaces that do not have a finite set of generators. Thus, in this chapter, we only prove the theorem for finitely generated vector spaces.

Theorem 1.5. *Given any finite family $S = (u_i)_{i\in I}$ generating a vector space E and any linearly independent subfamily $L = (u_j)_{j\in J}$ of S (where $J \subseteq I$), there is a basis B of E such that $L \subseteq B \subseteq S$.*

Proof. Consider the set of linearly independent families B such that $L \subseteq B \subseteq S$. Since this set is nonempty and finite, it has some maximal element (that is, a subfamily $B = (u_h)_{h\in H}$ of S with $H \subseteq I$ of maximum cardinality), say $B = (u_h)_{h\in H}$. We claim that B generates E. Indeed, if B does not generate E, then there is some $u_p \in S$ that is not a linear combination of vectors in B (since S generates E), with $p \notin H$. Then, by Lemma 1.4, the family $B' = (u_h)_{h\in H\cup\{p\}}$ is linearly independent, and since $L \subseteq B \subset B' \subseteq S$, this contradicts the maximality of B. Thus, B is a basis of E such that $L \subseteq B \subseteq S$. □

Remark: Theorem 1.5 also holds for vector spaces that are not finitely generated. In this case, the problem is to guarantee the existence of a maximal linearly independent family B such that $L \subseteq B \subseteq S$. The existence of such a maximal family can be shown using Zorn's lemma. A situation where the full generality of Theorem 1.5 is needed is the case of the vector space $\mathbb{R}$ over the field of coefficients $\mathbb{Q}$. The numbers 1 and $\sqrt{2}$ are linearly independent over $\mathbb{Q}$, so according to Theorem 1.5, the linearly independent family $L = (1, \sqrt{2})$ can be extended to a basis B of $\mathbb{R}$. Since $\mathbb{R}$ is uncountable and $\mathbb{Q}$ is countable, such a basis must be uncountable!

The notion of a basis can also be defined in terms of the notion of maximal linearly independent family, and minimal generating family.

Definition 1.6. Let $(v_i)_{i\in I}$ be a family of vectors in a vector space E. We say that $(v_i)_{i\in I}$ a *maximal linearly independent family of E* if it is linearly independent, and if for any vector $w \in E$, the family $(v_i)_{i\in I} \cup_k \{w\}$ obtained by adding w to the family $(v_i)_{i\in I}$ is linearly dependent. We say that $(v_i)_{i\in I}$ a *minimal generating family of E* if it spans E, and if for any index $p \in I$, the family $(v_i)_{i\in I-\{p\}}$ obtained by removing v_p from the family $(v_i)_{i\in I}$ does not span E.

The following proposition giving useful properties characterizing a basis is an immediate consequence of Lemma 1.4.

Proposition 1.6. *Given a vector space E, for any family $B = (v_i)_{i\in I}$ of vectors of E, the following properties are equivalent:*

(1) B is a basis of E.

(2) B is a maximal linearly independent family of E.

(3) B is a minimal generating family of E.

Proof. Assume (1). Since B is a basis, it is a linearly independent family. We claim that B is a maximal linearly independent family. If B is not a maximal linearly independent family, then there is some vector $w \in E$ such that the family B' obtained by adding w to B is linearly independent. However, since B is a basis of E, the vector w can be expressed as a linear combination of vectors in B, contradicting the fact that B' is linearly independent.

Conversely, assume (2). We claim that B spans E. If B does not span E, then there is some vector $w \in E$ which is not a linear combination of vectors in B. By Lemma 1.4, the family B' obtained by adding w to B is linearly independent. Since B is a proper subfamily of B', this contradicts the assumption that B is a maximal linearly independent family. Therefore, B must span E, and since B is also linearly independent, it is a basis of E.

Again, assume (1). Since B is a basis, it is a generating family of E. We claim that B is a minimal generating family. If B is not a minimal generating family, then there is a proper subfamily B' of B that spans E. Then, every $w \in B - B'$ can be expressed as a linear combination of vectors from B', contradicting the fact that B is linearly independent.

Conversely, assume (3). We claim that B is linearly independent. If B is not linearly independent, then some vector $w \in B$ can be expressed as a linear combination of vectors in $B' = B - \{w\}$. Since B generates E, the family B' also generates E, but B' is a proper subfamily of B, contradicting the minimality of B. Since B spans E and is linearly independent, it is a basis of E. □

The second key result of linear algebra that for any two bases $(u_i)_{i \in I}$ and $(v_j)_{j \in J}$ of a vector space E, the index sets I and J have the same cardinality. In particular, if E has a finite basis of n elements, every basis of E has n elements, and the integer n is called the *dimension* of the vector space E.

To prove the second key result, we can use the following *replacement lemma* due to Steinitz. This result shows the relationship between finite linearly independent families and finite families of generators of a vector space. We begin with a version of the lemma which is a bit informal, but easier to understand than the precise and more formal formulation given in Proposition 1.8. The technical difficulty has to do with the fact that some of the indices need to be renamed.

Proposition 1.7. *(Replacement lemma, version 1) Given a vector space E, let $(u_1, \ldots, u_m)$ be any finite linearly independent family in E, and let $(v_1, \ldots, v_n)$ be any finite family such that every u_i is a linear combination of $(v_1, \ldots, v_n)$. Then, we must have $m \leq n$, and there is a replacement of m of the vectors v_j by $(u_1, \ldots, u_m)$, such that after renaming some of the indices of the v_js, the families $(u_1, \ldots, u_m, v_{m+1}, \ldots, v_n)$ and $(v_1, \ldots, v_n)$ generate the same subspace of E.*

Proof. We proceed by induction on m. When $m = 0$, the family $(u_1, \ldots, u_m)$ is empty, and the proposition holds trivially. For the induction step, we have a linearly independent family $(u_1, \ldots, u_m, u_{m+1})$. Consider the linearly independent family $(u_1, \ldots, u_m)$. By the induction hypothesis, $m \leq n$, and there is a replacement of m of the vectors v_j by $(u_1, \ldots, u_m)$, such that after renaming some of the indices of the vs, the families $(u_1, \ldots, u_m, v_{m+1}, \ldots, v_n)$ and $(v_1, \ldots, v_n)$ generate the same subspace of E. The vector u_{m+1} can also be expressed as a linear combination of $(v_1, \ldots, v_n)$, and since $(u_1, \ldots, u_m, v_{m+1}, \ldots, v_n)$ and $(v_1, \ldots, v_n)$ generate the same subspace, u_{m+1} can be expressed as a linear combination of $(u_1, \ldots, u_m, v_{m+1}, \ldots,$

v_n), say

$$u_{m+1} = \sum_{i=1}^{m} \lambda_i u_i + \sum_{j=m+1}^{n} \lambda_j v_j.$$

We claim that $\lambda_j \neq 0$ for some j with $m+1 \leq j \leq n$, which implies that $m+1 \leq n$.

Otherwise, we would have

$$u_{m+1} = \sum_{i=1}^{m} \lambda_i u_i,$$

a nontrivial linear dependence of the u_i, which is impossible since $(u_1, \ldots, u_{m+1})$ are linearly independent.

Therefore $m+1 \leq n$, and after renaming indices if necessary, we may assume that $\lambda_{m+1} \neq 0$, so we get

$$v_{m+1} = -\sum_{i=1}^{m} (\lambda_{m+1}^{-1}\lambda_i) u_i - \lambda_{m+1}^{-1} u_{m+1} - \sum_{j=m+2}^{n} (\lambda_{m+1}^{-1}\lambda_j) v_j.$$

Observe that the families $(u_1, \ldots, u_m, v_{m+1}, \ldots, v_n)$ and $(u_1, \ldots, u_{m+1}, v_{m+2}, \ldots, v_n)$ generate the same subspace, since u_{m+1} is a linear combination of $(u_1, \ldots, u_m, v_{m+1}, \ldots, v_n)$ and v_{m+1} is a linear combination of $(u_1, \ldots, u_{m+1}, v_{m+2}, \ldots, v_n)$. Since $(u_1, \ldots, u_m, v_{m+1}, \ldots, v_n)$ and $(v_1, \ldots, v_n)$ generate the same subspace, we conclude that $(u_1, \ldots, u_{m+1}, v_{m+2}, \ldots, v_n)$ and and $(v_1, \ldots, v_n)$ generate the same subspace, which concludes the induction hypothesis. □

Here is an example illustrating the replacement lemma. Consider the sequences (u_1, u_2, u_3) and $(v_1, v_2, v_3, v_4, v_5)$ where (u_1, u_2, u_3) is a linearly independent family and with the u_is expressed in terms of the v_js as follows:

$$\begin{aligned} u_1 &= v_4 + v_5 \\ u_2 &= v_3 + v_4 - v_5 \\ u_3 &= v_1 + v_2 + v_3. \end{aligned}$$

From the first equation we get

$$v_4 = u_1 - v_5,$$

and by substituting in the second equation we have

$$u_2 = v_3 + v_4 - v_5 = v_3 + u_1 - v_5 - v_5 = u_1 + v_3 - 2v_5.$$

From the above equation we get

$$v_3 = -u_1 + u_2 + 2v_5,$$

and so

$$u_3 = v_1 + v_2 + v_3 = v_1 + v_2 - u_1 + u_2 + 2v_5.$$

Finally, we get

$$v_1 = u_1 - u_2 + u_3 - v_2 - 2v_5$$

Therefore we have

$$\begin{aligned} v_1 &= u_1 - u_2 + u_3 - v_2 - 2v_5 \\ v_3 &= -u_1 + u_2 + 2v_5 \\ v_4 &= u_1 - v_5, \end{aligned}$$

which shows that $(u_1, u_2, u_3, v_2, v_5)$ spans the same subspace as $(v_1, v_2, v_3, v_4, v_5)$. The vectors (v_1, v_3, v_4) have been replaced by (u_1, u_2, u_3), and the vectors left over are (v_2, v_5). We can rename them (v_4, v_5).

For the sake of completeness, here is a more formal statement of the replacement lemma (and its proof).

Proposition 1.8. *(Replacement lemma, version 2) Given a vector space E, let $(u_i)_{i\in I}$ be any finite linearly independent family in E, where $|I| = m$, and let $(v_j)_{j\in J}$ be any finite family such that every u_i is a linear combination of $(v_j)_{j\in J}$, where $|J| = n$. Then, there exists a set L and an injection $\rho\colon L \to J$ (a relabeling function) such that $L \cap I = \emptyset$, $|L| = n - m$, and the families $(u_i)_{i\in I} \cup (v_{\rho(l)})_{l\in L}$ and $(v_j)_{j\in J}$ generate the same subspace of E. In particular, $m \leq n$.*

Proof. We proceed by induction on $|I| = m$. When $m = 0$, the family $(u_i)_{i\in I}$ is empty, and the proposition holds trivially with $L = J$ (ρ is the identity). Assume $|I| = m + 1$. Consider the linearly independent family $(u_i)_{i\in(I-\{p\})}$, where p is any member of I. By the induction hypothesis, there exists a set L and an injection $\rho\colon L \to J$ such that $L \cap (I - \{p\}) = \emptyset$, $|L| = n - m$, and the families $(u_i)_{i\in(I-\{p\})} \cup (v_{\rho(l)})_{l\in L}$ and $(v_j)_{j\in J}$ generate the same subspace of E. If $p \in L$, we can replace L by $(L - \{p\}) \cup \{p'\}$ where p' does not belong to $I \cup L$, and replace ρ by the injection ρ' which agrees with ρ on $L - \{p\}$ and such that $\rho'(p') = \rho(p)$. Thus, we can always assume that $L \cap I = \emptyset$. Since u_p is a linear combination of $(v_j)_{j\in J}$ and the families $(u_i)_{i\in(I-\{p\})} \cup (v_{\rho(l)})_{l\in L}$ and $(v_j)_{j\in J}$ generate the same subspace of E, u_p is a linear combination of $(u_i)_{i\in(I-\{p\})} \cup (v_{\rho(l)})_{l\in L}$. Let

$$u_p = \sum_{i\in(I-\{p\})} \lambda_i u_i + \sum_{l\in L} \lambda_l v_{\rho(l)}. \tag{1}$$

If $\lambda_l = 0$ for all $l \in L$, we have

$$\sum_{i\in(I-\{p\})} \lambda_i u_i - u_p = 0,$$

contradicting the fact that $(u_i)_{i\in I}$ is linearly independent. Thus, $\lambda_l \neq 0$ for some $l \in L$, say $l = q$. Since $\lambda_q \neq 0$, we have

$$v_{\rho(q)} = \sum_{i\in(I-\{p\})} (-\lambda_q^{-1}\lambda_i) u_i + \lambda_q^{-1} u_p + \sum_{l\in(L-\{q\})} (-\lambda_q^{-1}\lambda_l) v_{\rho(l)}. \tag{2}$$

We claim that the families $(u_i)_{i\in(I-\{p\})} \cup (v_{\rho(l)})_{l\in L}$ and $(u_i)_{i\in I} \cup (v_{\rho(l)})_{l\in(L-\{q\})}$ generate the same subset of E. Indeed, the second family is obtained from the first by replacing $v_{\rho(q)}$ by u_p, and vice-versa, and u_p is a linear combination of $(u_i)_{i\in(I-\{p\})} \cup (v_{\rho(l)})_{l\in L}$, by (1), and $v_{\rho(q)}$ is a linear combination of $(u_i)_{i\in I} \cup (v_{\rho(l)})_{l\in(L-\{q\})}$, by (2). Thus, the families $(u_i)_{i\in I} \cup (v_{\rho(l)})_{l\in(L-\{q\})}$ and $(v_j)_{j\in J}$ generate the same subspace of E, and the proposition holds for $L - \{q\}$ and the restriction of the injection $\rho\colon L \to J$ to $L - \{q\}$, since $L \cap I = \emptyset$ and $|L| = n - m$ imply that $(L - \{q\}) \cap I = \emptyset$ and $|L - \{q\}| = n - (m + 1)$. □

The idea is that m of the vectors v_j can be *replaced* by the linearly independent u_i's in such a way that the same subspace is still generated. The purpose of the function $\rho\colon L \to J$ is to pick $n - m$ elements $j_1, \ldots, j_{n-m}$ of J and to relabel them $l_1, \ldots, l_{n-m}$ in such a way that these new indices do not clash with the indices in I; this way, the vectors $v_{j_1}, \ldots, v_{j_{n-m}}$ who "survive" (i.e. are not replaced) are relabeled $v_{l_1}, \ldots, v_{l_{n-m}}$, and the other m vectors v_j with $j \in J - \{j_1, \ldots, j_{n-m}\}$ are replaced by the u_i. The index set of this new family is $I \cup L$.

Actually, one can prove that Proposition 1.8 implies Theorem 1.5 when the vector space is finitely generated. Putting Theorem 1.5 and Proposition 1.8 together, we obtain the following fundamental theorem.

Theorem 1.9. *Let E be a finitely generated vector space. Any family $(u_i)_{i\in I}$ generating E contains a subfamily $(u_j)_{j\in J}$ which is a basis of E. Any linearly independent family $(u_i)_{i\in I}$ can be extended to a family $(u_j)_{j\in J}$ which is a basis of E (with $I \subseteq J$). Furthermore, for every two bases $(u_i)_{i\in I}$ and $(v_j)_{j\in J}$ of E, we have $|I| = |J| = n$ for some fixed integer $n \geq 0$.*

Proof. The first part follows immediately by applying Theorem 1.5 with $L = \emptyset$ and $S = (u_i)_{i\in I}$. For the second part, consider the family $S' = (u_i)_{i\in I} \cup (v_h)_{h\in H}$, where $(v_h)_{h\in H}$ is any finitely generated family generating E, and with $I \cap H = \emptyset$. Then, apply Theorem 1.5 to $L = (u_i)_{i\in I}$ and to S'. For the last statement, assume that $(u_i)_{i\in I}$ and $(v_j)_{j\in J}$ are bases of E. Since $(u_i)_{i\in I}$ is linearly independent and $(v_j)_{j\in J}$ spans E, Proposition 1.8 implies that $|I| \leq |J|$. A symmetric argument yields $|J| \leq |I|$. □

Remark: Theorem 1.9 also holds for vector spaces that are not finitely generated.

Definition 1.7. When a vector space E is not finitely generated, we say that E is of infinite dimension. The *dimension* of a finitely generated vector space E is the common dimension n of all of its bases and is denoted by $\dim(E)$.

Clearly, if the field $\mathbb{R}$ itself is viewed as a vector space, then every family (a) where $a \in \mathbb{R}$ and $a \neq 0$ is a basis. Thus $\dim(\mathbb{R}) = 1$. Note that $\dim(\{0\}) = 0$.

Definition 1.8. If E is a vector space of dimension $n \geq 1$, for any subspace U of E, if $\dim(U) = 1$, then U is called a *line*; if $\dim(U) = 2$, then U is called a *plane*; if $\dim(U) = n-1$, then U is called a *hyperplane*. If $\dim(U) = k$, then U is sometimes called a *k-plane*.

Let $(u_i)_{i\in I}$ be a basis of a vector space E. For any vector $v \in E$, since the family $(u_i)_{i\in I}$ generates E, there is a family $(\lambda_i)_{i\in I}$ of scalars in $\mathbb{R}$, such that

$$v = \sum_{i\in I} \lambda_i u_i.$$

A very important fact is that the family $(\lambda_i)_{i\in I}$ is **unique**.

Proposition 1.10. *Given a vector space E, let $(u_i)_{i\in I}$ be a family of vectors in E. Let $v \in E$, and assume that $v = \sum_{i\in I} \lambda_i u_i$. Then, the family $(\lambda_i)_{i\in I}$ of scalars such that $v = \sum_{i\in I} \lambda_i u_i$ is unique iff $(u_i)_{i\in I}$ is linearly independent.*

Proof. First, assume that $(u_i)_{i\in I}$ is linearly independent. If $(\mu_i)_{i\in I}$ is another family of scalars in $\mathbb{R}$ such that $v = \sum_{i\in I} \mu_i u_i$, then we have

$$\sum_{i\in I} (\lambda_i - \mu_i) u_i = 0,$$

and since $(u_i)_{i\in I}$ is linearly independent, we must have $\lambda_i - \mu_i = 0$ for all $i \in I$, that is, $\lambda_i = \mu_i$ for all $i \in I$. The converse is shown by contradiction. If $(u_i)_{i\in I}$ was linearly dependent, there would be a family $(\mu_i)_{i\in I}$ of scalars not all null such that

$$\sum_{i\in I} \mu_i u_i = 0$$

and $\mu_j \neq 0$ for some $j \in I$. But then,

$$v = \sum_{i\in I} \lambda_i u_i + 0 = \sum_{i\in I} \lambda_i u_i + \sum_{i\in I} \mu_i u_i = \sum_{i\in I} (\lambda_i + \mu_i) u_i,$$

with $\lambda_j \neq \lambda_j + \mu_j$ since $\mu_j \neq 0$, contradicting the assumption that $(\lambda_i)_{i\in I}$ is the unique family such that $v = \sum_{i\in I} \lambda_i u_i$. □

Definition 1.9. If $(u_i)_{i\in I}$ is a basis of a vector space E, for any vector $v \in E$, if $(x_i)_{i\in I}$ is the unique family of scalars in $\mathbb{R}$ such that

$$v = \sum_{i\in I} x_i u_i,$$

each x_i is called the *component (or coordinate) of index i of v with respect to the basis $(u_i)_{i\in I}$*.

Many interesting mathematical structures are vector spaces. A very important example is the set of linear maps between two vector spaces to be defined in the next section. Here is an example that will prepare us for the vector space of linear maps.

Example 1.6. Let X be any nonempty set and let E be a vector space. The set of all functions $f\colon X \to E$ can be made into a vector space as follows: Given any two functions $f\colon X \to E$ and $g\colon X \to E$, let $(f+g)\colon X \to E$ be defined such that

$$(f+g)(x) = f(x) + g(x)$$

for all $x \in X$, and for every $\lambda \in \mathbb{R}$, let $\lambda f\colon X \to E$ be defined such that

$$(\lambda f)(x) = \lambda f(x)$$

for all $x \in X$. The axioms of a vector space are easily verified. Now, let $E = \mathbb{R}$, and let I be the set of all nonempty subsets of X. For every $S \in I$, let $f_S\colon X \to E$ be the function such that $f_S(x) = 1$ iff $x \in S$, and $f_S(x) = 0$ iff $x \notin S$. We leave as an exercise to show that $(f_S)_{S \in I}$ is linearly independent.

1.6 Matrices

In Section 1.1 we introduced informally the notion of a matrix. In this section we define matrices precisely, and also introduce some operations on matrices. It turns out that matrices form a vector space equipped with a multiplication operation which is associative, but noncommutative. We will explain in Section 2.1 how matrices can be used to represent linear maps, defined in the next section.

Definition 1.10. If $K = \mathbb{R}$ or $K = \mathbb{C}$, an *$m \times n$-matrix over K* is a family $(a_{i\,j})_{1 \leq i \leq m,\ 1 \leq j \leq n}$ of scalars in K, represented by an array

$$\begin{pmatrix} a_{1\,1} & a_{1\,2} & \dots & a_{1\,n} \\ a_{2\,1} & a_{2\,2} & \dots & a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m\,1} & a_{m\,2} & \dots & a_{m\,n} \end{pmatrix}$$

In the special case where $m = 1$, we have a *row vector*, represented by

$$(a_{1\,1} \cdots a_{1\,n})$$

and in the special case where $n = 1$, we have a *column vector*, represented by

$$\begin{pmatrix} a_{1\,1} \\ \vdots \\ a_{m\,1} \end{pmatrix}.$$

In these last two cases, we usually omit the constant index 1 (first index in case of a row, second index in case of a column). The set of all $m \times n$-matrices is denoted by $\mathrm{M}_{m,n}(K)$ or $\mathrm{M}_{m,n}$. An $n \times n$-matrix is called a *square matrix of dimension n*. The set of all square matrices of dimension n is denoted by $\mathrm{M}_n(K)$, or M_n.

Remark: As defined, a matrix $A = (a_{i\,j})_{1 \leq i \leq m,\ 1 \leq j \leq n}$ is a *family*, that is, a function from $\{1, 2, \ldots, m\} \times \{1, 2, \ldots, n\}$ to K. As such, there is no reason to assume an ordering on the indices. Thus, the matrix A can be represented in many different ways as an array, by adopting different orders for the rows or the columns. However, it is customary (and usually convenient) to assume the natural ordering on the sets $\{1, 2, \ldots, m\}$ and $\{1, 2, \ldots, n\}$, and to represent A as an array according to this ordering of the rows and columns.

We define some operations on matrices as follows.

Definition 1.11. Given two $m \times n$ matrices $A = (a_{i\,j})$ and $B = (b_{i\,j})$, we define their *sum* $A + B$ as the matrix $C = (c_{i\,j})$ such that $c_{i\,j} = a_{i\,j} + b_{i\,j}$; that is,

$$\begin{pmatrix} a_{1\,1} & a_{1\,2} & \ldots & a_{1\,n} \\ a_{2\,1} & a_{2\,2} & \ldots & a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m\,1} & a_{m\,2} & \ldots & a_{m\,n} \end{pmatrix} + \begin{pmatrix} b_{1\,1} & b_{1\,2} & \ldots & b_{1\,n} \\ b_{2\,1} & b_{2\,2} & \ldots & b_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m\,1} & b_{m\,2} & \ldots & b_{m\,n} \end{pmatrix}$$

$$= \begin{pmatrix} a_{1\,1} + b_{1\,1} & a_{1\,2} + b_{1\,2} & \ldots & a_{1\,n} + b_{1\,n} \\ a_{2\,1} + b_{2\,1} & a_{2\,2} + b_{2\,2} & \ldots & a_{2\,n} + b_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m\,1} + b_{m\,1} & a_{m\,2} + b_{m\,2} & \ldots & a_{m\,n} + b_{m\,n} \end{pmatrix}.$$

For any matrix $A = (a_{i\,j})$, we let $-A$ be the matrix $(-a_{i\,j})$. Given a scalar $\lambda \in K$, we define the matrix λA as the matrix $C = (c_{i\,j})$ such that $c_{i\,j} = \lambda a_{i\,j}$; that is

$$\lambda \begin{pmatrix} a_{1\,1} & a_{1\,2} & \ldots & a_{1\,n} \\ a_{2\,1} & a_{2\,2} & \ldots & a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m\,1} & a_{m\,2} & \ldots & a_{m\,n} \end{pmatrix} = \begin{pmatrix} \lambda a_{1\,1} & \lambda a_{1\,2} & \ldots & \lambda a_{1\,n} \\ \lambda a_{2\,1} & \lambda a_{2\,2} & \ldots & \lambda a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda a_{m\,1} & \lambda a_{m\,2} & \ldots & \lambda a_{m\,n} \end{pmatrix}.$$

Given an $m \times n$ matrices $A = (a_{i\,k})$ and an $n \times p$ matrices $B = (b_{k\,j})$, we define their *product* AB as the $m \times p$ matrix $C = (c_{i\,j})$ such that

$$c_{i\,j} = \sum_{k=1}^{n} a_{i\,k} b_{k\,j},$$

for $1 \leq i \leq m$, and $1 \leq j \leq p$. In the product $AB = C$ shown below

$$\begin{pmatrix} a_{1\,1} & a_{1\,2} & \ldots & a_{1\,n} \\ a_{2\,1} & a_{2\,2} & \ldots & a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m\,1} & a_{m\,2} & \ldots & a_{m\,n} \end{pmatrix} \begin{pmatrix} b_{1\,1} & b_{1\,2} & \ldots & b_{1\,p} \\ b_{2\,1} & b_{2\,2} & \ldots & b_{2\,p} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n\,1} & b_{n\,2} & \ldots & b_{n\,p} \end{pmatrix} = \begin{pmatrix} c_{1\,1} & c_{1\,2} & \ldots & c_{1\,p} \\ c_{2\,1} & c_{2\,2} & \ldots & c_{2\,p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m\,1} & c_{m\,2} & \ldots & c_{m\,p} \end{pmatrix},$$

note that the entry of index i and j of the matrix AB obtained by multiplying the matrices A and B can be identified with the product of the row matrix corresponding to the i-th row of A with the column matrix corresponding to the j-column of B:

$$(a_{i\,1} \cdots a_{i\,n}) \begin{pmatrix} b_{1\,j} \\ \vdots \\ b_{n\,j} \end{pmatrix} = \sum_{k=1}^{n} a_{i\,k} b_{k\,j}.$$

Definition 1.12. The square matrix I_n of dimension n containing 1 on the diagonal and 0 everywhere else is called the *identity matrix*. It is denoted by

$$I_n = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

Definition 1.13. Given an $m \times n$ matrix $A = (a_{i\,j})$, its *transpose* $A^\top = (a^\top_{j\,i})$, is the $n \times m$-matrix such that $a^\top_{j\,i} = a_{i\,j}$, for all i, $1 \leq i \leq m$, and all j, $1 \leq j \leq n$.

The transpose of a matrix A is sometimes denoted by A^t, or even by ${}^tA$. Note that the transpose $A^\top$ of a matrix A has the property that the j-th row of $A^\top$ is the j-th column of A. In other words, transposition exchanges the rows and the columns of a matrix.

The following observation will be useful later on when we discuss the SVD. Given any $m \times n$ matrix A and any $n \times p$ matrix B, if we denote the columns of A by $A^1, \dots, A^n$ and the rows of B by $B_1, \dots, B_n$, then we have

$$AB = A^1 B_1 + \cdots + A^n B_n.$$

For every square matrix A of dimension n, it is immediately verified that $AI_n = I_n A = A$.

Definition 1.14. For any square matrix A of dimension n, if a matrix B such that $AB = BA = I_n$ exists, then it is unique, and it is called the *inverse* of A. The matrix B is also denoted by A^{-1}. An invertible matrix is also called a *nonsingular* matrix, and a matrix that is not invertible is called a *singular* matrix.

Using Proposition 1.15 and the fact that matrices represent linear maps, it can be shown that if a square matrix A has a left inverse, that is a matrix B such that $BA = I$, or a right inverse, that is a matrix C such that $AC = I$, then A is actually invertible; so $B = A^{-1}$ and $C = A^{-1}$. These facts also follow from Proposition 3.9.

It is immediately verified that the set $\mathrm{M}_{m,n}(K)$ of $m \times n$ matrices is a *vector space* under addition of matrices and multiplication of a matrix by a scalar. Consider the $m \times n$-matrices

$E_{i,j} = (e_{h\,k})$, defined such that $e_{i\,j} = 1$, and $e_{h\,k} = 0$, if $h \neq i$ or $k \neq j$. It is clear that every matrix $A = (a_{i\,j}) \in \mathrm{M}_{m,n}(K)$ can be written in a unique way as

$$A = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{i\,j} E_{i,j}.$$

Thus, the family $(E_{i,j})_{1 \leq i \leq m, 1 \leq j \leq n}$ is a basis of the vector space $\mathrm{M}_{m,n}(K)$, which has dimension mn.

Remark: Definition 1.10 and Definition 1.11 also make perfect sense when K is a (commutative) ring rather than a field. In this more general setting, the framework of vector spaces is too narrow, but we can consider structures over a commutative ring A satisfying all the axioms of Definition 1.2. Such structures are called *modules*. The theory of modules is (much) more complicated than that of vector spaces. For example, modules do not always have a basis, and other properties holding for vector spaces usually fail for modules. When a module has a basis, it is called a *free module*. For example, when A is a commutative ring, the structure A^n is a module such that the vectors e_i, with $(e_i)_i = 1$ and $(e_i)_j = 0$ for $j \neq i$, form a basis of A^n. Many properties of vector spaces still hold for A^n. Thus, A^n is a free module. As another example, when A is a commutative ring, $\mathrm{M}_{m,n}(A)$ is a free module with basis $(E_{i,j})_{1 \leq i \leq m, 1 \leq j \leq n}$. Polynomials over a commutative ring also form a free module of infinite dimension.

The properties listed in Proposition 1.11 are easily verified, although some of the computations are a bit tedious. A more conceptual proof is given in Proposition 2.1.

Proposition 1.11. *(1) Given any matrices $A \in \mathrm{M}_{m,n}(K)$, $B \in \mathrm{M}_{n,p}(K)$, and $C \in \mathrm{M}_{p,q}(K)$, we have*

$$(AB)C = A(BC);$$

that is, matrix multiplication is associative.

(2) Given any matrices $A, B \in \mathrm{M}_{m,n}(K)$, and $C, D \in \mathrm{M}_{n,p}(K)$, for all $\lambda \in K$, we have

$$\begin{aligned}(A + B)C &= AC + BC\\ A(C + D) &= AC + AD\\ (\lambda A)C &= \lambda(AC)\\ A(\lambda C) &= \lambda(AC),\end{aligned}$$

so that matrix multiplication $\cdot \colon \mathrm{M}_{m,n}(K) \times \mathrm{M}_{n,p}(K) \to \mathrm{M}_{m,p}(K)$ is bilinear.

The properties of Proposition 1.11 together with the fact that $AI_n = I_n A = A$ for all square $n \times n$ matrices show that $\mathrm{M}_n(K)$ is a ring with unit I_n (in fact, an associative algebra). This is a noncommutative ring with zero divisors, as shown by the following Example.

Example 1.7. For example, letting A, B be the 2×2-matrices

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix},$$

then

$$AB = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$

and

$$BA = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}.$$

Thus $AB \neq BA$, and $AB = 0$, even though both $A, B \neq 0$.

1.7 Linear Maps

Now that we understand vector spaces and how to generate them, we would like to be able to transform one vector space E into another vector space F. A function between two vector spaces that preserves the vector space structure is called a homomorphism of vector spaces, or *linear map*. Linear maps formalize the concept of linearity of a function.

Keep in mind that linear maps, which are transformations of space, are usually far more important than the spaces themselves.

In the rest of this section, we assume that all vector spaces are real vector spaces.

Definition 1.15. Given two vector spaces E and F, a *linear map* between E and F is a function $f : E \to F$ satisfying the following two conditions:

$$\begin{aligned} f(x + y) &= f(x) + f(y) && \text{for all } x, y \in E; \\ f(\lambda x) &= \lambda f(x) && \text{for all } \lambda \in \mathbb{R}, x \in E. \end{aligned}$$

Setting $x = y = 0$ in the first identity, we get $f(0) = 0$. The basic property of linear maps is that they transform linear combinations into linear combinations. Given any finite family $(u_i)_{i \in I}$ of vectors in E, given any family $(\lambda_i)_{i \in I}$ of scalars in $\mathbb{R}$, we have

$$f(\sum_{i \in I} \lambda_i u_i) = \sum_{i \in I} \lambda_i f(u_i).$$

The above identity is shown by induction on $|I|$ using the properties of Definition 1.15.

Example 1.8.

1. The map $f\colon \mathbb{R}^2 \to \mathbb{R}^2$ defined such that

$$\begin{aligned} x' &= x - y \\ y' &= x + y \end{aligned}$$

is a linear map. The reader should check that it is the composition of a rotation by $\pi/4$ with a magnification of ratio $\sqrt{2}$.

2. For any vector space E, the *identity map* $\mathrm{id}\colon E \to E$ given by

$$\mathrm{id}(u) = u \quad \text{for all } u \in E$$

is a linear map. When we want to be more precise, we write id_E instead of id.

3. The map $D\colon \mathbb{R}[X] \to \mathbb{R}[X]$ defined such that

$$D(f(X)) = f'(X),$$

where $f'(X)$ is the derivative of the polynomial $f(X)$, is a linear map.

4. The map $\Phi\colon \mathcal{C}([a,b]) \to \mathbb{R}$ given by

$$\Phi(f) = \int_a^b f(t)dt,$$

where $\mathcal{C}([a,b])$ is the set of continuous functions defined on the interval $[a,b]$, is a linear map.

5. The function $\langle -, - \rangle\colon \mathcal{C}([a,b]) \times \mathcal{C}([a,b]) \to \mathbb{R}$ given by

$$\langle f, g \rangle = \int_a^b f(t)g(t)dt,$$

is linear in each of the variable f, g. It also satisfies the properties $\langle f, g\rangle = \langle g, f\rangle$ and $\langle f, f\rangle = 0$ iff $f = 0$. It is an example of an *inner product*.

Definition 1.16. Given a linear map $f\colon E \to F$, we define its *image (or range)* $\mathrm{Im}\, f = f(E)$, as the set

$$\mathrm{Im}\, f = \{y \in F \mid (\exists x \in E)(y = f(x))\},$$

and its *Kernel (or nullspace)* $\mathrm{Ker}\, f = f^{-1}(0)$, as the set

$$\mathrm{Ker}\, f = \{x \in E \mid f(x) = 0\}.$$

The derivative map $D\colon \mathbb{R}[X] \to \mathbb{R}[X]$ from Example 1.8(3) has kernel the constant polynomials, so $\operatorname{Ker} D = \mathbb{R}$. If we consider the second derivative $D \circ D\colon \mathbb{R}[X] \to \mathbb{R}[X]$, then the kernel of $D \circ D$ consists of all polynomials of degree ≤ 1. The image of $D\colon \mathbb{R}[X] \to \mathbb{R}[X]$ is actually $\mathbb{R}[X]$ itself, because every polynomial $P(X) = a_0X^n + \cdots + a_{n-1}X + a_n$ of degree n is the derivative of the polynomial $Q(X)$ of degree $n+1$ given by

$$Q(X) = a_0\frac{X^{n+1}}{n+1} + \cdots + a_{n-1}\frac{X^2}{2} + a_nX.$$

On the other hand, if we consider the restriction of D to the vector space $\mathbb{R}[X]_n$ of polynomials of degree $\leq n$, then the kernel of D is still $\mathbb{R}$, but the image of D is the $\mathbb{R}[X]_{n-1}$, the vector space of polynomials of degree $\leq n-1$.

Proposition 1.12. *Given a linear map $f\colon E \to F$, the set* $\operatorname{Im} f$ *is a subspace of F and the set* $\operatorname{Ker} f$ *is a subspace of E. The linear map $f\colon E \to F$ is injective iff* $\operatorname{Ker} f = (0)$ *(where* (0) *is the trivial subspace* $\{0\}$*).*

Proof. Given any $x, y \in \operatorname{Im} f$, there are some $u, v \in E$ such that $x = f(u)$ and $y = f(v)$, and for all $\lambda, \mu \in \mathbb{R}$, we have

$$f(\lambda u + \mu v) = \lambda f(u) + \mu f(v) = \lambda x + \mu y,$$

and thus, $\lambda x + \mu y \in \operatorname{Im} f$, showing that $\operatorname{Im} f$ is a subspace of F.

Given any $x, y \in \operatorname{Ker} f$, we have $f(x) = 0$ and $f(y) = 0$, and thus,

$$f(\lambda x + \mu y) = \lambda f(x) + \mu f(y) = 0,$$

that is, $\lambda x + \mu y \in \operatorname{Ker} f$, showing that $\operatorname{Ker} f$ is a subspace of E.

First, assume that $\operatorname{Ker} f = (0)$. We need to prove that $f(x) = f(y)$ implies that $x = y$. However, if $f(x) = f(y)$, then $f(x) - f(y) = 0$, and by linearity of f we get $f(x - y) = 0$. Because $\operatorname{Ker} f = (0)$, we must have $x - y = 0$, that is $x = y$, so f is injective. Conversely, assume that f is injective. If $x \in \operatorname{Ker} f$, that is $f(x) = 0$, since $f(0) = 0$ we have $f(x) = f(0)$, and by injectivity, $x = 0$, which proves that $\operatorname{Ker} f = (0)$. Therefore, f is injective iff $\operatorname{Ker} f = (0)$. □

Since by Proposition 1.12, the image $\operatorname{Im} f$ of a linear map f is a subspace of F, we can define the *rank* $\operatorname{rk}(f)$ of f as the dimension of $\operatorname{Im} f$.

Definition 1.17. Given a linear map$f\colon E \to F$, the *rank* $\operatorname{rk}(f)$ of f is the dimension of the image $\operatorname{Im} f$ of f.

A fundamental property of bases in a vector space is that they allow the definition of linear maps as unique homomorphic extensions, as shown in the following proposition.

Proposition 1.13. *Given any two vector spaces E and F, given any basis $(u_i)_{i\in I}$ of E, given any other family of vectors $(v_i)_{i\in I}$ in F, there is a unique linear map $f\colon E \to F$ such that $f(u_i) = v_i$ for all $i \in I$. Furthermore, f is injective iff $(v_i)_{i\in I}$ is linearly independent, and f is surjective iff $(v_i)_{i\in I}$ generates F.*

Proof. If such a linear map $f\colon E \to F$ exists, since $(u_i)_{i\in I}$ is a basis of E, every vector $x \in E$ can written uniquely as a linear combination

$$x = \sum_{i\in I} x_i u_i,$$

and by linearity, we must have

$$f(x) = \sum_{i\in I} x_i f(u_i) = \sum_{i\in I} x_i v_i.$$

Define the function $f\colon E \to F$, by letting

$$f(x) = \sum_{i\in I} x_i v_i$$

for every $x = \sum_{i\in I} x_i u_i$. It is easy to verify that f is indeed linear, it is unique by the previous reasoning, and obviously, $f(u_i) = v_i$.

Now, assume that f is injective. Let $(\lambda_i)_{i\in I}$ be any family of scalars, and assume that

$$\sum_{i\in I} \lambda_i v_i = 0.$$

Since $v_i = f(u_i)$ for every $i \in I$, we have

$$f(\sum_{i\in I} \lambda_i u_i) = \sum_{i\in I} \lambda_i f(u_i) = \sum_{i\in I} \lambda_i v_i = 0.$$

Since f is injective iff $\operatorname{Ker} f = (0)$, we have

$$\sum_{i\in I} \lambda_i u_i = 0,$$

and since $(u_i)_{i\in I}$ is a basis, we have $\lambda_i = 0$ for all $i \in I$, which shows that $(v_i)_{i\in I}$ is linearly independent. Conversely, assume that $(v_i)_{i\in I}$ is linearly independent. Since $(u_i)_{i\in I}$ is a basis of E, every vector $x \in E$ is a linear combination $x = \sum_{i\in I} \lambda_i u_i$ of $(u_i)_{i\in I}$. If

$$f(x) = f(\sum_{i\in I} \lambda_i u_i) = 0,$$

then

$$\sum_{i\in I} \lambda_i v_i = \sum_{i\in I} \lambda_i f(u_i) = f(\sum_{i\in I} \lambda_i u_i) = 0,$$

and $\lambda_i = 0$ for all $i \in I$ because $(v_i)_{i\in I}$ is linearly independent, which means that $x = 0$. Therefore, $\operatorname{Ker} f = (0)$, which implies that f is injective. The part where f is surjective is left as a simple exercise. □

By the second part of Proposition 1.13, an injective linear map $f: E \to F$ sends a basis $(u_i)_{i\in I}$ to a linearly independent family $(f(u_i))_{i\in I}$ of F, which is also a basis when f is bijective. Also, when E and F have the same finite dimension n, $(u_i)_{i\in I}$ is a basis of E, and $f: E \to F$ is injective, then $(f(u_i))_{i\in I}$ is a basis of F (by Proposition 1.6).

The following simple proposition is also useful.

Proposition 1.14. *Given any two vector spaces E and F, with F nontrivial, given any family $(u_i)_{i\in I}$ of vectors in E, the following properties hold:*

(1) The family $(u_i)_{i\in I}$ generates E iff for every family of vectors $(v_i)_{i\in I}$ in F, there is at most one linear map $f: E \to F$ such that $f(u_i) = v_i$ for all $i \in I$.

(2) The family $(u_i)_{i\in I}$ is linearly independent iff for every family of vectors $(v_i)_{i\in I}$ in F, there is some linear map $f: E \to F$ such that $f(u_i) = v_i$ for all $i \in I$.

Proof. (1) If there is any linear map $f: E \to F$ such that $f(u_i) = v_i$ for all $i \in I$, since $(u_i)_{i\in I}$ generates E, every vector $x \in E$ can be written as some linear combination

$$x = \sum_{i\in I} x_i u_i,$$

and by linearity, we must have

$$f(x) = \sum_{i\in I} x_i f(u_i) = \sum_{i\in I} x_i v_i.$$

This shows that f is unique if it exists. Conversely, assume that $(u_i)_{i\in I}$ does not generate E. Since F is nontrivial, there is some some vector $y \in F$ such that $y \neq 0$. Since $(u_i)_{i\in I}$ does not generate E, there is some vector $w \in E$ that is not in the subspace generated by $(u_i)_{i\in I}$. By Theorem 1.9, there is a linearly independent subfamily $(u_i)_{i\in I_0}$ of $(u_i)_{i\in I}$ generating the same subspace. Since by hypothesis, $w \in E$ is not in the subspace generated by $(u_i)_{i\in I_0}$, by Lemma 1.4 and by Theorem 1.9 again, there is a basis $(e_j)_{j\in I_0\cup J}$ of E, such that $e_i = u_i$ for all $i \in I_0$, and $w = e_{j_0}$ for some $j_0 \in J$. Letting $(v_i)_{i\in I}$ be the family in F such that $v_i = 0$ for all $i \in I$, defining $f: E \to F$ to be the constant linear map with value 0, we have a linear map such that $f(u_i) = 0$ for all $i \in I$. By Proposition 1.13, there is a unique linear map $g: E \to F$ such that $g(w) = y$, and $g(e_j) = 0$ for all $j \in (I_0 \cup J) - \{j_0\}$. By definition of the basis $(e_j)_{j\in I_0\cup J}$ of E, we have $g(u_i) = 0$ for all $i \in I$, and since $f \neq g$, this contradicts the fact that there is at most one such map.

(2) If the family $(u_i)_{i\in I}$ is linearly independent, then by Theorem 1.9, $(u_i)_{i\in I}$ can be extended to a basis of E, and the conclusion follows by Proposition 1.13. Conversely, assume that $(u_i)_{i\in I}$ is linearly dependent. Then, there is some family $(\lambda_i)_{i\in I}$ of scalars (not all zero) such that

$$\sum_{i\in I} \lambda_i u_i = 0.$$

By the assumption, for any nonzero vector $y \in F$, for every $i \in I$, there is some linear map $f_i\colon E \to F$, such that $f_i(u_i) = y$, and $f_i(u_j) = 0$, for $j \in I - \{i\}$. Then, we would get

$$0 = f_i(\sum_{i\in I} \lambda_i u_i) = \sum_{i\in I} \lambda_i f_i(u_i) = \lambda_i y,$$

and since $y \neq 0$, this implies $\lambda_i = 0$ for every $i \in I$. Thus, $(u_i)_{i\in I}$ is linearly independent. □

Given vector spaces E, F, and G, and linear maps $f\colon E \to F$ and $g\colon F \to G$, it is easily verified that the composition $g \circ f\colon E \to G$ of f and g is a linear map.

Definition 1.18. A linear map $f\colon E \to F$ is an *isomorphism* iff there is a linear map $g\colon F \to E$, such that

$$g \circ f = \mathrm{id}_E \quad \text{and} \quad f \circ g = \mathrm{id}_F. \tag{$*$}$$

The map g in Definition 1.18 is unique. This is because if g and h both satisfy $g \circ f = \mathrm{id}_E$, $f \circ g = \mathrm{id}_F$, $h \circ f = \mathrm{id}_E$, and $f \circ h = \mathrm{id}_F$, then

$$g = g \circ \mathrm{id}_F = g \circ (f \circ h) = (g \circ f) \circ h = \mathrm{id}_E \circ h = h.$$

The map g satisfying $(*)$ above is called the *inverse* of f and it is also denoted by f^{-1}.

Observe that Proposition 1.13 shows that if $F = \mathbb{R}^n$, then we get an isomorphism between any vector space E of dimension $|J| = n$ and $\mathbb{R}^n$. Proposition 1.13 also implies that if E and F are two vector spaces, $(u_i)_{i\in I}$ is a basis of E, and $f\colon E \to F$ is a linear map which is an isomorphism, then the family $(f(u_i))_{i\in I}$ is a basis of F.

One can verify that if $f\colon E \to F$ is a bijective linear map, then its inverse $f^{-1}\colon F \to E$, as a function, is also a linear map, and thus f is an isomorphism.

Another useful corollary of Proposition 1.13 is this:

Proposition 1.15. *Let E be a vector space of finite dimension $n \geq 1$ and let $f\colon E \to E$ be any linear map. The following properties hold:*

(1) If f has a left inverse g, that is, if g is a linear map such that $g \circ f = \mathrm{id}$, then f is an isomorphism and $f^{-1} = g$.

(2) If f has a right inverse h, that is, if h is a linear map such that $f \circ h = \mathrm{id}$, then f is an isomorphism and $f^{-1} = h$.

Proof. (1) The equation $g \circ f = \mathrm{id}$ implies that f is injective; this is a standard result about functions (if $f(x) = f(y)$, then $g(f(x)) = g(f(y))$, which implies that $x = y$ since $g \circ f = \mathrm{id}$). Let $(u_1, \ldots, u_n)$ be any basis of E. By Proposition 1.13, since f is injective, $(f(u_1), \ldots, f(u_n))$ is linearly independent, and since E has dimension n, it is a basis of E (if $(f(u_1), \ldots, f(u_n))$ doesn't span E, then it can be extended to a basis of dimension

strictly greater than n, contradicting Theorem 1.9). Then, f is bijective, and by a previous observation its inverse is a linear map. We also have

$$g = g \circ \mathrm{id} = g \circ (f \circ f^{-1}) = (g \circ f) \circ f^{-1} = \mathrm{id} \circ f^{-1} = f^{-1}.$$

(2) The equation $f \circ h = \mathrm{id}$ implies that f is surjective; this is a standard result about functions (for any $y \in E$, we have $f(h(y)) = y$). Let $(u_1, \ldots, u_n)$ be any basis of E. By Proposition 1.13, since f is surjective, $(f(u_1), \ldots, f(u_n))$ spans E, and since E has dimension n, it is a basis of E (if $(f(u_1), \ldots, f(u_n))$ is not linearly independent, then because it spans E, it contains a basis of dimension strictly smaller than n, contradicting Theorem 1.9). Then, f is bijective, and by a previous observation its inverse is a linear map. We also have

$$h = \mathrm{id} \circ h = (f^{-1} \circ f) \circ h = f^{-1} \circ (f \circ h) = f^{-1} \circ \mathrm{id} = f^{-1}.$$

This completes the proof. □

Definition 1.19. The set of all linear maps between two vector spaces E and F is denoted by $\mathrm{Hom}(E, F)$ or by $\mathcal{L}(E; F)$ (the notation $\mathcal{L}(E; F)$ is usually reserved to the set of continuous linear maps, where E and F are normed vector spaces). When we wish to be more precise and specify the field K over which the vector spaces E and F are defined we write $\mathrm{Hom}_K(E, F)$.

The set $\mathrm{Hom}(E, F)$ is a vector space under the operations defined in Example 1.6, namely

$$(f + g)(x) = f(x) + g(x)$$

for all $x \in E$, and

$$(\lambda f)(x) = \lambda f(x)$$

for all $x \in E$. The point worth checking carefully is that λf is indeed a linear map, which uses the commutativity of $*$ in the field K (typically, $K = \mathbb{R}$ or $K = \mathbb{C}$). Indeed, we have

$$(\lambda f)(\mu x) = \lambda f(\mu x) = \lambda \mu f(x) = \mu \lambda f(x) = \mu(\lambda f)(x).$$

When E and F have finite dimensions, the vector space $\mathrm{Hom}(E, F)$ also has finite dimension, as we shall see shortly.

Definition 1.20. When $E = F$, a linear map $f \colon E \to E$ is also called an *endomorphism*. The space $\mathrm{Hom}(E, E)$ is also denoted by $\mathrm{End}(E)$.

It is also important to note that composition confers to $\mathrm{Hom}(E, E)$ a ring structure. Indeed, composition is an operation $\circ \colon \mathrm{Hom}(E, E) \times \mathrm{Hom}(E, E) \to \mathrm{Hom}(E, E)$, which is associative and has an identity id_E, and the distributivity properties hold:

$$(g_1 + g_2) \circ f = g_1 \circ f + g_2 \circ f;$$
$$g \circ (f_1 + f_2) = g \circ f_1 + g \circ f_2.$$

The ring $\mathrm{Hom}(E, E)$ is an example of a noncommutative ring.

It is easily seen that the set of bijective linear maps $f \colon E \to E$ is a group under composition.

Definition 1.21. Bijective linear maps $f\colon E \to E$ are also called *automorphisms*. The group of automorphisms of E is called the *general linear group (of E)*, and it is denoted by $\mathbf{GL}(E)$, or by $\mathrm{Aut}(E)$, or when $E = \mathbb{R}^n$, by $\mathbf{GL}(n, \mathbb{R})$, or even by $\mathbf{GL}(n)$.

1.8 Linear Forms and the Dual Space

We already observed that the field K itself ($K = \mathbb{R}$ or $K = \mathbb{C}$) is a vector space (over itself). The vector space $\mathrm{Hom}(E, K)$ of linear maps from E to the field K, the linear forms, plays a particular role. In this section, we only define linear forms and show that every finite-dimensional vector space has a dual basis. A more advanced presentation of dual spaces and duality is given in Chapter 8.

Definition 1.22. Given a vector space E, the vector space $\mathrm{Hom}(E, K)$ of linear maps from E to the field K is called the *dual space (or dual)* of E. The space $\mathrm{Hom}(E, K)$ is also denoted by E^*, and the linear maps in E^* are called *the linear forms*, or *covectors*. The dual space E^{**} of the space E^* is called the *bidual* of E.

As a matter of notation, linear forms $f\colon E \to K$ will also be denoted by starred symbol, such as u^*, x^*, *etc.*

If E is a vector space of finite dimension n and $(u_1, \ldots, u_n)$ is a basis of E, for any linear form $f^* \in E^*$, for every $x = x_1u_1 + \cdots + x_nu_n \in E$, by linearity we have

$$\begin{aligned} f^*(x) &= f^*(u_1)x_1 + \cdots + f^*(u_n)x_n \\ &= \lambda_1 x_1 + \cdots + \lambda_n x_n, \end{aligned}$$

with $\lambda_i = f^*(u_i) \in K$ for every i, $1 \leq i \leq n$. Thus, with respect to the basis $(u_1, \ldots, u_n)$, the linear form f^* is represented by the row vector

$$(\lambda_1 \ \cdots \ \lambda_n),$$

we have

$$f^*(x) = \begin{pmatrix} \lambda_1 & \cdots & \lambda_n \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix},$$

a linear combination of the coordinates of x, and we can view the linear form f^* as a *linear equation*. If we decide to use a column vector of coefficients

$$c = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix}$$

instead of a row vector, then the linear form f^* is defined by

$$f^*(x) = c^\top x.$$

The above notation is often used in machine learning.

Example 1.9. Given any differentiable function $f\colon \mathbb{R}^n \to \mathbb{R}$, by definition, for any $x \in \mathbb{R}^n$, the *total derivative* df_x of f at x is the linear form $df_x\colon \mathbb{R}^n \to \mathbb{R}$ defined so that for all $u = (u_1, \ldots, u_n) \in \mathbb{R}^n$,

$$df_x(u) = \begin{pmatrix} \dfrac{\partial f}{\partial x_1}(x) & \cdots & \dfrac{\partial f}{\partial x_n}(x) \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x)\, u_i.$$

Example 1.10. Let $\mathcal{C}([0,1])$ be the vector space of continuous functions $f\colon [0,1] \to \mathbb{R}$. The map $\mathcal{I}\colon \mathcal{C}([0,1]) \to \mathbb{R}$ given by

$$\mathcal{I}(f) = \int_0^1 f(x)dx \quad \text{for any } f \in \mathcal{C}([0,1])$$

is a linear form (integration).

Example 1.11. Consider the vector space $\mathrm{M}_n(\mathbb{R})$ of real $n \times n$ matrices. Let $\mathrm{tr}\colon \mathrm{M}_n(\mathbb{R}) \to \mathbb{R}$ be the function given by

$$\mathrm{tr}(A) = a_{11} + a_{22} + \cdots + a_{nn},$$

called the *trace* of A. It is a linear form. Let $s\colon \mathrm{M}_n(\mathbb{R}) \to \mathbb{R}$ be the function given by

$$s(A) = \sum_{i,j=1}^{n} a_{ij},$$

where $A = (a_{ij})$. It is immediately verified that s is a linear form.

Given a vector space E and any basis $(u_i)_{i\in I}$ for E, we can associate to each u_i a linear form $u_i^* \in E^*$, and the u_i^* have some remarkable properties.

Definition 1.23. Given a vector space E and any basis $(u_i)_{i\in I}$ for E, by Proposition 1.13, for every $i \in I$, there is a unique linear form u_i^* such that

$$u_i^*(u_j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j, \end{cases}$$

for every $j \in I$. The linear form u_i^* is called the *coordinate form* of index i w.r.t. the basis $(u_i)_{i\in I}$.

Remark: Given an index set I, authors often define the so called "Kronecker symbol" δ_{ij} such that

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

for all $i, j \in I$. Then, $u_i^*(u_j) = \delta_{ij}$.

The reason for the terminology *coordinate form* is as follows: If E has finite dimension and if $(u_1, \ldots, u_n)$ is a basis of E, for any vector

$$v = \lambda_1 u_1 + \cdots + \lambda_n u_n,$$

we have

$$\begin{aligned} u_i^*(v) &= u_i^*(\lambda_1 u_1 + \cdots + \lambda_n u_n) \\ &= \lambda_1 u_i^*(u_1) + \cdots + \lambda_i u_i^*(u_i) + \cdots + \lambda_n u_i^*(u_n) \\ &= \lambda_i, \end{aligned}$$

since $u_i^*(u_j) = \delta_{i\,j}$. Therefore, u_i^* is the linear function that returns the ith coordinate of a vector expressed over the basis $(u_1, \ldots, u_n)$.

The following theorem shows that in finite-dimension, every basis $(u_1, \ldots, u_n)$ of a vector space E yields a basis $(u_1^*, \ldots, u_n^*)$ of the dual space E^*, called a *dual basis*.

Theorem 1.16. *(Existence of dual bases) Let E be a vector space of dimension n. The following properties hold: For every basis $(u_1, \ldots, u_n)$ of E, the family of coordinate forms $(u_1^*, \ldots, u_n^*)$ is a basis of E^* (called the dual basis of $(u_1, \ldots, u_n)$).*

Proof. (a) If $v^* \in E^*$ is any linear form, consider the linear form

$$f^* = v^*(u_1)u_1^* + \cdots + v^*(u_n)u_n^*.$$

Observe that because $u_i^*(u_j) = \delta_{i\,j}$,

$$\begin{aligned} f^*(u_i) &= (v^*(u_1)u_1^* + \cdots + v^*(u_n)u_n^*)(u_i) \\ &= v^*(u_1)u_1^*(u_i) + \cdots + v^*(u_i)u_i^*(u_i) + \cdots + v^*(u_n)u_n^*(u_i) \\ &= v^*(u_i), \end{aligned}$$

and so f^* and v^* agree on the basis $(u_1, \ldots, u_n)$, which implies that

$$v^* = f^* = v^*(u_1)u_1^* + \cdots + v^*(u_n)u_n^*.$$

Therefore, $(u_1^*, \ldots, u_n^*)$ spans E^*. We claim that the covectors $u_1^*, \ldots, u_n^*$ are linearly independent. If not, we have a nontrivial linear dependence

$$\lambda_1 u_1^* + \cdots + \lambda_n u_n^* = 0,$$

and if we apply the above linear form to each u_i, using a familar computation, we get

$$0 = \lambda_i u_i^*(u_i) = \lambda_i,$$

proving that $u_1^*, \ldots, u_n^*$ are indeed linearly independent. Therefore, $(u_1^*, \ldots, u_n^*)$ is a basis of E^*. □

In particular, Theorem 1.16 shows a finite-dimensional vector space and its dual E^* have the same dimension.

1.9 Summary

The main concepts and results of this chapter are listed below:

- The notion of a *vector space*.
- *Families* of vectors.
- *Linear combinations* of vectors; *linear dependence* and *linear independence* of a family of vectors.
- Linear *subspaces*.
- *Spanning* (or *generating*) family; *generators*, *finitely generated subspace*; *basis of a subspace*.
- *Every linearly independent family can be extended to a basis* (Theorem 1.5).
- A family B of vectors is a basis iff it is a maximal linearly independent family iff it is a minimal generating family (Proposition 1.6).
- The replacement lemma (Proposition 1.8).
- Any two bases in a finitely generated vector space E have the *same number of elements*; this is the *dimension* of E (Theorem 1.9).
- *Hyperplanes*.
- Every vector has a *unique representation* over a basis (in terms of its coordinates).
- *matrices*
- *Column vectors*, *row vectors*.
- *Matrix operations*: addition, scalar multiplication, multiplication.
- The vector space $\mathrm{M}_{m,n}(K)$ of $m \times n$ matrices over the field K; The ring $\mathrm{M}_n(K)$ of $n \times n$ matrices over the field K.
- The notion of a *linear map*.
- The *image* $\mathrm{Im}\, f$ (or *range*) of a linear map f.
- The *kernel* $\mathrm{Ker}\, f$ (or *nullspace*) of a linear map f.
- The *rank* $\mathrm{rk}(f)$ of a linear map f.
- The image and the kernel of a linear map are subspaces. A linear map is injective iff its kernel is the trivial space (0) (Proposition 1.12).

- The *unique homomorphic extension property* of linear maps with respect to bases (Proposition 1.13).
- Linear forms (covectors) and the *dual space* E^*.
- Coordinate forms.
- The existence of *dual bases* (in finite dimension).

Chapter 2

Matrices and Linear Maps

2.1 Representation of Linear Maps by Matrices

Proposition 1.13 shows that given two vector spaces E and F and a basis $(u_j)_{j\in J}$ of E, every linear map $f\colon E \to F$ is uniquely determined by the family $(f(u_j))_{j\in J}$ of the images under f of the vectors in the basis $(u_j)_{j\in J}$.

If we also have a basis $(v_i)_{i\in I}$ of F, then every vector $f(u_j)$ can be written in a unique way as

$$f(u_j) = \sum_{i\in I} a_{i\,j} v_i,$$

where $j \in J$, for a family of scalars $(a_{i\,j})_{i\in I}$. Thus, with respect to the two bases $(u_j)_{j\in J}$ of E and $(v_i)_{i\in I}$ of F, the linear map f is completely determined by a "$I \times J$-matrix" $M(f) = (a_{i\,j})_{i\in I,\ j\in J}$.

Remark: Note that we intentionally assigned the index set J to the basis $(u_j)_{j\in J}$ of E, and the index set I to the basis $(v_i)_{i\in I}$ of F, so that the rows of the matrix $M(f)$ associated with $f\colon E \to F$ are indexed by I, and the columns of the matrix $M(f)$ are indexed by J. Obviously, this causes a mildly unpleasant reversal. If we had considered the bases $(u_i)_{i\in I}$ of E and $(v_j)_{j\in J}$ of F, we would obtain a $J \times I$-matrix $M(f) = (a_{j\,i})_{j\in J,\ i\in I}$. No matter what we do, there will be a reversal! We decided to stick to the bases $(u_j)_{j\in J}$ of E and $(v_i)_{i\in I}$ of F, so that we get an $I \times J$-matrix $M(f)$, knowing that we may occasionally suffer from this decision!

When I and J are finite, and say, when $|I| = m$ and $|J| = n$, the linear map f is determined by the matrix $M(f)$ whose entries in the j-th column are the components of the vector $f(u_j)$ over the basis $(v_1, \ldots, v_m)$, that is, the matrix

$$M(f) = \begin{pmatrix} a_{1\,1} & a_{1\,2} & \ldots & a_{1\,n} \\ a_{2\,1} & a_{2\,2} & \ldots & a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m\,1} & a_{m\,2} & \ldots & a_{m\,n} \end{pmatrix}$$

whose entry on row i and column j is a_{ij} ($1 \leq i \leq m$, $1 \leq j \leq n$).

We will now show that when E and F have finite dimension, linear maps can be very conveniently represented by matrices, and that composition of linear maps corresponds to matrix multiplication. We will follow rather closely an elegant presentation method due to Emil Artin.

Let E and F be two vector spaces, and assume that E has a finite basis $(u_1, \ldots, u_n)$ and that F has a finite basis $(v_1, \ldots, v_m)$. Recall that we have shown that every vector $x \in E$ can be written in a unique way as

$$x = x_1 u_1 + \cdots + x_n u_n,$$

and similarly every vector $y \in F$ can be written in a unique way as

$$y = y_1 v_1 + \cdots + y_m v_m.$$

Let $f\colon E \to F$ be a linear map between E and F. Then, for every $x = x_1 u_1 + \cdots + x_n u_n$ in E, by linearity, we have

$$f(x) = x_1 f(u_1) + \cdots + x_n f(u_n).$$

Let

$$f(u_j) = a_{1\,j} v_1 + \cdots + a_{m\,j} v_m,$$

or more concisely,

$$f(u_j) = \sum_{i=1}^{m} a_{i\,j} v_i,$$

for every j, $1 \leq j \leq n$. This can be expressed by writing the coefficients $a_{1j}, a_{2j}, \ldots, a_{mj}$ of $f(u_j)$ over the basis $(v_1, \ldots, v_m)$, as the jth column of a matrix, as shown below:

$$\begin{array}{c} \\ v_1 \\ v_2 \\ \vdots \\ v_m \end{array} \begin{array}{c} \begin{array}{cccc} f(u_1) & f(u_2) & \ldots & f(u_n) \end{array} \\ \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix} \end{array}$$

Then, substituting the right-hand side of each $f(u_j)$ into the expression for $f(x)$, we get

$$f(x) = x_1 (\sum_{i=1}^{m} a_{i\,1} v_i) + \cdots + x_n (\sum_{i=1}^{m} a_{i\,n} v_i),$$

which, by regrouping terms to obtain a linear combination of the v_i, yields

$$f(x) = (\sum_{j=1}^{n} a_{1\,j} x_j) v_1 + \cdots + (\sum_{j=1}^{n} a_{m\,j} x_j) v_m.$$

Thus, letting $f(x) = y = y_1v_1 + \cdots + y_mv_m$, we have

$$y_i = \sum_{j=1}^{n} a_{i\,j}x_j \tag{1}$$

for all i, $1 \leq i \leq m$.

To make things more concrete, let us treat the case where $n = 3$ and $m = 2$. In this case,

$$\begin{aligned} f(u_1) &= a_{11}v_1 + a_{21}v_2 \\ f(u_2) &= a_{12}v_1 + a_{22}v_2 \\ f(u_3) &= a_{13}v_1 + a_{23}v_2, \end{aligned}$$

which in matrix form is expressed by

$$\begin{array}{c} \\ v_1 \\ v_2 \end{array} \begin{array}{c} \begin{array}{ccc} f(u_1) & f(u_2) & f(u_3) \end{array} \\ \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} \end{array},$$

and for any $x = x_1u_1 + x_2u_2 + x_3u_3$, we have

$$\begin{aligned} f(x) &= f(x_1u_1 + x_2u_2 + x_3u_3) \\ &= x_1f(u_1) + x_2f(u_2) + x_3f(u_3) \\ &= x_1(a_{11}v_1 + a_{21}v_2) + x_2(a_{12}v_1 + a_{22}v_2) + x_3(a_{13}v_1 + a_{23}v_2) \\ &= (a_{11}x_1 + a_{12}x_2 + a_{13}x_3)v_1 + (a_{21}x_1 + a_{22}x_2 + a_{23}x_3)v_2. \end{aligned}$$

Consequently, since

$$y = y_1v_1 + y_2v_2,$$

we have

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\ y_2 &= a_{21}x_1 + a_{22}x_2 + a_{23}x_3. \end{aligned}$$

This agrees with the matrix equation

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

We now formalize the representation of linear maps by matrices.

Definition 2.1. Let E and F be two vector spaces, and let $(u_1, \ldots, u_n)$ be a basis for E, and $(v_1, \ldots, v_m)$ be a basis for F. Each vector $x \in E$ expressed in the basis $(u_1, \ldots, u_n)$ as $x = x_1u_1 + \cdots + x_nu_n$ is represented by the column matrix

$$M(x) = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

and similarly for each vector $y \in F$ expressed in the basis $(v_1, \ldots, v_m)$.

Every linear map $f\colon E \to F$ is represented by the matrix $M(f) = (a_{i\,j})$, where $a_{i\,j}$ is the i-th component of the vector $f(u_j)$ over the basis $(v_1, \ldots, v_m)$, i.e., where

$$f(u_j) = \sum_{i=1}^{m} a_{i\,j} v_i, \quad \text{for every } j,\ 1 \leq j \leq n.$$

The coefficients $a_{1j}, a_{2j}, \ldots, a_{mj}$ of $f(u_j)$ over the basis $(v_1, \ldots, v_m)$ form the jth column of the matrix $M(f)$ shown below:

$$\begin{array}{c} \\ v_1 \\ v_2 \\ \vdots \\ v_m \end{array} \begin{array}{c} \begin{array}{cccc} f(u_1) & f(u_2) & \ldots & f(u_n) \end{array} \\ \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix} \end{array}.$$

The matrix $M(f)$ associated with the linear map $f\colon E \to F$ is called the *matrix of f with respect to the bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_m)$*. When $E = F$ and the basis $(v_1, \ldots, v_m)$ is identical to the basis $(u_1, \ldots, u_n)$ of E, the matrix $M(f)$ associated with $f\colon E \to E$ (as above) is called the *matrix of f with respect to the basis $(u_1, \ldots, u_n)$*.

Remark: As in the remark after Definition 1.10, there is no reason to assume that the vectors in the bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_m)$ are ordered in any particular way. However, it is often convenient to assume the natural ordering. When this is so, authors sometimes refer to the matrix $M(f)$ as the matrix of f with respect to the *ordered bases* $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_m)$.

Let us now consider how the composition of linear maps is expressed in terms of bases.

Let E, F, and G, be three vectors spaces with respective bases $(u_1, \ldots, u_p)$ for E, $(v_1, \ldots, v_n)$ for F, and $(w_1, \ldots, w_m)$ for G. Let $g\colon E \to F$ and $f\colon F \to G$ be linear maps. As explained earlier, $g\colon E \to F$ is determined by the images of the basis vectors u_j, and $f\colon F \to G$ is determined by the images of the basis vectors v_k. We would like to understand how $f \circ g\colon E \to G$ is determined by the images of the basis vectors u_j.

Remark: Note that we are considering linear maps $g\colon E \to F$ and $f\colon F \to G$, instead of $f\colon E \to F$ and $g\colon F \to G$, which yields the composition $f \circ g\colon E \to G$ instead of $g \circ f\colon E \to G$. Our perhaps unusual choice is motivated by the fact that if f is represented by a matrix $M(f) = (a_{i\,k})$ and g is represented by a matrix $M(g) = (b_{k\,j})$, then $f \circ g\colon E \to G$ is represented by the product AB of the matrices A and B. If we had adopted the other choice where $f\colon E \to F$ and $g\colon F \to G$, then $g \circ f\colon E \to G$ would be represented by the product BA. Personally, we find it easier to remember the formula for the entry in row i and column of j of the product of two matrices when this product is written by AB, rather than BA. Obviously, this is a matter of taste! We will have to live with our perhaps unorthodox choice.

Thus, let

$$f(v_k) = \sum_{i=1}^{m} a_{i\,k} w_i,$$

for every k, $1 \leq k \leq n$, and let

$$g(u_j) = \sum_{k=1}^{n} b_{k\,j} v_k,$$

for every j, $1 \leq j \leq p$; in matrix form, we have

$$\begin{array}{c} \\ w_1 \\ w_2 \\ \vdots \\ w_m \end{array} \begin{array}{c} \begin{array}{cccc} f(v_1) & f(v_2) & \dots & f(v_n) \end{array} \\ \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} \end{array}$$

and

$$\begin{array}{c} \\ v_1 \\ v_2 \\ \vdots \\ v_n \end{array} \begin{array}{c} \begin{array}{cccc} g(u_1) & g(u_2) & \dots & g(u_p) \end{array} \\ \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1p} \\ b_{21} & b_{22} & \dots & b_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \dots & b_{np} \end{pmatrix} \end{array}$$

By previous considerations, for every

$$x = x_1 u_1 + \cdots + x_p u_p,$$

letting $g(x) = y = y_1 v_1 + \cdots + y_n v_n$, we have

$$y_k = \sum_{j=1}^{p} b_{k\,j} x_j \tag{2}$$

for all k, $1 \leq k \leq n$, and for every

$$y = y_1 v_1 + \cdots + y_n v_n,$$

letting $f(y) = z = z_1 w_1 + \cdots + z_m w_m$, we have

$$z_i = \sum_{k=1}^{n} a_{i\,k} y_k \tag{3}$$

for all i, $1 \leq i \leq m$. Then, if $y = g(x)$ and $z = f(y)$, we have $z = f(g(x))$, and in view of (2) and (3), we have

$$\begin{aligned} z_i &= \sum_{k=1}^{n} a_{i\,k} (\sum_{j=1}^{p} b_{k\,j} x_j) \\ &= \sum_{k=1}^{n} \sum_{j=1}^{p} a_{i\,k} b_{k\,j} x_j \\ &= \sum_{j=1}^{p} \sum_{k=1}^{n} a_{i\,k} b_{k\,j} x_j \\ &= \sum_{j=1}^{p} (\sum_{k=1}^{n} a_{i\,k} b_{k\,j}) x_j. \end{aligned}$$

Thus, defining $c_{i\,j}$ such that

$$c_{i\,j} = \sum_{k=1}^{n} a_{i\,k} b_{k\,j},$$

for $1 \leq i \leq m$, and $1 \leq j \leq p$, we have

$$z_i = \sum_{j=1}^{p} c_{i\,j} x_j \tag{4}$$

Identity (4) shows that the composition of linear maps corresponds to the product of matrices.

Then, given a linear map $f\colon E \to F$ represented by the matrix $M(f) = (a_{i\,j})$ w.r.t. the bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_m)$, by equations (1), namely

$$y_i = \sum_{j=1}^{n} a_{i\,j} x_j \quad 1 \leq i \leq m,$$

and the definition of matrix multiplication, the equation $y = f(x)$ corresponds to the matrix equation $M(y) = M(f)M(x)$, that is,

$$\begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} a_{1\,1} & \dots & a_{1\,n} \\ \vdots & \ddots & \vdots \\ a_{m\,1} & \dots & a_{m\,n} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

Recall that

$$\begin{pmatrix} a_{1\,1} & a_{1\,2} & \dots & a_{1\,n} \\ a_{2\,1} & a_{2\,2} & \dots & a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m\,1} & a_{m\,2} & \dots & a_{m\,n} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = x_1 \begin{pmatrix} a_{1\,1} \\ a_{2\,1} \\ \vdots \\ a_{m\,1} \end{pmatrix} + x_2 \begin{pmatrix} a_{1\,2} \\ a_{2\,2} \\ \vdots \\ a_{m\,2} \end{pmatrix} + \cdots + x_n \begin{pmatrix} a_{1\,n} \\ a_{2\,n} \\ \vdots \\ a_{m\,n} \end{pmatrix}.$$

Sometimes, it is necessary to incoporate the bases $(u_1, \dots, u_n)$ and $(v_1, \dots, v_m)$ in the notation for the matrix $M(f)$ expressing f with respect to these bases. This turns out to be a messy enterprise!

We propose the following course of action:

Definition 2.2. Write $\mathcal{U} = (u_1, \dots, u_n)$ and $\mathcal{V} = (v_1, \dots, v_m)$ for the bases of E and F, and denote by $M_{\mathcal{U},\mathcal{V}}(f)$ the *matrix of f with respect to the bases $\mathcal{U}$ and $\mathcal{V}$*. Furthermore, write $x_{\mathcal{U}}$ for the coordinates $M(x) = (x_1, \dots, x_n)$ of $x \in E$ w.r.t. the basis $\mathcal{U}$ and write $y_{\mathcal{V}}$ for the coordinates $M(y) = (y_1, \dots, y_m)$ of $y \in F$ w.r.t. the basis $\mathcal{V}$. Then,

$$y = f(x)$$

is expressed in matrix form by

$$y_{\mathcal{V}} = M_{\mathcal{U},\mathcal{V}}(f)\, x_{\mathcal{U}}.$$

When $\mathcal{U} = \mathcal{V}$, we abbreviate $M_{\mathcal{U},\mathcal{V}}(f)$ as $M_{\mathcal{U}}(f)$.

The above notation seems reasonable, but it has the slight disadvantage that in the expression $M_{\mathcal{U},\mathcal{V}}(f)x_{\mathcal{U}}$, the input argument $x_{\mathcal{U}}$ which is fed to the matrix $M_{\mathcal{U},\mathcal{V}}(f)$ does not appear next to the subscript $\mathcal{U}$ in $M_{\mathcal{U},\mathcal{V}}(f)$. We could have used the notation $M_{\mathcal{V},\mathcal{U}}(f)$, and some people do that. But then, we find a bit confusing that $\mathcal{V}$ comes before $\mathcal{U}$ when f maps from the space E with the basis $\mathcal{U}$ to the space F with the basis $\mathcal{V}$. So, we prefer to use the notation $M_{\mathcal{U},\mathcal{V}}(f)$.

Be aware that other authors such as Meyer [74] use the notation $[f]_{\mathcal{U},\mathcal{V}}$, and others such as Dummit and Foote [38] use the notation $M_{\mathcal{U}}^{\mathcal{V}}(f)$, instead of $M_{\mathcal{U},\mathcal{V}}(f)$. This gets worse! You may find the notation $M_{\mathcal{V}}^{\mathcal{U}}(f)$ (as in Lang [62]), or ${}_{\mathcal{U}}[f]_{\mathcal{V}}$, or other strange notations.

Let us illustrate the representation of a linear map by a matrix in a concrete situation. Let E be the vector space $\mathbb{R}[X]_4$ of polynomials of degree at most 4, let F be the vector

space $\mathbb{R}[X]_3$ of polynomials of degree at most 3, and let the linear map be the derivative map d: that is,

$$d(P+Q) = dP + dQ$$
$$d(\lambda P) = \lambda dP,$$

with $\lambda \in \mathbb{R}$. We choose $(1, x, x^2, x^3, x^4)$ as a basis of E and $(1, x, x^2, x^3)$ as a basis of F. Then, the 4×5 matrix D associated with d is obtained by expressing the derivative dx^i of each basis vector x^i for $i = 0, 1, 2, 3, 4$ over the basis $(1, x, x^2, x^3)$. We find

$$D = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{pmatrix}.$$

Then, if P denotes the polynomial

$$P = 3x^4 - 5x^3 + x^2 - 7x + 5,$$

we have

$$dP = 12x^3 - 15x^2 + 2x - 7,$$

the polynomial P is represented by the vector $(5, -7, 1, -5, 3)$ and dP is represented by the vector $(-7, 2, -15, 12)$, and we have

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} 5 \\ -7 \\ 1 \\ -5 \\ 3 \end{pmatrix} = \begin{pmatrix} -7 \\ 2 \\ -15 \\ 12 \end{pmatrix},$$

as expected! The kernel (nullspace) of d consists of the polynomials of degree 0, that is, the constant polynomials. Therefore $\dim(\operatorname{Ker} d) = 1$, and from

$$\dim(E) = \dim(\operatorname{Ker} d) + \dim(\operatorname{Im} d)$$

(see Theorem 3.6), we get $\dim(\operatorname{Im} d) = 4$ (since $\dim(E) = 5$).

For fun, let us figure out the linear map from the vector space $\mathbb{R}[X]_3$ to the vector space $\mathbb{R}[X]_4$ given by integration (finding the primitive, or anti-derivative) of x^i, for $i = 0, 1, 2, 3$). The 5×4 matrix S representing $\int$ with respect to the same bases as before is

$$S = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 1/4 \end{pmatrix}.$$

We verify that $DS = I_4$,

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 1/4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

as it should! The equation $DS = I_4$ show that S is injective and has D as a left inverse. However, $SD \neq I_5$, and instead

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 1/4 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

because constant polynomials (polynomials of degree 0) belong to the kernel of D.

The function that associates to a linear map $f\colon E \to F$ the matrix $M(f)$ w.r.t. the bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_m)$ has the property that matrix multiplication corresponds to composition of linear maps. This allows us to transfer properties of linear maps to matrices. Here is an illustration of this technique:

Proposition 2.1. *(1) Given any matrices $A \in \mathrm{M}_{m,n}(K)$, $B \in \mathrm{M}_{n,p}(K)$, and $C \in \mathrm{M}_{p,q}(K)$, we have*

$$(AB)C = A(BC);$$

that is, matrix multiplication is associative.

(2) Given any matrices $A, B \in \mathrm{M}_{m,n}(K)$, and $C, D \in \mathrm{M}_{n,p}(K)$, for all $\lambda \in K$, we have

$$\begin{aligned} (A + B)C &= AC + BC \\ A(C + D) &= AC + AD \\ (\lambda A)C &= \lambda(AC) \\ A(\lambda C) &= \lambda(AC), \end{aligned}$$

so that matrix multiplication $\cdot\colon \mathrm{M}_{m,n}(K) \times \mathrm{M}_{n,p}(K) \to \mathrm{M}_{m,p}(K)$ is bilinear.

Proof. (1) Every $m \times n$ matrix $A = (a_{i\,j})$ defines the function $f_A\colon K^n \to K^m$ given by

$$f_A(x) = Ax,$$

for all $x \in K^n$. It is immediately verified that f_A is linear and that the matrix $M(f_A)$ representing f_A over the canonical bases in K^n and K^m is equal to A. Then, formula (4) proves that

$$M(f_A \circ f_B) = M(f_A)M(f_B) = AB,$$

so we get

$$M((f_A \circ f_B) \circ f_C) = M(f_A \circ f_B)M(f_C) = (AB)C$$

and

$$M(f_A \circ (f_B \circ f_C)) = M(f_A)M(f_B \circ f_C) = A(BC),$$

and since composition of functions is associative, we have $(f_A \circ f_B) \circ f_C = f_A \circ (f_B \circ f_C)$, which implies that

$$(AB)C = A(BC).$$

(2) It is immediately verified that if $f_1, f_2 \in \mathrm{Hom}_K(E, F)$, $A, B \in \mathrm{M}_{m,n}(K)$, $(u_1, \ldots, u_n)$ is any basis of E, and $(v_1, \ldots, v_m)$ is any basis of F, then

$$\begin{aligned} M(f_1 + f_2) &= M(f_1) + M(f_2) \\ f_{A+B} &= f_A + f_B. \end{aligned}$$

Then we have

$$\begin{aligned} (A + B)C &= M(f_{A+B})M(f_C) \\ &= M(f_{A+B} \circ f_C) \\ &= M((f_A + f_B) \circ f_C)) \\ &= M((f_A \circ f_C) + (f_B \circ f_C)) \\ &= M(f_A \circ f_C) + M(f_B \circ f_C) \\ &= M(f_A)M(f_C) + M(f_B)M(f_C) \\ &= AC + BC. \end{aligned}$$

The equation $A(C + D) = AC + AD$ is proved in a similar fashion, and the last two equations are easily verified. We could also have verified all the identities by making matrix computations. □

Note that Proposition 2.1 implies that the vector space $\mathrm{M}_n(K)$ of square matrices is a (noncommutative) ring with unit I_n. (It even shows that $\mathrm{M}_n(K)$ is an associative *algebra*.)

The following proposition states the main properties of the mapping $f \mapsto M(f)$ between $\mathrm{Hom}(E, F)$ and $\mathrm{M}_{m,n}$. In short, it is an isomorphism of vector spaces.

Proposition 2.2. *Given three vector spaces E, F, G, with respective bases $(u_1, \ldots, u_p)$, $(v_1, \ldots, v_n)$, and $(w_1, \ldots, w_m)$, the mapping $M\colon \mathrm{Hom}(E, F) \to \mathrm{M}_{n,p}$ that associates the matrix $M(g)$ to a linear map $g\colon E \to F$ satisfies the following properties for all $x \in E$, all $g, h\colon E \to F$, and all $f\colon F \to G$:*

$$\begin{aligned} M(g(x)) &= M(g)M(x) \\ M(g + h) &= M(g) + M(h) \\ M(\lambda g) &= \lambda M(g) \\ M(f \circ g) &= M(f)M(g), \end{aligned}$$

where $M(x)$ is the column vector associated with the vector x and $M(g(x))$ is the column vector associated with $g(x)$, as explained in Definition 2.1.

Thus, $M\colon \mathrm{Hom}(E,F) \to \mathrm{M}_{n,p}$ is an isomorphism of vector spaces, and when $p = n$ and the basis $(v_1,\ldots,v_n)$ is identical to the basis $(u_1,\ldots,u_p)$, $M\colon \mathrm{Hom}(E,E) \to \mathrm{M}_n$ is an isomorphism of rings.

Proof. That $M(g(x)) = M(g)M(x)$ was shown just before stating the proposition, using identity (1). The identities $M(g+h) = M(g) + M(h)$ and $M(\lambda g) = \lambda M(g)$ are straightforward, and $M(f\circ g) = M(f)M(g)$ follows from (4) and the definition of matrix multiplication. The mapping $M\colon \mathrm{Hom}(E,F) \to \mathrm{M}_{n,p}$ is clearly injective, and since every matrix defines a linear map (see Proposition 2.1), it is also surjective, and thus bijective. In view of the above identities, it is an isomorphism (and similarly for $M\colon \mathrm{Hom}(E,E) \to \mathrm{M}_n$, where Proposition 2.1 is used to show that M_n is a ring). □

In view of Proposition 2.2, it seems preferable to represent vectors from a vector space of finite dimension as column vectors rather than row vectors. Thus, from now on, we will denote vectors of $\mathbb{R}^n$ (or more generally, of K^n) as columm vectors.

2.2 Change of Basis Matrix

It is important to observe that the isomorphism $M\colon \mathrm{Hom}(E,F) \to \mathrm{M}_{n,p}$ given by Proposition 2.2 depends on the choice of the bases $(u_1,\ldots,u_p)$ and $(v_1,\ldots,v_n)$, and similarly for the isomorphism $M\colon \mathrm{Hom}(E,E) \to \mathrm{M}_n$, which depends on the choice of the basis $(u_1,\ldots,u_n)$. Thus, it would be useful to know how a change of basis affects the representation of a linear map $f\colon E \to F$ as a matrix. The following simple proposition is needed.

Proposition 2.3. *Let E be a vector space, and let $(u_1,\ldots,u_n)$ be a basis of E. For every family $(v_1,\ldots,v_n)$, let $P = (a_{i\,j})$ be the matrix defined such that $v_j = \sum_{i=1}^n a_{i\,j}u_i$. The matrix P is invertible iff $(v_1,\ldots,v_n)$ is a basis of E.*

Proof. Note that we have $P = M(f)$, the matrix associated with the unique linear map $f\colon E \to E$ such that $f(u_i) = v_i$. By Proposition 1.13, f is bijective iff $(v_1,\ldots,v_n)$ is a basis of E. Furthermore, it is obvious that the identity matrix I_n is the matrix associated with the identity $\mathrm{id}\colon E \to E$ w.r.t. any basis. If f is an isomorphism, then $f\circ f^{-1} = f^{-1}\circ f = \mathrm{id}$, and by Proposition 2.2, we get $M(f)M(f^{-1}) = M(f^{-1})M(f) = I_n$, showing that P is invertible and that $M(f^{-1}) = P^{-1}$. □

Proposition 2.3 suggests the following definition.

Definition 2.3. Given a vector space E of dimension n, for any two bases $(u_1,\ldots,u_n)$ and $(v_1,\ldots,v_n)$ of E, let $P = (a_{i\,j})$ be the invertible matrix defined such that

$$v_j = \sum_{i=1}^n a_{i\,j}u_i,$$

which is also the matrix of the identity $\mathrm{id}\colon E \to E$ with respect to the bases $(v_1, \ldots, v_n)$ and $(u_1, \ldots, u_n)$, *in that order*. Indeed, we express each $\mathrm{id}(v_j) = v_j$ over the basis $(u_1, \ldots, u_n)$. The coefficients $a_{1j}, a_{2j}, \ldots, a_{nj}$ of v_j over the basis $(u_1, \ldots, u_n)$ form the jth column of the matrix P shown below:

$$\begin{array}{c} \\ u_1 \\ u_2 \\ \vdots \\ u_n \end{array} \begin{array}{c} \begin{array}{cccc} v_1 & v_2 & \ldots & v_n \end{array} \\ \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix} \end{array}.$$

The matrix P is called the *change of basis matrix from* $(u_1, \ldots, u_n)$ *to* $(v_1, \ldots, v_n)$.

Clearly, the change of basis matrix from $(v_1, \ldots, v_n)$ to $(u_1, \ldots, u_n)$ is P^{-1}. Since $P = (a_{ij})$ is the matrix of the identity $\mathrm{id}\colon E \to E$ with respect to the bases $(v_1, \ldots, v_n)$ and $(u_1, \ldots, u_n)$, given any vector $x \in E$, if $x = x_1u_1 + \cdots + x_nu_n$ over the basis $(u_1, \ldots, u_n)$ and $x = x'_1v_1 + \cdots + x'_nv_n$ over the basis $(v_1, \ldots, v_n)$, from Proposition 2.2, we have

$$\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} a_{1\,1} & \ldots & a_{1\,n} \\ \vdots & \ddots & \vdots \\ a_{n\,1} & \ldots & a_{n\,n} \end{pmatrix} \begin{pmatrix} x'_1 \\ \vdots \\ x'_n \end{pmatrix},$$

showing that the *old* coordinates (x_i) of x (over $(u_1, \ldots, u_n)$) are expressed in terms of the *new* coordinates (x'_i) of x (over $(v_1, \ldots, v_n)$).

Now we face the painful task of assigning a "good" notation incorporating the bases $\mathcal{U} = (u_1, \ldots, u_n)$ and $\mathcal{V} = (v_1, \ldots, v_n)$ into the notation for the change of basis matrix from $\mathcal{U}$ to $\mathcal{V}$. Because the change of basis matrix from $\mathcal{U}$ to $\mathcal{V}$ is the matrix of the identity map id_E *with respect to the bases* $\mathcal{V}$ *and* $\mathcal{U}$ *in that order*, we could denote it by $M_{\mathcal{V},\mathcal{U}}(\mathrm{id})$ (Meyer [74] uses the notation $[I]_{\mathcal{V},\mathcal{U}}$). We prefer to use an abbreviation for $M_{\mathcal{V},\mathcal{U}}(\mathrm{id})$.

Definition 2.4. The *change of basis matrix from* $\mathcal{U}$ *to* $\mathcal{V}$ is denoted

$$P_{\mathcal{V},\mathcal{U}}.$$

Note that

$$P_{\mathcal{U},\mathcal{V}} = P_{\mathcal{V},\mathcal{U}}^{-1}.$$

Then, if we write $x_{\mathcal{U}} = (x_1, \ldots, x_n)$ for the *old* coordinates of x with respect to the basis $\mathcal{U}$ and $x_{\mathcal{V}} = (x'_1, \ldots, x'_n)$ for the *new* coordinates of x with respect to the basis $\mathcal{V}$, we have

$$x_{\mathcal{U}} = P_{\mathcal{V},\mathcal{U}}\, x_{\mathcal{V}}, \quad x_{\mathcal{V}} = P_{\mathcal{V},\mathcal{U}}^{-1}\, x_{\mathcal{U}}.$$

The above may look backward, but remember that the matrix $M_{\mathcal{U},\mathcal{V}}(f)$ takes input expressed over the basis $\mathcal{U}$ to output expressed over the basis $\mathcal{V}$. Consequently, $P_{\mathcal{V},\mathcal{U}}$ takes input expressed over the basis $\mathcal{V}$ to output expressed over the basis $\mathcal{U}$, and $x_{\mathcal{U}} = P_{\mathcal{V},\mathcal{U}}\, x_{\mathcal{V}}$ matches this point of view!

Beware that some authors (such as Artin [6]) define the change of basis matrix from $\mathcal{U}$ to $\mathcal{V}$ as $P_{\mathcal{U},\mathcal{V}} = P_{\mathcal{V},\mathcal{U}}^{-1}$. Under this point of view, the old basis $\mathcal{U}$ is expressed in terms of the new basis $\mathcal{V}$. We find this a bit unnatural. Also, in practice, it seems that the new basis is often expressed in terms of the old basis, rather than the other way around.

Since the matrix $P = P_{\mathcal{V},\mathcal{U}}$ expresses the *new* basis $(v_1, \ldots, v_n)$ in terms of the *old* basis $(u_1, \ldots, u_n)$, we observe that the coordinates (x_i) of a vector x vary in the *opposite direction* of the change of basis. For this reason, vectors are sometimes said to be *contravariant*. However, this expression does not make sense! Indeed, a vector in an intrinsic quantity that does not depend on a specific basis. What makes sense is that the *coordinates* of a vector vary in a contravariant fashion.

Let us consider some concrete examples of change of bases.

Example 2.1. Let $E = F = \mathbb{R}^2$, with $u_1 = (1, 0)$, $u_2 = (0, 1)$, $v_1 = (1, 1)$ and $v_2 = (-1, 1)$. The change of basis matrix P from the basis $\mathcal{U} = (u_1, u_2)$ to the basis $\mathcal{V} = (v_1, v_2)$ is

$$P = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$$

and its inverse is

$$P^{-1} = \begin{pmatrix} 1/2 & 1/2 \\ -1/2 & 1/2 \end{pmatrix}.$$

The old coordinates (x_1, x_2) with respect to (u_1, u_2) are expressed in terms of the new coordinates (x_1', x_2') with respect to (v_1, v_2) by

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1' \\ x_2' \end{pmatrix},$$

and the new coordinates (x_1', x_2') with respect to (v_1, v_2) are expressed in terms of the old coordinates (x_1, x_2) with respect to (u_1, u_2) by

$$\begin{pmatrix} x_1' \\ x_2' \end{pmatrix} = \begin{pmatrix} 1/2 & 1/2 \\ -1/2 & 1/2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

Example 2.2. Let $E = F = \mathbb{R}[X]_3$ be the set of polynomials of degree at most 3, and consider the bases $\mathcal{U} = (1, x, x^2, x^3)$ and $\mathcal{V} = (B_0^3(x), B_1^3(x), B_2^3(x), B_3^3(x))$, where $B_0^3(x), B_1^3(x), B_2^3(x), B_3^3(x)$ are the *Bernstein polynomials* of degree 3, given by

$$B_0^3(x) = (1-x)^3 \qquad B_1^3(x) = 3(1-x)^2 x \qquad B_2^3(x) = 3(1-x)x^2 \qquad B_3^3(x) = x^3.$$

By expanding the Bernstein polynomials, we find that the change of basis matrix $P_{\mathcal{V},\mathcal{U}}$ is given by

$$P_{\mathcal{V},\mathcal{U}} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -3 & 3 & 0 & 0 \\ 3 & -6 & 3 & 0 \\ -1 & 3 & -3 & 1 \end{pmatrix}.$$

We also find that the inverse of $P_{\mathcal{V},\mathcal{U}}$ is

$$P_{\mathcal{V},\mathcal{U}}^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1/3 & 0 & 0 \\ 1 & 2/3 & 1/3 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

Therefore, the coordinates of the polynomial $2x^3 - x + 1$ over the basis $\mathcal{V}$ are

$$\begin{pmatrix} 1 \\ 2/3 \\ 1/3 \\ 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1/3 & 0 & 0 \\ 1 & 2/3 & 1/3 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ 0 \\ 2 \end{pmatrix},$$

and so

$$2x^3 - x + 1 = B_0^3(x) + \frac{2}{3}B_1^3(x) + \frac{1}{3}B_2^3(x) + 2B_3^3(x).$$

Our next example is the Haar wavelets, a fundamental tool in signal processing.

2.3 Haar Basis Vectors and a Glimpse at Wavelets

We begin by considering *Haar wavelets* in $\mathbb{R}^4$. Wavelets play an important role in audio and video signal processing, especially for *compressing* long signals into much smaller ones than still retain enough information so that when they are played, we can't see or hear any difference.

Consider the four vectors w_1, w_2, w_3, w_4 given by

$$w_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \qquad w_2 = \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix} \qquad w_3 = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix} \qquad w_4 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ -1 \end{pmatrix}.$$

Note that these vectors are pairwise orthogonal, so they are indeed linearly independent (we will see this in a later chapter). Let $\mathcal{W} = \{w_1, w_2, w_3, w_4\}$ be the *Haar basis*, and let $\mathcal{U} = \{e_1, e_2, e_3, e_4\}$ be the canonical basis of $\mathbb{R}^4$. The change of basis matrix $W = P_{\mathcal{W},\mathcal{U}}$ from $\mathcal{U}$ to $\mathcal{W}$ is given by

$$W = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & -1 \end{pmatrix},$$

and we easily find that the inverse of W is given by

$$W^{-1} = \begin{pmatrix} 1/4 & 0 & 0 & 0 \\ 0 & 1/4 & 0 & 0 \\ 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1/2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}.$$

So, the vector $v = (6, 4, 5, 1)$ over the basis $\mathcal{U}$ becomes $c = (c_1, c_2, c_3, c_4)$ over the Haar basis $\mathcal{W}$, with

$$\begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{pmatrix} = \begin{pmatrix} 1/4 & 0 & 0 & 0 \\ 0 & 1/4 & 0 & 0 \\ 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1/2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} 6 \\ 4 \\ 5 \\ 1 \end{pmatrix} = \begin{pmatrix} 4 \\ 1 \\ 1 \\ 2 \end{pmatrix}.$$

Given a signal $v = (v_1, v_2, v_3, v_4)$, we first *transform* v into its coefficients $c = (c_1, c_2, c_3, c_4)$ over the Haar basis by computing $c = W^{-1}v$. Observe that

$$c_1 = \frac{v_1 + v_2 + v_3 + v_4}{4}$$

is the overall *average* value of the signal v. The coefficient c_1 corresponds to the background of the image (or of the sound). Then, c_2 gives the coarse details of v, whereas, c_3 gives the details in the first part of v, and c_4 gives the details in the second half of v.

Reconstruction of the signal consists in computing $v = Wc$. The trick for good *compression* is to throw away some of the coefficients of c (set them to zero), obtaining a *compressed signal* $\widehat{c}$, and still retain enough crucial information so that the reconstructed signal $\widehat{v} = W\widehat{c}$ looks almost as good as the original signal v. Thus, the steps are:

$$\text{input } v \longrightarrow \text{coefficients } c = W^{-1}v \longrightarrow \text{compressed } \widehat{c} \longrightarrow \text{compressed } \widehat{v} = W\widehat{c}.$$

This kind of compression scheme makes modern video conferencing possible.

It turns out that there is a faster way to find $c = W^{-1}v$, without actually using W^{-1}. This has to do with the multiscale nature of Haar wavelets.

Given the original signal $v = (6, 4, 5, 1)$ shown in Figure 2.1, we compute averages and half differences obtaining Figure 2.2. We get the coefficients $c_3 = 1$ and $c_4 = 2$. Then,

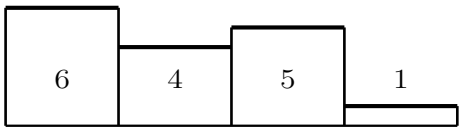


Figure 2.1: The original signal v

again we compute averages and half differences obtaining Figure 2.3. We get the coefficients $c_1 = 4$ and $c_2 = 1$. Note that the original signal v can be reconstruced from the two signals in Figure 2.2, and the signal on the left of Figure 2.2 can be reconstructed from the two signals in Figure 2.3.

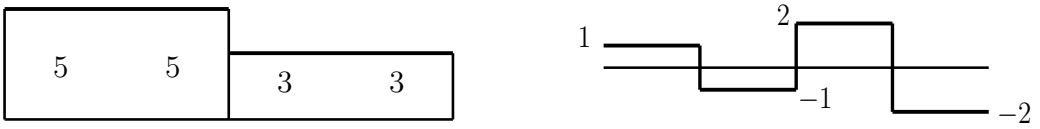


Figure 2.2: First averages and first half differences

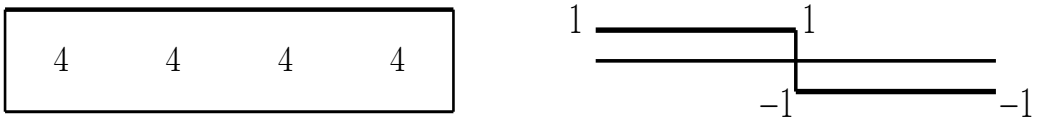


Figure 2.3: Second averages and second half differences

This method can be generalized to signals of any length 2^n. The previous case corresponds to $n = 2$. Let us consider the case $n = 3$. The *Haar basis* $(w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8)$ is given by the matrix

$$W = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & -1 & 0 & 0 & 0 \\ 1 & 1 & -1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 & -1 & 0 \\ 1 & -1 & 0 & -1 & 0 & 0 & 0 & 1 \\ 1 & -1 & 0 & -1 & 0 & 0 & 0 & -1 \end{pmatrix}.$$

The columns of this matrix are orthogonal, and it is easy to see that

$$W^{-1} = \mathrm{diag}(1/8, 1/8, 1/4, 1/4, 1/2, 1/2, 1/2, 1/2)W^\top.$$

A pattern is beginning to emerge. It looks like the second Haar basis vector w_2 is the "mother" of all the other basis vectors, except the first, whose purpose is to perform averaging. Indeed, in general, given

$$w_2 = (\underbrace{1, \ldots, 1}_{2^n}, \underbrace{-1, \ldots, -1}_{2^n}),$$

the other Haar basis vectors are obtained by a "scaling and shifting process." Starting from w_2, the scaling process generates the vectors

$$w_3, w_5, w_9, \ldots, w_{2^j+1}, \ldots, w_{2^{n-1}+1},$$

such that $w_{2^{j+1}+1}$ is obtained from w_{2^j+1} by forming two consecutive blocks of 1 and -1 of half the size of the blocks in w_{2^j+1}, and setting all other entries to zero. Observe that w_{2^j+1} has 2^j blocks of 2^{n-j} elements. The shifting process consists in shifting the blocks of 1 and -1 in w_{2^j+1} to the right by inserting a block of $(k-1)2^{n-j}$ zeros from the left, with $0 \leq j \leq n-1$ and $1 \leq k \leq 2^j$. Thus, we obtain the following formula for w_{2^j+k}:

$$w_{2^j+k}(i) = \begin{cases} 0 & 1 \leq i \leq (k-1)2^{n-j} \\ 1 & (k-1)2^{n-j}+1 \leq i \leq (k-1)2^{n-j}+2^{n-j-1} \\ -1 & (k-1)2^{n-j}+2^{n-j-1}+1 \leq i \leq k2^{n-j} \\ 0 & k2^{n-j}+1 \leq i \leq 2^n, \end{cases}$$

with $0 \leq j \leq n-1$ and $1 \leq k \leq 2^j$. Of course

$$w_1 = \underbrace{(1, \ldots, 1)}_{2^n}.$$

The above formulae look a little better if we change our indexing slightly by letting k vary from 0 to $2^j - 1$, and using the index j instead of 2^j.

Definition 2.5. The vectors of the *Haar basis* of dimension 2^n are denoted by

$$w_1, h_0^0, h_0^1, h_1^1, h_0^2, h_1^2, h_2^2, h_3^2, \ldots, h_k^j, \ldots, h_{2^{n-1}-1}^{n-1},$$

where

$$h_k^j(i) = \begin{cases} 0 & 1 \leq i \leq k2^{n-j} \\ 1 & k2^{n-j}+1 \leq i \leq k2^{n-j}+2^{n-j-1} \\ -1 & k2^{n-j}+2^{n-j-1}+1 \leq i \leq (k+1)2^{n-j} \\ 0 & (k+1)2^{n-j}+1 \leq i \leq 2^n, \end{cases}$$

with $0 \leq j \leq n-1$ and $0 \leq k \leq 2^j - 1$. The $2^n \times 2^n$ matrix whose columns are the vectors

$$w_1, h_0^0, h_0^1, h_1^1, h_0^2, h_1^2, h_2^2, h_3^2, \ldots, h_k^j, \ldots, h_{2^{n-1}-1}^{n-1},$$

(in that order), is called the *Haar matrix* of dimension 2^n, and is denoted by W_n.

It turns out that there is a way to understand these formulae better if we interpret a vector $u = (u_1, \ldots, u_m)$ as a piecewise linear function over the interval $[0, 1)$.

Definition 2.6. Given a vector $u = (u_1, \ldots, u_m)$, the *piecewise linear function* $\mathrm{plf}(u)$ is defined such that

$$\mathrm{plf}(u)(x) = u_i, \qquad \frac{i-1}{m} \leq x < \frac{i}{m},\ 1 \leq i \leq m.$$

In words, the function $\mathrm{plf}(u)$ has the value u_1 on the interval $[0, 1/m)$, the value u_2 on $[1/m, 2/m)$, *etc.*, and the value u_m on the interval $[(m-1)/m, 1)$.

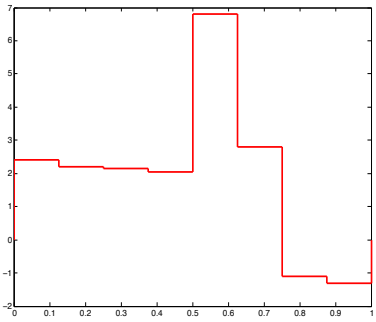


Figure 2.4: The piecewise linear function $\mathrm{plf}(u)$

For example, the piecewise linear function associated with the vector

$$u = (2.4, 2.2, 2.15, 2.05, 6.8, 2.8, -1.1, -1.3)$$

is shown in Figure 2.4.

Then, each basis vector h_k^j corresponds to the function

$$\psi_k^j = \mathrm{plf}(h_k^j).$$

In particular, for all n, the Haar basis vectors

$$h_0^0 = w_2 = \underbrace{(1, \ldots, 1, -1, \ldots, -1)}_{2^n}$$

yield the same piecewise linear function ψ given by

$$\psi(x) = \begin{cases} 1 & \text{if} \quad 0 \leq x < 1/2 \\ -1 & \text{if} \quad 1/2 \leq x < 1 \\ 0 & \text{otherwise,} \end{cases}$$

whose graph is shown in Figure 2.5. Then, it is easy to see that ψ_k^j is given by the simple expression

$$\psi_k^j(x) = \psi(2^j x - k), \quad 0 \leq j \leq n-1,\ 0 \leq k \leq 2^j - 1.$$

The above formula makes it clear that ψ_k^j is obtained from ψ by scaling and shifting.

Definition 2.7. The function $\phi_0^0 = \mathrm{plf}(w_1)$ is the piecewise linear function with the constant value 1 on $[0, 1)$, and the functions $\psi_k^j = \mathrm{plf}(h_k^j)$ together with ϕ_0^0 are known as the *Haar wavelets*.

Rather than using W^{-1} to convert a vector u to a vector c of coefficients over the Haar basis, and the matrix W to reconstruct the vector u from its Haar coefficients c, we can use faster algorithms that use averaging and differencing.

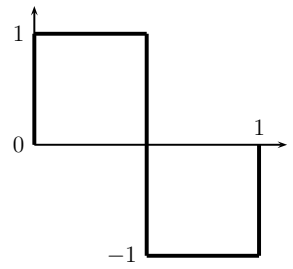


Figure 2.5: The Haar wavelet ψ

If c is a vector of Haar coefficients of dimension 2^n, we compute the sequence of vectors $u^0, u^1, \ldots, u^n$ as follows:

$$\begin{aligned} u^0 &= c \\ u^{j+1} &= u^j \\ u^{j+1}(2i-1) &= u^j(i) + u^j(2^j + i) \\ u^{j+1}(2i) &= u^j(i) - u^j(2^j + i), \end{aligned}$$

for $j = 0, \ldots, n-1$ and $i = 1, \ldots, 2^j$. The reconstructed vector (signal) is $u = u^n$.

If u is a vector of dimension 2^n, we compute the sequence of vectors $c^n, c^{n-1}, \ldots, c^0$ as follows:

$$\begin{aligned} c^n &= u \\ c^j &= c^{j+1} \\ c^j(i) &= (c^{j+1}(2i-1) + c^{j+1}(2i))/2 \\ c^j(2^j + i) &= (c^{j+1}(2i-1) - c^{j+1}(2i))/2, \end{aligned}$$

for $j = n-1, \ldots, 0$ and $i = 1, \ldots, 2^j$. The vector over the Haar basis is $c = c^0$.

We leave it as an exercise to implement the above programs in `Matlab` using two variables u and c, and by building iteratively 2^j. Here is an example of the conversion of a vector to its Haar coefficients for $n = 3$.

Given the sequence $u = (31, 29, 23, 17, -6, -8, -2, -4)$, we get the sequence

$$\begin{aligned} c^3 &= (31, 29, 23, 17, -6, -8, -2, -4) \\ c^2 &= (30, 20, -7, -3, 1, 3, 1, 1) \\ c^1 &= (25, -5, 5, -2, 1, 3, 1, 1) \\ c^0 &= (10, 15, 5, -2, 1, 3, 1, 1), \end{aligned}$$

so $c = (10, 15, 5, -2, 1, 3, 1, 1)$. Conversely, given $c = (10, 15, 5, -2, 1, 3, 1, 1)$, we get the sequence

$$\begin{aligned} u^0 &= (10, 15, 5, -2, 1, 3, 1, 1) \\ u^1 &= (25, -5, 5, -2, 1, 3, 1, 1) \\ u^2 &= (30, 20, -7, -3, 1, 3, 1, 1) \\ u^3 &= (31, 29, 23, 17, -6, -8, -2, -4), \end{aligned}$$

which gives back $u = (31, 29, 23, 17, -6, -8, -2, -4)$.

There is another recursive method for constucting the Haar matrix W_n of dimension 2^n that makes it clearer why the columns of W_n are pairwise orthogonal, and why the above algorithms are indeed correct (which nobody seems to prove!). If we split W_n into two $2^n \times 2^{n-1}$ matrices, then the second matrix containing the last 2^{n-1} columns of W_n has a very simple structure: it consists of the vector

$$\underbrace{(1, -1, 0, \ldots, 0)}_{2^n}$$

and $2^{n-1} - 1$ shifted copies of it, as illustrated below for $n = 3$:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 \end{pmatrix}.$$

Observe that this matrix can be obtained from the identity matrix $I_{2^{n-1}}$, in our example

$$I_4 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

by forming the $2^n \times 2^{n-1}$ matrix obtained by replacing each 1 by the column vector

$$\begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

and each zero by the column vector

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Now, the first half of W_n, that is the matrix consisting of the first 2^{n-1} columns of W_n, can be obtained from W_{n-1} by forming the $2^n \times 2^{n-1}$ matrix obtained by replacing each 1 by the column vector

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

each -1 by the column vector

$$\begin{pmatrix} -1 \\ -1 \end{pmatrix},$$

and each zero by the column vector

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

For $n = 3$, the first half of W_3 is the matrix

$$\begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & -1 \\ 1 & -1 & 0 & -1 \end{pmatrix}$$

which is indeed obtained from

$$W_2 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & -1 \end{pmatrix}$$

using the process that we just described.

These matrix manipulations can be described conveniently using a product operation on matrices known as the Kronecker product.

Definition 2.8. Given a $m \times n$ matrix $A = (a_{ij})$ and a $p \times q$ matrix $B = (b_{ij})$, the *Kronecker product* (or *tensor product*) $A \otimes B$ of A and B is the $mp \times nq$ matrix

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{pmatrix}.$$

It can be shown that $\otimes$ is associative and that

$$(A \otimes B)(C \otimes D) = AC \otimes BD$$
$$(A \otimes B)^\top = A^\top \otimes B^\top,$$

whenever AC and BD are well defined. Then, it is immediately verified that W_n is given by the following neat recursive equations:

$$W_n = \begin{pmatrix} W_{n-1} \otimes \begin{pmatrix} 1 \\ 1 \end{pmatrix} & I_{2^{n-1}} \otimes \begin{pmatrix} 1 \\ -1 \end{pmatrix} \end{pmatrix},$$

with $W_0 = (1)$. If we let

$$B_1 = 2\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

and for $n \geq 1$,

$$B_{n+1} = 2\begin{pmatrix} B_n & 0 \\ 0 & I_{2^n} \end{pmatrix},$$

then it is not hard to obtain a rigorous proof of the equation

$$W_n^\top W_n = B_n, \quad \text{for all } n \geq 1.$$

The above equation offers a clean justification of the fact that the columns of W_n are pairwise orthogonal.

Observe that the right block (of size $2^n \times 2^{n-1}$) shows clearly how the detail coefficients in the second half of the vector c are added and subtracted to the entries in the first half of the partially reconstructed vector after $n-1$ steps.

An important and attractive feature of the Haar basis is that it provides a *multiresolution analysis* of a signal. Indeed, given a signal u, if $c = (c_1, \ldots, c_{2^n})$ is the vector of its Haar coefficients, the coefficients with low index give coarse information about u, and the coefficients with high index represent fine information. For example, if u is an audio signal corresponding to a Mozart concerto played by an orchestra, c_1 corresponds to the "background noise," c_2 to the bass, c_3 to the first cello, c_4 to the second cello, c_5, c_6, c_7, c_7 to the violas, then the violins, *etc.* This multiresolution feature of wavelets can be exploited to compress a signal, that is, to use fewer coefficients to represent it. Here is an example.

Consider the signal

$$u = (2.4, 2.2, 2.15, 2.05, 6.8, 2.8, -1.1, -1.3),$$

whose Haar transform is

$$c = (2, 0.2, 0.1, 3, 0.1, 0.05, 2, 0.1).$$

The piecewise-linear curves corresponding to u and c are shown in Figure 2.6. Since some of the coefficients in c are small (smaller than or equal to 0.2) we can compress c by replacing them by 0. We get

$$c_2 = (2, 0, 0, 3, 0, 0, 2, 0),$$

and the reconstructed signal is

$$u_2 = (2, 2, 2, 2, 7, 3, -1, -1).$$

The piecewise-linear curves corresponding to u_2 and c_2 are shown in Figure 2.7.

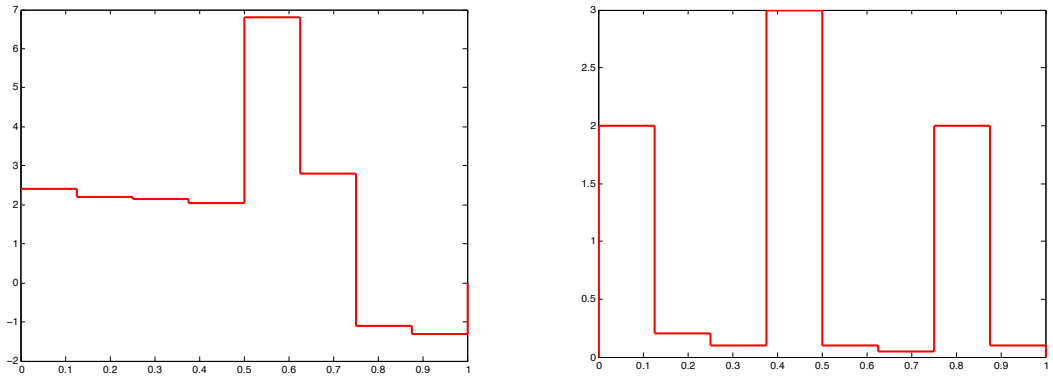


Figure 2.6: A signal and its Haar transform

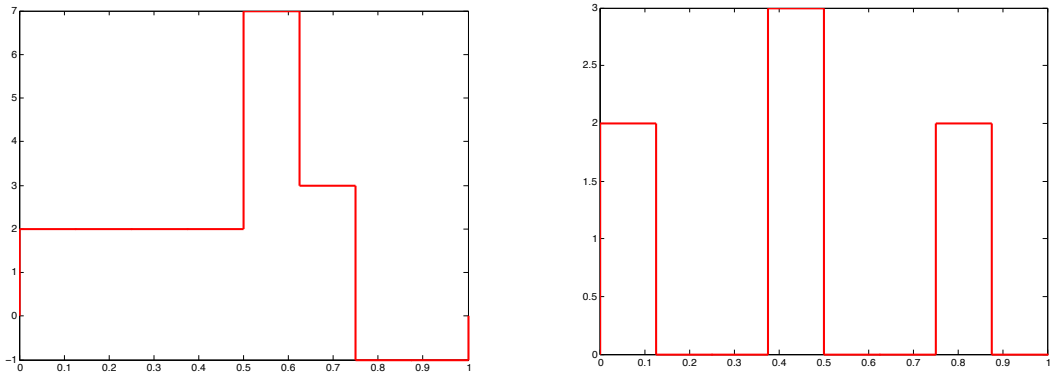


Figure 2.7: A compressed signal and its compressed Haar transform

An interesting (and amusing) application of the Haar wavelets is to the compression of audio signals. It turns out that if your type `load handel` in `Matlab` an audio file will be loaded in a vector denoted by y, and if you type `sound(y)`, the computer will play this piece of music. You can convert y to its vector of Haar coefficients c. The length of y is

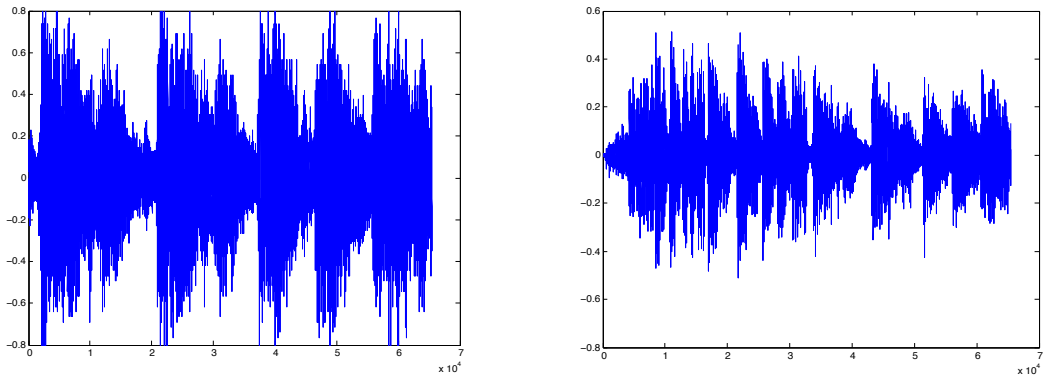


Figure 2.8: The signal "handel" and its Haar transform

73113, so first tuncate the tail of y to get a vector of length $65536 = 2^{16}$. A plot of the signals corresponding to y and c is shown in Figure 2.8. Then, run a program that sets all coefficients of c whose absolute value is less that 0.05 to zero. This sets 37272 coefficients to 0. The resulting vector c_2 is converted to a signal y_2. A plot of the signals corresponding to y_2 and c_2 is shown in Figure 2.9. When you type `sound(y2)`, you find that the music

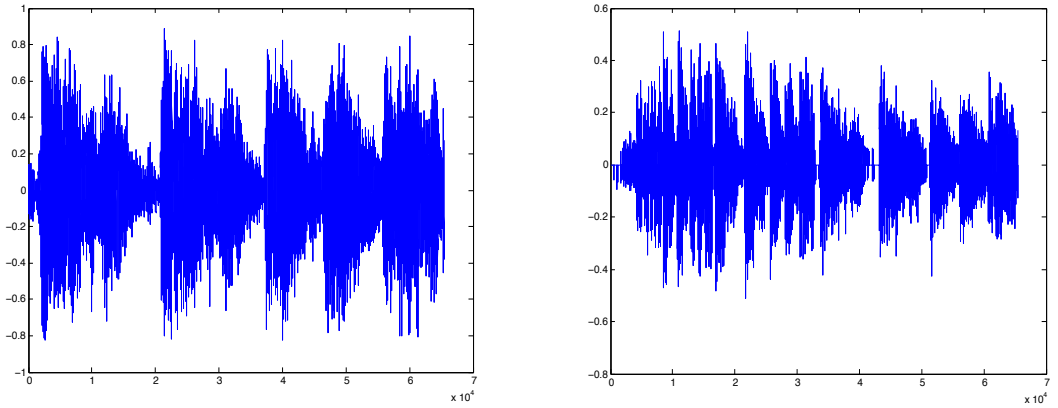


Figure 2.9: The compressed signal "handel" and its Haar transform

doesn't differ much from the original, although it sounds less crisp. You should play with other numbers greater than or less than 0.05. You should hear what happens when you type `sound(c)`. It plays the music corresponding to the Haar transform c of y, and it is quite funny.

Another neat property of the Haar transform is that it can be instantly generalized to matrices (even rectangular) without any extra effort! This allows for the compression of digital images. But first, we address the issue of normalization of the Haar coefficients. As

we observed earlier, the $2^n \times 2^n$ matrix W_n of Haar basis vectors has orthogonal columns, but its columns do not have unit length. As a consequence, $W_n^\top$ is not the inverse of W_n, but rather the matrix

$$W_n^{-1} = D_n W_n^\top$$

with $D_n = \mathrm{diag}\Big(2^{-n}, \underbrace{2^{-n}}_{2^0}, \underbrace{2^{-(n-1)}, 2^{-(n-1)}}_{2^1}, \underbrace{2^{-(n-2)}, \ldots, 2^{-(n-2)}}_{2^2}, \ldots, \underbrace{2^{-1}, \ldots, 2^{-1}}_{2^{n-1}}\Big)$.

Definition 2.9. The orthogonal matrix

$$H_n = W_n D_n^{\frac{1}{2}}$$

whose columns are the normalized Haar basis vectors, with

$$D_n^{\frac{1}{2}} = \mathrm{diag}\Big(2^{-\frac{n}{2}}, \underbrace{2^{-\frac{n}{2}}}_{2^0}, \underbrace{2^{-\frac{n-1}{2}}, 2^{-\frac{n-1}{2}}}_{2^1}, \underbrace{2^{-\frac{n-2}{2}}, \ldots, 2^{-\frac{n-2}{2}}}_{2^2}, \ldots, \underbrace{2^{-\frac{1}{2}}, \ldots, 2^{-\frac{1}{2}}}_{2^{n-1}}\Big)$$

is called the *normalized Haar transform matrix*. Given a vector (signal) u, we call $c = H_n^\top u$ the *normalized Haar coefficients* of u.

Because H_n is orthogonal, $H_n^{-1} = H_n^\top$.

Then, a moment of reflexion shows that we have to slightly modify the algorithms to compute $H_n^\top u$ and $H_n c$ as follows: When computing the sequence of u^js, use

$$u^{j+1}(2i-1) = (u^j(i) + u^j(2^j + i))/\sqrt{2}$$
$$u^{j+1}(2i) = (u^j(i) - u^j(2^j + i))/\sqrt{2},$$

and when computing the sequence of c^js, use

$$c^j(i) = (c^{j+1}(2i-1) + c^{j+1}(2i))/\sqrt{2}$$
$$c^j(2^j + i) = (c^{j+1}(2i-1) - c^{j+1}(2i))/\sqrt{2}.$$

Note that things are now more symmetric, at the expense of a division by $\sqrt{2}$. However, for long vectors, it turns out that these algorithms are numerically more stable.

Remark: Some authors (for example, Stollnitz, Derose and Salesin [100]) rescale c by $1/\sqrt{2^n}$ and u by $\sqrt{2^n}$. This is because the norm of the basis functions ψ_k^j is not equal to 1 (under the inner product $\langle f, g\rangle = \int_0^1 f(t)g(t)dt$). The normalized basis functions are the functions $\sqrt{2^j}\psi_k^j$.

Let us now explain the 2D version of the Haar transform. We describe the version using the matrix W_n, the method using H_n being identical (except that $H_n^{-1} = H_n^\top$, but this does not hold for W_n^{-1}). Given a $2^m \times 2^n$ matrix A, we can first convert the *rows* of A to their

Haar coefficients using the Haar transform W_n^{-1}, obtaining a matrix B, and then convert the *columns* of B to their Haar coefficients, using the matrix W_m^{-1}. Because columns and rows are exchanged in the first step,

$$B = A(W_n^{-1})^\top,$$

and in the second step $C = W_m^{-1}B$, thus, we have

$$C = W_m^{-1}A(W_n^{-1})^\top = D_m W_m^\top A W_n\, D_n.$$

In the other direction, given a matrix C of Haar coefficients, we reconstruct the matrix A (the image) by first applying W_m to the columns of C, obtaining B, and then $W_n^\top$ to the rows of B. Therefore

$$A = W_m C W_n^\top.$$

Of course, we dont actually have to invert W_m and W_n and perform matrix multiplications. We just have to use our algorithms using averaging and differencing. Here is an example.

If the data matrix (the image) is the 8×8 matrix

$$A = \begin{pmatrix} 64 & 2 & 3 & 61 & 60 & 6 & 7 & 57 \\ 9 & 55 & 54 & 12 & 13 & 51 & 50 & 16 \\ 17 & 47 & 46 & 20 & 21 & 43 & 42 & 24 \\ 40 & 26 & 27 & 37 & 36 & 30 & 31 & 33 \\ 32 & 34 & 35 & 29 & 28 & 38 & 39 & 25 \\ 41 & 23 & 22 & 44 & 45 & 19 & 18 & 48 \\ 49 & 15 & 14 & 52 & 53 & 11 & 10 & 56 \\ 8 & 58 & 59 & 5 & 4 & 62 & 63 & 1 \end{pmatrix},$$

then applying our algorithms, we find that

$$C = \begin{pmatrix} 32.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & -4 & 4 & -4 \\ 0 & 0 & 0 & 0 & 4 & -4 & 4 & -4 \\ 0 & 0 & 0.5 & 0.5 & 27 & -25 & 23 & -21 \\ 0 & 0 & -0.5 & -0.5 & -11 & 9 & -7 & 5 \\ 0 & 0 & 0.5 & 0.5 & -5 & 7 & -9 & 11 \\ 0 & 0 & -0.5 & -0.5 & 21 & -23 & 25 & -27 \end{pmatrix}.$$

As we can see, C has more zero entries than A; it is a compressed version of A. We can

further compress C by setting to 0 all entries of absolute value at most 0.5. Then, we get

$$C_2 = \begin{pmatrix} 32.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & -4 & 4 & -4 \\ 0 & 0 & 0 & 0 & 4 & -4 & 4 & -4 \\ 0 & 0 & 0 & 0 & 27 & -25 & 23 & -21 \\ 0 & 0 & 0 & 0 & -11 & 9 & -7 & 5 \\ 0 & 0 & 0 & 0 & -5 & 7 & -9 & 11 \\ 0 & 0 & 0 & 0 & 21 & -23 & 25 & -27 \end{pmatrix}.$$

We find that the reconstructed image is

$$A_2 = \begin{pmatrix} 63.5 & 1.5 & 3.5 & 61.5 & 59.5 & 5.5 & 7.5 & 57.5 \\ 9.5 & 55.5 & 53.5 & 11.5 & 13.5 & 51.5 & 49.5 & 15.5 \\ 17.5 & 47.5 & 45.5 & 19.5 & 21.5 & 43.5 & 41.5 & 23.5 \\ 39.5 & 25.5 & 27.5 & 37.5 & 35.5 & 29.5 & 31.5 & 33.5 \\ 31.5 & 33.5 & 35.5 & 29.5 & 27.5 & 37.5 & 39.5 & 25.5 \\ 41.5 & 23.5 & 21.5 & 43.5 & 45.5 & 19.5 & 17.5 & 47.5 \\ 49.5 & 15.5 & 13.5 & 51.5 & 53.5 & 11.5 & 9.5 & 55.5 \\ 7.5 & 57.5 & 59.5 & 5.5 & 3.5 & 61.5 & 63.5 & 1.5 \end{pmatrix},$$

which is pretty close to the original image matrix A.

It turns out that `Matlab` has a wonderful command, `image(X)` (also `imagesc(X)`, which often does a better job), which displays the matrix X has an image in which each entry is shown as a little square whose gray level is proportional to the numerical value of that entry (lighter if the value is higher, darker if the value is closer to zero; negative values are treated as zero). The images corresponding to A and C are shown in Figure 2.10. The

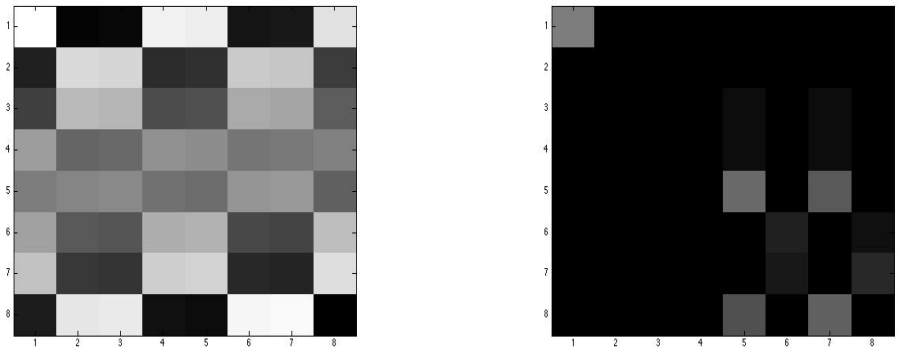


Figure 2.10: An image and its Haar transform

compressed images corresponding to A_2 and C_2 are shown in Figure 2.11. The compressed

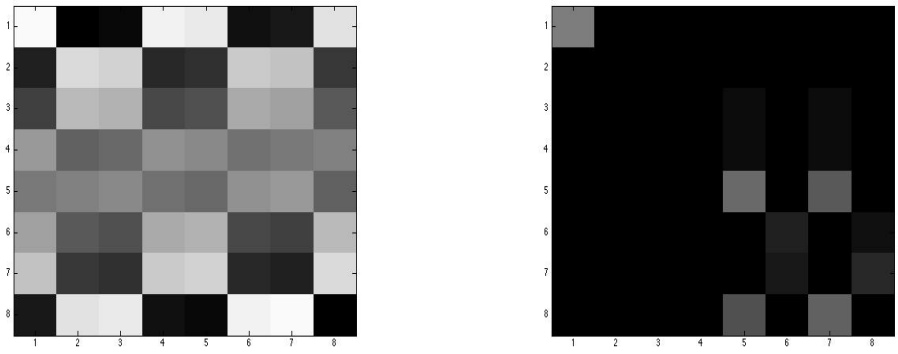


Figure 2.11: Compressed image and its Haar transform

versions appear to be indistinguishable from the originals!

If we use the normalized matrices H_m and H_n, then the equations relating the image matrix A and its normalized Haar transform C are

$$C = H_m^\top A H_n$$
$$A = H_m C H_n^\top.$$

The Haar transform can also be used to send large images progressively over the internet. Indeed, we can start sending the Haar coefficients of the matrix C starting from the coarsest coefficients (the first column from top down, then the second column, *etc.*), and at the receiving end we can start reconstructing the image as soon as we have received enough data.

Observe that instead of performing all rounds of averaging and differencing on each row and each column, we can perform partial encoding (and decoding). For example, we can perform a single round of averaging and differencing for each row and each column. The result is an image consisting of four subimages, where the top left quarter is a coarser version of the original, and the rest (consisting of three pieces) contain the finest detail coefficients. We can also perform two rounds of averaging and differencing, or three rounds, *etc.* The second round of averaging and differencing is applied to the top left quarter of the image. Generally, the kth round is applied to the $2^{m+1-k} \times 2^{n+1-k}$ submatrix consisting of the first 2^{m+1-k} rows and the first 2^{n+1-k} columns ($1 \leq k \leq n$) of the matrix obtained at the end of the previous round. This process is illustrated on the image shown in Figure 2.12. The result of performing one round, two rounds, three rounds, and nine rounds of averaging is shown in Figure 2.13. Since our images have size 512×512, nine rounds of averaging yields the Haar transform, displayed as the image on the bottom right. The original image has completely disappeared! We leave it as a fun exercise to modify the algorithms involving averaging and differencing to perform k rounds of averaging/differencing. The reconstruction algorithm is a little tricky.

Figure 2.12: Original drawing by Durer

A nice and easily accessible account of wavelets and their uses in image processing and computer graphics can be found in Stollnitz, Derose and Salesin [100]. A very detailed account is given in Strang and and Nguyen [103], but this book assumes a fair amount of background in signal processing.

We can find easily a basis of $2^n \times 2^n = 2^{2n}$ vectors w_{ij} ($2^n \times 2^n$ matrices) for the linear map that reconstructs an image from its Haar coefficients, in the sense that for any matrix C of Haar coefficients, the image matrix A is given by

$$A = \sum_{i=1}^{2^n} \sum_{j=1}^{2^n} c_{ij} w_{ij}.$$

Indeed, the matrix w_{ij} is given by the so-called outer product

$$w_{ij} = w_i (w_j)^\top.$$

Similarly, there is a basis of $2^n \times 2^n = 2^{2n}$ vectors h_{ij} ($2^n \times 2^n$ matrices) for the 2D Haar transform, in the sense that for any matrix A, its matrix C of Haar coefficients is given by

$$C = \sum_{i=1}^{2^n} \sum_{j=1}^{2^n} a_{ij} h_{ij}.$$

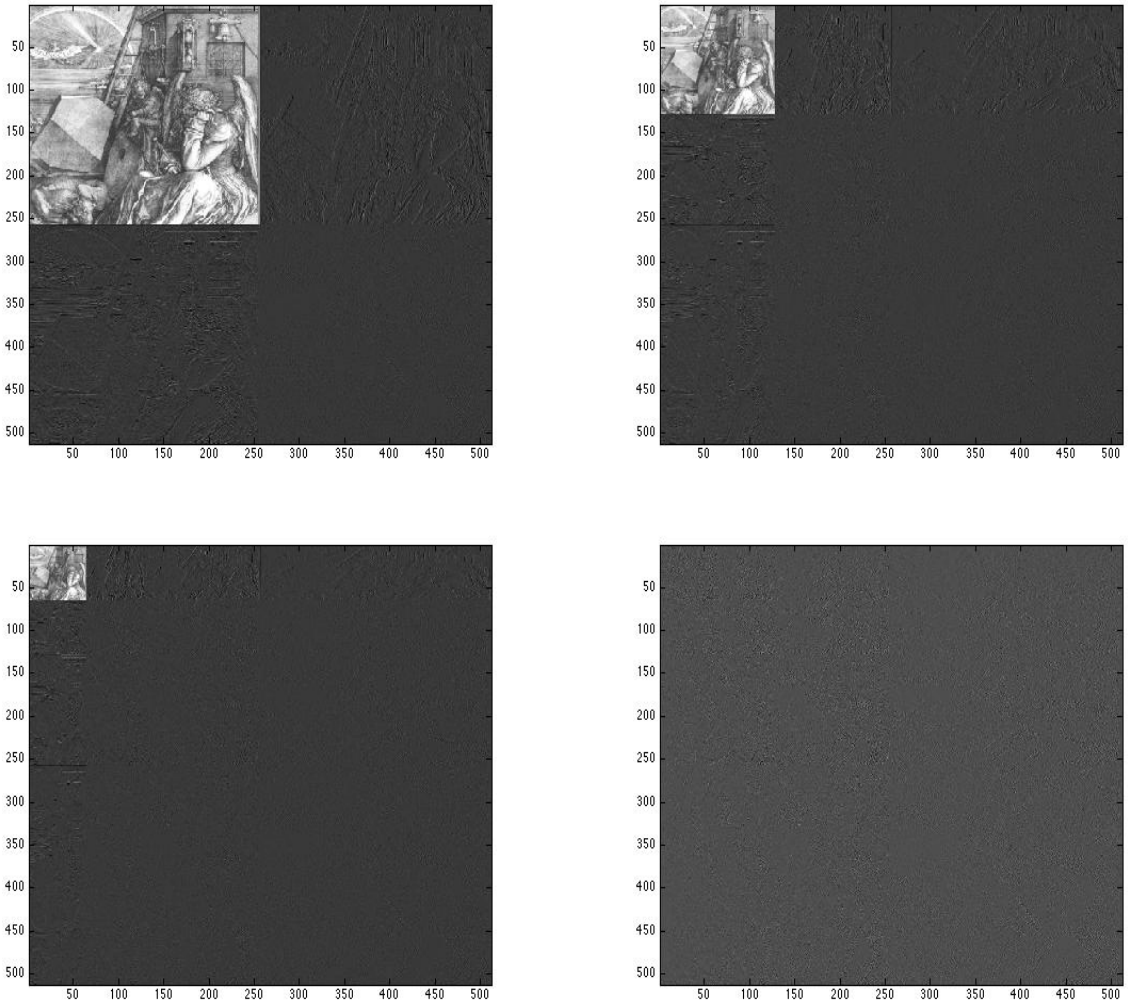


Figure 2.13: Haar tranforms after one, two, three, and nine rounds of averaging

If the columns of W^{-1} are $w'_1, \ldots, w'_{2^n}$, then

$$h_{ij} = w'_i (w'_j)^\top.$$

We leave it as exercise to compute the bases (w_{ij}) and (h_{ij}) for $n = 2$, and to display the corresponding images using the command `imagesc`.

2.4 The Effect of a Change of Bases on Matrices

The effect of a change of bases on the representation of a linear map is described in the following proposition.

Proposition 2.4. *Let E and F be vector spaces, let $\mathcal{U} = (u_1, \ldots, u_n)$ and $\mathcal{U}' = (u'_1, \ldots, u'_n)$ be two bases of E, and let $\mathcal{V} = (v_1, \ldots, v_m)$ and $\mathcal{V}' = (v'_1, \ldots, v'_m)$ be two bases of F. Let $P = P_{\mathcal{U}',\mathcal{U}}$ be the change of basis matrix from $\mathcal{U}$ to $\mathcal{U}'$, and let $Q = P_{\mathcal{V}',\mathcal{V}}$ be the change of basis matrix from $\mathcal{V}$ to $\mathcal{V}'$. For any linear map $f\colon E \to F$, let $M(f) = M_{\mathcal{U},\mathcal{V}}(f)$ be the matrix associated to f w.r.t. the bases $\mathcal{U}$ and $\mathcal{V}$, and let $M'(f) = M_{\mathcal{U}',\mathcal{V}'}(f)$ be the matrix associated to f w.r.t. the bases $\mathcal{U}'$ and $\mathcal{V}'$. We have*

$$M'(f) = Q^{-1} M(f) P,$$

or more explicitly

$$M_{\mathcal{U}',\mathcal{V}'}(f) = P^{-1}_{\mathcal{V}',\mathcal{V}} M_{\mathcal{U},\mathcal{V}}(f) P_{\mathcal{U}',\mathcal{U}} = P_{\mathcal{V},\mathcal{V}'} M_{\mathcal{U},\mathcal{V}}(f) P_{\mathcal{U}',\mathcal{U}}.$$

Proof. Since $f\colon E \to F$ can be written as $f = \mathrm{id}_F \circ f \circ \mathrm{id}_E$, since P is the matrix of id_E w.r.t. the bases $(u'_1, \ldots, u'_n)$ and $(u_1, \ldots, u_n)$, and Q^{-1} is the matrix of id_F w.r.t. the bases $(v_1, \ldots, v_m)$ and $(v'_1, \ldots, v'_m)$, by Proposition 2.2, we have $M'(f) = Q^{-1} M(f) P$. □

As a corollary, we get the following result.

Corollary 2.5. *Let E be a vector space, and let $\mathcal{U} = (u_1, \ldots, u_n)$ and $\mathcal{U}' = (u'_1, \ldots, u'_n)$ be two bases of E. Let $P = P_{\mathcal{U}',\mathcal{U}}$ be the change of basis matrix from $\mathcal{U}$ to $\mathcal{U}'$. For any linear map $f\colon E \to E$, let $M(f) = M_{\mathcal{U}}(f)$ be the matrix associated to f w.r.t. the basis $\mathcal{U}$, and let $M'(f) = M_{\mathcal{U}'}(f)$ be the matrix associated to f w.r.t. the basis $\mathcal{U}'$. We have*

$$M'(f) = P^{-1} M(f) P,$$

or more explicitly,

$$M_{\mathcal{U}'}(f) = P^{-1}_{\mathcal{U}',\mathcal{U}} M_{\mathcal{U}}(f) P_{\mathcal{U}',\mathcal{U}} = P_{\mathcal{U},\mathcal{U}'} M_{\mathcal{U}}(f) P_{\mathcal{U}',\mathcal{U}}.$$

Example 2.3. Let $E = \mathbb{R}^2$, $\mathcal{U} = (e_1, e_2)$ where $e_1 = (1, 0)$ and $e_2 = (0, 1)$ are the canonical basis vectors, let $\mathcal{V} = (v_1, v_2) = (e_1, e_1 - e_2)$, and let

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

The change of basis matrix $P = P_{\mathcal{V},\mathcal{U}}$ from $\mathcal{U}$ to $\mathcal{V}$ is

$$P = \begin{pmatrix} 1 & 1 \\ 0 & -1 \end{pmatrix},$$

and we check that

$$P^{-1} = P.$$

Therefore, in the basis $\mathcal{V}$, the matrix representing the linear map f defined by A is

$$A' = P^{-1}AP = PAP = \begin{pmatrix} 1 & 1 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} = D,$$

a diagonal matrix. In the basis $\mathcal{V}$, it is clear what the action of f is: it is a stretch by a factor of 2 in the v_1 direction and it is the identity in the v_2 direction. Observe that v_1 and v_2 are not orthogonal.

What happened is that we *diagonalized* the matrix A. The diagonal entries 2 and 1 are the *eigenvalues* of A (and f), and v_1 and v_2 are corresponding *eigenvectors*. We will come back to eigenvalues and eigenvectors later on.

The above example showed that the same linear map can be represented by different matrices. This suggests making the following definition:

Definition 2.10. Two $n \times n$ matrices A and B are said to be *similar* iff there is some invertible matrix P such that

$$B = P^{-1}AP.$$

It is easily checked that similarity is an equivalence relation. From our previous considerations, two $n \times n$ matrices A and B are similar iff they represent the same linear map with respect to two different bases. The following surprising fact can be shown: Every square matrix A is similar to its transpose $A^\top$. The proof requires advanced concepts (the Jordan form, or similarity invariants).

If $\mathcal{U} = (u_1, \ldots, u_n)$ and $\mathcal{V} = (v_1, \ldots, v_n)$ are two bases of E, the change of basis matrix

$$P = P_{\mathcal{V},\mathcal{U}} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$$

from $(u_1, \ldots, u_n)$ to $(v_1, \ldots, v_n)$ is the matrix whose jth column consists of the coordinates of v_j over the basis $(u_1, \ldots, u_n)$, which means that

$$v_j = \sum_{i=1}^{n} a_{ij} u_i.$$

It is natural to extend the matrix notation and to express the vector $\begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$ in E^n as the product of a matrix times the vector $\begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}$ in E^n, namely as

$$\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix},$$

but notice that the matrix involved is not P, but its transpose $P^\top$.

This observation has the following consequence: if $\mathcal{U} = (u_1, \ldots, u_n)$ and $\mathcal{V} = (v_1, \ldots, v_n)$ are two bases of E and if

$$\begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = A \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix},$$

that is,

$$v_i = \sum_{j=1}^{n} a_{ij} u_j,$$

for any vector $w \in E$, if

$$w = \sum_{i=1}^{n} x_i u_i = \sum_{k=1}^{n} y_k v_k,$$

then

$$\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = A^\top \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},$$

and so

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = (A^\top)^{-1} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

It is easy to see that $(A^\top)^{-1} = (A^{-1})^\top$. Also, if $\mathcal{U} = (u_1, \ldots, u_n)$, $\mathcal{V} = (v_1, \ldots, v_n)$, and $\mathcal{W} = (w_1, \ldots, w_n)$ are three bases of E, and if the change of basis matrix from $\mathcal{U}$ to $\mathcal{V}$ is $P = P_{\mathcal{V},\mathcal{U}}$ and the change of basis matrix from $\mathcal{V}$ to $\mathcal{W}$ is $Q = P_{\mathcal{W},\mathcal{V}}$, then

$$\begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = P^\top \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}, \quad \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} = Q^\top \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix},$$

so

$$\begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} = Q^\top P^\top \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} = (PQ)^\top \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix},$$

which means that the change of basis matrix $P_{\mathcal{W},\mathcal{U}}$ from $\mathcal{U}$ to $\mathcal{W}$ is PQ. This proves that

$$P_{\mathcal{W},\mathcal{U}} = P_{\mathcal{V},\mathcal{U}} P_{\mathcal{W},\mathcal{V}}.$$

Even though matrices are indispensable since they are *the* major tool in applications of linear algebra, one should not lose track of the fact that

linear maps are more fundamental, because they are intrinsic objects that do not depend on the choice of bases. Consequently, we advise the reader to try to think in terms of linear maps rather than reduce everthing to matrices.

In our experience, this is particularly effective when it comes to proving results about linear maps and matrices, where proofs involving linear maps are often more "conceptual." These proofs are usually more general because they do not depend on the fact that the dimension is finite. Also, instead of thinking of a matrix decomposition as a purely algebraic operation, it is often illuminating to view it as a *geometric decomposition*. This is the case of the SVD, which in geometric term says that every linear map can be factored as a rotation, followed by a rescaling along orthogonal axes, and then another rotation.

After all, a

a matrix is a representation of a linear map

and most decompositions of a matrix reflect the fact that with a *suitable choice of a basis (or bases)*, the linear map is a represented by a matrix having a special shape. The problem is then to find such bases.

Still, for the beginner, matrices have a certain irresistible appeal, and we confess that it takes a certain amount of practice to reach the point where it becomes more natural to deal with linear maps. We still recommend it! For example, try to translate a result stated in terms of matrices into a result stated in terms of linear maps. Whenever we tried this exercise, we learned something.

Also, always try to keep in mind that

linear maps are geometric in nature; they act on space.

2.5 Summary

The main concepts and results of this chapter are listed below:

- The representation of linear maps by *matrices*.
- The vector space of linear maps $\mathrm{Hom}_K(E, F)$.
- The *matrix representation mapping* $M : \mathrm{Hom}(E, F) \to \mathrm{M}_{n,p}$ and the representation isomorphism (Proposition 2.2).
- Haar basis vectors and a glimpse at *Haar wavelets*.
- *Kronecker product* (or *tensor product*) of matrices.
- *Change of basis matrix* and Proposition 2.4.

Chapter 3

Direct Sums, Affine Maps

3.1 Direct Products

There are some useful ways of forming new vector spaces from older ones.

Definition 3.1. Given $p \geq 2$ vector spaces $E_1, \ldots, E_p$, the product $F = E_1 \times \cdots \times E_p$ can be made into a vector space by defining addition and scalar multiplication as follows:

$$(u_1, \ldots, u_p) + (v_1, \ldots, v_p) = (u_1 + v_1, \ldots, u_p + v_p)$$
$$\lambda(u_1, \ldots, u_p) = (\lambda u_1, \ldots, \lambda u_p),$$

for all $u_i, v_i \in E_i$ and all $\lambda \in \mathbb{R}$. The zero vector of $E_1 \times \cdots \times E_p$ is the p-tuple

$$(\underbrace{0, \ldots, 0}_{p}),$$

where the ith zero is the zero vector of E_i.

With the above addition and multiplication, the vector space $F = E_1 \times \cdots \times E_p$ is called the *direct product* of the vector spaces $E_1, \ldots, E_p$.

As a special case, when $E_1 = \cdots = E_p = \mathbb{R}$, we find again the vector space $F = \mathbb{R}^p$. The *projection maps* $pr_i \colon E_1 \times \cdots \times E_p \to E_i$ given by

$$pr_i(u_1, \ldots, u_p) = u_i$$

are clearly linear. Similarly, the maps $\mathrm{in}_i \colon E_i \to E_1 \times \cdots \times E_p$ given by

$$\mathrm{in}_i(u_i) = (0, \ldots, 0, u_i, 0, \ldots, 0)$$

are injective and linear. If $\dim(E_i) = n_i$ and if $(e_1^i, \ldots, e_{n_i}^i)$ is a basis of E_i for $i = 1, \ldots, p$, then it is easy to see that the $n_1 + \cdots + n_p$ vectors

$$\begin{array}{ccc} (e_1^1, 0, \ldots, 0), & \ldots, & (e_{n_1}^1, 0, \ldots, 0), \\ \vdots & \vdots & \vdots \\ (0, \ldots, 0, e_1^i, 0, \ldots, 0), & \ldots, & (0, \ldots, 0, e_{n_i}^i, 0, \ldots, 0), \\ \vdots & \vdots & \vdots \\ (0, \ldots, 0, e_1^p), & \ldots, & (0, \ldots, 0, e_{n_p}^p) \end{array}$$

form a basis of $E_1 \times \cdots \times E_p$, and so

$$\dim(E_1 \times \cdots \times E_p) = \dim(E_1) + \cdots + \dim(E_p).$$

3.2 Sums and Direct Sums

Let us now consider a vector space E and p subspaces $U_1, \ldots, U_p$ of E. We have a map

$$a \colon U_1 \times \cdots \times U_p \to E$$

given by

$$a(u_1, \ldots, u_p) = u_1 + \cdots + u_p,$$

with $u_i \in U_i$ for $i = 1, \ldots, p$. It is clear that this map is linear, and so its image is a subspace of E denoted by

$$U_1 + \cdots + U_p$$

and called the *sum* of the subspaces $U_1, \ldots, U_p$. By definition,

$$U_1 + \cdots + U_p = \{u_1 + \cdots + u_p \mid u_i \in U_i,\ 1 \leq i \leq p\},$$

and it is immediately verified that $U_1 + \cdots + U_p$ is the smallest subspace of E containing $U_1, \ldots, U_p$. This also implies that $U_1 + \cdots + U_p$ does not depend on the order of the factors U_i; in particular,

$$U_1 + U_2 = U_2 + U_1.$$

If the map a is injective, then by Proposition 1.12 we have $\operatorname{Ker} a = \{(\underbrace{0, \ldots, 0}_{p})\}$ where each 0 is the zero vector of E, which means that if $u_i \in U_i$ for $i = 1, \ldots, p$ and if

$$u_1 + \cdots + u_p = 0,$$

then $(u_1, \ldots, u_p) = (0, \ldots, 0)$, that is, $u_1 = 0, \ldots, u_p = 0$. In this case, every $u \in U_1 + \cdots + U_p$ has a *unique* expression as a sum

$$u = u_1 + \cdots + u_p,$$

with $u_i \in U_i$, for $i = 1, \ldots, p$. Indeed, if

$$u = v_1 + \cdots + v_p = w_1 + \cdots + w_p,$$

with $v_i, w_i \in U_i$, for $i = 1, \ldots, p$, then we have

$$w_1 - v_1 + \cdots + w_p - v_p = 0,$$

and since $v_i, w_i \in U_i$ and each U_i is a subspace, $w_i - v_i \in U_i$. The injectivity of a implies that $w_i - v_i = 0$, that is, $w_i = v_i$ for $i = 1, \ldots, p$, which shows the uniqueness of the decomposition of u.

It is also clear that any p nonzero vectors $u_1, \ldots, u_p$ with $u_i \in U_i$ are linearly independent. To see this, assume that

$$\lambda_1 u_1 + \cdots + \lambda_p u_p = 0$$

for some $\lambda_i \in \mathbb{R}$. Since $u_i \in U_i$ and U_i is a subspace, $\lambda_i u_i \in U_i$, and the injectivity of a implies that $\lambda_i u_i = 0$, for $i = 1, \ldots, p$. Since $u_i \neq 0$, we must have $\lambda_i = 0$ for $i = 1, \ldots, p$; that is, $u_1, \ldots, u_p$ with $u_i \in U_i$ and $u_i \neq 0$ are linearly independent.

Observe that if a is injective, then we must have $U_i \cap U_j = (0)$ whenever $i \neq j$. However, this condition is generally not sufficient if $p \geq 3$. For example, if $E = \mathbb{R}^2$ and U_1 the line spanned by $e_1 = (1, 0)$, U_2 is the line spanned by $d = (1, 1)$, and U_3 is the line spanned by $e_2 = (0, 1)$, then $U_1 \cap U_2 = U_1 \cap U_3 = U_2 \cap U_3 = \{(0, 0)\}$, but $U_1 + U_2 = U_1 + U_3 = U_2 + U_3 = \mathbb{R}^2$, so $U_1 + U_2 + U_3$ is not a direct sum. For example, d is expressed in two different ways as

$$d = (1, 1) = (1, 0) + (0, 1) = e_1 + e_2.$$

Definition 3.2. For any vector space E and any $p \geq 2$ subspaces $U_1, \ldots, U_p$ of E, if the map a defined above is injective, then the sum $U_1 + \cdots + U_p$ is called a *direct sum* and it is denoted by

$$U_1 \oplus \cdots \oplus U_p.$$

The space E is the *direct sum* of the subspaces U_i if

$$E = U_1 \oplus \cdots \oplus U_p.$$

As in the case of a sum, $U_1 \oplus U_2 = U_2 \oplus U_1$. Observe that when the map a is injective, then it is a linear isomorphism between $U_1 \times \cdots \times U_p$ and $U_1 \oplus \cdots \oplus U_p$. The difference is that $U_1 \times \cdots \times U_p$ is defined even if the spaces U_i are not assumed to be subspaces of some common space.

If E is a direct sum $E = U_1 \oplus \cdots \oplus U_p$, since any p nonzero vectors $u_1, \ldots, u_p$ with $u_i \in U_i$ are linearly independent, if we pick a basis $(u_k)_{k \in I_j}$ in U_j for $j = 1, \ldots, p$, then $(u_i)_{i \in I}$ with $I = I_1 \cup \cdots \cup I_p$ is a basis of E. Intuitively, E is split into p independent subspaces.

Conversely, given a basis $(u_i)_{i\in I}$ of E, if we partition the index set I as $I = I_1 \cup \cdots \cup I_p$, then each subfamily $(u_k)_{k\in I_j}$ spans some subspace U_j of E, and it is immediately verified that we have a direct sum

$$E = U_1 \oplus \cdots \oplus U_p.$$

Let $f\colon E \to E$ be a linear map. If $f(U_j) \subseteq U_j$ we say that U_j is *invariant under* f. Assume that E is finite-dimensional, a direct sum $E = U_1 \oplus \cdots \oplus U_p$, and that each U_j is invariant under f. If we pick a basis $(u_i)_{i\in I}$ as above with $I = I_1 \cup \cdots \cup I_p$ and with each $(u_k)_{k\in I_j}$ a basis of U_j, since each U_j is invariant under f, the image $f(u_k)$ of every basis vector u_k with $k \in I_j$ belongs to U_j, so the matrix A representing f over the basis $(u_i)_{i\in I}$ is a *block diagonal* matrix of the form

$$A = \begin{pmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_p \end{pmatrix},$$

with each block A_j a $d_j \times d_j$-matrix with $d_j = \dim(U_j)$ and all other entries equal to 0. If $d_j = 1$ for $j = 1, \ldots, p$, the matrix A is a diagonal matrix.

There are natural injections from each U_i to E denoted by $\mathrm{in}_i\colon U_i \to E$.

Now, if $p = 2$, it is easy to determine the kernel of the map $a\colon U_1 \times U_2 \to E$. We have

$$a(u_1, u_2) = u_1 + u_2 = 0 \quad \text{iff} \quad u_1 = -u_2,\ u_1 \in U_1, u_2 \in U_2,$$

which implies that

$$\operatorname{Ker} a = \{(u, -u) \mid u \in U_1 \cap U_2\}.$$

Now, $U_1 \cap U_2$ is a subspace of E and the linear map $u \mapsto (u, -u)$ is clearly an isomorphism between $U_1 \cap U_2$ and $\operatorname{Ker} a$, so $\operatorname{Ker} a$ is isomorphic to $U_1 \cap U_2$. As a consequence, we get the following result:

Proposition 3.1. *Given any vector space E and any two subspaces U_1 and U_2, the sum $U_1 + U_2$ is a direct sum iff $U_1 \cap U_2 = (0)$.*

An interesting illustration of the notion of direct sum is the decomposition of a square matrix into its symmetric part and its skew-symmetric part. Recall that an $n \times n$ matrix $A \in \mathrm{M}_n$ is *symmetric* if $A^\top = A$, *skew -symmetric* if $A^\top = -A$. It is clear that

$$\mathbf{S}(n) = \{A \in \mathrm{M}_n \mid A^\top = A\} \quad \text{and} \quad \mathbf{Skew}(n) = \{A \in \mathrm{M}_n \mid A^\top = -A\}$$

are subspaces of M_n, and that $\mathbf{S}(n) \cap \mathbf{Skew}(n) = (0)$. Observe that for any matrix $A \in \mathrm{M}_n$, the matrix $H(A) = (A + A^\top)/2$ is symmetric and the matrix $S(A) = (A - A^\top)/2$ is skew-symmetric. Since

$$A = H(A) + S(A) = \frac{A + A^\top}{2} + \frac{A - A^\top}{2},$$

we see that $\mathrm{M}_n = \mathbf{S}(n) + \mathbf{Skew}(n)$, and since $\mathbf{S}(n) \cap \mathbf{Skew}(n) = (0)$, we have the direct sum

$$\mathrm{M}_n = \mathbf{S}(n) \oplus \mathbf{Skew}(n).$$

Remark: The vector space $\mathbf{Skew}(n)$ of skew-symmetric matrices is also denoted by $\mathfrak{so}(n)$. It is the *Lie algebra* of the group $\mathbf{SO}(n)$.

Proposition 3.1 can be generalized to any $p \geq 2$ subspaces at the expense of notation. The proof of the following proposition is left as an exercise.

Proposition 3.2. *Given any vector space E and any $p \geq 2$ subspaces $U_1, \ldots, U_p$, the following properties are equivalent:*

(1) The sum $U_1 + \cdots + U_p$ is a direct sum.

(2) We have

$$U_i \cap \left(\sum_{j=1, j \neq i}^{p} U_j \right) = (0), \quad i = 1, \ldots, p.$$

(3) We have

$$U_i \cap \left(\sum_{j=1}^{i-1} U_j \right) = (0), \quad i = 2, \ldots, p.$$

Because of the isomorphism

$$U_1 \times \cdots \times U_p \approx U_1 \oplus \cdots \oplus U_p,$$

we have

Proposition 3.3. *If E is any vector space, for any (finite-dimensional) subspaces $U_1, \ldots, U_p$ of E, we have*

$$\dim(U_1 \oplus \cdots \oplus U_p) = \dim(U_1) + \cdots + \dim(U_p).$$

If E is a direct sum

$$E = U_1 \oplus \cdots \oplus U_p,$$

since every $u \in E$ can be written in a unique way as

$$u = u_1 + \cdots + u_p$$

with $u_i \in U_i$ for $i = 1 \ldots, p$, we can define the maps $\pi_i \colon E \to U_i$, called *projections*, by

$$\pi_i(u) = \pi_i(u_1 + \cdots + u_p) = u_i.$$

It is easy to check that these maps are linear and satisfy the following properties:

$$\pi_j \circ \pi_i = \begin{cases} \pi_i & \text{if } i = j \\ 0 & \text{if } i \neq j, \end{cases}$$
$$\pi_1 + \cdots + \pi_p = \mathrm{id}_E.$$

For example, in the case of the direct sum

$$\mathrm{M}_n = \mathbf{S}(n) \oplus \mathbf{Skew}(n),$$

the projection onto $\mathbf{S}(n)$ is given by

$$\pi_1(A) = H(A) = \frac{A + A^\top}{2},$$

and the projection onto $\mathbf{Skew}(n)$ is given by

$$\pi_2(A) = S(A) = \frac{A - A^\top}{2}.$$

Clearly, $H(A)+S(A) = A$, $H(H(A)) = H(A)$, $S(S(A)) = S(A)$, and $H(S(A)) = S(H(A)) = 0$.

A function f such that $f \circ f = f$ is said to be *idempotent*. Thus, the projections π_i are idempotent. Conversely, the following proposition can be shown:

Proposition 3.4. *Let E be a vector space. For any $p \geq 2$ linear maps $f_i \colon E \to E$, if*

$$f_j \circ f_i = \begin{cases} f_i & \text{if } i = j \\ 0 & \text{if } i \neq j, \end{cases}$$
$$f_1 + \cdots + f_p = \mathrm{id}_E,$$

then if we let $U_i = f_i(E)$, we have a direct sum

$$E = U_1 \oplus \cdots \oplus U_p.$$

We also have the following proposition characterizing idempotent linear maps whose proof is also left as an exercise.

Proposition 3.5. *For every vector space E, if $f \colon E \to E$ is an idempotent linear map, i.e., $f \circ f = f$, then we have a direct sum*

$$E = \operatorname{Ker} f \oplus \operatorname{Im} f,$$

so that f is the projection onto its image $\operatorname{Im} f$.

We are now ready to prove a very crucial result relating the rank and the dimension of the kernel of a linear map.

3.3 The Rank-Nullity Theorem; Grassmann's Relation

We begin with the following theorem which shows that given a linear map $f\colon E \to F$, its domain E is the direct sum of its kernel $\operatorname{Ker} f$ with some isomorphic copy of its image $\operatorname{Im} f$.

Theorem 3.6. *(Rank-nullity theorem) Let $f\colon E \to F$ be a linear map. For any choice of a basis $(f_1, \ldots, f_r)$ of $\operatorname{Im} f$, let $(u_1, \ldots, u_r)$ be any vectors in E such that $f_i = f(u_i)$, for $i = 1, \ldots, r$. If $s\colon \operatorname{Im} f \to E$ is the unique linear map defined by $s(f_i) = u_i$, for $i = 1, \ldots, r$, then s is injective, $f \circ s = \mathrm{id}$, and we have a direct sum*

$$E = \operatorname{Ker} f \oplus \operatorname{Im} s$$

as illustrated by the following diagram:

$$\operatorname{Ker} f \longrightarrow E = \operatorname{Ker} f \oplus \operatorname{Im} s \underset{s}{\overset{f}{\rightleftarrows}} \operatorname{Im} f \subseteq F.$$

See Figure 3.1. As a consequence,

$$\dim(E) = \dim(\operatorname{Ker} f) + \dim(\operatorname{Im} f) = \dim(\operatorname{Ker} f) + \operatorname{rk}(f).$$

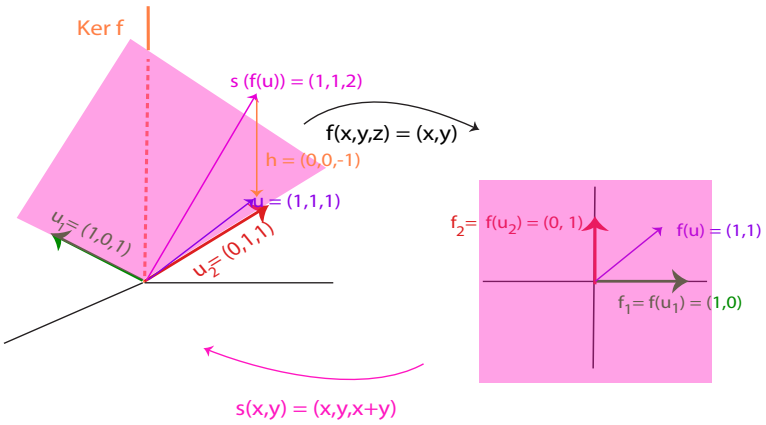


Figure 3.1: Let $f\colon E \to F$ be the linear map from $\mathbb{R}^3$ to $\mathbb{R}^2$ given by $f(x, y, z) = (x, y)$. Then $s\colon \mathbb{R}^2 \to \mathbb{R}^3$ is given by $s(x, y) = (x, y, x + y)$ and maps the pink $\mathbb{R}^2$ isomorphically onto the slanted pink plane of $\mathbb{R}^3$ whose equation is $-x - y + z = 0$. Theorem 3.6 shows that $\mathbb{R}^3$ is the direct sum of the plane $-x - y + z = 0$ and the kernel of f which the orange z-axis.

Proof. The vectors $u_1, \ldots, u_r$ must be linearly independent since otherwise we would have a nontrivial linear dependence

$$\lambda_1 u_1 + \cdots + \lambda_r u_r = 0,$$

and by applying f, we would get the nontrivial linear dependence

$$0 = \lambda_1 f(u_1) + \cdots + \lambda_r f(u_r) = \lambda_1 f_1 + \cdots + \lambda_r f_r,$$

contradicting the fact that $(f_1, \ldots, f_r)$ is a basis. Therefore, the unique linear map s given by $s(f_i) = u_i$, for $i = 1, \ldots, r$, is a linear isomorphism between $\operatorname{Im} f$ and its image, the subspace spanned by $(u_1, \ldots, u_r)$. It is also clear by definition that $f \circ s = \mathrm{id}$. For any $u \in E$, let

$$h = u - (s \circ f)(u).$$

Since $f \circ s = \mathrm{id}$, we have

$$f(h) = f(u - (s \circ f)(u)) = f(u) - (f \circ s \circ f)(u) = f(u) - (\mathrm{id} \circ f)(u) = f(u) - f(u) = 0,$$

which shows that $h \in \operatorname{Ker} f$. Since $h = u - (s \circ f)(u)$, it follows that

$$u = h + s(f(u)),$$

with $h \in \operatorname{Ker} f$ and $s(f(u)) \in \operatorname{Im} s$, which proves that

$$E = \operatorname{Ker} f + \operatorname{Im} s.$$

Now, if $u \in \operatorname{Ker} f \cap \operatorname{Im} s$, then $u = s(v)$ for some $v \in F$ and $f(u) = 0$ since $u \in \operatorname{Ker} f$. Since $u = s(v)$ and $f \circ s = \mathrm{id}$, we get

$$0 = f(u) = f(s(v)) = v,$$

and so $u = s(v) = s(0) = 0$. Thus, $\operatorname{Ker} f \cap \operatorname{Im} s = (0)$, which proves that we have a direct sum

$$E = \operatorname{Ker} f \oplus \operatorname{Im} s.$$

The equation

$$\dim(E) = \dim(\operatorname{Ker} f) + \dim(\operatorname{Im} f) = \dim(\operatorname{Ker} f) + \operatorname{rk}(f)$$

is an immediate consequence of the fact that the dimension is an additive property for direct sums, that by definition the rank of f is the dimension of the image of f, and that $\dim(\operatorname{Im} s) = \dim(\operatorname{Im} f)$, because s is an isomorphism between $\operatorname{Im} f$ and $\operatorname{Im} s$. □

Remark: The dimension $\dim(\operatorname{Ker} f)$ of the kernel of a linear map f is often called the *nullity* of f.

We now derive some important results using Theorem 3.6.

Proposition 3.7. *Given a vector space E, if U and V are any two subspaces of E, then*

$$\dim(U) + \dim(V) = \dim(U + V) + \dim(U \cap V),$$

an equation known as Grassmann's relation.

Proof. Recall that $U + V$ is the image of the linear map

$$a\colon U \times V \to E$$

given by

$$a(u, v) = u + v,$$

and that we proved earlier that the kernel $\operatorname{Ker} a$ of a is isomorphic to $U \cap V$. By Theorem 3.6,

$$\dim(U \times V) = \dim(\operatorname{Ker} a) + \dim(\operatorname{Im} a),$$

but $\dim(U \times V) = \dim(U) + \dim(V)$, $\dim(\operatorname{Ker} a) = \dim(U \cap V)$, and $\operatorname{Im} a = U + V$, so the Grassmann relation holds. □

The Grassmann relation can be very useful to figure out whether two subspace have a nontrivial intersection in spaces of dimension > 3. For example, it is easy to see that in $\mathbb{R}^5$, there are subspaces U and V with $\dim(U) = 3$ and $\dim(V) = 2$ such that $U \cap V = (0)$; for example, let U be generated by the vectors $(1, 0, 0, 0, 0), (0, 1, 0, 0, 0)$, $(0, 0, 1, 0, 0)$, and V be generated by the vectors $(0, 0, 0, 1, 0)$ and $(0, 0, 0, 0, 1)$. However, we claim that if $\dim(U) = 3$ and $\dim(V) = 3$, then $\dim(U \cap V) \geq 1$. Indeed, by the Grassmann relation, we have

$$\dim(U) + \dim(V) = \dim(U + V) + \dim(U \cap V),$$

namely

$$3 + 3 = 6 = \dim(U + V) + \dim(U \cap V),$$

and since $U + V$ is a subspace of $\mathbb{R}^5$, $\dim(U + V) \leq 5$, which implies

$$6 \leq 5 + \dim(U \cap V),$$

that is $1 \leq \dim(U \cap V)$.

As another consequence of Proposition 3.7, if U and V are two hyperplanes in a vector space of dimension n, so that $\dim(U) = n - 1$ and $\dim(V) = n - 1$, the reader should show that

$$\dim(U \cap V) \geq n - 2,$$

and so, if $U \neq V$, then

$$\dim(U \cap V) = n - 2.$$

Here is a characterization of direct sums that follows directly from Theorem 3.6.

Proposition 3.8. *If $U_1, \ldots, U_p$ are any subspaces of a finite dimensional vector space E, then*

$$\dim(U_1 + \cdots + U_p) \leq \dim(U_1) + \cdots + \dim(U_p),$$

and

$$\dim(U_1 + \cdots + U_p) = \dim(U_1) + \cdots + \dim(U_p)$$

iff the U_is form a direct sum $U_1 \oplus \cdots \oplus U_p$.

Proof. If we apply Theorem 3.6 to the linear map

$$a\colon U_1 \times \cdots \times U_p \to U_1 + \cdots + U_p$$

given by $a(u_1, \ldots, u_p) = u_1 + \cdots + u_p$, we get

$$\begin{aligned}\dim(U_1 + \cdots + U_p) &= \dim(U_1 \times \cdots \times U_p) - \dim(\operatorname{Ker} a)\\ &= \dim(U_1) + \cdots + \dim(U_p) - \dim(\operatorname{Ker} a),\end{aligned}$$

so the inequality follows. Since a is injective iff $\operatorname{Ker} a = (0)$, the U_is form a direct sum iff the second equation holds. □

Another important corollary of Theorem 3.6 is the following result:

Proposition 3.9. *Let E and F be two vector spaces with the same finite dimension $\dim(E) = \dim(F) = n$. For every linear map $f\colon E \to F$, the following properties are equivalent:*

(a) f is bijective.

(b) f is surjective.

(c) f is injective.

(d) $\operatorname{Ker} f = (0)$.

Proof. Obviously, (a) implies (b).

If f is surjective, then $\operatorname{Im} f = F$, and so $\dim(\operatorname{Im} f) = n$. By Theorem 3.6,

$$\dim(E) = \dim(\operatorname{Ker} f) + \dim(\operatorname{Im} f),$$

and since $\dim(E) = n$ and $\dim(\operatorname{Im} f) = n$, we get $\dim(\operatorname{Ker} f) = 0$, which means that $\operatorname{Ker} f = (0)$, and so f is injective (see Proposition 1.12). This proves that (b) implies (c).

If f is injective, then by Proposition 1.12, $\operatorname{Ker} f = (0)$, so (c) implies (d).

Finally, assume that $\operatorname{Ker} f = (0)$, so that $\dim(\operatorname{Ker} f) = 0$ and f is injective (by Proposition 1.12). By Theorem 3.6,

$$\dim(E) = \dim(\operatorname{Ker} f) + \dim(\operatorname{Im} f),$$

and since $\dim(\operatorname{Ker} f) = 0$, we get

$$\dim(\operatorname{Im} f) = \dim(E) = \dim(F),$$

which proves that f is also surjective, and thus bijective. This proves that (d) implies (a) and concludes the proof. □

One should be warned that Proposition 3.9 fails in infinite dimension.

Here are a few applications of Proposition 3.9. Let A be an $n \times n$ matrix and assume that A some right inverse B, which means that B is an $n \times n$ matrix such that

$$AB = I.$$

The linear map associated with A is surjective, since for every $u \in \mathbb{R}^n$, we have $A(Bu) = u$. By Proposition 3.9, this map is bijective so B is actually the inverse of A; in particular $BA = I$.

Similarly, assume that A has a left inverse B, so that

$$BA = I.$$

This time, the linear map associated with A is injective, because if $Au = 0$, then $BAu = B0 = 0$, and since $BA = I$ we get $u = 0$. Again, By Proposition 3.9, this map is bijective so B is actually the inverse of A; in particular $AB = I$.

Now, assume that the linear system $Ax = b$ has some solution for every b. Then the linear map associated with A is surjective and by Proposition 3.9, A is invertible.

Finally, assume that the linear system $Ax = b$ has at most one solution for every b. Then the linear map associated with A is injective and by Proposition 3.9, A is invertible.

We also have the following basic proposition about injective or surjective linear maps.

Proposition 3.10. *Let E and F be vector spaces, and let $f\colon E \to F$ be a linear map. If $f\colon E \to F$ is injective, then there is a surjective linear map $r\colon F \to E$ called a retraction, such that $r \circ f = \mathrm{id}_E$. If $f\colon E \to F$ is surjective, then there is an injective linear map $s\colon F \to E$ called a section, such that $f \circ s = \mathrm{id}_F$.*

Proof. Let $(u_i)_{i\in I}$ be a basis of E. Since $f\colon E \to F$ is an injective linear map, by Proposition 1.13, $(f(u_i))_{i\in I}$ is linearly independent in F. By Theorem 1.5, there is a basis $(v_j)_{j\in J}$ of F, where $I \subseteq J$, and where $v_i = f(u_i)$, for all $i \in I$. By Proposition 1.13, a linear map $r\colon F \to E$ can be defined such that $r(v_i) = u_i$, for all $i \in I$, and $r(v_j) = w$ for all $j \in (J - I)$, where w is any given vector in E, say $w = 0$. Since $r(f(u_i)) = u_i$ for all $i \in I$, by Proposition 1.13, we have $r \circ f = \mathrm{id}_E$.

Now, assume that $f\colon E \to F$ is surjective. Let $(v_j)_{j\in J}$ be a basis of F. Since $f\colon E \to F$ is surjective, for every $v_j \in F$, there is some $u_j \in E$ such that $f(u_j) = v_j$. Since $(v_j)_{j\in J}$ is a basis of F, by Proposition 1.13, there is a unique linear map $s\colon F \to E$ such that $s(v_j) = u_j$. Also, since $f(s(v_j)) = v_j$, by Proposition 1.13 (again), we must have $f \circ s = \mathrm{id}_F$. □

The converse of Proposition 3.10 is obvious.

The notion of rank of a linear map or of a matrix important, both theoretically and practically, since it is the key to the solvability of linear equations. We have the following simple proposition.

Proposition 3.11. *Given a linear map $f: E \to F$, the following properties hold:*

(i) $\mathrm{rk}(f) + \dim(\mathrm{Ker}\, f) = \dim(E)$.

(ii) $\mathrm{rk}(f) \leq \min(\dim(E), \dim(F))$.

Proof. Property (i) follows from Proposition 3.6. As for (ii), since $\mathrm{Im}\, f$ is a subspace of F, we have $\mathrm{rk}(f) \leq \dim(F)$, and since $\mathrm{rk}(f) + \dim(\mathrm{Ker}\, f) = \dim(E)$, we have $\mathrm{rk}(f) \leq \dim(E)$. □

The rank of a matrix is defined as follows.

Definition 3.3. Given a $m \times n$-matrix $A = (a_{i\,j})$, the *rank* $\mathrm{rk}(A)$ of the matrix A is the maximum number of linearly independent columns of A (viewed as vectors in $\mathbb{R}^m$).

In view of Proposition 1.6, the rank of a matrix A is the dimension of the subspace of $\mathbb{R}^m$ generated by the columns of A. Let E and F be two vector spaces, and let $(u_1, \ldots, u_n)$ be a basis of E, and $(v_1, \ldots, v_m)$ a basis of F. Let $f: E \to F$ be a linear map, and let $M(f)$ be its matrix w.r.t. the bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_m)$. Since the rank $\mathrm{rk}(f)$ of f is the dimension of $\mathrm{Im}\, f$, which is generated by $(f(u_1), \ldots, f(u_n))$, the rank of f is the maximum number of linearly independent vectors in $(f(u_1), \ldots, f(u_n))$, which is equal to the number of linearly independent columns of $M(f)$, since F and $\mathbb{R}^m$ are isomorphic. Thus, we have $\mathrm{rk}(f) = \mathrm{rk}(M(f))$, for every matrix representing f.

We will see later, using duality, that the rank of a matrix A is also equal to the maximal number of linearly independent rows of A.

3.4 Affine Maps

We showed in Section 1.7 that every linear map f must send the zero vector to the zero vector; that is,

$$f(0) = 0.$$

Yet, for any fixed nonzero vector $u \in E$ (where E is any vector space), the function t_u given by

$$t_u(x) = x + u, \quad \text{for all} \quad x \in E$$

shows up in pratice (for example, in robotics). Functions of this type are called *translations*. They are *not* linear for $u \neq 0$, since $t_u(0) = 0 + u = u$.

More generally, functions combining linear maps and translations occur naturally in many applications (robotics, computer vision, *etc.*), so it is necessary to understand some basic properties of these functions. For this, the notion of affine combination turns out to play a key role.

Recall from Section 1.7 that for any vector space E, given any family $(u_i)_{i\in I}$ of vectors $u_i \in E$, an *affine combination* of the family $(u_i)_{i\in I}$ is an expression of the form

$$\sum_{i\in I} \lambda_i u_i \quad \text{with} \quad \sum_{i\in I} \lambda_i = 1,$$

where $(\lambda_i)_{i\in I}$ is a family of scalars.

A linear combination places no restriction on the scalars involved, but an affine combination is a linear combination *with the restriction that the scalars λ_i must add up to* 1. Nevertheless, a linear combination can always be viewed as an affine combination using the following trick involving 0. For any family $(u_i)_{i\in I}$ of vectors in E and for *any* family of scalars $(\lambda_i)_{i\in I}$, we can write the linear combination $\sum_{i\in I} \lambda_i u_i$ as an affine combination as follows:

$$\sum_{i\in I} \lambda_i u_i = \sum_{i\in I} \lambda_i u_i + \Big(1 - \sum_{i\in I} \lambda_i\Big) 0.$$

Affine combinations are also called *barycentric combinations*.

Although this is not obvious at first glance, the condition that the scalars λ_i add up to 1 ensures that affine combinations are preserved under translations. To make this precise, consider functions $f\colon E \to F$, where E and F are two vector spaces, such that there is some *linear map* $h\colon E \to F$ and some fixed vector $b \in F$ (a *translation vector*), such that

$$f(x) = h(x) + b, \quad \text{for all} \quad x \in E.$$

The map f given by

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} 8/5 & -6/5 \\ 3/10 & 2/5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

is an example of the composition of a linear map with a translation.

We claim that functions of this type preserve affine combinations.

Proposition 3.12. *For any two vector spaces E and F, given any function $f\colon E \to F$ defined such that*

$$f(x) = h(x) + b, \quad \text{for all} \quad x \in E,$$

where $h\colon E \to F$ is a linear map and b is some fixed vector in F, for every affine combination $\sum_{i\in I} \lambda_i u_i$ (with $\sum_{i\in I} \lambda_i = 1$), we have

$$f\Big(\sum_{i\in I} \lambda_i u_i\Big) = \sum_{i\in I} \lambda_i f(u_i).$$

In other words, f preserves affine combinations.

Proof. By definition of f, using the fact that h is linear and the fact that $\sum_{i\in I}\lambda_i = 1$, we have

$$\begin{aligned} f\Big(\sum_{i\in}\lambda_i u_i\Big) &= h\Big(\sum_{i\in I}\lambda_i u_i\Big) + b \\ &= \sum_{i\in I}\lambda_i h(u_i) + 1b \\ &= \sum_{i\in I}\lambda_i h(u_i) + \Big(\sum_{i\in I}\lambda_i\Big)b \\ &= \sum_{i\in I}\lambda_i (h(u_i) + b) \\ &= \sum_{i\in I}\lambda_i f(u_i), \end{aligned}$$

as claimed. □

Observe how the fact that $\sum_{i\in I}\lambda_i = 1$ was used in a crucial way in line 3. Surprisingly, the converse of Proposition 3.12 also holds.

Proposition 3.13. *For any two vector spaces E and F, let $f\colon E \to F$ be any function that preserves affine combinations, i.e., for every affine combination $\sum_{i\in I}\lambda_i u_i$ (with $\sum_{i\in I}\lambda_i = 1$), we have*

$$f\Big(\sum_{i\in I}\lambda_i u_i\Big) = \sum_{i\in I}\lambda_i f(u_i).$$

Then, for any $a \in E$, the function $h\colon E \to F$ given by

$$h(x) = f(a + x) - f(a)$$

is a linear map independent of a, and

$$f(a + x) = h(x) + f(a), \quad \text{for all} \quad x \in E.$$

In particular, for $a = 0$, if we let $c = f(0)$, then

$$f(x) = h(x) + c, \quad \text{for all} \quad x \in E.$$

Proof. First, let us check that h is linear. Since f preserves affine combinations and since $a + u + v = (a + u) + (a + v) - a$ is an affine combination $(1 + 1 - 1 = 1)$, we have

$$\begin{aligned} h(u + v) &= f(a + u + v) - f(a) \\ &= f((a + u) + (a + v) - a) - f(a) \\ &= f(a + u) + f(a + v) - f(a) - f(a) \\ &= f(a + u) - f(a) + f(a + v) - f(a) \\ &= h(u) + h(v). \end{aligned}$$

This proves that

$$h(u+v) = h(u) + h(v), \quad u, v \in E.$$

Observe that $a + \lambda u = \lambda(a+u) + (1-\lambda)a$ is also an affine combination $(\lambda + 1 - \lambda = 1)$, so we have

$$\begin{aligned} h(\lambda u) &= f(a + \lambda u) - f(a) \\ &= f(\lambda(a+u) + (1-\lambda)a) - f(a) \\ &= \lambda f(a+u) + (1-\lambda)f(a) - f(a) \\ &= \lambda(f(a+u) - f(a)) \\ &= \lambda h(u). \end{aligned}$$

This proves that

$$h(\lambda u) = \lambda h(u), \quad u \in E,\ \lambda \in \mathbb{R}.$$

Therefore, h is indeed linear.

For any $b \in E$, since $b + u = (a+u) - a + b$ is an affine combination $(1 - 1 + 1 = 1)$, we have

$$\begin{aligned} f(b+u) - f(b) &= f((a+u) - a + b) - f(b) \\ &= f(a+u) - f(a) + f(b) - f(b) \\ &= f(a+u) - f(a), \end{aligned}$$

which proves that for all $a, b \in E$,

$$f(b+u) - f(b) = f(a+u) - f(a), \quad u \in E.$$

Therefore $h(x) = f(a+u) - f(a)$ does not depend on a, and it is obvious by the definition of h that

$$f(a+x) = h(x) + f(a), \quad \text{for all} \quad x \in E.$$

For $a = 0$, we obtain the last part of our proposition. □

We should think of a as a *chosen origin* in E. The function f maps the origin a in E to the origin $f(a)$ in F. Proposition 3.13 shows that the definition of h does not depend on the origin chosen in E. Also, since

$$f(x) = h(x) + c, \quad \text{for all} \quad x \in E$$

for some fixed vector $c \in F$, we see that f is the composition of the linear map h with the translation t_c (in F).

The unique linear map h as above is called the *linear map associated with* f and it is sometimes denoted by $\overrightarrow{f}$.

In view of Propositions 3.12 and 3.13, it is natural to make the following definition.

Definition 3.4. For any two vector spaces E and F, a function $f\colon E \to F$ is an *affine map* if f preserves affine combinations, *i.e.*, for every affine combination $\sum_{i\in I} \lambda_i u_i$ (with $\sum_{i\in I} \lambda_i = 1$), we have

$$f\Big(\sum_{i\in I} \lambda_i u_i\Big) = \sum_{i\in I} \lambda_i f(u_i).$$

Equivalently, a function $f\colon E \to F$ is an *affine map* if there is some linear map $h\colon E \to F$ (also denoted by $\overrightarrow{f}$) and some fixed vector $c \in F$ such that

$$f(x) = h(x) + c, \quad \text{for all} \quad x \in E.$$

Note that a linear map always maps the standard origin 0 in E to the standard origin 0 in F. However an affine map usually maps 0 to a nonzero vector $c = f(0)$. This is the "translation component" of the affine map.

When we deal with affine maps, it is often fruitful to think of the elements of E and F not only as vectors but also as *points*. In this point of view, *points can only be combined using affine combinations*, but vectors can be combined in an unrestricted fashion using linear combinations. We can also think of $u + v$ as the *result of translating the point u by the translation t_v*. These ideas lead to the definition of *affine spaces*.

The idea is that instead of a single space E, an affine space consists of two sets E and $\overrightarrow{E}$, where E is just an unstructured set of points, and $\overrightarrow{E}$ is a vector space. Furthermore, the vector space $\overrightarrow{E}$ acts on E. We can think of $\overrightarrow{E}$ as a set of *translations* specified by vectors, and given any point $a \in E$ and any vector (translation) $u \in \overrightarrow{E}$, the result of translating a by u is the point (not vector) $a + u$. Formally, we have the following definition.

Definition 3.5. An *affine space* is either the degenerate space reduced to the empty set, or a triple $\langle E, \overrightarrow{E}, + \rangle$ consisting of a nonempty set E (of *points*), a vector space $\overrightarrow{E}$ (of *translations*, or *free vectors*), and an action $+\colon E \times \overrightarrow{E} \to E$, satisfying the following conditions.

(A1) $a + 0 = a$, for every $a \in E$.

(A2) $(a + u) + v = a + (u + v)$, for every $a \in E$, and every $u, v \in \overrightarrow{E}$.

(A3) For any two points $a, b \in E$, there is a unique $u \in \overrightarrow{E}$ such that $a + u = b$.

The unique vector $u \in \overrightarrow{E}$ such that $a + u = b$ is denoted by $\overrightarrow{ab}$, or sometimes by $\mathbf{ab}$, or even by $b - a$. Thus, we also write

$$b = a + \overrightarrow{ab}$$

(or $b = a + \mathbf{ab}$, or even $b = a + (b - a)$).

It is important to note that *adding or rescaling points does not make sense*! However, using the fact that $\overrightarrow{E}$ acts on E is a special way (this action is transitive and faithful), it is possible to define rigorously the notion of *affine combinations* of points and to define affine spaces, affine maps, *etc.* However, this would lead us to far afield, and for our purposes it is enough to stick to vector spaces. Still, one should be aware that affine combinations really apply to points, and that points are not vectors!

If E and F are finite dimensional vector spaces with $\dim(E) = n$ and $\dim(F) = m$, then it is useful to represent an affine map with respect to bases in E in F. However, the translation part c of the affine map must be somewhow incorporated. There is a standard trick to do this which amounts to viewing an affine map as a linear map between spaces of dimension $n+1$ and $m+1$. We also have the extra flexibility of choosing origins $a \in E$ and $b \in F$.

Let $(u_1, \ldots, u_n)$ be a basis of E, $(v_1, \ldots, v_m)$ be a basis of F, and let $a \in E$ and $b \in F$ be any two fixed vectors viewed as *origins*. Our affine map f has the property that if $v = f(u)$, then

$$v - b = f(a + u - a) - b = f(a) - b + h(u - a).$$

So, if we let $y = v - b$, $x = u - a$, and $d = f(a) - b$, then

$$y = h(x) + d, \quad x \in E.$$

Over the basis $\mathcal{U} = (u_1, \ldots, u_n)$, we write

$$x = x_1 u_1 + \cdots + x_n u_n,$$

and over the basis $\mathcal{V} = (v_1, \ldots, v_m)$, we write

$$\begin{aligned} y &= y_1 v_1 + \cdots + y_m v_m, \\ d &= d_1 v_1 + \cdots + d_m v_m. \end{aligned}$$

Then, since

$$y = h(x) + d,$$

if we let A be the $m \times n$ matrix representing the linear map h, that is, the jth column of A consists of the coordinates of $h(u_j)$ over the basis $(v_1, \ldots, v_m)$, then we can write

$$y_{\mathcal{V}} = A x_{\mathcal{U}} + d_{\mathcal{V}}.$$

where $x_{\mathcal{U}} = (x_1, \ldots, x_n)^\top$, $y_{\mathcal{V}} = (y_1, \ldots, y_m)^\top$, and $d_{\mathcal{V}} = (d_1, \ldots, d_m)^\top$. The above is the matrix representation of our affine map f with respect to $(a, (u_1, \ldots, u_n))$ and $(b, (v_1, \ldots, v_m))$.

The reason for using the origins a and b is that it gives us more flexibility. In particular, we can choose $b = f(a)$, and then f behaves like a linear map with respect to the origins a and $b = f(a)$.

When $E = F$, if there is some $a \in E$ such that $f(a) = a$ (a is a *fixed point* of f), then we can pick $b = a$. Then, because $f(a) = a$, we get

$$v = f(u) = f(a + u - a) = f(a) + h(u - a) = a + h(u - a),$$

that is

$$v - a = h(u - a).$$

With respect to the new origin a, if we define x and y by

$$x = u - a$$
$$y = v - a,$$

then we get

$$y = h(x).$$

Therefore, f really behaves like a linear map, but *with respect to the new origin* a (not the *standard origin* 0). This is the case of a rotation around an axis that does not pass through the origin.

Remark: A pair $(a, (u_1, \ldots, u_n))$ where $(u_1, \ldots, u_n)$ is a basis of E and a is an origin chosen in E is called an *affine frame*.

We now describe the trick which allows us to incorporate the translation part d into the matrix A. We define the $(m + 1) \times (n + 1)$ matrix A' obtained by first adding d as the $(n + 1)$th column, and then $(\underbrace{0, \ldots, 0}_{n}, 1)$ as the $(m + 1)$th row:

$$A' = \begin{pmatrix} A & d \\ 0_n & 1 \end{pmatrix}.$$

Then, it is clear that

$$\begin{pmatrix} y \\ 1 \end{pmatrix} = \begin{pmatrix} A & d \\ 0_n & 1 \end{pmatrix} \begin{pmatrix} x \\ 1 \end{pmatrix}$$

iff

$$y = Ax + d.$$

This amounts to considering a point $x \in \mathbb{R}^n$ as a point $(x, 1)$ *in the (affine) hyperplane* H_{n+1} *in* $\mathbb{R}^{n+1}$ *of equation* $x_{n+1} = 1$. Then, an affine map is the restriction to the hyperplane H_{n+1} of the linear map $\widehat{f}$ from $\mathbb{R}^{n+1}$ to $\mathbb{R}^{m+1}$ corresponding to the matrix A' which maps H_{n+1} into H_{m+1} ($\widehat{f}(H_{n+1}) \subseteq H_{m+1}$). Figure 3.2 illustrates this process for $n = 2$.

For example, the map

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} 1 & 1 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 3 \\ 0 \end{pmatrix}$$

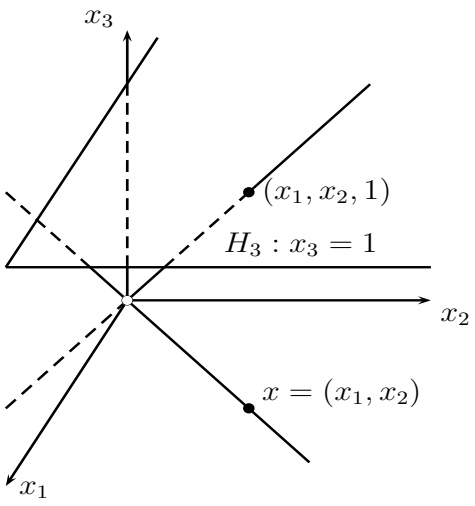


Figure 3.2: Viewing $\mathbb{R}^n$ as a hyperplane in $\mathbb{R}^{n+1}$ $(n = 2)$

defines an affine map f which is represented in $\mathbb{R}^3$ by

$$\begin{pmatrix} x_1 \\ x_2 \\ 1 \end{pmatrix} \mapsto \begin{pmatrix} 1 & 1 & 3 \\ 1 & 3 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ 1 \end{pmatrix}.$$

It is easy to check that the point $a = (6, -3)$ is fixed by f, which means that $f(a) = a$, so by translating the coordinate frame to the origin a, the affine map behaves like a linear map.

The idea of considering $\mathbb{R}^n$ as an hyperplane in $\mathbb{R}^{n+1}$ can be used to define *projective maps*.

3.5 Summary

The main concepts and results of this chapter are listed below:

- *Direct products, sums, direct sums*.
- *Projections*.
- The fundamental equation

$$\dim(E) = \dim(\operatorname{Ker} f) + \dim(\operatorname{Im} f) = \dim(\operatorname{Ker} f) + \operatorname{rk}(f)$$

  (Proposition 3.6).
- *Grassmann's relation*

$$\dim(U) + \dim(V) = \dim(U + V) + \dim(U \cap V).$$

- Characterizations of a bijective linear map $f\colon E \to F$.
- *Rank* of a matrix.
- *Affine Maps*.

Chapter 4

Determinants

4.1 Permutations, Signature of a Permutation

This chapter contains a review of determinants and their use in linear algebra. We begin with permutations and the signature of a permutation. Next, we define multilinear maps and alternating multilinear maps. Determinants are introduced as alternating multilinear maps taking the value 1 on the unit matrix (following Emil Artin). It is then shown how to compute a determinant using the Laplace expansion formula, and the connection with the usual definition is made. It is shown how determinants can be used to invert matrices and to solve (at least in theory!) systems of linear equations (the Cramer formulae). The determinant of a linear map is defined. We conclude by defining the characteristic polynomial of a matrix (and of a linear map) and by proving the celebrated Cayley–Hamilton theorem which states that every matrix is a "zero" of its characteristic polynomial (we give two proofs; one computational, the other one more conceptual).

Determinants can be defined in several ways. For example, determinants can be defined in a fancy way in terms of the exterior algebra (or alternating algebra) of a vector space. We will follow a more algorithmic approach due to Emil Artin. No matter which approach is followed, we need a few preliminaries about permutations on a finite set. We need to show that every permutation on n elements is a product of transpositions, and that the parity of the number of transpositions involved is an invariant of the permutation. Let $[n] = \{1, 2 \ldots, n\}$, where $n \in \mathbb{N}$, and $n > 0$.

Definition 4.1. A *permutation on n elements* is a bijection $\pi \colon [n] \to [n]$. When $n = 1$, the only function from [1] to [1] is the constant map: $1 \mapsto 1$. Thus, we will assume that $n \geq 2$. A *transposition* is a permutation $\tau \colon [n] \to [n]$ such that, for some $i < j$ (with $1 \leq i < j \leq n$), $\tau(i) = j$, $\tau(j) = i$, and $\tau(k) = k$, for all $k \in [n] - \{i, j\}$. In other words, a transposition exchanges two distinct elements $i, j \in [n]$.

If τ is a transposition, clearly, $\tau \circ \tau = \mathrm{id}$. We will also use the terminology product of permutations (or transpositions), as a synonym for composition of permutations. A permutation σ on n elements, say $\sigma(i) = k_i$ for $i = 1, \ldots, n$, can be represented in functional

notation by the $2 \times n$ array

$$\begin{pmatrix} 1 & \cdots & i & \cdots & n \\ k_1 & \cdots & k_i & \cdots & k_n \end{pmatrix}$$

known as *Cauchy two-line notation*. For example, we have the permutation σ denoted by

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 4 & 3 & 6 & 5 & 1 \end{pmatrix}.$$

A more concise notation often used in computer science and in combinatorics is to represent a permutation by its image, namely by the sequence

$$\sigma(1)\ \sigma(2)\ \cdots\ \sigma(n)$$

written as a row vector without commas separating the entries. The above is known as the *one-line notation*. For example, in the one-line notation, our previous permutation σ is represented by

$$2\ 4\ 3\ 6\ 5\ 1.$$

The reason for not enclosing the above sequence within parentheses is avoid confusion with the notation for cycles, for which is it customary to include parentheses.

Clearly, the composition of two permutations is a permutation and every permutation has an inverse which is also a permutation. Therefore, the set of permutations on $[n]$ is a *group* often denoted $\mathfrak{S}_n$ and called the *symmetric group* on n elements.

It is easy to show by induction that the group $\mathfrak{S}_n$ has $n!$ elements. The following proposition shows the importance of transpositions.

Proposition 4.1. *For every $n \geq 2$, every permutation $\pi\colon [n] \to [n]$ can be written as a nonempty composition of transpositions.*

Proof. We proceed by induction on n. If $n = 2$, there are exactly two permutations on $[2]$, the transposition τ exchanging 1 and 2, and the identity. However, $\mathrm{id}_2 = \tau^2$. Now, let $n \geq 3$. If $\pi(n) = n$, since by the induction hypothesis, the restriction of π to $[n-1]$ can be written as a product of transpositions, π itself can be written as a product of transpositions. If $\pi(n) = k \neq n$, letting τ be the transposition such that $\tau(n) = k$ and $\tau(k) = n$, it is clear that $\tau \circ \pi$ leaves n invariant, and by the induction hypothesis, we have $\tau \circ \pi = \tau_m \circ \ldots \circ \tau_1$ for some transpositions, and thus

$$\pi = \tau \circ \tau_m \circ \ldots \circ \tau_1,$$

a product of transpositions (since $\tau \circ \tau = \mathrm{id}_n$). □

Remark: When $\pi = \mathrm{id}_n$ is the identity permutation, we can agree that the composition of 0 transpositions is the identity. Proposition 4.1 shows that the transpositions generate the group of permutations $\mathfrak{S}_n$.

A transposition τ that exchanges two consecutive elements k and $k+1$ of $[n]$ $(1 \leq k \leq n-1)$ may be called a *basic* transposition. We leave it as a simple exercise to prove that every transposition can be written as a product of basic transpositions. In fact, the transposition that exchanges k and $k+p$ $(1 \leq p \leq n-k)$ can be realized using $2p-1$ basic transpositions. Therefore, the group of permutations $\mathfrak{S}_n$ is also generated by the basic transpositions.

Given a permutation written as a product of transpositions, we now show that the parity of the number of transpositions is an invariant. For this, we introduce the following function.

Definition 4.2. For every $n \geq 2$, let $\Delta\colon \mathbb{Z}^n \to \mathbb{Z}$ be the function given by

$$\Delta(x_1, \ldots, x_n) = \prod_{1 \leq i < j \leq n} (x_i - x_j).$$

More generally, for any permutation $\sigma \in \mathfrak{S}_n$, define $\Delta(x_{\sigma(1)}, \ldots, x_{\sigma(n)})$ by

$$\Delta(x_{\sigma(1)}, \ldots, x_{\sigma(n)}) = \prod_{1 \leq i < j \leq n} (x_{\sigma(i)} - x_{\sigma(j)}).$$

The expression $\Delta(x_1, \ldots, x_n)$ is often called the *discriminant* of $(x_1, \ldots, x_n)$.

It is clear that if the x_i are pairwise distinct, then $\Delta(x_1, \ldots, x_n) \neq 0$.

Proposition 4.2. *For every basic transposition τ of $[n]$ $(n \geq 2)$, we have*

$$\Delta(x_{\tau(1)}, \ldots, x_{\tau(n)}) = -\Delta(x_1, \ldots, x_n).$$

The above also holds for every transposition, and more generally, for every composition of transpositions $\sigma = \tau_p \circ \cdots \circ \tau_1$, we have

$$\Delta(x_{\sigma(1)}, \ldots, x_{\sigma(n)}) = (-1)^p \Delta(x_1, \ldots, x_n).$$

Consequently, for every permutation σ of $[n]$, the parity of the number p of transpositions involved in any decomposition of σ as $\sigma = \tau_p \circ \cdots \circ \tau_1$ is an invariant (only depends on σ).

Proof. Suppose τ exchanges x_k and x_{k+1}. The terms $x_i - x_j$ that are affected correspond to $i = k$, or $i = k+1$, or $j = k$, or $j = k+1$. The contribution of these terms in $\Delta(x_1, \ldots, x_n)$ is

$$\begin{aligned}(x_k - x_{k+1})[(x_k - x_{k+2}) \cdots (x_k - x_n)][(x_{k+1} - x_{k+2}) \cdots (x_{k+1} - x_n)] \\ [(x_1 - x_k) \cdots (x_{k-1} - x_k)][(x_1 - x_{k+1}) \cdots (x_{k-1} - x_{k+1})].\end{aligned}$$

When we exchange x_k and x_{k+1}, the first factor is multiplied by -1, the second and the third factor are exchanged, and the fourth and the fifth factor are exchanged, so the whole product $\Delta(x_1, \ldots, x_n)$ is is indeed multiplied by -1, that is,

$$\Delta(x_{\tau(1)}, \ldots, x_{\tau(n)}) = -\Delta(x_1, \ldots, x_n).$$

For the second statement, first we observe that since every transposition τ can be written as the composition of an odd number of basic transpositions (see the the remark following Proposition 4.1), we also have

$$\Delta(x_{\tau(1)}, \ldots, x_{\tau(n)}) = -\Delta(x_1, \ldots, x_n).$$

Next, we proceed by induction on the number p of transpositions involved in the decomposition of a permutation σ.

The base case $p = 1$ has just been proved. If $p \geq 2$, if we write $\omega = \tau_{p-1} \circ \cdots \circ \tau_1$, then $\sigma = \tau_p \circ \omega$ and

$$\begin{aligned}
\Delta(x_{\sigma(1)}, \ldots, x_{\sigma(n)}) &= \Delta(x_{\tau_p(\omega(1))}, \ldots, x_{\tau_p(\omega(n))}) \\
&= -\Delta(x_{\omega(1)}, \ldots, x_{\omega(n)}) \\
&= -(-1)^{p-1}\Delta(x_1, \ldots, x_n) \\
&= (-1)^p \Delta(x_1, \ldots, x_n),
\end{aligned}$$

where we used the induction hypothesis from the second to the third line, establishing the induction hypothesis. Since $\Delta(x_{\sigma(1)}, \ldots, x_{\sigma(n)})$ only depends on σ, the equation

$$\Delta(x_{\sigma(1)}, \ldots, x_{\sigma(n)}) = (-1)^p \Delta(x_1, \ldots, x_n).$$

shows that the parity $(-1)^p$ of the number of transpositions in any decomposition of σ is an invariant. □

In view of Proposition 4.2, the following definition makes sense:

Definition 4.3. For every permutation σ of $[n]$, the parity $\epsilon(\sigma)$ (or $\mathrm{sgn}(\sigma)$) of the number of transpositions involved in any decomposition of σ is called the *signature* (or *sign*) of σ.

Obviously $\epsilon(\tau) = -1$ for every transposition τ (since $(-1)^1 = -1$).

A simple way to compute the signature of a permutation is to count its number of inversions.

Definition 4.4. Given any permutation σ on n elements, we say that a pair (i, j) of indices $i, j \in \{1, \ldots, n\}$ such that $i < j$ and $\sigma(i) > \sigma(j)$ is an *inversion* of the permutation σ.

For example, the permutation σ given by

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 4 & 3 & 6 & 5 & 1 \end{pmatrix}$$

has seven inversions

$$(1,6),\ (2,3),\ (2,6),\ (3,6),\ (4,5),\ (4,6),\ (5,6).$$

Proposition 4.3. *The signature $\epsilon(\sigma)$ of any permutation σ is equal to the parity $(-1)^{I(\sigma)}$ of the number $I(\sigma)$ of inversions in σ.*

Proof. In the product

$$\Delta(x_{\sigma(1)}, \ldots, x_{\sigma(n)}) = \prod_{1 \leq i < j \leq n} (x_{\sigma(i)} - x_{\sigma(j)}),$$

the terms $x_{\sigma(i)} - x_{\sigma(j)}$ for which $\sigma(i) < \sigma(j)$ occur in $\Delta(x_1, \ldots, x_n)$, whereas the terms $x_{\sigma(i)} - x_{\sigma(j)}$ for which $\sigma(i) > \sigma(j)$ occur in $\Delta(x_1, \ldots, x_n)$ with a minus sign. Therefore, the number ν of terms in $\Delta(x_{\sigma(1)}, \ldots, x_{\sigma(n)})$ whose sign is the opposite of a term in $\Delta(x_1, \ldots, x_n)$, is equal to the number $I(\sigma)$ of inversions in σ, which implies that

$$\Delta(x_{\sigma(1)}, \ldots, x_{\sigma(n)}) = (-1)^{I(\sigma)} \Delta(x_1, \ldots, x_n).$$

By Proposition 4.2, the sign of $(-1)^{I(\sigma)}$ is equal to the signature of σ. □

For example, the permutation

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 4 & 3 & 6 & 5 & 1 \end{pmatrix}$$

has odd signature since it has seven inversions and $(-1)^7 = -1$.

Remark: When $\pi = \mathrm{id}_n$ is the identity permutation, since we agreed that the composition of 0 transpositions is the identity, it it still correct that $(-1)^0 = \epsilon(\mathrm{id}) = +1$. From proposition 4.2, it is immediate that $\epsilon(\pi' \circ \pi) = \epsilon(\pi')\epsilon(\pi)$. In particular, since $\pi^{-1} \circ \pi = \mathrm{id}_n$, we get $\epsilon(\pi^{-1}) = \epsilon(\pi)$.

We can now proceed with the definition of determinants.

4.2 Alternating Multilinear Maps

First, we define multilinear maps, symmetric multilinear maps, and alternating multilinear maps.

Remark: Most of the definitions and results presented in this section also hold when K is a commutative ring, and when we consider modules over K (free modules, when bases are needed).

Let $E_1, \ldots, E_n$, and F, be vector spaces over a field K, where $n \geq 1$.

Definition 4.5. A function $f\colon E_1 \times \ldots \times E_n \to F$ is a *multilinear map (or an n-linear map)* if it is linear in each argument, holding the others fixed. More explicitly, for every i,

$1 \leq i \leq n$, for all $x_1 \in E_1 \ldots$, $x_{i-1} \in E_{i-1}$, $x_{i+1} \in E_{i+1}, \ldots$, $x_n \in E_n$, for all $x, y \in E_i$, for all $\lambda \in K$,

$$\begin{aligned} f(x_1, \ldots, x_{i-1}, x + y, x_{i+1}, \ldots, x_n) &= f(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n) \\ &\quad + f(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_n), \\ f(x_1, \ldots, x_{i-1}, \lambda x, x_{i+1}, \ldots, x_n) &= \lambda f(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n). \end{aligned}$$

When $F = K$, we call f an *n-linear form (or multilinear form)*. If $n \geq 2$ and $E_1 = E_2 = \ldots = E_n$, an n-linear map $f \colon E \times \ldots \times E \to F$ is called *symmetric*, if $f(x_1, \ldots, x_n) = f(x_{\pi(1)}, \ldots, x_{\pi(n)})$, for every permutation π on $\{1, \ldots, n\}$. An n-linear map $f \colon E \times \ldots \times E \to F$ is called *alternating*, if $f(x_1, \ldots, x_n) = 0$ whenever $x_i = x_{i+1}$, for some i, $1 \leq i \leq n-1$ (in other words, when two adjacent arguments are equal). It does not harm to agree that when $n = 1$, a linear map is considered to be both symmetric and alternating, and we will do so.

When $n = 2$, a 2-linear map $f \colon E_1 \times E_2 \to F$ is called a *bilinear map*. We have already seen several examples of bilinear maps. Multiplication $\cdot \colon K \times K \to K$ is a bilinear map, treating K as a vector space over itself.

The operation $\langle -, - \rangle \colon E^* \times E \to K$ applying a linear form to a vector is a bilinear map.

Symmetric bilinear maps (and multilinear maps) play an important role in geometry (inner products, quadratic forms), and in differential calculus (partial derivatives).

A bilinear map is symmetric if $f(u, v) = f(v, u)$, for all $u, v \in E$.

Alternating multilinear maps satisfy the following simple but crucial properties.

Proposition 4.4. *Let $f \colon E \times \ldots \times E \to F$ be an n-linear alternating map, with $n \geq 2$. The following properties hold:*

(1)

$$f(\ldots, x_i, x_{i+1}, \ldots) = -f(\ldots, x_{i+1}, x_i, \ldots)$$

(2)

$$f(\ldots, x_i, \ldots, x_j, \ldots) = 0,$$

where $x_i = x_j$, and $1 \leq i < j \leq n$.

(3)

$$f(\ldots, x_i, \ldots, x_j, \ldots) = -f(\ldots, x_j, \ldots, x_i, \ldots),$$

where $1 \leq i < j \leq n$.

(4)

$$f(\ldots, x_i, \ldots) = f(\ldots, x_i + \lambda x_j, \ldots),$$

for any $\lambda \in K$, and where $i \neq j$.

Proof. (1) By multilinearity applied twice, we have

$$f(\ldots, x_i + x_{i+1}, x_i + x_{i+1}, \ldots) = f(\ldots, x_i, x_i, \ldots) + f(\ldots, x_i, x_{i+1}, \ldots) + f(\ldots, x_{i+1}, x_i, \ldots) + f(\ldots, x_{i+1}, x_{i+1}, \ldots),$$

and since f is alternating, this yields

$$0 = f(\ldots, x_i, x_{i+1}, \ldots) + f(\ldots, x_{i+1}, x_i, \ldots),$$

that is, $f(\ldots, x_i, x_{i+1}, \ldots) = -f(\ldots, x_{i+1}, x_i, \ldots)$.

(2) If $x_i = x_j$ and i and j are not adjacent, we can interchange x_i and x_{i+1}, and then x_i and x_{i+2}, etc, until x_i and x_j become adjacent. By (1),

$$f(\ldots, x_i, \ldots, x_j, \ldots) = \epsilon f(\ldots, x_i, x_j, \ldots),$$

where $\epsilon = +1$ or -1, but $f(\ldots, x_i, x_j, \ldots) = 0$, since $x_i = x_j$, and (2) holds.

(3) follows from (2) as in (1). (4) is an immediate consequence of (2). □

Proposition 4.4 will now be used to show a fundamental property of alternating multilinear maps. First, we need to extend the matrix notation a little bit. Let E be a vector space over K. Given an $n \times n$ matrix $A = (a_{ij})$ over K, we can define a map $L(A)\colon E^n \to E^n$ as follows:

$$\begin{aligned} L(A)_1(u) &= a_{11}u_1 + \cdots + a_{1n}u_n, \\ &\cdots \\ L(A)_n(u) &= a_{n1}u_1 + \cdots + a_{nn}u_n, \end{aligned}$$

for all $u_1, \ldots, u_n \in E$ and with $u = (u_1, \ldots, u_n)$. It is immediately verified that $L(A)$ is linear. Then, given two $n \times n$ matrice $A = (a_{ij})$ and $B = (b_{ij})$, by repeating the calculations establishing the product of matrices (just before Definition 1.10), we can show that

$$L(AB) = L(A) \circ L(B).$$

It is then convenient to use the matrix notation to describe the effect of the linear map $L(A)$, as

$$\begin{pmatrix} L(A)_1(u) \\ L(A)_2(u) \\ \vdots \\ L(A)_n(u) \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}.$$

Lemma 4.5. *Let $f\colon E \times \ldots \times E \to F$ be an n-linear alternating map. Let $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$ be two families of n vectors, such that,*

$$\begin{aligned} v_1 &= a_{11}u_1 + \cdots + a_{n1}u_n, \\ &\cdots \\ v_n &= a_{1n}u_1 + \cdots + a_{nn}u_n. \end{aligned}$$

Equivalently, letting

$$A = \begin{pmatrix} a_{1\,1} & a_{1\,2} & \dots & a_{1\,n} \\ a_{2\,1} & a_{2\,2} & \dots & a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n\,1} & a_{n\,2} & \dots & a_{n\,n} \end{pmatrix}$$

assume that we have

$$\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} = A^\top \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}.$$

Then,

$$f(v_1, \dots, v_n) = \Big(\sum_{\pi \in \mathfrak{S}_n} \epsilon(\pi) a_{\pi(1)\,1} \cdots a_{\pi(n)\,n} \Big) f(u_1, \dots, u_n),$$

where the sum ranges over all permutations π on $\{1, \dots, n\}$.

Proof. Expanding $f(v_1, \dots, v_n)$ by multilinearity, we get a sum of terms of the form

$$a_{\pi(1)\,1} \cdots a_{\pi(n)\,n} f(u_{\pi(1)}, \dots, u_{\pi(n)}),$$

for all possible functions $\pi \colon \{1, \dots, n\} \to \{1, \dots, n\}$. However, because f is alternating, only the terms for which π is a permutation are nonzero. By Proposition 4.1, every permutation π is a product of transpositions, and by Proposition 4.2, the parity $\epsilon(\pi)$ of the number of transpositions only depends on π. Then, applying Proposition 4.4 (3) to each transposition in π, we get

$$a_{\pi(1)\,1} \cdots a_{\pi(n)\,n} f(u_{\pi(1)}, \dots, u_{\pi(n)}) = \epsilon(\pi) a_{\pi(1)\,1} \cdots a_{\pi(n)\,n} f(u_1, \dots, u_n).$$

Thus, we get the expression of the lemma. □

The quantity

$$\det(A) = \sum_{\pi \in \mathfrak{S}_n} \epsilon(\pi) a_{\pi(1)\,1} \cdots a_{\pi(n)\,n}$$

is in fact the value of the determinant of A (which, as we shall see shortly, is also equal to the determinant of $A^\top$). However, working directly with the above definition is quite ackward, and we will proceed via a slightly indirect route

Remark: The reader might have been puzzled by the fact that it is the transpose matrix $A^\top$ rather than A itself that appears in Lemma 4.5. The reason is that if we want the generic term in the determinant to be

$$\epsilon(\pi) a_{\pi(1)\,1} \cdots a_{\pi(n)\,n},$$

where the permutation applies to the first index, then we have to express the v_js in terms of the u_is in terms of $A^\top$ as we did. Furthermore, since

$$v_j = a_{1\,j}u_1 + \cdots + a_{i\,j}u_i + \cdots + a_{n\,j}u_n,$$

we see that v_j corresponds to the jth column of the matrix A, and so the determinant is viewed as a function of the *columns* of A.

The literature is split on this point. Some authors prefer to define a determinant as we did. Others use A itself, which amounts to viewing det as a function of the rows, in which case we get the expression

$$\sum_{\sigma\in\mathfrak{S}_n} \epsilon(\sigma)a_{1\,\sigma(1)}\cdots a_{n\,\sigma(n)}.$$

Corollary 4.8 show that these two expressions are equal, so it doesn't matter which is chosen. This is a matter of taste.

4.3 Definition of a Determinant

Recall that the set of all square $n \times n$-matrices with coefficients in a field K is denoted by $\mathrm{M}_n(K)$.

Definition 4.6. A *determinant* is defined as any map

$$D\colon \mathrm{M}_n(K) \to K,$$

which, when viewed as a map on $(K^n)^n$, i.e., a map of the n columns of a matrix, is n-linear alternating and such that $D(I_n) = 1$ for the identity matrix I_n. Equivalently, we can consider a vector space E of dimension n, some fixed basis $(e_1, \ldots, e_n)$, and define

$$D\colon E^n \to K$$

as an n-linear alternating map such that $D(e_1, \ldots, e_n) = 1$.

First, we will show that such maps D exist, using an inductive definition that also gives a recursive method for computing determinants. Actually, we will define a family $(\mathcal{D}_n)_{n\geq 1}$ of (finite) sets of maps $D\colon \mathrm{M}_n(K) \to K$. Second, we will show that determinants are in fact uniquely defined, that is, we will show that each $\mathcal{D}_n$ consists of a single map. This will show the equivalence of the direct definition $\det(A)$ of Lemma 4.5 with the inductive definition $D(A)$. Finally, we will prove some basic properties of determinants, using the uniqueness theorem.

Given a matrix $A \in \mathrm{M}_n(K)$, we denote its n columns by $A^1, \ldots, A^n$. In order to describe the recursive process to define a determinant we need the notion of a minor.

Definition 4.7. Given any $n \times n$ matrix with $n \geq 2$, for any two indices i, j with $1 \leq i, j \leq n$, let A_{ij} be the $(n-1) \times (n-1)$ matrix obtained by deleting row i and columnn j from A and called a *minor*:

$$A_{ij} = \begin{pmatrix} & & & & \times & & \\ & & & & \times & & \\ \times & \times & \times & \times & \times & \times & \times \\ & & & & \times & & \\ & & & & \times & & \\ & & & & \times & & \\ & & & & \times & & \end{pmatrix}$$

For example, if

$$A = \begin{pmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{pmatrix}$$

then

$$A_{23} = \begin{pmatrix} 2 & -1 & 0 & 0 \\ 0 & -1 & -1 & 0 \\ 0 & 0 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix}.$$

Definition 4.8. For every $n \geq 1$, we define a finite set $\mathcal{D}_n$ of maps $D \colon \mathrm{M}_n(K) \to K$ inductively as follows:

When $n = 1$, $\mathcal{D}_1$ consists of the single map D such that, $D(A) = a$, where $A = (a)$, with $a \in K$.

Assume that $\mathcal{D}_{n-1}$ has been defined, where $n \geq 2$. Then, $\mathcal{D}_n$ consists of all the maps D such that, for some i, $1 \leq i \leq n$,

$$D(A) = (-1)^{i+1} a_{i\,1} D(A_{i\,1}) + \cdots + (-1)^{i+n} a_{i\,n} D(A_{i\,n}),$$

where for every j, $1 \leq j \leq n$, $D(A_{i\,j})$ is the result of applying any D in $\mathcal{D}_{n-1}$ to the minor $A_{i\,j}$.

We confess that the use of the same letter D for the member of $\mathcal{D}_n$ being defined, and for members of $\mathcal{D}_{n-1}$, may be slightly confusing. We considered using subscripts to distinguish, but this seems to complicate things unnecessarily. One should not worry too much anyway, since it will turn out that each $\mathcal{D}_n$ contains just one map.

Each $(-1)^{i+j} D(A_{i\,j})$ is called the *cofactor* of $a_{i\,j}$, and the inductive expression for $D(A)$ is called a *Laplace expansion of D according to the i-th row*. Given a matrix $A \in \mathrm{M}_n(K)$, each $D(A)$ is called a *determinant* of A.

We can think of each member of $\mathcal{D}_n$ as an *algorithm* to evaluate "the" determinant of A. The main point is that these algorithms, which recursively evaluate a determinant using all possible Laplace row expansions, all yield the same result, $\det(A)$.

We will prove shortly that $D(A)$ is uniquely defined (at the moment, it is not clear that $\mathcal{D}_n$ consists of a single map). Assuming this fact, given a $n \times n$-matrix $A = (a_{ij})$,

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}$$

its determinant is denoted by $D(A)$ or $\det(A)$, or more explicitly by

$$\det(A) = \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix}$$

First, let us first consider some examples.

Example 4.1.

1. When $n = 2$, if
$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$
expanding according to any row, we have
$$D(A) = ad - bc.$$

2. When $n = 3$, if
$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$
expanding according to the first row, we have
$$D(A) = a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$
that is,
$$D(A) = a_{11}(a_{22}a_{33} - a_{32}a_{23}) - a_{12}(a_{21}a_{33} - a_{31}a_{23}) + a_{13}(a_{21}a_{32} - a_{31}a_{22}),$$
which gives the explicit formula
$$D(A) = a_{11}a_{22}a_{33} + a_{21}a_{32}a_{13} + a_{31}a_{12}a_{23} - a_{11}a_{32}a_{23} - a_{21}a_{12}a_{33} - a_{31}a_{22}a_{13}.$$

We now show that each $D \in \mathcal{D}_n$ is a determinant (map).

Lemma 4.6. *For every $n \geq 1$, for every $D \in \mathcal{D}_n$ as defined in Definition 4.8, D is an alternating multilinear map such that $D(I_n) = 1$.*

Proof. By induction on n, it is obvious that $D(I_n) = 1$. Let us now prove that D is multilinear. Let us show that D is linear in each column. Consider any column k. Since

$$D(A) = (-1)^{i+1} a_{i\,1} D(A_{i\,1}) + \cdots + (-1)^{i+j} a_{i\,j} D(A_{i\,j}) + \cdots + (-1)^{i+n} a_{i\,n} D(A_{i\,n}),$$

if $j \neq k$, then by induction, $D(A_{i\,j})$ is linear in column k, and $a_{i\,j}$ does not belong to column k, so $(-1)^{i+j} a_{i\,j} D(A_{i\,j})$ is linear in column k. If $j = k$, then $D(A_{i\,j})$ does not depend on column $k = j$, since $A_{i\,j}$ is obtained from A by deleting row i and column $j = k$, and $a_{i\,j}$ belongs to column $j = k$. Thus, $(-1)^{i+j} a_{i\,j} D(A_{i\,j})$ is linear in column k. Consequently, in all cases, $(-1)^{i+j} a_{i\,j} D(A_{i\,j})$ is linear in column k, and thus, $D(A)$ is linear in column k.

Let us now prove that D is alternating. Assume that two adjacent columns of A are equal, say $A^k = A^{k+1}$. First, let $j \neq k$ and $j \neq k+1$. Then, the matrix $A_{i\,j}$ has two identical adjacent columns, and by the induction hypothesis, $D(A_{i\,j}) = 0$. The remaining terms of $D(A)$ are

$$(-1)^{i+k} a_{i\,k} D(A_{i\,k}) + (-1)^{i+k+1} a_{i\,k+1} D(A_{i\,k+1}).$$

However, the two matrices $A_{i\,k}$ and $A_{i\,k+1}$ are equal, since we are assuming that columns k and $k+1$ of A are identical, and since $A_{i\,k}$ is obtained from A by deleting row i and column k, and $A_{i\,k+1}$ is obtained from A by deleting row i and column $k+1$. Similarly, $a_{i\,k} = a_{i\,k+1}$, since columns k and $k+1$ of A are equal. But then,

$$(-1)^{i+k} a_{i\,k} D(A_{i\,k}) + (-1)^{i+k+1} a_{i\,k+1} D(A_{i\,k+1}) = (-1)^{i+k} a_{i\,k} D(A_{i\,k}) - (-1)^{i+k} a_{i\,k} D(A_{i\,k}) = 0.$$

This shows that D is alternating, and completes the proof. □

Lemma 4.6 shows the existence of determinants. We now prove their uniqueness.

Theorem 4.7. *For every $n \geq 1$, for every $D \in \mathcal{D}_n$, for every matrix $A \in \mathrm{M}_n(K)$, we have*

$$D(A) = \sum_{\pi \in \mathfrak{S}_n} \epsilon(\pi) a_{\pi(1)\,1} \cdots a_{\pi(n)\,n},$$

where the sum ranges over all permutations π on $\{1, \ldots, n\}$. As a consequence, $\mathcal{D}_n$ consists of a single map for every $n \geq 1$, and this map is given by the above explicit formula.

Proof. Consider the standard basis $(e_1, \ldots, e_n)$ of K^n, where $(e_i)_i = 1$ and $(e_i)_j = 0$, for $j \neq i$. Then, each column A^j of A corresponds to a vector v_j whose coordinates over the basis $(e_1, \ldots, e_n)$ are the components of A^j, that is, we can write

$$\begin{aligned} v_1 &= a_{1\,1} e_1 + \cdots + a_{n\,1} e_n, \\ &\cdots \\ v_n &= a_{1\,n} e_1 + \cdots + a_{n\,n} e_n. \end{aligned}$$

Since by Lemma 4.6, each D is a multilinear alternating map, by applying Lemma 4.5, we get

$$D(A) = D(v_1, \ldots, v_n) = \Big(\sum_{\pi \in \mathfrak{S}_n} \epsilon(\pi) a_{\pi(1)\,1} \cdots a_{\pi(n)\,n} \Big) D(e_1, \ldots, e_n),$$

where the sum ranges over all permutations π on $\{1, \ldots, n\}$. But $D(e_1, \ldots, e_n) = D(I_n)$, and by Lemma 4.6, we have $D(I_n) = 1$. Thus,

$$D(A) = \sum_{\pi \in \mathfrak{S}_n} \epsilon(\pi) a_{\pi(1)\,1} \cdots a_{\pi(n)\,n},$$

where the sum ranges over all permutations π on $\{1, \ldots, n\}$. □

From now on, we will favor the notation $\det(A)$ over $D(A)$ for the determinant of a square matrix.

Remark: There is a geometric interpretation of determinants which we find quite illuminating. Given n linearly independent vectors $(u_1, \ldots, u_n)$ in $\mathbb{R}^n$, the set

$$P_n = \{\lambda_1 u_1 + \cdots + \lambda_n u_n \mid 0 \leq \lambda_i \leq 1,\ 1 \leq i \leq n\}$$

is called a *parallelotope*. If $n = 2$, then P_2 is a *parallelogram* and if $n = 3$, then P_3 is a *parallelepiped*, a skew box having u_1, u_2, u_3 as three of its corner sides. Then, it turns out that $\det(u_1, \ldots, u_n)$ is the *signed volume* of the parallelotope P_n (where volume means n-dimensional volume). The sign of this volume accounts for the orientation of P_n in $\mathbb{R}^n$.

We can now prove some properties of determinants.

Corollary 4.8. *For every matrix $A \in \mathrm{M}_n(K)$, we have $\det(A) = \det(A^\top)$.*

Proof. By Theorem 4.7, we have

$$\det(A) = \sum_{\pi \in \mathfrak{S}_n} \epsilon(\pi) a_{\pi(1)\,1} \cdots a_{\pi(n)\,n},$$

where the sum ranges over all permutations π on $\{1, \ldots, n\}$. Since a permutation is invertible, every product

$$a_{\pi(1)\,1} \cdots a_{\pi(n)\,n}$$

can be rewritten as

$$a_{1\,\pi^{-1}(1)} \cdots a_{n\,\pi^{-1}(n)},$$

and since $\epsilon(\pi^{-1}) = \epsilon(\pi)$ and the sum is taken over all permutations on $\{1, \ldots, n\}$, we have

$$\sum_{\pi \in \mathfrak{S}_n} \epsilon(\pi) a_{\pi(1)\,1} \cdots a_{\pi(n)\,n} = \sum_{\sigma \in \mathfrak{S}_n} \epsilon(\sigma) a_{1\,\sigma(1)} \cdots a_{n\,\sigma(n)},$$

where π and σ range over all permutations. But it is immediately verified that

$$\det(A^\top) = \sum_{\sigma \in \mathfrak{S}_n} \epsilon(\sigma) a_{1\,\sigma(1)} \cdots a_{n\,\sigma(n)}.$$

□

A useful consequence of Corollary 4.8 is that the determinant of a matrix is also a multilinear alternating map of its rows. This fact, combined with the fact that the determinant of a matrix is a multilinear alternating map of its columns is often useful for finding short-cuts in computing determinants. We illustrate this point on the following example which shows up in polynomial interpolation.

Example 4.2. Consider the so-called *Vandermonde determinant*

$$V(x_1,\ldots,x_n) = \begin{vmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \ldots & x_n \\ x_1^2 & x_2^2 & \ldots & x_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{n-1} & x_2^{n-1} & \ldots & x_n^{n-1} \end{vmatrix}.$$

We claim that

$$V(x_1,\ldots,x_n) = \prod_{1\leq i<j\leq n}(x_j - x_i),$$

with $V(x_1,\ldots,x_n) = 1$, when $n = 1$. We prove it by induction on $n \geq 1$. The case $n = 1$ is obvious. Assume $n \geq 2$. We proceed as follows: multiply row $n-1$ by x_1 and substract it from row n (the last row), then multiply row $n-2$ by x_1 and subtract it from row $n-1$, etc, multiply row $i-1$ by x_1 and subtract it from row i, until we reach row 1. We obtain the following determinant:

$$V(x_1,\ldots,x_n) = \begin{vmatrix} 1 & 1 & \ldots & 1 \\ 0 & x_2 - x_1 & \ldots & x_n - x_1 \\ 0 & x_2(x_2 - x_1) & \ldots & x_n(x_n - x_1) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & x_2^{n-2}(x_2 - x_1) & \ldots & x_n^{n-2}(x_n - x_1) \end{vmatrix}$$

Now, expanding this determinant according to the first column and using multilinearity, we can factor $(x_i - x_1)$ from the column of index $i-1$ of the matrix obtained by deleting the first row and the first column, and thus

$$V(x_1,\ldots,x_n) = (x_2 - x_1)(x_3 - x_1)\cdots(x_n - x_1)V(x_2,\ldots,x_n),$$

which establishes the induction step.

Remark: Observe that

$$\Delta(x_1,\ldots,x_n) = V(x_n,\ldots,x_1),$$

where $\Delta(x_1,\ldots,x_n)$ is the discriminant of $(x_1,\ldots,x_n)$ introduced in Definition 4.2.

Lemma 4.5 can be reformulated nicely as follows.

Proposition 4.9. *Let $f\colon E \times \ldots \times E \to F$ be an n-linear alternating map. Let $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$ be two families of n vectors, such that*

$$\begin{aligned} v_1 &= a_{1\,1}u_1 + \cdots + a_{1\,n}u_n, \\ &\cdots \\ v_n &= a_{n\,1}u_1 + \cdots + a_{n\,n}u_n. \end{aligned}$$

Equivalently, letting

$$A = \begin{pmatrix} a_{1\,1} & a_{1\,2} & \ldots & a_{1\,n} \\ a_{2\,1} & a_{2\,2} & \ldots & a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n\,1} & a_{n\,2} & \ldots & a_{n\,n} \end{pmatrix}$$

assume that we have

$$\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} = A \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}.$$

Then,

$$f(v_1, \ldots, v_n) = \det(A) f(u_1, \ldots, u_n).$$

Proof. The only difference with Lemma 4.5 is that here, we are using $A^\top$ instead of A. Thus, by Lemma 4.5 and Corollary 4.8, we get the desired result. □

As a consequence, we get the very useful property that the determinant of a product of matrices is the product of the determinants of these matrices.

Proposition 4.10. *For any two $n \times n$-matrices A and B, we have* $\det(AB) = \det(A)\det(B)$.

Proof. We use Proposition 4.9 as follows: let $(e_1, \ldots, e_n)$ be the standard basis of K^n, and let

$$\begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} = AB \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}.$$

Then, we get

$$\det(w_1, \ldots, w_n) = \det(AB)\det(e_1, \ldots, e_n) = \det(AB),$$

since $\det(e_1, \ldots, e_n) = 1$. Now, letting

$$\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} = B \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix},$$

we get

$$\det(v_1, \ldots, v_n) = \det(B),$$

and since

$$\begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} = A \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix},$$

we get

$$\det(w_1, \ldots, w_n) = \det(A)\det(v_1, \ldots, v_n) = \det(A)\det(B). \qquad \square$$

It should be noted that all the results of this section, up to now, also hold when K is a commutative ring, and not necessarily a field. We can now characterize when an $n \times n$-matrix A is invertible in terms of its determinant $\det(A)$.

4.4 Inverse Matrices and Determinants

In the next two sections, K is a commutative ring, and when needed a field.

Definition 4.9. Let K be a commutative ring. Given a matrix $A \in \mathrm{M}_n(K)$, let $\widetilde{A} = (b_{ij})$ be the matrix defined such that

$$b_{ij} = (-1)^{i+j}\det(A_{ji}),$$

the cofactor of a_{ji}. The matrix $\widetilde{A}$ is called the *adjugate* of A, and each matrix A_{ji} is called a *minor* of the matrix A.

Note the reversal of the indices in

$$b_{ij} = (-1)^{i+j}\det(A_{ji}).$$

Thus, $\widetilde{A}$ is the transpose of the matrix of cofactors of elements of A.

We have the following proposition.

Proposition 4.11. *Let K be a commutative ring. For every matrix $A \in \mathrm{M}_n(K)$, we have*

$$A\widetilde{A} = \widetilde{A}A = \det(A)I_n.$$

As a consequence, A is invertible iff $\det(A)$ is invertible, and if so, $A^{-1} = (\det(A))^{-1}\widetilde{A}$.

Proof. If $\widetilde{A} = (b_{i\,j})$ and $A\widetilde{A} = (c_{i\,j})$, we know that the entry $c_{i\,j}$ in row i and column j of $A\widetilde{A}$ is

$$c_{i\,j} = a_{i\,1}b_{1\,j} + \cdots + a_{i\,k}b_{k\,j} + \cdots + a_{i\,n}b_{n\,j},$$

which is equal to

$$a_{i\,1}(-1)^{j+1}\det(A_{j\,1}) + \cdots + a_{i\,n}(-1)^{j+n}\det(A_{j\,n}).$$

If $j = i$, then we recognize the expression of the expansion of $\det(A)$ according to the i-th row:

$$c_{i\,i} = \det(A) = a_{i\,1}(-1)^{i+1}\det(A_{i\,1}) + \cdots + a_{i\,n}(-1)^{i+n}\det(A_{i\,n}).$$

If $j \neq i$, we can form the matrix A' by replacing the j-th row of A by the i-th row of A. Now, the matrix $A_{j\,k}$ obtained by deleting row j and column k from A is equal to the matrix $A'_{j\,k}$ obtained by deleting row j and column k from A', since A and A' only differ by the j-th row. Thus,

$$\det(A_{j\,k}) = \det(A'_{j\,k}),$$

and we have

$$c_{i\,j} = a_{i\,1}(-1)^{j+1}\det(A'_{j\,1}) + \cdots + a_{i\,n}(-1)^{j+n}\det(A'_{j\,n}).$$

However, this is the expansion of $\det(A')$ according to the j-th row, since the j-th row of A' is equal to the i-th row of A, and since A' has two identical rows i and j, because det is an alternating map of the rows (see an earlier remark), we have $\det(A') = 0$. Thus, we have shown that $c_{i\,i} = \det(A)$, and $c_{i\,j} = 0$, when $j \neq i$, and so

$$A\widetilde{A} = \det(A)I_n.$$

It is also obvious from the definition of $\widetilde{A}$, that

$$\widetilde{A}^{\top} = \widetilde{A^{\top}}.$$

Then, applying the first part of the argument to $A^{\top}$, we have

$$A^{\top}\widetilde{A^{\top}} = \det(A^{\top})I_n,$$

and since, $\det(A^{\top}) = \det(A)$, $\widetilde{A}^{\top} = \widetilde{A^{\top}}$, and $(\widetilde{A}A)^{\top} = A^{\top}\widetilde{A}^{\top}$, we get

$$\det(A)I_n = A^{\top}\widetilde{A^{\top}} = A^{\top}\widetilde{A}^{\top} = (\widetilde{A}A)^{\top},$$

that is,

$$(\widetilde{A}A)^{\top} = \det(A)I_n,$$

which yields

$$\widetilde{A}A = \det(A)I_n,$$

since $I_n^\top = I_n$. This proves that

$$A\widetilde{A} = \widetilde{A}A = \det(A)I_n.$$

As a consequence, if $\det(A)$ is invertible, we have $A^{-1} = (\det(A))^{-1}\widetilde{A}$. Conversely, if A is invertible, from $AA^{-1} = I_n$, by Proposition 4.10, we have $\det(A)\det(A^{-1}) = 1$, and $\det(A)$ is invertible. □

When K is a field, an element $a \in K$ is invertible iff $a \neq 0$. In this case, the second part of the proposition can be stated as A is invertible iff $\det(A) \neq 0$. Note in passing that this method of computing the inverse of a matrix is usually not practical.

We now consider some applications of determinants to linear independence and to solving systems of linear equations. Although these results hold for matrices over certain rings, their proofs require more sophisticated methods. Therefore, we assume again that K is a field (usually, $K = \mathbb{R}$ or $K = \mathbb{C}$).

Let A be an $n \times n$-matrix, x a column vectors of variables, and b another column vector, and let $A^1, \ldots, A^n$ denote the columns of A. Observe that the system of equation $Ax = b$,

$$\begin{pmatrix} a_{1\,1} & a_{1\,2} & \ldots & a_{1\,n} \\ a_{2\,1} & a_{2\,2} & \ldots & a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n\,1} & a_{n\,2} & \ldots & a_{n\,n} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

is equivalent to

$$x_1 A^1 + \cdots + x_j A^j + \cdots + x_n A^n = b,$$

since the equation corresponding to the i-th row is in both cases

$$a_{i\,1}x_1 + \cdots + a_{i\,j}x_j + \cdots + a_{i\,n}x_n = b_i.$$

First, we characterize linear independence of the column vectors of a matrix A in terms of its determinant.

Proposition 4.12. *Given an $n \times n$-matrix A over a field K, the columns $A^1, \ldots, A^n$ of A are linearly dependent iff $\det(A) = \det(A^1, \ldots, A^n) = 0$. Equivalently, A has rank n iff $\det(A) \neq 0$.*

Proof. First, assume that the columns $A^1, \ldots, A^n$ of A are linearly dependent. Then, there are $x_1, \ldots, x_n \in K$, such that

$$x_1 A^1 + \cdots + x_j A^j + \cdots + x_n A^n = 0,$$

where $x_j \neq 0$ for some j. If we compute

$$\det(A^1, \ldots, x_1 A^1 + \cdots + x_j A^j + \cdots + x_n A^n, \ldots, A^n) = \det(A^1, \ldots, 0, \ldots, A^n) = 0,$$

where 0 occurs in the j-th position, by multilinearity, all terms containing two identical columns A^k for $k \neq j$ vanish, and we get

$$x_j \det(A^1, \ldots, A^n) = 0.$$

Since $x_j \neq 0$ and K is a field, we must have $\det(A^1, \ldots, A^n) = 0$.

Conversely, we show that if the columns $A^1, \ldots, A^n$ of A are linearly independent, then $\det(A^1, \ldots, A^n) \neq 0$. If the columns $A^1, \ldots, A^n$ of A are linearly independent, then they form a basis of K^n, and we can express the standard basis $(e_1, \ldots, e_n)$ of K^n in terms of $A^1, \ldots, A^n$. Thus, we have

$$\begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = \begin{pmatrix} b_{1\,1} & b_{1\,2} & \ldots & b_{1\,n} \\ b_{2\,1} & b_{2\,2} & \ldots & b_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n\,1} & b_{n\,2} & \ldots & b_{n\,n} \end{pmatrix} \begin{pmatrix} A^1 \\ A^2 \\ \vdots \\ A^n \end{pmatrix},$$

for some matrix $B = (b_{i\,j})$, and by Proposition 4.9, we get

$$\det(e_1, \ldots, e_n) = \det(B) \det(A^1, \ldots, A^n),$$

and since $\det(e_1, \ldots, e_n) = 1$, this implies that $\det(A^1, \ldots, A^n) \neq 0$ (and $\det(B) \neq 0$). For the second assertion, recall that the rank of a matrix is equal to the maximum number of linearly independent columns, and the conclusion is clear. □

If we combine Proposition 4.12 with Proposition 8.12, we obtain the following criterion for finding the rank of a matrix.

Proposition 4.13. *Given any $m \times n$ matrix A over a field K (typically $K = \mathbb{R}$ or $K = \mathbb{C}$), the rank of A is the maximum natural number r such that there is an $r \times r$ submatrix B of A obtained by selecting r rows and r columns of A, and such that $\det(B) \neq 0$.*

4.5 Systems of Linear Equations and Determinants

We now characterize when a system of linear equations of the form $Ax = b$ has a unique solution.

Proposition 4.14. *Given an $n \times n$-matrix A over a field K, the following properties hold:*

(1) For every column vector b, there is a unique column vector x such that $Ax = b$ iff the only solution to $Ax = 0$ is the trivial vector $x = 0$, iff $\det(A) \neq 0$.

(2) If $\det(A) \neq 0$, the unique solution of $Ax = b$ is given by the expressions

$$x_j = \frac{\det(A^1, \ldots, A^{j-1}, b, A^{j+1}, \ldots, A^n)}{\det(A^1, \ldots, A^{j-1}, A^j, A^{j+1}, \ldots, A^n)},$$

known as Cramer's rules.

(3) The system of linear equations $Ax = 0$ has a nonzero solution iff $\det(A) = 0$.

Proof. Assume that $Ax = b$ has a single solution x_0, and assume that $Ay = 0$ with $y \neq 0$. Then,

$$A(x_0 + y) = Ax_0 + Ay = Ax_0 + 0 = b,$$

and $x_0 + y \neq x_0$ is another solution of $Ax = b$, contadicting the hypothesis that $Ax = b$ has a single solution x_0. Thus, $Ax = 0$ only has the trivial solution. Now, assume that $Ax = 0$ only has the trivial solution. This means that the columns $A^1, \ldots, A^n$ of A are linearly independent, and by Proposition 4.12, we have $\det(A) \neq 0$. Finally, if $\det(A) \neq 0$, by Proposition 4.11, this means that A is invertible, and then, for every b, $Ax = b$ is equivalent to $x = A^{-1}b$, which shows that $Ax = b$ has a single solution.

(2) Assume that $Ax = b$. If we compute

$$\det(A^1, \ldots, x_1A^1 + \cdots + x_jA^j + \cdots + x_nA^n, \ldots, A^n) = \det(A^1, \ldots, b, \ldots, A^n),$$

where b occurs in the j-th position, by multilinearity, all terms containing two identical columns A^k for $k \neq j$ vanish, and we get

$$x_j \det(A^1, \ldots, A^n) = \det(A^1, \ldots, A^{j-1}, b, A^{j+1}, \ldots, A^n),$$

for every j, $1 \leq j \leq n$. Since we assumed that $\det(A) = \det(A^1, \ldots, A^n) \neq 0$, we get the desired expression.

(3) Note that $Ax = 0$ has a nonzero solution iff $A^1, \ldots, A^n$ are linearly dependent (as observed in the proof of Proposition 4.12), which, by Proposition 4.12, is equivalent to $\det(A) = 0$. □

As pleasing as Cramer's rules are, it is usually impractical to solve systems of linear equations using the above expressions. However, these formula imply an interesting fact, which is that the solution of the system $Ax = b$ are continuous in A and b. If we assume that the entries in A are continuous functions $a_{ij}(t)$ and the entries in b are are also continuous functions $b_j(t)$ of a real parameter t, since determinants are polynomial functions of their entries, the expressions

$$x_j(t) = \frac{\det(A^1, \ldots, A^{j-1}, b, A^{j+1}, \ldots, A^n)}{\det(A^1, \ldots, A^{j-1}, A^j, A^{j+1}, \ldots, A^n)}$$

are ratios of polynomials, and thus are also continuous as long as $\det(A(t))$ is nonzero. Similarly, if the functions $a_{ij}(t)$ and $b_j(t)$ are differentiable, so are the $x_j(t)$.

4.6 Determinant of a Linear Map

We close this chapter with the notion of determinant of a linear map $f : E \to E$.

Given a vector space E of finite dimension n, given a basis $(u_1, \ldots, u_n)$ of E, for every linear map $f\colon E \to E$, if $M(f)$ is the matrix of f w.r.t. the basis $(u_1, \ldots, u_n)$, we can define $\det(f) = \det(M(f))$. If $(v_1, \ldots, v_n)$ is any other basis of E, and if P is the change of basis matrix, by Corollary 2.5, the matrix of f with respect to the basis $(v_1, \ldots, v_n)$ is $P^{-1}M(f)P$. Now, by proposition 4.10, we have

$$\det(P^{-1}M(f)P) = \det(P^{-1})\det(M(f))\det(P) = \det(P^{-1})\det(P)\det(M(f)) = \det(M(f)).$$

Thus, $\det(f)$ is indeed independent of the basis of E.

Definition 4.10. Given a vector space E of finite dimension, for any linear map $f\colon E \to E$, we define the *determinant* $\det(f)$ *of* f as the determinant $\det(M(f))$ of the matrix of f in any basis (since, from the discussion just before this definition, this determinant does not depend on the basis).

Then, we have the following proposition.

Proposition 4.15. *Given any vector space E of finite dimension n, a linear map $f\colon E \to E$ is invertible iff* $\det(f) \neq 0$.

Proof. The linear map $f\colon E \to E$ is invertible iff its matrix $M(f)$ in any basis is invertible (by Proposition 2.2), iff $\det(M(f)) \neq 0$, by Proposition 4.11. □

Given a vector space of finite dimension n, it is easily seen that the set of bijective linear maps $f\colon E \to E$ such that $\det(f) = 1$ is a group under composition. This group is a subgroup of the general linear group $\mathbf{GL}(E)$. It is called the *special linear group (of E)*, and it is denoted by $\mathbf{SL}(E)$, or when $E = K^n$, by $\mathbf{SL}(n, K)$, or even by $\mathbf{SL}(n)$.

4.7 The Cayley–Hamilton Theorem

We conclude this chapter with an interesting and important application of Proposition 4.11, the *Cayley–Hamilton theorem*. The results of this section apply to matrices over any commutative ring K. First, we need the concept of the characteristic polynomial of a matrix.

Definition 4.11. If K is any commutative ring, for every $n \times n$ matrix $A \in \mathrm{M}_n(K)$, the *characteristic polynomial* $P_A(X)$ of A is the determinant

$$P_A(X) = \det(XI - A).$$

The characteristic polynomial $P_A(X)$ is a polynomial in $K[X]$, the ring of polynomials in the indeterminate X with coefficients in the ring K. For example, when $n = 2$, if

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

then

$$P_A(X) = \begin{vmatrix} X-a & -b \\ -c & X-d \end{vmatrix} = X^2 - (a+d)X + ad - bc.$$

We can substitute the matrix A for the variable X in the polynomial $P_A(X)$, obtaining a *matrix* P_A. If we write

$$P_A(X) = X^n + c_1 X^{n-1} + \cdots + c_n,$$

then

$$P_A = A^n + c_1 A^{n-1} + \cdots + c_n I.$$

We have the following remarkable theorem.

Theorem 4.16. *(Cayley–Hamilton) If K is any commutative ring, for every $n \times n$ matrix $A \in \mathrm{M}_n(K)$, if we let*

$$P_A(X) = X^n + c_1 X^{n-1} + \cdots + c_n$$

be the characteristic polynomial of A, then

$$P_A = A^n + c_1 A^{n-1} + \cdots + c_n I = 0.$$

Proof. We can view the matrix $B = XI - A$ as a matrix with coefficients in the polynomial ring $K[X]$, and then we can form the matrix $\widetilde{B}$ which is the transpose of the matrix of cofactors of elements of B. Each entry in $\widetilde{B}$ is an $(n-1) \times (n-1)$ determinant, and thus a polynomial of degree a most $n-1$, so we can write $\widetilde{B}$ as

$$\widetilde{B} = X^{n-1} B_0 + X^{n-2} B_1 + \cdots + B_{n-1},$$

for some matrices $B_0, \ldots, B_{n-1}$ with coefficients in K. For example, when $n = 2$, we have

$$B = \begin{pmatrix} X-a & -b \\ -c & X-d \end{pmatrix}, \quad \widetilde{B} = \begin{pmatrix} X-d & b \\ c & X-a \end{pmatrix} = X \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} -d & b \\ c & -a \end{pmatrix}.$$

By Proposition 4.11, we have

$$B\widetilde{B} = \det(B) I = P_A(X) I.$$

On the other hand, we have

$$B\widetilde{B} = (XI - A)(X^{n-1} B_0 + X^{n-2} B_1 + \cdots + X^{n-j-1} B_j + \cdots + B_{n-1}),$$

and by multiplying out the right-hand side, we get

$$B\widetilde{B} = X^n D_0 + X^{n-1} D_1 + \cdots + X^{n-j} D_j + \cdots + D_n,$$

with

$$\begin{aligned} D_0 &= B_0 \\ D_1 &= B_1 - AB_0 \\ &\vdots \\ D_j &= B_j - AB_{j-1} \\ &\vdots \\ D_{n-1} &= B_{n-1} - AB_{n-2} \\ D_n &= -AB_{n-1}. \end{aligned}$$

Since

$$P_A(X)I = (X^n + c_1X^{n-1} + \cdots + c_n)I,$$

the equality

$$X^nD_0 + X^{n-1}D_1 + \cdots + D_n = (X^n + c_1X^{n-1} + \cdots + c_n)I$$

is an equality between two matrices, so it requires that all corresponding entries are equal, and since these are polynomials, the coefficients of these polynomials must be identical, which is equivalent to the set of equations

$$\begin{aligned} I &= B_0 \\ c_1I &= B_1 - AB_0 \\ &\vdots \\ c_jI &= B_j - AB_{j-1} \\ &\vdots \\ c_{n-1}I &= B_{n-1} - AB_{n-2} \\ c_nI &= -AB_{n-1}, \end{aligned}$$

for all j, with $1 \leq j \leq n-1$. If we multiply the first equation by A^n, the last by I, and generally the $(j+1)$th by A^{n-j}, when we add up all these new equations, we see that the right-hand side adds up to 0, and we get our desired equation

$$A^n + c_1A^{n-1} + \cdots + c_nI = 0,$$

as claimed. □

As a concrete example, when $n = 2$, the matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

satisfies the equation

$$A^2 - (a+d)A + (ad - bc)I = 0.$$

Most readers will probably find the proof of Theorem 4.16 rather clever but very mysterious and unmotivated. The conceptual difficulty is that we really need to understand how polynomials in one variable "act" on vectors, in terms of the matrix A. This can be done and yields a more "natural" proof. Actually, the reasoning is simpler and more general if we free ourselves from matrices and instead consider a finite-dimensional vector space E and some given linear map $f\colon E \to E$. Given any polynomial $p(X) = a_0X^n + a_1X^{n-1} + \cdots + a_n$ with coefficients in the field K, we define the *linear map* $p(f)\colon E \to E$ by

$$p(f) = a_0f^n + a_1f^{n-1} + \cdots + a_n\mathrm{id},$$

where $f^k = f \circ \cdots \circ f$, the k-fold composition of f with itself. Note that

$$p(f)(u) = a_0f^n(u) + a_1f^{n-1}(u) + \cdots + a_nu,$$

for every vector $u \in E$. Then, we define a new kind of scalar multiplication $\cdot\colon K[X] \times E \to E$ by polynomials as follows: for every polynomial $p(X) \in K[X]$, for every $u \in E$,

$$p(X) \cdot u = p(f)(u).$$

It is easy to verify that this is a "good action," which means that

$$\begin{aligned} p \cdot (u + v) &= p \cdot u + p \cdot v \\ (p + q) \cdot u &= p \cdot u + q \cdot u \\ (pq) \cdot u &= p \cdot (q \cdot u) \\ 1 \cdot u &= u, \end{aligned}$$

for all $p, q \in K[X]$ and all $u, v \in E$. With this new scalar multiplication, E is a $K[X]$-module.

If $p = \lambda$ is just a scalar in K (a polynomial of degree 0), then

$$\lambda \cdot u = (\lambda\mathrm{id})(u) = \lambda u,$$

which means that K acts on E by scalar multiplication as before. If $p(X) = X$ (the monomial X), then

$$X \cdot u = f(u).$$

Now, if we pick a basis $(e_1, \ldots, e_n)$, if a polynomial $p(X) \in K[X]$ has the property that

$$p(X) \cdot e_i = 0, \quad i = 1, \ldots, n,$$

then this means that $p(f)(e_i) = 0$ for $i = 1, \ldots, n$, which means that the linear map $p(f)$ vanishes on E. We can also check, as we did in Section 4.2, that if A and B are two $n \times n$ matrices and if $(u_1, \ldots, u_n)$ are any n vectors, then

$$A \cdot \left(B \cdot \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} \right) = (AB) \cdot \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}.$$

This suggests the plan of attack for our second proof of the Cayley–Hamilton theorem. For simplicity, we prove the theorem for vector spaces over a field. The proof goes through for a free module over a commutative ring.

Theorem 4.17. *(Cayley–Hamilton) For every finite-dimensional vector space over a field K, for every linear map $f\colon E \to E$, for every basis $(e_1, \ldots, e_n)$, if A is the matrix over f over the basis $(e_1, \ldots, e_n)$ and if*

$$P_A(X) = X^n + c_1 X^{n-1} + \cdots + c_n$$

is the characteristic polynomial of A, then

$$P_A(f) = f^n + c_1 f^{n-1} + \cdots + c_n \mathrm{id} = 0.$$

Proof. Since the columns of A consist of the vector $f(e_j)$ expressed over the basis $(e_1, \ldots, e_n)$, we have

$$f(e_j) = \sum_{i=1}^{n} a_{i\,j} e_i, \quad 1 \leq j \leq n.$$

Using our action of $K[X]$ on E, the above equations can be expressed as

$$X \cdot e_j = \sum_{i=1}^{n} a_{i\,j} \cdot e_i, \quad 1 \leq j \leq n,$$

which yields

$$\sum_{i=1}^{j-1} -a_{i\,j} \cdot e_i + (X - a_{j\,j}) \cdot e_j + \sum_{i=j+1}^{n} -a_{i\,j} \cdot e_i = 0, \qquad 1 \leq j \leq n.$$

Observe that the transpose of the characteristic polynomial shows up, so the above system can be written as

$$\begin{pmatrix} X - a_{1\,1} & -a_{2\,1} & \cdots & -a_{n\,1} \\ -a_{1\,2} & X - a_{2\,2} & \cdots & -a_{n\,2} \\ \vdots & \vdots & \vdots & \vdots \\ -a_{1\,n} & -a_{2\,n} & \cdots & X - a_{n\,n} \end{pmatrix} \cdot \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

If we let $B = XI - A^\top$, then as in the previous proof, if $\widetilde{B}$ is the transpose of the matrix of cofactors of B, we have

$$\widetilde{B}B = \det(B)I = \det(XI - A^\top)I = \det(XI - A)I = P_A I.$$

But then, since

$$B \cdot \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

and since $\widetilde{B}$ is matrix whose entries are polynomials in $K[X]$, it makes sense to multiply on the left by $\widetilde{B}$ and we get

$$\widetilde{B}\cdot B\cdot\begin{pmatrix}e_1\\ e_2\\ \vdots\\ e_n\end{pmatrix}=(\widetilde{B}B)\cdot\begin{pmatrix}e_1\\ e_2\\ \vdots\\ e_n\end{pmatrix}=P_AI\cdot\begin{pmatrix}e_1\\ e_2\\ \vdots\\ e_n\end{pmatrix}=\widetilde{B}\cdot\begin{pmatrix}0\\ 0\\ \vdots\\ 0\end{pmatrix}=\begin{pmatrix}0\\ 0\\ \vdots\\ 0\end{pmatrix};$$

that is,

$$P_A\cdot e_j=0,\quad j=1,\ldots,n,$$

which proves that $P_A(f)=0$, as claimed. □

If K is a field, then the characteristic polynomial of a linear map $f\colon E\to E$ is independent of the basis $(e_1,\ldots,e_n)$ chosen in E. To prove this, observe that the matrix of f over another basis will be of the form $P^{-1}AP$, for some inverible matrix P, and then

$$\begin{aligned}\det(XI-P^{-1}AP)&=\det(XP^{-1}IP-P^{-1}AP)\\ &=\det(P^{-1}(XI-A)P)\\ &=\det(P^{-1})\det(XI-A)\det(P)\\ &=\det(XI-A).\end{aligned}$$

Therefore, the characteristic polynomial of a linear map is intrinsic to f, and it is denoted by P_f.

The zeros (roots) of the characteristic polynomial of a linear map f are called the *eigenvalues* of f. They play an important role in theory and applications. We will come back to this topic later on.

4.8 Permanents

Recall that the explicit formula for the determinant of an $n\times n$ matrix is

$$\det(A)=\sum_{\pi\in\mathfrak{S}_n}\epsilon(\pi)a_{\pi(1)\,1}\cdots a_{\pi(n)\,n}.$$

If we drop the sign $\epsilon(\pi)$ of every permutation from the above formula, we obtain a quantity known as the *permanent*:

$$\mathrm{per}(A)=\sum_{\pi\in\mathfrak{S}_n}a_{\pi(1)\,1}\cdots a_{\pi(n)\,n}.$$

Permanents and determinants were investigated as early as 1812 by Cauchy. It is clear from the above definition that the permanent is a multilinear and symmetric form. We also have

$$\mathrm{per}(A)=\mathrm{per}(A^\top),$$

and the following unsigned version of the Laplace expansion formula:

$$\mathrm{per}(A) = a_{i\,1}\mathrm{per}(A_{i\,1}) + \cdots + a_{i\,j}\mathrm{per}(A_{i\,j}) + \cdots + a_{i\,n}\mathrm{per}(A_{i\,n}),$$

for $i = 1, \ldots, n$. However, unlike determinants which have a clear geometric interpretation as signed volumes, permanents do not have any natural geometric interpretation. Furthermore, determinants can be evaluated efficiently, for example using the conversion to row reduced echelon form, but computing the permanent is hard.

Permanents turn out to have various combinatorial interpretations. One of these is in terms of perfect matchings of bipartite graphs which we now discuss.

Recall that a *bipartite* (undirected) graph $G = (V, E)$ is a graph whose set of nodes V can be partionned into two nonempty disjoint subsets V_1 and V_2, such that every edge $e \in E$ has one endpoint in V_1 and one endpoint in V_2. An example of a bipatite graph with 14 nodes is shown in Figure 4.1; its nodes are partitioned into the two sets $\{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$ and $\{y_1, y_2, y_3, y_4, y_5, y_6, y_7\}$.

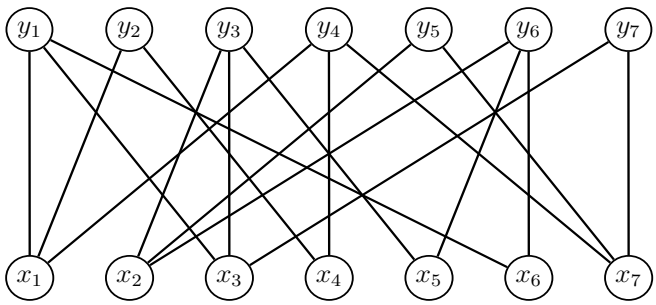


Figure 4.1: A bipartite graph G.

A *matching* in a graph $G = (V, E)$ (bipartite or not) is a set M of pairwise non-adjacent edges, which means that no two edges in M share a common vertex. A *perfect matching* is a matching such that every node in V is incident to some edge in the matching M (every node in V is an endpoint of some edge in M). Figure 4.2 shows a perfect matching (in red) in the bipartite graph G.

Obviously, a perfect matching in a bipartite graph can exist only if its set of nodes has a partition in two blocks of equal size, say $\{x_1, \ldots, x_m\}$ and $\{y_1, \ldots, y_m\}$. Then, there is a bijection between perfect matchings and bijections $\pi\colon \{x_1, \ldots, x_m\} \to \{y_1, \ldots, y_m\}$ such that $\pi(x_i) = y_j$ iff there is an edge between x_i and y_j.

Now, every bipartite graph G with a partition of its nodes into two sets of equal size as above is represented by an $m \times m$ matrix $A = (a_{ij})$ such that $a_{ij} = 1$ iff there is an edge between x_i and y_j, and $a_{ij} = 0$ otherwise. Using the interpretation of perfect machings as bijections $\pi\colon \{x_1, \ldots, x_m\} \to \{y_1, \ldots, y_m\}$, we see that *the permanent* $\mathrm{per}(A)$ *of the* $(0,1)$*-matrix* A *representing the bipartite graph* G *counts the number of perfect matchings in* G.

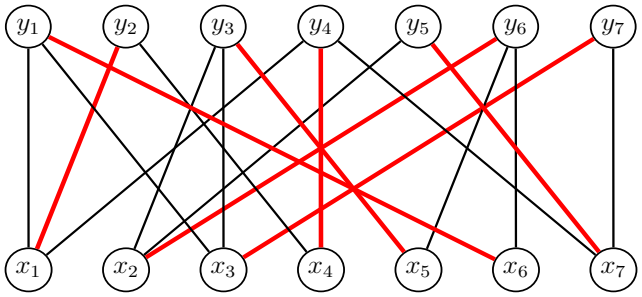


Figure 4.2: A perfect matching in the bipartite graph G.

In a famous paper published in 1979, Leslie Valiant proves that computing the permanent is a #P-complete problem. Such problems are suspected to be intractable. It is known that if a polynomial-time algorithm existed to solve a #P-complete problem, then we would have $P = NP$, which is believed to be very unlikely.

Another combinatorial interpretation of the permanent can be given in terms of systems of distinct representatives. Given a finite set S, let $(A_1, \ldots, A_n)$ be any sequence of nonempty subsets of S (not necessarily distinct). A *system of distinct representatives* (for short *SDR*) of the sets $A_1, \ldots, A_n$ is a sequence of n distinct elements $(a_1, \ldots, a_n)$, with $a_i \in A_i$ for $i = 1, \ldots, n$. The number of SDR's of a sequence of sets plays an important role in combinatorics. Now, if $S = \{1, 2, \ldots, n\}$ and if we associate to any sequence $(A_1, \ldots, A_n)$ of nonempty subsets of S the matrix $A = (a_{ij})$ defined such that $a_{ij} = 1$ if $j \in A_i$ and $a_{ij} = 0$ otherwise, then *the permanent* $\mathrm{per}(A)$ *counts the number of SDR's of the set* $A_1, \ldots, A_n$.

This interpretation of permanents in terms of SDR's can be used to prove bounds for the permanents of various classes of matrices. Interested readers are referred to van Lint and Wilson [108] (Chapters 11 and 12). In particular, a proof of a theorem known as *Van der Waerden conjecture* is given in Chapter 12. This theorem states that for any $n \times n$ matrix A with nonnegative entries in which all row-sums and column-sums are 1 (doubly stochastic matrices), we have

$$\mathrm{per}(A) \geq \frac{n!}{n^n},$$

with equality for the matrix in which all entries are equal to $1/n$.

4.9 Summary

The main concepts and results of this chapter are listed below:

- *permutations*, *transpositions*, *basics transpositions*.
- Every permutation can be written as a composition of permutations.

- The *parity* of the number of transpositions involved in any decomposition of a permutation σ is an invariant; it is the *signature* $\epsilon(\sigma)$ of the permutation σ.
- *Multilinear maps* (also called *n-linear maps*); *bilinear maps*.
- *Symmetric* and *alternating* multilinear maps.
- A basic property of alternating multilinear maps (Lemma 4.5) and the introduction of the formula expressing a determinant.
- Definition of a *determinant* as a multlinear alternating map $D\colon \mathrm{M}_n(K) \to K$ such that $D(I) = 1$.
- We define the set of algorithms $\mathcal{D}_n$, to compute the determinant of an $n \times n$ matrix.
- *Laplace expansion according to the ith row*; *cofactors*.
- We prove that the algorithms in $\mathcal{D}_n$ compute determinants (Lemma 4.6).
- We prove that all algorithms in $\mathcal{D}_n$ compute the same determinant (Theorem 4.7).
- We give an interpretation of determinants as *signed volumes*.
- We prove that $\det(A) = \det(A^\top)$.
- We prove that $\det(AB) = \det(A)\det(B)$.
- The *adjugate* $\widetilde{A}$ of a matrix A.
- Formula for the inverse in terms of the adjugate.
- A matrix A is invertible iff $\det(A) \neq 0$.
- Solving linear equations using *Cramer's rules*.
- Determinant of a linear map.
- The *characteristic polynomial* of a matrix.
- The *Cayley–Hamilton theorem*.
- The action of the polynomial ring induced by a linear map on a vector space.
- *Permanents*.
- Permanents count the number of perfect matchings in bipartite graphs.
- Computing the permanent is a #P-perfect problem (L. Valiant).
- Permanents count the number of SDRs of sequences of subsets of a given set.

4.10 Further Readings

Thorough expositions of the material covered in Chapter 1–3 and 4 can be found in Strang [102, 101], Lax [66], Lang [62], Artin [6], Mac Lane and Birkhoff [70], Hoffman and Kunze [58], Dummit and Foote [38], Bourbaki [19, 20], Van Der Waerden [107], Serre [95], Horn and Johnson [55], and Bertin [12]. These notions of linear algebra are nicely put to use in classical geometry, see Berger [9, 10], Tisseron [104] and Dieudonné [34].

Chapter 5

Gaussian Elimination, LU-Factorization, Cholesky Factorization, Reduced Row Echelon Form

5.1 Motivating Example: Curve Interpolation

Curve interpolation is a problem that arises frequently in computer graphics and in robotics (path planning). There are many ways of tackling this problem and in this section we will describe a solution using *cubic splines*. Such splines consist of cubic Bézier curves. They are often used because they are cheap to implement and give more flexibility than quadratic Bézier curves.

A *cubic Bézier curve* $C(t)$ (in $\mathbb{R}^2$ or $\mathbb{R}^3$) is specified by a list of four *control points* (b_0, b_2, b_2, b_3) and is given parametrically by the equation

$$C(t) = (1-t)^3\, b_0 + 3(1-t)^2 t\, b_1 + 3(1-t)t^2\, b_2 + t^3\, b_3.$$

Clearly, $C(0) = b_0$, $C(1) = b_3$, and for $t \in [0, 1]$, the point $C(t)$ belongs to the convex hull of the control points b_0, b_1, b_2, b_3. The polynomials

$$(1-t)^3, \quad 3(1-t)^2 t, \quad 3(1-t)t^2, \quad t^3$$

are the *Bernstein polynomials* of degree 3.

Typically, we are only interested in the curve segment corresponding to the values of t in the interval $[0, 1]$. Still, the placement of the control points drastically affects the shape of the curve segment, which can even have a self-intersection; See Figures 5.1, 5.2, 5.3 illustrating various configuations.

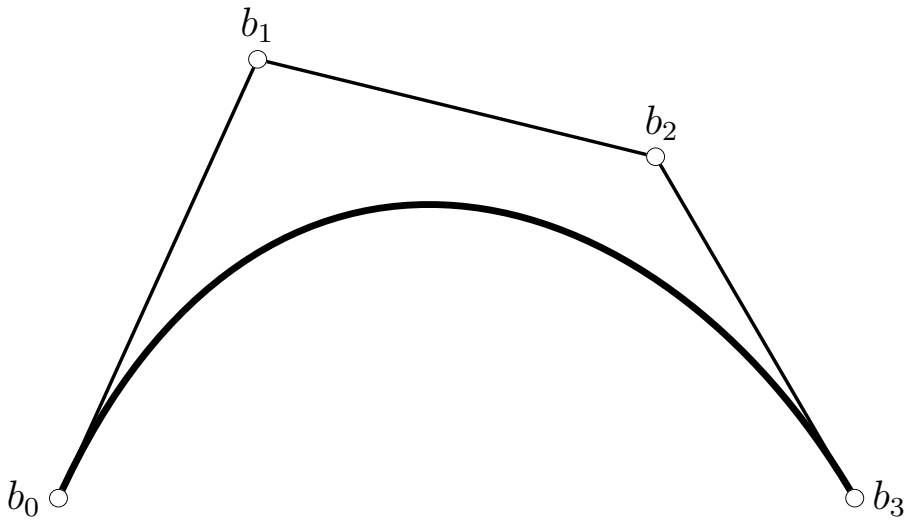


Figure 5.1: A "standard" Bézier curve

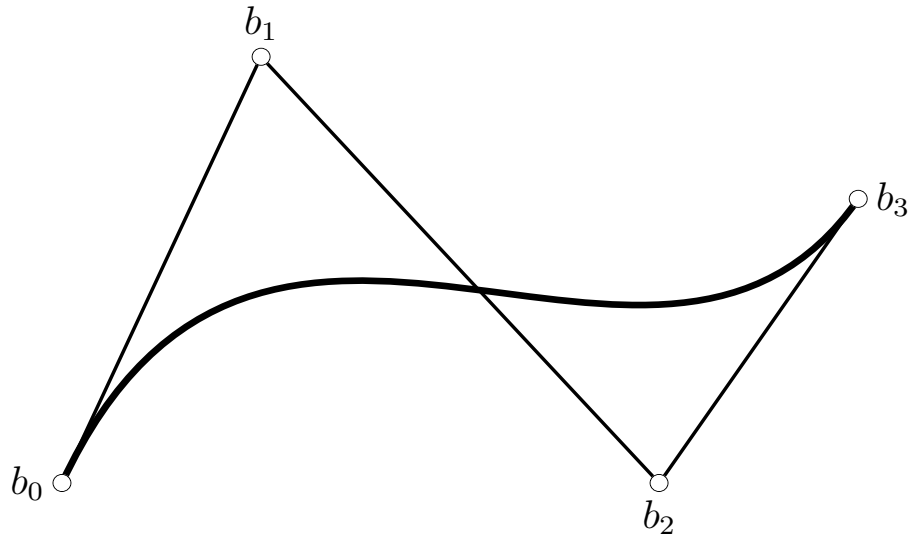


Figure 5.2: A Bézier curve with an inflexion point

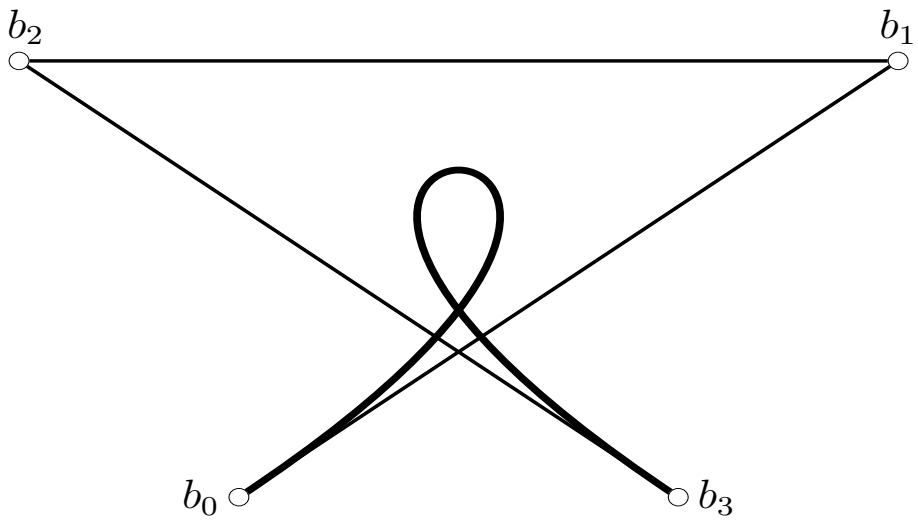


Figure 5.3: A self-intersecting Bézier curve

Interpolation problems require finding curves passing through some given data points and possibly satisfying some extra constraints.

A *Bézier spline curve* F is a curve which is made up of curve segments which are Bézier curves, say $C_1, \ldots, C_m$ $(m \geq 2)$. We will assume that F defined on $[0, m]$, so that for $i = 1, \ldots, m$,

$$F(t) = C_i(t - i + 1), \quad i - 1 \leq t \leq i.$$

Typically, some smoothness is required between any two junction points, that is, between any two points $C_i(1)$ and $C_{i+1}(0)$, for $i = 1, \ldots, m - 1$. We require that $C_i(1) = C_{i+1}(0)$ (C^0-*continuity*), and typically that the derivatives of C_i at 1 and of C_{i+1} at 0 agree up to second order derivatives. This is called C^2-*continuity*, and it ensures that the tangents agree as well as the curvatures.

There are a number of interpolation problems, and we consider one of the most common problems which can be stated as follows:

Problem: Given $N + 1$ data points $x_0, \ldots, x_N$, find a C^2 cubic spline curve F such that $F(i) = x_i$ for all i, $0 \leq i \leq N$ $(N \geq 2)$.

A way to solve this problem is to find $N + 3$ auxiliary points $d_{-1}, \ldots, d_{N+1}$, called *de Boor control points*, from which N Bézier curves can be found. Actually,

$$d_{-1} = x_0 \quad \text{and} \quad d_{N+1} = x_N$$

so we only need to find $N+1$ points $d_0, \ldots, d_N$.

It turns out that the C^2-continuity constraints on the N Bézier curves yield only $N-1$ equations, so d_0 and d_N can be chosen arbitrarily. In practice, d_0 and d_N are chosen according to various *end conditions*, such as prescribed velocities at x_0 and x_N. For the time being, we will assume that d_0 and d_N are given.

Figure 5.4 illustrates an interpolation problem involving $N+1 = 7+1 = 8$ data points. The control points d_0 and d_7 were chosen arbitrarily.

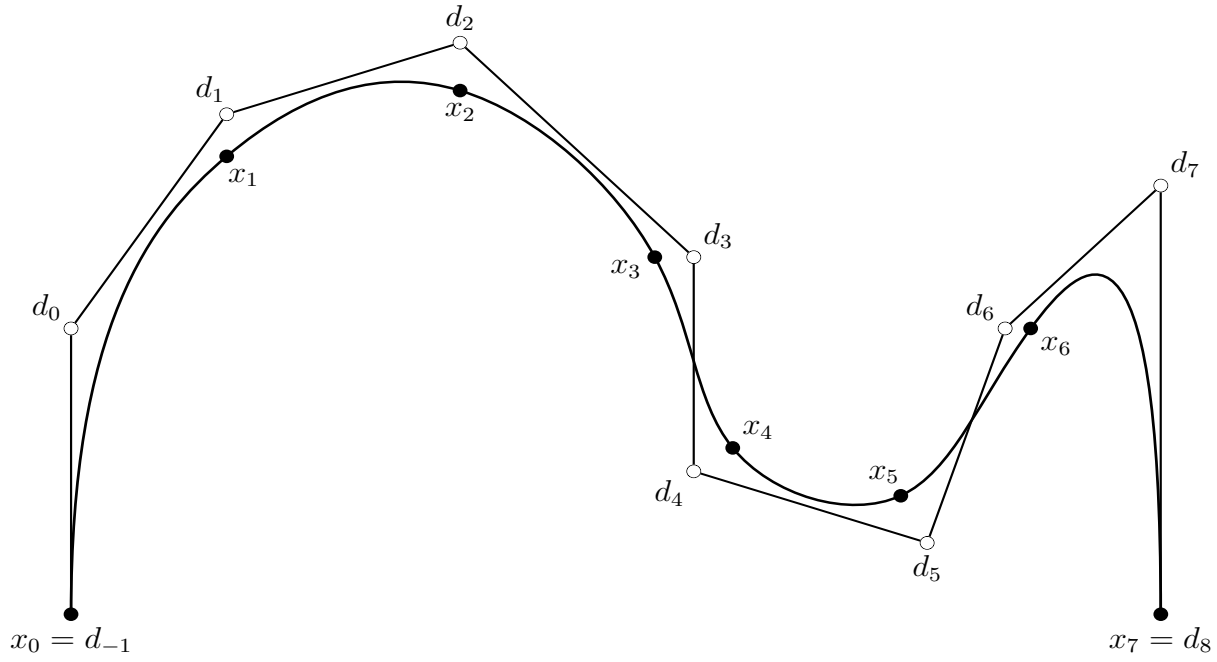


Figure 5.4: A C^2 cubic interpolation spline curve passing through the points $x_0, x_1, x_2, x_3, x_4, x_5, x_6, x_7$

It can be shown that $d_1, \ldots, d_{N-1}$ are given by the linear system

$$\begin{pmatrix} \frac{7}{2} & 1 & & & \\ 1 & 4 & 1 & & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & & 1 & 4 & 1 \\ & & & 1 & \frac{7}{2} \end{pmatrix} \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_{N-2} \\ d_{N-1} \end{pmatrix} = \begin{pmatrix} 6x_1 - \frac{3}{2}d_0 \\ 6x_2 \\ \vdots \\ 6x_{N-2} \\ 6x_{N-1} - \frac{3}{2}d_N \end{pmatrix}.$$

We will show later that the above matrix is invertible because it is strictly diagonally dominant.

Once the above system is solved, the Bézier cubics $C_1, \ldots, C_N$ are determined as follows (we assume $N \geq 2$): For $2 \leq i \leq N-1$, the control points $(b_0^i, b_1^i, b_2^i, b_3^i)$ of C_i are given by

$$\begin{aligned} b_0^i &= x_{i-1} \\ b_1^i &= \frac{2}{3}d_{i-1} + \frac{1}{3}d_i \\ b_2^i &= \frac{1}{3}d_{i-1} + \frac{2}{3}d_i \\ b_3^i &= x_i. \end{aligned}$$

The control points $(b_0^1, b_1^1, b_2^1, b_3^1)$ of C_1 are given by

$$\begin{aligned} b_0^1 &= x_0 \\ b_1^1 &= d_0 \\ b_2^1 &= \frac{1}{2}d_0 + \frac{1}{2}d_1 \\ b_3^1 &= x_1, \end{aligned}$$

and the control points $(b_0^N, b_1^N, b_2^N, b_3^N)$ of C_N are given by

$$\begin{aligned} b_0^N &= x_{N-1} \\ b_1^N &= \frac{1}{2}d_{N-1} + \frac{1}{2}d_N \\ b_2^N &= d_N \\ b_3^N &= x_N. \end{aligned}$$

We will now describe various methods for solving linear systems. Since the matrix of the above system is tridiagonal, there are specialized methods which are more efficient than the general methods. We will discuss a few of these methods.

5.2 Gaussian Elimination

Let A be an $n \times n$ matrix, let $b \in \mathbb{R}^n$ be an n-dimensional vector and assume that A is invertible. Our goal is to solve the system $Ax = b$. Since A is assumed to be invertible, we know that this system has a unique solution $x = A^{-1}b$. Experience shows that two counter-intuitive facts are revealed:

(1) One should avoid computing the inverse A^{-1} of A explicitly. This is because this would amount to solving the n linear systems $Au^{(j)} = e_j$ for $j = 1, \ldots, n$, where $e_j = (0, \ldots, 1, \ldots, 0)$ is the jth canonical basis vector of $\mathbb{R}^n$ (with a 1 is the jth slot). By doing so, we would replace the resolution of a single system by the resolution of n systems, and we would still have to multiply A^{-1} by b.

(2) One does not solve (large) linear systems by computing determinants (using Cramer's formulae). This is because this method requires a number of additions (resp. multiplications) proportional to $(n+1)!$ (resp. $(n+2)!$).

The key idea on which most direct methods (as opposed to iterative methods, that look for an approximation of the solution) are based is that if A is an upper-triangular matrix, which means that $a_{ij} = 0$ for $1 \leq j < i \leq n$ (resp. lower-triangular, which means that $a_{ij} = 0$ for $1 \leq i < j \leq n$), then computing the solution x is trivial. Indeed, say A is an upper-triangular matrix

$$A = \begin{pmatrix} a_{1\,1} & a_{1\,2} & \cdots & a_{1\,n-2} & a_{1\,n-1} & a_{1\,n} \\ 0 & a_{2\,2} & \cdots & a_{2\,n-2} & a_{2\,n-1} & a_{2\,n} \\ 0 & 0 & \ddots & \vdots & \vdots & \vdots \\ & & & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & a_{n-1\,n-1} & a_{n-1\,n} \\ 0 & 0 & \cdots & 0 & 0 & a_{n\,n} \end{pmatrix}.$$

Then, $\det(A) = a_{1\,1}a_{2\,2}\cdots a_{n\,n} \neq 0$, which implies that $a_{i\,i} \neq 0$ for $i = 1, \ldots, n$, and we can solve the system $Ax = b$ from bottom-up by *back-substitution*. That is, first we compute x_n from the last equation, next plug this value of x_n into the next to the last equation and compute x_{n-1} from it, *etc.* This yields

$$\begin{aligned} x_n &= a_{n\,n}^{-1} b_n \\ x_{n-1} &= a_{n-1\,n-1}^{-1}(b_{n-1} - a_{n-1\,n}x_n) \\ &\vdots \\ x_1 &= a_{1\,1}^{-1}(b_1 - a_{1\,2}x_2 - \cdots - a_{1\,n}x_n). \end{aligned}$$

Note that the use of determinants can be avoided to prove that if A is invertible then $a_{i\,i} \neq 0$ for $i = 1, \ldots, n$. Indeed, it can be shown directly (by induction) that an upper (or lower) triangular matrix is invertible iff all its diagonal entries are nonzero.

If A is lower-triangular, we solve the system from top-down by *forward-substitution*.

Thus, what we need is a method for transforming a matrix to an equivalent one in upper-triangular form. This can be done by *elimination*. Let us illustrate this method on the following example:

$$\begin{array}{rcrcrcr} 2x & + & y & + & z & = & 5 \\ 4x & - & 6y & & & = & -2 \\ -2x & + & 7y & + & 2z & = & 9. \end{array}$$

We can eliminate the variable x from the second and the third equation as follows: Subtract twice the first equation from the second and add the first equation to the third. We get the

new system

$$\begin{array}{rcrcrcr} 2x & + & y & + & z & = & 5 \\ & - & 8y & - & 2z & = & -12 \\ & & 8y & + & 3z & = & 14. \end{array}$$

This time, we can eliminate the variable y from the third equation by adding the second equation to the third:

$$\begin{array}{rcrcrcr} 2x & + & y & + & z & = & 5 \\ & - & 8y & - & 2z & = & -12 \\ & & & & z & = & 2. \end{array}$$

This last system is upper-triangular. Using back-substitution, we find the solution: $z = 2$, $y = 1$, $x = 1$.

Observe that we have performed only row operations. The general method is to iteratively eliminate variables using simple row operations (namely, adding or subtracting a multiple of a row to another row of the matrix) while simultaneously applying these operations to the vector b, to obtain a system, $MAx = Mb$, where MA is upper-triangular. Such a method is called *Gaussian elimination*. However, one extra twist is needed for the method to work in all cases: It may be necessary to permute rows, as illustrated by the following example:

$$\begin{array}{rcrcrl} x & + & y & + & z & = 1 \\ x & + & y & + & 3z & = 1 \\ 2x & + & 5y & + & 8z & = 1. \end{array}$$

In order to eliminate x from the second and third row, we subtract the first row from the second and we subtract twice the first row from the third:

$$\begin{array}{rcrcrl} x & + & y & + & z & = 1 \\ & & & & 2z & = 0 \\ & & 3y & + & 6z & = -1. \end{array}$$

Now, the trouble is that y does not occur in the second row; so, we can't eliminate y from the third row by adding or subtracting a multiple of the second row to it. The remedy is simple: Permute the second and the third row! We get the system:

$$\begin{array}{rcrcrl} x & + & y & + & z & = 1 \\ & & 3y & + & 6z & = -1 \\ & & & & 2z & = 0, \end{array}$$

which is already in triangular form. Another example where some permutations are needed is:

$$\begin{array}{rcrcrcr} & & & & z & = & 1 \\ -2x & + & 7y & + & 2z & = & 1 \\ 4x & - & 6y & & & = & -1. \end{array}$$

First, we permute the first and the second row, obtaining

$$\begin{array}{rcrcrcr} -2x & + & 7y & + & 2z & = & 1 \\ & & & & z & = & 1 \\ 4x & - & 6y & & & = & -1, \end{array}$$

and then, we add twice the first row to the third, obtaining:

$$\begin{array}{rcrcrcr} -2x & + & 7y & + & 2z & = & 1 \\ & & & & z & = & 1 \\ & & 8y & + & 4z & = & 1. \end{array}$$

Again, we permute the second and the third row, getting

$$\begin{array}{rcrcrcr} -2x & + & 7y & + & 2z & = & 1 \\ & & 8y & + & 4z & = & 1 \\ & & & & z & = & 1, \end{array}$$

an upper-triangular system. Of course, in this example, z is already solved and we could have eliminated it first, but for the general method, we need to proceed in a systematic fashion.

We now describe the method of *Gaussian Elimination* applied to a linear system $Ax = b$, where A is assumed to be invertible. We use the variable k to keep track of the stages of elimination. Initially, $k = 1$.

(1) The first step is to pick some nonzero entry $a_{i\,1}$ in the first column of A. Such an entry must exist, since A is invertible (otherwise, the first column of A would be the zero vector, and the columns of A would not be linearly independent. Equivalently, we would have $\det(A) = 0$). The actual choice of such an element has some impact on the numerical stability of the method, but this will be examined later. For the time being, we assume that some arbitrary choice is made. This chosen element is called the *pivot* of the elimination step and is denoted π_1 (so, in this first step, $\pi_1 = a_{i\,1}$).

(2) Next, we permute the row (i) corresponding to the pivot with the first row. Such a step is called *pivoting*. So, after this permutation, the first element of the first row is nonzero.

(3) We now eliminate the variable x_1 from all rows except the first by adding suitable multiples of the first row to these rows. More precisely we add $-a_{i\,1}/\pi_1$ times the first row to the ith row for $i = 2, \ldots, n$. At the end of this step, all entries in the first column are zero except the first.

(4) Increment k by 1. If $k = n$, stop. Otherwise, $k < n$, and then iteratively repeat steps (1), (2), (3) on the $(n - k + 1) \times (n - k + 1)$ subsystem obtained by deleting the first $k - 1$ rows and $k - 1$ columns from the current system.

If we let $A_1 = A$ and $A_k = (a_{i\,j}^{(k)})$ be the matrix obtained after $k-1$ elimination steps ($2 \leq k \leq n$), then the kth elimination step is applied to the matrix A_k of the form

$$A_k = \begin{pmatrix} a_{1\,1}^{(k)} & a_{1\,2}^{(k)} & \cdots & \cdots & \cdots & a_{1\,n}^{(k)} \\ & a_{2\,2}^{(k)} & \cdots & \cdots & \cdots & a_{2\,n}^{(k)} \\ & & \ddots & \vdots & & \vdots \\ & & & a_{k\,k}^{(k)} & \cdots & a_{k\,n}^{(k)} \\ & & & \vdots & & \vdots \\ & & & a_{n\,k}^{(k)} & \cdots & a_{n\,n}^{(k)} \end{pmatrix}.$$

Actually, note that

$$a_{i\,j}^{(k)} = a_{i\,j}^{(i)}$$

for all i, j with $1 \leq i \leq k-2$ and $i \leq j \leq n$, since the first $k-1$ rows remain unchanged after the $(k-1)$th step.

We will prove later that $\det(A_k) = \pm \det(A)$. Consequently, A_k is invertible. The fact that A_k is invertible iff A is invertible can also be shown without determinants from the fact that there is some invertible matrix M_k such that $A_k = M_k A$, as we will see shortly.

Since A_k is invertible, some entry $a_{i\,k}^{(k)}$ with $k \leq i \leq n$ is nonzero. Otherwise, the last $n-k+1$ entries in the first k columns of A_k would be zero, and the first k columns of A_k would yield k vectors in $\mathbb{R}^{k-1}$. But then, the first k columns of A_k would be linearly dependent and A_k would not be invertible, a contradiction.

So, one the entries $a_{i\,k}^{(k)}$ with $k \leq i \leq n$ can be chosen as pivot, and we permute the kth row with the ith row, obtaining the matrix $\alpha^{(k)} = (\alpha_{j\,l}^{(k)})$. The new pivot is $\pi_k = \alpha_{k\,k}^{(k)}$, and we zero the entries $i = k+1, \ldots, n$ in column k by adding $-\alpha_{i\,k}^{(k)}/\pi_k$ times row k to row i. At the end of this step, we have A_{k+1}. Observe that the first $k-1$ rows of A_k are identical to the first $k-1$ rows of A_{k+1}.

The process of Gaussian elimination is illustrated in schematic form below:

$$\begin{pmatrix} \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \end{pmatrix} \Longrightarrow \begin{pmatrix} \times & \times & \times & \times \\ \mathbf{0} & \times & \times & \times \\ \mathbf{0} & \times & \times & \times \\ \mathbf{0} & \times & \times & \times \end{pmatrix} \Longrightarrow \begin{pmatrix} \times & \times & \times & \times \\ 0 & \times & \times & \times \\ 0 & \mathbf{0} & \times & \times \\ 0 & \mathbf{0} & \times & \times \end{pmatrix} \Longrightarrow \begin{pmatrix} \times & \times & \times & \times \\ 0 & \times & \times & \times \\ 0 & 0 & \times & \times \\ 0 & 0 & \mathbf{0} & \times \end{pmatrix}.$$

5.3 Elementary Matrices and Row Operations

It is easy to figure out what kind of matrices perform the elementary row operations used during Gaussian elimination. The key point is that if $A = PB$, where A, B are $m \times n$ matrices and P is a square matrix of dimension m, if (as usual) we denote the rows of A and

B by $A_1, \ldots, A_m$ and $B_1, \ldots, B_m$, then the formula

$$a_{ij} = \sum_{k=1}^{m} p_{ik} b_{kj}$$

giving the (i, j)th entry in A shows that the ith row of A is a *linear combination* of the rows of B:

$$A_i = p_{i1} B_1 + \cdots + p_{im} B_m.$$

Therefore, *multiplication of a matrix on the left by a square matrix performs row operations*. Similarly, multiplication of a matrix on the right by a square matrix performs column operations

The permutation of the kth row with the ith row is achieved by multiplying A on the left by the *transposition matrix* $P(i, k)$, which is the matrix obtained from the identity matrix by permuting rows i and k, *i.e.*,

$$P(i,k) = \begin{pmatrix} 1 & & & & & & & & \\ & 1 & & & & & & & \\ & & 0 & & & & 1 & & \\ & & & 1 & & & & & \\ & & & & \ddots & & & & \\ & & & & & 1 & & & \\ & & 1 & & & & 0 & & \\ & & & & & & & 1 & \\ & & & & & & & & 1 \end{pmatrix}.$$

Observe that $\det(P(i, k)) = -1$. Furthermore, $P(i, k)$ is *symmetric* ($P(i, k)^\top = P(i, k)$), and

$$P(i, k)^{-1} = P(i, k).$$

During the permutation step (2), if row k and row i need to be permuted, the matrix A is multiplied on the left by the matrix P_k such that $P_k = P(i, k)$, else we set $P_k = I$.

Adding β times row j to row i (with $i \neq j$) is achieved by multiplying A on the left by the *elementary matrix*,

$$E_{i,j;\beta} = I + \beta e_{i\,j},$$

where

$$(e_{i\,j})_{k\,l} = \begin{cases} 1 & \text{if } k = i \text{ and } l = j \\ 0 & \text{if } k \neq i \text{ or } l \neq j, \end{cases}$$

i.e.,

$$E_{i,j;\beta} = \begin{pmatrix} 1 & & & & & & & & \\ & 1 & & & & & & & \\ & & 1 & & & & & & \\ & & & 1 & & & & & \\ & & & & \ddots & & & & \\ & & & & & 1 & & & \\ & & \beta & & & & 1 & & \\ & & & & & & & 1 & \\ & & & & & & & & 1 \end{pmatrix} \quad \text{or} \quad E_{i,j;\beta} = \begin{pmatrix} 1 & & & & & & & & \\ & 1 & & & & & & & \\ & & 1 & & & & \beta & & \\ & & & 1 & & & & & \\ & & & & \ddots & & & & \\ & & & & & 1 & & & \\ & & & & & & 1 & & \\ & & & & & & & 1 & \\ & & & & & & & & 1 \end{pmatrix}.$$

On the left, $i > j$, and on the right, $i < j$. Observe that the inverse of $E_{i,j;\beta} = I + \beta e_{i\,j}$ is $E_{i,j;-\beta} = I - \beta e_{i\,j}$ and that $\det(E_{i,j;\beta}) = 1$. Therefore, during step 3 (the elimination step), the matrix A is multiplied on the left by a product E_k of matrices of the form $E_{i,k;\beta_{i,k}}$, with $i > k$.

Consequently, we see that

$$A_{k+1} = E_k P_k A_k,$$

and then

$$A_k = E_{k-1} P_{k-1} \cdots E_1 P_1 A.$$

This justifies the claim made earlier that $A_k = M_k A$ for some invertible matrix M_k; we can pick

$$M_k = E_{k-1} P_{k-1} \cdots E_1 P_1,$$

a product of invertible matrices.

The fact that $\det(P(i,k)) = -1$ and that $\det(E_{i,j;\beta}) = 1$ implies immediately the fact claimed above: We always have

$$\det(A_k) = \pm \det(A).$$

Furthermore, since

$$A_k = E_{k-1} P_{k-1} \cdots E_1 P_1 A$$

and since Gaussian elimination stops for $k = n$, the matrix

$$A_n = E_{n-1} P_{n-1} \cdots E_2 P_2 E_1 P_1 A$$

is upper-triangular. Also note that if we let $M = E_{n-1} P_{n-1} \cdots E_2 P_2 E_1 P_1$, then $\det(M) = \pm 1$, and

$$\det(A) = \pm \det(A_n).$$

The matrices $P(i,k)$ and $E_{i,j;\beta}$ are called *elementary matrices*. We can summarize the above discussion in the following theorem:

Theorem 5.1. *(Gaussian Elimination) Let A be an $n \times n$ matrix (invertible or not). Then there is some invertible matrix M so that $U = MA$ is upper-triangular. The pivots are all nonzero iff A is invertible.*

Proof. We already proved the theorem when A is invertible, as well as the last assertion. Now, A is singular iff some pivot is zero, say at stage k of the elimination. If so, we must have $a_{i\,k}^{(k)} = 0$ for $i = k, \ldots, n$; but in this case, $A_{k+1} = A_k$ and we may pick $P_k = E_k = I$. □

Remark: Obviously, the matrix M can be computed as

$$M = E_{n-1}P_{n-1} \cdots E_2P_2E_1P_1,$$

but this expression is of no use. Indeed, what we need is M^{-1}; when no permutations are needed, it turns out that M^{-1} can be obtained immediately from the matrices E_k's, in fact, from their inverses, and no multiplications are necessary.

Remark: Instead of looking for an invertible matrix M so that MA is upper-triangular, we can look for an invertible matrix M so that MA is a diagonal matrix. Only a simple change to Gaussian elimination is needed. At every stage, k, after the pivot has been found and pivoting been performed, if necessary, in addition to adding suitable multiples of the kth row to the rows *below* row k in order to zero the entries in column k for $i = k+1, \ldots, n$, also add suitable multiples of the kth row to the rows *above* row k in order to zero the entries in column k for $i = 1, \ldots, k-1$. Such steps are also achieved by multiplying on the left by elementary matrices $E_{i,k;\beta_{i,k}}$, except that $i < k$, so that these matrices are not lower-triangular matrices. Nevertheless, at the end of the process, we find that $A_n = MA$, is a diagonal matrix.

This method is called the *Gauss-Jordan factorization*. Because it is more expensive than Gaussian elimination, this method is not used much in practice. However, Gauss-Jordan factorization can be used to compute the inverse of a matrix A. Indeed, we find the jth column of A^{-1} by solving the system $Ax^{(j)} = e_j$ (where e_j is the jth canonical basis vector of $\mathbb{R}^n$). By applying Gauss-Jordan, we are led to a system of the form $D_jx^{(j)} = M_je_j$, where D_j is a diagonal matrix, and we can immediately compute $x^{(j)}$.

It remains to discuss the choice of the pivot, and also conditions that guarantee that no permutations are needed during the Gaussian elimination process. We begin by stating a necessary and sufficient condition for an invertible matrix to have an LU-factorization (*i.e.*, Gaussian elimination does not require pivoting).

5.4 LU-Factorization

We say that an invertible matrix A has an *LU-factorization* if it can be written as $A = LU$, where U is upper-triangular invertible and L is lower-triangular, with $L_{i\,i} = 1$ for $i = 1, \ldots, n$.

A lower-triangular matrix with diagonal entries equal to 1 is called a *unit lower-triangular* matrix. Given an $n \times n$ matrix $A = (a_{ij})$, for any k with $1 \leq k \leq n$, let $A[1..k, 1..k]$ denote the submatrix of A whose entries are a_{ij}, where $1 \leq i, j \leq k$.

Proposition 5.2. *Let A be an invertible $n \times n$-matrix. Then, A has an LU-factorization $A = LU$ iff every matrix $A[1..k, 1..k]$ is invertible for $k = 1, \ldots, n$. Furthermore, when A has an LU-factorization, we have*

$$\det(A[1..k, 1..k]) = \pi_1 \pi_2 \cdots \pi_k, \quad k = 1, \ldots, n,$$

where π_k is the pivot obtained after $k-1$ elimination steps. Therefore, the kth pivot is given by

$$\pi_k = \begin{cases} a_{11} = \det(A[1..1, 1..1]) & \text{if } k = 1 \\ \dfrac{\det(A[1..k, 1..k])}{\det(A[1..k-1, 1..k-1])} & \text{if } k = 2, \ldots, n. \end{cases}$$

Proof. First, assume that $A = LU$ is an LU-factorization of A. We can write

$$A = \begin{pmatrix} A[1..k, 1..k] & A_2 \\ A_3 & A_4 \end{pmatrix} = \begin{pmatrix} L_1 & 0 \\ L_3 & L_4 \end{pmatrix} \begin{pmatrix} U_1 & U_2 \\ 0 & U_4 \end{pmatrix} = \begin{pmatrix} L_1 U_1 & L_1 U_2 \\ L_3 U_1 & L_3 U_2 + L_4 U_4 \end{pmatrix},$$

where L_1, L_4 are unit lower-triangular and U_1, U_4 are upper-triangular. Thus,

$$A[1..k, 1..k] = L_1 U_1,$$

and since U is invertible, U_1 is also invertible (the determinant of U is the product of the diagonal entries in U, which is the product of the diagonal entries in U_1 and U_4). As L_1 is invertible (since its diagonal entries are equal to 1), we see that $A[1..k, 1..k]$ is invertible for $k = 1, \ldots, n$.

Conversely, assume that $A[1..k, 1..k]$ is invertible for $k = 1, \ldots, n$. We just need to show that Gaussian elimination does not need pivoting. We prove by induction on k that the kth step does not need pivoting.

This holds for $k = 1$, since $A[1..1, 1..1] = (a_{1\,1})$, so $a_{1\,1} \neq 0$. Assume that no pivoting was necessary for the first $k-1$ steps ($2 \leq k \leq n-1$). In this case, we have

$$E_{k-1} \cdots E_2 E_1 A = A_k,$$

where $L = E_{k-1} \cdots E_2 E_1$ is a unit lower-triangular matrix and $A_k[1..k, 1..k]$ is upper-triangular, so that $LA = A_k$ can be written as

$$\begin{pmatrix} L_1 & 0 \\ L_3 & L_4 \end{pmatrix} \begin{pmatrix} A[1..k, 1..k] & A_2 \\ A_3 & A_4 \end{pmatrix} = \begin{pmatrix} U_1 & B_2 \\ 0 & B_4 \end{pmatrix},$$

where L_1 is unit lower-triangular and U_1 is upper-triangular. But then,

$$L_1 A[1..k, 1..k]) = U_1,$$

where L_1 is invertible (in fact, $\det(L_1) = 1$), and since by hypothesis $A[1..k, 1..k]$ is invertible, U_1 is also invertible, which implies that $(U_1)_{kk} \neq 0$, since U_1 is upper-triangular. Therefore, no pivoting is needed in step k, establishing the induction step. Since $\det(L_1) = 1$, we also have

$$\det(U_1) = \det(L_1 A[1..k, 1..k]) = \det(L_1)\det(A[1..k, 1..k]) = \det(A[1..k, 1..k]),$$

and since U_1 is upper-triangular and has the pivots $\pi_1, \ldots, \pi_k$ on its diagonal, we get

$$\det(A[1..k, 1..k]) = \pi_1\pi_2\cdots\pi_k, \quad k = 1, \ldots, n,$$

as claimed. □

Remark: The use of determinants in the first part of the proof of Proposition 5.2 can be avoided if we use the fact that a triangular matrix is invertible iff all its diagonal entries are nonzero.

Corollary 5.3. *(LU-Factorization) Let A be an invertible $n \times n$-matrix. If every matrix $A[1..k, 1..k]$ is invertible for $k = 1, \ldots, n$, then Gaussian elimination requires no pivoting and yields an LU-factorization $A = LU$.*

Proof. We proved in Proposition 5.2 that in this case Gaussian elimination requires no pivoting. Then, since every elementary matrix $E_{i,k;\beta}$ is lower-triangular (since we always arrange that the pivot π_k occurs above the rows that it operates on), since $E_{i,k;\beta}^{-1} = E_{i,k;-\beta}$ and the $E_k's$ are products of $E_{i,k;\beta_{i,k}}$'s, from

$$E_{n-1}\cdots E_2 E_1 A = U,$$

where U is an upper-triangular matrix, we get

$$A = LU,$$

where $L = E_1^{-1}E_2^{-1}\cdots E_{n-1}^{-1}$ is a lower-triangular matrix. Furthermore, as the diagonal entries of each $E_{i,k;\beta}$ are 1, the diagonal entries of each E_k are also 1. □

The reader should verify that the example below is indeed an LU-factorization.

$$\begin{pmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 4 & 3 & 1 & 0 \\ 3 & 4 & 1 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 2 \end{pmatrix}.$$

One of the main reasons why the existence of an LU-factorization for a matrix A is interesting is that if we need to solve *several* linear systems $Ax = b$ corresponding to the same matrix A, we can do this cheaply by solving the two triangular systems

$$Lw = b, \quad \text{and} \quad Ux = w.$$

There is a certain asymmetry in the LU-decomposition $A = LU$ of an invertible matrix A. Indeed, the diagonal entries of L are all 1, but this is generally false for U. This asymmetry can be eliminated as follows: if

$$D = \mathrm{diag}(u_{11}, u_{22}, \ldots, u_{nn})$$

is the diagonal matrix consisting of the diagonal entries in U (the pivots), then we if let $U' = D^{-1}U$, we can write

$$A = LDU',$$

where L is lower- triangular, U' is upper-triangular, all diagonal entries of both L and U' are 1, and D is a diagonal matrix of pivots. Such a decomposition is called an *LDU-factorization*. We will see shortly than if A is symmetric, then $U' = L^\top$.

As we will see a bit later, symmetric positive definite matrices satisfy the condition of Proposition 5.2. Therefore, linear systems involving symmetric positive definite matrices can be solved by Gaussian elimination without pivoting. Actually, it is possible to do better: This is the Cholesky factorization.

If a square invertible matrix A has an LU-factorization, then it is possible to find L and U while performing Gaussian elimination. Recall that at step k, we pick a pivot $\pi_k = a_{ik}^{(k)} \neq 0$ in the portion consisting of the entries of index $j \geq k$ of the k-th column of the matrix A_k obtained so far, we swap rows i and k if necessary (the pivoting step), and then we zero the entries of index $j = k+1, \ldots, n$ in column k. Schematically, we have the following steps:

$$\begin{pmatrix} \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ 0 & a_{ik}^{(k)} & \times & \times & \times \\ 0 & \times & \times & \times & \times \end{pmatrix} \overset{\text{pivot}}{\Longrightarrow} \begin{pmatrix} \times & \times & \times & \times & \times \\ 0 & a_{ik}^{(k)} & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \end{pmatrix} \overset{\text{elim}}{\Longrightarrow} \begin{pmatrix} \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ 0 & \mathbf{0} & \times & \times & \times \\ 0 & \mathbf{0} & \times & \times & \times \\ 0 & \mathbf{0} & \times & \times & \times \end{pmatrix}.$$

More precisely, after permuting row k and row i (the pivoting step), if the entries in column k below row k are $\alpha_{k+1k}, \ldots, \alpha_{nk}$, then we add $-\alpha_{jk}/\pi_k$ times row k to row j; this process is illustrated below:

$$\begin{pmatrix} a_{kk}^{(k)} \\ a_{k+1k}^{(k)} \\ \vdots \\ a_{ik}^{(k)} \\ \vdots \\ a_{nk}^{(k)} \end{pmatrix} \overset{\text{pivot}}{\Longrightarrow} \begin{pmatrix} a_{ik}^{(k)} \\ a_{k+1k}^{(k)} \\ \vdots \\ a_{kk}^{(k)} \\ \vdots \\ a_{nk}^{(k)} \end{pmatrix} = \begin{pmatrix} \pi_k \\ \alpha_{k+1k} \\ \vdots \\ \alpha_{ik} \\ \vdots \\ \alpha_{nk} \end{pmatrix} \overset{\text{elim}}{\Longrightarrow} \begin{pmatrix} \pi_k \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}.$$

Then, if we write $\ell_{jk} = \alpha_{jk}/\pi_k$ for $j = k+1, \ldots, n$, the kth column of L is

$$\begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \ell_{k+1k} \\ \vdots \\ \ell_{nk} \end{pmatrix}.$$

Observe that the signs of the multipliers $-\alpha_{jk}/\pi_k$ have been flipped. Thus, we obtain the unit lower triangular matrix

$$L = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \ell_{21} & 1 & 0 & \cdots & 0 \\ \ell_{31} & \ell_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \ell_{n1} & \ell_{n2} & \ell_{n3} & \cdots & 1 \end{pmatrix}.$$

It is easy to see (and this is proved in Theorem 5.5) that the inverse of L is obtained from L by flipping the signs of the ℓ_{ij}:

$$L^{-1} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -\ell_{21} & 1 & 0 & \cdots & 0 \\ -\ell_{31} & -\ell_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ -\ell_{n1} & -\ell_{n2} & -\ell_{n3} & \cdots & 1 \end{pmatrix}.$$

Furthermore, if the result of Gaussian elimination (without pivoting) is $U = E_{n-1} \cdots E_1 A$, then

$$E_k = \begin{pmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & -\ell_{k+1k} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & -\ell_{nk} & 0 & \cdots & 1 \end{pmatrix} \quad \text{and} \quad E_k^{-1} = \begin{pmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & \ell_{k+1k} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \ell_{nk} & 0 & \cdots & 1 \end{pmatrix},$$

so the kth column of E_k is the kth column of L^{-1}.

Here is an example illustrating the method. Given

$$A = A_1 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & -1 \end{pmatrix},$$

we have the following sequence of steps: The first pivot is $\pi_1 = 1$ in row 1, and we substract row 1 from rows 2, 3, and 4. We get

$$A_2 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & -2 & -1 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & -2 & -1 & -1 \end{pmatrix} \quad L_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}.$$

The next pivot is $\pi_2 = -2$ in row 2, and we subtract row 2 from row 4 (and add 0 times row 2 to row 3). We get

$$A_3 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & -2 & -1 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -2 \end{pmatrix} \quad L_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix}.$$

The next pivot is $\pi_3 = -2$ in row 3, and since the fourth entry in column 3 is already a zero, we add 0 times row 3 to row 4. We get

$$A_4 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & -2 & -1 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -2 \end{pmatrix} \quad L_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix}.$$

The procedure is finished, and we have

$$L = L_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix} \qquad U = A_4 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & -2 & -1 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -2 \end{pmatrix}.$$

It is easy to check that indeed

$$LU = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & -2 & -1 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & -1 \end{pmatrix} = A.$$

We now show how to extend the above method to deal with pivoting efficiently. This is the $PA = LU$ factorization.

5.5 $PA = LU$ Factorization

The following easy proposition shows that, in principle, A can be premultiplied by some permutation matrix P, so that PA can be converted to upper-triangular form without using

any pivoting. Permutations are discussed in some detail in Section 4.1, but for now we just need their definition. A *permutation matrix* is a square matrix that has a single 1 in every row and every column and zeros everywhere else. It is shown in Section 4.1 that every permutation matrix is a product of transposition matrices (the $P(i,k)$s), and that P is invertible with inverse $P^\top$.

Proposition 5.4. *Let A be an invertible $n \times n$-matrix. Then, there is some permutation matrix P so that $(PA)[1..k, 1..k]$ is invertible for $k = 1, \ldots, n$.*

Proof. The case $n = 1$ is trivial, and so is the case $n = 2$ (we swap the rows if necessary). If $n \geq 3$, we proceed by induction. Since A is invertible, its columns are linearly independent; in particular, its first $n-1$ columns are also linearly independent. Delete the last column of A. Since the remaining $n-1$ columns are linearly independent, there are also $n-1$ linearly independent rows in the corresponding $n \times (n-1)$ matrix. Thus, there is a permutation of these n rows so that the $(n-1) \times (n-1)$ matrix consisting of the first $n-1$ rows is invertible. But, then, there is a corresponding permutation matrix P_1, so that the first $n-1$ rows and columns of $P_1 A$ form an invertible matrix A'. Applying the induction hypothesis to the $(n-1) \times (n-1)$ matrix A', we see that there some permutation matrix P_2 (leaving the nth row fixed), so that $P_2 P_1 A[1..k, 1..k]$ is invertible, for $k = 1, \ldots, n-1$. Since A is invertible in the first place and P_1 and P_2 are invertible, $P_1 P_2 A$ is also invertible, and we are done. □

Remark: One can also prove Proposition 5.4 using a clever reordering of the Gaussian elimination steps suggested by Trefethen and Bau [105] (Lecture 21). Indeed, we know that if A is invertible, then there are permutation matrices P_i and products of elementary matrices E_i, so that

$$A_n = E_{n-1} P_{n-1} \cdots E_2 P_2 E_1 P_1 A,$$

where $U = A_n$ is upper-triangular. For example, when $n = 4$, we have $E_3 P_3 E_2 P_2 E_1 P_1 A = U$. We can define new matrices E_1', E_2', E_3' which are still products of elementary matrices so that we have

$$E_3' E_2' E_1' P_3 P_2 P_1 A = U.$$

Indeed, if we let $E_3' = E_3$, $E_2' = P_3 E_2 P_3^{-1}$, and $E_1' = P_3 P_2 E_1 P_2^{-1} P_3^{-1}$, we easily verify that each E_k' is a product of elementary matrices and that

$$E_3' E_2' E_1' P_3 P_2 P_1 = E_3 (P_3 E_2 P_3^{-1})(P_3 P_2 E_1 P_2^{-1} P_3^{-1}) P_3 P_2 P_1 = E_3 P_3 E_2 P_2 E_1 P_1.$$

It can also be proved that E_1', E_2', E_3' are lower triangular (see Theorem 5.5).

In general, we let

$$E_k' = P_{n-1} \cdots P_{k+1} E_k P_{k+1}^{-1} \cdots P_{n-1}^{-1},$$

and we have

$$E_{n-1}' \cdots E_1' P_{n-1} \cdots P_1 A = U,$$

where each E'_j is a lower triangular matrix (see Theorem 5.5).

It is remarkable that if pivoting steps are necessary during Gaussian elimination, a very simple modification of the algorithm for finding an LU-factorization yields the matrices L, U, and P, such that $PA = LU$. To describe this new method, since the diagonal entries of L are 1s, it is convenient to write

$$L = I + \Lambda.$$

Then, in assembling the matrix Λ while performing Gaussian elimination with pivoting, we make the same transposition on the rows of Λ (really Λ_{k-1}) that we make on the rows of A (really A_k) during a pivoting step involving row k and row i. We also assemble P by starting with the identity matrix and applying to P the same row transpositions that we apply to A and Λ. Here is an example illustrating this method. Given

$$A = A_1 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & -1 \end{pmatrix},$$

we have the following sequence of steps: We initialize $\Lambda_0 = 0$ and $P_0 = I_4$. The first pivot is $\pi_1 = 1$ in row 1, and we substract row 1 from rows 2, 3, and 4. We get

$$A_2 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & -2 & 0 \\ 0 & -2 & -1 & 1 \\ 0 & -2 & -1 & -1 \end{pmatrix} \quad \Lambda_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \quad P_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The next pivot is $\pi_2 = -2$ in row 3, so we permute row 2 and 3; we also apply this permutation to Λ and P:

$$A'_3 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & -2 & -1 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & -2 & -1 & -1 \end{pmatrix} \quad \Lambda'_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \quad P_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Next, we subtract row 2 from row 4 (and add 0 times row 2 to row 3). We get

$$A_3 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & -2 & -1 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -2 \end{pmatrix} \quad \Lambda_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix} \quad P_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The next pivot is $\pi_3 = -2$ in row 3, and since the fourth entry in column 3 is already a zero, we add 0 times row 3 to row 4. We get

$$A_4 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & -2 & -1 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -2 \end{pmatrix} \quad \Lambda_3 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix} \quad P_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The procedure is finished, and we have

$$L = \Lambda_3 + I = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix} \quad U = A_4 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & -2 & -1 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -2 \end{pmatrix} \quad P = P_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

It is easy to check that indeed

$$LU = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & -2 & -1 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & -1 \end{pmatrix}$$

and

$$PA = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & -1 \end{pmatrix}.$$

Using the idea in the remark before the above example, we can prove the theorem below which shows the correctness of the algorithm for computing P, L and U using a simple adaptation of Gaussian elimination.

We are not aware of a detailed proof of Theorem 5.5 in the standard texts. Although Golub and Van Loan [49] state a version of this theorem as their Theorem 3.1.4, they say that "The proof is a messy subscripting argument." Meyer [74] also provides a sketch of proof (see the end of Section 3.10). In view of this situation, we offer a complete proof. It does involve a lot of subscripts and superscripts, but in our opinion, it contains some interesting techniques that go far beyond symbol manipulation.

Theorem 5.5. *For every invertible $n \times n$-matrix A, the following hold:*

(1) There is some permutation matrix P, some upper-triangular matrix U, and some unit lower-triangular matrix L, so that $PA = LU$ (recall, $L_{ii} = 1$ for $i = 1, \ldots, n$). Furthermore, if $P = I$, then L and U are unique and they are produced as a result of Gaussian elimination without pivoting.

(2) If $E_{n-1} \ldots E_1 A = U$ is the result of Gaussian elimination without pivoting, write as usual $A_k = E_{k-1} \ldots E_1 A$ (with $A_k = (a_{ij}^{(k)})$), and let $\ell_{ik} = a_{ik}^{(k)}/a_{kk}^{(k)}$, with $1 \leq k \leq n-1$ and $k+1 \leq i \leq n$. Then

$$L = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \ell_{21} & 1 & 0 & \cdots & 0 \\ \ell_{31} & \ell_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \ell_{n1} & \ell_{n2} & \ell_{n3} & \cdots & 1 \end{pmatrix},$$

where the kth column of L is the kth column of E_k^{-1}, for $k = 1, \ldots, n-1$.

(3) If $E_{n-1}P_{n-1}\cdots E_1P_1A = U$ is the result of Gaussian elimination with some pivoting, write $A_k = E_{k-1}P_{k-1}\cdots E_1P_1A$, and define E_j^k, with $1 \leq j \leq n-1$ and $j \leq k \leq n-1$, such that, for $j = 1, \ldots, n-2$,

$$\begin{aligned} E_j^j &= E_j \\ E_j^k &= P_k E_j^{k-1} P_k, \quad \textit{for } k = j+1, \ldots, n-1, \end{aligned}$$

and

$$E_{n-1}^{n-1} = E_{n-1}.$$

Then,

$$\begin{aligned} E_j^k &= P_k P_{k-1} \cdots P_{j+1} E_j P_{j+1} \cdots P_{k-1} P_k \\ U &= E_{n-1}^{n-1} \cdots E_1^{n-1} P_{n-1} \cdots P_1 A, \end{aligned}$$

and if we set

$$\begin{aligned} P &= P_{n-1} \cdots P_1 \\ L &= (E_1^{n-1})^{-1} \cdots (E_{n-1}^{n-1})^{-1}, \end{aligned}$$

then

$$PA = LU.$$

Furthermore,

$$(E_j^k)^{-1} = I + \mathcal{E}_j^k, \quad 1 \leq j \leq n-1,\ j \leq k \leq n-1,$$

where $\mathcal{E}_j^k$ is a lower triangular matrix of the form

$$\mathcal{E}_j^k = \begin{pmatrix} 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & \ell_{j+1j}^{(k)} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \ell_{nj}^{(k)} & 0 & \cdots & 0 \end{pmatrix},$$

we have

$$E_j^k = I - \mathcal{E}_j^k,$$

and

$$\mathcal{E}_j^k = P_k \mathcal{E}_j^{k-1}, \quad 1 \leq j \leq n-2,\ j+1 \leq k \leq n-1,$$

where $P_k = I$ or else $P_k = P(k, i)$ for some i such that $k+1 \leq i \leq n$; if $P_k \neq I$, this means that $(E_j^k)^{-1}$ is obtained from $(E_j^{k-1})^{-1}$ by permuting the entries on row i and

k in column j. Because the matrices $(E_j^k)^{-1}$ are all lower triangular, the matrix L is also lower triangular.

In order to find L, define lower triangular matrices Λ_k of the form

$$\Lambda_k = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & 0 \\ \lambda_{21}^{(k)} & 0 & 0 & 0 & 0 & \vdots & \vdots & 0 \\ \lambda_{31}^{(k)} & \lambda_{32}^{(k)} & \ddots & 0 & 0 & \vdots & \vdots & 0 \\ \vdots & \vdots & \ddots & 0 & 0 & \vdots & \vdots & \vdots \\ \lambda_{k+11}^{(k)} & \lambda_{k+12}^{(k)} & \cdots & \lambda_{k+1k}^{(k)} & 0 & \cdots & \cdots & 0 \\ \lambda_{k+21}^{(k)} & \lambda_{k+22}^{(k)} & \cdots & \lambda_{k+2k}^{(k)} & 0 & \ddots & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda_{n1}^{(k)} & \lambda_{n2}^{(k)} & \cdots & \lambda_{nk}^{(k)} & 0 & \cdots & \cdots & 0 \end{pmatrix}$$

to assemble the columns of L iteratively as follows: let

$$(-\ell_{k+1k}^{(k)}, \ldots, -\ell_{nk}^{(k)})$$

be the last $n-k$ elements of the kth column of E_k, and define Λ_k inductively by setting

$$\Lambda_1 = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ \ell_{21}^{(1)} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \ell_{n1}^{(1)} & 0 & \cdots & 0 \end{pmatrix},$$

then for $k = 2, \ldots, n-1$, define

$$\Lambda'_k = P_k \Lambda_{k-1},$$

and

$$\Lambda_k = (I + \Lambda'_k)E_k^{-1} - I = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & 0 \\ \lambda_{21}^{\prime(k-1)} & 0 & 0 & 0 & 0 & \vdots & \vdots & 0 \\ \lambda_{31}^{\prime(k-1)} & \lambda_{32}^{\prime(k-1)} & \ddots & 0 & 0 & \vdots & \vdots & 0 \\ \vdots & \vdots & \ddots & 0 & 0 & \vdots & \vdots & \vdots \\ \lambda_{k1}^{\prime(k-1)} & \lambda_{k2}^{\prime(k-1)} & \cdots & \lambda_{k\,(k-1)}^{\prime(k-1)} & 0 & \cdots & \cdots & 0 \\ \lambda_{k+11}^{\prime(k-1)} & \lambda_{k+12}^{\prime(k-1)} & \cdots & \lambda_{k+1\,(k-1)}^{\prime(k-1)} & \ell_{k+1k}^{(k)} & \ddots & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda_{n1}^{\prime(k-1)} & \lambda_{n2}^{\prime(k-1)} & \cdots & \lambda_{n\,k-1}^{\prime(k-1)} & \ell_{nk}^{(k)} & \cdots & \cdots & 0 \end{pmatrix},$$

with $P_k = I$ or $P_k = P(k, i)$ for some $i > k$. This means that in assembling L, row k and row i of Λ_{k-1} need to be permuted when a pivoting step permuting row k and row i of A_k is required. Then

$$I + \Lambda_k = (E_1^k)^{-1} \cdots (E_k^k)^{-1}$$
$$\Lambda_k = \mathcal{E}_1^k \cdots \mathcal{E}_k^k,$$

for $k = 1, \ldots, n-1$, and therefore

$$L = I + \Lambda_{n-1}.$$

Proof. (1) The only part that has not been proved is the uniqueness part (when $P = I$). Assume that A is invertible and that $A = L_1U_1 = L_2U_2$, with L_1, L_2 unit lower-triangular and U_1, U_2 upper-triangular. Then, we have

$$L_2^{-1}L_1 = U_2U_1^{-1}.$$

However, it is obvious that L_2^{-1} is lower-triangular and that U_1^{-1} is upper-triangular, and so $L_2^{-1}L_1$ is lower-triangular and $U_2U_1^{-1}$ is upper-triangular. Since the diagonal entries of L_1 and L_2 are 1, the above equality is only possible if $U_2U_1^{-1} = I$, that is, $U_1 = U_2$, and so $L_1 = L_2$.

(2) When $P = I$, we have $L = E_1^{-1}E_2^{-1} \cdots E_{n-1}^{-1}$, where E_k is the product of $n - k$ elementary matrices of the form $E_{i,k;-\ell_i}$, where $E_{i,k;-\ell_i}$ subtracts ℓ_i times row k from row i, with $\ell_{ik} = a_{ik}^{(k)}/a_{kk}^{(k)}$, $1 \leq k \leq n-1$, and $k+1 \leq i \leq n$. Then, it is immediately verified that

$$E_k = \begin{pmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & -\ell_{k+1k} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & -\ell_{nk} & 0 & \cdots & 1 \end{pmatrix},$$

and that

$$E_k^{-1} = \begin{pmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & \ell_{k+1k} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \ell_{nk} & 0 & \cdots & 1 \end{pmatrix}.$$

If we define L_k by

$$L_k = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & \vdots & 0 \\ \ell_{21} & 1 & 0 & 0 & 0 & \vdots & 0 \\ \ell_{31} & \ell_{32} & \ddots & 0 & 0 & \vdots & 0 \\ \vdots & \vdots & \ddots & 1 & 0 & \vdots & 0 \\ \ell_{k+11} & \ell_{k+12} & \cdots & \ell_{k+1k} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & 0 & \vdots & 0 \\ \ell_{n1} & \ell_{n2} & \cdots & \ell_{nk} & 0 & \cdots & 1 \end{pmatrix}$$

for $k = 1, \ldots, n-1$, we easily check that $L_1 = E_1^{-1}$, and that

$$L_k = L_{k-1}E_k^{-1}, \quad 2 \leq k \leq n-1,$$

because multiplication on the right by E_k^{-1} adds ℓ_i times column i to column k (of the matrix L_{k-1}) with $i > k$, and column i of L_{k-1} has only the nonzero entry 1 as its ith element. Since

$$L_k = E_1^{-1} \cdots E_k^{-1}, \quad 1 \leq k \leq n-1,$$

we conclude that $L = L_{n-1}$, proving our claim about the shape of L.

(3) First, we prove by induction on k that

$$A_{k+1} = E_k^k \cdots E_1^k P_k \cdots P_1 A, \quad k = 1, \ldots, n-2.$$

For $k = 1$, we have $A_2 = E_1 P_1 A = E_1^1 P_1 A$, since $E_1^1 = E_1$, so our assertion holds trivially.

Now, if $k \geq 2$,

$$A_{k+1} = E_k P_k A_k,$$

and by the induction hypothesis,

$$A_k = E_{k-1}^{k-1} \cdots E_2^{k-1} E_1^{k-1} P_{k-1} \cdots P_1 A.$$

Because P_k is either the identity or a transposition, $P_k^2 = I$, so by inserting occurrences of $P_k P_k$ as indicated below we can write

$$\begin{aligned} A_{k+1} &= E_k P_k A_k \\ &= E_k P_k E_{k-1}^{k-1} \cdots E_2^{k-1} E_1^{k-1} P_{k-1} \cdots P_1 A \\ &= E_k P_k E_{k-1}^{k-1} (P_k P_k) \cdots (P_k P_k) E_2^{k-1} (P_k P_k) E_1^{k-1} (P_k P_k) P_{k-1} \cdots P_1 A \\ &= E_k (P_k E_{k-1}^{k-1} P_k) \cdots (P_k E_2^{k-1} P_k)(P_k E_1^{k-1} P_k) P_k P_{k-1} \cdots P_1 A. \end{aligned}$$

Observe that P_k has been "moved" to the right of the elimination steps. However, by definition,

$$\begin{aligned} E_j^k &= P_k E_j^{k-1} P_k, \quad j = 1, \ldots, k-1 \\ E_k^k &= E_k, \end{aligned}$$

so we get

$$A_{k+1} = E_k^k E_{k-1}^k \cdots E_2^k E_1^k P_k \cdots P_1 A,$$

establishing the induction hypothesis. For $k = n-2$, we get

$$U = A_{n-1} = E_{n-1}^{n-1} \cdots E_1^{n-1} P_{n-1} \cdots P_1 A,$$

as claimed, and the factorization $PA = LU$ with

$$\begin{aligned} P &= P_{n-1} \cdots P_1 \\ L &= (E_1^{n-1})^{-1} \cdots (E_{n-1}^{n-1})^{-1} \end{aligned}$$

is clear,

Since for $j = 1, \ldots, n-2$, we have $E_j^j = E_j$,

$$E_j^k = P_k E_j^{k-1} P_k, \quad k = j+1, \ldots, n-1,$$

since $E_{n-1}^{n-1} = E_{n-1}$ and $P_k^{-1} = P_k$, we get $(E_j^j)^{-1} = E_j^{-1}$ for $j = 1, \ldots, n-1$, and for $j = 1, \ldots, n-2$, we have

$$(E_j^k)^{-1} = P_k (E_j^{k-1})^{-1} P_k, \quad k = j+1, \ldots, n-1.$$

Since

$$(E_j^{k-1})^{-1} = I + \mathcal{E}_j^{k-1}$$

and $P_k = P(k,i)$ is a transposition, $P_k^2 = I$, so we get

$$(E_j^k)^{-1} = P_k (E_j^{k-1})^{-1} P_k = P_k (I + \mathcal{E}_j^{k-1}) P_k = P_k^2 + P_k \, \mathcal{E}_j^{k-1} \, P_k = I + P_k \, \mathcal{E}_j^{k-1} \, P_k.$$

Therfore, we have

$$(E_j^k)^{-1} = I + P_k \, \mathcal{E}_j^{k-1} \, P_k, \quad 1 \leq j \leq n-2, \; j+1 \leq k \leq n-1.$$

We prove for $j = 1, \ldots, n-1$, that for $k = j, \ldots, n-1$, each $\mathcal{E}_j^k$ is a lower triangular matrix of the form

$$\mathcal{E}_j^k = \begin{pmatrix} 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & \ell_{j+1j}^{(k)} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \ell_{nj}^{(k)} & 0 & \cdots & 0 \end{pmatrix},$$

and that

$$\mathcal{E}_j^k = P_k \, \mathcal{E}_j^{k-1}, \quad 1 \leq j \leq n-2, \; j+1 \leq k \leq n-1,$$

with $P_k = I$ or $P_k = P(k,i)$ for some i such that $k+1 \leq i \leq n$.

For each j $(1 \leq j \leq n-1)$ we proceed by induction on $k = j, \ldots, n-1$. Since $(E_j^j)^{-1} = E_j^{-1}$ and since E_j^{-1} is of the above form, the base case holds.

For the induction step, we only need to consider the case where $P_k = P(k,i)$ is a transposition, since the case where $P_k = I$ is trivial. We have to figure out what $P_k\, \mathcal{E}_j^{k-1}\, P_k = P(k,i)\, \mathcal{E}_j^{k-1}\, P(k,i)$ is. However, since

$$\mathcal{E}_j^{k-1} = \begin{pmatrix} 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & \ell_{j+1j}^{(k-1)} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \ell_{nj}^{(k-1)} & 0 & \cdots & 0 \end{pmatrix},$$

and because $k+1 \leq i \leq n$ and $j \leq k-1$, multiplying $\mathcal{E}_j^{k-1}$ on the right by $P(k,i)$ will permute *columns* i and k, which are columns of zeros, so

$$P(k,i)\, \mathcal{E}_j^{k-1}\, P(k,i) = P(k,i)\, \mathcal{E}_j^{k-1},$$

and thus,

$$(E_j^k)^{-1} = I + P(k,i)\, \mathcal{E}_j^{k-1},$$

which shows that

$$\mathcal{E}_j^k = P(k,i)\, \mathcal{E}_j^{k-1}.$$

We also know that multiplying $(\mathcal{E}_j^{k-1})^{-1}$ on the left by $P(k,i)$ will permute *rows* i and k, which shows that $\mathcal{E}_j^k$ has the desired form, as claimed. Since all $\mathcal{E}_j^k$ are strictly lower triangular, all $(E_j^k)^{-1} = I + \mathcal{E}_j^k$ are lower triangular, so the product

$$L = (E_1^{n-1})^{-1} \cdots (E_{n-1}^{n-1})^{-1}$$

is also lower triangular.

From the beginning of part (3), we know that

$$L = (E_1^{n-1})^{-1} \cdots (E_{n-1}^{n-1})^{-1}.$$

We prove by induction on k that

$$\begin{aligned} I + \Lambda_k &= (E_1^k)^{-1} \cdots (E_k^k)^{-1} \\ \Lambda_k &= \mathcal{E}_1^k \cdots \mathcal{E}_k^k, \end{aligned}$$

for $k = 1, \ldots, n-1$.

If $k = 1$, we have $E_1^1 = E_1$ and

$$E_1 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -\ell_{21}^{(1)} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -\ell_{n1}^{(1)} & 0 & \cdots & 1 \end{pmatrix}.$$

We get

$$(E_1^{-1})^{-1} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \ell_{21}^{(1)} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \ell_{n1}^{(1)} & 0 & \cdots & 1 \end{pmatrix} = I + \Lambda_1,$$

Since $(E_1^{-1})^{-1} = I + \mathcal{E}_1^1$, we also get $\Lambda_1 = \mathcal{E}_1^1$, and the base step holds.

Since $(E_j^k)^{-1} = I + \mathcal{E}_j^k$ with

$$\mathcal{E}_j^k = \begin{pmatrix} 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & \ell_{j+1j}^{(k)} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \ell_{nj}^{(k)} & 0 & \cdots & 0 \end{pmatrix},$$

as in part (2) for the computation involving the products of L_k's, we get

$$(E_1^{k-1})^{-1} \cdots (E_{k-1}^{k-1})^{-1} = I + \mathcal{E}_1^{k-1} \cdots \mathcal{E}_{k-1}^{k-1}, \quad 2 \leq k \leq n. \tag{$*$}$$

Similarly, from the fact that $\mathcal{E}_j^{k-1} P(k,i) = \mathcal{E}_j^{k-1}$ if $i \geq k+1$ and $j \leq k-1$ and since

$$(E_j^k)^{-1} = I + P_k \mathcal{E}_j^{k-1}, \quad 1 \leq j \leq n-2,\ j+1 \leq k \leq n-1,$$

we get

$$(E_1^k)^{-1} \cdots (E_{k-1}^k)^{-1} = I + P_k \mathcal{E}_1^{k-1} \cdots \mathcal{E}_{k-1}^{k-1}, \quad 2 \leq k \leq n-1. \tag{$**$}$$

By the induction hypothesis,

$$I + \Lambda_{k-1} = (E_1^{k-1})^{-1} \cdots (E_{k-1}^{k-1})^{-1},$$

and from $(*)$, we get

$$\Lambda_{k-1} = \mathcal{E}_1^{k-1} \cdots \mathcal{E}_{k-1}^{k-1}.$$

Using $(**)$, we deduce that

$$(E_1^k)^{-1} \cdots (E_{k-1}^k)^{-1} = I + P_k \Lambda_{k-1}.$$

Since $E_k^k = E_k$, we obtain

$$(E_1^k)^{-1} \cdots (E_{k-1}^k)^{-1}(E_k^k)^{-1} = (I + P_k\Lambda_{k-1})E_k^{-1}.$$

However, by definition

$$I + \Lambda_k = (I + P_k\Lambda_{k-1})E_k^{-1},$$

which proves that

$$I + \Lambda_k = (E_1^k)^{-1} \cdots (E_{k-1}^k)^{-1}(E_k^k)^{-1}, \tag{$\dagger$}$$

and finishes the induction step for the proof of this formula.

If we apply equation $(*)$ again with $k+1$ in place of k, we have

$$(E_1^k)^{-1} \cdots (E_k^k)^{-1} = I + \mathcal{E}_1^k \cdots \mathcal{E}_k^k,$$

and together with $(\dagger)$, we obtain,

$$\Lambda_k = \mathcal{E}_1^k \cdots \mathcal{E}_k^k,$$

also finishing the induction step for the proof of this formula. For $k = n-1$ in $(\dagger)$, we obtain the desired equation: $L = I + \Lambda_{n-1}$. □

We emphasize again that part (3) of Theorem 5.5 shows the remarkable fact that in assembling the matrix L while performing Gaussian elimination with pivoting, the only change to the algorithm is to make the same transposition on the rows of Λ_{k-1} that we make on the rows of A (really A_k) during a pivoting step involving row k and row i. We can also assemble P by starting with the identity matrix and applying to P the same row transpositions that we apply to A and Λ. Here is an example illustrating this method.

Consider the matrix

$$A = \begin{pmatrix} 1 & 2 & -3 & 4 \\ 4 & 8 & 12 & -8 \\ 2 & 3 & 2 & 1 \\ -3 & -1 & 1 & -4 \end{pmatrix}.$$

We set $P_0 = I_4$, and we can also set $\Lambda_0 = 0$. The first step is to permute row 1 and row 2, using the pivot 4. We also apply this permutation to P_0:

$$A_1' = \begin{pmatrix} 4 & 8 & 12 & -8 \\ 1 & 2 & -3 & 4 \\ 2 & 3 & 2 & 1 \\ -3 & -1 & 1 & -4 \end{pmatrix} \quad P_1 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Next, we subtract 1/4 times row 1 from row 2, 1/2 times row 1 from row 3, and add 3/4 times row 1 to row 4, and start assembling Λ:

$$A_2 = \begin{pmatrix} 4 & 8 & 12 & -8 \\ 0 & 0 & -6 & 6 \\ 0 & -1 & -4 & 5 \\ 0 & 5 & 10 & -10 \end{pmatrix} \quad \Lambda_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1/4 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 \\ -3/4 & 0 & 0 & 0 \end{pmatrix} \quad P_1 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Next we permute row 2 and row 4, using the pivot 5. We also apply this permutation to Λ and P:

$$A'_3 = \begin{pmatrix} 4 & 8 & 12 & -8 \\ 0 & 5 & 10 & -10 \\ 0 & -1 & -4 & 5 \\ 0 & 0 & -6 & 6 \end{pmatrix} \quad \Lambda'_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ -3/4 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 \\ 1/4 & 0 & 0 & 0 \end{pmatrix} \quad P_2 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

Next we add $1/5$ times row 2 to row 3, and update Λ'_2:

$$A_3 = \begin{pmatrix} 4 & 8 & 12 & -8 \\ 0 & 5 & 10 & -10 \\ 0 & 0 & -2 & 3 \\ 0 & 0 & -6 & 6 \end{pmatrix} \quad \Lambda_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ -3/4 & 0 & 0 & 0 \\ 1/2 & -1/5 & 0 & 0 \\ 1/4 & 0 & 0 & 0 \end{pmatrix} \quad P_2 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

Next we permute row 3 and row 4, using the pivot -6. We also apply this permutation to Λ and P:

$$A'_4 = \begin{pmatrix} 4 & 8 & 12 & -8 \\ 0 & 5 & 10 & -10 \\ 0 & 0 & -6 & 6 \\ 0 & 0 & -2 & 3 \end{pmatrix} \quad \Lambda'_3 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ -3/4 & 0 & 0 & 0 \\ 1/4 & 0 & 0 & 0 \\ 1/2 & -1/5 & 0 & 0 \end{pmatrix} \quad P_3 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Finally, we subtract $1/3$ times row 3 from row 4, and update Λ'_3:

$$A_4 = \begin{pmatrix} 4 & 8 & 12 & -8 \\ 0 & 5 & 10 & -10 \\ 0 & 0 & -6 & 6 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \Lambda_3 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ -3/4 & 0 & 0 & 0 \\ 1/4 & 0 & 0 & 0 \\ 1/2 & -1/5 & 1/3 & 0 \end{pmatrix} \quad P_3 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Consequently, adding the identity to Λ_3, we obtain

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -3/4 & 1 & 0 & 0 \\ 1/4 & 0 & 1 & 0 \\ 1/2 & -1/5 & 1/3 & 1 \end{pmatrix}, \quad U = \begin{pmatrix} 4 & 8 & 12 & -8 \\ 0 & 5 & 10 & -10 \\ 0 & 0 & -6 & 6 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

We check that

$$PA = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 2 & -3 & 4 \\ 4 & 8 & 12 & -8 \\ 2 & 3 & 2 & 1 \\ -3 & -1 & 1 & -4 \end{pmatrix} = \begin{pmatrix} 4 & 8 & 12 & -8 \\ -3 & -1 & 1 & -4 \\ 1 & 2 & -3 & 4 \\ 2 & 3 & 2 & 1 \end{pmatrix},$$

and that

$$LU = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -3/4 & 1 & 0 & 0 \\ 1/4 & 0 & 1 & 0 \\ 1/2 & -1/5 & 1/3 & 1 \end{pmatrix} \begin{pmatrix} 4 & 8 & 12 & -8 \\ 0 & 5 & 10 & -10 \\ 0 & 0 & -6 & 6 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 4 & 8 & 12 & -8 \\ -3 & -1 & 1 & -4 \\ 1 & 2 & -3 & 4 \\ 2 & 3 & 2 & 1 \end{pmatrix} = PA.$$

Note that if one willing to overwrite the lower triangular part of the evolving matrix A, one can store the evolving Λ there, since these entries will eventually be zero anyway! There is also no need to save explicitly the permutation matrix P. One could instead record the permutation steps in an extra column (record the vector $(\pi(1), \ldots, \pi(n))$ corresponding to the permutation π applied to the rows). We let the reader write such a bold and space-efficient version of LU-decomposition!

As a corollary of Theorem 5.5(1), we can show the following result.

Proposition 5.6. *If an invertible symmetric matrix A has an LU-decomposition, then A has a factorization of the form*

$$A = LDL^\top,$$

where L is a lower-triangular matrix whose diagonal entries are equal to 1, and where D consists of the pivots. Furthermore, such a decomposition is unique.

Proof. If A has an LU-factorization, then it has an LDU factorization

$$A = LDU,$$

where L is lower-triangular, U is upper-triangular, and the diagonal entries of both L and U are equal to 1. Since A is symmetric, we have

$$LDU = A = A^\top = U^\top D L^\top,$$

with $U^\top$ lower-triangular and $DL^\top$ upper-triangular. By the uniqueness of LU-factorization (part (1) of Theorem 5.5), we must have $L = U^\top$ (and $DU = DL^\top$), thus $U = L^\top$, as claimed. □

Remark: It can be shown that Gaussian elimination + back-substitution requires $n^3/3 + O(n^2)$ additions, $n^3/3 + O(n^2)$ multiplications and $n^2/2 + O(n)$ divisions.

5.6 Dealing with Roundoff Errors; Pivoting Strategies

Let us now briefly comment on the choice of a pivot. Although theoretically, any pivot can be chosen, the possibility of roundoff errors implies that it is not a good idea to pick very small pivots. The following example illustrates this point. Consider the linear system

$$\begin{aligned} 10^{-4}x &+ y = 1 \\ x &+ y = 2. \end{aligned}$$

Since 10^{-4} is nonzero, it can be taken as pivot, and we get

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

Thus, the exact solution is

$$x = \frac{10^4}{10^4 - 1}, \quad y = \frac{10^4 - 2}{10^4 - 1}.$$

However, if roundoff takes place on the fourth digit, then $10^4 - 1 = 9999$ and $10^4 - 2 = 9998$ will be rounded off both to 9990, and then the solution is $x = 0$ and $y = 1$, very far from the exact solution where $x \approx 1$ and $y \approx 1$. The problem is that we picked a very small pivot. If instead we permute the equations, the pivot is 1, and after elimination, we get the system

$$\begin{array}{rcl} x + y & = & 2 \\ (1 - 10^{-4})y & = & 1 - 2 \times 10^{-4}. \end{array}$$

This time, $1 - 10^{-4} = 0.9999$ and $1 - 2 \times 10^{-4} = 0.9998$ are rounded off to 0.999 and the solution is $x = 1, y = 1$, much closer to the exact solution.

To remedy this problem, one may use the strategy of *partial pivoting*. This consists of choosing during step k $(1 \leq k \leq n - 1)$ one of the entries $a_{i\,k}^{(k)}$ such that

$$|a_{i\,k}^{(k)}| = \max_{k \leq p \leq n} |a_{p\,k}^{(k)}|.$$

By maximizing the value of the pivot, we avoid dividing by undesirably small pivots.

Remark: A matrix, A, is called *strictly column diagonally dominant* iff

$$|a_{j\,j}| > \sum_{i=1,\, i \neq j}^{n} |a_{i\,j}|, \quad \text{for } j = 1, \ldots, n$$

(resp. *strictly row diagonally dominant* iff

$$|a_{i\,i}| > \sum_{j=1,\, j \neq i}^{n} |a_{i\,j}|, \quad \text{for } i = 1, \ldots, n.)$$

It has been known for a long time (before 1900, say by Hadamard) that if a matrix A is strictly column diagonally dominant (resp. strictly row diagonally dominant), then it is invertible. (This is a good exercise, try it!) It can also be shown that if A is strictly column diagonally dominant, then Gaussian elimination with partial pivoting does not actually require pivoting (See Problem 21.6 in Trefethen and Bau [105], or Question 2.19 in Demmel [33]).

Another strategy, called *complete pivoting*, consists in choosing some entry $a_{i\,j}^{(k)}$, where $k \leq i, j \leq n$, such that

$$|a_{i\,j}^{(k)}| = \max_{k \leq p,q \leq n} |a_{p\,q}^{(k)}|.$$

However, in this method, if the chosen pivot is not in column k, it is also necessary to permute columns. This is achieved by multiplying on the right by a permutation matrix. However, complete pivoting tends to be too expensive in practice, and partial pivoting is the method of choice.

A special case where the LU-factorization is particularly efficient is the case of tridiagonal matrices, which we now consider.

5.7 Gaussian Elimination of Tridiagonal Matrices

Consider the tridiagonal matrix

$$A = \begin{pmatrix} b_1 & c_1 & & & & \\ a_2 & b_2 & c_2 & & & \\ & a_3 & b_3 & c_3 & & \\ & \ddots & \ddots & \ddots & & \\ & & a_{n-2} & b_{n-2} & c_{n-2} & \\ & & & a_{n-1} & b_{n-1} & c_{n-1} \\ & & & & a_n & b_n \end{pmatrix}.$$

Define the sequence

$$\delta_0 = 1, \quad \delta_1 = b_1, \quad \delta_k = b_k\delta_{k-1} - a_k c_{k-1}\delta_{k-2}, \quad 2 \leq k \leq n.$$

Proposition 5.7. *If A is the tridiagonal matrix above, then $\delta_k = \det(A[1..k, 1..k])$ for $k = 1, \ldots, n$.*

Proof. By expanding $\det(A[1..k, 1..k])$ with respect to its last row, the proposition follows by induction on k. □

Theorem 5.8. *If A is the tridiagonal matrix above and $\delta_k \neq 0$ for $k = 1, \ldots, n$, then A has the following LU-factorization:*

$$A = \begin{pmatrix} 1 & & & & & \\ a_2\dfrac{\delta_0}{\delta_1} & 1 & & & & \\ & a_3\dfrac{\delta_1}{\delta_2} & 1 & & & \\ & & \ddots & \ddots & & \\ & & & a_{n-1}\dfrac{\delta_{n-3}}{\delta_{n-2}} & 1 & \\ & & & & a_n\dfrac{\delta_{n-2}}{\delta_{n-1}} & 1 \end{pmatrix} \begin{pmatrix} \dfrac{\delta_1}{\delta_0} & c_1 & & & & \\ & \dfrac{\delta_2}{\delta_1} & c_2 & & & \\ & & \dfrac{\delta_3}{\delta_2} & c_3 & & \\ & & & \ddots & \ddots & \\ & & & & \dfrac{\delta_{n-1}}{\delta_{n-2}} & c_{n-1} \\ & & & & & \dfrac{\delta_n}{\delta_{n-1}} \end{pmatrix}.$$

Proof. Since $\delta_k = \det(A[1..k, 1..k]) \neq 0$ for $k = 1, \ldots, n$, by Theorem 5.5 (and Proposition 5.2), we know that A has a unique LU-factorization. Therefore, it suffices to check that the proposed factorization works. We easily check that

$$\begin{aligned} (LU)_{k\,k+1} &= c_k, \quad 1 \leq k \leq n-1 \\ (LU)_{k\,k-1} &= a_k, \quad 2 \leq k \leq n \\ (LU)_{k\,l} &= 0, \quad |k-l| \geq 2 \\ (LU)_{1\,1} &= \frac{\delta_1}{\delta_0} = b_1 \\ (LU)_{k\,k} &= \frac{a_k c_{k-1}\delta_{k-2} + \delta_k}{\delta_{k-1}} = b_k, \quad 2 \leq k \leq n, \end{aligned}$$

since $\delta_k = b_k\delta_{k-1} - a_k c_{k-1}\delta_{k-2}$. □

It follows that there is a simple method to solve a linear system $Ax = d$ where A is tridiagonal (and $\delta_k \neq 0$ for $k = 1, \ldots, n$). For this, it is convenient to "squeeze" the diagonal matrix Δ defined such that $\Delta_{k\,k} = \delta_k/\delta_{k-1}$ into the factorization so that $A = (L\Delta)(\Delta^{-1}U)$, and if we let

$$z_1 = \frac{c_1}{b_1}, \quad z_k = c_k\,\frac{\delta_{k-1}}{\delta_k}, \quad 2 \leq k \leq n-1, \quad z_n = \frac{\delta_n}{\delta_{n-1}} = b_n - a_n z_{n-1},$$

$A = (L\Delta)(\Delta^{-1}U)$ is written as

$$A = \begin{pmatrix} \frac{c_1}{z_1} & & & & & \\ a_2 & \frac{c_2}{z_2} & & & & \\ & a_3 & \frac{c_3}{z_3} & & & \\ & & \ddots & \ddots & & \\ & & & a_{n-1} & \frac{c_{n-1}}{z_{n-1}} & \\ & & & & a_n & z_n \end{pmatrix} \begin{pmatrix} 1 & z_1 & & & & \\ & 1 & z_2 & & & \\ & & 1 & z_3 & & \\ & & & \ddots & \ddots & \\ & & & 1 & z_{n-2} & \\ & & & & 1 & z_{n-1} \\ & & & & & 1 \end{pmatrix}.$$

As a consequence, the system $Ax = d$ can be solved by constructing three sequences: First, the sequence

$$z_1 = \frac{c_1}{b_1}, \quad z_k = \frac{c_k}{b_k - a_k z_{k-1}}, \quad k = 2, \ldots, n-1, \quad z_n = b_n - a_n z_{n-1},$$

corresponding to the recurrence $\delta_k = b_k\delta_{k-1} - a_k c_{k-1}\delta_{k-2}$ and obtained by dividing both sides of this equation by δ_{k-1}, next

$$w_1 = \frac{d_1}{b_1}, \quad w_k = \frac{d_k - a_k w_{k-1}}{b_k - a_k z_{k-1}}, \quad k = 2, \ldots, n,$$

corresponding to solving the system $L\Delta w = d$, and finally

$$x_n = w_n, \quad x_k = w_k - z_k x_{k+1}, \quad k = n-1, n-2, \ldots, 1,$$

corresponding to solving the system $\Delta^{-1}Ux = w$.

Remark: It can be verified that this requires $3(n-1)$ additions, $3(n-1)$ multiplications, and $2n$ divisions, a total of $8n-6$ operations, which is much less that the $O(2n^3/3)$ required by Gaussian elimination in general.

We now consider the special case of symmetric positive definite matrices (SPD matrices).

5.8 SPD Matrices and the Cholesky Decomposition

Recall that an $n \times n$ symmetric matrix A is *positive definite* iff

$$x^\top Ax > 0 \quad \text{for all } x \in \mathbb{R}^n \text{ with } x \neq 0.$$

Equivalently, A is symmetric positive definite iff all its eigenvalues are strictly positive. The following facts about a symmetric positive definite matrice A are easily established (some left as an exercise):

(1) The matrix A is invertible. (Indeed, if $Ax = 0$, then $x^\top Ax = 0$, which implies $x = 0$.)

(2) We have $a_{ii} > 0$ for $i = 1, \ldots, n$. (Just observe that for $x = e_i$, the ith canonical basis vector of $\mathbb{R}^n$, we have $e_i^\top Ae_i = a_{ii} > 0$.)

(3) For every $n \times n$ invertible matrix Z, the matrix $Z^\top AZ$ is symmetric positive definite iff A is symmetric positive definite.

(4) The set of $n \times n$ symmetric positive definite matrices is convex. This means that if A and B are two $n \times n$ symmetric positive definite matrices, then for any λ such that $0 \leq \lambda \leq 1$, the matrix $(1 - \lambda)A + \lambda B$ is also symmetric positive definite. Clearly since A and B are symmetric, $(1 - \lambda)A + \lambda B$ is also symmetric. For any nonzero $x \in \mathbb{R}^n$, we have $x^\top Ax > 0$ and $x^\top Bx > 0$, so

$$x^\top((1 - \lambda)A + \lambda B)x = (1 - \lambda)x^\top Ax + \lambda x^\top Bx > 0,$$

because $0 \leq \lambda \leq 1$, so $1-\lambda \geq 0$ and $\lambda \geq 0$, and $1-\lambda$ and λ can't be zero simultaneously.

(5) The set of $n \times n$ symmetric positive definite matrices is a cone. This means that if A is symmetric positive definite and if $\lambda > 0$ is any real, then λA is symmetric positive definite. Clearly λA is symmetric, and for nonzero $x \in \mathbb{R}^n$, we have $x^\top Ax > 0$, and since $\lambda > 0$, we have $x^\top \lambda Ax = \lambda x^\top Ax > 0$.

It is instructive to characterize when a 2×2 symmetric matrix A is positive definite. Write

$$A = \begin{pmatrix} a & c \\ c & b \end{pmatrix}.$$

Then we have

$$\begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} a & c \\ c & b \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = ax^2 + 2cxy + by^2.$$

If the above expression is stritcly positive for all nonzero vectors $\binom{x}{y}$, then for $x = 1, y = 0$ we get $a > 0$ and for $x = 0, y = 1$ we get $b > 0$. Then we can write

$$\begin{aligned} ax^2 + 2cxy + by^2 &= \left(\sqrt{a}x + \frac{c}{\sqrt{a}}y\right)^2 + by^2 - \frac{c^2}{a}y^2 \\ &= \left(\sqrt{a}x + \frac{c}{\sqrt{a}}y\right)^2 + \frac{1}{a}\left(ab - c^2\right)y^2. \end{aligned}$$

Since $a > 0$, if $ab - c^2 \leq 0$, then we can choose $y > 0$ so that the second term is negative or zero, and we can set $x = -(c/a)y$ to make the first term zero, in which case $ax^2 + 2cxy + by^2 \leq 0$, so we must have $ab - c^2 > 0$.

Conversely, if $a > 0, b > 0$ and $ab > c^2$, then for any $(x, y) \neq (0, 0)$, if $y = 0$ then $x \neq 0$ and the first term is positive, and if $y \neq 0$ then the second term is positive. Therefore, the symmetric matrix A is positive definite iff

$$a > 0,\ b > 0,\ ab > c^2. \tag{$*$}$$

Note that $ab - c^2 = \det(A)$, so the third condition says that $\det(A) > 0$.

Observe that the condition $b > 0$ is redundant, since if $a > 0$ and $ab > c^2$, then we must have $b > 0$ (and similarly $b > 0$ and $ab > c^2$ implies that $a > 0$).

We can try to visualize the space of 2×2 symmetric positive definite matrices in $\mathbb{R}^3$, by viewing (a, b, c) as the coordinates along the x, y, z axes. Then the locus determined by the strict inequalities in $(*)$ corresponds to the region on the side of the cone of equation $xy = z^2$ that does not contain the origin and for which $x > 0$ and $y > 0$. For $z = \delta$ fixed, the equation $xy = \delta^2$ define a hyperbola in the plane $z = \delta$. The cone of equation $xy = z^2$ consists of the lines through the origin that touch the hyperbola $xy = 1$ in the plane $z = 1$. We only consider the branch of this hyperbola for which $x > 0$ and $y > 0$.

It is not hard to show that the inverse of a symmetric positive definite matrix is also symmetric positive definite, but the product of two symmetric positive definite matrices may *not* be symmetric positive definite, as the following example shows:

$$\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & -1\sqrt{2} \\ -1/\sqrt{2} & 3/\sqrt{2} \end{pmatrix} = \begin{pmatrix} 0 & 2/\sqrt{2} \\ -1/\sqrt{2} & 5/\sqrt{2} \end{pmatrix}.$$

According to the above criterion, the two matrices on the left-hand side are symmetric positive definite, but the matrix on the right-hand side is not even symmetric, and

$$\begin{pmatrix} -6 & 1 \end{pmatrix} \begin{pmatrix} 0 & 2/\sqrt{2} \\ -1/\sqrt{2} & 5/\sqrt{2} \end{pmatrix} \begin{pmatrix} -6 \\ 1 \end{pmatrix} = \begin{pmatrix} -6 & 1 \end{pmatrix} \begin{pmatrix} 2/\sqrt{2} \\ 11/\sqrt{2} \end{pmatrix} = -1/\sqrt{5},$$

even though its eigenvalues are both real and positive.

Next, we prove that a symmetric positive definite matrix has a special LU-factorization of the form $A = BB^\top$, where B is a lower-triangular matrix whose diagonal elements are strictly positive. This is the *Cholesky factorization*.

First, we note that a symmetric positive definite matrix satisfies the condition of Proposition 5.2.

Proposition 5.9. *If A is a symmetric positive definite matrix, then $A[1..k, 1..k]$ is symmetric positive definite, and thus invertible for $k = 1, \ldots, n$.*

Proof. Since A is symmetric, each $A[1..k, 1..k]$ is also symmetric. If $w \in \mathbb{R}^k$, with $1 \leq k \leq n$, we let $x \in \mathbb{R}^n$ be the vector with $x_i = w_i$ for $i = 1, \ldots, k$ and $x_i = 0$ for $i = k+1, \ldots, n$. Now, since A is symmetric positive definite, we have $x^\top Ax > 0$ for all $x \in \mathbb{R}^n$ with $x \neq 0$. This holds in particular for all vectors x obtained from nonzero vectors $w \in \mathbb{R}^k$ as defined earlier, and clearly

$$x^\top Ax = w^\top A[1..k, 1..k]\, w,$$

which implies that $A[1..k, 1..k]$ is positive definite. Thus, $A[1..k, 1..k]$ is also invertible. □

Proposition 5.9 can be strengthened as follows: *A symmetric matrix A is positive definite iff* $\det(A[1..k, 1..k]) > 0$ *for* $k = 1, \ldots, n$.

The above fact is known as *Sylvester's criterion*. We will prove it after establishing the Cholesky factorization.

Let A be an $n \times n$ symmetric positive definite matrix and write

$$A = \begin{pmatrix} a_{11} & W^\top \\ W & C \end{pmatrix},$$

where C is an $(n-1) \times (n-1)$ symmetric matrix and W is an $(n-1) \times 1$ matrix. Since A is symmetric positive definite, $a_{11} > 0$, and we can compute $\alpha = \sqrt{a_{11}}$. The trick is that we can factor A uniquely as

$$A = \begin{pmatrix} a_{11} & W^\top \\ W & C \end{pmatrix} = \begin{pmatrix} \alpha & 0 \\ W/\alpha & I \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & C - WW^\top/a_{11} \end{pmatrix} \begin{pmatrix} \alpha & W^\top/\alpha \\ 0 & I \end{pmatrix},$$

i.e., as $A = B_1 A_1 B_1^\top$, where B_1 is lower-triangular with positive diagonal entries. Thus, B_1 is invertible, and by fact (3) above, A_1 is also symmetric positive definite.

Remark: The matrix $C - WW^\top/a_{11}$ is known as the *Schur complement* of the matrix (a_{11}).

Theorem 5.10. *(Cholesky Factorization) Let A be a symmetric positive definite matrix. Then, there is some lower-triangular matrix B so that $A = BB^\top$. Furthermore, B can be chosen so that its diagonal elements are strictly positive, in which case B is unique.*

Proof. We proceed by induction on the dimension n of A. For $n = 1$, we must have $a_{11} > 0$, and if we let $\alpha = \sqrt{a_{11}}$ and $B = (\alpha)$, the theorem holds trivially. If $n \geq 2$, as we explained above, again we must have $a_{11} > 0$, and we can write

$$A = \begin{pmatrix} a_{11} & W^\top \\ W & C \end{pmatrix} = \begin{pmatrix} \alpha & 0 \\ W/\alpha & I \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & C - WW^\top/a_{11} \end{pmatrix} \begin{pmatrix} \alpha & W^\top/\alpha \\ 0 & I \end{pmatrix} = B_1 A_1 B_1^\top,$$

where $\alpha = \sqrt{a_{11}}$, the matrix B_1 is invertible and

$$A_1 = \begin{pmatrix} 1 & 0 \\ 0 & C - WW^\top/a_{11} \end{pmatrix}$$

is symmetric positive definite. However, this implies that $C - WW^\top/a_{11}$ is also symmetric positive definite (consider $x^\top A_1 x$ for every $x \in \mathbb{R}^n$ with $x \neq 0$ and $x_1 = 0$). Thus, we can apply the induction hypothesis to $C - WW^\top/a_{11}$ (which is an $(n-1) \times (n-1)$ matrix), and we find a unique lower-triangular matrix L with positive diagonal entries so that

$$C - WW^\top/a_{11} = LL^\top.$$

But then, we get

$$\begin{aligned} A &= \begin{pmatrix} \alpha & 0 \\ W/\alpha & I \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & C - WW^\top/a_{11} \end{pmatrix} \begin{pmatrix} \alpha & W^\top/\alpha \\ 0 & I \end{pmatrix} \\ &= \begin{pmatrix} \alpha & 0 \\ W/\alpha & I \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & LL^\top \end{pmatrix} \begin{pmatrix} \alpha & W^\top/\alpha \\ 0 & I \end{pmatrix} \\ &= \begin{pmatrix} \alpha & 0 \\ W/\alpha & I \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & L \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & L^\top \end{pmatrix} \begin{pmatrix} \alpha & W^\top/\alpha \\ 0 & I \end{pmatrix} \\ &= \begin{pmatrix} \alpha & 0 \\ W/\alpha & L \end{pmatrix} \begin{pmatrix} \alpha & W^\top/\alpha \\ 0 & L^\top \end{pmatrix}. \end{aligned}$$

Therefore, if we let

$$B = \begin{pmatrix} \alpha & 0 \\ W/\alpha & L \end{pmatrix},$$

we have a unique lower-triangular matrix with positive diagonal entries and $A = BB^\top$.

The uniqueness of the Cholesky decomposition can also be established using the uniqueness of an LU-decomposition. Indeed, if $A = B_1 B_1^\top = B_2 B_2^\top$ where B_1 and B_2 are lower triangular with positive diagonal entries, if we let Δ_1 (resp. Δ_2) be the diagonal matrix consisting of the diagonal entries of B_1 (resp. B_2) so that $(\Delta_k)_{ii} = (B_k)_{ii}$ for $k = 1, 2$, then we have two LU-decompositions

$$A = (B_1 \Delta_1^{-1})(\Delta_1 B_1^\top) = (B_2 \Delta_2^{-1})(\Delta_2 B_2^\top)$$

with $B_1\Delta_1^{-1}, B_2\Delta_2^{-1}$ unit lower triangular, and $\Delta_1 B_1^\top, \Delta_2 B_2^\top$ upper triangular. By uniquenes of LU-factorization (Theorem 5.5(1)), we have

$$B_1\Delta_1^{-1} = B_2\Delta_2^{-1}, \quad \Delta_1 B_1^\top = \Delta_2 B_2^\top,$$

and the second equation yields

$$B_1\Delta_1 = B_2\Delta_2. \tag{$*$}$$

The diagonal entries of $B_1\Delta_1$ are $(B_1)_{ii}^2$ and similarly the diagonal entries of $B_2\Delta_2$ are $(B_2)_{ii}^2$, so the above equation implies that

$$(B_1)_{ii}^2 = (B_2)_{ii}^2, \quad i = 1, \ldots, n.$$

Since the diagonal entries of both B_1 and B_2 are assumed to be positive, we must have

$$(B_1)_{ii} = (B_2)_{ii}, \quad i = 1, \ldots, n;$$

that is, $\Delta_1 = \Delta_2$, and since both are invertible, we conclude from $(*)$ that $B_1 = B_2$. □

The proof of Theorem 5.10 immediately yields an algorithm to compute B from A by solving for a lower triangular matrix B such that $A = BB^\top$. For $j = 1, \ldots, n$,

$$b_{j\,j} = \left(a_{j\,j} - \sum_{k=1}^{j-1} b_{j\,k}^2 \right)^{1/2},$$

and for $i = j+1, \ldots, n$ (and $j = 1, \ldots, n-1$)

$$b_{i\,j} = \left(a_{i\,j} - \sum_{k=1}^{j-1} b_{i\,k} b_{j\,k} \right) / b_{j\,j}.$$

The above formulae are used to compute the jth column of B from top-down, using the first $j-1$ columns of B previously computed, and the matrix A.

For example, if

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 & 3 & 3 \\ 1 & 2 & 3 & 4 & 4 & 4 \\ 1 & 2 & 3 & 4 & 5 & 5 \\ 1 & 2 & 3 & 4 & 5 & 6 \end{pmatrix},$$

we find that

$$B = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

The Cholesky factorization can be used to solve linear systems $Ax = b$ where A is symmetric positive definite: Solve the two systems $Bw = b$ and $B^\top x = w$.

Remark: It can be shown that this methods requires $n^3/6 + O(n^2)$ additions, $n^3/6 + O(n^2)$ multiplications, $n^2/2 + O(n)$ divisions, and $O(n)$ square root extractions. Thus, the Cholesky method requires half of the number of operations required by Gaussian elimination (since Gaussian elimination requires $n^3/3 + O(n^2)$ additions, $n^3/3 + O(n^2)$ multiplications, and $n^2/2 + O(n)$ divisions). It also requires half of the space (only B is needed, as opposed to both L and U). Furthermore, it can be shown that Cholesky's method is numerically stable (see Trefethen and Bau [105], Lecture 23).

Remark: If $A = BB^\top$, where B is any invertible matrix, then A is symmetric positive definite.

Proof. Obviously, $BB^\top$ is symmetric, and since B is invertible, $B^\top$ is invertible, and from

$$x^\top Ax = x^\top BB^\top x = (B^\top x)^\top B^\top x,$$

it is clear that $x^\top Ax > 0$ if $x \neq 0$. □

We now give three more criteria for a symmetric matrix to be positive definite.

Proposition 5.11. *Let A be any $n \times n$ symmetric matrix. The following conditions are equivalent:*

(a) A is positive definite.

(b) All principal minors of A are positive; that is: $\det(A[1..k, 1..k]) > 0$ for $k = 1, \ldots, n$ (Sylvester's criterion).

(c) A has an LU-factorization and all pivots are positive.

(d) A has an $LDL^\top$-factorization and all pivots in D are positive.

Proof. By Proposition 5.9, if A is symmetric positive definite, then each matrix $A[1..k, 1..k]$ is symmetric positive definite for $k = 1, \ldots, n$. By the Cholsesky decomposition, $A[1..k, 1..k] = Q^\top Q$ for some invertible matrix Q, so $\det(A[1..k, 1..k]) = \det(Q)^2 > 0$. This shows that (a) implies (b).

If $\det(A[1..k, 1..k]) > 0$ for $k = 1, \ldots, n$, then each $A[1..k, 1..k]$ is invertible. By Proposition 5.2, the matrix A has an LU-factorization, and since the pivots π_k are given by

$$\pi_k = \begin{cases} a_{11} = \det(A[1..1, 1..1]) & \text{if } k = 1 \\ \dfrac{\det(A[1..k, 1..k])}{\det(A[1..k-1, 1..k-1])} & \text{if } k = 2, \ldots, n, \end{cases}$$

we see that $\pi_k > 0$ for $k = 1, \ldots, n$. Thus (b) implies (c).

Assume A has an LU-factorization and that the pivots are all positive. Since A is symmetric, this implies that A has a factorization of the form

$$A = LDL^\top,$$

with L lower-triangular with 1's on its diagonal, and where D is a diagonal matrix with positive entries on the diagonal (the pivots). This shows that (c) implies (d).

Given a factorization $A = LDL^\top$ with all pivots in D positive, if we form the diagonal matrix

$$\sqrt{D} = \mathrm{diag}(\sqrt{\pi_1}, \ldots, \sqrt{\pi_n})$$

and if we let $B = L\sqrt{D}$, then we have

$$A = BB^\top,$$

with B lower-triangular and invertible. By the remark before Proposition 5.11, A is positive definite. Hence, (d) implies (a). □

Criterion (c) yields a simple computational test to check whether a symmetric matrix is positive definite. There is one more criterion for a symmetric matrix to be positive definite: its eigenvalues must be positive. We will have to learn about the spectral theorem for symmetric matrices to establish this criterion.

For more on the stability analysis and efficient implementation methods of Gaussian elimination, LU-factoring and Cholesky factoring, see Demmel [33], Trefethen and Bau [105], Ciarlet [30], Golub and Van Loan [49], Meyer [74], Strang [101, 102], and Kincaid and Cheney [59].

5.9 Reduced Row Echelon Form (RREF)

Gaussian elimination described in Section 5.2 can also be applied to rectangular matrices. This yields a method for determining whether a system $Ax = b$ is solvable, and a description of all the solutions when the system is solvable, for any rectangular $m \times n$ matrix A.

It turns out that the discussion is simpler if we rescale all pivots to be 1, and for this we need a third kind of elementary matrix. For any $\lambda \neq 0$, let $E_{i,\lambda}$ be the $n \times n$ diagonal matrix

$$E_{i,\lambda} = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & \lambda & & & \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{pmatrix},$$

with $(E_{i,\lambda})_{ii} = \lambda$ $(1 \leq i \leq n)$. Note that $E_{i,\lambda}$ is also given by

$$E_{i,\lambda} = I + (\lambda - 1)e_{i\,i},$$

and that $E_{i,\lambda}$ is invertible with

$$E_{i,\lambda}^{-1} = E_{i,\lambda^{-1}}.$$

Now, after $k-1$ elimination steps, if the bottom portion

$$(a_{kk}^{(k)}, a_{k+1k}^{(k)}, \ldots, a_{mk}^{(k)})$$

of the kth column of the current matrix A_k is nonzero so that a pivot π_k can be chosen, after a permutation of rows if necessary, we also divide row k by π_k to obtain the pivot 1, and not only do we zero all the entries $i = k+1, \ldots, m$ in column k, but also all the entries $i = 1, \ldots, k-1$, so that the only nonzero entry in column k is a 1 in row k. These row operations are achieved by multiplication on the left by elementary matrices.

If $a_{kk}^{(k)} = a_{k+1k}^{(k)} = \cdots = a_{mk}^{(k)} = 0$, we move on to column $k+1$.

When the kth column contains a pivot, the kth stage of the procedure for converting a matrix to *rref* consists of the following three steps illustrated below:

$$\begin{pmatrix} 1 & \times & 0 & \times & \times & \times & \times \\ 0 & 0 & 1 & \times & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & 0 & a_{ik}^{(k)} & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times & \times \end{pmatrix} \text{pivot} \Longrightarrow \begin{pmatrix} 1 & \times & 0 & \times & \times & \times & \times \\ 0 & 0 & 1 & \times & \times & \times & \times \\ 0 & 0 & 0 & a_{ik}^{(k)} & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times & \times \end{pmatrix} \text{rescale} \Longrightarrow$$

$$\begin{pmatrix} 1 & \times & 0 & \times & \times & \times & \times \\ 0 & 0 & 1 & \times & \times & \times & \times \\ 0 & 0 & 0 & 1 & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times & \times \end{pmatrix} \text{elim} \Longrightarrow \begin{pmatrix} 1 & \times & 0 & \mathbf{0} & \times & \times & \times \\ 0 & 0 & 1 & \mathbf{0} & \times & \times & \times \\ 0 & 0 & 0 & \mathbf{1} & \times & \times & \times \\ 0 & 0 & 0 & \mathbf{0} & \times & \times & \times \\ 0 & 0 & 0 & \mathbf{0} & \times & \times & \times \\ 0 & 0 & 0 & \mathbf{0} & \times & \times & \times \end{pmatrix}.$$

If the kth column does not contain a pivot, we simply move on to the next column.

The result is that after performing such elimination steps, we obtain a matrix that has a special shape known as a *reduced row echelon matrix*, for short *rref*.

Here is an example illustrating this process: Starting from the matrix

$$A_1 = \begin{pmatrix} 1 & 0 & 2 & 1 & 5 \\ 1 & 1 & 5 & 2 & 7 \\ 1 & 2 & 8 & 4 & 12 \end{pmatrix}$$

we perform the following steps

$$A_1 \longrightarrow A_2 = \begin{pmatrix} 1 & 0 & 2 & 1 & 5 \\ 0 & 1 & 3 & 1 & 2 \\ 0 & 2 & 6 & 3 & 7 \end{pmatrix},$$

by subtracting row 1 from row 2 and row 3;

$$A_2 \longrightarrow \begin{pmatrix} 1 & 0 & 2 & 1 & 5 \\ 0 & 2 & 6 & 3 & 7 \\ 0 & 1 & 3 & 1 & 2 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 0 & 2 & 1 & 5 \\ 0 & 1 & 3 & 3/2 & 7/2 \\ 0 & 1 & 3 & 1 & 2 \end{pmatrix} \longrightarrow A_3 = \begin{pmatrix} 1 & 0 & 2 & 1 & 5 \\ 0 & 1 & 3 & 3/2 & 7/2 \\ 0 & 0 & 0 & -1/2 & -3/2 \end{pmatrix},$$

after choosing the pivot 2 and permuting row 2 and row 3, dividing row 2 by 2, and subtracting row 2 from row 3;

$$A_3 \longrightarrow \begin{pmatrix} 1 & 0 & 2 & 1 & 5 \\ 0 & 1 & 3 & 3/2 & 7/2 \\ 0 & 0 & 0 & 1 & 3 \end{pmatrix} \longrightarrow A_4 = \begin{pmatrix} 1 & 0 & 2 & 0 & 2 \\ 0 & 1 & 3 & 0 & -1 \\ 0 & 0 & 0 & 1 & 3 \end{pmatrix},$$

after dividing row 3 by $-1/2$, subtracting row 3 from row 1, and subtracting $(3/2) \times$ row 3 from row 2.

It is clear that columns $1, 2$ and 4 are linearly independent, that column 3 is a linear combination of columns 1 and 2, and that column 5 is a linear combinations of columns $1, 2, 4$.

In general, the sequence of steps leading to a reduced echelon matrix is not unique. For example, we could have chosen 1 instead of 2 as the second pivot in matrix A_2. Nevertheless, the reduced row echelon matrix obtained from any given matrix is unique; that is, it does not depend on the the sequence of steps that are followed during the reduction process. This fact is not so easy to prove rigorously, but we will do it later.

If we want to solve a linear system of equations of the form $Ax = b$, we apply elementary row operations to both the matrix A and the right-hand side b. To do this conveniently, we form the *augmented matrix* (A, b), which is the $m \times (n+1)$ matrix obtained by adding b as an extra column to the matrix A. For example if

$$A = \begin{pmatrix} 1 & 0 & 2 & 1 \\ 1 & 1 & 5 & 2 \\ 1 & 2 & 8 & 4 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 5 \\ 7 \\ 12 \end{pmatrix},$$

then the augmented matrix is

$$(A, b) = \begin{pmatrix} 1 & 0 & 2 & 1 & 5 \\ 1 & 1 & 5 & 2 & 7 \\ 1 & 2 & 8 & 4 & 12 \end{pmatrix}.$$

Now, for any matrix M, since

$$M(A, b) = (MA, Mb),$$

performing elementary row operations on (A, b) is equivalent to simultaneously performing operations on both A and b. For example, consider the system

$$\begin{array}{rcrcrcrcl} x_1 & & & + & 2x_3 & + & x_4 & = & 5 \\ x_1 & + & x_2 & + & 5x_3 & + & 2x_4 & = & 7 \\ x_1 & + & 2x_2 & + & 8x_3 & + & 4x_4 & = & 12. \end{array}$$

Its augmented matrix is the matrix

$$(A, b) = \begin{pmatrix} 1 & 0 & 2 & 1 & 5 \\ 1 & 1 & 5 & 2 & 7 \\ 1 & 2 & 8 & 4 & 12 \end{pmatrix}$$

considered above, so the reduction steps applied to this matrix yield the system

$$\begin{array}{rcrcrcl} x_1 & & + & 2x_3 & & = & 2 \\ & x_2 & + & 3x_3 & & = & -1 \\ & & & & x_4 & = & 3. \end{array}$$

This reduced system has the same set of solutions as the original, and obviously x_3 can be chosen arbitrarily. Therefore, our system has infinitely many solutions given by

$$x_1 = 2 - 2x_3, \quad x_2 = -1 - 3x_3, \quad x_4 = 3,$$

where x_3 is arbitrary.

The following proposition shows that the set of solutions of a system $Ax = b$ is preserved by any sequence of row operations.

Proposition 5.12. *Given any $m \times n$ matrix A and any vector $b \in \mathbb{R}^m$, for any sequence of elementary row operations $E_1, \ldots, E_k$, if $P = E_k \cdots E_1$ and $(A', b') = P(A, b)$, then the solutions of $Ax = b$ are the same as the solutions of $A'x = b'$.*

Proof. Since each elementary row operation E_i is invertible, so is P, and since $(A', b') = P(A, b)$, then $A' = PA$ and $b' = Pb$. If x is a solution of the original system $Ax = b$, then multiplying both sides by P we get $PAx = Pb$; that is, $A'x = b'$, so x is a solution of the new system. Conversely, assume that x is a solution of the new system, that is $A'x = b'$. Then, because $A' = PA$, $b' = PB$, and P is invertible, we get

$$Ax = P^{-1}A'x = P^{-1}b' = b,$$

so x is a solution of the original system $Ax = b$. □

Another important fact is this:

Proposition 5.13. *Given a $m \times n$ matrix A, for any sequence of row operations $E_1, \ldots, E_k$, if $P = E_k \cdots E_1$ and $B = PA$, then the subspaces spanned by the rows of A and the rows of B are identical. Therefore, A and B have the same row rank. Furthermore, the matrices A and B also have the same (column) rank.*

Proof. Since $B = PA$, from a previous observation, the rows of B are linear combinations of the rows of A, so the span of the rows of B is a subspace of the span of the rows of A. Since P is invertible, $A = P^{-1}B$, so by the same reasoning the span of the rows of A is a subspace of the span of the rows of B. Therefore, the subspaces spanned by the rows of A and the rows of B are identical, which implies that A and B have the same row rank.

Proposition 5.12 implies that the systems $Ax = 0$ and $Bx = 0$ have the same solutions. Since Ax is a linear combinations of the columns of A and Bx is a linear combinations of the columns of B, the maximum number of linearly independent columns in A is equal to the maximum number of linearly independent columns in B; that is, A and B have the same rank. □

Remark: The subspaces spanned by the columns of A and B can be different! However, their dimension must be the same.

Of course, we know from Proposition 8.11 that the row rank is equal to the column rank. We will see that the reduction to row echelon form provides another proof of this important fact. Let us now define precisely what is a reduced row echelon matrix.

Definition 5.1. A $m \times n$ matrix A is a *reduced row echelon matrix* iff the following conditions hold:

(a) The first nonzero entry in every row is 1. This entry is called a *pivot*.

(b) The first nonzero entry of row $i+1$ is to the right of the first nonzero entry of row i.

(c) The entries above a pivot are zero.

If a matrix satisfies the above conditions, we also say that it is in *reduced row echelon form*, for short *rref*.

Note that condition (b) implies that the entries below a pivot are also zero. For example, the matrix

$$A = \begin{pmatrix} 1 & 6 & 0 & 1 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

is a reduced row echelon matrix. In general, a matrix in *rref* has the following shape:

$$\begin{pmatrix} 1 & 0 & 0 & \times & \times & 0 & 0 & \times \\ 0 & 1 & 0 & \times & \times & 0 & 0 & \times \\ 0 & 0 & 1 & \times & \times & 0 & 0 & \times \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & \times \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & \times \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

if the last row consists of zeros, or

$$\begin{pmatrix} 1 & 0 & 0 & \times & \times & 0 & 0 & \times & 0 & \times \\ 0 & 1 & 0 & \times & \times & 0 & 0 & \times & 0 & \times \\ 0 & 0 & 1 & \times & \times & 0 & 0 & \times & 0 & \times \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & \times & 0 & \times \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & \times & \times & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \times \end{pmatrix}$$

if the last row contains a pivot.

The following proposition shows that every matrix can be converted to a reduced row echelon form using row operations.

Proposition 5.14. *Given any $m \times n$ matrix A, there is a sequence of row operations $E_1, \ldots, E_k$ such that if $P = E_k \cdots E_1$, then $U = PA$ is a reduced row echelon matrix.*

Proof. We proceed by induction on m. If $m = 1$, then either all entries on this row are zero, so $A = 0$, or if a_j is the first nonzero entry in A, let $P = (a_j^{-1})$ (a 1×1 matrix); clearly, PA is a reduced row echelon matrix.

Let us now assume that $m \geq 2$. If $A = 0$ we are done, so let us assume that $A \neq 0$. Since $A \neq 0$, there is a leftmost column j which is nonzero, so pick any pivot $\pi = a_{ij}$ in the jth column, permute row i and row 1 if necessary, multiply the new first row by π^{-1}, and clear out the other entries in column j by subtracting suitable multiples of row 1. At the end of this process, we have a matrix A_1 that has the following shape:

$$A_1 = \begin{pmatrix} 0 & \cdots & 0 & 1 & * & \cdots & * \\ 0 & \cdots & 0 & 0 & * & \cdots & * \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & * & \cdots & * \end{pmatrix},$$

where $*$ stands for an arbitrary scalar, or more concisely

$$A_1 = \begin{pmatrix} 0 & 1 & B \\ 0 & 0 & D \end{pmatrix},$$

where D is a $(m-1) \times (n-j)$ matrix. If $j = n$, we are done. Otherwise, by the induction hypothesis applied to D, there is a sequence of row operations that converts D to a reduced row echelon matrix R', and these row operations do not affect the first row of A_1, which means that A_1 is reduced to a matrix of the form

$$R = \begin{pmatrix} 0 & 1 & B \\ 0 & 0 & R' \end{pmatrix}.$$

Because R' is a reduced row echelon matrix, the matrix R satisfies conditions (a) and (b) of the reduced row echelon form. Finally, the entries above all pivots in R' can be cleared out by subtracting suitable multiples of the rows of R' containing a pivot. The resulting matrix also satisfies condition (c), and the induction step is complete. □

Remark: There is a `Matlab` function named `rref` that converts any matrix to its reduced row echelon form.

If A is any matrix and if R is a reduced row echelon form of A, the second part of Proposition 5.13 can be sharpened a little. Namely, *the rank of A is equal to the number of pivots in R.*

This is because the structure of a reduced row echelon matrix makes it clear that its rank is equal to the number of pivots.

Given a system of the form $Ax = b$, we can apply the reduction procedure to the augmented matrix (A, b) to obtain a reduced row echelon matrix (A', b') such that the system

$A'x = b'$ has the same solutions as the original system $Ax = b$. The advantage of the reduced system $A'x = b'$ is that there is a simple test to check whether this system is solvable, and to find its solutions if it is solvable.

Indeed, if any row of the matrix A' is zero and if the corresponding entry in b' is nonzero, then it is a pivot and we have the "equation"

$$0 = 1,$$

which means that the system $A'x = b'$ has no solution. On the other hand, if there is no pivot in b', then for every row i in which $b'_i \neq 0$, there is some column j in A' where the entry on row i is 1 (a pivot). Consequently, we can assign arbitrary values to the variable x_k if column k does not contain a pivot, and then solve for the pivot variables.

For example, if we consider the reduced row echelon matrix

$$(A', b') = \begin{pmatrix} 1 & 6 & 0 & 1 & 0 \\ 0 & 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

there is no solution to $A'x = b'$ because the third equation is $0 = 1$. On the other hand, the reduced system

$$(A', b') = \begin{pmatrix} 1 & 6 & 0 & 1 & 1 \\ 0 & 0 & 1 & 2 & 3 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

has solutions. We can pick the variables x_2, x_4 corresponding to nonpivot columns arbitrarily, and then solve for x_3 (using the second equation) and x_1 (using the first equation).

The above reasoning proved the following theorem:

Theorem 5.15. *Given any system $Ax = b$ where A is a $m \times n$ matrix, if the augmented matrix (A, b) is a reduced row echelon matrix, then the system $Ax = b$ has a solution iff there is no pivot in b. In that case, an arbitrary value can be assigned to the variable x_j if column j does not contain a pivot.*

Nonpivot variables are often called *free variables*.

Putting Proposition 5.14 and Theorem 5.15 together we obtain a criterion to decide whether a system $Ax = b$ has a solution: Convert the augmented system (A, b) to a row reduced echelon matrix (A', b') and check whether b' has no pivot.

Remark: When writing a program implementing row reduction, we may stop when the last column of the matrix A is reached. In this case, the test whether the system $Ax = b$ is solvable is that the row-reduced matrix A' has no zero row of index $i > r$ such that $b'_i \neq 0$ (where r is the number of pivots, and b' is the row-reduced right-hand side).

If we have a *homogeneous system* $Ax = 0$, which means that $b = 0$, of course $x = 0$ is always a solution, but Theorem 5.15 implies that if the system $Ax = 0$ has more variables than equations, then it has some nonzero solution (we call it a *nontrivial solution*).

Proposition 5.16. *Given any homogeneous system $Ax = 0$ of m equations in n variables, if $m < n$, then there is a nonzero vector $x \in \mathbb{R}^n$ such that $Ax = 0$.*

Proof. Convert the matrix A to a reduced row echelon matrix A'. We know that $Ax = 0$ iff $A'x = 0$. If r is the number of pivots of A', we must have $r \leq m$, so by Theorem 5.15 we may assign arbitrary values to $n - r > 0$ nonpivot variables and we get nontrivial solutions. □

Theorem 5.15 can also be used to characterize when a square matrix is invertible. First, note the following simple but important fact:

If a square $n \times n$ matrix A is a row reduced echelon matrix, then either A is the identity or the bottom row of A is zero.

Proposition 5.17. *Let A be a square matrix of dimension n. The following conditions are equivalent:*

(a) The matrix A can be reduced to the identity by a sequence of elementary row operations.

(b) The matrix A is a product of elementary matrices.

(c) The matrix A is invertible.

(d) The system of homogeneous equations $Ax = 0$ has only the trivial solution $x = 0$.

Proof. First, we prove that (a) implies (b). If (a) can be reduced to the identity by a sequence of row operations $E_1, \ldots, E_p$, this means that $E_p \cdots E_1 A = I$. Since each E_i is invertible, we get

$$A = E_1^{-1} \cdots E_p^{-1},$$

where each E_i^{-1} is also an elementary row operation, so (b) holds. Now if (b) holds, since elementary row operations are invertible, A is invertible, and (c) holds. If A is invertible, we already observed that the homogeneous system $Ax = 0$ has only the trivial solution $x = 0$, because from $Ax = 0$, we get $A^{-1}Ax = A^{-1}0$; that is, $x = 0$. It remains to prove that (d) implies (a), and for this we prove the contrapositive: if (a) does not hold, then (d) does not hold.

Using our basic observation about reducing square matrices, if A does not reduce to the identity, then A reduces to a row echelon matrix A' whose bottom row is zero. Say $A' = PA$, where P is a product of elementary row operations. Because the bottom row of A' is zero, the system $A'x = 0$ has at most $n - 1$ nontrivial equations, and by Proposition 5.16, this system has a nontrivial solution x. But then, $Ax = P^{-1}A'x = 0$ with $x \neq 0$, contradicting the fact that the system $Ax = 0$ is assumed to have only the trivial solution. Therefore, (d) implies (a) and the proof is complete. □

Proposition 5.17 yields a method for computing the inverse of an invertible matrix A: reduce A to the identity using elementary row operations, obtaining

$$E_p \cdots E_1 A = I.$$

Multiplying both sides by A^{-1} we get

$$A^{-1} = E_p \cdots E_1.$$

From a practical point of view, we can build up the product $E_p \cdots E_1$ by reducing to row echelon form the augmented $n \times 2n$ matrix (A, I_n) obtained by adding the n columns of the identity matrix to A. This is just another way of performing the Gauss–Jordan procedure.

Here is an example: let us find the inverse of the matrix

$$A = \begin{pmatrix} 5 & 4 \\ 6 & 5 \end{pmatrix}.$$

We form the 2×4 block matrix

$$(A, I) = \begin{pmatrix} 5 & 4 & 1 & 0 \\ 6 & 5 & 0 & 1 \end{pmatrix}$$

and apply elementary row operations to reduce A to the identity. For example:

$$(A, I) = \begin{pmatrix} 5 & 4 & 1 & 0 \\ 6 & 5 & 0 & 1 \end{pmatrix} \longrightarrow \begin{pmatrix} 5 & 4 & 1 & 0 \\ 1 & 1 & -1 & 1 \end{pmatrix}$$

by subtracting row 1 from row 2,

$$\begin{pmatrix} 5 & 4 & 1 & 0 \\ 1 & 1 & -1 & 1 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 0 & 5 & -4 \\ 1 & 1 & -1 & 1 \end{pmatrix}$$

by subtracting $4\times$ row 2 from row 1,

$$\begin{pmatrix} 1 & 0 & 5 & -4 \\ 1 & 1 & -1 & 1 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 0 & 5 & -4 \\ 0 & 1 & -6 & 5 \end{pmatrix} = (I, A^{-1}),$$

by subtracting row 1 from row 2. Thus

$$A^{-1} = \begin{pmatrix} 5 & -4 \\ -6 & 5 \end{pmatrix}.$$

Proposition 5.17 can also be used to give an elementary proof of the fact that if a square matrix A has a left inverse B (resp. a right inverse B), so that $BA = I$ (resp. $AB = I$), then A is invertible and $A^{-1} = B$. This is an interesting exercise, try it!

For the sake of completeness, we prove that the reduced row echelon form of a matrix is unique. The neat proof given below is borrowed and adapted from W. Kahan.

Proposition 5.18. *Let A be any $m \times n$ matrix. If U and V are two reduced row echelon matrices obtained from A by applying two sequences of elementary row operations $E_1, \ldots, E_p$ and $F_1, \ldots, F_q$, so that*

$$U = E_p \cdots E_1 A \quad \textit{and} \quad V = F_q \cdots F_1 A,$$

then $U = V$ and $E_p \cdots E_1 = F_q \cdots F_1$. In other words, the reduced row echelon form of any matrix is unique.

Proof. Let

$$C = E_p \cdots E_1 F_1^{-1} \cdots F_q^{-1}$$

so that

$$U = CV \quad \text{and} \quad V = C^{-1}U.$$

We prove by induction on n that $U = V$ (and $C = I$).

Let ℓ_j denote the jth column of the identity matrix I_n, and let $u_j = U\ell_j$, $v_j = V\ell_j$, $c_j = C\ell_j$, and $a_j = A\ell_j$, be the jth column of U, V, C, and A respectively.

First, I claim that $u_j = 0$ iff $v_j = 0$, iff $a_j = 0$.

Indeed, if $v_j = 0$, then (because $U = CV$) $u_j = Cv_j = 0$, and if $u_j = 0$, then $v_j = C^{-1}u_j = 0$. Since $A = E_p \cdots E_1 U$, we also get $a_j = 0$ iff $u_j = 0$.

Therefore, we may simplify our task by striking out columns of zeros from U, V, and A, since they will have corresponding indices. We still use n to denote the number of columns of A. Observe that because U and V are reduced row echelon matrices with no zero columns, we must have $u_1 = v_1 = \ell_1$.

Claim. If U and V are reduced row echelon matrices without zero columns such that $U = CV$, for all $k \geq 1$, if $k \leq n$, then ℓ_k occurs in U iff ℓ_k occurs in V, and if ℓ_k does occurs in U, then

1. ℓ_k occurs for the same index j_k in both U and V;
2. the first j_k columns of U and V match;
3. the subsequent columns in U and V (of index $> j_k$) whose elements beyond the kth all vanish also match;
4. the first k columns of C match the first k columns of I_n.

We prove this claim by induction on k.

For the base case $k = 1$, we already know that $u_1 = v_1 = \ell_1$. We also have

$$c_1 = C\ell_1 = Cv_1 = u_1 = \ell_1.$$

If $v_j = \lambda \ell_1$ for some $\mu \in \mathbb{R}$, then

$$u_j = U\ell_1 = CV\ell_1 = Cv_j = \lambda C\ell_1 = \lambda\ell_1 = v_j.$$

A similar argument using C^{-1} shows that if $u_j = \lambda\ell_1$, then $v_j = u_j$. Therefore, all the columns of U and V proportional to ℓ_1 match, which establishes the base case. Observe that if ℓ_2 appears in U, then it must appear in both U and V for the same index, and if not then $U = V$.

Next us now prove the induction step; this is only necessary if ℓ_{k+1} appears in both U, in wich case, by (3) of the induction hypothesis, it appears in both U and V for the same index, say j_{k+1}. Thus $u_{j_{k+1}} = v_{j_{k+1}} = \ell_{k+1}$. It follows that

$$c_{k+1} = C\ell_{k+1} = Cv_{j_{k+1}} = u_{j_{k+1}} = \ell_{k+1},$$

so the first $k+1$ columns of C match the first $k+1$ columns of I_n.

Consider any subsequent column v_j (with $j > j_{k+1}$) whose elements beyond the $(k+1)$th all vanish. Then, v_j is a linear combination of columns of V to the left of v_j, so

$$u_j = Cv_j = v_j.$$

because the first $k+1$ columns of C match the first column of I_n. Similarly, any subsequent column u_j (with $j > j_{k+1}$) whose elements beyond the $(k+1)$th all vanish is equal to v_j. Therefore, all the subsequent columns in U and V (of index $> j_{k+1}$) whose elements beyond the $(k+1)$th all vanish also match, which completes the induction hypothesis.

We can now prove that $U = V$ (recall that we may assume that U and V have no zero columns). We noted earlier that $u_1 = v_1 = \ell_1$, so there is a largest $k \leq n$ such that ℓ_k occurs in U. Then, the previous claim implies that all the columns of U and V match, which means that $U = V$. □

The reduction to row echelon form also provides a method to describe the set of solutions of a linear system of the form $Ax = b$.

5.10 Solving Linear Systems Using RREF

First, we have the following simple result.

Proposition 5.19. *Let A be any $m \times n$ matrix and let $b \in \mathbb{R}^m$ be any vector. If the system $Ax = b$ has a solution, then the set Z of all solutions of this system is the set*

$$Z = x_0 + \mathrm{Ker}\,(A) = \{x_0 + x \mid Ax = 0\},$$

where $x_0 \in \mathbb{R}^n$ is any solution of the system $Ax = b$, which means that $Ax_0 = b$ (x_0 is called a special solution), and where $\mathrm{Ker}\,(A) = \{x \in \mathbb{R}^n \mid Ax = 0\}$, the set of solutions of the homogeneous system associated with $Ax = b$.

Proof. Assume that the system $Ax = b$ is solvable and let x_0 and x_1 be any two solutions so that $Ax_0 = b$ and $Ax_1 = b$. Subtracting the first equation from the second, we get

$$A(x_1 - x_0) = 0,$$

which means that $x_1 - x_0 \in \mathrm{Ker}\,(A)$. Therefore, $Z \subseteq x_0 + \mathrm{Ker}\,(A)$, where x_0 is a special solution of $Ax = b$. Conversely, if $Ax_0 = b$, then for any $z \in \mathrm{Ker}\,(A)$, we have $Az = 0$, and so

$$A(x_0 + z) = Ax_0 + Az = b + 0 = b,$$

which shows that $x_0 + \mathrm{Ker}\,(A) \subseteq Z$. Therefore, $Z = x_0 + \mathrm{Ker}\,(A)$. □

Given a linear system $Ax = b$, reduce the augmented matrix (A, b) to its row echelon form (A', b'). As we showed before, the system $Ax = b$ has a solution iff b' contains no pivot. Assume that this is the case. Then, if (A', b') has r pivots, which means that A' has r pivots since b' has no pivot, we know that the first r columns of I_m appear in A'.

We can permute the columns of A' and renumber the variables in x correspondingly so that the first r columns of I_m match the first r columns of A', and then our reduced echelon matrix is of the form (R, b') with

$$R = \begin{pmatrix} I_r & F \\ 0_{m-r,r} & 0_{m-r,n-r} \end{pmatrix}$$

and

$$b' = \begin{pmatrix} d \\ 0_{m-r} \end{pmatrix},$$

where F is a $r \times (n-r)$ matrix and $d \in \mathbb{R}^r$. Note that R has $m - r$ zero rows.

Then, because

$$\begin{pmatrix} I_r & F \\ 0_{m-r,r} & 0_{m-r,n-r} \end{pmatrix} \begin{pmatrix} d \\ 0_{n-r} \end{pmatrix} = \begin{pmatrix} d \\ 0_{m-r} \end{pmatrix} = b',$$

we see that

$$x_0 = \begin{pmatrix} d \\ 0_{n-r} \end{pmatrix}$$

is a special solution of $Rx = b'$, and thus to $Ax = b$. In other words, we get a special solution by assigning the first r components of b' to the pivot variables and setting the nonpivot variables (the *free variables*) to zero.

We can also find a basis of the kernel (nullspace) of A using F. If $x = (u, v)$ is in the kernel of A, with $u \in \mathbb{R}^r$ and $v \in \mathbb{R}^{n-r}$, then x is also in the kernel of R, which means that $Rx = 0$; that is,

$$\begin{pmatrix} I_r & F \\ 0_{m-r,r} & 0_{m-r,n-r} \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} u + Fv \\ 0_{m-r} \end{pmatrix} = \begin{pmatrix} 0_r \\ 0_{m-r} \end{pmatrix}.$$

Therefore, $u = -Fv$, and $\mathrm{Ker}\,(A)$ consists of all vectors of the form

$$\begin{pmatrix} -Fv \\ v \end{pmatrix} = \begin{pmatrix} -F \\ I_{n-r} \end{pmatrix} v,$$

for any arbitrary $v \in \mathbb{R}^{n-r}$. It follows that the $n - r$ columns of the matrix

$$N = \begin{pmatrix} -F \\ I_{n-r} \end{pmatrix}$$

form a basis of the kernel of A. This is because N contains the identity matrix I_{n-r} as a submatrix, so the columns of N are linearly independent. In summary, if $N^1, \ldots, N^{n-r}$ are the columns of N, then the general solution of the equation $Ax = b$ is given by

$$x = \begin{pmatrix} d \\ 0_{n-r} \end{pmatrix} + x_{r+1}N^1 + \cdots + x_n N^{n-r},$$

where $x_{r+1}, \ldots, x_n$ are the free variables; that is, the nonpivot variables.

In the general case where the columns corresponding to pivots are mixed with the columns corresponding to free variables, we find the special solution as follows. Let $i_1 < \cdots < i_r$ be the indices of the columns corresponding to pivots. Then, assign b'_k to the pivot variable x_{i_k} for $k = 1, \ldots, r$, and set all other variables to 0. To find a basis of the kernel, we form the $n - r$ vectors N^k obtained as follows. Let $j_1 < \cdots < j_{n-r}$ be the indices of the columns corresponding to free variables. For every column j_k corresponding to a free variable ($1 \leq k \leq n - r$), form the vector N^k defined so that the entries $N^k_{i_1}, \ldots, N^k_{i_r}$ are equal to the negatives of the first r entries in column j_k (flip the sign of these entries); let $N^k_{j_k} = 1$, and set all other entries to zero. Schematically, if the column of index j_k (corresponding to the free variable x_{j_k}) is

$$\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_r \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

then the vector N^k is given by

$$\begin{array}{c|c} 1 & 0 \\ \vdots & \vdots \\ i_1 - 1 & 0 \\ i_1 & -\alpha_1 \\ i_1 + 1 & 0 \\ \vdots & \vdots \\ i_r - 1 & 0 \\ i_r & -\alpha_r \\ i_r + 1 & 0 \\ \vdots & \vdots \\ j_k - 1 & 0 \\ j_k & 1 \\ j_k + 1 & 0 \\ \vdots & \vdots \\ n & 0 \end{array}.$$

The presence of the 1 in position j_k guarantees that $N^1, \ldots, N^{n-r}$ are linearly independent.

An illustration of the above method, consider the problem of finding a basis of the subspace V of $n \times n$ matrices $A \in \mathrm{M}_n(\mathbb{R})$ satisfying the following properties:

1. The sum of the entries in every row has the same value (say c_1);

2. The sum of the entries in every column has the same value (say c_2).

It turns out that $c_1 = c_2$ and that the $2n-2$ equations corresponding to the above conditions are linearly independent. We leave the proof of these facts as an interesting exercise. By the duality theorem, the dimension of the space V of matrices satisying the above equations is $n^2 - (2n-2)$. Let us consider the case $n = 4$. There are 6 equations, and the space V has dimension 10. The equations are

$$\begin{aligned}
a_{11} + a_{12} + a_{13} + a_{14} - a_{21} - a_{22} - a_{23} - a_{24} &= 0 \\
a_{21} + a_{22} + a_{23} + a_{24} - a_{31} - a_{32} - a_{33} - a_{34} &= 0 \\
a_{31} + a_{32} + a_{33} + a_{34} - a_{41} - a_{42} - a_{43} - a_{44} &= 0 \\
a_{11} + a_{21} + a_{31} + a_{41} - a_{12} - a_{22} - a_{32} - a_{42} &= 0 \\
a_{12} + a_{22} + a_{32} + a_{42} - a_{13} - a_{23} - a_{33} - a_{43} &= 0 \\
a_{13} + a_{23} + a_{33} + a_{43} - a_{14} - a_{24} - a_{34} - a_{44} &= 0,
\end{aligned}$$

and the corresponding matrix is

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 \end{pmatrix}.$$

The result of performing the reduction to row echelon form yields the following matrix in rref:

$$U = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & 0 & -1 & -1 & -1 & 2 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & -1 & -1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & -1 & -1 & 0 & -1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & -1 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \end{pmatrix}$$

The list *pivlist* of indices of the pivot variables and the list *freelist* of indices of the free variables is given by

$$\begin{aligned} \textit{pivlist} &= (1, 2, 3, 4, 5, 9), \\ \textit{freelist} &= (6, 7, 8, 10, 11, 12, 13, 14, 15, 16). \end{aligned}$$

After applying the algorithm to find a basis of the kernel of U, we find the following 16×10 matrix

$$BK = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & -2 & -1 & -1 & -1 \\ -1 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & -1 & 0 & 0 & -1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & -1 & 0 & 0 & -1 & 1 & 1 & 1 & 0 \\ -1 & -1 & -1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

The reader should check that that in each column j of BK, the lowest 1 belongs to the row whose index is the jth element in *freelist*, and that in each column j of BK, the signs of

the entries whose indices belong to *pivlist* are the fipped signs of the 6 entries in the column U corresponding to the jth index in *freelist*. We can now read off from BK the 4×4 matrices that form a basis of V: every column of BK corresponds to a matrix whose rows have been concatenated. We get the following 10 matrices:

$$M_1 = \begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad M_2 = \begin{pmatrix} 1 & 0 & -1 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad M_3 = \begin{pmatrix} 1 & 0 & 0 & -1 \\ -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$M_4 = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad M_5 = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad M_6 = \begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$M_7 = \begin{pmatrix} -2 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}, \quad M_8 = \begin{pmatrix} -1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \quad M_9 = \begin{pmatrix} -1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

$$M_{10} = \begin{pmatrix} -1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Recall that a *magic square* is a square matrix that satisfies the two conditions about the sum of the entries in each row and in each column to be the same number, and also the additional two constraints that the main descending and the main ascending diagonals add up to this common number. Furthermore, the entries are also required to be positive integers. For $n = 4$, the additional two equations are

$$\begin{aligned} a_{22} + a_{33} + a_{44} - a_{12} - a_{13} - a_{14} &= 0 \\ a_{41} + a_{32} + a_{23} - a_{11} - a_{12} - a_{13} &= 0, \end{aligned}$$

and the 8 equations stating that a matrix is a magic square are linearly independent. Again, by running row elimination, we get a basis of the "generalized magic squares" whose entries are not restricted to be positive integers. We find a basis of 8 matrices. For $n = 3$, we find a basis of 3 matrices.

A magic square is said to be *normal* if its entries are precisely the integers $1, 2 \ldots, n^2$. Then, since the sum of these entries is

$$1 + 2 + 3 + \cdots + n^2 = \frac{n^2(n^2+1)}{2},$$

and since each row (and column) sums to the same number, this common value (the *magic sum*) is

$$\frac{n(n^2+1)}{2}.$$

It is easy to see that there are no normal magic squares for $n = 2$. For $n = 3$, the magic sum is 15, for $n = 4$, it is 34, and for $n = 5$, it is 65.

In the case $n = 3$, we have the additional condition that the rows and columns add up to 15, so we end up with a solution parametrized by two numbers x_1, x_2; namely,

$$\begin{pmatrix} x_1 + x_2 - 5 & 10 - x_2 & 10 - x_1 \\ 20 - 2x_1 - x_2 & 5 & 2x_1 + x_2 - 10 \\ x_1 & x_2 & 15 - x_1 - x_2 \end{pmatrix}.$$

Thus, in order to find a normal magic square, we have the additional inequality constraints

$$\begin{aligned} x_1 + x_2 &> 5 \\ x_1 &< 10 \\ x_2 &< 10 \\ 2x_1 + x_2 &< 20 \\ 2x_1 + x_2 &> 10 \\ x_1 &> 0 \\ x_2 &> 0 \\ x_1 + x_2 &< 15, \end{aligned}$$

and all 9 entries in the matrix must be distinct. After a tedious case analysis, we discover the remarkable fact that there is a unique normal magic square (up to rotations and reflections):

$$\begin{pmatrix} 2 & 7 & 6 \\ 9 & 5 & 1 \\ 4 & 3 & 8 \end{pmatrix}.$$

It turns out that there are 880 different normal magic squares for $n = 4$, and 275, 305, 224 normal magic squares for $n = 5$ (up to rotations and reflections). Even for $n = 4$, it takes a fair amount of work to enumerate them all! Finding the number of magic squares for $n > 5$ is an open problem!

5.11 Elementary Matrices and Columns Operations

Instead of performing elementary row operations on a matrix A, we can perform elementary columns operations, which means that we multiply A by elementary matrices on the right. As elementary row and column operations, $P(i, k)$, $E_{i,j;\beta}$, $E_{i,\lambda}$ perform the following actions:

1. As a row operation, $P(i,k)$ permutes row i and row k.
2. As a column operation, $P(i,k)$ permutes column i and column k.
3. The inverse of $P(i,k)$ is $P(i,k)$ itself.
4. As a row operation, $E_{i,j;\beta}$ adds β times row j to row i.
5. As a column operation, $E_{i,j;\beta}$ adds β times column i to column j (note the switch in the indices).
6. The inverse of $E_{i,j;\beta}$ is $E_{i,j;-\beta}$.
7. As a row operation, $E_{i,\lambda}$ multiplies row i by λ.
8. As a column operation, $E_{i,\lambda}$ multiplies column i by λ.
9. The inverse of $E_{i,\lambda}$ is $E_{i,\lambda^{-1}}$.

We can define the notion of a reduced column echelon matrix and show that every matrix can be reduced to a unique reduced column echelon form. Now, given any $m \times n$ matrix A, if we first convert A to its reduced row echelon form R, it is easy to see that we can apply elementary column operations that will reduce R to a matrix of the form

$$\begin{pmatrix} I_r & 0_{r,n-r} \\ 0_{m-r,r} & 0_{m-r,n-r} \end{pmatrix},$$

where r is the number of pivots (obtained during the row reduction). Therefore, for every $m \times n$ matrix A, there exist two sequences of elementary matrices $E_1, \ldots, E_p$ and $F_1, \ldots, F_q$, such that

$$E_p \cdots E_1 A F_1 \cdots F_q = \begin{pmatrix} I_r & 0_{r,n-r} \\ 0_{m-r,r} & 0_{m-r,n-r} \end{pmatrix}.$$

The matrix on the right-hand side is called the *rank normal form* of A. Clearly, r is the rank of A. It is easy to see that the rank normal form also yields a proof of the fact that A and its transpose $A^\top$ have the same rank.

5.12 Transvections and Dilatations

In this section, we characterize the linear isomorphisms of a vector space E that leave every vector in some hyperplane fixed. These maps turn out to be the linear maps that are represented in some suitable basis by elementary matrices of the form $E_{i,j;\beta}$ (transvections) or $E_{i,\lambda}$ (dilatations). Furthermore, the transvections generate the group $\mathbf{SL}(E)$, and the dilatations generate the group $\mathbf{GL}(E)$.

Let H be any hyperplane in E, and pick some (nonzero) vector $v \in E$ such that $v \notin H$, so that

$$E = H \oplus Kv.$$

Assume that $f\colon E \to E$ is a linear isomorphism such that $f(u) = u$ for all $u \in H$, and that f is not the identity. We have

$$f(v) = h + \alpha v, \quad \text{for some } h \in H \text{ and some } \alpha \in K,$$

with $\alpha \neq 0$, because otherwise we would have $f(v) = h = f(h)$ since $h \in H$, contradicting the injectivity of f ($v \neq h$ since $v \notin H$). For any $x \in E$, if we write

$$x = y + tv, \quad \text{for some } y \in H \text{ and some } t \in K,$$

then

$$f(x) = f(y) + f(tv) = y + tf(v) = y + th + t\alpha v,$$

and since $\alpha x = \alpha y + t\alpha v$, we get

$$\begin{aligned} f(x) - \alpha x &= (1-\alpha)y + th \\ f(x) - x &= t(h + (\alpha - 1)v). \end{aligned}$$

Observe that if E is finite-dimensional, by picking a basis of E consisting of v and basis vectors of H, then the matrix of f is a lower triangular matrix whose diagonal entries are all 1 except the first entry which is equal to α. Therefore, $\det(f) = \alpha$.

Case 1. $\alpha \neq 1$.

We have $f(x) = \alpha x$ iff $(1-\alpha)y + th = 0$ iff

$$y = \frac{t}{\alpha - 1} h.$$

Then, if we let $w = h + (\alpha - 1)v$, for $y = (t/(\alpha - 1))h$, we have

$$x = y + tv = \frac{t}{\alpha - 1} h + tv = \frac{t}{\alpha - 1}(h + (\alpha - 1)v) = \frac{t}{\alpha - 1} w,$$

which shows that $f(x) = \alpha x$ iff $x \in Kw$. Note that $w \notin H$, since $\alpha \neq 1$ and $v \notin H$. Therefore,

$$E = H \oplus Kw,$$

and f is the identity on H and a magnification by α on the line $D = Kw$.

Definition 5.2. Given a vector space E, for any hyperplane H in E, any nonzero vector $u \in E$ such that $u \notin H$, and any scalar $\alpha \neq 0, 1$, a linear map f such that $f(x) = x$ for all $x \in H$ and $f(x) = \alpha x$ for every $x \in D = Ku$ is called a *dilatation of hyperplane H, direction D, and scale factor α.*

If π_H and π_D are the projections of E onto H and D, then we have

$$f(x) = \pi_H(x) + \alpha\pi_D(x).$$

The inverse of f is given by

$$f^{-1}(x) = \pi_H(x) + \alpha^{-1}\pi_D(x).$$

When $\alpha = -1$, we have $f^2 = \mathrm{id}$, and f is a symmetry about the hyperplane H in the direction D.

Case 2. $\alpha = 1$.

In this case,

$$f(x) - x = th,$$

that is, $f(x) - x \in Kh$ for all $x \in E$. Assume that the hyperplane H is given as the kernel of some linear form φ, and let $a = \varphi(v)$. We have $a \neq 0$, since $v \notin H$. For any $x \in E$, we have

$$\varphi(x - a^{-1}\varphi(x)v) = \varphi(x) - a^{-1}\varphi(x)\varphi(v) = \varphi(x) - \varphi(x) = 0,$$

which shows that $x - a^{-1}\varphi(x)v \in H$ for all $x \in E$. Since every vector in H is fixed by f, we get

$$\begin{aligned} x - a^{-1}\varphi(x)v &= f(x - a^{-1}\varphi(x)v) \\ &= f(x) - a^{-1}\varphi(x)f(v), \end{aligned}$$

so

$$f(x) = x + \varphi(x)(f(a^{-1}v) - a^{-1}v).$$

Since $f(z) - z \in Kh$ for all $z \in E$, we conclude that $u = f(a^{-1}v) - a^{-1}v = \beta h$ for some $\beta \in K$, so $\varphi(u) = 0$, and we have

$$f(x) = x + \varphi(x)u, \quad \varphi(u) = 0. \tag{$*$}$$

A linear map defined as above is denoted by $\tau_{\varphi,u}$.

Conversely for any linear map $f = \tau_{\varphi,u}$ given by equation $(*)$, where φ is a nonzero linear form and u is some vector $u \in E$ such that $\varphi(u) = 0$, if $u = 0$ then f is the identity, so assume that $u \neq 0$. If so, we have $f(x) = x$ iff $\varphi(x) = 0$, that is, iff $x \in H$. We also claim that the inverse of f is obtained by changing u to $-u$. Actually, we check the slightly more general fact that

$$\tau_{\varphi,u} \circ \tau_{\varphi,v} = \tau_{\varphi,u+v}.$$

Indeed, using the fact that $\varphi(v) = 0$, we have

$$\begin{aligned} \tau_{\varphi,u}(\tau_{\varphi,v}(x)) &= \tau_{\varphi,v}(x) + \varphi(\tau_{\varphi,v}(v))u \\ &= \tau_{\varphi,v}(x) + (\varphi(x) + \varphi(x)\varphi(v))u \\ &= \tau_{\varphi,v}(x) + \varphi(x)u \\ &= x + \varphi(x)v + \varphi(x)u \\ &= x + \varphi(x)(u + v). \end{aligned}$$

For $v = -u$, we have $\tau_{\varphi,u+v} = \varphi_{\varphi,0} = \mathrm{id}$, so $\tau_{\varphi,u}^{-1} = \tau_{\varphi,-u}$, as claimed.

Therefore, we proved that every linear isomorphism of E that leaves every vector in some hyperplane H fixed and has the property that $f(x) - x \in H$ for all $x \in E$ is given by a map $\tau_{\varphi,u}$ as defined by equation $(*)$, where φ is some nonzero linear form defining H and u is some vector in H. We have $\tau_{\varphi,u} = \mathrm{id}$ iff $u = 0$.

Definition 5.3. Given any hyperplane H in E, for any nonzero nonlinear form $\varphi \in E^*$ defining H (which means that $H = \mathrm{Ker}\,(\varphi)$) and any nonzero vector $u \in H$, the linear map $\tau_{\varphi,u}$ given by

$$\tau_{\varphi,u}(x) = x + \varphi(x)u, \quad \varphi(u) = 0,$$

for all $x \in E$ is called a *transvection of hyperplane H and direction u*. The map $\tau_{\varphi,u}$ leaves every vector in H fixed, and $f(x) - x \in Ku$ for all $x \in E$.

The above arguments show the following result.

Proposition 5.20. *Let $f\colon E \to E$ be a bijective linear map and assume that $f \neq \mathrm{id}$ and that $f(x) = x$ for all $x \in H$, where H is some hyperplane in E. If there is some nonzero vector $u \in E$ such that $u \notin H$ and $f(u) - u \in H$, then f is a transvection of hyperplane H; otherwise, f is a dilatation of hyperplane H.*

Proof. Using the notation as above, for some $v \notin H$, we have $f(v) = h + \alpha v$ with $\alpha \neq 0$, and write $u = y + tv$ with $y \in H$ and $t \neq 0$ since $u \notin H$. If $f(u) - u \in H$, from

$$f(u) - u = t(h + (\alpha - 1)v),$$

we get $(\alpha - 1)v \in H$, and since $v \notin H$, we must have $\alpha = 1$, and we proved that f is a transvection. Otherwise, $\alpha \neq 0, 1$, and we proved that f is a dilatation. □

If E is finite-dimensional, then $\alpha = \det(f)$, so we also have the following result.

Proposition 5.21. *Let $f\colon E \to E$ be a bijective linear map of a finite-dimensional vector space E and assume that $f \neq \mathrm{id}$ and that $f(x) = x$ for all $x \in H$, where H is some hyperplane in E. If $\det(f) = 1$, then f is a transvection of hyperplane H; otherwise, f is a dilatation of hyperplane H.*

Suppose that f is a dilatation of hyperplane H and direction u, and say $\det(f) = \alpha \neq 0, 1$. Pick a basis $(u, e_2, \ldots, e_n)$ of E where $(e_2, \ldots, e_n)$ is a basis of H. Then, the matrix of f is of the form

$$\begin{pmatrix} \alpha & 0 & \cdots & 0 \\ 0 & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix},$$

which is an elementary matrix of the form $E_{1,\alpha}$. Conversely, it is clear that every elementary matrix of the form $E_{i,\alpha}$ with $\alpha \neq 0, 1$ is a dilatation.

Now, assume that f is a transvection of hyperplane H and direction $u \in H$. Pick some $v \notin H$, and pick some basis $(u, e_3, \ldots, e_n)$ of H, so that $(v, u, e_3, \ldots, e_n)$ is a basis of E. Since $f(v) - v \in Ku$, the matrix of f is of the form

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ \alpha & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix},$$

which is an elementary matrix of the form $E_{2,1;\alpha}$. Conversely, it is clear that every elementary matrix of the form $E_{i,j;\alpha}$ ($\alpha \neq 0$) is a transvection.

The following proposition is an interesting exercise that requires good mastery of the elementary row operations $E_{i,j;\beta}$.

Proposition 5.22. *Given any invertible $n \times n$ matrix A, there is a matrix S such that*

$$SA = \begin{pmatrix} I_{n-1} & 0 \\ 0 & \alpha \end{pmatrix} = E_{n,\alpha},$$

with $\alpha = \det(A)$, and where S is a product of elementary matrices of the form $E_{i,j;\beta}$; that is, S is a composition of transvections.

Surprisingly, every transvection is the composition of two dilatations!

Proposition 5.23. *If the field K is not of charateristic 2, then every transvection f of hyperplane H can be written as $f = d_2 \circ d_1$, where d_1, d_2 are dilatations of hyperplane H, where the direction of d_1 can be chosen arbitrarily.*

Proof. Pick some dilalation d_1 of hyperplane H and scale factor $\alpha \neq 0, 1$. Then, $d_2 = f \circ d_1^{-1}$ leaves every vector in H fixed, and $\det(d_2) = \alpha^{-1} \neq 1$. By Proposition 5.21, the linear map d_2 is a dilatation of hyperplane H, and we have $f = d_2 \circ d_1$, as claimed. □

Observe that in Proposition 5.23, we can pick $\alpha = -1$; that is, every transvection of hyperplane H is the compositions of two symmetries about the hyperplane H, one of which can be picked arbitrarily.

Remark: Proposition 5.23 holds as long as $K \neq \{0, 1\}$.

The following important result is now obtained.

Theorem 5.24. *Let E be any finite-dimensional vector space over a field K of characteristic not equal to 2. Then, the group $\mathbf{SL}(E)$ is generated by the transvections, and the group $\mathbf{GL}(E)$ is generated by the dilatations.*

Proof. Consider any $f \in \mathbf{SL}(E)$, and let A be its matrix in any basis. By Proposition 5.22, there is a matrix S such that

$$SA = \begin{pmatrix} I_{n-1} & 0 \\ 0 & \alpha \end{pmatrix} = E_{n,\alpha},$$

with $\alpha = \det(A)$, and where S is a product of elementary matrices of the form $E_{i,j;\beta}$. Since $\det(A) = 1$, we have $\alpha = 1$, and the result is proved. Otherwise, $E_{n,\alpha}$ is a dilatation, S is a product of transvections, and by Proposition 5.23, every transvection is the composition of two dilatations, so the second result is also proved. □

We conclude this section by proving that any two transvections are conjugate in $\mathbf{GL}(E)$. Let $\tau_{\varphi,u}$ $(u \neq 0)$ be a transvection and let $g \in \mathbf{GL}(E)$ be any invertible linear map. We have

$$\begin{aligned}(g \circ \tau_{\varphi,u} \circ g^{-1})(x) &= g(g^{-1}(x) + \varphi(g^{-1}(x))u) \\ &= x + \varphi(g^{-1}(x))g(u).\end{aligned}$$

Let us find the hyperplane determined by the linear form $x \mapsto \varphi(g^{-1}(x))$. This is the set of vectors $x \in E$ such that $\varphi(g^{-1}(x)) = 0$, which holds iff $g^{-1}(x) \in H$ iff $x \in g(H)$. Therefore, $\mathrm{Ker}\,(\varphi \circ g^{-1}) = g(H) = H'$, and we have $g(u) \in g(H) = H'$, so $g \circ \tau_{\varphi,u} \circ g^{-1}$ is the transvection of hyperplane $H' = g(H)$ and direction $u' = g(u)$ (with $u' \in H'$).

Conversely, let $\tau_{\psi,u'}$ be some transvection $(u' \neq 0)$. Pick some vector v, v' such that $\varphi(v) = \psi(v') = 1$, so that

$$E = H \oplus Kv = H' \oplus v'.$$

There is a linear map $g \in \mathbf{GL}(E)$ such that $g(u) = u'$, $g(v) = v'$, and $g(H) = H'$. To define g, pick a basis $(v, u, e_2, \ldots, e_{n-1})$ where $(u, e_2, \ldots, e_{n-1})$ is a basis of H and pick a basis $(v', u', e'_2, \ldots, e'_{n-1})$ where $(u', e'_2, \ldots, e'_{n-1})$ is a basis of H'; then g is defined so that $g(v) = v'$, $g(u) = u'$, and $g(e_i) = g(e'_i)$, for $i = 2, \ldots, n-1$. If $n = 2$, then e_i and e'_i are missing. Then, we have

$$(g \circ \tau_{\varphi,u} \circ g^{-1})(x) = x + \varphi(g^{-1}(x))u'.$$

Now, $\varphi \circ g^{-1}$ also determines the hyperplane $H' = g(H)$, so we have $\varphi \circ g^{-1} = \lambda\psi$ for some nonzero λ in K. Since $v' = g(v)$, we get

$$\varphi(v) = \varphi \circ g^{-1}(v') = \lambda\psi(v'),$$

and since $\varphi(v) = \psi(v') = 1$, we must have $\lambda = 1$. It follows that

$$(g \circ \tau_{\varphi,u} \circ g^{-1})(x) = x + \psi(x)u' = \tau_{\psi,u'}(x).$$

In summary, we proved almost all parts the following result.

Proposition 5.25. *Let E be any finite-dimensional vector space. For every transvection $\tau_{\varphi,u}$ ($u \neq 0$) and every linear map $g \in \mathbf{GL}(E)$, the map $g \circ \tau_{\varphi,u} \circ g^{-1}$ is the transvection of hyperplane $g(H)$ and direction $g(u)$ (that is, $g \circ \tau_{\varphi,u} \circ g^{-1} = \tau_{\varphi \circ g^{-1}, g(u)}$). For every other transvection $\tau_{\psi,u'}$ ($u' \neq 0$), there is some $g \in \mathbf{GL}(E)$ such $\tau_{\psi,u'} = g \circ \tau_{\varphi,u} \circ g^{-1}$; in other words any two transvections ($\neq \mathrm{id}$) are conjugate in $\mathbf{GL}(E)$. Moreover, if $n \geq 3$, then the linear isomorphim g as above can be chosen so that $g \in \mathbf{SL}(E)$.*

Proof. We just need to prove that if $n \geq 3$, then for any two transvections $\tau_{\varphi,u}$ and $\tau_{\psi,u'}$ ($u, u' \neq 0$), there is some $g \in \mathbf{SL}(E)$ such that $\tau_{\psi,u'} = g \circ \tau_{\varphi,u} \circ g^{-1}$. As before, we pick a basis $(v, u, e_2, \ldots, e_{n-1})$ where $(u, e_2, \ldots, e_{n-1})$ is a basis of H, we pick a basis $(v', u', e'_2, \ldots, e'_{n-1})$ where $(u', e'_2, \ldots, e'_{n-1})$ is a basis of H', and we define g as the unique linear map such that $g(v) = v'$, $g(u) = u'$, and $g(e_i) = e'_i$, for $i = 1, \ldots, n-1$. But, in this case, both H and $H' = g(H)$ have dimension at least 2, so in any basis of H' including u', there is some basis vector e'_2 independent of u', and we can rescale e'_2 in such a way that the matrix of g over the two bases has determinant $+1$. □

5.13 Summary

The main concepts and results of this chapter are listed below:

- One does not solve (large) linear systems by computing determinants.
- *Upper-triangular* (*lower-triangular*) matrices.
- Solving by *back-substitution* (*forward-substitution*).
- *Gaussian elimination*.
- Permuting rows.
- The *pivot* of an elimination step; *pivoting*.
- *Transposition matrix*; *elementary matrix*.
- The *Gaussian elimination theorem* (Theorem 5.1).
- *Gauss-Jordan factorization*.
- *LU-factorization*; Necessary and sufficient condition for the existence of an LU-factorization (Proposition 5.2).
- *LDU-factorization*.
- "$PA = LU$ theorem" (Theorem 5.5).
- $LDL^\top$*-factorization* of a symmetric matrix.

- Avoiding small pivots: *partial pivoting*; *complete pivoting*.
- Gaussian elimination of tridiagonal matrices.
- LU-factorization of tridiagonal matrices.
- *Symmetric positive definite* matrices (SPD matrices).
- *Cholesky factorization* (Theorem 5.10).
- Criteria for a symmetric matrix to be positive definite; *Sylvester's criterion*.
- *Reduced row echelon form*.
- Reduction of a rectangular matrix to its row echelon form.
- Using the reduction to row echelon form to decide whether a system $Ax = b$ is solvable, and to find its solutions, using a *special* solution and a basis of the *homogeneous system* $Ax = 0$.
- *Magic squares*.
- *transvections and dilatations*.

Chapter 6

Vector Norms and Matrix Norms

6.1 Normed Vector Spaces

In order to define how close two vectors or two matrices are, and in order to define the convergence of sequences of vectors or matrices, we can use the notion of a norm. Recall that $\mathbb{R}_+ = \{x \in \mathbb{R} \mid x \geq 0\}$. Also recall that if $z = a + ib \in \mathbb{C}$ is a complex number, with $a, b \in \mathbb{R}$, then $\overline{z} = a - ib$ and $|z| = \sqrt{z\overline{z}} = \sqrt{a^2 + b^2}$ ($|z|$ is the *modulus* of z).

Definition 6.1. Let E be a vector space over a field K, where K is either the field $\mathbb{R}$ of reals, or the field $\mathbb{C}$ of complex numbers. A *norm* on E is a function $\| \| \colon E \to \mathbb{R}_+$, assigning a nonnegative real number $\|u\|$ to any vector $u \in E$, and satisfying the following conditions for all $x, y, z \in E$:

(N1) $\|x\| \geq 0$, and $\|x\| = 0$ iff $x = 0$. (positivity)

(N2) $\|\lambda x\| = |\lambda| \, \|x\|$. (homogeneity (or scaling))

(N3) $\|x + y\| \leq \|x\| + \|y\|$. (triangle inequality)

A vector space E together with a norm $\| \|$ is called a *normed vector space*.

By (N2), setting $\lambda = -1$, we obtain

$$\|-x\| = \|(-1)x\| = |-1| \, \|x\| = \|x\| \, ;$$

that is, $\|-x\| = \|x\|$. From (N3), we have

$$\|x\| = \|x - y + y\| \leq \|x - y\| + \|y\| \, ,$$

which implies that

$$\|x\| - \|y\| \leq \|x - y\| \, .$$

By exchanging x and y and using the fact that by (N2),

$$\|y - x\| = \|-(x - y)\| = \|x - y\| \, ,$$

we also have

$$\|y\| - \|x\| \leq \|x - y\|.$$

Therefore,

$$|\|x\| - \|y\|| \leq \|x - y\|, \quad \text{for all } x, y \in E. \tag{$*$}$$

Observe that setting $\lambda = 0$ in (N2), we deduce that $\|0\| = 0$ without assuming (N1). Then, by setting $y = 0$ in $(*)$, we obtain

$$|\|x\|| \leq \|x\|, \quad \text{for all } x \in E.$$

Therefore, the condition $\|x\| \geq 0$ in (N1) follows from (N2) and (N3), and (N1) can be replaced by the weaker condition

(N1') For all $x \in E$, if $\|x\| = 0$ then $x = 0$,

A function $\| \| : E \to \mathbb{R}$ satisfying axioms (N2) and (N3) is called a *seminorm*. From the above discussion, a seminorm also has the properties

$\|x\| \geq 0$ for all $x \in E$, and $\|0\| = 0$.

However, there may be nonzero vectors $x \in E$ such that $\|x\| = 0$. Let us give some examples of normed vector spaces.

Example 6.1.

1. Let $E = \mathbb{R}$, and $\|x\| = |x|$, the absolute value of x.
2. Let $E = \mathbb{C}$, and $\|z\| = |z|$, the modulus of z.
3. Let $E = \mathbb{R}^n$ (or $E = \mathbb{C}^n$). There are three standard norms. For every $(x_1, \ldots, x_n) \in E$, we have the norm $\|x\|_1$, defined such that,

$$\|x\|_1 = |x_1| + \cdots + |x_n|,$$

we have the *Euclidean norm* $\|x\|_2$, defined such that,

$$\|x\|_2 = \left(|x_1|^2 + \cdots + |x_n|^2\right)^{\frac{1}{2}},$$

and the *sup*-norm $\|x\|_\infty$, defined such that,

$$\|x\|_\infty = \max\{|x_i| \mid 1 \leq i \leq n\}.$$

More generally, we define the *ℓ_p-norm* (for $p \geq 1$) by

$$\|x\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}.$$

There are other norms besides the ℓ_p-norms. Here are some examples.

1. For $E = \mathbb{R}^2$,
$$\|(u_1, u_2)\| = |u_1| + 2|u_2|.$$
2. For $E = \mathbb{R}^2$,
$$\|(u_1, u_2)\| = \left((u_1 + u_2)^2 + u_1^2\right)^{1/2}.$$
3. For $E = \mathbb{C}^2$,
$$\|(u_1, u_2)\| = |u_1 + iu_2| + |u_1 - iu_2|.$$

The reader should check that they satisfy all the axioms of a norm.

Some work is required to show the triangle inequality for the ℓ_p-norm.

Proposition 6.1. *If E is a finite-dimensional vector space over $\mathbb{R}$ or $\mathbb{C}$, for every real number $p \geq 1$, the ℓ_p-norm is indeed a norm.*

Proof. The cases $p = 1$ and $p = \infty$ are easy and left to the reader. If $p > 1$, then let $q > 1$ such that
$$\frac{1}{p} + \frac{1}{q} = 1.$$
We will make use of the following fact: for all $\alpha, \beta \in \mathbb{R}$, if $\alpha, \beta \geq 0$, then
$$\alpha\beta \leq \frac{\alpha^p}{p} + \frac{\beta^q}{q}. \tag{$*$}$$
To prove the above inequality, we use the fact that the exponential function $t \mapsto e^t$ satisfies the following convexity inequality:
$$e^{\theta x + (1-\theta)y} \leq \theta e^x + (1 - \theta)e^y,$$
for all $x, y \in \mathbb{R}$ and all θ with $0 \leq \theta \leq 1$.

Since the case $\alpha\beta = 0$ is trivial, let us assume that $\alpha > 0$ and $\beta > 0$. If we replace θ by $1/p$, x by $p \log \alpha$ and y by $q \log \beta$, then we get
$$e^{\frac{1}{p}p\log\alpha + \frac{1}{q}q\log\beta} \leq \frac{1}{p}e^{p\log\alpha} + \frac{1}{q}e^{q\log\beta},$$
which simplifies to
$$\alpha\beta \leq \frac{\alpha^p}{p} + \frac{\beta^q}{q},$$
as claimed.

We will now prove that for any two vectors $u, v \in E$, we have
$$\sum_{i=1}^{n} |u_i v_i| \leq \|u\|_p \|v\|_q. \tag{$**$}$$

Since the above is trivial if $u = 0$ or $v = 0$, let us assume that $u \neq 0$ and $v \neq 0$. Then, the inequality $(*)$ with $\alpha = |u_i|/\|u\|_p$ and $\beta = |v_i|/\|v\|_q$ yields

$$\frac{|u_i v_i|}{\|u\|_p \|v\|_q} \leq \frac{|u_i|^p}{p\|u\|_p^p} + \frac{|v_i|^q}{q\|u\|_q^q},$$

for $i = 1, \ldots, n$, and by summing up these inequalities, we get

$$\sum_{i=1}^{n} |u_i v_i| \leq \|u\|_p \|v\|_q,$$

as claimed. To finish the proof, we simply have to prove that property (N3) holds, since (N1) and (N2) are clear. Now, for $i = 1, \ldots, n$, we can write

$$(|u_i| + |v_i|)^p = |u_i|(|u_i| + |v_i|)^{p-1} + |v_i|(|u_i| + |v_i|)^{p-1},$$

so that by summing up these equations we get

$$\sum_{i=1}^{n}(|u_i| + |v_i|)^p = \sum_{i=1}^{n} |u_i|(|u_i| + |v_i|)^{p-1} + \sum_{i=1}^{n} |v_i|(|u_i| + |v_i|)^{p-1},$$

and using the inequality $(**)$, we get

$$\sum_{i=1}^{n}(|u_i| + |v_i|)^p \leq (\|u\|_p + \|v\|_p)\Big(\sum_{i=1}^{n}(|u_i| + |v_i|)^{(p-1)q}\Big)^{1/q}.$$

However, $1/p + 1/q = 1$ implies $pq = p + q$, that is, $(p-1)q = p$, so we have

$$\sum_{i=1}^{n}(|u_i| + |v_i|)^p \leq (\|u\|_p + \|v\|_p)\Big(\sum_{i=1}^{n}(|u_i| + |v_i|)^{p}\Big)^{1/q},$$

which yields

$$\Big(\sum_{i=1}^{n}(|u_i| + |v_i|)^p\Big)^{1/p} \leq \|u\|_p + \|v\|_p.$$

Since $|u_i + v_i| \leq |u_i| + |v_i|$, the above implies the triangle inequality $\|u + v\|_p \leq \|u\|_p + \|v\|_p$, as claimed. □

For $p > 1$ and $1/p + 1/q = 1$, the inequality

$$\sum_{i=1}^{n} |u_i v_i| \leq \Big(\sum_{i=1}^{n} |u_i|^p\Big)^{1/p}\Big(\sum_{i=1}^{n} |v_i|^q\Big)^{1/q}$$

is known as *Hölder's inequality*. For $p = 2$, it is the *Cauchy–Schwarz inequality*.

Actually, if we define the *Hermitian inner product* $\langle -, - \rangle$ on $\mathbb{C}^n$ by

$$\langle u, v \rangle = \sum_{i=1}^{n} u_i \overline{v}_i,$$

where $u = (u_1, \ldots, u_n)$ and $v = (v_1, \ldots, v_n)$, then

$$|\langle u, v \rangle| \leq \sum_{i=1}^{n} |u_i \overline{v}_i| = \sum_{i=1}^{n} |u_i v_i|,$$

so Hölder's inequality implies the inequality

$$|\langle u, v \rangle| \leq \|u\|_p \|v\|_q$$

also called *Hölder's inequality*, which, for $p = 2$ is the standard Cauchy–Schwarz inequality. The triangle inequality for the ℓ_p-norm,

$$\left(\sum_{i=1}^{n} (|u_i + v_i|)^p \right)^{1/p} \leq \left(\sum_{i=1}^{n} |u_i|^p \right)^{1/p} + \left(\sum_{i=1}^{n} |v_i|^q \right)^{1/q},$$

is known as *Minkowski's inequality*.

When we restrict the Hermitian inner product to real vectors, $u, v \in \mathbb{R}^n$, we get the *Euclidean inner product*

$$\langle u, v \rangle = \sum_{i=1}^{n} u_i v_i.$$

It is very useful to observe that if we represent (as usual) $u = (u_1, \ldots, u_n)$ and $v = (v_1, \ldots, v_n)$ (in $\mathbb{R}^n$) by column vectors, then their Euclidean inner product is given by

$$\langle u, v \rangle = u^\top v = v^\top u,$$

and when $u, v \in \mathbb{C}^n$, their Hermitian inner product is given by

$$\langle u, v \rangle = v^* u = \overline{u^* v}.$$

In particular, when $u = v$, in the complex case we get

$$\|u\|_2^2 = u^* u,$$

and in the real case, this becomes

$$\|u\|_2^2 = u^\top u.$$

As convenient as these notations are, we still recommend that you do not abuse them; the notation $\langle u, v \rangle$ is more intrinsic and still "works" when our vector space is infinite dimensional.

The following proposition is easy to show.

Proposition 6.2. *The following inequalities hold for all $x \in \mathbb{R}^n$ (or $x \in \mathbb{C}^n$):*

$$\|x\|_\infty \leq \|x\|_1 \leq n\|x\|_\infty,$$
$$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty,$$
$$\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2.$$

Proposition 6.2 is actually a special case of a very important result: in a finite-dimensional vector space, any two norms are equivalent.

Definition 6.2. Given any (real or complex) vector space E, two norms $\| \|_a$ and $\| \|_b$ are *equivalent* iff there exists some positive reals $C_1, C_2 > 0$, such that

$$\|u\|_a \leq C_1 \|u\|_b \quad \text{and} \quad \|u\|_b \leq C_2 \|u\|_a, \text{ for all } u \in E.$$

Given any norm $\| \|$ on a vector space of dimension n, for any basis $(e_1, \ldots, e_n)$ of E, observe that for any vector $x = x_1e_1 + \cdots + x_ne_n$, we have

$$\|x\| = \|x_1e_1 + \cdots + x_ne_n\| \leq |x_1| \|e_1\| + \cdots + |x_n| \|e_n\| \leq C(|x_1| + \cdots + |x_n|) = C \|x\|_1,$$

with $C = \max_{1 \leq i \leq n} \|e_i\|$ and

$$\|x\|_1 = \|x_1e_1 + \cdots + x_ne_n\| = |x_1| + \cdots + |x_n|.$$

The above implies that

$$| \|u\| - \|v\| | \leq \|u - v\| \leq C \|u - v\|_1,$$

which means that the map $u \mapsto \|u\|$ is *continuous* with respect to the norm $\| \|_1$.

Let S_1^{n-1} be the unit sphere with respect to the norm $\| \|_1$, namely

$$S_1^{n-1} = \{x \in E \mid \|x\|_1 = 1\}.$$

Now, S_1^{n-1} is a closed and bounded subset of a finite-dimensional vector space, so by Heine–Borel (or equivalently, by Bolzano–Weiertrass), S_1^{n-1} is compact. On the other hand, it is a well known result of analysis that any continuous real-valued function on a nonempty compact set has a minimum and a maximum, and that they are achieved. Using these facts, we can prove the following important theorem:

Theorem 6.3. *If E is any real or complex vector space of finite dimension, then any two norms on E are equivalent.*

Proof. It is enough to prove that any norm $\| \|$ is equivalent to the 1-norm. We already proved that the function $x \mapsto \|x\|$ is continuous with respect to the norm $\| \|_1$ and we observed that the unit sphere S_1^{n-1} is compact. Now, we just recalled that because the function $f \colon x \mapsto \|x\|$ is continuous and because S_1^{n-1} is compact, the function f has a minimum m and a maximum

M, and because $\|x\|$ is never zero on S_1^{n-1}, we must have $m > 0$. Consequently, we just proved that if $\|x\|_1 = 1$, then

$$0 < m \leq \|x\| \leq M,$$

so for any $x \in E$ with $x \neq 0$, we get

$$m \leq \|x/\|x\|_1\| \leq M,$$

which implies

$$m\|x\|_1 \leq \|x\| \leq M\|x\|_1.$$

Since the above inequality holds trivially if $x = 0$, we just proved that $\| \|$ and $\| \|_1$ are equivalent, as claimed. □

Next, we will consider norms on matrices.

6.2 Matrix Norms

For simplicity of exposition, we will consider the vector spaces $\mathrm{M}_n(\mathbb{R})$ and $\mathrm{M}_n(\mathbb{C})$ of square $n \times n$ matrices. Most results also hold for the spaces $\mathrm{M}_{m,n}(\mathbb{R})$ and $\mathrm{M}_{m,n}(\mathbb{C})$ of rectangular $m \times n$ matrices. Since $n \times n$ matrices can be multiplied, the idea behind matrix norms is that they should behave "well" with respect to matrix multiplication.

Definition 6.3. A *matrix norm* $\| \|$ on the space of square $n \times n$ matrices in $\mathrm{M}_n(K)$, with $K = \mathbb{R}$ or $K = \mathbb{C}$, is a norm on the vector space $\mathrm{M}_n(K)$, with the additional property called *submultiplicativity* that

$$\|AB\| \leq \|A\|\,\|B\|,$$

for all $A, B \in \mathrm{M}_n(K)$. A norm on matrices satisfying the above property is often called a *submultiplicative* matrix norm.

Since $I^2 = I$, from $\|I\| = \|I^2\| \leq \|I\|^2$, we get $\|I\| \geq 1$, for every matrix norm.

Before giving examples of matrix norms, we need to review some basic definitions about matrices. Given any matrix $A = (a_{ij}) \in \mathrm{M}_{m,n}(\mathbb{C})$, the *conjugate* $\overline{A}$ of A is the matrix such that

$$\overline{A}_{ij} = \overline{a}_{ij}, \quad 1 \leq i \leq m,\ 1 \leq j \leq n.$$

The *transpose* of A is the $n \times m$ matrix $A^\top$ such that

$$A^\top_{ij} = a_{ji}, \quad 1 \leq i \leq m,\ 1 \leq j \leq n.$$

The *adjoint* of A is the $n \times m$ matrix A^* such that

$$A^* = \overline{(A^\top)} = (\overline{A})^\top.$$

When A is a real matrix, $A^* = A^\top$. A matrix $A \in \mathrm{M}_n(\mathbb{C})$ is *Hermitian* if

$$A^* = A.$$

If A is a real matrix ($A \in \mathrm{M}_n(\mathbb{R})$), we say that A is *symmetric* if

$$A^\top = A.$$

A matrix $A \in \mathrm{M}_n(\mathbb{C})$ is *normal* if

$$AA^* = A^*A,$$

and if A is a real matrix, it is *normal* if

$$AA^\top = A^\top A.$$

A matrix $U \in \mathrm{M}_n(\mathbb{C})$ is *unitary* if

$$UU^* = U^*U = I.$$

A real matrix $Q \in \mathrm{M}_n(\mathbb{R})$ is *orthogonal* if

$$QQ^\top = Q^\top Q = I.$$

Given any matrix $A = (a_{ij}) \in \mathrm{M}_n(\mathbb{C})$, the *trace* $\mathrm{tr}(A)$ of A is the sum of its diagonal elements

$$\mathrm{tr}(A) = a_{11} + \cdots + a_{nn}.$$

It is easy to show that the trace is a linear map, so that

$$\mathrm{tr}(\lambda A) = \lambda \mathrm{tr}(A)$$

and

$$\mathrm{tr}(A + B) = \mathrm{tr}(A) + \mathrm{tr}(B).$$

Moreover, if A is an $m \times n$ matrix and B is an $n \times m$ matrix, it is not hard to show that

$$\mathrm{tr}(AB) = \mathrm{tr}(BA).$$

We also review eigenvalues and eigenvectors. We content ourselves with definition involving matrices. A more general treatment will be given later on (see Chapter 12).

Definition 6.4. Given any square matrix $A \in \mathrm{M}_n(\mathbb{C})$, a complex number $\lambda \in \mathbb{C}$ is an *eigenvalue* of A if there is some *nonzero* vector $u \in \mathbb{C}^n$, such that

$$Au = \lambda u.$$

If λ is an eigenvalue of A, then the *nonzero* vectors $u \in \mathbb{C}^n$ such that $Au = \lambda u$ are called *eigenvectors of A associated with λ*; together with the zero vector, these eigenvectors form a subspace of $\mathbb{C}^n$ denoted by $E_\lambda(A)$, and called the *eigenspace associated with λ*.

Remark: Note that Definition 6.4 *requires an eigenvector to be nonzero*. A somewhat unfortunate consequence of this requirement is that the set of eigenvectors is *not* a subspace, since the zero vector is missing! On the positive side, whenever eigenvectors are involved, there is no need to say that they are nonzero. The fact that eigenvectors are nonzero is implicitly used in all the arguments involving them, so it seems safer (but perhaps not as elegant) to stituplate that eigenvectors should be nonzero.

If A is a square real matrix $A \in \mathrm{M}_n(\mathbb{R})$, then we restrict Definition 6.4 to real eigenvalues $\lambda \in \mathbb{R}$ and real eigenvectors. However, it should be noted that although every complex matrix always has at least some complex eigenvalue, a real matrix may not have any real eigenvalues. For example, the matrix

$$A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

has the complex eigenvalues i and $-i$, but no real eigenvalues. Thus, typically, even for real matrices, we consider complex eigenvalues.

Observe that $\lambda \in \mathbb{C}$ is an eigenvalue of A
iff $Au = \lambda u$ for some nonzero vector $u \in \mathbb{C}^n$
iff $(\lambda I - A)u = 0$
iff the matrix $\lambda I - A$ defines a linear map which has a nonzero kernel, that is,
iff $\lambda I - A$ not invertible.

However, from Proposition 4.11, $\lambda I - A$ is not invertible iff

$$\det(\lambda I - A) = 0.$$

Now, $\det(\lambda I - A)$ is a polynomial of degree n in the indeterminate λ, in fact, of the form

$$\lambda^n - \mathrm{tr}(A)\lambda^{n-1} + \cdots + (-1)^n \det(A).$$

Thus, we see that the eigenvalues of A are the zeros (also called *roots*) of the above polynomial. Since every complex polynomial of degree n has exactly n roots, counted with their multiplicity, we have the following definition:

Definition 6.5. Given any square $n \times n$ matrix $A \in \mathrm{M}_n(\mathbb{C})$, the polynomial

$$\det(\lambda I - A) = \lambda^n - \mathrm{tr}(A)\lambda^{n-1} + \cdots + (-1)^n \det(A)$$

is called the *characteristic polynomial* of A. The n (not necessarily distinct) roots $\lambda_1, \ldots, \lambda_n$ of the characteristic polynomial are all the *eigenvalues* of A and constitute the *spectrum* of A. We let

$$\rho(A) = \max_{1 \leq i \leq n} |\lambda_i|$$

be the largest modulus of the eigenvalues of A, called the *spectral radius* of A.

Since the eigenvalue $\lambda_1, \ldots, \lambda_n$ of A are the zeros of the polynomial

$$\det(\lambda I - A) = \lambda^n - \mathrm{tr}(A)\lambda^{n-1} + \cdots + (-1)^n \det(A),$$

we deduce (see Section 12.1 for details) that

$$\begin{aligned} \mathrm{tr}(A) &= \lambda_1 + \cdots + \lambda_n \\ \det(A) &= \lambda_1 \cdots \lambda_n. \end{aligned}$$

Proposition 6.4. *For any matrix norm $\| \|$ on $\mathrm{M}_n(\mathbb{C})$ and for any square $n \times n$ matrix $A \in \mathrm{M}_n(\mathbb{C})$, we have*

$$\rho(A) \leq \|A\|.$$

Proof. Let λ be some eigenvalue of A for which $|\lambda|$ is maximum, that is, such that $|\lambda| = \rho(A)$. If u ($\neq 0$) is any eigenvector associated with λ and if U is the $n \times n$ matrix whose columns are all u, then $Au = \lambda u$ implies

$$AU = \lambda U,$$

and since

$$|\lambda| \, \|U\| = \|\lambda U\| = \|AU\| \leq \|A\| \, \|U\|$$

and $U \neq 0$, we have $\|U\| \neq 0$, and get

$$\rho(A) = |\lambda| \leq \|A\|,$$

as claimed. □

Proposition 6.4 also holds for any real matrix norm $\| \|$ on $\mathrm{M}_n(\mathbb{R})$ but the proof is more subtle and requires the notion of induced norm. We prove it after giving Definition 6.7.

Now, it turns out that if A is a real $n \times n$ symmetric matrix, then the eigenvalues of A are all real and there is some orthogonal matrix Q such that

$$A = Q\mathrm{diag}(\lambda_1, \ldots, \lambda_n)Q^\top,$$

where $\mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ denotes the matrix whose only nonzero entries (if any) are its diagonal entries, which are the (real) eigenvalues of A. Similarly, if A is a complex $n \times n$ Hermitian matrix, then the eigenvalues of A are all real and there is some unitary matrix U such that

$$A = U\mathrm{diag}(\lambda_1, \ldots, \lambda_n)U^*,$$

where $\mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ denotes the matrix whose only nonzero entries (if any) are its diagonal entries, which are the (real) eigenvalues of A.

We now return to matrix norms. We begin with the so-called *Frobenius norm*, which is just the norm $\| \|_2$ on $\mathbb{C}^{n^2}$, where the $n \times n$ matrix A is viewed as the vector obtained by concatenating together the rows (or the columns) of A. The reader should check that for any $n \times n$ complex matrix $A = (a_{ij})$,

$$\left(\sum_{i,j=1}^{n} |a_{ij}|^2 \right)^{1/2} = \sqrt{\mathrm{tr}(A^*A)} = \sqrt{\mathrm{tr}(AA^*)}.$$

Definition 6.6. The *Frobenius norm* $\| \|_F$ is defined so that for every square $n \times n$ matrix $A \in \mathrm{M}_n(\mathbb{C})$,

$$\|A\|_F = \left(\sum_{i,j=1}^{n} |a_{ij}|^2 \right)^{1/2} = \sqrt{\mathrm{tr}(AA^*)} = \sqrt{\mathrm{tr}(A^*A)}.$$

The following proposition show that the Frobenius norm is a matrix norm satisfying other nice properties.

Proposition 6.5. *The Frobenius norm* $\| \|_F$ *on* $\mathrm{M}_n(\mathbb{C})$ *satisfies the following properties:*

(1) It is a matrix norm; that is, $\|AB\|_F \leq \|A\|_F \|B\|_F$*, for all* $A, B \in \mathrm{M}_n(\mathbb{C})$*.*

(2) It is unitarily invariant, which means that for all unitary matrices U, V*, we have*

$$\|A\|_F = \|UA\|_F = \|AV\|_F = \|UAV\|_F .$$

(3) $\sqrt{\rho(A^*A)} \leq \|A\|_F \leq \sqrt{n}\sqrt{\rho(A^*A)}$*, for all* $A \in \mathrm{M}_n(\mathbb{C})$*.*

Proof. (1) The only property that requires a proof is the fact $\|AB\|_F \leq \|A\|_F \|B\|_F$. This follows from the Cauchy–Schwarz inequality:

$$\begin{aligned} \|AB\|_F^2 &= \sum_{i,j=1}^{n} \left| \sum_{k=1}^{n} a_{ik} b_{kj} \right|^2 \\ &\leq \sum_{i,j=1}^{n} \left(\sum_{h=1}^{n} |a_{ih}|^2 \right) \left(\sum_{k=1}^{n} |b_{kj}|^2 \right) \\ &= \left(\sum_{i,h=1}^{n} |a_{ih}|^2 \right) \left(\sum_{k,j=1}^{n} |b_{kj}|^2 \right) = \|A\|_F^2 \|B\|_F^2 . \end{aligned}$$

(2) We have

$$\|A\|_F^2 = \mathrm{tr}(A^*A) = \mathrm{tr}(VV^*A^*A) = \mathrm{tr}(V^*A^*AV) = \|AV\|_F^2 ,$$

and

$$\|A\|_F^2 = \mathrm{tr}(A^*A) = \mathrm{tr}(A^*U^*UA) = \|UA\|_F^2 .$$

The identity

$$\|A\|_F = \|UAV\|_F$$

follows from the previous two.

(3) It is well known that the trace of a matrix is equal to the sum of its eigenvalues. Furthermore, A^*A is symmetric positive semidefinite (which means that its eigenvalues are nonnegative), so $\rho(A^*A)$ is the largest eigenvalue of A^*A and

$$\rho(A^*A) \leq \mathrm{tr}(A^*A) \leq n\rho(A^*A),$$

which yields (3) by taking square roots. □

Remark: The Frobenius norm is also known as the *Hilbert-Schmidt norm* or the *Schur norm*. So many famous names associated with such a simple thing!

We now give another method for obtaining matrix norms using subordinate norms. First, we need a proposition that shows that in a finite-dimensional space, the linear map induced by a matrix is bounded, and thus continuous.

Proposition 6.6. *For every norm $\|\ \|$ on $\mathbb{C}^n$ (or $\mathbb{R}^n$), for every matrix $A \in \mathrm{M}_n(\mathbb{C})$ (or $A \in \mathrm{M}_n(\mathbb{R})$), there is a real constant $C_A \geq 0$, such that*

$$\|Au\| \leq C_A \|u\|,$$

for every vector $u \in \mathbb{C}^n$ (or $u \in \mathbb{R}^n$ if A is real).

Proof. For every basis $(e_1, \ldots, e_n)$ of $\mathbb{C}^n$ (or $\mathbb{R}^n$), for every vector $u = u_1 e_1 + \cdots + u_n e_n$, we have

$$\begin{aligned}\|Au\| &= \|u_1 A(e_1) + \cdots + u_n A(e_n)\| \\ &\leq |u_1| \|A(e_1)\| + \cdots + |u_n| \|A(e_n)\| \\ &\leq C_1(|u_1| + \cdots + |u_n|) = C_1 \|u\|_1,\end{aligned}$$

where $C_1 = \max_{1 \leq i \leq n} \|A(e_i)\|$. By Theorem 6.3, the norms $\|\ \|$ and $\|\ \|_1$ are equivalent, so there is some constant $C_2 > 0$ so that $\|u\|_1 \leq C_2 \|u\|$ for all u, which implies that

$$\|Au\| \leq C_A \|u\|,$$

where $C_A = C_1 C_2$. □

Proposition 6.6 says that every linear map on a finite-dimensional space is *bounded*. This implies that every linear map on a finite-dimensional space is continuous. Actually, it is not hard to show that a linear map on a normed vector space E is bounded iff it is continuous, regardless of the dimension of E.

Proposition 6.6 implies that for every matrix $A \in \mathrm{M}_n(\mathbb{C})$ (or $A \in \mathrm{M}_n(\mathbb{R})$),

$$\sup_{\substack{x \in \mathbb{C}^n \\ x \neq 0}} \frac{\|Ax\|}{\|x\|} \leq C_A.$$

Now, since $\|\lambda u\| = |\lambda| \|u\|$, for every nonzero vector x, we have

$$\frac{\|Ax\|}{\|x\|} = \frac{\|x\| \|A(x/\|x\|)\|}{\|x\|} = \|A(x/\|x\|)\|,$$

which implies that

$$\sup_{\substack{x \in \mathbb{C}^n \\ x \neq 0}} \frac{\|Ax\|}{\|x\|} = \sup_{\substack{x \in \mathbb{C}^n \\ \|x\| = 1}} \|Ax\|.$$

Similarly

$$\sup_{\substack{x\in\mathbb{R}^n\\ x\neq 0}} \frac{\|Ax\|}{\|x\|} = \sup_{\substack{x\in\mathbb{R}^n\\ \|x\|=1}} \|Ax\|\,.$$

The above considerations justify the following definition.

Definition 6.7. If $\|\ \|$ is any norm on $\mathbb{C}^n$, we define the function $\|\ \|$ on $\mathrm{M}_n(\mathbb{C})$ by

$$\|A\| = \sup_{\substack{x\in\mathbb{C}^n\\ x\neq 0}} \frac{\|Ax\|}{\|x\|} = \sup_{\substack{x\in\mathbb{C}^n\\ \|x\|=1}} \|Ax\|\,.$$

The function $A \mapsto \|A\|$ is called the *subordinate matrix norm* or *operator norm* induced by the norm $\|\ \|$.

It is easy to check that the function $A \mapsto \|A\|$ is indeed a norm, and by definition, it satisfies the property

$$\|Ax\| \leq \|A\|\,\|x\|\,, \quad \text{for all } x \in \mathbb{C}^n.$$

A norm $\|\ \|$ on $\mathrm{M}_n(\mathbb{C})$ satisfying the above property is said to be *subordinate* to the vector norm $\|\ \|$ on $\mathbb{C}^n$. As a consequence of the above inequality, we have

$$\|ABx\| \leq \|A\|\,\|Bx\| \leq \|A\|\,\|B\|\,\|x\|\,,$$

for all $x \in \mathbb{C}^n$, which implies that

$$\|AB\| \leq \|A\|\,\|B\| \quad \text{for all } A, B \in \mathrm{M}_n(\mathbb{C}),$$

showing that $A \mapsto \|A\|$ is a matrix norm (it is submultiplicative).

Observe that the operator norm is also defined by

$$\|A\| = \inf\{\lambda \in \mathbb{R} \mid \|Ax\| \leq \lambda\,\|x\|\,, \text{ for all } x \in \mathbb{C}^n\}.$$

Since the function $x \mapsto \|Ax\|$ is continuous (because $|\,\|Ay\| - \|Ax\|\,| \leq \|Ay - Ax\| \leq C_A\,\|x - y\|$) and the unit sphere $S^{n-1} = \{x \in \mathbb{C}^n \mid \|x\| = 1\}$ is compact, there is some $x \in \mathbb{C}^n$ such that $\|x\| = 1$ and

$$\|Ax\| = \|A\|\,.$$

Equivalently, there is some $x \in \mathbb{C}^n$ such that $x \neq 0$ and

$$\|Ax\| = \|A\|\,\|x\|\,.$$

The definition of an operator norm also implies that

$$\|I\| = 1.$$

The above shows that the Frobenius norm is not a subordinate matrix norm (why?). The notion of subordinate norm can be slightly generalized.

Definition 6.8. If $K = \mathbb{R}$ or $K = \mathbb{C}$, for any norm $\|\ \|$ on $\mathrm{M}_{m,n}(K)$, and for any two norms $\|\ \|_a$ on K^n and $\|\ \|_b$ on K^m, we say that the norm $\|\ \|$ is *subordinate* to the norms $\|\ \|_a$ and $\|\ \|_b$ if

$$\|Ax\|_b \leq \|A\|\, \|x\|_a \quad \text{for all } A \in \mathrm{M}_{m,n}(K) \text{ and all } x \in K^n.$$

Remark: For any norm $\|\ \|$ on $\mathbb{C}^n$, we can define the function $\|\ \|_{\mathbb{R}}$ on $\mathrm{M}_n(\mathbb{R})$ by

$$\|A\|_{\mathbb{R}} = \sup_{\substack{x \in \mathbb{R}^n \\ x \neq 0}} \frac{\|Ax\|}{\|x\|} = \sup_{\substack{x \in \mathbb{R}^n \\ \|x\|=1}} \|Ax\|.$$

The function $A \mapsto \|A\|_{\mathbb{R}}$ is a matrix norm on $\mathrm{M}_n(\mathbb{R})$, and

$$\|A\|_{\mathbb{R}} \leq \|A\|,$$

for all real matrices $A \in \mathrm{M}_n(\mathbb{R})$. However, it is possible to construct vector norms $\|\ \|$ on $\mathbb{C}^n$ and *real* matrices A such that

$$\|A\|_{\mathbb{R}} < \|A\|.$$

In order to avoid this kind of difficulties, we define subordinate matrix norms over $\mathrm{M}_n(\mathbb{C})$. Luckily, it turns out that $\|A\|_{\mathbb{R}} = \|A\|$ for the vector norms, $\|\ \|_1$, $\|\ \|_2$, and $\|\ \|_\infty$.

We now prove Proposition 6.4 for real matrix norms.

Proposition 6.7. *For any matrix norm $\|\ \|$ on $\mathrm{M}_n(\mathbb{R})$ and for any square $n \times n$ matrix $A \in \mathrm{M}_n(\mathbb{R})$, we have*

$$\rho(A) \leq \|A\|.$$

Proof. We follow the proof in Denis Serre's book [95]. If A is a real matrix, the problem is that the eigenvectors associated with the eigenvalue of maximum modulus may be complex. We use a trick based on the fact that for every matrix A (real or complex),

$$\rho(A^k) = (\rho(A))^k,$$

which is left as an exercise (use Proposition 12.5 which shows that if $(\lambda_1, \ldots, \lambda_n)$ are the (not necessarily distinct) eigenvalues of A, then $(\lambda_1^k, \ldots, \lambda_n^k)$ are the eigenvalues of A^k, for $k \geq 1$).

Pick any complex matrix norm $\|\ \|_c$ on $\mathbb{C}^n$ (for example, the Frobenius norm, or any subordinate matrix norm induced by a norm on $\mathbb{C}^n$). The restriction of $\|\ \|_c$ to real matrices is a real norm that we also denote by $\|\ \|_c$. Now, by Theorem 6.3, since $\mathrm{M}_n(\mathbb{R})$ has finite dimension n^2, there is some constant $C > 0$ so that

$$\|B\|_c \leq C\, \|B\|, \quad \text{for all} \quad B \in \mathrm{M}_n(\mathbb{R}).$$

Furthermore, for every $k \geq 1$ and for every real $n \times n$ matrix A, by Proposition 6.4, $\rho(A^k) \leq \left\|A^k\right\|_c$, and because $\|\ \|$ is a matrix norm, $\left\|A^k\right\| \leq \|A\|^k$, so we have

$$(\rho(A))^k = \rho(A^k) \leq \left\|A^k\right\|_c \leq C \left\|A^k\right\| \leq C\, \|A\|^k,$$

for all $k \geq 1$. It follows that

$$\rho(A) \leq C^{1/k} \|A\|, \quad \text{for all} \quad k \geq 1.$$

However because $C > 0$, we have $\lim_{k\mapsto\infty} C^{1/k} = 1$ (we have $\lim_{k\mapsto\infty} \frac{1}{k}\log(C) = 0$). Therefore, we conclude that

$$\rho(A) \leq \|A\|,$$

as desired. □

We now determine explicitly what are the subordinate matrix norms associated with the vector norms $\| \|_1, \| \|_2$, and $\| \|_\infty$.

Proposition 6.8. *For every square matrix $A = (a_{ij}) \in \mathrm{M}_n(\mathbb{C})$, we have*

$$\|A\|_1 = \sup_{\substack{x\in\mathbb{C}^n \\ \|x\|_1=1}} \|Ax\|_1 = \max_j \sum_{i=1}^n |a_{ij}|$$

$$\|A\|_\infty = \sup_{\substack{x\in\mathbb{C}^n \\ \|x\|_\infty=1}} \|Ax\|_\infty = \max_i \sum_{j=1}^n |a_{ij}|$$

$$\|A\|_2 = \sup_{\substack{x\in\mathbb{C}^n \\ \|x\|_2=1}} \|Ax\|_2 = \sqrt{\rho(A^*A)} = \sqrt{\rho(AA^*)}.$$

Furthermore, $\|A^*\|_2 = \|A\|_2$, the norm $\| \|_2$ is unitarily invariant, which means that

$$\|A\|_2 = \|UAV\|_2$$

for all unitary matrices U, V, and if A is a normal matrix, then $\|A\|_2 = \rho(A)$.

Proof. For every vector u, we have

$$\|Au\|_1 = \sum_i \left| \sum_j a_{ij}u_j \right| \leq \sum_j |u_j| \sum_i |a_{ij}| \leq \left(\max_j \sum_i |a_{ij}| \right) \|u\|_1,$$

which implies that

$$\|A\|_1 \leq \max_j \sum_{i=1}^n |a_{ij}|.$$

It remains to show that equality can be achieved. For this let j_0 be some index such that

$$\max_j \sum_i |a_{ij}| = \sum_i |a_{ij_0}|,$$

and let $u_i = 0$ for all $i \neq j_0$ and $u_{j_0} = 1$.

In a similar way, we have

$$\|Au\|_\infty = \max_i \left| \sum_j a_{ij} u_j \right| \leq \left(\max_i \sum_j |a_{ij}| \right) \|u\|_\infty ,$$

which implies that

$$\|A\|_\infty \leq \max_i \sum_{j=1}^{n} |a_{ij}|.$$

To achieve equality, let i_0 be some index such that

$$\max_i \sum_j |a_{ij}| = \sum_j |a_{i_0 j}|.$$

The reader should check that the vector given by

$$u_j = \begin{cases} \frac{\overline{a}_{i_0 j}}{|a_{i_0 j}|} & \text{if } a_{i_0 j} \neq 0 \\ 1 & \text{if } a_{i_0 j} = 0 \end{cases}$$

works.

We have

$$\|A\|_2^2 = \sup_{\substack{x \in \mathbb{C}^n \\ x^* x = 1}} \|Ax\|_2^2 = \sup_{\substack{x \in \mathbb{C}^n \\ x^* x = 1}} x^* A^* A x.$$

Since the matrix A^*A is symmetric, it has real eigenvalues and it can be diagonalized with respect to an orthogonal matrix. These facts can be used to prove that the function $x \mapsto x^*A^*Ax$ has a maximum on the sphere $x^*x = 1$ equal to the largest eigenvalue of A^*A, namely, $\rho(A^*A)$. We postpone the proof until we discuss optimizing quadratic functions. Therefore,

$$\|A\|_2 = \sqrt{\rho(A^*A)}.$$

Let use now prove that $\rho(A^*A) = \rho(AA^*)$. First, assume that $\rho(A^*A) > 0$. In this case, there is some eigenvector $u\,(\neq 0)$ such that

$$A^*Au = \rho(A^*A)u,$$

and since $\rho(A^*A) > 0$, we must have $Au \neq 0$. Since $Au \neq 0$,

$$AA^*(Au) = \rho(A^*A)Au$$

which means that $\rho(A^*A)$ is an eigenvalue of AA^*, and thus

$$\rho(A^*A) \leq \rho(AA^*).$$

Because $(A^*)^* = A$, by replacing A by A^*, we get

$$\rho(AA^*) \leq \rho(A^*A),$$

and so $\rho(A^*A) = \rho(AA^*)$.

If $\rho(A^*A) = 0$, then we must have $\rho(AA^*) = 0$, since otherwise by the previous reasoning we would have $\rho(A^*A) = \rho(AA^*) > 0$. Hence, in all case

$$\|A\|_2^2 = \rho(A^*A) = \rho(AA^*) = \|A^*\|_2^2 .$$

For any unitary matrices U and V, it is an easy exercise to prove that V^*A^*AV and A^*A have the same eigenvalues, so

$$\|A\|_2^2 = \rho(A^*A) = \rho(V^*A^*AV) = \|AV\|_2^2 ,$$

and also

$$\|A\|_2^2 = \rho(A^*A) = \rho(A^*U^*UA) = \|UA\|_2^2 .$$

Finally, if A is a normal matrix ($AA^* = A^*A$), it can be shown that there is some unitary matrix U so that

$$A = UDU^*,$$

where $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix consisting of the eigenvalues of A, and thus

$$A^*A = (UDU^*)^*UDU^* = UD^*U^*UDU^* = UD^*DU^*.$$

However, $D^*D = \mathrm{diag}(|\lambda_1|^2, \ldots, |\lambda_n|^2)$, which proves that

$$\rho(A^*A) = \rho(D^*D) = \max_i |\lambda_i|^2 = (\rho(A))^2,$$

so that $\|A\|_2 = \rho(A)$. □

The norm $\|A\|_2$ is often called the *spectral norm*. Observe that property (3) of proposition 6.5 says that

$$\|A\|_2 \leq \|A\|_F \leq \sqrt{n}\, \|A\|_2 ,$$

which shows that the Frobenius norm is an upper bound on the spectral norm. The Frobenius norm is much easier to compute than the spectal norm.

The reader will check that the above proof still holds if the matrix A is real, confirming the fact that $\|A\|_{\mathbb{R}} = \|A\|$ for the vector norms $\| \|_1, \| \|_2$, and $\| \|_\infty$. It is also easy to verify that the proof goes through for *rectangular* matrices, with the same formulae. Similarly, the Frobenius norm is also a norm on rectangular matrices. For these norms, whenever AB makes sense, we have

$$\|AB\| \leq \|A\|\, \|B\| .$$

Remark: Let $(E, \| \|)$ and $(F, \| \|)$ be two normed vector spaces (for simplicity of notation, we use the same symbol $\| \|$ for the norms on E and F; this should not cause any confusion).

Recall that a function $f\colon E \to F$ is *continuous* if for every $a \in E$, for every $\epsilon > 0$, there is some $\eta > 0$ such that for all $x \in E$,

$$\text{if} \quad \|x - a\| \leq \eta \quad \text{then} \quad \|f(x) - f(a)\| \leq \epsilon.$$

It is not hard to show that a *linear map* $f\colon E \to F$ is continuous iff there is some constant $C \geq 0$ such that

$$\|f(x)\| \leq C\,\|x\| \text{ for all } x \in E.$$

If so, we say that f is *bounded* (or a *linear bounded operator*). We let $\mathcal{L}(E;F)$ denote the set of all continuous (equivalently, bounded) linear maps from E to F. Then, we can define the *operator norm* (or *subordinate norm*) $\|\ \|$ on $\mathcal{L}(E;F)$ as follows: for every $f \in \mathcal{L}(E;F)$,

$$\|f\| = \sup_{\substack{x \in E \\ x \neq 0}} \frac{\|f(x)\|}{\|x\|} = \sup_{\substack{x \in E \\ \|x\| = 1}} \|f(x)\|,$$

or equivalently by

$$\|f\| = \inf\{\lambda \in \mathbb{R} \mid \|f(x)\| \leq \lambda\,\|x\|, \text{ for all } x \in E\}.$$

It is not hard to show that the map $f \mapsto \|f\|$ is a norm on $\mathcal{L}(E;F)$ satisfying the property

$$\|f(x)\| \leq \|f\|\,\|x\|$$

for all $x \in E$, and that if $f \in \mathcal{L}(E;F)$ and $g \in \mathcal{L}(F;G)$, then

$$\|g \circ f\| \leq \|g\|\,\|f\|.$$

Operator norms play an important role in functional analysis, especially when the spaces E and F are *complete*.

The following proposition will be needed when we deal with the condition number of a matrix.

Proposition 6.9. *Let $\|\ \|$ be any matrix norm and let B be a matrix such that $\|B\| < 1$.*

(1) If $\|\ \|$ is a subordinate matrix norm, then the matrix $I + B$ is invertible and

$$\left\|(I + B)^{-1}\right\| \leq \frac{1}{1 - \|B\|}.$$

(2) If a matrix of the form $I + B$ is singular, then $\|B\| \geq 1$ for every matrix norm (not necessarily subordinate).

Proof. (1) Observe that $(I + B)u = 0$ implies $Bu = -u$, so

$$\|u\| = \|Bu\|.$$

Recall that

$$\|Bu\| \leq \|B\| \, \|u\|$$

for every subordinate norm. Since $\|B\| < 1$, if $u \neq 0$, then

$$\|Bu\| < \|u\|,$$

which contradicts $\|u\| = \|Bu\|$. Therefore, we must have $u = 0$, which proves that $I + B$ is injective, and thus bijective, i.e., invertible. Then, we have

$$(I + B)^{-1} + B(I + B)^{-1} = (I + B)(I + B)^{-1} = I,$$

so we get

$$(I + B)^{-1} = I - B(I + B)^{-1},$$

which yields

$$\left\|(I + B)^{-1}\right\| \leq 1 + \|B\| \left\|(I + B)^{-1}\right\|,$$

and finally,

$$\left\|(I + B)^{-1}\right\| \leq \frac{1}{1 - \|B\|}.$$

(2) If $I + B$ is singular, then -1 is an eigenvalue of B, and by Proposition 6.4, we get $\rho(B) \leq \|B\|$, which implies $1 \leq \rho(B) \leq \|B\|$. □

The following result is needed to deal with the convergence of sequences of powers of matrices.

Proposition 6.10. *For every matrix $A \in \mathrm{M}_n(\mathbb{C})$ and for every $\epsilon > 0$, there is some subordinate matrix norm $\| \|$ such that*

$$\|A\| \leq \rho(A) + \epsilon.$$

Proof. By Theorem 12.4, there exists some invertible matrix U and some upper triangular matrix T such that

$$A = UTU^{-1},$$

and say that

$$T = \begin{pmatrix} \lambda_1 & t_{12} & t_{13} & \cdots & t_{1n} \\ 0 & \lambda_2 & t_{23} & \cdots & t_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda_{n-1} & t_{n-1\,n} \\ 0 & 0 & \cdots & 0 & \lambda_n \end{pmatrix},$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A. For every $\delta \neq 0$, define the diagonal matrix

$$D_\delta = \mathrm{diag}(1, \delta, \delta^2, \ldots, \delta^{n-1}),$$

and consider the matrix

$$(UD_\delta)^{-1} A (UD_\delta) = D_\delta^{-1} T D_\delta = \begin{pmatrix} \lambda_1 & \delta t_{12} & \delta^2 t_{13} & \cdots & \delta^{n-1} t_{1n} \\ 0 & \lambda_2 & \delta t_{23} & \cdots & \delta^{n-2} t_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda_{n-1} & \delta t_{n-1\,n} \\ 0 & 0 & \cdots & 0 & \lambda_n \end{pmatrix}.$$

Now, define the function $\| \| \colon \mathrm{M}_n(\mathbb{C}) \to \mathbb{R}$ by

$$\|B\| = \left\| (UD_\delta)^{-1} B (UD_\delta) \right\|_\infty,$$

for every $B \in \mathrm{M}_n(\mathbb{C})$. Then it is easy to verify that the above function is the matrix norm subordinate to the vector norm

$$v \mapsto \left\| (UD_\delta)^{-1} v \right\|_\infty.$$

Furthermore, for every $\epsilon > 0$, we can pick δ so that

$$\sum_{j=i+1}^{n} |\delta^{j-i} t_{ij}| \leq \epsilon, \quad 1 \leq i \leq n-1,$$

and by definition of the norm $\| \|_\infty$, we get

$$\|A\| \leq \rho(A) + \epsilon,$$

which shows that the norm that we have constructed satisfies the required properties. □

Note that equality is generally not possible; consider the matrix

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix},$$

for which $\rho(A) = 0 < \|A\|$, since $A \neq 0$.

6.3 Condition Numbers of Matrices

Unfortunately, there exist linear systems $Ax = b$ whose solutions are not stable under small perturbations of either b or A. For example, consider the system

$$\begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}.$$

The reader should check that it has the solution $x = (1, 1, 1, 1)$. If we perturb slightly the right-hand side, obtaining the new system

$$\begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \begin{pmatrix} x_1 + \Delta x_1 \\ x_2 + \Delta x_2 \\ x_3 + \Delta x_3 \\ x_4 + \Delta x_4 \end{pmatrix} = \begin{pmatrix} 32.1 \\ 22.9 \\ 33.1 \\ 30.9 \end{pmatrix},$$

the new solutions turns out to be $x = (9.2, -12.6, 4.5, -1.1)$. In other words, a relative error of the order $1/200$ in the data (here, b) produces a relative error of the order $10/1$ in the solution, which represents an amplification of the relative error of the order 2000.

Now, let us perturb the matrix slightly, obtaining the new system

$$\begin{pmatrix} 10 & 7 & 8.1 & 7.2 \\ 7.08 & 5.04 & 6 & 5 \\ 8 & 5.98 & 9.98 & 9 \\ 6.99 & 4.99 & 9 & 9.98 \end{pmatrix} \begin{pmatrix} x_1 + \Delta x_1 \\ x_2 + \Delta x_2 \\ x_3 + \Delta x_3 \\ x_4 + \Delta x_4 \end{pmatrix} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}.$$

This time, the solution is $x = (-81, 137, -34, 22)$. Again, a small change in the data alters the result rather drastically. Yet, the original system is symmetric, has determinant 1, and has integer entries. The problem is that the matrix of the system is badly conditioned, a concept that we will now explain.

Given an invertible matrix A, first, assume that we perturb b to $b+\delta b$, and let us analyze the change between the two exact solutions x and $x + \delta x$ of the two systems

$$\begin{aligned} Ax &= b \\ A(x + \delta x) &= b + \delta b. \end{aligned}$$

We also assume that we have some norm $\| \|$ and we use the subordinate matrix norm on matrices. From

$$\begin{aligned} Ax &= b \\ Ax + A\delta x &= b + \delta b, \end{aligned}$$

we get

$$\delta x = A^{-1}\delta b,$$

and we conclude that

$$\begin{aligned} \|\delta x\| &\leq \left\|A^{-1}\right\| \|\delta b\| \\ \|b\| &\leq \|A\| \|x\|. \end{aligned}$$

Consequently, the relative error in the result $\|\delta x\| / \|x\|$ is bounded in terms of the relative error $\|\delta b\| / \|b\|$ in the data as follows:

$$\frac{\|\delta x\|}{\|x\|} \leq \left(\|A\| \left\|A^{-1}\right\| \right) \frac{\|\delta b\|}{\|b\|}.$$

Now let us assume that A is perturbed to $A+\delta A$, and let us analyze the change between the exact solutions of the two systems

$$\begin{aligned} Ax &= b \\ (A+\Delta A)(x+\Delta x) &= b. \end{aligned}$$

The second equation yields $Ax + A\Delta x + \Delta A(x+\Delta x) = b$, and by subtracting the first equation we get

$$\Delta x = -A^{-1}\Delta A(x+\Delta x).$$

It follows that

$$\|\Delta x\| \leq \left\|A^{-1}\right\| \|\Delta A\| \|x+\Delta x\|,$$

which can be rewritten as

$$\frac{\|\Delta x\|}{\|x+\Delta x\|} \leq \left(\|A\| \left\|A^{-1}\right\| \right) \frac{\|\Delta A\|}{\|A\|}.$$

Observe that the above reasoning is valid even if the matrix $A+\Delta A$ is singular, as long as $x+\Delta x$ is a solution of the second system. Furthermore, if $\|\Delta A\|$ is small enough, it is not unreasonable to expect that the ratio $\|\Delta x\| / \|x+\Delta x\|$ is close to $\|\Delta x\| / \|x\|$. This will be made more precise later.

In summary, for each of the two perturbations, we see that the relative error in the result is bounded by the relative error in the data, *multiplied the number* $\|A\| \|A^{-1}\|$. In fact, this factor turns out to be optimal and this suggests the following definition:

Definition 6.9. For any subordinate matrix norm $\| \|$, for any invertible matrix A, the number

$$\mathrm{cond}(A) = \|A\| \left\|A^{-1}\right\|$$

is called the *condition number* of A relative to $\| \|$.

The condition number $\mathrm{cond}(A)$ measures the sensitivity of the linear system $Ax = b$ to variations in the data b and A; a feature referred to as the *condition* of the system. Thus, when we says that a linear system is *ill-conditioned*, we mean that the condition number of its matrix is large. We can sharpen the preceding analysis as follows:

Proposition 6.11. *Let A be an invertible matrix and let x and $x+\delta x$ be the solutions of the linear systems*

$$\begin{aligned} Ax &= b \\ A(x+\delta x) &= b+\delta b. \end{aligned}$$

If $b \neq 0$, then the inequality

$$\frac{\|\delta x\|}{\|x\|} \leq \mathrm{cond}(A)\frac{\|\delta b\|}{\|b\|}$$

holds and is the best possible. This means that for a given matrix A, there exist some vectors $b \neq 0$ and $\delta b \neq 0$ for which equality holds.

Proof. We already proved the inequality. Now, because $\| \|$ is a subordinate matrix norm, there exist some vectors $x \neq 0$ and $\delta b \neq 0$ for which

$$\left\|A^{-1}\delta b\right\| = \left\|A^{-1}\right\| \|\delta b\| \quad \text{and} \quad \|Ax\| = \|A\| \|x\|.$$

□

Proposition 6.12. *Let A be an invertible matrix and let x and $x + \Delta x$ be the solutions of the two systems*

$$\begin{aligned} Ax &= b \\ (A + \Delta A)(x + \Delta x) &= b. \end{aligned}$$

If $b \neq 0$, then the inequality

$$\frac{\|\Delta x\|}{\|x + \Delta x\|} \leq \operatorname{cond}(A) \frac{\|\Delta A\|}{\|A\|}$$

holds and is the best possible. This means that given a matrix A, there exist a vector $b \neq 0$ and a matrix $\Delta A \neq 0$ for which equality holds. Furthermore, if $\|\Delta A\|$ is small enough (for instance, if $\|\Delta A\| < 1/\|A^{-1}\|$), we have

$$\frac{\|\Delta x\|}{\|x\|} \leq \operatorname{cond}(A) \frac{\|\Delta A\|}{\|A\|}(1 + O(\|\Delta A\|));$$

in fact, we have

$$\frac{\|\Delta x\|}{\|x\|} \leq \operatorname{cond}(A) \frac{\|\Delta A\|}{\|A\|} \left(\frac{1}{1 - \|A^{-1}\| \|\Delta A\|}\right).$$

Proof. The first inequality has already been proved. To show that equality can be achieved, let w be any vector such that $w \neq 0$ and

$$\left\|A^{-1}w\right\| = \left\|A^{-1}\right\| \|w\|,$$

and let $\beta \neq 0$ be any real number. Now, the vectors

$$\begin{aligned} \Delta x &= -\beta A^{-1} w \\ x + \Delta x &= w \\ b &= (A + \beta I) w \end{aligned}$$

and the matrix

$$\Delta A = \beta I$$

sastisfy the equations

$$\begin{aligned} Ax &= b \\ (A + \Delta A)(x + \Delta x) &= b \\ \|\Delta x\| = |\beta| \left\|A^{-1}w\right\| &= \|\Delta A\| \left\|A^{-1}\right\| \|x + \Delta x\|. \end{aligned}$$

Finally, we can pick β so that $-\beta$ is not equal to any of the eigenvalues of A, so that $A + \Delta A = A + \beta I$ is invertible and b is is nonzero.

If $\|\Delta A\| < 1/\|A^{-1}\|$, then

$$\left\|A^{-1}\Delta A\right\| \leq \left\|A^{-1}\right\| \|\Delta A\| < 1,$$

so by Proposition 6.9, the matrix $I + A^{-1}\Delta A$ is invertible and

$$\left\|(I + A^{-1}\Delta A)^{-1}\right\| \leq \frac{1}{1 - \|A^{-1}\Delta A\|} \leq \frac{1}{1 - \|A^{-1}\| \|\Delta A\|}.$$

Recall that we proved earlier that

$$\Delta x = -A^{-1}\Delta A(x + \Delta x),$$

and by adding x to both sides and moving the right-hand side to the left-hand side yields

$$(I + A^{-1}\Delta A)(x + \Delta x) = x,$$

and thus

$$x + \Delta x = (I + A^{-1}\Delta A)^{-1}x,$$

which yields

$$\begin{aligned}\Delta x &= ((I + A^{-1}\Delta A)^{-1} - I)x = (I + A^{-1}\Delta A)^{-1}(I - (I + A^{-1}\Delta A))x \\ &= -(I + A^{-1}\Delta A)^{-1}A^{-1}(\Delta A)x.\end{aligned}$$

From this and

$$\left\|(I + A^{-1}\Delta A)^{-1}\right\| \leq \frac{1}{1 - \|A^{-1}\| \|\Delta A\|},$$

we get

$$\|\Delta x\| \leq \frac{\|A^{-1}\| \|\Delta A\|}{1 - \|A^{-1}\| \|\Delta A\|} \|x\|,$$

which can be written as

$$\frac{\|\Delta x\|}{\|x\|} \leq \mathrm{cond}(A)\frac{\|\Delta A\|}{\|A\|}\left(\frac{1}{1 - \|A^{-1}\| \|\Delta A\|}\right),$$

which is the kind of inequality that we were seeking. □

Remark: If A and b are perturbed simultaneously, so that we get the "perturbed" system

$$(A + \Delta A)(x + \delta x) = b + \delta b,$$

it can be shown that if $\|\Delta A\| < 1/\|A^{-1}\|$ (and $b \neq 0$), then

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\mathrm{cond}(A)}{1 - \|A^{-1}\|\,\|\Delta A\|}\left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right);$$

see Demmel [33], Section 2.2 and Horn and Johnson [55], Section 5.8.

We now list some properties of condition numbers and figure out what $\mathrm{cond}(A)$ is in the case of the spectral norm (the matrix norm induced by $\|\ \|_2$). First, we need to introduce a very important factorization of matrices, the *singular value decomposition*, for short, *SVD*.

It can be shown that given any $n \times n$ matrix $A \in \mathrm{M}_n(\mathbb{C})$, there exist two unitary matrices U and V, and a *real* diagonal matrix $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$, with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$, such that

$$A = V\Sigma U^*.$$

The nonnegative numbers $\sigma_1, \ldots, \sigma_n$ are called the *singular values* of A.

If A is a real matrix, the matrices U and V are orthogonal matrices. The factorization $A = V\Sigma U^*$ implies that

$$A^*A = U\Sigma^2 U^* \quad \text{and} \quad AA^* = V\Sigma^2 V^*,$$

which shows that $\sigma_1^2, \ldots, \sigma_n^2$ are the eigenvalues of *both* A^*A and AA^*, that the columns of U are corresponding eivenvectors for A^*A, and that the columns of V are corresponding eivenvectors for AA^*. Since σ_1^2 is the largest eigenvalue of A^*A (and AA^*), note that $\sqrt{\rho(A^*A)} = \sqrt{\rho(AA^*)} = \sigma_1$; that is, *the spectral norm $\|A\|_2$ of a matrix A is equal to the largest singular value of A.* Equivalently, the spectral norm $\|A\|_2$ of a matrix A is equal to the ℓ_∞-norm of its vector of singular values,

$$\|A\|_2 = \max_{1 \leq i \leq n} \sigma_i = \|(\sigma_1, \ldots, \sigma_n)\|_\infty.$$

Since the Frobenius norm of a matrix A is defined by $\|A\|_F = \sqrt{\mathrm{tr}(A^*A)}$ and since

$$\mathrm{tr}(A^*A) = \sigma_1^2 + \cdots + \sigma_n^2$$

where $\sigma_1^2, \ldots, \sigma_n^2$ are the eigenvalues of A^*A, we see that

$$\|A\|_F = (\sigma_1^2 + \cdots + \sigma_n^2)^{1/2} = \|(\sigma_1, \ldots, \sigma_n)\|_2.$$

This shows that *the Frobenius norm of a matrix is given by the ℓ_2-norm of its vector of singular values.*

In the case of a normal matrix if $\lambda_1, \ldots, \lambda_n$ are the (complex) eigenvalues of A, then

$$\sigma_i = |\lambda_i|, \quad 1 \leq i \leq n.$$

Proposition 6.13. *For every invertible matrix $A \in \mathrm{M}_n(\mathbb{C})$, the following properties hold:*

(1)

$$\begin{aligned}\operatorname{cond}(A) &\geq 1,\\ \operatorname{cond}(A) &= \operatorname{cond}(A^{-1})\\ \operatorname{cond}(\alpha A) &= \operatorname{cond}(A) \quad \textit{for all } \alpha \in \mathbb{C} - \{0\}.\end{aligned}$$

(2) If $\operatorname{cond}_2(A)$ *denotes the condition number of* A *with respect to the spectral norm, then*

$$\operatorname{cond}_2(A) = \frac{\sigma_1}{\sigma_n},$$

where $\sigma_1 \geq \cdots \geq \sigma_n$ *are the singular values of* A*.*

(3) If the matrix A *is normal, then*

$$\operatorname{cond}_2(A) = \frac{|\lambda_1|}{|\lambda_n|},$$

where $\lambda_1, \ldots, \lambda_n$ *are the eigenvalues of* A *sorted so that* $|\lambda_1| \geq \cdots \geq |\lambda_n|$*.*

(4) If A *is a unitary or an orthogonal matrix, then*

$$\operatorname{cond}_2(A) = 1.$$

(5) The condition number $\operatorname{cond}_2(A)$ *is invariant under unitary transformations, which means that*

$$\operatorname{cond}_2(A) = \operatorname{cond}_2(UA) = \operatorname{cond}_2(AV),$$

for all unitary matrices U *and* V*.*

Proof. The properties in (1) are immediate consequences of the properties of subordinate matrix norms. In particular, $AA^{-1} = I$ implies

$$1 = \|I\| \leq \|A\| \left\|A^{-1}\right\| = \operatorname{cond}(A).$$

(2) We showed earlier that $\|A\|_2^2 = \rho(A^*A)$, which is the square of the modulus of the largest eigenvalue of A^*A. Since we just saw that the eigenvalues of A^*A are $\sigma_1^2 \geq \cdots \geq \sigma_n^2$, where $\sigma_1, \ldots, \sigma_n$ are the singular values of A, we have

$$\|A\|_2 = \sigma_1.$$

Now, if A is invertible, then $\sigma_1 \geq \cdots \geq \sigma_n > 0$, and it is easy to show that the eigenvalues of $(A^*A)^{-1}$ are $\sigma_n^{-2} \geq \cdots \geq \sigma_1^{-2}$, which shows that

$$\left\|A^{-1}\right\|_2 = \sigma_n^{-1},$$

and thus

$$\mathrm{cond}_2(A) = \frac{\sigma_1}{\sigma_n}.$$

(3) This follows from the fact that $\|A\|_2 = \rho(A)$ for a normal matrix.

(4) If A is a unitary matrix, then $A^*A = AA^* = I$, so $\rho(A^*A) = 1$, and $\|A\|_2 = \sqrt{\rho(A^*A)} = 1$. We also have $\|A^{-1}\|_2 = \|A^*\|_2 = \sqrt{\rho(AA^*)} = 1$, and thus $\mathrm{cond}(A) = 1$.

(5) This follows immediately from the unitary invariance of the spectral norm. □

Proposition 6.13 (4) shows that unitary and orthogonal transformations are very well-conditioned, and part (5) shows that unitary transformations preserve the condition number.

In order to compute $\mathrm{cond}_2(A)$, we need to compute the top and bottom singular values of A, which may be hard. The inequality

$$\|A\|_2 \leq \|A\|_F \leq \sqrt{n}\,\|A\|_2\,,$$

may be useful in getting an approximation of $\mathrm{cond}_2(A) = \|A\|_2\,\|A^{-1}\|_2$, if A^{-1} can be determined.

Remark: There is an interesting geometric characterization of $\mathrm{cond}_2(A)$. If $\theta(A)$ denotes the least angle between the vectors Au and Av as u and v range over all pairs of orthonormal vectors, then it can be shown that

$$\mathrm{cond}_2(A) = \cot(\theta(A)/2)).$$

Thus, if A is nearly singular, then there will be some orthonormal pair u, v such that Au and Av are nearly parallel; the angle $\theta(A)$ will the be small and $\cot(\theta(A)/2))$ will be large. For more details, see Horn and Johnson [55] (Section 5.8 and Section 7.4).

It should be noted that in general (if A is not a normal matrix) a matrix could have a very large condition number even if all its eigenvalues are identical! For example, if we consider the $n \times n$ matrix

$$A = \begin{pmatrix} 1 & 2 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 2 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & 2 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 1 & 2 & 0 \\ 0 & 0 & \dots & 0 & 0 & 1 & 2 \\ 0 & 0 & \dots & 0 & 0 & 0 & 1 \end{pmatrix},$$

it turns out that $\mathrm{cond}_2(A) \geq 2^{n-1}$.

A classical example of matrix with a very large condition number is the *Hilbert matrix* $H^{(n)}$, the $n \times n$ matrix with

$$H^{(n)}_{ij} = \left(\frac{1}{i+j-1}\right).$$

For example, when $n = 5$,

$$H^{(5)} = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} \\ \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} & \frac{1}{9} \end{pmatrix}.$$

It can be shown that

$$\mathrm{cond}_2(H^{(5)}) \approx 4.77 \times 10^5.$$

Hilbert introduced these matrices in 1894 while studying a problem in approximation theory. The Hilbert matrix $H^{(n)}$ is symmetric positive definite. A closed-form formula can be given for its determinant (it is a special form of the so-called *Cauchy determinant*). The inverse of $H^{(n)}$ can also be computed explicitly! It can be shown that

$$\mathrm{cond}_2(H^{(n)}) = O((1+\sqrt{2})^{4n}/\sqrt{n}).$$

Going back to our matrix

$$A = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix},$$

which is a symmetric, positive, definite, matrix, it can be shown that its eigenvalues, which in this case are also its singular values because A is SPD, are

$$\lambda_1 \approx 30.2887 > \lambda_2 \approx 3.858 > \lambda_3 \approx 0.8431 > \lambda_4 \approx 0.01015,$$

so that

$$\mathrm{cond}_2(A) = \frac{\lambda_1}{\lambda_4} \approx 2984.$$

The reader should check that for the perturbation of the right-hand side b used earlier, the relative errors $\|\delta x\|/\|x\|$ and $\|\delta x\|/\|x\|$ satisfy the inequality

$$\frac{\|\delta x\|}{\|x\|} \leq \mathrm{cond}(A)\frac{\|\delta b\|}{\|b\|}$$

and comes close to equality.

6.4 An Application of Norms: Solving Inconsistent Linear Systems

The problem of solving an inconsistent linear system $Ax = b$ often arises in practice. This is a system where b does not belong to the column space of A, usually with more equations than variables. Thus, such a system has no solution. Yet, we would still like to "solve" such a system, at least approximately.

Such systems often arise when trying to fit some data. For example, we may have a set of 3D data points

$$\{p_1, \ldots, p_n\},$$

and we have reason to believe that these points are nearly coplanar. We would like to find a plane that best fits our data points. Recall that the equation of a plane is

$$\alpha x + \beta y + \gamma z + \delta = 0,$$

with $(\alpha, \beta, \gamma) \neq (0, 0, 0)$. Thus, every plane is either not parallel to the x-axis ($\alpha \neq 0$) or not parallel to the y-axis ($\beta \neq 0$) or not parallel to the z-axis ($\gamma \neq 0$).

Say we have reasons to believe that the plane we are looking for is not parallel to the z-axis. If we are wrong, in the least squares solution, one of the coefficients, α, β, will be very large. If $\gamma \neq 0$, then we may assume that our plane is given by an equation of the form

$$z = ax + by + d,$$

and we would like this equation to be satisfied for all the p_i's, which leads to a system of n equations in 3 unknowns a, b, d, with $p_i = (x_i, y_i, z_i)$;

$$\begin{aligned} ax_1 + by_1 + d &= z_1 \\ \vdots \qquad & \quad \vdots \\ ax_n + by_n + d &= z_n. \end{aligned}$$

However, if n is larger than 3, such a system generally has *no solution*. Since the above system can't be solved exactly, we can try to find a solution (a, b, d) that *minimizes the least-squares error*

$$\sum_{i=1}^{n} (ax_i + by_i + d - z_i)^2.$$

This is what Legendre and Gauss figured out in the early 1800's!

In general, given a linear system

$$Ax = b,$$

we solve the *least squares problem*: minimize $\|Ax - b\|_2^2$.

Fortunately, every $n \times m$-matrix A can be written as

$$A = VDU^\top$$

where U and V are orthogonal and D is a rectangular diagonal matrix with non-negative entries (*singular value decomposition, or SVD*); see Chapter 15.

The SVD can be used to solve an inconsistent system. It is shown in Chapter 16 that there is a vector x of smallest norm minimizing $\|Ax - b\|_2$. It is given by the (Penrose) *pseudo-inverse* of A (itself given by the SVD).

It has been observed that solving in the least-squares sense may give too much weight to "outliers," that is, points clearly outside the best-fit plane. In this case, it is preferable to minimize (the ℓ_1-norm)

$$\sum_{i=1}^{n} |ax_i + by_i + d - z_i|.$$

This does not appear to be a linear problem, but we can use a trick to convert this minimization problem into a linear program (which means a problem involving linear constraints).

Note that $|x| = \max\{x, -x\}$. So, by introducing new variables $e_1, \ldots, e_n$, our minimization problem is equivalent to the linear program (LP):

$$\begin{array}{ll} \text{minimize} & e_1 + \cdots + e_n \\ \text{subject to} & ax_i + by_i + d - z_i \leq e_i \\ & -(ax_i + by_i + d - z_i) \leq e_i \\ & 1 \leq i \leq n. \end{array}$$

Observe that the constraints are equivalent to

$$e_i \geq |ax_i + by_i + d - z_i|, \qquad 1 \leq i \leq n.$$

For an optimal solution, we must have equality, since otherwise we could decrease some e_i and get an even better solution. Of course, we are no longer dealing with "pure" linear algebra, since our constraints are inequalities.

We prefer not getting into linear programming right now, but the above example provides a good reason to learn more about linear programming!

6.5 Summary

The main concepts and results of this chapter are listed below:

- *Norms* and *normed vector spaces*.

- The *triangle inequality*.
- The *Euclidean norm*; the ℓ_p*-norms*.
- *Hölder's inequality*; the *Cauchy–Schwarz inequality*; *Minkowski's inequality*.
- *Hermitian inner product* and *Euclidean inner product*.
- *Equivalent* norms.
- *All norms on a finite-dimensional vector space are equivalent* (Theorem 6.3).
- *Matrix norms*.
- *Hermitian, symmetric* and *normal* matrices. *Orthogonal* and *unitary* matrices.
- The *trace* of a matrix.
- *Eigenvalues* and *eigenvectors* of a matrix.
- The *characteristic polynomial* of a matrix.
- The *spectral radius* $\rho(A)$ of a matrix A.
- The *Frobenius norm*.
- The Frobenius norm is a *unitarily invariant* matrix norm.
- *Bounded* linear maps.
- *Subordinate matrix norms*.
- Characterization of the subordinate matrix norms for the vector norms $\| \|_1$, $\| \|_2$, and $\| \|_\infty$.
- The *spectral norm*.
- For every matrix $A \in \mathrm{M}_n(\mathbb{C})$ and for every $\epsilon > 0$, there is some subordinate matrix norm $\| \|$ such that $\|A\| \leq \rho(A) + \epsilon$.
- *Condition numbers* of matrices.
- Perturbation analysis of linear systems.
- The *singular value decomposition* (SVD).
- Properties of conditions numbers. Characterization of $\mathrm{cond}_2(A)$ in terms of the largest and smallest singular values of A.
- The *Hilbert matrix*: a very badly conditioned matrix.
- Solving inconsistent linear systems by the method of *least-squares*; *linear programming*.

Chapter 7

Iterative Methods for Solving Linear Systems

7.1 Convergence of Sequences of Vectors and Matrices

In Chapter 5 we have discussed some of the main methods for solving systems of linear equations. These methods are *direct methods*, in the sense that they yield exact solutions (assuming infinite precision!).

Another class of methods for solving linear systems consists in approximating solutions using *iterative methods*. The basic idea is this: Given a linear system $Ax = b$ (with A a square invertible matrix), find another matrix B and a vector c, such that

1. The matrix $I - B$ is invertible
2. The unique solution $\widetilde{x}$ of the system $Ax = b$ is identical to the unique solution $\widetilde{u}$ of the system
$$u = Bu + c,$$

and then, starting from any vector u_0, compute the sequence (u_k) given by

$$u_{k+1} = Bu_k + c, \quad k \in \mathbb{N}.$$

Under certain conditions (to be clarified soon), the sequence (u_k) converges to a limit $\widetilde{u}$ which is the unique solution of $u = Bu + c$, and thus of $Ax = b$.

Consequently, it is important to find conditions that ensure the convergence of the above sequences and to have tools to compare the "rate" of convergence of these sequences. Thus, we begin with some general results about the convergence of sequences of vectors and matrices.

Let $(E, \| \|)$ be a normed vector space. Recall that a sequence (u_k) of vectors $u_k \in E$ *converges to a limit* $u \in E$, if for every $\epsilon > 0$, there some natural number N such that

$$\|u_k - u\| \leq \epsilon, \quad \textit{for all } k \geq N.$$

We write

$$u = \lim_{k \mapsto \infty} u_k.$$

If E is a finite-dimensional vector space and $\dim(E) = n$, we know from Theorem 6.3 that any two norms are equivalent, and if we choose the norm $\| \|_\infty$, we see that the convergence of the sequence of vectors u_k is equivalent to the convergence of the n sequences of scalars formed by the components of these vectors (over any basis). The same property applies to the finite-dimensional vector space $\mathrm{M}_{m,n}(K)$ of $m \times n$ matrices (with $K = \mathbb{R}$ or $K = \mathbb{C}$), which means that the convergence of a sequence of matrices $A_k = (a_{ij}^{(k)})$ is equivalent to the convergence of the $m \times n$ sequences of scalars $(a_{ij}^{(k)})$, with i, j fixed $(1 \leq i \leq m,\ 1 \leq j \leq n)$.

The first theorem below gives a necessary and sufficient condition for the sequence (B^k) of powers of a matrix B to converge to the zero matrix. Recall that the spectral radius $\rho(B)$ of a matrix B is the maximum of the moduli $|\lambda_i|$ of the eigenvalues of B.

Theorem 7.1. *For any square matrix B, the following conditions are equivalent:*

(1) $\lim_{k \mapsto \infty} B^k = 0$,

(2) $\lim_{k \mapsto \infty} B^k v = 0$, *for all vectors v,*

(3) $\rho(B) < 1$,

(4) $\|B\| < 1$, *for some subordinate matrix norm $\| \|$.*

Proof. Assume (1) and let $\| \|$ be a vector norm on E and $\| \|$ be the corresponding matrix norm. For every vector $v \in E$, because $\| \|$ is a matrix norm, we have

$$\|B^k v\| \leq \|B^k\| \|v\|,$$

and since $\lim_{k \mapsto \infty} B^k = 0$ means that $\lim_{k \mapsto \infty} \|B^k\| = 0$, we conclude that $\lim_{k \mapsto \infty} \|B^k v\| = 0$, that is, $\lim_{k \mapsto \infty} B^k v = 0$. This proves that (1) implies (2).

Assume (2). If We had $\rho(B) \geq 1$, then there would be some eigenvector u $(\neq 0)$ and some eigenvalue λ such that

$$Bu = \lambda u, \quad |\lambda| = \rho(B) \geq 1,$$

but then the sequence $(B^k u)$ would not converge to 0, because $B^k u = \lambda^k u$ and $|\lambda^k| = |\lambda|^k \geq 1$. It follows that (2) implies (3).

Assume that (3) holds, that is, $\rho(B) < 1$. By Proposition 6.10, we can find $\epsilon > 0$ small enough that $\rho(B) + \epsilon < 1$, and a subordinate matrix norm $\| \|$ such that

$$\|B\| \leq \rho(B) + \epsilon,$$

which is (4).

Finally, assume (4). Because $\| \|$ is a matrix norm,

$$\|B^k\| \leq \|B\|^k,$$

and since $\|B\| < 1$, we deduce that (1) holds. □

The following proposition is needed to study the rate of convergence of iterative methods.

Proposition 7.2. *For every square matrix B and every matrix norm $\| \|$, we have*

$$\lim_{k \mapsto \infty} \|B^k\|^{1/k} = \rho(B).$$

Proof. We know from Proposition 6.4 that $\rho(B) \leq \|B\|$, and since $\rho(B) = (\rho(B^k))^{1/k}$, we deduce that

$$\rho(B) \leq \|B^k\|^{1/k} \quad \text{for all } k \geq 1,$$

and so

$$\rho(B) \leq \lim_{k \mapsto \infty} \|B^k\|^{1/k}.$$

Now, let us prove that for every $\epsilon > 0$, there is some integer $N(\epsilon)$ such that

$$\|B^k\|^{1/k} \leq \rho(B) + \epsilon \quad \text{for all } k \geq N(\epsilon),$$

which proves that

$$\lim_{k \mapsto \infty} \|B^k\|^{1/k} \leq \rho(B),$$

and our proposition.

For any given $\epsilon > 0$, let B_ϵ be the matrix

$$B_\epsilon = \frac{B}{\rho(B) + \epsilon}.$$

Since $\|B_\epsilon\| < 1$, Theorem 7.1 implies that $\lim_{k \mapsto \infty} B_\epsilon^k = 0$. Consequently, there is some integer $N(\epsilon)$ such that for all $k \geq N(\epsilon)$, we have

$$\|B^k\| = \frac{\|B^k\|}{(\rho(B) + \epsilon)^k} \leq 1,$$

which implies that

$$\|B^k\|^{1/k} \leq \rho(B) + \epsilon,$$

as claimed. □

We now apply the above results to the convergence of iterative methods.

7.2 Convergence of Iterative Methods

Recall that iterative methods for solving a linear system $Ax = b$ (with A invertible) consists in finding some matrix B and some vector c, such that $I - B$ is invertible, and the unique solution $\widetilde{x}$ of $Ax = b$ is equal to the unique solution $\widetilde{u}$ of $u = Bu + c$. Then, starting from any vector u_0, compute the sequence (u_k) given by

$$u_{k+1} = Bu_k + c, \quad k \in \mathbb{N},$$

and say that the iterative method is *convergent* iff

$$\lim_{k \mapsto \infty} u_k = \widetilde{u},$$

for *every* initial vector u_0.

Here is a fundamental criterion for the convergence of any iterative methods based on a matrix B, called the *matrix of the iterative method*.

Theorem 7.3. *Given a system $u = Bu + c$ as above, where $I - B$ is invertible, the following statements are equivalent:*

(1) The iterative method is convergent.

(2) $\rho(B) < 1$.

(3) $\|B\| < 1$, for some subordinate matrix norm $\| \|$.

Proof. Define the vector e_k (*error vector*) by

$$e_k = u_k - \widetilde{u},$$

where $\widetilde{u}$ is the unique solution of the system $u = Bu + c$. Clearly, the iterative method is convergent iff

$$\lim_{k \mapsto \infty} e_k = 0.$$

We claim that

$$e_k = B^k e_0, \quad k \geq 0,$$

where $e_0 = u_0 - \widetilde{u}$.

This is proved by induction on k. The base case $k = 0$ is trivial. By the induction hypothesis, $e_k = B^k e_0$, and since $u_{k+1} = Bu_k + c$, we get

$$u_{k+1} - \widetilde{u} = Bu_k + c - \widetilde{u},$$

and because $\widetilde{u} = B\widetilde{u} + c$ and $e_k = B^k e_0$ (by the induction hypothesis), we obtain

$$u_{k+1} - \widetilde{u} = Bu_k - B\widetilde{u} = B(u_k - \widetilde{u}) = Be_k = BB^k e_0 = B^{k+1} e_0,$$

proving the induction step. Thus, the iterative method converges iff

$$\lim_{k \mapsto \infty} B^k e_0 = 0.$$

Consequently, our theorem follows by Theorem 7.1. □

The next proposition is needed to compare the rate of convergence of iterative methods. It shows that *asymptotically, the error vector $e_k = B^k e_0$ behaves at worst like $(\rho(B))^k$.*

Proposition 7.4. *Let $\|\ \|$ be any vector norm, let B be a matrix such that $I - B$ is invertible, and let $\widetilde{u}$ be the unique solution of $u = Bu + c$.*

(1) If (u_k) is any sequence defined iteratively by

$$u_{k+1} = Bu_k + c, \quad k \in \mathbb{N},$$

then

$$\lim_{k \mapsto \infty} \left[\sup_{\|u_0 - \widetilde{u}\| = 1} \|u_k - \widetilde{u}\|^{1/k} \right] = \rho(B).$$

(2) Let B_1 and B_2 be two matrices such that $I - B_1$ and $I - B_2$ are invertibe, assume that both $u = B_1 u + c_1$ and $u = B_2 u + c_2$ have the same unique solution $\widetilde{u}$, and consider any two sequences (u_k) and (v_k) defined inductively by

$$u_{k+1} = B_1 u_k + c_1$$
$$v_{k+1} = B_2 v_k + c_2,$$

with $u_0 = v_0$. If $\rho(B_1) < \rho(B_2)$, then for any $\epsilon > 0$, there is some integer $N(\epsilon)$, such that for all $k \geq N(\epsilon)$, we have

$$\sup_{\|u_0 - \widetilde{u}\| = 1} \left[\frac{\|v_k - \widetilde{u}\|}{\|u_k - \widetilde{u}\|} \right]^{1/k} \geq \frac{\rho(B_2)}{\rho(B_1) + \epsilon}.$$

Proof. Let $\|\ \|$ be the subordinate matrix norm. Recall that

$$u_k - \widetilde{u} = B^k e_0,$$

with $e_0 = u_0 - \widetilde{u}$. For every $k \in \mathbb{N}$, we have

$$(\rho(B_1))^k = \rho(B_1^k) \leq \|B_1^k\| = \sup_{\|e_0\| = 1} \|B_1^k e_0\|,$$

which implies

$$\rho(B_1) = \sup_{\|e_0\| = 1} \|B_1^k e_0\|^{1/k} = \|B_1^k\|^{1/k},$$

and statement (1) follows from Proposition 7.2.

Because $u_0 = v_0$, we have

$$u_k - \widetilde{u} = B_1^k e_0$$
$$v_k - \widetilde{u} = B_2^k e_0,$$

with $e_0 = u_0 - \widetilde{u} = v_0 - \widetilde{u}$. Again, by Proposition 7.2, for every $\epsilon > 0$, there is some natural number $N(\epsilon)$ such that if $k \geq N(\epsilon)$, then

$$\sup_{\|e_0\|=1} \|B_1^k e_0\|^{1/k} \leq \rho(B_1) + \epsilon.$$

Furthermore, for all $k \geq N(\epsilon)$, there exists a vector $e_0 = e_0(k)$ such that

$$\|e_0\| = 1 \quad \text{and} \quad \|B_2^k e_0\|^{1/k} = \|B_2^k\|^{1/k} \geq \rho(B_2),$$

which implies statement (2). □

In light of the above, we see that when we investigate new iterative methods, we have to deal with the following two problems:

1. Given an iterative method with matrix B, determine whether the method is convergent. This involves determining whether $\rho(B) < 1$, or equivalently whether there is a subordinate matrix norm such that $\|B\| < 1$. By Proposition 6.9, this implies that $I - B$ is invertible (since $\| - B\| = \|B\|$, Proposition 6.9 applies).

2. Given two convergent iterative methods, compare them. The iterative method which is faster is that whose matrix has the smaller spectral radius.

We now discuss three iterative methods for solving linear systems:

1. Jacobi's method

2. Gauss-Seidel's method

3. The relaxation method.

7.3 Description of the Methods of Jacobi, Gauss-Seidel, and Relaxation

The methods described in this section are instances of the following scheme: Given a linear system $Ax = b$, with A invertible, suppose we can write A in the form

$$A = M - N,$$

with M invertible, and "easy to invert," which means that M is close to being a diagonal or a triangular matrix (perhaps by blocks). Then, $Au = b$ is equivalent to

$$Mu = Nu + b,$$

that is,

$$u = M^{-1}Nu + M^{-1}b.$$

Therefore, we are in the situation described in the previous sections with $B = M^{-1}N$ and $c = M^{-1}b$. In fact, since $A = M - N$, we have

$$B = M^{-1}N = M^{-1}(M - A) = I - M^{-1}A,$$

which shows that $I - B = M^{-1}A$ is invertible. The iterative method associated with the matrix $B = M^{-1}N$ is given by

$$u_{k+1} = M^{-1}Nu_k + M^{-1}b, \quad k \geq 0,$$

starting from any arbitrary vector u_0. From a practical point of view, we do not invert M, and instead we solve iteratively the systems

$$Mu_{k+1} = Nu_k + b, \quad k \geq 0.$$

Various methods correspond to various ways of choosing M and N from A. The first two methods choose M and N as disjoint submatrices of A, but the relaxation method allows some overlapping of M and N.

To describe the various choices of M and N, it is convenient to write A in terms of three submatrices D, E, F, as

$$A = D - E - F,$$

where the only nonzero entries in D are the diagonal entries in A, the only nonzero entries in E are entries in A below the the diagonal, and the only nonzero entries in F are entries in A above the diagonal. More explicitly, if

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1\,n-1} & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n-1} & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n-1} & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n-1\,1} & a_{n-1\,2} & a_{n-1\,3} & \cdots & a_{n-1\,n-1} & a_{n-1\,n} \\ a_{n\,1} & a_{n\,2} & a_{n\,3} & \cdots & a_{n\,n-1} & a_{n\,n} \end{pmatrix},$$

then

$$D = \begin{pmatrix} a_{11} & 0 & 0 & \cdots & 0 & 0 \\ 0 & a_{22} & 0 & \cdots & 0 & 0 \\ 0 & 0 & a_{33} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{n-1\,n-1} & 0 \\ 0 & 0 & 0 & \cdots & 0 & a_{n\,n} \end{pmatrix},$$

$$-E = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ a_{21} & 0 & 0 & \cdots & 0 & 0 \\ a_{31} & a_{32} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ a_{n-1\,1} & a_{n-1\,2} & a_{n-1\,3} & \ddots & 0 & 0 \\ a_{n\,1} & a_{n\,2} & a_{n\,3} & \cdots & a_{n\,n-1} & 0 \end{pmatrix}, \quad -F = \begin{pmatrix} 0 & a_{12} & a_{13} & \cdots & a_{1\,n-1} & a_{1n} \\ 0 & 0 & a_{23} & \cdots & a_{2\,n-1} & a_{2n} \\ 0 & 0 & 0 & \ddots & a_{3\,n-1} & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & a_{n-1\,n} \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix}.$$

In *Jacobi's method*, we assume that all diagonal entries in A are nonzero, and we pick

$$M = D$$
$$N = E + F,$$

so that

$$B = M^{-1}N = D^{-1}(E + F) = I - D^{-1}A.$$

As a matter of notation, we let

$$J = I - D^{-1}A = D^{-1}(E + F),$$

which is called *Jacobi's matrix*. The corresponding method, *Jacobi's iterative method*, computes the sequence (u_k) using the recurrence

$$u_{k+1} = D^{-1}(E + F)u_k + D^{-1}b, \quad k \geq 0.$$

In practice, we iteratively solve the systems

$$Du_{k+1} = (E + F)u_k + b, \quad k \geq 0.$$

If we write $u_k = (u_1^k, \ldots, u_n^k)$, we solve iteratively the following system:

$$\begin{array}{rcccccccc} a_{11}u_1^{k+1} & = & & -a_{12}u_2^k & -a_{13}u_3^k & \cdots & -a_{1n}u_n^k & + b_1 \\ a_{22}u_2^{k+1} & = & -a_{21}u_1^k & & -a_{23}u_3^k & \cdots & -a_{2n}u_n^k & + b_2 \\ \vdots & \vdots & \vdots & & & & & \\ a_{n-1\,n-1}u_{n-1}^{k+1} & = & -a_{n-1\,1}u_1^k & \cdots & -a_{n-1\,n-2}u_{n-2}^k & & -a_{n-1\,n}u_n^k & + b_{n-1} \\ a_{n\,n}u_n^{k+1} & = & -a_{n\,1}u_1^k & -a_{n\,2}u_2^k & \cdots & -a_{n\,n-1}u_{n-1}^k & & + b_n \end{array}.$$

Observe that we can try to "speed up" the method by using the new value u_1^{k+1} instead of u_1^k in solving for u_2^{k+2} using the second equations, and more generally, use $u_1^{k+1}, \ldots, u_{i-1}^{k+1}$ instead of $u_1^k, \ldots, u_{i-1}^k$ in solving for u_i^{k+1} in the ith equation. This observation leads to the system

$$\begin{array}{rcccccccc} a_{11}u_1^{k+1} & = & & -a_{12}u_2^k & -a_{13}u_3^k & \cdots & -a_{1n}u_n^k & + b_1 \\ a_{22}u_2^{k+1} & = & -a_{21}u_1^{k+1} & & -a_{23}u_3^k & \cdots & -a_{2n}u_n^k & + b_2 \\ \vdots & \vdots & \vdots & & & & & \\ a_{n-1\,n-1}u_{n-1}^{k+1} & = & -a_{n-1\,1}u_1^{k+1} & \cdots & -a_{n-1\,n-2}u_{n-2}^{k+1} & & -a_{n-1\,n}u_n^k & + b_{n-1} \\ a_{n\,n}u_n^{k+1} & = & -a_{n\,1}u_1^{k+1} & -a_{n\,2}u_2^{k+1} & \cdots & -a_{n\,n-1}u_{n-1}^{k+1} & & + b_n, \end{array}$$

which, in matrix form, is written

$$Du_{k+1} = Eu_{k+1} + Fu_k + b.$$

Because D is invertible and E is lower triangular, the matrix $D - E$ is invertible, so the above equation is equivalent to

$$u_{k+1} = (D - E)^{-1}Fu_k + (D - E)^{-1}b, \quad k \geq 0.$$

The above corresponds to choosing M and N to be

$$\begin{aligned} M &= D - E \\ N &= F, \end{aligned}$$

and the matrix B is given by

$$B = M^{-1}N = (D - E)^{-1}F.$$

Since $M = D - E$ is invertible, we know that $I - B = M^{-1}A$ is also invertible.

The method that we just described is the *iterative method of Gauss-Seidel*, and the matrix B is called the *matrix of Gauss-Seidel* and denoted by $\mathcal{L}_1$, with

$$\mathcal{L}_1 = (D - E)^{-1}F.$$

One of the advantages of the method of Gauss-Seidel is that is requires only half of the memory used by Jacobi's method, since we only need

$$u_1^{k+1}, \ldots, u_{i-1}^{k+1}, u_{i+1}^{k}, \ldots, u_n^{k}$$

to compute u_i^{k+1}. We also show that in certain important cases (for example, if A is a tridiagonal matrix), the method of Gauss-Seidel converges faster than Jacobi's method (in this case, they both converge or diverge simultaneously).

The new ingredient in the *relaxation method* is to incorporate part of the matrix D into N: we define M and N by

$$\begin{aligned} M &= \frac{D}{\omega} - E \\ N &= \frac{1 - \omega}{\omega}D + F, \end{aligned}$$

where $\omega \neq 0$ is a real parameter to be suitably chosen. Actually, we show in Section 7.4 that for the relaxation method to converge, we must have $\omega \in (0, 2)$. Note that the case $\omega = 1$ corresponds to the method of Gauss-Seidel.

If we assume that all diagonal entries of D are nonzero, the matrix M is invertible. The matrix B is denoted by $\mathcal{L}_\omega$ and called the *matrix of relaxation*, with

$$\mathcal{L}_\omega = \left(\frac{D}{\omega} - E\right)^{-1}\left(\frac{1-\omega}{\omega}D + F\right) = (D - \omega E)^{-1}((1-\omega)D + \omega F).$$

The number ω is called the *parameter of relaxation*. When $\omega > 1$, the relaxation method is known as *successive overrelaxation*, abbreviated as *SOR*.

At first glance, the relaxation matrix $\mathcal{L}_\omega$ seems at lot more complicated than the Gauss-Seidel matrix $\mathcal{L}_1$, but the iterative system associated with the relaxation method is very similar to the method of Gauss-Seidel, and is quite simple. Indeed, the system associated with the relaxation method is given by

$$\left(\frac{D}{\omega} - E\right)u_{k+1} = \left(\frac{1-\omega}{\omega}D + F\right)u_k + b,$$

which is equivalent to

$$(D - \omega E)u_{k+1} = ((1-\omega)D + \omega F)u_k + \omega b,$$

and can be written

$$Du_{k+1} = Du_k - \omega(Du_k - Eu_{k+1} - Fu_k - b).$$

Explicitly, this is the system

$$\begin{aligned}
a_{11}u_1^{k+1} &= a_{11}u_1^k - \omega(a_{11}u_1^k + a_{12}u_2^k + a_{13}u_3^k + \cdots + a_{1n-2}u_{n-2}^k + a_{1n-1}u_{n-1}^k + a_{1n}u_n^k - b_1)\\
a_{22}u_2^{k+1} &= a_{22}u_2^k - \omega(a_{21}u_1^{k+1} + a_{22}u_2^k + a_{23}u_3^k + \cdots + a_{2n-2}u_{n-2}^k + a_{2n-1}u_{n-1}^k + a_{2n}u_n^k - b_2)\\
&\vdots\\
a_{n\,n}u_n^{k+1} &= a_{n\,n}u_n^k - \omega(a_{n\,1}u_1^{k+1} + a_{n\,2}u_2^{k+1} + \cdots + a_{n\,n-2}u_{n-2}^{k+1} + a_{n\,n-1}u_{n-1}^{k+1} + a_{n\,n}u_n^k - b_n).
\end{aligned}$$

What remains to be done is to find conditions that ensure the convergence of the relaxation method (and the Gauss-Seidel method), that is:

1. Find conditions on ω, namely some interval $I \subseteq \mathbb{R}$ so that $\omega \in I$ implies $\rho(\mathcal{L}_\omega) < 1$; we will prove that $\omega \in (0, 2)$ is a necessary condition.

2. Find if there exist some *optimal value* ω_0 of $\omega \in I$, so that

$$\rho(\mathcal{L}_{\omega_0}) = \inf_{\omega \in I} \rho(\mathcal{L}_\omega).$$

We will give partial answers to the above questions in the next section.

It is also possible to extend the methods of this section by using *block decompositions* of the form $A = D - E - F$, where D, E, and F consist of blocks, and with D an invertible block-diagonal matrix.

7.4 Convergence of the Methods of Jacobi, Gauss-Seidel, and Relaxation

We begin with a general criterion for the convergence of an iterative method associated with a (complex) Hermitian, positive, definite matrix, $A = M - N$. Next, we apply this result to the relaxation method.

Proposition 7.5. *Let A be any Hermitian, positive, definite matrix, written as*

$$A = M - N,$$

with M invertible. Then, $M^* + N$ is Hermitian, and if it is positive, definite, then

$$\rho(M^{-1}N) < 1,$$

so that the iterative method converges.

Proof. Since $M = A + N$ and A is Hermitian, $A^* = A$, so we get

$$M^* + N = A^* + N^* + N = A + N + N^* = M + N^* = (M^* + N)^*,$$

which shows that $M^* + N$ is indeed Hermitian.

Because A is symmetric, positive, definite, the function

$$v \mapsto (v^* A v)^{1/2}$$

from $\mathbb{C}^n$ to $\mathbb{R}$ is a vector norm $\| \ \|$, and let $\| \ \|$ also denote its subordinate matrix norm. We prove that

$$\|M^{-1}N\| < 1,$$

which, by Theorem 7.1 proves that $\rho(M^{-1}N) < 1$. By definition

$$\|M^{-1}N\| = \|I - M^{-1}A\| = \sup_{\|v\|=1} \|v - M^{-1}Av\|,$$

which leads us to evaluate $\|v - M^{-1}Av\|$ when $\|v\| = 1$. If we write $w = M^{-1}Av$, using the facts that $\|v\| = 1$, $v = A^{-1}Mw$, $A^* = A$, and $A = M - N$, we have

$$\begin{aligned}
\|v - w\|^2 &= (v - w)^* A (v - w) \\
&= \|v\|^2 - v^* A w - w^* A v + w^* A w \\
&= 1 - w^* M^* w - w^* M w + w^* A w \\
&= 1 - w^* (M^* + N) w.
\end{aligned}$$

Now, since we assumed that $M^* + N$ is positive definite, if $w \neq 0$, then $w^*(M^* + N)w > 0$, and we conclude that

$$\text{if} \quad \|v\| = 1 \quad \text{then} \quad \|v - M^{-1}Av\| < 1.$$

Finally, the function

$$v \mapsto \|v - M^{-1}Av\|$$

is continuous as a composition of continuous functions, therefore it achieves its maximum on the compact subset $\{v \in \mathbb{C}^n \mid \|v\| = 1\}$, which proves that

$$\sup_{\|v\|=1} \|v - M^{-1}Av\| < 1,$$

and completes the proof. □

Now, as in the previous sections, we assume that A is written as $A = D - E - F$, with D invertible, possibly in block form. The next theorem provides a sufficient condition (which turns out to be also necessary) for the relaxation method to converge (and thus, for the method of Gauss-Seidel to converge). This theorem is known as the *Ostrowski-Reich theorem*.

Theorem 7.6. *If $A = D - E - F$ is Hermitian, positive, definite, and if $0 < \omega < 2$, then the relaxation method converges. This also holds for a block decomposition of A.*

Proof. Recall that for the relaxation method, $A = M - N$ with

$$M = \frac{D}{\omega} - E$$

$$N = \frac{1-\omega}{\omega}D + F,$$

and because $D^* = D$, $E^* = F$ (since A is Hermitian) and $\omega \neq 0$ is real, we have

$$M^* + N = \frac{D^*}{\omega} - E^* + \frac{1-\omega}{\omega}D + F = \frac{2-\omega}{\omega}D.$$

If D consists of the diagonal entries of A, then we know from Section 5.7 that these entries are all positive, and since $\omega \in (0, 2)$, we see that the matrix $((2-\omega)/\omega)D$ is positive definite. If D consists of diagonal blocks of A, because A is positive, definite, by choosing vectors z obtained by picking a nonzero vector for each block of D and padding with zeros, we see that each block of D is positive, definite, and thus D itself is positive definite. Therefore, in all cases, $M^* + N$ is positive, definite, and we conclude by using Proposition 7.5. □

Remark: What if we allow the parameter ω to be a nonzero complex number $\omega \in \mathbb{C}$? In this case, we get

$$M^* + N = \frac{D^*}{\overline{\omega}} - E^* + \frac{1-\omega}{\omega}D + F = \left(\frac{1}{\omega} + \frac{1}{\overline{\omega}} - 1\right)D.$$

But,

$$\frac{1}{\omega} + \frac{1}{\overline{\omega}} - 1 = \frac{\omega + \overline{\omega} - \omega\overline{\omega}}{\omega\overline{\omega}} = \frac{1 - (\omega - 1)(\overline{\omega} - 1)}{|\omega|^2} = \frac{1 - |\omega - 1|^2}{|\omega|^2},$$

so the relaxation method also converges for $\omega \in \mathbb{C}$, provided that

$$|\omega - 1| < 1.$$

This condition reduces to $0 < \omega < 2$ if ω is real.

Unfortunately, Theorem 7.6 does not apply to Jacobi's method, but in special cases, Proposition 7.5 can be used to prove its convergence. On the positive side, if a matrix is strictly column (or row) diagonally dominant, then it can be shown that the method of Jacobi and the method of Gauss-Seidel both converge. The relaxation method also converges if $\omega \in (0, 1]$, but this is not a very useful result because the speed-up of convergence usually occurs for $\omega > 1$.

We now prove that, without any assumption on $A = D - E - F$, other than the fact that A and D are invertible, in order for the relaxation method to converge, we must have $\omega \in (0, 2)$.

Proposition 7.7. *Given any matrix $A = D - E - F$, with A and D invertible, for any $\omega \neq 0$, we have*

$$\rho(\mathcal{L}_\omega) \geq |\omega - 1|.$$

Therefore, the relaxation method (possibly by blocks) does not converge unless $\omega \in (0, 2)$. If we allow ω to be complex, then we must have

$$|\omega - 1| < 1$$

for the relaxation method to converge.

Proof. Observe that the product $\lambda_1 \cdots \lambda_n$ of the eigenvalues of $\mathcal{L}_\omega$, which is equal to $\det(\mathcal{L}_\omega)$, is given by

$$\lambda_1 \cdots \lambda_n = \det(\mathcal{L}_\omega) = \frac{\det\left(\dfrac{1-\omega}{\omega} D + F\right)}{\det\left(\dfrac{D}{\omega} - E\right)} = (1 - \omega)^n.$$

It follows that

$$\rho(\mathcal{L}_\omega) \geq |\lambda_1 \cdots \lambda_n|^{1/n} = |\omega - 1|.$$

The proof is the same if $\omega \in \mathbb{C}$. □

We now consider the case where A is a *tridiagonal matrix*, possibly by blocks. In this case, we obtain precise results about the spectral radius of J and $\mathcal{L}_\omega$, and as a consequence, about the convergence of these methods. We also obtain some information about the rate of convergence of these methods. We begin with the case $\omega = 1$, which is technically easier to deal with. The following proposition gives us the precise relationship between the spectral radii $\rho(J)$ and $\rho(\mathcal{L}_1)$ of the Jacobi matrix and the Gauss-Seidel matrix.

Proposition 7.8. *Let A be a tridiagonal matrix (possibly by blocks). If $\rho(J)$ is the spectral radius of the Jacobi matrix and $\rho(\mathcal{L}_1)$ is the spectral radius of the Gauss-Seidel matrix, then we have*

$$\rho(\mathcal{L}_1) = (\rho(J))^2.$$

Consequently, the method of Jacobi and the method of Gauss-Seidel both converge or both diverge simultaneously (even when A is tridiagonal by blocks); when they converge, the method of Gauss-Seidel converges faster than Jacobi's method.

Proof. We begin with a preliminary result. Let $A(\mu)$ with a tridiagonal matrix by block of the form

$$A(\mu) = \begin{pmatrix} A_1 & \mu^{-1}C_1 & 0 & 0 & \cdots & 0 \\ \mu B_1 & A_2 & \mu^{-1}C_2 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \cdots & \vdots \\ \vdots & \cdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \mu B_{p-2} & A_{p-1} & \mu^{-1}C_{p-1} \\ 0 & \cdots & \cdots & 0 & \mu B_{p-1} & A_p \end{pmatrix},$$

then

$$\det(A(\mu)) = \det(A(1)), \quad \mu \neq 0.$$

To prove this fact, form the block diagonal matrix

$$P(\mu) = \operatorname{diag}(\mu I_1, \mu^2 I_2, \ldots, \mu^p I_p),$$

where I_j is the identity matrix of the same dimension as the block A_j. Then, it is easy to see that

$$A(\mu) = P(\mu)A(1)P(\mu)^{-1},$$

and thus,

$$\det(A(\mu)) = \det(P(\mu)A(1)P(\mu)^{-1}) = \det(A(1)).$$

Since the Jacobi matrix is $J = D^{-1}(E + F)$, the eigenvalues of J are the zeros of the characteristic polynomial

$$p_J(\lambda) = \det(\lambda I - D^{-1}(E + F)),$$

and thus, they are also the zeros of the polynomial

$$q_J(\lambda) = \det(\lambda D - E - F) = \det(D)p_J(\lambda).$$

Similarly, since the Gauss-Seidel matrix is $\mathcal{L}_1 = (D - E)^{-1}F$, the zeros of the characteristic polynomial

$$p_{\mathcal{L}_1}(\lambda) = \det(\lambda I - (D - E)^{-1}F)$$

are also the zeros of the polynomial

$$q_{\mathcal{L}_1}(\lambda) = \det(\lambda D - \lambda E - F) = \det(D - E)p_{\mathcal{L}_1}(\lambda).$$

Since A is tridiagonal (or tridiagonal by blocks), using our preliminary result with $\mu = \lambda \neq 0$, we get

$$q_{\mathcal{L}_1}(\lambda^2) = \det(\lambda^2 D - \lambda^2 E - F) = \det(\lambda^2 D - \lambda E - \lambda F) = \lambda^n q_J(\lambda).$$

By continuity, the above equation also holds for $\lambda = 0$. But then, we deduce that:

1. For any $\beta \neq 0$, if β is an eigenvalue of $\mathcal{L}_1$, then $\beta^{1/2}$ and $-\beta^{1/2}$ are both eigenvalues of J, where $\beta^{1/2}$ is one of the complex square roots of β.

2. For any $\alpha \neq 0$, if α and $-\alpha$ are both eigenvalues of J, then α^2 is an eigenvalue of $\mathcal{L}_1$.

The above immediately implies that $\rho(\mathcal{L}_1) = (\rho(J))^2$. □

We now consider the more general situation where ω is any real in $(0, 2)$.

Proposition 7.9. *Let A be a tridiagonal matrix (possibly by blocks), and assume that the eigenvalues of the Jacobi matrix are all real. If $\omega \in (0, 2)$, then the method of Jacobi and the method of relaxation both converge or both diverge simultaneously (even when A is tridiagonal by blocks). When they converge, the function $\omega \mapsto \rho(\mathcal{L}_\omega)$ (for $\omega \in (0, 2)$) has a unique minimum equal to $\omega_0 - 1$ for*

$$\omega_0 = \frac{2}{1 + \sqrt{1 - (\rho(J))^2}},$$

where $1 < \omega_0 < 2$ if $\rho(J) > 0$. We also have $\rho(\mathcal{L}_1) = (\rho(J))^2$, as before.

Proof. The proof is very technical and can be found in Serre [95] and Ciarlet [30]. As in the proof of the previous proposition, we begin by showing that the eigenvalues of the matrix $\mathcal{L}_\omega$ are the zeros of the polynomnial

$$q_{\mathcal{L}_\omega}(\lambda) = \det\left(\frac{\lambda + \omega - 1}{\omega} D - \lambda E - F\right) = \det\left(\frac{D}{\omega} - E\right) p_{\mathcal{L}_\omega}(\lambda),$$

where $p_{\mathcal{L}_\omega}(\lambda)$ is the characteristic polynomial of $\mathcal{L}_\omega$. Then, using the preliminary fact from Proposition 7.8, it is easy to show that

$$q_{\mathcal{L}_\omega}(\lambda^2) = \lambda^n q_J\left(\frac{\lambda^2 + \omega - 1}{\lambda\omega}\right),$$

for all $\lambda \in \mathbb{C}$, with $\lambda \neq 0$. This time, we cannot extend the above equation to $\lambda = 0$. This leads us to consider the equation

$$\frac{\lambda^2 + \omega - 1}{\lambda\omega} = \alpha,$$

which is equivalent to

$$\lambda^2 - \alpha\omega\lambda + \omega - 1 = 0,$$

for all $\lambda \neq 0$. Since $\lambda \neq 0$, the above equivalence does not hold for $\omega = 1$, but this is not a problem since the case $\omega = 1$ has already been considered in the previous proposition. Then, we can show the following:

1. For any $\beta \neq 0$, if β is an eigenvalue of $\mathcal{L}_\omega$, then

$$\frac{\beta + \omega - 1}{\beta^{1/2}\omega}, \quad -\frac{\beta + \omega - 1}{\beta^{1/2}\omega}$$

   are eigenvalues of J.

2. For every $\alpha \neq 0$, if α and $-\alpha$ are eigenvalues of J, then $\mu_+(\alpha, \omega)$ and $\mu_-(\alpha, \omega)$ are eigenvalues of $\mathcal{L}_\omega$, where $\mu_+(\alpha, \omega)$ and $\mu_-(\alpha, \omega)$ are the squares of the roots of the equation

$$\lambda^2 - \alpha\omega\lambda + \omega - 1 = 0.$$

It follows that

$$\rho(\mathcal{L}_\omega) = \max_{\lambda \,|\, p_J(\lambda)=0} \{\max(|\mu_+(\alpha, \omega)|, |\mu_-(\alpha, \omega)|)\},$$

and since we are assuming that J has real roots, we are led to study the function

$$M(\alpha, \omega) = \max\{|\mu_+(\alpha, \omega)|, |\mu_-(\alpha, \omega)|\},$$

where $\alpha \in \mathbb{R}$ and $\omega \in (0, 2)$. Actually, because $M(-\alpha, \omega) = M(\alpha, \omega)$, it is only necessary to consider the case where $\alpha \geq 0$.

Note that for $\alpha \neq 0$, the roots of the equation

$$\lambda^2 - \alpha\omega\lambda + \omega - 1 = 0.$$

are

$$\frac{\alpha\omega \pm \sqrt{\alpha^2\omega^2 - 4\omega + 4}}{2}.$$

In turn, this leads to consider the roots of the equation

$$\omega^2\alpha^2 - 4\omega + 4 = 0,$$

which are

$$\frac{2(1 \pm \sqrt{1 - \alpha^2})}{\alpha^2},$$

for $\alpha \neq 0$. Since we have

$$\frac{2(1 + \sqrt{1 - \alpha^2})}{\alpha^2} = \frac{2(1 + \sqrt{1 - \alpha^2})(1 - \sqrt{1 - \alpha^2})}{\alpha^2(1 - \sqrt{1 - \alpha^2})} = \frac{2}{1 - \sqrt{1 - \alpha^2}}$$

and

$$\frac{2(1 - \sqrt{1 - \alpha^2})}{\alpha^2} = \frac{2(1 + \sqrt{1 - \alpha^2})(1 - \sqrt{1 - \alpha^2})}{\alpha^2(1 + \sqrt{1 - \alpha^2})} = \frac{2}{1 + \sqrt{1 - \alpha^2}},$$

these roots are

$$\omega_0(\alpha) = \frac{2}{1 + \sqrt{1 - \alpha^2}}, \quad \omega_1(\alpha) = \frac{2}{1 - \sqrt{1 - \alpha^2}}.$$

Observe that the expression for $\omega_0(\alpha)$ is exactly the expression in the statement of our proposition! The rest of the proof consists in analyzing the variations of the function $M(\alpha, \omega)$ by considering various cases for α. In the end, we find that the minimum of $\rho(\mathcal{L}_\omega)$ is obtained for $\omega_0(\rho(J))$. The details are tedious and we omit them. The reader will find complete proofs in Serre [95] and Ciarlet [30]. □

Combining the results of Theorem 7.6 and Proposition 7.9, we obtain the following result which gives precise information about the spectral radii of the matrices J, $\mathcal{L}_1$, and $\mathcal{L}_\omega$.

Proposition 7.10. *Let A be a tridiagonal matrix (possibly by blocks) which is Hermitian, positive, definite. Then, the methods of Jacobi, Gauss-Seidel, and relaxation, all converge for $\omega \in (0, 2)$. There is a unique optimal relaxation parameter*

$$\omega_0 = \frac{2}{1 + \sqrt{1 - (\rho(J))^2}},$$

such that

$$\rho(\mathcal{L}_{\omega_0}) = \inf_{0<\omega<2} \rho(\mathcal{L}_\omega) = \omega_0 - 1.$$

Furthermore, if $\rho(J) > 0$, then

$$\rho(\mathcal{L}_{\omega_0}) < \rho(\mathcal{L}_1) = (\rho(J))^2 < \rho(J),$$

and if $\rho(J) = 0$, then $\omega_0 = 1$ and $\rho(\mathcal{L}_1) = \rho(J) = 0$.

Proof. In order to apply Proposition 7.9, we have to check that $J = D^{-1}(E + F)$ has real eigenvalues. However, if α is any eigenvalue of J and if u is any corresponding eigenvector, then

$$D^{-1}(E + F)u = \alpha u$$

implies that

$$(E + F)u = \alpha D u,$$

and since $A = D - E - F$, the above shows that $(D - A)u = \alpha D u$, that is,

$$Au = (1 - \alpha)Du.$$

Consequently,

$$u^* Au = (1 - \alpha)u^* Du,$$

and since A and D are hermitian, positive, definite, we have $u^* Au > 0$ and $u^* Du > 0$ if $u \neq 0$, which proves that $\alpha \in \mathbb{R}$. The rest follows from Theorem 7.6 and Proposition 7.9. □

Remark: It is preferable to overestimate rather than underestimate the relaxation parameter when the optimum relaxation parameter is not known exactly.

7.5 Summary

The main concepts and results of this chapter are listed below:

- Iterative methods. Splitting A as $A = M - N$.
- *Convergence of a sequence of vectors or matrices*.
- A criterion for the convergence of the sequence (B^k) of powers of a matrix B to zero in terms of the spectral radius $\rho(B)$.
- A characterization of the spectral radius $\rho(B)$ as the limit of the sequence $(\|B^k\|^{1/k})$.
- A criterion of the convergence of iterative methods.
- Asymptotic behavior of iterative methods.
- Splitting A as $A = D-E-F$, and the methods of *Jacobi*, *Gauss-Seidel*, and *relaxation* (and *SOR*).
- The *Jacobi matrix*, $J = D^{-1}(E + F)$.
- The *Gauss-Seidel matrix*, $\mathcal{L}_2 = (D - E)^{-1}F$.
- The *matrix of relaxation*, $\mathcal{L}_\omega = (D - \omega E)^{-1}((1 - \omega)D + \omega F)$.
- Convergence of iterative methods: a general result when $A = M - N$ is Hermitian, positive, definite.
- A sufficient condition for the convergence of the methods of Jacobi, Gauss-Seidel, and relaxation. The *Ostrowski-Reich Theorem*: A is symmetric, positive, definite, and $\omega \in (0, 2)$.
- A necessary condition for the convergence of the methods of Jacobi , Gauss-Seidel, and relaxation: $\omega \in (0, 2)$.
- The case of tridiagonal matrices (possibly by blocks). Simultaneous convergence or divergence of Jacobi's method and Gauss-Seidel's method, and comparison of the spectral radii of $\rho(J)$ and $\rho(\mathcal{L}_1)$: $\rho(\mathcal{L}_1) = (\rho(J))^2$.
- The case of tridiagonal, Hermitian, positive, definite matrices (possibly by blocks). The methods of Jacobi, Gauss-Seidel, and relaxation, all converge.
- In the above case, there is a unique optimal relaxation parameter for which $\rho(\mathcal{L}_{\omega_0}) < \rho(\mathcal{L}_1) = (\rho(J))^2 < \rho(J)$ (if $\rho(J) \neq 0$).

Chapter 8

The Dual Space and Duality

8.1 The Dual Space E^* and Linear Forms

In Section 1.8 we defined linear forms, the dual space $E^* = \mathrm{Hom}(E, K)$ of a vector space E, and showed the existence of dual bases for vector spaces of finite dimension.

In this chapter, we take a deeper look at the connection between a space E and its dual space E^*. As we will see shortly, every linear map $f\colon E \to F$ gives rise to a linear map $f^\top\colon F^* \to E^*$, and it turns out that in a suitable basis, the matrix of $f^\top$ is the transpose of the matrix of f. Thus, the notion of dual space provides a conceptual explanation of the phenomena associated with transposition.

But it does more, because it allows us to view a linear equation as an element of the dual space E^*, and thus to view subspaces of E as solutions of sets of linear equations and vice-versa. The relationship between subspaces and sets of linear forms is the essence of *duality*, a term which is often used loosely, but can be made precise as a bijection between the set of subspaces of a given vector space E and the set of subspaces of its dual E^*. In this correspondence, a subspace V of E yields the subspace V^0 of E^* consisting of all linear forms that vanish on V (that is, have the value zero for all input in V).

Consider the following set of two "linear equations" in $\mathbb{R}^3$,

$$x - y + z = 0$$
$$x - y - z = 0,$$

and let us find out what is their set V of common solutions $(x, y, z) \in \mathbb{R}^3$. By subtracting the second equation from the first, we get $2z = 0$, and by adding the two equations, we find that $2(x - y) = 0$, so the set V of solutions is given by

$$y = x$$
$$z = 0.$$

This is a one dimensional subspace of $\mathbb{R}^3$. Geometrically, this is the line of equation $y = x$ in the plane $z = 0$.

Now, why did we say that the above equations are linear? This is because, as functions of (x, y, z), both maps $f_1 \colon (x, y, z) \mapsto x - y + z$ and $f_2 \colon (x, y, z) \mapsto x - y - z$ are linear. The set of all such linear functions from $\mathbb{R}^3$ to $\mathbb{R}$ is a vector space; we used this fact to form linear combinations of the "equations" f_1 and f_2. Observe that the dimension of the subspace V is 1. The ambient space has dimension $n = 3$ and there are two "independent" equations f_1, f_2, so it appears that the dimension $\dim(V)$ of the subspace V defined by m independent equations is

$$\dim(V) = n - m,$$

which is indeed a general fact (proved in Theorem 8.1).

More generally, in $\mathbb{R}^n$, a linear equation is determined by an n-tuple $(a_1, \ldots, a_n) \in \mathbb{R}^n$, and the solutions of this linear equation are given by the n-tuples $(x_1, \ldots, x_n) \in \mathbb{R}^n$ such that

$$a_1 x_1 + \cdots + a_n x_n = 0;$$

these solutions constitute the kernel of the linear map $(x_1, \ldots, x_n) \mapsto a_1 x_1 + \cdots + a_n x_n$. The above considerations assume that we are working in the canonical basis $(e_1, \ldots, e_n)$ of $\mathbb{R}^n$, but we can define "linear equations" independently of bases and in any dimension, by viewing them as elements of the vector space $\mathrm{Hom}(E, K)$ of linear maps from E to the field K.

Definition 8.1. Given a vector space E, the vector space $\mathrm{Hom}(E, K)$ of linear maps from E to the field K is called the *dual space (or dual)* of E. The space $\mathrm{Hom}(E, K)$ is also denoted by E^*, and the linear maps in E^* are called *the linear forms*, or *covectors*. The dual space E^{**} of the space E^* is called the *bidual* of E.

As a matter of notation, linear forms $f \colon E \to K$ will also be denoted by starred symbol, such as u^*, x^*, *etc.*

Given a vector space E and any basis $(u_i)_{i \in I}$ for E, we can associate to each u_i a linear form $u_i^* \in E^*$, and the u_i^* have some remarkable properties.

Definition 8.2. Given a vector space E and any basis $(u_i)_{i \in I}$ for E, by Proposition 1.13, for every $i \in I$, there is a unique linear form u_i^* such that

$$u_i^*(u_j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j, \end{cases}$$

for every $j \in I$. The linear form u_i^* is called the *coordinate form* of index i w.r.t. the basis $(u_i)_{i \in I}$.

The reason for the terminology *coordinate form* was explained in Section 1.8.

We proved in Theorem 1.16 that if $(u_1, \ldots, u_n)$ is a basis of E, then $(u_1^*, \ldots, u_n^*)$ is a basis of E^* called the *dual basis*.

If $(u_1, \ldots, u_n)$ is a basis of $\mathbb{R}^n$ (more generally K^n), it is possible to find explicitly the dual basis $(u_1^*, \ldots, u_n^*)$, where each u_i^* is represented by a row vector. For example, consider the columns of the Bézier matrix

$$B_4 = \begin{pmatrix} 1 & -3 & 3 & -1 \\ 0 & 3 & -6 & 3 \\ 0 & 0 & 3 & -3 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Since the form u_1^* is defined by the conditions $u_1^*(u_1) = 1, u_1^*(u_2) = 0, u_1^*(u_3) = 0, u_1^*(u_4) = 0$, it is represented by a row vector $(\lambda_1 \ \lambda_2 \ \lambda_3 \ \lambda_4)$ such that

$$\begin{pmatrix} \lambda_1 & \lambda_2 & \lambda_3 & \lambda_4 \end{pmatrix} \begin{pmatrix} 1 & -3 & 3 & -1 \\ 0 & 3 & -6 & 3 \\ 0 & 0 & 3 & -3 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix}.$$

This implies that u_1^* is the first row of the inverse of B_4. Since

$$B_4^{-1} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1/3 & 2/3 & 1 \\ 0 & 0 & 1/3 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

the linear forms $(u_1^*, u_2^*, u_3^*, u_4^*)$ correspond to the rows of B_4^{-1}. In particular, u_1^* is represented by $(1 \ 1 \ 1 \ 1)$.

The above method works for any n. Given any basis $(u_1, \ldots, u_n)$ of $\mathbb{R}^n$, if P is the $n \times n$ matrix whose jth column is u_j, then the dual form u_i^* is given by the ith row of the matrix P^{-1}.

When E is of finite dimension n and $(u_1, \ldots, u_n)$ is a basis of E, by Theorem 8.1 (1), the family $(u_1^*, \ldots, u_n^*)$ is a basis of the dual space E^*. Let us see how the coordinates of a linear form $\varphi^* \in E^*$ over the dual basis $(u_1^*, \ldots, u_n^*)$ vary under a change of basis.

Let $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$ be two bases of E, and let $P = (a_{i\,j})$ be the change of basis matrix from $(u_1, \ldots, u_n)$ to $(v_1, \ldots, v_n)$, so that

$$v_j = \sum_{i=1}^{n} a_{i\,j} u_i,$$

and let $P^{-1} = (b_{i\,j})$ be the inverse of P, so that

$$u_i = \sum_{j=1}^{n} b_{j\,i} v_j.$$

Since $u_i^*(u_j) = \delta_{i\,j}$ and $v_i^*(v_j) = \delta_{i\,j}$, we get

$$v_j^*(u_i) = v_j^*(\sum_{k=1}^{n} b_{k\,i} v_k) = b_{j\,i},$$

and thus

$$v_j^* = \sum_{i=1}^{n} b_{j\,i} u_i^*,$$

and

$$u_i^* = \sum_{j=1}^{n} a_{i\,j} v_j^*.$$

This means that the change of basis from the dual basis $(u_1^*, \ldots, u_n^*)$ to the dual basis $(v_1^*, \ldots, v_n^*)$ is $(P^{-1})^\top$. Since

$$\varphi^* = \sum_{i=1}^{n} \varphi_i u_i^* = \sum_{i=1}^{n} \varphi_i' v_i^*,$$

we get

$$\varphi_j' = \sum_{i=1}^{n} a_{i\,j} \varphi_i,$$

so the new coordinates φ_j' are expressed in terms of the old coordinates φ_i using the matrix $P^\top$. If we use the row vectors $(\varphi_1, \ldots, \varphi_n)$ and $(\varphi_1', \ldots, \varphi_n')$, we have

$$(\varphi_1', \ldots, \varphi_n') = (\varphi_1, \ldots, \varphi_n)P.$$

Comparing with the change of basis

$$v_j = \sum_{i=1}^{n} a_{i\,j} u_i,$$

we note that this time, the coordinates (φ_i) of the linear form φ^* change in the *same direction* as the change of basis. For this reason, we say that the coordinates of linear forms are *covariant*. By abuse of language, it is often said that linear forms are *covariant*, which explains why the term *covector* is also used for a linear form.

Observe that if $(e_1, \ldots, e_n)$ is a basis of the vector space E, then, as a linear map from E to K, every linear form $f \in E^*$ is represented by a $1 \times n$ matrix, that is, by a *row vector*

$$(\lambda_1 \cdots \lambda_n),$$

with respect to the basis $(e_1, \ldots, e_n)$ of E, and 1 of K, where $f(e_i) = \lambda_i$. A vector $u = \sum_{i=1}^{n} u_i e_i \in E$ is represented by a $n \times 1$ matrix, that is, by a *column vector*

$$\begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix},$$

and the action of f on u, namely $f(u)$, is represented by the matrix product

$$\begin{pmatrix}\lambda_1 & \cdots & \lambda_n\end{pmatrix}\begin{pmatrix}u_1\\ \vdots\\ u_n\end{pmatrix} = \lambda_1 u_1 + \cdots + \lambda_n u_n.$$

On the other hand, with respect to the dual basis $(e_1^*, \ldots, e_n^*)$ of E^*, the linear form f is represented by the column vector

$$\begin{pmatrix}\lambda_1\\ \vdots\\ \lambda_n\end{pmatrix}.$$

Remark: In many texts using tensors, vectors are often indexed with lower indices. If so, it is more convenient to write the coordinates of a vector x over the basis $(u_1, \ldots, u_n)$ as (x^i), using an upper index, so that

$$x = \sum_{i=1}^{n} x^i u_i,$$

and in a change of basis, we have

$$v_j = \sum_{i=1}^{n} a_j^i u_i$$

and

$$x^i = \sum_{j=1}^{n} a_j^i x'^j.$$

Dually, linear forms are indexed with upper indices. Then, it is more convenient to write the coordinates of a covector φ^* over the dual basis $(u^{*1}, \ldots, u^{*n})$ as (φ_i), using a lower index, so that

$$\varphi^* = \sum_{i=1}^{n} \varphi_i u^{*i}$$

and in a change of basis, we have

$$u^{*i} = \sum_{j=1}^{n} a_j^i v^{*j}$$

and

$$\varphi'_j = \sum_{i=1}^{n} a_j^i \varphi_i.$$

With these conventions, the index of summation appears once in upper position and once in lower position, and the summation sign can be safely omitted, a trick due to *Einstein*. For example, we can write

$$\varphi'_j = a_j^i \varphi_i$$

as an abbreviation for

$$\varphi'_j = \sum_{i=1}^{n} a^i_j \varphi_i.$$

For another example of the use of Einstein's notation, if the vectors $(v_1, \ldots, v_n)$ are linear combinations of the vectors $(u_1, \ldots, u_n)$, with

$$v_i = \sum_{j=1}^{n} a_{ij} u_j, \quad 1 \leq i \leq n,$$

then the above equations are witten as

$$v_i = a^j_i u_j, \quad 1 \leq i \leq n.$$

Thus, in Einstein's notation, the $n \times n$ matrix (a_{ij}) is denoted by (a^j_i), a $(1,1)$-*tensor*.

Beware that some authors view a matrix as a mapping between *coordinates*, in which case the matrix (a_{ij}) is denoted by (a^i_j).

8.2 Pairing and Duality Between E and E^*

Given a linear form $u^* \in E^*$ and a vector $v \in E$, the result $u^*(v)$ of applying u^* to v is also denoted by $\langle u^*, v\rangle$. This defines a binary operation $\langle -, -\rangle \colon E^* \times E \to K$ satisfying the following properties:

$$\begin{aligned}
\langle u^*_1 + u^*_2, v\rangle &= \langle u^*_1, v\rangle + \langle u^*_2, v\rangle \\
\langle u^*, v_1 + v_2\rangle &= \langle u^*, v_1\rangle + \langle u^*, v_2\rangle \\
\langle \lambda u^*, v\rangle &= \lambda \langle u^*, v\rangle \\
\langle u^*, \lambda v\rangle &= \lambda \langle u^*, v\rangle.
\end{aligned}$$

The above identities mean that $\langle -, -\rangle$ is a *bilinear map*, since it is linear in each argument. It is often called the *canonical pairing* between E^* and E. In view of the above identities, given any fixed vector $v \in E$, the map $\mathrm{eval}_v \colon E^* \to K$ (*evaluation at* v) defined such that

$$\mathrm{eval}_v(u^*) = \langle u^*, v\rangle = u^*(v) \quad \text{for every } u^* \in E^*$$

is a linear map from E^* to K, that is, eval_v is a linear form in E^{**}. Again, from the above identities, the map $\mathrm{eval}_E \colon E \to E^{**}$, defined such that

$$\mathrm{eval}_E(v) = \mathrm{eval}_v \quad \text{for every } v \in E,$$

is a linear map. Observe that

$$\mathrm{eval}_E(v)(u^*) = \langle u^*, v\rangle = u^*(v), \quad \text{for all } v \in E \text{ and all } u^* \in E^*.$$

We shall see that the map eval_E is injective, and that it is an isomorphism when E has finite dimension.

We now formalize the notion of the set V^0 of linear equations vanishing on all vectors in a given subspace $V \subseteq E$, and the notion of the set U^0 of common solutions of a given set $U \subseteq E^*$ of linear equations. The duality theorem (Theorem 8.1) shows that the dimensions of V and V^0, and the dimensions of U and U^0, are related in a crucial way. It also shows that, in finite dimension, the maps $V \mapsto V^0$ and $U \mapsto U^0$ are inverse bijections from subspaces of E to subspaces of E^*.

Definition 8.3. Given a vector space E and its dual E^*, we say that a vector $v \in E$ and a linear form $u^* \in E^*$ are *orthogonal* iff $\langle u^*, v\rangle = 0$. Given a subspace V of E and a subspace U of E^*, we say that V *and* U *are orthogonal* iff $\langle u^*, v\rangle = 0$ for every $u^* \in U$ and every $v \in V$. Given a subset V of E (resp. a subset U of E^*), the *orthogonal* V^0 *of* V is the subspace V^0 of E^* defined such that

$$V^0 = \{u^* \in E^* \mid \langle u^*, v\rangle = 0, \text{ for every } v \in V\}$$

(resp. the *orthogonal* U^0 *of* U is the subspace U^0 of E defined such that

$$U^0 = \{v \in E \mid \langle u^*, v\rangle = 0, \text{ for every } u^* \in U\}).$$

The subspace $V^0 \subseteq E^*$ is also called the *annihilator* of V. The subspace $U^0 \subseteq E$ annihilated by $U \subseteq E^*$ does not have a special name. It seems reasonable to call it the *linear subspace (or linear variety) defined by* U.

Informally, V^0 is the *set of linear equations that vanish on* V, and U^0 is *the set of common zeros of all linear equations in* U. We can also define V^0 by

$$V^0 = \{u^* \in E^* \mid V \subseteq \operatorname{Ker} u^*\}$$

and U^0 by

$$U^0 = \bigcap_{u^* \in U} \operatorname{Ker} u^*.$$

Observe that $E^0 = \{0\} = (0)$, and $\{0\}^0 = E^*$. Furthermore, if $V_1 \subseteq V_2 \subseteq E$, then $V_2^0 \subseteq V_1^0 \subseteq E^*$, and if $U_1 \subseteq U_2 \subseteq E^*$, then $U_2^0 \subseteq U_1^0 \subseteq E$.

Proof. Indeed, if $V_1 \subseteq V_2 \subseteq E$, then for any $f^* \in V_2^0$ we have $f^*(v) = 0$ for all $v \in V_2$, and thus $f^*(v) = 0$ for all $v \in V_1$, so $f^* \in V_1^0$. Similarly, if $U_1 \subseteq U_2 \subseteq E^*$, then for any $v \in U_2^0$, we have $f^*(v) = 0$ for all $f^* \in U_2$, so $f^*(v) = 0$ for all $f^* \in U_1$, which means that $v \in U_1^0$. □

Here are some examples. Let $E = \mathrm{M}_2(\mathbb{R})$, the space of real 2×2 matrices, and let V be the subspace of $\mathrm{M}_2(\mathbb{R})$ spanned by the matrices

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

We check immediately that the subspace V consists of all matrices of the form

$$\begin{pmatrix} b & a \\ a & c \end{pmatrix},$$

that is, all symmetric matrices. The matrices

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

in V satisfy the equation

$$a_{12} - a_{21} = 0,$$

and all scalar multiples of these equations, so V^0 is the subspace of E^* spanned by the linear form given by $u^*(a_{11}, a_{12}, a_{21}, a_{22}) = a_{12} - a_{21}$. By the duality theorem (Theorem 8.1) we have

$$\dim(V^0) = \dim(E) - \dim(V) = 4 - 3 = 1.$$

The above example generalizes to $E = \mathrm{M}_n(\mathbb{R})$ for any $n \geq 1$, but this time, consider the space U of linear forms asserting that a matrix A is symmetric; these are the linear forms spanned by the $n(n-1)/2$ equations

$$a_{ij} - a_{ji} = 0, \quad 1 \leq i < j \leq n;$$

Note there are no constraints on diagonal entries, and half of the equations

$$a_{ij} - a_{ji} = 0, \quad 1 \leq i \neq j \leq n$$

are redundant. It is easy to check that the equations (linear forms) for which $i < j$ are linearly independent. To be more precise, let U be the space of linear forms in E^* spanned by the linear forms

$$u^*_{ij}(a_{11}, \ldots, a_{1n}, a_{21}, \ldots, a_{2n}, \ldots, a_{n1}, \ldots, a_{nn}) = a_{ij} - a_{ji}, \quad 1 \leq i < j \leq n.$$

The dimension of U is $n(n-1)/2$. Then, the set U^0 of common solutions of these equations is the space $\mathbf{S}(n)$ of symmetric matrices. By the duality theorem (Theorem 8.1), this space has dimension

$$\frac{n(n+1)}{2} = n^2 - \frac{n(n-1)}{2}.$$

We leave it as an exercise to find a basis of $\mathbf{S}(n)$.

If $E = \mathrm{M}_n(\mathbb{R})$, consider the subspace U of linear forms in E^* spanned by the linear forms

$$u^*_{ij}(a_{11}, \ldots, a_{1n}, a_{21}, \ldots, a_{2n}, \ldots, a_{n1}, \ldots, a_{nn}) = a_{ij} + a_{ji}, \quad 1 \leq i < j \leq n$$
$$u^*_{ii}(a_{11}, \ldots, a_{1n}, a_{21}, \ldots, a_{2n}, \ldots, a_{n1}, \ldots, a_{nn}) = a_{ii}, \quad 1 \leq i \leq n.$$

It is easy to see that these linear forms are linearly independent, so $\dim(U) = n(n+1)/2$. The space U^0 of matrices $A \in \mathrm{M}_n(\mathbb{R})$ satifying all of the above equations is clearly the space $\mathbf{Skew}(n)$ of skew-symmetric matrices. By the duality theorem (Theorem 8.1), the dimension of U^0 is

$$\frac{n(n-1)}{2} = n^2 - \frac{n(n+1)}{2}.$$

We leave it as an exercise to find a basis of $\mathbf{Skew}(n)$.

For yet another example with $E = \mathrm{M}_n(\mathbb{R})$, for any $A \in \mathrm{M}_n(\mathbb{R})$, consider the linear form in E^* given by

$$\mathrm{tr}(A) = a_{11} + a_{22} + \cdots + a_{nn},$$

called the *trace* of A. The subspace U^0 of E consisting of all matrices A such that $\mathrm{tr}(A) = 0$ is a space of dimension $n^2 - 1$. We leave it as an exercise to find a basis of this space.

The dimension equations

$$\dim(V) + \dim(V^0) = \dim(E)$$
$$\dim(U) + \dim(U^0) = \dim(E)$$

are always true (if E is finite-dimensional). This is part of the duality theorem (Theorem 8.1).

In contrast with the previous examples, given a matrix $A \in \mathrm{M}_n(\mathbb{R})$, the equations asserting that $A^\top A = I$ are not linear constraints. For example, for $n = 2$, we have

$$\begin{aligned} a_{11}^2 + a_{21}^2 &= 1 \\ a_{21}^2 + a_{22}^2 &= 1 \\ a_{11}a_{12} + a_{21}a_{22} &= 0. \end{aligned}$$

Remarks:

(1) The notation V^0 (resp. U^0) for the orthogonal of a subspace V of E (resp. a subspace U of E^*) is not universal. Other authors use the notation $V^\perp$ (resp. $U^\perp$). However, the notation $V^\perp$ is also used to denote the orthogonal complement of a subspace V with respect to an inner product on a space E, in which case $V^\perp$ is a subspace of E and not a subspace of E^* (see Chapter 9). To avoid confusion, we prefer using the notation V^0.

(2) Since linear forms can be viewed as linear equations (at least in finite dimension), given a subspace (or even a subset) U of E^*, we can define the set $\mathcal{Z}(U)$ of *common zeros* of the equations in U by

$$\mathcal{Z}(U) = \{v \in E \mid u^*(v) = 0, \text{ for all } u^* \in U\}.$$

Of course $\mathcal{Z}(U) = U^0$, but the notion $\mathcal{Z}(U)$ can be generalized to more general kinds of equations, namely polynomial equations. In this more general setting, U is a set of *polynomials* in n variables with coefficients in a field K (where $n = \dim(E)$). Sets of the form $\mathcal{Z}(U)$ are called *algebraic varieties*. Linear forms correspond to the special case where homogeneous polynomials of degree 1 are considered.

If V is a subset of E, it is natural to associate with V the *set of polynomials in* $K[X_1, \ldots, X_n]$ *that vanish on* V. This set, usually denoted $\mathcal{I}(V)$, has some special properties that make it an *ideal*. If V is a linear subspace of E, it is natural to restrict our attention to the space V^0 of linear forms that vanish on V, and in this case we identify $\mathcal{I}(V)$ and V^0 (although technically, $\mathcal{I}(V)$ is no longer an ideal).

For any arbitrary set of polynomials $U \subseteq K[X_1, \ldots, X_n]$ (resp. subset $V \subseteq E$), the relationship between $\mathcal{I}(\mathcal{Z}(U))$ and U (resp. $\mathcal{Z}(\mathcal{I}(V))$ and V) is generally not simple, even though we always have

$$U \subseteq \mathcal{I}(\mathcal{Z}(U)) \quad (\text{resp.} \quad V \subseteq \mathcal{Z}(\mathcal{I}(V))).$$

However, when the field K is algebraically closed, then $\mathcal{I}(\mathcal{Z}(U))$ is equal to the *radical* of the ideal U, a famous result due to Hilbert known as the *Nullstellensatz* (see Lang [62] or Dummit and Foote [38]). The study of algebraic varieties is the main subject of *algebraic geometry*, a beautiful but formidable subject. For a taste of algebraic geometry, see Lang [62] or Dummit and Foote [38].

The duality theorem (Theorem 8.1) shows that the situation is much simpler if we restrict our attention to linear subspaces; in this case

$$U = \mathcal{I}(\mathcal{Z}(U)) \quad \text{and} \quad V = \mathcal{Z}(\mathcal{I}(V)).$$

We claim that $V \subseteq V^{00}$ for every subspace V of E, and that $U \subseteq U^{00}$ for every subspace U of E^*.

Proof. Indeed, for any $v \in V$, to show that $v \in V^{00}$ we need to prove that $u^*(v) = 0$ for all $u^* \in V^0$. However, V^0 consists of all linear forms u^* such that $u^*(y) = 0$ for *all* $y \in V$; in particular, for a fixed $v \in V$, we have $u^*(v) = 0$ for all $u^* \in V^0$, as required.

Similarly, for any $u^* \in U$, to show that $u^* \in U^{00}$ we need to prove that $u^*(v) = 0$ for all $v \in U^0$. However, U^0 consists of all vectors v such that $f^*(v) = 0$ for *all* $f^* \in U$; in particular, for a fixed $u^* \in U$, we have $u^*(v) = 0$ for all $v \in U^0$, as required. □

We will see shortly that in finite dimension, we have $V = V^{00}$ and $U = U^{00}$.

8.3 The Duality Theorem

Given a vector space E of dimension $n \geq 1$ and a subspace U of E, by Theorem 1.9, every basis $(u_1, \ldots, u_m)$ of U can be extended to a basis $(u_1, \ldots, u_n)$ of E. We have the following important theorem adapted from E. Artin [5] (Chapter 1).

Theorem 8.1. *(Duality theorem) Let E be a vector space of dimension n. The following properties hold:*

(a) For every basis $(u_1, \ldots, u_n)$ of E, the family of coordinate forms $(u_1^*, \ldots, u_n^*)$ is a basis of E^* (called the dual basis of $(u_1, \ldots, u_n)$).

(b) For every subspace V of E, we have $V^{00} = V$.

(c) For every pair of subspaces V and W of E such that $E = V \oplus W$, with V of dimension m, for every basis $(u_1, \ldots, u_n)$ of E such that $(u_1, \ldots, u_m)$ is a basis of V and $(u_{m+1}, \ldots, u_n)$ is a basis of W, the family $(u_1^*, \ldots, u_m^*)$ is a basis of the orthogonal W^0 of W in E^*, so that

$$\dim(W) + \dim(W^0) = \dim(E).$$

Furthermore, we have $W^{00} = W$.

(d) For every subspace U of E^*, we have

$$\dim(U) + \dim(U^0) = \dim(E),$$

where U^0 is the orthogonal of U in E, and $U^{00} = U$.

Proof. (a) This part was proved in Theorem 1.16.

(b) Clearly, we have $V \subseteq V^{00}$. If $V \neq V^{00}$, then let $(u_1, \ldots, u_p)$ be a basis of V^{00} such that $(u_1, \ldots, u_m)$ is a basis of V, with $m < p$. Since $u_{m+1} \in V^{00}$, u_{m+1} is orthogonal to every linear form in V^0. Now, we have $u_{m+1}^*(u_i) = 0$ for all $i = 1, \ldots, m$, and thus $u_{m+1}^* \in V^0$. However, $u_{m+1}^*(u_{m+1}) = 1$, contradicting the fact that u_{m+1} is orthogonal to every linear form in V^0. Thus, $V = V^{00}$.

(c) Every linear form $f^* \in W^0$ is orthogonal to every u_j for $j = m+1, \ldots, n$, and thus, $f^*(u_j) = 0$ for $j = m+1, \ldots, n$. For such a linear form $f^* \in W^0$, let

$$g^* = f^*(u_1)u_1^* + \cdots + f^*(u_m)u_m^*.$$

We have $g^*(u_i) = f^*(u_i)$, for every i, $1 \leq i \leq m$. Furthermore, by definition, g^* vanishes on all u_j with $j = m+1, \ldots, n$. Thus, f^* and g^* agree on the basis $(u_1, \ldots, u_n)$ of E, and so $g^* = f^*$. This shows that $(u_1^*, \ldots, u_m^*)$ generates W^0, and since it is also a linearly independent family, $(u_1^*, \ldots, u_m^*)$ is a basis of W^0. It is then obvious that $\dim(W) + \dim(W^0) = \dim(E)$, and by part (b), we have $W^{00} = W$.

(d) Let $(f_1^*, \ldots, f_m^*)$ be a basis of U. Note that the map $h\colon E \to K^m$ defined such that

$$h(v) = (f_1^*(v), \ldots, f_m^*(v))$$

for every $v \in E$ is a linear map, and that its kernel $\operatorname{Ker} h$ is precisely U^0. Then, by Proposition 3.6,

$$n = \dim(E) = \dim(\operatorname{Ker} h) + \dim(\operatorname{Im} h) \leq \dim(U^0) + m,$$

since $\dim(\operatorname{Im} h) \leq m$. Thus, $n - \dim(U^0) \leq m$. By (c), we have $\dim(U^0) + \dim(U^{00}) = \dim(E) = n$, so we get $\dim(U^{00}) \leq m$. However, it is clear that $U \subseteq U^{00}$, which implies $m = \dim(U) \leq \dim(U^{00})$, so $\dim(U) = \dim(U^{00}) = m$, and we must have $U = U^{00}$. □

Part (a) of Theorem 8.1 shows that

$$\dim(E) = \dim(E^*),$$

and if $(u_1, \ldots, u_n)$ is a basis of E, then $(u_1^*, \ldots, u_n^*)$ is a basis of the dual space E^* called the *dual basis* of $(u_1, \ldots, u_n)$.

Define the function $\mathcal{E}$ ($\mathcal{E}$ for equations) from subspaces of E to subspaces of E^* and the function $\mathcal{Z}$ ($\mathcal{Z}$ for zeros) from subspaces of E^* to subspaces of E by

$$\begin{aligned} \mathcal{E}(V) &= V^0, \quad V \subseteq E \\ \mathcal{Z}(U) &= U^0, \quad U \subseteq E^*. \end{aligned}$$

By part (c) and (d) of theorem 8.1,

$$\begin{aligned} (\mathcal{Z} \circ \mathcal{E})(V) &= V^{00} = V \\ (\mathcal{E} \circ \mathcal{Z})(U) &= U^{00} = U, \end{aligned}$$

so $\mathcal{Z} \circ \mathcal{E} = \mathrm{id}$ and $\mathcal{E} \circ \mathcal{Z} = \mathrm{id}$, and the maps $\mathcal{E}$ and $\mathcal{V}$ are inverse bijections. These maps set up a *duality* between subspaces of E and subspaces of E^*. In particular, every subspace $V \subseteq E$ of dimension m is the set of common zeros of the space of linear forms (equations) V^0, which has dimension $n - m$. This confirms the claim we made about the dimension of the subpsace defined by a set of linear equations.

One should be careful that this bijection does not hold if E has infinite dimension. Some restrictions on the dimensions of U and V are needed.

However, even if E is infinite-dimensional, the identity $V = V^{00}$ holds for every subspace V of E. The proof is basically the same but uses an infinite basis of V^{00} extending a basis of V.

Suppose that V is a subspace of $\mathbb{R}^n$ of dimension m and that $(v_1, \ldots, v_m)$ is a basis of V. To find a basis of V^0, we first extend $(v_1, \ldots, v_m)$ to a basis $(v_1, \ldots, v_n)$ of $\mathbb{R}^n$, and then by

part (c) of Theorem 8.1, we know that $(v_{m+1}^*, \ldots, v_n^*)$ is a basis of V^0. For example, suppose that V is the subspace of $\mathbb{R}^4$ spanned by the two linearly independent vectors

$$v_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \quad v_2 = \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix},$$

the first two vectors of the Haar basis in $\mathbb{R}^4$. The four columns of the Haar matrix

$$W = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & -1 \end{pmatrix}$$

form a basis of $\mathbb{R}^4$, and the inverse of W is given by

$$W^{-1} = \begin{pmatrix} 1/4 & 0 & 0 & 0 \\ 0 & 1/4 & 0 & 0 \\ 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1/2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} = \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & -1/4 & -1/4 \\ 1/2 & -1/2 & 0 & 0 \\ 0 & 0 & 1/2 & -1/2 \end{pmatrix}.$$

Since the dual basis $(v_1^*, v_2^*, v_3^*, v_4^*)$ is given by the row of W^{-1}, the last two rows of W^{-1},

$$\begin{pmatrix} 1/2 & -1/2 & 0 & 0 \\ 0 & 0 & 1/2 & -1/2 \end{pmatrix},$$

form a basis of V^0. We also obtain a basis by rescaling by the factor $1/2$, so the linear forms given by the row vectors

$$\begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}$$

form a basis of V^0, the space of linear forms (linear equations) that vanish on the subspace V.

The method that we described to find V^0 requires first extending a basis of V and then inverting a matrix, but there is a more direct method. Indeed, let A be the $n \times m$ matrix whose columns are the basis vectors $(v_1, \ldots, v_m)$ of V. Then, a linear form u represented by a row vector belongs to V^0 iff $uv_i = 0$ for $i = 1, \ldots, m$ iff

$$uA = 0$$

iff

$$A^\top u^\top = 0.$$

Therefore, all we need to do is to find a basis of the nullspace of $A^\top$. This can be done quite effectively using the reduction of a matrix to reduced row echelon form (rref); see Section 5.9.

Let us now consider the problem of finding a basis of the hyperplane H in $\mathbb{R}^n$ defined by the equation

$$c_1x_1 + \cdots + c_nx_n = 0.$$

More precisely, if $u^*(x_1, \ldots, x_n)$ is the linear form in $(\mathbb{R}^n)^*$ given by $u^*(x_1, \ldots, x_n) = c_1x_1 + \cdots + c_nx_n$, then the hyperplane H is the kernel of u^*. Of course we assume that some c_j is nonzero, in which case the linear form u^* spans a one-dimensional subspace U of $(\mathbb{R}^n)^*$, and $U^0 = H$ has dimension $n-1$.

Since u^* is not the linear form which is identically zero, there is a smallest positive index $j \leq n$ such that $c_j \neq 0$, so our linear form is really $u^*(x_1, \ldots, x_n) = c_jx_j + \cdots + c_nx_n$. We claim that the following $n-1$ vectors (in $\mathbb{R}^n$) form a basis of H:

$$\begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ j-1 \\ j \\ j+1 \\ j+2 \\ \vdots \\ n \end{array}
\begin{array}{c}
\begin{array}{ccccccccc} 1 & 2 & \ldots & j-1 & j & j+1 & \ldots & n-1 \end{array} \\
\begin{pmatrix}
1 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\
0 & 1 & \ldots & 0 & 0 & 0 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & 1 & 0 & 0 & \ldots & 0 \\
0 & 0 & \ldots & 0 & -c_{j+1}/c_j & -c_{j+2}/c_j & \ldots & -c_n/c_j \\
0 & 0 & \ldots & 0 & 1 & 0 & \ldots & 0 \\
0 & 0 & \ldots & 0 & 0 & 1 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 1
\end{pmatrix}
\end{array}.$$

Observe that the $(n-1)\times(n-1)$ matrix obtained by deleting row j is the identity matrix, so the columns of the above matrix are linearly independent. A simple calculation also shows that the linear form $u^*(x_1, \ldots, x_n) = c_jx_j + \cdots + c_nx_n$ vanishes on every column of the above matrix. For a concrete example in $\mathbb{R}^6$, if $u^*(x_1, \ldots, x_6) = x_3 + 2x_4 + 3x_5 + 4x_6$, we obtain the basis for the hyperplane H of equation

$$x_3 + 2x_4 + 3x_5 + 4x_6 = 0$$

given by the following matrix:

$$\begin{pmatrix}
1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 \\
0 & 0 & -2 & -3 & -4 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1
\end{pmatrix}.$$

Conversely, given a hyperplane H in $\mathbb{R}^n$ given as the span of $n-1$ linearly vectors $(u_1, \ldots, u_{n-1})$, it is possible using determinants to find a linear form $(\lambda_1, \ldots, \lambda_n)$ that vanishes

on H. In the case $n = 2$, we are looking for a row vector $(\lambda_1, \lambda_2, \lambda_3)$ such that if

$$u = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \quad \text{and} \quad v = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}$$

are two linearly independent vectors, then

$$\begin{pmatrix} u_1 & u_2 & u_2 \\ v_1 & v_2 & v_2 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

and the cross-product $u \times v$ of u and v given by

$$u \times v = \begin{pmatrix} u_2 v_3 - u_3 v_2 \\ u_3 v_1 - u_1 v_3 \\ u_1 v_2 - u_2 v_1 \end{pmatrix}$$

is a solution.

Here is another example illustrating the power of Theorem 8.1. Let $E = \mathrm{M}_n(\mathbb{R})$, and consider the equations asserting that the sum of the entries in every row of a matrix $A \in \mathrm{M}_n(\mathbb{R})$ is equal to the same number. We have $n - 1$ equations

$$\sum_{j=1}^{n} (a_{ij} - a_{i+1j}) = 0, \quad 1 \leq i \leq n - 1,$$

and it is easy to see that they are linearly independent. Therefore, the space U of linear forms in E^* spanned by the above linear forms (equations) has dimension $n - 1$, and the space U^0 of matrices sastisfying all these equations has dimension $n^2 - n + 1$. It is not so obvious to find a basis for this space.

We will now pin down the relationship between a vector space E and its bidual E^{**}.

Proposition 8.2. *Let E be a vector space. The following properties hold:*

(a) The linear map $\mathrm{eval}_E \colon E \to E^{**}$ defined such that

$$\mathrm{eval}_E(v) = \mathrm{eval}_v \quad \textit{for all } v \in E,$$

that is, $\mathrm{eval}_E(v)(u^*) = \langle u^*, v \rangle = u^*(v)$ for every $u^* \in E^*$, is injective.

(b) When E is of finite dimension n, the linear map $\mathrm{eval}_E \colon E \to E^{**}$ is an isomorphism (called the canonical isomorphism).

Proof. (a) Let $(u_i)_{i\in I}$ be a basis of E, and let $v = \sum_{i\in I} v_i u_i$. If $\mathrm{eval}_E(v) = 0$, then in particular $\mathrm{eval}_E(v)(u_i^*) = 0$ for all u_i^*, and since

$$\mathrm{eval}_E(v)(u_i^*) = \langle u_i^*, v\rangle = v_i,$$

we have $v_i = 0$ for all $i \in I$, that is, $v = 0$, showing that $\mathrm{eval}_E\colon E \to E^{**}$ is injective.

If E is of finite dimension n, by Theorem 8.1, for every basis $(u_1, \ldots, u_n)$, the family $(u_1^*, \ldots, u_n^*)$ is a basis of the dual space E^*, and thus the family $(u_1^{**}, \ldots, u_n^{**})$ is a basis of the bidual E^{**}. This shows that $\dim(E) = \dim(E^{**}) = n$, and since by part (a), we know that $\mathrm{eval}_E\colon E \to E^{**}$ is injective, in fact, $\mathrm{eval}_E\colon E \to E^{**}$ is bijective (by Proposition 3.9). □

When E is of finite dimension and $(u_1, \ldots, u_n)$ is a basis of E, in view of the canonical isomorphism $\mathrm{eval}_E\colon E \to E^{**}$, the basis $(u_1^{**}, \ldots, u_n^{**})$ of the bidual is identified with $(u_1, \ldots, u_n)$.

Proposition 8.2 can be reformulated very fruitfully in terms of pairings, a remarkably useful concept discovered by Pontrjagin in 1931 (adapted from E. Artin [5], Chapter 1). Given two vector spaces E and F over a field K, we say that a function $\varphi\colon E \times F \to K$ is *bilinear* if for every $v \in V$, the map $u \mapsto \varphi(u, v)$ (from E to K) is linear, and for every $u \in E$, the map $v \mapsto \varphi(u, v)$ (from F to K) is linear.

Definition 8.4. Given two vector spaces E and F over K, a *pairing between E and F* is a bilinear map $\varphi\colon E \times F \to K$. Such a pairing is *nondegenerate* iff

(1) for every $u \in E$, if $\varphi(u, v) = 0$ for all $v \in F$, then $u = 0$, and

(2) for every $v \in F$, if $\varphi(u, v) = 0$ for all $u \in E$, then $v = 0$.

A pairing $\varphi\colon E \times F \to K$ is often denoted by $\langle -, -\rangle\colon E \times F \to K$. For example, the map $\langle -, -\rangle\colon E^* \times E \to K$ defined earlier is a nondegenerate pairing (use the proof of (a) in Proposition 8.2). If $E = F$ and $K = \mathbb{R}$, any inner product on E is a nondegenerate pairing (because an inner product is positive definite); see Chapter 9.

Given a pairing $\varphi\colon E \times F \to K$, we can define two maps $l_\varphi\colon E \to F^*$ and $r_\varphi\colon F \to E^*$ as follows: For every $u \in E$, we define the linear form $l_\varphi(u)$ in F^* such that

$$l_\varphi(u)(y) = \varphi(u, y) \quad \text{for every } y \in F,$$

and for every $v \in F$, we define the linear form $r_\varphi(v)$ in E^* such that

$$r_\varphi(v)(x) = \varphi(x, v) \quad \text{for every } x \in E.$$

We have the following useful proposition.

Proposition 8.3. *Given two vector spaces E and F over K, for every nondegenerate pairing $\varphi\colon E \times F \to K$ between E and F, the maps $l_\varphi\colon E \to F^*$ and $r_\varphi\colon F \to E^*$ are linear and injective. Furthermore, if E and F have finite dimension, then this dimension is the same and $l_\varphi\colon E \to F^*$ and $r_\varphi\colon F \to E^*$ are bijections.*

Proof. The maps $l_\varphi\colon E \to F^*$ and $r_\varphi\colon F \to E^*$ are linear because a pairing is bilinear. If $l_\varphi(u) = 0$ (the null form), then

$$l_\varphi(u)(v) = \varphi(u, v) = 0 \quad \text{for every } v \in F,$$

and since φ is nondegenerate, $u = 0$. Thus, $l_\varphi\colon E \to F^*$ is injective. Similarly, $r_\varphi\colon F \to E^*$ is injective. When F has finite dimension n, we have seen that F and F^* have the same dimension. Since $l_\varphi\colon E \to F^*$ is injective, we have $m = \dim(E) \leq \dim(F) = n$. The same argument applies to E, and thus $n = \dim(F) \leq \dim(E) = m$. But then, $\dim(E) = \dim(F)$, and $l_\varphi\colon E \to F^*$ and $r_\varphi\colon F \to E^*$ are bijections. □

When E has finite dimension, the nondegenerate pairing $\langle -, - \rangle\colon E^* \times E \to K$ yields another proof of the existence of a natural isomorphism between E and E^{**}. When $E = F$, the nondegenerate pairing induced by an inner product on E yields a natural isomorphism between E and E^* (see Section 9.2).

Interesting nondegenerate pairings arise in exterior algebra and differential geometry. We now show the relationship between hyperplanes and linear forms.

8.4 Hyperplanes and Linear Forms

Actually, Proposition 8.4 below follows from parts (c) and (d) of Theorem 8.1, but we feel that it is also interesting to give a more direct proof.

Proposition 8.4. *Let E be a vector space. The following properties hold:*

(a) Given any nonnull linear form $f^* \in E^*$, its kernel $H = \operatorname{Ker} f^*$ is a hyperplane.

(b) For any hyperplane H in E, there is a (nonnull) linear form $f^* \in E^*$ such that $H = \operatorname{Ker} f^*$.

(c) Given any hyperplane H in E and any (nonnull) linear form $f^* \in E^*$ such that $H = \operatorname{Ker} f^*$, for every linear form $g^* \in E^*$, $H = \operatorname{Ker} g^*$ iff $g^* = \lambda f^*$ for some $\lambda \neq 0$ in K.

Proof. (a) If $f^* \in E^*$ is nonnull, there is some vector $v_0 \in E$ such that $f^*(v_0) \neq 0$. Let $H = \operatorname{Ker} f^*$. For every $v \in E$, we have

$$f^*\left(v - \frac{f^*(v)}{f^*(v_0)} v_0\right) = f^*(v) - \frac{f^*(v)}{f^*(v_0)} f^*(v_0) = f^*(v) - f^*(v) = 0.$$

Thus,

$$v - \frac{f^*(v)}{f^*(v_0)}v_0 = h \in H,$$

and

$$v = h + \frac{f^*(v)}{f^*(v_0)}v_0,$$

that is, $E = H + Kv_0$. Also, since $f^*(v_0) \neq 0$, we have $v_0 \notin H$, that is, $H \cap Kv_0 = 0$. Thus, $E = H \oplus Kv_0$, and H is a hyperplane.

(b) If H is a hyperplane, $E = H \oplus Kv_0$ for some $v_0 \notin H$. Then, every $v \in E$ can be written in a unique way as $v = h + \lambda v_0$. Thus, there is a well-defined function $f^*\colon E \to K$, such that, $f^*(v) = \lambda$, for every $v = h + \lambda v_0$. We leave as a simple exercise the verification that f^* is a linear form. Since $f^*(v_0) = 1$, the linear form f^* is nonnull. Also, by definition, it is clear that $\lambda = 0$ iff $v \in H$, that is, $\operatorname{Ker} f^* = H$.

(c) Let H be a hyperplane in E, and let $f^* \in E^*$ be any (nonnull) linear form such that $H = \operatorname{Ker} f^*$. Clearly, if $g^* = \lambda f^*$ for some $\lambda \neq 0$, then $H = \operatorname{Ker} g^*$. Conversely, assume that $H = \operatorname{Ker} g^*$ for some nonnull linear form g^*. From (a), we have $E = H \oplus Kv_0$, for some v_0 such that $f^*(v_0) \neq 0$ and $g^*(v_0) \neq 0$. Then, observe that

$$g^* - \frac{g^*(v_0)}{f^*(v_0)}f^*$$

is a linear form that vanishes on H, since both f^* and g^* vanish on H, but also vanishes on Kv_0. Thus, $g^* = \lambda f^*$, with

$$\lambda = \frac{g^*(v_0)}{f^*(v_0)}.$$

□

We leave as an exercise the fact that every subspace $V \neq E$ of a vector space E is the intersection of all hyperplanes that contain V. We now consider the notion of transpose of a linear map and of a matrix.

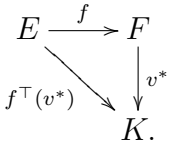
8.5 Transpose of a Linear Map and of a Matrix

Given a linear map $f\colon E \to F$, it is possible to define a map $f^\top\colon F^* \to E^*$ which has some interesting properties.

Definition 8.5. Given a linear map $f\colon E \to F$, the *transpose* $f^\top\colon F^* \to E^*$ *of* f is the linear map defined such that

$$f^\top(v^*) = v^* \circ f, \quad \text{for every } v^* \in F^*,$$

as shown in the diagram below:



Equivalently, the linear map $f^\top \colon F^* \to E^*$ is defined such that

$$\langle v^*, f(u)\rangle = \langle f^\top(v^*), u\rangle,$$

for all $u \in E$ and all $v^* \in F^*$.

It is easy to verify that the following properties hold:

$$(f+g)^\top = f^\top + g^\top$$
$$(g \circ f)^\top = f^\top \circ g^\top$$
$$\mathrm{id}_E^\top = \mathrm{id}_{E^*}.$$

Note the reversal of composition on the right-hand side of $(g \circ f)^\top = f^\top \circ g^\top$.

The equation $(g \circ f)^\top = f^\top \circ g^\top$ implies the following useful proposition.

Proposition 8.5. *If $f \colon E \to F$ is any linear map, then the following properties hold:*

(1) If f is injective, then $f^\top$ is surjective.

(2) If f is surjective, then $f^\top$ is injective.

Proof. If $f \colon E \to F$ is injective, then it has a retraction $r \colon F \to E$ such that $r \circ f = \mathrm{id}_E$, and if $f \colon E \to F$ is surjective, then it has a section $s \colon F \to E$ such that $f \circ s = \mathrm{id}_F$. Now, if $f \colon E \to F$ is injective, then we have

$$(r \circ f)^\top = f^\top \circ r^\top = \mathrm{id}_{E^*},$$

which implies that $f^\top$ is surjective, and if f is surjective, then we have

$$(f \circ s)^\top = s^\top \circ f^\top = \mathrm{id}_{F^*},$$

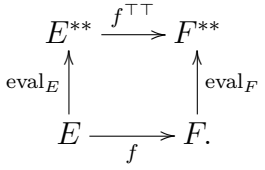
which implies that $f^\top$ is injective. □

We also have the following property showing the naturality of the eval map.

Proposition 8.6. *For any linear map $f \colon E \to F$, we have*

$$f^{\top\top} \circ \mathrm{eval}_E = \mathrm{eval}_F \circ f,$$

or equivalently the following diagram commutes:



Proof. For every $u \in E$ and every $\varphi \in F^*$, we have

$$\begin{aligned}(f^{\top\top} \circ \mathrm{eval}_E)(u)(\varphi) &= \langle f^{\top\top}(\mathrm{eval}_E(u)), \varphi\rangle \\ &= \langle \mathrm{eval}_E(u), f^{\top}(\varphi)\rangle \\ &= \langle f^{\top}(\varphi), u\rangle \\ &= \langle \varphi, f(u)\rangle \\ &= \langle \mathrm{eval}_F(f(u)), \varphi\rangle \\ &= \langle (\mathrm{eval}_F \circ f)(u), \varphi\rangle \\ &= (\mathrm{eval}_F \circ f)(u)(\varphi),\end{aligned}$$

which proves that $f^{\top\top} \circ \mathrm{eval}_E = \mathrm{eval}_F \circ f$, as claimed. □

If E and F are finite-dimensional, then eval_E and eval_F are isomorphisms, so Proposition 8.6 shows that

$$f^{\top\top} = \mathrm{eval}_F^{-1} \circ f \circ \mathrm{eval}_E. \qquad (*)$$

The above equation is often interpreted as follows: if we identify E with its bidual E^{**} and F with its bidual F^{**}, then $f^{\top\top} = f$. This is an abuse of notation; the rigorous statement is $(*)$.

As a corollary of Proposition 8.6, if $\dim(E)$ is finite, then we have

$$\mathrm{Ker}\,(f^{\top\top}) = \mathrm{eval}_E(\mathrm{Ker}\,(f)).$$

Proof. Indeed, if E is finite-dimensional, the map $\mathrm{eval}_E \colon E \to E^{**}$ is an isomorphism, so every $\varphi \in E^{**}$ is of the form $\varphi = \mathrm{eval}_E(u)$ for some $u \in E$, the map $\mathrm{eval}_F \colon F \to F^{**}$ is injective, and we have

$$\begin{aligned} f^{\top\top}(\varphi) = 0 \quad &\text{iff} \quad f^{\top\top}(\mathrm{eval}_E(u)) = 0 \\ &\text{iff} \quad \mathrm{eval}_F(f(u)) = 0 \\ &\text{iff} \quad f(u) = 0 \\ &\text{iff} \quad u \in \mathrm{Ker}\,(f) \\ &\text{iff} \quad \varphi \in \mathrm{eval}_E(\mathrm{Ker}\,(f)),\end{aligned}$$

which proves that $\mathrm{Ker}\,(f^{\top\top}) = \mathrm{eval}_E(\mathrm{Ker}\,(f))$. □

The following proposition shows the relationship between orthogonality and transposition.

Proposition 8.7. *Given a linear map $f \colon E \to F$, for any subspace V of E, we have*

$$f(V)^0 = (f^{\top})^{-1}(V^0) = \{w^* \in F^* \mid f^{\top}(w^*) \in V^0\}.$$

As a consequence,

$$\mathrm{Ker}\, f^{\top} = (\mathrm{Im}\, f)^0 \quad \textit{and} \quad \mathrm{Ker}\, f = (\mathrm{Im}\, f^{\top})^0.$$

Proof. We have

$$\langle w^*, f(v)\rangle = \langle f^\top(w^*), v\rangle,$$

for all $v \in E$ and all $w^* \in F^*$, and thus, we have $\langle w^*, f(v)\rangle = 0$ for every $v \in V$, i.e. $w^* \in f(V)^0$ iff $\langle f^\top(w^*), v\rangle = 0$ for every $v \in V$ iff $f^\top(w^*) \in V^0$, i.e. $w^* \in (f^\top)^{-1}(V^0)$, proving that

$$f(V)^0 = (f^\top)^{-1}(V^0).$$

Since we already observed that $E^0 = (0)$, letting $V = E$ in the above identity we obtain that

$$\operatorname{Ker} f^\top = (\operatorname{Im} f)^0.$$

From the equation

$$\langle w^*, f(v)\rangle = \langle f^\top(w^*), v\rangle,$$

we deduce that $v \in (\operatorname{Im} f^\top)^0$ iff $\langle f^\top(w^*), v\rangle = 0$ for all $w^* \in F^*$ iff $\langle w^*, f(v)\rangle = 0$ for all $w^* \in F^*$. Assume that $v \in (\operatorname{Im} f^\top)^0$. If we pick a basis $(w_i)_{i\in I}$ of F, then we have the linear forms $w_i^*\colon F \to K$ such that $w_i^*(w_j) = \delta_{ij}$, and since we must have $\langle w_i^*, f(v)\rangle = 0$ for all $i \in I$ and $(w_i)_{i\in I}$ is a basis of F, we conclude that $f(v) = 0$, and thus $v \in \operatorname{Ker} f$ (this is because $\langle w_i^*, f(v)\rangle$ is the coefficient of $f(v)$ associated with the basis vector w_i). Conversely, if $v \in \operatorname{Ker} f$, then $\langle w^*, f(v)\rangle = 0$ for all $w^* \in F^*$, so we conclude that $v \in (\operatorname{Im} f^\top)^0$. Therefore, $v \in (\operatorname{Im} f^\top)^0$ iff $v \in \operatorname{Ker} f$; that is,

$$\operatorname{Ker} f = (\operatorname{Im} f^\top)^0,$$

as claimed. □

The following theorem shows the relationship between the rank of f and the rank of $f^\top$.

Theorem 8.8. *Given a linear map $f\colon E \to F$, the following properties hold.*

(a) The dual $(\operatorname{Im} f)^*$ of $\operatorname{Im} f$ is isomorphic to $\operatorname{Im} f^\top = f^\top(F^*)$; that is,

$$(\operatorname{Im} f)^* \approx \operatorname{Im} f^\top.$$

(b) If F is finite dimensional, then $\operatorname{rk}(f) = \operatorname{rk}(f^\top)$.

Proof. (a) Consider the linear maps

$$E \xrightarrow{p} \operatorname{Im} f \xrightarrow{j} F,$$

where $E \xrightarrow{p} \operatorname{Im} f$ is the surjective map induced by $E \xrightarrow{f} F$, and $\operatorname{Im} f \xrightarrow{j} F$ is the injective inclusion map of $\operatorname{Im} f$ into F. By definition, $f = j \circ p$. To simplify the notation, let $I = \operatorname{Im} f$. By Proposition 8.5, since $E \xrightarrow{p} I$ is surjective, $I^* \xrightarrow{p^\top} E^*$ is injective, and since $\operatorname{Im} f \xrightarrow{j} F$ is injective, $F^* \xrightarrow{j^\top} I^*$ is surjective. Since $f = j \circ p$, we also have

$$f^\top = (j \circ p)^\top = p^\top \circ j^\top,$$

and since $F^* \xrightarrow{j^\top} I^*$ is surjective, and $I^* \xrightarrow{p^\top} E^*$ is injective, we have an isomorphism between $(\operatorname{Im} f)^*$ and $f^\top(F^*)$.

(b) We already noted that part (a) of Theorem 8.1 shows that $\dim(F) = \dim(F^*)$, for every vector space F of finite dimension. Consequently, $\dim(\operatorname{Im} f) = \dim((\operatorname{Im} f)^*)$, and thus, by part (a) we have $\operatorname{rk}(f) = \operatorname{rk}(f^\top)$.

When both E and F are finite-dimensional, there is also a simple proof of (b) that doesn't use the result of part (a). By Theorem 8.1(c)

$$\dim(\operatorname{Im} f) + \dim((\operatorname{Im} f)^0) = \dim(F),$$

and by Theorem 3.6

$$\dim(\operatorname{Ker} f^\top) + \dim(\operatorname{Im} f^\top) = \dim(F^*).$$

Furthermore, by Proposition 8.7, we have

$$\operatorname{Ker} f^\top = (\operatorname{Im} f)^0,$$

and since F is finite-dimensional $\dim(F) = \dim(F^*)$, so we deduce

$$\dim(\operatorname{Im} f) + \dim((\operatorname{Im} f)^0) = \dim((\operatorname{Im} f)^0) + \dim(\operatorname{Im} f^\top),$$

which yields $\dim(\operatorname{Im} f) = \dim(\operatorname{Im} f^\top)$; that is, $\operatorname{rk}(f) = \operatorname{rk}(f^\top)$. □

Remarks:

1. If $\dim(E)$ is finite, following an argument of Dan Guralnik, we can also prove that $\operatorname{rk}(f) = \operatorname{rk}(f^\top)$ as follows.

   We know from Proposition 8.7 applied to $f^\top \colon F^* \to E^*$ that

   $$\operatorname{Ker}(f^{\top\top}) = (\operatorname{Im} f^\top)^0,$$

   and we showed as a consequence of Proposition 8.6 that

   $$\operatorname{Ker}(f^{\top\top}) = \operatorname{eval}_E(\operatorname{Ker}(f)).$$

   It follows (since eval_E is an isomorphism) that

   $$\dim((\operatorname{Im} f^\top)^0) = \dim(\operatorname{Ker}(f^{\top\top})) = \dim(\operatorname{Ker}(f)) = \dim(E) - \dim(\operatorname{Im} f),$$

   and since

   $$\dim(\operatorname{Im} f^\top) + \dim((\operatorname{Im} f^\top)^0) = \dim(E),$$

   we get

   $$\dim(\operatorname{Im} f^\top) = \dim(\operatorname{Im} f).$$

2. As indicated by Dan Guralnik, if $\dim(E)$ is finite, the above result can be used to prove that
$$\operatorname{Im} f^\top = (\operatorname{Ker}(f))^0.$$
From
$$\langle f^\top(\varphi), u\rangle = \langle \varphi, f(u)\rangle$$
for all $\varphi \in F^*$ and all $u \in E$, we see that if $u \in \operatorname{Ker}(f)$, then $\langle f^\top(\varphi), u\rangle = \langle \varphi, 0\rangle = 0$, which means that $f^\top(\varphi) \in (\operatorname{Ker}(f))^0$, and thus, $\operatorname{Im} f^\top \subseteq (\operatorname{Ker}(f))^0$. For the converse, since $\dim(E)$ is finite, we have
$$\dim((\operatorname{Ker}(f))^0) = \dim(E) - \dim(\operatorname{Ker}(f)) = \dim(\operatorname{Im} f),$$
but we just proved that $\dim(\operatorname{Im} f^\top) = \dim(\operatorname{Im} f)$, so we get
$$\dim((\operatorname{Ker}(f))^0) = \dim(\operatorname{Im} f^\top),$$
and since $\operatorname{Im} f^\top \subseteq (\operatorname{Ker}(f))^0$, we obtain
$$\operatorname{Im} f^\top = (\operatorname{Ker}(f))^0,$$
as claimed. Now, since $(\operatorname{Ker}(f))^{00} = \operatorname{Ker}(f)$, the above equation yields another proof of the fact that
$$\operatorname{Ker}(f) = (\operatorname{Im} f^\top)^0,$$
when E is finite-dimensional.

3. The equation
$$\operatorname{Im} f^\top = (\operatorname{Ker}(f))^0$$
is actually valid even if when E if infinite-dimensional, but we will not prove this here.

The following proposition can be shown, but it requires a generalization of the duality theorem, so is proof is omitted.

Proposition 8.9. *If $f\colon E \to F$ is any linear map, then the following identities hold:*
$$\begin{aligned}
\operatorname{Im} f^\top &= (\operatorname{Ker}(f))^0\\
\operatorname{Ker}(f^\top) &= (\operatorname{Im} f)^0\\
\operatorname{Im} f &= (\operatorname{Ker}(f^\top)^0\\
\operatorname{Ker}(f) &= (\operatorname{Im} f^\top)^0.
\end{aligned}$$

The following proposition shows the relationship between the matrix representing a linear map $f\colon E \to F$ and the matrix representing its transpose $f^\top\colon F^* \to E^*$.

Proposition 8.10. *Let E and F be two vector spaces, and let $(u_1, \ldots, u_n)$ be a basis for E and $(v_1, \ldots, v_m)$ be a basis for F. Given any linear map $f\colon E \to F$, if $M(f)$ is the $m \times n$-matrix representing f w.r.t. the bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_m)$ the $n \times m$-matrix $M(f^\top)$ representing $f^\top\colon F^* \to E^*$ w.r.t. the dual bases $(v_1^*, \ldots, v_m^*)$ and $(u_1^*, \ldots, u_n^*)$ is the transpose $M(f)^\top$ of $M(f)$.*

Proof. Recall that the entry $a_{i\,j}$ in row i and column j of $M(f)$ is the i-th coordinate of $f(u_j)$ over the basis $(v_1, \ldots, v_m)$. By definition of v_i^*, we have $\langle v_i^*, f(u_j)\rangle = a_{i\,j}$. The entry $a_{j\,i}^\top$ in row j and column i of $M(f^\top)$ is the j-th coordinate of

$$f^\top(v_i^*) = a_{1\,i}^\top u_1^* + \cdots + a_{j\,i}^\top u_j^* + \cdots + a_{n\,i}^\top u_n^*$$

over the basis $(u_1^*, \ldots, u_n^*)$, which is just $a_{j\,i}^\top = f^\top(v_i^*)(u_j) = \langle f^\top(v_i^*), u_j\rangle$. Since

$$\langle v_i^*, f(u_j)\rangle = \langle f^\top(v_i^*), u_j\rangle,$$

we have $a_{i\,j} = a_{j\,i}^\top$, proving that $M(f^\top) = M(f)^\top$. □

We now can give a very short proof of the fact that the rank of a matrix is equal to the rank of its transpose.

Proposition 8.11. *Given a $m \times n$ matrix A over a field K, we have* $\mathrm{rk}(A) = \mathrm{rk}(A^\top)$.

Proof. The matrix A corresponds to a linear map $f\colon K^n \to K^m$, and by Theorem 8.8, $\mathrm{rk}(f) = \mathrm{rk}(f^\top)$. By Proposition 8.10, the linear map $f^\top$ corresponds to $A^\top$. Since $\mathrm{rk}(A) = \mathrm{rk}(f)$, and $\mathrm{rk}(A^\top) = \mathrm{rk}(f^\top)$, we conclude that $\mathrm{rk}(A) = \mathrm{rk}(A^\top)$. □

Thus, given an $m \times n$-matrix A, the maximum number of linearly independent columns is equal to the maximum number of linearly independent rows. There are other ways of proving this fact that do not involve the dual space, but instead some elementary transformations on rows and columns.

Proposition 8.11 immediately yields the following criterion for determining the rank of a matrix:

Proposition 8.12. *Given any $m \times n$ matrix A over a field K (typically $K = \mathbb{R}$ or $K = \mathbb{C}$), the rank of A is the maximum natural number r such that there is an invertible $r \times r$ submatrix of A obtained by selecting r rows and r columns of A.*

For example, the 3×2 matrix

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix}$$

has rank 2 iff one of the three 2×2 matrices

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \quad \begin{pmatrix} a_{11} & a_{12} \\ a_{31} & a_{32} \end{pmatrix} \quad \begin{pmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix}$$

is invertible. We saw in Chapter 4 that this is equivalent to the fact the determinant of one of the above matrices is nonzero. This is not a very efficient way of finding the rank of a matrix. We will see that there are better ways using various decompositions such as LU, QR, or SVD.

8.6 The Four Fundamental Subspaces

Given a linear map $f\colon E \to F$ (where E and F are finite-dimensional), Proposition 8.7 revealed that the four spaces

$$\operatorname{Im} f,\ \operatorname{Im} f^\top,\ \operatorname{Ker} f,\ \operatorname{Ker} f^\top$$

play a special role. They are often called the *fundamental subspaces* associated with f. These spaces are related in an intimate manner, since Proposition 8.7 shows that

$$\begin{aligned} \operatorname{Ker} f &= (\operatorname{Im} f^\top)^0 \\ \operatorname{Ker} f^\top &= (\operatorname{Im} f)^0, \end{aligned}$$

and Theorem 8.8 shows that

$$\operatorname{rk}(f) = \operatorname{rk}(f^\top).$$

It is instructive to translate these relations in terms of matrices (actually, certain linear algebra books make a big deal about this!). If $\dim(E) = n$ and $\dim(F) = m$, given any basis $(u_1, \ldots, u_n)$ of E and a basis $(v_1, \ldots, v_m)$ of F, we know that f is represented by an $m \times n$ matrix $A = (a_{i\,j})$, where the jth column of A is equal to $f(u_j)$ over the basis $(v_1, \ldots, v_m)$. Furthermore, the transpose map $f^\top$ is represented by the $n \times m$ matrix $A^\top$ (with respect to the dual bases). Consequently, the four fundamental spaces

$$\operatorname{Im} f,\ \operatorname{Im} f^\top,\ \operatorname{Ker} f,\ \operatorname{Ker} f^\top$$

correspond to

(1) The *column space* of A, denoted by $\operatorname{Im} A$ or $\mathcal{R}(A)$; this is the subspace of $\mathbb{R}^m$ spanned by the columns of A, which corresponds to the image $\operatorname{Im} f$ of f.

(2) The *kernel* or *nullspace* of A, denoted by $\operatorname{Ker} A$ or $\mathcal{N}(A)$; this is the subspace of $\mathbb{R}^n$ consisting of all vectors $x \in \mathbb{R}^n$ such that $Ax = 0$.

(3) The *row space* of A, denoted by $\operatorname{Im} A^\top$ or $\mathcal{R}(A^\top)$; this is the subspace of $\mathbb{R}^n$ spanned by the rows of A, or equivalently, spanned by the columns of $A^\top$, which corresponds to the image $\operatorname{Im} f^\top$ of $f^\top$.

(4) The *left kernel* or *left nullspace* of A denoted by $\operatorname{Ker} A^\top$ or $\mathcal{N}(A^\top)$; this is the kernel (nullspace) of $A^\top$, the subspace of $\mathbb{R}^m$ consisting of all vectors $y \in \mathbb{R}^m$ such that $A^\top y = 0$, or equivalently, $y^\top A = 0$.

Recall that the dimension r of $\operatorname{Im} f$, which is also equal to the dimension of the column space $\operatorname{Im} A = \mathcal{R}(A)$, is the *rank* of A (and f). Then, some our previous results can be reformulated as follows:

1. The column space $\mathcal{R}(A)$ of A has dimension r.
2. The nullspace $\mathcal{N}(A)$ of A has dimension $n - r$.
3. The row space $\mathcal{R}(A^\top)$ has dimension r.
4. The left nullspace $\mathcal{N}(A^\top)$ of A has dimension $m - r$.

The above statements constitute what Strang calls the *Fundamental Theorem of Linear Algebra, Part I* (see Strang [102]).

The two statements

$$\operatorname{Ker} f = (\operatorname{Im} f^\top)^0$$
$$\operatorname{Ker} f^\top = (\operatorname{Im} f)^0$$

translate to

(1) The nullspace of A is the orthogonal of the row space of A.

(2) The left nullspace of A is the orthogonal of the column space of A.

The above statements constitute what Strang calls the *Fundamental Theorem of Linear Algebra, Part II* (see Strang [102]).

Since vectors are represented by column vectors and linear forms by row vectors (over a basis in E or F), a vector $x \in \mathbb{R}^n$ is orthogonal to a linear form y iff

$$yx = 0.$$

Then, a vector $x \in \mathbb{R}^n$ is orthogonal to the row space of A iff x is orthogonal to every row of A, namely $Ax = 0$, which is equivalent to the fact that x belong to the nullspace of A. Similarly, the column vector $y \in \mathbb{R}^m$ (representing a linear form over the dual basis of F^*) belongs to the nullspace of $A^\top$ iff $A^\top y = 0$, iff $y^\top A = 0$, which means that the linear form given by $y^\top$ (over the basis in F) is orthogonal to the column space of A.

Since (2) is equivalent to the fact that the column space of A is equal to the orthogonal of the left nullspace of A, we get the following criterion for the solvability of an equation of the form $Ax = b$:

The equation $Ax = b$ has a solution iff for all $y \in \mathbb{R}^m$, if $A^\top y = 0$, then $y^\top b = 0$.

Indeed, the condition on the right-hand side says that b is orthogonal to the left nullspace of A; that is, b belongs to the column space of A.

This criterion can be cheaper to check that checking directly that b is spanned by the columns of A. For example, if we consider the system

$$\begin{aligned} x_1 - x_2 &= b_1 \\ x_2 - x_3 &= b_2 \\ x_3 - x_1 &= b_3 \end{aligned}$$

which, in matrix form, is written $Ax = b$ as below:

$$\begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix},$$

we see that the rows of the matrix A add up to 0. In fact, it is easy to convince ourselves that the left nullspace of A is spanned by $y = (1, 1, 1)$, and so the system is solvable iff $y^\top b = 0$, namely

$$b_1 + b_2 + b_3 = 0.$$

Note that the above criterion can also be stated negatively as follows:

The equation $Ax = b$ has no solution iff there is some $y \in \mathbb{R}^m$ such that $A^\top y = 0$ and $y^\top b \neq 0$.

Since $A^\top y = 0$ iff $y^\top A = 0$, we can view $y^\top$ as a row vector representing a linear form, and $y^\top A = 0$ asserts that the linear form $y^\top$ vanishes on the columns $A^1, \ldots, A^n$ of A but does not vanish on b. Since the linear form $y^\top$ defines the hyperplane H of equation $y^\top z = 0$ (with $z \in \mathbb{R}^m$), geometrically the equation $Ax = b$ has no solution iff there is a hyperplane H containing $A^1, \ldots, A^n$ and not containing b.

8.7 Summary

The main concepts and results of this chapter are listed below:

- The *dual space* E^* and *linear forms* (*covector*). The *bidual* E^{**}.
- The *bilinear pairing* $\langle -, - \rangle \colon E^* \times E \to K$ (the *canonical pairing*).
- *Evaluation at v*: $\mathrm{eval}_v \colon E^* \to K$.
- The map $\mathrm{eval}_E \colon E \to E^{**}$.
- *Othogonality* between a subspace V of E and a subspace U of E^*; the *orthogonal* V^0 and the *orthogonal* U^0.

- *Coordinate forms*.
- The *Duality theorem* (Theorem 8.1).
- The *dual basis* of a basis.
- The isomorphism $\mathrm{eval}_E\colon E \to E^{**}$ when $\dim(E)$ is finite.
- *Pairing* between two vector spaces; *nondegenerate pairing*; Proposition 8.3.
- Hyperplanes and linear forms.
- The *transpose* $f^\top\colon F^* \to E^*$ of a linear map $f\colon E \to F$.
- The fundamental identities:

$$\operatorname{Ker} f^\top = (\operatorname{Im} f)^0 \quad \text{and} \quad \operatorname{Ker} f = (\operatorname{Im} f^\top)^0$$

  (Proposition 8.7).
- If F is finite-dimensional, then

$$\mathrm{rk}(f) = \mathrm{rk}(f^\top).$$

  (Theorem 8.8).
- The matrix of the transpose map $f^\top$ is equal to the transpose of the matrix of the map f (Proposition 8.10).
- For any $m \times n$ matrix A,

$$\mathrm{rk}(A) = \mathrm{rk}(A^\top).$$

- Characterization of the rank of a matrix in terms of a maximal invertible submatrix (Proposition 8.12).
- The *four fundamental subspaces*:

$$\operatorname{Im} f,\ \operatorname{Im} f^\top,\ \operatorname{Ker} f,\ \operatorname{Ker} f^\top.$$

- The *column space*, the *nullspace*, the *row space*, and the *left nullspace* (of a matrix).
- Criterion for the solvability of an equation of the form $Ax = b$ in terms of the left nullspace.

Chapter 9

Euclidean Spaces

Rien n'est beau que le vrai.
—**Hermann Minkowski**

9.1 Inner Products, Euclidean Spaces

So far, the framework of vector spaces allows us to deal with ratios of vectors and linear combinations, but there is no way to express the notion of length of a line segment or to talk about orthogonality of vectors. A Euclidean structure allows us to deal with *metric notions* such as orthogonality and length (or distance).

This chapter covers the bare bones of Euclidean geometry. Deeper aspects of Euclidean geometry are investigated in Chapter 10. One of our main goals is to give the basic properties of the transformations that preserve the Euclidean structure, rotations and reflections, since they play an important role in practice. Euclidean geometry is the study of properties invariant under certain affine maps called *rigid motions*. Rigid motions are the maps that preserve the distance between points.

We begin by defining inner products and Euclidean spaces. The Cauchy–Schwarz inequality and the Minkowski inequality are shown. We define orthogonality of vectors and of subspaces, orthogonal bases, and orthonormal bases. We prove that every finite-dimensional Euclidean space has orthonormal bases. The first proof uses duality, and the second one the Gram–Schmidt orthogonalization procedure. The QR-decomposition for invertible matrices is shown as an application of the Gram–Schmidt procedure. Linear isometries (also called orthogonal transformations) are defined and studied briefly. We conclude with a short section in which some applications of Euclidean geometry are sketched. One of the most important applications, the method of least squares, is discussed in Chapter 16.

For a more detailed treatment of Euclidean geometry, see Berger [9, 10], Snapper and Troyer [97], or any other book on geometry, such as Pedoe [80], Coxeter [32], Fresnel [43], Tisseron [104], or Cagnac, Ramis, and Commeau [25]. Serious readers should consult Emil

Artin's famous book [5], which contains an in-depth study of the orthogonal group, as well as other groups arising in geometry. It is still worth consulting some of the older classics, such as Hadamard [51, 52] and Rouché and de Comberousse [81]. The first edition of [51] was published in 1898, and finally reached its thirteenth edition in 1947! In this chapter it is assumed that all vector spaces are defined over the field $\mathbb{R}$ of real numbers unless specified otherwise (in a few cases, over the complex numbers $\mathbb{C}$).

First, we define a Euclidean structure on a vector space. Technically, a Euclidean structure over a vector space E is provided by a symmetric bilinear form on the vector space satisfying some extra properties. Recall that a bilinear form $\varphi\colon E \times E \to \mathbb{R}$ is *definite* if for every $u \in E$, $u \neq 0$ implies that $\varphi(u, u) \neq 0$, and *positive* if for every $u \in E$, $\varphi(u, u) \geq 0$.

Definition 9.1. A *Euclidean space* is a real vector space E equipped with a symmetric bilinear form $\varphi\colon E \times E \to \mathbb{R}$ that is *positive definite*. More explicitly, $\varphi\colon E \times E \to \mathbb{R}$ satisfies the following axioms:

$$\begin{aligned}
\varphi(u_1 + u_2, v) &= \varphi(u_1, v) + \varphi(u_2, v),\\
\varphi(u, v_1 + v_2) &= \varphi(u, v_1) + \varphi(u, v_2),\\
\varphi(\lambda u, v) &= \lambda\varphi(u, v),\\
\varphi(u, \lambda v) &= \lambda\varphi(u, v),\\
\varphi(u, v) &= \varphi(v, u),\\
u &\neq 0 \text{ implies that } \varphi(u, u) > 0.
\end{aligned}$$

The real number $\varphi(u, v)$ is also called the *inner product (or scalar product) of u and v*. We also define the *quadratic form associated with φ* as the function $\Phi\colon E \to \mathbb{R}_+$ such that

$$\Phi(u) = \varphi(u, u),$$

for all $u \in E$.

Since φ is bilinear, we have $\varphi(0, 0) = 0$, and since it is positive definite, we have the stronger fact that

$$\varphi(u, u) = 0 \quad \text{iff} \quad u = 0,$$

that is, $\Phi(u) = 0$ iff $u = 0$.

Given an inner product $\varphi\colon E \times E \to \mathbb{R}$ on a vector space E, we also denote $\varphi(u, v)$ by

$$u \cdot v \quad \text{or} \quad \langle u, v\rangle \quad \text{or} \quad (u|v),$$

and $\sqrt{\Phi(u)}$ by $\|u\|$.

Example 9.1. The standard example of a Euclidean space is $\mathbb{R}^n$, under the inner product $\cdot$ defined such that

$$(x_1, \ldots, x_n) \cdot (y_1, \ldots, y_n) = x_1y_1 + x_2y_2 + \cdots + x_ny_n.$$

This Euclidean space is denoted by $\mathbb{E}^n$.

There are other examples.

Example 9.2. For instance, let E be a vector space of dimension 2, and let (e_1, e_2) be a basis of E. If $a > 0$ and $b^2 - ac < 0$, the bilinear form defined such that

$$\varphi(x_1e_1 + y_1e_2,\, x_2e_1 + y_2e_2) = ax_1x_2 + b(x_1y_2 + x_2y_1) + cy_1y_2$$

yields a Euclidean structure on E. In this case,

$$\Phi(xe_1 + ye_2) = ax^2 + 2bxy + cy^2.$$

Example 9.3. Let $\mathcal{C}[a, b]$ denote the set of continuous functions $f\colon [a, b] \to \mathbb{R}$. It is easily checked that $\mathcal{C}[a, b]$ is a vector space of infinite dimension. Given any two functions $f, g \in \mathcal{C}[a, b]$, let

$$\langle f, g\rangle = \int_a^b f(t)g(t)dt.$$

We leave as an easy exercise that $\langle -, -\rangle$ is indeed an inner product on $\mathcal{C}[a, b]$. In the case where $a = -\pi$ and $b = \pi$ (or $a = 0$ and $b = 2\pi$, this makes basically no difference), one should compute

$$\langle \sin px, \sin qx\rangle, \quad \langle \sin px, \cos qx\rangle, \quad \text{and} \quad \langle \cos px, \cos qx\rangle,$$

for all natural numbers $p, q \geq 1$. The outcome of these calculations is what makes Fourier analysis possible!

Example 9.4. Let $E = \mathrm{M}_n(\mathbb{R})$ be the vector space of real $n \times n$ matrices. If we view a matrix $A \in \mathrm{M}_n(\mathbb{R})$ as a "long" column vector obtained by concatenating together its columns, we can define the inner product of two matrices $A, B \in \mathrm{M}_n(\mathbb{R})$ as

$$\langle A, B\rangle = \sum_{i,j=1}^{n} a_{ij}b_{ij},$$

which can be conveniently written as

$$\langle A, B\rangle = \mathrm{tr}(A^\top B) = \mathrm{tr}(B^\top A).$$

Since this can be viewed as the Euclidean product on $\mathbb{R}^{n^2}$, it is an inner product on $\mathrm{M}_n(\mathbb{R})$. The corresponding norm

$$\|A\|_F = \sqrt{\mathrm{tr}(A^\top A)}$$

is the Frobenius norm (see Section 6.2).

Let us observe that φ can be recovered from Φ. Indeed, by bilinearity and symmetry, we have

$$\begin{aligned}
\Phi(u + v) &= \varphi(u + v,\, u + v)\\
&= \varphi(u,\, u + v) + \varphi(v,\, u + v)\\
&= \varphi(u,\, u) + 2\varphi(u,\, v) + \varphi(v,\, v)\\
&= \Phi(u) + 2\varphi(u,\, v) + \Phi(v).
\end{aligned}$$

Thus, we have

$$\varphi(u,\, v) = \frac{1}{2}[\Phi(u+v) - \Phi(u) - \Phi(v)].$$

We also say that φ is the *polar form of* Φ.

If E is finite-dimensional and if $\varphi\colon E \times E \to \mathbb{R}$ is a bilinear form on E, given any basis $(e_1, \ldots, e_n)$ of E, we can write $x = \sum_{i=1}^n x_i e_i$ and $y = \sum_{j=1}^n y_j e_j$, and we have

$$\varphi(x, y) = \varphi\Big(\sum_{i=1}^n x_i e_i, \sum_{j=1}^n y_j e_j\Big) = \sum_{i,j=1}^n x_i y_j \varphi(e_i, e_j).$$

If we let G be the matrix $G = (\varphi(e_i, e_j))$, and if x and y are the column vectors associated with $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$, then we can write

$$\varphi(x, y) = x^\top G y = y^\top G^\top x.$$

Note that we are committing an abuse of notation, since $x = \sum_{i=1}^n x_i e_i$ is a vector in E, but the column vector associated with $(x_1, \ldots, x_n)$ belongs to $\mathbb{R}^n$. To avoid this minor abuse, we could denote the column vector associated with $(x_1, \ldots, x_n)$ by $\mathbf{x}$ (and similarly $\mathbf{y}$ for the column vector associated with $(y_1, \ldots, y_n)$), in wich case the "correct" expression for $\varphi(x, y)$ is

$$\varphi(x, y) = \mathbf{x}^\top G \mathbf{y}.$$

However, in view of the isomorphism between E and $\mathbb{R}^n$, to keep notation as simple as possible, we will use x and y instead of $\mathbf{x}$ and $\mathbf{y}$.

Also observe that φ is symmetric iff $G = G^\top$, and φ is positive definite iff the matrix G is positive definite, that is,

$$x^\top G x > 0 \quad \text{for all } x \in \mathbb{R}^n,\ x \neq 0.$$

The matrix G associated with an inner product is called the *Gram matrix* of the inner product with respect to the basis $(e_1, \ldots, e_n)$.

Conversely, if A is a symmetric positive definite $n \times n$ matrix, it is easy to check that the bilinear form

$$\langle x, y\rangle = x^\top A y$$

is an inner product. If we make a change of basis from the basis $(e_1, \ldots, e_n)$ to the basis $(f_1, \ldots, f_n)$, and if the change of basis matrix is P (where the jth column of P consists of the coordinates of f_j over the basis $(e_1, \ldots, e_n)$), then with respect to coordinates x' and y' over the basis $(f_1, \ldots, f_n)$, we have

$$x^\top G y = x'^\top P^\top G P y',$$

so the matrix of our inner product over the basis $(f_1, \ldots, f_n)$ is $P^\top G P$. We summarize these facts in the following proposition.

Proposition 9.1. *Let E be a finite-dimensional vector space, and let $(e_1, \ldots, e_n)$ be a basis of E.*

1. *For any inner product $\langle -, - \rangle$ on E, if $G = (\langle e_i, e_j \rangle)$ is the Gram matrix of the inner product $\langle -, - \rangle$ w.r.t. the basis $(e_1, \ldots, e_n)$, then G is symmetric positive definite.*

2. *For any change of basis matrix P, the Gram matrix of $\langle -, - \rangle$ with respect to the new basis is $P^\top GP$.*

3. *If A is any $n \times n$ symmetric positive definite matrix, then*

$$\langle x, y \rangle = x^\top Ay$$

is an inner product on E.

We will see later that a symmetric matrix is positive definite iff its eigenvalues are all positive.

One of the very important properties of an inner product φ is that the map $u \mapsto \sqrt{\Phi(u)}$ is a norm.

Proposition 9.2. *Let E be a Euclidean space with inner product φ, and let Φ be the corresponding quadratic form. For all $u, v \in E$, we have the* Cauchy–Schwarz inequality

$$\varphi(u, v)^2 \leq \Phi(u)\Phi(v),$$

the equality holding iff u and v are linearly dependent.

We also have the Minkowski inequality

$$\sqrt{\Phi(u + v)} \leq \sqrt{\Phi(u)} + \sqrt{\Phi(v)},$$

the equality holding iff u and v are linearly dependent, where in addition if $u \neq 0$ and $v \neq 0$, then $u = \lambda v$ for some $\lambda > 0$.

Proof. For any vectors $u, v \in E$, we define the function $T \colon \mathbb{R} \to \mathbb{R}$ such that

$$T(\lambda) = \Phi(u + \lambda v),$$

for all $\lambda \in \mathbb{R}$. Using bilinearity and symmetry, we have

$$\begin{aligned} \Phi(u + \lambda v) &= \varphi(u + \lambda v,\, u + \lambda v) \\ &= \varphi(u,\, u + \lambda v) + \lambda\varphi(v,\, u + \lambda v) \\ &= \varphi(u,\, u) + 2\lambda\varphi(u,\, v) + \lambda^2\varphi(v,\, v) \\ &= \Phi(u) + 2\lambda\varphi(u,\, v) + \lambda^2\Phi(v). \end{aligned}$$

Since φ is positive definite, Φ is nonnegative, and thus $T(\lambda) \geq 0$ for all $\lambda \in \mathbb{R}$. If $\Phi(v) = 0$, then $v = 0$, and we also have $\varphi(u,\, v) = 0$. In this case, the Cauchy–Schwarz inequality is trivial, and $v = 0$ and u are linearly dependent.

Now, assume $\Phi(v) > 0$. Since $T(\lambda) \geq 0$, the quadratic equation

$$\lambda^2\Phi(v) + 2\lambda\varphi(u,\, v) + \Phi(u) = 0$$

cannot have distinct real roots, which means that its discriminant

$$\Delta = 4(\varphi(u,\, v)^2 - \Phi(u)\Phi(v))$$

is null or negative, which is precisely the Cauchy–Schwarz inequality

$$\varphi(u, v)^2 \leq \Phi(u)\Phi(v).$$

If

$$\varphi(u, v)^2 = \Phi(u)\Phi(v)$$

then there are two cases. If $\Phi(v) = 0$, then $v = 0$ and u and v are linearly dependent. If $\Phi(v) \neq 0$, then the above quadratic equation has a double root λ_0, and we have $\Phi(u+\lambda_0 v) = 0$. Since φ is positive definite, $\Phi(u + \lambda_0 v) = 0$ implies that $u + \lambda_0 v = 0$, which shows that u and v are linearly dependent. Conversely, it is easy to check that we have equality when u and v are linearly dependent.

The Minkowski inequality

$$\sqrt{\Phi(u + v)} \leq \sqrt{\Phi(u)} + \sqrt{\Phi(v)}$$

is equivalent to

$$\Phi(u + v) \leq \Phi(u) + \Phi(v) + 2\sqrt{\Phi(u)\Phi(v)}.$$

However, we have shown that

$$2\varphi(u,\, v) = \Phi(u + v) - \Phi(u) - \Phi(v),$$

and so the above inequality is equivalent to

$$\varphi(u,\, v) \leq \sqrt{\Phi(u)\Phi(v)},$$

which is trivial when $\varphi(u,\, v) \leq 0$, and follows from the Cauchy–Schwarz inequality when $\varphi(u,\, v) \geq 0$. Thus, the Minkowski inequality holds. Finally, assume that $u \neq 0$ and $v \neq 0$, and that

$$\sqrt{\Phi(u + v)} = \sqrt{\Phi(u)} + \sqrt{\Phi(v)}.$$

When this is the case, we have

$$\varphi(u,\, v) = \sqrt{\Phi(u)\Phi(v)},$$

and we know from the discussion of the Cauchy–Schwarz inequality that the equality holds iff u and v are linearly dependent. The Minkowski inequality is an equality when u or v is null. Otherwise, if $u \neq 0$ and $v \neq 0$, then $u = \lambda v$ for some $\lambda \neq 0$, and since

$$\varphi(u,\, v) = \lambda\varphi(v,\, v) = \sqrt{\Phi(u)\Phi(v)},$$

by positivity, we must have $\lambda > 0$. □

Note that the Cauchy–Schwarz inequality can also be written as

$$|\varphi(u,v)| \leq \sqrt{\Phi(u)}\sqrt{\Phi(v)}.$$

Remark: It is easy to prove that the Cauchy–Schwarz and the Minkowski inequalities still hold for a symmetric bilinear form that is positive, but not necessarily definite (i.e., $\varphi(u,v) \geq 0$ for all $u, v \in E$). However, u and v need not be linearly dependent when the equality holds.

The Minkowski inequality

$$\sqrt{\Phi(u+v)} \leq \sqrt{\Phi(u)} + \sqrt{\Phi(v)}$$

shows that the map $u \mapsto \sqrt{\Phi(u)}$ satisfies the convexity inequality (also known as triangle inequality), condition (N3) of Definition 6.1, and since φ is bilinear and positive definite, it also satisfies conditions (N1) and (N2) of Definition 6.1, and thus it is a *norm* on E. The norm induced by φ is called the *Euclidean norm induced by* φ.

Note that the Cauchy–Schwarz inequality can be written as

$$|u \cdot v| \leq \|u\|\|v\|,$$

and the Minkowski inequality as

$$\|u+v\| \leq \|u\| + \|v\|.$$

Remark: One might wonder if every norm on a vector space is induced by some Euclidean inner product. In general, this is false, but remarkably, there is a simple necessary and sufficient condition, which is that the norm must satisfy the *parallelogram law*:

$$\|u+v\|^2 + \|u-v\|^2 = 2(\|u\|^2 + \|v\|^2).$$

If $\langle -, - \rangle$ is an inner product, then we have

$$\begin{aligned} \|u+v\|^2 &= \|u\|^2 + \|v\|^2 + 2\langle u, v\rangle \\ \|u-v\|^2 &= \|u\|^2 + \|v\|^2 - 2\langle u, v\rangle, \end{aligned}$$

and by adding and subtracting these identities, we get the parallelogram law and the equation

$$\langle u, v\rangle = \frac{1}{4}(\|u+v\|^2 - \|u-v\|^2),$$

which allows us to recover $\langle -, - \rangle$ from the norm.

Conversely, if $\| \|$ is a norm satisfying the parallelogram law, and if it comes from an inner product, then this inner product must be given by

$$\langle u, v\rangle = \frac{1}{4}(\|u+v\|^2 - \|u-v\|^2).$$

We need to prove that the above form is indeed symmetric and bilinear.

Symmetry holds because $\|u-v\| = \|-(u-v)\| = \|v-u\|$. Let us prove additivity in the variable u. By the parallelogram law, we have

$$2(\|x+z\|^2 + \|y\|^2) = \|x+y+z\|^2 + \|x-y+z\|^2$$

which yields

$$\begin{aligned}\|x+y+z\|^2 &= 2(\|x+z\|^2 + \|y\|^2) - \|x-y+z\|^2\\ \|x+y+z\|^2 &= 2(\|y+z\|^2 + \|x\|^2) - \|y-x+z\|^2,\end{aligned}$$

where the second formula is obtained by swapping x and y. Then by adding up these equations, we get

$$\|x+y+z\|^2 = \|x\|^2 + \|y\|^2 + \|x+z\|^2 + \|y+z\|^2 - \frac{1}{2}\|x-y+z\|^2 - \frac{1}{2}\|y-x+z\|^2.$$

Replacing z by $-z$ in the above equation, we get

$$\|x+y-z\|^2 = \|x\|^2 + \|y\|^2 + \|x-z\|^2 + \|y-z\|^2 - \frac{1}{2}\|x-y-z\|^2 - \frac{1}{2}\|y-x-z\|^2,$$

Since $\|x-y+z\| = \|-(x-y+z)\| = \|y-x-z\|$ and $\|y-x+z\| = \|-(y-x+z)\| = \|x-y-z\|$, by subtracting the last two equations, we get

$$\begin{aligned}\langle x+y, z\rangle &= \frac{1}{4}(\|x+y+z\|^2 - \|x+y-z\|^2)\\ &= \frac{1}{4}(\|x+z\|^2 - \|x-z\|^2) + \frac{1}{4}(\|y+z\|^2 - \|y-z\|^2)\\ &= \langle x, z\rangle + \langle y, z\rangle,\end{aligned}$$

as desired.

Proving that

$$\langle \lambda x, y\rangle = \lambda\langle x, y\rangle \quad \text{for all } \lambda \in \mathbb{R}$$

is a little tricky. The strategy is to prove the identity for $\lambda \in \mathbb{Z}$, then to promote it to $\mathbb{Q}$, and then to $\mathbb{R}$ by continuity.

Since

$$\begin{aligned}\langle -u, v\rangle &= \frac{1}{4}(\|-u+v\|^2 - \|-u-v\|^2)\\ &= \frac{1}{4}(\|u-v\|^2 - \|u+v\|^2)\\ &= -\langle u, v\rangle,\end{aligned}$$

the property holds for $\lambda = -1$. By linearity and by induction, for any $n \in \mathbb{N}$ with $n \geq 1$, writing $n = n - 1 + 1$, we get

$$\langle \lambda x, y\rangle = \lambda\langle x, y\rangle \quad \text{for all } \lambda \in \mathbb{N},$$

and since the above also holds for $\lambda = -1$, it holds for all $\lambda \in \mathbb{Z}$. For $\lambda = p/q$ with $p, q \in \mathbb{Z}$ and $q \neq 0$, we have

$$q\langle (p/q)u, v\rangle = \langle pu, v\rangle = p\langle u, v\rangle,$$

which shows that

$$\langle (p/q)u, v\rangle = (p/q)\langle u, v\rangle,$$

and thus

$$\langle \lambda x, y\rangle = \lambda\langle x, y\rangle \quad \text{for all } \lambda \in \mathbb{Q}.$$

To finish the proof, we use the fact that a norm is a continuous map $x \mapsto \|x\|$. Then, the continuous function $t \mapsto \frac{1}{t}\langle tu, v\rangle$ defined on $\mathbb{R} - \{0\}$ agrees with $\langle u, v\rangle$ on $\mathbb{Q} - \{0\}$, so it is equal to $\langle u, v\rangle$ on $\mathbb{R} - \{0\}$. The case $\lambda = 0$ is trivial, so we are done.

We now define orthogonality.

9.2 Orthogonality, Duality, Adjoint of a Linear Map

An inner product on a vector space gives the ability to define the notion of orthogonality. Families of nonnull pairwise orthogonal vectors must be linearly independent. They are called orthogonal families. In a vector space of finite dimension it is always possible to find orthogonal bases. This is very useful theoretically and practically. Indeed, in an orthogonal basis, finding the coordinates of a vector is very cheap: It takes an inner product. Fourier series make crucial use of this fact. When E has finite dimension, we prove that the inner product on E induces a natural isomorphism between E and its dual space E^*. This allows us to define the adjoint of a linear map in an intrinsic fashion (i.e., independently of bases). It is also possible to orthonormalize any basis (certainly when the dimension is finite). We give two proofs, one using duality, the other more constructive using the Gram–Schmidt orthonormalization procedure.

Definition 9.2. Given a Euclidean space E, any two vectors $u, v \in E$ are *orthogonal, or perpendicular*, if $u \cdot v = 0$. Given a family $(u_i)_{i\in I}$ of vectors in E, we say that $(u_i)_{i\in I}$ is *orthogonal* if $u_i \cdot u_j = 0$ for all $i, j \in I$, where $i \neq j$. We say that the family $(u_i)_{i\in I}$ is *orthonormal* if $u_i \cdot u_j = 0$ for all $i, j \in I$, where $i \neq j$, and $\|u_i\| = u_i \cdot u_i = 1$, for all $i \in I$. For any subset F of E, the set

$$F^{\perp} = \{v \in E \mid u \cdot v = 0, \text{ for all } u \in F\},$$

of all vectors orthogonal to all vectors in F, is called the *orthogonal complement of* F.

Since inner products are positive definite, observe that for any vector $u \in E$, we have

$$u \cdot v = 0 \quad \text{for all } v \in E \quad \text{iff} \quad u = 0.$$

It is immediately verified that the orthogonal complement $F^\perp$ of F is a subspace of E.

Example 9.5. Going back to Example 9.3 and to the inner product

$$\langle f, g\rangle = \int_{-\pi}^{\pi} f(t)g(t)dt$$

on the vector space $\mathcal{C}[-\pi, \pi]$, it is easily checked that

$$\langle \sin px, \sin qx\rangle = \begin{cases} \pi & \text{if } p = q,\ p, q \geq 1, \\ 0 & \text{if } p \neq q,\ p, q \geq 1, \end{cases}$$

$$\langle \cos px, \cos qx\rangle = \begin{cases} \pi & \text{if } p = q,\ p, q \geq 1, \\ 0 & \text{if } p \neq q,\ p, q \geq 0, \end{cases}$$

and

$$\langle \sin px, \cos qx\rangle = 0,$$

for all $p \geq 1$ and $q \geq 0$, and of course, $\langle 1, 1\rangle = \int_{-\pi}^{\pi} dx = 2\pi$.

As a consequence, the family $(\sin px)_{p\geq 1} \cup (\cos qx)_{q\geq 0}$ is orthogonal. It is not orthonormal, but becomes so if we divide every trigonometric function by $\sqrt{\pi}$, and 1 by $\sqrt{2\pi}$.

Proposition 9.3. *Given a Euclidean space E, for any family $(u_i)_{i\in I}$ of nonnull vectors in E, if $(u_i)_{i\in I}$ is orthogonal, then it is linearly independent.*

Proof. Assume there is a linear dependence

$$\sum_{j\in J} \lambda_j u_j = 0$$

for some $\lambda_j \in \mathbb{R}$ and some finite subset J of I. By taking the inner product with u_i for any $i \in J$, and using the the bilinearity of the inner product and the fact that $u_i \cdot u_j = 0$ whenever $i \neq j$, we get

$$\begin{aligned} 0 = u_i \cdot 0 &= u_i \cdot \left(\sum_{j\in J} \lambda_j u_j \right) \\ &= \sum_{j\in J} \lambda_j (u_i \cdot u_j) = \lambda_i (u_i \cdot u_i), \end{aligned}$$

so

$$\lambda_i (u_i \cdot u_i) = 0, \qquad \text{for all } i \in J,$$

and since $u_i \neq 0$ and an inner product is positive definite, $u_i \cdot u_i \neq 0$, so we obtain

$$\lambda_i = 0, \qquad \text{for all } i \in J,$$

which shows that the family $(u_i)_{i\in I}$ is linearly independent. □

We leave the following simple result as an exercise.

Proposition 9.4. *Given a Euclidean space E, any two vectors $u, v \in E$ are orthogonal iff*

$$\|u+v\|^2 = \|u\|^2 + \|v\|^2.$$

One of the most useful features of orthonormal bases is that they afford a very simple method for computing the coordinates of a vector over any basis vector. Indeed, assume that $(e_1, \ldots, e_m)$ is an orthonormal basis. For any vector

$$x = x_1 e_1 + \cdots + x_m e_m,$$

if we compute the inner product $x \cdot e_i$, we get

$$x \cdot e_i = x_1 e_1 \cdot e_i + \cdots + x_i e_i \cdot e_i + \cdots + x_m e_m \cdot e_i = x_i,$$

since

$$e_i \cdot e_j = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j \end{cases}$$

is the property characterizing an orthonormal family. Thus,

$$x_i = x \cdot e_i,$$

which means that $x_i e_i = (x \cdot e_i) e_i$ is the orthogonal projection of x onto the subspace generated by the basis vector e_i. If the basis is orthogonal but not necessarily orthonormal, then

$$x_i = \frac{x \cdot e_i}{e_i \cdot e_i} = \frac{x \cdot e_i}{\|e_i\|^2}.$$

All this is true even for an infinite orthonormal (or orthogonal) basis $(e_i)_{i \in I}$.

However, remember that every vector x is expressed as a linear combination

$$x = \sum_{i \in I} x_i e_i$$

where the family of scalars $(x_i)_{i \in I}$ has **finite support**, which means that $x_i = 0$ for all $i \in I - J$, where J is a finite set. Thus, even though the family $(\sin px)_{p \geq 1} \cup (\cos qx)_{q \geq 0}$ is orthogonal (it is not orthonormal, but becomes so if we divide every trigonometric function by $\sqrt{\pi}$, and 1 by $\sqrt{2\pi}$; we won't because it looks messy!), the fact that a function $f \in \mathcal{C}^0[-\pi, \pi]$ can be written as a Fourier series as

$$f(x) = a_0 + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx)$$

does not mean that $(\sin px)_{p\geq 1} \cup (\cos qx)_{q\geq 0}$ is a basis of this vector space of functions, because in general, the families (a_k) and (b_k) **do not** have finite support! In order for this infinite linear combination to make sense, it is necessary to prove that the partial sums

$$a_0 + \sum_{k=1}^{n}(a_k \cos kx + b_k \sin kx)$$

of the series converge to a limit when n goes to infinity. This requires a topology on the space.

A very important property of Euclidean spaces of finite dimension is that the inner product induces a canonical bijection (i.e., independent of the choice of bases) between the vector space E and its dual E^*. The reason is that an inner product $\cdot\colon E \times E \to \mathbb{R}$ defines a nondegenerate pairing, as defined in Definition 8.4. Indeed, if $u \cdot v = 0$ for all $v \in E$ then $u = 0$, and similarly if $u \cdot v = 0$ for all $u \in E$ then $v = 0$ (since an inner product is positive definite and symmetric). By Proposition 8.3, there is a canonical isomorphism between E and E^*. We feel that the reader will appreciate if we exhibit this mapping explicitly and reprove that it is an isomorphism.

The mapping from E to E^* is defined as follows. For any vector $u \in E$, let $\varphi_u\colon E \to \mathbb{R}$ be the map defined such that

$$\varphi_u(v) = u \cdot v, \quad \text{for all } v \in E.$$

Since the inner product is bilinear, the map φ_u is a linear form in E^*. Thus, we have a map $\flat\colon E \to E^*$, defined such that

$$\flat(u) = \varphi_u.$$

Theorem 9.5. *Given a Euclidean space E, the map $\flat\colon E \to E^*$ defined such that*

$$\flat(u) = \varphi_u$$

is linear and injective. When E is also of finite dimension, the map $\flat\colon E \to E^*$ is a canonical isomorphism.

Proof. That $\flat\colon E \to E^*$ is a linear map follows immediately from the fact that the inner product is bilinear. If $\varphi_u = \varphi_v$, then $\varphi_u(w) = \varphi_v(w)$ for all $w \in E$, which by definition of φ_u means that $u \cdot w = v \cdot w$ for all $w \in E$, which by bilinearity is equivalent to

$$(v - u) \cdot w = 0$$

for all $w \in E$, which implies that $u = v$, since the inner product is positive definite. Thus, $\flat\colon E \to E^*$ is injective. Finally, when E is of finite dimension n, we know that E^* is also of dimension n, and then $\flat\colon E \to E^*$ is bijective. □

The inverse of the isomorphism $\flat\colon E \to E^*$ is denoted by $\sharp\colon E^* \to E$.

As a consequence of Theorem 9.5, if E is a Euclidean space of finite dimension, every linear form $f \in E^*$ corresponds to a unique $u \in E$ such that

$$f(v) = u \cdot v,$$

for every $v \in E$. In particular, if f is not the null form, the kernel of f, which is a hyperplane H, is precisely the set of vectors that are orthogonal to u.

Remarks:

(1) The "musical map" $\flat\colon E \to E^*$ is not surjective when E has infinite dimension. The result can be salvaged by restricting our attention to continuous linear maps, and by assuming that the vector space E is a *Hilbert space* (i.e., E is a complete normed vector space w.r.t. the Euclidean norm). This is the famous "little" Riesz theorem (or Riesz representation theorem).

(2) Theorem 9.5 still holds if the inner product on E is replaced by a nondegenerate symmetric bilinear form φ. We say that a symmetric bilinear form $\varphi\colon E \times E \to \mathbb{R}$ is *nondegenerate* if for every $u \in E$,

$$\text{if} \quad \varphi(u, v) = 0 \quad \text{for all } v \in E, \quad \text{then} \quad u = 0.$$

For example, the symmetric bilinear form on $\mathbb{R}^4$ (the Lorentz form) defined such that

$$\varphi((x_1, x_2, x_3, x_4), (y_1, y_2, y_3, y_4)) = x_1y_1 + x_2y_2 + x_3y_3 - x_4y_4$$

is nondegenerate. However, there are nonnull vectors $u \in \mathbb{R}^4$ such that $\varphi(u, u) = 0$, which is impossible in a Euclidean space. Such vectors are called *isotropic*.

Example 9.6. Consider $\mathbb{R}^n$ with its usual Euclidean inner product. Given any differentiable function $f\colon U \to \mathbb{R}$, where U is some open subset of $\mathbb{R}^n$, by definition, for any $x \in U$, the *total derivative* df_x of f at x is the linear form defined so that for all $u = (u_1, \ldots, u_n) \in \mathbb{R}^n$,

$$df_x(u) = \begin{pmatrix} \dfrac{\partial f}{\partial x_1}(x) & \cdots & \dfrac{\partial f}{\partial x_n}(x) \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x)\, u_i.$$

The unique vector $v \in \mathbb{R}^n$ such that

$$v \cdot u = df_x(u) \quad \text{for all } u \in \mathbb{R}^n$$

is the transpose of the *Jacobian matrix* of f at x, the $1 \times n$ matrix

$$\begin{pmatrix} \dfrac{\partial f}{\partial x_1}(x) & \cdots & \dfrac{\partial f}{\partial x_n}(x) \end{pmatrix}.$$

This is the *gradient* $\mathrm{grad}(f)_x$ of f at x, given by

$$\mathrm{grad}(f)_x = \begin{pmatrix} \dfrac{\partial f}{\partial x_1}(x) \\ \vdots \\ \dfrac{\partial f}{\partial x_n}(x) \end{pmatrix}.$$

Example 9.7. Given any two vectors $u, v \in \mathbb{R}^3$, let $c(u, v)$ be the linear form given by

$$c(u, v)(w) = \det(u, v, w) \quad \text{for all } w \in \mathbb{R}^3.$$

Since

$$\begin{aligned} \det(u, v, w) &= \begin{vmatrix} u_1 & v_1 & w_1 \\ u_2 & v_2 & w_2 \\ u_3 & v_3 & w_3 \end{vmatrix} = w_1 \begin{vmatrix} u_2 & v_2 \\ u_3 & v_3 \end{vmatrix} - w_2 \begin{vmatrix} u_1 & v_1 \\ u_3 & v_3 \end{vmatrix} + w_3 \begin{vmatrix} u_1 & v_1 \\ u_2 & v_2 \end{vmatrix} \\ &= w_1(u_2v_3 - u_3v_2) + w_2(u_3v_1 - u_1v_3) + w_3(u_1v_2 - u_2v_1), \end{aligned}$$

we see that the unique vector $z \in \mathbb{R}^3$ such that

$$z \cdot w = c(u, v)(w) = \det(u, v, w) \quad \text{for all } w \in \mathbb{R}^3$$

is the vector

$$z = \begin{pmatrix} u_2v_3 - u_3v_2 \\ u_3v_1 - u_1v_3 \\ u_1v_2 - u_2v_1 \end{pmatrix}.$$

This is just the *cross-product* $u \times v$ of u and v. Since $\det(u, v, u) = \det(u, v, v) = 0$, we see that $u \times v$ is orthogonal to both u and v. The above allows us to generalize the cross-product to $\mathbb{R}^n$. Given any $n - 1$ vectors $u_1, \ldots, u_{n-1} \in \mathbb{R}^n$, the cross-product $u_1 \times \cdots \times u_{n-1}$ is the unique vector in $\mathbb{R}^n$ such that

$$(u_1 \times \cdots \times u_{n-1}) \cdot w = \det(u_1, \ldots, u_{n-1}, w) \quad \text{for all } w \in \mathbb{R}^n.$$

Example 9.8. Consider the vector space $\mathrm{M}_n(\mathbb{R})$ of real $n \times n$ matrices with the inner product

$$\langle A, B \rangle = \mathrm{tr}(A^\top B).$$

Let $s \colon \mathrm{M}_n(\mathbb{R}) \to \mathbb{R}$ be the function given by

$$s(A) = \sum_{i,j=1}^{n} a_{ij},$$

where $A = (a_{ij})$. It is immediately verified that s is a linear form. It is easy to check that the unique matrix Z such that

$$\langle Z, A \rangle = s(A) \quad \text{for all } A \in \mathrm{M}_n(\mathbb{R})$$

is the matrix $Z = \mathbf{ones}(n, n)$ whose entries are all equal to 1.

The existence of the isomorphism $\flat\colon E \to E^*$ is crucial to the existence of adjoint maps. The importance of adjoint maps stems from the fact that the linear maps arising in physical problems are often self-adjoint, which means that $f = f^*$. Moreover, self-adjoint maps can be diagonalized over orthonormal bases of eigenvectors. This is the key to the solution of many problems in mechanics, and engineering in general (see Strang [101]).

Let E be a Euclidean space of finite dimension n, and let $f\colon E \to E$ be a linear map. For every $u \in E$, the map

$$v \mapsto u \cdot f(v)$$

is clearly a linear form in E^*, and by Theorem 9.5, there is a unique vector in E denoted by $f^*(u)$ such that

$$f^*(u) \cdot v = u \cdot f(v),$$

for every $v \in E$. The following simple proposition shows that the map f^* is linear.

Proposition 9.6. *Given a Euclidean space E of finite dimension, for every linear map $f\colon E \to E$, there is a unique linear map $f^*\colon E \to E$ such that*

$$f^*(u) \cdot v = u \cdot f(v),$$

for all $u, v \in E$. The map f^* is called the adjoint of f (w.r.t. to the inner product).

Proof. Given $u_1, u_2 \in E$, since the inner product is bilinear, we have

$$(u_1 + u_2) \cdot f(v) = u_1 \cdot f(v) + u_2 \cdot f(v),$$

for all $v \in E$, and

$$(f^*(u_1) + f^*(u_2)) \cdot v = f^*(u_1) \cdot v + f^*(u_2) \cdot v,$$

for all $v \in E$, and since by assumption,

$$f^*(u_1) \cdot v = u_1 \cdot f(v) \quad \text{and} \quad f^*(u_2) \cdot v = u_2 \cdot f(v),$$

for all $v \in E$, we get

$$(f^*(u_1) + f^*(u_2)) \cdot v = (u_1 + u_2) \cdot f(v),$$

for all $v \in E$. Since $\flat$ is bijective, this implies that

$$f^*(u_1 + u_2) = f^*(u_1) + f^*(u_2).$$

Similarly,

$$(\lambda u) \cdot f(v) = \lambda(u \cdot f(v)),$$

for all $v \in E$, and

$$(\lambda f^*(u)) \cdot v = \lambda(f^*(u) \cdot v),$$

for all $v \in E$, and since by assumption,

$$f^*(u) \cdot v = u \cdot f(v),$$

for all $v \in E$, we get

$$(\lambda f^*(u)) \cdot v = \lambda(u \cdot f(v)) = (\lambda u) \cdot f(v)$$

for all $v \in E$. Since $\flat$ is bijective, this implies that

$$f^*(\lambda u) = \lambda f^*(u).$$

Thus, f^* is indeed a linear map, and it is unique, since $\flat$ is a bijection. □

Linear maps $f\colon E \to E$ such that $f = f^*$ are called *self-adjoint* maps. They play a very important role because they have real eigenvalues, and because orthonormal bases arise from their eigenvectors. Furthermore, many physical problems lead to self-adjoint linear maps (in the form of symmetric matrices).

Remark: Proposition 9.6 still holds if the inner product on E is replaced by a nondegenerate symmetric bilinear form φ.

Linear maps such that $f^{-1} = f^*$, or equivalently

$$f^* \circ f = f \circ f^* = \mathrm{id},$$

also play an important role. They are *linear isometries*, or *isometries*. Rotations are special kinds of isometries. Another important class of linear maps are the linear maps satisfying the property

$$f^* \circ f = f \circ f^*,$$

called *normal linear maps*. We will see later on that normal maps can always be diagonalized over orthonormal bases of eigenvectors, but this will require using a Hermitian inner product (over $\mathbb{C}$).

Given two Euclidean spaces E and F, where the inner product on E is denoted by $\langle -, - \rangle_1$ and the inner product on F is denoted by $\langle -, - \rangle_2$, given any linear map $f\colon E \to F$, it is immediately verified that the proof of Proposition 9.6 can be adapted to show that there is a unique linear map $f^*\colon F \to E$ such that

$$\langle f(u), v \rangle_2 = \langle u, f^*(v) \rangle_1$$

for all $u \in E$ and all $v \in F$. The linear map f^* is also called the *adjoint of* f.

The following properties immediately follow from the definition of the adjoint map:

(1) For any linear map $f\colon E \to F$, we have

$$f^{**} = f.$$

(2) For any two linear maps $f, g\colon E \to F$ and any scalar $\lambda \in \mathbb{R}$:

$$(f+g)^* = f^* + g^*$$
$$(\lambda f)^* = \lambda f^*.$$

(3) If E, F, G are Euclidean spaces with respective inner products $\langle -,-\rangle_1, \langle -,-\rangle_2$, and $\langle -,-\rangle_3$, and if $f\colon E \to F$ and $g\colon F \to G$ are two linear maps, then

$$(g \circ f)^* = f^* \circ g^*.$$

Remark: Given any basis for E and any basis for F, it is possible to characterize the matrix of the adjoint f^* of f in terms of the matrix of f, and the symmetric matrices defining the inner products. We will do so with respect to orthonormal bases. Also, since inner products are symmetric, the adjoint f^* of f is also characterized by

$$f(u) \cdot v = u \cdot f^*(v),$$

for all $u, v \in E$.

We can also use Theorem 9.5 to show that any Euclidean space of finite dimension has an orthonormal basis.

Proposition 9.7. *Given any nontrivial Euclidean space E of finite dimension $n \geq 1$, there is an orthonormal basis $(u_1, \ldots, u_n)$ for E.*

Proof. We proceed by induction on n. When $n = 1$, take any nonnull vector $v \in E$, which exists, since we assumed E nontrivial, and let

$$u = \frac{v}{\|v\|}.$$

If $n \geq 2$, again take any nonnull vector $v \in E$, and let

$$u_1 = \frac{v}{\|v\|}.$$

Consider the linear form φ_{u_1} associated with u_1. Since $u_1 \neq 0$, by Theorem 9.5, the linear form φ_{u_1} is nonnull, and its kernel is a hyperplane H. Since $\varphi_{u_1}(w) = 0$ iff $u_1 \cdot w = 0$, the hyperplane H is the orthogonal complement of $\{u_1\}$. Furthermore, since $u_1 \neq 0$ and the inner product is positive definite, $u_1 \cdot u_1 \neq 0$, and thus, $u_1 \notin H$, which implies that $E = H \oplus \mathbb{R}u_1$. However, since E is of finite dimension n, the hyperplane H has dimension $n-1$, and by the induction hypothesis, we can find an orthonormal basis $(u_2, \ldots, u_n)$ for H. Now, because H and the one dimensional space $\mathbb{R}u_1$ are orthogonal and $E = H \oplus \mathbb{R}u_1$, it is clear that $(u_1, \ldots, u_n)$ is an orthonormal basis for E. □

There is a more constructive way of proving Proposition 9.7, using a procedure known as the *Gram–Schmidt orthonormalization procedure*. Among other things, the Gram–Schmidt orthonormalization procedure yields the *QR-decomposition for matrices*, an important tool in numerical methods.

Proposition 9.8. *Given any nontrivial Euclidean space E of finite dimension $n \geq 1$, from any basis $(e_1, \ldots, e_n)$ for E we can construct an orthonormal basis $(u_1, \ldots, u_n)$ for E, with the property that for every k, $1 \leq k \leq n$, the families $(e_1, \ldots, e_k)$ and $(u_1, \ldots, u_k)$ generate the same subspace.*

Proof. We proceed by induction on n. For $n = 1$, let

$$u_1 = \frac{e_1}{\|e_1\|}.$$

For $n \geq 2$, we also let

$$u_1 = \frac{e_1}{\|e_1\|},$$

and assuming that $(u_1, \ldots, u_k)$ is an orthonormal system that generates the same subspace as $(e_1, \ldots, e_k)$, for every k with $1 \leq k < n$, we note that the vector

$$u'_{k+1} = e_{k+1} - \sum_{i=1}^{k} (e_{k+1} \cdot u_i)\, u_i$$

is nonnull, since otherwise, because $(u_1, \ldots, u_k)$ and $(e_1, \ldots, e_k)$ generate the same subspace, $(e_1, \ldots, e_{k+1})$ would be linearly dependent, which is absurd, since $(e_1, \ldots, e_n)$ is a basis. Thus, the norm of the vector u'_{k+1} being nonzero, we use the following construction of the vectors u_k and u'_k:

$$u'_1 = e_1, \qquad u_1 = \frac{u'_1}{\|u'_1\|},$$

and for the inductive step

$$u'_{k+1} = e_{k+1} - \sum_{i=1}^{k} (e_{k+1} \cdot u_i)\, u_i, \qquad u_{k+1} = \frac{u'_{k+1}}{\|u'_{k+1}\|},$$

where $1 \leq k \leq n-1$. It is clear that $\|u_{k+1}\| = 1$, and since $(u_1, \ldots, u_k)$ is an orthonormal system, we have

$$u'_{k+1} \cdot u_i = e_{k+1} \cdot u_i - (e_{k+1} \cdot u_i) u_i \cdot u_i = e_{k+1} \cdot u_i - e_{k+1} \cdot u_i = 0,$$

for all i with $1 \leq i \leq k$. This shows that the family $(u_1, \ldots, u_{k+1})$ is orthonormal, and since $(u_1, \ldots, u_k)$ and $(e_1, \ldots, e_k)$ generates the same subspace, it is clear from the definition of u_{k+1} that $(u_1, \ldots, u_{k+1})$ and $(e_1, \ldots, e_{k+1})$ generate the same subspace. This completes the induction step and the proof of the proposition. □

Note that u'_{k+1} is obtained by subtracting from e_{k+1} the projection of e_{k+1} itself onto the orthonormal vectors $u_1, \ldots, u_k$ that have already been computed. Then, u'_{k+1} is normalized.

Remarks:

(1) The QR-decomposition can now be obtained very easily, but we postpone this until Section 9.4.

(2) We could compute u'_{k+1} using the formula

$$u'_{k+1} = e_{k+1} - \sum_{i=1}^{k} \left(\frac{e_{k+1} \cdot u'_i}{\|u'_i\|^2} \right) u'_i,$$

and normalize the vectors u'_k at the end. This time, we are subtracting from e_{k+1} the projection of e_{k+1} itself onto the orthogonal vectors $u'_1, \ldots, u'_k$. This might be preferable when writing a computer program.

(3) The proof of Proposition 9.8 also works for a countably infinite basis for E, producing a countably infinite orthonormal basis.

Example 9.9. If we consider polynomials and the inner product

$$\langle f, g \rangle = \int_{-1}^{1} f(t)g(t)dt,$$

applying the Gram–Schmidt orthonormalization procedure to the polynomials

$$1, x, x^2, \ldots, x^n, \ldots,$$

which form a basis of the polynomials in one variable with real coefficients, we get a family of orthonormal polynomials $Q_n(x)$ related to the *Legendre polynomials*.

The Legendre polynomials $P_n(x)$ have many nice properties. They are orthogonal, but their norm is not always 1. The Legendre polynomials $P_n(x)$ can be defined as follows. Letting f_n be the function

$$f_n(x) = (x^2 - 1)^n,$$

we define $P_n(x)$ as follows:

$$P_0(x) = 1, \quad \text{and} \quad P_n(x) = \frac{1}{2^n n!} f_n^{(n)}(x),$$

where $f_n^{(n)}$ is the nth derivative of f_n.

They can also be defined inductively as follows:

$$\begin{aligned} P_0(x) &= 1, \\ P_1(x) &= x, \\ P_{n+1}(x) &= \frac{2n+1}{n+1}\, xP_n(x) - \frac{n}{n+1}\, P_{n-1}(x). \end{aligned}$$

Here is an explicit summation fo $P_n(x)$ (thanks to Jocelyn Qaintance for telling me about this formula):

$$P_n(x) = \frac{1}{2^n} \sum_{k=0}^{\lfloor n/2 \rfloor} (-1)^k \binom{n}{k} \binom{2n-2k}{n} x^{n-2k}.$$

The polynomials Q_n are related to the Legendre polynomials P_n as follows:

$$Q_n(x) = \sqrt{\frac{2n+1}{2}} P_n(x).$$

Example 9.10. Consider polynomials over $[-1, 1]$, with the symmetric bilinear form

$$\langle f, g \rangle = \int_{-1}^{1} \frac{1}{\sqrt{1-t^2}} f(t)g(t)dt.$$

We leave it as an exercise to prove that the above defines an inner product. It can be shown that the polynomials $T_n(x)$ given by

$$T_n(x) = \cos(n \arccos x), \quad n \geq 0,$$

(equivalently, with $x = \cos\theta$, we have $T_n(\cos\theta) = \cos(n\theta)$) are orthogonal with respect to the above inner product. These polynomials are the *Chebyshev polynomials*. Their norm is not equal to 1. Instead, we have

$$\langle T_n, T_n \rangle = \begin{cases} \frac{\pi}{2} & \text{if } n > 0, \\ \pi & \text{if } n = 0. \end{cases}$$

Using the identity $(\cos\theta + i\sin\theta)^n = \cos n\theta + i \sin n\theta$ and the binomial formula, we obtain the following expression for $T_n(x)$:

$$T_n(x) = \sum_{k=0}^{\lfloor n/2 \rfloor} \binom{n}{2k} (x^2-1)^k x^{n-2k}.$$

The Chebyshev polynomials are defined inductively as follows:

$$\begin{aligned} T_0(x) &= 1 \\ T_1(x) &= x \\ T_{n+1}(x) &= 2xT_n(x) - T_{n-1}(x), \quad n \geq 1. \end{aligned}$$

Using these recurrence equations, we can show that

$$T_n(x) = \frac{(x - \sqrt{x^2-1})^n + (x + \sqrt{x^2-1})^n}{2}.$$

The polynomial T_n has n distinct roots in the interval $[-1, 1]$. The Chebyshev polynomials play an important role in approximation theory. They are used as an approximation to a best polynomial approximation of a continuous function under the sup-norm (∞-norm).

The inner products of the last two examples are special cases of an inner product of the form

$$\langle f, g\rangle = \int_{-1}^{1} W(t)f(t)g(t)dt,$$

where $W(t)$ is a *weight function*. If W is a nonzero continuous function such that $W(x) \geq 0$ on $(-1, 1)$, then the above bilinear form is indeed positive definite. Families of orthogonal polynomials used in approximation theory and in physics arise by a suitable choice of the weight function W. Besides the previous two examples, the *Hermite polynomials* correspond to $W(x) = e^{-x^2}$, the *Laguerre polynomials* to $W(x) = e^{-x}$, and the *Jacobi polynomials* to $W(x) = (1-x)^\alpha(1+x)^\beta$, with $\alpha, \beta > -1$. Comprehensive treatments of orthogonal polynomials can be found in Lebedev [67], Sansone [84], and Andrews, Askey and Roy [2].

As a consequence of Proposition 9.7 (or Proposition 9.8), given any Euclidean space of finite dimension n, if $(e_1, \ldots, e_n)$ is an orthonormal basis for E, then for any two vectors $u = u_1e_1 + \cdots + u_ne_n$ and $v = v_1e_1 + \cdots + v_ne_n$, the inner product $u \cdot v$ is expressed as

$$u \cdot v = (u_1e_1 + \cdots + u_ne_n) \cdot (v_1e_1 + \cdots + v_ne_n) = \sum_{i=1}^{n} u_iv_i,$$

and the norm $\|u\|$ as

$$\|u\| = \|u_1e_1 + \cdots + u_ne_n\| = \left(\sum_{i=1}^{n} u_i^2\right)^{1/2}.$$

The fact that a Euclidean space always has an orthonormal basis implies that any Gram matrix G can be written as

$$G = Q^\top Q,$$

for some invertible matrix Q. Indeed, we know that in a change of basis matrix, a Gram matrix G becomes $G' = P^\top GP$. If the basis corresponding to G' is orthonormal, then $G' = I$, so $G = (P^{-1})^\top P^{-1}$.

We can also prove the following proposition regarding orthogonal spaces.

Proposition 9.9. *Given any nontrivial Euclidean space E of finite dimension $n \geq 1$, for any subspace F of dimension k, the orthogonal complement $F^\perp$ of F has dimension $n - k$, and $E = F \oplus F^\perp$. Furthermore, we have $F^{\perp\perp} = F$.*

Proof. From Proposition 9.7, the subspace F has some orthonormal basis $(u_1, \ldots, u_k)$. This linearly independent family $(u_1, \ldots, u_k)$ can be extended to a basis $(u_1, \ldots, u_k, v_{k+1}, \ldots, v_n)$, and by Proposition 9.8, it can be converted to an orthonormal basis $(u_1, \ldots, u_n)$, which contains $(u_1, \ldots, u_k)$ as an orthonormal basis of F. Now, any vector $w = w_1u_1 + \cdots + w_nu_n \in E$ is orthogonal to F iff $w \cdot u_i = 0$, for every i, where $1 \leq i \leq k$, iff $w_i = 0$ for every i, where $1 \leq i \leq k$. Clearly, this shows that $(u_{k+1}, \ldots, u_n)$ is a basis of $F^\perp$, and thus $E = F \oplus F^\perp$, and $F^\perp$ has dimension $n - k$. Similarly, any vector $w = w_1u_1 + \cdots + w_nu_n \in E$ is orthogonal to $F^\perp$ iff $w \cdot u_i = 0$, for every i, where $k+1 \leq i \leq n$, iff $w_i = 0$ for every i, where $k+1 \leq i \leq n$. Thus, $(u_1, \ldots, u_k)$ is a basis of $F^{\perp\perp}$, and $F^{\perp\perp} = F$. □

9.3 Linear Isometries (Orthogonal Transformations)

In this section we consider linear maps between Euclidean spaces that preserve the Euclidean norm. These transformations, sometimes called *rigid motions*, play an important role in geometry.

Definition 9.3. Given any two nontrivial Euclidean spaces E and F of the same finite dimension n, a function $f\colon E \to F$ is *an orthogonal transformation, or a linear isometry*, if it is linear and

$$\|f(u)\| = \|u\|, \quad \text{for all } u \in E.$$

Remarks:

(1) A linear isometry is often defined as a linear map such that

$$\|f(v) - f(u)\| = \|v - u\|,$$

for all $u, v \in E$. Since the map f is linear, the two definitions are equivalent. The second definition just focuses on preserving the distance between vectors.

(2) Sometimes, a linear map satisfying the condition of Definition 9.3 is called a *metric map*, and a linear isometry is defined as a *bijective* metric map.

An isometry (without the word linear) is sometimes defined as a function $f\colon E \to F$ (not necessarily linear) such that

$$\|f(v) - f(u)\| = \|v - u\|,$$

for all $u, v \in E$, i.e., as a function that preserves the distance. This requirement turns out to be very strong. Indeed, the next proposition shows that all these definitions are equivalent when E and F are of finite dimension, and for functions such that $f(0) = 0$.

Proposition 9.10. *Given any two nontrivial Euclidean spaces E and F of the same finite dimension n, for every function $f\colon E \to F$, the following properties are equivalent:*

(1) f is a linear map and $\|f(u)\| = \|u\|$, for all $u \in E$;

(2) $\|f(v) - f(u)\| = \|v - u\|$, *for all* $u, v \in E$, *and* $f(0) = 0$;

(3) $f(u) \cdot f(v) = u \cdot v$, *for all* $u, v \in E$.

Furthermore, such a map is bijective.

Proof. Clearly, (1) implies (2), since in (1) it is assumed that f is linear.

Assume that (2) holds. In fact, we shall prove a slightly stronger result. We prove that if

$$\|f(v) - f(u)\| = \|v - u\|$$

for all $u, v \in E$, then for any vector $\tau \in E$, the function $g\colon E \to F$ defined such that

$$g(u) = f(\tau + u) - f(\tau)$$

for all $u \in E$ is a linear map such that $g(0) = 0$ and (3) holds. Clearly, $g(0) = f(\tau) - f(\tau) = 0$.

Note that from the hypothesis

$$\|f(v) - f(u)\| = \|v - u\|$$

for all $u, v \in E$, we conclude that

$$\begin{aligned} \|g(v) - g(u)\| &= \|f(\tau + v) - f(\tau) - (f(\tau + u) - f(\tau))\|, \\ &= \|f(\tau + v) - f(\tau + u)\|, \\ &= \|\tau + v - (\tau + u)\|, \\ &= \|v - u\|, \end{aligned}$$

for all $u, v \in E$. Since $g(0) = 0$, by setting $u = 0$ in

$$\|g(v) - g(u)\| = \|v - u\|,$$

we get

$$\|g(v)\| = \|v\|$$

for all $v \in E$. In other words, g preserves both the distance and the norm.

To prove that g preserves the inner product, we use the simple fact that

$$2u \cdot v = \|u\|^2 + \|v\|^2 - \|u - v\|^2$$

for all $u, v \in E$. Then, since g preserves distance and norm, we have

$$\begin{aligned} 2g(u) \cdot g(v) &= \|g(u)\|^2 + \|g(v)\|^2 - \|g(u) - g(v)\|^2 \\ &= \|u\|^2 + \|v\|^2 - \|u - v\|^2 \\ &= 2u \cdot v, \end{aligned}$$

and thus $g(u) \cdot g(v) = u \cdot v$, for all $u, v \in E$, which is (3). In particular, if $f(0) = 0$, by letting $\tau = 0$, we have $g = f$, and f preserves the scalar product, i.e., (3) holds.

Now assume that (3) holds. Since E is of finite dimension, we can pick an orthonormal basis $(e_1, \ldots, e_n)$ for E. Since f preserves inner products, $(f(e_1), \ldots, f(e_n))$ is also orthonormal, and since F also has dimension n, it is a basis of F. Then note that for any $u = u_1 e_1 + \cdots + u_n e_n$, we have

$$u_i = u \cdot e_i,$$

for all i, $1 \leq i \leq n$. Thus, we have

$$f(u) = \sum_{i=1}^{n} (f(u) \cdot f(e_i)) f(e_i),$$

and since f preserves inner products, this shows that

$$f(u) = \sum_{i=1}^{n} (u \cdot e_i) f(e_i) = \sum_{i=1}^{n} u_i f(e_i),$$

which shows that f is linear. Obviously, f preserves the Euclidean norm, and (3) implies (1).

Finally, if $f(u) = f(v)$, then by linearity $f(v - u) = 0$, so that $\|f(v - u)\| = 0$, and since f preserves norms, we must have $\|v - u\| = 0$, and thus $u = v$. Thus, f is injective, and since E and F have the same finite dimension, f is bijective. □

Remarks:

(i) The dimension assumption is needed only to prove that (3) implies (1) when f is not known to be linear, and to prove that f is surjective, but the proof shows that (1) implies that f is injective.

(ii) The implication that (3) implies (1) holds if we also assume that f is surjective, even if E has infinite dimension.

In (2), when f does not satisfy the condition $f(0) = 0$, the proof shows that f is an affine map. Indeed, taking any vector τ as an origin, the map g is linear, and

$$f(\tau + u) = f(\tau) + g(u) \quad \text{for all } u \in E.$$

By Proposition 3.13, this shows that f is affine with associated linear map g.

This fact is worth recording as the following proposition.

Proposition 9.11. *Given any two nontrivial Euclidean spaces E and F of the same finite dimension n, for every function $f\colon E \to F$, if*

$$\|f(v) - f(u)\| = \|v - u\| \quad \textit{for all } u, v \in E,$$

then f is an affine map, and its associated linear map g is an isometry.

In view of Proposition 9.10, we usually abbreviate "linear isometry" as "isometry," unless we wish to emphasize that we are dealing with a map between vector spaces.

We are now going to take a closer look at the isometries $f\colon E \to E$ of a Euclidean space of finite dimension.

9.4 The Orthogonal Group, Orthogonal Matrices

In this section we explore some of the basic properties of the orthogonal group and of orthogonal matrices.

Proposition 9.12. *Let E be any Euclidean space of finite dimension n, and let $f\colon E \to E$ be any linear map. The following properties hold:*

(1) The linear map $f\colon E \to E$ is an isometry iff

$$f \circ f^* = f^* \circ f = \mathrm{id}.$$

(2) For every orthonormal basis $(e_1, \ldots, e_n)$ of E, if the matrix of f is A, then the matrix of f^* is the transpose $A^\top$ of A, and f is an isometry iff A satisfies the identities

$$A\,A^\top = A^\top A = I_n,$$

where I_n denotes the identity matrix of order n, iff the columns of A form an orthonormal basis of $\mathbb{R}^n$, iff the rows of A form an orthonormal basis of $\mathbb{R}^n$.

Proof. (1) The linear map $f\colon E \to E$ is an isometry iff

$$f(u) \cdot f(v) = u \cdot v,$$

for all $u, v \in E$, iff

$$f^*(f(u)) \cdot v = f(u) \cdot f(v) = u \cdot v$$

for all $u, v \in E$, which implies

$$(f^*(f(u)) - u) \cdot v = 0$$

for all $u, v \in E$. Since the inner product is positive definite, we must have

$$f^*(f(u)) - u = 0$$

for all $u \in E$, that is,

$$f^* \circ f = f \circ f^* = \mathrm{id}.$$

The converse is established by doing the above steps backward.

(2) If $(e_1, \ldots, e_n)$ is an orthonormal basis for E, let $A = (a_{i\,j})$ be the matrix of f, and let $B = (b_{i\,j})$ be the matrix of f^*. Since f^* is characterized by

$$f^*(u) \cdot v = u \cdot f(v)$$

for all $u, v \in E$, using the fact that if $w = w_1 e_1 + \cdots + w_n e_n$ we have $w_k = w \cdot e_k$ for all k, $1 \leq k \leq n$, letting $u = e_i$ and $v = e_j$, we get

$$b_{j\,i} = f^*(e_i) \cdot e_j = e_i \cdot f(e_j) = a_{i\,j},$$

for all i, j, $1 \leq i, j \leq n$. Thus, $B = A^\top$. Now, if X and Y are arbitrary matrices over the basis $(e_1, \ldots, e_n)$, denoting as usual the jth column of X by X^j, and similarly for Y, a simple calculation shows that

$$X^\top Y = (X^i \cdot Y^j)_{1 \leq i,j \leq n}.$$

Then it is immediately verified that if $X = Y = A$, then

$$A^\top A = A\, A^\top = I_n$$

iff the column vectors $(A^1, \ldots, A^n)$ form an orthonormal basis. Thus, from (1), we see that (2) is clear (also because the rows of A are the columns of $A^\top$). □

Proposition 9.12 shows that the inverse of an isometry f is its adjoint f^*. Recall that the set of all real $n \times n$ matrices is denoted by $\mathrm{M}_n(\mathbb{R})$. Proposition 9.12 also motivates the following definition.

Definition 9.4. A real $n \times n$ matrix is an *orthogonal matrix* if

$$A\, A^\top = A^\top A = I_n.$$

Remark: It is easy to show that the conditions $A\, A^\top = I_n$, $A^\top A = I_n$, and $A^{-1} = A^\top$, are equivalent. Given any two orthonormal bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$, if P is the change of basis matrix from $(u_1, \ldots, u_n)$ to $(v_1, \ldots, v_n)$, since the columns of P are the coordinates of the vectors v_j with respect to the basis $(u_1, \ldots, u_n)$, and since $(v_1, \ldots, v_n)$ is orthonormal, the columns of P are orthonormal, and by Proposition 9.12 (2), the matrix P is orthogonal.

The proof of Proposition 9.10 (3) also shows that if f is an isometry, then the image of an orthonormal basis $(u_1, \ldots, u_n)$ is an orthonormal basis. Students often ask why ortho*gonal* matrices are not called ortho*normal* matrices, since their columns (and rows) are orthonormal bases! I have no good answer, but isometries do preserve orthogonality, and orthogonal matrices correspond to isometries.

Recall that the determinant $\det(f)$ of a linear map $f\colon E \to E$ is independent of the choice of a basis in E. Also, for every matrix $A \in \mathrm{M}_n(\mathbb{R})$, we have $\det(A) = \det(A^\top)$, and for any two $n \times n$ matrices A and B, we have $\det(AB) = \det(A)\det(B)$. Then, if f is an isometry, and A is its matrix with respect to any orthonormal basis, $A\,A^\top = A^\top A = I_n$ implies that $\det(A)^2 = 1$, that is, either $\det(A) = 1$, or $\det(A) = -1$. It is also clear that the isometries of a Euclidean space of dimension n form a group, and that the isometries of determinant $+1$ form a subgroup. This leads to the following definition.

Definition 9.5. Given a Euclidean space E of dimension n, the set of isometries $f\colon E \to E$ forms a subgroup of $\mathbf{GL}(E)$ denoted by $\mathbf{O}(E)$, or $\mathbf{O}(n)$ when $E = \mathbb{R}^n$, called the *orthogonal group (of E)*. For every isometry f, we have $\det(f) = \pm 1$, where $\det(f)$ denotes the determinant of f. The isometries such that $\det(f) = 1$ are called *rotations, or proper isometries, or proper orthogonal transformations*, and they form a subgroup of the special linear group $\mathbf{SL}(E)$ (and of $\mathbf{O}(E)$), denoted by $\mathbf{SO}(E)$, or $\mathbf{SO}(n)$ when $E = \mathbb{R}^n$, called the *special orthogonal group (of E)*. The isometries such that $\det(f) = -1$ are called *improper isometries, or improper orthogonal transformations, or flip transformations*.

As an immediate corollary of the Gram–Schmidt orthonormalization procedure, we obtain the QR-decomposition for invertible matrices.

9.5 QR-Decomposition for Invertible Matrices

Now that we have the definition of an orthogonal matrix, we can explain how the Gram–Schmidt orthonormalization procedure immediately yields the QR-decomposition for matrices.

Proposition 9.13. *Given any real $n \times n$ matrix A, if A is invertible, then there is an orthogonal matrix Q and an upper triangular matrix R with positive diagonal entries such that $A = QR$.*

Proof. We can view the columns of A as vectors $A^1, \ldots, A^n$ in $\mathbb{E}^n$. If A is invertible, then they are linearly independent, and we can apply Proposition 9.8 to produce an orthonormal basis using the Gram–Schmidt orthonormalization procedure. Recall that we construct vectors Q^k and Q'^k as follows:

$$Q'^1 = A^1, \qquad Q^1 = \frac{Q'^1}{\|Q'^1\|},$$

and for the inductive step

$$Q'^{k+1} = A^{k+1} - \sum_{i=1}^{k} (A^{k+1} \cdot Q^i)\, Q^i, \qquad Q^{k+1} = \frac{Q'^{k+1}}{\|Q'^{k+1}\|},$$

where $1 \leq k \leq n-1$. If we express the vectors A^k in terms of the Q^i and Q'^i, we get the triangular system

$$\begin{aligned}
A^1 &= \|Q'^1\| Q^1, \\
&\vdots \\
A^j &= (A^j \cdot Q^1)\, Q^1 + \cdots + (A^j \cdot Q^i)\, Q^i + \cdots + \|Q'^j\| Q^j, \\
&\vdots \\
A^n &= (A^n \cdot Q^1)\, Q^1 + \cdots + (A^n \cdot Q^{n-1})\, Q^{n-1} + \|Q'^n\| Q^n.
\end{aligned}$$

Letting $r_{k\,k} = \|Q'^k\|$, and $r_{i\,j} = A^j \cdot Q^i$ (the reversal of i and j on the right-hand side *is* intentional!), where $1 \leq k \leq n$, $2 \leq j \leq n$, and $1 \leq i \leq j-1$, and letting $q_{i\,j}$ be the ith component of Q^j, we note that $a_{i\,j}$, the ith component of A^j, is given by

$$a_{i\,j} = r_{1\,j} q_{i\,1} + \cdots + r_{i\,j} q_{i\,i} + \cdots + r_{j\,j} q_{i\,j} = q_{i\,1} r_{1\,j} + \cdots + q_{i\,i} r_{i\,j} + \cdots + q_{i\,j} r_{j\,j}.$$

If we let $Q = (q_{i\,j})$, the matrix whose columns are the components of the Q^j, and $R = (r_{i\,j})$, the above equations show that $A = QR$, where R is upper triangular. The diagonal entries $r_{k\,k} = \|Q'^k\| = A^k \cdot Q^k$ are indeed positive. □

The reader should try the above procedure on some concrete examples for 2×2 and 3×3 matrices.

Remarks:

(1) Because the diagonal entries of R are positive, it can be shown that Q and R are unique.

(2) The QR-decomposition holds even when A is not invertible. In this case, R has some zero on the diagonal. However, a different proof is needed. We will give a nice proof using Householder matrices (see Proposition 10.3, and also Strang [101, 102], Golub and Van Loan [49], Trefethen and Bau [105], Demmel [33], Kincaid and Cheney [59], or Ciarlet [30]).

Example 9.11. Consider the matrix

$$A = \begin{pmatrix} 0 & 0 & 5 \\ 0 & 4 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

We leave as an exercise to show that $A = QR$, with

$$Q = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \qquad \text{and} \qquad R = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 4 & 1 \\ 0 & 0 & 5 \end{pmatrix}.$$

Example 9.12. Another example of QR-decomposition is

$$A = \begin{pmatrix} 1 & 1 & 2 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & 1 \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \end{pmatrix} \begin{pmatrix} \sqrt{2} & 1/\sqrt{2} & \sqrt{2} \\ 0 & 1/\sqrt{2} & \sqrt{2} \\ 0 & 0 & 1 \end{pmatrix}.$$

The QR-decomposition yields a rather efficient and numerically stable method for solving systems of linear equations. Indeed, given a system $Ax = b$, where A is an $n \times n$ invertible matrix, writing $A = QR$, since Q is orthogonal, we get

$$Rx = Q^\top b,$$

and since R is upper triangular, we can solve it by Gaussian elimination, by solving for the last variable x_n first, substituting its value into the system, then solving for x_{n-1}, etc. The QR-decomposition is also very useful in solving least squares problems (we will come back to this later on), and for finding eigenvalues. It can be easily adapted to the case where A is a rectangular $m \times n$ matrix with independent columns (thus, $n \leq m$). In this case, Q is not quite orthogonal. It is an $m \times n$ matrix whose columns are orthogonal, and R is an invertible $n \times n$ upper triangular matrix with positive diagonal entries. For more on QR, see Strang [101, 102], Golub and Van Loan [49], Demmel [33], Trefethen and Bau [105], or Serre [95].

It should also be said that the Gram–Schmidt orthonormalization procedure that we have presented is not very stable numerically, and instead, one should use the *modified Gram–Schmidt method*. To compute Q'^{k+1}, instead of projecting A^{k+1} onto $Q^1, \ldots, Q^k$ in a single step, it is better to perform k projections. We compute $Q_1^{k+1}, Q_2^{k+1}, \ldots, Q_k^{k+1}$ as follows:

$$\begin{aligned} Q_1^{k+1} &= A^{k+1} - (A^{k+1} \cdot Q^1)\, Q^1, \\ Q_{i+1}^{k+1} &= Q_i^{k+1} - (Q_i^{k+1} \cdot Q^{i+1})\, Q^{i+1}, \end{aligned}$$

where $1 \leq i \leq k-1$. It is easily shown that $Q'^{k+1} = Q_k^{k+1}$. The reader is urged to code this method.

A somewhat surprising consequence of the QR-decomposition is a famous determinantal inequality due to Hadamard.

Proposition 9.14. *(Hadamard) For any real $n \times n$ matrix $A = (a_{ij})$, we have*

$$|\det(A)| \leq \prod_{i=1}^{n} \Big(\sum_{j=1}^{n} a_{ij}^2 \Big)^{1/2} \quad \textit{and} \quad |\det(A)| \leq \prod_{j=1}^{n} \Big(\sum_{i=1}^{n} a_{ij}^2 \Big)^{1/2}.$$

Moreover, equality holds iff either A has a zero column in the left inequality or a zero row in the right inequality, or A is orthogonal.

Proof. If $\det(A) = 0$, then the inequality is trivial. In addition, if the righthand side is also 0, then either some column or some row is zero. If $\det(A) \neq 0$, then we can factor A as $A = QR$, with Q is orthogonal and $R = (r_{ij})$ upper triangular with positive diagonal entries. Then, since Q is orthogonal $\det(Q) = \pm 1$, so

$$|\det(A)| = |\det(Q)|\,|\det(R)| = \prod_{j=1} r_{jj}.$$

Now, as Q is orthogonal, it preserves the Euclidean norm, so

$$\sum_{i=1}^{n} a_{ij}^2 = \left\|A^j\right\|_2^2 = \left\|QR^j\right\|_2^2 = \left\|R^j\right\|_2^2 = \sum_{i=1}^{n} r_{ij}^2 \geq r_{jj}^2,$$

which implies that

$$|\det(A)| = \prod_{j=1}^{n} r_{jj} \leq \prod_{j=1}^{n} \left\|R^j\right\|_2 = \prod_{j=1}^{n} \Big(\sum_{i=1}^{n} a_{ij}^2\Big)^{1/2}.$$

The other inequality is obtained by replacing A by $A^\top$. Finally, if $\det(A) \neq 0$ and equality holds, then we must have

$$r_{jj} = \left\|A^j\right\|_2, \quad 1 \leq j \leq n,$$

which can only occur is R is orthogonal. □

Another version of Hadamard's inequality applies to symmetric positive semidefinite matrices.

Proposition 9.15. *(Hadamard) For any real $n \times n$ matrix $A = (a_{ij})$, if A is symmetric positive semidefinite, then we have*

$$\det(A) \leq \prod_{i=1}^{n} a_{ii}.$$

Moreover, if A is positive definite, then equality holds iff A is a diagonal matrix.

Proof. If $\det(A) = 0$, the inequality is trivial. Otherwise, A is positive definite, and by Theorem 5.10 (the Cholesky Factorization), there is a unique upper triangular matrix B with positive diagonal entries such that

$$A = B^\top B.$$

Thus, $\det(A) = \det(B^\top B) = \det(B^\top)\det(B) = \det(B)^2$. If we apply the Hadamard inequality (Proposition 9.15) to B, we obtain

$$\det(B) \leq \prod_{j=1}^{n} \Big(\sum_{i=1}^{n} b_{ij}^2\Big)^{1/2}. \qquad (*)$$

However, the diagonal entries a_{jj} of $A = B^\top B$ are precisely the square norms $\|B^j\|_2^2 = \sum_{i=1}^n b_{ij}^2$, so by squaring $(*)$, we obtain

$$\det(A) = \det(B)^2 \leq \prod_{j=1}^n \Big(\sum_{i=1}^n b_{ij}^2 \Big) = \prod_{j=1}^n a_{jj}.$$

If $\det(A) \neq 0$ and equality holds, then B must be orthogonal, which implies that B is a diagonal matrix, and so is A. □

We derived the second Hadamard inequality (Proposition 9.15) from the first (Proposition 9.14). We leave it as an exercise to prove that the first Hadamard inequality can be deduced from the second Hadamard inequality.

9.6 Some Applications of Euclidean Geometry

Euclidean geometry has applications in computational geometry, in particular Voronoi diagrams and Delaunay triangulations. In turn, Voronoi diagrams have applications in motion planning (see O'Rourke [78]).

Euclidean geometry also has applications to matrix analysis. Recall that a real $n \times n$ matrix A is *symmetric* if it is equal to its transpose $A^\top$. One of the most important properties of symmetric matrices is that they have real eigenvalues and that they can be diagonalized by an orthogonal matrix (see Chapter 13). This means that for every symmetric matrix A, there is a diagonal matrix D and an orthogonal matrix P such that

$$A = PDP^\top.$$

Even though it is not always possible to diagonalize an arbitrary matrix, there are various decompositions involving orthogonal matrices that are of great practical interest. For example, for every real matrix A, there is the *QR-decomposition*, which says that a real matrix A can be expressed as

$$A = QR,$$

where Q is orthogonal and R is an upper triangular matrix. This can be obtained from the Gram–Schmidt orthonormalization procedure, as we saw in Section 9.5, or better, using Householder matrices, as shown in Section 10.2. There is also the *polar decomposition*, which says that a real matrix A can be expressed as

$$A = QS,$$

where Q is orthogonal and S is symmetric positive semidefinite (which means that the eigenvalues of S are nonnegative). Such a decomposition is important in continuum mechanics and in robotics, since it separates stretching from rotation. Finally, there is the wonderful

singular value decomposition, abbreviated as SVD, which says that a real matrix A can be expressed as

$$A = VDU^\top,$$

where U and V are orthogonal and D is a diagonal matrix with nonnegative entries (see Chapter 15). This decomposition leads to the notion of *pseudo-inverse*, which has many applications in engineering (least squares solutions, etc). For an excellent presentation of all these notions, we highly recommend Strang [102, 101], Golub and Van Loan [49], Demmel [33], Serre [95], and Trefethen and Bau [105].

The method of least squares, invented by Gauss and Legendre around 1800, is another great application of Euclidean geometry. Roughly speaking, the method is used to solve inconsistent linear systems $Ax = b$, where the number of equations is greater than the number of variables. Since this is generally impossible, the method of least squares consists in finding a solution x minimizing the Euclidean norm $\|Ax - b\|^2$, that is, the sum of the squares of the "errors." It turns out that there is always a unique solution x^+ of smallest norm minimizing $\|Ax - b\|^2$, and that it is a solution of the square system

$$A^\top Ax = A^\top b,$$

called the system of *normal equations*. The solution x^+ can be found either by using the QR-decomposition in terms of Householder transformations, or by using the notion of pseudo-inverse of a matrix. The pseudo-inverse can be computed using the SVD decomposition. Least squares methods are used extensively in computer vision More details on the method of least squares and pseudo-inverses can be found in Chapter 16.

9.7 Summary

The main concepts and results of this chapter are listed below:

- Bilinear forms; *positive definite* bilinear forms.
- *inner products*, *scalar products*, *Euclidean spaces*.
- *quadratic form* associated with a bilinear form.
- The Euclidean space $\mathbb{E}^n$.
- The *polar form* of a quadratic form.
- *Gram matrix* associated with an inner product.
- The *Cauchy–Schwarz inequality*; the *Minkowski inequality*.
- The *parallelogram law*.

- *Orthogonality*, *orthogonal complement* $F^{\perp}$; *orthonormal family*.
- The *musical isomorphisms* $\flat\colon E \to E^*$ and $\sharp\colon E^* \to E$ (when E is finite-dimensional); Theorem 9.5.
- The *adjoint* of a linear map (with respect to an inner product).
- Existence of an orthonormal basis in a finite-dimensional Euclidean space (Proposition 9.7).
- The *Gram–Schmidt orthonormalization procedure* (Proposition 9.8).
- The *Legendre* and the *Chebyshev* polynomials.
- *Linear isometries* (*orthogonal transformations*, *rigid motions*).
- The *orthogonal group*, *orthogonal matrices*.
- The matrix representing the adjoint f^* of a linear map f is the transpose of the matrix representing f.
- The *orthogonal group* $\mathbf{O}(n)$ and the *special orthogonal group* $\mathbf{SO}(n)$.
- *QR-decomposition* for invertible matrices.
- The *Hadamard inequality* for arbitrary real matrices.
- The *Hadamard inequality* for symmetric positive semidefinite matrices.

Chapter 10

QR-Decomposition for Arbitrary Matrices

10.1 Orthogonal Reflections

Hyperplane reflections are represented by matrices called Householder matrices. These matrices play an important role in numerical methods, for instance for solving systems of linear equations, solving least squares problems, for computing eigenvalues, and for transforming a symmetric matrix into a tridiagonal matrix. We prove a simple geometric lemma that immediately yields the QR-decomposition of arbitrary matrices in terms of Householder matrices.

Orthogonal symmetries are a very important example of isometries. First let us review the definition of projections. Given a vector space E, let F and G be subspaces of E that form a direct sum $E = F \oplus G$. Since every $u \in E$ can be written uniquely as $u = v + w$, where $v \in F$ and $w \in G$, we can define the two *projections* $p_F \colon E \to F$ and $p_G \colon E \to G$ such that $p_F(u) = v$ and $p_G(u) = w$. It is immediately verified that p_G and p_F are linear maps, and that $p_F^2 = p_F$, $p_G^2 = p_G$, $p_F \circ p_G = p_G \circ p_F = 0$, and $p_F + p_G = \mathrm{id}$.

Definition 10.1. Given a vector space E, for any two subspaces F and G that form a direct sum $E = F \oplus G$, the *symmetry (or reflection) with respect to F and parallel to G* is the linear map $s \colon E \to E$ defined such that

$$s(u) = 2p_F(u) - u,$$

for every $u \in E$.

Because $p_F + p_G = \mathrm{id}$, note that we also have

$$s(u) = p_F(u) - p_G(u)$$

and

$$s(u) = u - 2p_G(u),$$

$s^2 = \mathrm{id}$, s is the identity on F, and $s = -\mathrm{id}$ on G. We now assume that E is a Euclidean space of finite dimension.

Definition 10.2. Let E be a Euclidean space of finite dimension n. For any two subspaces F and G, if F and G form a direct sum $E = F \oplus G$ and F and G are orthogonal, i.e., $F = G^{\perp}$, the *orthogonal symmetry (or reflection) with respect to F and parallel to G* is the linear map $s\colon E \to E$ defined such that

$$s(u) = 2p_F(u) - u,$$

for every $u \in E$. When F is a hyperplane, we call s a *hyperplane symmetry with respect to F (or reflection about F)*, and when G is a plane (and thus $\dim(F) = n - 2$), we call s a *flip about F*.

A reflection about a hyperplane F is shown in Figure 10.1.

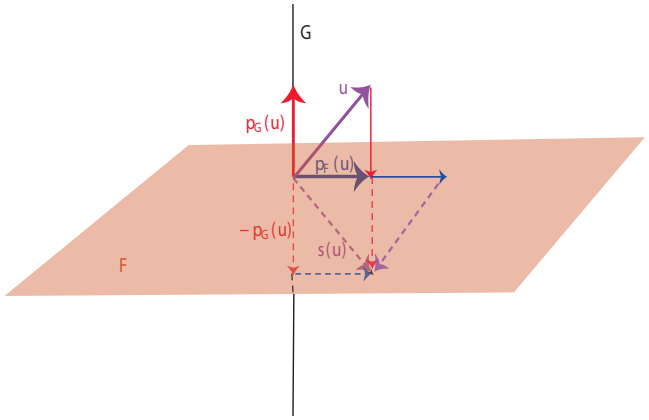


Figure 10.1: A reflection about the peach hyperplane F. Note that u is purple, $p_F(u)$ is blue and $p_G(u)$ is red.

For any two vectors $u, v \in E$, it is easily verified using the bilinearity of the inner product that

$$\|u + v\|^2 - \|u - v\|^2 = 4(u \cdot v).$$

Then, since

$$u = p_F(u) + p_G(u)$$

and

$$s(u) = p_F(u) - p_G(u),$$

since F and G are orthogonal, it follows that

$$p_F(u) \cdot p_G(v) = 0,$$

and thus,

$$\|s(u)\| = \|u\|,$$

so that s is an isometry.

Using Proposition 9.8, it is possible to find an orthonormal basis $(e_1, \ldots, e_n)$ of E consisting of an orthonormal basis of F and an orthonormal basis of G. Assume that F has dimension p, so that G has dimension $n - p$. With respect to the orthonormal basis $(e_1, \ldots, e_n)$, the symmetry s has a matrix of the form

$$\begin{pmatrix} I_p & 0 \\ 0 & -I_{n-p} \end{pmatrix}.$$

Thus, $\det(s) = (-1)^{n-p}$, and s is a rotation iff $n - p$ is even. In particular, when F is a hyperplane H, we have $p = n - 1$ and $n - p = 1$, so that s is an improper orthogonal transformation. When $F = \{0\}$, we have $s = -\mathrm{id}$, which is called the *symmetry with respect to the origin*. The symmetry with respect to the origin is a rotation iff n is even, and an improper orthogonal transformation iff n is odd. When n is odd, we observe that every improper orthogonal transformation is the composition of a rotation with the symmetry with respect to the origin. When G is a plane, $p = n - 2$, and $\det(s) = (-1)^2 = 1$, so that a flip about F is a rotation. In particular, when $n = 3$, F is a line, and a flip about the line F is indeed a rotation of measure π.

Remark: Given any two orthogonal subspaces F, G forming a direct sum $E = F \oplus G$, let f be the symmetry with respect to F and parallel to G, and let g be the symmetry with respect to G and parallel to F. We leave as an exercise to show that

$$f \circ g = g \circ f = -\mathrm{id}.$$

When $F = H$ is a hyperplane, we can give an explicit formula for $s(u)$ in terms of any nonnull vector w orthogonal to H. Indeed, from

$$u = p_H(u) + p_G(u),$$

since $p_G(u) \in G$ and G is spanned by w, which is orthogonal to H, we have

$$p_G(u) = \lambda w$$

for some $\lambda \in \mathbb{R}$, and we get

$$u \cdot w = \lambda \|w\|^2,$$

and thus

$$p_G(u) = \frac{(u \cdot w)}{\|w\|^2}\, w.$$

Since

$$s(u) = u - 2p_G(u),$$

we get

$$s(u) = u - 2\,\frac{(u \cdot w)}{\|w\|^2}\,w.$$

Such reflections are represented by matrices called *Householder matrices*, and they play an important role in numerical matrix analysis (see Kincaid and Cheney [59] or Ciarlet [30]). Householder matrices are symmetric and orthogonal. It is easily checked that over an orthonormal basis $(e_1, \ldots, e_n)$, a hyperplane reflection about a hyperplane H orthogonal to a nonnull vector w is represented by the matrix

$$H = I_n - 2\,\frac{WW^\top}{\|W\|^2} = I_n - 2\,\frac{WW^\top}{W^\top W},$$

where W is the column vector of the coordinates of w over the basis $(e_1, \ldots, e_n)$, and I_n is the identity $n \times n$ matrix. Since

$$p_G(u) = \frac{(u \cdot w)}{\|w\|^2}\,w,$$

the matrix representing p_G is

$$\frac{WW^\top}{W^\top W},$$

and since $p_H + p_G = \mathrm{id}$, the matrix representing p_H is

$$I_n - \frac{WW^\top}{W^\top W}.$$

These formulae can be used to derive a formula for a rotation of $\mathbb{R}^3$, given the direction w of its axis of rotation and given the angle θ of rotation.

The following fact is the key to the proof that every isometry can be decomposed as a product of reflections.

Proposition 10.1. *Let E be any nontrivial Euclidean space. For any two vectors $u, v \in E$, if $\|u\| = \|v\|$, then there is a hyperplane H such that the reflection s about H maps u to v, and if $u \neq v$, then this reflection is unique.*

Proof. If $u = v$, then any hyperplane containing u does the job. Otherwise, we must have $H = \{v - u\}^\perp$, and by the above formula,

$$s(u) = u - 2\,\frac{(u \cdot (v-u))}{\|(v-u)\|^2}\,(v-u) = u + \frac{2\|u\|^2 - 2u \cdot v}{\|(v-u)\|^2}\,(v-u),$$

and since

$$\|(v-u)\|^2 = \|u\|^2 + \|v\|^2 - 2u \cdot v$$

and $\|u\| = \|v\|$, we have

$$\|(v-u)\|^2 = 2\|u\|^2 - 2u \cdot v,$$

and thus, $s(u) = v$. □

If E is a complex vector space and the inner product is Hermitian, Proposition 10.1 is false. The problem is that the vector $v-u$ does not work unless the inner product $u \cdot v$ is real! The proposition can be salvaged enough to yield the QR-decomposition in terms of Householder transformations; see Gallier [44].

We now show that hyperplane reflections can be used to obtain another proof of the QR-decomposition.

10.2 QR-Decomposition Using Householder Matrices

First, we state the result geometrically. When translated in terms of Householder matrices, we obtain the fact advertised earlier that every matrix (not necessarily invertible) has a QR-decomposition.

Proposition 10.2. *Let E be a nontrivial Euclidean space of dimension n. For any orthonormal basis $(e_1, \ldots, e_n)$ and for any n-tuple of vectors $(v_1, \ldots, v_n)$, there is a sequence of n isometries $h_1, \ldots, h_n$ such that h_i is a hyperplane reflection or the identity, and if $(r_1, \ldots, r_n)$ are the vectors given by*

$$r_j = h_n \circ \cdots \circ h_2 \circ h_1(v_j),$$

then every r_j is a linear combination of the vectors $(e_1, \ldots, e_j)$, $1 \leq j \leq n$. Equivalently, the matrix R whose columns are the components of the r_j over the basis $(e_1, \ldots, e_n)$ is an upper triangular matrix. Furthermore, the h_i can be chosen so that the diagonal entries of R are nonnegative.

Proof. We proceed by induction on n. For $n = 1$, we have $v_1 = \lambda e_1$ for some $\lambda \in \mathbb{R}$. If $\lambda \geq 0$, we let $h_1 = \mathrm{id}$, else if $\lambda < 0$, we let $h_1 = -\mathrm{id}$, the reflection about the origin.

For $n \geq 2$, we first have to find h_1. Let

$$r_{1,1} = \|v_1\|.$$

If $v_1 = r_{1,1}e_1$, we let $h_1 = \mathrm{id}$. Otherwise, there is a unique hyperplane reflection h_1 such that

$$h_1(v_1) = r_{1,1}\, e_1,$$

defined such that

$$h_1(u) = u - 2\,\frac{(u \cdot w_1)}{\|w_1\|^2}\, w_1$$

for all $u \in E$, where

$$w_1 = r_{1,1}\, e_1 - v_1.$$

The map h_1 is the reflection about the hyperplane H_1 orthogonal to the vector $w_1 = r_{1,1}\, e_1 - v_1$. Letting

$$r_1 = h_1(v_1) = r_{1,1}\, e_1,$$

it is obvious that r_1 belongs to the subspace spanned by e_1, and $r_{1,1} = \|v_1\|$ is nonnegative.

Next, assume that we have found k linear maps $h_1, \ldots, h_k$, hyperplane reflections or the identity, where $1 \leq k \leq n-1$, such that if $(r_1, \ldots, r_k)$ are the vectors given by

$$r_j = h_k \circ \cdots \circ h_2 \circ h_1(v_j),$$

then every r_j is a linear combination of the vectors $(e_1, \ldots, e_j)$, $1 \leq j \leq k$. The vectors $(e_1, \ldots, e_k)$ form a basis for the subspace denoted by U'_k, the vectors $(e_{k+1}, \ldots, e_n)$ form a basis for the subspace denoted by U''_k, the subspaces U'_k and U''_k are orthogonal, and $E = U'_k \oplus U''_k$. Let

$$u_{k+1} = h_k \circ \cdots \circ h_2 \circ h_1(v_{k+1}).$$

We can write

$$u_{k+1} = u'_{k+1} + u''_{k+1},$$

where $u'_{k+1} \in U'_k$ and $u''_{k+1} \in U''_k$. Let

$$r_{k+1,k+1} = \|u''_{k+1}\|.$$

If $u''_{k+1} = r_{k+1,k+1}\, e_{k+1}$, we let $h_{k+1} = \mathrm{id}$. Otherwise, there is a unique hyperplane reflection h_{k+1} such that

$$h_{k+1}(u''_{k+1}) = r_{k+1,k+1}\, e_{k+1},$$

defined such that

$$h_{k+1}(u) = u - 2\,\frac{(u \cdot w_{k+1})}{\|w_{k+1}\|^2}\, w_{k+1}$$

for all $u \in E$, where

$$w_{k+1} = r_{k+1,k+1}\, e_{k+1} - u''_{k+1}.$$

The map h_{k+1} is the reflection about the hyperplane H_{k+1} orthogonal to the vector $w_{k+1} = r_{k+1,k+1}\, e_{k+1} - u''_{k+1}$. However, since $u''_{k+1}, e_{k+1} \in U''_k$ and U'_k is orthogonal to U''_k, the subspace U'_k is contained in H_{k+1}, and thus, the vectors $(r_1, \ldots, r_k)$ and u'_{k+1}, which belong to U'_k, are invariant under h_{k+1}. This proves that

$$h_{k+1}(u_{k+1}) = h_{k+1}(u'_{k+1}) + h_{k+1}(u''_{k+1}) = u'_{k+1} + r_{k+1,k+1}\, e_{k+1}$$

is a linear combination of $(e_1, \ldots, e_{k+1})$. Letting

$$r_{k+1} = h_{k+1}(u_{k+1}) = u'_{k+1} + r_{k+1,k+1}\, e_{k+1},$$

since $u_{k+1} = h_k \circ \cdots \circ h_2 \circ h_1(v_{k+1})$, the vector

$$r_{k+1} = h_{k+1} \circ \cdots \circ h_2 \circ h_1(v_{k+1})$$

is a linear combination of $(e_1, \ldots, e_{k+1})$. The coefficient of r_{k+1} over e_{k+1} is $r_{k+1,k+1} = \|u''_{k+1}\|$, which is nonnegative. This concludes the induction step, and thus the proof. □

Remarks:

(1) Since every h_i is a hyperplane reflection or the identity,

$$\rho = h_n \circ \cdots \circ h_2 \circ h_1$$

is an isometry.

(2) If we allow negative diagonal entries in R, the last isometry h_n may be omitted.

(3) Instead of picking $r_{k,k} = \|u''_k\|$, which means that

$$w_k = r_{k,k}\, e_k - u''_k,$$

where $1 \leq k \leq n$, it might be preferable to pick $r_{k,k} = -\|u''_k\|$ if this makes $\|w_k\|^2$ larger, in which case

$$w_k = r_{k,k}\, e_k + u''_k.$$

Indeed, since the definition of h_k involves division by $\|w_k\|^2$, it is desirable to avoid division by very small numbers.

(4) The method also applies to any m-tuple of vectors $(v_1, \ldots, v_m)$, where m is not necessarily equal to n (the dimension of E). In this case, R is an upper triangular $n \times m$ matrix we leave the minor adjustments to the method as an exercise to the reader (if $m > n$, the last $m - n$ vectors are unchanged).

Proposition 10.2 directly yields the QR-decomposition in terms of Householder transformations (see Strang [101, 102], Golub and Van Loan [49], Trefethen and Bau [105], Kincaid and Cheney [59], or Ciarlet [30]).

Theorem 10.3. *For every real $n \times n$ matrix A, there is a sequence $H_1, \ldots, H_n$ of matrices, where each H_i is either a Householder matrix or the identity, and an upper triangular matrix R such that*

$$R = H_n \cdots H_2 H_1 A.$$

As a corollary, there is a pair of matrices Q, R, where Q is orthogonal and R is upper triangular, such that $A = QR$ (a QR-decomposition of A). Furthermore, R can be chosen so that its diagonal entries are nonnegative.

Proof. The jth column of A can be viewed as a vector v_j over the canonical basis $(e_1, \ldots, e_n)$ of $\mathbb{E}^n$ (where $(e_j)_i = 1$ if $i = j$, and 0 otherwise, $1 \leq i, j \leq n$). Applying Proposition 10.2 to $(v_1, \ldots, v_n)$, there is a sequence of n isometries $h_1, \ldots, h_n$ such that h_i is a hyperplane reflection or the identity, and if $(r_1, \ldots, r_n)$ are the vectors given by

$$r_j = h_n \circ \cdots \circ h_2 \circ h_1(v_j),$$

then every r_j is a linear combination of the vectors $(e_1, \ldots, e_j)$, $1 \leq j \leq n$. Letting R be the matrix whose columns are the vectors r_j, and H_i the matrix associated with h_i, it is clear that

$$R = H_n \cdots H_2 H_1 A,$$

where R is upper triangular and every H_i is either a Householder matrix or the identity. However, $h_i \circ h_i = \mathrm{id}$ for all i, $1 \leq i \leq n$, and so

$$v_j = h_1 \circ h_2 \circ \cdots \circ h_n(r_j)$$

for all j, $1 \leq j \leq n$. But $\rho = h_1 \circ h_2 \circ \cdots \circ h_n$ is an isometry represented by the orthogonal matrix $Q = H_1 H_2 \cdots H_n$. It is clear that $A = QR$, where R is upper triangular. As we noted in Proposition 10.2, the diagonal entries of R can be chosen to be nonnegative. □

Remarks:

(1) Letting

$$A_{k+1} = H_k \cdots H_2 H_1 A,$$

with $A_1 = A$, $1 \leq k \leq n$, the proof of Proposition 10.2 can be interpreted in terms of the computation of the sequence of matrices $A_1, \ldots, A_{n+1} = R$. The matrix A_{k+1} has the shape

$$A_{k+1} = \begin{pmatrix} \times & \times & \times & u_1^{k+1} & \times & \times & \times & \times \\ 0 & \times & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \times & u_k^{k+1} & \times & \times & \times & \times \\ 0 & 0 & 0 & u_{k+1}^{k+1} & \times & \times & \times & \times \\ 0 & 0 & 0 & u_{k+2}^{k+1} & \times & \times & \times & \times \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & u_{n-1}^{k+1} & \times & \times & \times & \times \\ 0 & 0 & 0 & u_n^{k+1} & \times & \times & \times & \times \end{pmatrix},$$

where the $(k+1)$th column of the matrix is the vector

$$u_{k+1} = h_k \circ \cdots \circ h_2 \circ h_1(v_{k+1}),$$

and thus

$$u'_{k+1} = \left(u_1^{k+1}, \ldots, u_k^{k+1}\right)$$

and

$$u''_{k+1} = \left(u^{k+1}_{k+1}, u^{k+1}_{k+2}, \ldots, u^{k+1}_{n}\right).$$

If the last $n-k-1$ entries in column $k+1$ are all zero, there is nothing to do, and we let $H_{k+1} = I$. Otherwise, we kill these $n-k-1$ entries by multiplying A_{k+1} on the left by the Householder matrix H_{k+1} sending

$$\left(0, \ldots, 0, u^{k+1}_{k+1}, \ldots, u^{k+1}_{n}\right) \quad \text{to} \quad (0, \ldots, 0, r_{k+1,k+1}, 0, \ldots, 0),$$

where $r_{k+1,k+1} = \|(u^{k+1}_{k+1}, \ldots, u^{k+1}_{n})\|$.

(2) If A is invertible and the diagonal entries of R are positive, it can be shown that Q and R are unique.

(3) If we allow negative diagonal entries in R, the matrix H_n may be omitted ($H_n = I$).

(4) The method allows the computation of the determinant of A. We have

$$\det(A) = (-1)^m r_{1,1} \cdots r_{n,n},$$

where m is the number of Householder matrices (not the identity) among the H_i.

(5) The "condition number" of the matrix A is preserved (see Strang [102], Golub and Van Loan [49], Trefethen and Bau [105], Kincaid and Cheney [59], or Ciarlet [30]). This is very good for numerical stability.

(6) The method also applies to a rectangular $m \times n$ matrix. In this case, R is also an $m \times n$ matrix (and it is upper triangular).

10.3 Summary

The main concepts and results of this chapter are listed below:

- *Symmetry (or reflection) with respect to F and parallel to G.*
- *Orthogonal symmetry (or reflection) with respect to F and parallel to G; reflections, flips.*
- Hyperplane reflections and *Householder matrices*.
- A key fact about reflections (Proposition 10.1).
- *QR-decomposition in terms of Householder transformations* (Theorem 10.3).

Chapter 11

Hermitian Spaces

11.1 Sesquilinear and Hermitian Forms, Pre-Hilbert Spaces and Hermitian Spaces

In this chapter we generalize the basic results of Euclidean geometry presented in Chapter 9 to vector spaces over the complex numbers. Such a generalization is inevitable, and not simply a luxury. For example, linear maps may not have real eigenvalues, but they always have complex eigenvalues. Furthermore, some very important classes of linear maps can be diagonalized if they are extended to the complexification of a real vector space. This is the case for orthogonal matrices, and, more generally, normal matrices. Also, complex vector spaces are often the natural framework in physics or engineering, and they are more convenient for dealing with Fourier series. However, some complications arise due to complex conjugation.

Recall that for any complex number $z \in \mathbb{C}$, if $z = x + iy$ where $x, y \in \mathbb{R}$, we let $\Re z = x$, the real part of z, and $\Im z = y$, the imaginary part of z. We also denote the conjugate of $z = x + iy$ by $\overline{z} = x - iy$, and the absolute value (or length, or modulus) of z by $|z|$. Recall that $|z|^2 = z\overline{z} = x^2 + y^2$.

There are many natural situations where a map $\varphi\colon E \times E \to \mathbb{C}$ is linear in its first argument and only semilinear in its second argument, which means that $\varphi(u, \mu v) = \overline{\mu}\varphi(u, v)$, as opposed to $\varphi(u, \mu v) = \mu\varphi(u, v)$. For example, the natural inner product to deal with functions $f\colon \mathbb{R} \to \mathbb{C}$, especially Fourier series, is

$$\langle f, g\rangle = \int_{-\pi}^{\pi} f(x)\overline{g(x)}dx,$$

which is semilinear (but not linear) in g. Thus, when generalizing a result from the real case of a Euclidean space to the complex case, we always have to check very carefully that our proofs do not rely on linearity in the second argument. Otherwise, we need to revise our proofs, and sometimes the result is simply wrong!

Before defining the natural generalization of an inner product, it is convenient to define semilinear maps.

Definition 11.1. Given two vector spaces E and F over the complex field $\mathbb{C}$, a function $f\colon E \to F$ is *semilinear* if

$$\begin{aligned} f(u+v) &= f(u) + f(v), \\ f(\lambda u) &= \overline{\lambda} f(u), \end{aligned}$$

for all $u, v \in E$ and all $\lambda \in \mathbb{C}$.

Remark: Instead of defining semilinear maps, we could have defined the vector space $\overline{E}$ as the vector space with the same carrier set E whose addition is the same as that of E, but whose multiplication by a complex number is given by

$$(\lambda, u) \mapsto \overline{\lambda} u.$$

Then it is easy to check that a function $f\colon E \to \mathbb{C}$ is semilinear iff $f\colon \overline{E} \to \mathbb{C}$ is linear.

We can now define sesquilinear forms and Hermitian forms.

Definition 11.2. Given a complex vector space E, a function $\varphi\colon E \times E \to \mathbb{C}$ is a *sesquilinear form* if it is linear in its first argument and semilinear in its second argument, which means that

$$\begin{aligned} \varphi(u_1 + u_2, v) &= \varphi(u_1, v) + \varphi(u_2, v), \\ \varphi(u, v_1 + v_2) &= \varphi(u, v_1) + \varphi(u, v_2), \\ \varphi(\lambda u, v) &= \lambda \varphi(u, v), \\ \varphi(u, \mu v) &= \overline{\mu} \varphi(u, v), \end{aligned}$$

for all $u, v,\ u_1, u_2,\ v_1, v_2 \in E$, and all $\lambda, \mu \in \mathbb{C}$. A function $\varphi\colon E \times E \to \mathbb{C}$ is a *Hermitian form* if it is sesquilinear and if

$$\varphi(v, u) = \overline{\varphi(u, v)}$$

for all all $u, v \in E$.

Obviously, $\varphi(0, v) = \varphi(u, 0) = 0$. Also note that if $\varphi\colon E \times E \to \mathbb{C}$ is sesquilinear, we have

$$\varphi(\lambda u + \mu v, \lambda u + \mu v) = |\lambda|^2 \varphi(u, u) + \lambda \overline{\mu} \varphi(u, v) + \overline{\lambda} \mu \varphi(v, u) + |\mu|^2 \varphi(v, v),$$

and if $\varphi\colon E \times E \to \mathbb{C}$ is Hermitian, we have

$$\varphi(\lambda u + \mu v, \lambda u + \mu v) = |\lambda|^2 \varphi(u, u) + 2\Re(\lambda \overline{\mu} \varphi(u, v)) + |\mu|^2 \varphi(v, v).$$

Note that restricted to real coefficients, a sesquilinear form is bilinear (we sometimes say $\mathbb{R}$-bilinear). The function $\Phi\colon E \to \mathbb{C}$ defined such that $\Phi(u) = \varphi(u,u)$ for all $u \in E$ is called the *quadratic form* associated with φ.

The standard example of a Hermitian form on $\mathbb{C}^n$ is the map φ defined such that

$$\varphi((x_1,\ldots,x_n),(y_1,\ldots,y_n)) = x_1\overline{y_1} + x_2\overline{y_2} + \cdots + x_n\overline{y_n}.$$

This map is also positive definite, but before dealing with these issues, we show the following useful proposition.

Proposition 11.1. *Given a complex vector space E, the following properties hold:*

(1) A sesquilinear form $\varphi\colon E \times E \to \mathbb{C}$ is a Hermitian form iff $\varphi(u,u) \in \mathbb{R}$ for all $u \in E$.

(2) If $\varphi\colon E \times E \to \mathbb{C}$ is a sesquilinear form, then

$$\begin{aligned} 4\varphi(u,v) = {} & \varphi(u+v,u+v) - \varphi(u-v,u-v) \\ & + i\varphi(u+iv,u+iv) - i\varphi(u-iv,u-iv), \end{aligned}$$

and

$$2\varphi(u,v) = (1+i)(\varphi(u,u)+\varphi(v,v)) - \varphi(u-v,u-v) - i\varphi(u-iv,u-iv).$$

These are called polarization identities.

Proof. (1) If φ is a Hermitian form, then

$$\varphi(v,u) = \overline{\varphi(u,v)}$$

implies that

$$\varphi(u,u) = \overline{\varphi(u,u)},$$

and thus $\varphi(u,u) \in \mathbb{R}$. If φ is sesquilinear and $\varphi(u,u) \in \mathbb{R}$ for all $u \in E$, then

$$\varphi(u+v,u+v) = \varphi(u,u) + \varphi(u,v) + \varphi(v,u) + \varphi(v,v),$$

which proves that

$$\varphi(u,v) + \varphi(v,u) = \alpha,$$

where α is real, and changing u to iu, we have

$$i(\varphi(u,v) - \varphi(v,u)) = \beta,$$

where β is real, and thus

$$\varphi(u,v) = \frac{\alpha - i\beta}{2} \quad \text{and} \quad \varphi(v,u) = \frac{\alpha + i\beta}{2},$$

proving that φ is Hermitian.

(2) These identities are verified by expanding the right-hand side, and we leave them as an exercise. □

Proposition 11.1 shows that a sesquilinear form is completely determined by the quadratic form $\Phi(u) = \varphi(u, u)$, even if φ is not Hermitian. This is false for a real bilinear form, unless it is symmetric. For example, the bilinear form $\varphi\colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ defined such that

$$\varphi((x_1, y_1), (x_2, y_2)) = x_1 y_2 - x_2 y_1$$

is not identically zero, and yet it is null on the diagonal. However, a real symmetric bilinear form is indeed determined by its values on the diagonal, as we saw in Chapter 9.

As in the Euclidean case, Hermitian forms for which $\varphi(u, u) \geq 0$ play an important role.

Definition 11.3. Given a complex vector space E, a Hermitian form $\varphi\colon E \times E \to \mathbb{C}$ is *positive* if $\varphi(u, u) \geq 0$ for all $u \in E$, and *positive definite* if $\varphi(u, u) > 0$ for all $u \neq 0$. A pair $\langle E, \varphi \rangle$ where E is a complex vector space and φ is a Hermitian form on E is called a *pre-Hilbert space* if φ is positive, and a *Hermitian (or unitary) space* if φ is positive definite.

We warn our readers that some authors, such as Lang [64], define a pre-Hilbert space as what we define as a Hermitian space. We prefer following the terminology used in Schwartz [90] and Bourbaki [21]. The quantity $\varphi(u, v)$ is usually called the *Hermitian product* of u and v. We will occasionally call it the inner product of u and v.

Given a pre-Hilbert space $\langle E, \varphi \rangle$, as in the case of a Euclidean space, we also denote $\varphi(u, v)$ by

$$u \cdot v \quad \text{or} \quad \langle u, v \rangle \quad \text{or} \quad (u|v),$$

and $\sqrt{\Phi(u)}$ by $\|u\|$.

Example 11.1. The complex vector space $\mathbb{C}^n$ under the Hermitian form

$$\varphi((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = x_1\overline{y_1} + x_2\overline{y_2} + \cdots + x_n\overline{y_n}$$

is a Hermitian space.

Example 11.2. Let l^2 denote the set of all countably infinite sequences $x = (x_i)_{i \in \mathbb{N}}$ of complex numbers such that $\sum_{i=0}^{\infty} |x_i|^2$ is defined (i.e., the sequence $\sum_{i=0}^{n} |x_i|^2$ converges as $n \to \infty$). It can be shown that the map $\varphi\colon l^2 \times l^2 \to \mathbb{C}$ defined such that

$$\varphi\left((x_i)_{i \in \mathbb{N}}, (y_i)_{i \in \mathbb{N}}\right) = \sum_{i=0}^{\infty} x_i \overline{y_i}$$

is well defined, and l^2 is a Hermitian space under φ. Actually, l^2 is even a Hilbert space.

Example 11.3. Let $\mathcal{C}_{\text{piece}}[a, b]$ be the set of piecewise bounded continuous functions $f\colon [a, b] \to \mathbb{C}$ under the Hermitian form

$$\langle f, g \rangle = \int_a^b f(x)\overline{g(x)}dx.$$

It is easy to check that this Hermitian form is positive, but it is not definite. Thus, under this Hermitian form, $\mathcal{C}_{\text{piece}}[a, b]$ is only a pre-Hilbert space.

Example 11.4. Let $\mathcal{C}[a,b]$ be the set of complex-valued continuous functions $f\colon [a,b] \to \mathbb{C}$ under the Hermitian form

$$\langle f, g\rangle = \int_a^b f(x)\overline{g(x)}dx.$$

It is easy to check that this Hermitian form is positive definite. Thus, $\mathcal{C}[a,b]$ is a Hermitian space.

Example 11.5. Let $E = \mathrm{M}_n(\mathbb{C})$ be the vector space of complex $n \times n$ matrices. If we view a matrix $A \in \mathrm{M}_n(\mathbb{C})$ as a "long" column vector obtained by concatenating together its columns, we can define the Hermitian product of two matrices $A, B \in \mathrm{M}_n(\mathbb{C})$ as

$$\langle A, B\rangle = \sum_{i,j=1}^{n} a_{ij}\bar{b}_{ij},$$

which can be conveniently written as

$$\langle A, B\rangle = \mathrm{tr}(A^\top \overline{B}) = \mathrm{tr}(B^* A).$$

Since this can be viewed as the standard Hermitian product on $\mathbb{C}^{n^2}$, it is a Hermitian product on $\mathrm{M}_n(\mathbb{C})$. The corresponding norm

$$\|A\|_F = \sqrt{\mathrm{tr}(A^* A)}$$

is the Frobenius norm (see Section 6.2).

If E is finite-dimensional and if $\varphi\colon E \times E \to \mathbb{R}$ is a sequilinear form on E, given any basis $(e_1, \ldots, e_n)$ of E, we can write $x = \sum_{i=1}^n x_i e_i$ and $y = \sum_{j=1}^n y_j e_j$, and we have

$$\varphi(x,y) = \varphi\Big(\sum_{i=1}^n x_i e_i, \sum_{j=1}^n y_j e_j\Big) = \sum_{i,j=1}^n x_i \overline{y}_j \varphi(e_i, e_j).$$

If we let $G = (g_{ij})$ be the matrix given by $g_{ij} = \varphi(e_j, e_i)$, and if x and y are the column vectors associated with $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$, then we can write

$$\varphi(x,y) = x^\top G^\top \overline{y} = y^* G x,$$

where $\overline{y}$ corresponds to $(\overline{y}_1, \ldots, \overline{y}_n)$. As in Section 9.1, we are committing the slight abuse of notation of letting x denote both the vector $x = \sum_{i=1}^n x_i e_i$ and the column vector associated with $(x_1, \ldots, x_n)$ (and similarly for y). The "correct" expression for $\varphi(x,y)$ is

$$\varphi(x,y) = \mathbf{y}^* G \mathbf{x} = \mathbf{x}^\top G^\top \overline{\mathbf{y}}.$$

Observe that in $\varphi(x,y) = y^* G x$, the matrix involved is the transpose of the matrix $(\varphi(e_i, e_j))$. The reason for this is that we want G to be positive definite when φ is positive definite, not $G^\top$.

Furthermore, observe that φ is Hermitian iff $G = G^*$, and φ is positive definite iff the matrix G is positive definite, that is,

$$(Gx)^\top \overline{x} = x^* G x > 0 \quad \text{for all } x \in \mathbb{C}^n,\ x \neq 0.$$

The matrix G associated with a Hermitian product is called the *Gram matrix* of the Hermitian product with respect to the basis $(e_1, \ldots, e_n)$.

Conversely, if A is a Hermitian positive definite $n \times n$ matrix, it is easy to check that the Hermitian form

$$\langle x, y \rangle = y^* A x$$

is positive definite. If we make a change of basis from the basis $(e_1, \ldots, e_n)$ to the basis $(f_1, \ldots, f_n)$, and if the change of basis matrix is P (where the jth column of P consists of the coordinates of f_j over the basis $(e_1, \ldots, e_n)$), then with respect to coordinates x' and y' over the basis $(f_1, \ldots, f_n)$, we have

$$y^* G x = (y')^* P^* G P x',$$

so the matrix of our inner product over the basis $(f_1, \ldots, f_n)$ is P^*GP. We summarize these facts in the following proposition.

Proposition 11.2. *Let E be a finite-dimensional vector space, and let $(e_1, \ldots, e_n)$ be a basis of E.*

1. *For any Hermitian inner product $\langle -, - \rangle$ on E, if $G = (g_{ij})$ with $g_{ij} = \langle e_j, e_i \rangle$ is the Gram matrix of the Hermitian product $\langle -, - \rangle$ w.r.t. the basis $(e_1, \ldots, e_n)$, then G is Hermitian positive definite.*

2. *For any change of basis matrix P, the Gram matrix of $\langle -, - \rangle$ with respect to the new basis is P^*GP.*

3. *If A is any $n \times n$ Hermitian positive definite matrix, then*

$$\langle x, y \rangle = y^* A x$$

is a Hermitian product on E.

We will see later that a Hermitian matrix is positive definite iff its eigenvalues are all positive.

The following result reminiscent of the first polarization identity of Proposition 11.1 can be used to prove that two linear maps are identical.

Proposition 11.3. *Given any Hermitian space E with Hermitian product $\langle -, - \rangle$, for any linear map $f \colon E \to E$, if $\langle f(x), x \rangle = 0$ for all $x \in E$, then $f = 0$.*

Proof. Compute $\langle f(x+y), x+y\rangle$ and $\langle f(x-y), x-y\rangle$:

$$\langle f(x+y), x+y\rangle = \langle f(x), x\rangle + \langle f(x), y\rangle + \langle f(y), x\rangle + \langle y, y\rangle$$
$$\langle f(x-y), x-y\rangle = \langle f(x), x\rangle - \langle f(x), y\rangle - \langle f(y), x\rangle + \langle y, y\rangle;$$

then, subtract the second equation from the first, to obtain

$$\langle f(x+y), x+y\rangle - \langle f(x-y), x-y\rangle = 2(\langle f(x), y\rangle + \langle f(y), x\rangle).$$

If $\langle f(u), u\rangle = 0$ for all $u \in E$, we get

$$\langle f(x), y\rangle + \langle f(y), x\rangle = 0 \quad \text{for all } x, y \in E.$$

Then, the above equation also holds if we replace x by ix, and we obtain

$$i\langle f(x), y\rangle - i\langle f(y), x\rangle = 0, \quad \text{for all } x, y \in E,$$

so we have

$$\langle f(x), y\rangle + \langle f(y), x\rangle = 0$$
$$\langle f(x), y\rangle - \langle f(y), x\rangle = 0,$$

which implies that $\langle f(x), y\rangle = 0$ for all $x, y \in E$. Since $\langle -, -\rangle$ is positive definite, we have $f(x) = 0$ for all $x \in E$; that is, $f = 0$. □

One should be careful not to apply Proposition 11.3 to a linear map on a real Euclidean space, because it is false! The reader should find a counterexample.

The Cauchy–Schwarz inequality and the Minkowski inequalities extend to pre-Hilbert spaces and to Hermitian spaces.

Proposition 11.4. *Let $\langle E, \varphi\rangle$ be a pre-Hilbert space with associated quadratic form Φ. For all $u, v \in E$, we have the* Cauchy–Schwarz inequality

$$|\varphi(u, v)| \leq \sqrt{\Phi(u)}\sqrt{\Phi(v)}.$$

Furthermore, if $\langle E, \varphi\rangle$ is a Hermitian space, the equality holds iff u and v are linearly dependent.

We also have the Minkowski inequality

$$\sqrt{\Phi(u+v)} \leq \sqrt{\Phi(u)} + \sqrt{\Phi(v)}.$$

Furthermore, if $\langle E, \varphi\rangle$ is a Hermitian space, the equality holds iff u and v are linearly dependent, where in addition, if $u \neq 0$ and $v \neq 0$, then $u = \lambda v$ for some real λ such that $\lambda > 0$.

Proof. For all $u, v \in E$ and all $\mu \in \mathbb{C}$, we have observed that

$$\varphi(u + \mu v, u + \mu v) = \varphi(u, u) + 2\Re(\overline{\mu}\varphi(u, v)) + |\mu|^2\varphi(v, v).$$

Let $\varphi(u, v) = \rho e^{i\theta}$, where $|\varphi(u, v)| = \rho$ ($\rho \geq 0$). Let $F\colon \mathbb{R} \to \mathbb{R}$ be the function defined such that

$$F(t) = \Phi(u + te^{i\theta}v),$$

for all $t \in \mathbb{R}$. The above shows that

$$F(t) = \varphi(u, u) + 2t|\varphi(u, v)| + t^2\varphi(v, v) = \Phi(u) + 2t|\varphi(u, v)| + t^2\Phi(v).$$

Since φ is assumed to be positive, we have $F(t) \geq 0$ for all $t \in \mathbb{R}$. If $\Phi(v) = 0$, we must have $\varphi(u, v) = 0$, since otherwise, $F(t)$ could be made negative by choosing t negative and small enough. If $\Phi(v) > 0$, in order for $F(t)$ to be nonnegative, the equation

$$\Phi(u) + 2t|\varphi(u, v)| + t^2\Phi(v) = 0$$

must not have distinct real roots, which is equivalent to

$$|\varphi(u, v)|^2 \leq \Phi(u)\Phi(v).$$

Taking the square root on both sides yields the Cauchy–Schwarz inequality.

For the second part of the claim, if φ is positive definite, we argue as follows. If u and v are linearly dependent, it is immediately verified that we get an equality. Conversely, if

$$|\varphi(u, v)|^2 = \Phi(u)\Phi(v),$$

then there are two cases. If $\Phi(v) = 0$, since φ is positive definite, we must have $v = 0$, so u and v are linearly dependent. Otherwise, the equation

$$\Phi(u) + 2t|\varphi(u, v)| + t^2\Phi(v) = 0$$

has a double root t_0, and thus

$$\Phi(u + t_0e^{i\theta}v) = 0.$$

Since φ is positive definite, we must have

$$u + t_0e^{i\theta}v = 0,$$

which shows that u and v are linearly dependent.

If we square the Minkowski inequality, we get

$$\Phi(u + v) \leq \Phi(u) + \Phi(v) + 2\sqrt{\Phi(u)}\sqrt{\Phi(v)}.$$

However, we observed earlier that

$$\Phi(u + v) = \Phi(u) + \Phi(v) + 2\Re(\varphi(u, v)).$$

Thus, it is enough to prove that

$$\Re(\varphi(u,v)) \leq \sqrt{\Phi(u)}\sqrt{\Phi(v)},$$

but this follows from the Cauchy–Schwarz inequality

$$|\varphi(u,v)| \leq \sqrt{\Phi(u)}\sqrt{\Phi(v)}$$

and the fact that $\Re z \leq |z|$.

If φ is positive definite and u and v are linearly dependent, it is immediately verified that we get an equality. Conversely, if equality holds in the Minkowski inequality, we must have

$$\Re(\varphi(u,v)) = \sqrt{\Phi(u)}\sqrt{\Phi(v)},$$

which implies that

$$|\varphi(u,v)| = \sqrt{\Phi(u)}\sqrt{\Phi(v)},$$

since otherwise, by the Cauchy–Schwarz inequality, we would have

$$\Re(\varphi(u,v)) \leq |\varphi(u,v)| < \sqrt{\Phi(u)}\sqrt{\Phi(v)}.$$

Thus, equality holds in the Cauchy–Schwarz inequality, and

$$\Re(\varphi(u,v)) = |\varphi(u,v)|.$$

But then, we proved in the Cauchy–Schwarz case that u and v are linearly dependent. Since we also just proved that $\varphi(u,v)$ is real and nonnegative, the coefficient of proportionality between u and v is indeed nonnegative. □

As in the Euclidean case, if $\langle E, \varphi\rangle$ is a Hermitian space, the Minkowski inequality

$$\sqrt{\Phi(u+v)} \leq \sqrt{\Phi(u)} + \sqrt{\Phi(v)}$$

shows that the map $u \mapsto \sqrt{\Phi(u)}$ is a *norm* on E. The norm induced by φ is called the *Hermitian norm induced by* φ. We usually denote $\sqrt{\Phi(u)}$ by $\|u\|$, and the Cauchy–Schwarz inequality is written as

$$|u \cdot v| \leq \|u\|\|v\|.$$

Since a Hermitian space is a normed vector space, it is a topological space under the topology induced by the norm (a basis for this topology is given by the open balls $B_0(u,\rho)$ of center u and radius $\rho > 0$, where

$$B_0(u,\rho) = \{v \in E \mid \|v-u\| < \rho\}.$$

If E has finite dimension, every linear map is continuous; see Chapter 6 (or Lang [64, 65], Dixmier [35], or Schwartz [90, 91]). The Cauchy–Schwarz inequality

$$|u \cdot v| \leq \|u\|\|v\|$$

shows that $\varphi\colon E \times E \to \mathbb{C}$ is continuous, and thus, that $\| \ \|$ is continuous.

If $\langle E, \varphi\rangle$ is only pre-Hilbertian, $\|u\|$ is called a *seminorm*. In this case, the condition

$$\|u\| = 0 \quad \text{implies} \quad u = 0$$

is not necessarily true. However, the Cauchy–Schwarz inequality shows that if $\|u\| = 0$, then $u \cdot v = 0$ for all $v \in E$.

Remark: As in the case of real vector spaces, a norm on a complex vector space is induced by some positive definite Hermitian product $\langle -, -\rangle$ iff it satisfies the *parallelogram law*:

$$\|u + v\|^2 + \|u - v\|^2 = 2(\|u\|^2 + \|v\|^2).$$

This time, the Hermitian product is recovered using the polarization identity from Proposition 11.1:

$$4\langle u, v\rangle = \|u + v\|^2 - \|u - v\|^2 + i\,\|u + iv\|^2 - i\,\|u - iv\|^2\,.$$

It is easy to check that $\langle u, u\rangle = \|u\|^2$, and

$$\begin{aligned}\langle v, u\rangle &= \overline{\langle u, v\rangle}\\ \langle iu, v\rangle &= i\langle u, v\rangle,\end{aligned}$$

so it is enough to check linearity in the variable u, and only for real scalars. This is easily done by applying the proof from Section 9.1 to the real and imaginary part of $\langle u, v\rangle$; the details are left as an exercise.

We will now basically mirror the presentation of Euclidean geometry given in Chapter 9 rather quickly, leaving out most proofs, except when they need to be seriously amended.

11.2 Orthogonality, Duality, Adjoint of a Linear Map

In this section we assume that we are dealing with Hermitian spaces. We denote the Hermitian inner product by $u \cdot v$ or $\langle u, v\rangle$. The concepts of orthogonality, orthogonal family of vectors, orthonormal family of vectors, and orthogonal complement of a set of vectors are unchanged from the Euclidean case (Definition 9.2).

For example, the set $\mathcal{C}[-\pi, \pi]$ of continuous functions $f\colon [-\pi, \pi] \to \mathbb{C}$ is a Hermitian space under the product

$$\langle f, g\rangle = \int_{-\pi}^{\pi} f(x)\overline{g(x)}dx,$$

and the family $(e^{ikx})_{k\in\mathbb{Z}}$ is orthogonal.

Proposition 9.3 and 9.4 hold without any changes. It is easy to show that

$$\left\|\sum_{i=1}^{n} u_i\right\|^2 = \sum_{i=1}^{n} \|u_i\|^2 + \sum_{1\le i<j\le n} 2\Re(u_i \cdot u_j).$$

Analogously to the case of Euclidean spaces of finite dimension, the Hermitian product induces a canonical bijection (i.e., independent of the choice of bases) between the vector space E and the space E^*. This is one of the places where conjugation shows up, but in this case, troubles are minor.

Given a Hermitian space E, for any vector $u \in E$, let $\varphi_u^l \colon E \to \mathbb{C}$ be the map defined such that

$$\varphi_u^l(v) = \overline{u \cdot v}, \quad \text{for all } v \in E.$$

Similarly, for any vector $v \in E$, let $\varphi_v^r \colon E \to \mathbb{C}$ be the map defined such that

$$\varphi_v^r(u) = u \cdot v, \quad \text{for all } u \in E.$$

Since the Hermitian product is linear in its first argument u, the map φ_v^r is a linear form in E^*, and since it is semilinear in its second argument v, the map φ_u^l is also a linear form in E^*. Thus, we have two maps $\flat^l \colon E \to E^*$ and $\flat^r \colon E \to E^*$, defined such that

$$\flat^l(u) = \varphi_u^l, \quad \text{and} \quad \flat^r(v) = \varphi_v^r.$$

Actually, $\varphi_u^l = \varphi_u^r$ and $\flat^l = \flat^r$. Indeed, for all $u, v \in E$, we have

$$\begin{aligned} \flat^l(u)(v) &= \varphi_u^l(v) \\ &= \overline{u \cdot v} \\ &= v \cdot u \\ &= \varphi_u^r(v) \\ &= \flat^r(u)(v). \end{aligned}$$

Therefore, we use the notation φ_u for both φ_u^l and φ_u^r, and $\flat$ for both $\flat^l$ and $\flat^r$.

Theorem 11.5. *let E be a Hermitian space E. The map $\flat \colon E \to E^*$ defined such that*

$$\flat(u) = \varphi_u^l = \varphi_u^r \quad \textit{for all } u \in E$$

is semilinear and injective. When E is also of finite dimension, the map $\flat \colon \overline{E} \to E^*$ is a canonical isomorphism.

Proof. That $\flat \colon E \to E^*$ is a semilinear map follows immediately from the fact that $\flat = \flat^r$, and that the Hermitian product is semilinear in its second argument. If $\varphi_u = \varphi_v$, then $\varphi_u(w) = \varphi_v(w)$ for all $w \in E$, which by definition of φ_u and φ_v means that

$$w \cdot u = w \cdot v$$

for all $w \in E$, which by semilinearity on the right is equivalent to

$$w \cdot (v - u) = 0 \quad \text{for all } w \in E,$$

which implies that $u = v$, since the Hermitian product is positive definite. Thus, $\flat \colon E \to E^*$ is injective. Finally, when E is of finite dimension n, E^* is also of dimension n, and then $\flat \colon E \to E^*$ is bijective. Since $\flat$ is semilinar, the map $\flat \colon \overline{E} \to E^*$ is an isomorphism. □

The inverse of the isomorphism $\flat\colon \overline{E} \to E^*$ is denoted by $\sharp\colon E^* \to \overline{E}$.

As a corollary of the isomorphism $\flat\colon \overline{E} \to E^*$, if E is a Hermitian space of finite dimension, then every linear form $f \in E^*$ corresponds to a unique $v \in E$, such that

$$f(u) = u \cdot v, \quad \text{for every } u \in E.$$

In particular, if f is not the null form, the kernel of f, which is a hyperplane H, is precisely the set of vectors that are orthogonal to v.

Remarks:

1. The "musical map" $\flat\colon \overline{E} \to E^*$ is not surjective when E has infinite dimension. This result can be salvaged by restricting our attention to continuous linear maps, and by assuming that the vector space E is a *Hilbert space*.

2. *Dirac's "bra-ket" notation.* Dirac invented a notation widely used in quantum mechanics for denoting the linear form $\varphi_u = \flat(u)$ associated to the vector $u \in E$ *via* the duality induced by a Hermitian inner product. Dirac's proposal is to denote the vectors u in E by $|u\rangle$, and call them *kets*; the notation $|u\rangle$ is pronounced "ket u." Given two kets (vectors) $|u\rangle$ and $|v\rangle$, their inner product is denoted by

$$\langle u|v\rangle$$

(instead of $|u\rangle \cdot |v\rangle$). The notation $\langle u|v\rangle$ for the inner product of $|u\rangle$ and $|v\rangle$ anticipates duality. Indeed, we define the dual (usually called adjoint) *bra* u of ket u, denoted by $\langle u|$, as the linear form whose value on any ket v is given by the inner product, so

$$\langle u|(|v\rangle) = \langle u|v\rangle.$$

Thus, bra $u = \langle u|$ is Dirac's notation for our $\flat(u)$. Since the map $\flat$ is semi-linear, we have

$$\langle \lambda u| = \overline{\lambda}\langle u|.$$

Using the bra-ket notation, given an orthonormal basis $(|u_1\rangle, \ldots, |u_n\rangle)$, ket v (a vector) is written as

$$|v\rangle = \sum_{i=1}^{n} \langle v|u_i\rangle |u_i\rangle,$$

and the corresponding linear form bra v is written as

$$\langle v| = \sum_{i=1}^{n} \overline{\langle v|u_i\rangle} \langle u_i| = \sum_{i=1}^{n} \langle u_i|v\rangle \langle u_i|$$

over the dual basis $(\langle u_1|, \ldots, \langle u_n|)$. As cute as it looks, we do not recommend using the Dirac notation.

The existence of the isomorphism $\flat\colon \overline{E} \to E^*$ is crucial to the existence of adjoint maps. Indeed, Theorem 11.5 allows us to define the adjoint of a linear map on a Hermitian space. Let E be a Hermitian space of finite dimension n, and let $f\colon E \to E$ be a linear map. For every $u \in E$, the map

$$v \mapsto \overline{u \cdot f(v)}$$

is clearly a linear form in E^*, and by Theorem 11.5, there is a unique vector in E denoted by $f^*(u)$, such that

$$\overline{f^*(u) \cdot v} = \overline{u \cdot f(v)},$$

that is,

$$f^*(u) \cdot v = u \cdot f(v), \quad \text{for every } v \in E.$$

The following proposition shows that the map f^* is linear.

Proposition 11.6. *Given a Hermitian space E of finite dimension, for every linear map $f\colon E \to E$ there is a unique linear map $f^*\colon E \to E$ such that*

$$f^*(u) \cdot v = u \cdot f(v),$$

for all $u, v \in E$. The map f^* is called the adjoint of f (w.r.t. to the Hermitian product).

Proof. Careful inspection of the proof of Proposition 9.6 reveals that it applies unchanged. The only potential problem is in proving that $f^*(\lambda u) = \lambda f^*(u)$, but everything takes place in the first argument of the Hermitian product, and there, we have linearity. □

The fact that

$$v \cdot u = \overline{u \cdot v}$$

implies that the adjoint f^* of f is also characterized by

$$f(u) \cdot v = u \cdot f^*(v),$$

for all $u, v \in E$.

Given two Hermitian spaces E and F, where the Hermitian product on E is denoted by $\langle -, - \rangle_1$ and the Hermitian product on F is denoted by $\langle -, - \rangle_2$, given any linear map $f\colon E \to F$, it is immediately verified that the proof of Proposition 11.6 can be adapted to show that there is a unique linear map $f^*\colon F \to E$ such that

$$\langle f(u), v \rangle_2 = \langle u, f^*(v) \rangle_1$$

for all $u \in E$ and all $v \in F$. The linear map f^* is also called the *adjoint* of f.

As in the euclidean case, the following properties immediately follow from the definition of the adjoint map:

(1) For any linear map $f\colon E \to F$, we have

$$f^{**} = f.$$

(2) For any two linear maps $f, g\colon E \to F$ and any scalar $\lambda \in \mathbb{R}$:

$$(f+g)^* = f^* + g^*$$
$$(\lambda f)^* = \overline{\lambda} f^*.$$

(3) If E, F, G are Hermitian spaces with respective inner products $\langle -,-\rangle_1, \langle -,-\rangle_2$, and $\langle -,-\rangle_3$, and if $f\colon E \to F$ and $g\colon F \to G$ are two linear maps, then

$$(g \circ f)^* = f^* \circ g^*.$$

As in the Euclidean case, a linear map $f\colon E \to E$ (where E is a finite-dimensional Hermitian space) is *self-adjoint* if $f = f^*$. The map f is *positive semidefinite* iff

$$\langle f(x), x\rangle \geq 0 \quad \text{all } x \in E;$$

positive definite iff

$$\langle f(x), x\rangle > 0 \quad \text{all } x \in E,\ x \neq 0.$$

An interesting corollary of Proposition 11.3 is that a positive semidefinite linear map must be self-adjoint. In fact, we can prove a slightly more general result.

Proposition 11.7. *Given any finite-dimensional Hermitian space E with Hermitian product $\langle -,-\rangle$, for any linear map $f\colon E \to E$, if $\langle f(x), x\rangle \in \mathbb{R}$ for all $x \in E$, then f is self-adjoint. In particular, any positive semidefinite linear map $f\colon E \to E$ is self-adjoint.*

Proof. Since $\langle f(x), x\rangle \in \mathbb{R}$ for all $x \in E$, we have

$$\begin{aligned}\langle f(x), x\rangle &= \overline{\langle f(x), x\rangle} \\ &= \langle x, f(x)\rangle \\ &= \langle f^*(x), x\rangle,\end{aligned}$$

so we have

$$\langle (f - f^*)(x), x\rangle = 0 \quad \text{all } x \in E,$$

and Proposition 11.3 implies that $f - f^* = 0$. □

Beware that Proposition 11.7 is false if E is a real Euclidean space.

As in the Euclidean case, Theorem 11.5 can be used to show that any Hermitian space of finite dimension has an orthonormal basis. The proof is unchanged.

Proposition 11.8. *Given any nontrivial Hermitian space E of finite dimension $n \geq 1$, there is an orthonormal basis $(u_1, \ldots, u_n)$ for E.*

The *Gram–Schmidt orthonormalization procedure* also applies to Hermitian spaces of finite dimension, without any changes from the Euclidean case!

Proposition 11.9. *Given a nontrivial Hermitian space E of finite dimension $n \geq 1$, from any basis $(e_1, \ldots, e_n)$ for E we can construct an orthonormal basis $(u_1, \ldots, u_n)$ for E with the property that for every k, $1 \leq k \leq n$, the families $(e_1, \ldots, e_k)$ and $(u_1, \ldots, u_k)$ generate the same subspace.*

Remark: The remarks made after Proposition 9.8 also apply here, except that in the QR-decomposition, Q is a unitary matrix.

As a consequence of Proposition 9.7 (or Proposition 11.9), given any Hermitian space of finite dimension n, if $(e_1, \ldots, e_n)$ is an orthonormal basis for E, then for any two vectors $u = u_1 e_1 + \cdots + u_n e_n$ and $v = v_1 e_1 + \cdots + v_n e_n$, the Hermitian product $u \cdot v$ is expressed as

$$u \cdot v = (u_1 e_1 + \cdots + u_n e_n) \cdot (v_1 e_1 + \cdots + v_n e_n) = \sum_{i=1}^{n} u_i \overline{v_i},$$

and the norm $\|u\|$ as

$$\|u\| = \|u_1 e_1 + \cdots + u_n e_n\| = \left(\sum_{i=1}^{n} |u_i|^2 \right)^{1/2}.$$

The fact that a Hermitian space always has an orthonormal basis implies that any Gram matrix G can be written as

$$G = Q^* Q,$$

for some invertible matrix Q. Indeed, we know that in a change of basis matrix, a Gram matrix G becomes $G' = (\overline{P})^* G \overline{P}$. If the basis corresponding to G' is orthonormal, then $G' = I$, so $G = (\overline{P}^{-1})^* \overline{P}^{-1}$.

Proposition 9.9 also holds unchanged.

Proposition 11.10. *Given any nontrivial Hermitian space E of finite dimension $n \geq 1$, for any subspace F of dimension k, the orthogonal complement $F^\perp$ of F has dimension $n - k$, and $E = F \oplus F^\perp$. Furthermore, we have $F^{\perp\perp} = F$.*

11.3 Linear Isometries (Also Called Unitary Transformations)

In this section we consider linear maps between Hermitian spaces that preserve the Hermitian norm. All definitions given for Euclidean spaces in Section 9.3 extend to Hermitian spaces, except that orthogonal transformations are called unitary transformation, but Proposition 9.10 extends only with a modified condition (2). Indeed, the old proof that (2) implies (3) does not work, and the implication is in fact false! It can be repaired by strengthening condition (2). For the sake of completeness, we state the Hermitian version of Definition 9.3.

Definition 11.4. Given any two nontrivial Hermitian spaces E and F of the same finite dimension n, a function $f: E \to F$ is *a unitary transformation, or a linear isometry*, if it is linear and

$$\|f(u)\| = \|u\|, \quad \text{for all } u \in E.$$

Proposition 9.10 can be salvaged by strengthening condition (2).

Proposition 11.11. *Given any two nontrivial Hermitian spaces E and F of the same finite dimension n, for every function $f: E \to F$, the following properties are equivalent:*

(1) f is a linear map and $\|f(u)\| = \|u\|$, for all $u \in E$;

(2) $\|f(v) - f(u)\| = \|v - u\|$ and $f(iu) = if(u)$, for all $u, v \in E$.

(3) $f(u) \cdot f(v) = u \cdot v$, for all $u, v \in E$.

Furthermore, such a map is bijective.

Proof. The proof that (2) implies (3) given in Proposition 9.10 needs to be revised as follows. We use the polarization identity

$$2\varphi(u, v) = (1 + i)(\|u\|^2 + \|v\|^2) - \|u - v\|^2 - i\|u - iv\|^2.$$

Since $f(iv) = if(v)$, we get $f(0) = 0$ by setting $v = 0$, so the function f preserves distance and norm, and we get

$$\begin{aligned} 2\varphi(f(u), f(v)) &= (1 + i)(\|f(u)\|^2 + \|f(v)\|^2) - \|f(u) - f(v)\|^2 \\ &\quad - i\|f(u) - if(v)\|^2 \\ &= (1 + i)(\|f(u)\|^2 + \|f(v)\|^2) - \|f(u) - f(v)\|^2 \\ &\quad - i\|f(u) - f(iv)\|^2 \\ &= (1 + i)(\|u\|^2 + \|v\|^2) - \|u - v\|^2 - i\|u - iv\|^2 \\ &= 2\varphi(u, v), \end{aligned}$$

which shows that f preserves the Hermitian inner product, as desired. The rest of the proof is unchanged. □

Remarks:

(i) In the Euclidean case, we proved that the assumption

$$\|f(v) - f(u)\| = \|v - u\| \quad \text{for all } u, v \in E \text{ and } f(0) = 0 \tag{2'}$$

implies (3). For this we used the polarization identity

$$2u \cdot v = \|u\|^2 + \|v\|^2 - \|u - v\|^2.$$

In the Hermitian case the polarization identity involves the complex number i. In fact, the implication $(2')$ implies (3) is false in the Hermitian case! Conjugation $z \mapsto \overline{z}$ satisfies $(2')$ since

$$|\overline{z_2} - \overline{z_1}| = |\overline{z_2 - z_1}| = |z_2 - z_1|,$$

and yet, it is not linear!

(ii) If we modify (2) by changing the second condition by now requiring that there be some $\tau \in E$ such that

$$f(\tau + iu) = f(\tau) + i(f(\tau + u) - f(\tau))$$

for all $u \in E$, then the function $g\colon E \to E$ defined such that

$$g(u) = f(\tau + u) - f(\tau)$$

satisfies the old conditions of (2), and the implications (2) $\to$ (3) and (3) $\to$ (1) prove that g is linear, and thus that f is affine. In view of the first remark, some condition involving i is needed on f, in addition to the fact that f is distance-preserving.

11.4 The Unitary Group, Unitary Matrices

In this section, as a mirror image of our treatment of the isometries of a Euclidean space, we explore some of the fundamental properties of the unitary group and of unitary matrices. As an immediate corollary of the Gram–Schmidt orthonormalization procedure, we obtain the QR-decomposition for invertible matrices. In the Hermitian framework, the matrix of the adjoint of a linear map is not given by the transpose of the original matrix, but by its conjugate.

Definition 11.5. Given a complex $m \times n$ matrix A, the *transpose* $A^\top$ *of* A is the $n \times m$ matrix $A^\top = \left(a_{ij}^\top\right)$ defined such that

$$a_{ij}^\top = a_{ji},$$

and the *conjugate* $\overline{A}$ *of* A is the $m \times n$ matrix $\overline{A} = (b_{ij})$ defined such that

$$b_{ij} = \overline{a}_{ij}$$

for all i, j, $1 \leq i \leq m$, $1 \leq j \leq n$. The *adjoint* A^* *of* A is the matrix defined such that

$$A^* = \overline{(A^\top)} = \left(\overline{A}\right)^\top.$$

Proposition 11.12. *Let E be any Hermitian space of finite dimension n, and let $f\colon E \to E$ be any linear map. The following properties hold:*

(1) The linear map $f: E \to E$ is an isometry iff

$$f \circ f^* = f^* \circ f = \mathrm{id}.$$

(2) For every orthonormal basis $(e_1, \ldots, e_n)$ of E, if the matrix of f is A, then the matrix of f^* is the adjoint A^* of A, and f is an isometry iff A satisfies the identities

$$A\,A^* = A^*A = I_n,$$

where I_n denotes the identity matrix of order n, iff the columns of A form an orthonormal basis of $\mathbb{C}^n$, iff the rows of A form an orthonormal basis of $\mathbb{C}^n$.

Proof. (1) The proof is identical to that of Proposition 9.12 (1).

(2) If $(e_1, \ldots, e_n)$ is an orthonormal basis for E, let $A = (a_{i\,j})$ be the matrix of f, and let $B = (b_{i\,j})$ be the matrix of f^*. Since f^* is characterized by

$$f^*(u) \cdot v = u \cdot f(v)$$

for all $u, v \in E$, using the fact that if $w = w_1 e_1 + \cdots + w_n e_n$, we have $w_k = w \cdot e_k$, for all k, $1 \leq k \leq n$; letting $u = e_i$ and $v = e_j$, we get

$$b_{j\,i} = f^*(e_i) \cdot e_j = e_i \cdot f(e_j) = \overline{f(e_j) \cdot e_i} = \overline{a_{i\,j}},$$

for all i, j, $1 \leq i, j \leq n$. Thus, $B = A^*$. Now, if X and Y are arbitrary matrices over the basis $(e_1, \ldots, e_n)$, denoting as usual the jth column of X by X^j, and similarly for Y, a simple calculation shows that

$$Y^*X = (X^j \cdot Y^i)_{1 \leq i,j \leq n}.$$

Then it is immediately verified that if $X = Y = A$, then $A^*A = A\,A^* = I_n$ iff the column vectors $(A^1, \ldots, A^n)$ form an orthonormal basis. Thus, from (1), we see that (2) is clear. □

Proposition 9.12 shows that the inverse of an isometry f is its adjoint f^*. Proposition 9.12 also motivates the following definition.

Definition 11.6. A complex $n \times n$ matrix is a *unitary matrix* if

$$A\,A^* = A^*A = I_n.$$

Remarks:

(1) The conditions $A\,A^* = I_n$, $A^*A = I_n$, and $A^{-1} = A^*$ are equivalent. Given any two orthonormal bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$, if P is the change of basis matrix from $(u_1, \ldots, u_n)$ to $(v_1, \ldots, v_n)$, it is easy to show that the matrix P is unitary. The proof of Proposition 11.11 (3) also shows that if f is an isometry, then the image of an orthonormal basis $(u_1, \ldots, u_n)$ is an orthonormal basis.

(2) Using the explicit formula for the determinant, we see immediately that

$$\det(\overline{A}) = \overline{\det(A)}.$$

If f is a unitary transformation and A is its matrix with respect to any orthonormal basis, from $AA^* = I$, we get

$$\det(AA^*) = \det(A)\det(A^*) = \det(A)\overline{\det(A^\top)} = \det(A)\overline{\det(A)} = |\det(A)|^2,$$

and so $|\det(A)| = 1$. It is clear that the isometries of a Hermitian space of dimension n form a group, and that the isometries of determinant $+1$ form a subgroup.

This leads to the following definition.

Definition 11.7. Given a Hermitian space E of dimension n, the set of isometries $f: E \to E$ forms a subgroup of $\mathbf{GL}(E, \mathbb{C})$ denoted by $\mathbf{U}(E)$, or $\mathbf{U}(n)$ when $E = \mathbb{C}^n$, called the *unitary group (of E)*. For every isometry f we have $|\det(f)| = 1$, where $\det(f)$ denotes the determinant of f. The isometries such that $\det(f) = 1$ are called *rotations, or proper isometries, or proper unitary transformations*, and they form a subgroup of the special linear group $\mathbf{SL}(E, \mathbb{C})$ (and of $\mathbf{U}(E)$), denoted by $\mathbf{SU}(E)$, or $\mathbf{SU}(n)$ when $E = \mathbb{C}^n$, called the *special unitary group (of E)*. The isometries such that $\det(f) \neq 1$ are called *improper isometries, or improper unitary transformations, or flip transformations*.

A very important example of unitary matrices is provided by Fourier matrices (up to a factor of $\sqrt{n}$), matrices that arise in the various versions of the discrete Fourier transform. For more on this topic, see the problems, and Strang [101, 103].

Now that we have the definition of a unitary matrix, we can explain how the Gram–Schmidt orthonormalization procedure immediately yields the QR-decomposition for matrices.

Proposition 11.13. *Given any $n \times n$ complex matrix A, if A is invertible, then there is a unitary matrix Q and an upper triangular matrix R with positive diagonal entries such that $A = QR$.*

The proof is absolutely the same as in the real case!

We have the following version of the Hadamard inequality for complex matrices. The proof is essentially the same as in the Euclidean case but it uses Proposition 11.13 instead of Proposition 9.13.

Proposition 11.14. *(Hadamard) For any complex $n \times n$ matrix $A = (a_{ij})$, we have*

$$|\det(A)| \leq \prod_{i=1}^{n}\left(\sum_{j=1}^{n}|a_{ij}|^2\right)^{1/2} \quad \textit{and} \quad |\det(A)| \leq \prod_{j=1}^{n}\left(\sum_{i=1}^{n}|a_{ij}|^2\right)^{1/2}.$$

Moreover, equality holds iff either A has a zero column in the left inequality or a zero row in the right inequality, or A is unitary.

We also have the following version of Proposition 9.15 for Hermitian matrices. The proof of Proposition 9.15 goes through because the Cholesky decomposition for a Hermitian positive definite A matrix holds in the form $A = B^*B$, where B is upper triangular with positive diagonal entries. The details are left to the reader.

Proposition 11.15. *(Hadamard) For any complex $n \times n$ matrix $A = (a_{ij})$, if A is Hermitian positive semidefinite, then we have*

$$\det(A) \leq \prod_{i=1}^{n} a_{ii}.$$

Moreover, if A is positive definite, then equality holds iff A is a diagonal matrix.

11.5 Orthogonal Projections and Involutions

In this section, we assume that the field K is not a field of characteristic 2. Recall that a linear map $f: E \to E$ is an *involution* iff $f^2 = \mathrm{id}$, and is *idempotent* iff $f^2 = f$. We know from Proposition 3.5 that if f is idempotent, then

$$E = \mathrm{Im}(f) \oplus \mathrm{Ker}\,(f),$$

and that the restriction of f to its image is the identity. For this reason, a linear involution is called a *projection*. The connection between involutions and projections is given by the following simple proposition.

Proposition 11.16. *For any linear map $f: E \to E$, we have $f^2 = \mathrm{id}$ iff $\frac{1}{2}(\mathrm{id} - f)$ is a projection iff $\frac{1}{2}(\mathrm{id} + f)$ is a projection; in this case, f is equal to the difference of the two projections $\frac{1}{2}(\mathrm{id} + f)$ and $\frac{1}{2}(\mathrm{id} - f)$.*

Proof. We have

$$\left(\frac{1}{2}(\mathrm{id} - f)\right)^2 = \frac{1}{4}(\mathrm{id} - 2f + f^2)$$

so

$$\left(\frac{1}{2}(\mathrm{id} - f)\right)^2 = \frac{1}{2}(\mathrm{id} - f) \quad \text{iff} \quad f^2 = \mathrm{id}.$$

We also have

$$\left(\frac{1}{2}(\mathrm{id} + f)\right)^2 = \frac{1}{4}(\mathrm{id} + 2f + f^2),$$

so

$$\left(\frac{1}{2}(\mathrm{id} + f)\right)^2 = \frac{1}{2}(\mathrm{id} + f) \quad \text{iff} \quad f^2 = \mathrm{id}.$$

Oviously, $f = \frac{1}{2}(\mathrm{id} + f) - \frac{1}{2}(\mathrm{id} - f)$. □

Let $U^+ = \mathrm{Ker}\,(\frac{1}{2}(\mathrm{id} - f))$ and let $U^- = \mathrm{Im}(\frac{1}{2}(\mathrm{id} - f))$. If $f^2 = \mathrm{id}$, then

$$(\mathrm{id} + f) \circ (\mathrm{id} - f) = \mathrm{id} - f^2 = \mathrm{id} - \mathrm{id} = 0,$$

which implies that

$$\mathrm{Im}\left(\frac{1}{2}(\mathrm{id} + f)\right) \subseteq \mathrm{Ker}\left(\frac{1}{2}(\mathrm{id} - f)\right).$$

Conversely, if $u \in \mathrm{Ker}\left(\frac{1}{2}(\mathrm{id} - f)\right)$, then $f(u) = u$, so

$$\frac{1}{2}(\mathrm{id} + f)(u) = \frac{1}{2}(u + u) = u,$$

and thus

$$\mathrm{Ker}\left(\frac{1}{2}(\mathrm{id} - f)\right) \subseteq \mathrm{Im}\left(\frac{1}{2}(\mathrm{id} + f)\right).$$

Therefore,

$$U^+ = \mathrm{Ker}\left(\frac{1}{2}(\mathrm{id} - f)\right) = \mathrm{Im}\left(\frac{1}{2}(\mathrm{id} + f)\right),$$

and so, $f(u) = u$ on U^+ and $f(u) = -u$ on U^-. The involutions of E that are unitary transformations are characterized as follows.

Proposition 11.17. *Let $f \in \mathbf{GL}(E)$ be an involution. The following properties are equivalent:*

(a) The map f is unitary; that is, $f \in \mathbf{U}(E)$.

(b) The subspaces $U^- = \mathrm{Im}(\frac{1}{2}(\mathrm{id} - f))$ and $U^+ = \mathrm{Im}(\frac{1}{2}(\mathrm{id} + f))$ are orthogonal.

Furthermore, if E is finite-dimensional, then (a) and (b) are equivalent to

(c) The map is self-adjoint; that is, $f = f^*$.

Proof. If f is unitary, then from $\langle f(u), f(v)\rangle = \langle u, v\rangle$ for all $u, v \in E$, we see that if $u \in U^+$ and $v \in U^-$, we get

$$\langle u, v\rangle = \langle f(u), f(v)\rangle = \langle u, -v\rangle = -\langle u, v\rangle,$$

so $2\langle u, v\rangle = 0$, which implies $\langle u, v\rangle = 0$, that is, U^+ and U^- are orthogonal. Thus, (a) implies (b).

Conversely, if (b) holds, since $f(u) = u$ on U^+ and $f(u) = -u$ on U^-, we see that $\langle f(u), f(v)\rangle = \langle u, v\rangle$ if $u, v \in U^+$ or if $u, v \in U^-$. Since $E = U^+ \oplus U^-$ and since U^+ and U^- are orthogonal, we also have $\langle f(u), f(v)\rangle = \langle u, v\rangle$ for all $u, v \in E$, and (b) implies (a).

If E is finite-dimensional, the adjoint f^* of f exists, and we know that $f^{-1} = f^*$. Since f is an involution, $f^2 = \mathrm{id}$, which implies that $f^* = f^{-1} = f$. □

A unitary involution is the identity on $U^+ = \mathrm{Im}(\frac{1}{2}(\mathrm{id} + f))$, and $f(v) = -v$ for all $v \in U^- = \mathrm{Im}(\frac{1}{2}(\mathrm{id} - f))$. Furthermore, E is an orthogonal direct sum $E = U^+ \oplus U^-$. We say that f is an *orthogonal reflection* about U^+. In the special case where U^+ is a hyperplane, we say that f is a *hyperplane reflection*. We already studied hyperplane reflections in the Euclidean case; see Chapter 10.

If $f\colon E \to E$ is a projection ($f^2 = f$), then

$$(\mathrm{id} - 2f)^2 = \mathrm{id} - 4f + 4f^2 = \mathrm{id} - 4f + 4f = \mathrm{id},$$

so $\mathrm{id} - 2f$ is an involution. As a consequence, we get the following result.

Proposition 11.18. *If $f\colon E \to E$ is a projection ($f^2 = f$), then* $\mathrm{Ker}\,(f)$ *and* $\mathrm{Im}(f)$ *are orthogonal iff $f^* = f$.*

Proof. Apply Proposition 11.17 to $g = \mathrm{id} - 2f$. Since $\mathrm{id} - g = 2f$ we have

$$U^+ = \mathrm{Ker}\left(\frac{1}{2}(\mathrm{id} - g)\right) = \mathrm{Ker}\,(f)$$

and

$$U^- = \mathrm{Im}\left(\frac{1}{2}(\mathrm{id} - g)\right) = \mathrm{Im}(f),$$

which proves the proposition. □

A projection such that $f = f^*$ is called an *orthogonal projection*.

If $(a_1 \ldots, a_k)$ are k linearly independent vectors in $\mathbb{R}^n$, let us determine the matrix P of the orthogonal projection onto the subspace of $\mathbb{R}^n$ spanned by $(a_1, \ldots, a_k)$. Let A be the $n \times k$ matrix whose jth column consists of the coordinates of the vector a_j over the canonical basis $(e_1, \ldots, e_n)$.

Any vector in the subspace $(a_1, \ldots, a_k)$ is a linear combination of the form Ax, for some $x \in \mathbb{R}^k$. Given any $y \in \mathbb{R}^n$, the orthogonal projection $Py = Ax$ of y onto the subspace spanned by $(a_1, \ldots, a_k)$ is the vector Ax such that $y - Ax$ is orthogonal to the subspace spanned by $(a_1, \ldots, a_k)$ (prove it). This means that $y - Ax$ is orthogonal to every a_j, which is expressed by

$$A^\top(y - Ax) = 0;$$

that is,

$$A^\top Ax = A^\top y.$$

The matrix $A^\top A$ is invertible because A has full rank k, thus we get

$$x = (A^\top A)^{-1} A^\top y,$$

and so

$$Py = Ax = A(A^\top A)^{-1} A^\top y.$$

Therefore, the matrix P of the projection onto the subspace spanned by $(a_1 \ldots, a_k)$ is given by

$$P = A(A^\top A)^{-1}A^\top.$$

The reader should check that $P^2 = P$ and $P^\top = P$.

11.6 Dual Norms

In the remark following the proof of Proposition 6.8, we explained that if $(E, \| \|)$ and $(F, \| \|)$ are two normed vector spaces and if we let $\mathcal{L}(E; F)$ denote the set of all continuous (equivalently, bounded) linear maps from E to F, then, we can define the *operator norm* (or *subordinate norm*) $\| \|$ on $\mathcal{L}(E; F)$ as follows: for every $f \in \mathcal{L}(E; F)$,

$$\|f\| = \sup_{\substack{x \in E \\ x \neq 0}} \frac{\|f(x)\|}{\|x\|} = \sup_{\substack{x \in E \\ \|x\|=1}} \|f(x)\|.$$

In particular, if $F = \mathbb{C}$, then $\mathcal{L}(E; F) = E'$ is the *dual space* of E, and we get the operator norm denoted by $\| \|_*$ given by

$$\|f\|_* = \sup_{\substack{x \in E \\ \|x\|=1}} |f(x)|.$$

The norm $\| \|_*$ is called the *dual norm* of $\| \|$ on E'.

Let us now assume that E is a finite-dimensional Hermitian space, in which case $E' = E^*$. Theorem 11.5 implies that for every linear form $f \in E^*$, there is a unique vector $y \in E$ so that

$$f(x) = \langle x, y \rangle,$$

for all $x \in E$, and so we can write

$$\|f\|_* = \sup_{\substack{x \in E \\ \|x\|=1}} |\langle x, y \rangle|.$$

The above suggests defining a norm $\| \|^D$ on E.

Definition 11.8. If E is a finite-dimensional Hermitian space and $\| \|$ is any norm on E, for any $y \in E$ we let

$$\|y\|^D = \sup_{\substack{x \in E \\ \|x\|=1}} |\langle x, y \rangle|,$$

be the *dual norm* of $\| \|$ (on E). If E is a real Euclidean space, then the dual norm is defined by

$$\|y\|^D = \sup_{\substack{x \in E \\ \|x\|=1}} \langle x, y \rangle$$

for all $y \in E$.

Beware that $\| \ \|$ is generally *not* the Hermitian norm associated with the Hermitian innner product. The dual norm shows up in convex programming; see Boyd and Vandenberghe [22], Chapters 2, 3, 6, 9.

The fact that $\| \ \|^D$ is a norm follows from the fact that $\| \ \|_*$ is a norm and can also be checked directly. It is worth noting that the triangle inequality for $\| \ \|^D$ comes "for free," in the sense that it holds for any function $p\colon E \to \mathbb{R}$. Indeed, we have

$$\begin{aligned} p^D(x+y) &= \sup_{p(z)=1} |\langle z, x+y\rangle| \\ &= \sup_{p(z)=1} (|\langle z, x\rangle + \langle z, y\rangle|) \\ &\leq \sup_{p(z)=1} (|\langle z, x\rangle| + |\langle z, y\rangle|) \\ &\leq \sup_{p(z)=1} |\langle z, x\rangle| + \sup_{p(z)=1} |\langle z, y\rangle| \\ &= p^D(x) + p^D(y). \end{aligned}$$

If $p\colon E \to \mathbb{R}$ is a function such that

(1) $p(x) \geq 0$ for all $x \in E$, and $p(x) = 0$ iff $x = 0$;

(2) $p(\lambda x) = |\lambda| p(x)$, for all $x \in E$ and all $\lambda \in \mathbb{C}$;

(3) p is continuous, in the sense that for some basis $(e_1, \ldots, e_n)$ of E, the function

$$(x_1, \ldots, x_n) \mapsto p(x_1 e_1 + \cdots + x_n e_n)$$

from $\mathbb{C}^n$ to $\mathbb{R}$ is continuous;

then we say that p is a *pre-norm*. Obviously, every norm is a pre-norm, but a pre-norm may not satisfy the triangle inequality. However, we just showed that the dual norm of any pre-norm is actually a norm.

Since E is finite dimensional, the unit sphere $S^{n-1} = \{x \in E \mid \|x\| = 1\}$ is compact, so there is some $x_0 \in S^{n-1}$ such that

$$\|y\|^D = |\langle x_0, y\rangle|.$$

If $\langle x_0, y\rangle = \rho e^{i\theta}$, with $\rho \geq 0$, then

$$|\langle e^{-i\theta} x_0, y\rangle| = |e^{-i\theta}\langle x_0, y\rangle| = |e^{-i\theta}\rho e^{i\theta}| = \rho,$$

so

$$\|y\|^D = \rho = |\langle e^{-i\theta} x_0, y\rangle|,$$

with $\left\|e^{-i\theta} x_0\right\| = \|x_0\| = 1$. On the other hand,

$$\Re\langle x, y\rangle \leq |\langle x, y\rangle|,$$

so we get

$$\|y\|^D = \sup_{\substack{x\in E\\ \|x\|=1}} |\langle x, y\rangle| = \sup_{\substack{x\in E\\ \|x\|=1}} \Re\langle x, y\rangle.$$

Proposition 11.19. *For all $x, y \in E$, we have*

$$|\langle x, y\rangle| \leq \|x\| \, \|y\|^D$$
$$|\langle x, y\rangle| \leq \|x\|^D \, \|y\| .$$

Proof. If $x = 0$, then $\langle x, y\rangle = 0$ and these inequalities are trivial. If $x \neq 0$, since $\|x/\|x\|\| = 1$, by definition of $\|y\|^D$, we have

$$|\langle x/\|x\|, y\rangle| \leq \sup_{\|z\|=1} |\langle z, y\rangle| = \|y\|^D ,$$

which yields

$$|\langle x, y\rangle| \leq \|x\| \, \|y\|^D .$$

The second inequality holds because $|\langle x, y\rangle| = |\langle y, x\rangle|$. □

It is not hard to show that

$$\|y\|_1^D = \|y\|_\infty$$
$$\|y\|_\infty^D = \|y\|_1$$
$$\|y\|_2^D = \|y\|_2 .$$

Thus, the Euclidean norm is autodual. More generally, if $p, q \geq 1$ and $1/p + 1/q = 1$, we have

$$\|y\|_p^D = \|y\|_q .$$

It can also be shown that the dual of the spectral norm is the trace norm (or nuclear norm) from Section 15.3. We close this section by stating the following duality theorem.

Theorem 11.20. *If E is a finite-dimensional Hermitian space, then for any norm $\|\ \|$ on E, we have*

$$\|y\|^{DD} = \|y\|$$

for all $y \in E$.

Proof. By Proposition 11.19, we have

$$|\langle x, y\rangle| \leq \|x\|^D \, \|y\| ,$$

so we get

$$\|y\|^{DD} = \sup_{\|x\|^D=1} |\langle x, y\rangle| \leq \|y\| , \quad \text{for all } y \in E.$$

It remains to prove that

$$\|y\| \leq \|y\|^{DD}, \quad \text{for all } y \in E.$$

Proofs of this fact can be found in Horn and Johnson [55] (Section 5.5), and in Serre [95] (Chapter 7). The proof makes use of the fact that a nonempty, closed, convex set has a supporting hyperplane through each of its boundary points, a result known as *Minkowski's lemma*. This result is a consequence of the *Hahn–Banach theorem*; see Gallier [44]. We give the proof in the case where E is a real Euclidean space. Some minor modifications have to be made when dealing with complex vector spaces and are left as an exercise.

Since the unit ball $B = \{z \in E \mid \|z\| \leq 1\}$ is closed and convex, the Minkowski lemma says for every x such that $\|x\| = 1$, there is an affine map g, of the form

$$g(z) = \langle z, w\rangle - \langle x, w\rangle$$

with $\|w\| = 1$, such that $g(x) = 0$ and $g(z) \leq 0$ for all z such that $\|z\| \leq 1$. Then, it is clear that

$$\sup_{\|z\|=1} \langle z, w\rangle = \langle x, w\rangle,$$

and so

$$\|w\|^D = \langle x, w\rangle.$$

It follows that

$$\|x\|^{DD} \geq \langle w/\|w\|^D, x\rangle = \frac{\langle x, w\rangle}{\|w\|^D} = 1 = \|x\|$$

for all x such that $\|x\| = 1$. By homogeneity, this is true for all $y \in E$, which completes the proof in the real case. When E is a complex vector space, we have to view the unit ball B as a closed convex set in $\mathbb{R}^{2n}$ and we use the fact that there is real affine map of the form

$$g(z) = \Re\langle z, w\rangle - \Re\langle x, w\rangle$$

such that $g(x) = 0$ and $g(z) \leq 0$ for all z with $\|z\| = 1$, so that $\|w\|^D = \Re\langle x, w\rangle$. □

More details on dual norms and unitarily invariant norms can be found in Horn and Johnson [55] (Chapters 5 and 7).

11.7 Summary

The main concepts and results of this chapter are listed below:

- *Semilinear maps*.
- *Sesquilinear forms*; *Hermitian forms*.
- *Quadratic form* associated with a sesquilinear form.

- *Polarization identities*.
- *Positive* and *positive definite* Hermitian forms; *pre-Hilbert spaces*, *Hermitian spaces*.
- *Gram matrix* associated with a Hermitian product.
- The *Cauchy–Schwarz inequality* and the *Minkowski inequality*.
- *Hermitian inner product*, *Hermitian norm*.
- The *parallelogram law*.
- The musical isomorphisms $\flat\colon \overline{E} \to E^*$ and $\sharp\colon E^* \to \overline{E}$; Theorem 11.5 ($E$ is finite-dimensional).
- The *adjoint* of a linear map (with respect to a Hermitian inner product).
- Existence of orthonormal bases in a Hermitian space (Proposition 11.8).
- *Gram–Schmidt orthonormalization procedure*.
- *Linear isometries* (*unitary transformations*).
- The *unitary group*, *unitary matrices*.
- The *unitary group* $\mathbf{U}(n)$;
- The *special unitary group* $\mathbf{SU}(n)$.
- QR-Decomposition for invertible matrices.
- The *Hadamard inequality* for complex matrices.
- The *Hadamard inequality* for Hermitian positive semidefinite matrices.
- Orthogonal projections and involutions; orthogonal reflections.
- Dual norms.

Chapter 12

Eigenvectors and Eigenvalues

12.1 Eigenvectors and Eigenvalues of a Linear Map

Given a finite-dimensional vector space E, let $f\colon E \to E$ be any linear map. If, by luck, there is a basis $(e_1, \ldots, e_n)$ of E with respect to which f is represented by a *diagonal matrix*

$$D = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \lambda_n \end{pmatrix},$$

then the action of f on E is very simple; in every "direction" e_i, we have

$$f(e_i) = \lambda_i e_i.$$

We can think of f as a transformation that stretches or shrinks space along the direction $e_1, \ldots, e_n$ (at least if E is a real vector space). In terms of matrices, the above property translates into the fact that there is an invertible matrix P and a diagonal matrix D such that a matrix A can be factored as

$$A = PDP^{-1}.$$

When this happens, we say that f (or A) is *diagonalizable*, the λ_is are called the *eigenvalues* of f, and the e_is are *eigenvectors* of f. For example, we will see that every symmetric matrix can be diagonalized. Unfortunately, not every matrix can be diagonalized. For example, the matrix

$$A_1 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

can't be diagonalized. Sometimes, a matrix fails to be diagonalizable because its eigenvalues do not belong to the field of coefficients, such as

$$A_2 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

whose eigenvalues are $\pm i$. This is not a serious problem because A_2 can be diagonalized over the complex numbers. However, A_1 is a "fatal" case! Indeed, its eigenvalues are both 1 and the problem is that A_1 does not have enough eigenvectors to span E.

The next best thing is that there is a basis with respect to which f is represented by an *upper triangular* matrix. In this case we say that f can be *triangularized*, or that f is *triangulable*. As we will see in Section 12.2, if all the eigenvalues of f belong to the field of coefficients K, then f can be triangularized. In particular, this is the case if $K = \mathbb{C}$.

Now, an alternative to triangularization is to consider the representation of f with respect to *two* bases $(e_1, \ldots, e_n)$ and $(f_1, \ldots, f_n)$, rather than a single basis. In this case, if $K = \mathbb{R}$ or $K = \mathbb{C}$, it turns out that we can even pick these bases to be *orthonormal*, and we get a diagonal matrix Σ with *nonnegative entries*, such that

$$f(e_i) = \sigma_i f_i, \quad 1 \leq i \leq n.$$

The nonzero σ_is are the *singular values* of f, and the corresponding representation is the *singular value decomposition*, or *SVD*. The SVD plays a very important role in applications, and will be considered in detail later.

In this section, we focus on the possibility of diagonalizing a linear map, and we introduce the relevant concepts to do so. Given a vector space E over a field K, let id denote the identity map on E.

The notion of eigenvalue of a linear map $f\colon E \to E$ defined on an infinite-dimensional space E is quite subtle because it cannot be defined in terms of eigenvectors as in the finite-dimensional case. The problem is that the map $\lambda \,\mathrm{id} - f$ (with $\lambda \in \mathbb{C}$) could be noninvertible (because it is not surjective) and yet injective. In finite dimension this cannot happen, so until further notice we *assume that E is of finite dimension n.*

Definition 12.1. Given any vector space E of finite dimension n and any linear map $f\colon E \to E$, a scalar $\lambda \in K$ is called an *eigenvalue, or proper value, or characteristic value of f* if there is some *nonzero* vector $u \in E$ such that

$$f(u) = \lambda u.$$

Equivalently, λ is an eigenvalue of f if $\mathrm{Ker}\,(\lambda \,\mathrm{id} - f)$ is nontrivial (i.e., $\mathrm{Ker}\,(\lambda \,\mathrm{id} - f) \neq \{0\}$) iff $\lambda \,\mathrm{id} - f$ is *not* invertible (this is where the fact that E is finite-dimensional is used; a linear map from E to istelf is injective iff it is invertible). A vector $u \in E$ is called an *eigenvector, or proper vector, or characteristic vector of f* if $u \neq 0$ and if there is some $\lambda \in K$ such that

$$f(u) = \lambda u;$$

the scalar λ is then an eigenvalue, and we say that u is an *eigenvector associated with* λ. Given any eigenvalue $\lambda \in K$, the nontrivial subspace $\mathrm{Ker}\,(\lambda \,\mathrm{id} - f)$ consists of all the eigenvectors associated with λ together with the zero vector; this subspace is denoted by $E_\lambda(f)$, or $E(\lambda, f)$, or even by E_λ, and is called the *eigenspace associated with λ, or proper subspace associated with λ.*

Note that distinct eigenvectors may correspond to the same eigenvalue, but distinct eigenvalues correspond to disjoint sets of eigenvectors.

Remark: As we emphasized in the remark following Definition 6.4, we *require an eigenvector to be nonzero*. This requirement seems to have more benefits than inconvenients, even though it may considered somewhat inelegant because the set of all eigenvectors associated with an eigenvalue is not a subspace since the zero vector is excluded.

The next proposition shows that the eigenvalues of a linear map $f\colon E \to E$ are the roots of a polynomial associated with f.

Proposition 12.1. *Let E be any vector space of finite dimension n and let f be any linear map $f\colon E \to E$. The eigenvalues of f are the roots (in K) of the polynomial*

$$\det(\lambda \operatorname{id} - f).$$

Proof. A scalar $\lambda \in K$ is an eigenvalue of f iff there is some vector $u \neq 0$ in E such that

$$f(u) = \lambda u$$

iff

$$(\lambda \operatorname{id} - f)(u) = 0$$

iff $(\lambda \operatorname{id} - f)$ is not invertible iff, by Proposition 4.15,

$$\det(\lambda \operatorname{id} - f) = 0. \qquad \square$$

In view of the importance of the polynomial $\det(\lambda \operatorname{id} - f)$, we have the following definition.

Definition 12.2. Given any vector space E of dimension n, for any linear map $f\colon E \to E$, the polynomial $P_f(X) = \chi_f(X) = \det(X \operatorname{id} - f)$ is called the *characteristic polynomial of f*. For any square matrix A, the polynomial $P_A(X) = \chi_A(X) = \det(XI - A)$ is called the *characteristic polynomial of A*.

Note that we already encountered the characteristic polynomial in Section 4.7; see Definition 4.11.

Given any basis $(e_1, \ldots, e_n)$, if $A = M(f)$ is the matrix of f w.r.t. $(e_1, \ldots, e_n)$, we can compute the characteristic polynomial $\chi_f(X) = \det(X \operatorname{id} - f)$ of f by expanding the following determinant:

$$\det(XI - A) = \begin{vmatrix} X - a_{1\,1} & -a_{1\,2} & \ldots & -a_{1\,n} \\ -a_{2\,1} & X - a_{2\,2} & \ldots & -a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ -a_{n\,1} & -a_{n\,2} & \ldots & X - a_{n\,n} \end{vmatrix}.$$

If we expand this determinant, we find that

$$\chi_A(X) = \det(XI - A) = X^n - (a_{1\,1} + \cdots + a_{n\,n})X^{n-1} + \cdots + (-1)^n \det(A).$$

The sum $\mathrm{tr}(A) = a_{1\,1} + \cdots + a_{n\,n}$ of the diagonal elements of A is called the *trace of* A. Since we proved in Section 4.7 that the characteristic polynomial only depends on the linear map f, the above shows that $\mathrm{tr}(A)$ has the same value for all matrices A representing f. Thus, the *trace of a linear map* is well-defined; we have $\mathrm{tr}(f) = \mathrm{tr}(A)$ for any matrix A representing f.

Remark: The characteristic polynomial of a linear map is sometimes defined as $\det(f - X\,\mathrm{id})$. Since

$$\det(f - X\,\mathrm{id}) = (-1)^n \det(X\,\mathrm{id} - f),$$

this makes essentially no difference but the version $\det(X\,\mathrm{id} - f)$ has the small advantage that the coefficient of X^n is $+1$.

If we write

$$\chi_A(X) = \det(XI - A) = X^n - \tau_1(A)X^{n-1} + \cdots + (-1)^k\tau_k(A)X^{n-k} + \cdots + (-1)^n\tau_n(A),$$

then we just proved that

$$\tau_1(A) = \mathrm{tr}(A) \quad \text{and} \quad \tau_n(A) = \det(A).$$

It is also possible to express $\tau_k(A)$ in terms of determinants of certain submatrices of A. For any nonempty subset, $I \subseteq \{1, \ldots, n\}$, say $I = \{i_1 < \ldots < i_k\}$, let $A_{I,I}$ be the $k \times k$ submatrix of A whose jth column consists of the elements $a_{i_h\, i_j}$, where $h = 1, \ldots, k$. Equivalently, $A_{I,I}$ is the matrix obtained from A by first selecting the columns whose indices belong to I, and then the rows whose indices also belong to I. Then, it can be shown that

$$\tau_k(A) = \sum_{\substack{I \subseteq \{1,\ldots,n\} \\ |I| = k}} \det(A_{I,I}).$$

If all the roots, $\lambda_1, \ldots, \lambda_n$, of the polynomial $\det(XI - A)$ belong to the field K, then we can write

$$\chi_A(X) = \det(XI - A) = (X - \lambda_1) \cdots (X - \lambda_n),$$

where some of the λ_is may appear more than once. Consequently,

$$\chi_A(X) = \det(XI - A) = X^n - \sigma_1(\lambda)X^{n-1} + \cdots + (-1)^k\sigma_k(\lambda)X^{n-k} + \cdots + (-1)^n\sigma_n(\lambda),$$

where

$$\sigma_k(\lambda) = \sum_{\substack{I \subseteq \{1,\ldots,n\} \\ |I| = k}} \prod_{i \in I} \lambda_i,$$

the *kth elementary symmetric polynomial (or function)* of the λ_i's, where $\lambda = (\lambda_1, \ldots, \lambda_n)$. The elementary symmetric polynomial $\sigma_k(\lambda)$ is often denoted $E_k(\lambda)$, but this notation may be confusing in the context of linear algebra. For $n = 5$, the elementary symmetric polynomials are listed below:

$$\begin{aligned}
\sigma_0(\lambda) &= 1\\
\sigma_1(\lambda) &= \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5\\
\sigma_2(\lambda) &= \lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_1\lambda_4 + \lambda_1\lambda_5 + \lambda_2\lambda_3 + \lambda_2\lambda_4 + \lambda_2\lambda_5\\
&\quad + \lambda_3\lambda_4 + \lambda_3\lambda_5 + \lambda_4\lambda_5\\
\sigma_3(\lambda) &= \lambda_3\lambda_4\lambda_5 + \lambda_2\lambda_4\lambda_5 + \lambda_2\lambda_3\lambda_5 + \lambda_2\lambda_3\lambda_4 + \lambda_1\lambda_4\lambda_5\\
&\quad + \lambda_1\lambda_3\lambda_5 + \lambda_1\lambda_3\lambda_4 + \lambda_1\lambda_2\lambda_5 + \lambda_1\lambda_2\lambda_4 + \lambda_1\lambda_2\lambda_3\\
\sigma_4(\lambda) &= \lambda_1\lambda_2\lambda_3\lambda_4 + \lambda_1\lambda_2\lambda_3\lambda_5 + \lambda_1\lambda_2\lambda_4\lambda_5 + \lambda_1\lambda_3\lambda_4\lambda_5 + \lambda_2\lambda_3\lambda_4\lambda_5\\
\sigma_5(\lambda) &= \lambda_1\lambda_2\lambda_3\lambda_4\lambda_5.
\end{aligned}$$

Since

$$\begin{aligned}
\chi_A(X) &= X^n - \tau_1(A)X^{n-1} + \cdots + (-1)^k\tau_k(A)X^{n-k} + \cdots + (-1)^n\tau_n(A)\\
&= X^n - \sigma_1(\lambda)X^{n-1} + \cdots + (-1)^k\sigma_k(\lambda)X^{n-k} + \cdots + (-1)^n\sigma_n(\lambda),
\end{aligned}$$

we have

$$\sigma_k(\lambda) = \tau_k(A), \quad k = 1, \ldots, n,$$

and in particular, the product of the eigenvalues of f is equal to $\det(A) = \det(f)$, and the sum of the eigenvalues of f is equal to the trace $\mathrm{tr}(A) = \mathrm{tr}(f)$, of f; for the record,

$$\begin{aligned}
\mathrm{tr}(f) &= \lambda_1 + \cdots + \lambda_n\\
\det(f) &= \lambda_1 \cdots \lambda_n,
\end{aligned}$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of f (and A), where some of the λ_is may appear more than once. In particular, f is not invertible iff it admits 0 has an eigenvalue.

Remark: Depending on the field K, the characteristic polynomial $\chi_A(X) = \det(XI - A)$ may or may not have roots in K. This motivates considering *algebraically closed fields*, which are fields K such that every polynomial with coefficients in K has all its root in K. For example, over $K = \mathbb{R}$, not every polynomial has real roots. If we consider the matrix

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix},$$

then the characteristic polynomial $\det(XI - A)$ has no real roots unless $\theta = k\pi$. However, over the field $\mathbb{C}$ of complex numbers, every polynomial has roots. For example, the matrix above has the roots $\cos\theta \pm i\sin\theta = e^{\pm i\theta}$.

It is possible to show that every linear map f over a complex vector space E must have some (complex) eigenvalue without having recourse to determinants (and the characteristic polynomial). Let $n = \dim(E)$, pick any nonzero vector $u \in E$, and consider the sequence

$$u, f(u), f^2(u), \ldots, f^n(u).$$

Since the above sequence has $n+1$ vectors and E has dimension n, these vectors must be linearly dependent, so there are some complex numbers $c_0, \ldots, c_m$, not all zero, such that

$$c_0 f^m(u) + c_1 f^{m-1}(u) + \cdots + c_m u = 0,$$

where $m \leq n$ is the largest integer such that the coefficient of $f^m(u)$ is nonzero (m must exits since we have a nontrivial linear dependency). Now, because the field $\mathbb{C}$ is algebraically closed, the polynomial

$$c_0 X^m + c_1 X^{m-1} + \cdots + c_m$$

can be written as a product of linear factors as

$$c_0 X^m + c_1 X^{m-1} + \cdots + c_m = c_0 (X - \lambda_1) \cdots (X - \lambda_m)$$

for some complex numbers $\lambda_1, \ldots, \lambda_m \in \mathbb{C}$, not necessarily distinct. But then, since $c_0 \neq 0$,

$$c_0 f^m(u) + c_1 f^{m-1}(u) + \cdots + c_m u = 0$$

is equivalent to

$$(f - \lambda_1 \,\mathrm{id}) \circ \cdots \circ (f - \lambda_m \,\mathrm{id})(u) = 0.$$

If all the linear maps $f - \lambda_i \,\mathrm{id}$ were injective, then $(f - \lambda_1 \,\mathrm{id}) \circ \cdots \circ (f - \lambda_m \,\mathrm{id})$ would be injective, contradicting the fact that $u \neq 0$. Therefore, some linear map $f - \lambda_i \,\mathrm{id}$ must have a nontrivial kernel, which means that there is some $v \neq 0$ so that

$$f(v) = \lambda_i v;$$

that is, λ_i is some eigenvalue of f and v is some eigenvector of f.

As nice as the above argument is, it does not provide a method for *finding* the eigenvalues of f, and even if we prefer avoiding determinants as a much as possible, we are forced to deal with the characteristic polynomial $\det(X \,\mathrm{id} - f)$.

Definition 12.3. Let A be an $n \times n$ matrix over a field K. Assume that all the roots of the characteristic polynomial $\chi_A(X) = \det(XI - A)$ of A belong to K, which means that we can write

$$\det(XI - A) = (X - \lambda_1)^{k_1} \cdots (X - \lambda_m)^{k_m},$$

where $\lambda_1, \ldots, \lambda_m \in K$ are the distinct roots of $\det(XI - A)$ and $k_1 + \cdots + k_m = n$. The integer k_i is called the *algebraic multiplicity* of the eigenvalue λ_i, and the dimension of the eigenspace $E_{\lambda_i} = \mathrm{Ker}(\lambda_i I - A)$ is called the *geometric multiplicity* of λ_i. We denote the algebraic multiplicity of λ_i by $\mathrm{alg}(\lambda_i)$, and its geometric multiplicity by $\mathrm{geo}(\lambda_i)$.

By definition, the sum of the algebraic multiplicities is equal to n, but the sum of the geometric multiplicities can be strictly smaller.

Proposition 12.2. *Let A be an $n \times n$ matrix over a field K and assume that all the roots of the characteristic polynomial $\chi_A(X) = \det(XI - A)$ of A belong to K. For every eigenvalue λ_i of A, the geometric multiplicity of λ_i is always less than or equal to its algebraic multiplicity, that is,*

$$\mathrm{geo}(\lambda_i) \leq \mathrm{alg}(\lambda_i).$$

Proof. To see this, if n_i is the dimension of the eigenspace E_{λ_i} associated with the eigenvalue λ_i, we can form a basis of K^n obtained by picking a basis of E_{λ_i} and completing this linearly independent family to a basis of K^n. With respect to this new basis, our matrix is of the form

$$A' = \begin{pmatrix} \lambda_i I_{n_i} & B \\ 0 & D \end{pmatrix}$$

and a simple determinant calculation shows that

$$\det(XI - A) = \det(XI - A') = (X - \lambda_i)^{n_i} \det(XI_{n-n_i} - D).$$

Therefore, $(X - \lambda_i)^{n_i}$ divides the characteristic polynomial of A', and thus, the characteristic polynomial of A. It follows that n_i is less than or equal to the algebraic multiplicity of λ_i. □

The following proposition shows an interesting property of eigenspaces.

Proposition 12.3. *Let E be any vector space of finite dimension n and let f be any linear map. If $u_1, \ldots, u_m$ are eigenvectors associated with pairwise distinct eigenvalues $\lambda_1, \ldots, \lambda_m$, then the family $(u_1, \ldots, u_m)$ is linearly independent.*

Proof. Assume that $(u_1, \ldots, u_m)$ is linearly dependent. Then, there exists $\mu_1, \ldots, \mu_k \in K$ such that

$$\mu_1 u_{i_1} + \cdots + \mu_k u_{i_k} = 0,$$

where $1 \leq k \leq m$, $\mu_i \neq 0$ for all i, $1 \leq i \leq k$, $\{i_1, \ldots, i_k\} \subseteq \{1, \ldots, m\}$, and no proper subfamily of $(u_{i_1}, \ldots, u_{i_k})$ is linearly dependent (in other words, we consider a dependency relation with k minimal). Applying f to this dependency relation, we get

$$\mu_1 \lambda_{i_1} u_{i_1} + \cdots + \mu_k \lambda_{i_k} u_{i_k} = 0,$$

and if we multiply the original dependency relation by λ_{i_1} and subtract it from the above, we get

$$\mu_2(\lambda_{i_2} - \lambda_{i_1})u_{i_2} + \cdots + \mu_k(\lambda_{i_k} - \lambda_{i_1})u_{i_k} = 0,$$

which is a nontrivial linear dependency among a proper subfamily of $(u_{i_1}, \ldots, u_{i_k})$ since the λ_j are all distinct and the μ_i are nonzero, a contradiction. □

Thus, from Proposition 12.3, if $\lambda_1, \ldots, \lambda_m$ are all the pairwise distinct eigenvalues of f (where $m \leq n$), we have a direct sum

$$E_{\lambda_1} \oplus \cdots \oplus E_{\lambda_m}$$

of the eigenspaces E_{λ_i}. Unfortunately, it is not always the case that

$$E = E_{\lambda_1} \oplus \cdots \oplus E_{\lambda_m}.$$

When

$$E = E_{\lambda_1} \oplus \cdots \oplus E_{\lambda_m},$$

we say that f is *diagonalizable* (and similarly for any matrix associated with f). Indeed, picking a basis in each E_{λ_i}, we obtain a matrix which is a diagonal matrix consisting of the eigenvalues, each λ_i occurring a number of times equal to the dimension of E_{λ_i}. This happens if the algebraic multiplicity and the geometric multiplicity of every eigenvalue are equal. In particular, when the characteristic polynomial has n distinct roots, then f is diagonalizable. It can also be shown that symmetric matrices have real eigenvalues and can be diagonalized.

For a negative example, we leave as exercise to show that the matrix

$$M = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

cannot be diagonalized, even though 1 is an eigenvalue. The problem is that the eigenspace of 1 only has dimension 1. The matrix

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

cannot be diagonalized either, because it has no real eigenvalues, unless $\theta = k\pi$. However, over the field of complex numbers, it can be diagonalized.

12.2 Reduction to Upper Triangular Form

Unfortunately, not every linear map on a complex vector space can be diagonalized. The next best thing is to "triangularize," which means to find a basis over which the matrix has zero entries below the main diagonal. Fortunately, such a basis always exist.

We say that a square matrix A is an *upper triangular matrix* if it has the following shape,

$$\begin{pmatrix} a_{1\,1} & a_{1\,2} & a_{1\,3} & \ldots & a_{1\,n-1} & a_{1\,n} \\ 0 & a_{2\,2} & a_{2\,3} & \ldots & a_{2\,n-1} & a_{2\,n} \\ 0 & 0 & a_{3\,3} & \ldots & a_{3\,n-1} & a_{3\,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & a_{n-1\,n-1} & a_{n-1\,n} \\ 0 & 0 & 0 & \ldots & 0 & a_{n\,n} \end{pmatrix},$$

i.e., $a_{i\,j} = 0$ whenever $j < i$, $1 \leq i, j \leq n$.

Theorem 12.4. *Given any finite dimensional vector space over a field K, for any linear map $f\colon E \to E$, there is a basis $(u_1, \ldots, u_n)$ with respect to which f is represented by an upper triangular matrix (in $\mathrm{M}_n(K)$) iff all the eigenvalues of f belong to K. Equivalently, for every $n \times n$ matrix $A \in \mathrm{M}_n(K)$, there is an invertible matrix P and an upper triangular matrix T (both in $\mathrm{M}_n(K)$) such that*

$$A = PTP^{-1}$$

iff all the eigenvalues of A belong to K.

Proof. If there is a basis $(u_1, \ldots, u_n)$ with respect to which f is represented by an upper triangular matrix T in $\mathrm{M}_n(K)$, then since the eigenvalues of f are the diagonal entries of T, all the eigenvalues of f belong to K.

For the converse, we proceed by induction on the dimension n of E. For $n = 1$ the result is obvious. If $n > 1$, since by assumption f has all its eigenvalue in K, pick some eigenvalue $\lambda_1 \in K$ of f, and let u_1 be some corresponding (nonzero) eigenvector. We can find $n-1$ vectors $(v_2, \ldots, v_n)$ such that $(u_1, v_2, \ldots, v_n)$ is a basis of E, and let F be the subspace of dimension $n-1$ spanned by $(v_2, \ldots, v_n)$. In the basis $(u_1, v_2 \ldots, v_n)$, the matrix of f is of the form

$$U = \begin{pmatrix} \lambda_1 & a_{1\,2} & \ldots & a_{1\,n} \\ 0 & a_{2\,2} & \ldots & a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n\,2} & \ldots & a_{n\,n} \end{pmatrix},$$

since its first column contains the coordinates of $\lambda_1 u_1$ over the basis $(u_1, v_2, \ldots, v_n)$. If we let $p\colon E \to F$ be the projection defined such that $p(u_1) = 0$ and $p(v_i) = v_i$ when $2 \leq i \leq n$, the linear map $g\colon F \to F$ defined as the restriction of $p \circ f$ to F is represented by the $(n-1) \times (n-1)$ matrix $V = (a_{i\,j})_{2 \leq i,j \leq n}$ over the basis $(v_2, \ldots, v_n)$. We need to prove that all the eigenvalues of g belong to K. However, since the first column of U has a single nonzero entry, we get

$$\chi_U(X) = \det(XI - U) = (X - \lambda_1)\det(XI - V) = (X - \lambda_1)\chi_V(X),$$

where $\chi_U(X)$ is the characteristic polynomial of U and $\chi_V(X)$ is the characteristic polynomial of V. It follows that $\chi_V(X)$ divides $\chi_U(X)$, and since all the roots of $\chi_U(X)$ are in K, all the roots of $\chi_V(X)$ are also in K. Consequently, we can apply the induction hypothesis, and there is a basis $(u_2, \ldots, u_n)$ of F such that g is represented by an upper triangular matrix $(b_{i\,j})_{1 \leq i,j \leq n-1}$. However,

$$E = Ku_1 \oplus F,$$

and thus $(u_1, \ldots, u_n)$ is a basis for E. Since p is the projection from $E = Ku_1 \oplus F$ onto F and $g\colon F \to F$ is the restriction of $p \circ f$ to F, we have

$$f(u_1) = \lambda_1 u_1$$

and

$$f(u_{i+1}) = a_{1\,i}u_1 + \sum_{j=1}^{i} b_{i\,j}u_{j+1}$$

for some $a_{1\,i} \in K$, when $1 \leq i \leq n-1$. But then the matrix of f with respect to $(u_1, \ldots, u_n)$ is upper triangular.

For the matrix version, we assume that A is the matrix of f with respect to some basis, Then, we just proved that there is a change of basis matrix P such that $A = PTP^{-1}$ where T is upper triangular. □

If $A = PTP^{-1}$ where T is upper triangular, note that the diagonal entries of T are the eigenvalues $\lambda_1, \ldots, \lambda_n$ of A. Indeed, A and T have the same characteristic polynomial. Also, if A is a real matrix whose eigenvalues are all real, then P can be chosen to real, and if A is a rational matrix whose eigenvalues are all rational, then P can be chosen rational. Since any polynomial over $\mathbb{C}$ has all its roots in $\mathbb{C}$, Theorem 12.4 implies that every complex $n \times n$ matrix can be triangularized.

If λ is an eigenvalue of the matrix A and if u is an eigenvector associated with λ, from

$$Au = \lambda u,$$

we obtain

$$A^2u = A(Au) = A(\lambda u) = \lambda Au = \lambda^2 u,$$

which shows that λ^2 is an eigenvalue of A^2 for the eigenvector u. An obvious induction shows that λ^k is an eigenvalue of A^k for the eigenvector u, for all $k \geq 1$. Now, if all eigenvalues $\lambda_1, \ldots, \lambda_n$ of A are in K, it follows that $\lambda_1^k, \ldots, \lambda_n^k$ are eigenvalues of A^k. However, it is not obvious that A^k does not have other eigenvalues. In fact, this can't happen, and this can be proved using Theorem 12.4.

Proposition 12.5. *Given any $n \times n$ matrix $A \in \mathrm{M}_n(K)$ with coefficients in a field K, if all eigenvalues $\lambda_1, \ldots, \lambda_n$ of A are in K, then for every polynomial $q(X) \in K[X]$, the eigenvalues of $q(A)$ are exactly $(q(\lambda_1), \ldots, q(\lambda_n))$.*

Proof. By Theorem 12.4, there is an upper triangular matrix T and an invertible matrix P (both in $\mathrm{M}_n(K)$) such that

$$A = PTP^{-1}.$$

Since A and T are similar, they have the same eigenvalues (with the same multiplicities), so the diagonal entries of T are the eigenvalues of A. Since

$$A^k = PT^kP^{-1}, \quad k \geq 1,$$

for any polynomial $q(X) = c_0X^m + \cdots + c_{m-1}X + c_m$, we have

$$\begin{aligned} q(A) &= c_0A^m + \cdots + c_{m-1}A + c_mI \\ &= c_0PT^mP^{-1} + \cdots + c_{m-1}PTP^{-1} + c_mPIP^{-1} \\ &= P(c_0T^m + \cdots + c_{m-1}T + c_mI)P^{-1} \\ &= Pq(T)P^{-1}. \end{aligned}$$

Furthermore, it is easy to check that $q(T)$ is upper triangular and that its diagonal entries are $q(\lambda_1), \ldots, q(\lambda_n)$, where $\lambda_1, \ldots, \lambda_n$ are the diagonal entries of T, namely the eigenvalues of A. It follows that $q(\lambda_1), \ldots, q(\lambda_n)$ are the eigenvalues of $q(A)$. □

If E is a Hermitian space (see Chapter 11), the proof of Theorem 12.4 can be easily adapted to prove that there is an *orthonormal* basis $(u_1, \ldots, u_n)$ with respect to which the matrix of f is upper triangular. This is usually known as *Schur's lemma*.

Theorem 12.6. *(Schur decomposition) Given any linear map $f\colon E \to E$ over a complex Hermitian space E, there is an orthonormal basis $(u_1, \ldots, u_n)$ with respect to which f is represented by an upper triangular matrix. Equivalently, for every $n \times n$ matrix $A \in \mathrm{M}_n(\mathbb{C})$, there is a unitary matrix U and an upper triangular matrix T such that*

$$A = UTU^*.$$

If A is real and if all its eigenvalues are real, then there is an orthogonal matrix Q and a real upper triangular matrix T such that

$$A = QTQ^\top.$$

Proof. During the induction, we choose F to be the orthogonal complement of $\mathbb{C}u_1$ and we pick orthonormal bases (use Propositions 11.10 and 11.9). If E is a real Euclidean space and if the eigenvalues of f are all real, the proof also goes through with real matrices (use Propositions 9.9 and 9.8). □

Using, Theorem 12.6, we can derive two very important results.

Proposition 12.7. *If A is a Hermitian matrix (i.e. $A^* = A$), then its eigenvalues are real and A can be diagonalized with respect to an orthonormal basis of eigenvectors. In matrix terms, there is a unitary matrix U and a real diagonal matrix D such that $A = UDU^*$. If A is a real symmetric matrix (i.e. $A^\top = A$), then its eigenvalues are real and A can be diagonalized with respect to an orthonormal basis of eigenvectors. In matrix terms, there is an orthogonal matrix Q and a real diagonal matrix D such that $A = QDQ^\top$.*

Proof. By Theorem 12.6, we can write $A = UTU^*$ where $T = (t_{ij})$ is upper triangular and U is a unitary matrix. If $A^* = A$, we get

$$UTU^* = UT^*U^*,$$

which implies that $T = T^*$. Since T is an upper triangular matrix, T^* is a lower triangular matrix, which implies that T is a diagonal matrix. Furthermore, since $T = T^*$, we have $t_{ii} = \overline{t_{ii}}$ for $i = 1, \ldots, n$, which means that the t_{ii} are real, so T is indeed a real diagonal matrix, say D.

If we apply this result to a (real) symmetric matrix A, we obtain the fact that all the eigenvalues of a symmetric matrix are real, and by applying Theorem 12.6 again, we conclude that $A = QDQ^\top$, where Q is orthogonal and D is a real diagonal matrix. □

More general versions of Proposition 12.7 are proved in Chapter 13.

When a real matrix A has complex eigenvalues, there is a version of Theorem 12.6 involving only real matrices provided that we allow T to be block upper-triangular (the diagonal entries may be 2×2 matrices or real entries).

Theorem 12.6 is not a very practical result but it is a useful theoretical result to cope with matrices that cannot be diagonalized. For example, it can be used to prove that *every* complex matrix is the limit of a sequence of diagonalizable matrices that have distinct eigenvalues!

Remark: There is another way to prove Proposition 12.5 that does not use Theorem 12.4, but instead uses the fact that given any field K, there is field extension $\overline{K}$ of K ($K \subseteq \overline{K}$) such that every polynomial $q(X) = c_0X^m + \cdots + c_{m-1}X + c_m$ (of degree $m \geq 1$) with coefficients $c_i \in K$ factors as

$$q(X) = c_0(X - \alpha_1) \cdots (X - \alpha_n), \quad \alpha_i \in \overline{K}, i = 1, \ldots, n.$$

The field $\overline{K}$ is called an *algebraically closed field* (and an algebraic closure of K).

Assume that all eigenvalues $\lambda_1, \ldots, \lambda_n$ of A belong to K. Let $q(X)$ be any polynomial (in $K[X]$) and let $\mu \in \overline{K}$ be any eigenvalue of $q(A)$ (this means that μ is a zero of the characteristic polynomial $\chi_{q(A)}(X) \in K[X]$ of $q(A)$. Since $\overline{K}$ is algebraically closed, $\chi_{q(A)}(X)$ has all its root in $\overline{K}$). We claim that $\mu = q(\lambda_i)$ for some eigenvalue λ_i of A.

Proof. (After Lax [66], Chapter 6). Since $\overline{K}$ is algebraically closed, the polynomial $\mu - q(X)$ factors as

$$\mu - q(X) = c_0(X - \alpha_1) \cdots (X - \alpha_n),$$

for some $\alpha_i \in \overline{K}$. Now, $\mu I - q(A)$ is a matrix in $\mathrm{M}_n(\overline{K})$, and since μ is an eigenvalue of $q(A)$, it must be singular. We have

$$\mu I - q(A) = c_0(A - \alpha_1 I) \cdots (A - \alpha_n I),$$

and since the left-hand side is singular, so is the right-hand side, which implies that some factor $A - \alpha_i I$ is singular. This means that α_i is an eigenvalue of A, say $\alpha_i = \lambda_i$. As $\alpha_i = \lambda_i$ is a zero of $\mu - q(X)$, we get

$$\mu = q(\lambda_i),$$

which proves that μ is indeed of the form $q(\lambda_i)$ for some eigenvalue λ_i of A. □

12.3 Location of Eigenvalues

If A is an $n \times n$ complex (or real) matrix A, it would be useful to know, even roughly, where the eigenvalues of A are located in the complex plane $\mathbb{C}$. The Gershgorin discs provide some precise information about this.

Definition 12.4. For any complex $n \times n$ matrix A, for $i = 1, \ldots, n$, let

$$R'_i(A) = \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i\,j}|$$

and let

$$G(A) = \bigcup_{i=1}^{n} \{z \in \mathbb{C} \mid |z - a_{i\,i}| \leq R'_i(A)\}.$$

Each disc $\{z \in \mathbb{C} \mid |z - a_{i\,i}| \leq R'_i(A)\}$ is called a *Gershgorin disc* and their union $G(A)$ is called the *Gershgorin domain*.

Although easy to prove, the following theorem is very useful:

Theorem 12.8. *(Gershgorin's disc theorem) For any complex $n \times n$ matrix A, all the eigenvalues of A belong to the Gershgorin domain $G(A)$. Furthermore the following properties hold:*

(1) If A is strictly row diagonally dominant, that is

$$|a_{i\,i}| > \sum_{j=1,\, j \neq i}^{n} |a_{i\,j}|, \quad \text{for } i = 1, \ldots, n,$$

then A is invertible.

(2) If A is strictly row diagonally dominant, and if $a_{i\,i} > 0$ for $i = 1, \ldots, n$, then every eigenvalue of A has a strictly positive real part.

Proof. Let λ be any eigenvalue of A and let u be a corresponding eigenvector (recall that we must have $u \neq 0$). Let k be an index such that

$$|u_k| = \max_{1 \leq i \leq n} |u_i|.$$

Since $Au = \lambda u$, we have

$$(\lambda - a_{k\,k})u_k = \sum_{\substack{j=1 \\ j \neq k}}^{n} a_{k\,j} u_j,$$

which implies that

$$|\lambda - a_{k\,k}||u_k| \leq \sum_{\substack{j=1\\ j\neq k}}^{n} |a_{k\,j}||u_j| \leq |u_k| \sum_{\substack{j=1\\ j\neq k}}^{n} |a_{k\,j}|$$

and since $u \neq 0$ and $|u_k| = \max_{1\leq i\leq n} |u_i|$, we must have $|u_k| \neq 0$, and it follows that

$$|\lambda - a_{k\,k}| \leq \sum_{\substack{j=1\\ j\neq k}}^{n} |a_{k\,j}| = R'_k(A),$$

and thus

$$\lambda \in \{z \in \mathbb{C} \mid |z - a_{k\,k}| \leq R'_k(A)\} \subseteq G(A),$$

as claimed.

(1) Strict row diagonal dominance implies that 0 does not belong to any of the Gershgorin discs, so all eigenvalues of A are nonzero, and A is invertible.

(2) If A is strictly row diagonally dominant and $a_{i\,i} > 0$ for $i = 1, \ldots, n$, then each of the Gershgorin discs lies strictly in the right half-plane, so every eigenvalue of A has a strictly positive real part. □

In particular, Theorem 12.8 implies that if a symmetric matrix is strictly row diagonally dominant and has strictly positive diagonal entries, then it is positive definite. Theorem 12.8 is sometimes called the *Gershgorin–Hadamard theorem*.

Since A and $A^\top$ have the same eigenvalues (even for complex matrices) we also have a version of Theorem 12.8 for the discs of radius

$$C'_j(A) = \sum_{\substack{i=1\\ i\neq j}}^{n} |a_{i\,j}|,$$

whose domain is denoted by $G(A^\top)$. Thus we get the following:

Theorem 12.9. *For any complex $n \times n$ matrix A, all the eigenvalues of A belong to the intersection of the Gershgorin domains, $G(A) \cap G(A^\top)$. Furthermore the following properties hold:*

(1) If A is strictly column diagonally dominant, that is

$$|a_{i\,i}| > \sum_{i=1,\, i\neq j}^{n} |a_{i\,j}|, \quad \textit{for } j = 1, \ldots, n,$$

then A is invertible.

(2) If A is strictly column diagonally dominant, and if $a_{ii} > 0$ for $i = 1, \ldots, n$, then every eigenvalue of A has a strictly positive real part.

There are refinements of Gershgorin's theorem and eigenvalue location results involving other domains besides discs; for more on this subject, see Horn and Johnson [55], Sections 6.1 and 6.2.

Remark: Neither strict row diagonal dominance nor strict column diagonal dominance are necessary for invertibility. Also, if we relax all strict inequalities to inequalities, then row diagonal dominance (or column diagonal dominance) is not a sufficient condition for invertibility.

12.4 Summary

The main concepts and results of this chapter are listed below:

- *Diagonal matrix*.
- *Eigenvalues, eigenvectors*; the *eigenspace associated* with an eigenvalue.
- The *characteristic polynomial*.
- The *trace*.
- *algebraic and geometric multiplicity*.
- Eigenspaces associated with distinct eigenvalues form a direct sum (Proposition 12.3).
- Reduction of a matrix to an upper-triangular matrix.
- *Schur decomposition*.
- The *Gershgorin's discs* can be used to locate the eigenvalues of a complex matrix; see Theorems 12.8 and 12.9.

Chapter 13

Spectral Theorems in Euclidean and Hermitian Spaces

13.1 Introduction

The goal of this chapter is to show that there are nice normal forms for symmetric matrices, skew-symmetric matrices, orthogonal matrices, and normal matrices. The spectral theorem for symmetric matrices states that symmetric matrices have real eigenvalues and that they can be diagonalized over an orthonormal basis. The spectral theorem for Hermitian matrices states that Hermitian matrices also have real eigenvalues and that they can be diagonalized over a complex orthonormal basis. Normal real matrices can be block diagonalized over an orthonormal basis with blocks having size at most two, and there are refinements of this normal form for skew-symmetric and orthogonal matrices.

13.2 Normal Linear Maps

We begin by studying normal maps, to understand the structure of their eigenvalues and eigenvectors. This section and the next two were inspired by Lang [62], Artin [6], Mac Lane and Birkhoff [70], Berger [9], and Bertin [12].

Definition 13.1. Given a Euclidean space E, a linear map $f\colon E \to E$ is *normal* if

$$f \circ f^* = f^* \circ f.$$

A linear map $f\colon E \to E$ is *self-adjoint* if $f = f^*$, *skew-self-adjoint* if $f = -f^*$, and *orthogonal* if $f \circ f^* = f^* \circ f = \mathrm{id}$.

Obviously, a self-adjoint, skew-self-adjoint, or orthogonal linear map is a normal linear map. Our first goal is to show that for every normal linear map $f\colon E \to E$, there is an orthonormal basis (w.r.t. $\langle -, - \rangle$) such that the matrix of f over this basis has an especially

nice form: It is a block diagonal matrix in which the blocks are either one-dimensional matrices (i.e., single entries) or two-dimensional matrices of the form

$$\begin{pmatrix} \lambda & \mu \\ -\mu & \lambda \end{pmatrix}.$$

This normal form can be further refined if f is self-adjoint, skew-self-adjoint, or orthogonal. As a first step, we show that f and f^* have the same kernel when f is normal.

Proposition 13.1. *Given a Euclidean space E, if $f\colon E \to E$ is a normal linear map, then* $\operatorname{Ker} f = \operatorname{Ker} f^*$.

Proof. First, let us prove that

$$\langle f(u), f(v)\rangle = \langle f^*(u), f^*(v)\rangle$$

for all $u, v \in E$. Since f^* is the adjoint of f and $f \circ f^* = f^* \circ f$, we have

$$\begin{aligned} \langle f(u), f(u)\rangle &= \langle u, (f^* \circ f)(u)\rangle, \\ &= \langle u, (f \circ f^*)(u)\rangle, \\ &= \langle f^*(u), f^*(u)\rangle. \end{aligned}$$

Since $\langle -, -\rangle$ is positive definite,

$$\begin{aligned} \langle f(u), f(u)\rangle = 0 \quad &\text{iff} \quad f(u) = 0, \\ \langle f^*(u), f^*(u)\rangle = 0 \quad &\text{iff} \quad f^*(u) = 0, \end{aligned}$$

and since

$$\langle f(u), f(u)\rangle = \langle f^*(u), f^*(u)\rangle,$$

we have

$$f(u) = 0 \quad \text{iff} \quad f^*(u) = 0.$$

Consequently, $\operatorname{Ker} f = \operatorname{Ker} f^*$. □

The next step is to show that for every linear map $f\colon E \to E$ there is some subspace W of dimension 1 or 2 such that $f(W) \subseteq W$. When $\dim(W) = 1$, the subspace W is actually an eigenspace for some real eigenvalue of f. Furthermore, when f is normal, there is a subspace W of dimension 1 or 2 such that $f(W) \subseteq W$ and $f^*(W) \subseteq W$. The difficulty is that the eigenvalues of f are not necessarily real. One way to get around this problem is to complexify both the vector space E and the inner product $\langle -, -\rangle$.

Every real vector space E can be embedded into a complex vector space $E_{\mathbb{C}}$, and every linear map $f\colon E \to E$ can be extended to a linear map $f_{\mathbb{C}}\colon E_{\mathbb{C}} \to E_{\mathbb{C}}$.

Definition 13.2. Given a real vector space E, let $E_{\mathbb{C}}$ be the structure $E \times E$ under the addition operation

$$(u_1,\, u_2) + (v_1,\, v_2) = (u_1 + v_1,\, u_2 + v_2),$$

and let multiplication by a complex scalar $z = x + iy$ be defined such that

$$(x + iy) \cdot (u,\, v) = (xu - yv,\, yu + xv).$$

The space $E_{\mathbb{C}}$ is called the *complexification* of E.

It is easily shown that the structure $E_{\mathbb{C}}$ is a complex vector space. It is also immediate that

$$(0,\, v) = i(v,\, 0),$$

and thus, identifying E with the subspace of $E_{\mathbb{C}}$ consisting of all vectors of the form $(u,\, 0)$, we can write

$$(u,\, v) = u + iv.$$

Observe that if $(e_1, \ldots, e_n)$ is a basis of E (a real vector space), then $(e_1, \ldots, e_n)$ is also a basis of $E_{\mathbb{C}}$ (recall that e_i is an abreviation for $(e_i, 0)$).

A linear map $f\colon E \to E$ is extended to the linear map $f_{\mathbb{C}}\colon E_{\mathbb{C}} \to E_{\mathbb{C}}$ defined such that

$$f_{\mathbb{C}}(u + iv) = f(u) + if(v).$$

For any basis $(e_1, \ldots, e_n)$ of E, the matrix $M(f)$ representing f over $(e_1, \ldots, e_n)$ is identical to the matrix $M(f_{\mathbb{C}})$ representing $f_{\mathbb{C}}$ over $(e_1, \ldots, e_n)$, where we view $(e_1, \ldots, e_n)$ as a basis of $E_{\mathbb{C}}$. As a consequence, $\det(zI - M(f)) = \det(zI - M(f_{\mathbb{C}}))$, which means that f and $f_{\mathbb{C}}$ have the same characteristic polynomial (which has real coefficients). We know that every polynomial of degree n with real (or complex) coefficients always has n complex roots (counted with their multiplicity), and the roots of $\det(zI - M(f_{\mathbb{C}}))$ that are real (if any) are the eigenvalues of f.

Next, we need to extend the inner product on E to an inner product on $E_{\mathbb{C}}$.

The inner product $\langle -, - \rangle$ on a Euclidean space E is extended to the Hermitian positive definite form $\langle -, - \rangle_{\mathbb{C}}$ on $E_{\mathbb{C}}$ as follows:

$$\langle u_1 + iv_1, u_2 + iv_2 \rangle_{\mathbb{C}} = \langle u_1, u_2 \rangle + \langle v_1, v_2 \rangle + i(\langle v_1, u_2 \rangle - \langle u_1, v_2 \rangle).$$

It is easily verified that $\langle -, - \rangle_{\mathbb{C}}$ is indeed a Hermitian form that is positive definite, and it is clear that $\langle -, - \rangle_{\mathbb{C}}$ agrees with $\langle -, - \rangle$ on real vectors. Then, given any linear map $f\colon E \to E$, it is easily verified that the map $f^*_{\mathbb{C}}$ defined such that

$$f^*_{\mathbb{C}}(u + iv) = f^*(u) + if^*(v)$$

for all $u, v \in E$ is the adjoint of $f_{\mathbb{C}}$ w.r.t. $\langle -, - \rangle_{\mathbb{C}}$.

Assuming again that E is a Hermitian space, observe that Proposition 13.1 also holds. We deduce the following corollary.

Proposition 13.2. *Given a Hermitian space E, for any normal linear map $f: E \to E$, we have* $\mathrm{Ker}\,(f) \cap \mathrm{Im}(f) = (0)$.

Proof. Assume $v \in \mathrm{Ker}\,(f) \cap \mathrm{Im}(f) = (0)$, which means that $v = f(u)$ for some $u \in E$, and $f(v) = 0$. By Proposition 13.1, $\mathrm{Ker}\,(f) = \mathrm{Ker}\,(f^*)$, so $f(v) = 0$ implies that $f^*(v) = 0$. Consequently,

$$\begin{aligned} 0 &= \langle f^*(v), u\rangle \\ &= \langle v, f(u)\rangle \\ &= \langle v, v\rangle, \end{aligned}$$

and thus, $v = 0$. □

We also have the following crucial proposition relating the eigenvalues of f and f^*.

Proposition 13.3. *Given a Hermitian space E, for any normal linear map $f: E \to E$, a vector u is an eigenvector of f for the eigenvalue λ (in $\mathbb{C}$) iff u is an eigenvector of f^* for the eigenvalue $\overline{\lambda}$.*

Proof. First, it is immediately verified that the adjoint of $f - \lambda\,\mathrm{id}$ is $f^* - \overline{\lambda}\,\mathrm{id}$. Furthermore, $f - \lambda\,\mathrm{id}$ is normal. Indeed,

$$\begin{aligned} (f - \lambda\,\mathrm{id}) \circ (f - \lambda\,\mathrm{id})^* &= (f - \lambda\,\mathrm{id}) \circ (f^* - \overline{\lambda}\,\mathrm{id}), \\ &= f \circ f^* - \overline{\lambda} f - \lambda f^* + \lambda\overline{\lambda}\,\mathrm{id}, \\ &= f^* \circ f - \lambda f^* - \overline{\lambda} f + \overline{\lambda}\lambda\,\mathrm{id}, \\ &= (f^* - \overline{\lambda}\,\mathrm{id}) \circ (f - \lambda\,\mathrm{id}), \\ &= (f - \lambda\,\mathrm{id})^* \circ (f - \lambda\,\mathrm{id}). \end{aligned}$$

Applying Proposition 13.1 to $f - \lambda\,\mathrm{id}$, for every nonnull vector u, we see that

$$(f - \lambda\,\mathrm{id})(u) = 0 \quad \text{iff} \quad (f^* - \overline{\lambda}\,\mathrm{id})(u) = 0,$$

which is exactly the statement of the proposition. □

The next proposition shows a very important property of normal linear maps: Eigenvectors corresponding to distinct eigenvalues are orthogonal.

Proposition 13.4. *Given a Hermitian space E, for any normal linear map $f: E \to E$, if u and v are eigenvectors of f associated with the eigenvalues λ and μ (in $\mathbb{C}$) where $\lambda \neq \mu$, then $\langle u, v\rangle = 0$.*

Proof. Let us compute $\langle f(u), v\rangle$ in two different ways. Since v is an eigenvector of f for μ, by Proposition 13.3, v is also an eigenvector of f^* for $\overline{\mu}$, and we have

$$\langle f(u), v\rangle = \langle \lambda u, v\rangle = \lambda\langle u, v\rangle$$

and

$$\langle f(u), v\rangle = \langle u, f^*(v)\rangle = \langle u, \overline{\mu}v\rangle = \mu\langle u, v\rangle,$$

where the last identity holds because of the semilinearity in the second argument, and thus

$$\lambda\langle u, v\rangle = \mu\langle u, v\rangle,$$

that is,

$$(\lambda - \mu)\langle u, v\rangle = 0,$$

which implies that $\langle u, v\rangle = 0$, since $\lambda \neq \mu$. □

We can also show easily that the eigenvalues of a self-adjoint linear map are real.

Proposition 13.5. *Given a Hermitian space E, all the eigenvalues of any self-adjoint linear map $f\colon E \to E$ are real.*

Proof. Let z (in $\mathbb{C}$) be an eigenvalue of f and let u be an eigenvector for z. We compute $\langle f(u), u\rangle$ in two different ways. We have

$$\langle f(u), u\rangle = \langle zu, u\rangle = z\langle u, u\rangle,$$

and since $f = f^*$, we also have

$$\langle f(u), u\rangle = \langle u, f^*(u)\rangle = \langle u, f(u)\rangle = \langle u, zu\rangle = \overline{z}\langle u, u\rangle.$$

Thus,

$$z\langle u, u\rangle = \overline{z}\langle u, u\rangle,$$

which implies that $z = \overline{z}$, since $u \neq 0$, and z is indeed real. □

There is also a version of Proposition 13.5 for a (real) Euclidean space E and a self-adjoint map $f\colon E \to E$.

Proposition 13.6. *Given a Euclidean space E, if $f\colon E \to E$ is any self-adjoint linear map, then every eigenvalue λ of $f_{\mathbb{C}}$ is real and is actually an eigenvalue of f (which means that there is some real eigenvector $u \in E$ such that $f(u) = \lambda u$). Therefore, all the eigenvalues of f are real.*

Proof. Let $E_{\mathbb{C}}$ be the complexification of E, $\langle -, -\rangle_{\mathbb{C}}$ the complexification of the inner product $\langle -, -\rangle$ on E, and $f_{\mathbb{C}}\colon E_{\mathbb{C}} \to E_{\mathbb{C}}$ the complexification of $f\colon E \to E$. By definition of $f_{\mathbb{C}}$ and $\langle -, -\rangle_{\mathbb{C}}$, if f is self-adjoint, we have

$$\begin{aligned}
\langle f_{\mathbb{C}}(u_1 + iv_1), u_2 + iv_2\rangle_{\mathbb{C}} &= \langle f(u_1) + if(v_1), u_2 + iv_2\rangle_{\mathbb{C}} \\
&= \langle f(u_1), u_2\rangle + \langle f(v_1), v_2\rangle + i(\langle u_2, f(v_1)\rangle - \langle f(u_1), v_2\rangle) \\
&= \langle u_1, f(u_2)\rangle + \langle v_1, f(v_2)\rangle + i(\langle f(u_2), v_1\rangle - \langle u_1, f(v_2)\rangle) \\
&= \langle u_1 + iv_1, f(u_2) + if(v_2)\rangle_{\mathbb{C}} \\
&= \langle u_1 + iv_1, f_{\mathbb{C}}(u_2 + iv_2)\rangle_{\mathbb{C}},
\end{aligned}$$

which shows that $f_{\mathbb{C}}$ is also self-adjoint with respect to $\langle -, - \rangle_{\mathbb{C}}$.

As we pointed out earlier, f and $f_{\mathbb{C}}$ have the same characteristic polynomial $\det(zI - f_{\mathbb{C}}) = \det(zI - f)$, which is a polynomial with real coefficients. Proposition 13.5 shows that the zeros of $\det(zI - f_{\mathbb{C}}) = \det(zI - f)$ are all real, and for each real zero λ of $\det(zI - f)$, the linear map $\lambda \mathrm{id} - f$ is singular, which means that there is some nonzero $u \in E$ such that $f(u) = \lambda u$. Therefore, all the eigenvalues of f are real. □

Given any subspace W of a Euclidean space E, recall that the *orthogonal complement* $W^{\perp}$ of W is the subspace defined such that

$$W^{\perp} = \{u \in E \mid \langle u, w \rangle = 0, \text{ for all } w \in W\}.$$

Recall from Proposition 9.9 that $E = W \oplus W^{\perp}$ (this can be easily shown, for example, by constructing an orthonormal basis of E using the Gram–Schmidt orthonormalization procedure). The same result also holds for Hermitian spaces; see Proposition 11.10.

As a warm up for the proof of Theorem 13.10, let us prove that every self-adjoint map on a Euclidean space can be diagonalized with respect to an orthonormal basis of eigenvectors.

Theorem 13.7. *(Spectral theorem for self-adjoint linear maps on a Euclidean space) Given a Euclidean space E of dimension n, for every self-adjoint linear map $f \colon E \to E$, there is an orthonormal basis $(e_1, \ldots, e_n)$ of eigenvectors of f such that the matrix of f w.r.t. this basis is a diagonal matrix*

$$\begin{pmatrix} \lambda_1 & & \cdots & \\ & \lambda_2 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \cdots & \lambda_n \end{pmatrix},$$

with $\lambda_i \in \mathbb{R}$.

Proof. We proceed by induction on the dimension n of E as follows. If $n = 1$, the result is trivial. Assume now that $n \geq 2$. From Proposition 13.6, all the eigenvalues of f are real, so pick some eigenvalue $\lambda \in \mathbb{R}$, and let w be some eigenvector for λ. By dividing w by its norm, we may assume that w is a unit vector. Let W be the subspace of dimension 1 spanned by w. Clearly, $f(W) \subseteq W$. We claim that $f(W^{\perp}) \subseteq W^{\perp}$, where $W^{\perp}$ is the orthogonal complement of W.

Indeed, for any $v \in W^{\perp}$, that is, if $\langle v, w \rangle = 0$, because f is self-adjoint and $f(w) = \lambda w$, we have

$$\begin{aligned} \langle f(v), w \rangle &= \langle v, f(w) \rangle \\ &= \langle v, \lambda w \rangle \\ &= \lambda \langle v, w \rangle = 0 \end{aligned}$$

since $\langle v, w\rangle = 0$. Therefore,

$$f(W^\perp) \subseteq W^\perp.$$

Clearly, the restriction of f to $W^\perp$ is self-adjoint, and we conclude by applying the induction hypothesis to $W^\perp$ (whose dimension is $n-1$). □

We now come back to normal linear maps. One of the key points in the proof of Theorem 13.7 is that we found a subspace W with the property that $f(W) \subseteq W$ implies that $f(W^\perp) \subseteq W^\perp$. In general, this does not happen, but normal maps satisfy a stronger property which ensures that such a subspace exists.

The following proposition provides a condition that will allow us to show that a normal linear map can be diagonalized. It actually holds for any linear map. We found the inspiration for this proposition in Berger [9].

Proposition 13.8. *Given a Hermitian space E, for any linear map $f\colon E \to E$ and any subspace W of E, if $f(W) \subseteq W$, then $f^*\big(W^\perp\big) \subseteq W^\perp$. Consequently, if $f(W) \subseteq W$ and $f^*(W) \subseteq W$, then $f\big(W^\perp\big) \subseteq W^\perp$ and $f^*\big(W^\perp\big) \subseteq W^\perp$.*

Proof. If $u \in W^\perp$, then

$$\langle w, u\rangle = 0 \quad \text{for all } w \in W.$$

However,

$$\langle f(w), u\rangle = \langle w, f^*(u)\rangle,$$

and $f(W) \subseteq W$ implies that $f(w) \in W$. Since $u \in W^\perp$, we get

$$0 = \langle f(w), u\rangle = \langle w, f^*(u)\rangle,$$

which shows that $\langle w, f^*(u)\rangle = 0$ for all $w \in W$, that is, $f^*(u) \in W^\perp$. Therefore, we have $f^*(W^\perp) \subseteq W^\perp$.

We just proved that if $f(W) \subseteq W$, then $f^*\big(W^\perp\big) \subseteq W^\perp$. If we also have $f^*(W) \subseteq W$, then by applying the above fact to f^*, we get $f^{**}(W^\perp) \subseteq W^\perp$, and since $f^{**} = f$, this is just $f(W^\perp) \subseteq W^\perp$, which proves the second statement of the proposition. □

It is clear that the above proposition also holds for Euclidean spaces.

Although we are ready to prove that for every normal linear map f (over a Hermitian space) there is an orthonormal basis of eigenvectors (see Theorem 13.11 below), we now return to real Euclidean spaces.

If $f\colon E \to E$ is a linear map and $w = u + iv$ is an eigenvector of $f_{\mathbb{C}}\colon E_{\mathbb{C}} \to E_{\mathbb{C}}$ for the eigenvalue $z = \lambda + i\mu$, where $u, v \in E$ and $\lambda, \mu \in \mathbb{R}$, since

$$f_{\mathbb{C}}(u + iv) = f(u) + if(v)$$

and

$$f_{\mathbb{C}}(u + iv) = (\lambda + i\mu)(u + iv) = \lambda u - \mu v + i(\mu u + \lambda v),$$

we have

$$f(u) = \lambda u - \mu v \quad \text{and} \quad f(v) = \mu u + \lambda v,$$

from which we immediately obtain

$$f_{\mathbb{C}}(u - iv) = (\lambda - i\mu)(u - iv),$$

which shows that $\overline{w} = u - iv$ is an eigenvector of $f_{\mathbb{C}}$ for $\overline{z} = \lambda - i\mu$. Using this fact, we can prove the following proposition.

Proposition 13.9. *Given a Euclidean space E, for any normal linear map $f\colon E \to E$, if $w = u + iv$ is an eigenvector of $f_{\mathbb{C}}$ associated with the eigenvalue $z = \lambda + i\mu$ (where $u, v \in E$ and $\lambda, \mu \in \mathbb{R}$), if $\mu \neq 0$ (i.e., z is not real) then $\langle u, v\rangle = 0$ and $\langle u, u\rangle = \langle v, v\rangle$, which implies that u and v are linearly independent, and if W is the subspace spanned by u and v, then $f(W) = W$ and $f^*(W) = W$. Furthermore, with respect to the (orthogonal) basis (u, v), the restriction of f to W has the matrix*

$$\begin{pmatrix} \lambda & \mu \\ -\mu & \lambda \end{pmatrix}.$$

If $\mu = 0$, then λ is a real eigenvalue of f, and either u or v is an eigenvector of f for λ. If W is the subspace spanned by u if $u \neq 0$, or spanned by $v \neq 0$ if $u = 0$, then $f(W) \subseteq W$ and $f^*(W) \subseteq W$.

Proof. Since $w = u + iv$ is an eigenvector of $f_{\mathbb{C}}$, by definition it is nonnull, and either $u \neq 0$ or $v \neq 0$. From the fact stated just before Proposition 13.9, $u - iv$ is an eigenvector of $f_{\mathbb{C}}$ for $\lambda - i\mu$. It is easy to check that $f_{\mathbb{C}}$ is normal. However, if $\mu \neq 0$, then $\lambda + i\mu \neq \lambda - i\mu$, and from Proposition 13.4, the vectors $u + iv$ and $u - iv$ are orthogonal w.r.t. $\langle -, - \rangle_{\mathbb{C}}$, that is,

$$\langle u + iv, u - iv\rangle_{\mathbb{C}} = \langle u, u\rangle - \langle v, v\rangle + 2i\langle u, v\rangle = 0.$$

Thus, we get $\langle u, v\rangle = 0$ and $\langle u, u\rangle = \langle v, v\rangle$, and since $u \neq 0$ or $v \neq 0$, u and v are linearly independent. Since

$$f(u) = \lambda u - \mu v \quad \text{and} \quad f(v) = \mu u + \lambda v$$

and since by Proposition 13.3 $u + iv$ is an eigenvector of $f^*_{\mathbb{C}}$ for $\lambda - i\mu$, we have

$$f^*(u) = \lambda u + \mu v \quad \text{and} \quad f^*(v) = -\mu u + \lambda v,$$

and thus $f(W) = W$ and $f^*(W) = W$, where W is the subspace spanned by u and v.

When $\mu = 0$, we have

$$f(u) = \lambda u \quad \text{and} \quad f(v) = \lambda v,$$

and since $u \neq 0$ or $v \neq 0$, either u or v is an eigenvector of f for λ. If W is the subspace spanned by u if $u \neq 0$, or spanned by v if $u = 0$, it is obvious that $f(W) \subseteq W$ and $f^*(W) \subseteq W$. Note that $\lambda = 0$ is possible, and this is why $\subseteq$ cannot be replaced by $=$. □

The beginning of the proof of Proposition 13.9 actually shows that for every linear map $f: E \to E$ there is some subspace W such that $f(W) \subseteq W$, where W has dimension 1 or 2. In general, it doesn't seem possible to prove that $W^\perp$ is invariant under f. However, this happens when f is normal.

We can finally prove our first main theorem.

Theorem 13.10. *(Main spectral theorem) Given a Euclidean space E of dimension n, for every normal linear map $f: E \to E$, there is an orthonormal basis $(e_1, \ldots, e_n)$ such that the matrix of f w.r.t. this basis is a block diagonal matrix of the form*

$$\begin{pmatrix} A_1 & & \cdots & \\ & A_2 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \cdots & A_p \end{pmatrix}$$

such that each block A_j is either a one-dimensional matrix (i.e., a real scalar) or a two-dimensional matrix of the form

$$A_j = \begin{pmatrix} \lambda_j & -\mu_j \\ \mu_j & \lambda_j \end{pmatrix},$$

where $\lambda_j, \mu_j \in \mathbb{R}$, with $\mu_j > 0$.

Proof. We proceed by induction on the dimension n of E as follows. If $n = 1$, the result is trivial. Assume now that $n \geq 2$. First, since $\mathbb{C}$ is algebraically closed (i.e., every polynomial has a root in $\mathbb{C}$), the linear map $f_{\mathbb{C}}: E_{\mathbb{C}} \to E_{\mathbb{C}}$ has some eigenvalue $z = \lambda + i\mu$ (where $\lambda, \mu \in \mathbb{R}$). Let $w = u + iv$ be some eigenvector of $f_{\mathbb{C}}$ for $\lambda + i\mu$ (where $u, v \in E$). We can now apply Proposition 13.9.

If $\mu = 0$, then either u or v is an eigenvector of f for $\lambda \in \mathbb{R}$. Let W be the subspace of dimension 1 spanned by $e_1 = u/\|u\|$ if $u \neq 0$, or by $e_1 = v/\|v\|$ otherwise. It is obvious that $f(W) \subseteq W$ and $f^*(W) \subseteq W$. The orthogonal $W^\perp$ of W has dimension $n - 1$, and by Proposition 13.8, we have $f\big(W^\perp\big) \subseteq W^\perp$. But the restriction of f to $W^\perp$ is also normal, and we conclude by applying the induction hypothesis to $W^\perp$.

If $\mu \neq 0$, then $\langle u, v\rangle = 0$ and $\langle u, u\rangle = \langle v, v\rangle$, and if W is the subspace spanned by $u/\|u\|$ and $v/\|v\|$, then $f(W) = W$ and $f^*(W) = W$. We also know that the restriction of f to W has the matrix

$$\begin{pmatrix} \lambda & \mu \\ -\mu & \lambda \end{pmatrix}$$

with respect to the basis $(u/\|u\|, v/\|v\|)$. If $\mu < 0$, we let $\lambda_1 = \lambda$, $\mu_1 = -\mu$, $e_1 = u/\|u\|$, and $e_2 = v/\|v\|$. If $\mu > 0$, we let $\lambda_1 = \lambda$, $\mu_1 = \mu$, $e_1 = v/\|v\|$, and $e_2 = u/\|u\|$. In all cases, it is easily verified that the matrix of the restriction of f to W w.r.t. the orthonormal basis (e_1, e_2) is

$$A_1 = \begin{pmatrix} \lambda_1 & -\mu_1 \\ \mu_1 & \lambda_1 \end{pmatrix},$$

where $\lambda_1, \mu_1 \in \mathbb{R}$, with $\mu_1 > 0$. However, $W^\perp$ has dimension $n-2$, and by Proposition 13.8, $f(W^\perp) \subseteq W^\perp$. Since the restriction of f to $W^\perp$ is also normal, we conclude by applying the induction hypothesis to $W^\perp$. □

After this relatively hard work, we can easily obtain some nice normal forms for the matrices of self-adjoint, skew-self-adjoint, and orthogonal linear maps. However, for the sake of completeness (and since we have all the tools to so do), we go back to the case of a Hermitian space and show that normal linear maps can be diagonalized with respect to an orthonormal basis. The proof is a slight generalization of the proof of Theorem 13.6.

Theorem 13.11. *(Spectral theorem for normal linear maps on a Hermitian space) Given a Hermitian space E of dimension n, for every normal linear map $f: E \to E$ there is an orthonormal basis $(e_1, \ldots, e_n)$ of eigenvectors of f such that the matrix of f w.r.t. this basis is a diagonal matrix*

$$\begin{pmatrix} \lambda_1 & & \cdots & \\ & \lambda_2 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \cdots & \lambda_n \end{pmatrix},$$

where $\lambda_j \in \mathbb{C}$.

Proof. We proceed by induction on the dimension n of E as follows. If $n = 1$, the result is trivial. Assume now that $n \geq 2$. Since $\mathbb{C}$ is algebraically closed (i.e., every polynomial has a root in $\mathbb{C}$), the linear map $f: E \to E$ has some eigenvalue $\lambda \in \mathbb{C}$, and let w be some unit eigenvector for λ. Let W be the subspace of dimension 1 spanned by w. Clearly, $f(W) \subseteq W$. By Proposition 13.3, w is an eigenvector of f^* for $\overline{\lambda}$, and thus $f^*(W) \subseteq W$. By Proposition 13.8, we also have $f(W^\perp) \subseteq W^\perp$. The restriction of f to $W^\perp$ is still normal, and we conclude by applying the induction hypothesis to $W^\perp$ (whose dimension is $n-1$). □

Thus, in particular, self-adjoint, skew-self-adjoint, and orthogonal linear maps can be diagonalized with respect to an orthonormal basis of eigenvectors. In this latter case, though, an orthogonal map is called a *unitary* map. Also, Proposition 13.5 shows that the eigenvalues of a self-adjoint linear map are real. It is easily shown that skew-self-adjoint maps have eigenvalues that are pure imaginary or null, and that unitary maps have eigenvalues of absolute value 1.

Remark: There is a converse to Theorem 13.11, namely, if there is an orthonormal basis $(e_1, \ldots, e_n)$ of eigenvectors of f, then f is normal. We leave the easy proof as an exercise.

13.3 Self-Adjoint, Skew-Self-Adjoint, and Orthogonal Linear Maps

We begin with self-adjoint maps.

Theorem 13.12. *Given a Euclidean space E of dimension n, for every self-adjoint linear map $f: E \to E$, there is an orthonormal basis $(e_1, \ldots, e_n)$ of eigenvectors of f such that the matrix of f w.r.t. this basis is a diagonal matrix*

$$\begin{pmatrix} \lambda_1 & & \cdots & \\ & \lambda_2 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \cdots & \lambda_n \end{pmatrix},$$

where $\lambda_i \in \mathbb{R}$.

Proof. We already proved this; see Theorem 13.6. However, it is instructive to give a more direct method not involving the complexification of $\langle -, - \rangle$ and Proposition 13.5.

Since $\mathbb{C}$ is algebraically closed, $f_{\mathbb{C}}$ has some eigenvalue $\lambda + i\mu$, and let $u + iv$ be some eigenvector of $f_{\mathbb{C}}$ for $\lambda + i\mu$, where $\lambda, \mu \in \mathbb{R}$ and $u, v \in E$. We saw in the proof of Proposition 13.9 that

$$f(u) = \lambda u - \mu v \quad \text{and} \quad f(v) = \mu u + \lambda v.$$

Since $f = f^*$,

$$\langle f(u), v \rangle = \langle u, f(v) \rangle$$

for all $u, v \in E$. Applying this to

$$f(u) = \lambda u - \mu v \quad \text{and} \quad f(v) = \mu u + \lambda v,$$

we get

$$\langle f(u), v \rangle = \langle \lambda u - \mu v, v \rangle = \lambda \langle u, v \rangle - \mu \langle v, v \rangle$$

and

$$\langle u, f(v) \rangle = \langle u, \mu u + \lambda v \rangle = \mu \langle u, u \rangle + \lambda \langle u, v \rangle,$$

and thus we get

$$\lambda \langle u, v \rangle - \mu \langle v, v \rangle = \mu \langle u, u \rangle + \lambda \langle u, v \rangle,$$

that is,

$$\mu(\langle u, u \rangle + \langle v, v \rangle) = 0,$$

which implies $\mu = 0$, since either $u \neq 0$ or $v \neq 0$. Therefore, λ is a real eigenvalue of f.

Now, going back to the proof of Theorem 13.10, only the case where $\mu = 0$ applies, and the induction shows that all the blocks are one-dimensional. □

Theorem 13.12 implies that if $\lambda_1, \ldots, \lambda_p$ are the distinct real eigenvalues of f, and E_i is the eigenspace associated with λ_i, then

$$E = E_1 \oplus \cdots \oplus E_p,$$

where E_i and E_j are orthogonal for all $i \neq j$.

Remark: Another way to prove that a self-adjoint map has a real eigenvalue is to use a little bit of calculus. We learned such a proof from Herman Gluck. The idea is to consider the real-valued function $\Phi\colon E \to \mathbb{R}$ defined such that

$$\Phi(u) = \langle f(u), u\rangle$$

for every $u \in E$. This function is C^∞, and if we represent f by a matrix A over some orthonormal basis, it is easy to compute the gradient vector

$$\nabla\Phi(X) = \left(\frac{\partial\Phi}{\partial x_1}(X), \ldots, \frac{\partial\Phi}{\partial x_n}(X)\right)$$

of Φ at X. Indeed, we find that

$$\nabla\Phi(X) = (A + A^\top)X,$$

where X is a column vector of size n. But since f is self-adjoint, $A = A^\top$, and thus

$$\nabla\Phi(X) = 2AX.$$

The next step is to find the maximum of the function Φ on the sphere

$$S^{n-1} = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid x_1^2 + \cdots + x_n^2 = 1\}.$$

Since S^{n-1} is compact and Φ is continuous, and in fact C^∞, Φ takes a maximum at some X on S^{n-1}. But then it is well known that at an extremum X of Φ we must have

$$d\Phi_X(Y) = \langle\nabla\Phi(X), Y\rangle = 0$$

for all tangent vectors Y to S^{n-1} at X, and so $\nabla\Phi(X)$ is orthogonal to the tangent plane at X, which means that

$$\nabla\Phi(X) = \lambda X$$

for some $\lambda \in \mathbb{R}$. Since $\nabla\Phi(X) = 2AX$, we get

$$2AX = \lambda X,$$

and thus $\lambda/2$ is a real eigenvalue of A (i.e., of f).

Next, we consider skew-self-adjoint maps.

Theorem 13.13. *Given a Euclidean space E of dimension n, for every skew-self-adjoint linear map $f\colon E \to E$ there is an orthonormal basis $(e_1, \ldots, e_n)$ such that the matrix of f w.r.t. this basis is a block diagonal matrix of the form*

$$\begin{pmatrix} A_1 & & \cdots & \\ & A_2 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \cdots & A_p \end{pmatrix}$$

such that each block A_j is either 0 or a two-dimensional matrix of the form

$$A_j = \begin{pmatrix} 0 & -\mu_j \\ \mu_j & 0 \end{pmatrix},$$

where $\mu_j \in \mathbb{R}$, with $\mu_j > 0$. In particular, the eigenvalues of $f_{\mathbb{C}}$ are pure imaginary of the form $\pm i\mu_j$ or 0.

Proof. The case where $n = 1$ is trivial. As in the proof of Theorem 13.10, $f_{\mathbb{C}}$ has some eigenvalue $z = \lambda + i\mu$, where $\lambda, \mu \in \mathbb{R}$. We claim that $\lambda = 0$. First, we show that

$$\langle f(w), w\rangle = 0$$

for all $w \in E$. Indeed, since $f = -f^*$, we get

$$\langle f(w), w\rangle = \langle w, f^*(w)\rangle = \langle w, -f(w)\rangle = -\langle w, f(w)\rangle = -\langle f(w), w\rangle,$$

since $\langle -, -\rangle$ is symmetric. This implies that

$$\langle f(w), w\rangle = 0.$$

Applying this to u and v and using the fact that

$$f(u) = \lambda u - \mu v \quad \text{and} \quad f(v) = \mu u + \lambda v,$$

we get

$$0 = \langle f(u), u\rangle = \langle \lambda u - \mu v, u\rangle = \lambda\langle u, u\rangle - \mu\langle u, v\rangle$$

and

$$0 = \langle f(v), v\rangle = \langle \mu u + \lambda v, v\rangle = \mu\langle u, v\rangle + \lambda\langle v, v\rangle,$$

from which, by addition, we get

$$\lambda(\langle v, v\rangle + \langle v, v\rangle) = 0.$$

Since $u \neq 0$ or $v \neq 0$, we have $\lambda = 0$.

Then, going back to the proof of Theorem 13.10, unless $\mu = 0$, the case where u and v are orthogonal and span a subspace of dimension 2 applies, and the induction shows that all the blocks are two-dimensional or reduced to 0. □

Remark: One will note that if f is skew-self-adjoint, then $if_{\mathbb{C}}$ is self-adjoint w.r.t. $\langle -, - \rangle_{\mathbb{C}}$. By Proposition 13.5, the map $if_{\mathbb{C}}$ has real eigenvalues, which implies that the eigenvalues of $f_{\mathbb{C}}$ are pure imaginary or 0.

Finally, we consider orthogonal linear maps.

Theorem 13.14. *Given a Euclidean space E of dimension n, for every orthogonal linear map $f\colon E \to E$ there is an orthonormal basis $(e_1, \ldots, e_n)$ such that the matrix of f w.r.t. this basis is a block diagonal matrix of the form*

$$\begin{pmatrix} A_1 & & \cdots & \\ & A_2 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \cdots & A_p \end{pmatrix}$$

such that each block A_j is either 1, -1, or a two-dimensional matrix of the form

$$A_j = \begin{pmatrix} \cos\theta_j & -\sin\theta_j \\ \sin\theta_j & \cos\theta_j \end{pmatrix}$$

where $0 < \theta_j < \pi$. In particular, the eigenvalues of $f_{\mathbb{C}}$ are of the form $\cos\theta_j \pm i \sin\theta_j$, 1, or -1.

Proof. The case where $n = 1$ is trivial. As in the proof of Theorem 13.10, $f_{\mathbb{C}}$ has some eigenvalue $z = \lambda + i\mu$, where $\lambda, \mu \in \mathbb{R}$. It is immediately verified that $f \circ f^* = f^* \circ f = \mathrm{id}$ implies that $f_{\mathbb{C}} \circ f_{\mathbb{C}}^* = f_{\mathbb{C}}^* \circ f_{\mathbb{C}} = \mathrm{id}$, so the map $f_{\mathbb{C}}$ is unitary. In fact, the eigenvalues of $f_{\mathbb{C}}$ have absolute value 1. Indeed, if z (in $\mathbb{C}$) is an eigenvalue of $f_{\mathbb{C}}$, and u is an eigenvector for z, we have

$$\langle f_{\mathbb{C}}(u), f_{\mathbb{C}}(u) \rangle = \langle zu, zu \rangle = z\overline{z}\langle u, u \rangle$$

and

$$\langle f_{\mathbb{C}}(u), f_{\mathbb{C}}(u) \rangle = \langle u, (f_{\mathbb{C}}^* \circ f_{\mathbb{C}})(u) \rangle = \langle u, u \rangle,$$

from which we get

$$z\overline{z}\langle u, u \rangle = \langle u, u \rangle.$$

Since $u \neq 0$, we have $z\overline{z} = 1$, i.e., $|z| = 1$. As a consequence, the eigenvalues of $f_{\mathbb{C}}$ are of the form $\cos\theta \pm i \sin\theta$, 1, or -1. The theorem then follows immediately from Theorem 13.10, where the condition $\mu > 0$ implies that $\sin\theta_j > 0$, and thus, $0 < \theta_j < \pi$. □

It is obvious that we can reorder the orthonormal basis of eigenvectors given by Theorem 13.14, so that the matrix of f w.r.t. this basis is a block diagonal matrix of the form

$$\begin{pmatrix} A_1 & \cdots & & & \\ \vdots & \ddots & \vdots & & \vdots \\ & \cdots & A_r & & \\ & & & -I_q & \\ \cdots & & & & I_p \end{pmatrix}$$

where each block A_j is a two-dimensional rotation matrix $A_j \neq \pm I_2$ of the form

$$A_j = \begin{pmatrix} \cos\theta_j & -\sin\theta_j \\ \sin\theta_j & \cos\theta_j \end{pmatrix}$$

with $0 < \theta_j < \pi$.

The linear map f has an eigenspace $E(1, f) = \mathrm{Ker}\,(f - \mathrm{id})$ of dimension p for the eigenvalue 1, and an eigenspace $E(-1, f) = \mathrm{Ker}\,(f + \mathrm{id})$ of dimension q for the eigenvalue -1. If $\det(f) = +1$ (f is a rotation), the dimension q of $E(-1, f)$ must be even, and the entries in $-I_q$ can be paired to form two-dimensional blocks, if we wish. In this case, every rotation in $\mathbf{SO}(n)$ has a matrix of the form

$$\begin{pmatrix} A_1 & \cdots & & \\ \vdots & \ddots & \vdots & \\ & \cdots & A_m & \\ \cdots & & & I_{n-2m} \end{pmatrix}$$

where the first m blocks A_j are of the form

$$A_j = \begin{pmatrix} \cos\theta_j & -\sin\theta_j \\ \sin\theta_j & \cos\theta_j \end{pmatrix}$$

with $0 < \theta_j \leq \pi$.

Theorem 13.14 can be used to prove a version of the Cartan–Dieudonné theorem.

Theorem 13.15. *Let E be a Euclidean space of dimension $n \geq 2$. For every isometry $f \in \mathbf{O}(E)$, if $p = \dim(E(1, f)) = \dim(\mathrm{Ker}\,(f - \mathrm{id}))$, then f is the composition of $n - p$ reflections, and $n - p$ is minimal.*

Proof. From Theorem 13.14 there are r subspaces $F_1, \ldots, F_r$, each of dimension 2, such that

$$E = E(1, f) \oplus E(-1, f) \oplus F_1 \oplus \cdots \oplus F_r,$$

and all the summands are pairwise orthogonal. Furthermore, the restriction r_i of f to each F_i is a rotation $r_i \neq \pm\mathrm{id}$. Each 2D rotation r_i can be written a the composition $r_i = s'_i \circ s_i$ of two reflections s_i and s'_i about lines in F_i (forming an angle $\theta_i/2$). We can extend s_i and s'_i to hyperplane reflections in E by making them the identity on $F_i^{\perp}$. Then,

$$s'_r \circ s_r \circ \cdots \circ s'_1 \circ s_1$$

agrees with f on $F_1 \oplus \cdots \oplus F_r$ and is the identity on $E(1, f) \oplus E(-1, f)$. If $E(-1, f)$ has an orthonormal basis of eigenvectors $(v_1, \ldots, v_q)$, letting s''_j be the reflection about the hyperplane $(v_j)^{\perp}$, it is clear that

$$s''_q \circ \cdots \circ s''_1$$

agrees with f on $E(-1, f)$ and is the identity on $E(1, f) \oplus F_1 \oplus \cdots \oplus F_r$. But then,

$$f = s''_q \circ \cdots \circ s''_1 \circ s'_r \circ s_r \circ \cdots \circ s'_1 \circ s_1,$$

the composition of $2r + q = n - p$ reflections.

If

$$f = s_t \circ \cdots \circ s_1,$$

for t reflections s_i, it is clear that

$$F = \bigcap_{i=1}^{t} E(1, s_i) \subseteq E(1, f),$$

where $E(1, s_i)$ is the hyperplane defining the reflection s_i. By the Grassmann relation, if we intersect $t \leq n$ hyperplanes, the dimension of their intersection is at least $n - t$. Thus, $n - t \leq p$, that is, $t \geq n - p$, and $n - p$ is the smallest number of reflections composing f. □

As a corollary of Theorem 13.15, we obtain the following fact: If the dimension n of the Euclidean space E is odd, then every rotation $f \in \mathbf{SO}(E)$ admits 1 has an eigenvalue.

Proof. The characteristic polynomial $\det(XI - f)$ of f has odd degree n and has real coefficients, so it must have some real root λ. Since f is an isometry, its n eigenvalues are of the form, $+1, -1$, and $e^{\pm i\theta}$, with $0 < \theta < \pi$, so $\lambda = \pm 1$. Now, the eigenvalues $e^{\pm i\theta}$ appear in conjugate pairs, and since n is odd, the number of real eigenvalues of f is odd. This implies that $+1$ is an eigenvalue of f, since otherwise -1 would be the only real eigenvalue of f, and since its multiplicity is odd, we would have $\det(f) = -1$, contradicting the fact that f is a rotation. □

When $n = 3$, we obtain the result due to Euler which says that every 3D rotation R has an invariant axis D, and that restricted to the plane orthogonal to D, it is a 2D rotation. Furthermore, if (a, b, c) is a unit vector defining the axis D of the rotation R and if the angle of the rotation is θ, if B is the skew-symmetric matrix

$$B = \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix},$$

then it can be shown that

$$R = I + \sin\theta B + (1 - \cos\theta)B^2.$$

The theorems of this section and of the previous section can be immediately applied to matrices.

13.4 Normal and Other Special Matrices

First, we consider real matrices. Recall the following definitions.

Definition 13.3. Given a real $m \times n$ matrix A, the *transpose* $A^\top$ *of* A is the $n \times m$ matrix $A^\top = (a^\top_{ij})$ defined such that

$$a^\top_{ij} = a_{ji}$$

for all i, j, $1 \leq i \leq m$, $1 \leq j \leq n$. A real $n \times n$ matrix A is

- *normal* if
$$A\,A^\top = A^\top A,$$
- *symmetric* if
$$A^\top = A,$$
- *skew-symmetric* if
$$A^\top = -A,$$
- *orthogonal* if
$$A\,A^\top = A^\top A = I_n.$$

Recall from Proposition 9.12 that when E is a Euclidean space and $(e_1, \ldots, e_n)$ is an orthonormal basis for E, if A is the matrix of a linear map $f\colon E \to E$ w.r.t. the basis $(e_1, \ldots, e_n)$, then $A^\top$ is the matrix of the adjoint f^* of f. Consequently, a normal linear map has a normal matrix, a self-adjoint linear map has a symmetric matrix, a skew-self-adjoint linear map has a skew-symmetric matrix, and an orthogonal linear map has an orthogonal matrix. Similarly, if E and F are Euclidean spaces, $(u_1, \ldots, u_n)$ is an orthonormal basis for E, and $(v_1, \ldots, v_m)$ is an orthonormal basis for F, if a linear map $f\colon E \to F$ has the matrix A w.r.t. the bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_m)$, then its adjoint f^* has the matrix $A^\top$ w.r.t. the bases $(v_1, \ldots, v_m)$ and $(u_1, \ldots, u_n)$.

Furthermore, if $(u_1, \ldots, u_n)$ is another orthonormal basis for E and P is the change of basis matrix whose columns are the components of the u_i w.r.t. the basis $(e_1, \ldots, e_n)$, then P is orthogonal, and for any linear map $f\colon E \to E$, if A is the matrix of f w.r.t $(e_1, \ldots, e_n)$ and B is the matrix of f w.r.t. $(u_1, \ldots, u_n)$, then

$$B = P^\top A P.$$

As a consequence, Theorems 13.10 and 13.12–13.14 can be restated as follows.

Theorem 13.16. *For every normal matrix A there is an orthogonal matrix P and a block diagonal matrix D such that $A = PD\,P^\top$, where D is of the form*

$$D = \begin{pmatrix} D_1 & & \dots & \\ & D_2 & \dots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \dots & D_p \end{pmatrix}$$

such that each block D_j is either a one-dimensional matrix (i.e., a real scalar) or a two-dimensional matrix of the form

$$D_j = \begin{pmatrix} \lambda_j & -\mu_j \\ \mu_j & \lambda_j \end{pmatrix},$$

where $\lambda_j, \mu_j \in \mathbb{R}$, with $\mu_j > 0$.

Theorem 13.17. *For every symmetric matrix A there is an orthogonal matrix P and a diagonal matrix D such that $A = PD\,P^\top$, where D is of the form*

$$D = \begin{pmatrix} \lambda_1 & & \dots & \\ & \lambda_2 & \dots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \dots & \lambda_n \end{pmatrix},$$

where $\lambda_i \in \mathbb{R}$.

Theorem 13.18. *For every skew-symmetric matrix A there is an orthogonal matrix P and a block diagonal matrix D such that $A = PD\,P^\top$, where D is of the form*

$$D = \begin{pmatrix} D_1 & & \dots & \\ & D_2 & \dots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \dots & D_p \end{pmatrix}$$

such that each block D_j is either 0 or a two-dimensional matrix of the form

$$D_j = \begin{pmatrix} 0 & -\mu_j \\ \mu_j & 0 \end{pmatrix},$$

where $\mu_j \in \mathbb{R}$, with $\mu_j > 0$. In particular, the eigenvalues of A are pure imaginary of the form $\pm i\mu_j$, or 0.

Theorem 13.19. *For every orthogonal matrix A there is an orthogonal matrix P and a block diagonal matrix D such that $A = PD\,P^\top$, where D is of the form*

$$D = \begin{pmatrix} D_1 & & \dots & \\ & D_2 & \dots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \dots & D_p \end{pmatrix}$$

such that each block D_j is either 1, -1, *or a two-dimensional matrix of the form*

$$D_j = \begin{pmatrix} \cos\theta_j & -\sin\theta_j \\ \sin\theta_j & \cos\theta_j \end{pmatrix}$$

where $0 < \theta_j < \pi$. *In particular, the eigenvalues of* A *are of the form* $\cos\theta_j \pm i \sin\theta_j$, 1, *or* -1.

We now consider complex matrices.

Definition 13.4. Given a complex $m \times n$ matrix A, the *transpose* $A^\top$ *of* A is the $n \times m$ matrix $A^\top = \left(a^\top_{i\,j}\right)$ defined such that

$$a^\top_{i\,j} = a_{j\,i}$$

for all i, j, $1 \leq i \leq m$, $1 \leq j \leq n$. The *conjugate* $\overline{A}$ *of* A is the $m \times n$ matrix $\overline{A} = (b_{i\,j})$ defined such that

$$b_{i\,j} = \overline{a}_{i\,j}$$

for all i, j, $1 \leq i \leq m$, $1 \leq j \leq n$. Given an $m \times n$ complex matrix A, the *adjoint* A^* *of* A is the matrix defined such that

$$A^* = \overline{(A^\top)} = (\overline{A})^\top.$$

A complex $n \times n$ matrix A is

- *normal* if
$$AA^* = A^*A,$$
- *Hermitian* if
$$A^* = A,$$
- *skew-Hermitian* if
$$A^* = -A,$$
- *unitary* if
$$AA^* = A^*A = I_n.$$

Recall from Proposition 11.12 that when E is a Hermitian space and $(e_1, \ldots, e_n)$ is an orthonormal basis for E, if A is the matrix of a linear map $f\colon E \to E$ w.r.t. the basis $(e_1, \ldots, e_n)$, then A^* is the matrix of the adjoint f^* of f. Consequently, a normal linear map has a normal matrix, a self-adjoint linear map has a Hermitian matrix, a skew-self-adjoint linear map has a skew-Hermitian matrix, and a unitary linear map has a unitary matrix.

Similarly, if E and F are Hermitian spaces, $(u_1, \ldots, u_n)$ is an orthonormal basis for E, and $(v_1, \ldots, v_m)$ is an orthonormal basis for F, if a linear map $f\colon E \to F$ has the matrix A w.r.t. the bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_m)$, then its adjoint f^* has the matrix A^* w.r.t. the bases $(v_1, \ldots, v_m)$ and $(u_1, \ldots, u_n)$.

Furthermore, if $(u_1, \ldots, u_n)$ is another orthonormal basis for E and P is the change of basis matrix whose columns are the components of the u_i w.r.t. the basis $(e_1, \ldots, e_n)$, then P is unitary, and for any linear map $f\colon E \to E$, if A is the matrix of f w.r.t $(e_1, \ldots, e_n)$ and B is the matrix of f w.r.t. $(u_1, \ldots, u_n)$, then

$$B = P^* A P.$$

Theorem 13.11 can be restated in terms of matrices as follows. We can also say a little more about eigenvalues (easy exercise left to the reader).

Theorem 13.20. *For every complex normal matrix A there is a unitary matrix U and a diagonal matrix D such that $A = UDU^*$. Furthermore, if A is Hermitian, then D is a real matrix; if A is skew-Hermitian, then the entries in D are pure imaginary or null; and if A is unitary, then the entries in D have absolute value* 1.

13.5 Conditioning of Eigenvalue Problems

The following $n \times n$ matrix

$$A = \begin{pmatrix} 0 & & & & & \\ 1 & 0 & & & & \\ & 1 & 0 & & & \\ & & \ddots & \ddots & & \\ & & & 1 & 0 & \\ & & & & 1 & 0 \end{pmatrix}$$

has the eigenvalue 0 with multiplicity n. However, if we perturb the top rightmost entry of A by ϵ, it is easy to see that the characteristic polynomial of the matrix

$$A(\epsilon) = \begin{pmatrix} 0 & & & & & \epsilon \\ 1 & 0 & & & & \\ & 1 & 0 & & & \\ & & \ddots & \ddots & & \\ & & & 1 & 0 & \\ & & & & 1 & 0 \end{pmatrix}$$

is $X^n - \epsilon$. It follows that if $n = 40$ and $\epsilon = 10^{-40}$, $A(10^{-40})$ has the eigenvalues $e^{k2\pi i/40}10^{-1}$ with $k = 1, \ldots, 40$. Thus, we see that a very small change ($\epsilon = 10^{-40}$) to the matrix A causes

a significant change to the eigenvalues of A (from 0 to $e^{k2\pi i/40}10^{-1}$). Indeed, the relative error is 10^{-39}. Worse, due to machine precision, since very small numbers are treated as 0, the error on the computation of eigenvalues (for example, of the matrix $A(10^{-40})$) can be very large.

This phenomenon is similar to the phenomenon discussed in Section 6.3 where we studied the effect of a small pertubation of the coefficients of a linear system $Ax = b$ on its solution. In Section 6.3, we saw that the behavior of a linear system under small perturbations is governed by the condition number $\mathrm{cond}(A)$ of the matrix A. In the case of the eigenvalue problem (finding the eigenvalues of a matrix), we will see that the conditioning of the problem depends on the condition number of the change of basis matrix P used in reducing the matrix A to its diagonal form $D = P^{-1}AP$, rather than on the condition number of A itself. The following proposition in which we assume that A is diagonalizable and that the matrix norm $\| \|$ satisfies a special condition (satisfied by the operator norms $\| \|_p$ for $p = 1, 2, \infty$), is due to Bauer and Fike (1960).

Proposition 13.21. *Let $A \in \mathrm{M}_n(\mathbb{C})$ be a diagonalizable matrix, P be an invertible matrix and, D be a diagonal matrix $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ such that*

$$A = PDP^{-1},$$

and let $\| \|$ be a matrix norm such that

$$\|\mathrm{diag}(\alpha_1, \ldots, \alpha_n)\| = \max_{1 \leq i \leq n} |\alpha_i|,$$

for every diagonal matrix. Then, for every perturbation matrix δA, if we write

$$B_i = \{z \in \mathbb{C} \mid |z - \lambda_i| \leq \mathrm{cond}(P)\, \|\delta A\|\},$$

for every eigenvalue λ of $A + \delta A$, we have

$$\lambda \in \bigcup_{k=1}^{n} B_k.$$

Proof. Let λ be any eigenvalue of the matrix $A + \delta A$. If $\lambda = \lambda_j$ for some j, then the result is trivial. Thus, assume that $\lambda \neq \lambda_j$ for $j = 1, \ldots, n$. In this case, the matrix $D - \lambda I$ is invertible (since its eigenvalues are $\lambda - \lambda_j$ for $j = 1, \ldots, n$), and we have

$$\begin{aligned} P^{-1}(A + \delta A - \lambda I)P &= D - \lambda I + P^{-1}(\delta A)P \\ &= (D - \lambda I)(I + (D - \lambda I)^{-1}P^{-1}(\delta A)P). \end{aligned}$$

Since λ is an eigenvalue of $A + \delta A$, the matrix $A + \delta A - \lambda I$ is singular, so the matrix

$$I + (D - \lambda I)^{-1}P^{-1}(\delta A)P$$

must also be singular. By Proposition 6.9(2), we have

$$1 \leq \left\|(D - \lambda I)^{-1} P^{-1} (\delta A) P\right\|,$$

and since $\| \|$ is a matrix norm,

$$\left\|(D - \lambda I)^{-1} P^{-1} (\delta A) P\right\| \leq \left\|(D - \lambda I)^{-1}\right\| \left\|P^{-1}\right\| \|\delta A\| \|P\|,$$

so we have

$$1 \leq \left\|(D - \lambda I)^{-1}\right\| \left\|P^{-1}\right\| \|\delta A\| \|P\|.$$

Now, $(D - \lambda I)^{-1}$ is a diagonal matrix with entries $1/(\lambda_i - \lambda)$, so by our assumption on the norm,

$$\left\|(D - \lambda I)^{-1}\right\| = \frac{1}{\min_i(|\lambda_i - \lambda|)}.$$

As a consequence, since there is some index k for which $\min_i(|\lambda_i - \lambda|) = |\lambda_k - \lambda|$, we have

$$\left\|(D - \lambda I)^{-1}\right\| = \frac{1}{|\lambda_k - \lambda|},$$

and we obtain

$$|\lambda - \lambda_k| \leq \left\|P^{-1}\right\| \|\delta A\| \|P\| = \mathrm{cond}(P) \|\delta A\|,$$

which proves our result. □

Proposition 13.21 implies that for any diagonalizable matrix A, if we define $\Gamma(A)$ by

$$\Gamma(A) = \inf\{\mathrm{cond}(P) \mid P^{-1} A P = D\},$$

then for every eigenvalue λ of $A + \delta A$, we have

$$\lambda \in \bigcup_{k=1}^{n} \{z \in \mathbb{C}^n \mid |z - \lambda_k| \leq \Gamma(A) \|\delta A\|\}.$$

The number $\Gamma(A)$ is called the *conditioning of A relative to the eigenvalue problem*. If A is a normal matrix, since by Theorem 13.20, A can be diagonalized with respect to a unitary matrix U, and since for the spectral norm $\|U\|_2 = 1$, we see that $\Gamma(A) = 1$. Therefore, normal matrices are very well conditionned w.r.t. the eigenvalue problem. In fact, for every eigenvalue λ of $A + \delta A$ (with A normal), we have

$$\lambda \in \bigcup_{k=1}^{n} \{z \in \mathbb{C}^n \mid |z - \lambda_k| \leq \|\delta A\|_2\}.$$

If A and $A+\delta A$ are both symmetric (or Hermitian), there are sharper results; see Proposition 13.27.

Note that the matrix $A(\epsilon)$ from the beginning of the section is not normal.

13.6 Rayleigh Ratios and the Courant-Fischer Theorem

A fact that is used frequently in optimization problems is that the eigenvalues of a symmetric matrix are characterized in terms of what is known as the *Rayleigh ratio*, defined by

$$R(A)(x) = \frac{x^\top Ax}{x^\top x}, \quad x \in \mathbb{R}^n, x \neq 0.$$

The following proposition is often used to prove the correctness of various optimization or approximation problems (for example PCA).

Proposition 13.22. *(Rayleigh–Ritz) If A is a symmetric $n \times n$ matrix with eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and if $(u_1, \ldots, u_n)$ is any orthonormal basis of eigenvectors of A, where u_i is a unit eigenvector associated with λ_i, then*

$$\max_{x \neq 0} \frac{x^\top Ax}{x^\top x} = \lambda_n$$

(with the maximum attained for $x = u_n$), and

$$\max_{x \neq 0, x \in \{u_{n-k+1}, \ldots, u_n\}^\perp} \frac{x^\top Ax}{x^\top x} = \lambda_{n-k}$$

(with the maximum attained for $x = u_{n-k}$), where $1 \leq k \leq n-1$. Equivalently, if V_k is the subspace spanned by $(u_1, \ldots, u_k)$, then

$$\lambda_k = \max_{x \neq 0, x \in V_k} \frac{x^\top Ax}{x^\top x}, \quad k = 1, \ldots, n.$$

Proof. First, observe that

$$\max_{x \neq 0} \frac{x^\top Ax}{x^\top x} = \max_x \{x^\top Ax \mid x^\top x = 1\},$$

and similarly,

$$\max_{x \neq 0, x \in \{u_{n-k+1}, \ldots, u_n\}^\perp} \frac{x^\top Ax}{x^\top x} = \max_x \left\{x^\top Ax \mid (x \in \{u_{n-k+1}, \ldots, u_n\}^\perp) \wedge (x^\top x = 1)\right\}.$$

Since A is a symmetric matrix, its eigenvalues are real and it can be diagonalized with respect to an orthonormal basis of eigenvectors, so let $(u_1, \ldots, u_n)$ be such a basis. If we write

$$x = \sum_{i=1}^n x_i u_i,$$

a simple computation shows that

$$x^\top Ax = \sum_{i=1}^{n} \lambda_i x_i^2.$$

If $x^\top x = 1$, then $\sum_{i=1}^{n} x_i^2 = 1$, and since we assumed that $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, we get

$$x^\top Ax = \sum_{i=1}^{n} \lambda_i x_i^2 \leq \lambda_n \left(\sum_{i=1}^{n} x_i^2 \right) = \lambda_n.$$

Thus,

$$\max_x \left\{ x^\top Ax \mid x^\top x = 1 \right\} \leq \lambda_n,$$

and since this maximum is achieved for $e_n = (0, 0, \ldots, 1)$, we conclude that

$$\max_x \left\{ x^\top Ax \mid x^\top x = 1 \right\} = \lambda_n.$$

Next, observe that $x \in \{u_{n-k+1}, \ldots, u_n\}^\perp$ and $x^\top x = 1$ iff $x_{n-k+1} = \cdots = x_n = 0$ and $\sum_{i=1}^{n-k} x_i^2 = 1$. Consequently, for such an x, we have

$$x^\top Ax = \sum_{i=1}^{n-k} \lambda_i x_i^2 \leq \lambda_{n-k} \left(\sum_{i=1}^{n-k} x_i^2 \right) = \lambda_{n-k}.$$

Thus,

$$\max_x \left\{ x^\top Ax \mid (x \in \{u_{n-k+1}, \ldots, u_n\}^\perp) \wedge (x^\top x = 1) \right\} \leq \lambda_{n-k},$$

and since this maximum is achieved for $e_{n-k} = (0, \ldots, 0, 1, 0, \ldots, 0)$ with a 1 in position $n - k$, we conclude that

$$\max_x \left\{ x^\top Ax \mid (x \in \{u_{n-k+1}, \ldots, u_n\}^\perp) \wedge (x^\top x = 1) \right\} = \lambda_{n-k},$$

as claimed. □

For our purposes, we need the version of Proposition 13.22 applying to min instead of max, whose proof is obtained by a trivial modification of the proof of Proposition 13.22.

Proposition 13.23. *(Rayleigh–Ritz) If A is a symmetric $n \times n$ matrix with eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and if $(u_1, \ldots, u_n)$ is any orthonormal basis of eigenvectors of A, where u_i is a unit eigenvector associated with λ_i, then*

$$\min_{x \neq 0} \frac{x^\top Ax}{x^\top x} = \lambda_1$$

(with the minimum attained for $x = u_1$), and

$$\min_{x \neq 0, x \in \{u_1, \ldots, u_{i-1}\}^\perp} \frac{x^\top Ax}{x^\top x} = \lambda_i$$

(with the minimum attained for $x = u_i$), where $2 \leq i \leq n$. Equivalently, if $W_k = V_{k-1}^{\perp}$ denotes the subspace spanned by $(u_k, \ldots, u_n)$ (with $V_0 = (0)$), then

$$\lambda_k = \min_{x \neq 0, x \in W_k} \frac{x^\top A x}{x^\top x} = \min_{x \neq 0, x \in V_{k-1}^{\perp}} \frac{x^\top A x}{x^\top x}, \quad k = 1, \ldots, n.$$

Propositions 13.22 and 13.23 together are known the *Rayleigh–Ritz theorem*.

As an application of Propositions 13.22 and 13.23, we prove a proposition which allows us to compare the eigenvalues of two symmetric matrices A and $B = R^\top A R$, where R is a rectangular matrix satisfying the equation $R^\top R = I$.

First, we need a definition.

Definition 13.5. Given an $n \times n$ symmetric matrix A and an $m \times m$ symmetric B, with $m \leq n$, if $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ are the eigenvalues of A and $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_m$ are the eigenvalues of B, then we say that the eigenvalues of B *interlace* the eigenvalues of A if

$$\lambda_i \leq \mu_i \leq \lambda_{n-m+i}, \quad i = 1, \ldots, m.$$

For example, if $n = 5$ and $m = 3$, we have

$$\begin{aligned} \lambda_1 &\leq \mu_1 \leq \lambda_3 \\ \lambda_2 &\leq \mu_2 \leq \lambda_4 \\ \lambda_3 &\leq \mu_3 \leq \lambda_5. \end{aligned}$$

Proposition 13.24. *Let A be an $n \times n$ symmetric matrix, R be an $n \times m$ matrix such that $R^\top R = I$ (with $m \leq n$), and let $B = R^\top A R$ (an $m \times m$ matrix). The following properties hold:*

(a) The eigenvalues of B interlace the eigenvalues of A.

(b) If $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ are the eigenvalues of A and $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_m$ are the eigenvalues of B, and if $\lambda_i = \mu_i$, then there is an eigenvector v of B with eigenvalue μ_i such that Rv is an eigenvector of A with eigenvalue λ_i.

Proof. (a) Let $(u_1, \ldots, u_n)$ be an orthonormal basis of eigenvectors for A, and let $(v_1, \ldots, v_m)$ be an orthonormal basis of eigenvectors for B. Let U_j be the subspace spanned by $(u_1, \ldots, u_j)$ and let V_j be the subspace spanned by $(v_1, \ldots, v_j)$. For any i, the subspace V_i has dimension i and the subspace $R^\top U_{i-1}$ has dimension at most $i - 1$. Therefore, there is some nonzero vector $v \in V_i \cap (R^\top U_{i-1})^{\perp}$, and since

$$v^\top R^\top u_j = (Rv)^\top u_j = 0, \quad j = 1, \ldots, i-1,$$

we have $Rv \in (U_{i-1})^{\perp}$. By Proposition 13.23 and using the fact that $R^\top R = I$, we have

$$\lambda_i \leq \frac{(Rv)^\top A R v}{(Rv)^\top R v} = \frac{v^\top B v}{v^\top v}.$$

On the other hand, by Proposition 13.22,

$$\mu_i = \max_{x \neq 0, x \in \{v_{i+1}, \ldots, v_n\}^\perp} \frac{x^\top B x}{x^\top x} = \max_{x \neq 0, x \in \{v_1, \ldots, v_i\}} \frac{x^\top B x}{x^\top x},$$

so

$$\frac{w^\top B w}{w^\top w} \leq \mu_i \quad \text{for all } w \in V_i,$$

and since $v \in V_i$, we have

$$\lambda_i \leq \frac{v^\top B v}{v^\top v} \leq \mu_i, \quad i = 1, \ldots, m.$$

We can apply the same argument to the symmetric matrices $-A$ and $-B$, to conclude that

$$-\lambda_{n-m+i} \leq -\mu_i,$$

that is,

$$\mu_i \leq \lambda_{n-m+i}, \quad i = 1, \ldots, m.$$

Therefore,

$$\lambda_i \leq \mu_i \leq \lambda_{n-m+i}, \quad i = 1, \ldots, m,$$

as desired.

(b) If $\lambda_i = \mu_i$, then

$$\lambda_i = \frac{(Rv)^\top A R v}{(Rv)^\top R v} = \frac{v^\top B v}{v^\top v} = \mu_i,$$

so v must be an eigenvector for B and Rv must be an eigenvector for A, both for the eigenvalue $\lambda_i = \mu_i$. □

Proposition 13.24 immediately implies the *Poincaré separation theorem*. It can be used in situations, such as in quantum mechanics, where one has information about the inner products $u_i^\top A u_j$.

Proposition 13.25. *(Poincaré separation theorem) Let A be a $n \times n$ symmetric (or Hermitian) matrix, let r be some integer with $1 \leq r \leq n$, and let $(u_1, \ldots, u_r)$ be r orthonormal vectors. Let $B = (u_i^\top A u_j)$ (an $r \times r$ matrix), let $\lambda_1(A) \leq \ldots \leq \lambda_n(A)$ be the eigenvalues of A and $\lambda_1(B) \leq \ldots \leq \lambda_r(B)$ be the eigenvalues of B; then we have*

$$\lambda_k(A) \leq \lambda_k(B) \leq \lambda_{k+n-r}(A), \quad k = 1, \ldots, r.$$

Observe that Proposition 13.24 implies that

$$\lambda_1 + \cdots + \lambda_m \leq \mathrm{tr}(R^\top A R) \leq \lambda_{n-m+1} + \cdots + \lambda_n.$$

If P_1 is the the $n \times (n-1)$ matrix obtained from the identity matrix by dropping its last column, we have $P_1^\top P_1 = I$, and the matrix $B = P_1^\top A P_1$ is the matrix obtained from A by deleting its last row and its last column. In this case, the interlacing result is

$$\lambda_1 \leq \mu_1 \leq \lambda_2 \leq \mu_2 \leq \cdots \leq \mu_{n-2} \leq \lambda_{n-1} \leq \mu_{n-1} \leq \lambda_n,$$

a genuine interlacing. We obtain similar results with the matrix P_{n-r} obtained by dropping the last $n-r$ columns of the identity matrix and setting $B = P_{n-r}^\top A P_{n-r}$ (B is the $r \times r$ matrix obtained from A by deleting its last $n-r$ rows and columns). In this case, we have the following interlacing inequalities known as *Cauchy interlacing theorem*:

$$\lambda_k \leq \mu_k \leq \lambda_{k+n-r}, \quad k = 1, \ldots, r. \tag{$*$}$$

Another useful tool to prove eigenvalue equalities is the Courant–Fischer characterization of the eigenvalues of a symmetric matrix, also known as the Min-max (and Max-min) theorem.

Theorem 13.26. *(Courant–Fischer) Let A be a symmetric $n \times n$ matrix with eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. If $\mathcal{V}_k$ denotes the set of subspaces of $\mathbb{R}^n$ of dimension k, then*

$$\lambda_k = \max_{W \in \mathcal{V}_{n-k+1}} \min_{x \in W, x \neq 0} \frac{x^\top A x}{x^\top x}$$

$$\lambda_k = \min_{W \in \mathcal{V}_k} \max_{x \in W, x \neq 0} \frac{x^\top A x}{x^\top x}.$$

Proof. Let us consider the second equality, the proof of the first equality being similar. Let $(u_1, \ldots, u_n)$ be any orthonormal basis of eigenvectors of A, where u_i is a unit eigenvector associated with λ_i. Observe that the space V_k spanned by $(u_1, \ldots, u_k)$ has dimension k, and by Proposition 13.22, we have

$$\lambda_k = \max_{x \neq 0, x \in V_k} \frac{x^\top A x}{x^\top x} \geq \min_{W \in \mathcal{V}_k} \max_{x \in W, x \neq 0} \frac{x^\top A x}{x^\top x}.$$

Therefore, we need to prove the reverse inequality; that is, we have to show that

$$\lambda_k \leq \max_{x \neq 0, x \in W} \frac{x^\top A x}{x^\top x}, \quad \text{for all} \quad W \in \mathcal{V}_k.$$

Now, for any $W \in \mathcal{V}_k$, if we can prove that $W \cap V_{k-1}^\perp \neq (0)$, then for any nonzero $v \in W \cap V_{k-1}^\perp$, by Proposition 13.23 , we have

$$\lambda_k = \min_{x \neq 0, x \in V_{k-1}^\perp} \frac{x^\top A x}{x^\top x} \leq \frac{v^\top A v}{v^\top v} \leq \max_{x \in W, x \neq 0} \frac{x^\top A x}{x^\top x}.$$

It remains to prove that $\dim(W \cap V_{k-1}^\perp) \geq 1$. However, $\dim(V_{k-1}) = k-1$, so $\dim(V_{k-1}^\perp) = n-k+1$, and by hypothesis $\dim(W) = k$. By the Grassmann relation,

$$\dim(W) + \dim(V_{k-1}^\perp) = \dim(W \cap V_{k-1}^\perp) + \dim(W + V_{k-1}^\perp),$$

and since $\dim(W + V_{k-1}^{\perp}) \leq \dim(\mathbb{R}^n) = n$, we get

$$k + n - k + 1 \leq \dim(W \cap V_{k-1}^{\perp}) + n;$$

that is, $1 \leq \dim(W \cap V_{k-1}^{\perp})$, as claimed. □

The Courant–Fischer theorem yields the following useful result about perturbing the eigenvalues of a symmetric matrix due to Hermann Weyl.

Proposition 13.27. *Given two $n \times n$ symmetric matrices A and $B = A + \delta A$, if $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_n$ are the eigenvalues of A and $\beta_1 \leq \beta_2 \leq \cdots \leq \beta_n$ are the eigenvalues of B, then*

$$|\alpha_k - \beta_k| \leq \rho(\delta A) \leq \|\delta A\|_2, \quad k = 1, \ldots, n.$$

Proof. Let $\mathcal{V}_k$ be defined as in the Courant–Fischer theorem and let V_k be the subspace spanned by the k eigenvectors associated with $\lambda_1, \ldots, \lambda_k$. By the Courant–Fischer theorem applied to B, we have

$$\begin{aligned}
\beta_k &= \min_{W \in \mathcal{V}_k} \max_{x \in W, x \neq 0} \frac{x^\top B x}{x^\top x} \\
&\leq \max_{x \in V_k} \frac{x^\top B x}{x^\top x} \\
&= \max_{x \in V_k} \left(\frac{x^\top A x}{x^\top x} + \frac{x^\top \delta\, A x}{x^\top x} \right) \\
&\leq \max_{x \in V_k} \frac{x^\top A x}{x^\top x} + \max_{x \in V_k} \frac{x^\top \delta A\, x}{x^\top x}.
\end{aligned}$$

By Proposition 13.22, we have

$$\alpha_k = \max_{x \in V_k} \frac{x^\top A x}{x^\top x},$$

so we obtain

$$\begin{aligned}
\beta_k &\leq \max_{x \in V_k} \frac{x^\top A x}{x^\top x} + \max_{x \in V_k} \frac{x^\top \delta A\, x}{x^\top x} \\
&= \alpha_k + \max_{x \in V_k} \frac{x^\top \delta A\, x}{x^\top x} \\
&\leq \alpha_k + \max_{x \in \mathbb{R}^n} \frac{x^\top \delta A\, x}{x^\top x}.
\end{aligned}$$

Now, by Proposition 13.22 and Proposition 6.7, we have

$$\max_{x \in \mathbb{R}^n} \frac{x^\top \delta A\, x}{x^\top x} = \max_i \lambda_i(\delta A) \leq \rho(\delta A) \leq \|\delta A\|_2,$$

where $\lambda_i(\delta A)$ denotes the ith eigenvalue of δA, which implies that

$$\beta_k \leq \alpha_k + \rho(\delta A) \leq \alpha_k + \|\delta A\|_2 .$$

By exchanging the roles of A and B, we also have

$$\alpha_k \leq \beta_k + \rho(\delta A) \leq \beta_k + \|\delta A\|_2 ,$$

and thus,

$$|\alpha_k - \beta_k| \leq \rho(\delta A) \leq \|\delta A\|_2 , \quad k = 1, \ldots, n,$$

as claimed. □

Proposition 13.27 also holds for Hermitian matrices.

A pretty result of Wielandt and Hoffman asserts that

$$\sum_{k=1}^{n} (\alpha_k - \beta_k)^2 \leq \|\delta A\|_F^2 ,$$

where $\| \ \|_F$ is the Frobenius norm. However, the proof is significantly harder than the above proof; see Lax [66].

The Courant–Fischer theorem can also be used to prove some famous inequalities due to Hermann Weyl. Given two symmetric (or Hermitian) matrices A and B, let $\lambda_i(A), \lambda_i(B)$, and $\lambda_i(A + B)$ denote the ith eigenvalue of A, B, and $A + B$, respectively, arranged in nondecreasing order.

Proposition 13.28. *(Weyl) Given two symmetric (or Hermitian) $n \times n$ matrices A and B, the following inequalities hold: For all i, j, k with $1 \leq i, j, k \leq n$:*

1. *If $i + j = k + 1$, then*
$$\lambda_i(A) + \lambda_j(B) \leq \lambda_k(A + B).$$
2. *If $i + j = k + n$, then*
$$\lambda_k(A + B) \leq \lambda_i(A) + \lambda_j(B).$$

Proof. Observe that the first set of inequalities is obtained form the second set by replacing A by $-A$ and B by $-B$, so it is enough to prove the second set of inequalities. By the Courant–Fischer theorem, there is a subspace H of dimension $n - k + 1$ such that

$$\lambda_k(A + B) = \min_{x \in H, x \neq 0} \frac{x^\top (A + B)x}{x^\top x}.$$

Similarly, there exist a subspace F of dimension i and a subspace G of dimension j such that

$$\lambda_i(A) = \max_{x \in F, x \neq 0} \frac{x^\top Ax}{x^\top x}, \quad \lambda_j(B) = \max_{x \in G, x \neq 0} \frac{x^\top Bx}{x^\top x}.$$

We claim that $F \cap G \cap H \neq (0)$. To prove this, we use the Grassmann relation twice. First,

$$\dim(F \cap G \cap H) = \dim(F) + \dim(G \cap H) - \dim(F + (G \cap H)) \geq \dim(F) + \dim(G \cap H) - n,$$

and second,

$$\dim(G \cap H) = \dim(G) + \dim(H) - \dim(G + H) \geq \dim(G) + \dim(H) - n,$$

so

$$\dim(F \cap G \cap H) \geq \dim(F) + \dim(G) + \dim(H) - 2n.$$

However,

$$\dim(F) + \dim(G) + \dim(H) = i + j + n - k + 1$$

and $i + j = k + n$, so we have

$$\dim(F \cap G \cap H) \geq i + j + n - k + 1 - 2n = k + n + n - k + 1 - 2n = 1,$$

which shows that $F \cap G \cap H \neq (0)$. Then, for any unit vector $z \in F \cap G \cap H \neq (0)$, we have

$$\lambda_k(A + B) \leq z^\top (A + B)z, \quad \lambda_i(A) \geq z^\top Az, \quad \lambda_j(B) \geq z^\top Bz,$$

establishing the desired inequality $\lambda_k(A + B) \leq \lambda_i(A) + \lambda_j(B)$. □

In the special case $i = j = k$, we obtain

$$\lambda_1(A) + \lambda_1(B) \leq \lambda_1(A + B), \quad \lambda_n(A + B) \leq \lambda_n(A) + \lambda_n(B).$$

It follows that λ_1 is concave, while λ_n is convex.

If $i = 1$ and $j = k$, we obtain

$$\lambda_1(A) + \lambda_k(B) \leq \lambda_k(A + B),$$

and if $i = k$ and $j = n$, we obtain

$$\lambda_k(A + B) \leq \lambda_k(A) + \lambda_n(B),$$

and combining them, we get

$$\lambda_1(A) + \lambda_k(B) \leq \lambda_k(A + B) \leq \lambda_k(A) + \lambda_n(B).$$

In particular, if B is positive semidefinite, since its eigenvalues are nonnegative, we obtain the following inequality known as the *monotonicity theorem* for symmetric (or Hermitian) matrices: if A and B are symmetric (or Hermitian) and B is positive semidefinite, then

$$\lambda_k(A) \leq \lambda_k(A + B) \quad k = 1, \ldots, n.$$

The reader is referred to Horn and Johnson [55] (Chapters 4 and 7) for a very complete treatment of matrix inequalities and interlacing results, and also to Lax [66] and Serre [95].

We now have all the tools to present the important *singular value decomposition* (SVD) and the *polar form* of a matrix. However, we prefer to first illustrate how the material of this section can be used to discretize boundary value problems, and we give a brief introduction to the finite elements method.

13.7 Summary

The main concepts and results of this chapter are listed below:

- *Normal* linear maps, *self-adjoint* linear maps, *skew-self-adjoint* linear maps, and *orthogonal* linear maps.
- Properties of the eigenvalues and eigenvectors of a normal linear map.
- The *complexification* of a real vector space, of a linear map, and of a Euclidean inner product.
- The eigenvalues of a self-adjoint map in a Hermitian space are *real*.
- The eigenvalues of a self-adjoint map in a Euclidean space are *real*.
- Every self-adjoint linear map on a Euclidean space has an orthonormal basis of eigenvectors.
- Every normal linear map on a Euclidean space can be block diagonalized (blocks of size at most 2×2) with respect to an orthonormal basis of eigenvectors.
- Every normal linear map on a Hermitian space can be diagonalized with respect to an orthonormal basis of eigenvectors.
- The spectral theorems for self-adjoint, skew-self-adjoint, and orthogonal linear maps (on a Euclidean space).
- The spectral theorems for normal, symmetric, skew-symmetric, and orthogonal (real) matrices.
- The spectral theorems for normal, Hermitian, skew-Hermitian, and unitary (complex) matrices.
- The conditioning of eigenvalue problems.
- The *Rayleigh ratio* and the *Rayleigh–Ritz theorem*.
- *Interlacing inequalities* and the *Cauchy interlacing theorem*.
- The *Poincaré separation theorem*.
- The *Courant–Fischer theorem*.
- Inequalities involving perturbations of the eigenvalues of a symmetric matrix.
- The *Weyl inequalities*.

Chapter 14

Variational Approximation of Boundary-Value Problems; Introduction to the Finite Elements Method

14.1 A One-Dimensional Problem: Bending of a Beam

Consider a beam of unit length supported at its ends in 0 and 1, stretched along its axis by a force P, and subjected to a transverse load $f(x)dx$ per element dx, as illustrated in Figure 14.1.

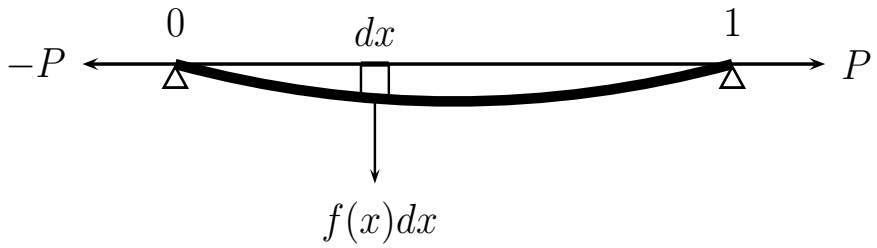


Figure 14.1: Vertical deflection of a beam

The bending moment $u(x)$ at the absissa x is the solution of a boundary problem (BP) of the form

$$\begin{aligned} -u''(x) + c(x)u(x) &= f(x), \quad 0 < x < 1 \\ u(0) &= \alpha \\ u(1) &= \beta, \end{aligned}$$

where $c(x) = P/(EI(x))$, where E is the Young's modulus of the material of which the beam is made and $I(x)$ is the principal moment of inertia of the cross-section of the beam at the abcissa x, and with $\alpha = \beta = 0$. For this problem, we may assume that $c(x) \geq 0$ for all $x \in [0, 1]$.

Remark: The vertical deflection $w(x)$ of the beam and the bending moment $u(x)$ are related by the equation

$$u(x) = -EI\frac{d^2w}{dx^2}.$$

If we seek a solution $u \in C^2([0, 1])$, that is, a function whose first and second derivatives exist and are continuous, then it can be shown that the problem has a unique solution (assuming c and f to be continuous functions on $[0, 1]$).

Except in very rare situations, this problem has no closed-form solution, so we are led to seek approximations of the solutions.

One one way to proceed is to use the *finite difference method*, where we discretize the problem and replace derivatives by differences. Another way is to use a variational approach. In this approach, we follow a somewhat surprising path in which we come up with a so-called "weak formulation" of the problem, by using a trick based on integrating by parts!

First, let us observe that we can always assume that $\alpha = \beta = 0$, by looking for a solution of the form $u(x) - (\alpha(1 - x) + \beta x)$. This turns out to be crucial when we integrate by parts. There are a lot of subtle mathematical details involved to make what follows rigorous, but here, we will take a "relaxed" approach.

First, we need to specify the space of "weak solutions." This will be the vector space V of continuous functions f on $[0, 1]$, with $f(0) = f(1) = 0$, and which are piecewise continuously differentiable on $[0, 1]$. This means that there is a finite number of points $x_0, \ldots, x_{N+1}$ with $x_0 = 0$ and $x_{N+1} = 1$, such that $f'(x_i)$ is undefined for $i = 1, \ldots, N$, but otherwise f' is defined and continuous on each interval (x_i, x_{i+1}) for $i = 0, \ldots, N$.[1] The space V becomes a Euclidean vector space under the inner product

$$\langle f, g\rangle_V = \int_0^1 (f(x)g(x) + f'(x)g'(x))dx,$$

for all $f, g \in V$. The associated norm is

$$\|f\|_V = \left(\int_0^1 (f(x)^2 + f'(x)^2)dx\right)^{1/2}.$$

Assume that u is a solution of our original boundary problem (BP), so that

$$\begin{aligned} -u''(x) + c(x)u(x) &= f(x), \quad 0 < x < 1 \\ u(0) &= 0 \\ u(1) &= 0. \end{aligned}$$

[1] We also assume that $f'(x)$ has a limit when x tends to a boundary of (x_i, x_{i+1}).

Multiply the differential equation by any arbitrary *test function* $v \in V$, obtaining

$$-u''(x)v(x) + c(x)u(x)v(x) = f(x)v(x), \tag{$*$}$$

and integrate this equation! We get

$$-\int_0^1 u''(x)v(x)dx + \int_0^1 c(x)u(x)v(x)dx = \int_0^1 f(x)v(x)dx. \tag{$\dagger$}$$

Now, the trick is to use integration by parts on the first term. Recall that

$$(u'v)' = u''v + u'v',$$

and to be careful about discontinuities, write

$$\int_0^1 u''(x)v(x)dx = \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} u''(x)v(x)dx.$$

Using integration by parts, we have

$$\begin{aligned}
\int_{x_i}^{x_{i+1}} u''(x)v(x)dx &= \int_{x_i}^{x_{i+1}} (u'(x)v(x))'dx - \int_{x_i}^{x_{i+1}} u'(x)v'(x)dx \\
&= [u'(x)v(x)]_{x=x_i}^{x=x_{i+1}} - \int_{x_i}^{x_{i+1}} u'(x)v'(x)dx \\
&= u'(x_{i+1})v(x_{i+1}) - u'(x_i)v(x_i) - \int_{x_i}^{x_{i+1}} u'(x)v'(x)dx.
\end{aligned}$$

It follows that

$$\begin{aligned}
\int_0^1 u''(x)v(x)dx &= \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} u''(x)v(x)dx \\
&= \sum_{i=0}^{N} \left(u'(x_{i+1})v(x_{i+1}) - u'(x_i)v(x_i) - \int_{x_i}^{x_{i+1}} u'(x)v'(x)dx \right) \\
&= u'(1)v(1) - u'(0)v(0) - \int_0^1 u'(x)v'(x)dx.
\end{aligned}$$

However, the test function v satisfies the boundary conditions $v(0) = v(1) = 0$ (recall that $v \in V$), so we get

$$\int_0^1 u''(x)v(x)dx = -\int_0^1 u'(x)v'(x)dx.$$

Consequently, the equation ($\dagger$) becomes

$$\int_0^1 u'(x)v'(x)dx + \int_0^1 c(x)u(x)v(x)dx = \int_0^1 f(x)v(x)dx,$$

or

$$\int_0^1 (u'v' + cuv)dx = \int_0^1 fvdx, \quad \text{for all } v \in V. \tag{$**$}$$

Thus, it is natural to introduce the bilinear form $a\colon V \times V \to \mathbb{R}$ given by

$$a(u,v) = \int_0^1 (u'v' + cuv)dx, \quad \text{for all } u, v \in V,$$

and the linear form $\widetilde{f}\colon V \to \mathbb{R}$ given by

$$\widetilde{f}(v) = \int_0^1 f(x)v(x)dx, \quad \text{for all } v \in V.$$

Then, $(**)$ becomes

$$a(u,v) = \widetilde{f}(v), \quad \text{for all } v \in V.$$

We also introduce the *energy function* J given by

$$J(v) = \frac{1}{2}a(v,v) - \widetilde{f}(v) \quad v \in V.$$

Then, we have the following theorem.

Theorem 14.1. *Let u be any solution of the boundary problem (BP).*

(1) Then we have

$$a(u,v) = \widetilde{f}(v), \quad \text{for all } v \in V, \tag{WF}$$

where

$$a(u,v) = \int_0^1 (u'v' + cuv)dx, \quad \textit{for all } u, v \in V,$$

and

$$\widetilde{f}(v) = \int_0^1 f(x)v(x)dx, \quad \textit{for all } v \in V.$$

(2) If $c(x) \geq 0$ for all $x \in [0,1]$, then a function $u \in V$ is a solution of (WF) iff u minimizes $J(v)$, that is,

$$J(u) = \inf_{v \in V} J(v),$$

with

$$J(v) = \frac{1}{2}a(v,v) - \widetilde{f}(v) \quad v \in V.$$

Furthermore, u is unique.

Proof. We already proved (1).

To prove (2), first we show that

$$\|v\|_V^2 \leq 2a(v,v), \quad \text{for all } v \in V.$$

For this, it suffices to prove that

$$\|v\|_V^2 \leq 2\int_0^1 (f'(x))^2 dx, \quad \text{for all } v \in V.$$

However, by Cauchy-Schwarz for functions, for every $x \in [0,1]$, we have

$$|v(x)| = \left|\int_0^x v'(t)dt\right| \leq \int_0^1 |v'(t)|dt \leq \left(\int_0^1 |v'(t)|^2 dt\right)^{1/2},$$

and so

$$\|v\|_V^2 = \int_0^1 ((v(x))^2 + (v'(x))^2)dx \leq 2\int_0^1 (v'(x))^2 dx \leq 2a(v,v),$$

since

$$a(v,v) = \int_0^1 ((v')^2 + cv^2)dx.$$

Next, it is easy to check that

$$J(u+v) - J(u) = a(u,v) - \widetilde{f}(v) + \frac{1}{2}a(v,v), \quad \text{for all } u, v \in V.$$

Then, if u is a solution of (WF), we deduce that

$$J(u+v) - J(u) = \frac{1}{2}a(v,v) \geq \frac{1}{4}\|v\|_V \geq 0 \quad \text{for all } v \in V.$$

since $a(u,v) - \widetilde{f}(v) = 0$ for all $v \in V$. Therefore, J achieves a minimun for u.

We also have

$$J(u+\theta v) - J(u) = \theta(a(u,v) - f(v)) + \frac{\theta^2}{2}a(v,v) \quad \text{for all } \theta \in \mathbb{R},$$

and so $J(u+\theta v) - J(u) \geq 0$ for all $\theta \in \mathbb{R}$. Consequently, if J achieves a minimum for u, then $a(u,v) = \widetilde{f}(v)$, which means that u is a solution of (WF).

Finally, assuming that $c(x) \geq 0$, we claim that if $v \in V$ and $v \neq 0$, then $a(v,v) > 0$. This is because if $a(v,v) = 0$, since

$$\|v\|_V^2 \leq 2a(v,v) \quad \text{for all } v \in V,$$

we would have $\|v\|_V = 0$, that is, $v = 0$. Then, if $v \neq 0$, from

$$J(u+v) - J(u) = \frac{1}{2}a(v,v) \quad \text{for all } v \in V$$

we see that $J(u+v) > J(u)$, so the minimum u is unique □

Theorem 14.1 shows that every solution u of our boundary problem (BP) is a solution (in fact, unique) of the equation (WF).

The equation (WF) is called the *weak form* or *variational equation* associated with the boundary problem. This idea to derive these equations is due to *Ritz and Galerkin*.

Now, the natural question is whether the variational equation (WF) has a solution, and whether this solution, if it exists, is also a solution of the boundary problem (it must belong to $C^2([0,1])$, which is far from obvious). Then, (BP) and (WF) would be equivalent.

Some fancy tools of analysis can be used to prove these assertions. The first difficulty is that the vector space V is not the right space of solutions, because in order for the variational problem to have a solution, it must be complete. So, we must construct a completion of the vector space V. This can be done and we get the *Sobolev space* $H_0^1(0,1)$. Then, the question of the regularity of the "weak solution" can also be tackled.

We will not worry about all this. Instead, let us find *approximations* of the problem (WF). Instead of using the infinite-dimensional vector space V, we consider *finite-dimensional* subspaces V_a (with $\dim(V_a) = n$) of V, and we consider the *discrete problem*:

Find a function $u^{(a)} \in V_a$, such that

$$a(u^{(a)}, v) = \widetilde{f}(v), \quad \text{for all } v \in V_a. \tag{DWF}$$

Since V_a is finite dimensional (of dimension n), let us pick a basis of functions $(w_1, \ldots, w_n)$ in V_a, so that every function $u \in V_a$ can we written as

$$u = u_1 w_1 + \cdots + u_n w_n.$$

Then, the equation (DWF) holds iff

$$a(u, w_j) = \widetilde{f}(w_j), \quad j = 1, \ldots, n,$$

and by plugging $u_1 w_1 + \cdots + u_n w_n$ for u, we get a system of k linear equations

$$\sum_{i=1}^{n} a(w_i, w_j) u_i = \widetilde{f}(w_j), \quad 1 \leq j \leq n.$$

Because $a(v,v) \geq \frac{1}{2}\|v\|_{V_a}$, the bilinear form a is symmetric positive definite, and thus the matrix $(a(w_i, w_j))$ is symmetric positive definite, and thus invertible. Therefore, (DWF) has a solution given by a *linear system*!

From a practical point of view, we have to compute the integrals

$$a_{ij} = a(w_i, w_j) = \int_0^1 (w_i' w_j' + c w_i w_j) dx,$$

and

$$b_j = \widetilde{f}(w_j) = \int_0^1 f(x) w_j(x) dx.$$

However, if the basis functions are simple enough, this can be done "by hand." Otherwise, numerical integration methods must be used, but there are some good ones.

Let us also remark that the proof of Theorem 14.1 also shows that the unique solution of (DWF) is the unique minimizer of J over all functions in V_a. It is also possible to compare the approximate solution $u^{(a)} \in V_a$ with the exact solution $u \in V$.

Theorem 14.2. *Suppose $c(x) \geq 0$ for all $x \in [0,1]$. For every finite-dimensional subspace V_a ($\dim(V_a) = n$) of V, for every basis $(w_1, \ldots, w_n)$ of V_a, the following properties hold:*

(1) There is a unique function $u^{(a)} \in V_a$ such that

$$a(u^{(a)}, v) = \widetilde{f}(v), \quad \text{for all } v \in V_a, \tag{DWF}$$

and if $u^{(a)} = u_1 w_1 + \cdots + u_n w_n$, then $\mathbf{u} = (u_1, \ldots, u_n)$ is the solution of the linear system

$$A\mathbf{u} = b, \tag{$*$}$$

with $A = (a_{ij}) = (a(w_i, w_j))$ and $b_j = \widetilde{f}(w_j)$, $1 \leq i, j \leq n$. Furthermore, the matrix $A = (a_{ij})$ is symmetric positive definite.

(2) The unique solution $u^{(a)} \in V_a$ of (DWF) is the unique minimizer of J over V_a, that is,

$$J(u^{(a)}) = \inf_{v \in V_a} J(v),$$

(3) There is a constant C independent of V_a and of the unique solution $u \in V$ of (WF), such that

$$\left\|u - u^{(a)}\right\|_V \leq C \inf_{v \in V_a} \|u - v\|_V .$$

We proved (1) and (2), but we will omit the proof of (3) which can be found in Ciarlet [30].

Let us now give examples of the subspaces V_a used in practice. They usually consist of piecewise polynomial functions.

Pick an integer $N \geq 1$ and subdivide $[0,1]$ into $N+1$ intervals $[x_i, x_{i+1}]$, where

$$x_i = hi, \quad h = \frac{1}{N+1}, \quad i = 0, \ldots, N+1.$$

We will use the following fact: every polynomial $P(x)$ of degree $2m+1$ ($m \geq 0$) is completely determined by its values as well as the values of its first m derivatives at two distinct points $\alpha, \beta \in \mathbb{R}$.

There are various ways to prove this. One way is to use the Bernstein basis, because the kth derivative of a polynomial is given by a formula in terms of its control points. For example, for $m = 1$, every degree 3 polynomial can be written as

$$P(x) = (1-x)^3\, b_0 + 3(1-x)^2 x\, b_1 + 3(1-x)x^2\, b_2 + x^3\, b_3,$$

with $b_0, b_1, b_2, b_3 \in \mathbb{R}$, and we showed that

$$\begin{aligned} P'(0) &= 3(b_1 - b_0) \\ P'(1) &= 3(b_3 - b_2). \end{aligned}$$

Given $P(0)$ and $P(1)$, we determine b_0 and b_3, and from $P'(0)$ and $P'(1)$, we determine b_1 and b_2.

In general, for a polynomial of degree m written as

$$P(x) = \sum_{j=0}^{m} b_j B_j^m(x)$$

in terms of the Bernstein basis $(B_0^m(x), \ldots, B_m^m(x))$ with

$$B_j^m(x) = \binom{m}{j}(1-x)^{m-j}x^j,$$

it can be shown that the kth derivative of P at zero is given by

$$P^{(k)}(0) = m(m-1)\cdots(m-k+1)\Big(\sum_{i=0}^{k}\binom{k}{i}(-1)^{k-i}\, b_i\Big),$$

and there is a similar formula for $P^{(k)}(1)$.

Actually, we need to use the Bernstein basis of polynomials $B_k^m[r,s]$, where

$$B_j^m[r,s](x) = \binom{m}{j}\left(\frac{s-x}{s-r}\right)^{m-j}\left(\frac{x-r}{s-r}\right)^{j},$$

with $r < s$, in which case

$$P^{(k)}(0) = \frac{m(m-1)\cdots(m-k+1)}{(s-r)^k}\Big(\sum_{i=0}^{k}\binom{k}{i}(-1)^{k-i}\, b_i\Big),$$

with a similar formula for $P^{(k)}(1)$. In our case, we set $r = x_i, s = x_{i+1}$.

Now, if the $2m+2$ values

$$P(0), P^{(1)}(0), \ldots, P^{(m)}(0), P(1), P^{(1)}(1), \ldots, P^{(m)}(1)$$

are given, we obtain a triangular system that determines uniquely the $2m+2$ control points $b_0, \ldots, b_{2m+1}$.

Recall that $C^m([0,1])$ denotes the set of C^m functions f on $[0,1]$, which means that $f, f^{(1)}, \ldots, f^{(m)}$ exist are are continuous on $[0,1]$.

We define the vector space V_N^m as the subspace of $C^m([0,1])$ consisting of all functions f such that

1. $f(0) = f(1) = 0$.

2. The restriction of f to $[x_i, x_{i+1}]$ is a polynomial of degree $2m+1$, for $i = 0, \ldots, N$.

Observe that the functions in V_N^0 are the piecewise affine functions f with $f(0) = f(1) = 0$; an example is shown in Figure 14.2.

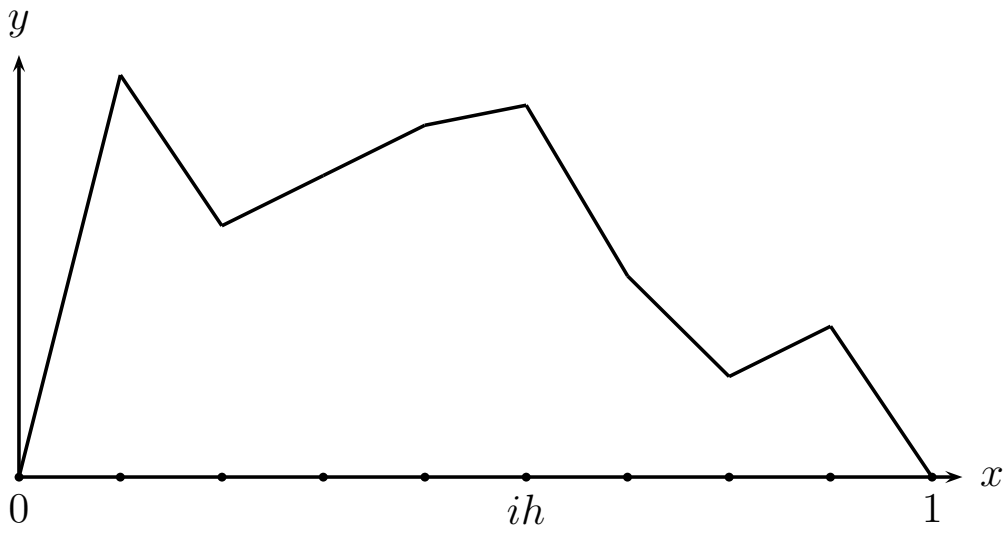


Figure 14.2: A piecewise affine function

This space has dimension N, and a basis consists of the "hat functions" w_i, where the only two nonflat parts of the graph of w_i are the line segments from $(x_{i-1}, 0)$ to $(x_i, 1)$, and from $(x_i, 1)$ to $(x_{i+1}, 0)$, for $i = 1, \ldots, N$, see Figure 14.3.

The basis functions w_i have a small support, which is good because in computing the integrals giving $a(w_i, w_j)$, we find that we get a tridiagonal matrix. They also have the nice property that every function $v \in V_N^0$ has the following expression on the basis (w_i):

$$v(x) = \sum_{i=1}^{N} v(ih) w_i(x), \quad x \in [0,1].$$

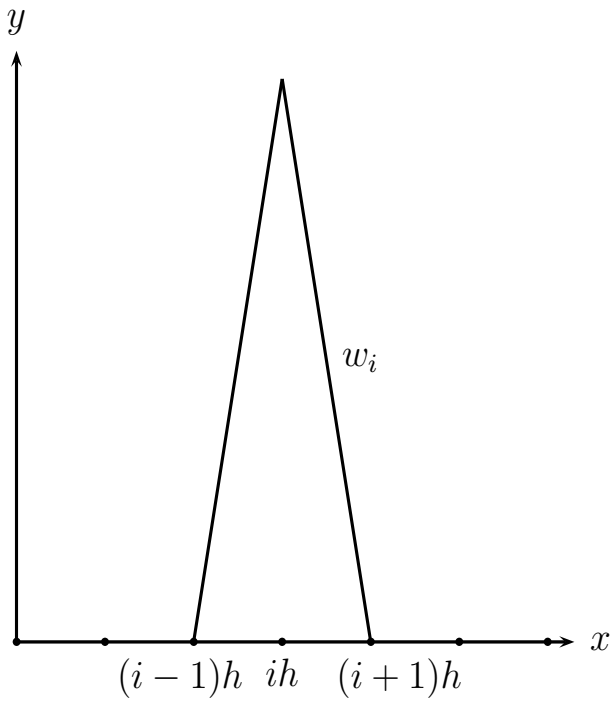


Figure 14.3: A basis "hat function"

In general, it it not hard to see that V_N^m has dimension $mN + 2(m-1)$.

Going back to our problem (the bending of a beam), assuming that c and f are constant functions, it is not hard to show that the linear system $(*)$ becomes

$$\frac{1}{h}\begin{pmatrix} 2+\frac{2c}{3}h^2 & -1+\frac{c}{6}h^2 & & & & \\ -1+\frac{c}{6}h^2 & 2+\frac{2c}{3}h^2 & -1+\frac{c}{6}h^2 & & & \\ \ddots & \ddots & \ddots & & & \\ & -1+\frac{c}{6}h^2 & 2+\frac{2c}{3}h^2 & -1+\frac{c}{6}h^2 & & \\ & & & -1+\frac{c}{6}h^2 & 2+\frac{2c}{3}h^2 & \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-1} \\ u_N \end{pmatrix} = h\begin{pmatrix} f \\ f \\ \vdots \\ f \\ f \end{pmatrix}.$$

We can also find a basis of $2N+2$ cubic functions for V_N^1 consisting of functions with small support. This basis consists of the N functions w_i^0 and of the $N+2$ functions w_i^1

uniquely determined by the following conditions:

$$\begin{aligned} w_i^0(x_j) &= \delta_{ij}, \quad 1 \leq j \leq N,\ 1 \leq i \leq N \\ (w_i^0)'(x_j) &= 0, \quad 0 \leq j \leq N+1,\ 1 \leq i \leq N \\ w_i^1(x_j) &= 0, \quad 1 \leq j \leq N,\ 0 \leq i \leq N+1 \\ (w_i^1)'(x_j) &= \delta_{ij}, \quad 0 \leq j \leq N+1,\ 0 \leq i \leq N+1 \end{aligned}$$

with $\delta_{ij} = 1$ iff $i = j$ and $\delta_{ij} = 0$ if $i \neq j$. Some of these functions are displayed in Figure 14.4. The function w_i^0 is given explicitly by

$$w_i^0(x) = \frac{1}{h^3}(x - (i-1)h)^2((2i+1)h - 2x), \quad (i-1)h \leq x \leq ih,$$

$$w_i^0(x) = \frac{1}{h^3}((i+1)h - x)^2(2x - (2i-1)h), \quad ih \leq x \leq (i+1)h,$$

for $i = 1, \ldots, N$. The function w_j^1 is given explicitly by

$$w_j^1(x) = -\frac{1}{h^2}(ih - x)(x - (i-1)h)^2, \quad (i-1)h \leq x \leq ih,$$

and

$$w_j^1(x) = \frac{1}{h^2}((i+1)h - x)^2(x - ih), \quad ih \leq x \leq (i+1)h,$$

for $j = 0, \ldots, N+1$. Furthermore, for every function $v \in V_N^1$, we have

$$v(x) = \sum_{i=1}^{N} v(ih)w_i^0(x) + \sum_{j=0}^{N+1} v'jih)w_j^1(x), \quad x \in [0,1].$$

If we order these basis functions as

$$w_0^1, w_1^0, w_1^1, w_2^0, w_2^1, \ldots, w_N^0, w_N^1, w_{N+1}^1,$$

we find that if $c = 0$, the matrix A of the system $(*)$ is tridiagonal by blocks, where the blocks are 2×2, 2×1, or 1×2 matrices, and with single entries in the top left and bottom right corner. A different order of the basis vectors would mess up the tridiagonal block structure of A. We leave the details as an exercise.

Let us now take a quick look at a two-dimensional problem, the bending of an elastic membrane.

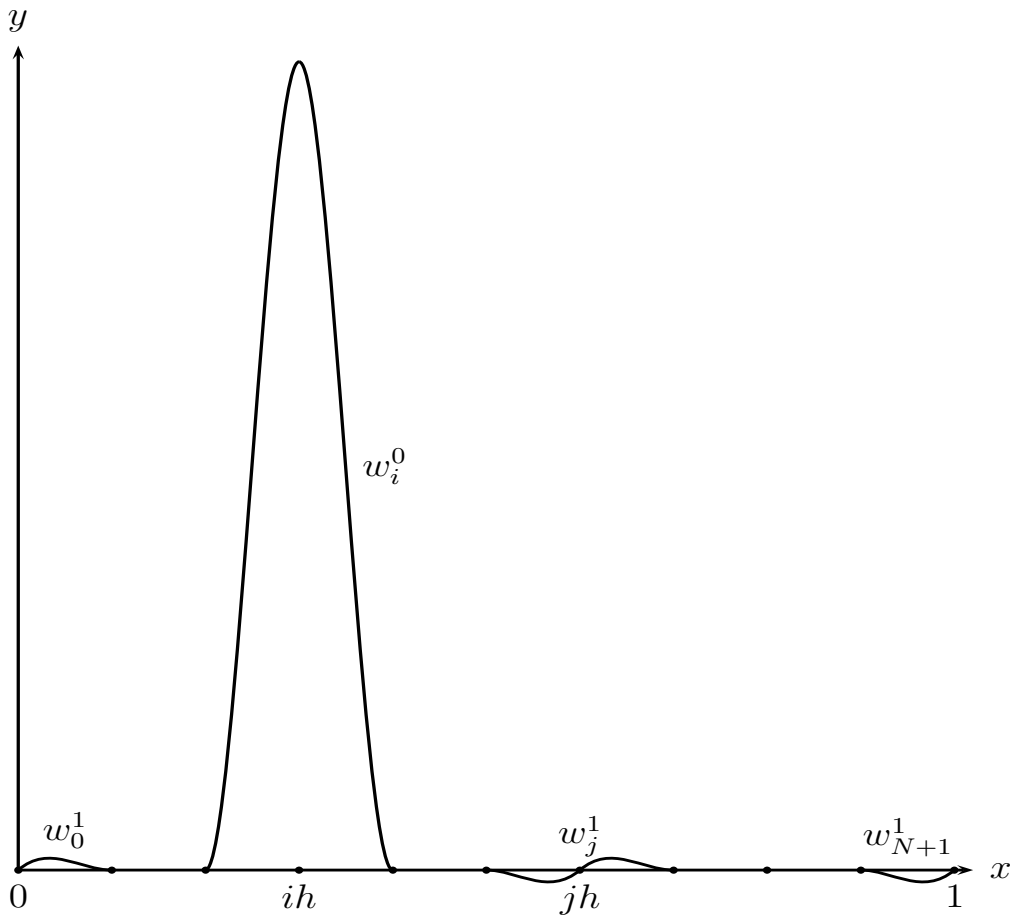


Figure 14.4: The basis functions w_i^0 and w_j^1

14.2 A Two-Dimensional Problem: An Elastic Membrane

Consider an elastic membrane attached to a round contour whose projection on the (x_1, x_2)-plane is the boundary Γ of an open, connected, bounded region Ω in the (x_1, x_2)-plane, as illustrated in Figure 14.5. In other words, we view the membrane as a surface consisting of the set of points (x, z) given by an equation of the form

$$z = u(x),$$

with $x = (x_1, x_2) \in \overline{\Omega}$, where $u\colon \overline{\Omega} \to \mathbb{R}$ is some sufficiently regular function, and we think of $u(x)$ as the vertical displacement of this membrane.

We assume that this membrane is under the action of a vertical force $\tau f(x)dx$ per surface element in the horizontal plane (where τ is the tension of the membrane). The problem is

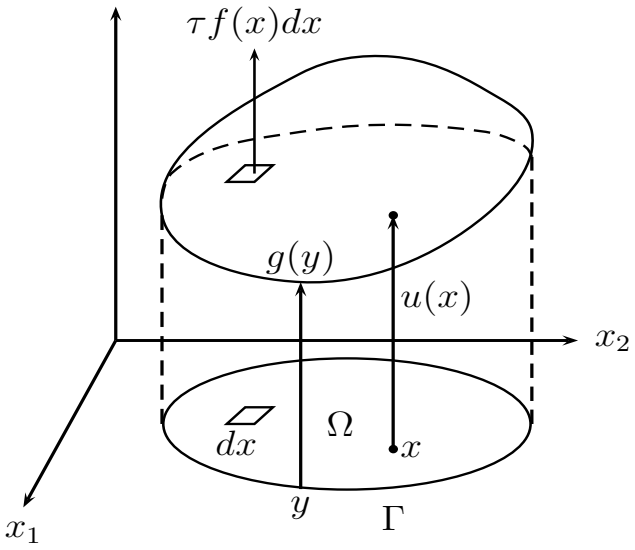


Figure 14.5: An elastic membrane

to find the vertical displacement u as a function of x, for $x \in \overline{\Omega}$. It can be shown (under some assumptions on Ω, Γ, and f), that $u(x)$ is given by a PDE with boundary condition, of the form

$$\begin{aligned} -\Delta u(x) &= f(x), \quad x \in \Omega \\ u(x) &= g(x), \quad x \in \Gamma, \end{aligned}$$

where $g\colon \Gamma \to \mathbb{R}$ represents the height of the contour of the membrane. We are looking for a function u in $C^2(\Omega) \cap C^1(\overline{\Omega})$. The operator Δ is the *Laplacian*, and it is given by

$$\Delta u(x) = \frac{\partial^2 u}{\partial x_1^2}(x) + \frac{\partial^2 u}{\partial x_2^2}(x).$$

This is an example of a *boundary problem*, since the solution u of the PDE must satisfy the condition $u(x) = g(x)$ on the boundary of the domain Ω. The above equation is known as *Poisson's equation*, and when $f = 0$ as *Laplace's equation*.

It can be proved that if the data f, g and Γ are sufficiently smooth, then the problem has a unique solution.

To get a weak formulation of the problem, first we have to make the boundary condition homogeneous, which means that $g(x) = 0$ on Γ. It turns out that g can be extended to the whole of $\overline{\Omega}$ as some sufficiently smooth function $\widehat{h}$, so we can look for a solution of the form $u - \widehat{h}$, but for simplicity, let us assume that the contour of Ω lies in a plane parallel to the

(x_1, x_2)- plane, so that $g = 0$. We let V be the subspace of $C^2(\Omega) \cap C^1(\overline{\Omega})$ consisting of functions v such that $v = 0$ on Γ.

As before, we multiply the PDE by a test function $v \in V$, getting

$$-\Delta u(x)v(x) = f(x)v(x),$$

and we "integrate by parts." In this case, this means that we use a version of Stokes formula known as *Green's first identity*, which says that

$$\int_\Omega -\Delta u\, v\, dx = \int_\Omega (\operatorname{grad} u) \cdot (\operatorname{grad} v)\, dx - \int_\Gamma (\operatorname{grad} u) \cdot n\, v d\sigma$$

(where n denotes the outward pointing unit normal to the surface). Because $v = 0$ on Γ, the integral $\int_\Gamma$ drops out, and we get an equation of the form

$$a(u, v) = \widetilde{f}(v) \quad \text{for all } v \in V,$$

where a is the bilinear form given by

$$a(u, v) = \int_\Omega \left(\frac{\partial u}{\partial x_1} \frac{\partial v}{\partial x_1} + \frac{\partial u}{\partial x_2} \frac{\partial v}{\partial x_2} \right) dx$$

and $\widetilde{f}$ is the linear form given by

$$\widetilde{f}(v) = \int_\Omega f v dx.$$

We get the same equation as in section 14.2, but over a set of functions defined on a two-dimensional domain. As before, we can choose a finite-dimensional subspace V_a of V and consider the discrete problem with respect to V_a. Again, if we pick a basis $(w_1, \ldots, w_n)$ of V_a, a vector $u = u_1 w_1 + \cdots + u_n w_n$ is a solution of the Weak Formulation of our problem iff $\mathbf{u} = (u_1, \ldots, u_n)$ is a solution of the linear system

$$A\mathbf{u} = b,$$

with $A = (a(w_i, w_j))$ and $b = (\widetilde{f}(w_j))$. However, the integrals that give the entries in A and b are much more complicated.

An approach to deal with this problem is the *method of finite elements*. The idea is to also discretize the boundary curve Γ. If we assume that Γ is a *polygonal line*, then we can *triangulate* the domain Ω, and then we consider spaces of functions which are piecewise defined on the triangles of the triangulation of Ω. The simplest functions are piecewise affine and look like tents erected above groups of triangles. Again, we can define base functions with small support, so that the matrix A is tridiagonal by blocks.

The finite element method is a vast subject and it is presented in many books of various degrees of difficulty and obscurity. Let us simply state three important requirements of the finite element method:

1. "Good" triangulations must be found. This in itself is a vast research topic. Delaunay triangulations are good candidates.

2. "Good" spaces of functions must be found; typically piecewise polynomials and splines.

3. "Good" bases consisting of functions will small support must be found, so that integrals can be easily computed and sparse banded matrices arise.

We now consider boundary problems where the solution varies with time.

14.3 Time-Dependent Boundary Problems: The Wave Equation

Consider a homogeneous string (or rope) of constant cross-section, of length L, and stretched (in a vertical plane) between its two ends which are assumed to be fixed and located along the x-axis at $x = 0$ and at $x = L$.

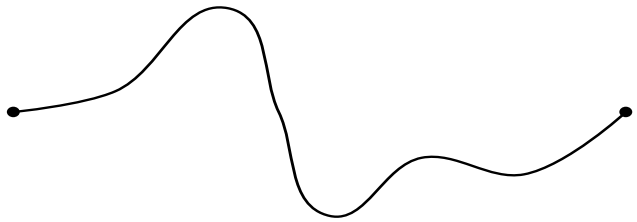


Figure 14.6: A vibrating string

The string is subjected to a transverse force $\tau f(x)dx$ per element of length dx (where τ is the tension of the string). We would like to investigate the small displacements of the string in the vertical plane, that is, how it vibrates.

Thus, we seek a function $u(x,t)$ defined for $t \geq 0$ and $x \in [0,L]$, such that $u(x,t)$ represents the vertical deformation of the string at the abscissa x and at time t.

It can be shown that u must satisfy the following PDE

$$\frac{1}{c^2}\frac{\partial^2 u}{\partial t^2}(x,t) - \frac{\partial^2 u}{\partial x^2}(x,t) = f(x,t), \quad 0 < x < L,\ t > 0,$$

with $c = \sqrt{\tau/\rho}$, where ρ is the linear density of the string, known as the *one-dimensional wave equation*.

Furthermore, the initial shape of the string is known at $t = 0$, as well as the distribution of the initial velocities along the string; in other words, there are two functions $u_{i,0}$ and $u_{i,1}$ such that

$$\begin{aligned} u(x,0) &= u_{i,0}(x), \quad 0 \leq x \leq L, \\ \frac{\partial u}{\partial t}(x,0) &= u_{i,1}(x), \quad 0 \leq x \leq L. \end{aligned}$$

For example, if the string is simply released from its given starting position, we have $u_{i,1} = 0$. Lastly, because the ends of the string are fixed, we must have

$$u(0,t) = u(L,t) = 0, \quad t \geq 0.$$

Consequently, we look for a function $u\colon \mathbb{R}_+ \times [0,L] \to \mathbb{R}$ satisfying the following conditions:

$$\begin{aligned} \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2}(x,t) - \frac{\partial^2 u}{\partial x^2}(x,t) &= f(x,t), \quad 0 < x < L,\ t > 0, \\ u(0,t) = u(L,t) &= 0, \quad t \geq 0 \quad \text{(boundary condition)}, \\ u(x,0) &= u_{i,0}(x), \quad 0 \leq x \leq L \quad \text{(intitial condition)}, \\ \frac{\partial u}{\partial t}(x,0) &= u_{i,1}(x), \quad 0 \leq x \leq L \quad \text{(intitial condition)}. \end{aligned}$$

This is an example of a *time-dependent boundary-value problem*, with two *initial conditions*.

To simplify the problem, assume that $f = 0$, which amounts to neglecting the effect of gravity. In this case, our PDE becomes

$$\frac{1}{c^2}\frac{\partial^2 u}{\partial t^2}(x,t) - \frac{\partial^2 u}{\partial x^2}(x,t) = 0, \quad 0 < x < L,\ t > 0,$$

Let us try our trick of multiplying by a test function v depending only on x, C^1 on $[0,L]$, and such that $v(0) = v(L) = 0$, and integrate by parts. We get the equation

$$\int_0^L \frac{\partial^2 u}{\partial t^2}(x,t)v(x)dx - c^2 \int_0^L \frac{\partial^2 u}{\partial x^2}(x,t)v(x)dx = 0.$$

For the first term, we get

$$\begin{aligned} \int_0^L \frac{\partial^2 u}{\partial t^2}(x,t)v(x)dx &= \int_0^L \frac{\partial^2}{\partial t^2}[u(x,t)v(x)]dx \\ &= \frac{d^2}{dt^2}\int_0^L u(x,t)v(x)dx \\ &= \frac{d^2}{dt^2}\langle u, v\rangle, \end{aligned}$$

where $\langle u, v\rangle$ is the inner product in $L^2([0, L])$. The fact that it is legitimate to move $\partial^2/\partial t^2$ outside of the integral needs to be justified rigorously, but we won't do it here.

For the second term, we get

$$-\int_0^L \frac{\partial^2 u}{\partial x^2}(x,t)v(x)dx = -\left[\frac{\partial u}{\partial x}(x,t)v(x)\right]_{x=0}^{x=L} + \int_0^L \frac{\partial u}{\partial x}(x,t)\frac{dv}{dx}(x)dx,$$

and because $v \in V$, we have $v(0) = v(L) = 0$, so we obtain

$$-\int_0^L \frac{\partial^2 u}{\partial x^2}(x,t)v(x)dx = \int_0^L \frac{\partial u}{\partial x}(x,t)\frac{dv}{dx}(x)dx.$$

Our integrated equation becomes

$$\frac{d^2}{dt^2}\langle u, v\rangle + c^2\int_0^L \frac{\partial u}{\partial x}(x,t)\frac{dv}{dx}(x)dx = 0, \quad \text{for all } v \in V \quad \text{and all } t \geq 0.$$

It is natural to introduce the bilinear form $a\colon V \times V \to \mathbb{R}$ given by

$$a(u, v) = \int_0^L \frac{\partial u}{\partial x}(x,t)\frac{\partial v}{\partial x}(x,t)dx,$$

where, for every $t \in \mathbb{R}_+$, the functions $u(x,t)$ and (v,t) belong to V. Actually, we have to replace V by the subspace of the Sobolev space $H_0^1(0, L)$ consisting of the functions such that $v(0) = v(L) = 0$. Then, the weak formulation (variational formulation) of our problem is this:

Find a function $u \in V$ such that

$$\begin{aligned}\frac{d^2}{dt^2}\langle u, v\rangle + a(u, v) &= 0, \qquad \text{for all } v \in V \quad \text{and all } t \geq 0\\ u(x, 0) &= u_{i,0}(x), \quad 0 \leq x \leq L \quad \text{(intitial condition)},\\ \frac{\partial u}{\partial t}(x, 0) &= u_{i,1}(x), \quad 0 \leq x \leq L \quad \text{(intitial condition)}.\end{aligned}$$

It can be shown that there is a positive constant $\alpha > 0$ such that

$$a(u, u) \geq \alpha \|u\|_{H_0^1}^2 \quad \text{for all } v \in V$$

(Poincaré's inequality), which shows that a is positive definite on V. The above method is known as the method of *Rayleigh-Ritz*.

A study of the above equation requires some sophisticated tools of analysis which go far beyond the scope of these notes. Let us just say that there is a countable sequence of solutions with separated variables of the form

$$u_k^{(1)} = \sin\left(\frac{k\pi x}{L}\right)\cos\left(\frac{k\pi ct}{L}\right), \quad u_k^{(2)} = \sin\left(\frac{k\pi x}{L}\right)\sin\left(\frac{k\pi ct}{L}\right), \quad k \in \mathbb{N}_+,$$

called *modes* (or *normal modes*). Complete solutions of the problem are series obtained by combining the normal modes, and they are of the form

$$u(x,t) = \sum_{k=1}^{\infty} \sin\left(\frac{k\pi x}{L}\right)\left(A_k \cos\left(\frac{k\pi ct}{L}\right) + B_k \sin\left(\frac{k\pi ct}{L}\right)\right),$$

where the coefficients A_k, B_k are determined from the Fourier series of $u_{i,0}$ and $u_{i,1}$.

We now consider discrete approximations of our problem. As before, consider a finite dimensional subspace V_a of V and assume that we have approximations $u_{a,0}$ and $u_{a,1}$ of $u_{i,0}$ and $u_{i,1}$. If we pick a basis $(w_1, \ldots, w_n)$ of V_a, then we can write our unknown function $u(x,t)$ as

$$u(x,t) = u_1(t)w_1 + \cdots + u_n(t)w_n,$$

where $u_1, \ldots, u_n$ are functions of t. Then, if we write $\mathbf{u} = (u_1, \ldots, u_n)$, the discrete version of our problem is

$$\begin{aligned} A\frac{d^2\mathbf{u}}{dt^2} + K\mathbf{u} &= 0, \\ u(x,0) &= u_{a,0}(x), \quad 0 \le x \le L, \\ \frac{\partial u}{\partial t}(x,0) &= u_{a,1}(x), \quad 0 \le x \le L, \end{aligned}$$

where $A = (\langle w_i, w_j\rangle)$ and $K = (a(w_i, w_j))$ are two symmetric matrices, called the *mass matrix* and the *stiffness matrix*, respectively. In fact, because a and the inner product $\langle -, -\rangle$ are positive definite, these matrices are also positive definite.

We have made some progress since we now have a system of ODE's, and we can solve it by analogy with the scalar case. So, we look for solutions of the form $\mathbf{U}\cos\omega t$ (or $\mathbf{U}\sin\omega t$), where $\mathbf{U}$ is an n-dimensional vector. We find that we should have

$$(K - \omega^2 A)\mathbf{U}\cos\omega t = 0,$$

which implies that ω must be a solution of the equation

$$K\mathbf{U} = \omega^2 A\mathbf{U}.$$

Thus, we have to find some λ such that

$$K\mathbf{U} = \lambda A\mathbf{U},$$

a problem known as a *generalized eigenvalue problem*, since the ordinary eigenvalue problem for K is

$$K\mathbf{U} = \lambda\mathbf{U}.$$

Fortunately, because A is SPD, we can reduce this generalized eigenvalue problem to a standard eigenvalue problem. A good way to do so is to use a Cholesky decomposition of A as

$$A = LL^\top,$$

where L is a lower triangular matrix (see Theorem 5.10). Because A is SPD, it is invertible, so L is also invertible, and

$$K\mathbf{U} = \lambda A\mathbf{U} = \lambda LL^\top\mathbf{U}$$

yields

$$L^{-1}K\mathbf{U} = \lambda L^\top\mathbf{U},$$

which can also be written as

$$L^{-1}K(L^\top)^{-1}L^\top\mathbf{U} = \lambda L^\top\mathbf{U}.$$

Then, if we make the change of variable

$$\mathbf{Y} = L^\top\mathbf{U},$$

using the fact $(L^\top)^{-1} = (L^{-1})^\top$, the above equation is equivalent to

$$L^{-1}K(L^{-1})^\top\mathbf{Y} = \lambda\mathbf{Y},$$

a standard eigenvalue problem for the matrix $\widehat{K} = L^{-1}K(L^{-1})^\top$. Furthermore, we know from Section 5.7 that since K is SPD and L^{-1} is invertible, the matrix $\widehat{K} = L^{-1}K(L^{-1})^\top$ is also SPD.

Consequently, $\widehat{K}$ has positive real eigenvalues $(\omega_1^2, \ldots, \omega_n^2)$ (not necessarily distinct) and it can be diagonalized with respect to an orthonormal basis of eigenvectors, say $\mathbf{Y}^1, \ldots, \mathbf{Y}^n$. Then, since $\mathbf{Y} = L^\top\mathbf{U}$, the vectors

$$\mathbf{U}^i = (L^\top)^{-1}\mathbf{Y}^i, \quad i = 1, \ldots, n,$$

are linearly independent and are solutions of the generalized eigenvalue problem; that is,

$$K\mathbf{U}^i = \omega_i^2 A\mathbf{U}^i, \quad i = 1, \ldots, n.$$

More is true. Because the vectors $\mathbf{Y}^1, \ldots, \mathbf{Y}^n$ are orthonormal, and because $\mathbf{Y}^i = L^\top\mathbf{U}^i$, from

$$(\mathbf{Y}^i)^\top\mathbf{Y}^j = \delta_{ij},$$

we get

$$(\mathbf{U}^i)^\top LL^\top\mathbf{U}^j = \delta_{ij}, \quad 1 \leq i, j \leq n,$$

and since $A = LL^\top$, this yields

$$(\mathbf{U}^i)^\top A\mathbf{U}^j = \delta_{ij}, \quad 1 \leq i, j \leq n.$$

This suggests defining the functions $U^i \in V_a$ by

$$U^i = \sum_{k=1}^{n} \mathbf{U}^i_k w_k.$$

Then, it immediate to check that

$$a(U^i, U^j) = (\mathbf{U}^i)^\top A \mathbf{U}^j = \delta_{ij},$$

which means that the functions $(U^1, \ldots, U^n)$ form an orthonormal basis of V_a for the inner product a. The functions $U^i \in V_a$ are called *modes* (or *modal vectors*).

As a final step, let us look again for a solution of our discrete weak formulation of the problem, this time expressing the unknown solution $u(x,t)$ over the modal basis $(U^1, \ldots, U^n)$, say

$$u = \sum_{j=1}^{n} \widetilde{u}_j(t) U^j,$$

where each $\widetilde{u}_j$ is a function of t. Because

$$u = \sum_{j=1}^{n} \widetilde{u}_j(t) U^j = \sum_{j=1}^{n} \widetilde{u}_j(t) \left(\sum_{k=1}^{n} \mathbf{U}^j_k w_k \right) = \sum_{k=1}^{n} \left(\sum_{j=1}^{n} \widetilde{u}_j(t) \mathbf{U}^j_k \right) w_k,$$

if we write $\mathbf{u} = (u_1, \ldots, u_n)$ with $u_k = \sum_{j=1}^{n} \widetilde{u}_j(t) \mathbf{U}^j_k$ for $k = 1, \ldots, n$, we see that

$$\mathbf{u} = \sum_{j=1}^{n} \widetilde{u}_j \mathbf{U}^j,$$

so using the fact that

$$K \mathbf{U}^j = \omega_j^2 A \mathbf{U}^j, \quad j = 1, \ldots, n,$$

the equation

$$A \frac{d^2 \mathbf{u}}{dt^2} + K \mathbf{u} = 0$$

yields

$$\sum_{j=1}^{n} [(\widetilde{u}_j)'' + \omega_j^2 \widetilde{u}_j] A \mathbf{U}^j = 0.$$

Since A is invertible and since $(\mathbf{U}^1, \ldots, \mathbf{U}^n)$ are linearly independent, the vectors $(A\mathbf{U}^1, \ldots, A\mathbf{U}^n)$ are linearly independent, and consequently we get the system of n ODEs'

$$(\widetilde{u}_j)'' + \omega_j^2 \widetilde{u}_j = 0, \quad 1 \leq j \leq n.$$

Each of these equation has a well-known solution of the form

$$\widetilde{u}_j = A_j \cos \omega_j t + B_j \sin \omega_j t.$$

Therefore, the solution of our approximation problem is given by

$$u = \sum_{j=1}^{n} (A_j \cos \omega_j t + B_j \sin \omega_j t) U^j,$$

and the constants A_j, B_j are obtained from the intial conditions

$$u(x,0) = u_{a,0}(x), \quad 0 \leq x \leq L,$$
$$\frac{\partial u}{\partial t}(x,0) = u_{a,1}(x), \quad 0 \leq x \leq L,$$

by expressing $u_{a,0}$ and $u_{a,1}$ on the modal basis $(U^1, \ldots, U^n)$. Furthermore, the modal functions $(U^1, \ldots, U^n)$ form an orthonormal basis of V_a for the inner product a.

If we use the vector space V_N^0 of piecewise affine functions, we find that the matrices A and K are familar! Indeed,

$$A = \frac{1}{h} \begin{pmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{pmatrix}$$

and

$$K = \frac{h}{6} \begin{pmatrix} 4 & 1 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 1 & 4 \end{pmatrix}.$$

To conclude this section, let us discuss briefly the wave equation for an elastic membrane, as described in Section 14.2. This time, we look for a function $u \colon \mathbb{R}_+ \times \Omega \to \mathbb{R}$ satisfying the following conditions:

$$\begin{aligned} \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2}(x,t) - \Delta u(x,t) &= f(x,t), \quad x \in \Omega,\ t > 0, \\ u(x,t) &= 0, \quad x \in \Gamma, \quad t \geq 0 \quad \text{(boundary condition)}, \\ u(x,0) &= u_{i,0}(x), \quad x \in \Omega \quad \text{(intitial condition)}, \\ \frac{\partial u}{\partial t}(x,0) &= u_{i,1}(x), \quad x \in \Omega \quad \text{(intitial condition)}. \end{aligned}$$

Assuming that $f = 0$, we look for solutions in the subspace V of the Sobolev space $H_0^1(\overline{\Omega})$ consisting of functions v such that $v = 0$ on Γ. Multiplying by a test function $v \in V$ and using Green's first identity, we get the weak formulation of our problem:

Find a function $u \in V$ such that

$$\begin{aligned}\frac{d^2}{dt^2}\langle u, v\rangle + a(u,v) &= 0, \qquad \text{for all } v \in V \quad \text{and all } t \geq 0\\ u(x,0) &= u_{i,0}(x), \quad x \in \Omega \quad \text{(intitial condition)},\\ \frac{\partial u}{\partial t}(x,0) &= u_{i,1}(x), \quad x \in \Omega \quad \text{(intitial condition)},\end{aligned}$$

where $a\colon V \times V \to \mathbb{R}$ is the bilinear form given by

$$a(u,v) = \int_\Omega \left(\frac{\partial u}{\partial x_1}\frac{\partial v}{\partial x_1} + \frac{\partial u}{\partial x_2}\frac{\partial v}{\partial x_2}\right) dx,$$

and

$$\langle u, v\rangle = \int_\Omega uv dx.$$

As usual, we find approximations of our problem by using finite dimensional subspaces V_a of V. Picking some basis $(w_1, \ldots, w_n)$ of V_a, and triangulating Ω, as before, we obtain the equation

$$\begin{aligned}A\frac{d^2\mathbf{u}}{dt^2} + K\mathbf{u} &= 0,\\ u(x,0) &= u_{a,0}(x), \quad x \in \Gamma,\\ \frac{\partial u}{\partial t}(x,0) &= u_{a,1}(x), \quad x \in \Gamma,\end{aligned}$$

where $A = (\langle w_i, w_j\rangle)$ and $K = (a(w_i, w_j))$ are two symmetric positive definite matrices.

In principle, the problem is solved, but, it may be difficult to find good spaces V_a, good triangulations of Ω, and good bases of V_a, to be able to compute the matrices A and K, and to ensure that they are sparse.

Chapter 15

Singular Value Decomposition and Polar Form

15.1 Singular Value Decomposition for Square Matrices

In this section, we assume that we are dealing with real Euclidean spaces. Let $f\colon E \to E$ be any linear map. In general, it may not be possible to diagonalize f. We show that every linear map can be diagonalized if we are willing to use *two* orthonormal bases. This is the celebrated *singular value decomposition (SVD)*. A close cousin of the SVD is the *polar form* of a linear map, which shows how a linear map can be decomposed into its purely rotational component (perhaps with a flip) and its purely stretching part.

The key observation is that $f^* \circ f$ is self-adjoint, since

$$\langle (f^* \circ f)(u), v\rangle = \langle f(u), f(v)\rangle = \langle u, (f^* \circ f)(v)\rangle.$$

Similarly, $f \circ f^*$ is self-adjoint.

The fact that $f^* \circ f$ and $f \circ f^*$ are self-adjoint is very important, because it implies that $f^* \circ f$ and $f \circ f^*$ can be diagonalized and that they have real eigenvalues. In fact, these eigenvalues are all nonnegative. Indeed, if u is an eigenvector of $f^* \circ f$ for the eigenvalue λ, then

$$\langle (f^* \circ f)(u), u\rangle = \langle f(u), f(u)\rangle$$

and

$$\langle (f^* \circ f)(u), u\rangle = \lambda\langle u, u\rangle,$$

and thus

$$\lambda\langle u, u\rangle = \langle f(u), f(u)\rangle,$$

which implies that $\lambda \geq 0$, since $\langle -, -\rangle$ is positive definite. A similar proof applies to $f \circ f^*$. Thus, the eigenvalues of $f^* \circ f$ are of the form $\sigma_1^2, \ldots, \sigma_r^2$ or 0, where $\sigma_i > 0$, and similarly for $f \circ f^*$.

The above considerations also apply to any linear map $f\colon E \to F$ betwen two Euclidean spaces $(E, \langle -, - \rangle_1)$ and $(F, \langle -, - \rangle_2)$. Recall that the adjoint $f^*\colon F \to E$ of f is the unique linear map f^* such that

$$\langle f(u), v\rangle_2 = \langle u, f^*(v)\rangle_1, \quad \text{for all } u \in E \text{ and all } v \in F.$$

Then, $f^* \circ f$ and $f \circ f^*$ are self-adjoint (the proof is the same as in the previous case), and the eigenvalues of $f^* \circ f$ and $f \circ f^*$ are nonnegative. If λ is an eigenvalue of $f^* \circ f$ and $u (\neq 0)$ is a corresponding eigenvector, we have

$$\langle (f^* \circ f)(u), u\rangle_1 = \langle f(u), f(u)\rangle_2,$$

and also

$$\langle (f^* \circ f)(u), u\rangle_1 = \lambda \langle u, u\rangle_1,$$

so

$$\lambda \langle u, u\rangle_1, = \langle f(u), f(u)\rangle_2,$$

which implies that $\lambda \geq 0$. A similar proof applies to $f \circ f^*$. The situation is even better, since we will show shortly that $f^* \circ f$ and $f \circ f^*$ have the same nonzero eigenvalues.

Remark: Given any two linear maps $f\colon E \to F$ and $g\colon F \to E$, where $\dim(E) = n$ and $\dim(F) = m$, it can be shown that

$$\lambda^m \det(\lambda\, I_n - g \circ f) = \lambda^n \det(\lambda\, I_m - f \circ g),$$

and thus $g \circ f$ and $f \circ g$ always have the same nonzero eigenvalues!

Definition 15.1. Given any linear map $f\colon E \to F$, the square roots $\sigma_i > 0$ of the positive eigenvalues of $f^* \circ f$ (and $f \circ f^*$) are called the *singular values of* f.

Definition 15.2. A self-adjoint linear map $f\colon E \to E$ whose eigenvalues are nonnegative is called *positive semidefinite* (or *positive*), and if f is also invertible, f is said to be *positive definite*. In the latter case, every eigenvalue of f is strictly positive.

If $f\colon E \to F$ is any linear map, we just showed that $f^* \circ f$ and $f \circ f^*$ are positive semidefinite self-adjoint linear maps. This fact has the remarkable consequence that every linear map has two important decompositions:

1. The polar form.

2. The singular value decomposition (SVD).

The wonderful thing about the singular value decomposition is that there exist two orthonormal bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_m)$ such that, with respect to these bases, f is a diagonal matrix consisting of the singular values of f, or 0. Thus, in some sense, f can always be diagonalized with respect to *two* orthonormal bases. The SVD is also a useful tool for solving overdetermined linear systems in the least squares sense and for data analysis, as we show later on.

First, we show some useful relationships between the kernels and the images of f, f^*, $f^* \circ f$, and $f \circ f^*$. Recall that if $f\colon E \to F$ is a linear map, the *image* $\operatorname{Im} f$ *of* f is the subspace $f(E)$ of F, and the *rank of* f is the dimension $\dim(\operatorname{Im} f)$ of its image. Also recall that (Theorem 3.6)

$$\dim(\operatorname{Ker} f) + \dim(\operatorname{Im} f) = \dim(E),$$

and that (Propositions 9.9 and 11.10) for every subspace W of E,

$$\dim(W) + \dim(W^{\perp}) = \dim(E).$$

Proposition 15.1. *Given any two Euclidean spaces E and F, where E has dimension n and F has dimension m, for any linear map $f\colon E \to F$, we have*

$$\begin{aligned}
\operatorname{Ker} f &= \operatorname{Ker}(f^* \circ f),\\
\operatorname{Ker} f^* &= \operatorname{Ker}(f \circ f^*),\\
\operatorname{Ker} f &= (\operatorname{Im} f^*)^{\perp},\\
\operatorname{Ker} f^* &= (\operatorname{Im} f)^{\perp},\\
\dim(\operatorname{Im} f) &= \dim(\operatorname{Im} f^*),
\end{aligned}$$

and f, f^*, $f^* \circ f$, and $f \circ f^*$ have the same rank.

Proof. To simplify the notation, we will denote the inner products on E and F by the same symbol $\langle -, - \rangle$ (to avoid subscripts). If $f(u) = 0$, then $(f^* \circ f)(u) = f^*(f(u)) = f^*(0) = 0$, and so $\operatorname{Ker} f \subseteq \operatorname{Ker}(f^* \circ f)$. By definition of f^*, we have

$$\langle f(u), f(u) \rangle = \langle (f^* \circ f)(u), u \rangle$$

for all $u \in E$. If $(f^* \circ f)(u) = 0$, since $\langle -, - \rangle$ is positive definite, we must have $f(u) = 0$, and so $\operatorname{Ker}(f^* \circ f) \subseteq \operatorname{Ker} f$. Therefore,

$$\operatorname{Ker} f = \operatorname{Ker}(f^* \circ f).$$

The proof that $\operatorname{Ker} f^* = \operatorname{Ker}(f \circ f^*)$ is similar.

By definition of f^*, we have

$$\langle f(u), v \rangle = \langle u, f^*(v) \rangle \quad \text{for all } u \in E \text{ and all } v \in F. \tag{$*$}$$

This immediately implies that

$$\operatorname{Ker} f = (\operatorname{Im} f^*)^\perp \quad \text{and} \quad \operatorname{Ker} f^* = (\operatorname{Im} f)^\perp.$$

Let us explain why $\operatorname{Ker} f = (\operatorname{Im} f^*)^\perp$, the proof of the other equation being similar.

Because the inner product is positive definite, for every $u \in E$, we have
$u \in \operatorname{Ker} f$
iff $f(u) = 0$
iff $\langle f(u), v\rangle = 0$ for all v,
by $(*)$ iff $\langle u, f^*(v)\rangle = 0$ for all v,
iff $u \in (\operatorname{Im} f^*)^\perp$.

Since

$$\dim(\operatorname{Im} f) = n - \dim(\operatorname{Ker} f)$$

and

$$\dim(\operatorname{Im} f^*) = n - \dim((\operatorname{Im} f^*)^\perp),$$

from

$$\operatorname{Ker} f = (\operatorname{Im} f^*)^\perp$$

we also have

$$\dim(\operatorname{Ker} f) = \dim((\operatorname{Im} f^*)^\perp),$$

from which we obtain

$$\dim(\operatorname{Im} f) = \dim(\operatorname{Im} f^*).$$

Since

$$\dim(\operatorname{Ker}(f^* \circ f)) + \dim(\operatorname{Im}(f^* \circ f)) = \dim(E),$$

$\operatorname{Ker}(f^* \circ f) = \operatorname{Ker} f$ and $\operatorname{Ker} f = (\operatorname{Im} f^*)^\perp$, we get

$$\dim((\operatorname{Im} f^*)^\perp) + \dim(\operatorname{Im}(f^* \circ f)) = \dim(E).$$

Since

$$\dim((\operatorname{Im} f^*)^\perp) + \dim(\operatorname{Im} f^*) = \dim(E),$$

we deduce that

$$\dim(\operatorname{Im} f^*) = \dim(\operatorname{Im}(f^* \circ f)).$$

A similar proof shows that

$$\dim(\operatorname{Im} f) = \dim(\operatorname{Im}(f \circ f^*)).$$

Consequently, f, f^*, $f^* \circ f$, and $f \circ f^*$ have the same rank. □

We will now prove that every square matrix has an SVD. Stronger results can be obtained if we first consider the polar form and then derive the SVD from it (there are uniqueness properties of the polar decomposition). For our purposes, uniqueness results are not as important so we content ourselves with existence results, whose proofs are simpler. Readers interested in a more general treatment are referred to [44].

The early history of the singular value decomposition is described in a fascinating paper by Stewart [98]. The SVD is due to Beltrami and Camille Jordan independently (1873, 1874). Gauss is the grandfather of all this, for his work on least squares (1809, 1823) (but Legendre also published a paper on least squares!). Then come Sylvester, Schmidt, and Hermann Weyl. Sylvester's work was apparently "opaque." He gave a computational method to find an SVD. Schmidt's work really has to do with integral equations and symmetric and asymmetric kernels (1907). Weyl's work has to do with perturbation theory (1912). Autonne came up with the polar decomposition (1902, 1915). Eckart and Young extended SVD to rectangular matrices (1936, 1939).

Theorem 15.2. *(Singular value decomposition) For every real $n \times n$ matrix A there are two orthogonal matrices U and V and a diagonal matrix D such that $A = VDU^\top$, where D is of the form*

$$D = \begin{pmatrix} \sigma_1 & & \dots & \\ & \sigma_2 & \dots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \dots & \sigma_n \end{pmatrix},$$

where $\sigma_1, \dots, \sigma_r$ are the singular values of f, i.e., the (positive) square roots of the nonzero eigenvalues of $A^\top A$ and $A\,A^\top$, and $\sigma_{r+1} = \cdots = \sigma_n = 0$. The columns of U are eigenvectors of $A^\top A$, and the columns of V are eigenvectors of $A\,A^\top$.

Proof. Since $A^\top A$ is a symmetric matrix, in fact, a positive semidefinite matrix, there exists an orthogonal matrix U such that

$$A^\top A = UD^2U^\top,$$

with $D = \mathrm{diag}(\sigma_1, \dots, \sigma_r, 0, \dots, 0)$, where $\sigma_1^2, \dots, \sigma_r^2$ are the nonzero eigenvalues of $A^\top A$, and where r is the rank of A; that is, $\sigma_1, \dots, \sigma_r$ are the singular values of A. It follows that

$$U^\top A^\top AU = (AU)^\top AU = D^2,$$

and if we let f_j be the jth column of AU for $j = 1, \dots, n$, then we have

$$\langle f_i, f_j \rangle = \sigma_i^2 \delta_{ij}, \quad 1 \leq i, j \leq r$$

and

$$f_j = 0, \quad r+1 \leq j \leq n.$$

If we define $(v_1, \dots, v_r)$ by

$$v_j = \sigma_j^{-1} f_j, \quad 1 \leq j \leq r,$$

then we have

$$\langle v_i, v_j\rangle = \delta_{ij}, \quad 1 \leq i, j \leq r,$$

so complete $(v_1, \ldots, v_r)$ into an orthonormal basis $(v_1, \ldots, v_r, v_{r+1}, \ldots, v_n)$ (for example, using Gram–Schmidt). Now, since $f_j = \sigma_j v_j$ for $j = 1 \ldots, r$, we have

$$\langle v_i, f_j\rangle = \sigma_j \langle v_i, v_j\rangle = \sigma_j \delta_{i,j}, \quad 1 \leq i \leq n,\ 1 \leq j \leq r$$

and since $f_j = 0$ for $j = r+1, \ldots, n$,

$$\langle v_i, f_j\rangle = 0 \quad 1 \leq i \leq n,\ r+1 \leq j \leq n.$$

If V is the matrix whose columns are $v_1, \ldots, v_n$, then V is orthogonal and the above equations prove that

$$V^\top A U = D,$$

which yields $A = VDU^\top$, as required.

The equation $A = VDU^\top$ implies that

$$A^\top A = UD^2U^\top, \quad AA^\top = VD^2V^\top,$$

which shows that $A^\top A$ and $AA^\top$ have the same eigenvalues, that the columns of U are eigenvectors of $A^\top A$, and that the columns of V are eigenvectors of $AA^\top$. □

Theorem 15.2 suggests the following definition.

Definition 15.3. A triple (U, D, V) such that $A = V D U^\top$, where U and V are orthogonal and D is a diagonal matrix whose entries are nonnegative (it is positive semidefinite) is called a *singular value decomposition (SVD) of* A.

The proof of Theorem 15.2 shows that there are two orthonormal bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$, where $(u_1, \ldots, u_n)$ are eigenvectors of $A^\top A$ and $(v_1, \ldots, v_n)$ are eigenvectors of $AA^\top$. Furthermore, $(u_1, \ldots, u_r)$ is an orthonormal basis of $\operatorname{Im} A^\top$, $(u_{r+1}, \ldots, u_n)$ is an orthonormal basis of $\operatorname{Ker} A$, $(v_1, \ldots, v_r)$ is an orthonormal basis of $\operatorname{Im} A$, and $(v_{r+1}, \ldots, v_n)$ is an orthonormal basis of $\operatorname{Ker} A^\top$.

Using a remark made in Chapter 2, if we denote the columns of U by $u_1, \ldots, u_n$ and the columns of V by $v_1, \ldots, v_n$, then we can write

$$A = V D U^\top = \sigma_1 v_1 u_1^\top + \cdots + \sigma_r v_r u_r^\top.$$

As a consequence, if r is a lot smaller than n (we write $r \ll n$), we see that A can be reconstructed from U and V using a much smaller number of elements. This idea will be used to provide "low-rank" approximations of a matrix. The idea is to keep only the k top singular values for some suitable $k \ll r$ for which $\sigma_{k+1}, \ldots \sigma_r$ are very small.

Remarks:

(1) In Strang [102] the matrices U, V, D are denoted by $U = Q_2$, $V = Q_1$, and $D = \Sigma$, and an SVD is written as $A = Q_1 \Sigma Q_2^\top$. This has the advantage that Q_1 comes before Q_2 in $A = Q_1 \Sigma Q_2^\top$. This has the disadvantage that A maps the columns of Q_2 (eigenvectors of $A^\top A$) to multiples of the columns of Q_1 (eigenvectors of $A\,A^\top$).

(2) Algorithms for actually computing the SVD of a matrix are presented in Golub and Van Loan [49], Demmel [33], and Trefethen and Bau [105], where the SVD and its applications are also discussed quite extensively.

(3) The SVD also applies to complex matrices. In this case, for every complex $n \times n$ matrix A, there are two unitary matrices U and V and a diagonal matrix D such that

$$A = V D\, U^*,$$

where D is a diagonal matrix consisting of real entries $\sigma_1, \ldots, \sigma_n$, where $\sigma_1, \ldots, \sigma_r$ are the singular values of A, i.e., the positive square roots of the nonzero eigenvalues of A^*A and $A\,A^*$, and $\sigma_{r+1} = \ldots = \sigma_n = 0$.

A notion closely related to the SVD is the polar form of a matrix.

Definition 15.4. A pair (R, S) such that $A = RS$ with R orthogonal and S symmetric positive semidefinite is called a *polar decomposition of* A.

Theorem 15.2 implies that for every real $n \times n$ matrix A, there is some orthogonal matrix R and some positive semidefinite symmetric matrix S such that

$$A = RS.$$

This is easy to show and we will prove it below. Furthermore, R, S are unique if A is invertible, but this is harder to prove.

For example, the matrix

$$A = \frac{1}{2}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

is both orthogonal and symmetric, and $A = RS$ with $R = A$ and $S = I$, which implies that some of the eigenvalues of A are negative.

Remark: In the complex case, the polar decomposition states that for every complex $n \times n$ matrix A, there is some unitary matrix U and some positive semidefinite Hermitian matrix H such that

$$A = UH.$$

It is easy to go from the polar form to the SVD, and conversely.

Given an SVD decomposition $A = V D U^\top$, let $R = V U^\top$ and $S = U D U^\top$. It is clear that R is orthogonal and that S is positive semidefinite symmetric, and

$$RS = V U^\top U D U^\top = V D U^\top = A.$$

Going the other way, given a polar decomposition $A = R_1 S$, where R_1 is orthogonal and S is positive semidefinite symmetric, there is an orthogonal matrix R_2 and a positive semidefinite diagonal matrix D such that $S = R_2 D R_2^\top$, and thus

$$A = R_1 R_2 D R_2^\top = V D U^\top,$$

where $V = R_1 R_2$ and $U = R_2$ are orthogonal.

The eigenvalues and the singular values of a matrix are typically not related in any obvious way. For example, the $n \times n$ matrix

$$A = \begin{pmatrix} 1 & 2 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 2 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & 2 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 1 & 2 & 0 \\ 0 & 0 & \dots & 0 & 0 & 1 & 2 \\ 0 & 0 & \dots & 0 & 0 & 0 & 1 \end{pmatrix}$$

has the eigenvalue 1 with multiplicity n, but its singular values, $\sigma_1 \geq \cdots \geq \sigma_n$, which are the positive square roots of the eigenvalues of the matrix $B = A^\top A$ with

$$B = \begin{pmatrix} 1 & 2 & 0 & 0 & \dots & 0 & 0 \\ 2 & 5 & 2 & 0 & \dots & 0 & 0 \\ 0 & 2 & 5 & 2 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 2 & 5 & 2 & 0 \\ 0 & 0 & \dots & 0 & 2 & 5 & 2 \\ 0 & 0 & \dots & 0 & 0 & 2 & 5 \end{pmatrix}$$

have a wide spread, since

$$\frac{\sigma_1}{\sigma_n} = \mathrm{cond}_2(A) \geq 2^{n-1}.$$

If A is a complex $n \times n$ matrix, the eigenvalues $\lambda_1, \dots, \lambda_n$ and the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$ of A are not unrelated, since

$$\sigma_1^2 \cdots \sigma_n^2 = \det(A^* A) = |\det(A)|^2$$

and

$$|\lambda_1| \cdots |\lambda_n| = |\det(A)|,$$

so we have

$$|\lambda_1| \cdots |\lambda_n| = \sigma_1 \cdots \sigma_n.$$

More generally, Hermann Weyl proved the following remarkable theorem:

Theorem 15.3. *(Weyl's inequalities, 1949) For any complex $n \times n$ matrix, A, if $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ are the eigenvalues of A and $\sigma_1, \ldots, \sigma_n \in \mathbb{R}_+$ are the singular values of A, listed so that $|\lambda_1| \geq \cdots \geq |\lambda_n|$ and $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$, then*

$$\begin{aligned} &|\lambda_1| \cdots |\lambda_n| = \sigma_1 \cdots \sigma_n \quad \textit{and} \\ &|\lambda_1| \cdots |\lambda_k| \leq \sigma_1 \cdots \sigma_k, \quad \textit{for} \quad k = 1, \ldots, n-1. \end{aligned}$$

A proof of Theorem 15.3 can be found in Horn and Johnson [56], Chapter 3, Section 3.3, where more inequalities relating the eigenvalues and the singular values of a matrix are given.

Theorem 15.2 can be easily extended to rectangular $m \times n$ matrices, as we show in the next section (for various versions of the SVD for rectangular matrices, see Strang [102] Golub and Van Loan [49], Demmel [33], and Trefethen and Bau [105]).

15.2 Singular Value Decomposition for Rectangular Matrices

Here is the generalization of Theorem 15.2 to rectangular matrices.

Theorem 15.4. *(Singular value decomposition) For every real $m \times n$ matrix A, there are two orthogonal matrices U $(n \times n)$ and V $(m \times m)$ and a diagonal $m \times n$ matrix D such that $A = V D U^\top$, where D is of the form*

$$D = \begin{pmatrix} \sigma_1 & & \cdots & \\ & \sigma_2 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \cdots & \sigma_n \\ 0 & \vdots & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \vdots & \cdots & 0 \end{pmatrix} \quad \textit{or} \quad D = \begin{pmatrix} \sigma_1 & & \cdots & & 0 & \cdots & 0 \\ & \sigma_2 & \cdots & & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & 0 & \vdots & 0 \\ & & \cdots & \sigma_m & 0 & \cdots & 0 \end{pmatrix},$$

where $\sigma_1, \ldots, \sigma_r$ are the singular values of f, i.e. the (positive) square roots of the nonzero eigenvalues of $A^\top A$ and $A A^\top$, and $\sigma_{r+1} = \ldots = \sigma_p = 0$, where $p = \min(m, n)$. The columns of U are eigenvectors of $A^\top A$, and the columns of V are eigenvectors of $A A^\top$.

Proof. As in the proof of Theorem 15.2, since $A^\top A$ is symmetric positive semidefinite, there exists an $n \times n$ orthogonal matrix U such that

$$A^\top A = U\Sigma^2 U^\top,$$

with $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0)$, where $\sigma_1^2, \ldots, \sigma_r^2$ are the nonzero eigenvalues of $A^\top A$, and where r is the rank of A. Observe that $r \leq \min\{m, n\}$, and AU is an $m \times n$ matrix. It follows that

$$U^\top A^\top AU = (AU)^\top AU = \Sigma^2,$$

and if we let $f_j \in \mathbb{R}^m$ be the jth column of AU for $j = 1, \ldots, n$, then we have

$$\langle f_i, f_j \rangle = \sigma_i^2 \delta_{ij}, \quad 1 \leq i, j \leq r$$

and

$$f_j = 0, \quad r+1 \leq j \leq n.$$

If we define $(v_1, \ldots, v_r)$ by

$$v_j = \sigma_j^{-1} f_j, \quad 1 \leq j \leq r,$$

then we have

$$\langle v_i, v_j \rangle = \delta_{ij}, \quad 1 \leq i, j \leq r,$$

so complete $(v_1, \ldots, v_r)$ into an orthonormal basis $(v_1, \ldots, v_r, v_{r+1}, \ldots, v_m)$ (for example, using Gram–Schmidt).

Now, since $f_j = \sigma_j v_j$ for $j = 1 \ldots, r$, we have

$$\langle v_i, f_j \rangle = \sigma_j \langle v_i, v_j \rangle = \sigma_j \delta_{i,j}, \quad 1 \leq i \leq m,\ 1 \leq j \leq r$$

and since $f_j = 0$ for $j = r+1, \ldots, n$, we have

$$\langle v_i, f_j \rangle = 0 \quad 1 \leq i \leq m,\ r+1 \leq j \leq n.$$

If V is the matrix whose columns are $v_1, \ldots, v_m$, then V is an $m \times m$ orthogonal matrix and if $m \geq n$, we let

$$D = \begin{pmatrix} \Sigma \\ 0_{m-n} \end{pmatrix} = \begin{pmatrix} \sigma_1 & & \ldots & \\ & \sigma_2 & \ldots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \ldots & \sigma_n \\ 0 & \vdots & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \vdots & \ldots & 0 \end{pmatrix},$$

else if $n \geq m$, then we let

$$D = \begin{pmatrix} \sigma_1 & & \ldots & & 0 & \ldots & 0 \\ & \sigma_2 & \ldots & & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & 0 & \vdots & 0 \\ & & \ldots & \sigma_m & 0 & \ldots & 0 \end{pmatrix}.$$

In either case, the above equations prove that

$$V^\top AU = D,$$

which yields $A = VDU^\top$, as required.

The equation $A = VDU^\top$ implies that

$$A^\top A = UD^\top DU^\top = U\mathrm{diag}(\sigma_1^2, \ldots, \sigma_r^2, \underbrace{0, \ldots, 0}_{n-r})U^\top$$

and

$$AA^\top = VDD^\top V^\top = V\mathrm{diag}(\sigma_1^2, \ldots, \sigma_r^2, \underbrace{0, \ldots, 0}_{m-r})V^\top,$$

which shows that $A^\top A$ and $AA^\top$ have the same nonzero eigenvalues, that the columns of U are eigenvectors of $A^\top A$, and that the columns of V are eigenvectors of $AA^\top$. □

A triple (U, D, V) such that $A = VDU^\top$ is called a *singular value decomposition (SVD) of* A.

Even though the matrix D is an $m \times n$ rectangular matrix, since its only nonzero entries are on the descending diagonal, we still say that D is a diagonal matrix.

If we view A as the representation of a linear map $f\colon E \to F$, where $\dim(E) = n$ and $\dim(F) = m$, the proof of Theorem 15.4 shows that there are two orthonormal bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_m)$ for E and F, respectively, where $(u_1, \ldots, u_n)$ are eigenvectors of $f^* \circ f$ and $(v_1, \ldots, v_m)$ are eigenvectors of $f \circ f^*$. Furthermore, $(u_1, \ldots, u_r)$ is an orthonormal basis of $\mathrm{Im}\, f^*$, $(u_{r+1}, \ldots, u_n)$ is an orthonormal basis of $\mathrm{Ker}\, f$, $(v_1, \ldots, v_r)$ is an orthonormal basis of $\mathrm{Im}\, f$, and $(v_{r+1}, \ldots, v_m)$ is an orthonormal basis of $\mathrm{Ker}\, f^*$.

The SVD of matrices can be used to define the pseudo-inverse of a rectangular matrix; we will do so in Chapter 16. The reader may also consult Strang [102], Demmel [33], Trefethen and Bau [105], and Golub and Van Loan [49].

One of the spectral theorems states that a symmetric matrix can be diagonalized by an orthogonal matrix. There are several numerical methods to compute the eigenvalues of a symmetric matrix A. One method consists in *tridiagonalizing* A, which means that there exists some orthogonal matrix P and some symmetric tridiagonal matrix T such that $A = PTP^\top$. In fact, this can be done using Householder transformations. It is then possible to compute the eigenvalues of T using a bisection method based on Sturm sequences. One can also use Jacobi's method. For details, see Golub and Van Loan [49], Chapter 8, Demmel [33], Trefethen and Bau [105], Lecture 26, or Ciarlet [30]. Computing the SVD of a matrix A is more involved. Most methods begin by finding orthogonal matrices U and V and a *bidiagonal* matrix B such that $A = VBU^\top$. This can also be done using Householder transformations. Observe that $B^\top B$ is symmetric tridiagonal. Thus, in principle, the previous method to diagonalize a symmetric tridiagonal matrix can be applied. However, it is unwise to compute

$B^\top B$ explicitly, and more subtle methods are used for this last step. Again, see Golub and Van Loan [49], Chapter 8, Demmel [33], and Trefethen and Bau [105], Lecture 31.

The polar form has applications in continuum mechanics. Indeed, in any deformation it is important to separate stretching from rotation. This is exactly what QS achieves. The orthogonal part Q corresponds to rotation (perhaps with an additional reflection), and the symmetric matrix S to stretching (or compression). The real eigenvalues $\sigma_1, \ldots, \sigma_r$ of S are the stretch factors (or compression factors) (see Marsden and Hughes [71]). The fact that S can be diagonalized by an orthogonal matrix corresponds to a natural choice of axes, the principal axes.

The SVD has applications to data compression, for instance in image processing. The idea is to retain only singular values whose magnitudes are significant enough. The SVD can also be used to determine the rank of a matrix when other methods such as Gaussian elimination produce very small pivots. One of the main applications of the SVD is the computation of the pseudo-inverse. Pseudo-inverses are the key to the solution of various optimization problems, in particular the method of least squares. This topic is discussed in the next chapter (Chapter 16). Applications of the material of this chapter can be found in Strang [102, 101]; Ciarlet [30]; Golub and Van Loan [49], which contains many other references; Demmel [33]; and Trefethen and Bau [105].

15.3 Ky Fan Norms and Schatten Norms

The singular values of a matrix can be used to define various norms on matrices which have found recent applications in quantum information theory and in spectral graph theory. Following Horn and Johnson [56] (Section 3.4) we can make the following definitions:

Definition 15.5. For any matrix $A \in \mathrm{M}_{m,n}(\mathbb{C})$, let $q = \min\{m, n\}$, and if $\sigma_1 \geq \cdots \geq \sigma_q$ are the singular values of A, for any k with $1 \leq k \leq q$, let

$$N_k(A) = \sigma_1 + \cdots + \sigma_k,$$

called the *Ky Fan k-norm* of A.

More generally, for any $p \geq 1$ and any k with $1 \leq k \leq q$, let

$$N_{k;p}(A) = (\sigma_1^p + \cdots + \sigma_k^p)^{1/p},$$

called the *Ky Fan p-k-norm* of A. When $k = q$, $N_{q;p}$ is also called the *Schatten p-norm*.

Observe that when $k = 1$, $N_1(A) = \sigma_1$, and the Ky Fan norm N_1 is simply the *spectral norm* from Chapter 6, which is the subordinate matrix norm associated with the Euclidean norm. When $k = q$, the Ky Fan norm N_q is given by

$$N_q(A) = \sigma_1 + \cdots + \sigma_q = \mathrm{tr}((A^*A)^{1/2})$$

and is called the *trace norm* or *nuclear norm*. When $p = 2$ and $k = q$, the Ky Fan $N_{q;2}$ norm is given by

$$N_{k;2}(A) = (\sigma_1^2 + \cdots + \sigma_q^2)^{1/2} = \sqrt{\mathrm{tr}(A^*A)} = \|A\|_F\,,$$

which is the *Frobenius norm* of A.

It can be shown that N_k and $N_{k;p}$ are unitarily invariant norms, and that when $m = n$, they are matrix norms; see Horn and Johnson [56] (Section 3.4, Corollary 3.4.4 and Problem 3).

15.4 Summary

The main concepts and results of this chapter are listed below:

- For any linear map $f\colon E \to E$ on a Euclidean space E, the maps $f^* \circ f$ and $f \circ f^*$ are self-adjoint and positive semidefinite.
- The *singular values* of a linear map.
- *Positive semidefinite* and *positive definite* self-adjoint maps.
- Relationships between $\mathrm{Im}\, f$, $\mathrm{Ker}\, f$, $\mathrm{Im}\, f^*$, and $\mathrm{Ker}\, f^*$.
- The *singular value decomposition theorem* for square matrices (Theorem 15.2).
- The *SVD* of matrix.
- The *polar decomposition* of a matrix.
- The *Weyl inequalities*.
- The *singular value decomposition theorem* for $m \times n$ matrices (Theorem 15.4).
- Ky Fan k-norms, Ky Fan p-k-norms, Schatten p-norms.

Chapter 16

Applications of SVD and Pseudo-Inverses

> De tous les principes qu'on peut proposer pour cet objet, je pense qu'il n'en est pas de plus général, de plus exact, ni d'une application plus facile, que celui dont nous avons fait usage dans les recherches précédentes, et qui consiste à rendre *minimum* la somme des carrés des erreurs. Par ce moyen il s'établit entre les erreurs une sorte d'équilibre qui, empêchant les extrêmes de prévaloir, est très propre à faire connaitre l'état du système le plus proche de la vérité.
>
> —**Legendre, 1805,** *Nouvelles Méthodes pour la détermination des Orbites des Comètes*

16.1 Least Squares Problems and the Pseudo-Inverse

This chapter presents several applications of SVD. The first one is the pseudo-inverse, which plays a crucial role in solving linear systems by the method of least squares. The second application is data compression. The third application is principal component analysis (PCA), whose purpose is to identify patterns in data and understand the variance–covariance structure of the data. The fourth application is the best affine approximation of a set of data, a problem closely related to PCA.

The method of least squares is a way of "solving" an overdetermined system of linear equations

$$Ax = b,$$

i.e., a system in which A is a rectangular $m \times n$ matrix with more equations than unknowns (when $m > n$). Historically, the method of least squares was used by Gauss and Legendre to solve problems in astronomy and geodesy. The method was first published by Legendre in 1805 in a paper on methods for determining the orbits of comets. However, Gauss had already used the method of least squares as early as 1801 to determine the orbit of the asteroid

Ceres, and he published a paper about it in 1810 after the discovery of the asteroid Pallas. Incidentally, it is in that same paper that Gaussian elimination using pivots is introduced.

The reason why more equations than unknowns arise in such problems is that repeated measurements are taken to minimize errors. This produces an overdetermined and often inconsistent system of linear equations. For example, Gauss solved a system of eleven equations in six unknowns to determine the orbit of the asteroid Pallas. As a concrete illustration, suppose that we observe the motion of a small object, assimilated to a point, in the plane. From our observations, we suspect that this point moves along a straight line, say of equation $y = dx + c$. Suppose that we observed the moving point at three different locations (x_1, y_1), (x_2, y_2), and (x_3, y_3). Then we should have

$$\begin{aligned} c + dx_1 &= y_1, \\ c + dx_2 &= y_2, \\ c + dx_3 &= y_3. \end{aligned}$$

If there were no errors in our measurements, these equations would be compatible, and c and d would be determined by only two of the equations. However, in the presence of errors, the system may be inconsistent. Yet we would like to find c and d!

The idea of the method of least squares is to determine (c, d) such that it minimizes the sum of the squares of the errors, namely,

$$(c + dx_1 - y_1)^2 + (c + dx_2 - y_2)^2 + (c + dx_3 - y_3)^2.$$

In general, for an overdetermined $m \times n$ system $Ax = b$, what Gauss and Legendre discovered is that there are solutions x minimizing

$$\|Ax - b\|_2^2$$

(where $\|u\|_2^2 = u_1^2 + \cdots + u_n^2$, the square of the Euclidean norm of the vector $u = (u_1, \ldots, u_n)$), and that these solutions are given by the square $n \times n$ system

$$A^\top Ax = A^\top b,$$

called the *normal equations*. Furthermore, when the columns of A are linearly independent, it turns out that $A^\top A$ is invertible, and so x is unique and given by

$$x = (A^\top A)^{-1} A^\top b.$$

Note that $A^\top A$ is a symmetric matrix, one of the nice features of the normal equations of a least squares problem. For instance, the normal equations for the above problem are

$$\begin{pmatrix} 3 & x_1 + x_2 + x_3 \\ x_1 + x_2 + x_3 & x_1^2 + x_2^2 + x_3^2 \end{pmatrix} \begin{pmatrix} c \\ d \end{pmatrix} = \begin{pmatrix} y_1 + y_2 + y_3 \\ x_1 y_1 + x_2 y_2 + x_3 y_3 \end{pmatrix}.$$

In fact, given any real $m \times n$ matrix A, there is always a unique x^+ of minimum norm that minimizes $\|Ax - b\|_2^2$, even when the columns of A are linearly dependent. How do we prove this, and how do we find x^+?

Theorem 16.1. *Every linear system $Ax = b$, where A is an $m \times n$ matrix, has a unique least squares solution x^+ of smallest norm.*

Proof. Geometry offers a nice proof of the existence and uniqueness of x^+. Indeed, we can interpret b as a point in the Euclidean (affine) space $\mathbb{R}^m$, and the image subspace of A (also called the column space of A) as a subspace U of $\mathbb{R}^m$ (passing through the origin). Then, it is clear that

$$\inf_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 = \inf_{y \in U} \|y - b\|_2^2,$$

with $U = \operatorname{Im} A$, and we claim that x minimizes $\|Ax - b\|_2^2$ iff $Ax = p$, where p the orthogonal projection of b onto the subspace U.

Recall from Section 10.1 that the orthogonal projection $p_U \colon U \oplus U^\perp \to U$ is the linear map given by

$$p_U(u + v) = u,$$

with $u \in U$ and $v \in U^\perp$. If we let $p = p_U(b) \in U$, then for any point $y \in U$, the vectors $\overrightarrow{py} = y - p \in U$ and $\overrightarrow{bp} = p - b \in U^\perp$ are orthogonal, which implies that

$$\|\overrightarrow{by}\|_2^2 = \|\overrightarrow{bp}\|_2^2 + \|\overrightarrow{py}\|_2^2,$$

where $\overrightarrow{by} = y - b$. Thus, p is indeed the unique point in U that minimizes the distance from b to any point in U.

Thus, the problem has been reduced to proving that there is a unique x^+ of minimum norm such that $Ax^+ = p$, with $p = p_U(b) \in U$, the orthogonal projection of b onto U. We use the fact that

$$\mathbb{R}^n = \operatorname{Ker} A \oplus (\operatorname{Ker} A)^\perp.$$

Consequently, every $x \in \mathbb{R}^n$ can be written uniquely as $x = u + v$, where $u \in \operatorname{Ker} A$ and $v \in (\operatorname{Ker} A)^\perp$, and since u and v are orthogonal,

$$\|x\|_2^2 = \|u\|_2^2 + \|v\|_2^2.$$

Furthermore, since $u \in \operatorname{Ker} A$, we have $Au = 0$, and thus $Ax = p$ iff $Av = p$, which shows that the solutions of $Ax = p$ for which x has minimum norm must belong to $(\operatorname{Ker} A)^\perp$. However, the restriction of A to $(\operatorname{Ker} A)^\perp$ is injective. This is because if $Av_1 = Av_2$, where $v_1, v_2 \in (\operatorname{Ker} A)^\perp$, then $A(v_2 - v_2) = 0$, which implies $v_2 - v_1 \in \operatorname{Ker} A$, and since $v_1, v_2 \in (\operatorname{Ker} A)^\perp$, we also have $v_2 - v_1 \in (\operatorname{Ker} A)^\perp$, and consequently, $v_2 - v_1 = 0$. This shows that there is a unique x^+ of minimum norm such that $Ax^+ = p$, and that x^+ must belong to $(\operatorname{Ker} A)^\perp$. By our previous reasoning, x^+ is the unique vector of minimum norm minimizing $\|Ax - b\|_2^2$. □

The proof also shows that x minimizes $\|Ax - b\|_2^2$ iff $\overrightarrow{pb} = b - Ax$ is orthogonal to U, which can be expressed by saying that $b - Ax$ is orthogonal to every column of A. However, this is equivalent to

$$A^\top (b - Ax) = 0, \quad \text{i.e.,} \quad A^\top Ax = A^\top b.$$

Finally, it turns out that the minimum norm least squares solution x^+ can be found in terms of the pseudo-inverse A^+ of A, which is itself obtained from any SVD of A.

Definition 16.1. Given any nonzero $m \times n$ matrix A of rank r, if $A = VDU^\top$ is an SVD of A such that

$$D = \begin{pmatrix} \Lambda & 0_{r,n-r} \\ 0_{m-r,r} & 0_{m-r,n-r} \end{pmatrix},$$

with

$$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$$

an $r \times r$ diagonal matrix consisting of the nonzero singular values of A, then if we let D^+ be the $n \times m$ matrix

$$D^+ = \begin{pmatrix} \Lambda^{-1} & 0_{r,m-r} \\ 0_{n-r,r} & 0_{n-r,m-r} \end{pmatrix},$$

with

$$\Lambda^{-1} = \mathrm{diag}(1/\lambda_1, \ldots, 1/\lambda_r),$$

the *pseudo-inverse* of A is defined by

$$A^+ = UD^+V^\top.$$

If $A = 0_{m,n}$ is the zero matrix, we set $A^+ = 0_{n,m}$. Observe that D^+ is obtained from D by inverting the nonzero diagonal entries of D, leaving all zeros in place, and then transposing the matrix. The pseudo-inverse of a matrix is also known as the *Moore–Penrose pseudo-inverse*.

Actually, it seems that A^+ depends on the specific choice of U and V in an SVD (U, D, V) for A, but the next theorem shows that this is not so.

Theorem 16.2. *The least squares solution of smallest norm of the linear system $Ax = b$, where A is an $m \times n$ matrix, is given by*

$$x^+ = A^+b = UD^+V^\top b.$$

Proof. First, assume that A is a (rectangular) diagonal matrix D, as above. Then, since x minimizes $\|Dx - b\|_2^2$ iff Dx is the projection of b onto the image subspace F of D, it is fairly obvious that $x^+ = D^+b$. Otherwise, we can write

$$A = VDU^\top,$$

where U and V are orthogonal. However, since V is an isometry,

$$\|Ax - b\|_2 = \|VDU^\top x - b\|_2 = \|DU^\top x - V^\top b\|_2.$$

Letting $y = U^\top x$, we have $\|x\|_2 = \|y\|_2$, since U is an isometry, and since U is surjective, $\|Ax - b\|_2$ is minimized iff $\|Dy - V^\top b\|_2$ is minimized, and we have shown that the least solution is

$$y^+ = D^+ V^\top b.$$

Since $y = U^\top x$, with $\|x\|_2 = \|y\|_2$, we get

$$x^+ = UD^+ V^\top b = A^+ b.$$

Thus, the pseudo-inverse provides the optimal solution to the least squares problem. □

By Proposition 16.2 and Theorem 16.1, $A^+ b$ is uniquely defined by every b, and thus A^+ depends only on A.

Proposition 16.3. *When A has full rank, the pseudo-inverse A^+ can be expressed as $A^+ = (A^\top A)^{-1} A^\top$ when $m \geq n$, and as $A^+ = A^\top (AA^\top)^{-1}$ when $n \geq m$. In the first case ($m \geq n$), observe that $A^+ A = I$, so A^+ is a left inverse of A; in the second case ($n \geq m$), we have $AA^+ = I$, so A^+ is a right inverse of A.*

Proof. If $m \geq n$ and A has full rank rank n, we have

$$A = V \begin{pmatrix} \Lambda \\ 0_{m-n,n} \end{pmatrix} U^\top$$

with Λ an $n \times n$ diagonal invertible matrix (with positive entries), so

$$A^+ = U \begin{pmatrix} \Lambda^{-1} & 0_{n,m-n} \end{pmatrix} V^\top.$$

We find that

$$A^\top A = U \begin{pmatrix} \Lambda & 0_{n,m-n} \end{pmatrix} V^\top V \begin{pmatrix} \Lambda \\ 0_{m-n,n} \end{pmatrix} U^\top = U\Lambda^2 U^\top,$$

which yields

$$(A^\top A)^{-1} A^\top = U\Lambda^{-2} U^\top U \begin{pmatrix} \Lambda & 0_{n,m-n} \end{pmatrix} V^\top V = U \begin{pmatrix} \Lambda^{-1} & 0_{n,m-n} \end{pmatrix} V^\top = A^+.$$

Therefore, if $m \geq n$ and A has full rank rank n, then

$$A^+ = (A^\top A)^{-1} A^\top.$$

If $n \geq m$ and A has full rank rank m, then

$$A = V \begin{pmatrix} \Lambda & 0_{m,n-m} \end{pmatrix} U^\top$$

with Λ an $m \times m$ diagonal invertible matrix (with positive entries), so

$$A^+ = U \begin{pmatrix} \Lambda^{-1} \\ 0_{n-m,m} \end{pmatrix} V^\top.$$

We find that

$$AA^\top = V\begin{pmatrix}\Lambda & 0_{m,n-m}\end{pmatrix}U^\top U\begin{pmatrix}\Lambda\\ 0_{n-m,m}\end{pmatrix}V^\top = V\Lambda^2V^\top,$$

which yields

$$A^\top(AA^\top)^{-1} = U\begin{pmatrix}\Lambda\\ 0_{n-m,m}\end{pmatrix}V^\top V\Lambda^{-2}V^\top = U\begin{pmatrix}\Lambda^{-1}\\ 0_{n-m,m}\end{pmatrix}V^\top = A^+.$$

Therefore, if $n \geq m$ and A has full rank rank m, then $A^+ = A^\top(AA^\top)^{-1}$. □

16.2 Properties of the Pseudo-Inverse

Let $A = V\Sigma U^\top$ be an SVD for any $m \times n$ matrix A. It is easy to check that

$$\begin{aligned} AA^+A &= A,\\ A^+AA^+ &= A^+,\end{aligned}$$

and both AA^+ and A^+A are symmetric matrices. In fact,

$$AA^+ = V\Sigma U^\top U\Sigma^+V^\top = V\Sigma\Sigma^+V^\top = V\begin{pmatrix}I_r & 0\\ 0 & 0_{m-r}\end{pmatrix}V^\top$$

and

$$A^+A = U\Sigma^+V^\top V\Sigma U^\top = U\Sigma^+\Sigma U^\top = U\begin{pmatrix}I_r & 0\\ 0 & 0_{n-r}\end{pmatrix}U^\top.$$

We immediately get

$$\begin{aligned}(AA^+)^2 &= AA^+,\\ (A^+A)^2 &= A^+A,\end{aligned}$$

so both AA^+ and A^+A are orthogonal projections (since they are both symmetric).

Proposition 16.4. *The matrix AA^+ is the orthogonal projection onto the range of A and A^+A is the orthogonal projection onto* $\mathrm{Ker}(A)^\perp = \mathrm{Im}(A^\top)$*, the range of $A^\top$.*

Proof. Obviously, we have $\mathrm{range}(AA^+) \subseteq \mathrm{range}(A)$, and for any $y = Ax \in \mathrm{range}(A)$, since $AA^+A = A$, we have

$$AA^+y = AA^+Ax = Ax = y,$$

so the image of AA^+ is indeed the range of A. It is also clear that $\mathrm{Ker}(A) \subseteq \mathrm{Ker}(A^+A)$, and since $AA^+A = A$, we also have $\mathrm{Ker}(A^+A) \subseteq \mathrm{Ker}(A)$, and so

$$\mathrm{Ker}(A^+A) = \mathrm{Ker}(A).$$

Since A^+A is symmetric, $\mathrm{range}(A^+A) = \mathrm{range}((A^+A)^\top) = \mathrm{Ker}(A^+A)^\perp = \mathrm{Ker}(A)^\perp$, as claimed. □

Proposition 16.5. *The set* $\mathrm{range}(A) = \mathrm{range}(AA^+)$ *consists of all vectors* $y \in \mathbb{R}^m$ *such that*

$$V^\top y = \begin{pmatrix} z \\ 0 \end{pmatrix},$$

with $z \in \mathbb{R}^r$.

Proof. Indeed, if $y = Ax$, then

$$V^\top y = V^\top Ax = V^\top V \Sigma U^\top x = \Sigma U^\top x = \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0_{m-r} \end{pmatrix} U^\top x = \begin{pmatrix} z \\ 0 \end{pmatrix},$$

where Σ_r is the $r \times r$ diagonal matrix $\mathrm{diag}(\sigma_1, \ldots, \sigma_r)$. Conversely, if $V^\top y = \left(\begin{smallmatrix} z \\ 0 \end{smallmatrix}\right)$, then $y = V \left(\begin{smallmatrix} z \\ 0 \end{smallmatrix}\right)$, and

$$\begin{aligned} AA^+ y &= V \begin{pmatrix} I_r & 0 \\ 0 & 0_{m-r} \end{pmatrix} V^\top y \\ &= V \begin{pmatrix} I_r & 0 \\ 0 & 0_{m-r} \end{pmatrix} V^\top V \begin{pmatrix} z \\ 0 \end{pmatrix} \\ &= V \begin{pmatrix} I_r & 0 \\ 0 & 0_{m-r} \end{pmatrix} \begin{pmatrix} z \\ 0 \end{pmatrix} \\ &= V \begin{pmatrix} z \\ 0 \end{pmatrix} = y, \end{aligned}$$

which shows that y belongs to the range of A. □

Similarly, we have the following result.

Proposition 16.6. *The set* $\mathrm{range}(A^+A) = \mathrm{Ker}(A)^\perp$ *consists of all vectors* $y \in \mathbb{R}^n$ *such that*

$$U^\top y = \begin{pmatrix} z \\ 0 \end{pmatrix},$$

with $z \in \mathbb{R}^r$.

Proof. If $y = A^+Au$, then

$$y = A^+Au = U \begin{pmatrix} I_r & 0 \\ 0 & 0_{n-r} \end{pmatrix} U^\top u = U \begin{pmatrix} z \\ 0 \end{pmatrix},$$

for some $z \in \mathbb{R}^r$. Conversely, if $U^\top y = \left(\begin{smallmatrix} z \\ 0 \end{smallmatrix}\right)$, then $y = U \left(\begin{smallmatrix} z \\ 0 \end{smallmatrix}\right)$, and so

$$\begin{aligned} A^+AU \begin{pmatrix} z \\ 0 \end{pmatrix} &= U \begin{pmatrix} I_r & 0 \\ 0 & 0_{n-r} \end{pmatrix} U^\top U \begin{pmatrix} z \\ 0 \end{pmatrix} \\ &= U \begin{pmatrix} I_r & 0 \\ 0 & 0_{n-r} \end{pmatrix} \begin{pmatrix} z \\ 0 \end{pmatrix} \\ &= U \begin{pmatrix} z \\ 0 \end{pmatrix} = y, \end{aligned}$$

which shows that $y \in \mathrm{range}(A^+A)$. □

If A is a symmetric matrix, then in general, there is no SVD $V\Sigma U^\top$ of A with $V = U$. However, if A is positive semidefinite, then the eigenvalues of A are nonnegative, and so the nonzero eigenvalues of A are equal to the singular values of A and SVDs of A are of the form

$$A = V\Sigma V^\top.$$

Analogous results hold for complex matrices, but in this case, V and U are unitary matrices and AA^+ and A^+A are Hermitian orthogonal projections.

If A is a normal matrix, which means that $AA^\top = A^\top A$, then there is an intimate relationship between SVD's of A and block diagonalizations of A. As a consequence, the pseudo-inverse of a normal matrix A can be obtained directly from a block diagonalization of A.

If A is a (real) normal matrix, then we know from Theorem 13.16 that A can be block diagonalized with respect to an orthogonal matrix U as

$$A = U\Lambda U^\top,$$

where Λ is the (real) block diagonal matrix

$$\Lambda = \mathrm{diag}(B_1, \ldots, B_n),$$

consisting either of 2×2 blocks of the form

$$B_j = \begin{pmatrix} \lambda_j & -\mu_j \\ \mu_j & \lambda_j \end{pmatrix}$$

with $\mu_j \neq 0$, or of one-dimensional blocks $B_k = (\lambda_k)$. Then we have the following proposition:

Proposition 16.7. *For any (real) normal matrix A and any block diagonalization $A = U\Lambda U^\top$ of A as above, the pseudo-inverse of A is given by*

$$A^+ = U\Lambda^+ U^\top,$$

where Λ^+ is the pseudo-inverse of Λ. Furthermore, if

$$\Lambda = \begin{pmatrix} \Lambda_r & 0 \\ 0 & 0 \end{pmatrix},$$

where Λ_r has rank r, then

$$\Lambda^+ = \begin{pmatrix} \Lambda_r^{-1} & 0 \\ 0 & 0 \end{pmatrix}.$$

Proof. Assume that $B_1, \ldots, B_p$ are 2×2 blocks and that $\lambda_{2p+1}, \ldots, \lambda_n$ are the scalar entries. We know that the numbers $\lambda_j \pm i\mu_j$, and the λ_{2p+k} are the eigenvalues of A. Let $\rho_{2j-1} =$

$\rho_{2j} = \sqrt{\lambda_j^2 + \mu_j^2}$ for $j = 1, \ldots, p$, $\rho_{2p+j} = \lambda_j$ for $j = 1, \ldots, n-2p$, and assume that the blocks are ordered so that $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_n$. Then it is easy to see that

$$UU^\top = U^\top U = U\Lambda U^\top U\Lambda^\top U^\top = U\Lambda\Lambda^\top U^\top,$$

with

$$\Lambda\Lambda^\top = \mathrm{diag}(\rho_1^2, \ldots, \rho_n^2),$$

so the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$ of A, which are the nonnegative square roots of the eigenvalues of $AA^\top$, are such that

$$\sigma_j = \rho_j, \quad 1 \leq j \leq n.$$

We can define the diagonal matrices

$$\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0),$$

where $r = \mathrm{rank}(A)$, $\sigma_1 \geq \cdots \geq \sigma_r > 0$ and

$$\Theta = \mathrm{diag}(\sigma_1^{-1}B_1, \ldots, \sigma_{2p}^{-1}B_p, 1, \ldots, 1),$$

so that Θ is an orthogonal matrix and

$$\Lambda = \Theta\Sigma = (B_1, \ldots, B_p, \lambda_{2p+1}, \ldots, \lambda_r, 0, \ldots, 0).$$

But then we can write

$$A = U\Lambda U^\top = U\Theta\Sigma U^\top,$$

and we if let $V = U\Theta$, since U is orthogonal and Θ is also orthogonal, V is also orthogonal and $A = V\Sigma U^\top$ *is an SVD for* A. Now we get

$$A^+ = U\Sigma^+ V^\top = U\Sigma^+\Theta^\top U^\top.$$

However, since Θ is an orthogonal matrix, $\Theta^\top = \Theta^{-1}$, and a simple calculation shows that

$$\Sigma^+\Theta^\top = \Sigma^+\Theta^{-1} = \Lambda^+,$$

which yields the formula

$$A^+ = U\Lambda^+ U^\top.$$

Also observe that if we write

$$\Lambda_r = (B_1, \ldots, B_p, \lambda_{2p+1}, \ldots, \lambda_r),$$

then Λ_r is invertible and

$$\Lambda^+ = \begin{pmatrix} \Lambda_r^{-1} & 0 \\ 0 & 0 \end{pmatrix}.$$

Therefore, the pseudo-inverse of a normal matrix can be computed directly from any block diagonalization of A, as claimed. □

The following properties, due to Penrose, characterize the pseudo-inverse of a matrix. We have already proved that the pseudo-inverse satisfies these equations. For a proof of the converse, see Kincaid and Cheney [59].

Proposition 16.8. *Given any $m \times n$ matrix A (real or complex), the pseudo-inverse A^+ of A is the unique $n \times m$ matrix satisfying the following properties:*

$$\begin{aligned} AA^+A &= A,\\ A^+AA^+ &= A^+,\\ (AA^+)^\top &= AA^+,\\ (A^+A)^\top &= A^+A. \end{aligned}$$

If A is an $m \times n$ matrix of rank n (and so $m \geq n$), it is immediately shown that the QR-decomposition in terms of Householder transformations applies as follows:

There are n $m \times m$ matrices $H_1, \ldots, H_n$, Householder matrices or the identity, and an upper triangular $m \times n$ matrix R of rank n such that

$$A = H_1 \cdots H_n R.$$

Then, because each H_i is an isometry,

$$\|Ax - b\|_2 = \|Rx - H_n \cdots H_1 b\|_2,$$

and the least squares problem $Ax = b$ is equivalent to the system

$$Rx = H_n \cdots H_1 b.$$

Now, the system

$$Rx = H_n \cdots H_1 b$$

is of the form

$$\begin{pmatrix} R_1 \\ 0_{m-n} \end{pmatrix} x = \begin{pmatrix} c \\ d \end{pmatrix},$$

where R_1 is an invertible $n \times n$ matrix (since A has rank n), $c \in \mathbb{R}^n$, and $d \in \mathbb{R}^{m-n}$, and the least squares solution of smallest norm is

$$x^+ = R_1^{-1} c.$$

Since R_1 is a triangular matrix, it is very easy to invert R_1.

The method of least squares is one of the most effective tools of the mathematical sciences. There are entire books devoted to it. Readers are advised to consult Strang [102], Golub and Van Loan [49], Demmel [33], and Trefethen and Bau [105], where extensions and applications of least squares (such as weighted least squares and recursive least squares) are described. Golub and Van Loan [49] also contains a very extensive bibliography, including a list of books on least squares.

16.3 Data Compression and SVD

Among the many applications of SVD, a very useful one is *data compression*, notably for images. In order to make precise the notion of closeness of matrices, we use the notion of *matrix norm*. This concept is defined in Chapter 6 and the reader may want to review it before reading any further.

Given an $m \times n$ matrix of rank r, we would like to find a best approximation of A by a matrix B of rank $k \leq r$ (actually, $k < r$) so that $\|A - B\|_2$ (or $\|A - B\|_F$) is minimized.

Proposition 16.9. *Let A be an $m \times n$ matrix of rank r and let $VDU^\top = A$ be an SVD for A. Write u_i for the columns of U, v_i for the columns of V, and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p$ for the singular values of A ($p = \min(m, n)$). Then a matrix of rank $k < r$ closest to A (in the $\| \|_2$ norm) is given by*

$$A_k = \sum_{i=1}^{k} \sigma_i v_i u_i^\top = V \operatorname{diag}(\sigma_1, \ldots, \sigma_k) U^\top$$

and $\|A - A_k\|_2 = \sigma_{k+1}$.

Proof. By construction, A_k has rank k, and we have

$$\|A - A_k\|_2 = \Big\| \sum_{i=k+1}^{p} \sigma_i v_i u_i^\top \Big\|_2 = \big\| V \operatorname{diag}(0, \ldots, 0, \sigma_{k+1}, \ldots, \sigma_p) U^\top \big\|_2 = \sigma_{k+1}.$$

It remains to show that $\|A - B\|_2 \geq \sigma_{k+1}$ for all rank-k matrices B. Let B be any rank-k matrix, so its kernel has dimension $n - k$. The subspace U_{k+1} spanned by $(u_1, \ldots, u_{k+1})$ has dimension $k + 1$, and because the sum of the dimensions of the kernel of B and of U_{k+1} is $(n - k) + k + 1 = n + 1$, these two subspaces must intersect in a subspace of dimension at least 1. Pick any unit vector h in $\operatorname{Ker}(B) \cap U_{k+1}$. Then since $Bh = 0$, we have

$$\|A - B\|_2^2 \geq \|(A - B)h\|_2^2 = \|Ah\|_2^2 = \big\|VDU^\top h\big\|_2^2 = \big\|DU^\top h\big\|_2^2 \geq \sigma_{k+1}^2 \big\|U^\top h\big\|_2^2 = \sigma_{k+1}^2,$$

which proves our claim. □

Note that A_k can be stored using $(m + n)k$ entries, as opposed to mn entries. When $k \ll m$, this is a substantial gain.

A nice example of the use of Proposition 16.9 in image compression is given in Demmel [33], Chapter 3, Section 3.2.3, pages 113–115; see the Matlab demo.

An interesting topic that we have not addressed is the actual computation of an SVD. This is a very interesting but tricky subject. Most methods reduce the computation of an SVD to the diagonalization of a well-chosen symmetric matrix (which is not $A^\top A$). Interested readers should read Section 5.4 of Demmel's excellent book [33], which contains an overview of most known methods and an extensive list of references.

16.4 Principal Components Analysis (PCA)

Suppose we have a set of data consisting of n points $X_1, \ldots, X_n$, with each $X_i \in \mathbb{R}^d$ *viewed as a row vector*.

Think of the X_i's as persons, and if $X_i = (x_{i\,1}, \ldots, x_{i\,d})$, each $x_{i\,j}$ is the value of some *feature* (or *attribute*) of that person. For example, the X_i's could be mathematicians, $d = 2$, and the first component, $x_{i\,1}$, of X_i could be the year that X_i was born, and the second component, $x_{i\,2}$, the length of the beard of X_i in centimeters. Here is a small data set:

Name	year	length
Carl Friedrich Gauss	1777	0
Camille Jordan	1838	12
Adrien-Marie Legendre	1752	0
Bernhard Riemann	1826	15
David Hilbert	1862	2
Henri Poincaré	1854	5
Emmy Noether	1882	0
Karl Weierstrass	1815	0
Eugenio Beltrami	1835	2
Hermann Schwarz	1843	20

We usually form the $n \times d$ matrix X whose ith row is X_i, with $1 \leq i \leq n$. Then the jth column is denoted by C_j $(1 \leq j \leq d)$. It is sometimes called a *feature vector*, but this terminology is far from being universally accepted. In fact, many people in computer vision call the data points X_i feature vectors!

The purpose of *principal components analysis*, for short *PCA*, is to identify patterns in data and understand the *variance–covariance* structure of the data. This is useful for the following tasks:

1. Data reduction: Often much of the variability of the data can be accounted for by a smaller number of *principal components*.

2. Interpretation: PCA can show relationships that were not previously suspected.

Given a vector (a *sample* of measurements) $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, recall that the *mean* (or *average*) $\overline{x}$ of x is given by

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

We let $x - \overline{x}$ denote the *centered data point*

$$x - \overline{x} = (x_1 - \overline{x}, \ldots, x_n - \overline{x}).$$

In order to *measure the spread* of the x_i's around the mean, we define the *sample variance* (for short, *variance*) $\mathrm{var}(x)$ (or s^2) of the sample x by

$$\mathrm{var}(x) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}.$$

There is a reason for using $n-1$ instead of n. The above definition makes $\mathrm{var}(x)$ an unbiased estimator of the variance of the random variable being sampled. However, we don't need to worry about this. Curious readers will find an explanation of these peculiar definitions in Epstein [39] (Chapter 14, Section 14.5), or in any decent statistics book.

Given two vectors $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$, the *sample covariance* (for short, *covariance*) of x and y is given by

$$\mathrm{cov}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}.$$

The covariance of x and y measures how x and y vary from the mean with respect to each other. Obviously, $\mathrm{cov}(x, y) = \mathrm{cov}(y, x)$ and $\mathrm{cov}(x, x) = \mathrm{var}(x)$.

Note that

$$\mathrm{cov}(x, y) = \frac{(x - \overline{x})^\top (y - \overline{y})}{n-1}.$$

We say that x and y are *uncorrelated* iff $\mathrm{cov}(x, y) = 0$.

Finally, given an $n \times d$ matrix X of n points X_i, for PCA to be meaningful, it will be necessary to translate the origin to the *centroid* (or *center of gravity*) μ of the X_i's, defined by

$$\mu = \frac{1}{n}(X_1 + \cdots + X_n).$$

Observe that if $\mu = (\mu_1, \ldots, \mu_d)$, then μ_j is the mean of the vector C_j (the jth column of X).

We let $X - \mu$ denote the *matrix* whose ith row is the centered data point $X_i - \mu$ ($1 \leq i \leq n$). Then, the *sample covariance matrix* (for short, *covariance matrix*) of X is the $d \times d$ symmetric matrix

$$\Sigma = \frac{1}{n-1}(X - \mu)^\top (X - \mu) = (\mathrm{cov}(C_i, C_j)).$$

Remark: The factor $\frac{1}{n-1}$ is irrelevant for our purposes and can be ignored.

Here is the matrix $X - \mu$ in the case of our bearded mathematicians: Since

$$\mu_1 = 1828.4, \quad \mu_2 = 5.6,$$

we get

Name	year	length
Carl Friedrich Gauss	−51.4	−5.6
Camille Jordan	9.6	6.4
Adrien-Marie Legendre	−76.4	−5.6
Bernhard Riemann	−2.4	9.4
David Hilbert	33.6	−3.6
Henri Poincaré	25.6	−0.6
Emmy Noether	53.6	−5.6
Karl Weierstrass	13.4	−5.6
Eugenio Beltrami	6.6	−3.6
Hermann Schwarz	14.6	14.4

We can think of the vector C_j as representing the features of X in the direction e_j (the jth canonical basis vector in $\mathbb{R}^d$, namely $e_j = (0, \ldots, 1, \ldots 0)$, with a 1 in the jth position).

If $v \in \mathbb{R}^d$ is a unit vector, we wish to consider the projection of the data points $X_1, \ldots, X_n$ onto the line spanned by v. Recall from Euclidean geometry that if $x \in \mathbb{R}^d$ is any vector and $v \in \mathbb{R}^d$ is a unit vector, the projection of x onto the line spanned by v is

$$\langle x, v\rangle v.$$

Thus, with respect to the basis v, the projection of x has coordinate $\langle x, v\rangle$. If x is represented by a row vector and v by a column vector, then

$$\langle x, v\rangle = xv.$$

Therefore, the vector $Y \in \mathbb{R}^n$ consisting of the coordinates of the projections of $X_1, \ldots, X_n$ onto the line spanned by v is given by $Y = Xv$, and this is the linear combination

$$Xv = v_1 C_1 + \cdots + v_d C_d$$

of the columns of X (with $v = (v_1, \ldots, v_d)$).

Observe that because μ_j is the mean of the vector C_j (the jth column of X), we get

$$\overline{Y} = \overline{Xv} = v_1\mu_1 + \cdots + v_d\mu_d,$$

and so the centered point $Y - \overline{Y}$ is given by

$$Y - \overline{Y} = v_1(C_1 - \mu_1) + \cdots + v_d(C_d - \mu_d) = (X - \mu)v.$$

Furthermore, if $Y = Xv$ and $Z = Xw$, then

$$\begin{aligned} \mathrm{cov}(Y, Z) &= \frac{((X - \mu)v)^\top (X - \mu)w}{n - 1} \\ &= v^\top \frac{1}{n-1}(X - \mu)^\top (X - \mu)w \\ &= v^\top \Sigma w, \end{aligned}$$

where Σ is the covariance matrix of X. Since $Y - \overline{Y}$ has zero mean, we have

$$\mathrm{var}(Y) = \mathrm{var}(Y - \overline{Y}) = v^\top \frac{1}{n-1}(X-\mu)^\top (X-\mu)v.$$

The above suggests that we should move the origin to the centroid μ of the X_i's and consider the matrix $X - \mu$ of the centered data points $X_i - \mu$.

From now on, beware that we denote the columns of $X - \mu$ by $C_1, \ldots, C_d$ and that Y denotes the *centered* point $Y = (X-\mu)v = \sum_{j=1}^d v_j C_j$, where v is a unit vector.

Basic idea of PCA: The principal components of X are *uncorrelated* projections Y of the data points X_1, ..., X_n onto some directions v (where the v's are unit vectors) such that $\mathrm{var}(Y)$ is maximal.

This suggests the following definition:

Definition 16.2. Given an $n \times d$ matrix X of data points $X_1, \ldots, X_n$, if μ is the centroid of the X_i's, then a *first principal component of X (first PC)* is a centered point $Y_1 = (X-\mu)v_1$, the projection of $X_1, \ldots, X_n$ onto a direction v_1 such that $\mathrm{var}(Y_1)$ is maximized, where v_1 is a unit vector (recall that $Y_1 = (X-\mu)v_1$ is a linear combination of the C_j's, the columns of $X - \mu$).

More generally, if $Y_1, \ldots, Y_k$ are k principal components of X along some unit vectors $v_1, \ldots, v_k$, where $1 \leq k < d$, a *$(k+1)$th principal component of X ($(k+1)$th PC)* is a centered point $Y_{k+1} = (X-\mu)v_{k+1}$, the projection of $X_1, \ldots, X_n$ onto some direction v_{k+1} such that $\mathrm{var}(Y_{k+1})$ is maximized, subject to $\mathrm{cov}(Y_h, Y_{k+1}) = 0$ for all h with $1 \leq h \leq k$, and where v_{k+1} is a unit vector (recall that $Y_h = (X-\mu)v_h$ is a linear combination of the C_j's). The v_h are called *principal directions*.

The following proposition is the key to the main result about PCA:

Proposition 16.10. *If A is a symmetric $d \times d$ matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ and if $(u_1, \ldots, u_d)$ is any orthonormal basis of eigenvectors of A, where u_i is a unit eigenvector associated with λ_i, then*

$$\max_{x \neq 0} \frac{x^\top A x}{x^\top x} = \lambda_1$$

(with the maximum attained for $x = u_1$) and

$$\max_{x \neq 0, x \in \{u_1, \ldots, u_k\}^\perp} \frac{x^\top A x}{x^\top x} = \lambda_{k+1}$$

(with the maximum attained for $x = u_{k+1}$), where $1 \leq k \leq d-1$.

Proof. First, observe that

$$\max_{x\neq 0}\frac{x^\top Ax}{x^\top x} = \max_x\{x^\top Ax \mid x^\top x = 1\},$$

and similarly,

$$\max_{x\neq 0, x\in\{u_1,\ldots,u_k\}^\perp}\frac{x^\top Ax}{x^\top x} = \max_x\left\{x^\top Ax \mid (x\in\{u_1,\ldots,u_k\}^\perp)\wedge(x^\top x = 1)\right\}.$$

Since A is a symmetric matrix, its eigenvalues are real and it can be diagonalized with respect to an orthonormal basis of eigenvectors, so let $(u_1,\ldots,u_d)$ be such a basis. If we write

$$x = \sum_{i=1}^d x_i u_i,$$

a simple computation shows that

$$x^\top Ax = \sum_{i=1}^d \lambda_i x_i^2.$$

If $x^\top x = 1$, then $\sum_{i=1}^d x_i^2 = 1$, and since we assumed that $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_d$, we get

$$x^\top Ax = \sum_{i=1}^d \lambda_i x_i^2 \leq \lambda_1\left(\sum_{i=1}^d x_i^2\right) = \lambda_1.$$

Thus,

$$\max_x\left\{x^\top Ax \mid x^\top x = 1\right\}\leq\lambda_1,$$

and since this maximum is achieved for $e_1 = (1,0,\ldots,0)$, we conclude that

$$\max_x\left\{x^\top Ax \mid x^\top x = 1\right\} = \lambda_1.$$

Next, observe that $x\in\{u_1,\ldots,u_k\}^\perp$ and $x^\top x = 1$ iff $x_1 = \cdots = x_k = 0$ and $\sum_{i=1}^d x_i = 1$. Consequently, for such an x, we have

$$x^\top Ax = \sum_{i=k+1}^d \lambda_i x_i^2 \leq \lambda_{k+1}\left(\sum_{i=k+1}^d x_i^2\right) = \lambda_{k+1}.$$

Thus,

$$\max_x\left\{x^\top Ax \mid (x\in\{u_1,\ldots,u_k\}^\perp)\wedge(x^\top x = 1)\right\}\leq\lambda_{k+1},$$

and since this maximum is achieved for $e_{k+1} = (0,\ldots,0,1,0,\ldots,0)$ with a 1 in position $k+1$, we conclude that

$$\max_x\left\{x^\top Ax \mid (x\in\{u_1,\ldots,u_k\}^\perp)\wedge(x^\top x = 1)\right\} = \lambda_{k+1},$$

as claimed. □

The quantity

$$\frac{x^\top Ax}{x^\top x}$$

is known as the *Rayleigh–Ritz ratio* and Proposition 16.10 is often known as part of the *Rayleigh–Ritz theorem*.

Proposition 16.10 also holds if A is a Hermitian matrix and if we replace $x^\top Ax$ by x^*Ax and $x^\top x$ by x^*x. The proof is unchanged, since a Hermitian matrix has real eigenvalues and is diagonalized with respect to an orthonormal basis of eigenvectors (with respect to the Hermitian inner product).

We then have the following fundamental result showing how *the SVD of X yields the PCs*:

Theorem 16.11. *(SVD yields PCA) Let X be an $n \times d$ matrix of data points $X_1, \ldots, X_n$, and let μ be the centroid of the X_i's. If $X - \mu = VDU^\top$ is an SVD decomposition of $X - \mu$ and if the main diagonal of D consists of the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$, then the centered points $Y_1, \ldots, Y_d$, where*

$$Y_k = (X - \mu)u_k = k\text{th column of } VD$$

and u_k is the kth column of U, are d principal components of X. Furthermore,

$$\mathrm{var}(Y_k) = \frac{\sigma_k^2}{n-1}$$

and $\mathrm{cov}(Y_h, Y_k) = 0$, whenever $h \neq k$ and $1 \leq k, h \leq d$.

Proof. Recall that for any unit vector v, the centered projection of the points $X_1, \ldots, X_n$ onto the line of direction v is $Y = (X - \mu)v$ and that the variance of Y is given by

$$\mathrm{var}(Y) = v^\top \frac{1}{n-1}(X - \mu)^\top (X - \mu)v.$$

Since $X - \mu = VDU^\top$, we get

$$\begin{aligned}
\mathrm{var}(Y) &= v^\top \frac{1}{(n-1)}(X - \mu)^\top (X - \mu)v \\
&= v^\top \frac{1}{(n-1)} UDV^\top VDU^\top v \\
&= v^\top U \frac{1}{(n-1)} D^2 U^\top v.
\end{aligned}$$

Similarly, if $Y = (X - \mu)v$ and $Z = (X - \mu)w$, then the covariance of Y and Z is given by

$$\mathrm{cov}(Y, Z) = v^\top U \frac{1}{(n-1)} D^2 U^\top w.$$

Obviously, $U\frac{1}{(n-1)}D^2U^\top$ is a symmetric matrix whose eigenvalues are $\frac{\sigma_1^2}{n-1} \geq \cdots \geq \frac{\sigma_d^2}{n-1}$, and the columns of U form an orthonormal basis of unit eigenvectors.

We proceed by induction on k. For the base case, $k = 1$, maximizing $\mathrm{var}(Y)$ is equivalent to maximizing

$$v^\top U\frac{1}{(n-1)}D^2U^\top v,$$

where v is a unit vector. By Proposition 16.10, the maximum of the above quantity is the largest eigenvalue of $U\frac{1}{(n-1)}D^2U^\top$, namely $\frac{\sigma_1^2}{n-1}$, and it is achieved for u_1, the first columnn of U. Now we get

$$Y_1 = (X - \mu)u_1 = VDU^\top u_1,$$

and since the columns of U form an orthonormal basis, $U^\top u_1 = e_1 = (1, 0, \ldots, 0)$, and so Y_1 is indeed the first column of VD.

By the induction hypothesis, the centered points $Y_1, \ldots, Y_k$, where $Y_h = (X - \mu)u_h$ and $u_1, \ldots, u_k$ are the first k columns of U, are k principal components of X. Because

$$\mathrm{cov}(Y, Z) = v^\top U\frac{1}{(n-1)}D^2U^\top w,$$

where $Y = (X - \mu)v$ and $Z = (X - \mu)w$, the condition $\mathrm{cov}(Y_h, Z) = 0$ for $h = 1, \ldots, k$ is equivalent to the fact that w belongs to the orthogonal complement of the subspace spanned by $\{u_1, \ldots, u_k\}$, and maximizing $\mathrm{var}(Z)$ subject to $\mathrm{cov}(Y_h, Z) = 0$ for $h = 1, \ldots, k$ is equivalent to maximizing

$$w^\top U\frac{1}{(n-1)}D^2U^\top w,$$

where w is a unit vector orthogonal to the subspace spanned by $\{u_1, \ldots, u_k\}$. By Proposition 16.10, the maximum of the above quantity is the $(k+1)$th eigenvalue of $U\frac{1}{(n-1)}D^2U^\top$, namely $\frac{\sigma_{k+1}^2}{n-1}$, and it is achieved for u_{k+1}, the $(k+1)$th columnn of U. Now we get

$$Y_{k+1} = (X - \mu)u_{k+1} = VDU^\top u_{k+1},$$

and since the columns of U form an orthonormal basis, $U^\top u_{k+1} = e_{k+1}$, and Y_{k+1} is indeed the $(k+1)$th column of VD, which completes the proof of the induction step. □

The d columns $u_1, \ldots, u_d$ of U are usually called the *principal directions* of $X - \mu$ (and X). We note that not only do we have $\mathrm{cov}(Y_h, Y_k) = 0$ whenever $h \neq k$, but the directions $u_1, \ldots, u_d$ along which the data are projected are mutually orthogonal. Also, if r is the rank of the matrix X, then the columns of index $k \geq r + 1$ in D are zero, so the columns of index $k \geq r + 1$ in VD are also zero, and we have $Y_k = 0$ for $k \geq r + 1$. Thus the principal components Y_k only yield useful information if $k \geq r = \mathrm{rank}(X)$.

We know from our study of SVD that $\sigma_1^2, \ldots, \sigma_d^2$ are the eigenvalues of the symmetric positive semidefinite matrix $(X - \mu)^\top(X - \mu)$ and that $u_1, \ldots, u_d$ are corresponding eigenvectors. Numerically, it is preferable to use SVD on $X - \mu$ rather than to compute explicitly

$(X - \mu)^\top (X - \mu)$ and then diagonalize it. Indeed, the explicit computation of $A^\top A$ from a matrix A can be numerically quite unstable, and good SVD algorithms avoid computing $A^\top A$ explicitly.

In general, since an SVD of X is not unique, *the principal directions $u_1, \ldots, u_d$ are not unique*. This can happen when a data set has some *rotational symmetries*, and in such a case, PCA is not a very good method for analyzing the data set.

16.5 Best Affine Approximation

A problem very close to PCA (and based on least squares) is to *best approximate a data set of n points $X_1, \ldots, X_n$, with $X_i \in \mathbb{R}^d$, by a p-dimensional affine subspace A of $\mathbb{R}^d$, with* $1 \leq p \leq d-1$ (the terminology rank $d - p$ is also used).

First, consider $p = d - 1$. Then $A = A_1$ is an affine hyperplane (in $\mathbb{R}^d$), and it is given by an equation of the form

$$a_1 x_1 + \cdots + a_d x_d + c = 0.$$

By *best approximation*, we mean that $(a_1, \ldots, a_d, c)$ solves the homogeneous linear system

$$\begin{pmatrix} x_{1\,1} & \cdots & x_{1\,d} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ x_{n\,1} & \cdots & x_{n\,d} & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_d \\ c \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}$$

in the *least squares sense, subject to the condition that $a = (a_1, \ldots, a_d)$ is a unit vector*, that is, $a^\top a = 1$, where $X_i = (x_{i\,1}, \cdots, x_{i\,d})$.

If we form the symmetric matrix

$$\begin{pmatrix} x_{1\,1} & \cdots & x_{1\,d} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ x_{n\,1} & \cdots & x_{n\,d} & 1 \end{pmatrix}^\top \begin{pmatrix} x_{1\,1} & \cdots & x_{1\,d} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ x_{n\,1} & \cdots & x_{n\,d} & 1 \end{pmatrix}$$

involved in the normal equations, we see that the bottom row (and last column) of that matrix is

$$n\mu_1 \quad \cdots \quad n\mu_d \quad n,$$

where $n\mu_j = \sum_{i=1}^n x_{i\,j}$ is n times the mean of the column C_j of X.

Therefore, if $(a_1, \ldots, a_d, c)$ is a least squares solution, that is, a solution of the normal equations, we must have

$$n\mu_1 a_1 + \cdots + n\mu_d a_d + nc = 0,$$

that is,

$$a_1 \mu_1 + \cdots + a_d \mu_d + c = 0,$$

which means that the *hyperplane A_1 must pass through the centroid μ of the data points* $X_1, \ldots, X_n$. Then we can rewrite the original system with respect to the centered data $X_i - \mu$, and we find that the variable c drops out and we get the system

$$(X - \mu)a = 0,$$

where $a = (a_1, \ldots, a_d)$.

Thus, we are looking for a unit vector a solving $(X - \mu)a = 0$ in the least squares sense, that is, some a such that $a^\top a = 1$ minimizing

$$a^\top (X - \mu)^\top (X - \mu)a.$$

Compute some SVD $VDU^\top$ of $X - \mu$, where the main diagonal of D consists of the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$ of $X - \mu$ arranged in descending order. Then

$$a^\top (X - \mu)^\top (X - \mu)a = a^\top U D^2 U^\top a,$$

where $D^2 = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ is a diagonal matrix, so pick a to be *the last column in U* (corresponding to the smallest eigenvalue σ_d^2 of $(X - \mu)^\top (X - \mu)$). This is a solution to our best fit problem.

Therefore, if U_{d-1} is the linear hyperplane defined by a, that is,

$$U_{d-1} = \{u \in \mathbb{R}^d \mid \langle u, a\rangle = 0\},$$

where a is the last column in U for some SVD $VDU^\top$ of $X - \mu$, we have shown that the affine hyperplane $A_1 = \mu + U_{d-1}$ is a best approximation of the data set $X_1, \ldots, X_n$ in the least squares sense.

Is is easy to show that this hyperplane $A_1 = \mu + U_{d-1}$ minimizes the sum of the square distances of each X_i to its orthogonal projection onto A_1. Also, since U_{d-1} is the orthogonal complement of a, the last column of U, we see that U_{d-1} is spanned by the first $d-1$ columns of U, that is, the first $d-1$ principal directions of $X - \mu$.

All this can be generalized to a *best $(d-k)$-dimensional affine subspace A_k approximating $X_1, \ldots, X_n$ in the least squares sense* $(1 \leq k \leq d-1)$. Such an affine subspace A_k is cut out by k independent hyperplanes H_i (with $1 \leq i \leq k$), each given by some equation

$$a_{i\,1} x_1 + \cdots + a_{i\,d} x_d + c_i = 0.$$

If we write $a_i = (a_{i\,1}, \cdots, a_{i\,d})$, to say that the H_i are independent means that $a_1, \ldots, a_k$ are linearly independent. In fact, we may assume that $a_1, \ldots, a_k$ form an *orthonormal system*.

Then, finding a best $(d-k)$-dimensional affine subspace A_k amounts to solving the homogeneous linear system

$$\begin{pmatrix} X & \mathbf{1} & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & X & \mathbf{1} \end{pmatrix} \begin{pmatrix} a_1 \\ c_1 \\ \vdots \\ a_k \\ c_k \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix},$$

in the least squares sense, subject to the conditions $a_i^\top a_j = \delta_{i\,j}$, for all i, j with $1 \leq i, j \leq k$, where the matrix of the system is a block diagonal matrix consisting of k diagonal blocks $(X, \mathbf{1})$, where $\mathbf{1}$ denotes the column vector $(1, \ldots, 1) \in \mathbb{R}^n$.

Again, it is easy to see that each hyperplane H_i *must pass through the centroid* μ of $X_1, \ldots, X_n$, and by switching to the centered data $X_i - \mu$ we get the system

$$\begin{pmatrix} X-\mu & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X-\mu \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_k \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix},$$

with $a_i^\top a_j = \delta_{i\,j}$ for all i, j with $1 \leq i, j \leq k$.

If $VDU^\top = X - \mu$ is an SVD decomposition, it is easy to see that a least squares solution of this system is given by the last k columns of U, assuming that the main diagonal of D consists of the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$ of $X - \mu$ arranged in descending order. But now the $(d-k)$-dimensional subspace U_{d-k} cut out by the hyperplanes defined by $a_1, \ldots, a_k$ is simply the orthogonal complement of $(a_1, \ldots, a_k)$, which is the subspace spanned by the first $d - k$ columns of U.

So the best $(d-k)$-dimensional affine subpsace A_k approximating $X_1, \ldots, X_n$ in the least squares sense is

$$A_k = \mu + U_{d-k},$$

where U_{d-k} is the linear subspace spanned by the first $d-k$ principal directions of $X-\mu$, that is, the first $d-k$ columns of U. Consequently, we get the following interesting interpretation of PCA (actually, principal directions):

Theorem 16.12. *Let X be an $n \times d$ matrix of data points $X_1, \ldots, X_n$, and let μ be the centroid of the X_i's. If $X - \mu = VDU^\top$ is an SVD decomposition of $X - \mu$ and if the main diagonal of D consists of the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$, then a best $(d-k)$-dimensional affine approximation A_k of $X_1, \ldots, X_n$ in the least squares sense is given by*

$$A_k = \mu + U_{d-k},$$

where U_{d-k} is the linear subspace spanned by the first $d-k$ columns of U, the first $d-k$ principal directions of $X - \mu$ $(1 \leq k \leq d-1)$.

There are many applications of PCA to data compression, dimension reduction, and pattern analysis. The basic idea is that in many cases, given a data set $X_1, \ldots, X_n$, with $X_i \in \mathbb{R}^d$, only a "small" subset of $m < d$ of the features is needed to describe the data set accurately.

If $u_1, \ldots, u_d$ are the principal directions of $X - \mu$, then the first m projections of the data (the first m principal components, i.e., the first m columns of VD) onto the first m principal directions represent the data without much loss of information. Thus, instead of using the

original data points $X_1, \ldots, X_n$, with $X_i \in \mathbb{R}^d$, we can use their projections onto the first m principal directions $Y_1, \ldots, Y_m$, where $Y_i \in \mathbb{R}^m$ and $m < d$, obtaining a compressed version of the original data set.

For example, PCA is used in computer vision for *face recognition*. Sirovitch and Kirby (1987) seem to be the first to have had the idea of using PCA to compress facial images. They introduced the term *eigenpicture* to refer to the principal directions, u_i. However, an explicit face recognition algorithm was given only later, by Turk and Pentland (1991). They renamed eigenpictures as *eigenfaces*.

For details on the topic of eigenfaces, see Forsyth and Ponce [42] (Chapter 22, Section 22.3.2), where you will also find exact references to Turk and Pentland's papers.

Another interesting application of PCA is to the *recognition of handwritten digits*. Such an application is described in Hastie, Tibshirani, and Friedman, [53] (Chapter 14, Section 14.5.1).

16.6 Summary

The main concepts and results of this chapter are listed below:

- *Least squares problems*.
- Existence of a least squares solution of smallest norm (Theorem 16.1).
- The *pseudo-inverse* A^+ of a matrix A.
- The least squares solution of smallest norm is given by the pseudo-inverse (Theorem 16.2)
- Projection properties of the pseudo-inverse.
- The pseudo-inverse of a normal matrix.
- The *Penrose characterization* of the pseudo-inverse.
- Data compression and SVD.
- Best approximation of rank $< r$ of a matrix.
- *Principal component analysis*.
- Review of basic statistical concepts: *mean, variance, covariance*, *covariance matrix*.
- Centered data, *centroid*.
- The *principal components* (*PCA*).

- The *Rayleigh–Ritz theorem* (Theorem 16.10).
- The main theorem: *SVD yields PCA* (Theorem 16.11).
- Best affine approximation.
- SVD yields a best affine approximation (Theorem 16.12).
- Face recognition, eigenfaces.

Chapter 17

Annihilating Polynomials and the Primary Decomposition

17.1 Annihilating Polynomials and the Minimal Polynomial

In Section 4.7, we explained that if $f\colon E \to E$ is a linear map on a K-vector space E, then for any polynomial $p(X) = a_0X^d + a_1X^{d-1} + \cdots + a_d$ with coefficients in the field K, we can define the *linear map* $p(f)\colon E \to E$ by

$$p(f) = a_0f^d + a_1f^{d-1} + \cdots + a_d\mathrm{id},$$

where $f^k = f \circ \cdots \circ f$, the k-fold composition of f with itself. Note that

$$p(f)(u) = a_0f^d(u) + a_1f^{d-1}(u) + \cdots + a_du,$$

for every vector $u \in E$. Then, we showed that if E is finite-dimensional and if $\chi_f(X) = \det(XI - f)$ is the characteristic polynomial of f, by the Cayley–Hamilton Theorem, we have

$$\chi_f(f) = 0.$$

This fact suggests looking at the set of all polynomials $p(X)$ such that

$$p(f) = 0.$$

Such polynomials are called *annihilating polynomials* of f, the set of all these poynomials, denoted $\mathrm{Ann}(f)$, is called the *annihilator* of f, and the Cayley-Hamilton Theorem shows that it is nontrivial, since it contains a polynomial of positive degree. It turns out that $\mathrm{Ann}(f)$ contains a polynomial m_f of smallest degree that generates $\mathrm{Ann}(f)$, and this polynomial divides the characteristic polynomial. Furthermore, the polynomial m_f encapsulates a lot of information about f, in particular whether f can be diagonalized.

In order to understand the structure of $\mathrm{Ann}(f)$, we need to review some basic properties of polynomials. The first crucial notion is that of an ideal.

Definition 17.1. Given a commutative ring A with unit 1, *an ideal of* A is a nonempty subset $\mathfrak{I}$ of A satisfying the following properties:

(ID1) If $a, b \in \mathfrak{I}$, then $b - a \in \mathfrak{I}$.

(ID2) If $a \in \mathfrak{I}$, then $ax \in \mathfrak{I}$ for all $x \in A$.

An ideal $\mathfrak{I}$ is a *principal ideal* if there is some $a \in \mathfrak{I}$, called a *generator*, such that

$$\mathfrak{I} = \{ax \mid x \in A\}.$$

In this case, we usually write $\mathfrak{I} = aA$, or $\mathfrak{I} = (a)$. The ideal $\mathfrak{I} = (0) = \{0\}$ is called the *null ideal* (or *zero ideal*).

Given a field K, any nonzero polynomial $p(X) \in K[X]$ has some monomial of highest degree a_0X^n with $a_0 \neq 0$, and the integer $n = \deg(p) \geq 0$ is called the *degree* of p. It is convenient to set the degree of the zero polynomial (denoted by 0) to be

$$\deg(0) = -\infty.$$

A polynomial $p(X)$ such that the coefficient a_0 of its monomial of highest degree is 1 is called a *monic* polynomial.

The following proposition is a fundamental result about polynomials over a field.

Proposition 17.1. *If K is a field, then every polynomial ideal $\mathfrak{I} \subseteq K[X]$ is a principal ideal. As a consequence, if $\mathfrak{I}$ is not the zero ideal, then there is a unique monic polynomial*

$$p(X) = X^n + a_1X^{n-1} + \cdots + a_{n-1}X + a_n$$

in $\mathfrak{I}$ such that $\mathfrak{I} = (p)$.

Proof. This result is not hard to prove if we recall that polynomials can divided: Given any two nonzero polynomials $f, g \in K[X]$, there are unique polynomials q, r such that

$$f = gq + r, \quad \text{and} \quad \deg(r) < \deg(g).$$

If $\mathfrak{I}$ is not the zero ideal, there is some polynomial of smallest degree in $\mathfrak{I}$, and since K is a field, by suitable multiplication by a scalar, we can make sure that this polynomial is monic. Thus, let f be a monic polynomial of smallest degree in $\mathfrak{I}$. By (ID2), it is clear that $(f) \subseteq \mathfrak{I}$. Now, let $g \in \mathfrak{I}$. Using the Euclidean algorithm, there exist unique $q, r \in K[X]$ such that

$$g = qf + r \quad \text{and} \quad \deg(r) < \deg(f).$$

If $r \neq 0$, there is some $\lambda \neq 0$ in K such that λr is a monic polynomial, and since $\lambda r = \lambda g - \lambda qf$, with $f, g \in \mathfrak{I}$, by (ID1) and (ID2), we have $\lambda r \in \mathfrak{I}$, where $\deg(\lambda r) < \deg(f)$ and λr is a monic polynomial, contradicting the minimality of the degree of f. Thus, $r = 0$, and $g \in (f)$. The uniqueness of the monic polynomial f is left as an exercise. □

We will also need to know that the greatest common divisor of polynomials exist. Given any two nonzero polynomials $f, g \in K[X]$, recall that f divides g if $g = fq$ for some $q \in K[X]$.

Definition 17.2. Given any two nonzero polynomials $f, g \in K[X]$, a polynomial $d \in K[X]$ is a *greatest common divisor of f and g* (for short, a *gcd of f and g*) if d divides f and g and whenever $h \in K[X]$ divides f and g, then h divides d. We say that f and g are *relatively prime* if 1 is a gcd of f and g.

Note that f and g are relatively prime iff all of their gcd's are constants (scalars in K), or equivalently, if f, g have no common divisor q of degree $\deg(q) \geq 1$.

We can characterize gcd's of polynomials as follows.

Proposition 17.2. *Let K be a field and let $f, g \in K[X]$ be any two nonzero polynomials. For every polynomial $d \in K[X]$, the following properties are equivalent:*

(1) The polynomial d is a gcd of f and g.

(2) The polynomial d divides f and g and there exist $u, v \in K[X]$ such that

$$d = uf + vg.$$

(3) The ideals $(f), (g)$, and (d) satisfy the equation

$$(d) = (f) + (g).$$

In addition, $d \neq 0$, and d is unique up to multiplication by a nonzero scalar in K.

As a consequence of Proposition 17.2, two nonzero polynomials $f, g \in K[X]$ are relatively prime iff there exist $u, v \in K[X]$ such that

$$uf + vg = 1.$$

The identity

$$d = uf + vg$$

of part (2) of Lemma 17.2 is often called the *Bezout identity*. An important consequence of the Bezout identity is the following result.

Proposition 17.3. *(Euclid's proposition) Let K be a field and let $f, g, h \in K[X]$ be any nonzero polynomials. If f divides gh and f is relatively prime to g, then f divides h.*

Proposition 17.3 can be generalized to any number of polynomials.

Proposition 17.4. *Let K be a field and let $f, g_1, \ldots, g_m \in K[X]$ be some nonzero polynomials. If f and g_i are relatively prime for all i, $1 \leq i \leq m$, then f and $g_1 \cdots g_m$ are relatively prime.*

Definition 17.2 is generalized to any finite number of polynomials as follows.

Definition 17.3. Given any nonzero polynomials $f_1, \ldots, f_n \in K[X]$, where $n \geq 2$, a polynomial $d \in K[X]$ is a *greatest common divisor of* $f_1, \ldots, f_n$ (for short, a *gcd of* $f_1, \ldots, f_n$) if d divides each f_i and whenever $h \in K[X]$ divides each f_i, then h divides d. We say that $f_1, \ldots, f_n$ are *relatively prime* if 1 is a gcd of $f_1, \ldots, f_n$.

It is easily shown that Proposition 17.2 can be generalized to any finite number of polynomials.

Proposition 17.5. *Let K be a field and let $f_1, \ldots, f_n \in K[X]$ be any $n \geq 2$ nonzero polynomials. For every polynomial $d \in K[X]$, the following properties are equivalent:*

(1) The polynomial d is a gcd of $f_1, \ldots, f_n$.

(2) The polynomial d divides each f_i and there exist $u_1, \ldots, u_n \in K[X]$ such that

$$d = u_1 f_1 + \cdots + u_n f_n.$$

(3) The ideals (f_i), and (d) satisfy the equation

$$(d) = (f_1) + \cdots + (f_n).$$

In addition, $d \neq 0$, and d is unique up to multiplication by a nonzero scalar in K.

As a consequence of Proposition 17.5, any $n \geq 2$ nonzero polynomials $f_1, \ldots, f_n \in K[X]$ are relatively prime iff there exist $u_1, \ldots, u_n \in K[X]$ such that

$$u_1 f_1 + \cdots + u_n f_n = 1,$$

the *Bezout identity*.

We will also need to know that every nonzero polynomial (over a field) can be factored into irreducible polynomials, which are the generalization of the prime numbers to polynomials.

Definition 17.4. Given a field K, a polynomial $p \in K[X]$ is *irreducible or indecomposable or prime* if $\deg(p) \geq 1$ and if p is not divisible by any polynomial $q \in K[X]$ such that $1 \leq \deg(q) < \deg(p)$. Equivalently, p is irreducible if $\deg(p) \geq 1$ and if $p = q_1 q_2$, then either $q_1 \in K$ or $q_2 \in K$ (and of course, $q_1 \neq 0$, $q_2 \neq 0$).

Every polynomial $aX + b$ of degree 1 is irreducible. Over the field $\mathbb{R}$, the polynomial $X^2 + 1$ is irreducible (why?), but $X^3 + 1$ is not irreducible, since

$$X^3 + 1 = (X + 1)(X^2 - X + 1).$$

The polynomial $X^2 - X + 1$ is irreducible over $\mathbb{R}$ (why?). It would seem that $X^4 + 1$ is irreducible over $\mathbb{R}$, but in fact,

$$X^4 + 1 = (X^2 - \sqrt{2}X + 1)(X^2 + \sqrt{2}X + 1).$$

However, in view of the above factorization, $X^4 + 1$ is irreducible over $\mathbb{Q}$.

It can be shown that the irreducible polynomials over $\mathbb{R}$ are the polynomials of degree 1, or the polynomials of degree 2 of the form $aX^2 + bX + c$, for which $b^2 - 4ac < 0$ (i.e., those having no real roots). This is not easy to prove! Over the complex numbers $\mathbb{C}$, the only irreducible polynomials are those of degree 1. This is a version of a fact often referred to as the "Fundamental theorem of Algebra."

Observe that the definition of irreducibilty implies that any finite number of distinct irreducible polynomials are relatively prime.

The following fundamental result can be shown

Theorem 17.6. *Given any field K, for every nonzero polynomial*

$$f = a_d X^d + a_{d-1} X^{d-1} + \cdots + a_0$$

of degree $d = \deg(f) \geq 1$ in $K[X]$, there exists a unique set $\{\langle p_1, k_1\rangle, \ldots, \langle p_m, k_m\rangle\}$ such that

$$f = a_d p_1^{k_1} \cdots p_m^{k_m},$$

where the $p_i \in K[X]$ are distinct irreducible monic polynomials, the k_i are (not necessarily distinct) integers, and with $m \geq 1$, $k_i \geq 1$.

We can now return to minimal polynomials. Given a linear map $f\colon E \to E$, it is easy to check that the set $\mathrm{Ann}(f)$ of polynomials that annihilate f is an ideal. Furthermore, when E is finite-dimensional, the Cayley-Hamilton Theorem implies that $\mathrm{Ann}(f)$ is not the zero ideal. Therefore, by Proposition 17.1, there is a unique monic polynomial m_f that generates $\mathrm{Ann}(f)$.

Definition 17.5. If $f\colon E \to E$, is linear map on a finite-dimensional vector space E, the unique monic polynomial $m_f(X)$ that generates the ideal $\mathrm{Ann}(f)$ of polynomials which annihilate f (the *annihilator* of f) is called the *minimal polynomial* of f.

The minimal polynomial m_f of f is the monic polynomial of smallest degree that annihilates f. Thus, the minimal polynomial divides the characteristic polynomial χ_f, and $\deg(m_f) \geq 1$. For simplicity of notation, we often write m instead of m_f.

If A is any $n \times n$ matrix, the set $\mathrm{Ann}(A)$ of polynomials that annihilate A is the set of polynomials

$$p(X) = a_0 X^d + a_1 X^{d-1} + \cdots + a_{d-1} X + a_d$$

such that

$$a_0 A^d + a_1 A^{d-1} + \cdots + a_{d-1} A + a_d I = 0.$$

It is clear that $\mathrm{Ann}(A)$ is a nonzero ideal and its unique monic generator is called the *minimal polynomial* of A. We check immediately that if Q is an invertible matrix, then A and $Q^{-1}AQ$ have the same minimal polynomial. Also, if A is the matrix of f with respect to some basis, then f and A have the same minimal polynomial.

The zeros (in K) of the minimal polynomial of f and the eigenvalues of f (in K) are intimately related.

Proposition 17.7. *Let $f\colon E \to E$ be a linear map on some finite-dimensional vector space E. Then, $\lambda \in K$ is a zero of the minimal polynomial $m_f(X)$ of f iff λ is an eigenvalue of f iff λ is a zero of $\chi_f(X)$. Therefore, the minimal and the characteristic polynomials have the same zeros (in K), except for multiplicities.*

Proof. First, assume that $m(\lambda) = 0$ (with $\lambda \in K$, and writing m instead of m_f). If so, using polynomial division, m can be factored as

$$m = (X - \lambda)q,$$

with $\deg(q) < \deg(m)$. Since m is the minimal polynomial, $q(f) \neq 0$, so there is some nonzero vector $v \in E$ such that $u = q(f)(v) \neq 0$. But then, because m is the minimal polynomial,

$$\begin{aligned} 0 &= m(f)(v) \\ &= (f - \lambda\mathrm{id})(q(f)(v)) \\ &= (f - \lambda\mathrm{id})(u), \end{aligned}$$

which shows that λ is an eigenvalue of f.

Conversely, assume that $\lambda \in K$ is an eigenvalue of f. This means that for some $u \neq 0$, we have $f(u) = \lambda u$. Now, it is easy to show that

$$m(f)(u) = m(\lambda)u,$$

and since m is the minimal polynomial of f, we have $m(f)(u) = 0$, so $m(\lambda)u = 0$, and since $u \neq 0$, we must have $m(\lambda) = 0$. □

If we assume that f is diagonalizable, then its eigenvalues are all in K, and if $\lambda_1, \ldots, \lambda_k$ are the distinct eigenvalues of f, then by Proposition 17.7, the minimal polynomial m of f must be a product of powers of the polynomials $(X - \lambda_i)$. Actually, we claim that

$$m = (X - \lambda_1) \cdots (X - \lambda_k).$$

For this, we just have to show that m annihilates f. However, for any eigenvector u of f, one of the linear maps $f - \lambda_i\mathrm{id}$ sends u to 0, so

$$m(f)(u) = (f - \lambda_1\mathrm{id}) \circ \cdots \circ (f - \lambda_k\mathrm{id})(u) = 0.$$

Since E is spanned by the eigenvectors of f, we conclude that

$$m(f) = 0.$$

Therefore, if a linear map is diagonalizable, then its minimal polynomial is a product of distinct factors of degree 1. It turns out that the converse is true, but this will take a little work to establish it.

17.2 Minimal Polynomials of Diagonalizable Linear Maps

In this section, we prove that if the minimal polynomial m_f of a linear map f is of the form

$$m_f = (X - \lambda_1) \cdots (X - \lambda_k)$$

for disctinct scalars $\lambda_1, \ldots, \lambda_k \in K$, then f is diagonalizable. This is a powerful result that has a number of implications. We need of few properties of invariant subspaces.

Given a linear map $f\colon E \to E$, recall that a subspace W of E is *invariant under* f if $f(u) \in W$ for all $u \in W$.

Proposition 17.8. *Let W be a subspace of E invariant under the linear map $f\colon E \to E$ (where E is finite-dimensional). Then, the minimal polynomial of the restriction $f \mid W$ of f to W divides the minimal polynomial of f, and the characteristic polynomial of $f \mid W$ divides the characteristic polynomial of f.*

Sketch of proof. The key ingredient is that we can pick a basis $(e_1, \ldots, e_n)$ of E in which $(e_1, \ldots, e_k)$ is a basis of W. Then, the matrix of f over this basis is a block matrix of the form

$$A = \begin{pmatrix} B & C \\ 0 & D \end{pmatrix},$$

where B is a $k \times k$ matrix, D is a $(n-k) \times (n-k)$ matrix, and C is a $k \times (n-k)$ matrix. Then

$$\det(XI - A) = \det(XI - B)\det(XI - D),$$

which implies the statement about the characteristic polynomials. Furthermore,

$$A^i = \begin{pmatrix} B^i & C_i \\ 0 & D^i \end{pmatrix},$$

for some $k \times (n-k)$ matrix C_i. It follows that any polynomial which annihilates A also annihilates B and D. So, the minimal polynomial of B divides the minimal polynomial of A. □

For the next step, there are at least two ways to proceed. We can use an old-fashion argument using Lagrange interpolants, or use a slight generalization of the notion of annihilator. We pick the second method because it illustrates nicely the power of principal ideals.

What we need is the notion of conductor (also called transporter).

Definition 17.6. Let $f\colon E \to E$ be a linear map on a finite-dimensional vector space E, let W be an invariant subspace of f, and let u be any vector in E. The set $S_f(u, W)$ consisting of all polynomials $q \in K[X]$ such that $q(f)(u) \in W$ is called the *f-conductor of u into W*.

Observe that the minimal polynomial m of f always belongs to $S_f(u, W)$, so this is a nontrivial set. Also, if $W = (0)$, then $S_f(u, (0))$ is just the annihilator of f. The crucial property of $S_f(u, W)$ is that it is an ideal.

Proposition 17.9. *If W is an invariant subspace for f, then for each $u \in E$, the f-conductor $S_f(u, W)$ is an ideal in $K[X]$.*

We leave the proof as a simple exercise, using the fact that if W invariant under f, then W is invariant under every polynomial $q(f)$ in f.

Since $S_f(u, W)$ is an ideal, it is generated by a unique monic polynomial q of smallest degree, and because the minimal polynomial m_f of f is in $S_f(u, W)$, the polynomial q divides m.

Proposition 17.10. *Let $f\colon E \to E$ be a linear map on a finite-dimensional space E, and assume that the minimal polynomial m of f is of the form*

$$m = (X - \lambda_1)^{r_1} \cdots (X - \lambda_k)^{r_k},$$

where the eigenvalues $\lambda_1, \ldots, \lambda_k$ of f belong to K. If W is a proper subspace of E which is invariant under f, then there is a vector $u \in E$ with the following properties:

(a) $u \notin W$;

(b) $(f - \lambda \mathrm{id})(u) \in W$, for some eigenvalue λ of f.

Proof. Observe that (a) and (b) together assert that the f-conductor of u into W is a polynomial of the form $X - \lambda_i$. Pick any vector $v \in E$ not in W, and let g be the conductor of v into W. Since g divides m and $v \notin W$, the polynomial g is not a constant, and thus it is of the form

$$g = (X - \lambda_1)^{s_1} \cdots (X - \lambda_k)^{s_k},$$

with at least some $s_i > 0$. Choose some index j such that $s_j > 0$. Then $X - \lambda_j$ is a factor of g, so we can write

$$g = (X - \lambda_j)q.$$

By definition of g, the vector $u = q(f)(v)$ cannot be in W, since otherwise g would not be of minimal degree. However,

$$\begin{aligned}(f - \lambda_j \mathrm{id})(u) &= (f - \lambda_j \mathrm{id})(q(f)(v)) \\ &= g(f)(v)\end{aligned}$$

is in W, which concludes the proof. □

We can now prove the main result of this section.

Theorem 17.11. *Let $f : E \to E$ be a linear map on a finite-dimensional space E. Then f is diagonalizable iff its minimal polynomial m is of the form*

$$m = (X - \lambda_1) \cdots (X - \lambda_k),$$

where $\lambda_1, \ldots, \lambda_k$ are distinct elements of K.

Proof. We already showed in Section 17.2 that if f is diagonalizable, then its minimal polynomial is of the above form (where $\lambda_1, \ldots, \lambda_k$ are the distinct eigenvalues of f).

For the converse, let W be the subspace spanned by all the eigenvectors of f. If $W \neq E$, since W is invariant under f, by Proposition 17.10, there is some vector $u \notin W$ such that for some λ_j, we have

$$(f - \lambda_j \mathrm{id})(u) \in W.$$

Let $v = (f - \lambda_j \mathrm{id})(u) \in W$. Since $v \in W$, we can write

$$v = w_1 + \cdots + w_k$$

where $f(w_i) = \lambda_i w_i$ (either $w_i = 0$ or w_i is an eigenvector for λ_i), and so, for every polynomial h, we have

$$h(f)(v) = h(\lambda_1)w_1 + \cdots + h(\lambda_k)w_k,$$

which shows that $h(f)(v) \in W$ for every polynomial h. We can write

$$m = (X - \lambda_j)q$$

for some polynomial q, and also

$$q - q(\lambda_j) = p(X - \lambda_j)$$

for some polynomial p. We know that $p(f)(v) \in W$, and since m is the minimal polynomial of f, we have

$$0 = m(f)(u) = (f - \lambda_j \mathrm{id})(q(f)(u)),$$

which implies that $q(f)(u) \in W$ (either $q(f)(u) = 0$, or it is an eigenvector associated with λ_j). However,

$$q(f)(u) - q(\lambda_j)u = p(f)((f - \lambda_j \mathrm{id})(u)) = p(f)(v),$$

and since $p(f)(v) \in W$ and $q(f)(u) \in W$, we conclude that $q(\lambda_j)u \in W$. But, $u \notin W$, which implies that $q(\lambda_j) = 0$, so λ_j is a double root of m, a contradiction. Therefore, we must have $W = E$. □

Remark: Proposition 17.10 can be used to give a quick proof of Theorem 12.4.

Using Theorem 17.11, we can give a short proof about commuting diagonalizable linear maps. If $\mathcal{F}$ is a family of linear maps on a vector space E, we say that $\mathcal{F}$ is a *commuting family* iff $f \circ g = g \circ f$ for all $f, g \in \mathcal{F}$.

Proposition 17.12. *Let $\mathcal{F}$ be a nonempty finite commuting family of diagonalizable linear maps on a finite-dimensional vector space E. There exists a basis of E such that every linear map in $\mathcal{F}$ is represented in that basis by a diagonal matrix.*

Proof. We proceed by induction on $n = \dim(E)$. If $n = 1$, there is nothing to prove. If $n > 1$, there are two cases. If all linear maps in $\mathcal{F}$ are of the form $\lambda \mathrm{id}$ for some $\lambda \in K$, then the proposition holds trivially. In the second case, let $f \in \mathcal{F}$ be some linear map in $\mathcal{F}$ which is not a scalar multiple of the identity. In this case, f has at least two distinct eigenvalues $\lambda_1, \ldots, \lambda_k$, and because f is diagonalizable, E is the direct sum of the corresponding eigenspaces $E_{\lambda_1}, \ldots, E_{\lambda_k}$. For every index i, the eigenspace E_{λ_i} is invariant under f and under every other linear map g in $\mathcal{F}$, since for any $g \in \mathcal{F}$ and any $u \in E_{\lambda_i}$, because f and g commute, we have

$$f(g(u)) = g(f(u)) = g(\lambda_i u) = \lambda_i g(u)$$

so $g(u) \in E_{\lambda_i}$. Let $\mathcal{F}_i$ be the family obtained by restricting each $f \in \mathcal{F}$ to E_{λ_i}. By proposition 17.8, the minimal polynomial of every linear map $f \mid E_{\lambda_i}$ in $\mathcal{F}_i$ divides the minimal polynomial m_f of f, and since f is diagonalizable, m_f is a product of distinct linear factors, so the minimal polynomial of $f \mid E_{\lambda_i}$ is also a product of distinct linear factors. By Theorem 17.11, the linear map $f \mid E_{\lambda_i}$ is diagonalizable. Since $k > 1$, we have $\dim(E_{\lambda_i}) < \dim(E)$ for $i = 1, \ldots, k$, and by the induction hypothesis, for each i there is a basis of E_{λ_i} over which $f \mid E_{\lambda_i}$ is represented by a diagonal matrix. Since the above argument holds for all i, by combining the bases of the E_{λ_i}, we obtain a basis of E such that the matrix of every linear map $f \in \mathcal{F}$ is represented by a diagonal matrix. □

Remark: Proposition 17.12 also holds for infinite commuting familes $\mathcal{F}$ of diagonalizable linear maps, because E being finite dimensional, there is a finite subfamily of linearly independent linear maps in $\mathcal{F}$ spanning $\mathcal{F}$.

There is also an analogous result for commuting families of linear maps represented by upper triangular matrices. To prove this, we need the following proposition.

Proposition 17.13. *Let $\mathcal{F}$ be a nonempty finite commuting family of triangulable linear maps on a finite-dimensional vector space E. Let W be a proper subspace of E which is invariant under $\mathcal{F}$. Then there exists a vector $u \in E$ such that:*

1. *$u \notin W$.*
2. *For every $f \in \mathcal{F}$, the vector $f(u)$ belongs to the subspace $W \oplus Ku$ spanned by W and u.*

Proof. By renaming the elements of $\mathcal{F}$ if necessary, we may assume that $(f_1, \ldots, f_r)$ is a basis of the subspace of $\mathrm{End}(E)$ spanned by $\mathcal{F}$. We prove by induction on r that there exists some vector $u \in E$ such that

1. $u \notin W$.

2. $(f_i - \alpha_i \mathrm{id})(u) \in W$ for $i = 1, \ldots, r$, for some scalars $\alpha_i \in K$.

Consider the base case $r = 1$. Since f_1 is triangulable, its eigenvalues all belong to K since they are the diagonal entries of the triangular matrix associated with f_1 (this is the easy direction of Theorem 12.4), so the minimal polynomial of f_1 is of the form

$$m = (X - \lambda_1)^{r_1} \cdots (X - \lambda_k)^{r_k},$$

where the eigenvalues $\lambda_1, \ldots, \lambda_k$ of f_1 belong to K. We conclude by applying Proposition 17.10.

Next, assume that $r \geq 2$ and that the induction hypothesis holds for $f_1, \ldots, f_{r-1}$. Thus, there is a vector $u_{r-1} \in E$ such that

1. $u_{r-1} \notin W$.

2. $(f_i - \alpha_i \mathrm{id})(u_{r-1}) \in W$ for $i = 1, \ldots, r-1$, for some scalars $\alpha_i \in K$.

Let

$$V_{r-1} = \{w \in E \mid (f_i - \alpha_i \mathrm{id})(w) \in W,\ i = 1, \ldots, r-1\}.$$

Clearly, $W \subseteq V_{r-1}$ and $u_{r-1} \in V_{r-1}$. We claim that V_{r-1} is invariant under $\mathcal{F}$. This is because, for any $v \in V_{r-1}$ and any $f \in \mathcal{F}$, since f and f_i commute, we have

$$(f_i - \alpha_i \mathrm{id})(f(v)) = f(f_i - \alpha_i \mathrm{id})(v)), \quad 1 \leq i \leq r-1.$$

Now, $(f_i - \alpha_i \mathrm{id})(v) \in W$ because $v \in V_{r-1}$, and W is invariant under $\mathcal{F}$ so $f(f_i - \alpha_i \mathrm{id})(v)) \in W$, that is, $(f_i - \alpha_i \mathrm{id})(f(v)) \in W$.

Consider the restriction g_r of f_r to V_{r-1}. The minimal polynomial of g_r divides the minimal polynomial of f_r, and since f_r is triangulable, just as we saw for f_1, the minimal polynomial of f_r is of the form

$$m = (X - \lambda_1)^{r_1} \cdots (X - \lambda_k)^{r_k},$$

where the eigenvalues $\lambda_1, \ldots, \lambda_k$ of f_r belong to K, so the minimal polynomial of g_r is of the same form. By Proposition 17.10, there is some vector $u_r \in V_{r-1}$ such that

1. $u_r \notin W$.

2. $(g_r - \alpha_r \mathrm{id})(u_r) \in W$ for some scalars $\alpha_r \in K$.

Now, since $u_r \in V_{r-1}$, we have $(f_i - \alpha_i \mathrm{id})(u_r) \in W$ for $i = 1, \ldots, r-1$, so $(f_i - \alpha_i \mathrm{id})(u_r) \in W$ for $i = 1, \ldots, r$ (since g_r is the restriction of f_r), which concludes the proof of the induction step. Finally, since every $f \in \mathcal{F}$ is the linear combination of $(f_1, \ldots, f_r)$, condition (2) of the inductive claim implies condition (2) of the proposition. □

We can now prove the following result.

Proposition 17.14. *Let $\mathcal{F}$ be a nonempty finite commuting family of triangulable linear maps on a finite-dimensional vector space E. There exists a basis of E such that every linear map in $\mathcal{F}$ is represented in that basis by an upper triangular matrix.*

Proof. Let $n = \dim(E)$. We construct inductively a basis $(u_1, \ldots, u_n)$ of E such that if W_i is the subspace spanned by $(u_1 \ldots, u_i)$, then for every $f \in \mathcal{F}$,

$$f(u_i) = a^f_{1i} u_1 + \cdots + a^f_{ii} u_i,$$

for some $a^f_{ij} \in K$; that is, $f(u_i)$ belongs to the subspace W_i.

We begin by applying Proposition 17.13 to the subspace $W_0 = (0)$ to get u_1 so that for all $f \in \mathcal{F}$,

$$f(u_1) = \alpha^f_1 u_1.$$

For the induction step, since W_i invariant under $\mathcal{F}$, we apply Proposition 17.13 to the subspace W_i, to get $u_{i+1} \in E$ such that

1. $u_{i+1} \notin W_i$.

2. For every $f \in \mathcal{F}$, the vector $f(u_{i+1})$ belong to the subspace spanned by W_i and u_{i+1}.

Condition (1) implies that $(u_1, \ldots, u_i, u_{i+1})$ is linearly independent, and condition (2) means that for every $f \in \mathcal{F}$,

$$f(u_{i+1}) = a^f_{1i+1} u_1 + \cdots + a^f_{i+1i+1} u_{i+1},$$

for some $a^f_{i+1j} \in K$, establishing the induction step. After n steps, each $f \in \mathcal{F}$ is represented by an upper triangular matrix. □

Observe that if $\mathcal{F}$ consists of a single linear map f and if the minimal polynomial of f is of the form

$$m = (X - \lambda_1)^{r_1} \cdots (X - \lambda_k)^{r_k},$$

with all $\lambda_i \in K$, using Proposition 17.10 instead of Proposition 17.13, the proof of Proposition 17.14 yields another proof of Theorem 12.4.

17.3 The Primary Decomposition Theorem

If $f\colon E \to E$ is a linear map and $\lambda \in K$ is an eigenvalue of f, recall that the eigenspace E_λ associated with λ is the kernel of the linear map $\lambda \mathrm{id} - f$. If all the eigenvalues $\lambda_1 \ldots, \lambda_k$ of f are in K, it may happen that

$$E = E_{\lambda_1} \oplus \cdots \oplus E_{\lambda_k},$$

but in general there are not enough eigenvectors to span E. What if we generalize the notion of eigenvector and look for (nonzero) vectors u such that

$$(\lambda \mathrm{id} - f)^r(u) = 0, \quad \text{for some } r \geq 1?$$

Then, it turns out that if the minimal polynomial of f is of the form

$$m = (X - \lambda_1)^{r_1} \cdots (X - \lambda_k)^{r_k},$$

then $r = r_i$ does the job for λ_i; that is, if we let

$$W_i = \mathrm{Ker}\,(\lambda_i \mathrm{id} - f)^{r_i},$$

then

$$E = W_1 \oplus \cdots \oplus W_k.$$

This result is very nice but seems to require that the eigenvalues of f all belong to K. Actually, it is a special case of a more general result involving the factorization of the minimal polynomial m into its irreducible monic factors (See Theorem 17.6),

$$m = p_1^{r_1} \cdots p_k^{r_k},$$

where the p_i are distinct irreducible monic polynomials over K.

Theorem 17.15. *(Primary Decomposition Theorem) Let $f\colon E \to E$ be a linear map on the finite-dimensional vector space E over the field K. Write the minimal polynomial m of f as*

$$m = p_1^{r_1} \cdots p_k^{r_k},$$

where the p_i are distinct irreducible monic polynomials over K, and the r_i are positive integers. Let

$$W_i = \mathrm{Ker}\,(p_i^{r_i}(f)), \quad i = 1, \ldots, k.$$

Then

(a) $E = W_1 \oplus \cdots \oplus W_k$.

(b) Each W_i is invariant under f.

(c) The minimal polynomial of the restriction $f \mid W_i$ of f to W_i is $p_i^{r_i}$.

Proof. The trick is to construct projections π_i using the polynomials $p_j^{r_j}$ so that the range of π_i is equal to W_i. Let

$$g_i = m/p_i^{r_i} = \prod_{j \neq i} p_j^{r_j}.$$

Note that

$$p_i^{r_i} g_i = m.$$

Since $p_1, \ldots, p_k$ are irreducible and distinct, they are relatively prime. Then, using Proposition 17.4, it is easy to show that $g_1, \ldots, g_k$ are relatively prime. Otherwise, some irreducible polynomial p would divide all of $g_1, \ldots, g_k$, so by Proposition 17.4 it would be equal to one of the irreducible factors p_i. But, that p_i is missing from g_i, a contradiction. Therefore, by Proposition 17.5, there exist some polynomials $h_1, \ldots, h_k$ such that

$$g_1 h_1 + \cdots + g_k h_k = 1.$$

Let $q_i = g_i h_i$ and let $\pi_i = q_i(f) = g_i(f) h_i(f)$. We have

$$q_1 + \cdots + q_k = 1,$$

and since m divides $q_i q_j$ for $i \neq j$, we get

$$\begin{aligned} \pi_1 + \cdots + \pi_k &= \mathrm{id} \\ \pi_i \pi_j &= 0, \quad i \neq j. \end{aligned}$$

(We implicitly used the fact that if p, q are two polynomials, the linear maps $p(f) \circ q(f)$ and $q(f) \circ p(f)$ are the same since $p(f)$ and $q(f)$ are polynomials in the powers of f, which commute.) Composing the first equation with π_i and using the second equation, we get

$$\pi_i^2 = \pi_i.$$

Therefore, the π_i are projections, and E is the direct sum of the images of the π_i. Indeed, every $u \in E$ can be expressed as

$$u = \pi_1(u) + \cdots + \pi_k(u).$$

Also, if

$$\pi_1(u) + \cdots + \pi_k(u) = 0,$$

then by applying π_i we get

$$0 = \pi_i^2(u) = \pi_i(u), \quad i = 1, \ldots k.$$

To finish proving (a), we need to show that

$$W_i = \mathrm{Ker}\,(p_i^{r_i}(f)) = \pi_i(E).$$

If $v \in \pi_i(E)$, then $v = \pi_i(u)$ for some $u \in E$, so

$$\begin{aligned} p_i^{r_i}(f)(v) &= p_i^{r_i}(f)(\pi_i(u)) \\ &= p_i^{r_i}(f)g_i(f)h_i(f)(u) \\ &= h_i(f)p_i^{r_i}(f)g_i(f)(u) \\ &= h_i(f)m(f)(u) = 0, \end{aligned}$$

because m is the minimal polynomial of f. Therefore, $v \in W_i$.

Conversely, assume that $v \in W_i = \operatorname{Ker}(p_i^{r_i}(f))$. If $j \neq i$, then $g_j h_j$ is divisible by $p_i^{r_i}$, so

$$g_j(f)h_j(f)(v) = \pi_j(v) = 0, \quad j \neq i.$$

Then, since $\pi_1 + \cdots + \pi_k = \mathrm{id}$, we have $v = \pi_i v$, which shows that v is in the range of π_i. Therefore, $W_i = \operatorname{Im}(\pi_i)$, and this finishes the proof of (a).

If $p_i^{r_i}(f)(u) = 0$, then $p_i^{r_i}(f)(f(u)) = f(p_i^{r_i}(f)(u)) = 0$, so (b) holds.

If we write $f_i = f \mid W_i$, then $p_i^{r_i}(f_i) = 0$, because $p_i^{r_i}(f) = 0$ on W_i (its kernel). Therefore, the minimal polynomial of f_i divides $p_i^{r_i}$. Conversely, let q be any polynomial such that $q(f_i) = 0$ (on W_i). Since $m = p_i^{r_i} g_i$, the fact that $m(f)(u) = 0$ for all $u \in E$ shows that

$$p_i^{r_i}(f)(g_i(f)(u)) = 0, \quad u \in E,$$

and thus $\operatorname{Im}(g_i(f)) \subseteq \operatorname{Ker}(p_i^{r_i}(f)) = W_i$. Consequently, since $q(f)$ is zero on W_i,

$$q(f)g_i(f) = 0 \quad \text{for all } u \in E.$$

But then, qg_i is divisible by the minimal polynomial $m = p_i^{r_i} g_i$ of f, and since $p_i^{r_i}$ and g_i are relatively prime, by Euclid's Proposition, $p_i^{r_i}$ must divide q. This finishes the proof that the minimal polynomial of f_i is $p_i^{r_i}$, which is (c). □

If all the eigenvalues of f belong to the field K, we obtain the following result.

Theorem 17.16. *(Primary Decomposition Theorem, Version 2) Let $f\colon E \to E$ be a linear map on the finite-dimensional vector space E over the field K. If all the eigenvalues $\lambda_1, \ldots, \lambda_k$ of f belong to K, write*

$$m = (X - \lambda_1)^{r_1} \cdots (X - \lambda_k)^{r_k}$$

for the minimal polynomial of f,

$$\chi_f = (X - \lambda_1)^{n_1} \cdots (X - \lambda_k)^{n_k}$$

for the characteristic polynomial of f, with $1 \leq r_i \leq n_i$, and let

$$W_i = \operatorname{Ker}(\lambda_i \mathrm{id} - f)^{r_i}, \quad i = 1, \ldots, k.$$

Then

(a) $E = W_1 \oplus \cdots \oplus W_k$.

(b) Each W_i *is invariant under* f.

(c) $\dim(W_i) = n_i$.

(d) The minimal polynomial of the restriction $f \mid W_i$ *of* f *to* W_i *is* $(X - \lambda_i)^{r_i}$.

Proof. Parts (a), (b) and (d) have already been proved in Theorem 17.16, so it remains to prove (c). Since W_i is invariant under f, let f_i be the restriction of f to W_i. The characteristic polynomial χ_{f_i} of f_i divides $\chi(f)$, and since $\chi(f)$ has all its roots in K, so does $\chi_i(f)$. By Theorem 12.4, there is a basis of W_i in which f_i is represented by an upper triangular matrix, and since $(\lambda_i \mathrm{id} - f)^{r_i} = 0$, the diagonal entries of this matrix are equal to λ_i. Consequently,

$$\chi_{f_i} = (X - \lambda_i)^{\dim(W_i)},$$

and since χ_{f_i} divides $\chi(f)$, we conclude hat

$$\dim(W_i) \leq n_i, \quad i = 1, \ldots, k.$$

Because E is the direct sum of the W_i, we have $\dim(W_1) + \cdots + \dim(W_k) = n$, and since $n_1 + \cdots + n_k = n$, we must have

$$\dim(W_i) = n_i, \quad i = 1, \ldots, k,$$

proving (c). □

Definition 17.7. If $\lambda \in K$ is an eigenvalue of f, we define a *generalized eigenvector* of f as a nonzero vector $u \in E$ such that

$$(\lambda \mathrm{id} - f)^r(u) = 0, \quad \text{for some } r \geq 1.$$

The *index* of λ is defined as the smallest $r \geq 1$ such that

$$\mathrm{Ker}\,(\lambda \mathrm{id} - f)^r = \mathrm{Ker}\,(\lambda \mathrm{id} - f)^{r+1}.$$

It is clear that $\mathrm{Ker}\,(\lambda \mathrm{id} - f)^i \subseteq \mathrm{Ker}\,(\lambda \mathrm{id} - f)^{i+1}$ for all $i \geq 1$. By Theorem 17.16(d), if $\lambda = \lambda_i$, the index of λ_i is equal to r_i.

Another important consequence of Theorem 17.16 is that f can be written as the sum of a diagonalizable and a nilpotent linear map (which commute). If we write

$$D = \lambda_1 \pi_1 + \cdots + \lambda_k \pi_k,$$

where π_i is the projection from E onto the subspace W_i defined in the proof of Theorem 17.15, since

$$\pi_1 + \cdots + \pi_k = \mathrm{id},$$

we have

$$f = f\pi_1 + \cdots + f\pi_k,$$

and so we get

$$f - D = (f - \lambda_1 \mathrm{id})\pi_1 + \cdots + (f - \lambda_k \mathrm{id})\pi_k.$$

Since the π_i are polynomials in f, they commute with f, and if we write $N = f - D$, using the properties of the π_i, we get

$$N^r = (f - \lambda_1 \mathrm{id})^r \pi_1 + \cdots + (f - \lambda_k \mathrm{id})^r \pi_k.$$

Therefore, if $r = \max\{r_i\}$, we have $(f - \lambda_k \mathrm{id})^r = 0$ for $i = 1, \ldots, k$, which implies that

$$N^r = 0.$$

A linear map $g\colon E \to E$ is said to be *nilpotent* if there is some positive integer r such that $g^r = 0$.

Since N is a polynomial in f, it commutes with f, and thus with D. From

$$D = \lambda_1\pi_1 + \cdots + \lambda_k\pi_k,$$

and

$$\pi_1 + \cdots + \pi_k = \mathrm{id},$$

we see that

$$\begin{aligned} D - \lambda_i \mathrm{id} &= \lambda_1\pi_1 + \cdots + \lambda_k\pi_k - \lambda_i(\pi_1 + \cdots + \pi_k) \\ &= (\lambda_1 - \lambda_i)\pi_1 + \cdots + (\lambda_{i-1} - \lambda_i)\pi_{i-1} + (\lambda_{i+1} - \lambda_i)\pi_{i+1} + \cdots + (\lambda_k - \lambda_i)\pi_k. \end{aligned}$$

Since the projections π_j with $j \neq i$ vanish on W_i, the above equation implies that $D - \lambda_i \mathrm{id}$ vanishes on W_i and that $(D - \lambda_j \mathrm{id})(W_i) \subseteq W_i$, and thus that the minimal polynomial of D is

$$(X - \lambda_1) \cdots (X - \lambda_k).$$

Since the λ_i are distinct, by Theorem 17.11, the linear map D is diagonalizable, so we have shown that when all the eigenvalues of f belong to K, there exist a diagonalizable linear map D and a nilpotent linear map N, such that

$$\begin{aligned} f &= D + N \\ DN &= ND, \end{aligned}$$

and N and D are polynomials in f.

A decomposition of f as above is called a *Jordan decomposition*. In fact, we can prove more: The maps D and N are uniquely determined by f.

Theorem 17.17. *(Jordan Decomposition) Let $f: E \to E$ be a linear map on the finite-dimensional vector space E over the field K. If all the eigenvalues $\lambda_1, \ldots, \lambda_k$ of f belong to K, then there exist a diagonalizable linear map D and a nilpotent linear map N such that*

$$f = D + N$$
$$DN = ND.$$

Furthermore, D and N are uniquely determined by the above equations and they are polynomials in f.

Proof. We already proved the existence part. Suppose we also have $f = D' + N'$, with $D'N' = N'D'$, where D' is diagonalizable, N' is nilpotent, and both are polynomials in f. We need to prove that $D = D'$ and $N = N'$.

Since D' and N' commute with one another and $f = D' + N'$, we see that D' and N' commute with f. Then, D' and N' commute with any polynomial in f; hence they commute with D and N. From

$$D + N = D' + N',$$

we get

$$D - D' = N' - N,$$

and D, D', N, N' commute with one another. Since D and D' are both diagonalizable and commute, by Proposition 17.12, they are simultaneousy diagonalizable, so $D - D'$ is diagonalizable. Since N and N' commute, by the binomial formula, for any $r \geq 1$,

$$(N' - N)^r = \sum_{j=0}^{r} (-1)^j \binom{r}{j} (N')^{r-j} N^j.$$

Since both N and N' are nilpotent, we have $N^{r_1} = 0$ and $(N')^{r_2} = 0$, for some $r_1, r_2 > 0$, so for $r \geq r_1 + r_2$, the right-hand side of the above expression is zero, which shows that $N' - N$ is nilpotent. (In fact, it is easy that $r_1 = r_2 = n$ works). It follows that $D - D' = N' - N$ is both diagonalizable and nilpotent. Clearly, the minimal polynomial of a nilpotent linear map is of the form X^r for some $r > 0$ (and $r \leq \dim(E)$). But $D - D'$ is diagonalizable, so its minimal polynomial has simple roots, which means that $r = 1$. Therefore, the minimal polynomial of $D - D'$ is X, which says that $D - D' = 0$, and then $N = N'$. □

If K is an algebraically closed field, then Theorem 17.17 holds. This is the case when $K = \mathbb{C}$. This theorem reduces the study of linear maps (from E to itself) to the study of nilpotent operators. There is a special normal form for such operators which is discussed in the next section.

17.4 Nilpotent Linear Maps and Jordan Form

This section is devoted to a normal form for nilpotent maps. We follow Godement's exposition [47]. Let $f\colon E \to E$ be a nilpotent linear map on a finite-dimensional vector space over a field K, and assume that f is not the zero map. Then, there is a smallest positive integer $r \geq 1$ such $f^r \neq 0$ and $f^{r+1} = 0$. Clearly, the polynomial X^{r+1} annihilates f, and it is the minimal polynomial of f since $f^r \neq 0$. It follows that $r + 1 \leq n = \dim(E)$. Let us define the subspaces N_i by

$$N_i = \mathrm{Ker}\,(f^i), \quad i \geq 0.$$

Note that $N_0 = (0)$, $N_1 = \mathrm{Ker}\,(f)$, and $N_{r+1} = E$. Also, it is obvious that

$$N_i \subseteq N_{i+1}, \quad i \geq 0.$$

Proposition 17.18. *Given a nilpotent linear map f with $f^r \neq 0$ and $f^{r+1} = 0$ as above, the inclusions in the following sequence are strict:*

$$(0) = N_0 \subset N_1 \subset \cdots \subset N_r \subset N_{r+1} = E.$$

Proof. We proceed by contradiction. Assume that $N_i = N_{i+1}$ for some i with $0 \leq i \leq r$. Since $f^{r+1} = 0$, for every $u \in E$, we have

$$0 = f^{r+1}(u) = f^{i+1}(f^{r-i}(u)),$$

which shows that $f^{r-i}(u) \in N_{i+1}$. Since $N_i = N_{i+1}$, we get $f^{r-i}(u) \in N_i$, and thus $f^r(u) = 0$. Since this holds for all $u \in E$, we see that $f^r = 0$, a contradiction. □

Proposition 17.19. *Given a nilpotent linear map f with $f^r \neq 0$ and $f^{r+1} = 0$, for any integer i with $1 \leq i \leq r$, for any subspace U of E, if $U \cap N_i = (0)$, then $f(U) \cap N_{i-1} = (0)$, and the restriction of f to U is an isomorphism onto $f(U)$.*

Proof. Pick $v \in f(U) \cap N_{i-1}$. We have $v = f(u)$ for some $u \in U$ and $f^{i-1}(v) = 0$, which means that $f^i(u) = 0$. Then, $u \in U \cap N_i$, so $u = 0$ since $U \cap N_i = (0)$, and $v = f(u) = 0$. Therefore, $f(U) \cap N_{i-1} = (0)$. The restriction of f to U is obviously surjective on $f(U)$. Suppose that $f(u) = 0$ for some $u \in U$. Then $u \in U \cap N_1 \subseteq U \cap N_i = (0)$ (since $i \geq 1$), so $u = 0$, which proves that f is also injective on U. □

Proposition 17.20. *Given a nilpotent linear map f with $f^r \neq 0$ and $f^{r+1} = 0$, there exists a sequence of subspace $U_1, \ldots, U_{r+1}$ of E with the following properties:*

(1) $N_i = N_{i-1} \oplus U_i$, for $i = 1, \ldots, r+1$.

(2) We have $f(U_i) \subseteq U_{i-1}$, and the restriction of f to U_i is an injection, for $i = 2, \ldots, r+1$.

Proof. We proceed inductively, by defining the sequence $U_{r+1}, U_r, \ldots, U_1$. We pick U_{r+1} to be any supplement of N_r in $N_{r+1} = E$, so that

$$E = N_{r+1} = N_r \oplus U_{r+1}.$$

Since $f^{r+1} = 0$ and $N_r = \mathrm{Ker}\,(f^r)$, we have $f(U_{r+1}) \subseteq N_r$, and by Proposition 17.19, as $U_{r+1} \cap N_r = (0)$, we have $f(U_{r+1}) \cap N_{r-1} = (0)$. As a consequence, we can pick a supplement U_r of N_{r-1} in N_r so that $f(U_{r+1}) \subseteq U_r$. We have

$$N_r = N_{r-1} \oplus U_r \quad \textit{and} \quad f(U_{r+1}) \subseteq U_r.$$

By Proposition 17.19, f is an injection from U_{r+1} to U_r. Assume inductively that $U_{r+1}, \ldots, U_i$ have been defined for $i \geq 2$ and that they satisfy (1) and (2). Since

$$N_i = N_{i-1} \oplus U_i,$$

we have $U_i \subseteq N_i$, so $f^{i-1}(f(U_i)) = f^i(U_i) = (0)$, which implies that $f(U_i) \subseteq N_{i-1}$. Also, since $U_i \cap N_{i-1} = (0)$, by Proposition 17.19, we have $f(U_i) \cap N_{i-2} = (0)$. It follows that there is a supplement U_{i-1} of N_{i-2} in N_{i-1} that contains $f(U_i)$. We have

$$N_{i-1} = N_{i-2} \oplus U_{i-1} \quad \textit{and} \quad f(U_i) \subseteq U_{i-1}.$$

The fact that f is an injection from U_i into U_{i-1} follows from Proposition 17.19. Therefore, the induction step is proved. The construction stops when $i = 1$. □

Because $N_0 = (0)$ and $N_{r+1} = E$, we see that E is the direct sum of the U_i:

$$E = U_1 \oplus \cdots \oplus U_{r+1},$$

with $f(U_i) \subseteq U_{i-1}$, and f an injection from U_i to U_{i-1}, for $i = r+1, \ldots, 2$. By a clever choice of bases in the U_i, we obtain the following nice theorem.

Theorem 17.21. *For any nilpotent linear map $f\colon E \to E$ on a finite-dimensional vector space E of dimension n over a field K, there is a basis of E such that the matrix N of f is of the form*

$$N = \begin{pmatrix} 0 & \nu_1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \nu_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \nu_n \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix},$$

where $\nu_i = 1$ or $\nu_i = 0$.

Proof. First, apply Proposition 17.20 to obtain a direct sum $E = \bigoplus_{i=1}^{r+1} U_i$. Then, we define a basis of E inductively as follows. First, we choose a basis

$$e_1^{r+1}, \ldots, e_{n_{r+1}}^{r+1}$$

of U_{r+1}. Next, for $i = r+1, \ldots, 2$, given the basis

$$e_1^i, \ldots, e_{n_i}^i$$

of U_i, since f is injective on U_i and $f(U_i) \subseteq U_{i-1}$, the vectors $f(e_1^i), \ldots, f(e_{n_i}^i)$ are linearly independent, so we define a basis of U_{i-1} by completing $f(e_1^i), \ldots, f(e_{n_i}^i)$ to a basis in U_{i-1}:

$$e_1^{i-1}, \ldots, e_{n_i}^{i-1}, e_{n_i+1}^{i-1}, \ldots, e_{n_{i-1}}^{i-1}$$

with

$$e_j^{i-1} = f(e_j^i), \quad j = 1 \ldots, n_i.$$

Since $U_1 = N_1 = \mathrm{Ker}\,(f)$, we have

$$f(e_j^1) = 0, \quad j = 1, \ldots, n_1.$$

These basis vectors can be arranged as the rows of the following matrix:

$$\begin{pmatrix}
e_1^{r+1} & \cdots & e_{n_{r+1}}^{r+1} & & & & & & & & & \\
\vdots & & \vdots & & & & & & & & & \\
e_1^r & \cdots & e_{n_{r+1}}^r & e_{n_{r+1}+1}^r & \cdots & e_{n_r}^r & & & & & & \\
\vdots & & \vdots & \vdots & & \vdots & & & & & & \\
e_1^{r-1} & \cdots & e_{n_{r+1}}^{r-1} & e_{n_{r+1}+1}^{r-1} & \cdots & e_{n_r}^{r-1} & e_{n_r+1}^{r-1} & \cdots & e_{n_{r-1}}^{r-1} & & & \\
\vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots & & & \\
\vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots & & & \\
e_1^1 & \cdots & e_{n_{r+1}}^1 & e_{n_{r+1}+1}^1 & \cdots & e_{n_r}^1 & e_{n_r+1}^1 & \cdots & e_{n_{r-1}}^1 & \cdots & \cdots & e_{n_1}^1
\end{pmatrix}$$

Finally, we define the basis $(e_1, \ldots, e_n)$ by listing each column of the above matrix from the bottom-up, starting with column one, then column two, *etc.* This means that we list the vectors e_j^i in the following order:

For $j = 1, \ldots, n_{r+1}$, list $e_j^1, \ldots, e_j^{r+1}$;

In general, for $i = r, \ldots, 1$,

for $j = n_{i+1} + 1, \ldots, n_i$, list $e_j^1, \ldots, e_j^i$.

Then, because $f(e_j^1) = 0$ and $e_j^{i-1} = f(e_j^i)$ for $i \geq 2$, either

$$f(e_i) = 0 \quad \text{or} \quad f(e_i) = e_{i-1},$$

which proves the theorem. □

As an application of Theorem 17.21, we obtain the *Jordan form* of a linear map.

Definition 17.8. A *Jordan block* is an $r \times r$ matrix $J_r(\lambda)$, of the form

$$J_r(\lambda) = \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & 1 \\ 0 & 0 & 0 & \cdots & \lambda \end{pmatrix},$$

where $\lambda \in K$, with $J_1(\lambda) = (\lambda)$ if $r = 1$. A *Jordan matrix*, J, is an $n \times n$ block diagonal matrix of the form

$$J = \begin{pmatrix} J_{r_1}(\lambda_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & J_{r_m}(\lambda_m) \end{pmatrix},$$

where each $J_{r_k}(\lambda_k)$ is a Jordan block associated with some $\lambda_k \in K$, and with $r_1 + \cdots + r_m = n$.

To simplify notation, we often write $J(\lambda)$ for $J_r(\lambda)$. Here is an example of a Jordan matrix with four blocks:

$$J = \begin{pmatrix} \lambda & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mu & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mu \end{pmatrix}.$$

Theorem 17.22. *(Jordan form) Let E be a vector space of dimension n over a field K and let $f\colon E \to E$ be a linear map. The following properties are equivalent:*

(1) The eigenvalues of f all belong to K (i.e. the roots of the characteristic polynomial χ_f all belong to K).

(2) There is a basis of E in which the matrix of f is a Jordan matrix.

Proof. Assume (1). First we apply Theorem 17.16, and we get a direct sum $E = \bigoplus_{j=1}^{k} W_k$, such that the restriction of $g_i = f - \lambda_j \mathrm{id}$ to W_i is nilpotent. By Theorem 17.21, there is a basis of W_i such that the matrix of the restriction of g_i is of the form

$$G_i = \begin{pmatrix} 0 & \nu_1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \nu_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \nu_{n_i} \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix},$$

where $\nu_i = 1$ or $\nu_i = 0$. Furthermore, over any basis, $\lambda_i \mathrm{id}$ is represented by the diagonal matrix D_i with λ_i on the diagonal. Then, it is clear that we can split $D_i + G_i$ into Jordan blocks by forming a Jordan block for every uninterrupted chain of 1s. By Putting the bases of the W_i together, we obtain a matrix in Jordan form for f.

Now, assume (2). If f can be represented by a Jordan matrix, it is obvious that the diagonal entries are the eigenvalues of f, so they all belong to K. □

Observe that Theorem 17.22 applies if $K = \mathbb{C}$. It turns out that there are uniqueness properties of the Jordan blocks, but we will use more powerful machinery to prove this.

Part II

Preliminaries for Optimization Theory

Chapter 18

Topology

18.1 Metric Spaces and Normed Vector Spaces

This chapter contains a review of basic topological concepts. First metric spaces are defined. Next normed vector spaces are defined. Closed and open sets are defined, and their basic properties are stated. The general concept of a topological space is defined. The closure and the interior of a subset are defined. The subspace topology and the product topology are defined. Continuous maps and homeomorphisms are defined. Limits of sequences are defined. Continuous linear maps and multilinear maps are defined and studied briefly. The chapter ends with the definition of a normed affine space.

Most spaces considered in this book have a topological structure given by a metric or a norm, and we first review these notions. We begin with metric spaces. Recall that $\mathbb{R}_+ = \{x \in \mathbb{R} \mid x \geq 0\}$.

Definition 18.1. A *metric space* is a set E together with a function $d\colon E \times E \to \mathbb{R}_+$, called a *metric, or distance*, assigning a nonnegative real number $d(x, y)$ to any two points $x, y \in E$, and satisfying the following conditions for all $x, y, z \in E$:

(D1) $d(x, y) = d(y, x)$. (symmetry)

(D2) $d(x, y) \geq 0$, and $d(x, y) = 0$ iff $x = y$. (positivity)

(D3) $d(x, z) \leq d(x, y) + d(y, z)$. (triangle inequality)

Geometrically, Condition (D3) expresses the fact that in a triangle with vertices x, y, z, the length of any side is bounded by the sum of the lengths of the other two sides. From (D3), we immediately get

$$|d(x, y) - d(y, z)| \leq d(x, z).$$

Let us give some examples of metric spaces. Recall that the *absolute value* $|x|$ of a real number $x \in \mathbb{R}$ is defined such that $|x| = x$ if $x \geq 0$, $|x| = -x$ if $x < 0$, and for a complex number $x = a + ib$, by $|x| = \sqrt{a^2 + b^2}$.

Example 18.1.

1. Let $E = \mathbb{R}$, and $d(x, y) = |x - y|$, the absolute value of $x - y$. This is the so-called natural metric on $\mathbb{R}$.

2. Let $E = \mathbb{R}^n$ (or $E = \mathbb{C}^n$). We have the *Euclidean metric*

$$d_2(x, y) = \left(|x_1 - y_1|^2 + \cdots + |x_n - y_n|^2\right)^{\frac{1}{2}},$$

   the distance between the points $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$.

3. For every set E, we can define the *discrete metric*, defined such that $d(x, y) = 1$ iff $x \neq y$, and $d(x, x) = 0$.

4. For any $a, b \in \mathbb{R}$ such that $a < b$, we define the following sets:

   $[a, b] = \{x \in \mathbb{R} \mid a \leq x \leq b\}$, (closed interval)

   $(a, b) = \{x \in \mathbb{R} \mid a < x < b\}$, (open interval)

   $[a, b) = \{x \in \mathbb{R} \mid a \leq x < b\}$, (interval closed on the left, open on the right)

   $(a, b] = \{x \in \mathbb{R} \mid a < x \leq b\}$, (interval open on the left, closed on the right)

   Let $E = [a, b]$, and $d(x, y) = |x - y|$. Then $([a, b], d)$ is a metric space.

We will need to define the notion of proximity in order to define convergence of limits and continuity of functions. For this, we introduce some standard "small neighborhoods."

Definition 18.2. Given a metric space E with metric d, for every $a \in E$, for every $\rho \in \mathbb{R}$, with $\rho > 0$, the set

$$B(a, \rho) = \{x \in E \mid d(a, x) \leq \rho\}$$

is called the *closed ball of center a and radius ρ*, the set

$$B_0(a, \rho) = \{x \in E \mid d(a, x) < \rho\}$$

is called the *open ball of center a and radius ρ*, and the set

$$S(a, \rho) = \{x \in E \mid d(a, x) = \rho\}$$

is called the *sphere of center a and radius ρ*. It should be noted that ρ is finite (i.e., not $+\infty$). A subset X of a metric space E is *bounded* if there is a closed ball $B(a, \rho)$ such that $X \subseteq B(a, \rho)$.

Clearly, $B(a, \rho) = B_0(a, \rho) \cup S(a, \rho)$.

Example 18.2.

1. In $E = \mathbb{R}$ with the distance $|x - y|$, an open ball of center a and radius ρ is the open interval $(a - \rho, a + \rho)$.

2. In $E = \mathbb{R}^2$ with the Euclidean metric, an open ball of center a and radius ρ is the set of points inside the disk of center a and radius ρ, excluding the boundary points on the circle.

3. In $E = \mathbb{R}^3$ with the Euclidean metric, an open ball of center a and radius ρ is the set of points inside the sphere of center a and radius ρ, excluding the boundary points on the sphere.

One should be aware that intuition can be misleading in forming a geometric image of a closed (or open) ball. For example, if d is the discrete metric, a closed ball of center a and radius $\rho < 1$ consists only of its center a, and a closed ball of center a and radius $\rho \geq 1$ consists of the entire space!

If $E = [a, b]$, and $d(x, y) = |x - y|$, as in Example 18.1, an open ball $B_0(a, \rho)$, with $\rho < b - a$, is in fact the interval $[a, a + \rho)$, which is closed on the left.

We now consider a very important special case of metric spaces, normed vector spaces. Normed vector spaces have already been defined in Chapter 6 (Definition 6.1) but for the reader's convenience we repeat the definition.

Definition 18.3. Let E be a vector space over a field K, where K is either the field $\mathbb{R}$ of reals, or the field $\mathbb{C}$ of complex numbers. A *norm on* E is a function $\| \| \colon E \to \mathbb{R}_+$, assigning a nonnegative real number $\|u\|$ to any vector $u \in E$, and satisfying the following conditions for all $x, y, z \in E$:

(N1) $\|x\| \geq 0$, and $\|x\| = 0$ iff $x = 0$. (positivity)

(N2) $\|\lambda x\| = |\lambda| \, \|x\|$. (homogeneity (or scaling))

(N3) $\|x + y\| \leq \|x\| + \|y\|$. (triangle inequality)

A vector space E together with a norm $\| \|$ is called a *normed vector space*.

We showed in Chapter 6 that

$$\|-x\| = \|x\|,$$

and from (N3), we get

$$|\|x\| - \|y\|| \leq \|x - y\|.$$

Given a normed vector space E, if we define d such that

$$d(x, y) = \|x - y\|,$$

it is easily seen that d is a metric. Thus, every normed vector space is immediately a metric space. Note that the metric associated with a norm is invariant under translation, that is,

$$d(x+u,\, y+u) = d(x,\, y).$$

For this reason, we can restrict ourselves to open or closed balls of center 0.

Examples of normed vector spaces were given in Example 6.1. We repeat the most important examples.

Example 18.3. Let $E = \mathbb{R}^n$ (or $E = \mathbb{C}^n$). There are three standard norms. For every $(x_1, \ldots, x_n) \in E$, we have the norm $\|x\|_1$, defined such that,

$$\|x\|_1 = |x_1| + \cdots + |x_n|,$$

we have the *Euclidean norm* $\|x\|_2$, defined such that,

$$\|x\|_2 = \left(|x_1|^2 + \cdots + |x_n|^2\right)^{\frac{1}{2}},$$

and the *sup*-norm $\|x\|_\infty$, defined such that,

$$\|x\|_\infty = \max\{|x_i| \mid 1 \leq i \leq n\}.$$

More generally, we define the ℓ_p*-norm* (for $p \geq 1$) by

$$\|x\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}.$$

We proved in Proposition 6.1 that the ℓ_p-norms are indeed norms. The closed unit balls centered at $(0, 0)$ for $\| \|_1$, $\| \|_2$, and $\| \|_\infty$, along with the containment relationships, are shown in Figures 18.1 and 18.2. Figures 18.3 and 18.4 illustrate the situation in $\mathbb{R}^3$.

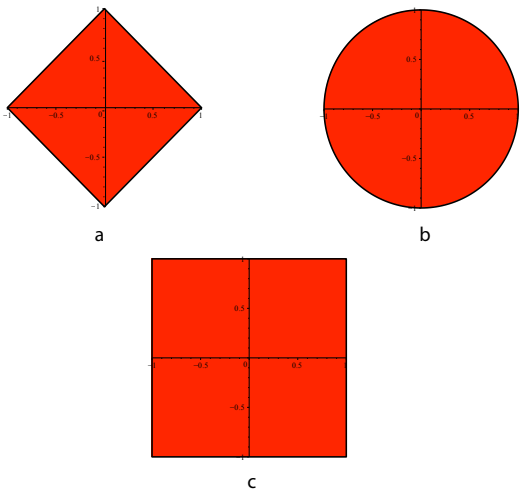


Figure 18.1: Figure (a) shows the diamond shaped closed ball associated with $\| \|_1$. Figure (b) shows the closed unit disk associated with $\| \|_2$, while Figure (c) illustrates the closed unit ball associated with $\| \|_\infty$.

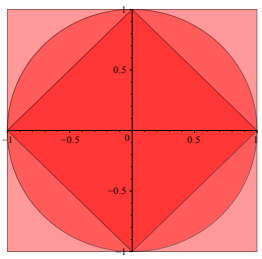


Figure 18.2: The relationship between the closed unit balls centered at $(0, 0)$.

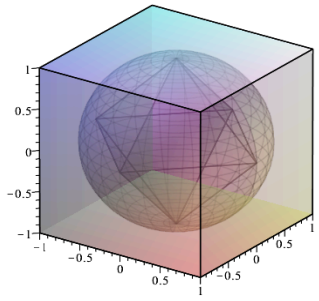


Figure 18.4: The relationship between the closed unit balls centered at $(0, 0, 0)$.

In a normed vector space, we define a closed ball or an open ball of radius ρ as a closed ball or an open ball of center 0. We may use the notation $B(\rho)$ and $B_0(\rho)$.

We will now define the crucial notions of open sets and closed sets, and of a topological space.

Definition 18.4. Let E be a metric space with metric d. A subset $U \subseteq E$ is an *open set* in E if either $U = \emptyset$, or for every $a \in U$, there is some open ball $B_0(a, \rho)$ such that, $B_0(a, \rho) \subseteq U$.[1] A subset $F \subseteq E$ is a *closed set* in E if its complement $E - F$ is open in E. See Figure 18.5.

The set E itself is open, since for every $a \in E$, every open ball of center a is contained in E. In $E = \mathbb{R}^n$, given n intervals $[a_i, b_i]$, with $a_i < b_i$, it is easy to show that the open n-cube

$$\{(x_1, \ldots, x_n) \in E \mid a_i < x_i < b_i,\ 1 \leq i \leq n\}$$

is an open set. In fact, it is possible to find a metric for which such open n-cubes are open balls! Similarly, we can define the closed n-cube

$$\{(x_1, \ldots, x_n) \in E \mid a_i \leq x_i \leq b_i,\ 1 \leq i \leq n\},$$

[1] Recall that $\rho > 0$.

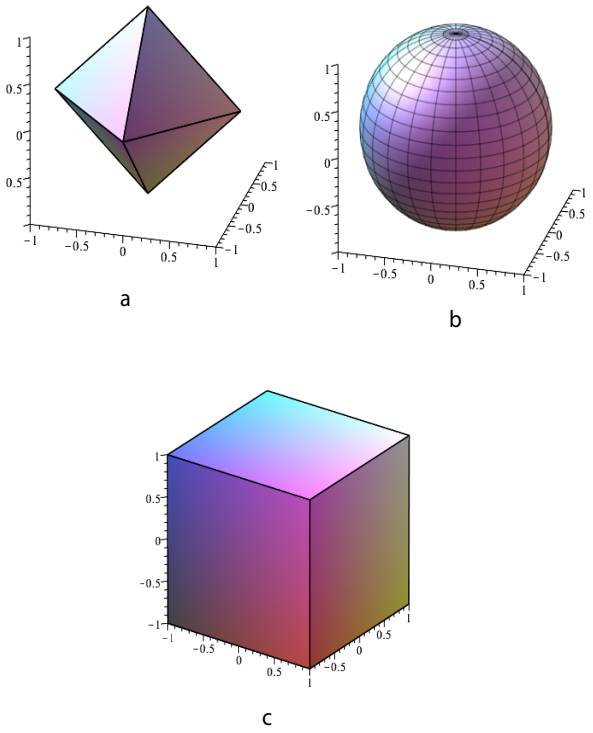


Figure 18.3: Figure (a) shows the octahedral shaped closed ball associated with $\| \|_1$. Figure (b) shows the closed spherical associated with $\| \|_2$, while Figure (c) illustrates the closed unit ball associated with $\| \|_\infty$.

which is a closed set.

The open sets satisfy some important properties that lead to the definition of a topological space.

Proposition 18.1. *Given a metric space E with metric d, the family $\mathcal{O}$ of all open sets defined in Definition 18.4 satisfies the following properties:*

(O1) For every finite family $(U_i)_{1 \leq i \leq n}$ of sets $U_i \in \mathcal{O}$, we have $U_1 \cap \cdots \cap U_n \in \mathcal{O}$, i.e., $\mathcal{O}$ is closed under finite intersections.

(O2) For every arbitrary family $(U_i)_{i \in I}$ of sets $U_i \in \mathcal{O}$, we have $\bigcup_{i \in I} U_i \in \mathcal{O}$, i.e., $\mathcal{O}$ is closed under arbitrary unions.

(O3) $\emptyset \in \mathcal{O}$, and $E \in \mathcal{O}$, i.e., $\emptyset$ and E belong to $\mathcal{O}$.

Furthermore, for any two distinct points $a \neq b$ in E, there exist two open sets U_a and U_b such that, $a \in U_a$, $b \in U_b$, and $U_a \cap U_b = \emptyset$.

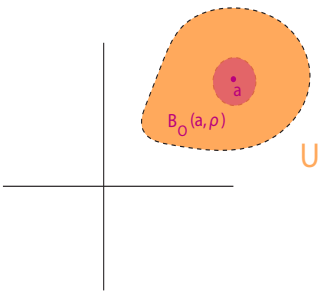


Figure 18.5: An open set U in $E = \mathbb{R}^2$ under the standard Euclidean metric. Any point in the peach set U is surrounded by a small raspberry open set which lies within U.

Proof. It is straightforward. For the last point, letting $\rho = d(a, b)/3$ (in fact $\rho = d(a, b)/2$ works too), we can pick $U_a = B_0(a, \rho)$ and $U_b = B_0(b, \rho)$. By the triangle inequality, we must have $U_a \cap U_b = \emptyset$. □

The above proposition leads to the very general concept of a topological space.

One should be careful that, in general, the family of open sets is not closed under infinite intersections. For example, in $\mathbb{R}$ under the metric $|x - y|$, letting $U_n = (-1/n, +1/n)$, each U_n is open, but $\bigcap_n U_n = \{0\}$, which is not open.

18.2 Topological Spaces

Motivated by Proposition 18.1, a topological space is defined in terms of a family of sets satisfing the properties of open sets stated in that proposition.

Definition 18.5. Given a set E, a *topology on E (or a topological structure on E)*, is defined as a family $\mathcal{O}$ of subsets of E called *open sets*, and satisfying the following three properties:

(1) For every finite family $(U_i)_{1 \leq i \leq n}$ of sets $U_i \in \mathcal{O}$, we have $U_1 \cap \cdots \cap U_n \in \mathcal{O}$, i.e., $\mathcal{O}$ is closed under finite intersections.

(2) For every arbitrary family $(U_i)_{i \in I}$ of sets $U_i \in \mathcal{O}$, we have $\bigcup_{i \in I} U_i \in \mathcal{O}$, i.e., $\mathcal{O}$ is closed under arbitrary unions.

(3) $\emptyset \in \mathcal{O}$, and $E \in \mathcal{O}$, i.e., $\emptyset$ and E belong to $\mathcal{O}$.

A set E together with a topology $\mathcal{O}$ on E is called a *topological space*. Given a topological space $(E, \mathcal{O})$, a subset F of E is a *closed set* if $F = E - U$ for some open set $U \in \mathcal{O}$, i.e., F is the complement of some open set.

It is possible that an open set is also a closed set. For example, $\emptyset$ and E are both open and closed. When a topological space contains a proper nonempty subset U which is both open and closed, the space E is said to be *disconnected*.

A topological space $(E, \mathcal{O})$ is said to satisfy the *Hausdorff separation axiom (or T_2-separation axiom)* if for any two distinct points $a \neq b$ in E, there exist two open sets U_a and U_b such that, $a \in U_a$, $b \in U_b$, and $U_a \cap U_b = \emptyset$. When the T_2-separation axiom is satisfied, we also say that $(E, \mathcal{O})$ is a *Hausdorff space*.

As shown by Proposition 18.1, any metric space is a topological Hausdorff space, the family of open sets being in fact the family of arbitrary unions of open balls. Similarly, any normed vector space is a topological Hausdorff space, the family of open sets being the family of arbitrary unions of open balls. The topology $\mathcal{O}$ consisting of all subsets of E is called the *discrete topology*.

Remark: Most (if not all) spaces used in analysis are Hausdorff spaces. Intuitively, the Hausdorff separation axiom says that there are enough "small" open sets. Without this axiom, some counter-intuitive behaviors may arise. For example, a sequence may have more than one limit point (or a compact set may not be closed). Nevertheless, non-Hausdorff topological spaces arise naturally in algebraic geometry. But even there, some substitute for separation is used.

One of the reasons why topological spaces are important is that the definition of a topology only involves a certain family $\mathcal{O}$ of sets, and not **how** such family is generated from a metric or a norm. For example, different metrics or different norms can define the same family of open sets. Many topological properties only depend on the family $\mathcal{O}$ and not on the specific metric or norm. But the fact that a topology is definable from a metric or a norm is important, because it usually implies nice properties of a space. All our examples will be spaces whose topology is defined by a metric or a norm.

By taking complements, we can state properties of the closed sets dual to those of Definition 18.5. Thus, $\emptyset$ and E are closed sets, and the closed sets are closed under finite unions and arbitrary intersections.

It is also worth noting that the Hausdorff separation axiom implies that for every $a \in E$, the set $\{a\}$ is closed. Indeed, if $x \in E - \{a\}$, then $x \neq a$, and so there exist open sets U_a and U_x such that $a \in U_a$, $x \in U_x$, and $U_a \cap U_x = \emptyset$. See Figure 18.6. Thus, for every $x \in E - \{a\}$, there is an open set U_x containing x and contained in $E - \{a\}$, showing by (O3) that $E - \{a\}$ is open, and thus that the set $\{a\}$ is closed.

Given a topological space $(E, \mathcal{O})$, given any subset A of E, since $E \in \mathcal{O}$ and E is a closed set, the family $\mathcal{C}_A = \{F \mid A \subseteq F, F \text{ a closed set}\}$ of closed sets containing A is nonempty, and since any arbitrary intersection of closed sets is a closed set, the intersection $\bigcap \mathcal{C}_A$ of the sets in the family $\mathcal{C}_A$ is the smallest closed set containing A. By a similar reasoning, the union of all the open subsets contained in A is the largest open set contained in A.

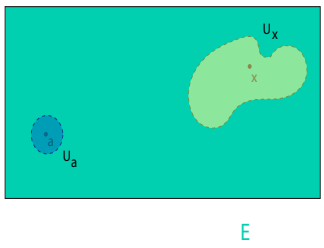


Figure 18.6: A schematic illustration of the Hausdorff separation property.

Definition 18.6. Given a topological space $(E, \mathcal{O})$, given any subset A of E, the smallest closed set containing A is denoted by $\overline{A}$, and is called the *closure, or adherence of* A. See Figure 18.7. A subset A of E is *dense in* E if $\overline{A} = E$. The largest open set contained in A is denoted by $\overset{\circ}{A}$, and is called the *interior of* A. See Figure 18.8. The set $\mathrm{Fr}\, A = \overline{A} \cap \overline{E - A}$ is called the *boundary (or frontier) of* A. We also denote the boundary of A by ∂A. See Figure 18.9.

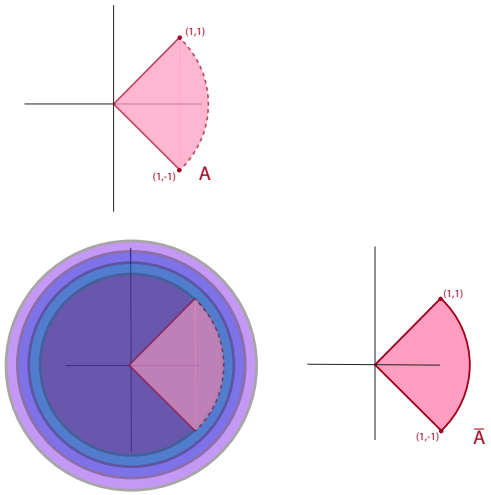


Figure 18.7: The topological space $(E, \mathcal{O})$ is $\mathbb{R}^2$ with topology induced by the Euclidean metric. The subset A is the section $B_0(1)$ in the first and fourth quadrants bound by the lines $y = x$ and $y = -x$. The closure of A is obtained by the intersection of A with the closed unit ball.

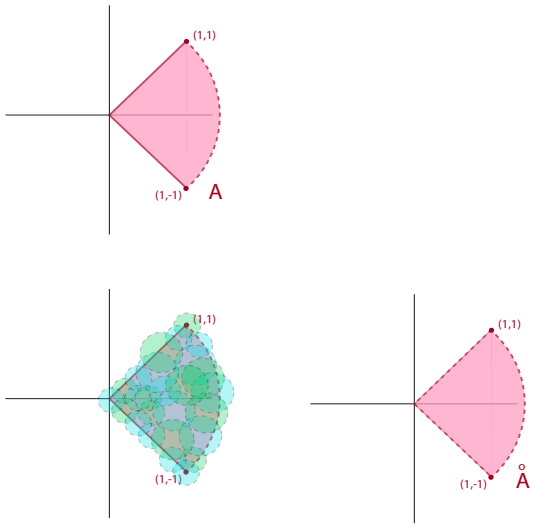


Figure 18.8: The topological space $(E, \mathcal{O})$ is $\mathbb{R}^2$ with topology induced by the Euclidean metric. The subset A is the section $B_0(1)$ in the first and fourth quadrants bound by the lines $y = x$ and $y = -x$. The interior of A is obtained by the covering A with small open balls.

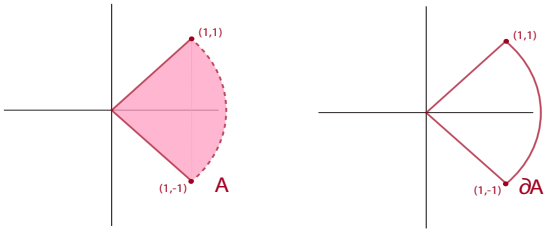


Figure 18.9: The topological space $(E, \mathcal{O})$ is $\mathbb{R}^2$ with topology induced by the Euclidean metric. The subset A is the section $B_0(1)$ in the first and fourth quadrants bound by the lines $y = x$ and $y = -x$. The boundary of A is $\overline{A} - \overset{\circ}{A}$.

Remark: The notation $\overline{A}$ for the closure of a subset A of E is somewhat unfortunate, since $\overline{A}$ is often used to denote the set complement of A in E. Still, we prefer it to more cumbersome notations such as $\mathrm{clo}(A)$, and we denote the complement of A in E by $E - A$ (or sometimes, A^c).

By definition, it is clear that a subset A of E is closed iff $A = \overline{A}$. The set $\mathbb{Q}$ of rationals is dense in $\mathbb{R}$. It is easily shown that $\overline{A} = \overset{\circ}{A} \cup \partial A$ and $\overset{\circ}{A} \cap \partial A = \emptyset$. Another useful characterization of $\overline{A}$ is given by the following proposition.

Proposition 18.2. *Given a topological space $(E, \mathcal{O})$, given any subset A of E, the closure $\overline{A}$ of A is the set of all points $x \in E$ such that for every open set U containing x, then $U \cap A \neq \emptyset$. See Figure 18.10.*

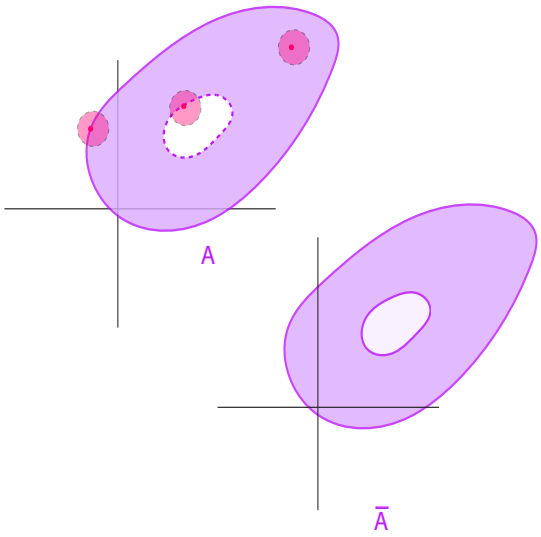


Figure 18.10: The topological space $(E, \mathcal{O})$ is $\mathbb{R}^2$ with topology induced by the Euclidean metric. The purple subset A is illustrated with three red points, each in its closure since the open ball centered at each point has nontrivial intersection with A.

Proof. If $A = \emptyset$, since $\emptyset$ is closed, the proposition holds trivially. Thus, assume that $A \neq \emptyset$. First, assume that $x \in \overline{A}$. Let U be any open set such that $x \in U$. If $U \cap A = \emptyset$, since U is open, then $E - U$ is a closed set containing A, and since $\overline{A}$ is the intersection of all closed sets containing A, we must have $x \in E - U$, which is impossible. Conversely, assume that $x \in E$ is a point such that for every open set U containing x, then $U \cap A \neq \emptyset$. Let F be any closed subset containing A. If $x \notin F$, since F is closed, then $U = E - F$ is an open set such that $x \in U$, and $U \cap A = \emptyset$, a contradiction. Thus, we have $x \in F$ for every closed set containing A, that is, $x \in \overline{A}$. □

Often, it is necessary to consider a subset A of a topological space E, and to view the subset A as a topological space. The following proposition shows how to define a topology on a subset.

Proposition 18.3. *Given a topological space $(E, \mathcal{O})$, given any subset A of E, let*

$$\mathcal{U} = \{U \cap A \mid U \in \mathcal{O}\}$$

be the family of all subsets of A obtained as the intersection of any open set in $\mathcal{O}$ with A. The following properties hold.

(1) The space $(A,\mathcal{U})$ is a topological space.

(2) If E is a metric space with metric d, then the restriction $d_A\colon A \times A \to \mathbb{R}_+$ of the metric d to A defines a metric space. Furthermore, the topology induced by the metric d_A agrees with the topology defined by $\mathcal{U}$, as above.

Proof. Left as an exercise. □

Proposition 18.3 suggests the following definition.

Definition 18.7. Given a topological space $(E, \mathcal{O})$, given any subset A of E, the *subspace topology on A induced by $\mathcal{O}$* is the family $\mathcal{U}$ of open sets defined such that

$$\mathcal{U} = \{U \cap A \mid U \in \mathcal{O}\}$$

is the family of all subsets of A obtained as the intersection of any open set in $\mathcal{O}$ with A. We say that $(A,\mathcal{U})$ has the *subspace topology*. If (E, d) is a metric space, the restriction $d_A\colon A \times A \to \mathbb{R}_+$ of the metric d to A is called the *subspace metric*.

For example, if $E = \mathbb{R}^n$ and d is the Euclidean metric, we obtain the subspace topology on the closed n-cube

$$\{(x_1, \ldots, x_n) \in E \mid a_i \leq x_i \leq b_i,\ 1 \leq i \leq n\}.$$

See Figure 18.11,

One should realize that every open set $U \in \mathcal{O}$ which is entirely contained in A is also in the family $\mathcal{U}$, but $\mathcal{U}$ may contain open sets that are not in $\mathcal{O}$. For example, if $E = \mathbb{R}$ with $|x-y|$, and $A = [a, b]$, then sets of the form $[a, c)$, with $a < c < b$ belong to $\mathcal{U}$, but they are not open sets for $\mathbb{R}$ under $|x-y|$. However, there is agreement in the following situation.

Proposition 18.4. *Given a topological space $(E, \mathcal{O})$, given any subset A of E, if $\mathcal{U}$ is the subspace topology, then the following properties hold.*

(1) If A is an open set $A \in \mathcal{O}$, then every open set $U \in \mathcal{U}$ is an open set $U \in \mathcal{O}$.

(2) If A is a closed set in E, then every closed set w.r.t. the subspace topology is a closed set w.r.t. $\mathcal{O}$.

Proof. Left as an exercise. □

The concept of product topology is also useful. We have the following proposition.

Proposition 18.5. *Given n topological spaces $(E_i, \mathcal{O}_i)$, let $\mathcal{B}$ be the family of subsets of $E_1 \times \cdots \times E_n$ defined as follows:*

$$\mathcal{B} = \{U_1 \times \cdots \times U_n \mid U_i \in \mathcal{O}_i,\ 1 \leq i \leq n\},$$

and let $\mathcal{P}$ be the family consisting of arbitrary unions of sets in $\mathcal{B}$, including $\emptyset$. Then, $\mathcal{P}$ is a topology on $E_1 \times \cdots \times E_n$.

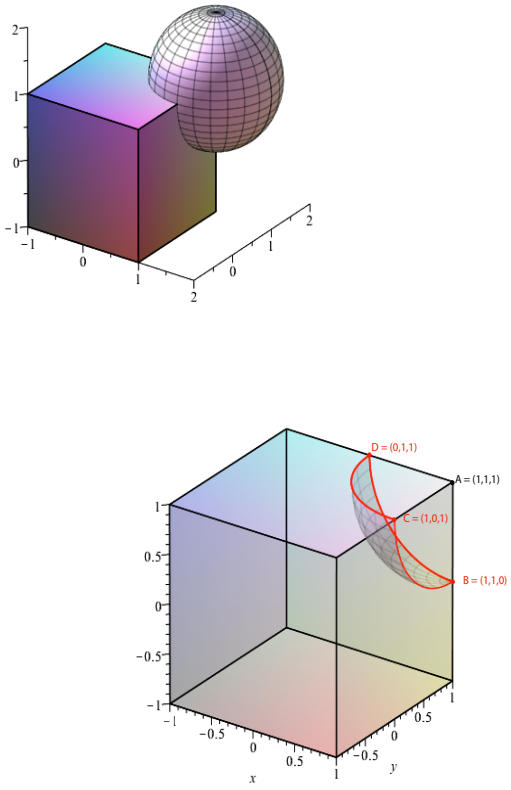


Figure 18.11: An example of an open set in the subspace topology for $\{(x, y, z) \in \mathbb{R}^3 \mid -1 \leq x \leq 1, -1 \leq y \leq 1, -1 \leq z \leq 1\}$. The open set is the corner region $ABCD$ and is obtained by intersection the cube $B_0((1, 1, 1), 1)$.

Proof. Left as an exercise. □

Definition 18.8. Given n topological spaces $(E_i, \mathcal{O}_i)$, the *product topology on* $E_1 \times \cdots \times E_n$ is the family $\mathcal{P}$ of subsets of $E_1 \times \cdots \times E_n$ defined as follows: if

$$\mathcal{B} = \{U_1 \times \cdots \times U_n \mid U_i \in \mathcal{O}_i,\ 1 \leq i \leq n\},$$

then $\mathcal{P}$ is the family consisting of arbitrary unions of sets in $\mathcal{B}$, including $\emptyset$. See Figure 18.12.

If each (E_i, d_{E_i}) is a metric space, there are three natural metrics that can be defined on $E_1 \times \cdots \times E_n$:

$$\begin{aligned}
d_1((x_1, \ldots, x_n), (y_1, \ldots, y_n)) &= d_{E_1}(x_1, y_1) + \cdots + d_{E_n}(x_n, y_n),\\
d_2((x_1, \ldots, x_n), (y_1, \ldots, y_n)) &= \left((d_{E_1}(x_1, y_1))^2 + \cdots + (d_{E_n}(x_n, y_n))^2\right)^{\frac{1}{2}},\\
d_\infty((x_1, \ldots, x_n), (y_1, \ldots, y_n)) &= \max\{d_{E_1}(x_1, y_1), \ldots, d_{E_n}(x_n, y_n)\}.
\end{aligned}$$

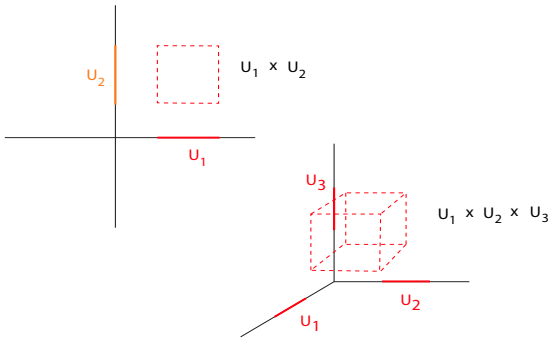


Figure 18.12: Examples of open sets in the product topology for $\mathbb{R}^2$ and $\mathbb{R}^3$ induced by the Euclidean metric.

It is easy to show that

$$\begin{aligned} d_\infty((x_1,\ldots,x_n),(y_1,\ldots,y_n)) &\leq d_2((x_1,\ldots,x_n),(y_1,\ldots,y_n)) \leq d_1((x_1,\ldots,x_n),(y_1,\ldots,y_n)) \\ &\leq n d_\infty((x_1,\ldots,x_n),(y_1,\ldots,y_n)), \end{aligned}$$

so these distances define the same topology, which is the product topology.

If each $(E_i, \| \|_{E_i})$ is a normed vector space, there are three natural norms that can be defined on $E_1 \times \cdots \times E_n$:

$$\begin{aligned} \|(x_1,\ldots,x_n)\|_1 &= \|x_1\|_{E_1} + \cdots + \|x_n\|_{E_n}, \\ \|(x_1,\ldots,x_n)\|_2 &= \left(\|x_1\|_{E_1}^2 + \cdots + \|x_n\|_{E_n}^2\right)^{\frac{1}{2}}, \\ \|(x_1,\ldots,x_n)\|_\infty &= \max\left\{\|x_1\|_{E_1},\ldots,\|x_n\|_{E_n}\right\}. \end{aligned}$$

It is easy to show that

$$\|(x_1,\ldots,x_n)\|_\infty \leq \|(x_1,\ldots,x_n)\|_2 \leq \|(x_1,\ldots,x_n)\|_1 \leq n\|(x_1,\ldots,x_n)\|_\infty,$$

so these norms define the same topology, which is the product topology. It can also be verified that when $E_i = \mathbb{R}$, with the standard topology induced by $|x-y|$, the topology product on $\mathbb{R}^n$ is the standard topology induced by the Euclidean norm.

Definition 18.9. Two metrics d_1 and d_2 on a space E are *equivalent* if they induce the same topology $\mathcal{O}$ on E (i.e., they define the same family $\mathcal{O}$ of open sets). Similarly, two norms $\| \|_1$ and $\| \|_2$ on a space E are *equivalent* if they induce the same topology $\mathcal{O}$ on E.

Remark: Given a topological space $(E, \mathcal{O})$, it is often useful, as in Proposition 18.5, to define the topology $\mathcal{O}$ in terms of a subfamily $\mathcal{B}$ of subsets of E. We say that a family $\mathcal{B}$ of

subsets of E is a *basis for the topology* $\mathcal{O}$, if $\mathcal{B}$ is a subset of $\mathcal{O}$, and if every open set U in $\mathcal{O}$ can be obtained as some union (possibly infinite) of sets in $\mathcal{B}$ (agreeing that the empty union is the empty set).

For example, given any metric space (E, d), $\mathcal{B} = \{B_0(a, \rho) \mid a \in E, \rho > 0\}$. In particular, if $d = \| \ \|_2$, the open intervals form a basis for $\mathbb{R}$, while the open disks form a basis for $\mathbb{R}^2$. The open rectangles also form a basis for $\mathbb{R}^2$ with the standard topology. See Figure 18.13.

It is immediately verified that if a family $\mathcal{B} = (U_i)_{i \in I}$ is a basis for the topology of $(E, \mathcal{O})$, then $E = \bigcup_{i \in I} U_i$, and the intersection of any two sets $U_i, U_j \in \mathcal{B}$ is the union of some sets in the family $\mathcal{B}$ (again, agreeing that the empty union is the empty set). Conversely, a family $\mathcal{B}$ with these properties is the basis of the topology obtained by forming arbitrary unions of sets in $\mathcal{B}$.

A *subbasis for* $\mathcal{O}$ is a family $\mathcal{S}$ of subsets of E, such that the family $\mathcal{B}$ of all finite intersections of sets in $\mathcal{S}$ (including E itself, in case of the empty intersection) is a basis of $\mathcal{O}$. See Figure 18.13

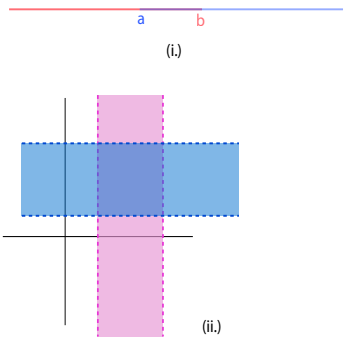


Figure 18.13: Figure $(i.)$ shows that the set of infinite open intervals forms a subbasis for $\mathbb{R}$. Figure $(ii.)$ shows that the infinite open strips form a subbasis for $\mathbb{R}^2$.

The following proposition gives useful criteria for determining whether a family of open subsets is a basis of a topological space.

Proposition 18.6. *Given a topological space $(E, \mathcal{O})$ and a family $\mathcal{B}$ of open subsets in $\mathcal{O}$ the following properties hold:*

(1) The family $\mathcal{B}$ is a basis for the topology $\mathcal{O}$ iff for every open set $U \in \mathcal{O}$ and every $x \in U$, there is some $B \in \mathcal{B}$ such that $x \in B$ and $B \subseteq U$. See Figure 18.14.

(2) The family $\mathcal{B}$ is a basis for the topology $\mathcal{O}$ iff

(a) For every $x \in E$, there is some $B \in \mathcal{B}$ such that $x \in B$.

(b) For any two open subsets, $B_1, B_2 \in \mathcal{B}$, for every $x \in E$, if $x \in B_1 \cap B_2$, then there is some $B_3 \in \mathcal{B}$ such that $x \in B_3$ and $B_3 \subseteq B_1 \cap B_2$. See Figure 18.15.

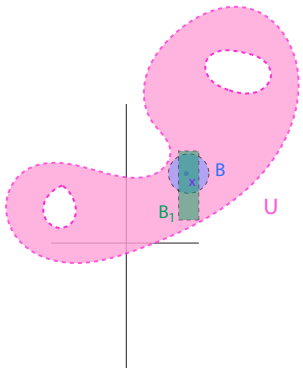


Figure 18.14: Given an open subset U of $\mathbb{R}^2$ and $x \in U$, there exists an open ball B containing x with $B \subset U$. There also exists an open rectangle B_1 containing x with $B_1 \subset U$.

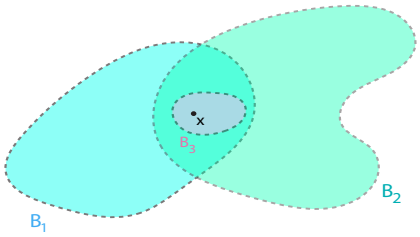


Figure 18.15: A schematic illustration of Condition (b) in Proposition 18.6.

We now consider the fundamental property of continuity.

18.3 Continuous Functions, Limits

Definition 18.10. Let $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ be topological spaces, and let $f\colon E \to F$ be a function. For every $a \in E$, we say that f *is continuous at* a, if for every open set $V \in \mathcal{O}_F$ containing $f(a)$, there is some open set $U \in \mathcal{O}_E$ containing a, such that, $f(U) \subseteq V$. See Figure 18.16. We say that f *is continuous* if it is continuous at every $a \in E$.

Define a *neighborhood of* $a \in E$ as any subset N of E containing some open set $O \in \mathcal{O}$ such that $a \in O$. Now, if f is continuous at a and N is any neighborhood of $f(a)$, there is some open set $V \subseteq N$ containing $f(a)$, and since f is continuous at a, there is some open

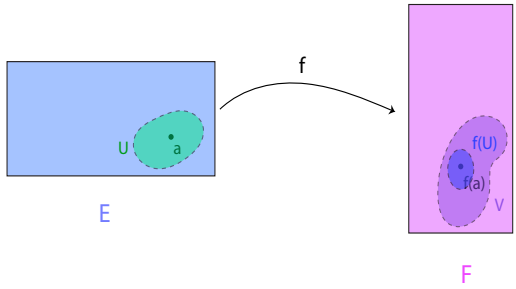


Figure 18.16: A schematic illustration of Definition 18.10.

set U containing a, such that $f(U) \subseteq V$. Since $V \subseteq N$, the open set U is a subset of $f^{-1}(N)$ containing a, and $f^{-1}(N)$ is a neighborhood of a. Conversely, if $f^{-1}(N)$ is a neighborhood of a whenever N is any neighborhood of $f(a)$, it is immediate that f is continuous at a. See Figure 18.17.

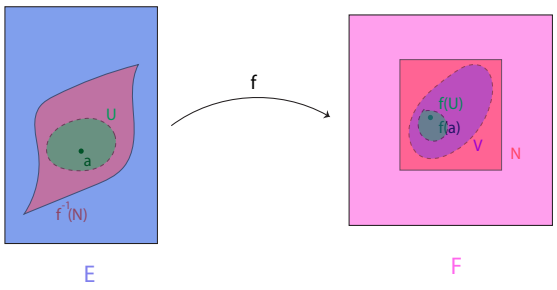


Figure 18.17: A schematic illustration of the neighborhood condition.

It is easy to see that Definition 18.10 is equivalent to the following statements.

Proposition 18.7. *Let $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ be topological spaces, and let $f\colon E \to F$ be a function. For every $a \in E$, the function f is continuous at $a \in E$ iff for every neighborhood N of $f(a) \in F$, then $f^{-1}(N)$ is a neighborhood of a. The function f is continuous on E iff $f^{-1}(V)$ is an open set in $\mathcal{O}_E$ for every open set $V \in \mathcal{O}_F$.*

If E and F are metric spaces defined by metrics d_1 and d_2, we can show easily that f is continuous at a iff

for every $\epsilon > 0$, there is some $\eta > 0$, such that, for every $x \in E$,

$$\text{if } d_1(a, x) \leq \eta, \text{ then } d_2(f(a), f(x)) \leq \epsilon.$$

Similarly, if E and F are normed vector spaces defined by norms $\| \|_1$ and $\| \|_2$, we can show easily that f is continuous at a iff

for every $\epsilon > 0$, there is some $\eta > 0$, such that, for every $x \in E$,

$$\text{if } \|x - a\|_1 \leq \eta, \text{ then } \|f(x) - f(a)\|_2 \leq \epsilon.$$

It is worth noting that continuity is a topological notion, in the sense that equivalent metrics (or equivalent norms) define exactly the same notion of continuity.

If $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ are topological spaces, and $f\colon E \to F$ is a function, for every nonempty subset $A \subseteq E$ of E, we say that f *is continuous on* A if the restriction of f to A is continuous with respect to $(A, \mathcal{U})$ and $(F, \mathcal{O}_F)$, where $\mathcal{U}$ is the subspace topology induced by $\mathcal{O}_E$ on A.

Given a product $E_1 \times \cdots \times E_n$ of topological spaces, as usual, we let $\pi_i\colon E_1 \times \cdots \times E_n \to E_i$ be the projection function such that, $\pi_i(x_1, \ldots, x_n) = x_i$. It is immediately verified that each π_i is continuous.

Given a topological space $(E, \mathcal{O})$, we say that a point $a \in E$ is *isolated* if $\{a\}$ is an open set in $\mathcal{O}$. Then if $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ are topological spaces, any function $f\colon E \to F$ is continuous at every isolated point $a \in E$. In the discrete topology, every point is isolated.

In a nontrivial normed vector space $(E, \| \|)$ (with $E \neq \{0\}$), no point is isolated. To show this, we show that every open ball $B_0(u, \rho,)$ contains some vectors different from u. Indeed, since E is nontrivial, there is some $v \in E$ such that $v \neq 0$, and thus $\lambda = \|v\| > 0$ (by (N1)). Let

$$w = u + \frac{\rho}{\lambda + 1} v.$$

Since $v \neq 0$ and $\rho > 0$, we have $w \neq u$. Then,

$$\|w - u\| = \left\| \frac{\rho}{\lambda + 1} v \right\| = \frac{\rho\lambda}{\lambda + 1} < \rho,$$

which shows that $\|w - u\| < \rho$, for $w \neq u$.

The following proposition is easily shown.

Proposition 18.8. *Given topological spaces $(E, \mathcal{O}_E)$, $(F, \mathcal{O}_F)$, and $(G, \mathcal{O}_G)$, and two functions $f\colon E \to F$ and $g\colon F \to G$, if f is continuous at $a \in E$ and g is continuous at $f(a) \in F$, then $g \circ f\colon E \to G$ is continuous at $a \in E$. Given n topological spaces $(F_i, \mathcal{O}_i)$, for every function $f\colon E \to F_1 \times \cdots \times F_n$, then f is continuous at $a \in E$ iff every $f_i\colon E \to F_i$ is continuous at a, where $f_i = \pi_i \circ f$.*

One can also show that in a metric space (E, d), the distance $d\colon E \times E \to \mathbb{R}$ is continuous, where $E \times E$ has the product topology. By the triangle inequality, we have

$$d(x, y) \leq d(x, x_0) + d(x_0, y_0) + d(y_0, y) = d(x_0, y_0) + d(x_0, x) + d(y_0, y)$$

and

$$d(x_0, y_0) \leq d(x_0, x) + d(x, y) + d(y, y_0) = d(x, y) + d(x_0, x) + d(y_0, y).$$

Consequently,

$$|d(x,y) - d(x_0,y_0)| \leq d(x_0,x) + d(y_0,y),$$

which proves that d is continuous at (x_0,y_0). In fact this shows that d is uniformly continuous; see Definition 18.14.

Similarly, for a normed vector space $(E, \| \|)$, the norm $\| \| \colon E \to \mathbb{R}$ is (uniformly) continuous.

Given a function $f \colon E_1 \times \cdots \times E_n \to F$, we can fix $n-1$ of the arguments, say $a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n$, and view f as a function of the remaining argument,

$$x_i \mapsto f(a_1, \ldots, a_{i-1}, x_i, a_{i+1}, \ldots, a_n),$$

where $x_i \in E_i$. If f is continuous, it is clear that each f_i is continuous.

One should be careful that the converse is false! For example, consider the function $f \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, defined such that,

$$f(x,y) = \frac{xy}{x^2+y^2} \quad \text{if } (x,y) \neq (0,0), \quad \text{and} \quad f(0,0) = 0.$$

The function f is continuous on $\mathbb{R} \times \mathbb{R} - \{(0,0)\}$, but on the line $y = mx$, with $m \neq 0$, we have $f(x,y) = \frac{m}{1+m^2} \neq 0$, and thus, on this line, $f(x,y)$ does not approach 0 when (x,y) approaches $(0,0)$. See Figure 18.18.

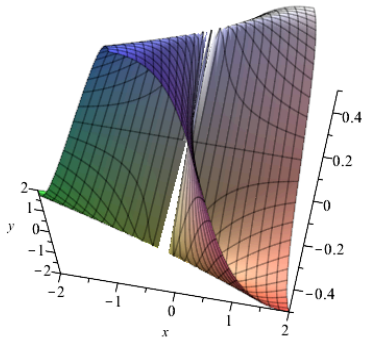


Figure 18.18: The graph of $f(x,y) = \frac{xy}{x^2+y^2}$ for $(x,y) \neq (0,0)$. The bottom of this graph, which shows the approach along the line $y = -x$, does not have a z value of 0.

The following proposition is useful for showing that real-valued functions are continuous.

Proposition 18.9. *If E is a topological space, and $(\mathbb{R}, |x-y|)$ the reals under the standard topology, for any two functions $f \colon E \to \mathbb{R}$ and $g \colon E \to \mathbb{R}$, for any $a \in E$, for any $\lambda \in \mathbb{R}$, if f and g are continuous at a, then $f+g$, λf, $f \cdot g$, are continuous at a, and f/g is continuous at a if $g(a) \neq 0$.*

Proof. Left as an exercise. □

Using Proposition 18.9, we can show easily that every real polynomial function is continuous.

The notion of isomorphism of topological spaces is defined as follows.

Definition 18.11. Let $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ be topological spaces, and let $f\colon E \to F$ be a function. We say that f *is a homeomorphism between* E *and* F if f is bijective, and both $f\colon E \to F$ and $f^{-1}\colon F \to E$ are continuous.

One should be careful that a bijective continuous function $f\colon E \to F$ is not necessarily a homeomorphism. For example, if $E = \mathbb{R}$ with the discrete topology, and $F = \mathbb{R}$ with the standard topology, the identity is not a homeomorphism. Another interesting example involving a parametric curve is given below. Let $L\colon \mathbb{R} \to \mathbb{R}^2$ be the function, defined such that,

$$L_1(t) = \frac{t(1+t^2)}{1+t^4},$$
$$L_2(t) = \frac{t(1-t^2)}{1+t^4}.$$

If we think of $(x(t), y(t)) = (L_1(t), L_2(t))$ as a geometric point in $\mathbb{R}^2$, the set of points $(x(t), y(t))$ obtained by letting t vary in $\mathbb{R}$ from $-\infty$ to $+\infty$, defines a curve having the shape of a "figure eight", with self-intersection at the origin, called the "lemniscate of Bernoulli". See Figure 18.19. The map L is continuous, and in fact bijective, but its inverse L^{-1} is not continuous. Indeed, when we approach the origin on the branch of the curve in the upper left quadrant (i.e., points such that, $x \leq 0$, $y \geq 0$), then t goes to $-\infty$, and when we approach the origin on the branch of the curve in the lower right quadrant (i.e., points such that, $x \geq 0$, $y \leq 0$), then t goes to $+\infty$.

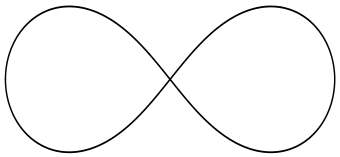


Figure 18.19: The lemniscate of Bernoulli

We also review the concept of limit of a sequence. Given any set E, a *sequence* is any function $x\colon \mathbb{N} \to E$, usually denoted by $(x_n)_{n\in\mathbb{N}}$, or $(x_n)_{n\geq 0}$, or even by (x_n).

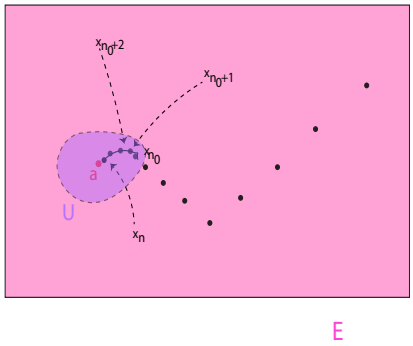


Figure 18.20: A schematic illustration of Definition 18.12.

Definition 18.12. Given a topological space $(E, \mathcal{O})$, we say that *a sequence* $(x_n)_{n\in\mathbb{N}}$ *converges to some* $a \in E$ if for every open set U containing a, there is some $n_0 \geq 0$, such that, $x_n \in U$, for all $n \geq n_0$. We also say that *a is a limit of* $(x_n)_{n\in\mathbb{N}}$. See Figure 18.20.

When E is a metric space with metric d, it is easy to show that this is equivalent to the fact that,

for every $\epsilon > 0$, there is some $n_0 \geq 0$, such that, $d(x_n, a) \leq \epsilon$, for all $n \geq n_0$.

When E is a normed vector space with norm $\| \|$, it is easy to show that this is equivalent to the fact that,

for every $\epsilon > 0$, there is some $n_0 \geq 0$, such that, $\|x_n - a\| \leq \epsilon$, for all $n \geq n_0$.

The following proposition shows the importance of the Hausdorff separation axiom.

Proposition 18.10. *Given a topological space* $(E, \mathcal{O})$*, if the Hausdorff separation axiom holds, then every sequence has at most one limit.*

Proof. Left as an exercise. □

It is worth noting that the notion of limit is topological, in the sense that a sequence converge to a limit b iff it converges to the same limit b in any equivalent metric (and similarly for equivalent norms).

If E is a metric space and if A is a subset of E, there is a convenient way of showing that a point $x \in E$ belongs to the closure $\overline{A}$ of A in terms of sequences.

Proposition 18.11. *Given any metric space* (E, d)*, for any subset* A *of* E *and any point* $x \in E$*, we have* $x \in \overline{A}$ *iff there is a sequence* (a_n) *of points* $a_n \in A$ *converging to* x*.*

Proof. If the sequence (a_n) of points $a_n \in A$ converges to x, then for every open subset U of E containing x, there is some n_0 such that $a_n \in U$ for all $n \geq n_0$, so $U \cap A \neq \emptyset$, and Proposition 18.2 implies that $x \in \overline{A}$.

Conversely, assume that $x \in \overline{A}$. Then for every $n \geq 1$, consider the open ball $B_0(x, 1/n)$. By Proposition 18.2, we have $B_0(x, 1/n) \cap A \neq \emptyset$, so we can pick some $a_n \in B_0(x, 1/n) \cap A$. This, way, we define a sequence (a_n) of points in A, and by construction $d(x, a_n) < 1/n$ for all $n \geq 1$, so the sequence (a_n) converges to x. □

We still need one more concept of limit for functions.

Definition 18.13. Let $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ be topological spaces, let A be some nonempty subset of E, and let $f\colon A \to F$ be a function. For any $a \in \overline{A}$ and any $b \in F$, we say that $f(x)$ *approaches* b *as* x *approaches* a *with values in* A if for every open set $V \in \mathcal{O}_F$ containing b, there is some open set $U \in \mathcal{O}_E$ containing a, such that, $f(U \cap A) \subseteq V$. See Figure 18.21. This is denoted by

$$\lim_{x \to a, x \in A} f(x) = b.$$

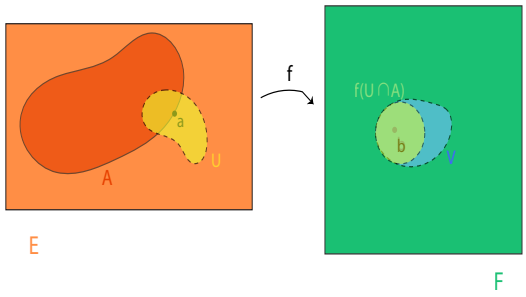


Figure 18.21: A schematic illustration of Definition 18.13.

First, note that by Proposition 18.2, since $a \in \overline{A}$, for every open set U containing a, we have $U \cap A \neq \emptyset$, and the definition is nontrivial. Also, even if $a \in A$, the value $f(a)$ of f at a plays no role in this definition. When E and F are metric space with metrics d_1 and d_2, it can be shown easily that the definition can be stated as follows:

For every $\epsilon > 0$, there is some $\eta > 0$, such that, for every $x \in A$,

$$\text{if } d_1(x, a) \leq \eta, \text{ then } d_2(f(x), b) \leq \epsilon.$$

When E and F are normed vector spaces with norms $\| \|_1$ and $\| \|_2$, it can be shown easily that the definition can be stated as follows:

For every $\epsilon > 0$, there is some $\eta > 0$, such that, for every $x \in A$,

$$\text{if } \|x - a\|_1 \leq \eta, \text{ then } \|f(x) - b\|_2 \leq \epsilon.$$

We have the following result relating continuity at a point and the previous notion.

Proposition 18.12. *Let $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ be two topological spaces, and let $f\colon E \to F$ be a function. For any $a \in E$, the function f is continuous at a iff $f(x)$ approaches $f(a)$ when x approaches a (with values in E).*

Proof. Left as a trivial exercise. □

Another important proposition relating the notion of convergence of a sequence to continuity, is stated without proof.

Proposition 18.13. *Let $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ be two topological spaces, and let $f\colon E \to F$ be a function.*

(1) If f is continuous, then for every sequence $(x_n)_{n\in\mathbb{N}}$ in E, if (x_n) converges to a, then $(f(x_n))$ converges to $f(a)$.

(2) If E is a metric space, and $(f(x_n))$ converges to $f(a)$ whenever (x_n) converges to a, for every sequence $(x_n)_{n\in\mathbb{N}}$ in E, then f is continuous.

A special case of Definition 18.13 will be used when E and F are (nontrivial) normed vector spaces with norms $\| \|_1$ and $\| \|_2$. Let U be any nonempty open subset of E. We showed earlier that E has no isoled points and that every set $\{v\}$ is closed, for every $v \in E$. Since E is nontrivial, for every $v \in U$, there is a nontrivial open ball contained in U (an open ball not reduced to its center). Then, for every $v \in U$, $A = U - \{v\}$ is open and nonempty, and clearly, $v \in \overline{A}$. For any $v \in U$, if $f(x)$ approaches b when x approaches v with values in $A = U - \{v\}$, we say that $f(x)$ *approaches b when x approaches v with values $\neq v$ in U*. This is denoted by

$$\lim_{x\to v, x\in U, x\neq v} f(x) = b.$$

Remark: Variations of the above case show up in the following case: $E = \mathbb{R}$, and F is some arbitrary topological space. Let A be some nonempty subset of $\mathbb{R}$, and let $f\colon A \to F$ be some function. For any $a \in A$, we say that f *is continuous on the right at a* if

$$\lim_{x\to a, x\in A\cap[a,+\infty[} f(x) = f(a).$$

We can define *continuity on the left* at a in a similar fashion.

Let us consider another variation. Let A be some nonempty subset of $\mathbb{R}$, and let $f\colon A \to F$ be some function. For any $a \in A$, we say that f *has a discontinuity of the first kind at a* if

$$\lim_{x\to a, x\in A\cap\,]-\infty,a[} f(x) = f(a_-)$$

and

$$\lim_{x \to a, x \in A \cap]a, +\infty[} f(x) = f(a_+)$$

both exist, and either $f(a_-) \neq f(a)$, or $f(a_+) \neq f(a)$.

Note that it is possible that $f(a_-) = f(a_+)$, but f is still discontinuous at a if this common value differs from $f(a)$. Functions defined on a nonempty subset of $\mathbb{R}$, and that are continuous, except for some points of discontinuity of the first kind, play an important role in analysis.

In a metric space, there is another important notion of continuity, namely uniform continuity.

Definition 18.14. Given two metric spaces, (E, d_E) and (F, d_F), a function, $f\colon E \to F$, is *uniformly continuous* if for every $\epsilon > 0$, there is some $\eta > 0$, such that, for all $a, b \in E$,

$$\text{if} \quad d_E(a, b) \leq \eta \quad \text{then} \quad d_F(f(a), f(b)) \leq \epsilon.$$

See Figures 18.22 and 18.23.

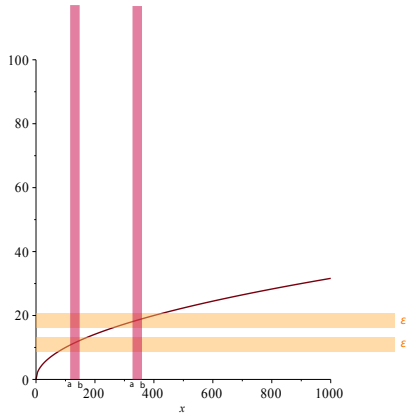


Figure 18.22: The real valued function $f(x) = \sqrt{x}$ is uniformly continuous over $(0, \infty)$. Fix ϵ. If the x values lie within the rose colored η strip, the y values always lie within the peach ϵ strip.

As we saw earlier, the metric on a metric space is uniformly continuous, and the norm on a normed metric space is uniformly continuous.

Before considering differentials, we need to look at the continuity of linear maps.

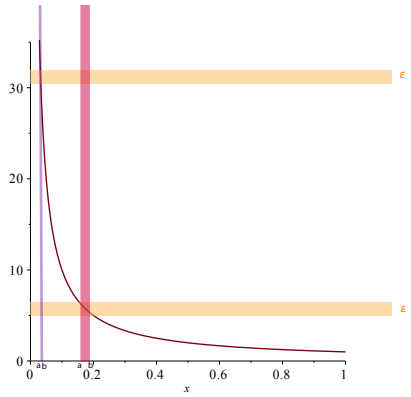


Figure 18.23: The real valued function $f(x) = 1/x$ is not uniformly continuous over $(0, \infty)$. Fix ϵ. In order for the y values to lie within the peach epsilon strip, the widths of the eta strips decrease as $x \to 0$.

18.4 Continuous Linear and Multilinear Maps

If E and F are normed vector spaces, we first characterize when a linear map $f\colon E \to F$ is continuous.

Proposition 18.14. *Given two normed vector spaces E and F, for any linear map $f\colon E \to F$, the following conditions are equivalent:*

(1) The function f is continuous at 0.

(2) There is a constant $k \geq 0$ such that,

$$\|f(u)\| \leq k, \text{ for every } u \in E \text{ such that } \|u\| \leq 1.$$

(3) There is a constant $k \geq 0$ such that,

$$\|f(u)\| \leq k\|u\|, \text{ for every } u \in E.$$

(4) The function f is continuous at every point of E.

Proof. Assume (1). Then for every $\epsilon > 0$, there is some $\eta > 0$ such that, for every $u \in E$, if $\|u\| \leq \eta$, then $\|f(u)\| \leq \epsilon$. Pick $\epsilon = 1$, so that there is some $\eta > 0$ such that, if $\|u\| \leq \eta$, then $\|f(u)\| \leq 1$. If $\|u\| \leq 1$, then $\|\eta u\| \leq \eta\|u\| \leq \eta$, and so, $\|f(\eta u)\| \leq 1$, that is, $\eta\|f(u)\| \leq 1$, which implies $\|f(u)\| \leq \eta^{-1}$. Thus, Condition (2) holds with $k = \eta^{-1}$.

Assume that (2) holds. If $u = 0$, then by linearity, $f(0) = 0$, and thus $\|f(0)\| \leq k\|0\|$ holds trivially for all $k \geq 0$. If $u \neq 0$, then $\|u\| > 0$, and since

$$\left\| \frac{u}{\|u\|} \right\| = 1,$$

we have

$$\left\| f\left(\frac{u}{\|u\|}\right)\right\| \leq k,$$

which implies that

$$\|f(u)\| \leq k\|u\|.$$

Thus, Condition (3) holds.

If (3) holds, then for all $u, v \in E$, we have

$$\|f(v) - f(u)\| = \|f(v-u)\| \leq k\|v-u\|.$$

If $k = 0$, then f is the zero function, and continuity is obvious. Otherwise, if $k > 0$, for every $\epsilon > 0$, if $\|v-u\| \leq \frac{\epsilon}{k}$, then $\|f(v-u)\| \leq \epsilon$, which shows continuity at every $u \in E$. Finally, it is obvious that (4) implies (1). □

Among other things, Proposition 18.14 shows that a linear map is continuous iff the image of the unit (closed) ball is bounded. Since a continuous linear map satisfies the condition $\|f(u)\| \leq k\|u\|$ (for some $k \geq 0$), it is also uniformly continuous.

If E and F are normed vector spaces, the set of all continuous linear maps $f\colon E \to F$ is denoted by $\mathcal{L}(E;F)$.

Using Proposition 18.14, we can define a norm on $\mathcal{L}(E;F)$ which makes it into a normed vector space. This definition has already been given in Chapter 6 (Definition 6.7) but for the reader's convenience, we repeat it here.

Definition 18.15. Given two normed vector spaces E and F, for every continuous linear map $f\colon E \to F$, we define the *norm* $\|f\|$ *of* f as

$$\|f\| = \inf\{k \geq 0 \mid \|f(x)\| \leq k\|x\|, \text{ for all } x \in E\} = \sup\{\|f(x)\| \mid \|x\| \leq 1\}.$$

From Definition 18.15, for every continuous linear map $f \in \mathcal{L}(E;F)$, we have

$$\|f(x)\| \leq \|f\|\|x\|,$$

for every $x \in E$. It is easy to verify that $\mathcal{L}(E;F)$ is a normed vector space under the norm of Definition 18.15. Furthermore, if E, F, G, are normed vector spaces, and $f\colon E \to F$ and $g\colon F \to G$ are continuous linear maps, we have

$$\|g \circ f\| \leq \|g\|\|f\|.$$

We can now show that when $E = \mathbb{R}^n$ or $E = \mathbb{C}^n$, with any of the norms $\| \ \|_1$, $\| \ \|_2$, or $\| \ \|_\infty$, then every linear map $f\colon E \to F$ is continuous.

Proposition 18.15. *If $E = \mathbb{R}^n$ or $E = \mathbb{C}^n$, with any of the norms $\| \ \|_1$, $\| \ \|_2$, or $\| \ \|_\infty$, and F is any normed vector space, then every linear map $f\colon E \to F$ is continuous.*

Proof. Let $(e_1, \ldots, e_n)$ be the standard basis of $\mathbb{R}^n$ (a similar proof applies to $\mathbb{C}^n$). In view of Proposition 6.2, it is enough to prove the proposition for the norm

$$\|x\|_\infty = \max\{|x_i| \mid 1 \leq i \leq n\}.$$

We have,

$$\|f(v) - f(u)\| = \|f(v-u)\| = \left\| f\Big(\sum_{1 \leq i \leq n} (v_i - u_i) e_i \Big) \right\| = \left\| \sum_{1 \leq i \leq n} (v_i - u_i) f(e_i) \right\|,$$

and so,

$$\|f(v) - f(u)\| \leq \Big(\sum_{1 \leq i \leq n} \|f(e_i)\| \Big) \max_{1 \leq i \leq n} |v_i - u_i| = \Big(\sum_{1 \leq i \leq n} \|f(e_i)\| \Big) \|v - u\|_\infty.$$

By the argument used in Proposition 18.14 to prove that (3) implies (4), f is continuous. □

Actually, we proved in Theorem 6.3 that if E is a vector space of finite dimension, then any two norms are equivalent, so that they define the same topology. This fact together with Proposition 18.15 prove the following:

Theorem 18.16. *If E is a vector space of finite dimension (over $\mathbb{R}$ or $\mathbb{C}$), then all norms are equivalent (define the same topology). Furthermore, for any normed vector space F, every linear map $f\colon E \to F$ is continuous.*

If E is a normed vector space of infinite dimension, a linear map $f\colon E \to F$ may not be continuous. As an example, let E be the infinite vector space of all polynomials over $\mathbb{R}$. Let

$$\|P(X)\| = \sup_{0 \leq x \leq 1} |P(x)|.$$

We leave as an exercise to show that this is indeed a norm. Let $F = \mathbb{R}$, and let $f\colon E \to F$ be the map defined such that, $f(P(X)) = P(3)$. It is clear that f is linear. Consider the sequence of polynomials

$$P_n(X) = \left(\frac{X}{2} \right)^n.$$

It is clear that $\|P_n\| = \left(\frac{1}{2} \right)^n$, and thus, the sequence P_n has the null polynomial as a limit. However, we have

$$f(P_n(X)) = P_n(3) = \left(\frac{3}{2} \right)^n,$$

and the sequence $f(P_n(X))$ diverges to $+\infty$. Consequently, in view of Proposition 18.13 (1), f is not continuous.

We now consider the continuity of multilinear maps. We treat explicitly bilinear maps, the general case being a straightforward extension.

Proposition 18.17. *Given normed vector spaces E, F and G, for any bilinear map $f\colon E \times E \to G$, the following conditions are equivalent:*

(1) The function f is continuous at $\langle 0, 0\rangle$.

2) There is a constant $k \geq 0$ such that,

$$\|f(u, v)\| \leq k, \text{ for all } u, v \in E \text{ such that } \|u\|, \|v\| \leq 1.$$

3) There is a constant $k \geq 0$ such that,

$$\|f(u, v)\| \leq k\|u\|\|v\|, \text{ for all } u, v \in E.$$

4) The function f is continuous at every point of $E \times F$.

Proof. It is similar to that of Proposition 18.14, with a small subtlety in proving that (3) implies (4), namely that two different η's that are not independent are needed. □

In contrast to continuous linear maps, which must be uniformly continuous, nonzero continuous bilinear maps are **not** uniformly continuous. Let $f\colon E \times F \to G$ be a continuous bilinear map such that $f(a, b) \neq 0$ for some $a \in E$ and some $b \in F$. Consider the sequences (u_n) and (v_n) (with $n \geq 1$) given by

$$\begin{aligned} u_n &= (x_n, y_n) = (na, nb) \\ v_n &= (x'_n, y'_n) = \left(\left(n + \frac{1}{n}\right) a, \left(n + \frac{1}{n}\right) b\right). \end{aligned}$$

Obviously

$$\|v_n - u_n\| \leq \frac{1}{n}(\|a\| + \|b\|),$$

so $\lim_{n\mapsto\infty} \|v_n - u_n\| = 0$. On the other hand

$$f(x'_n, y'_n) - f(x_n, y_n) = \left(2 + \frac{1}{n^2}\right) f(a, b),$$

and thus $\lim_{n\mapsto\infty} \|f(x'_n, y'_n) - f(x_n, y_n)\| = 2\,\|f(a, b)\| \neq 0$, which shows that f is not uniformly continuous, because if this was the case, this limit would be zero.

If E, F, and G, are normed vector spaces, we denote the set of all continuous bilinear maps $f\colon E \times F \to G$ by $\mathcal{L}_2(E, F; G)$. Using Proposition 18.17, we can define a norm on $\mathcal{L}_2(E, F; G)$ which makes it into a normed vector space.

Definition 18.16. Given normed vector spaces E, F, and G, for every continuous bilinear map $f\colon E \times F \to G$, we define the *norm* $\|f\|$ *of* f as

$$\begin{aligned}\|f\| &= \inf\{k \geq 0 \mid \|f(x,y)\| \leq k\|x\|\|y\|, \text{ for all } x, y \in E\} \\ &= \sup\{\|f(x,y)\| \mid \|x\|, \|y\| \leq 1\}.\end{aligned}$$

From Definition 18.15, for every continuous bilinear map $f \in \mathcal{L}_2(E, F; G)$, we have

$$\|f(x,y)\| \leq \|f\|\|x\|\|y\|,$$

for all $x, y \in E$. It is easy to verify that $\mathcal{L}_2(E, F; G)$ is a normed vector space under the norm of Definition 18.16.

Given a bilinear map $f\colon E \times F \to G$, for every $u \in E$, we obtain a linear map denoted $fu\colon F \to G$, defined such that, $fu(v) = f(u, v)$. Furthermore, since

$$\|f(x,y)\| \leq \|f\|\|x\|\|y\|,$$

it is clear that fu is continuous. We can then consider the map $\varphi\colon E \to \mathcal{L}(F; G)$, defined such that, $\varphi(u) = fu$, for any $u \in E$, or equivalently, such that,

$$\varphi(u)(v) = f(u, v).$$

Actually, it is easy to show that φ is linear and continuous, and that $\|\varphi\| = \|f\|$. Thus, $f \mapsto \varphi$ defines a map from $\mathcal{L}_2(E, F; G)$ to $\mathcal{L}(E; \mathcal{L}(F; G))$. We can also go back from $\mathcal{L}(E; \mathcal{L}(F; G))$ to $\mathcal{L}_2(E, F; G)$. We summarize all this in the following proposition.

Proposition 18.18. *Let E, F, G be three normed vector spaces. The map $f \mapsto \varphi$, from $\mathcal{L}_2(E, F; G)$ to $\mathcal{L}(E; \mathcal{L}(F; G))$, defined such that, for every $f \in \mathcal{L}_2(E, F; G)$,*

$$\varphi(u)(v) = f(u, v),$$

is an isomorphism of vector spaces, and furthermore, $\|\varphi\| = \|f\|$.

As a corollary of Proposition 18.18, we get the following proposition which will be useful when we define second-order derivatives.

Proposition 18.19. *Let E, F be normed vector spaces. The map app from $\mathcal{L}(E; F) \times E$ to F, defined such that, for every $f \in \mathcal{L}(E; F)$, for every $u \in E$,*

$$app(f, u) = f(u),$$

is a continuous bilinear map.

Remark: If E and F are nontrivial, it can be shown that $\|\text{app}\| = 1$. It can also be shown that composition

$$\circ\colon \mathcal{L}(E;F) \times \mathcal{L}(F;G) \to \mathcal{L}(E;G),$$

is bilinear and continuous.

The above propositions and definition generalize to arbitrary n-multilinear maps, with $n \geq 2$. Proposition 18.17 extends in the obvious way to any n-multilinear map $f\colon E_1 \times \cdots \times E_n \to F$, but condition (3) becomes:

There is a constant $k \geq 0$ such that,

$$\|f(u_1, \ldots, u_n)\| \leq k\|u_1\| \cdots \|u_n\|, \text{ for all } u_1 \in E_1, \ldots, u_n \in E_n.$$

Definition 18.16 also extends easily to

$$\begin{aligned}\|f\| &= \inf\{k \geq 0 \mid \|f(x_1, \ldots, x_n)\| \leq k\|x_1\| \cdots \|x_n\|, \text{ for all } x_i \in E_i, 1 \leq i \leq n\} \\ &= \sup\{\|f(x_1, \ldots, x_n)\| \mid \|x_n\|, \ldots, \|x_n\| \leq 1\}.\end{aligned}$$

Proposition 18.18 is also easily extended, and we get an isomorphism between continuous n-multilinear maps in $\mathcal{L}_n(E_1, \ldots, E_n; F)$, and continuous linear maps in

$$\mathcal{L}(E_1; \mathcal{L}(E_2; \ldots; \mathcal{L}(E_n; F)))$$

An obvious extension of Proposition 18.19 also holds.

Complete metric spaces and complete normed vector spaces are important tools in analysis and optimization theory, so we include some sections covering the basics.

18.5 Complete Metric Spaces and Banach Spaces

Definition 18.17. Given a metric space, (E, d), a sequence, $(x_n)_{n\in\mathbb{N}}$, in E is a *Cauchy sequence* if the following condition holds: for every $\epsilon > 0$, there is some $p \geq 0$, such that, for all $m, n \geq p$, then $d(x_m, x_n) \leq \epsilon$.

If every Cauchy sequence in (E, d) converges we say that (E, d) is a *complete metric space*. A normed vector space $(E, \| \|)$ over $\mathbb{R}$ (or $\mathbb{C}$) which is a complete metric space for the distance $d(u, v) = \|v - u\|$, is called a *Banach space*.

The standard example of a complete metric space is the set $\mathbb{R}$ of real numbers. As a matter of fact, the set $\mathbb{R}$ can be defined as the "completion" of the set $\mathbb{Q}$ of rationals. The spaces $\mathbb{R}^n$ and $\mathbb{C}^n$ under their standard topology are complete metric spaces.

It can be shown that every normed vector space of finite dimension is a Banach space (is complete). It can also be shown that if E and F are normed vector spaces, and F is a

Banach space, then $\mathcal{L}(E;F)$ is a Banach space. If E, F and G are normed vector spaces, and G is a Banach space, then $\mathcal{L}_2(E,F;G)$ is a Banach space.

An arbitrary metric space (E,d) is not necessarily complete, but there is a construction of a metric space $(\widehat{E},\widehat{d})$ such that $\widehat{E}$ is complete, and there is a continuous (injective) distance-preserving map $\varphi\colon E \to \widehat{E}$ such that $\varphi(E)$ is dense in $\widehat{E}$. This is a generalization of the construction of the set $\mathbb{R}$ of real numbers from the set $\mathbb{Q}$ of rational numbers in terms of Cauchy sequences. This construction can be immediately adapted to a normed vector space $(E, \| \|)$ to embed $(E, \| \|)$ into a complete normed vector space $(\widehat{E}, \| \|_{\widehat{E}})$ (a Banach space). This construction is used heavily in integration theory, where E is a set of functions.

18.6 Completion of a Metric Space

In order to prove a kind of uniqueness result for the completion $(\widehat{E},\widehat{d})$ of a metric space (E,d), we need the following result about extending a uniformly continuous function.

Recall that E_0 is dense in E iff $\overline{E_0} = E$. Since E is a metric space, by Proposition 18.11, this means that for every $x \in E$, there is some sequence (x_n) converging to x, with $x_n \in E_0$.

Theorem 18.20. *Let E and F be two metric spaces, let E_0 be a dense subspace of E, and let $f_0\colon E_0 \to F$ be a continuous function. If f_0 is uniformly continuous and if F is complete, then there is a unique uniformly continuous function $f\colon E \to F$ extending f_0.*

Proof. We follow Schwartz's proof; see Schwartz [89] (Chapter XI, Section 3, Theorem 1).

Step 1. We begin by constructing a function $f\colon E \to F$ extending f_0. Since E_0 is dense in E, for every $x \in E$, there is some sequence (x_n) converging to x, with $x_n \in E_0$. Then the sequence (x_n) is a Cauchy sequence in E. We claim that $(f_0(x_n))$ is a Cauchy sequence in F.

Proof of the claim. For every $\epsilon > 0$, since f_0 is uniformly continuous, there is some $\eta > 0$ such that for all $(y,z) \in E_0$, if $d(y,z) \leq \eta$, then $d(f_0(y), f_0(z)) \leq \epsilon$. Since (x_n) is a Cauchy sequence with $x_n \in E_0$, there is some integer $p > 0$ such that if $m, n \geq p$, then $d(x_m, x_n) \leq \eta$, thus $d(f_0(x_m), f_0(x_n)) \leq \epsilon$, which proves that $(f_0(x_n))$ is a Cauchy sequence in F. □

Since F is complete and $(f_0(x_n))$ is a Cauchy sequence in F, the sequence $(f_0(x_n))$ converges to some element of F; denote this element by $f(x)$.

Step 2. Let us now show that $f(x)$ does not depend on the sequence (x_n) converging to x. Suppose that (x'_n) and (x''_n) are two sequences of elements in E_0 converging to x. Then the mixed sequence

$$x'_0, x''_0, x'_1, x''_1, \ldots, x'_n, x''_n, \ldots,$$

also converges to x. It follows that the sequence

$$f_0(x'_0), f_0(x''_0), f_0(x'_1), f_0(x''_1), \ldots, f_0(x'_n), f_0(x''_n), \ldots,$$

is a Cauchy sequence in F, and since F is complete, it converges to some element of F, which implies that the sequences $(f_0(x'_n))$ and $(f_0(x''_n))$ converge to the same limit.

As a summary, we have defined a function $f\colon E \to F$ by

$$f(x) = \lim_{n\mapsto\infty} f_0(x_n).$$

for any sequence (x_n) converging to x, with $x_n \in E_0$.

Step 3. The function f extends f_0. Since every element $x \in E_0$ is the limit of the constant sequence (x_n) with $x_n = x$ for all $n \geq 0$, by definition $f(x)$ is the limit of the sequence $(f_0(x_n))$, which is the constant sequence with value $f_0(x)$, so $f(x) = f_0(x)$; that is, f extends f_0.

Step 4. We now prove that f is uniformly continuous. Since f_0 is uniformly continuous, for every $\epsilon > 0$, there is some $\eta > 0$ such that if $a, b \in E_0$ and $d(a, b) \leq \eta$, then $d(f_0(a), f_0(b)) \leq \epsilon$. Consider any two points $x, y \in E$ such that $d(x, y) \leq \eta/2$. We claim that $d(f(x), f(y)) \leq \epsilon$, which shows that f is uniformly continuous.

Let (x_n) be a sequence of points in E_0 converging to x, and let (y_n) be a sequence of points in E_0 converging to y. By the triangle inequality,

$$d(x_n, y_n) \leq d(x_n, x) + d(x, y) + d(y, y_n) = d(x, y) + d(x_n, x) + d(y_n, y),$$

and since (x_n) converges to x and (y_n) converges to y, there is some integer $p > 0$ such that for all $n \geq p$, we have $d(x_n, x) \leq \eta/4$ and $d(y_n, y) \leq \eta/4$, and thus

$$d(x_n, y_n) \leq d(x, y) + \frac{\eta}{2}.$$

Since we assumed that $d(x, y) \leq \eta/2$, we get $d(x_n, y_n) \leq \eta$ for all $n \geq p$, and by uniform continuity of f_0, we get

$$d(f_0(x_n), f_0(y_n)) \leq \epsilon$$

for all $n \geq p$. Since the distance function on F is also continuous, and since $(f_0(x_n))$ converges to $f(x)$ and $(f_0(y_n))$ converges to $f(y)$, we deduce that the sequence $(d(f_0(x_n), f_0(y_n)))$ converges to $d(f(x), f(y))$. This implies that $d(f(x), f(y)) \leq \epsilon$, as desired.

Step 5. It remains to prove that f is unique. Since E_0 is dense in E, for every $x \in E$, there is some sequence (x_n) converging to x, with $x_n \in E_0$. Since f extends f_0 and since f is continuous, we get

$$f(x) = \lim_{n\mapsto\infty} f_0(x_n),$$

which only depends on f_0 and x, and shows that f is unique. □

Remark: It can be shown that the theorem no longer holds if we either omit the hypothesis that F is complete or omit that f_0 is uniformly continuous.

For example, if $E_0 \neq E$ and if we let $F = E_0$ and f_0 be the identity function, it is easy to see that f_0 cannot be extended to a continuous function from E to E_0 (for any $x \in E - E_0$, any continous extension f of f_0 would satisfy $f(x) = x$, which is absurd since $x \notin E_0$).

If f_0 is continuous but not uniformly continuous, a counter-example can be given by using $E = \overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$ made into a metric space, $E_0 = \mathbb{R}$, $F = \mathbb{R}$, and f_0 the identity function; for details, see Schwartz [89] (Chapter XI, Section 3, page 134).

Definition 18.18. If (E, d_E) and (F, d_F) are two metric spaces, then a function $f\colon E \to F$ is *distance-preserving*, or an *isometry*, if

$$d_F(f(x), f(y)) = d_E(x, y), \quad \text{for all for all } x, y \in E.$$

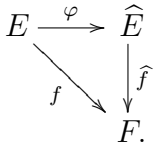
Observe that an isometry must be injective, because if $f(x) = f(y)$, then $d_F(f(x), f(y)) = 0$, and since $d_F(f(x), f(y)) = d_E(x, y)$, we get $d_E(x, y) = 0$, but $d_E(x, y) = 0$ implies that $x = y$. Also, an isometry is uniformly continuous (since we can pick $\eta = \epsilon$ to satisfy the condition of uniform continuity). However, an isometry is not necessarily surjective.

We now give a construction of the completion of a metric space. This construction is just a generalization of the classical construction of $\mathbb{R}$ from $\mathbb{Q}$ using Cauchy sequences.

Theorem 18.21. *Let (E, d) be any metric space. There is a complete metric space $(\widehat{E}, \widehat{d})$ called a completion of (E, d), and a distance-preserving (uniformly continuous) map $\varphi\colon E \to \widehat{E}$ such that $\varphi(E)$ is dense in $\widehat{E}$, and the following extension property holds: for every complete metric space F and for every uniformly continuous function $f\colon E \to F$, there is a unique uniformly continuous function $\widehat{f}\colon \widehat{E} \to F$ such that*

$$f = \widehat{f} \circ \varphi,$$

as illustrated in the following diagram.



As a consequence, for any two completions $(\widehat{E}_1, \widehat{d}_1)$ and $(\widehat{E}_2, \widehat{d}_2)$ of (E, d), there is a unique bijective isometry betwen $(\widehat{E}_1, \widehat{d}_1)$ and $(\widehat{E}_2, \widehat{d}_2)$.

Proof. Consider the set $\mathcal{E}$ of all Cauchy sequences (x_n) in E, and define the relation $\sim$ on $\mathcal{E}$ as follows:

$$(x_n) \sim (y_n) \quad \text{iff} \quad \lim_{n \mapsto \infty} d(x_n, y_n) = 0.$$

It is easy to check that $\sim$ is an equivalence relation on $\mathcal{E}$, and let $\widehat{E} = \mathcal{E}/\sim$ be the quotient set, that is, the set of equivalence classes modulo $\sim$. Our goal is to show that we can endow

$\widehat{E}$ with a distance that makes it into a complete metric space satisfying the conditions of the theorem. We proceed in several steps.

Step 1. First, let us construct the function $\varphi\colon E \to \widehat{E}$. For every $a \in E$, we have the constant sequence (a_n) such that $a_n = a$ for all $n \geq 0$, which is obviously a Cauchy sequence. Let $\varphi(a) \in \widehat{E}$ be the equivalence class $[(a_n)]$ of the constant sequence (a_n) with $a_n = a$ for all n. By definition of $\sim$, the equivalence class $\varphi(a)$ is also the equivalence class of all sequences converging to a. The map $a \mapsto \varphi(a)$ is injective because a metric space is Hausdorff, so if $a \neq b$, then a sequence converging to a does not converge to b. After having defined a distance on $\widehat{E}$, we will check that φ is an isometry.

Step 2. Let us now define a distance on $\widehat{E}$. Let $\alpha = [(a_n)]$ and $\beta = [(b_n)]$ be two equivalence classes of Cauchy sequences in E. The triangle inequality implies that

$$d(a_m, b_m) \leq d(a_m, a_n) + d(a_n, b_n) + d(b_n, b_m) = d(a_n, b_n) + d(a_m, a_n) + d(b_m, b_n)$$

and

$$d(a_n, b_n) \leq d(a_n, a_m) + d(a_m, b_m) + d(b_m, b_n) = d(a_m, b_m) + d(a_m, a_n) + d(b_m, b_n),$$

which implies that

$$|d(a_m, b_m) - d(a_n, b_n)| \leq d(a_m, a_n) + d(b_m, b_n).$$

Since (a_n) and (b_n) are Cauchy sequences, it follows that $(d(a_n, b_n))$ is a Cauchy sequence of nonnegative reals. Since $\mathbb{R}$ is complete, the sequence $(d(a_n, b_n))$ has a limit, which we denote by $\widehat{d}(\alpha, \beta)$; that is, we set

$$\widehat{d}(\alpha, \beta) = \lim_{n \mapsto \infty} d(a_n, b_n), \quad \alpha = [(a_n)],\ \beta = [(b_n)].$$

Step 3. Let us check that $\widehat{d}(\alpha, \beta)$ does not depend on the Cauchy sequences (a_n) and (b_n) chosen in the equivalence classes α and β.

If $(a_n) \sim (a'_n)$ and $(b_n) \sim (b'_n)$, then $\lim_{n \mapsto \infty} d(a_n, a'_n) = 0$ and $\lim_{n \mapsto \infty} d(b_n, b'_n) = 0$, and since

$$d(a'_n, b'_n) \leq d(a'_n, a_n) + d(a_n, b_n) + d(b_n, b'_n) = d(a_n, b_n) + d(a_n, a'_n) + d(b_n, b'_n)$$

and

$$d(a_n, b_n) \leq d(a_n, a'_n) + d(a'_n, b'_n) + d(b'_n, b_n) = d(a'_n, b'_n) + d(a_n, a'_n) + d(b_n, b'_n)$$

we have

$$|d(a_n, b_n) - d(a'_n, b'_n)| \leq d(a_n, a'_n) + d(b_n, b'_n),$$

so we have $\lim_{n \mapsto \infty} d(a'_n, b'_n) = \lim_{n \mapsto \infty} d(a_n, b_n) = \widehat{d}(\alpha, \beta)$. Therefore, $\widehat{d}(\alpha, \beta)$ is indeed well defined.

Step 4. Let us check that φ is indeed an isometry.

Given any two elements $\varphi(a)$ and $\varphi(b)$ in $\widehat{E}$, since they are the equivalence classes of the constant sequences (a_n) and (b_n) such that $a_n = a$ and $b_n = b$ for all n, the constant sequence $(d(a_n, b_n))$ with $d(a_n, b_n) = d(a, b)$ for all n converges to $d(a, b)$, so by definition $\widehat{d}(\varphi(a), \varphi(b)) = \lim_{n\mapsto\infty} d(a_n, b_n) = d(a, b)$, which shows that φ is an isometry.

Step 5. Let us verify that $\widehat{d}$ is a metric on $\widehat{E}$. By definition it is obvious that $\widehat{d}(\alpha, \beta) = \widehat{d}(\beta, \alpha)$. If α and β are two distinct equivalence classes, then for any Cauchy sequence (a_n) in the equivalence class α and for any Cauchy sequence (b_n) in the equivalence class β, the sequences (a_n) and (b_n) are inequivalent, which means that $\lim_{n\mapsto\infty} d(a_n, b_n) \neq 0$, that is, $\widehat{d}(\alpha, \beta) \neq 0$. Obviously, $\widehat{d}(\alpha, \alpha) = 0$.

For any equivalence classes $\alpha = [(a_n)]$, $\beta = [(b_n)]$, and $\gamma = [(c_n)]$, we have the triangle inequality

$$d(a_n, c_n) \leq d(a_n, b_n) + d(b_n, c_n),$$

so by continuity of the distance function, by passing to the limit, we obtain

$$\widehat{d}(\alpha, \gamma) \leq \widehat{d}(\alpha, \beta) + \widehat{d}(\beta, \gamma),$$

which is the triangle inequality for $\widehat{d}$. Therefore, $\widehat{d}$ is a distance on $\widehat{E}$.

Step 6. Let us prove that $\varphi(E)$ is dense in $\widehat{E}$. For any $\alpha = [(a_n)]$, let (x_n) be the constant sequence such that $x_k = a_n$ for all $k \geq 0$, so that $\varphi(a_n) = [(x_n)]$. Then we have

$$\widehat{d}(\alpha, \varphi(a_n)) = \lim_{m\mapsto\infty} d(a_m, a_n) \leq \sup_{p,q\geq n} d(a_p, a_q).$$

Since (a_n) is a Cauchy sequence, $\sup_{p,q\geq n} d(a_p, a_q)$ tends to 0 as n goes to infinity, so

$$\lim_{n\mapsto\infty} d(\alpha, \varphi(a_n)) = 0,$$

which means that the sequence $(\varphi(a_n))$ converge to α, and $\varphi(E)$ is indeed dense in $\widehat{E}$.

Step 7. Finally, let us prove that the metric space $\widehat{E}$ is complete.

Let (α_n) be a Cauchy sequence in $\widehat{E}$. Since $\varphi(E)$ is dense in $\widehat{E}$, for every $n > 0$, there some $a_n \in E$ such that

$$\widehat{d}(\alpha_n, \varphi(a_n)) \leq \frac{1}{n}.$$

Since

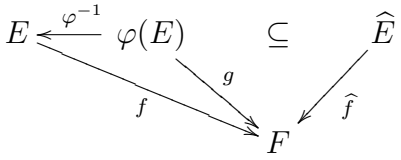
$$\widehat{d}(\varphi(a_m), \varphi(a_n)) \leq \widehat{d}(\varphi(a_m), \alpha_m) + \widehat{d}(\alpha_m, \alpha_n) + \widehat{d}(\alpha_n, \varphi(a_n)) \leq \widehat{d}(\alpha_m, \alpha_n) + \frac{1}{m} + \frac{1}{n},$$

and since (α_m) is a Cauchy sequence, so is $(\varphi(a_n))$, and as φ is an isometry, the sequence (a_n) is a Cauchy sequence in E. Let $\alpha \in \widehat{E}$ be the equivalence class of (a_n). Since

$$\widehat{d}(\alpha, \varphi(a_n)) = \lim_{m\mapsto\infty} d(a_m, a_n)$$

and (a_n) is a Cauchy sequence, we deduce that the sequence $(\varphi(a_n))$ converges to α, and since $d(\alpha_n, \varphi(a_n)) \leq 1/n$ for all $n > 0$, the sequence (α_n) also converges to α.

Step 8. Let us prove the extension property. Let F be any complete metric space and let $f\colon E \to F$ be any uniformly continuous function. The function $\varphi\colon E \to \widehat{E}$ is an isometry and a bijection between E and its image $\varphi(E)$, so its inverse $\varphi^{-1}\colon \varphi(E) \to E$ is also an isometry, and thus is uniformly continuous. If we let $g = f \circ \varphi^{-1}$, then $g\colon \varphi(E) \to F$ is a uniformly continuous function, and $\varphi(E)$ is dense in $\widehat{E}$, so by Theorem 18.20 there is a unique uniformly continuous function $\widehat{f}\colon \widehat{E} \to F$ extending $g = f \circ \varphi^{-1}$; see the diagram below:



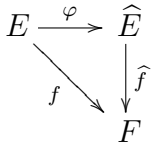
This means that

$$\widehat{f}|\varphi(E) = f \circ \varphi^{-1},$$

which implies that

$$(\widehat{f}|\varphi(E)) \circ \varphi = f,$$

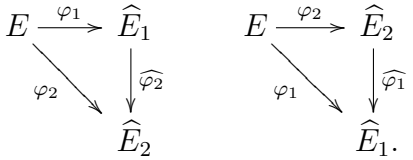
that is, $f = \widehat{f} \circ \varphi$, as illustrated in the diagram below:



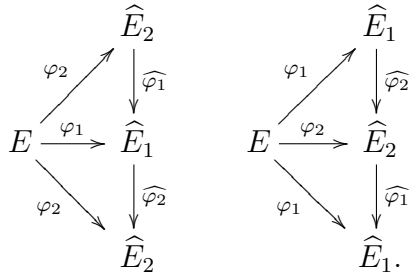
If $h\colon \widehat{E} \to F$ is any other uniformly continuous function such that $f = h \circ \varphi$, then $g = f \circ \varphi^{-1} = h|\varphi(E)$, so h is a uniformly continuous function extending g, and by Theorem 18.20, we have have $h = \widehat{f}$, so $\widehat{f}$ is indeed unique.

Step 9. Uniqueness of the completion $(\widehat{E}, \widehat{d})$ up to a bijective isometry.

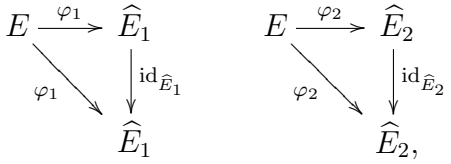
Let $(\widehat{E}_1, \widehat{d}_1)$ and $(\widehat{E}_2, \widehat{d}_2)$ be any two completions of (E, d). Then we have two uniformly continuous isometries $\varphi_1\colon E \to \widehat{E}_1$ and $\varphi_2\colon E \to \widehat{E}_2$, so by the unique extension property, there exist unique uniformly continuous maps $\widehat{\varphi_2}\colon \widehat{E}_1 \to \widehat{E}_2$ and $\widehat{\varphi_1}\colon \widehat{E}_2 \to \widehat{E}_1$ such that the following diagrams commute:



Consequently we have the following commutative diagrams:



However, $\mathrm{id}_{\widehat{E}_1}$ and $\mathrm{id}_{\widehat{E}_2}$ are uniformly continuous functions making the following diagrams commute



so by the uniqueness of extensions we must have

$$\widehat{\varphi_1} \circ \widehat{\varphi_2} = \mathrm{id}_{\widehat{E}_1} \quad \text{and} \quad \widehat{\varphi_2} \circ \widehat{\varphi_1} = \mathrm{id}_{\widehat{E}_2}.$$

This proves that $\widehat{\varphi_1}$ and $\widehat{\varphi_2}$ are mutual inverses. Now, since $\varphi_2 = \widehat{\varphi_2} \circ \varphi_1$, we have

$$\widehat{\varphi_2}|\varphi_1(E) = \varphi_2 \circ \varphi_1^{-1},$$

and since φ_1^{-1} and φ_2 are isometries, so is $\widehat{\varphi_2}|\varphi_1(E)$. But we saw earlier that $\widehat{\varphi_2}$ is the uniform continuous extension of $\widehat{\varphi_2}|\varphi_1(E)$ and $\varphi_1(E)$ is dense in $\widehat{E}_1$, so for any two elements $\alpha, \beta \in \widehat{E}_1$, if (a_n) and (b_n) are sequences in $\varphi_1(E)$ converging to α and β, we have

$$\widehat{d}_2((\widehat{\varphi_2}|\varphi_1(E))(a_n), ((\widehat{\varphi_2}|\varphi_1(E))(b_n)) = \widehat{d}_1(a_n, b_n),$$

and by passing to the limit we get

$$\widehat{d}_2(\widehat{\varphi_2}(\alpha), \widehat{\varphi_2}(\beta)) = \widehat{d}_1(\alpha, \beta),$$

which shows that $\widehat{\varphi_2}$ is an isometry (similarly, $\widehat{\varphi_1}$ is an isometry). □

Remarks:

1. Except for Step 8 and Step 9, the proof of Theorem 18.21 is the proof given in Schwartz [89] (Chapter XI, Section 4, Theorem 1), and Kormogorov and Fomin [60] (Chapter 2, Section 7, Theorem 4).

2. The construction of $\widehat{E}$ relies on the completeness of $\mathbb{R}$, and so it cannot be used to construct $\mathbb{R}$ from $\mathbb{Q}$. However, this construction can be modified to yield a construction of $\mathbb{R}$ from $\mathbb{Q}$.

We show in Section 18.7 that Theorem 18.21 yields a construction of the completion of a normed vector space.

18.7 Completion of a Normed Vector Space

An easy corollary of Theorem 18.21 and Theorem 18.20 is that every normed vector space can be embedded in a complete normed vector space, that is, a Banach space.

Theorem 18.22. *If $(E, \| \|)$ is a normed vector space, then its completion $(\widehat{E}, \widehat{d})$ as a metric space (where E is given the metric $d(x, y) = \|x - y\|$) can be given a unique vector space structure extending the vector space structure on E, and a norm $\| \|_{\widehat{E}}$, so that $(\widehat{E}, \| \|_{\widehat{E}})$ is a Banach space, and the metric $\widehat{d}$ is associated with the norm $\| \|_{\widehat{E}}$. Furthermore, the isometry $\varphi\colon E \to \widehat{E}$ is a linear isometry.*

Proof. The addition operation $+\colon E \times E \to E$ is uniformly continuous because

$$\|(u' + v') - (u'' + v'')\| \leq \|u' - u''\| + \|v' - v''\|.$$

It is not hard to show that $\widehat{E} \times \widehat{E}$ is a complete metric space and that $E \times E$ is dense in $\widehat{E} \times \widehat{E}$. Then, by Theorem 18.20, the uniformly continuous function $+$ has a unique continuous extension $+\colon \widehat{E} \times \widehat{E} \to \widehat{E}$.

The map $\cdot\colon \mathbb{R} \times E \to E$ is not uniformly continuous, but for any fixed $\lambda \in \mathbb{R}$, the map $L_\lambda\colon E \to E$ given by $L_\lambda(u) = \lambda \cdot u$ is uniformly continuous, so by Theorem 18.20 the function L_λ has a unique continuous extension $L_\lambda\colon \widehat{E} \to \widehat{E}$, which we use to define the scalar multiplication $\cdot\colon \mathbb{R} \times \widehat{E} \to \widehat{E}$. It is easily checked that with the above addition and scalar multiplication, $\widehat{E}$ is a vector space.

Since the norm $\| \|$ on E is uniformly continuous, it has a unique continuous extension $\| \|_{\widehat{E}}\colon \widehat{E} \to \mathbb{R}_+$. The identities $\|u + v\| \leq \|u\| + \|v\|$ and $\|\lambda u\| \leq |\lambda| \, \|u\|$ extend to $\widehat{E}$ by continuity. The equation

$$d(u, v) = \|u - v\|$$

also extends to $\widehat{E}$ by continuity and yields

$$\widehat{d}(\alpha, \beta) = \|\alpha - \beta\|_{\widehat{E}},$$

which shows that $\| \|_{\widehat{E}}$ is indeed a norm, and that the metric $\widehat{d}$ is associated to it. Finally, it is easy to verify that the map φ is linear. The uniqueness of the structure of normed vector space follows from the uniqueness of continuous extensions in Theorem 18.20. □

Theorem 18.22 and Theorem 18.20 will be used to show that every Hermitian space can be embedded in a Hilbert space.

We refer the readers to the references cited at the end of this chapter for a discussion of the concepts of compactness and connecteness. They are important, but of less immediate concern.

18.8 The Contraction Mapping Theorem

If (E, d) is a nonempty complete metric space, every map, $f\colon E \to E$, for which there is some k such that $0 \leq k < 1$ and

$$d(f(x), f(y)) \leq kd(x, y) \quad \text{for all } x, y \in E$$

has the very important property that it has a unique fixed point, that is, there is a unique, $a \in E$, such that $f(a) = a$.

Definition 18.19. Let (E, d) be a metric space. A map $f\colon E \to E$ is a *contraction* (or a *contraction mapping*) if there is some real number k such that $0 \leq k < 1$ and

$$d(f(u), f(v)) \leq kd(u, v) \quad \text{for all } u, v \in E.$$

The number k is often called a *Lipschitz constant*.

Furthermore, the fixed point of a contraction mapping can be computed as the limit of a fast converging sequence.

The fixed point property of contraction mappings is used to show some important theorems of analysis, such as the implicit function theorem and the existence of solutions to certain differential equations. It can also be used to show the existence of fractal sets defined in terms of iterated function systems. Since the proof is quite simple, we prove the fixed point property of contraction mappings. First, observe that a contraction mapping is (uniformly) continuous.

Theorem 18.23. *(Contraction Mapping Theorem) If (E, d) is a nonempty complete metric space, every contraction mapping, $f\colon E \to E$, has a unique fixed point. Furthermore, for every $x_0 \in E$, if we define the sequence $(x_n)_{\geq 0}$ such that $x_{n+1} = f(x_n)$ for all $n \geq 0$, then $(x_n)_{n\geq 0}$ converges to the unique fixed point of f.*

Proof. First we prove that f has at most one fixed point. Indeed, if $f(a) = a$ and $f(b) = b$, since

$$d(a, b) = d(f(a), f(b)) \leq kd(a, b)$$

and $0 \leq k < 1$, we must have $d(a, b) = 0$, that is, $a = b$.

Next we prove that (x_n) is a Cauchy sequence. Observe that

$$\begin{aligned} d(x_2, x_1) &\leq kd(x_1, x_0), \\ d(x_3, x_2) &\leq kd(x_2, x_1) \leq k^2 d(x_1, x_0), \\ \vdots &\qquad\quad \vdots \\ d(x_{n+1}, x_n) &\leq kd(x_n, x_{n-1}) \leq \cdots \leq k^n d(x_1, x_0). \end{aligned}$$

Thus, we have

$$\begin{aligned} d(x_{n+p}, x_n) &\leq d(x_{n+p}, x_{n+p-1}) + d(x_{n+p-1}, x_{n+p-2}) + \cdots + d(x_{n+1}, x_n) \\ &\leq (k^{p-1} + k^{p-2} + \cdots + k + 1)k^n d(x_1, x_0) \\ &\leq \frac{k^n}{1-k}\, d(x_1, x_0). \end{aligned}$$

We conclude that $d(x_{n+p}, x_n)$ converges to 0 when n goes to infinity, which shows that (x_n) is a Cauchy sequence. Since E is complete, the sequence (x_n) has a limit, a. Since f is continuous, the sequence $(f(x_n))$ converges to $f(a)$. But $x_{n+1} = f(x_n)$ converges to a and so $f(a) = a$, the unique fixed point of f. □

The above theorem is also called the *Banach fixed point theorem*. Note that no matter how the starting point x_0 of the sequence (x_n) is chosen, (x_n) converges to the unique fixed point of f. Also, the convergence is fast, since

$$d(x_n, a) \leq \frac{k^n}{1-k}\, d(x_1, x_0).$$

18.9 Futher Readings

A thorough treatment of general topology can be found in Munkres [77, 76], Dixmier [35], Lang [65], Schwartz [90, 89], Bredon [23], and the classic, Seifert and Threlfall [94].

18.10 Summary

The main concepts and results of this chapter are listed below:

- *Metric space*, *distance*, *metric*.
- *Euclidean metric*, *discrete metric*.
- *Closed ball*, *open ball*, *sphere*, *bounded subset*.
- *Normed vector space*, *norm*.
- *Open* and *closed* sets.
- *Topology*, *topological space*.
- *Hausdorff separation axiom*, *Hausdorff space*.
- *Discrete topology*.
- *Closure*, *dense subset*, *interior*, *frontier* or *boundary*.

- *Subspace topology*.
- *Product topology*.
- *Basis of a topology*, *subbasis of a topology*.
- *Continuous* functions.
- *Neighborhood* of a point.
- *Homeomorphisms*.
- *Limits* of sequences.
- *Continuous linear maps*.
- The *norm* of a continuous linear map.
- *Continuous bilinear maps*.
- The *norm* of a continuous bilinear map.
- The isomorphism between $\mathcal{L}(E, F; G)$ and $\mathcal{L}(E, \mathcal{L}(F; G))$.
- *Cauchy sequences*
- *Complete metric spaces* and *Banach spaces*.
- *Completion* of a metric space or of a normed vector space.
- *Contractions*.
- *The contraction mapping theorem*.

Chapter 19

Differential Calculus

19.1 Directional Derivatives, Total Derivatives

This chapter contains a review of basic notions of differential calculus. First, we review the definition of the derivative of a function $f\colon \mathbb{R} \to \mathbb{R}$. Next, we define directional derivatives and the total derivative of a function $f\colon E \to F$ between normed vector spaces. Basic properties of derivatives are shown, including the chain rule. We show how derivatives are represented by Jacobian matrices. The mean value theorem is stated, as well as the implicit function theorem and the inverse function theorem. Diffeomorphisms and local diffeomorphisms are defined. Higher-order derivatives are defined, as well as the Hessian. Schwarz's lemma (about the commutativity of partials) is stated. Several versions of Taylor's formula are stated, and a famous formula due to Faà di Bruno's is given.

We first review the notion of the derivative of a real-valued function whose domain is an open subset of $\mathbb{R}$.

Let $f\colon A \to \mathbb{R}$, where A is a nonempty open subset of $\mathbb{R}$, and consider any $a \in A$. The main idea behind the concept of the derivative of f at a, denoted by $f'(a)$, is that locally around a (that is, in some small open set $U \subseteq A$ containing a), the function f is approximated linearly[1] by the map

$$x \mapsto f(a) + f'(a)(x - a).$$

As pointed out by Dieudonné in the early 1960s, it is an "unfortunate accident" that if V is vector space of dimension one, then there is a bijection between the space V^* of linear forms defined on V and the field of scalars. As a consequence, the derivative of a real-valued function f defined on an open subset A of the reals can be defined as the scalar $f'(a)$ (for any $a \in A$). But as soon as f is a function of several arguments, the scalar interpretation of the derivative breaks down.

[1] Actually, the approximation is affine, but everybody commits this abuse of language.

Part of the difficulty in extending the idea of derivative to more complex spaces is to give an adequate notion of linear approximation. The key idea is to use linear maps. This could be carried out in terms of matrices but it turns out that this neither shortens nor simplifies proofs. In fact, this is often the opposite.

We admit that the more intrinsic definition of the notion of derivative f'_a at a point a of a function $f: E \to F$ between two normed vector spaces E and F as a linear map requires a greater effort to be grasped, but we feel that the advantages of this definition outweight its degree of abstraction. In particular, it yields a clear notion of the derivative of a function $f: \mathrm{M}_m(\mathbb{R}) \to \mathrm{M}_n(\mathbb{R})$ defined from $m \times m$ matrices to $n \times n$ matrices (many definitions make use of partial derivatives with respect to matrices that do make any sense). But more importantly, the definition of the derivative as a linear map makes it clear that whether the space E or the space F is infinite dimensional does not matter. This is important in optimization theory where the natural space of solutions of the problem is often an infinite dimensional function space. Of course, to carry out computations one need to pick finite bases and to use Jacobian matrices, but this is a different matter.

Let us now review the formal definition of the derivative of a real-valued function.

Definition 19.1. Let A be any nonempty open subset of $\mathbb{R}$, and let $a \in A$. For any function $f: A \to \mathbb{R}$, the *derivative of f at $a \in A$* is the limit (if it exists)

$$\lim_{h \to 0,\, h \in U} \frac{f(a+h) - f(a)}{h},$$

where $U = \{h \in \mathbb{R} \mid a + h \in A,\, h \neq 0\}$. This limit is denoted by $f'(a)$, or $\mathrm{D}f(a)$, or $\frac{df}{dx}(a)$. If $f'(a)$ exists for every $a \in A$, we say that f *is differentiable on A*. In this case, the map $a \mapsto f'(a)$ is denoted by f', or $\mathrm{D}f$, or $\frac{df}{dx}$.

Note that since A is assumed to be open, $A - \{a\}$ is also open, and since the function $h \mapsto a + h$ is continuous and U is the inverse image of $A - \{a\}$ under this function, U is indeed open and the definition makes sense.

We can also define $f'(a)$ as follows: there is some function ϵ, such that,

$$f(a+h) = f(a) + f'(a) \cdot h + \epsilon(h)h,$$

whenever $a + h \in A$, where $\epsilon(h)$ is defined for all h such that $a + h \in A$, and

$$\lim_{h \to 0,\, h \in U} \epsilon(h) = 0.$$

Remark: We can also define the notion of *derivative of f at a on the left*, and *derivative of f at a on the right*. For example, we say that the *derivative of f at a on the left* is the limit $f'(a_-)$ (if it exists)

$$f'(a_-) = \lim_{h \to 0,\, h \in U} \frac{f(a+h) - f(a)}{h},$$

where $U = \{h \in \mathbb{R} \mid a + h \in A,\ h < 0\}$.

If a function f as in Definition 19.1 has a derivative $f'(a)$ at a, then it is continuous at a. If f is differentiable on A, then f is continuous on A. The composition of differentiable functions is differentiable.

Remark: A function f has a derivative $f'(a)$ at a iff the derivative of f on the left at a and the derivative of f on the right at a exist, and if they are equal. Also, if the derivative of f on the left at a exists, then f is continuous on the left at a (and similarly on the right).

We would like to extend the notion of derivative to functions $f\colon A \to F$, where E and F are normed vector spaces, and A is some nonempty open subset of E. The first difficulty is to make sense of the quotient

$$\frac{f(a+h) - f(a)}{h}.$$

Since F is a normed vector space, $f(a + h) - f(a)$ makes sense. But now, how do we define the quotient by a vector? Well, we don't!

A first possibility is to consider the *directional derivative* with respect to a vector $u \neq 0$ in E. We can consider the vector $f(a + tu) - f(a)$, where $t \in \mathbb{R}$. Now,

$$\frac{f(a+tu) - f(a)}{t}$$

makes sense.

The idea is that in E, the points of the form $a+tu$ for t in some small interval $[-\epsilon, +\epsilon]$ in $\mathbb{R}$ form a line segment $[r, s]$ in A containing a, and that the image of this line segment defines a small curve segment on $f(A)$. This curve segment is defined by the map $t \mapsto f(a + tu)$, from $[r, s]$ to F, and the directional derivative $\mathrm{D}_u f(a)$ defines the direction of the tangent line at a to this curve; see Figure 19.1. This leads us to the following definition.

Definition 19.2. Let E and F be two normed vector spaces, let A be a nonempty open subset of E, and let $f\colon A \to F$ be any function. For any $a \in A$, for any $u \neq 0$ in E, the *directional derivative of f at a w.r.t. the vector u*, denoted by $\mathrm{D}_u f(a)$, is the limit (if it exists)

$$\mathrm{D}_u f(a) = \lim_{t \to 0,\, t \in U} \frac{f(a+tu) - f(a)}{t},$$

where $U = \{t \in \mathbb{R} \mid a + tu \in A,\ t \neq 0\}$ (or $U = \{t \in \mathbb{C} \mid a + tu \in A,\ t \neq 0\}$).

Since the map $t \mapsto a + tu$ is continuous, and since $A - \{a\}$ is open, the inverse image U of $A - \{a\}$ under the above map is open, and the definition of the limit in Definition 19.2 makes sense. The directional derivative is sometimes called the *Gâteaux derivative*.

Remark: Since the notion of limit is purely topological, the existence and value of a directional derivative is independent of the choice of norms in E and F, as long as they are equivalent norms.

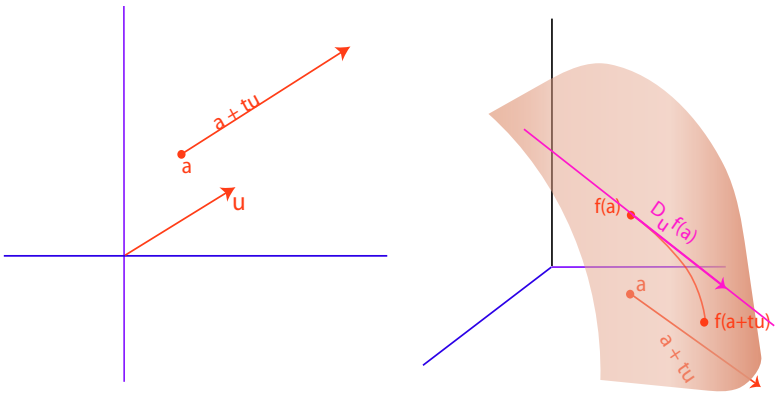


Figure 19.1: Let $f\colon \mathbb{R}^2 \to \mathbb{R}$. The graph of f is the peach surface in $\mathbb{R}^3$, and $t \mapsto f(a+tu)$ is the embedded orange curve connecting $f(a)$ to $f(a+tu)$. Then $\mathrm{D}_u f(a)$ is the slope of the pink tangent line in the direction of u.

In the special case where $E = \mathbb{R}$ and $F = \mathbb{R}$, and we let $u = 1$ (i.e., the real number 1, viewed as a vector), it is immediately verified that $\mathrm{D}_1 f(a) = f'(a)$, in the sense of Definition 19.1. When $E = \mathbb{R}$ (or $E = \mathbb{C}$) and F is any normed vector space, the derivative $\mathrm{D}_1 f(a)$, also denoted by $f'(a)$, provides a suitable generalization of the notion of derivative.

However, when E has dimension ≥ 2, directional derivatives present a serious problem, which is that their definition is not sufficiently uniform. Indeed, there is no reason to believe that the directional derivatives w.r.t. all nonnull vectors u share something in common. As a consequence, a function can have all directional derivatives at a, and yet not be continuous at a. Two functions may have all directional derivatives in some open sets, and yet their composition may not.

Example 19.1. Let $f\colon \mathbb{R}^2 \to \mathbb{R}$ be the function given by

$$f(x,y) = \begin{cases} \frac{x^2 y}{x^4+y^2} & \text{if } (x,y) \neq (0,0) \\ 0 & \text{if } (x,y) = (0,0). \end{cases}$$

For any $u \neq 0$, letting $u = \begin{pmatrix} h \\ k \end{pmatrix}$, we have

$$\frac{f(0+tu) - f(0)}{t} = \frac{h^2 k}{t^2 h^4 + k^2},$$

so that

$$\mathrm{D}_u f(0,0) = \begin{cases} \frac{h^2}{k} & \text{if } k \neq 0 \\ 0 & \text{if } k = 0. \end{cases}$$

Thus, $\mathrm{D}_u f(0,0)$ exists for all $u \neq 0$.

On the other hand, if $\mathrm{D}f(0,0)$ existed, it would be a linear map $\mathrm{D}f(0,0)\colon \mathbb{R}^2 \to \mathbb{R}$ represented by a row matrix $(\alpha\ \beta)$, and we would have $\mathrm{D}_u f(0,0) = \mathrm{D}f(0,0)(u) = \alpha h + \beta k$, but the explicit formula for $\mathrm{D}_u f(0,0)$ is not linear. As a matter of fact, the function f is not continuous at $(0,0)$. For example, on the parabola $y = x^2$, $f(x,y) = \frac{1}{2}$, and when we approach the origin on this parabola, the limit is $\frac{1}{2}$, but $f(0,0) = 0$.

To avoid the problems arising with directional derivatives we introduce a more uniform notion.

Given two normed spaces E and F, recall that a linear map $f\colon E \to F$ is *continuous* iff there is some constant $C \geq 0$ such that

$$\|f(u)\| \leq C\,\|u\| \quad \text{for all } u \in E.$$

Definition 19.3. Let E and F be two normed vector spaces, let A be a nonempty open subset of E, and let $f\colon A \to F$ be any function. For any $a \in A$, we say that f is *differentiable at* $a \in A$ if there is a *linear continuous* map $L\colon E \to F$ and a function $h \mapsto \epsilon(h)$, such that

$$f(a+h) = f(a) + L(h) + \epsilon(h)\|h\|$$

for every $a + h \in A$, where $\epsilon(h)$ is defined for every h such that $a + h \in A$, and

$$\lim_{h \to 0,\, h \in U} \epsilon(h) = 0,$$

where $U = \{h \in E \mid a + h \in A,\, h \neq 0\}$. The linear map L is denoted by $\mathrm{D}f(a)$, or $\mathrm{D}f_a$, or $df(a)$, or df_a, or $f'(a)$, and it is called the *Fréchet derivative*, or *derivative*, or *total derivative*, or *total differential*, or *differential*, of f at a; see Figure 19.2.

Since the map $h \mapsto a + h$ from E to E is continuous, and since A is open in E, the inverse image U of $A - \{a\}$ under the above map is open in E, and it makes sense to say that

$$\lim_{h \to 0,\, h \in U} \epsilon(h) = 0.$$

Note that for every $h \in U$, since $h \neq 0$, $\epsilon(h)$ is uniquely determined since

$$\epsilon(h) = \frac{f(a+h) - f(a) - L(h)}{\|h\|},$$

and that the value $\epsilon(0)$ plays absolutely no role in this definition. The condition for f to be differentiable at a amounts to the fact that

$$\lim_{h \mapsto 0} \frac{\|f(a+h) - f(a) - L(h)\|}{\|h\|} = 0$$

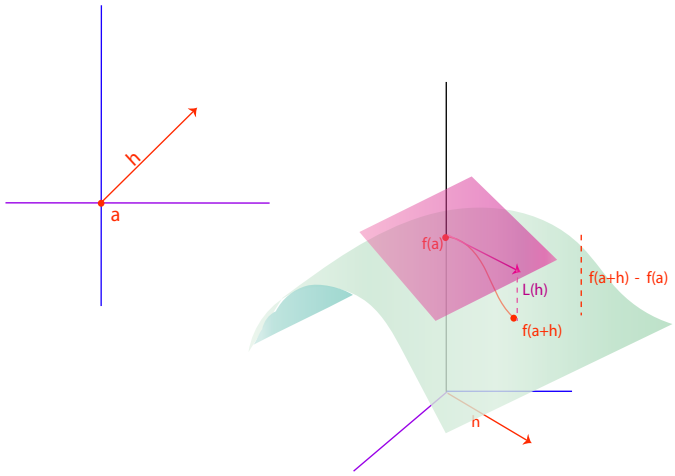


Figure 19.2: Let $f\colon \mathbb{R}^2 \to \mathbb{R}$. The graph of f is the green surface in $\mathbb{R}^3$. The linear map $L = \mathrm{D}f(a)$ is the pink tangent plane. For any vector $h \in \mathbb{R}^2$, $L(h)$ is approximately equal to $f(a+h) - f(a)$. Note that $L(h)$ is also the direction tangent to the curve $t \mapsto f(a+tu)$.

as $h \neq 0$ approaches 0, when $a + h \in A$. However, it does no harm to assume that $\epsilon(0) = 0$, and we will assume this from now on.

Again, we note that the derivative $\mathrm{D}f(a)$ of f at a provides an affine approximation of f, locally around a.

Remarks:

(1) Since the notion of limit is purely topological, the existence and value of a derivative is independent of the choice of norms in E and F, as long as they are equivalent norms.

(2) If $h\colon (-a, a) \to \mathbb{R}$ is a real-valued function defined on some open interval containing 0, we say that h *is* $o(t)$ *for* $t \to 0$, and we write $h(t) = o(t)$, if

$$\lim_{t \mapsto 0,\, t \neq 0} \frac{h(t)}{t} = 0.$$

With this notation (the *little o notation*), the function f is differentiable at a iff

$$f(a+h) - f(a) - L(h) = o(\|h\|),$$

which is also written as

$$f(a+h) = f(a) + L(h) + o(\|h\|).$$

The following proposition shows that our new definition is consistent with the definition of the directional derivative and that *the continuous linear map L is unique*, if it exists.

Proposition 19.1. *Let E and F be two normed spaces, let A be a nonempty open subset of E, and let $f\colon A \to F$ be any function. For any $a \in A$, if $\mathrm{D}f(a)$ is defined, then f is continuous at a and f has a directional derivative $\mathrm{D}_u f(a)$ for every $u \neq 0$ in E. Furthermore,*

$$\mathrm{D}_u f(a) = \mathrm{D}f(a)(u)$$

and thus, $\mathrm{D}f(a)$ is uniquely defined.

Proof. If $L = \mathrm{D}f(a)$ exists, then for any nonzero vector $u \in E$, because A is open, for any $t \in \mathbb{R} - \{0\}$ (or $t \in \mathbb{C} - \{0\}$) small enough, $a + tu \in A$, so

$$\begin{aligned} f(a+tu) &= f(a) + L(tu) + \epsilon(tu)\|tu\| \\ &= f(a) + tL(u) + |t|\epsilon(tu)\|u\| \end{aligned}$$

which implies that

$$L(u) = \frac{f(a+tu) - f(a)}{t} - \frac{|t|}{t}\epsilon(tu)\|u\|,$$

and since $\lim_{t \mapsto 0} \epsilon(tu) = 0$, we deduce that

$$L(u) = \mathrm{D}f(a)(u) = \mathrm{D}_u f(a).$$

Because

$$f(a+h) = f(a) + L(h) + \epsilon(h)\|h\|$$

for all h such that $\|h\|$ is small enough, L is continuous, and $\lim_{h \mapsto 0} \epsilon(h)\|h\| = 0$, we have $\lim_{h \mapsto 0} f(a+h) = f(a)$, that is, f is continuous at a. □

When E is of finite dimension, every linear map is continuous (see Proposition 6.6 or Theorem 18.16), and this assumption is then redundant.

Although this may not be immediately obvious, the reason for requiring the linear map $\mathrm{D}f_a$ to be continuous is to ensure that if a function f is differentiable at a, then it is continuous at a. This is certainly a desirable property of a differentiable function. In finite dimension this holds, but in infinite dimension this is not the case. The following proposition shows that if $\mathrm{D}f_a$ exists at a and if f is continuous at a, then $\mathrm{D}f_a$ must be a continuous map. So if a function is differentiable at a, then it is continuous iff the linear map $\mathrm{D}f_a$ is continuous. We chose to include the second condition rather that the first in the definition of a differentiable function.

Proposition 19.2. *Let E and F be two normed spaces, let A be a nonempty open subset of E, and let $f\colon A \to F$ be any function. For any $a \in A$, if $\mathrm{D}f_a$ is defined, then f is continuous at a iff $\mathrm{D}f_a$ is a continuous linear map.*

Proof. Proposition 19.1 shows that if $\mathrm{D}f_a$ is defined and continuous then f is continuous at a. Conversely, assume that $\mathrm{D}f_a$ exists and that f is continuous at a. Since f is continuous at a and since $\mathrm{D}f_a$ exists, for any $\eta > 0$ there is some ρ with $0 < \rho < 1$ such that if $\|h\| \leq \rho$ then

$$\|f(a+h) - f(a)\| \leq \frac{\eta}{2},$$

and

$$\|f(a+h) - f(a) - \mathrm{D}_a(h)\| \leq \frac{\eta}{2}\|h\| \leq \frac{\eta}{2},$$

so we have

$$\begin{aligned}\|\mathrm{D}_a(h)\| &= \|\mathrm{D}_a(h) - (f(a+h) - f(a)) + f(a+h) - f(a)\| \\ &\leq \|f(a+h) - f(a) - \mathrm{D}_a(h)\| + \|f(a+h) - f(a)\| \\ &\leq \frac{\eta}{2} + \frac{\eta}{2} = \eta,\end{aligned}$$

which proves that $\mathrm{D}f_a$ is continuous at 0. By Proposition 18.14, $\mathrm{D}f_a$ is a continuous linear map. □

As an example, consider the map $f\colon \mathrm{M}_n(\mathbb{R}) \to \mathrm{M}_n(\mathbb{R})$ given by

$$f(A) = A^\top A - I,$$

where $\mathrm{M}_n(\mathbb{R})$ denotes the vector space of all $n \times n$ matrices with real entries equipped with any matrix norm, since they are all equivalent; for example, pick the Frobenius norm $\|A\|_F = \sqrt{\mathrm{tr}(A^\top A)}$. We claim that

$$Df(A)(H) = A^\top H + H^\top A, \quad \text{for all } A \text{ and } H \text{ in } \mathrm{M}_n(\mathbb{R}).$$

We have

$$\begin{aligned}f(A+H) - f(A) - (A^\top H + H^\top A) &= (A+H)^\top(A+H) - I - (A^\top A - I) - A^\top H - H^\top A \\ &= A^\top A + A^\top H + H^\top A + H^\top H - A^\top A - A^\top H - H^\top A \\ &= H^\top H.\end{aligned}$$

It follows that

$$\epsilon(H) = \frac{f(A+H) - f(A) - (A^\top H + H^\top A)}{\|H\|} = \frac{H^\top H}{\|H\|},$$

and since our norm is the Frobenius norm,

$$\|\epsilon(H)\| = \left\|\frac{H^\top H}{\|H\|}\right\| \leq \frac{\left\|H^\top\right\| \|H\|}{\|H\|} = \left\|H^\top\right\| = \|H\|,$$

so

$$\lim_{H \mapsto 0} \epsilon(H) = 0,$$

and we conclude that

$$Df(A)(H) = A^\top H + H^\top A.$$

If $\mathrm{D}f(a)$ exists for every $a \in A$, we get a map $\mathrm{D}f\colon A \to \mathcal{L}(E;F)$, called the *derivative of f on A*, and also denoted by df. Here $\mathcal{L}(E;F)$ denotes the vector space of continuous linear maps from E to F.

We now consider a number of standard results about derivatives. A function $f\colon E \to F$ is said to be *affine* if there is some linear map $\overrightarrow{f} : E \to F$ and some fixed vector $c \in F$, such that

$$f(u) = \overrightarrow{f}(u) + c$$

for all $u \in E$. We call $\overrightarrow{f}$ the *linear map associated with f*.

Proposition 19.3. *Given two normed spaces E and F, if $f\colon E \to F$ is a constant function, then $\mathrm{D}f(a) = 0$, for every $a \in E$. If $f\colon E \to F$ is a continuous affine map, then $\mathrm{D}f(a) = \overrightarrow{f}$, for every $a \in E$, where $\overrightarrow{f}$ denotes the linear map associated with f.*

Proposition 19.4. *Given a normed space E and a normed vector space F, for any two functions $f, g\colon E \to F$, for every $a \in E$, if $\mathrm{D}f(a)$ and $\mathrm{D}g(a)$ exist, then $\mathrm{D}(f+g)(a)$ and $\mathrm{D}(\lambda f)(a)$ exist, and*

$$\mathrm{D}(f+g)(a) = \mathrm{D}f(a) + \mathrm{D}g(a),$$
$$\mathrm{D}(\lambda f)(a) = \lambda \mathrm{D}f(a).$$

Given two normed vector spaces $(E_1, \| \|_1)$ and $(E_2, \| \|_2)$, there are three natural and equivalent norms that can be used to make $E_1 \times E_2$ into a normed vector space:

1. $\|(u_1, u_2)\|_1 = \|u_1\|_1 + \|u_2\|_2$.
2. $\|(u_1, u_2)\|_2 = (\|u_1\|_1^2 + \|u_2\|_2^2)^{1/2}$.
3. $\|(u_1, u_2)\|_\infty = \max(\|u_1\|_1, \|u_2\|_2)$.

We usually pick the first norm. If E_1, E_2, and F are three normed vector spaces, recall that a bilinear map $f\colon E_1 \times E_2 \to F$ is *continuous* iff there is some constant $C \geq 0$ such that

$$\|f(u_1, u_2)\| \leq C \|u_1\|_1 \|u_2\|_2 \quad \text{for all } u_1 \in E_1 \text{ and all } u_2 \in E_2.$$

Proposition 19.5. *Given three normed vector spaces E_1, E_2, and F, for any continuous bilinear map $f\colon E_1 \times E_2 \to F$, for every $(a, b) \in E_1 \times E_2$, $\mathrm{D}f(a, b)$ exists, and for every $u \in E_1$ and $v \in E_2$,*

$$\mathrm{D}f(a, b)(u, v) = f(u, b) + f(a, v).$$

Proof. Since f is bilinear, a simple computation implies that

$$\begin{aligned} f((a,b)+(u,v)) - f(a,b) - (f(u,b)+f(a,v)) &= f(a+u,b+v) - f(a,b) - f(u,b) - f(a,v) \\ &= f(a+u,b) + f(a+u,v) - f(a,b) - f(u,b) - f(a,v) \\ &= f(a,b) + f(u,b) + f(a,v) + f(u,v) - f(a,b) - f(u,b) - f(a,v) \\ &= f(u,v). \end{aligned}$$

We define

$$\epsilon(u,v) = \frac{f((a,b)+(u,v)) - f(a,b) - (f(u,b)+f(a,v))}{\|(u,v)\|_1},$$

and observe that the continuity of f implies

$$\begin{aligned} \|f((a,b)+(u,v)) - f(a,b) - (f(u,b)+f(a,v))\| &= \|f(u,v)\| \\ &\leq C\,\|u\|_1\,\|v\|_2 \leq C\left(\|u\|_1 + \|v\|_2\right)^2. \end{aligned}$$

Hence

$$\|\epsilon(u,v)\| = \left\| \frac{f(u,v)}{\|(u,v)\|_1} \right\| = \frac{\|f(u,v)\|}{\|(u,v)\|_1} \leq \frac{C\left(\|u\|_1 + \|v\|_2\right)^2}{\|u\|_1 + \|v\|_2} = C\left(\|u\|_1 + \|v\|_2\right) = C\,\|(u,v)\|_1,$$

which in turn implies

$$\lim_{(u,v)\mapsto(0,0)} \epsilon(u,v) = 0.$$

□

We now state the very useful *chain rule*.

Theorem 19.6. *Given three normed spaces E, F, and G, let A be an open set in E, and let B an open set in F. For any functions $f\colon A \to F$ and $g\colon B \to G$, such that $f(A) \subseteq B$, for any $a \in A$, if $\mathrm{D}f(a)$ exists and $\mathrm{D}g(f(a))$ exists, then $\mathrm{D}(g \circ f)(a)$ exists, and*

$$\mathrm{D}(g \circ f)(a) = \mathrm{D}g(f(a)) \circ \mathrm{D}f(a).$$

Proof. Since f is differentiable at a and g is differentiable at $b = f(a)$ for every η such that $0 < \eta < 1$ there is some $\rho > 0$ such that for all s, t, if $\|s\| \leq \rho$ and $\|t\| \leq \rho$ then

$$\begin{aligned} f(a+s) &= f(a) + \mathrm{D}f_a(s) + \epsilon_1(s) \\ g(b+t) &= g(b) + \mathrm{D}g_b(t) + \epsilon_2(t), \end{aligned}$$

with $\|\epsilon_1(s)\| \leq \eta\,\|s\|$ and $\|\epsilon_2(t)\| \leq \eta\,\|t\|$. Since $\mathrm{D}f_a$ and $\mathrm{D}g_b$ are continuous, we have

$$\|\mathrm{D}f_a(s)\| \leq \|\mathrm{D}f_a\|\,\|s\| \quad \text{and} \quad \|\mathrm{D}g_b(t)\| \leq \|\mathrm{D}g_b\|\,\|t\|,$$

which, since $\|\epsilon_1(s)\| \leq \eta \|s\|$ and $\eta < 1$, implies that

$$\|\mathrm{D}f_a(s) + \epsilon_1(s)\| \leq \|\mathrm{D}f_a\| \|s\| + \|\epsilon_1(s)\| \leq \|\mathrm{D}f_a\| \|s\| + \eta \|s\| \leq (\|\mathrm{D}f_a\| + 1) \|s\|.$$

Consequently, if $\|s\| < \rho/(\|\mathrm{D}f_a\| + 1)$, we have

$$\|\epsilon_2(\mathrm{D}f_a(s) + \epsilon_1(s))\| \leq \eta(\|\mathrm{D}f_a\| + 1) \|s\| \tag{$*_1$}$$

and

$$\|\mathrm{D}g_b(\epsilon_1(s))\| \leq \|\mathrm{D}g_b\| \|\epsilon_1(s)\| \leq \eta \|\mathrm{D}g_b\| \|s\|. \tag{$*_2$}$$

Then since $b = f(a)$, using the above we have

$$\begin{aligned}(g \circ f)(a + s) &= g(f(a + s)) = g(b + \mathrm{D}f_a(s) + \epsilon_1(s)) \\ &= g(b) + \mathrm{D}g_b(\mathrm{D}f_a(s) + \epsilon_1(s)) + \epsilon_2(\mathrm{D}f_a(s) + \epsilon_1(s)) \\ &= g(b) + (\mathrm{D}g_b \circ \mathrm{D}f_a)(s) + \mathrm{D}g_b(\epsilon_1(s)) + \epsilon_2(\mathrm{D}f_a(s) + \epsilon_1(s)).\end{aligned}$$

Now by $(*_1)$ and $(*_2)$ we have

$$\begin{aligned}\|\mathrm{D}g_b(\epsilon_1(s)) + \epsilon_2(\mathrm{D}f_a(s) + \epsilon_1(s))\| &\leq \|\mathrm{D}g_b(\epsilon_1(s))\| + \|\epsilon_2(\mathrm{D}f_a(s) + \epsilon_1(s))\| \\ &\leq \eta \|\mathrm{D}g_b\| \|s\| + \eta(\|\mathrm{D}f_a\| + 1) \|s\| \\ &= \eta(\|\mathrm{D}f_a\| + \|\mathrm{D}g_b\| + 1) \|s\|,\end{aligned}$$

so if we write $\epsilon_3(s) = \mathrm{D}g_b(\epsilon_1(s)) + \epsilon_2(\mathrm{D}f_a(s) + \epsilon_1(s))$ we proved that

$$(g \circ f)(a + s) = g(b) + (\mathrm{D}g_b \circ \mathrm{D}f_a)(s) + \epsilon_3(s)$$

with $\epsilon_3(s) \leq \eta(\|\mathrm{D}f_a\| + \|\mathrm{D}g_b\| + 1) \|s\|$, which proves that $\mathrm{D}g_b \circ \mathrm{D}f_a$ is the derivative of $g \circ f$ at a. Since $\mathrm{D}f_a$ and $\mathrm{D}g_b$ are continuous, so is $\mathrm{D}g_b \circ \mathrm{D}f_a$, which proves our proposition. □

Theorem 19.6 has many interesting consequences. We mention two corollaries.

Proposition 19.7. *Given three normed vector spaces E, F, and G, for any open subset A in E, for any $a \in A$, let $f\colon A \to F$ such that $\mathrm{D}f(a)$ exists, and let $g\colon F \to G$ be a continuous affine map. Then, $\mathrm{D}(g \circ f)(a)$ exists, and*

$$\mathrm{D}(g \circ f)(a) = \overrightarrow{g} \circ \mathrm{D}f(a),$$

where $\overrightarrow{g}$ is the linear map associated with the affine map g.

Proposition 19.8. *Given two normed vector spaces E and F, let A be some open subset in E, let B be some open subset in F, let $f\colon A \to B$ be a bijection from A to B, and assume that $\mathrm{D}f$ exists on A and that $\mathrm{D}f^{-1}$ exists on B. Then, for every $a \in A$,*

$$\mathrm{D}f^{-1}(f(a)) = (\mathrm{D}f(a))^{-1}.$$

Proposition 19.8 has the remarkable consequence that the two vector spaces E and F have the same dimension. In other words, a local property, the existence of a bijection f between an open set A of E and an open set B of F, such that f is differentiable on A and f^{-1} is differentiable on B, implies a global property, that the two vector spaces E and F have the same dimension.

Let us mention two more rules about derivatives that are used all the time.

Let $\iota\colon \mathbf{GL}(n,\mathbb{C}) \to \mathrm{M}_n(\mathbb{C})$ be the function (inversion) defined on invertible $n\times n$ matrices by

$$\iota(A) = A^{-1}.$$

Observe that $\mathbf{GL}(n,\mathbb{C})$ is indeed an open subset of the normed vector space $\mathrm{M}_n(\mathbb{C})$ of complex $n \times n$ matrices, since its complement is the closed set of matrices $A \in \mathrm{M}_n(\mathbb{C})$ satisfying $\det(A) = 0$. Then we have

$$d\iota_A(H) = -A^{-1}HA^{-1},$$

for all $A \in \mathbf{GL}(n,\mathbb{C})$ and for all $H \in \mathrm{M}_n(\mathbb{C})$.

To prove the preceding line observe that for H with sufficiently small norm, we have

$$\begin{aligned}\iota(A+H) - \iota(A) + A^{-1}HA^{-1} &= (A+H)^{-1} - A^{-1} + A^{-1}HA^{-1}\\ &= (A+H)^{-1}[I - (A+H)A^{-1} + (A+H)A^{-1}HA^{-1}]\\ &= (A+H)^{-1}[I - I - HA^{-1} + HA^{-1} + HA^{-1}HA^{-1}]\\ &= (A+H)^{-1}HA^{-1}HA^{-1}.\end{aligned}$$

Consequently, we get

$$\epsilon(H) = \frac{\iota(A+H) - \iota(A) + A^{-1}HA^{-1}}{\|H\|} = \frac{(A+H)^{-1}HA^{-1}HA^{-1}}{\|H\|},$$

and since

$$\left\|(A+H)^{-1}HA^{-1}HA^{-1}\right\| \leq \|H\|^2 \left\|A^{-1}\right\|^2 \left\|(A+H)^{-1}\right\|,$$

it is clear that $\lim_{H\mapsto 0} \epsilon(H) = 0$, which proves that

$$d\iota_A(H) = -A^{-1}HA^{-1}.$$

In particular, if $A = I$, then $d\iota_I(H) = -H$.

Next, if $f\colon \mathrm{M}_n(\mathbb{C}) \to \mathrm{M}_n(\mathbb{C})$ and $g\colon \mathrm{M}_n(\mathbb{C}) \to \mathrm{M}_n(\mathbb{C})$ are differentiable matrix functions, then

$$d(fg)_A(B) = df_A(B)g(A) + f(A)dg_A(B),$$

for all $A, B \in \mathbf{M}_n(\mathbb{C})$. This is known as the *product rule*.

When E is of finite dimension n, for any basis, $(u_1,\ldots,u_n)$, of E, we can define the directional derivatives with respect to the vectors in the basis $(u_1,\ldots,u_n)$ (actually, we can also do it for an infinite basis). This way we obtain the definition of partial derivatives, as follows:

Definition 19.4. For any two normed spaces E and F, if E is of finite dimension n, for every basis $(u_1, \ldots, u_n)$ for E, for every $a \in E$, for every function $f\colon E \to F$, the directional derivatives $\mathrm{D}_{u_j} f(a)$ (if they exist) are called the *partial derivatives of f with respect to the basis $(u_1, \ldots, u_n)$*. The partial derivative $\mathrm{D}_{u_j} f(a)$ is also denoted by $\partial_j f(a)$, or $\dfrac{\partial f}{\partial x_j}(a)$.

The notation $\dfrac{\partial f}{\partial x_j}(a)$ for a partial derivative, although customary and going back to Leibniz, is a "logical obscenity." Indeed, the variable x_j really has nothing to do with the formal definition. This is just another of these situations where tradition is just too hard to overthrow!

We now consider the situation where the normed vector space F is a finite direct sum $F = F_1 \oplus \cdots \oplus F_m$.

Proposition 19.9. *Given normed vector spaces E and $F = F_1 \oplus \cdots \oplus F_m$, given any open subset A of E, for any $a \in A$, for any function $f\colon A \to F$, letting $f = (f_1, \ldots, f_m)$, $\mathrm{D}f(a)$ exists iff every $\mathrm{D}f_i(a)$ exists, and*

$$\mathrm{D}f(a) = in_1 \circ \mathrm{D}f_1(a) + \cdots + in_m \circ \mathrm{D}f_m(a).$$

Proof. The proposition is a simple application of Theorem 19.6. □

In the special case where F is a normed vector space of finite dimension m, for any basis $(v_1, \ldots, v_m)$ of F, every vector $x \in F$ can be expressed uniquely as

$$x = x_1 v_1 + \cdots + x_m v_m,$$

where $(x_1, \ldots, x_m) \in K^m$, the coordinates of x in the basis $(v_1, \ldots, v_m)$ (where $K = \mathbb{R}$ or $K = \mathbb{C}$). Thus, letting F_i be the standard normed vector space K with its natural structure, we note that F is isomorphic to the direct sum $F = K \oplus \cdots \oplus K$. Then, every function $f\colon E \to F$ is represented by m functions $(f_1, \ldots, f_m)$, where $f_i\colon E \to K$ (where $K = \mathbb{R}$ or $K = \mathbb{C}$), and

$$f(x) = f_1(x) v_1 + \cdots + f_m(x) v_m,$$

for every $x \in E$. The following proposition is an immediate corollary of Proposition 19.9.

Proposition 19.10. *For any two normed vector spaces E and F, if F is of finite dimension m, for any basis $(v_1, \ldots, v_m)$ of F, a function $f\colon E \to F$ is differentiable at a iff each f_i is differentiable at a, and*

$$\mathrm{D}f(a)(u) = \mathrm{D}f_1(a)(u) v_1 + \cdots + \mathrm{D}f_m(a)(u) v_m,$$

for every $u \in E$.

We now consider the situation where E is a finite direct sum. Given a normed vector space $E = E_1 \oplus \cdots \oplus E_n$ and a normed vector space F, given any open subset A of E, for any $c = (c_1, \ldots, c_n) \in A$, we define the continuous functions $i_j^c \colon E_j \to E$, such that

$$i_j^c(x) = (c_1, \ldots, c_{j-1}, x, c_{j+1}, \ldots, c_n).$$

For any function $f \colon A \to F$, we have functions $f \circ i_j^c \colon E_j \to F$, defined on $(i_j^c)^{-1}(A)$, which contains c_j. If $\mathrm{D}(f \circ i_j^c)(c_j)$ exists, we call it the *partial derivative of f w.r.t. its jth argument, at c*. We also denote this derivative by $\mathrm{D}_j f(c)$. Note that $\mathrm{D}_j f(c) \in \mathcal{L}(E_j; F)$.

This notion is a generalization of the notion defined in Definition 19.4. In fact, when E is of dimension n, and a basis $(u_1, \ldots, u_n)$ has been chosen, we can write $E = E_1 \oplus \cdots \oplus E_n$, for some obvious E_j (as explained just after Proposition 19.9), and then

$$\mathrm{D}_j f(c)(\lambda u_j) = \lambda \partial_j f(c),$$

and the two notions are consistent. We will use freely the notation $\partial_j f(c)$ instead of $\mathrm{D}_j f(c)$.

The notion $\partial_j f(c)$ introduced in Definition 19.4 is really that of the vector derivative, whereas $\mathrm{D}_j f(c)$ is the corresponding linear map. Although perhaps confusing, we identify the two notions. The following proposition holds.

Proposition 19.11. *Given a normed vector space $E = E_1 \oplus \cdots \oplus E_n$, and a normed vector space F, given any open subset A of E, for any function $f \colon A \to F$, for every $c \in A$, if $\mathrm{D}f(c)$ exists, then each $\mathrm{D}_j f(c)$ exists, and*

$$\mathrm{D}f(c)(u_1, \ldots, u_n) = \mathrm{D}_1 f(c)(u_1) + \cdots + \mathrm{D}_n f(c)(u_n),$$

for every $u_i \in E_i$, $1 \leq i \leq n$. The same result holds for the finite product $E_1 \times \cdots \times E_n$.

Proof. If $i_j \colon E_j \to E$ is the linear map given by

$$i_j(x) = (0, \ldots, 0, x, 0, \ldots, 0),$$

then

$$i_j^c(x) = (c_1, \ldots, c_{j-1}, 0, c_{j+1}, \ldots, c_n) + i_j(x),$$

which shows that i_j^c is affine, so $\mathrm{D}i_j^c(x) = i_j$. The proposition is then a simple application of Theorem 19.6. □

19.2 Jacobian Matrices

If both E and F are of finite dimension, for any basis $(u_1, \ldots, u_n)$ of E and any basis $(v_1, \ldots, v_m)$ of F, every function $f \colon E \to F$ is determined by m functions $f_i \colon E \to \mathbb{R}$ (or $f_i \colon E \to \mathbb{C}$), where

$$f(x) = f_1(x)v_1 + \cdots + f_m(x)v_m,$$

for every $x \in E$. From Proposition 19.1, we have

$$\mathrm{D}f(a)(u_j) = \mathrm{D}_{u_j} f(a) = \partial_j f(a),$$

and from Proposition 19.10, we have

$$\mathrm{D}f(a)(u_j) = \mathrm{D}f_1(a)(u_j)v_1 + \cdots + \mathrm{D}f_i(a)(u_j)v_i + \cdots + \mathrm{D}f_m(a)(u_j)v_m,$$

that is,

$$\mathrm{D}f(a)(u_j) = \partial_j f_1(a)v_1 + \cdots + \partial_j f_i(a)v_i + \cdots + \partial_j f_m(a)v_m.$$

Since the j-th column of the $m \times n$-matrix representing $\mathrm{D}f(a)$ w.r.t. the bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_m)$ is equal to the components of the vector $\mathrm{D}f(a)(u_j)$ over the basis $(v_1, \ldots, v_m)$, the linear map $\mathrm{D}f(a)$ is determined by the $m \times n$-matrix $J(f)(a) = (\partial_j f_i(a))$, (or $J(f)(a) = (\partial f_i/\partial x_j)(a)$):

$$J(f)(a) = \begin{pmatrix} \partial_1 f_1(a) & \partial_2 f_1(a) & \ldots & \partial_n f_1(a) \\ \partial_1 f_2(a) & \partial_2 f_2(a) & \ldots & \partial_n f_2(a) \\ \vdots & \vdots & \ddots & \vdots \\ \partial_1 f_m(a) & \partial_2 f_m(a) & \ldots & \partial_n f_m(a) \end{pmatrix}$$

or

$$J(f)(a) = \begin{pmatrix} \dfrac{\partial f_1}{\partial x_1}(a) & \dfrac{\partial f_1}{\partial x_2}(a) & \ldots & \dfrac{\partial f_1}{\partial x_n}(a) \\ \dfrac{\partial f_2}{\partial x_1}(a) & \dfrac{\partial f_2}{\partial x_2}(a) & \ldots & \dfrac{\partial f_2}{\partial x_n}(a) \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial f_m}{\partial x_1}(a) & \dfrac{\partial f_m}{\partial x_2}(a) & \ldots & \dfrac{\partial f_m}{\partial x_n}(a) \end{pmatrix}$$

This matrix is called the *Jacobian matrix* of $\mathrm{D}f$ at a. When $m = n$, the determinant, $\det(J(f)(a))$, of $J(f)(a)$ is called the *Jacobian* of $\mathrm{D}f(a)$. From a previous remark, we know that this determinant in fact only depends on $\mathrm{D}f(a)$, and not on specific bases. However, partial derivatives give a means for computing it.

When $E = \mathbb{R}^n$ and $F = \mathbb{R}^m$, for any function $f\colon \mathbb{R}^n \to \mathbb{R}^m$, it is easy to compute the partial derivatives $(\partial f_i/\partial x_j)(a)$. We simply treat the function $f_i\colon \mathbb{R}^n \to \mathbb{R}$ as a function of its j-th argument, leaving the others fixed, and compute the derivative as in Definition 19.1, that is, the usual derivative.

Example 19.2. For example, consider the function $f\colon \mathbb{R}^2 \to \mathbb{R}^2$, defined such that

$$f(r, \theta) = (r\cos(\theta), r\sin(\theta)).$$

Then, we have

$$J(f)(r, \theta) = \begin{pmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{pmatrix}$$

and the Jacobian (determinant) has value $\det(J(f)(r, \theta)) = r$.

In the case where $E = \mathbb{R}$ (or $E = \mathbb{C}$), for any function $f\colon \mathbb{R} \to F$ (or $f\colon \mathbb{C} \to F$), the Jacobian matrix of $\mathrm{D}f(a)$ is a column vector. In fact, this column vector is just $\mathrm{D}_1 f(a)$. Then, for every $\lambda \in \mathbb{R}$ (or $\lambda \in \mathbb{C}$),

$$\mathrm{D}f(a)(\lambda) = \lambda \mathrm{D}_1 f(a).$$

This case is sufficiently important to warrant a definition.

Definition 19.5. Given a function $f\colon \mathbb{R} \to F$ (or $f\colon \mathbb{C} \to F$), where F is a normed vector space, the vector

$$\mathrm{D}f(a)(1) = \mathrm{D}_1 f(a)$$

is called the *vector derivative or velocity vector (in the real case)* at a. We usually identify $\mathrm{D}f(a)$ with its Jacobian matrix $\mathrm{D}_1 f(a)$, which is the column vector corresponding to $\mathrm{D}_1 f(a)$. By abuse of notation, we also let $\mathrm{D}f(a)$ denote the vector $\mathrm{D}f(a)(1) = \mathrm{D}_1 f(a)$.

When $E = \mathbb{R}$, the physical interpretation is that f defines a (parametric) curve that is the trajectory of some particle moving in $\mathbb{R}^m$ as a function of time, and the vector $\mathrm{D}_1 f(a)$ is the *velocity* of the moving particle $f(t)$ at $t = a$; see Figure 19.3.

It is often useful to consider functions $f\colon [a, b] \to F$ from a closed interval $[a, b] \subseteq \mathbb{R}$ to a normed vector space F, and its derivative $\mathrm{D}f(a)$ on $[a, b]$, even though $[a, b]$ is not open. In this case, as in the case of a real-valued function, we define the right derivative $\mathrm{D}_1 f(a_+)$ at a, and the left derivative $\mathrm{D}_1 f(b_-)$ at b, and we assume their existence.

Example 19.3.

1. When $A = (0, 1)$ and $F = \mathbb{R}^3$, a function $f\colon (0, 1) \to \mathbb{R}^3$ defines a (parametric) curve in $\mathbb{R}^3$. If $f = (f_1, f_2, f_3)$, its Jacobian matrix at $a \in \mathbb{R}$ is
$$J(f)(a) = \begin{pmatrix} \dfrac{\partial f_1}{\partial t}(a) \\ \dfrac{\partial f_2}{\partial t}(a) \\ \dfrac{\partial f_3}{\partial t}(a) \end{pmatrix}.$$
See Figure 19.3.
The velocity vectors $J(f)(a) = \begin{pmatrix} -\sin(t) \\ \cos(t) \\ 1 \end{pmatrix}$ are represented by the blue arrows.

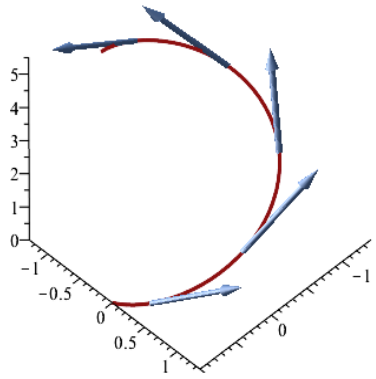


Figure 19.3: The red space curve $f(t) = (\cos(t), \sin(t), t)$.

2. When $E = \mathbb{R}^2$ and $F = \mathbb{R}^3$, a function $\varphi\colon \mathbb{R}^2 \to \mathbb{R}^3$ defines a parametric surface. Letting $\varphi = (f, g, h)$, its Jacobian matrix at $a \in \mathbb{R}^2$ is

$$J(\varphi)(a) = \begin{pmatrix} \dfrac{\partial f}{\partial u}(a) & \dfrac{\partial f}{\partial v}(a) \\ \dfrac{\partial g}{\partial u}(a) & \dfrac{\partial g}{\partial v}(a) \\ \dfrac{\partial h}{\partial u}(a) & \dfrac{\partial h}{\partial v}(a) \end{pmatrix}.$$

   See Figure 19.4. The Jacobian matrix is $J(f)(a) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 2u & 2v \end{pmatrix}$. The first column is the vector tangent to the pink u-direction curve, while the second column is the vector tangent to the blue v-direction curve.

3. When $E = \mathbb{R}^3$ and $F = \mathbb{R}$, for a function $f\colon \mathbb{R}^3 \to \mathbb{R}$, the Jacobian matrix at $a \in \mathbb{R}^3$ is

$$J(f)(a) = \begin{pmatrix} \dfrac{\partial f}{\partial x}(a) & \dfrac{\partial f}{\partial y}(a) & \dfrac{\partial f}{\partial z}(a) \end{pmatrix}.$$

More generally, when $f\colon \mathbb{R}^n \to \mathbb{R}$, the Jacobian matrix at $a \in \mathbb{R}^n$ is the row vector

$$J(f)(a) = \begin{pmatrix} \dfrac{\partial f}{\partial x_1}(a) & \cdots & \dfrac{\partial f}{\partial x_n}(a) \end{pmatrix}.$$

Its transpose is a column vector called the *gradient* of f at a, denoted by $\mathrm{grad} f(a)$ or $\nabla f(a)$. Then, given any $v \in \mathbb{R}^n$, note that

$$\mathrm{D}f(a)(v) = \frac{\partial f}{\partial x_1}(a)\, v_1 + \cdots + \frac{\partial f}{\partial x_n}(a)\, v_n = \mathrm{grad} f(a) \cdot v,$$

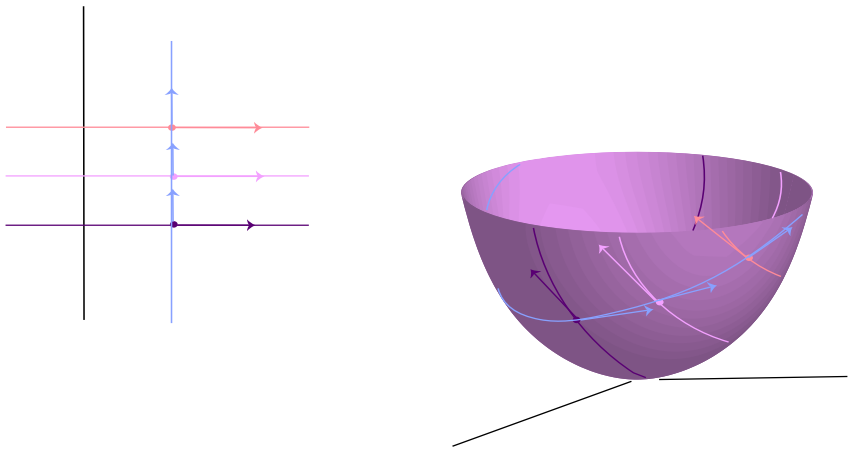


Figure 19.4: The parametric surface $x = u, y = v, z = u^2 + v^2$.

the scalar product of $\mathrm{grad} f(a)$ and v.

Example 19.4. Consider the quadratic function $f\colon \mathbb{R}^n \to \mathbb{R}$ given by

$$f(x) = x^\top A x, \quad x \in \mathbb{R}^n,$$

where A is a real $n \times n$ symmetric matrix. We claim that

$$df_u(h) = 2u^\top A h \quad \text{for all } u, h \in \mathbb{R}^n.$$

Since A is symmetric, we have

$$\begin{aligned} f(u+h) &= (u^\top + h^\top)A(u+h) \\ &= u^\top A u + u^\top A h + h^\top A u + h^\top A h \\ &= u^\top A u + 2u^\top A h + h^\top A h, \end{aligned}$$

so we have

$$f(u+h) - f(u) - 2u^\top A h = h^\top A h.$$

If we write

$$\epsilon(h) = \frac{h^\top A h}{\|h\|}$$

for $h \notin 0$ where $\| \|$ is the 2-norm, by Cauchy–Schwarz we have

$$|\epsilon(h)| \leq \frac{\|h\| \, \|Ah\|}{\|h\|} \leq \frac{\|h\|^2 \, \|A\|}{\|h\|} = \|h\| \, \|A\|,$$

which shows that $\lim_{h \mapsto 0} \epsilon(h) = 0$. Therefore,

$$df_u(h) = 2u^\top Ah \quad \text{for all } u, h \in \mathbb{R}^n,$$

as claimed. This formula shows that the gradient ∇f_u of f at u is given by

$$\nabla f_u = 2Au.$$

As a first corollary we obtain the gradient of a function of the form

$$f(x) = \frac{1}{2} x^\top Ax - b^\top x,$$

where A is a symmetric $n \times n$ matrix and b is some vector $b \in \mathbb{R}^n$. Since the derivative of a linear function is itself, we obtain

$$df_u(h) = u^\top Ah - b^\top h,$$

and the gradient of f is given by

$$\nabla f_u = Au - b.$$

As a second corollary we obtain the gradient of the function

$$f(x) = \|Ax - b\|_2^2 = (Ax - b)^\top (Ax - b) = (x^\top A^\top - b^\top)(Ax - b)$$

which is the function to minimize in a least squares problem, where A is an $m \times n$ matrix. We have

$$f(x) = x^\top A^\top Ax - x^\top A^\top b - b^\top Ax + b^\top b = x^\top A^\top Ax - 2b^\top Ax + b^\top b,$$

and since the derivative of a constant function is 0 and the derivative of a linear function is itself, we get

$$df_u(h) = 2u^\top A^\top Ah - 2b^\top Ah.$$

Consequently, the gradient of f is given by

$$\nabla f_u = 2A^\top Au - 2A^\top b.$$

When E, F, and G have finite dimensions, and $(u_1, \ldots, u_p)$ is a basis for E, $(v_1, \ldots, v_n)$ is a basis for F, and $(w_1, \ldots, w_m)$ is a basis for G, if A is an open subset of E, B is an open subset of F, for any functions $f \colon A \to F$ and $g \colon B \to G$, such that $f(A) \subseteq B$, for any $a \in A$, letting $b = f(a)$, and $h = g \circ f$, if $\mathrm{D}f(a)$ exists and $\mathrm{D}g(b)$ exists, by Theorem 19.6, the Jacobian matrix $J(h)(a) = J(g \circ f)(a)$ w.r.t. the bases $(u_1, \ldots, u_p)$ and $(w_1, \ldots, w_m)$ is

the product of the Jacobian matrices $J(g)(b)$ w.r.t. the bases $(v_1, \ldots, v_n)$ and $(w_1, \ldots, w_m)$, and $J(f)(a)$ w.r.t. the bases $(u_1, \ldots, u_p)$ and $(v_1, \ldots, v_n)$:

$$J(h)(a) = \begin{pmatrix} \partial_1 g_1(b) & \partial_2 g_1(b) & \ldots & \partial_n g_1(b) \\ \partial_1 g_2(b) & \partial_2 g_2(b) & \ldots & \partial_n g_2(b) \\ \vdots & \vdots & \ddots & \vdots \\ \partial_1 g_m(b) & \partial_2 g_m(b) & \ldots & \partial_n g_m(b) \end{pmatrix} \begin{pmatrix} \partial_1 f_1(a) & \partial_2 f_1(a) & \ldots & \partial_p f_1(a) \\ \partial_1 f_2(a) & \partial_2 f_2(a) & \ldots & \partial_p f_2(a) \\ \vdots & \vdots & \ddots & \vdots \\ \partial_1 f_n(a) & \partial_2 f_n(a) & \ldots & \partial_p f_n(a) \end{pmatrix}$$

or

$$J(h)(a) = \begin{pmatrix} \dfrac{\partial g_1}{\partial y_1}(b) & \dfrac{\partial g_1}{\partial y_2}(b) & \ldots & \dfrac{\partial g_1}{\partial y_n}(b) \\ \dfrac{\partial g_2}{\partial y_1}(b) & \dfrac{\partial g_2}{\partial y_2}(b) & \ldots & \dfrac{\partial g_2}{\partial y_n}(b) \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial g_m}{\partial y_1}(b) & \dfrac{\partial g_m}{\partial y_2}(b) & \ldots & \dfrac{\partial g_m}{\partial y_n}(b) \end{pmatrix} \begin{pmatrix} \dfrac{\partial f_1}{\partial x_1}(a) & \dfrac{\partial f_1}{\partial x_2}(a) & \ldots & \dfrac{\partial f_1}{\partial x_p}(a) \\ \dfrac{\partial f_2}{\partial x_1}(a) & \dfrac{\partial f_2}{\partial x_2}(a) & \ldots & \dfrac{\partial f_2}{\partial x_p}(a) \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial f_n}{\partial x_1}(a) & \dfrac{\partial f_n}{\partial x_2}(a) & \ldots & \dfrac{\partial f_n}{\partial x_p}(a) \end{pmatrix}.$$

Thus, we have the familiar formula

$$\frac{\partial h_i}{\partial x_j}(a) = \sum_{k=1}^{k=n} \frac{\partial g_i}{\partial y_k}(b) \frac{\partial f_k}{\partial x_j}(a).$$

Given two normed vector spaces E and F of finite dimension, given an open subset A of E, if a function $f\colon A \to F$ is differentiable at $a \in A$, then its Jacobian matrix is well defined.

One should be warned that the converse is false. There are functions such that all the partial derivatives exist at some $a \in A$, but yet, the function is not differentiable at a, and not even continuous at a. For example, consider the function $f\colon \mathbb{R}^2 \to \mathbb{R}$, defined such that $f(0,0) = 0$, and

$$f(x,y) = \frac{x^2 y}{x^4 + y^2} \qquad \text{if } (x,y) \neq (0,0).$$

For any $u \neq 0$, letting $u = \begin{pmatrix} h \\ k \end{pmatrix}$, we have

$$\frac{f(0+tu) - f(0)}{t} = \frac{h^2 k}{t^2 h^4 + k^2},$$

so that

$$\mathrm{D}_u f(0,0) = \begin{cases} \frac{h^2}{k} & \text{if } k \neq 0 \\ 0 & \text{if } k = 0. \end{cases}$$

Thus, $\mathrm{D}_u f(0,0)$ exists for all $u \neq 0$. On the other hand, if $\mathrm{D}f(0,0)$ existed, it would be a linear map $\mathrm{D}f(0,0)\colon \mathbb{R}^2 \to \mathbb{R}$ represented by a row matrix $(\alpha \ \ \beta)$, and we would have

$\mathrm{D}_u f(0,0) = \mathrm{D}f(0,0)(u) = \alpha h + \beta k$, but the explicit formula for $\mathrm{D}_u f(0,0)$ is not linear. As a matter of fact, the function f is not continuous at $(0,0)$. For example, on the parabola $y = x^2$, $f(x,y) = \frac{1}{2}$, and when we approach the origin on this parabola, the limit is $\frac{1}{2}$, when in fact, $f(0,0) = 0$.

However, there are sufficient conditions on the partial derivatives for $\mathrm{D}f(a)$ to exist, namely, continuity of the partial derivatives.

If f is differentiable on A, then f defines a function $\mathrm{D}f\colon A \to \mathcal{L}(E;F)$. It turns out that the continuity of the partial derivatives on A is a necessary and sufficient condition for $\mathrm{D}f$ to exist and to be continuous on A.

If $f\colon [a,b] \to \mathbb{R}$ is a function which is continuous on $[a,b]$ and differentiable on $]a,b]$, then there is some c with $a < c < b$ such that

$$f(b) - f(a) = (b-a)f'(c).$$

This result is known as the *mean value theorem* and is a generalization of *Rolle's theorem*, which corresponds to the case where $f(a) = f(b)$.

Unfortunately, the mean value theorem fails for vector-valued functions. For example, the function $f\colon [0,2\pi] \to \mathbb{R}^2$ given by

$$f(t) = (\cos t, \sin t)$$

is such that $f(2\pi) - f(0) = (0,0)$, yet its derivative $f'(t) = (-\sin t, \cos t)$ does not vanish in $(0,2\pi)$.

A suitable generalization of the mean value theorem to vector-valued functions is possible if we consider an inequality (an upper bound) instead of an equality. This generalized version of the mean value theorem plays an important role in the proof of several major results of differential calculus.

If E is an vector space (over $\mathbb{R}$ or $\mathbb{C}$), given any two points $a, b \in E$, the *closed segment* $[a,b]$ is the set of all points $a + \lambda(b-a)$, where $0 \leq \lambda \leq 1$, $\lambda \in \mathbb{R}$, and the *open segment* (a,b) is the set of all points $a + \lambda(b-a)$, where $0 < \lambda < 1$, $\lambda \in \mathbb{R}$.

Lemma 19.12. *Let E and F be two normed vector spaces, let A be an open subset of E, and let $f\colon A \to F$ be a continuous function on A. Given any $a \in A$ and any $h \neq 0$ in E, if the closed segment $[a, a+h]$ is contained in A, if $f\colon A \to F$ is differentiable at every point of the open segment $(a, a+h)$, and*

$$\sup_{x \in (a,a+h)} \|\mathrm{D}f(x)\| \leq M,$$

for some $M \geq 0$, then

$$\|f(a+h) - f(a)\| \leq M\|h\|.$$

As a corollary, if $L\colon E \to F$ is a continuous linear map, then

$$\|f(a+h) - f(a) - L(h)\| \leq M\|h\|,$$

where $M = \sup_{x\in(a,a+h)} \|\mathrm{D}f(x) - L\|$.

The above lemma is sometimes called the "mean value theorem." Lemma 19.12 can be used to show the following important result.

Theorem 19.13. *Given two normed vector spaces E and F, where E is of finite dimension n, and where $(u_1, \ldots, u_n)$ is a basis of E, given any open subset A of E, given any function $f\colon A \to F$, the derivative $\mathrm{D}f\colon A \to \mathcal{L}(E;F)$ is defined and continuous on A iff every partial derivative $\partial_j f$ (or $\dfrac{\partial f}{\partial x_j}$) is defined and continuous on A, for all j, $1 \leq j \leq n$. As a corollary, if F is of finite dimension m, and $(v_1, \ldots, v_m)$ is a basis of F, the derivative $\mathrm{D}f\colon A \to \mathcal{L}(E;F)$ is defined and continuous on A iff every partial derivative $\partial_j f_i$ (or $\dfrac{\partial f_i}{\partial x_j}$) is defined and continuous on A, for all i, j, $1 \leq i \leq m$, $1 \leq j \leq n$.*

Theorem 19.13 gives a necessary and sufficient condition for the existence and continuity of the derivative of a function on an open set. It should be noted that a more general version of Theorem 19.13 holds, assuming that $E = E_1 \oplus \cdots \oplus E_n$, or $E = E_1 \times \cdots \times E_n$, and using the more general partial derivatives $\mathrm{D}_j f$ introduced before Proposition 19.11.

Definition 19.6. Given two normed vector spaces E and F, and an open subset A of E, we say that a function $f\colon A \to F$ is *of class C^0 on A or a C^0-function on A* if f is continuous on A. We say that $f\colon A \to F$ is *of class C^1 on A or a C^1-function on A* if $\mathrm{D}f$ exists and is continuous on A.

Since the existence of the derivative on an open set implies continuity, a C^1-function is of course a C^0-function. Theorem 19.13 gives a necessary and sufficient condition for a function f to be a C^1-function (when E is of finite dimension). It is easy to show that the composition of C^1-functions (on appropriate open sets) is a C^1-function.

19.3 The Implicit and The Inverse Function Theorems

Given three normed vector spaces E, F, and G, given a function $f\colon E \times F \to G$, given any $c \in G$, it may happen that the equation

$$f(x, y) = c$$

has the property that, for some open sets $A \subseteq E$, and $B \subseteq F$, there is a function $g\colon A \to B$, such that

$$f(x, g(x)) = c,$$

for all $x \in A$. Such a situation is usually very rare, but if some solution $(a, b) \in E \times F$ such that $f(a, b) = c$ is known, under certain conditions, for some small open sets $A \subseteq E$ containing a and $B \subseteq F$ containing b, the existence of a unique $g\colon A \to B$, such that

$$f(x, g(x)) = c,$$

for all $x \in A$, can be shown. Under certain conditions, it can also be shown that g is continuous, and differentiable. Such a theorem, known as the *implicit function theorem*, can be shown. We state a version of this result below. The proof is fairly involved, and uses a fixed-point theorem for contracting mappings in complete metric spaces; it is given in Schwartz [91].

Theorem 19.14. *Let E, F, and G, be normed vector spaces, let Ω be an open subset of $E \times F$, let $f\colon \Omega \to G$ be a function defined on Ω, let $(a, b) \in \Omega$, let $c \in G$, and assume that $f(a, b) = c$. If the following assumptions hold:*

(1) The function $f\colon \Omega \to G$ is continuous on Ω;

(2) F is a complete normed vector space (and so is G);

(3) $\dfrac{\partial f}{\partial y}(x, y)$ exists for every $(x, y) \in \Omega$, and $\dfrac{\partial f}{\partial y}\colon \Omega \to \mathcal{L}(F; G)$ is continuous;

(4) $\dfrac{\partial f}{\partial y}(a, b)$ is a bijection of $\mathcal{L}(F; G)$, and $\Big(\dfrac{\partial f}{\partial y}(a, b)\Big)^{-1} \in \mathcal{L}(G; F)$;

then the following properties hold:

(a) There exist some open subset $A \subseteq E$ containing a and some open subset $B \subseteq F$ containing b, such that $A \times B \subseteq \Omega$, and for every $x \in A$, the equation $f(x, y) = c$ has a single solution $y = g(x)$, and thus, there is a unique function $g\colon A \to B$ such that $f(x, g(x)) = c$, for all $x \in A$;

(b) The function $g\colon A \to B$ is continuous.

If we also assume that

(5) The derivative $\mathrm{D}f(a, b)$ exists;

then

(c) The derivative $\mathrm{D}g(a)$ exists, and

$$\mathrm{D}g(a) = -\Big(\frac{\partial f}{\partial y}(a, b)\Big)^{-1} \circ \frac{\partial f}{\partial x}(a, b);$$

and if in addition

(6) $\dfrac{\partial f}{\partial x}\colon \Omega \to \mathcal{L}(E;G)$ *is also continuous (and thus, in view of (3),* f *is* C^1 *on* Ω*);*

then

(d) The derivative $\mathrm{D}g\colon A \to \mathcal{L}(E;F)$ *is continuous, and*

$$\mathrm{D}g(x) = -\Big(\frac{\partial f}{\partial y}(x, g(x))\Big)^{-1} \circ \frac{\partial f}{\partial x}(x, g(x)),$$

for all $x \in A$.

The implicit function theorem plays an important role in the calculus of variations. We now consider another very important notion, that of a (local) diffeomorphism.

Definition 19.7. Given two topological spaces E and F, and an open subset A of E, we say that a function $f\colon A \to F$ is a *local homeomorphism from* A *to* F if for every $a \in A$, there is an open set $U \subseteq A$ containing a and an open set V containing $f(a)$ such that f is a homeomorphism from U to $V = f(U)$. If B is an open subset of F, we say that $f\colon A \to F$ is a *(global) homeomorphism from* A *to* B if f is a homeomorphism from A to $B = f(A)$. If E and F are normed vector spaces, we say that $f\colon A \to F$ is a *local diffeomorphism from* A *to* F if for every $a \in A$, there is an open set $U \subseteq A$ containing a and an open set V containing $f(a)$ such that f is a bijection from U to V, f is a C^1-function on U, and f^{-1} is a C^1-function on $V = f(U)$. We say that $f\colon A \to F$ is a *(global) diffeomorphism from* A *to* B if f is a homeomorphism from A to $B = f(A)$, f is a C^1-function on A, and f^{-1} is a C^1-function on B.

Note that a local diffeomorphism is a local homeomorphism. Also, as a consequence of Proposition 19.8, if f is a diffeomorphism on A, then $\mathrm{D}f(a)$ is a bijection for every $a \in A$. The following theorem can be shown. In fact, there is a fairly simple proof using Theorem 19.14.

Theorem 19.15. *(Inverse Function Theorem) Let* E *and* F *be complete normed spaces, let* A *be an open subset of* E*, and let* $f\colon A \to F$ *be a* C^1*-function on* A*. The following properties hold:*

(1) For every $a \in A$*, if* $\mathrm{D}f(a)$ *is a linear isomorphism (which means that both* $\mathrm{D}f(a)$ *and* $(\mathrm{D}f(a))^{-1}$ *are linear and continuous),*[2] *then there exist some open subset* $U \subseteq A$ *containing* a*, and some open subset* V *of* F *containing* $f(a)$*, such that* f *is a diffeomorphism from* U *to* $V = f(U)$*. Furthermore,*

$$\mathrm{D}f^{-1}(f(a)) = (\mathrm{D}f(a))^{-1}.$$

For every neighborhood N *of* a*, the image* $f(N)$ *of* N *is a neighborhood of* $f(a)$*, and for every open ball* $U \subseteq A$ *of center* a*, the image* $f(U)$ *of* U *contains some open ball of center* $f(a)$*.*

[2]Actually, since E and F are Banach spaces, by the Open Mapping Theorem, it is sufficient to assume that $\mathrm{D}f(a)$ is continuous and bijective; see Lang [64].

(2) If $\mathrm{D}f(a)$ *is invertible for every* $a \in A$*, then* $B = f(A)$ *is an open subset of* F*, and* f *is a local diffeomorphism from* A *to* B*. Furthermore, if* f *is injective, then* f *is a diffeomorphism from* A *to* B*.*

Proofs of the Inverse function theorem can be found in Lang [64], Abraham and Marsden [1], Schwartz [91], and Cartan [26]. Part (1) of Theorem 19.15 is often referred to as the "(local) inverse function theorem." It plays an important role in the study of manifolds and (ordinary) differential equations.

If E and F are both of finite dimension, and some bases have been chosen, the invertibility of $\mathrm{D}f(a)$ is equivalent to the fact that the Jacobian determinant $\det(J(f)(a))$ is nonnull. The case where $\mathrm{D}f(a)$ is just injective or just surjective is also important for defining manifolds, using implicit definitions.

Definition 19.8. Let E and F be normed vector spaces, where E and F are of finite dimension (or both E and F are complete), and let A be an open subset of E. For any $a \in A$, a C^1-function $f\colon A \to F$ is an *immersion at* a if $\mathrm{D}f(a)$ is injective. A C^1-function $f\colon A \to F$ is a *submersion at* a if $\mathrm{D}f(a)$ is surjective. A C^1-function $f\colon A \to F$ is an *immersion on* A (resp. *a submersion on* A) if $\mathrm{D}f(a)$ is injective (resp. surjective) for every $a \in A$.

When E and F are finite dimensional with $\dim(E) = n$ and $\dim(F) = m$, if $m \geq n$, then f is an immersion iff the Jacobian matrix, $J(f)(a)$, has full rank n for all $a \in E$ and if $n \geq m$, then f is a submersion iff the Jacobian matrix, $J(f)(a)$, has full rank m for all $a \in E$. For example, $f\colon \mathbb{R} \to \mathbb{R}^2$ defined by $f(t) = (\cos(t), \sin(t))$ is an immersion since $J(f)(t) = \begin{pmatrix} -\sin(t) \\ \cos(t) \end{pmatrix}$ has rank 1 for all t. On the other hand, $f\colon \mathbb{R} \to \mathbb{R}^2$ defined by $f(t) = (t^2, t^2)$ is not an immersion since $J(f)(t) = \begin{pmatrix} 2t \\ 2t \end{pmatrix}$ vanishes at $t = 0$. See Figure 19.5. An example of a submersion is given by the projection map $f\colon \mathbb{R}^2 \to \mathbb{R}$, where $f(x, y) = x$, since $J(f)(x, y) = \begin{pmatrix} 1 & 0 \end{pmatrix}$.

The following results can be shown.

Proposition 19.16. *Let* A *be an open subset of* $\mathbb{R}^n$*, and let* $f\colon A \to \mathbb{R}^m$ *be a function. For every* $a \in A$*,* $f\colon A \to \mathbb{R}^m$ *is a submersion at* a *iff there exists an open subset* U *of* A *containing* a*, an open subset* $W \subseteq \mathbb{R}^{n-m}$*, and a diffeomorphism* $\varphi\colon U \to f(U) \times W$*, such that,*

$$f = \pi_1 \circ \varphi,$$

where $\pi_1\colon f(U) \times W \to f(U)$ *is the first projection. Equivalently,*

$$(f \circ \varphi^{-1})(y_1, \ldots, y_m, \ldots, y_n) = (y_1, \ldots, y_m).$$

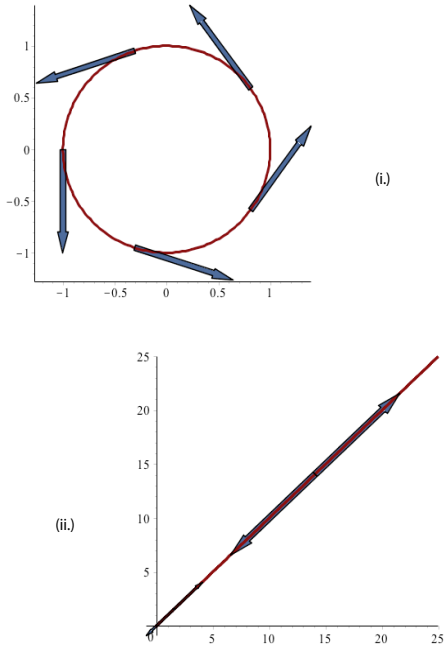


Figure 19.5: Figure $(i.)$ is the immersion of $\mathbb{R}$ into $\mathbb{R}^2$ given by $f(t) = (\cos(t), \sin(t))$. Figure $(ii.)$, the parametric curve $f(t) = (t^2, t^2)$, is not an immersion since the tangent vanishes at the origin.

$$\begin{array}{ccc} U \subseteq A & \xrightarrow{\varphi} & f(U) \times W \\ & \searrow^{f} & \downarrow{\scriptstyle \pi_1} \\ & & f(U) \subseteq \mathbb{R}^m \end{array}$$

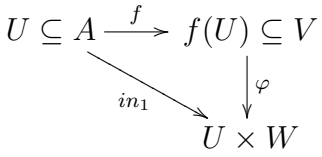
Futhermore, the image of every open subset of A under f is an open subset of F. (The same result holds for $\mathbb{C}^n$ and $\mathbb{C}^m$).

Proposition 19.17. *Let A be an open subset of $\mathbb{R}^n$, and let $f\colon A \to \mathbb{R}^m$ be a function. For every $a \in A$, $f\colon A \to \mathbb{R}^m$ is an immersion at a iff there exists an open subset U of A containing a, an open subset V containing $f(a)$ such that $f(U) \subseteq V$, an open subset W containing 0 such that $W \subseteq \mathbb{R}^{m-n}$, and a diffeomorphism $\varphi\colon V \to U \times W$, such that,*

$$\varphi \circ f = in_1,$$

where $in_1\colon U \to U \times W$ is the injection map such that $in_1(u) = (u, 0)$, or equivalently,

$$(\varphi \circ f)(x_1, \ldots, x_n) = (x_1, \ldots, x_n, 0, \ldots, 0).$$



(The same result holds for $\mathbb{C}^n$ and $\mathbb{C}^m$).

We now briefly consider second-order and higher-order derivatives.

19.4 Second-Order and Higher-Order Derivatives

Given two normed vector spaces E and F, and some open subset A of E, if $\mathrm{D}f(a)$ is defined for every $a \in A$, then we have a mapping $\mathrm{D}f\colon A \to \mathcal{L}(E;F)$. Since $\mathcal{L}(E;F)$ is a normed vector space, if $\mathrm{D}f$ exists on an open subset U of A containing a, we can consider taking the derivative of $\mathrm{D}f$ at some $a \in A$. If $\mathrm{D}(\mathrm{D}f)(a)$ exists for every $a \in A$, we get a mapping $\mathrm{D}^2 f\colon A \to \mathcal{L}(E;\mathcal{L}(E;F))$, where $\mathrm{D}^2 f(a) = \mathrm{D}(\mathrm{D}f)(a)$, for every $a \in A$. If $\mathrm{D}^2 f(a)$ exists, then for every $u \in E$,

$$\mathrm{D}^2 f(a)(u) = \mathrm{D}(\mathrm{D}f)(a)(u) = \mathrm{D}_u(\mathrm{D}f)(a) \in \mathcal{L}(E;F).$$

Recall from Proposition 18.19, that the map app from $\mathcal{L}(E;F) \times E$ to F, defined such that for every $L \in \mathcal{L}(E;F)$, for every $v \in E$,

$$\mathrm{app}(L,v) = L(v),$$

is a continuous bilinear map. Thus, in particular, given a fixed $v \in E$, the linear map $\mathrm{app}_v\colon \mathcal{L}(E;F) \to F$, defined such that $\mathrm{app}_v(L) = L(v)$, is a continuous map.

Also recall from Proposition 19.7, that if $h\colon A \to G$ is a function such that $\mathrm{D}h(a)$ exits, and $k\colon G \to H$ is a continuous linear map, then, $\mathrm{D}(k \circ h)(a)$ exists, and

$$k(\mathrm{D}h(a)(u)) = \mathrm{D}(k \circ h)(a)(u),$$

that is,

$$k(\mathrm{D}_u h(a)) = \mathrm{D}_u(k \circ h)(a),$$

Applying these two facts to $h = \mathrm{D}f$, and to $k = \mathrm{app}_v$, we have

$$\mathrm{D}_u(\mathrm{D}f)(a)(v) = \mathrm{D}_u(\mathrm{app}_v \circ \mathrm{D}f)(a).$$

But $(\mathrm{app}_v \circ \mathrm{D}f)(x) = \mathrm{D}f(x)(v) = \mathrm{D}_v f(x)$, for every $x \in A$, that is, $\mathrm{app}_v \circ \mathrm{D}f = \mathrm{D}_v f$ on A. So, we have

$$\mathrm{D}_u(\mathrm{D}f)(a)(v) = \mathrm{D}_u(\mathrm{D}_v f)(a),$$

and since $\mathrm{D}^2 f(a)(u) = \mathrm{D}_u(\mathrm{D}f)(a)$, we get

$$\mathrm{D}^2 f(a)(u)(v) = \mathrm{D}_u(\mathrm{D}_v f)(a).$$

Thus, when $\mathrm{D}^2 f(a)$ exists, $\mathrm{D}_u(\mathrm{D}_v f)(a)$ exists, and

$$\mathrm{D}^2 f(a)(u)(v) = \mathrm{D}_u(\mathrm{D}_v f)(a),$$

for all $u, v \in E$. We also denote $\mathrm{D}_u(\mathrm{D}_v f)(a)$ by $\mathrm{D}^2_{u,v} f(a)$, or $\mathrm{D}_u\mathrm{D}_v f(a)$.

Recall from Proposition 18.18, that the map from $\mathcal{L}_2(E, E; F)$ to $\mathcal{L}(E; \mathcal{L}(E; F))$ defined such that $g \mapsto \varphi$ iff for every $g \in \mathcal{L}_2(E, E; F)$,

$$\varphi(u)(v) = g(u, v),$$

is an isomorphism of vector spaces. Thus, we will consider $\mathrm{D}^2 f(a) \in \mathcal{L}(E; \mathcal{L}(E; F))$ as a continuous bilinear map in $\mathcal{L}_2(E, E; F)$, and we will write $\mathrm{D}^2 f(a)(u, v)$, instead of $\mathrm{D}^2 f(a)(u)(v)$.

Then, the above discussion can be summarized by saying that when $\mathrm{D}^2 f(a)$ is defined, we have

$$\mathrm{D}^2 f(a)(u, v) = \mathrm{D}_u\mathrm{D}_v f(a).$$

When E has finite dimension and $(e_1, \ldots, e_n)$ is a basis for E, we denote $\mathrm{D}_{e_j}\mathrm{D}_{e_i} f(a)$ by $\dfrac{\partial^2 f}{\partial x_i \partial x_j}(a)$, when $i \neq j$, and we denote $\mathrm{D}_{e_i}\mathrm{D}_{e_i} f(a)$ by $\dfrac{\partial^2 f}{\partial x_i^2}(a)$.

The following important lemma attributed to Schwarz can be shown, using Lemma 19.12. Given a bilinear map $f\colon E \times E \to F$, recall that f is *symmetric*, if

$$f(u, v) = f(v, u),$$

for all $u, v \in E$.

Lemma 19.18. *(Schwarz's lemma) Given two normed vector spaces E and F, given any open subset A of E, given any $f\colon A \to F$, for every $a \in A$, if $\mathrm{D}^2 f(a)$ exists, then $\mathrm{D}^2 f(a) \in \mathcal{L}_2(E, E; F)$ is a continuous symmetric bilinear map. As a corollary, if E is of finite dimension n, and $(e_1, \ldots, e_n)$ is a basis for E, we have*

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(a) = \frac{\partial^2 f}{\partial x_j \partial x_i}(a).$$

Remark: There is a variation of the above lemma which does not assume the existence of $\mathrm{D}^2 f(a)$, but instead assumes that $\mathrm{D}_u\mathrm{D}_v f$ and $\mathrm{D}_v\mathrm{D}_u f$ exist on an open subset containing a and are continuous at a, and concludes that $\mathrm{D}_u\mathrm{D}_v f(a) = \mathrm{D}_v\mathrm{D}_u f(a)$. This is just a different result which does not imply Lemma 19.18, and is not a consequence of Lemma 19.18.

When $E = \mathbb{R}^2$, the only existence of $\dfrac{\partial^2 f}{\partial x \partial y}(a)$ and $\dfrac{\partial^2 f}{\partial y \partial x}(a)$ is not sufficient to insure the existence of $\mathrm{D}^2 f(a)$.

When E if of finite dimension n and $(e_1, \ldots, e_n)$ is a basis for E, if $\mathrm{D}^2 f(a)$ exists, for every $u = u_1 e_1 + \cdots + u_n e_n$ and $v = v_1 e_1 + \cdots + v_n e_n$ in E, since $\mathrm{D}^2 f(a)$ is a symmetric bilinear form, we have

$$\mathrm{D}^2 f(a)(u, v) = \sum_{i=1,j=1}^{n} u_i v_j \frac{\partial^2 f}{\partial x_i \partial x_j}(a),$$

which can be written in matrix form as:

$$\mathrm{D}^2 f(a)(u, v) = U^\top \begin{pmatrix} \dfrac{\partial^2 f}{\partial x_1^2}(a) & \dfrac{\partial^2 f}{\partial x_1 \partial x_2}(a) & \cdots & \dfrac{\partial^2 f}{\partial x_1 \partial x_n}(a) \\ \dfrac{\partial^2 f}{\partial x_1 \partial x_2}(a) & \dfrac{\partial^2 f}{\partial x_2^2}(a) & \cdots & \dfrac{\partial^2 f}{\partial x_2 \partial x_n}(a) \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_1 \partial x_n}(a) & \dfrac{\partial^2 f}{\partial x_2 \partial x_n}(a) & \cdots & \dfrac{\partial^2 f}{\partial x_n^2}(a) \end{pmatrix} V$$

where U is the column matrix representing u, and V is the column matrix representing v, over the basis $(e_1, \ldots, e_n)$.

The above symmetric matrix is called the *Hessian of f at a*. If F itself is of finite dimension, and $(v_1, \ldots, v_m)$ is a basis for F, then $f = (f_1, \ldots, f_m)$, and each component $\mathrm{D}^2 f(a)_i(u, v)$ of $\mathrm{D}^2 f(a)(u, v)$ $(1 \leq i \leq m)$, can be written as

$$\mathrm{D}^2 f(a)_i(u, v) = U^\top \begin{pmatrix} \dfrac{\partial^2 f_i}{\partial x_1^2}(a) & \dfrac{\partial^2 f_i}{\partial x_1 \partial x_2}(a) & \cdots & \dfrac{\partial^2 f_i}{\partial x_1 \partial x_n}(a) \\ \dfrac{\partial^2 f_i}{\partial x_1 \partial x_2}(a) & \dfrac{\partial^2 f_i}{\partial x_2^2}(a) & \cdots & \dfrac{\partial^2 f_i}{\partial x_2 \partial x_n}(a) \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f_i}{\partial x_1 \partial x_n}(a) & \dfrac{\partial^2 f_i}{\partial x_2 \partial x_n}(a) & \cdots & \dfrac{\partial^2 f_i}{\partial x_n^2}(a) \end{pmatrix} V$$

Thus, we could describe the vector $\mathrm{D}^2 f(a)(u, v)$ in terms of an $mn \times mn$-matrix consisting of m diagonal blocks, which are the above Hessians, and the row matrix $(U^\top, \ldots, U^\top)$ (m times) and the column matrix consisting of m copies of V. In particular, if $m = 1$, that is, $F = \mathbb{R}$ or $F = \mathbb{C}$, then the Hessian matrix is an $n \times n$ matrix.

We now indicate briefly how higher-order derivatives are defined. Let $m \geq 2$. Given a function $f \colon A \to F$ as before, for any $a \in A$, if the derivatives $\mathrm{D}^i f$ exist on A for all i, $1 \leq i \leq m-1$, by induction, $\mathrm{D}^{m-1} f$ can be considered to be a continuous function $\mathrm{D}^{m-1} f \colon A \to \mathcal{L}_{m-1}(E^{m-1}; F)$, and we define

$$\mathrm{D}^m f(a) = \mathrm{D}(\mathrm{D}^{m-1} f)(a).$$

Then, $\mathrm{D}^m f(a)$ can be identified with a continuous m-multilinear map in $\mathcal{L}_m(E^m; F)$. We can then show (as we did before), that if $\mathrm{D}^m f(a)$ is defined, then

$$\mathrm{D}^m f(a)(u_1, \ldots, u_m) = \mathrm{D}_{u_1} \ldots \mathrm{D}_{u_m} f(a).$$

When E if of finite dimension n and $(e_1, \ldots, e_n)$ is a basis for E, if $\mathrm{D}^m f(a)$ exists, for every $j_1, \ldots, j_m \in \{1, \ldots, n\}$, we denote $\mathrm{D}_{e_{j_m}} \ldots \mathrm{D}_{e_{j_1}} f(a)$ by

$$\frac{\partial^m f}{\partial x_{j_1} \ldots \partial x_{j_m}}(a).$$

Given a m-multilinear map $f \in \mathcal{L}_m(E^m; F)$, recall that f is *symmetric* if

$$f(u_{\pi(1)}, \ldots, u_{\pi(m)}) = f(u_1, \ldots, u_m),$$

for all $u_1, \ldots, u_m \in E$, and all permutations π on $\{1, \ldots, m\}$. Then, the following generalization of Schwarz's lemma holds.

Lemma 19.19. *Given two normed vector spaces E and F, given any open subset A of E, given any $f\colon A \to F$, for every $a \in A$, for every $m \geq 1$, if $\mathrm{D}^m f(a)$ exists, then $\mathrm{D}^m f(a) \in \mathcal{L}_m(E^m; F)$ is a continuous symmetric m-multilinear map. As a corollary, if E is of finite dimension n, and $(e_1, \ldots, e_n)$ is a basis for E, we have*

$$\frac{\partial^m f}{\partial x_{j_1} \ldots \partial x_{j_m}}(a) = \frac{\partial^m f}{\partial x_{\pi(j_1)} \ldots \partial x_{\pi(j_m)}}(a),$$

for every $j_1, \ldots, j_m \in \{1, \ldots, n\}$, and for every permutation π on $\{1, \ldots, m\}$.

If E is of finite dimension n, and $(e_1, \ldots, e_n)$ is a basis for E, $\mathrm{D}^m f(a)$ is a symmetric m-multilinear map, and we have

$$\mathrm{D}^m f(a)(u_1, \ldots, u_m) = \sum_j u_{1,j_1} \cdots u_{m,j_m} \frac{\partial^m f}{\partial x_{j_1} \ldots \partial x_{j_m}}(a),$$

where j ranges over all functions $j\colon \{1, \ldots, m\} \to \{1, \ldots, n\}$, for any m vectors

$$u_j = u_{j,1} e_1 + \cdots + u_{j,n} e_n.$$

The concept of C^1-function is generalized to the concept of C^m-function, and Theorem 19.13 can also be generalized.

Definition 19.9. Given two normed vector spaces E and F, and an open subset A of E, for any $m \geq 1$, we say that a function $f\colon A \to F$ is *of class C^m on A or a C^m-function on A* if $\mathrm{D}^k f$ exists and is continuous on A for every k, $1 \leq k \leq m$. We say that $f\colon A \to F$ is *of class C^∞ on A or a C^∞-function on A* if $\mathrm{D}^k f$ exists and is continuous on A for every $k \geq 1$. A C^∞-function (on A) is also called a *smooth function* (on A). A *C^m-diffeomorphism $f\colon A \to B$ between A and B* (where A is an open subset of E and B is an open subset of B) is a bijection between A and $B = f(A)$, such that both $f\colon A \to B$ and its inverse $f^{-1}\colon B \to A$ are C^m-functions.

Equivalently, f is a C^m-function on A if f is a C^1-function on A and $\mathrm{D}f$ is a C^{m-1}-function on A.

We have the following theorem giving a necessary and sufficient condition for f to a C^m-function on A. A generalization to the case where $E = E_1 \oplus \cdots \oplus E_n$ also holds.

Theorem 19.20. *Given two normed vector spaces E and F, where E is of finite dimension n, and where $(u_1, \ldots, u_n)$ is a basis of E, given any open subset A of E, given any function $f\colon A \to F$, for any $m \geq 1$, the derivative $\mathrm{D}^m f$ is a C^m-function on A iff every partial derivative $\mathrm{D}_{u_{j_k}} \ldots \mathrm{D}_{u_{j_1}} f$ (or $\dfrac{\partial^k f}{\partial x_{j_1} \ldots \partial x_{j_k}}(a)$) is defined and continuous on A, for all k, $1 \leq k \leq m$, and all $j_1, \ldots, j_k \in \{1, \ldots, n\}$. As a corollary, if F is of finite dimension p, and $(v_1, \ldots, v_p)$ is a basis of F, the derivative $\mathrm{D}^m f$ is defined and continuous on A iff every partial derivative $\mathrm{D}_{u_{j_k}} \ldots \mathrm{D}_{u_{j_1}} f_i$ (or $\dfrac{\partial^k f_i}{\partial x_{j_1} \ldots \partial x_{j_k}}(a)$) is defined and continuous on A, for all k, $1 \leq k \leq m$, for all i, $1 \leq i \leq p$, and all $j_1, \ldots, j_k \in \{1, \ldots, n\}$.*

When $E = \mathbb{R}$ (or $E = \mathbb{C}$), for any $a \in E$, $\mathrm{D}^m f(a)(1, \ldots, 1)$ is a vector in F, called the mth-order vector derivative. As in the case $m = 1$, we will usually identify the multilinear map $\mathrm{D}^m f(a)$ with the vector $\mathrm{D}^m f(a)(1, \ldots, 1)$. Some notational conventions can also be introduced to simplify the notation of higher-order derivatives, and we discuss such conventions very briefly.

Recall that when E is of finite dimension n, and $(e_1, \ldots, e_n)$ is a basis for E, $\mathrm{D}^m f(a)$ is a symmetric m-multilinear map, and we have

$$\mathrm{D}^m f(a)(u_1, \ldots, u_m) = \sum_j u_{1,j_1} \cdots u_{m,j_m} \frac{\partial^m f}{\partial x_{j_1} \ldots \partial x_{j_m}}(a),$$

where j ranges over all functions $j\colon \{1, \ldots, m\} \to \{1, \ldots, n\}$, for any m vectors

$$u_j = u_{j,1} e_1 + \cdots + u_{j,n} e_n.$$

We can then group the various occurrences of ∂x_{j_k} corresponding to the same variable x_{j_k}, and this leads to the notation

$$\Big(\frac{\partial}{\partial x_1}\Big)^{\alpha_1} \Big(\frac{\partial}{\partial x_2}\Big)^{\alpha_2} \cdots \Big(\frac{\partial}{\partial x_n}\Big)^{\alpha_n} f(a),$$

where $\alpha_1 + \alpha_2 + \cdots + \alpha_n = m$.

If we denote $(\alpha_1, \ldots, \alpha_n)$ simply by α, then we denote

$$\Big(\frac{\partial}{\partial x_1}\Big)^{\alpha_1} \Big(\frac{\partial}{\partial x_2}\Big)^{\alpha_2} \cdots \Big(\frac{\partial}{\partial x_n}\Big)^{\alpha_n} f$$

by

$$\partial^\alpha f, \quad \text{or} \quad \Big(\frac{\partial}{\partial x}\Big)^{\alpha} f.$$

If $\alpha = (\alpha_1, \ldots, \alpha_n)$, we let $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_n$, $\alpha! = \alpha_1! \cdots \alpha_n!$, and if $h = (h_1, \ldots, h_n)$, we denote $h_1^{\alpha_1} \cdots h_n^{\alpha_n}$ by h^α.

In the next section, we survey various versions of Taylor's formula.

19.5 Taylor's Formula, Faà di Bruno's Formula

We discuss, without proofs, several versions of Taylor's formula. The hypotheses required in each version become increasingly stronger. The first version can be viewed as a generalization of the notion of derivative. Given an m-linear map $f\colon E^m \to F$, for any vector $h \in E$, we abbreviate

$$f(\underbrace{h, \ldots, h}_{m})$$

by $f(h^m)$. The version of Taylor's formula given next is sometimes referred to as the *formula of Taylor–Young*.

Theorem 19.21. *(Taylor–Young) Given two normed vector spaces E and F, for any open subset $A \subseteq E$, for any function $f\colon A \to F$, for any $a \in A$, if $\mathrm{D}^k f$ exists in A for all k, $1 \leq k \leq m-1$, and if $\mathrm{D}^m f(a)$ exists, then we have:*

$$f(a+h) = f(a) + \frac{1}{1!}\mathrm{D}^1 f(a)(h) + \cdots + \frac{1}{m!}\mathrm{D}^m f(a)(h^m) + \|h\|^m \epsilon(h),$$

for any h such that $a + h \in A$, and where $\lim_{h \to 0,\, h \neq 0} \epsilon(h) = 0$.

The above version of Taylor's formula has applications to the study of relative maxima (or minima) of real-valued functions. It is also used to study the local properties of curves and surfaces.

The next version of Taylor's formula can be viewed as a generalization of Lemma 19.12. It is sometimes called the *Taylor formula with Lagrange remainder* or *generalized mean value theorem*.

Theorem 19.22. *(Generalized mean value theorem) Let E and F be two normed vector spaces, let A be an open subset of E, and let $f\colon A \to F$ be a function on A. Given any $a \in A$ and any $h \neq 0$ in E, if the closed segment $[a, a+h]$ is contained in A, $\mathrm{D}^k f$ exists in A for all k, $1 \leq k \leq m$, $\mathrm{D}^{m+1} f(x)$ exists at every point x of the open segment $]a, a+h[$, and*

$$\max_{x \in (a, a+h)} \left\| \mathrm{D}^{m+1} f(x) \right\| \leq M,$$

for some $M \geq 0$, then

$$\left\| f(a+h) - f(a) - \left(\frac{1}{1!}\mathrm{D}^1 f(a)(h) + \cdots + \frac{1}{m!}\mathrm{D}^m f(a)(h^m) \right) \right\| \leq M \frac{\|h\|^{m+1}}{(m+1)!}.$$

As a corollary, if $L\colon E^{m+1} \to F$ is a continuous $(m+1)$-linear map, then

$$\left\| f(a+h) - f(a) - \Big(\frac{1}{1!}\mathrm{D}^1 f(a)(h) + \cdots + \frac{1}{m!}\mathrm{D}^m f(a)(h^m) + \frac{L(h^{m+1})}{(m+1)!}\Big)\right\| \leq M\frac{\|h\|^{m+1}}{(m+1)!},$$

where $M = \max_{x\in(a,a+h)} \|\mathrm{D}^{m+1}f(x) - L\|$.

The above theorem is sometimes stated under the slightly stronger assumption that f is a C^m-function on A. If $f\colon A \to \mathbb{R}$ is a real-valued function, Theorem 19.22 can be refined a little bit. This version is often called the *formula of Taylor–Maclaurin*.

Theorem 19.23. *(Taylor–Maclaurin) Let E be a normed vector space, let A be an open subset of E, and let $f\colon A \to \mathbb{R}$ be a real-valued function on A. Given any $a \in A$ and any $h \neq 0$ in E, if the closed segment $[a, a+h]$ is contained in A, if $\mathrm{D}^k f$ exists in A for all k, $1 \leq k \leq m$, and $\mathrm{D}^{m+1}f(x)$ exists at every point x of the open segment $]a, a+h[$, then there is some $\theta \in \mathbb{R}$, with $0 < \theta < 1$, such that*

$$f(a+h) = f(a) + \frac{1}{1!}\mathrm{D}^1 f(a)(h) + \cdots + \frac{1}{m!}\mathrm{D}^m f(a)(h^m) + \frac{1}{(m+1)!}\mathrm{D}^{m+1}f(a+\theta h)(h^{m+1}).$$

We also mention for "mathematical culture," a version with integral remainder, in the case of a real-valued function. This is usually called *Taylor's formula with integral remainder*.

Theorem 19.24. *(Taylor's formula with integral remainder) Let E be a normed vector space, let A be an open subset of E, and let $f\colon A \to \mathbb{R}$ be a real-valued function on A. Given any $a \in A$ and any $h \neq 0$ in E, if the closed segment $[a, a+h]$ is contained in A, and if f is a C^{m+1}-function on A, then we have*

$$f(a+h) = f(a) + \frac{1}{1!}\mathrm{D}^1 f(a)(h) + \cdots + \frac{1}{m!}\mathrm{D}^m f(a)(h^m) + \int_0^1 \frac{(1-t)^m}{m!}\Big[\mathrm{D}^{m+1}f(a+th)(h^{m+1})\Big]dt.$$

The advantage of the above formula is that it gives an explicit remainder. We now examine briefly the situation where E is of finite dimension n, and $(e_1, \ldots, e_n)$ is a basis for E. In this case, we get a more explicit expression for the expression

$$\sum_{i=0}^{k=m} \frac{1}{k!}\mathrm{D}^k f(a)(h^k)$$

involved in all versions of Taylor's formula, where by convention, $\mathrm{D}^0 f(a)(h^0) = f(a)$. If $h = h_1 e_1 + \cdots + h_n e_n$, then we have

$$\sum_{k=0}^{k=m} \frac{1}{k!}\mathrm{D}^k f(a)(h^k) = \sum_{k_1+\cdots+k_n\leq m} \frac{h_1^{k_1}\cdots h_n^{k_n}}{k_1!\cdots k_n!}\Big(\frac{\partial}{\partial x_1}\Big)^{k_1}\cdots\Big(\frac{\partial}{\partial x_n}\Big)^{k_n} f(a),$$

which, using the abbreviated notation introduced at the end of Section 19.4, can also be written as

$$\sum_{k=0}^{k=m} \frac{1}{k!} \mathrm{D}^k f(a)(h^k) = \sum_{|\alpha| \leq m} \frac{h^\alpha}{\alpha!} \partial^\alpha f(a).$$

The advantange of the above notation is that it is the same as the notation used when $n = 1$, i.e., when $E = \mathbb{R}$ (or $E = \mathbb{C}$). Indeed, in this case, the Taylor–Maclaurin formula reads as:

$$f(a+h) = f(a) + \frac{h}{1!}\mathrm{D}^1 f(a) + \cdots + \frac{h^m}{m!}\mathrm{D}^m f(a) + \frac{h^{m+1}}{(m+1)!}\mathrm{D}^{m+1} f(a+\theta h),$$

for some $\theta \in \mathbb{R}$, with $0 < \theta < 1$, where $\mathrm{D}^k f(a)$ is the value of the k-th derivative of f at a (and thus, as we have already said several times, this is the kth-order vector derivative, which is just a scalar, since $F = \mathbb{R}$).

In the above formula, the assumptions are that $f\colon [a, a+h] \to \mathbb{R}$ is a C^m-function on $[a, a+h]$, and that $\mathrm{D}^{m+1} f(x)$ exists for every $x \in (a, a+h)$.

Taylor's formula is useful to study the local properties of curves and surfaces. In the case of a curve, we consider a function $f\colon [r, s] \to F$ from a closed interval $[r, s]$ of $\mathbb{R}$ to some vector space F, the derivatives $\mathrm{D}^k f(a)(h^k)$ correspond to vectors $h^k \mathrm{D}^k f(a)$, where $\mathrm{D}^k f(a)$ is the kth vector derivative of f at a (which is really $\mathrm{D}^k f(a)(1, \ldots, 1)$), and for any $a \in (r, s)$, Theorem 19.21 yields the following formula:

$$f(a+h) = f(a) + \frac{h}{1!}\mathrm{D}^1 f(a) + \cdots + \frac{h^m}{m!}\mathrm{D}^m f(a) + h^m \epsilon(h),$$

for any h such that $a + h \in (r, s)$, and where $\lim_{h \to 0,\, h \neq 0} \epsilon(h) = 0$.

In the case of functions $f\colon \mathbb{R}^n \to \mathbb{R}$, it is convenient to have formulae for the Taylor–Young formula and the Taylor–Maclaurin formula in terms of the gradient and the Hessian. Recall that the *gradient* $\nabla f(a)$ of f at $a \in \mathbb{R}^n$ is the column vector

$$\nabla f(a) = \begin{pmatrix} \dfrac{\partial f}{\partial x_1}(a) \\ \dfrac{\partial f}{\partial x_2}(a) \\ \vdots \\ \dfrac{\partial f}{\partial x_n}(a) \end{pmatrix},$$

and that

$$f'(a)(u) = \mathrm{D}f(a)(u) = \nabla f(a) \cdot u,$$

for any $u \in \mathbb{R}^n$ (where $\cdot$ means inner product). The above equation shows that *the direction of the gradient $\nabla f(a)$ is the direction of maximal increase of the function f at a* and that

$\|\nabla f(a)\|$ *is the rate of change of* f *in its direction of maximal increase*. This is the reason why methods of "gradient descent" pick the direction *opposite* to the gradient (we are trying to minimize f).

The *Hessian matrix* $\nabla^2 f(a)$ of f at $a \in \mathbb{R}^n$ is the $n \times n$ symmetric matrix

$$\nabla^2 f(a) = \begin{pmatrix} \dfrac{\partial^2 f}{\partial x_1^2}(a) & \dfrac{\partial^2 f}{\partial x_1 \partial x_2}(a) & \cdots & \dfrac{\partial^2 f}{\partial x_1 \partial x_n}(a) \\ \dfrac{\partial^2 f}{\partial x_1 \partial x_2}(a) & \dfrac{\partial^2 f}{\partial x_2^2}(a) & \cdots & \dfrac{\partial^2 f}{\partial x_2 \partial x_n}(a) \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_1 \partial x_n}(a) & \dfrac{\partial^2 f}{\partial x_2 \partial x_n}(a) & \cdots & \dfrac{\partial^2 f}{\partial x_n^2}(a) \end{pmatrix},$$

and we have

$$\mathrm{D}^2 f(a)(u, v) = u^\top \nabla^2 f(a)\, v = u \cdot \nabla^2 f(a) v = \nabla^2 f(a) u \cdot v,$$

for all $u, v \in \mathbb{R}^n$. Then, we have the following three formulations of the formula of Taylor–Young of order 2:

$$f(a+h) = f(a) + \mathrm{D}f(a)(h) + \frac{1}{2}\mathrm{D}^2 f(a)(h,h) + \|h\|^2 \epsilon(h)$$
$$f(a+h) = f(a) + \nabla f(a) \cdot h + \frac{1}{2}(h \cdot \nabla^2 f(a) h) + (h \cdot h)\epsilon(h)$$
$$f(a+h) = f(a) + (\nabla f(a))^\top h + \frac{1}{2}(h^\top \nabla^2 f(a)\, h) + (h^\top h)\epsilon(h),$$

with $\lim_{h \mapsto 0} \epsilon(h) = 0$.

One should keep in mind that only the first formula is intrinsic (i.e., does not depend on the choice of a basis), whereas the other two depend on the basis and the inner product chosen on $\mathbb{R}^n$. As an exercise, the reader should write similar formulae for the Taylor–Maclaurin formula of order 2.

Another application of Taylor's formula is the derivation of a formula which gives the m-th derivative of the composition of two functions, usually known as "Faà di Bruno's formula." This formula is useful when dealing with geometric continuity of splines curves and surfaces.

Proposition 19.25. *Given any normed vector space* E*, for any function* $f \colon \mathbb{R} \to \mathbb{R}$ *and any function* $g \colon \mathbb{R} \to E$*, for any* $a \in \mathbb{R}$*, letting* $b = f(a)$*,* $f^{(i)}(a) = \mathrm{D}^i f(a)$*, and* $g^{(i)}(b) = \mathrm{D}^i g(b)$*, for any* $m \geq 1$*, if* $f^{(i)}(a)$ *and* $g^{(i)}(b)$ *exist for all* i*,* $1 \leq i \leq m$*, then* $(g \circ f)^{(m)}(a) = \mathrm{D}^m(g \circ f)(a)$ *exists and is given by the following formula:*

$$(g \circ f)^{(m)}(a) = \sum_{0 \leq j \leq m} \sum_{\substack{i_1 + i_2 + \cdots + i_m = j \\ i_1 + 2i_2 + \cdots + m i_m = m \\ i_1, i_2, \cdots, i_m \geq 0}} \frac{m!}{i_1! \cdots i_m!}\, g^{(j)}(b) \left(\frac{f^{(1)}(a)}{1!}\right)^{i_1} \cdots \left(\frac{f^{(m)}(a)}{m!}\right)^{i_m}.$$

When $m = 1$, the above simplifies to the familiar formula

$$(g \circ f)'(a) = g'(b)f'(a),$$

and for $m = 2$, we have

$$(g \circ f)^{(2)}(a) = g^{(2)}(b)(f^{(1)}(a))^2 + g^{(1)}(b)f^{(2)}(a).$$

19.6 Futher Readings

A thorough treatment of differential calculus can be found in Munkres [76], Lang [65], Schwartz [91], Cartan [26], and Avez [7]. The techniques of differential calculus have many applications, especially to the geometry of curves and surfaces and to differential geometry in general. For this, we recommend do Carmo [36, 37] (two beautiful classics on the subject), Kreyszig [61], Stoker [99], Gray [50], Berger and Gostiaux [11], Milnor [75], Lang [63], Warner [111] and Choquet-Bruhat [28].

19.7 Summary

The main concepts and results of this chapter are listed below:

- *Directional derivative* ($\mathrm{D}_u f(a)$).
- *Total derivative*, *Fréchet derivative*, *derivative*, *total differential*, *differential* ($df(a), df_a$).
- *Partial derivatives*.
- *Affine* functions.
- The *chain rule*.
- *Jacobian matrices* ($J(f)(a)$) *Jacobians*.
- *Gradient* of a function ($\operatorname{grad} f(a)$, $\nabla f(a)$).
- *Mean value theorem*.
- C^0*-functions,* C^1*-functions*.
- The *implicit function theorem*.
- *Local homeomorphisms, local diffeomorphisms*, *diffeomorphisms*.
- The *inverse function theorem*.

- *Immersions, submersions.*
- Second-order derivatives.
- *Schwarz's lemma.*
- *Hessian matrix.*
- C^∞*-functions*, *smooth functions.*
- *Taylor–Young's formula.*
- Generalized mean value theorem.
- *Taylor–MacLaurin's formula.*
- *Taylor's formula with integral remainder.*
- *Faà di Bruno's formula.*

Chapter 20

Extrema of Real-Valued Functions

20.1 Local Extrema, Constrained Local Extrema, and Lagrange Multipliers

Let $J: E \to \mathbb{R}$ be a real-valued function defined on a normed vector space E (or more generally, any topological space). Ideally we would like to find where the function J reaches a minimum or a maximum value, at least locally. In this chapter we will usually use the notations $dJ(u)$ or $J'(u)$ (or dJ_u or J'_u) for the derivative of J at u, instead of $\mathrm{D}J(u)$. Our presentation follows very closely that of Ciarlet [30] (Chapter 7), which we find to be one of the clearest.

Definition 20.1. If $J: E \to \mathbb{R}$ is a real-valued function defined on a normed vector space E, we say that J has a *local minimum* (or *relative minimum*) at the point $u \in E$ if there is some open subset $W \subseteq E$ containing u such that

$$J(u) \leq J(w) \quad \text{for all } w \in W.$$

Similarly, we say that J has a *local maximum* (or *relative maximum*) at the point $u \in E$ if there is some open subset $W \subseteq E$ containing u such that

$$J(u) \geq J(w) \quad \text{for all } w \in W.$$

In either case, we say that J has a *local extremum* (or *relative extremum*) at u. We say that J has a *strict local minimum* (resp. *strict local maximum*) at the point $u \in E$ if there is some open subset $W \subseteq E$ containing u such that

$$J(u) < J(w) \quad \text{for all } w \in W - \{u\}$$

(resp.

$$J(u) > J(w) \quad \text{for all } w \in W - \{u\}).$$

By abuse of language, we often say that the point u itself "is a local minimum" or a "local maximum," even though, strictly speaking, this does not make sense.

We begin with a well-known necessary condition for a local extremum.

Proposition 20.1. *Let E be a normed vector space and let $J: \Omega \to \mathbb{R}$ be a function, with Ω some open subset of E. If the function J has a local extremum at some point $u \in \Omega$ and if J is differentiable at u, then*

$$dJ_u = J'(u) = 0.$$

Proof. Pick any $v \in E$. Since Ω is open, for t small enough we have $u + tv \in \Omega$, so there is an open interval $I \subseteq \mathbb{R}$ such that the function φ given by

$$\varphi(t) = J(u + tv)$$

for all $t \in I$ is well-defined. By applying the chain rule, we see that φ is differentiable at $t = 0$, and we get

$$\varphi'(0) = dJ_u(v).$$

Without loss of generality, assume that u is a local minimum. Then we have

$$\varphi'(0) = \lim_{t \mapsto 0_-} \frac{\varphi(t) - \varphi(0)}{t} \leq 0$$

and

$$\varphi'(0) = \lim_{t \mapsto 0_+} \frac{\varphi(t) - \varphi(0)}{t} \geq 0,$$

which shows that $\varphi'(0) = dJ_u(v) = 0$. As $v \in E$ is arbitrary, we conclude that $dJ_u = 0$. □

A point $u \in \Omega$ such that $J'(u) = 0$ is called a *critical point* of J.

If $E = \mathbb{R}^n$, then the condition $dJ_u = 0$ is equivalent to the system

$$\begin{aligned} \frac{\partial J}{\partial x_1}(u_1, \ldots, u_n) &= 0 \\ &\vdots \\ \frac{\partial J}{\partial x_n}(u_1, \ldots, u_n) &= 0. \end{aligned}$$

The condition of Proposition 20.1 is only a *necessary* condition for the existences of an extremum, but not a sufficient condition. Here are some counter-examples. If $f: \mathbb{R} \to \mathbb{R}$ is the function given by $f(x) = x^3$, since $f'(x) = 3x^2$, we have $f'(0) = 0$, but 0 is neither a minimum nor a maximum of f. If $g: \mathbb{R}^2 \to \mathbb{R}$ is the function given by $g(x, y) = x^2 - y^2$, then $g'_{(x,y)} = (2x \;\; -2y)$, so $g'_{(0,0)} = (0 \;\; 0)$, yet near $(0, 0)$ the function g takes negative and positive values.

In many practical situations, we need to look for local extrema of a function J *under additional constraints*. This situation can be formalized conveniently as follows: We have a function $J\colon \Omega \to \mathbb{R}$ defined on some open subset Ω of a normed vector space, but we also have some subset U of Ω, and we are looking for the local extrema of J *with respect to the set* U.

The elements $u \in U$ are often called *feasible solutions* of the optimization problem consisting in finding the local extrema of some objective function J with respect to some subset U of Ω defined by a set of constraints. Note that in most cases, U is *not* open. In fact, U is usually closed.

Definition 20.2. If $J\colon \Omega \to \mathbb{R}$ is a real-valued function defined on some open subset Ω of a normed vector space E and if U is some subset of Ω, we say that J has a *local minimum* (or *relative minimum*) at the point $u \in U$ *with respect to* U if there is some open subset $W \subseteq \Omega$ containing u such that

$$J(u) \leq J(w) \quad \text{for all } w \in U \cap W.$$

Similarly, we say that J has a *local maximum* (or *relative maximum*) at the point $u \in U$ *with respect to* U if there is some open subset $W \subseteq \Omega$ containing u such that

$$J(u) \geq J(w) \quad \text{for all } w \in U \cap W.$$

In either case, we say that J has a *local extremum* at u *with respect to* U.

It is very important to note that the hypothesis that Ω *is open* is crucial for the validity of Proposition 20.1. For example, if J is the identity function on $\mathbb{R}$ and $U = [0, 1]$, a closed subset, then $J'(x) = 1$ for all $x \in [0, 1]$, even though J has a minimum at $x = 0$ and a maximum at $x = 1$.

Therefore, in order to find sufficient conditions for a function $J\colon \Omega \to \mathbb{R}$ to have a local extremum with respect to a subset U of Ω (where Ω is open), we need to somehow incorporate the definition of U into these conditions. This can be done in two cases:

(1) The set U is defined by a set of equations,

$$U = \{x \in \Omega \mid \varphi_i(x) = 0,\ 1 \leq i \leq m\},$$

where the functions $\varphi_i\colon \Omega \to \mathbb{R}$ are continuous (and usually differentiable).

(2) The set U is defined by a set of inequalities,

$$U = \{x \in \Omega \mid \varphi_i(x) \leq 0,\ 1 \leq i \leq m\},$$

where the functions $\varphi_i\colon \Omega \to \mathbb{R}$ are continuous (and usually differentiable).

In (1), the equations $\varphi_i(x) = 0$ are called *equality constraints*, and in (2), the inequalities $\varphi_i(x) \leq 0$ are called *inequality constraints*.

An inequality constraint of the form $\varphi_i(x) \geq 0$ is equivalent to the inequality constraint $-\varphi_x(x) \leq 0$. An equality constraint $\varphi_i(x) = 0$ is equivalent to the conjunction of the two inequality constraints $\varphi_i(x) \leq 0$ and $-\varphi_i(x) \leq 0$, so the case of inequality constraints subsumes the case of equality constraints. However, the case of equality constraints is easier to deal with, and in this chapter we will restrict our attention to this case.

If the functions φ_i are convex and Ω is convex, then U is convex. This is a very important case that we will discuss later. In particular, if the functions φ_i are affine, then the equality constraints can be written as $Ax = b$, and the inequality constraints as $Ax \leq b$, for some $m \times n$ matrix A and some vector $b \in \mathbb{R}^m$. We will also discuss the case of affine constraints later.

In the case of equality constraints, a necessary condition for a local extremum with respect to U can be given in terms of *Lagrange multipliers*. In the case of inequality constraints, there is also a necessary condition for a local extremum with respect to U in terms of generalized Lagrange multipliers and the *Karush–Kuhn–Tucker* conditions. This will be discussed in the next chapter.

We begin by considering the case where $\Omega \subseteq E_1 \times E_2$ is an open subset of a product of normed vector spaces and where U is the zero locus of some continuous function $\varphi\colon \Omega \to E_2$, which means that

$$U = \{(u_1, u_2) \in \Omega \mid \varphi(u_1, u_2) = 0\}.$$

For the sake of brevity, we say that J has a *constrained local extremum* at u instead of saying that J has a *local extremum* at the point $u \in U$ *with respect to* U. Fortunately, there is a necessary condition for constrained local extrema in terms of *Lagrange multipliers*.

Theorem 20.2. *(Necessary condition for a constrained extremum) Let $\Omega \subseteq E_1 \times E_2$ be an open subset of a product of normed vector spaces, with E_1 a Banach space (E_1 is complete), let $\varphi\colon \Omega \to E_2$ be a C^1-function (which means that $d\varphi(\omega)$ exists and is continuous for all $\omega \in \Omega$), and let*

$$U = \{(u_1, u_2) \in \Omega \mid \varphi(u_1, u_2) = 0\}.$$

Moreover, let $u = (u_1, u_2) \in U$ be a point such that

$$\frac{\partial \varphi}{\partial x_2}(u_1, u_2) \in \mathcal{L}(E_2; E_2) \quad \textit{and} \quad \left(\frac{\partial \varphi}{\partial x_2}(u_1, u_2)\right)^{-1} \in \mathcal{L}(E_2; E_2),$$

and let $J\colon \Omega \to \mathbb{R}$ be a function which is differentiable at u. If J has a constrained local extremum at u, then there is a continuous linear form $\Lambda(u) \in \mathcal{L}(E_2; \mathbb{R})$ such that

$$dJ(u) + \Lambda(u) \circ d\varphi(u) = 0.$$

Proof. The plan of attack is to use the implicit function theorem; Theorem 19.14. Observe that the assumptions of Theorem 19.14 are indeed met. Therefore, there exist some open subsets $U_1 \subseteq E_1$, $U_2 \subseteq E_2$, and a continuous function $g\colon U_1 \to U_2$ with $(u_1, u_2) \in U_1 \times U_2 \subseteq \Omega$ and such that

$$\varphi(v_1, g(v_1)) = 0$$

for all $v_1 \in U_1$. Moreover, g is differentiable at $u_1 \in U_1$ and

$$dg(u_1) = -\left(\frac{\partial \varphi}{\partial x_2}(u)\right)^{-1} \circ \frac{\partial \varphi}{\partial x_1}(u).$$

It follows that the restriction of J to $(U_1 \times U_2) \cap U$ yields a function G of a single variable, with

$$G(v_1) = J(v_1, g(v_1))$$

for all $v_1 \in U_1$. Now, the function G is differentiable at u_1 and it has a local extremum at u_1 on U_1, so Proposition 20.1 implies that

$$dG(u_1) = 0.$$

By the chain rule,

$$\begin{aligned} dG(u_1) &= \frac{\partial J}{\partial x_1}(u) + \frac{\partial J}{\partial x_2}(u) \circ dg(u_1) \\ &= \frac{\partial J}{\partial x_1}(u) - \frac{\partial J}{\partial x_2}(u) \circ \left(\frac{\partial \varphi}{\partial x_2}(u)\right)^{-1} \circ \frac{\partial \varphi}{\partial x_1}(u). \end{aligned}$$

From $dG(u_1) = 0$, we deduce

$$\frac{\partial J}{\partial x_1}(u) = \frac{\partial J}{\partial x_2}(u) \circ \left(\frac{\partial \varphi}{\partial x_2}(u)\right)^{-1} \circ \frac{\partial \varphi}{\partial x_1}(u),$$

and since we also have

$$\frac{\partial J}{\partial x_2}(u) = \frac{\partial J}{\partial x_2}(u) \circ \left(\frac{\partial \varphi}{\partial x_2}(u)\right)^{-1} \circ \frac{\partial \varphi}{\partial x_2}(u),$$

if we let

$$\Lambda(u) = -\frac{\partial J}{\partial x_2}(u) \circ \left(\frac{\partial \varphi}{\partial x_2}(u)\right)^{-1},$$

then we get

$$\begin{aligned} dJ(u) &= \frac{\partial J}{\partial x_1}(u) + \frac{\partial J}{\partial x_2}(u) \\ &= \frac{\partial J}{\partial x_2}(u) \circ \left(\frac{\partial \varphi}{\partial x_2}(u)\right)^{-1} \circ \left(\frac{\partial \varphi}{\partial x_1}(u) + \frac{\partial \varphi}{\partial x_2}(u)\right) \\ &= -\Lambda(u) \circ d\varphi(u), \end{aligned}$$

which yields $dJ(u) + \Lambda(u) \circ d\varphi(u) = 0$, as claimed. □

In most applications, we have $E_1 = \mathbb{R}^{n-m}$ and $E_2 = \mathbb{R}^m$ for some integers m, n such that $1 \leq m < n$, Ω is an open subset of $\mathbb{R}^n$, $J\colon \Omega \to \mathbb{R}$, and we have m functions $\varphi_i\colon \Omega \to \mathbb{R}$ defining the subset

$$U = \{v \in \Omega \mid \varphi_i(v) = 0,\ 1 \leq i \leq m\}.$$

Theorem 20.2 yields the following necessary condition:

Theorem 20.3. *(Necessary condition for a constrained extremum in terms of Lagrange multipliers) Let Ω be an open subset of $\mathbb{R}^n$, consider m C^1-functions $\varphi_i\colon \Omega \to \mathbb{R}$ (with $1 \leq m < n$), let*

$$U = \{v \in \Omega \mid \varphi_i(v) = 0,\ 1 \leq i \leq m\},$$

and let $u \in U$ be a point such that the derivatives $d\varphi_i(u) \in \mathcal{L}(\mathbb{R}^n; \mathbb{R})$ are linearly independent; equivalently, assume that the $m \times n$ matrix $\big((\partial\varphi_i/\partial x_j)(u)\big)$ has rank m. If $J\colon \Omega \to \mathbb{R}$ is a function which is differentiable at $u \in U$ and if J has a local constrained extremum at u, then there exist m numbers $\lambda_i(u) \in \mathbb{R}$, uniquely defined, such that

$$dJ(u) + \lambda_1(u)d\varphi_1(u) + \cdots + \lambda_m(u)d\varphi_m(u) = 0;$$

equivalently,

$$\nabla J(u) + \lambda_1(u)\nabla\varphi_1(u) + \cdots + \lambda_1(u)\nabla\varphi_m(u) = 0.$$

Proof. The linear independence of the m linear forms $d\varphi_i(u)$ is equivalent to the fact that the $m \times n$ matrix $A = \big((\partial\varphi_i/\partial x_j)(u)\big)$ has rank m. By reordering the columns, we may assume that the first m columns are linearly independent. If we let $\varphi\colon \Omega \to \mathbb{R}^m$ be the function defined by

$$\varphi(v) = (\varphi_1(v), \ldots, \varphi_m(v))$$

for all $v \in \Omega$, then we see that $\partial\varphi/\partial x_2(u)$ is invertible and both $\partial\varphi/\partial x_2(u)$ and its inverse are continuous, so that Theorem 20.2 applies, and there is some (continuous) linear form $\Lambda(u) \in \mathcal{L}(\mathbb{R}^m; \mathbb{R})$ such that

$$dJ(u) + \Lambda(u) \circ d\varphi(u) = 0.$$

However, $\Lambda(u)$ is defined by some m-tuple $(\lambda_1(u), \ldots, \lambda_m(u)) \in \mathbb{R}^m$, and in view of the definition of φ, the above equation is equivalent to

$$dJ(u) + \lambda_1(u)d\varphi_1(u) + \cdots + \lambda_m(u)d\varphi_m(u) = 0.$$

The uniqueness of the $\lambda_i(u)$ is a consequence of the linear independence of the $d\varphi_i(u)$. □

The numbers $\lambda_i(u)$ involved in Theorem 20.3 are called the *Lagrange multipliers* associated with the constrained extremum u (again, with some minor abuse of language). The linear independence of the linear forms $d\varphi_i(u)$ is equivalent to the fact that the Jacobian matrix $\big((\partial\varphi_i/\partial x_j)(u)\big)$ of $\varphi = (\varphi_1, \ldots, \varphi_m)$ at u has rank m. If $m = 1$, the linear independence of the $d\varphi_i(u)$ reduces to the condition $\nabla\varphi_1(u) \neq 0$.

A fruitful way to reformulate the use of Lagrange multipliers is to introduce the notion of the *Lagrangian* associated with our constrained extremum problem. This is the function $L\colon \Omega \times \mathbb{R}^m \to \mathbb{R}$ given by

$$L(v, \lambda) = J(v) + \lambda_1 \varphi_1(v) + \cdots + \lambda_m \varphi_m(v),$$

with $\lambda = (\lambda_1, \ldots, \lambda_m)$. Then, observe that there exists some $\mu = (\mu_1, \ldots, \mu_m)$ and some $u \in U$ such that

$$dJ(u) + \mu_1 d\varphi_1(u) + \cdots + \mu_m d\varphi_m(u) = 0$$

if and only if

$$dL(u, \mu) = 0,$$

or equivalently

$$\nabla L(u, \mu) = 0;$$

that is, iff (u, λ) is a *critical point* of the Lagrangian L.

Indeed $dL(u, \mu) = 0$ if equivalent to

$$\begin{aligned}
\frac{\partial L}{\partial v}(u, \mu) &= 0 \\
\frac{\partial L}{\partial \lambda_1}(u, \mu) &= 0 \\
&\vdots \\
\frac{\partial L}{\partial \lambda_m}(u, \mu) &= 0,
\end{aligned}$$

and since

$$\frac{\partial L}{\partial v}(u, \mu) = dJ(u) + \mu_1 d\varphi_1(u) + \cdots + \mu_m d\varphi_m(u)$$

and

$$\frac{\partial L}{\partial \lambda_i}(u, \mu) = \varphi_i(u),$$

we get

$$dJ(u) + \mu_1 d\varphi_1(u) + \cdots + \mu_m d\varphi_m(u) = 0$$

and

$$\varphi_1(u) = \cdots = \varphi_m(u) = 0,$$

that is, $u \in U$.

If we write out explicitly the condition

$$dJ(u) + \mu_1 d\varphi_1(u) + \cdots + \mu_m d\varphi_m(u) = 0,$$

we get the $n \times m$ system

$$\begin{aligned}
\frac{\partial J}{\partial x_1}(u) + \lambda_1 \frac{\partial \varphi_1}{\partial x_1}(u) + \cdots + \lambda_m \frac{\partial \varphi_m}{\partial x_1}(u) &= 0 \\
&\vdots \\
\frac{\partial J}{\partial x_n}(u) + \lambda_1 \frac{\partial \varphi_1}{\partial x_n}(u) + \cdots + \lambda_m \frac{\partial \varphi_m}{\partial x_n}(u) &= 0,
\end{aligned}$$

and it is important to note that the matrix of this system is the *transpose* of the Jacobian matrix of φ at u. If we write $\mathrm{Jac}(J)(u) = \big((\partial \varphi_i/\partial x_j)(u)\big)$ for the Jacobian matrix of J (at u), then the above system is written in matrix form as

$$\nabla J(u) + (\mathrm{Jac}(J)(u))^\top \lambda = 0,$$

where λ is viewed as a column vector, and the Lagrangian is equal to

$$L(u, \lambda) = J(u) + (\varphi_1(u), \ldots, \varphi_m(u))\lambda.$$

Remark: If the Jacobian matrix $\mathrm{Jac}(J)(v) = \big((\partial \varphi_i/\partial x_j)(v)\big)$ has rank m for all $v \in U$ (which is equivalent to the linear independence of the linear forms $d\varphi_i(v)$), then we say that $0 \in \mathbb{R}^m$ is a *regular value* of φ. In this case, it is known that

$$U = \{v \in \Omega \mid \varphi(v) = 0\}$$

is a *smooth submanifold of dimension* $n - m$ *of* $\mathbb{R}^n$. Furthermore, the set

$$T_v U = \{w \in \mathbb{R}^n \mid d\varphi_i(v)(w) = 0,\ 1 \leq i \leq m\} = \bigcap_{i=1}^{m} \mathrm{Ker}\, d\varphi_i(v)$$

is the *tangent space* to U at v (a vector space of dimension $n - m$). Then, the condition

$$dJ(v) + \mu_1 d\varphi_1(v) + \cdots + \mu_m d\varphi_m(v) = 0$$

implies that $dJ(v)$ vanishes on the tangent space $T_v U$. Conversely, if $dJ(v)(w) = 0$ for all $w \in T_v U$, this means that $dJ(v)$ is orthogonal (in the sense of Definition 8.3) to $T_v U$. Since (by Theorem 8.1 (b)) the orthogonal of $T_v U$ is the space of linear forms spanned by $d\varphi_1(v), \ldots, d\varphi_m(v)$, it follows that $dJ(v)$ must be a linear combination of the $d\varphi_i(v)$. Therefore, when 0 is a regular value of φ, Theorem 20.3 asserts that if $u \in U$ is a local extremum of J, then $dJ(u)$ must vanish on the tangent space $T_u U$. We can say even more. The subset $Z(J)$ of Ω given by

$$Z(J) = \{v \in \Omega \mid J(v) = J(u)\}$$

(the *level set of level* $J(u)$) is a hypersurface in Ω, and if $dJ(u) \neq 0$, the zero locus of $dJ(u)$ is the tangent space $T_uZ(J)$ to $Z(J)$ at u (a vector space of dimension $n-1$), where

$$T_uZ(J) = \{w \in \mathbb{R}^n \mid dJ(u)(w) = 0\}.$$

Consequently, Theorem 20.3 asserts that

$$T_uU \subseteq T_uZ(J);$$

this is a geometric condition.

The beauty of the Lagrangian is that the constraints $\{\varphi_i(v) = 0\}$ have been incorporated into the function $L(v, \lambda)$, and that the necessary condition for the existence of a constrained local extremum of J is reduced to the necessary condition for the existence of a local extremum of the *unconstrained* L.

However, one should be careful to check that the assumptions of Theorem 20.3 are satisfied (in particular, the linear independence of the linear forms $d\varphi_i$). For example, let $J\colon \mathbb{R}^3 \to \mathbb{R}$ be given by

$$J(x, y, z) = x + y + z^2$$

and $g\colon \mathbb{R}^3 \to \mathbb{R}$ by

$$g(x, y, z) = x^2 + y^2.$$

Since $g(x, y, z) = 0$ iff $x = y = 0$, we have $U = \{(0, 0, z) \mid z \in \mathbb{R}\}$ and the restriction of J to U is given by

$$J(0, 0, z) = z^2,$$

which has a minimum for $z = 0$. However, a "blind" use of Lagrange multipliers would require that there is some λ so that

$$\frac{\partial J}{\partial x}(0, 0, z) = \lambda \frac{\partial g}{\partial x}(0, 0, z), \quad \frac{\partial J}{\partial y}(0, 0, z) = \lambda \frac{\partial g}{\partial y}(0, 0, z), \quad \frac{\partial J}{\partial z}(0, 0, z) = \lambda \frac{\partial g}{\partial z}(0, 0, z),$$

and since

$$\frac{\partial g}{\partial x}(x, y, z) = 2x, \quad \frac{\partial g}{\partial y}(x, y, z) = 2y, \quad \frac{\partial g}{\partial z}(0, 0, z) = 0,$$

the partial derivatives above all vanish for $x = y = 0$, so at a local extremum we should also have

$$\frac{\partial J}{\partial x}(0, 0, z) = 0, \quad \frac{\partial J}{\partial y}(0, 0, z) = 0, \quad \frac{\partial J}{\partial z}(0, 0, z) = 0,$$

but this is absurd since

$$\frac{\partial J}{\partial x}(x, y, z) = 1, \quad \frac{\partial J}{\partial y}(x, y, z) = 1, \quad \frac{\partial J}{\partial z}(x, y, z) = 2z.$$

The reader should enjoy finding the reason for the flaw in the argument.

One should also keep in mind that Theorem 20.3 gives only a necessary condition. The (u, λ) may *not* correspond to local extrema! Thus, it is always necessary to analyze the local behavior of J near a critical point u. This is generally difficult, but in the case where J is affine or quadratic and the constraints are affine or quadratic, this is possible (although not always easy).

Let us apply the above method to the following example in which $E_1 = \mathbb{R}$, $E_2 = \mathbb{R}$, $\Omega = \mathbb{R}^2$, and

$$\begin{aligned} J(x_1, x_2) &= -x_2 \\ \varphi(x_1, x_2) &= x_1^2 + x_2^2 - 1. \end{aligned}$$

Observe that

$$U = \{(x_1, x_2) \in \mathbb{R}^2 \mid x_1^2 + x_2^2 = 1\}$$

is the unit circle, and since

$$\nabla\varphi(x_1, x_2) = \begin{pmatrix} 2x_1 \\ 2x_2 \end{pmatrix},$$

it is clear that $\nabla\varphi(x_1, x_2) \neq 0$ for every point $= (x_1, x_2)$ on the unit circle. If we form the Lagrangian

$$L(x_1, x_2, \lambda) = -x_2 + \lambda(x_1^2 + x_2^2 - 1),$$

Theorem 20.3 says that a necessary condition for J to have a constrained local extremum is that $\nabla L(x_1, x_2, \lambda) = 0$, so the following equations must hold:

$$\begin{aligned} 2\lambda x_1 &= 0 \\ -1 + 2\lambda x_2 &= 0 \\ x_1^2 + x_2^2 &= 1. \end{aligned}$$

The second equation implies that $\lambda \neq 0$, and then the first yields $x_1 = 0$, so the third yields $x_2 = \pm 1$, and we get two solutions:

$$\begin{aligned} \lambda &= \frac{1}{2}, && (x_1, x_2) = (0, 1) \\ \lambda &= -\frac{1}{2}, && (x_1', x_2') = (0, -1). \end{aligned}$$

We can check immediately that the first solution is a minimum and the second is a maximum. The reader should look for a geometric interpretation of this problem.

Let us now consider the case in which J is a quadratic function of the form

$$J(v) = \frac{1}{2} v^\top A v - v^\top b,$$

where A is an $n \times n$ symmetric matrix, $b \in \mathbb{R}^n$, and the constraints are given by a linear system of the form

$$Cv = d,$$

where C is an $m \times n$ matrix with $m < n$ and $d \in \mathbb{R}^m$. We also assume that C has rank m. In this case, the function φ is given by

$$\varphi(v) = (Cv - d)^\top,$$

because we view $\varphi(v)$ as a row vector (and v as a column vector), and since

$$d\varphi(v)(w) = C^\top w,$$

the condition that the Jacobian matrix of φ at u have rank m is satisfied. The Lagrangian of this problem is

$$L(v, \lambda) = \frac{1}{2} v^\top A v - v^\top b + (Cv - d)^\top \lambda = \frac{1}{2} v^\top A v - v^\top b + \lambda^\top (Cv - d),$$

where λ is viewed as a column vector. Now, because A is a symmetric matrix, it is easy to show that

$$\nabla L(v, \lambda) = \begin{pmatrix} Av - b + C^\top \lambda \\ Cv - d \end{pmatrix}.$$

Therefore, the necessary condition for contrained local extrema is

$$\begin{aligned} Av + C^\top \lambda &= b \\ Cv &= d, \end{aligned}$$

which can be expressed in matrix form as

$$\begin{pmatrix} A & C^\top \\ C & 0 \end{pmatrix} \begin{pmatrix} v \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ d \end{pmatrix},$$

where the matrix of the system is a symmetric matrix. We should not be surprised to find the system of Section 22, except for some renaming of the matrices and vectors involved. As we know from Section 22.2, the function J has a minimum iff A is positive definite, so in general, if A is only a symmetric matrix, the critical points of the Lagrangian do *not* correspond to extrema of J.

We now investigate conditions for the existence of extrema involving the second derivative of J.

20.2 Using Second Derivatives to Find Extrema

For the sake of brevity, we consider only the case of local minima; analogous results are obtained for local maxima (replace J by $-J$, since $\max_u J(u) = -\min_u -J(u)$). We begin with a necessary condition for an unconstrained local minimum.

Proposition 20.4. *Let E be a normed vector space and let $J: \Omega \to \mathbb{R}$ be a function, with Ω some open subset of E. If the function J is differentiable in Ω, if J has a second derivative $\mathrm{D}^2 J(u)$ at some point $u \in \Omega$, and if J has a local minimum at u, then*

$$\mathrm{D}^2 J(u)(w, w) \geq 0 \quad \textit{for all } w \in E.$$

Proof. Pick any nonzero vector $w \in E$. Since Ω is open, for t small enough, $u + tw \in \Omega$ and $J(u + tw) \geq J(u)$, so there is some open interval $I \subseteq \mathbb{R}$ such that

$$u + tw \in \Omega \quad \text{and} \quad J(u + tw) \geq J(u)$$

for all $t \in I$. Using the Taylor–Young formula and the fact that we must have $dJ(u) = 0$ since J has a local minimum at u, we get

$$0 \leq J(u + tw) - J(u) = \frac{t^2}{2}\mathrm{D}^2 J(u)(w, w) + t^2 \left\|w\right\|^2 \epsilon(tw),$$

with $\lim_{t \mapsto 0} \epsilon(tw) = 0$, which implies that

$$\mathrm{D}^2 J(u)(w, w) \geq 0.$$

Since the argument holds for all $w \in E$ (trivially if $w = 0$), the proposition is proved. □

One should be cautioned that there is no converse to the previous proposition. For example, the function $f: x \mapsto x^3$ has no local minimum at 0, yet $df(0) = 0$ and $\mathrm{D}^2 f(0)(u, v) = 0$. Similarly, the reader should check that the function $f: \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x, y) = x^2 - 3y^3$$

has no local minimum at $(0, 0)$; yet $df(0, 0) = 0$ and $\mathrm{D}^2 f(0, 0)(u, v) = 2u^2 \geq 0$.

When $E = \mathbb{R}^n$, Proposition 20.4 says that a necessary condition for having a local minimum is that the Hessian $\nabla^2 J(u)$ be positive semidefinite (it is always symmetric).

We now give sufficient conditions for the existence of a local minimum.

Theorem 20.5. *Let E be a normed vector space, let $J: \Omega \to \mathbb{R}$ be a function with Ω some open subset of E, and assume that J is differentiable in Ω and that $dJ(u) = 0$ at some point $u \in \Omega$. The following properties hold:*

(1) If $\mathrm{D}^2 J(u)$ exists and if there is some number $\alpha \in \mathbb{R}$ such that $\alpha > 0$ and

$$\mathrm{D}^2 J(u)(w, w) \geq \alpha \left\|w\right\|^2 \quad \textit{for all } w \in E,$$

then J has a strict local minimum at u.

(2) If $\mathrm{D}^2J(v)$ *exists for all* $v \in \Omega$ *and if there is a ball* $B \subseteq \Omega$ *centered at* u *such that*

$$\mathrm{D}^2J(v)(w,w) \geq 0 \quad \textit{for all } v \in B \textit{ and all } w \in E,$$

then J *has a local minimum at* u.

Proof. (1) Using the formula of Taylor–Young, for every vector w small enough, we can write

$$\begin{aligned} J(u+w) - J(u) &= \frac{1}{2}\mathrm{D}^2J(u)(w,w) + \|w\|^2 \epsilon(w) \\ &\geq \left(\frac{1}{2}\alpha + \epsilon(w)\right)\|w\|^2 \end{aligned}$$

with $\lim_{w \mapsto 0} \epsilon(w) = 0$. Consequently if we pick $r > 0$ small enough that $|\epsilon(w)| < \alpha$ for all w with $\|w\| < r$, then $J(u+w) > J(u)$ for all $u + w \in B$, where B is the open ball of center u and radius r. This proves that J has a local strict minimum at u.

(2) The formula of Taylor–Maclaurin shows that for all $u + w \in B$, we have

$$J(u+w) = J(u) + \frac{1}{2}\mathrm{D}^2J(v)(w,w) \geq J(u),$$

for some $v \in (u, w+w)$. □

There are no converses of the two assertions of Theorem 20.5. However, there is a condition on $\mathrm{D}^2J(u)$ that implies the condition of Part (1). Since this condition is easier to state when $E = \mathbb{R}^n$, we begin with this case.

Recall that a $n \times n$ symmetric matrix A is *positive definite* if $x^\top Ax > 0$ for all $x \in \mathbb{R}^n - \{0\}$. In particular, A must be invertible.

Proposition 20.6. *For any symmetric matrix* A*, if* A *is positive definite, then there is some* $\alpha > 0$ *such that*

$$x^\top Ax \geq \alpha \|x\|^2 \quad \textit{for all } x \in \mathbb{R}^n.$$

Proof. Pick any norm in $\mathbb{R}^n$ (recall that all norms on $\mathbb{R}^n$ are equivalent). Since the unit sphere $S^{n-1} = \{x \in \mathbb{R}^n \mid \|x\| = 1\}$ is compact and since the function $f(x) = x^\top Ax$ is never zero on S^{n-1}, the function f has a minimum $\alpha > 0$ on S^{n-1}. Using the usual trick that $x = \|x\|\,(x/\|x\|)$ for every nonzero vector $x \in \mathbb{R}^n$ and the fact that the inequality of the proposition is trivial for $x = 0$, from

$$x^\top Ax \geq \alpha \quad \text{for all } x \text{ with } \|x\| = 1,$$

we get

$$x^\top Ax \geq \alpha \|x\|^2 \quad \text{for all } x \in \mathbb{R}^n,$$

as claimed. □

We can combine Theorem 20.5 and Proposition 20.6 to obtain a useful sufficient condition for the existence of a strict local minimum. First let us introduce some terminology.

Definition 20.3. Given a function $J: \Omega \to \mathbb{R}$ as before, say that a point $u \in \Omega$ is a *nondegenerate critical point* if $dJ(u) = 0$ and if the Hessian matrix $\nabla^2 J(u)$ is invertible.

Proposition 20.7. *Let $J: \Omega \to \mathbb{R}$ be a function defined on some open subset $\Omega \subseteq \mathbb{R}^n$. If J is differentiable in Ω and if some point $u \in \Omega$ is a nondegenerate critical point such that $\nabla^2 J(u)$ is positive definite, then J has a strict local minimum at u.*

Remark: It is possible to generalize Proposition 20.7 to infinite-dimensional spaces by finding a suitable generalization of the notion of a nondegenerate critical point. Firstly, we assume that E is a Banach space (a complete normed vector space). Then, we define the dual E' of E as the set of continuous linear forms on E, so that $E' = \mathcal{L}(E; \mathbb{R})$. Following Lang, we use the notation E' for the space of continuous linear forms to avoid confusion with the space $E^* = \mathrm{Hom}(E, \mathbb{R})$ of all linear maps from E to $\mathbb{R}$. A continuous bilinear map $\varphi: E \times E \to \mathbb{R}$ in $\mathcal{L}_2(E, E; \mathbb{R})$ yields a map Φ from E to E' given by

$$\Phi(u) = \varphi_u,$$

where $\varphi_u \in E'$ is the linear form defined by

$$\varphi_u(v) = \varphi(u, v).$$

It is easy to check that φ_u is continuous and that the map Φ is continuous. Then, we say that φ is *nondegenerate* iff $\Phi: E \to E'$ is an isomorphism of Banach spaces, which means that Φ is invertible and that both Φ and Φ^{-1} are continuous linear maps. Given a function $J: \Omega \to \mathbb{R}$ differentiable on Ω as before (where Ω is an open subset of E), if $\mathrm{D}^2 J(u)$ exists for some $u \in \Omega$, we say that u is a *nondegenerate critical point* if $dJ(u) = 0$ and if $\mathrm{D}^2 J(u)$ is nondegenerate. Of course, $\mathrm{D}^2 J(u)$ is positive definite if $\mathrm{D}^2 J(u)(w, w) > 0$ for all $w \in E - \{0\}$.

Using the above definition, Proposition 20.6 can be generalized to a nondegenerate positive definite bilinear form (on a Banach space) and Theorem 20.7 can also be generalized to the situation where $J: \Omega \to \mathbb{R}$ is defined on an open subset of a Banach space. For details and proofs, see Cartan [26] (Part I Chapter 8) and Avez [7] (Chapter 8 and Chapter 10).

In the next section we make use of convexity; both on the domain Ω and on the function J itself.

20.3 Using Convexity to Find Extrema

We begin by reviewing the definition of a convex set and of a convex function.

Definition 20.4. Given any real vector space E, we say that a subset C of E is *convex* if either $C = \emptyset$ or if for every pair of points $u, v \in C$, the line segment connecting u and v is contained in C, i.e.,

$$(1-\lambda)u + \lambda v \in C \quad \text{for all } \lambda \in \mathbb{R} \text{ such that } 0 \leq \lambda \leq 1.$$

Given any two points $u\, v \in E$, the *line segment* $[u, v]$ is the set

$$[u, v] = \{(1-\lambda)u + \lambda v \in E \mid \lambda \in \mathbb{R},\ 0 \leq \lambda \leq 1\}.$$

Clearly, a nonempty set C is convex iff $[u, v] \subseteq C$ whenever $u, v \in C$. See Figure 20.1 for an example of a convex set.

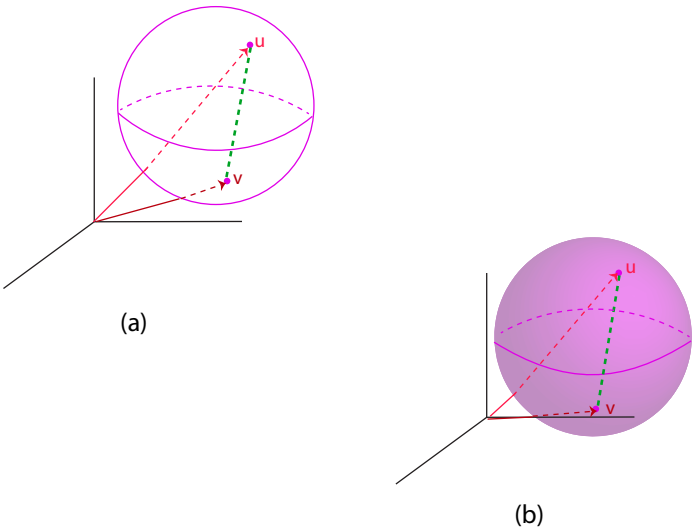


Figure 20.1: Figure (a) shows that a sphere is not convex in $\mathbb{R}^3$ since the dashed green line does not lie on its surface. Figure (b) shows that a solid ball is convex in $\mathbb{R}^3$.

Definition 20.5. If C is a nonempty convex subset of E, a function $f\colon C \to \mathbb{R}$ is *convex* (on C) if for every pair of points $u, v \in C$,

$$f((1-\lambda)u + \lambda v) \leq (1-\lambda)f(u) + \lambda f(v) \quad \text{for all } \lambda \in \mathbb{R} \text{ such that } 0 \leq \lambda \leq 1;$$

the function f is *strictly convex* (on C) if for every pair of distinct points $u, v \in C$ $(u \neq v)$,

$$f((1-\lambda)u + \lambda v) < (1-\lambda)f(u) + \lambda f(v) \quad \text{for all } \lambda \in \mathbb{R} \text{ such that } 0 < \lambda < 1;$$

see Figure 20.2. The *epigraph*[1] $\mathbf{epi}(f)$ of a function $f\colon A \to \mathbb{R}$ defined on some subset A of $\mathbb{R}^n$ is the subset of $\mathbb{R}^{n+1}$ defined as

$$\mathbf{epi}(f) = \{(x, y) \in \mathbb{R}^{n+1} \mid f(x) \leq y,\ x \in A\}.$$

[1] "Epi" means above.

A function $f\colon C \to \mathbb{R}$ defined on a convex subset C is *concave* (resp. *strictly concave*) if $(-f)$ is convex (resp. strictly convex).

It is obvious that a function f if convex iff its epigraph $\mathbf{epi}(f)$ is a convex subset of $\mathbb{R}^{n+1}$.

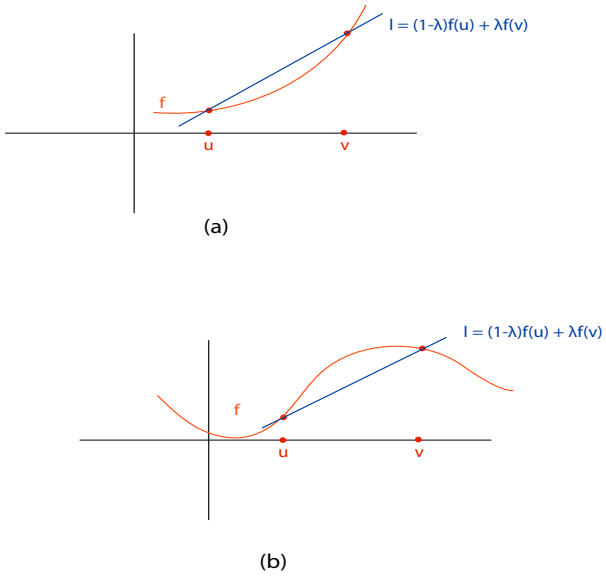


Figure 20.2: Figures (a) and (b) are the graphs of real valued functions. Figure (a) is the graph of convex function since the blue line lies above the graph of f. Figure (b) shows the graph of a function which is not convex.

Subspaces $V \subseteq E$ of a vector space E are convex; *affine subspaces*, that is, sets of the form $u + V$, where V is a subspace of E and $u \in E$, are convex. Balls (open or closed) are convex. Given any linear form $\varphi\colon E \to \mathbb{R}$, for any scalar $c \in \mathbb{R}$, the *closed half–spaces*

$$H^+_{\varphi,c} = \{u \in E \mid \varphi(u) \geq c\}, \qquad H^-_{\varphi,c} = \{u \in E \mid \varphi(u) \leq c\},$$

are convex. Any intersection of half–spaces is convex. More generally, any intersection of convex sets is convex.

Linear forms are convex functions (but not strictly convex). Any norm $\| \| : E \to \mathbb{R}_+$ is a convex function. The max function,

$$\max(x_1, \ldots, x_n) = \max\{x_1, \ldots, x_n\}$$

is convex on $\mathbb{R}^n$. The exponential $x \mapsto e^{cx}$ is strictly convex for any $c \neq 0$ $(c \in \mathbb{R})$. The logarithm function is concave on $\mathbb{R}_+ - \{0\}$, and the *log-determinant function* $\log\det$ is concave on the set of symmetric positive definite matrices. This function plays an important

role in convex optimization. An excellent exposition of convexity and its applications to optimization can be found in Boyd [22].

Here is a necessary condition for a function to have a local minimum with respect to a convex subset U.

Theorem 20.8. *(Necessary condition for a local minimum on a convex subset) Let $J: \Omega \to \mathbb{R}$ be a function defined on some open subset Ω of a normed vector space E and let $U \subseteq \Omega$ be a nonempty convex subset. Given any $u \in U$, if $dJ(u)$ exists and if J has a local minimum in u with respect to U, then*

$$dJ(u)(v-u) \geq 0 \quad \textit{for all } v \in U.$$

Proof. Let $v = u + w$ be an arbitrary point in U. Since U is convex, we have $u + tw \in U$ for all t such that $0 \leq t \leq 1$. Since $dJ(u)$ exists, we can write

$$J(u+tw) - J(u) = dJ(u)(tw) + \|tw\|\,\epsilon(tw)$$

with $\lim_{t \mapsto 0} \epsilon(tw) = 0$. However, because $0 \leq t \leq 1$,

$$J(u+tw) - J(u) = t(dJ(u)(w) + \|w\|\,\epsilon(tw))$$

and since u is a local minimum with respect to U, we have $J(u+tw) - J(u) \geq 0$, so we get

$$t(dJ(u)(w) + \|w\|\,\epsilon(tw)) \geq 0.$$

The above implies that $dJ(u)(w) \geq 0$, because otherwise we could pick $t > 0$ small enough so that

$$dJ(u)(w) + \|w\|\,\epsilon(tw) < 0,$$

a contradiction. Since the argument holds for all $v = u + w \in U$, the theorem is proved. □

Observe that the convexity of U is a substitute for the use of Lagrange multipliers, but we now have to deal with an *inequality* instead of an equality.

Consider the special case where U is a subspace of E. In this case since $u \in U$ we have $2u \in U$, and for any $u + w \in U$, we must have $2u - (u + w) = u - w \in U$. The previous theorem implies that $dJ(u)(w) \geq 0$ and $dJ(u)(-w) \geq 0$, that is, $dJ(u)(w) \leq 0$, so $dJ(u) = 0$. Since the argument holds for $w \in U$ (because U is a subspace, if $u, w \in U$, then $u + w \in U$), we conclude that

$$dJ(u)(w) = 0 \quad \text{for all } w \in U.$$

We will now characterize convex functions when they have a first derivative or a second derivative.

Proposition 20.9. *(Convexity and first derivative) Let $f: \Omega \to \mathbb{R}$ be a function differentiable on some open subset Ω of a normed vector space E and let $U \subseteq \Omega$ be a nonempty convex subset.*

(1) The function f is convex on U iff

$$f(v) \geq f(u) + df(u)(v-u) \quad \textit{for all } u, v \in U.$$

(2) The function f is strictly convex on U iff

$$f(v) > f(u) + df(u)(v-u) \quad \textit{for all } u, v \in U \textit{ with } u \neq v.$$

See Figure 20.3.

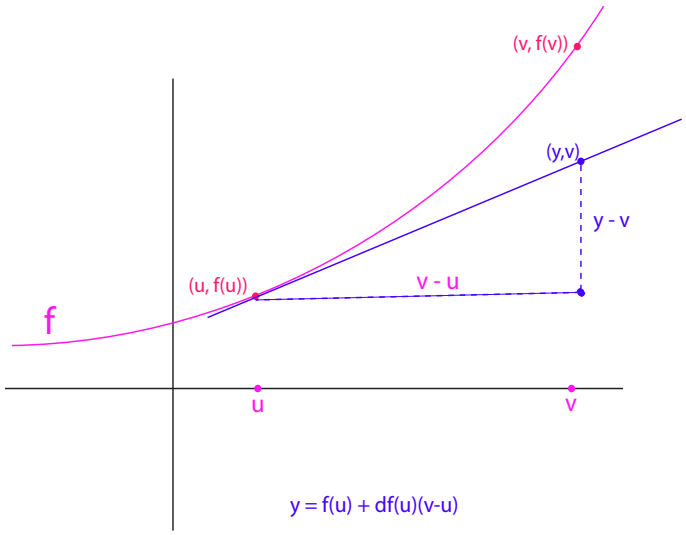


Figure 20.3: An illustration of a convex valued function f. Since f is convex it always lies above its tangent line.

Proof. Let $u, v \in U$ be any two distinct points and pick $\lambda \in \mathbb{R}$ with $0 < \lambda < 1$. If the function f is convex, then

$$f((1-\lambda)u + \lambda v) \leq (1-\lambda)f(u) + \lambda f(v),$$

which yields

$$\frac{f((1-\lambda)u + \lambda v) - f(u)}{\lambda} \leq f(v) - f(u).$$

It follows that

$$df(u)(v-u) = \lim_{\lambda \mapsto 0} \frac{f((1-\lambda)u + \lambda v) - f(u)}{\lambda} \leq f(v) - f(u).$$

If f is strictly convex, the above reasoning does not work, because a strict inequality is not necessarily preserved by "passing to the limit." We have recourse to the following trick: For any ω such that $0 < \omega < 1$, observe that

$$(1-\lambda)u + \lambda v = u + \lambda(v-u) = \frac{\omega-\lambda}{\omega}u + \frac{\lambda}{\omega}(u + \omega(v-u)).$$

If we assume that $0 < \lambda \leq \omega$, the convexity of f yields

$$f(u + \lambda(v-u)) \leq \frac{\omega-\lambda}{\omega}f(u) + \frac{\lambda}{\omega}f(u + \omega(v-u)).$$

If we subtract $f(u)$ to both sides, we get

$$\frac{f(u + \lambda(v-u)) - f(u)}{\lambda} \leq \frac{f(u + \omega(v-u)) - f(u)}{\omega}.$$

Now, since $0 < \omega < 1$ and f is strictly convex,

$$f(u + \omega(v-u)) = f((1-\omega)u + \omega v) < (1-\omega)f(u) + \omega f(v),$$

which implies that

$$\frac{f(u + \omega(v-u)) - f(u)}{\omega} < f(v) - f(u),$$

and thus we get

$$\frac{f(u + \lambda(v-u)) - f(u)}{\lambda} \leq \frac{f(u + \omega(v-u)) - f(u)}{\omega} < f(v) - f(u).$$

If we let λ go to 0, by passing to the limit we get

$$df(u)(v-u) \leq \frac{f(u + \omega(v-u)) - f(u)}{\omega} < f(v) - f(u),$$

which yields the desired strict inequality.

Let us now consider the converse of (1); that is, assume that

$$f(v) \geq f(u) + df(u)(v-u) \quad \text{for all } u, v \in U.$$

For any two distinct points $u, v \in U$ and for any λ with $0 < \lambda < 1$, we get

$$\begin{aligned} f(v) &\geq f(v + \lambda(v-u)) - \lambda df(v + \lambda(u-v))(u-v) \\ f(u) &\geq f(v + \lambda(u-v)) + (1-\lambda)df(v + \lambda(u-v))(u-v), \end{aligned}$$

and if we multiply the first inequality by $1-\lambda$ and the second inequality by λ and them add up the resulting inequalities, we get

$$(1-\lambda)f(v) + \lambda f(u) \geq f(v + \lambda(u-v)) = f((1-\lambda)v + \lambda u),$$

which proves that f is convex.

The proof of the converse of (2) is similar, except that the inequalities are replaced by strict inequalities. □

We now establish a convexity criterion using the second derivative of f. This criterion is often easier to check than the previous one.

Proposition 20.10. *(Convexity and second derivative) Let $f\colon \Omega \to \mathbb{R}$ be a function twice differentiable on some open subset Ω of a normed vector space E and let $U \subseteq \Omega$ be a nonempty convex subset.*

(1) The function f is convex on U iff

$$\mathrm{D}^2 f(u)(v-u, v-u) \geq 0 \quad \textit{for all } u, v \in U.$$

(2) If

$$\mathrm{D}^2 f(u)(v-u, v-u) > 0 \quad \textit{for all } u, v \in U \textit{ with } u \neq v,$$

then f is strictly convex.

Proof. First,assume that the inequality in Condition (1) is satisfied. For any two distinct points $u, v \in U$, the formula of Taylor–Maclaurin yields

$$\begin{aligned} f(v) - f(u) - df(u)(v-u) &= \frac{1}{2}\mathrm{D}^2 f(w)(v-u, v-u) \\ &= \frac{\rho^2}{2}\mathrm{D}^2 f(w)(v-w, v-w), \end{aligned}$$

for some $w = (1-\lambda)u + \lambda v = u + \lambda(v-u)$ with $0 < \lambda < 1$, and with $\rho = 1/(1-\lambda) > 0$, so that $v - u = \rho(v-w)$. Since $\mathrm{D}^2 f(u)(v-w, v-w) \geq 0$ for all $u, w \in U$, we conclude by applying Proposition 20.9(1).

Similarly, if (2) holds, the above reasoning and Proposition 20.9(2) imply that f is strictly convex.

To prove the necessary condition in (1), define $g\colon \Omega \to \mathbb{R}$ by

$$g(v) = f(v) - df(u)(v),$$

where $u \in U$ is any point considered fixed. If f is convex, since

$$g(v) - g(u) = f(v) - f(u) - df(u)(v-u),$$

Proposition 20.9 implies that $f(v) - f(u) - df(u)(v-u) \geq 0$, which implies that g has a local minimum at u with respect to all $v \in U$. Therefore, we have $dg(u) = 0$. Observe that g is twice differentiable in Ω and $\mathrm{D}^2 g(u) = \mathrm{D}^2 f(u)$, so the formula of Taylor–Young yields for every $v = u + w \in U$ and all t with $0 \leq t \leq 1$,

$$\begin{aligned} 0 \leq g(u+tw) - g(u) &= \frac{t^2}{2}\mathrm{D}^2 f(u)(tw, tw) + \|tw\|^2 \epsilon(tw) \\ &= \frac{t^2}{2}(\mathrm{D}^2 f(u)(w, w) + 2\|w\|^2 \epsilon(wt)), \end{aligned}$$

with $\lim_{t \mapsto 0} \epsilon(wt) = 0$, and for t small enough, we must have $\mathrm{D}^2 f(u)(w, w) \geq 0$, as claimed. □

The converse of Proposition 20.10 (2) is false as we see by considering the function f given by $f(x) = x^4$.

Example 20.1. On the other hand, if f is a quadratic function of the form

$$f(u) = \frac{1}{2}u^\top Au - u^\top b$$

where A is a symmetric matrix, we know that

$$df(u)(v) = v^\top(Au - b),$$

so

$$\begin{aligned} f(v) - f(u) - df(u)(v-u) &= \frac{1}{2}v^\top Av - v^\top b - \frac{1}{2}u^\top Au + u^\top b - (v-u)^\top(Au - b) \\ &= \frac{1}{2}v^\top Av - \frac{1}{2}u^\top Au - (v-u)^\top Au \\ &= \frac{1}{2}v^\top Av + \frac{1}{2}u^\top Au - v^\top Au \\ &= \frac{1}{2}(v-u)^\top A(v-u). \end{aligned}$$

Therefore, Proposition 20.9 implies that if A is positive semidefinite, then f is convex and if A is positive definite, then f is strictly convex. The converse follows by Proposition 20.10.

We conclude this section by applying our previous theorems to convex functions defined on convex subsets. In this case, local minima (resp. local maxima) are global minima (resp. global maxima).

Definition 20.6. Let $f\colon E \to \mathbb{R}$ be any function defined on some normed vector space (or more generally, any set). For any $u \in E$, we say that f has a *minimum* in u (resp. *maximum* in u) if

$$f(u) \leq f(v) \text{ (resp. } f(u) \geq f(v)) \quad \text{for all } v \in E.$$

We say that f has a *strict minimum* in u (resp. *strict maximum* in u) if

$$f(u) < f(v) \text{ (resp. } f(u) > f(v)) \quad \text{for all } v \in E - \{u\}.$$

If $U \subseteq E$ is a subset of E and $u \in U$, we say that f has a *minimum* in u (resp. *strict minimum* in u) *with respect to* U if

$$f(u) \leq f(v) \quad \text{for all } v \in U \quad (\text{resp. } f(u) < f(v) \quad \text{for all } v \in U - \{u\}),$$

and similarly for a *maximum* in u (resp. *strict maximum* in u) *with respect to* U with $\leq$ changed to $\geq$ and $<$ to $>$.

Sometimes, we say *global* maximum (or minimum) to stress that a maximum (or a minimum) is not simply a local maximum (or minimum).

Theorem 20.11. *Given any normed vector space E, let U be any nonempty convex subset of E.*

(1) For any convex function $J: U \to \mathbb{R}$, for any $u \in U$, if J has a local minimum at u in U, then J has a (global) minimum at u in U.

(2) Any strict convex function $J: U \to \mathbb{R}$ has at most one minimum (in U), and if it does, then it is a strict minimum (in U).

(3) Let $J: \Omega \to \mathbb{R}$ be any function defined on some open subset Ω of E with $U \subseteq \Omega$ and assume that J is convex on U. For any point $u \in U$, if $dJ(u)$ exists, then J has a minimum in u with respect to U iff

$$dJ(u)(v-u) \geq 0 \quad \textit{for all } v \in U.$$

(4) If the convex subset U in (3) is open, then the above condition is equivalent to

$$dJ(u) = 0.$$

Proof. (1) Let $v = u + w$ be any arbitrary point in U. Since J is convex, for all t with $0 \leq t \leq 1$, we have

$$J(u + tw) = J(u + t(v-u)) \leq (1-t)J(u) + tJ(v),$$

which yields

$$J(u + tw) - J(u) \leq t(J(v) - J(u)).$$

Because J has a local minimum in u, there is some t_0 with $0 < t_0 < 1$ such that

$$0 \leq J(u + t_0 w) - J(u),$$

which implies that $J(v) - J(u) \geq 0$.

(2) If J is strictly convex, the above reasoning with $w \neq 0$ shows that there is some t_0 with $0 < t_0 < 1$ such that

$$0 \leq J(u + t_0 w) - J(u) < t_0(J(v) - J(u)),$$

which shows that u is a strict global minimum (in U), and thus that it is unique.

(3) We already know from Theorem 20.8 that the condition $dJ(u)(v-u) \geq 0$ for all $v \in U$ is necessary (even if J is not convex). Conversely, because J is convex, careful inspection of the proof of part (1) of Proposition 20.9 shows that only the fact that $dJ(u)$ exists in needed to prove that

$$J(v) - J(u) \geq dJ(u)(v-u) \quad \text{for all } v \in U,$$

and if

$$dJ(u)(v-u) \geq 0 \quad \text{for all } v \in U,$$

then

$$J(v) - J(u) \geq 0 \quad \text{for all } v \in U,$$

as claimed.

(4) If U is open, then for every $u \in U$ we can find an open ball B centered at u of radius ϵ small enough so that $B \subseteq U$. Then, for any $w \neq 0$ such that $\|w\| < \epsilon$, we have both $v = u + w \in B$ and $v' = u - w \in B$, so condition (3) implies that

$$dJ(u)(w) \geq 0 \quad \text{and} \quad dJ(u)(-w) \geq 0,$$

which yields

$$dJ(u)(w) = 0.$$

Since the above holds for all $w \neq 0$ such such that $\|w\| < \epsilon$ and since $dJ(u)$ is linear, we leave it to the reader to fill in the details of the proof that $dJ(u) = 0$. □

Theorem 20.11 can be used to rederive the fact that the least squares solutions of a linear system $Ax = b$ (where A is an $m \times n$ matrix) are given by the normal equation

$$A^\top Ax = A^\top b.$$

For this, we consider the quadratic function

$$J(v) = \frac{1}{2}\|Av - b\|_2^2 - \frac{1}{2}\|b\|_2^2,$$

and our least squares problem is equivalent to finding the minima of J on $\mathbb{R}^n$. A computation reveals that

$$\begin{aligned} J(v) &= \frac{1}{2}\|Av - b\|_2^2 - \frac{1}{2}\|b\|_2^2 \\ &= \frac{1}{2}(Av - b)^\top (Av - b) - \frac{1}{2}b^\top b \\ &= \frac{1}{2}(v^\top A^\top - b^\top)(Av - b) - \frac{1}{2}b^\top b \\ &= \frac{1}{2}v^\top A^\top Av - v^\top A^\top b, \end{aligned}$$

and so

$$dJ(u) = A^\top Au - A^\top b.$$

Since $A^\top A$ is positive semidefinite, the function J is convex, and Theorem 20.11(4) implies that the minima of J are the solutions of the equation

$$A^\top Au - A^\top b = 0.$$

The considerations in this chapter reveal the need to find methods for finding the zeros of the derivative map

$$dJ : \Omega \to E',$$

where Ω is some open subset of a normed vector space E and E' is the space of all continuous linear forms on E (a subspace of E^*). Generalizations of *Newton's method* yield such methods and they are the objet of the next chapter.

20.4 Summary

The main concepts and results of this chapter are listed below:

- *Local minimum*, *local maximum*, *local extremum*, *strict local minimum*, *strict local maximum*.
- Necessary condition for a local extremum involving the derivative; *critical point*.
- *Local minimum with respect to a subset U*, *local maximum with respect to a subset U*, *local extremum with respect to a subset U*.
- *Constrained local extremum*.
- Necessary condition for a constrained extremum.
- Necessary condition for a constrained extremum in terms of *Lagrange multipliers*.
- *Lagrangian*.
- *Critical points of a Lagrangian*.
- Necessary condition of an unconstrained local minimum involving the second-order derivative.
- Sufficient condition for a local minimum involving the second-order derivative.
- A sufficient condition involving *nondegenerate critical points*.
- *Convex sets*, *convex functions*, *concave functions*, *strictly convex functions*, *strictly concave functions*,
- Necessary condition for a local minimum on a convex set involving the derivative.
- Convexity of a function involving a condition on its first derivative.
- Convexity of a function involving a condition on its second derivative.
- Minima of convex functions on convex sets.

Chapter 21

Newton's Method and Its Generalizations

21.1 Newton's Method for Real Functions of a Real Argument

In Chapter 20 we investigated the problem of determining when a function $J: \Omega \to \mathbb{R}$ defined on some open subset Ω of a normed vector space E has a local extremum. Proposition 20.1 gives a necessary condition when J is differentiable: if J has a local extremum at $u \in \Omega$, then we must have

$$J'(u) = 0.$$

Thus we are led to the problem of finding the zeros of the derivative

$$J': \Omega \to E',$$

where $E' = \mathcal{L}(E; \mathbb{R})$ is the set of linear continuous functions from E to $\mathbb{R}$; that is, the *dual* of E, as defined in the remark after Proposition 20.7.

This leads us to consider the problem in a more general form, namely: Given a function $f: \Omega \to Y$ from an open subset Ω of a normed vector space X to a normed vector space Y, find

(i) Sufficient conditions which guarantee the *existence of a zero* of the function f; that is, an element $a \in \Omega$ such that $f(a) = 0$.

(ii) An *algorithm* for approximating such an a, that is, a sequence (x_k) of points of Ω whose limit is a.

When $X = Y = \mathbb{R}$, we can use *Newton's method*. We pick some initial element $x_0 \in \mathbb{R}$ "close enough" to a zero a of f, and we define the sequence (x_k) by

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)},$$

for all $k \geq 0$, provided that $f'(x_k) \neq 0$. The idea is to define x_{k+1} as the intersection of the x-axis with the tangent line to the graph of the function $x \mapsto f(x)$ at the point $(x_k, f(x_k))$. Indeed, the equation of this tangent line is

$$y - f(x_k) = f'(x_k)(x - x_k),$$

and its intersection with the x-axis is obtained for $y = 0$, which yields

$$x = x_k - \frac{f(x_k)}{f'(x_k)},$$

as claimed.

For example, if $\alpha > 0$ and $f(x) = x^2 - \alpha$, Newton's method yields the sequence

$$x_{k+1} = \frac{1}{2}\left(x_k + \frac{\alpha}{x_k}\right)$$

to compute the square root $\sqrt{\alpha}$ of α. It can be shown that the method converges to $\sqrt{\alpha}$ for any $x_0 > 0$. Actually, the method also converges when $x_0 < 0$! Find out what is the limit.

The case of a real function suggests the following method for finding the zeros of a function $f\colon \Omega \to Y$, with $\Omega \subseteq X$: given a starting point $x_0 \in \Omega$, the sequence (x_k) is defined by

$$x_{k+1} = x_k - (f'(x_k))^{-1}(f(x_k))$$

for all $k \geq 0$.

For the above to make sense, it must be ensured that

(1) All the points x_k remain within Ω.

(2) The function f is differentiable within Ω.

(3) The derivative $f'(x)$ is a bijection from X to Y for all $x \in \Omega$.

These are rather demanding conditions but there are sufficient conditions that guarantee that they are met. Another practical issue is that it may be very costly to compute $(f'(x_k))^{-1}$ at every iteration step. In the next section, we investigate generalizations of Newton's method which address the issues that we just discussed.

21.2 Generalizations of Newton's Method

Suppose that $f\colon \Omega \to \mathbb{R}^n$ is given by n functions $f_i\colon \Omega \to \mathbb{R}$, where $\Omega \subseteq \mathbb{R}^n$. In this case, finding a zero a of f is equivalent to solving the system

$$\begin{aligned} f_1(a_1 \ldots, a_n) &= 0 \\ f_2(a_1 \ldots, a_n) &= 0 \\ &\vdots \\ f_n(a_1 \ldots, a_n) &= 0. \end{aligned}$$

A single iteration of Newton's method consists in solving the linear system

$$(J(f)(x_k))\epsilon_k = -f(x_k),$$

and then setting

$$x_{k+1} = x_k + \epsilon_k,$$

where $J(f)(x_k) = (\frac{\partial f_i}{\partial x_j}(x_k))$ is the Jacobian matrix of f at x_k.

In general, it is very costly to compute $J(f)(x_k)$ at each iteration and then to solve the corresponding linear system. If the method converges, the consecutive vectors x_k should differ only a little, as also the corresponding matrices $J(f)(x_k)$. Thus, we are led to a variant of Newton's method which consists in keeping the same matrix for p consecutive steps (where p is some fixed integer ≥ 2):

$$\begin{aligned}
x_{k+1} &= x_k - (f'(x_0))^{-1}(f(x_k)), && 0 \leq k \leq p-1\\
x_{k+1} &= x_k - (f'(x_p))^{-1}(f(x_k)), && p \leq k \leq 2p-1\\
&\vdots\\
x_{k+1} &= x_k - (f'(x_{rp}))^{-1}(f(x_k)), && rp \leq k \leq (r+1)p-1\\
&\vdots
\end{aligned}$$

It is also possible to set $p = \infty$, that is, to use the same matrix $f'(x_0)$ for all iterations, which leads to iterations of the form

$$x_{k+1} = x_k - (f'(x_0))^{-1}(f(x_k)), \quad k \geq 0,$$

or even to replace $f'(x_0)$ by a particular matrix A_0 which is easy to invert:

$$x_{k+1} = x_k - A_0^{-1}f(x_k), \quad k \geq 0.$$

In the last two cases, if possible, we use an LU factorization of $f'(x_0)$ or A_0 to speed up the method. In some cases, it may even possible to set $A_0 = I$.

The above considerations lead us to the definition of a *generalized Newton method*, as in Ciarlet [30] (Chapter 7). Recall that a linear map $f \in \mathcal{L}(E; F)$ is called an *isomorphism* iff f is continuous, bijective, and f^{-1} is also continuous.

Definition 21.1. If X and Y are two normed vector spaces and if $f\colon \Omega \to Y$ is a function from some open subset Ω of X, a *generalized Newton method* for finding zeros of f consists of

(1) A sequence of families $(A_k(x))$ of linear isomorphisms from X to Y, for all $x \in \Omega$ and all integers $k \geq 0$;

(2) Some starting point $x_0 \in \Omega$;

(3) A sequence (x_k) of points of Ω defined by

$$x_{k+1} = x_k - (A_k(x_\ell))^{-1}(f(x_k)), \quad k \geq 0,$$

where for every integer $k \geq 0$, the integer ℓ satisfies the condition

$$0 \leq \ell \leq k.$$

The function $A_k(x)$ usually depends on f'.

Definition 21.1 gives us enough flexibility to capture all the situations that we have previously discussed:

$$\begin{array}{ll} A_k(x) = f'(x), & \ell = k \\ A_k(x) = f'(x), & \ell = \min\{rp, k\}, \text{ if } rp \leq k \leq (r+1)p - 1, r \geq 0 \\ A_k(x) = f'(x), & \ell = 0 \\ A_k(x) = A_0, & \end{array}$$

where A_0 is a linear isomorphism from X to Y. The first case corresponds to Newton's orginal method and the others to the variants that we just discussed. We could also have $A_k(x) = A_k$, a fixed linear isomorphism independent of $x \in \Omega$.

The following theorem inspired by the *Newton–Kantorovich theorem* gives sufficient conditions that guarantee that the sequence (x_k) constructed by a generalized Newton method converges to a zero of f close to x_0. Althoug quite technical, these conditions are not very surprising.

Theorem 21.1. *Let X be a Banach space, let $f\colon \Omega \to Y$ be differentiable on the open subset $\Omega \subseteq X$, and assume that there are constants $r, M, \beta > 0$ such that if we let*

$$B = \{x \in X \mid \|x - x_0\| \leq r\} \subseteq \Omega,$$

then

(1)

$$\sup_{k \geq 0} \sup_{x \in B} \left\| A_k^{-1}(x) \right\|_{\mathcal{L}(Y;X)} \leq M,$$

(2) $\beta < 1$ and

$$\sup_{k \geq 0} \sup_{x, x' \in B} \|f'(x) - A_k(x')\|_{\mathcal{L}(X;Y)} \leq \frac{\beta}{M}$$

(3)

$$\|f(x_0)\| \leq \frac{r}{M}(1 - \beta).$$

Then, the sequence (x_k) *defined by*

$$x_{k+1} = x_k - A_k^{-1}(x_\ell)(f(x_k)), \quad 0 \leq \ell \leq k$$

is entirely contained within B *and converges to a zero* a *of* f*, which is the only zero of* f *in* B*. Furthermore, the convergence is geometric, which means that*

$$\|x_k - a\| \leq \frac{\|x_1 - x_0\|}{1-\beta}\beta^k.$$

A proof of Theorem 21.1 can be found in Ciarlet [30] (Section 7.5). It is not really difficult but quite technical.

If we assume that we already know that some element $a \in \Omega$ is a zero of f, the next theorem gives sufficient conditions for a special version of a generalized Newton method to converge. For this special method, the linear isomorphisms $A_k(x)$ are independent of $x \in \Omega$.

Theorem 21.2. *Let* X *be a Banach space, and let* $f\colon \Omega \to Y$ *be differentiable on the open subset* $\Omega \subseteq X$*. If* $a \in \Omega$ *is a point such that* $f(a) = 0$*, if* $f'(a)$ *is a linear isomorphism, and if there is some* λ *with* $0 < \lambda < 1/2$ *such that*

$$\sup_{k\geq 0} \|A_k - f'(a)\|_{\mathcal{L}(X;Y)} \leq \frac{\lambda}{\|(f'(a))^{-1}\|_{\mathcal{L}(Y;X)}},$$

then there is a closed ball B *of center* a *such that for every* $x_0 \in B$*, the sequence* (x_k) *defined by*

$$x_{k+1} = x_k - A_k^{-1}(f(x_k)), \quad k \geq 0,$$

is entirely contained within B *and converges to* a*, which is the only zero of* f *in* B*. Furthermore, the convergence is geometric, which means that*

$$\|x_k - a\| \leq \beta^k \|x_0 - a\|,$$

for some $\beta < 1$*.*

A proof of Theorem 21.2 can be also found in Ciarlet [30] (Section 7.5).

For the sake of completeness, we state a version of the Newton–Kantorovich theorem, which corresponds to the case where $A_k(x) = f'(x)$. In this instance, a stronger result can be obtained especially regarding upper bounds, and we state a version due to Gragg and Tapia which appears in Problem 7.5-4 of Ciarlet [30].

Theorem 21.3. *(Newton–Kantorovich) Let* X *be a Banach space, and let* $f\colon \Omega \to Y$ *be differentiable on the open subset* $\Omega \subseteq X$*. Assume that there exist three positive constants* λ, μ, ν *and a point* $x_0 \in \Omega$ *such that*

$$0 < \lambda\mu\nu \leq \frac{1}{2},$$

and if we let

$$\begin{aligned}\rho^- &= \frac{1-\sqrt{1-2\lambda\mu\nu}}{\mu\nu}\\ \rho^+ &= \frac{1+\sqrt{1-2\lambda\mu\nu}}{\mu\nu}\\ B &= \{x \in X \mid \|x - x_0\| < \rho^-\}\\ \Omega^+ &= \{x \in \Omega \mid \|x - x_0\| < \rho^+\},\end{aligned}$$

then $\overline{B} \subseteq \Omega$, $f'(x_0)$ *is an isomorphism of* $\mathcal{L}(X;Y)$, *and*

$$\begin{aligned}\left\|(f'(x_0))^{-1}\right\| &\leq \mu,\\ \left\|(f'(x_0))^{-1} f(x_0)\right\| &\leq \lambda,\\ \sup_{x,y\in\Omega^+} \|f'(x) - f'(y)\| &\leq \nu \|x - y\|.\end{aligned}$$

Then, $f'(x)$ *is isomorphism of* $\mathcal{L}(X;Y)$ *for all* $x \in B$, *and the sequence defined by*

$$x_{k+1} = x_k - (f'(x_k))^{-1}(f(x_k)), \quad k \geq 0$$

is entirely contained within the ball B *and converges to a zero* a *of* f *which is the only zero of* f *in* Ω^+. *Finally, if we write* $\theta = \rho^-/\rho^+$, *then we have the following bounds:*

$$\begin{aligned}\|x_k - a\| &\leq \frac{2\sqrt{1-2\lambda\mu\nu}}{\lambda\mu\nu}\frac{\theta^{2k}}{1-\theta^{2k}}\|x_1 - x_0\| && \text{if } \lambda\mu\nu < \frac{1}{2}\\ \|x_k - a\| &\leq \frac{\|x_1 - x_0\|}{2^{k-1}} && \text{if } \lambda\mu\nu = \frac{1}{2},\end{aligned}$$

and

$$\frac{2\|x_{k+1} - x_k\|}{1+\sqrt{(1+4\theta^{2k}(1+\theta^{2k})^{-2})}} \leq \|x_k - a\| \leq \theta^{2k-1}\|x_k - x_{k-1}\|.$$

We can now specialize Theorems 21.1 and 21.2 to the search of zeros of the derivative $f'\colon \Omega \to E'$, of a function $f\colon \Omega \to \mathbb{R}$, with $\Omega \subseteq E$. The second derivative J'' of J is a continuous bilinear form $J''\colon E \times E \to \mathbb{R}$, but is is convenient to view it as a linear map in $\mathcal{L}(E, E')$; the continuous linear form $J''(u)$ is given by $J''(u)(v) = J''(u, v)$. In our next theorem, we assume that the $A_k(x)$ are isomorphisms in $\mathcal{L}(E, E')$.

Theorem 21.4. *Let* E *be a Banach space, let* $J\colon \Omega \to \mathbb{R}$ *be twice differentiable on the open subset* $\Omega \subseteq E$, *and assume that there are constants* $r, M, \beta > 0$ *such that if we let*

$$B = \{x \in E \mid \|x - x_0\| \leq r\} \subseteq \Omega,$$

then

(1)

$$\sup_{k\geq 0}\sup_{x\in B}\left\|A_k^{-1}(x)\right\|_{\mathcal{L}(E';E)}\leq M,$$

(2) $\beta<1$ *and*

$$\sup_{k\geq 0}\sup_{x,x'\in B}\|J''(x)-A_k(x')\|_{\mathcal{L}(E;E')}\leq\frac{\beta}{M}$$

(3)

$$\|J'(x_0)\|\leq\frac{r}{M}(1-\beta).$$

Then, the sequence (x_k) *defined by*

$$x_{k+1}=x_k-A_k^{-1}(x_\ell)(J'(x_k)),\quad 0\leq\ell\leq k$$

is entirely contained within B *and converges to a zero* a *of* J'*, which is the only zero of* J' *in* B*. Furthermore, the convergence is geometric, which means that*

$$\|x_k-a\|\leq\frac{\|x_1-x_0\|}{1-\beta}\beta^k.$$

In the next theorem, we assume that the $A_k(x)$ are isomorphisms in $\mathcal{L}(E,E')$ that are independent of $x\in\Omega$.

Theorem 21.5. *Let* E *be a Banach space, and let* $J\colon\Omega\to\mathbb{R}$ *be twice differentiable on the open subset* $\Omega\subseteq E$*. If* $a\in\Omega$ *is a point such that* $J'(a)=0$*, if* $J''(a)$ *is a linear isomorphism, and if there is some* λ *with* $0<\lambda<1/2$ *such that*

$$\sup_{k\geq 0}\|A_k-J''(a)\|_{\mathcal{L}(E;E')}\leq\frac{\lambda}{\|(J''(a))^{-1}\|_{\mathcal{L}(E';E)}},$$

then there is a closed ball B *of center* a *such that for every* $x_0\in B$*, the sequence* (x_k) *defined by*

$$x_{k+1}=x_k-A_k^{-1}(J'(x_k)),\quad k\geq 0,$$

is entirely contained within B *and converges to* a*, which is the only zero of* J' *in* B*. Furthermore, the convergence is geometric, which means that*

$$\|x_k-a\|\leq\beta^k\|x_0-a\|,$$

for some $\beta<1$*.*

When $E=\mathbb{R}^n$, the Newton method given by Theorem 21.4 yield an itereation step of the form

$$x_{k+1}=x_k-A_k^{-1}(x_\ell)\nabla J(x_k),\quad 0\leq\ell\leq k,$$

where $\nabla J(x_k)$ is the gradient of J at x_k (here, we identify E' with $\mathbb{R}^n$). In particular, Newton's original method picks $A_k = J''$, and the iteration step is of the form

$$x_{k+1} = x_k - (\nabla^2 J(x_k))^{-1} \nabla J(x_k), \quad k \geq 0,$$

where $\nabla^2 J(x_k)$ is the Hessian of J at x_k.

As remarked in [30] (Section 7.5), generalized Newton methods have a very wide range of applicability. For example, various versions of gradient descent methods can be viewed as instances of Newton methods.

Newton's method also plays an important role in convex optimization, in particular, interior-point methods. A variant of Newton's method dealing with equality constraints has been developed. We refer the reader to Boyd and Vandenberghe [22], Chapters 10 and 11, for a comprehensive exposition of these topics.

21.3 Summary

The main concepts and results of this chapter are listed below:

- Newton's method for functions $f : \mathbb{R} \to \mathbb{R}$.
- Generalized Newton methods.
- The *Newton-Kantorovich* theorem.

Chapter 22

Quadratic Optimization Problems

22.1 Quadratic Optimization: The Positive Definite Case

In this chapter, we consider two classes of quadratic optimization problems that appear frequently in engineering and in computer science (especially in computer vision):

1. Minimizing
$$Q(x) = \frac{1}{2}x^\top Ax - x^\top b$$
over all $x \in \mathbb{R}^n$, or subject to linear or affine constraints.

2. Minimizing
$$Q(x) = \frac{1}{2}x^\top Ax - x^\top b$$
over the unit sphere.

In both cases, A is a symmetric matrix. We also seek necessary and sufficient conditions for f to have a global minimum.

Many problems in physics and engineering can be stated as the minimization of some energy function, with or without constraints. Indeed, it is a fundamental principle of mechanics that nature acts so as to minimize energy. Furthermore, if a physical system is in a stable state of equilibrium, then the energy in that state should be minimal. For example, a small ball placed on top of a sphere is in an unstable equilibrium position. A small motion causes the ball to roll down. On the other hand, a ball placed inside and at the bottom of a sphere is in a stable equilibrium position, because the potential energy is minimal.

The simplest kind of energy function is a quadratic function. Such functions can be conveniently defined in the form
$$Q(x) = x^\top Ax - x^\top b,$$

where A is a symmetric $n \times n$ matrix, and x, b, are vectors in $\mathbb{R}^n$, viewed as column vectors. Actually, for reasons that will be clear shortly, it is preferable to put a factor $\frac{1}{2}$ in front of the quadratic term, so that

$$Q(x) = \frac{1}{2} x^\top A x - x^\top b.$$

The question is, under what conditions (on A) does $Q(x)$ have a global minimum, preferably unique?

We give a complete answer to the above question in two stages:

1. In this section, we show that if A is symmetric positive definite, then $Q(x)$ has a unique global minimum precisely when
$$Ax = b.$$
2. In Section 22.2, we give necessary and sufficient conditions in the general case, in terms of the pseudo-inverse of A.

We begin with the matrix version of Definition 15.2.

Definition 22.1. A symmetric *positive definite matrix* is a matrix whose eigenvalues are strictly positive, and a symmetric *positive semidefinite matrix* is a matrix whose eigenvalues are nonnegative.

Equivalent criteria are given in the following proposition.

Proposition 22.1. *Given any Euclidean space E of dimension n, the following properties hold:*

(1) Every self-adjoint linear map $f : E \to E$ is positive definite iff

$$\langle f(x), x\rangle > 0$$

for all $x \in E$ with $x \neq 0$.

(2) Every self-adjoint linear map $f : E \to E$ is positive semidefinite iff

$$\langle f(x), x\rangle \geq 0$$

for all $x \in E$.

Proof. (1) First, assume that f is positive definite. Recall that every self-adjoint linear map has an orthonormal basis $(e_1, \ldots, e_n)$ of eigenvectors, and let $\lambda_1, \ldots, \lambda_n$ be the corresponding eigenvalues. With respect to this basis, for every $x = x_1 e_1 + \cdots + x_n e_n \neq 0$, we have

$$\langle f(x), x\rangle = \Big\langle f\Big(\sum_{i=1}^n x_i e_i\Big), \sum_{i=1}^n x_i e_i\Big\rangle = \Big\langle \sum_{i=1}^n \lambda_i x_i e_i, \sum_{i=1}^n x_i e_i\Big\rangle = \sum_{i=1}^n \lambda_i x_i^2,$$

which is strictly positive, since $\lambda_i > 0$ for $i = 1, \ldots, n$, and $x_i^2 > 0$ for some i, since $x \neq 0$.

Conversely, assume that

$$\langle f(x), x\rangle > 0$$

for all $x \neq 0$. Then for $x = e_i$, we get

$$\langle f(e_i), e_i\rangle = \langle \lambda_i e_i, e_i\rangle = \lambda_i,$$

and thus $\lambda_i > 0$ for all $i = 1, \ldots, n$.

(2) As in (1), we have

$$\langle f(x), x\rangle = \sum_{i=1}^{n} \lambda_i x_i^2,$$

and since $\lambda_i \geq 0$ for $i = 1, \ldots, n$ because f is positive semidefinite, we have $\langle f(x), x\rangle \geq 0$, as claimed. The converse is as in (1) except that we get only $\lambda_i \geq 0$ since $\langle f(e_i), e_i\rangle \geq 0$. □

Some special notation is customary (especially in the field of convex optimization) to express that a symmetric matrix is positive definite or positive semidefinite.

Definition 22.2. Given any $n \times n$ symmetric matrix A we write $A \succeq 0$ if A is positive semidefinite and we write $A \succ 0$ if A is positive definite.

It should be noted that we can define the relation

$$A \succeq B$$

between any two $n \times n$ matrices (symmetric or not) iff $A - B$ is symmetric positive semidefinite. It is easy to check that this relation is actually a partial order on matrices, called the *positive semidefinite cone ordering*; for details, see Boyd and Vandenberghe [22], Section 2.4.

If A is symmetric positive definite, it is easily checked that A^{-1} is also symmetric positive definite. Also, if C is a symmetric positive definite $m \times m$ matrix and A is an $m \times n$ matrix of rank n (and so $m \geq n$ and the map $x \mapsto Ax$ is surjective onto $\mathbb{R}^m$), then $A^\top CA$ is symmetric positive definite.

We can now prove that

$$Q(x) = \frac{1}{2} x^\top Ax - x^\top b$$

has a global minimum when A is symmetric positive definite.

Proposition 22.2. *Given a quadratic function*

$$Q(x) = \frac{1}{2} x^\top Ax - x^\top b,$$

if A is symmetric positive definite, then $Q(x)$ has a unique global minimum for the solution of the linear system $Ax = b$. The minimum value of $Q(x)$ is

$$Q(A^{-1}b) = -\frac{1}{2} b^\top A^{-1} b.$$

Proof. Since A is positive definite, it is invertible, since its eigenvalues are all strictly positive. Let $x = A^{-1}b$, and compute $Q(y) - Q(x)$ for any $y \in \mathbb{R}^n$. Since $Ax = b$, we get

$$\begin{aligned} Q(y) - Q(x) &= \frac{1}{2}y^\top Ay - y^\top b - \frac{1}{2}x^\top Ax + x^\top b \\ &= \frac{1}{2}y^\top Ay - y^\top Ax + \frac{1}{2}x^\top Ax \\ &= \frac{1}{2}(y-x)^\top A(y-x). \end{aligned}$$

Since A is positive definite, the last expression is nonnegative, and thus

$$Q(y) \geq Q(x)$$

for all $y \in \mathbb{R}^n$, which proves that $x = A^{-1}b$ is a global minimum of $Q(x)$. A simple computation yields

$$Q(A^{-1}b) = -\frac{1}{2}b^\top A^{-1}b.$$

□

Remarks:

(1) The quadratic function $Q(x)$ is also given by

$$Q(x) = \frac{1}{2}x^\top Ax - b^\top x,$$

but the definition using $x^\top b$ is more convenient for the proof of Proposition 22.2.

(2) If $Q(x)$ contains a constant term $c \in \mathbb{R}$, so that

$$Q(x) = \frac{1}{2}x^\top Ax - x^\top b + c,$$

the proof of Proposition 22.2 still shows that $Q(x)$ has a unique global minimum for $x = A^{-1}b$, but the minimal value is

$$Q(A^{-1}b) = -\frac{1}{2}b^\top A^{-1}b + c.$$

Thus, when the energy function $Q(x)$ of a system is given by a quadratic function

$$Q(x) = \frac{1}{2}x^\top Ax - x^\top b,$$

where A is symmetric positive definite, finding the global minimum of $Q(x)$ is equivalent to solving the linear system $Ax = b$. Sometimes, it is useful to recast a linear problem $Ax = b$

as a variational problem (finding the minimum of some energy function). However, very often, a minimization problem comes with extra constraints that must be satisfied for all admissible solutions. For instance, we may want to minimize the quadratic function

$$Q(x_1, x_2) = \frac{1}{2}\left(x_1^2 + x_2^2\right)$$

subject to the constraint

$$2x_1 - x_2 = 5.$$

The solution for which $Q(x_1, x_2)$ is minimum is no longer $(x_1, x_2) = (0, 0)$, but instead, $(x_1, x_2) = (2, -1)$, as will be shown later.

Geometrically, the graph of the function defined by $z = Q(x_1, x_2)$ in $\mathbb{R}^3$ is a paraboloid of revolution P with axis of revolution Oz. The constraint

$$2x_1 - x_2 = 5$$

corresponds to the vertical plane H parallel to the z-axis and containing the line of equation $2x_1 - x_2 = 5$ in the xy-plane. Thus, the constrained minimum of Q is located on the parabola that is the intersection of the paraboloid P with the plane H.

A nice way to solve constrained minimization problems of the above kind is to use the method of *Lagrange multipliers* discussed in Section 20.1. But first, let us define precisely what kind of minimization problems we intend to solve.

Definition 22.3. The *quadratic constrained minimization problem* consists in minimizing a quadratic function

$$Q(x) = \frac{1}{2}x^\top A^{-1}x - b^\top x$$

subject to the linear constraints

$$B^\top x = f,$$

where A^{-1} is an $m \times m$ symmetric positive definite matrix, B is an $m \times n$ matrix of rank n (so that $m \geq n$), and where $b, x \in \mathbb{R}^m$ (viewed as column vectors), and $f \in \mathbb{R}^n$ (viewed as a column vector).

The reason for using A^{-1} instead of A is that the constrained minimization problem has an interpretation as a set of equilibrium equations in which the matrix that arises naturally is A (see Strang [101]). Since A and A^{-1} are both symmetric positive definite, this doesn't make any difference, but it seems preferable to stick to Strang's notation.

As explained in Section 20.1, the method of Lagrange multipliers consists in incorporating the n constraints $B^\top x = f$ into the quadratic function $Q(x)$, by introducing extra variables $\lambda = (\lambda_1, \ldots, \lambda_n)$ called *Lagrange multipliers*, one for each constraint. We form the *Lagrangian*

$$L(x, \lambda) = Q(x) + \lambda^\top(B^\top x - f) = \frac{1}{2}x^\top A^{-1}x - (b - B\lambda)^\top x - \lambda^\top f.$$

We know from Theorem 20.3 that a necessary condition for our constrained optimization problem to have a solution is that $\nabla L(x, \lambda) = 0$. Since

$$\frac{\partial L}{\partial x}(x, \lambda) = A^{-1}x - (b - B\lambda)$$
$$\frac{\partial L}{\partial \lambda}(x, \lambda) = B^\top x - f,$$

we obtain the system of linear equations

$$A^{-1}x + B\lambda = b,$$
$$B^\top x = f,$$

which can be written in matrix form as

$$\begin{pmatrix} A^{-1} & B \\ B^\top & 0 \end{pmatrix} \begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ f \end{pmatrix}.$$

We shall prove in Proposition 22.3 below that our constrained minimization problem has a unique solution actually given by the above system.

Note that the matrix of this system is symmetric. We solve it as follows. Eliminating x from the first equation

$$A^{-1}x + B\lambda = b,$$

we get

$$x = A(b - B\lambda),$$

and substituting into the second equation, we get

$$B^\top A(b - B\lambda) = f,$$

that is,

$$B^\top AB\lambda = B^\top Ab - f.$$

However, by a previous remark, since A is symmetric positive definite and the columns of B are linearly independent, $B^\top AB$ is symmetric positive definite, and thus invertible. Thus we obtain the solution

$$\lambda = (B^\top AB)^{-1}(B^\top Ab - f), \qquad x = A(b - B\lambda).$$

Note that this way of solving the system requires solving for the Lagrange multipliers first.

Letting $e = b - B\lambda$, we also note that the system

$$\begin{pmatrix} A^{-1} & B \\ B^\top & 0 \end{pmatrix} \begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ f \end{pmatrix}$$

is equivalent to the system

$$\begin{aligned} e &= b - B\lambda, \\ x &= Ae, \\ B^\top x &= f. \end{aligned}$$

The latter system is called the *equilibrium equations* by Strang [101]. Indeed, Strang shows that the equilibrium equations of many physical systems can be put in the above form. This includes spring-mass systems, electrical networks, and trusses, which are structures built from elastic bars. In each case, x, e, b, A, λ, f, and $K = B^\top AB$ have a physical interpretation. The matrix $K = B^\top AB$ is usually called the *stiffness matrix*. Again, the reader is referred to Strang [101].

In order to prove that our constrained minimization problem has a unique solution, we proceed to prove that the constrained minimization of $Q(x)$ subject to $B^\top x = f$ is equivalent to the unconstrained maximization of another function $-G(\lambda)$. We get $G(\lambda)$ by minimizing the Lagrangian $L(x, \lambda)$ treated as a function of x alone. The function $-G(\lambda)$ is the *dual function* of the Lagrangian $L(x, \lambda)$. Here we are encountering a special case of the notion of dual function defined in Section 30.5.

Since A^{-1} is symmetric positive definite and

$$L(x, \lambda) = \frac{1}{2} x^\top A^{-1} x - (b - B\lambda)^\top x - \lambda^\top f,$$

by Proposition 22.2 the global minimum (with respect to x) of $L(x, \lambda)$ is obtained for the solution x of

$$A^{-1}x = b - B\lambda,$$

that is, when

$$x = A(b - B\lambda),$$

and the minimum of $L(x, \lambda)$ is

$$\min_x L(x, \lambda) = -\frac{1}{2}(B\lambda - b)^\top A(B\lambda - b) - \lambda^\top f.$$

Letting

$$G(\lambda) = \frac{1}{2}(B\lambda - b)^\top A(B\lambda - b) + \lambda^\top f,$$

we will show in Proposition 22.3 that the solution of the constrained minimization of $Q(x)$ subject to $B^\top x = f$ is equivalent to the unconstrained maximization of $-G(\lambda)$. This is a special case of the duality discussed in Section 30.5.

Of course, since we minimized $L(x, \lambda)$ with respect to x, we have

$$L(x, \lambda) \geq -G(\lambda)$$

for all x and all λ. However, when the constraint $B^\top x = f$ holds, $L(x, \lambda) = Q(x)$, and thus for any admissible x, which means that $B^\top x = f$, we have

$$\min_x Q(x) \geq \max_\lambda -G(\lambda).$$

In order to prove that the unique minimum of the constrained problem $Q(x)$ subject to $B^\top x = f$ is the unique maximum of $-G(\lambda)$, we compute $Q(x) + G(\lambda)$.

Proposition 22.3. *The quadratic constrained minimization problem of Definition 22.3 has a unique solution (x, λ) given by the system*

$$\begin{pmatrix} A^{-1} & B \\ B^\top & 0 \end{pmatrix} \begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ f \end{pmatrix}.$$

Furthermore, the component λ of the above solution is the unique value for which $-G(\lambda)$ is maximum.

Proof. As we suggested earlier, let us compute $Q(x) + G(\lambda)$, assuming that the constraint $B^\top x = f$ holds. Eliminating f, since $b^\top x = x^\top b$ and $\lambda^\top B^\top x = x^\top B\lambda$, we get

$$\begin{aligned} Q(x) + G(\lambda) &= \frac{1}{2} x^\top A^{-1} x - b^\top x + \frac{1}{2}(B\lambda - b)^\top A(B\lambda - b) + \lambda^\top f \\ &= \frac{1}{2}(A^{-1}x + B\lambda - b)^\top A(A^{-1}x + B\lambda - b). \end{aligned}$$

Since A is positive definite, the last expression is nonnegative. In fact, it is null iff

$$A^{-1}x + B\lambda - b = 0,$$

that is,

$$A^{-1}x + B\lambda = b.$$

But then the unique constrained minimum of $Q(x)$ subject to $B^\top x = f$ is equal to the unique maximum of $-G(\lambda)$ exactly when $B^\top x = f$ and $A^{-1}x + B\lambda = b$, which proves the proposition. □

We can confirm that the maximum of $-G(\lambda)$, or equivalently the minimum of

$$G(\lambda) = \frac{1}{2}(B\lambda - b)^\top A(B\lambda - b) + \lambda^\top f,$$

corresponds to value of λ obtained by solving the system

$$\begin{pmatrix} A^{-1} & B \\ B^\top & 0 \end{pmatrix} \begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ f \end{pmatrix}.$$

Indeed, since

$$G(\lambda) = \frac{1}{2}\lambda^\top B^\top AB\lambda - \lambda^\top B^\top Ab + \lambda^\top f + \frac{1}{2} b^\top b,$$

Since $B^\top AB$ is symmetric positive defintite, by Proposition 22.2, the global minimum of $G(\lambda)$ is obtained when

$$B^\top AB\lambda - B^\top Ab + f = 0,$$

that is, $\lambda = (B^\top AB)^{-1}(B^\top Ab - f)$, as we found earlier.

Remarks:

(1) There is a form of duality going on in this situation. The constrained minimization of $Q(x)$ subject to $B^\top x = f$ is called the *primal problem*, and the unconstrained maximization of $-G(\lambda)$ is called the *dual problem*. Duality is the fact stated slightly loosely as

$$\min_x Q(x) = \max_\lambda -G(\lambda).$$

A general treatment of duality in constrained minimization problems is given in Section 30.5.

Recalling that $e = b - B\lambda$, since

$$G(\lambda) = \frac{1}{2}(B\lambda - b)^\top A(B\lambda - b) + \lambda^\top f,$$

we can also write

$$G(\lambda) = \frac{1}{2}e^\top Ae + \lambda^\top f.$$

This expression often represents the total potential energy of a system. Again, the optimal solution is the one that minimizes the potential energy (and thus maximizes $-G(\lambda)$).

(2) It is immediately verified that the equations of Proposition 22.3 are equivalent to the equations stating that the partial derivatives of the Lagrangian $L(x, \lambda)$ are null:

$$\frac{\partial L}{\partial x_i} = 0, \quad i = 1, \ldots, m,$$
$$\frac{\partial L}{\partial \lambda_j} = 0, \quad j = 1, \ldots, n.$$

Thus, the constrained minimum of $Q(x)$ subject to $B^\top x = f$ is an extremum of the Lagrangian $L(x, \lambda)$. As we showed in Proposition 22.3, this extremum corresponds to simultaneously minimizing $L(x, \lambda)$ with respect to x and maximizing $L(x, \lambda)$ with respect to λ. Geometrically, such a point is a *saddle point* for $L(x, \lambda)$. Saddle points are discussed in Section 30.5.

(3) The Lagrange multipliers sometimes have a natural physical meaning. For example, in the spring-mass system they correspond to node displacements. In some general sense, Lagrange multipliers are correction terms needed to satisfy equilibrium equations and the price paid for the constraints. For more details, see Strang [101].

Going back to the constrained minimization of $Q(x_1, x_2) = \frac{1}{2}(x_1^2 + x_2^2)$ subject to

$$2x_1 - x_2 = 5,$$

the Lagrangian is

$$L(x_1, x_2, \lambda) = \frac{1}{2}\big(x_1^2 + x_2^2\big) + \lambda(2x_1 - x_2 - 5),$$

and the equations stating that the Lagrangian has a saddle point are

$$\begin{aligned} x_1 + 2\lambda &= 0,\\ x_2 - \lambda &= 0,\\ 2x_1 - x_2 - 5 &= 0. \end{aligned}$$

We obtain the solution $(x_1, x_2, \lambda) = (2, -1, -1)$.

The use of Lagrange multipliers in optimization and variational problems is discussed extensively in Chapter 30.

Least squares methods and Lagrange multipliers are used to tackle many problems in computer graphics and computer vision; see Trucco and Verri [106], Metaxas [73], Jain, Katsuri, and Schunck [57], Faugeras [40], and Foley, van Dam, Feiner, and Hughes [41].

22.2 Quadratic Optimization: The General Case

In this section we complete the study initiated in Section 22.1 and give necessary and sufficient conditions for the quadratic function $\frac{1}{2}x^\top Ax - x^\top b$ to have a global minimum. We begin with the following simple fact:

Proposition 22.4. *If A is an invertible symmetric matrix, then the function*

$$f(x) = \frac{1}{2}x^\top Ax - x^\top b$$

has a minimum value iff $A \succeq 0$, in which case this optimal value is obtained for a unique value of x, namely $x^* = A^{-1}b$, and with

$$f(A^{-1}b) = -\frac{1}{2}b^\top A^{-1}b.$$

Proof. Observe that

$$\frac{1}{2}(x - A^{-1}b)^\top A(x - A^{-1}b) = \frac{1}{2}x^\top Ax - x^\top b + \frac{1}{2}b^\top A^{-1}b.$$

Thus,

$$f(x) = \frac{1}{2}x^\top Ax - x^\top b = \frac{1}{2}(x - A^{-1}b)^\top A(x - A^{-1}b) - \frac{1}{2}b^\top A^{-1}b.$$

If A has some negative eigenvalue, say $-\lambda$ (with $\lambda > 0$), if we pick any eigenvector u of A associated with λ, then for any $\alpha \in \mathbb{R}$ with $\alpha \neq 0$, if we let $x = \alpha u + A^{-1}b$, then since $Au = -\lambda u$, we get

$$\begin{aligned} f(x) &= \frac{1}{2}(x - A^{-1}b)^\top A(x - A^{-1}b) - \frac{1}{2}b^\top A^{-1}b \\ &= \frac{1}{2}\alpha u^\top A\alpha u - \frac{1}{2}b^\top A^{-1}b \\ &= -\frac{1}{2}\alpha^2\lambda \left\|u\right\|_2^2 - \frac{1}{2}b^\top A^{-1}b, \end{aligned}$$

and since α can be made as large as we want and $\lambda > 0$, we see that f has no minimum. Consequently, in order for f to have a minimum, we must have $A \succeq 0$. If $A \succeq 0$, since A is invertible, it is positive definite, so $(x - A^{-1}b)^\top A(x - A^{-1}b) > 0$ iff $x - A^{-1}b \neq 0$, and it is clear that the minimum value of f is achieved when $x - A^{-1}b = 0$, that is, $x = A^{-1}b$. □

Let us now consider the case of an arbitrary symmetric matrix A.

Proposition 22.5. *If A is a $n \times n$ symmetric matrix, then the function*

$$f(x) = \frac{1}{2}x^\top Ax - x^\top b$$

has a minimum value iff $A \succeq 0$ and $(I - AA^+)b = 0$, in which case this minimum value is

$$p^* = -\frac{1}{2}b^\top A^+ b.$$

Furthermore, if A is diagonlized as $A = U^\top \Sigma U$ (with U orthogonal), then the optimal value is achieved by all $x \in \mathbb{R}^n$ of the form

$$x = A^+ b + U^\top \begin{pmatrix} 0 \\ z \end{pmatrix},$$

for any $z \in \mathbb{R}^{n-r}$, where r is the rank of A.

Proof. The case that A is invertible is taken care of by Proposition 22.4, so we may assume that A is singular. If A has rank $r < n$, then we can diagonalize A as

$$A = U^\top \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} U,$$

where U is an orthogonal matrix and where Σ_r is an $r \times r$ diagonal invertible matrix. Then we have

$$\begin{aligned} f(x) &= \frac{1}{2}x^\top U^\top \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} Ux - x^\top U^\top Ub \\ &= \frac{1}{2}(Ux)^\top \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} Ux - (Ux)^\top Ub. \end{aligned}$$

If we write

$$Ux = \begin{pmatrix} y \\ z \end{pmatrix} \quad \text{and} \quad Ub = \begin{pmatrix} c \\ d \end{pmatrix},$$

with $y, c \in \mathbb{R}^r$ and $z, d \in \mathbb{R}^{n-r}$, we get

$$\begin{aligned} f(x) &= \frac{1}{2}(Ux)^\top \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} Ux - (Ux)^\top Ub \\ &= \frac{1}{2}(y^\top \; z^\top) \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} - (y^\top \; z^\top) \begin{pmatrix} c \\ d \end{pmatrix} \\ &= \frac{1}{2} y^\top \Sigma_r y - y^\top c - z^\top d. \end{aligned}$$

For $y = 0$, we get

$$f(x) = -z^\top d,$$

so if $d \neq 0$, the function f has no minimum. Therefore, if f has a minimum, then $d = 0$. However, $d = 0$ means that

$$Ub = \begin{pmatrix} c \\ 0 \end{pmatrix},$$

and we know from Proposition 16.5 that b is in the range of A (here, U is $V^\top$), which is equivalent to $(I - AA^+)b = 0$. If $d = 0$, then

$$f(x) = \frac{1}{2} y^\top \Sigma_r y - y^\top c,$$

and since Σ_r is invertible, by Proposition 22.4, the function f has a minimum iff $\Sigma_r \succeq 0$, which is equivalent to $A \succeq 0$.

Therefore, we have proved that if f has a minimum, then $(I - AA^+)b = 0$ and $A \succeq 0$. Conversely, if $(I - AA^+)b = 0$ and $A \succeq 0$, what we just did proves that f does have a minimum.

When the above conditions hold, since

$$A = U^\top \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} U$$

is positive semidefinite, the pseudo-inverse A^+ of A is given by

$$A^+ = U^\top \begin{pmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} U,$$

and by Proposition 22.4 the minimum is achieved if $y = \Sigma_r^{-1} c$, $z = 0$ and $d = 0$, that is, for x^* given by

$$Ux^* = \begin{pmatrix} \Sigma_r^{-1} c \\ 0 \end{pmatrix} \quad \text{and} \quad Ub = \begin{pmatrix} c \\ 0 \end{pmatrix},$$

from which we deduce that

$$x^* = U^\top \begin{pmatrix} \Sigma_r^{-1} c \\ 0 \end{pmatrix} = U^\top \begin{pmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} c \\ 0 \end{pmatrix} = U^\top \begin{pmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} Ub = A^+ b$$

and the minimum value of f is

$$f(x^*) = -\frac{1}{2} b^\top A^+ b.$$

For any $x \in \mathbb{R}^n$ of the form

$$x = A^+ b + U^\top \begin{pmatrix} 0 \\ z \end{pmatrix},$$

for any $z \in \mathbb{R}^{n-r}$, we have

$$\begin{aligned} f(x) &= \frac{1}{2} \left(A^+ b + U^\top \begin{pmatrix} 0 \\ z \end{pmatrix} \right)^\top A \left(A^+ b + U^\top \begin{pmatrix} 0 \\ z \end{pmatrix} \right) - \left(A^+ b + U^\top \begin{pmatrix} 0 \\ z \end{pmatrix} \right)^\top b \\ &= \frac{1}{2} (A^+ b)^\top A A^+ b + (0\ z^\top) U A A^+ b + \frac{1}{2} (0\ z^\top) U A U^\top \begin{pmatrix} 0 \\ z \end{pmatrix} - (A^+ b)^\top b - (0\ z^\top) U b \\ &= -\frac{1}{2} b^\top A^+ b + (0\ z^\top) U A A^+ b + \frac{1}{2} (0\ z^\top) U A U^\top \begin{pmatrix} 0 \\ z \end{pmatrix} - (0\ z^\top) U b. \end{aligned}$$

We have

$$\begin{aligned} (0\ z^\top) U A A^+ b &= (0\ z^\top) U U^\top \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} U U^\top \begin{pmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} U b \\ &= (0\ z^\top) \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} U b = 0, \end{aligned}$$

$$\begin{aligned} (0\ z^\top) U A U^\top \begin{pmatrix} 0 \\ z \end{pmatrix} &= (0\ z^\top) U U^\top \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} U U^\top \begin{pmatrix} 0 \\ z \end{pmatrix} \\ &= (0\ z^\top) \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ z \end{pmatrix} = 0, \end{aligned}$$

and

$$(0\ z^\top) U b = (0\ z^\top) \begin{pmatrix} c \\ 0 \end{pmatrix} = 0,$$

because $(I - AA^+)b = 0$, that is,

$$\begin{aligned} \left(\begin{pmatrix} I_r & 0 \\ 0 & I_{n-r} \end{pmatrix} - U^\top \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} U U^\top \begin{pmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} U \right) b &= \left(\begin{pmatrix} I_r & 0 \\ 0 & I_{n-r} \end{pmatrix} - U^\top \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} U \right) b \\ &= U^\top \begin{pmatrix} 0 & 0 \\ 0 & I_{n-r} \end{pmatrix} U b = 0, \end{aligned}$$

so if

$$Ub = \begin{pmatrix} c \\ d \end{pmatrix},$$

then $d = 0$. Therefore, $f(x) = -\frac{1}{2} b^\top A^+ b$. □

The problem of minimizing the function

$$f(x) = \frac{1}{2}x^\top Ax - x^\top b$$

in the case where we add either linear constraints of the form $C^\top x = 0$ or affine constraints of the form $C^\top x = t$ (where $t \in \mathbb{R}^m$ and $t \neq 0$) where C is an $n \times m$ matrix can be reduced to the unconstrained case using a QR-decomposition of C. Let us show how to do this for linear constraints of the form $C^\top x = 0$.

If we use a QR decomposition of C, by permuting the columns of C to make sure that the first r columns of C are linearly independent (where $r = \text{rank}(C)$), we may assume that

$$C = Q^\top \begin{pmatrix} R & S \\ 0 & 0 \end{pmatrix} \Pi,$$

where Q is an $n \times n$ orthogonal matrix, R is an $r \times r$ invertible upper triangular matrix, S is an $r \times (m-r)$ matrix, and Π is a permutation matrix (C has rank r). Then if we let

$$x = Q^\top \begin{pmatrix} y \\ z \end{pmatrix},$$

where $y \in \mathbb{R}^r$ and $z \in \mathbb{R}^{n-r}$, then $C^\top x = 0$ becomes

$$C^\top x = \Pi^\top \begin{pmatrix} R^\top & 0 \\ S^\top & 0 \end{pmatrix} Qx = \Pi^\top \begin{pmatrix} R^\top & 0 \\ S^\top & 0 \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} = 0,$$

which implies $y = 0$, and every solution of $C^\top x = 0$ is of the form

$$x = Q^\top \begin{pmatrix} 0 \\ z \end{pmatrix}.$$

Our original problem becomes

$$\begin{array}{ll} \text{minimize} & \dfrac{1}{2}(y^\top\ z^\top)QAQ^\top \begin{pmatrix} y \\ z \end{pmatrix} + (y^\top\ z^\top)Qb \\ \text{subject to} & y = 0,\ y \in \mathbb{R}^r,\ z \in \mathbb{R}^{n-r}. \end{array}$$

Thus, the constraint $C^\top x = 0$ has been simplifed to $y = 0$, and if we write

$$QAQ^\top = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix},$$

where G_{11} is an $r \times r$ matrix and G_{22} is an $(n-r) \times (n-r)$ matrix, and

$$Qb = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, \quad b_1 \in \mathbb{R}^r,\ b_2 \in \mathbb{R}^{n-r},$$

our problem becomes

$$\text{minimize } \frac{1}{2}z^\top G_{22}z + z^\top b_2, \quad z \in \mathbb{R}^{n-r},$$

the problem solved in Proposition 22.5.

Constraints of the form $C^\top x = t$ (where $t \neq 0$) can be handled in a similar fashion. In this case, we may assume that C is an $n \times m$ matrix with full rank (so that $m \leq n$) and $t \in \mathbb{R}^m$. Then we use a QR-decomposition of the form

$$C = P\begin{pmatrix} R \\ 0 \end{pmatrix},$$

where P is an orthogonal $n \times n$ matrix and R is an $m \times m$ invertible upper triangular matrix. If we write

$$x = P\begin{pmatrix} y \\ z \end{pmatrix},$$

where $y \in \mathbb{R}^m$ and $z \in \mathbb{R}^{n-m}$, the equation $C^\top x = t$ becomes

$$(R^\top\ 0)P^\top x = t,$$

that is,

$$(R^\top\ 0)\begin{pmatrix} y \\ z \end{pmatrix} = t,$$

which yields

$$R^\top y = t.$$

Since R is invertible, we get $y = (R^\top)^{-1}t$, and then it is easy to see that our original problem reduces to an unconstrained problem in terms of the matrix $P^\top AP$; the details are left as an exercise.

22.3 Maximizing a Quadratic Function on the Unit Sphere

In this section we discuss various quadratic optimization problems mostly arising from computer vision (image segmentation and contour grouping). These problems can be reduced to the following basic optimization problem: Given an $n \times n$ real symmetric matrix A

$$\begin{array}{ll} \text{maximize} & x^\top Ax \\ \text{subject to} & x^\top x = 1,\ x \in \mathbb{R}^n. \end{array}$$

In view of Proposition 16.10, the maximum value of $x^\top Ax$ on the unit sphere is equal to the largest eigenvalue λ_1 of the matrix A, and it is achieved for any unit eigenvector u_1 associated with λ_1.

A variant of the above problem often encountered in computer vision consists in minimizing $x^\top Ax$ on the ellipsoid given by an equation of the form

$$x^\top Bx = 1,$$

where B is a symmetric positive definite matrix. Since B is positive definite, it can be diagonalized as

$$B = QDQ^\top,$$

where Q is an orthogonal matrix and D is a diagonal matrix,

$$D = \mathrm{diag}(d_1, \ldots, d_n),$$

with $d_i > 0$, for $i = 1, \ldots, n$. If we define the matrices $B^{1/2}$ and $B^{-1/2}$ by

$$B^{1/2} = Q\,\mathrm{diag}\left(\sqrt{d_1}, \ldots, \sqrt{d_n}\right) Q^\top$$

and

$$B^{-1/2} = Q\,\mathrm{diag}\left(1/\sqrt{d_1}, \ldots, 1/\sqrt{d_n}\right) Q^\top,$$

it is clear that these matrices are symmetric, that $B^{-1/2}BB^{-1/2} = I$, and that $B^{1/2}$ and $B^{-1/2}$ are mutual inverses. Then, if we make the change of variable

$$x = B^{-1/2}y,$$

the equation $x^\top Bx = 1$ becomes $y^\top y = 1$, and the optimization problem

$$\begin{array}{ll} \text{maximize} & x^\top Ax \\ \text{subject to} & x^\top Bx = 1,\ x \in \mathbb{R}^n, \end{array}$$

is equivalent to the problem

$$\begin{array}{ll} \text{maximize} & y^\top B^{-1/2}AB^{-1/2}y \\ \text{subject to} & y^\top y = 1,\ y \in \mathbb{R}^n, \end{array}$$

where $y = B^{1/2}x$ and where $B^{-1/2}AB^{-1/2}$ is symmetric.

The complex version of our basic optimization problem in which A is a Hermitian matrix also arises in computer vision. Namely, given an $n \times n$ complex Hermitian matrix A,

$$\begin{array}{ll} \text{maximize} & x^* Ax \\ \text{subject to} & x^* x = 1,\ x \in \mathbb{C}^n. \end{array}$$

Again by Proposition 16.10, the maximum value of x^*Ax on the unit sphere is equal to the largest eigenvalue λ_1 of the matrix A and it is achieved for any unit eigenvector u_1 associated with λ_1.

Remark: It is worth pointing out that if A is a *skew-Hermitian* matrix, that is, if $A^* = -A$, then x^*Ax is *pure imaginary or zero*.

Indeed, since $z = x^*Ax$ is a scalar, we have $z^* = \overline{z}$ (the conjugate of z), so we have

$$\overline{x^*Ax} = (x^*Ax)^* = x^*A^*x = -x^*Ax,$$

so $\overline{x^*Ax} + x^*Ax = 2\mathrm{Re}(x^*Ax) = 0$, which means that x^*Ax is pure imaginary or zero.

In particular, if A is a real matrix and if A is *skew-symmetric*, then

$$x^\top Ax = 0.$$

Thus, for any real matrix (symmetric or not),

$$x^\top Ax = x^\top H(A)x,$$

where $H(A) = (A + A^\top)/2$, the symmetric part of A.

There are situations in which it is necessary to add linear constraints to the problem of maximizing a quadratic function on the sphere. This problem was completely solved by Golub [48] (1973). The problem is the following: Given an $n \times n$ real symmetric matrix A and an $n \times p$ matrix C,

$$\begin{array}{ll} \text{minimize} & x^\top Ax \\ \text{subject to} & x^\top x = 1,\ C^\top x = 0,\ x \in \mathbb{R}^n. \end{array}$$

As in Section 22.2, Golub shows that the linear constraint $C^\top x = 0$ can be eliminated as follows: If we use a QR decomposition of C, by permuting the columns, we may assume that

$$C = Q^\top \begin{pmatrix} R & S \\ 0 & 0 \end{pmatrix} \Pi,$$

where Q is an orthogonal $n \times n$ matrix, R is an $r \times r$ invertible upper triangular matrix, and S is an $r \times (p-r)$ matrix (assuming C has rank r). Then if we let

$$x = Q^\top \begin{pmatrix} y \\ z \end{pmatrix},$$

where $y \in \mathbb{R}^r$ and $z \in \mathbb{R}^{n-r}$, then $C^\top x = 0$ becomes

$$\Pi^\top \begin{pmatrix} R^\top & 0 \\ S^\top & 0 \end{pmatrix} Qx = \Pi^\top \begin{pmatrix} R^\top & 0 \\ S^\top & 0 \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} = 0,$$

which implies $y = 0$, and every solution of $C^\top x = 0$ is of the form

$$x = Q^\top \begin{pmatrix} 0 \\ z \end{pmatrix}.$$

Our original problem becomes

$$\begin{array}{ll} \text{minimize} & (y^\top\ z^\top)QAQ^\top \begin{pmatrix} y \\ z \end{pmatrix} \\ \text{subject to} & z^\top z = 1,\ z \in \mathbb{R}^{n-r}, \\ & y = 0,\ y \in \mathbb{R}^r. \end{array}$$

Thus, the constraint $C^\top x = 0$ has been simplified to $y = 0$, and if we write

$$QAQ^\top = \begin{pmatrix} G_{11} & G_{12} \\ G_{12}^\top & G_{22} \end{pmatrix},$$

our problem becomes

$$\begin{array}{ll} \text{minimize} & z^\top G_{22} z \\ \text{subject to} & z^\top z = 1,\ z \in \mathbb{R}^{n-r}, \end{array}$$

a standard eigenvalue problem.

Remark: There is a way of finding the eigenvalues of G_{22} which does not require the QR-factorization of C. Observe that if we let

$$J = \begin{pmatrix} 0 & 0 \\ 0 & I_{n-r} \end{pmatrix},$$

then

$$JQAQ^\top J = \begin{pmatrix} 0 & 0 \\ 0 & G_{22} \end{pmatrix},$$

and if we set

$$P = Q^\top JQ,$$

then

$$PAP = Q^\top JQAQ^\top JQ.$$

Now, $Q^\top JQAQ^\top JQ$ and $JQAQ^\top J$ have the same eigenvalues, so PAP and $JQAQ^\top J$ also have the same eigenvalues. It follows that the solutions of our optimization problem are among the eigenvalues of $K = PAP$, and at least r of those are 0. Using the fact that CC^+ is the projection onto the range of C, where C^+ is the pseudo-inverse of C, it can also be shown that

$$P = I - CC^+,$$

the projection onto the kernel of $C^\top$. So P can be computed directly in terms of C. In particular, when $n \geq p$ and C has full rank (the columns of C are linearly independent), then we know that $C^+ = (C^\top C)^{-1}C^\top$ and

$$P = I - C(C^\top C)^{-1}C^\top.$$

This fact is used by Cour and Shi [31] and implicitly by Yu and Shi [112].

The problem of adding affine constraints of the form $N^\top x = t$, where $t \neq 0$, also comes up in practice. At first glance, this problem may not seem harder than the linear problem in which $t = 0$, but it is. This problem was extensively studied in a paper by Gander, Golub, and von Matt [46] (1989).

Gander, Golub, and von Matt consider the following problem: Given an $(n+m)\times(n+m)$ real symmetric matrix A (with $n > 0$), an $(n+m)\times m$ matrix N with full rank, and a nonzero vector $t \in \mathbb{R}^m$ with $\left\|(N^\top)^+ t\right\| < 1$ (where $(N^\top)^+$ denotes the pseudo-inverse of $N^\top$),

$$\begin{array}{ll} \text{minimize} & x^\top A x \\ \text{subject to} & x^\top x = 1,\ N^\top x = t,\ x \in \mathbb{R}^{n+m}. \end{array}$$

The condition $\left\|(N^\top)^+ t\right\| < 1$ ensures that the problem has a solution and is not trivial. The authors begin by proving that the affine constraint $N^\top x = t$ can be eliminated. One way to do so is to use a QR decomposition of N. If

$$N = P \begin{pmatrix} R \\ 0 \end{pmatrix},$$

where P is an orthogonal $(n+m)\times(n+m)$ matrix and R is an $m \times m$ invertible upper triangular matrix, then if we observe that

$$\begin{aligned} x^\top A x &= x^\top P P^\top A P P^\top x, \\ N^\top x &= (R^\top\ 0) P^\top x = t, \\ x^\top x &= x^\top P P^\top x = 1, \end{aligned}$$

and if we write

$$P^\top A P = \begin{pmatrix} B & \Gamma^\top \\ \Gamma & C \end{pmatrix},$$

where B is an $m \times m$ symmetric matrix, C is an $n \times n$ symmetric matrix, Γ is an $m \times n$ matrix, and

$$P^\top x = \begin{pmatrix} y \\ z \end{pmatrix},$$

with $y \in \mathbb{R}^m$ and $z \in \mathbb{R}^n$, then we get

$$\begin{aligned} x^\top A x &= y^\top B y + 2 z^\top \Gamma y + z^\top C z, \\ R^\top y &= t, \\ y^\top y + z^\top z &= 1. \end{aligned}$$

Thus

$$y = (R^\top)^{-1} t,$$

and if we write

$$s^2 = 1 - y^\top y > 0$$

and

$$b = \Gamma y,$$

we get the simplified problem

$$\begin{array}{ll} \text{minimize} & z^\top C z + 2 z^\top b \\ \text{subject to} & z^\top z = s^2,\ z \in \mathbb{R}^m. \end{array}$$

Unfortunately, if $b \neq 0$, Proposition 16.10 is no longer applicable. It is still possible to find the minimum of the function $z^\top C z + 2 z^\top b$ using Lagrange multipliers, but such a solution is too involved to be presented here. Interested readers will find a thorough discussion in Gander, Golub, and von Matt [46].

22.4 Summary

The main concepts and results of this chapter are listed below:

- Quadratic optimization problems; *quadratic functions*.
- Symmetric *positive definite* and *positive semidefinite* matrices.
- The *positive semidefinite cone ordering*.
- Existence of a global minimum when A is symmetric positive definite.
- Constrained quadratic optimization problems.
- *Lagrange multipliers*; *Lagrangian*.
- *Primal* and *dual* problems.
- Quadratic optimization problems: the case of a symmetric invertible matrix A.
- Quadratic optimization problems: the general case of a symmetric matrix A.
- Adding linear constraints of the form $C^\top x = 0$.
- Adding affine constraints of the form $C^\top x = t$, with $t \neq 0$.
- Maximizing a quadratic function over the unit sphere.
- Maximizing a quadratic function over an ellipsoid.
- Maximizing a Hermitian quadratic form.
- Adding linear constraints of the form $C^\top x = 0$.
- Adding affine constraints of the form $N^\top x = t$, with $t \neq 0$.

Chapter 23

Schur Complements and Applications

23.1 Schur Complements

Schur complements arise naturally in the process of inverting block matrices of the form

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

and in characterizing when symmetric versions of these matrices are positive definite or positive semidefinite. These characterizations come up in various quadratic optimization problems; see Boyd and Vandenberghe [22], especially Appendix B. In the most general case, pseudo-inverses are also needed.

In this chapter we introduce Schur complements and describe several interesting ways in which they are used. Along the way we provide some details and proofs of some results from Appendix A.5 (especially Section A.5.5) of Boyd and Vandenberghe [22].

Let M be an $n \times n$ matrix written as a 2×2 block matrix

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix},$$

where A is a $p \times p$ matrix and D is a $q \times q$ matrix, with $n = p + q$ (so B is a $p \times q$ matrix and C is a $q \times p$ matrix). We can try to solve the linear system

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} c \\ d \end{pmatrix},$$

that is,

$$\begin{aligned} Ax + By &= c, \\ Cx + Dy &= d, \end{aligned}$$

by mimicking Gaussian elimination. If we assume that D is invertible, then we first solve for y, getting

$$y = D^{-1}(d - Cx),$$

and after substituting this expression for y in the first equation, we get

$$Ax + B(D^{-1}(d - Cx)) = c,$$

that is,

$$(A - BD^{-1}C)x = c - BD^{-1}d.$$

If the matrix $A - BD^{-1}C$ is invertible, then we obtain the solution to our system

$$\begin{aligned} x &= (A - BD^{-1}C)^{-1}(c - BD^{-1}d), \\ y &= D^{-1}(d - C(A - BD^{-1}C)^{-1}(c - BD^{-1}d)). \end{aligned}$$

If A is invertible, then by eliminating x first using the first equation, we obtain analogous formulas involving the matrix $D - CA^{-1}B$. The above formulas suggest that the matrices $A - BD^{-1}C$ and $D - CA^{-1}B$ play a special role and suggest the following definition:

Definition 23.1. Given any $n \times n$ block matrix of the form

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix},$$

where A is a $p \times p$ matrix and D is a $q \times q$ matrix, with $n = p + q$ (so B is a $p \times q$ matrix and C is a $q \times p$ matrix), if D is invertible, then the matrix $A - BD^{-1}C$ is called the *Schur complement* of D in M. If A is invertible, then the matrix $D - CA^{-1}B$ is called the *Schur complement* of A in M.

The above equations written as

$$\begin{aligned} x &= (A - BD^{-1}C)^{-1}c - (A - BD^{-1}C)^{-1}BD^{-1}d, \\ y &= -D^{-1}C(A - BD^{-1}C)^{-1}c \\ &\quad + (D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1})d, \end{aligned}$$

yield a formula for the inverse of M in terms of the Schur complement of D in M, namely

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{pmatrix}.$$

A moment of reflection reveals that

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & 0 \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} \end{pmatrix} \begin{pmatrix} I & -BD^{-1} \\ 0 & I \end{pmatrix},$$

and then

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} I & 0 \\ -D^{-1}C & I \end{pmatrix} \begin{pmatrix} (A - BD^{-1}C)^{-1} & 0 \\ 0 & D^{-1} \end{pmatrix} \begin{pmatrix} I & -BD^{-1} \\ 0 & I \end{pmatrix}.$$

By taking inverses, we obtain the following result.

Proposition 23.1. *If the matrix D is invertibke, then*

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} I & BD^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} A - BD^{-1}C & 0 \\ 0 & D \end{pmatrix} \begin{pmatrix} I & 0 \\ D^{-1}C & I \end{pmatrix}.$$

The above expression can be checked directly and has the advantage of requiring only the invertibility of D.

Remark: If A is invertible, then we can use the Schur complement $D - CA^{-1}B$ of A to obtain the following factorization of M:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} I & 0 \\ CA^{-1} & I \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & D - CA^{-1}B \end{pmatrix} \begin{pmatrix} I & A^{-1}B \\ 0 & I \end{pmatrix}.$$

If $D - CA^{-1}B$ is invertible, we can invert all three matrices above, and we get another formula for the inverse of M in terms of $(D - CA^{-1}B)$, namely,

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{pmatrix}.$$

If A, D and both Schur complements $A - BD^{-1}C$ and $D - CA^{-1}B$ are all invertible, by comparing the two expressions for M^{-1}, we get the (nonobvious) formula

$$(A - BD^{-1}C)^{-1} = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1}.$$

Using this formula, we obtain another expression for the inverse of M involving the Schur complements of A and D (see Horn and Johnson [55]):

Proposition 23.2. *If A, D and both Schur complements $A - BD^{-1}C$ and $D - CA^{-1}B$ are all invertible, then*

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{pmatrix}.$$

If we set $D = I$ and change B to $-B$, we get

$$(A + BC)^{-1} = A^{-1} - A^{-1}B(I - CA^{-1}B)^{-1}CA^{-1},$$

a formula known as the *matrix inversion lemma* (see Boyd and Vandenberghe [22], Appendix C.4, especially C.4.3).

23.2 Symmetric Positive Definite Matrices and Schur Complements

If we assume that our block matrix M is symmetric, so that A, D are symmetric and $C = B^\top$, then we see that M is expressed as

$$M = \begin{pmatrix} A & B \\ B^\top & D \end{pmatrix} = \begin{pmatrix} I & BD^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} A - BD^{-1}B^\top & 0 \\ 0 & D \end{pmatrix} \begin{pmatrix} I & BD^{-1} \\ 0 & I \end{pmatrix}^\top,$$

which shows that M is similar to a block diagonal matrix (obviously, the Schur complement, $A - BD^{-1}B^\top$, is symmetric). As a consequence, we have the following version of "Schur's trick" to check whether $M \succ 0$ for a symmetric matrix.

Proposition 23.3. *For any symmetric matrix M of the form*

$$M = \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix},$$

if C is invertible, then the following properties hold:

(1) $M \succ 0$ iff $C \succ 0$ and $A - BC^{-1}B^\top \succ 0$.

(2) If $C \succ 0$, then $M \succeq 0$ iff $A - BC^{-1}B^\top \succeq 0$.

Proof. (1) Observe that

$$\begin{pmatrix} I & BC^{-1} \\ 0 & I \end{pmatrix}^{-1} = \begin{pmatrix} I & -BC^{-1} \\ 0 & I \end{pmatrix},$$

and we know that for any symmetric matrix T and any invertible matrix N, the matrix T is positive definite ($T \succ 0$) iff $NTN^\top$ (which is obviously symmetric) is positive definite ($NTN^\top \succ 0$). But a block diagonal matrix is positive definite iff each diagonal block is positive definite, which concludes the proof.

(2) This is because for any symmetric matrix T and any invertible matrix N, we have $T \succeq 0$ iff $NTN^\top \succeq 0$. □

Another version of Proposition 23.3 using the Schur complement of A instead of the Schur complement of C also holds. The proof uses the factorization of M using the Schur complement of A (see Section 23.1).

Proposition 23.4. *For any symmetric matrix M of the form*

$$M = \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix},$$

if A is invertible then the following properties hold:

(1) $M \succ 0$ *iff* $A \succ 0$ *and* $C - B^\top A^{-1} B \succ 0$.

(2) If $A \succ 0$, *then* $M \succeq 0$ *iff* $C - B^\top A^{-1} B \succeq 0$.

When C is singular (or A is singular), it is still possible to characterize when a symmetric matrix M as above is positive semidefinite, but this requires using a version of the Schur complement involving the pseudo-inverse of C, namely $A - BC^+B^\top$ (or the Schur complement, $C - B^\top A^+ B$, of A). We use the criterion of Proposition 22.5, which tells us when a quadratic function of the form $\frac{1}{2}x^\top Px - x^\top b$ has a minimum and what this optimum value is (where P is a symmetric matrix).

23.3 Symmetric Positive Semidefinite Matrices and Schur Complements

We now return to our original problem, characterizing when a symmetric matrix

$$M = \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix}$$

is positive semidefinite.

Thus, we want to know when the function

$$f(x, y) = (x^\top, y^\top) \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = x^\top Ax + 2x^\top By + y^\top Cy$$

has a minimum with respect to both x and y. If we hold y constant, Proposition 22.5 implies that $f(x, y)$ has a minimum iff $A \succeq 0$ and $(I - AA^+)By = 0$, and then the minimum value is

$$f(x^*, y) = -y^\top B^\top A^+ By + y^\top Cy = y^\top (C - B^\top A^+ B)y.$$

Since we want $f(x, y)$ to be uniformly bounded from below for all x, y, we must have $(I - AA^+)B = 0$. Now, $f(x^*, y)$ has a minimum iff $C - B^\top A^+ B \succeq 0$. Therefore, we have established that $f(x, y)$ has a minimum over all x, y iff

$$A \succeq 0, \quad (I - AA^+)B = 0, \quad C - B^\top A^+ B \succeq 0.$$

Similar reasoning applies if we first minimize with respect to y and then with respect to x, but this time, the Schur complement $A - BC^+B^\top$ of C is involved. Putting all these facts together, we get our main result:

Theorem 23.5. *Given any symmetric matrix*

$$M = \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix}$$

the following conditions are equivalent:

(1) $M \succeq 0$ *(M is positive semidefinite).*

(2) $A \succeq 0, \quad (I - AA^+)B = 0, \quad C - B^\top A^+ B \succeq 0.$

(3) $C \succeq 0, \quad (I - CC^+)B^\top = 0, \quad A - BC^+B^\top \succeq 0.$

If $M \succeq 0$ as in Theorem 23.5, then it is easy to check that we have the following factorizations (using the fact that $A^+AA^+ = A^+$ and $C^+CC^+ = C^+$):

$$\begin{pmatrix} A & B \\ B^\top & C \end{pmatrix} = \begin{pmatrix} I & BC^+ \\ 0 & I \end{pmatrix} \begin{pmatrix} A - BC^+B^\top & 0 \\ 0 & C \end{pmatrix} \begin{pmatrix} I & 0 \\ C^+B^\top & I \end{pmatrix}$$

and

$$\begin{pmatrix} A & B \\ B^\top & C \end{pmatrix} = \begin{pmatrix} I & 0 \\ B^\top A^+ & I \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & C - B^\top A^+ B \end{pmatrix} \begin{pmatrix} I & A^+B \\ 0 & I \end{pmatrix}.$$

Part III

Linear Optimization

Chapter 24

Convex Sets, Cones, $\mathcal{H}$-Polyhedra

24.1 What is Linear Programming?

What is *linear programming*? At first glance, one might think that this is some style of computer programming. After all, there is imperative programming, functional programming, object-oriented programming *etc.* The term linear programming is somewhat misleading, because it really refers to a method for *planning* with linear constraints, or more accurately, an *optimization method* where both the objective function and the constraints are linear.[1]

Linear programming was created in the late 1940's, one of the key players being George Dantzing, who invented the simplex algorithm. Kantorovitch also did some pioneering work on linear programming as early as 1939. The term *linear programming* has a military connotation because in the early 1950's it was used as a synonym for plans or schedules for training troops, logistical supply, resource allocation, *etc.* Unfortunately the term linear programming is well established and we are stuck with it.

Interestingly, even though originally most applications of linear programming were in the field of economics and industrial engineering, linear programming has become an important tool in theoretical computer science and in the theory of algorithms. Indeed, linear programming is often an effective tool for designing approximation algorithms to solve hard problems (typically NP-hard problems). Linear programming is also the "baby version" of convex programming, a very effective methodology which has received much attention in recent years.

Our goal in these notes is to present the mathematical underpinnings of linear programming, in particular the existence of an optimal solution if a linear program is feasible and bounded, and the duality theorem in linear programming, one of the deepest results in this field. The duality theorem in linear programming also has significant algorithmic implications but we do not discuss this here. We present the simplex algorithm, the dual simplex algorithm, and the primal dual algorithm. We also describe the tableau formalism

[1] Again, we witness another unfortunate abuse of terminology; the constraints are in fact *affine*.

for running the simplex algorithm and its variants. A particularly nice feature of the tableau formalism is that the update of a tableau can be performed using elementary row operations identical to the operations used during the reduction of a matrix to row reduced echelon form (rref). What differs is the criterion for the choice of the pivot.

However, we do not discuss other methods such as the ellipsoid method or interior points methods. For these more algorithmic issues, we refer the reader to standard texts on linear programming. In our opinion, one of the clearest (and among the most concise!) is Matousek and Gardner [72]; Chvatal [29] and Schrijver [88] are classics. Papadimitriou and Steiglitz [79] offers a very crisp presentation in the broader context of combinatorial optimization, and Bertsimas and Tsitsiklis [17] and Vanderbei [109] are very complete.

Linear programming has to do with maximizing a linear cost function $c_1x_1 + \cdots + c_nx_n$ with respect to m "linear" inequalities of the form

$$a_{i1}x_1 + \cdots + a_{in}x_n \leq b_i.$$

These constraints can be put together into an $m \times n$ matrix $A = (a_{ij})$, and written more concisely as

$$Ax \leq b.$$

For technical reasons that will appear clearer later on, it is often preferable to add the nonnegativity constaints $x_i \geq 0$ for $i = 1, \ldots, n$. We write $x \geq 0$. It is easy to show that every linear program is equivalent to another one satisfying the constraints $x \geq 0$, at the expense of adding new variables that are also constrained to be nonnegative. Let $\mathcal{P}(A, b)$ be the set of *feasible solutions* of our linear program given by

$$\mathcal{P}(A, b) = \{x \in \mathbb{R}^n \mid Ax \leq b,\ x \geq 0\}.$$

Then, there are two basic questions:

(1) Is $\mathcal{P}(A, b)$ nonempty, that is, does our linear program have a chance to have a solution?

(2) Does the objective function $c_1x_1 + \cdots + c_nx_n$ have a maximum value on $\mathcal{P}(A, b)$?

The answer to both questions can be **no**. But if $\mathcal{P}(A, b)$ is nonempty and if the objective function is bounded above (on $\mathcal{P}(A, b)$), then it can be shown that the maximum of $c_1x_1 + \cdots + c_nx_n$ is achieved by some $x \in \mathcal{P}(A, b)$. Such a solution is called an *optimal solution*. Perhaps surprisingly, this result is not so easy to prove (unless one has the simplex method as its disposal). We will prove this result in full detail (see Proposition 25.1).

The reason why linear constraints are so important is that the domain of potential optimal solutions $\mathcal{P}(A, b)$ is *convex*. In fact, $\mathcal{P}(A, b)$ is a convex polyhedron which is the intersection of half-spaces cut out by affine hyperplanes. The objective function being linear is convex, and this is also a crucial fact. Thus, we are led to study convex sets, in particular those that arise from solutions of inequalities defined by affine forms, but also convex cones.

We give a brief introduction to these topics. As a reward, we provide several criteria for testing whether a system of inequalities

$$Ax \leq b,\ x \geq 0$$

has a solution or not in terms of versions of the *Farkas lemma* (see Proposition 30.3 and Proposition 27.4). Then we give a complete proof of the strong duality theorem for linear programming (see Theorem 27.7). We also discuss the complementary slackness conditions and show that they can be exploired to design an algorithm for solving a linear program that uses both the primal problem and its dual. This algorithm known as the *primal dual algorithm*, although not used much nowadays, has been the source of inspiration for a whole class of approximation algorithms also known as primal dual algorithms.

We hope that these notes will be a motivation for learning more about linear programming, convex optimization, but also convex geometry. The "bible" in convex optimization is Boyd and Vandenberghe [22], and one of the best sources for convex geometry is Ziegler [113]. This is a rather advanced text, so the reader may want to begin with Gallier [45].

24.2 Affine Subsets, Convex Sets, Affine Hyperplanes, Half-Spaces

We view $\mathbb{R}^n$ as consisting of *column vectors* ($n \times 1$ matrices). As usual, row vectors represent *linear forms*, that is linear maps $\varphi\colon \mathbb{R}^n \to \mathbb{R}$, in the sense that the row vector y (a $1 \times n$ matrix) represents the linear form φ if $\varphi(x) = yx$ for all $x \in \mathbb{R}^n$. We denote the space of linear forms (row vectors) by $(\mathbb{R}^n)^*$.

Recall that a *linear combination* of vectors in $\mathbb{R}^n$ is an expression

$$\lambda_1 x_1 + \cdots + \lambda_m x_m$$

where $x_1, \ldots, x_m \in \mathbb{R}^n$ and where $\lambda_1, \ldots, \lambda_m$ are *arbitrary* scalars in $\mathbb{R}$. Given a sequence of vectors $S = (x_1, \ldots, x_m)$ with $x_i \in \mathbb{R}^n$, the set of all linear combinations of the vectors in S is the smallest (linear) subspace containing S called the *linear span* of S, and denoted $\mathrm{span}(S)$. A *linear subspace* of $\mathbb{R}^n$ is any nonempty subset of $\mathbb{R}^n$ closed under linear combinations.

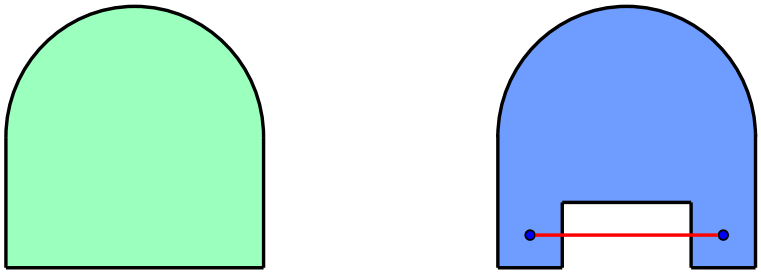
An *affine combination* of vectors in $\mathbb{R}^n$ is an expression

$$\lambda_1 x_1 + \cdots + \lambda_m x_m$$

where $x_1, \ldots, x_m \in \mathbb{R}^n$ and where $\lambda_1, \ldots, \lambda_m$ are scalars in $\mathbb{R}$ *satisfying the condition*

$$\lambda_1 + \cdots + \lambda_m = 1.$$

Given a sequence of vectors $S = (x_1, \ldots, x_m)$ with $x_i \in \mathbb{R}^n$, the set of all affine combinations of the vectors in S is the smallest affine subspace containing S called the *affine hull* of S and denoted $\mathrm{aff}(S)$.



(a) (b)

Figure 24.1: (a) A convex set; (b) A nonconvex set

Definition 24.1. An *affine subspace* A of $\mathbb{R}^n$ is any subset of $\mathbb{R}^n$ closed under affine combinations.

If A is a nonempty affine subset of $\mathbb{R}^n$, then it can be shown that $V_A = \{a - b \mid a, b \in A\}$ is a linear subspace of $\mathbb{R}^n$ called the *direction of* A, and that

$$A = a + V_A = \{a + v \mid v \in V_A\}$$

for *any* $a \in A$. The *dimension* of a nonempty affine subspace A is the dimension of its direction V_A.

Convex combinations are affine combinations $\lambda_1 x_1 + \cdots + \lambda_m x_m$ satisfying the extra condition that $\lambda_i \geq 0$ for $i = 1, \ldots, m$. A convex set is defined as follows.

Definition 24.2. A subset V of $\mathbb{R}^n$ is *convex* if for any two points $a, b \in V$, we have $c \in V$ for every point $c = (1 - \lambda)a + \lambda b$, with $0 \leq \lambda \leq 1$ ($\lambda \in \mathbb{R}$). Given any two points a, b, the notation $[a, b]$ is often used to denote the line segment between a and b, that is,

$$[a, b] = \{c \in \mathbb{R}^n \mid c = (1 - \lambda)a + \lambda b,\ 0 \leq \lambda \leq 1\},$$

and thus a set V is convex if $[a, b] \subseteq V$ for any two points $a, b \in V$ ($a = b$ is allowed). The *dimension* of a convex set V is the dimension of its affine hull $\mathrm{aff}(A)$.

The empty set is trivially convex, every one-point set $\{a\}$ is convex, and the entire affine space $\mathbb{R}^n$ is convex.

It is obvious that the intersection of any family (finite or infinite) of convex sets is convex.

Definition 24.3. Given any (nonempty) subset S of $\mathbb{R}^n$, the smallest convex set containing S is denoted by $\mathrm{conv}(S)$ and called the *convex hull of* S (it is the intersection of all convex sets containing S).

A good understanding of what $\mathrm{conv}(S)$ is, and good methods for computing it, are essential. We have the following simple but crucial result.

Proposition 24.1. *For any family $S = (a_i)_{i\in I}$ of points in $\mathbb{R}^n$, the set V of convex combinations $\sum_{i\in I} \lambda_i a_i$ (where $\sum_{i\in I} \lambda_i = 1$ and $\lambda_i \geq 0$) is the convex hull $\mathrm{conv}(S)$ of $S = (a_i)_{i\in I}$.*

It is natural to wonder whether Proposition 24.1 can be sharpened in two directions: (1) Is it possible to have a fixed bound on the number of points involved in the convex combinations? (2) Is it necessary to consider convex combinations of all points, or is it possible to consider only a subset with special properties?

The answer is yes in both cases. In Case 1, Carathéodory's theorem asserts that it is enough to consider convex combinations of $n + 1$ points. For example, in the plane $\mathbb{R}^2$, the convex hull of a set S of points is the union of all triangles (interior points included) with vertices in S. In Case 2, the theorem of Krein and Milman asserts that a convex set that is also compact is the convex hull of its extremal points (given a convex set S, a point $a \in S$ is extremal if $S - \{a\}$ is also convex).

We will not prove these theorems here, but we invite the reader to consult Gallier [45] or Berger [10].

Convex sets also arise as half-spaces cut out by affine hyperplanes.

Definition 24.4. An *affine form* $\varphi\colon \mathbb{R}^n \to \mathbb{R}$ is defined by some linear form $c \in (\mathbb{R}^n)^*$ and some scalar $\beta \in \mathbb{R}$ so that

$$\varphi(x) = cx + \beta \quad \text{for all } x \in \mathbb{R}^n.$$

If $c \neq 0$, the affine form φ specified by (c, β) defines the *affine hyperplane* (for short *hyperplane*) $H(\varphi)$ given by

$$H(\varphi) = \{x \in \mathbb{R}^n \mid \varphi(x) = 0\} = \{x \in \mathbb{R}^n \mid cx + \beta = 0\},$$

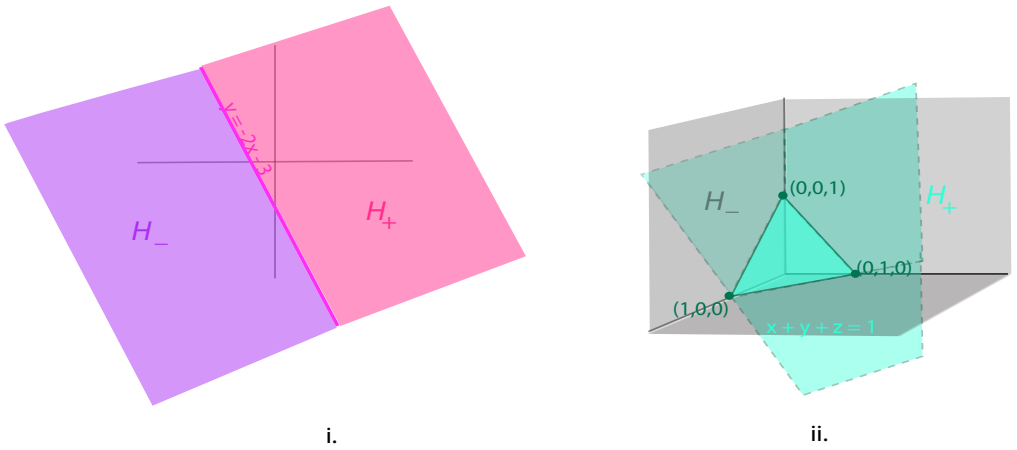
and the two *(closed) half-spaces*

$$H_+(\varphi) = \{x \in \mathbb{R}^n \mid \varphi(x) \geq 0\} = \{x \in \mathbb{R}^n \mid cx + \beta \geq 0\},$$
$$H_-(\varphi) = \{x \in \mathbb{R}^n \mid \varphi(x) \leq 0\} = \{x \in \mathbb{R}^n \mid cx + \beta \leq 0\}.$$

When $\beta = 0$, we call H a *linear hyperplane*.

Both $H_+(\varphi)$ and $H_-(\varphi)$ are convex and $H = H_+(\varphi) \cap H_-(\varphi)$.

For example, $\varphi\colon \mathbb{R}^2 \to \mathbb{R}$ with $\varphi(x, y) = 2x + y + 3$ is an affine form defining the line given by the equation $y = -2x - 3$. Another example of an affine form is $\varphi\colon \mathbb{R}^3 \to \mathbb{R}$ with $\varphi(x, y, z) = x + y + z - 1$; this affine form defines the plane given by the equation $x + y + z = 1$, which is the plane through the points $(0, 0, 1), (0, 1, 0)$, and $(1, 0, 0)$. Both of these hyperplanes are illustrated in Figure 24.2.



i. ii.

Figure 24.2: Figure i. illustrates the hyperplane $H(\varphi)$ for $\varphi(x, y) = 2x + y + 3$, while Figure ii. illustrates the hyperplane $H(\varphi)$ for $\varphi(x, y, z) = x + y + z - 1$.

For any two vector $x, y \in \mathbb{R}^n$ with $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ we write $x \leq y$ iff $x_i \leq y_i$ for $i = 1, \ldots, n$, and $x \geq y$ iff $y \leq x$. In particular $x \geq 0$ iff $x_i \geq 0$ for $i = 1, \ldots, n$.

Certain special types of convex sets called cones and $\mathcal{H}$-polyhedra play an important role. The set of feasible solutions of a linear program is an $\mathcal{H}$-polyhedron, and cones play a crucial role in the proof of Proposition 25.1 and in the Farkas–Minkowski proposition (Proposition 27.2).

24.3 Cones, Polyhedral Cones, and $\mathcal{H}$-Polyhedra

Cones and polyhedral cones are defined as follows.

Definition 24.5. Given a nonempty subset $S \subseteq \mathbb{R}^n$, the *cone* $C = \text{cone}(S)$ spanned by S is the convex set

$$\text{cone}(S) = \left\{ \sum_{i=1}^{k} \lambda_i u_i,\ u_i \in S, \lambda_i \in \mathbb{R},\ \lambda_i \geq 0 \right\},$$

of positive combinations of vectors from S. If S consists of a finite set of vector, the cone $C = \text{cone}(S)$ is called a *polyhedral cone*. Figure 24.3 illustrates a polyhedral cone.

Note that if some nonzero vector u belongs to a cone C, then $\lambda u \in C$ for all $\lambda \geq 0$, that is, the *ray* $\{\lambda u \mid \lambda \geq 0\}$ belongs to C.

Remark: The cones (and polyhedral cones) of Definition 24.5 are always convex. For this reason we use the simpler terminology cone instead of convex cone. However, there are more

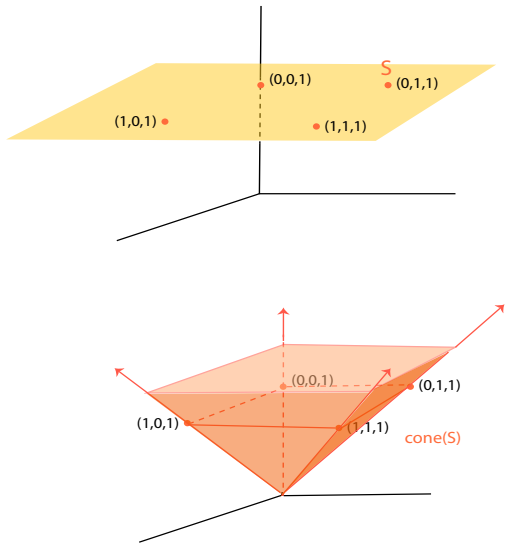


Figure 24.3: Let $S = \{(0,0,1),(1,0,1),(1,1,1),(0,1,1)\}$. The polyhedral cone, cone(S), is the solid "pyramid" with apex at the origin and square cross sections.

general kinds of cones that are not convex (for example, a union of polyhedral cones or the linear cone generated by the curve in Figure 24.4), and if we were dealing with those we would refer to the cones of Definition 24.5 as convex cones.

Definition 24.6. An *$\mathcal{H}$-polyhedron*, for short a *polyhedron*, is any subset $\mathcal{P} = \bigcap_{i=1}^{s} C_i$ of $\mathbb{R}^n$ defined as the intersection of a finite number s of closed half-spaces C_i. An example of an $\mathcal{H}$-polyhedron is shown in Figure 24.6. An *$\mathcal{H}$-polytope* is a bounded $\mathcal{H}$-polyhedron, which means that there is a closed ball $B_r(x)$ of center x and radius $r > 0$ such that $\mathcal{P} \subseteq B_r(x)$. An example of a $\mathcal{H}$-polytope is shown in Figure 24.5.

By convention, we agree that $\mathbb{R}^n$ itself is an $\mathcal{H}$-polyhedron.

Remark: The $\mathcal{H}$-polyhedra of Definition 24.6 are always convex. For this reason, as in the case of cones we use the simpler terminology $\mathcal{H}$-polyhedron instead of convex $\mathcal{H}$-polyhedron. In algebraic topology, there are more general polyhedra that are not convex.

It can be shown that an $\mathcal{H}$-polytope $\mathcal{P}$ is equal to the convex hull of finitely many points (the extreme points of $\mathcal{P}$). This is a nontrivial result whose proof takes a significant amount of work; see Gallier [45] and Ziegler [113].

An unbounded $\mathcal{H}$-polyhedron is not equal to the convex hull of finite set of points. To obtain an equivalent notion we introduce the notion of a $\mathcal{V}$-polyhedron.

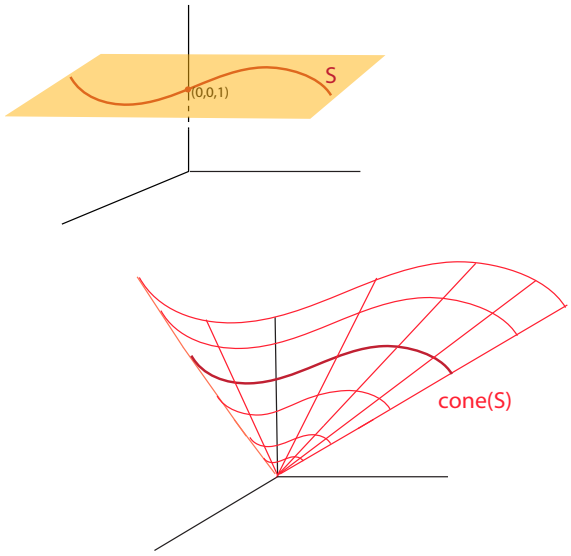


Figure 24.4: Let S be a planar curve in $z = 1$. The linear cone of S, consisting of all half rays connecting S to the origin, is not convex.

Definition 24.7. A $\mathcal{V}$-*polyhedron* is any convex subset $A \subseteq \mathbb{R}^n$ of the form

$$A = \mathrm{conv}(Y) + \mathrm{cone}(V) = \{a + v \mid a \in \mathrm{conv}(Y),\ v \in \mathrm{cone}(V)\},$$

where $Y \subseteq \mathbb{R}^n$ and $V \subseteq \mathbb{R}^n$ are *finite* (possibly empty).

When $V = \emptyset$ we simply have a *polytope*, and when $Y = \emptyset$ or $Y = \{0\}$, we simply have a cone.

It can be shown that every $\mathcal{H}$-polyhedron is a $\mathcal{V}$-polyhedron and conversely. This is one of the major theorems in the theory of polyhedra, and its proof is nontrivial. For a complete proof, see Gallier [45] and Ziegler [113].

Every polyhedral cone is closed. This is an important fact that is used in the proof of several other key results such as Proposition 25.1 and the Farkas–Minkowski proposition (Proposition 27.2).

Although it seems obvious that a polyhedral cone should be closed, a rigorous proof is not entirely trivial.

Indeed, the fact that a polyhedral cone is closed relies crucially on the fact that C is spanned by a finite number of vectors, because the cone generated by an infinite set may not be closed. For example, consider the closed disk $D \subseteq \mathbb{R}^2$ of center $(0, 1)$ and radius 1, which is tangent to the x-axis at the origin. Then the cone(D) consists of the open upper half-plane *plus* the origin $(0, 0)$, but this set is not closed.

Figure 24.5: An icosahedron is an example of an $\mathcal{H}$-polytope.

Proposition 24.2. *Every polyhedral cone C is closed.*

Proof. This is proved by showing that

1. Every primitive cone is closed.
2. A polyhedral cone C is the union of finitely many primitive cones, where a *primitive cone* is a polyhedral cone spanned by linearly independent vectors.

Assume that $(a_1, \ldots, a_m)$ are linearly independent vectors in $\mathbb{R}^n$, and consider any sequence $(x^{(k)})_{k\geq 0}$

$$x^{(k)} = \sum_{i=1}^{m} \lambda_i^{(k)} a_i$$

of vectors in the primitive cone $\text{cone}(\{a_1, \ldots, a_m\})$, which means that $\lambda_j^{(k)} \geq 0$ for $i = 1, \ldots, m$ and all $k \geq 0$. The vectors $x^{(k)}$ belong to the subspace U spanned by $(a_1, \ldots, a_m)$, and U is closed. Assume that the sequence $(x^{(k)})_{k\geq 0}$ converges to a limit $x \in \mathbb{R}^n$. Since U is closed and $x^{(k)} \in U$ for all $k \geq 0$, we have $x \in U$. If we write $x = x_1 a_1 + \cdots + x_m a_m$, we would like to prove that $x_i \geq 0$ for $i = 1, \ldots, m$. The sequence the $(x^{(k)})_{k\geq 0}$ converges to x iff

$$\lim_{k \mapsto \infty} \left\| x^{(k)} - x \right\| = 0,$$

iff

$$\lim_{k \mapsto \infty} \Big(\sum_{i=1}^{m} |\lambda_i^{(k)} - x_i|^2 \Big)^{1/2} = 0$$

iff

$$\lim_{k \mapsto \infty} \lambda_i^{(k)} = x_i, \quad i = 1, \ldots, m.$$

Since $\lambda_i^{(k)} \geq 0$ for $i = 1, \ldots, m$ and all $k \geq 0$, we have $x_i \geq 0$ for $i = 1, \ldots, m$, so $x \in \text{cone}(\{a_1, \ldots, a_m\})$.

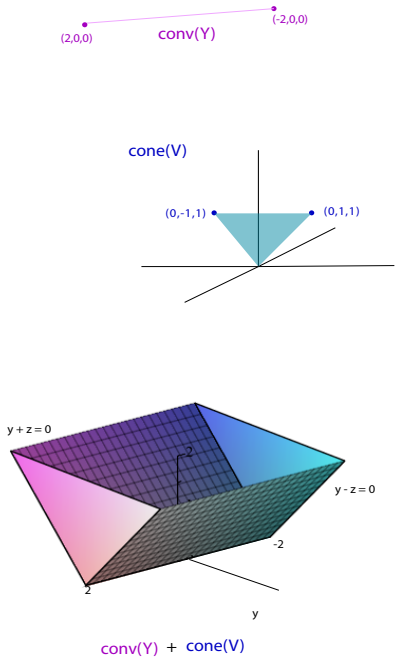


Figure 24.6: The "triangular trough" determined by the inequalities $y - z \leq 0$, $y + z \geq 0$, and $-2 \leq x \leq 2$ is an $\mathcal{H}$-polyhedron and an $\mathcal{V}$-polyhedron, where $Y = \{(2, 0, 0,), (-2, 0, 0)\}$ and $V = \{(0, 1, 1), (0, -1, 1)\}$.

Next, assume that x belongs to the polyhedral cone C. Consider a positive combination

$$x = \lambda_1 a_1 + \cdots + \lambda_k a_k, \tag{$*_1$}$$

for some nonzero $a_1, \ldots, a_k \in C$, with $\lambda_i \geq 0$ and with k *minimal*. Since k is minimal, we must have $\lambda_i > 0$ for $i = 1, \ldots, k$. We claim that $(a_1, \ldots, a_k)$ are linearly independent.

If not, there is some nontrivial linear combination

$$\mu_1 a_1 + \cdots + \mu_k a_k = 0, \tag{$*_2$}$$

and since the a_i are nonzero, $\mu_j \neq 0$ for some at least some j. We may assume that $\mu_j < 0$ for some j (otherwise, we consider the family $(-\mu_i)_{1 \leq i \leq k}$), so let

$$J = \{j \in \{1, \ldots, k\} \mid \mu_j < 0\}.$$

For any $t \in \mathbb{R}$, since $x = \lambda_1 a_1 + \cdots + \lambda_k a_k$, using $(*_2)$ we get

$$x = (\lambda_1 + t\mu_1)a_1 + \cdots + (\lambda_k + t\mu_k)a_k, \tag{$*_3$}$$

and if we pick

$$t = \min_{j \in J}\left(-\frac{\lambda_j}{\mu_j}\right) \geq 0,$$

we have $(\lambda_i + t\mu_i) \geq 0$ for $i = 1, \ldots, k$, but $\lambda_j + t\mu_j = 0$ for some $j \in J$, so $(*_3)$ is an expression of x with less that k nonzero coefficients, contadicting the minimality of k in $(*_1)$. Therefore, $(a_1, \ldots, a_k)$ are linearly independent.

Since a polyhedral cone C is spanned by finitely many vectors, there are finitely many primitive cones (corresponding to linearly independent subfamilies), and since every $x \in C$, belongs to some primitive cone, C is the union of a finite number of primitive cones. Since every primitive cone is closed, as a union of finitely many closed sets, C itself is closed.

The above facts are also proved in Matousek and Gardner [72] (Chapter 6, Section 5, Lemma 6.5.3, 6.5.4, and 6.5.5). □

Another way to prove that a polyhedral cone C is closed is to show that C is also a $\mathcal{H}$-polyhedron. This takes even more work; see Gallier [45] (Chapter 4, Section 4, Proposition 4.16). Yet another proof is given in Lax [66] (Chapter 13, Theorem 1).

Chapter 25

Linear Programs

25.1 Linear Programs, Feasible Solutions, Optimal Solutions

The purpose of linear programming is to solve the following type of optimization problem.

Definition 25.1. *A linear program* (P) is the following kind of optimization problem:

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \\ &\quad a_1 x \leq b_1 \\ &\quad \cdots \\ &\quad a_m x \leq b_m \\ &\quad x \geq 0, \end{aligned}$$

where $x \in \mathbb{R}^n$, $c, a_1, \ldots, a_m \in (\mathbb{R}^n)^*$, $b_1, \ldots, b_m \in \mathbb{R}$.

The linear form c defines the *objective function* $x \mapsto cx$ of the program (P) (from $\mathbb{R}^n$ to $\mathbb{R}$), and the inequalities $a_i x \leq b_i$ and $x_j \geq 0$ are called the *constraints* of the linear program (P).

If we define the $m \times n$ matrix

$$A = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix}$$

whose rows are the row vectors $a_1, \ldots, a_m$ and b as the column vector

$$b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix},$$

the m inequality constraints $a_i x \leq b_i$ can be written in matrix form as

$$Ax \leq b.$$

Thus the linear program (P) can also be stated as the linear program (P):

$$\begin{array}{ll} \text{maximize} & cx \\ \text{subject to} & Ax \leq b \text{ and } x \geq 0. \end{array}$$

Here is an explicit example of a linear program of type (P):

Example 25.1.

$$\begin{array}{ll} \text{maximize} & x_1 + x_2 \\ \text{subject to} & \\ & x_2 - x_1 \leq 1 \\ & x_1 + 6x_2 \leq 15 \\ & 4x_1 - x_2 \leq 10 \\ & x_1 \geq 0,\ x_2 \geq 0, \end{array}$$

and in matrix form

$$\begin{array}{ll} \text{maximize} & \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ \text{subject to} & \\ & \begin{pmatrix} -1 & 1 \\ 1 & 6 \\ 4 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \leq \begin{pmatrix} 1 \\ 15 \\ 10 \end{pmatrix} \\ & x_1 \geq 0,\ x_2 \geq 0. \end{array}$$

It turns out that $x_1 = 3, x_2 = 2$ yields the maximum of the objective function $x_1 + x_2$, which is 5. This is illustrated in Figure 25.1. Observe that the set of points that satisfy the above constraints is a convex region cut out by half planes determined by the lines of equations

$$\begin{aligned} x_2 - x_1 &= 1 \\ x_1 + 6x_2 &= 15 \\ 4x_1 - x_2 &= 10 \\ x_1 &= 0 \\ x_2 &= 0. \end{aligned}$$

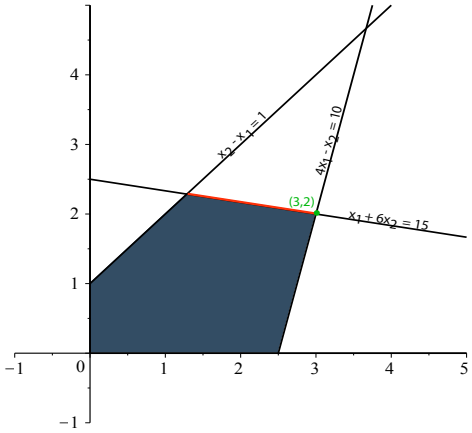


Figure 25.1: The $\mathcal{H}$-polyhedron associated with Example 25.1. The green point $(3, 2)$ is the unique optimal solution.

In general, each constraint $a_i x \leq b_i$ corresponds to the affine form φ_i given by $\varphi_i(x) = a_i x - b_i$ and defines the half-space $H_-(\varphi_i)$, and each inequality $x_j \geq 0$ defines the half-space $H_+(x_j)$. The intersection of these half-spaces is the set of solutions of all these constraints. It is a (possibly empty) $\mathcal{H}$-polyhedron denoted $\mathcal{P}(A, b)$.

Definition 25.2. If $\mathcal{P}(A, b) = \emptyset$, we say that the linear program (P) has *no feasible solution*, and otherwise any $x \in \mathcal{P}(A, b)$ is called a *feasible solution* of (P).

The linear program shown in Example 25.2 obtained by reversing the direction of the inequalities $x_2 - x_1 \leq 1$ and $4x_1 - x_2 \leq 10$ in the linear program of Example 25.1 has no feasible solution; see Figure 25.2.

Example 25.2.

$$\begin{aligned} &\text{maximize} && x_1 + x_2 \\ &\text{subject to} && \\ & && x_1 - x_2 \leq -1 \\ & && x_1 + 6x_2 \leq 15 \\ & && x_2 - 4x_1 \leq -10 \\ & && x_1 \geq 0,\ x_2 \geq 0. \end{aligned}$$

Assume $\mathcal{P}(A, b) \neq \emptyset$, so that the linear program (P) has a feasible solution. In this case, consider the image $\{cx \in \mathbb{R} \mid x \in \mathcal{P}(A, b)\}$ of $\mathcal{P}(A, b)$ under the objective function $x \mapsto cx$.

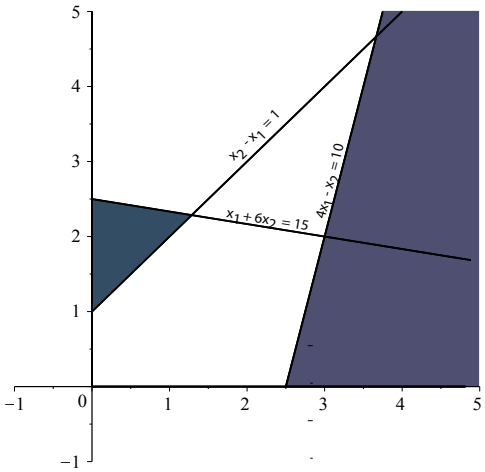


Figure 25.2: There is no $\mathcal{H}$-polyhedron associated with Example 25.2 since the blue and purple regions do not overlap.

Definition 25.3. If the set $\{cx \in \mathbb{R} \mid x \in \mathcal{P}(A, b)\}$ is unbounded above, then we say that the linear program (P) is *unbounded*.

The linear program shown in Example 25.3 obtained from the linear program of Example 25.1 by deleting the constraints $4x_1 - x_2 \leq 10$ and $x_1 + 6x_2 \leq 15$ is unbounded.

Example 25.3.

$$\begin{aligned} &\text{maximize} && x_1 + x_2 \\ &\text{subject to} && \\ & && x_2 - x_1 \leq 1 \\ & && x_1 \geq 0,\ x_2 \geq 0. \end{aligned}$$

Otherwise, we will prove shortly that if μ is the least upper bound of the set $\{cx \in \mathbb{R} \mid x \in \mathcal{P}(A, b)\}$, then there is some $p \in \mathcal{P}(A, b)$ such that

$$cp = \mu,$$

that is, the objective function $x \mapsto cx$ has a maximum value μ on $\mathcal{P}(A, b)$ which is achieved by some $p \in \mathcal{P}(A, b)$.

Definition 25.4. If the set $\{cx \in \mathbb{R} \mid x \in \mathcal{P}(A, b)\}$ is nonempty and bounded above, any point $p \in \mathcal{P}(A, b)$ such that $cp = \max\{cx \in \mathbb{R} \mid x \in \mathcal{P}(A, b)\}$ is called an *optimal solution* (or *optimum*) of (P). Optimal solutions are often denoted by an upper $*$; for example, p^*.

The linear program of Example 25.1 has a unique optimal solution $(3, 2)$, but observe that the linear program of Example 25.4 in which the objective function is $(1/6)x_1 + x_2$ has infinitely many optimal solutions; the maximum of the objective function is $15/6$ which occurs along the points of orange boundary line in Figure 25.1.

Example 25.4.

$$\begin{aligned} &\text{maximize} && \frac{1}{6}x_1 + x_2 \\ &\text{subject to} && \\ & && x_2 - x_1 \leq 1 \\ & && x_1 + 6x_2 \leq 15 \\ & && 4x_1 - x_2 \leq 10 \\ & && x_1 \geq 0,\ x_2 \geq 0. \end{aligned}$$

The proof that if the set $\{cx \in \mathbb{R} \mid x \in \mathcal{P}(A, b)\}$ is nonempty and bounded above, then there is an optimal solution $p \in \mathcal{P}(A, b)$, is not as trivial as it might seem. It relies on the fact that a polyhedral cone is closed, a fact that was shown in Section 24.3.

We also use a trick that makes the proof simpler, which is that a linear program (P) with inequality constraints $Ax \leq b$

$$\begin{aligned} &\text{maximize} && cx \\ &\text{subject to} && Ax \leq b \text{ and } x \geq 0, \end{aligned}$$

is equivalent to the linear program (P_2) with equality constraints

$$\begin{aligned} &\text{maximize} && \widehat{c}\,\widehat{x} \\ &\text{subject to} && \widehat{A}\widehat{x} = b \text{ and } \widehat{x} \geq 0, \end{aligned}$$

where $\widehat{A}$ is an $m \times (n + m)$ matrix, $\widehat{c}$ is a linear form in $(\mathbb{R}^{n+m})^*$, and $\widehat{x} \in \mathbb{R}^{n+m}$, given by

$$\widehat{A} = \begin{pmatrix} A & I_m \end{pmatrix}, \quad \widehat{c} = \begin{pmatrix} c & 0_m^\top \end{pmatrix}, \quad \text{and} \quad \widehat{x} = \begin{pmatrix} x \\ z \end{pmatrix},$$

with $x \in \mathbb{R}^n$ and $z \in \mathbb{R}^m$.

Indeed, $\widehat{A}\widehat{x} = b$ and $\widehat{x} \geq 0$ iff

$$Ax + z = b, \quad x \geq 0,\ z \geq 0,$$

iff

$$Ax \leq b, \quad x \geq 0,$$

and $\widehat{c}\,\widehat{x} = cx$.

The variables z are called *slack variables*, and a linear program of the form (P_2) is called a linear program in *standard form*.

The result of converting the linear program of Example 25.4 to standard form is the program shown in Example 25.5.

Example 25.5.

$$\begin{aligned}
&\text{maximize} && \frac{1}{6}x_1 + x_2 \\
&\text{subject to} && \\
& && x_2 - x_1 + z_1 = 1 \\
& && x_1 + 6x_2 + z_2 = 15 \\
& && 4x_1 - x_2 + z_3 = 10 \\
& && x_1 \geq 0,\ x_2 \geq 0,\ z_1 \geq 0,\ z_2 \geq 0,\ z_3 \geq 0.
\end{aligned}$$

We can now prove that if a linear program has a feasible solution and is bounded, then it has an optimal solution.

Proposition 25.1. *Let (P_2) be a linear program in standard form, with equality constraint $Ax = b$. If $\mathcal{P}(A, b)$ is nonempty and bounded above, and if μ is the least upper bound of the set $\{cx \in \mathbb{R} \mid x \in \mathcal{P}(A, b)\}$, then there is some $p \in \mathcal{P}(A, b)$ such that*

$$cp = \mu,$$

that is, the objective function $x \mapsto cx$ has a maximum value μ on $\mathcal{P}(A, b)$ which is achieved by some optimum solution $p \in \mathcal{P}(A, b)$.

Proof. Since $\mu = \sup\{cx \in \mathbb{R} \mid x \in \mathcal{P}(A, b)\}$, there is a sequence $(x^{(k)})_{k\geq 0}$ of vectors $x^{(k)} \in \mathcal{P}(A, b)$ such that $\lim_{k\mapsto\infty} cx^{(k)} = \mu$. In particular, if we write $x^{(k)} = (x_1^{(k)}, \ldots, x_n^{(k)})$ we have $x_j^{(k)} \geq 0$ for $j = 1, \ldots, n$ and for all $k \geq 0$. Let $\widetilde{A}$ be the $(m + 1) \times n$ matrix

$$\widetilde{A} = \begin{pmatrix} c \\ A \end{pmatrix},$$

and consider the sequence $(\widetilde{A}x^{(k)})_{k\geq 0}$ of vectors $\widetilde{A}x^{(k)} \in \mathbb{R}^{m+1}$. We have

$$\widetilde{A}x^{(k)} = \begin{pmatrix} c \\ A \end{pmatrix} x^{(k)} = \begin{pmatrix} cx^{(k)} \\ Ax^{(k)} \end{pmatrix} = \begin{pmatrix} cx^{(k)} \\ b \end{pmatrix},$$

since by hypothesis $x^{(k)} \in \mathcal{P}(A, b)$, and the constraints are $Ax = b$ and $x \geq 0$. Since by hypothesis $\lim_{k\mapsto\infty} cx^{(k)} = \mu$, the sequence $(\widetilde{A}x^{(k)})_{k\geq 0}$ converges to the vector $\begin{pmatrix} \mu \\ b \end{pmatrix}$. Now, observe that each vector $\widetilde{A}x^{(k)}$ can be written as the convex combination

$$\widetilde{A}x^{(k)} = \sum_{j=1}^{n} x_j^{(k)} \widetilde{A}^j,$$

with $x_j^{(k)} \geq 0$ and where $\widetilde{A}^j \in \mathbb{R}^{m+1}$ is the jth column of $\widetilde{A}$. Therefore, $\widetilde{A}x^{(k)}$ belongs to the polyheral cone

$$C = \text{cone}(\widetilde{A}^1, \ldots, \widetilde{A}^n) = \{\widetilde{A}x \mid x \in \mathbb{R}^n,\ x \geq 0\},$$

and since by Proposition 24.2 this cone is closed, $\lim_{k\geq\infty} \widetilde{A}x^{(k)} \in C$, which means that there is some $u \in \mathbb{R}^n$ with $u \geq 0$ such that

$$\begin{pmatrix} \mu \\ b \end{pmatrix} = \lim_{k\geq\infty} \widetilde{A}x^{(k)} = \widetilde{A}u = \begin{pmatrix} cu \\ Au \end{pmatrix},$$

that is, $cu = \mu$ and $Au = b$. Hence, u is an optimal solution of (P_2). □

The next question is, how do we find such an optimal solution? It turns out that for linear programs in standard form where the constraints are of the form $Ax = b$ and $x \geq 0$, there are always optimal solutions of a special type called basic feasible solutions.

25.2 Basic Feasible Solutions and Vertices

If the system $Ax = b$ has a solution and if some row of A is a linear combination of other rows, then the corresponding equation is redundant, so we may assume that the rows of A are linearly independent; that is, we may assume that A has rank m, so $m \leq n$.

If A is an $m \times n$ matrix, for any nonempty subset K of $\{1, \ldots, n\}$, let A_K be the submatrix of A consisting of the columns of A whose indices belong to K. We denote the jth column of the matrix A by A^j.

Definition 25.5. Given a linear program (P_2)

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax = b \text{ and } x \geq 0, \end{aligned}$$

where A has rank m, a vector $x \in \mathbb{R}^n$ is a *basic feasible solution* of (P) if $x \in \mathcal{P}(A, b) \neq \emptyset$, and if there is some subset K of $\{1, \ldots, n\}$ of size m such that

(1) The matrix A_K is invertible (that is, the columns of A_K are linearly independent).

(2) $x_j = 0$ for all $j \notin K$.

The subset K is called a *basis* of x. Every index $k \in K$ is called *basic*, and every index $j \notin K$ is called *nonbasic*. Similarly, the columns A^k corresponding to indices $k \in K$ are called *basic*, and the columns A^j corresponding to indices $j \notin K$ are called *nonbasic*. The variables corresponding to basic indices $k \in K$ ar called *basic variables*, and the variables corresponding to indices $j \notin K$ are called *nonbasic*.

For example, the linear program

$$\begin{array}{ll} \text{maximize} & x_1 + x_2 \\ \text{subject to} & x_1 + x_2 + x_3 = 1 \text{ and } x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0, \end{array} \qquad (*)$$

has three basic feasible solutions; the basic feasible solution $K = \{1\}$ corresponds to the point $(1, 0, 0)$; the basic feasible solution $K = \{2\}$ corresponds to the point $(0, 1, 0)$; the basic feasible solution $K = \{3\}$ corresponds to the point $(0, 0, 1)$. Each of these points corresponds to the vertices of the slanted purple triangle illustrated in Figure 25.3. The vertices $(1, 0, 0)$ and $(0, 1, 0)$ optimize the objective function with a value of 1.

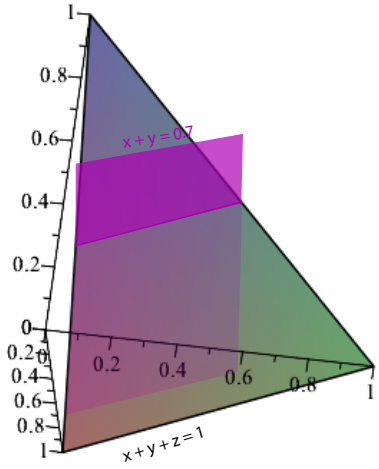


Figure 25.3: The $\mathcal{H}$-polytope associated with Linear Program $(*)$. The objective function (with $x_1 \to x$ and $x_2 \to y$) is represented by vertical planes parallel to the purple plane $x + y = 0.7$, and reaches it maximal value when $x + y = 1$.

We now show that if the standard linear program (P_2) as in Definition 25.5 has some feasible solution and is bounded above, then some basic feasible solution is an optimal solution. We follow Matousek and Gardner [72] (Chapter 4, Section 2, Theorem 4.2.3).

First we obtain a more convenient characterization of a basic feasible solution.

Proposition 25.2. *Given any standard linear program (P_2) where $Ax = b$ and A is an $m \times n$ matrix of rank m, for any feasible solution x, if $J_> = \{j \in \{1, \ldots, n\} \mid x_j > 0\}$, then x is a basic feasible solution iff the columns of the matrix $A_{J_>}$ are linearly independent.*

Proof. If x is a basic feasible solution then there is some subset $K \subseteq \{1, \ldots, n\}$ of size m such that the columns of A_K are linearly independent and $x_j = 0$ for all $j \notin K$, so by definition $J_> \subseteq K$, which implies that the columns of the matrix $A_{J_>}$ are linearly independent.

Conversely, assume that x is a feasible solution such that the columns of the matrix $A_{J_>}$ are linearly independent. If $|J_>| = m$, we are done since we can pick $K = J_>$ and then x

is a basic feasible solution. If $|J_>| < m$, we can extend $J_>$ to an m-element subset K by adding $m - |J_>|$ column indices so that the columns of A_K are linearly independent, which is possible since A has rank m. □

Next we prove that if a linear program in standard form has any feasible solution x_0 and is bounded above, then is has some basic feasible solution $\widetilde{x}$ which is as good as x_0, in the sense that $c\widetilde{x} \geq cx_0$.

Proposition 25.3. *Let (P_2) be any standard linear program with objective function cx, where $Ax = b$ and A is an $m \times n$ matrix of rank m. If (P_2) is bounded above and if x_0 is some feasible solution of (P_2), then there is some basic feasible solution $\widetilde{x}$ such that $c\widetilde{x} \geq cx_0$.*

Proof. Among the feasible solutions x such that $cx \geq cx_0$ (x_0 is one of them) pick one with the *maximum* number of coordinates x_j equal to 0, say $\widetilde{x}$. Let

$$K = J_> = \{j \in \{1, \ldots, n\} \mid \widetilde{x}_j > 0\}$$

and let $s = |K|$. We claim that $\widetilde{x}$ is a basic feasible solution, and by construction $c\widetilde{x} \geq cx_0$.

If the columns of A_K are linearly independent, then by Proposition 25.2 we know that $\widetilde{x}$ is a basic feasible solution and we are done.

Otherwise, the columns of A_K are linearly dependent, so there is some nonzero vector $v = (v_1, \ldots, v_s)$ such that $A_K\, v = 0$. Let $w \in \mathbb{R}^n$ be the vector obtained by extending v by setting $w_j = 0$ for all $j \notin K$. By construction,

$$Aw = A_K\, v = 0.$$

We will derive a contradiction by exhibiting a feasible solution $x(t_0)$ such that $cx(t_0) \geq cx_0$ with more zero coordinates than $\widetilde{x}$.

For this we claim that we may assume that w satisfies the following two conditions:

(1) $cw \geq 0$.

(2) There is some $j \in K$ such that $w_j < 0$.

If $cw = 0$ and if Condition (2) fails, since $w \neq 0$, we have $w_j > 0$ for some $j \in K$, in which case we can use $-w$, for which $w_j < 0$.

If $cw < 0$ then $c(-w) > 0$, so we may assume that $cw > 0$. If $w_j > 0$ for all $j \in K$, since $\widetilde{x}$ is feasible $\widetilde{x} \geq 0$, and so $x(t) = \widetilde{x} + tw \geq 0$ for all $t \geq 0$. Furthermore, since $Aw = 0$ and $\widetilde{x}$ is feasible, we have

$$Ax(t) = A\widetilde{x} + tAw = b,$$

and thus $x(t)$ is feasible for all $t \geq 0$. We also have

$$cx(t) = c\widetilde{x} + tcw.$$

Since $cw > 0$, as $t > 0$ goes to infinity the objective function $cx(t)$ also tends to infinity, contradicting the fact that is is bounded above. Therefore, some w satisfying Conditions (1) and (2) above must exist.

We show that there is some $t_0 > 0$ such that $cx(t_0) \geq cx_0$ and $x(t_0) = \widetilde{x} + t_0 w$ is feasible, yet $x(t_0)$ has more zero coordinates than $\widetilde{x}$, a contradiction.

Since $x(t) = \widetilde{x} + tw$, we have

$$x(t)_i = \widetilde{x}_i + tw_i,$$

so if we let $I = \{i \in \{1, \ldots, n\} \mid w_i < 0\} \subseteq K$, which is nonempty since w satisfies Condition (2) above, if we pick

$$t_0 = \min_{i \in I} \left\{ \frac{-\widetilde{x}_i}{w_i} \right\},$$

then $t_0 > 0$, because $w_i < 0$ for all $i \in I$, and by definition of K we have $\widetilde{x}_i > 0$ for all $i \in K$. By the definition of $t_0 > 0$ and since $\widetilde{x} \geq 0$, we have

$$x(t_0)_j = \widetilde{x}_j + t_0 w_j \geq 0 \quad \text{for all } j \in K,$$

so $x(t_0) \geq 0$, and $x(t_0)_i = 0$ for some $i \in I$. Since $Ax(t_0) = b$ (for any t), $x(t_0)$ is a feasible solution,

$$cx(t_0) = c\widetilde{x} + t_0 cw \geq cx_0 + t_0 cw \geq cx_0,$$

and $x(t_0)_i = 0$ for some $i \in I$, we see that $x(t_0)$ has more zero coordinates than $\widetilde{x}$, a contradiction. □

Proposition 25.3 implies the following important result.

Theorem 25.4. *Let (P_2) be any standard linear program with objective function cx, where $Ax = b$ and A is an $m \times n$ matrix of rank m. If (P_2) has some feasible solution and if it is bounded above, then some basic feasible solution $\widetilde{x}$ is an optimal solution of (P_2).*

Proof. By Proposition 25.3, for any feasible solution x there is some basic feasible solution $\widetilde{x}$ such that $cx \leq c\widetilde{x}$. But there are only finitely many basic feasible solutions, so one of them has to yield the maximum of the objective function. □

Geometrically, basic solutions are exactly the vertices of the polyhedron $\mathcal{P}(A, b)$, a notion that we now define.

Definition 25.6. Given an $\mathcal{H}$-polyhedron $\mathcal{P} \subseteq \mathbb{R}^n$, a *vertex* of $\mathcal{P}$ is a point $v \in \mathcal{P}$ with property that there is some nonzero linear form $c \in (\mathbb{R}^n)^*$ and some $\mu \in \mathbb{R}$, such that v is the unique point of $\mathcal{P}$ for which the map $x \mapsto cx$ has the maximum value μ ; that is, $cy < cv = \mu$ for all $y \in \mathcal{P} - \{v\}$. Geometrically this means that the hyperplane of equation $cy = \mu$ touches $\mathcal{P}$ exactly at v. More generally, a convex subset F of $\mathcal{P}$ is a *k-dimensional face* of $\mathcal{P}$ if F has dimension k and if there is some affine form $\varphi(x) = cx - \mu$ such that $cy = \mu$ for all $y \in F$, and $cy < \mu$ for all $y \in \mathcal{P} - F$. A 1-dimensional face is called an *edge*.

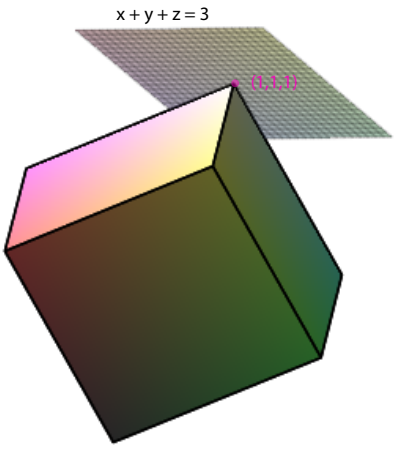


Figure 25.4: The cube centered at the origin with diagonal through $(-1,-1,-1)$ and $(1,1,1)$ has eight vertices. The vertex $(1,1,1)$ is associated with the linear form $x+y+z=3$.

The concept of a vertex is illustrated in Figure 25.4, while the concept of an edge is illustrated in Figure 25.5.

Since a k-dimensional face F of $\mathcal{P}$ is equal to the intersection of the hyperplane $H(\varphi)$ of equation $cx = \mu$ with $\mathcal{P}$, it is indeed convex and the notion of dimension makes sense. Observe that a 0-dimensional face of $\mathcal{P}$ is a vertex. If $\mathcal{P}$ has dimension d, then the $(d-1)$-dimensional faces of $\mathcal{P}$ are called its *facets*.

If (P) is a linear program in standard form, then its basic feasible solutions are exactly the vertices of the polyhedron $\mathcal{P}(A,b)$. To prove this fact we need the following simple proposition

Proposition 25.5. *Let $Ax = b$ be a linear system where A is an $m \times n$ matrix of rank m. For any subset $K \subseteq \{1,\ldots,n\}$ of size m, if A_K is invertible, then there is at most one basic feasible solution $x \in \mathbb{R}^n$ with $x_j = 0$ for all $j \notin K$ (of course, $x \geq 0$)*

Proof. In order for x to be feasible we must have $Ax = b$. Write $N = \{1,\ldots,n\} - K$, x_K for the vector consisting of the coordinates of x with indices in K, and x_N for the vector consisting of the coordinates of x with indices in N. Then

$$Ax = A_K x_K + A_N x_N = b.$$

In order for x to be a basic feasible solution we must have $x_N = 0$, so

$$A_K x_K = b.$$

Since by hypothesis A_K is invertible, $x_K = A_K^{-1} b$ is uniquely determined. If $x_K \geq 0$ then x is a basic feasible solution, otherwise it is not. This proves that there is at most one basic feasible solution $x \in \mathbb{R}^n$ with $x_j = 0$ for all $j \notin K$. □

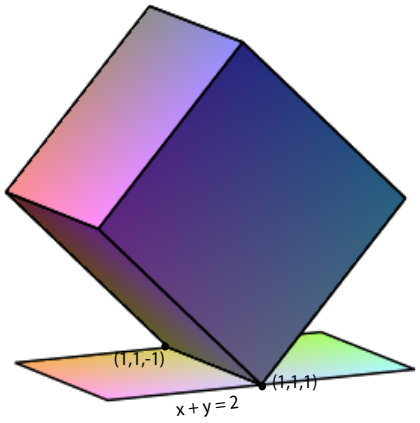


Figure 25.5: The cube centered at the origin with diagonal through $(-1,-1,-1)$ and $(1,1,1)$ has twelve edges. The vertex edge from $(1,1,-1)$ to $(1,1,1)$ is associated with the linear form $x+y=2$.

Theorem 25.6. *Let (P) be a linear program in standard form, where $Ax = b$ and A is an $m \times n$ matrix of rank m. For every $v \in \mathcal{P}(A,b)$, the following conditions are equivalent:*

(1) v is a vertex of the polyhedron $\mathcal{P}(A,b)$.

(2) v is a basic feasible solution of the linear program (P).

Proof. First, assume that v is a vertex of $\mathcal{P}(A,b)$, and let $\varphi(x) = cx - \mu$ be a linear form such that $cy < \mu$ for all $y \in \mathcal{P}(A,b)$ and $cv = \mu$. This means that v is the unique point of $\mathcal{P}(A,b)$ for which the objective function $x \mapsto cx$ has the maximum value μ on $\mathcal{P}(A,b)$, so by Theorem 25.4, since this maximum is achieved by some basic feasible solution, by uniqueness v must be a basic feasible solution.

Conversely, suppose v is a basic feasible solution of (P) corresponding to a subset $K \subseteq \{1,\ldots,n\}$ of size m. Let $\widehat{c} \in (\mathbb{R}^n)^*$ be the linear form defined by

$$\widehat{c}_j = \begin{cases} 0 & \text{if } j \in K \\ -1 & \text{if } j \notin K. \end{cases}$$

By construction $\widehat{c}\,v = 0$ and $\widehat{c}\,x \leq 0$ for any $x \geq 0$, hence the function $x \mapsto \widehat{c}\,x$ on $\mathcal{P}(A,B)$ has a maximum at v. Furthermore, $\widehat{c}\,x < 0$ for any $x \geq 0$ such that $x_j > 0$ for some $j \notin K$. However, by Proposition 25.5, the vector v is the only basic feasible solution such that $v_j = 0$ for all $j \notin K$, and therefore v is the only point of $\mathcal{P}(A,b)$ maximizing the function $x \mapsto \widehat{c}\,x$, so it is a vertex. □

In theory, to find an optimal solution we try all $\binom{n}{m}$ possible m-elements subsets K of $\{1, \ldots, n\}$ and solve for the corresponding unique solution x_K of $A_K x = b$. Then we check whether such a solution satisfies $x_K \geq 0$, compute cx_K, and return some feasible x_K for which the objective function is maximum. This is a totally impractible algorithm.

A practical algorithm is the *simplex algorithm*. Basically, the simplex algorithm tries to "climb" in the polyhderon $\mathcal{P}(A, b)$ from vertex to vertex along edges (using basic feasible solutions), trying to maximize the objective function. We present the simplex algorithm in the next chapter. The reader may also consult texts on linear programming. In particular, we recommend Matousek and Gardner [72], Chvatal [29], Papadimitriou and Steiglitz [79], Bertsimas and Tsitsiklis [17], Ciarlet [30], Schrijver [88], and Vanderbei [109].

Observe that Theorem 25.4 asserts that if a linear program (P) in standard form (where $Ax = b$ and A is an $m \times n$ matrix of rank m) has some feasible solution and is bounded above, then some basic feasible solution is an optimal solution. By Theorem 25.6, the polyhedron $\mathcal{P}(A, b)$ must have some vertex.

But suppose we only know that $\mathcal{P}(A, b)$ is nonempty; that is, we don't know that the objective function cx is bounded above. Does $\mathcal{P}(A, b)$ have some vertex?

The answer to the above question is *yes*, and this is important because the simplex algorithm needs an initial basic feasible solution to get started. Here we prove that if $\mathcal{P}(A, b)$ is nonempty, then it must contain a vertex. This proof still doesn't constructively yield a vertex, but we will see in the next chapter that the simplex algorithm always finds a vertex if there is one (provided that we use a pivot rule that prevents cycling).

Theorem 25.7. *Let (P) be a linear program in standard form, where $Ax = b$ and A is an $m \times n$ matrix of rank m. If $\mathcal{P}(A, b)$ is nonempty (there is a feasible solution), then $\mathcal{P}(A, b)$ has some vertex; equivalently, (P) has some basic feasible solution.*

Proof. The proof relies on a trick, which is to add slack variables $x_{n+1}, \ldots, x_{n+m}$ and use the new objective function $-(x_{n+1} + \cdots + x_{n+m})$.

If we let $\widehat{A}$ be the $m \times (m + n)$-matrix, and x, $\overline{x}$, and $\widehat{x}$ be the vectors given by

$$\widehat{A} = \begin{pmatrix} A & I_m \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n, \quad \overline{x} = \begin{pmatrix} x_{n+1} \\ \vdots \\ x_{n+m} \end{pmatrix} \in \mathbb{R}^m, \quad \widehat{x} = \begin{pmatrix} x \\ \overline{x} \end{pmatrix} \in \mathbb{R}^{n+m},$$

then consider the linear program $(\widehat{P})$ in standard form

$$\begin{aligned} &\text{maximize} && -(x_{n+1} + \cdots + x_{n+m}) \\ &\text{subject to} && \widehat{A}\widehat{x} = b \text{ and } \widehat{x} \geq 0. \end{aligned}$$

Since $x_i \geq 0$ for all i, the objective function $-(x_{n+1} + \cdots + x_{n+m})$ is bounded above by 0. The system $\widehat{A}\widehat{x} = b$ is equivalent to the system

$$Ax + \overline{x} = b,$$

so for every feasible solution $u \in \mathcal{P}(A, b)$, since $Au = b$, the vector $(u, 0_m)$ is also a feasible solution of $(\widehat{P})$, in fact an optimal solution since the value of the objective function $-(x_{n+1} + \cdots + x_{n+m})$ for $\overline{x} = 0$ is 0. By Proposition 25.3, the linear program $(\widehat{P})$ has some basic feasible solution (u^*, w^*) for which the value of the objective function is greater than or equal to the value of the objective function for $(u, 0_m)$, and since $(u, 0_m)$ is an optimal solution, (u^*, w^*) is also an optimal solution of $(\widehat{P})$. This implies that $w^* = 0$, since otherwise the objective function $-(x_{n+1} + \cdots + x_{n+m})$ would have a strictly negative value.

Therefore, $(u^*, 0_m)$ is a basic feasible solution of $(\widehat{P})$, and thus the columns corresponding to nonzero components of u^* are linearly independent. Some of the coordinates of u^* could be equal to 0, but since A has rank m we can add columns of A to obtain a basis K associated with u^*, and u^* is indeed a basic feasible solution of (P). □

The definition of a basic feasible solution can be adapted to linear programs where the constraints are of the form $Ax \leq b$, $x \geq 0$; see Matousek and Gardner [72] (Chapter 4, Section 4, Definition 4.4.2).

The most general type of linear program allows constraints of the form $a_i x \geq b_i$ or $a_i x = b_i$ besides constraints of the form $a_i x \leq b_i$. The variables x_i may also take negative values. It is always possible to convert such programs to the type considered in Definition 25.1. We proceed as follows.

Every constraint $a_i x \geq b_i$ is replaced by the constraint $-a_i x \leq -b_i$. Every equality constraint $a_i x = b_i$ is replaced by the two constraints $a_i x \leq b_i$ and $-a_i x \leq -b_i$.

If there are n variables x_i, we create n new variables y_i and n new variables z_i and replace every variable x_i by $y_i - z_i$. We also add the $2n$ constraints $y_i \geq 0$ and $z_i \geq 0$. If the constraints are given by the inequalities $Ax \leq b$, we now have constraints given by

$$\begin{pmatrix} A & -A \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} \leq b, \quad y \geq 0,\ z \geq 0.$$

We replace the objective function cx by $cy - cz$.

Remark: We also showed that we can replace the ineqality constraints $Ax \leq b$ by equality constraints $Ax = b$, by adding slack variables constrained to be nonnegative.

Chapter 26

The Simplex Algorithm

26.1 The Idea Behind the Simplex Algorithm

The simplex algorithm, due to Dantzig, applies to a linear program (P) in standard form, where the constraints are given by $Ax = b$ and $x \geq 0$, with A a $m \times n$ matrix of rank m, and with an objective function $c \mapsto cx$. This algorithm either reports that (P) has no feasible solution, or that (P) is unbounded, or yields an optimal solution. Geometrically, the algorithm climbs from vertex to vertex in the polyhedron $\mathcal{P}(A, b)$, trying to improve the value of the objective function. Since vertices correspond to basic feasible solutions, the simplex algorithm actually works with basic feasible solutions.

Recall that a basic feasible solution x is a feasible solution for which there is a subset $K \subseteq \{1, \ldots, n\}$ of size m such that the matrix A_K consisting of the columns of A whose indices belong to K are linearly independent, and that $x_j = 0$ for all $j \notin K$. We also let $J_>(x)$ be the set of indices

$$J_>(x) = \{j \in \{1, \ldots, n\} \mid x_j > 0\},$$

so for a basic feasible solution x associated with K, we have $J_>(x) \subseteq K$. In fact, by Proposition 25.2, a feasible solution x is a basic feasible solution iff the columns of $A_{J_>(x)}$ are linearly independent.

If $J_>(x)$ had cardinality m for all basic feasible solutions x, then the simplex algorithm would make progress at every step, in the sense that it would strictly increase the value of the objective function. Unfortunately, it is possible that $|J_>(x)| < m$ for certain basic feasible solutions, and in this case a step of the simplex algorithm may not increase the value of the objective function. Worse, in rare cases, it is possible that the algorithm enters an infinite loop. This phenomenon called *cycling* can be detected, but in this case the algorithm fails to give a conclusive answer.

Fortunately, there are ways of preventing the simplex algorithm from cycling (for example, Bland's rule discussed later), although proving that these rules work correctly is quite involved.

The potential "bad" behavior of a basic feasible solution is recorded in the following definition.

Definition 26.1. Given a linear program (P) in standard form where the constraints are given by $Ax = b$ and $x \geq 0$, with A an $m \times n$ matrix of rank m, a basic feasible solution x is *degenerate* if $|J_>(x)| < m$, otherwise it is *nondegenerate*.

The origin 0_n, if it is a basic feasible solution, is degenerate. For a less trivial example, $x = (0, 0, 0, 2)$ is a degenerate basic feasible solution of the following linear program in which $m = 2$ and $n = 4$.

Example 26.1.

$$\begin{aligned} &\text{maximize} && x_2 \\ &\text{subject to} && \\ & && -x_1 + x_2 + x_3 = 0 \\ & && x_1 + x_4 = 2 \\ & && x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0,\ x_4 \geq 0. \end{aligned}$$

The matrix A and the vector b are given by

$$A = \begin{pmatrix} -1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 2 \end{pmatrix},$$

and if $x = (0, 0, 0, 2)$, then $J_>(x) = \{4\}$. There are two ways of forming a set of two linearly independent columns of A containing the fourth column.

Given a basic feasible solution x associated with a subset K of size m, since the columns of the matrix A_K are linearly independent, by abuse of language we call the columns of A_K a *basis* of x.

If u is a vertex of (P), that is, a basic feasible solution of (P) associated with a basis K (of size m), in "normal mode," the simplex algorithm tries to move along an edge from the vertex u to an adjacent vertex v (with $u, v \in \mathcal{P}(A, b) \subseteq \mathbb{R}^n$) corresponding to a basic feasible solution whose basis is obtained by replacing one of the basic vectors A^k with $k \in K$ by another nonbasic vector A^j for some $j \notin K$, in such a way that the value of the objective function is increased.

Let us demonstrate this process on an example.

Example 26.2. Let (P) be the following linear program in standard form.

$$\begin{aligned} &\text{maximize} && x_1 + x_2 \\ &\text{subject to} && \\ & && -x_1 + x_2 + x_3 = 1 \\ & && x_1 + x_4 = 3 \\ & && x_2 + x_5 = 2 \\ & && x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0,\ x_4 \geq 0,\ x_5 \geq 0. \end{aligned}$$

The matrix A and the vector b are given by

$$A = \begin{pmatrix} -1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}.$$

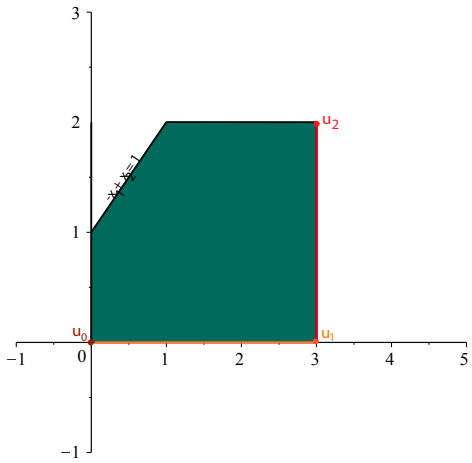


Figure 26.1: The planar $\mathcal{H}$-polyhedron associated with Example 26.2. The initial basic feasible solution is the origin. The simplex algorithm first moves along the horizontal orange line to feasible solution at vertex u_1. It then moves along the vertical red line to obtain the optimal feasible solution u_2.

The vector $u_0 = (0, 0, 1, 3, 2)$ corresponding to the basis $K = \{3, 4, 5\}$ is a basic feasible solution, and the corresponding value of the objective function is $0 + 0 = 0$. Since the columns (A^3, A^4, A^5) corresponding to $K = \{3, 4, 5\}$ are linearly independent we can express A^1 and A^2 as

$$\begin{aligned} A^1 &= -A^3 + A^4 \\ A^2 &= A^3 + A^5. \end{aligned}$$

Since

$$1A^3 + 3A^4 + 2A^5 = Au_0 = b,$$

for any $\theta \in \mathbb{R}$, we have

$$\begin{aligned} b &= 1A^3 + 3A^4 + 2A^5 - \theta A^1 + \theta A^1 \\ &= 1A^3 + 3A^4 + 2A^5 - \theta(-A^3 + A^4) + \theta A^1 \\ &= \theta A^1 + (1 + \theta)A^3 + (3 - \theta)A^4 + 2A^5, \end{aligned}$$

and

$$\begin{aligned} b &= 1A^3 + 3A^4 + 2A^5 - \theta A^2 + \theta A^2 \\ &= 1A^3 + 3A^4 + 2A^5 - \theta(A^3 + A^5) + \theta A^1 \\ &= \theta A^2 + (1-\theta)A^3 + 3A^4 + (2-\theta)A^5. \end{aligned}$$

In the first case, the vector $(\theta, 0, 1+\theta, 3-\theta, 2)$ is a feasible solution iff $0 \leq \theta \leq 3$, and the new value of the objective function is θ.

In the second case, the vector $(0, \theta, 1-\theta, 3, 2-\theta, 1)$ is a feasible solution iff $0 \leq \theta \leq 1$, and the new value of the objective function is also θ.

Consider the first case. It is natural to ask whether we can get another vertex and increase the objective function by setting to zero one of the coordinates of $(\theta, 0, 1+\theta, 3-\theta, 2)$, in this case the fouth one, by picking $\theta = 3$. This yields the feasible solution $(3, 0, 4, 0, 2)$, which corresponds to the basis (A^1, A^3, A^5), and so is indeed a basic feasible solution, with an improved value of the objective function equal to 3. Note that A^4 *left* the basis (A^3, A^4, A^5) and A^1 *entered* the new basis (A^1, A^3, A^5).

We can now express A^2 and A^4 in terms of the basis (A^1, A^3, A^5), which is easy to do since we already have A^1 and A^2 in term of (A^3, A^4, A^5), and A^1 and A^4 are swapped. Such a step is called a *pivoting step*. We obtain

$$\begin{aligned} A^2 &= A^3 + A^5 \\ A^4 &= A^1 + A^3. \end{aligned}$$

Then we repeat the process with $u_1 = (3, 0, 4, 0, 2)$ and the basis (A^1, A^3, A^5). We have

$$\begin{aligned} b &= 3A^1 + 4A^3 + 2A^5 - \theta A^2 + \theta A^2 \\ &= 3A^1 + 4A^3 + 2A^5 - \theta(A^3 + A^5) + \theta A^2 \\ &= 3A^1 + \theta A^2 + (4-\theta)A^3 + (2-\theta)A^5, \end{aligned}$$

and

$$\begin{aligned} b &= 3A^1 + 4A^3 + 2A^5 - \theta A^4 + \theta A^4 \\ &= 3A^1 + 4A^3 + 2A^5 - \theta(A^1 + A^3) + \theta A^4 \\ &= (3-\theta)A^1 + (4-\theta)A^3 + \theta A^4 + 2A^5. \end{aligned}$$

In the first case, the point $(3, \theta, 4-\theta, 0, 2-\theta)$ is a feasible solution iff $0 \leq \theta \leq 2$, and the new value of the objective function is $3+\theta$. In the second case, the point $(3-\theta, 0, 4-\theta, \theta, 2)$ is a feasible solution iff $0 \leq \theta \leq 3$, and the new value of the objective function is $3-\theta$. To increase the objective function we must choose the first case and we pick $\theta = 2$. Then, we get the feasible solution $u_2 = (3, 2, 2, 0, 0)$, which corresponds to the basis (A^1, A^2, A^3), and thus is a basic feasible solution. The new value of the objective function is 5.

Next we express A^4 and A^5 in terms of the basis (A^1, A^2, A^3). Again this is easy to do since we just swapped A^5 and A^2 (a pivoting step), and we get

$$A^5 = A^2 - A^3$$
$$A^4 = A^1 + A^3.$$

We repeat the process with $u_2 = (3, 2, 2, 0, 0)$ and the basis (A^1, A^2, A^3). We have

$$\begin{aligned} b &= 3A^1 + 2A^2 + 2A^3 - \theta A^4 + \theta A^4 \\ &= 3A^1 + 2A^2 + 2A^3 - \theta(A^1 + A^3) + \theta A^4 \\ &= (3 - \theta)A^1 + 2A^2 + (2 - \theta)A^3 + \theta A^4, \end{aligned}$$

and

$$\begin{aligned} b &= 3A^1 + 2A^2 + 2A^3 - \theta A^5 + \theta A^5 \\ &= 3A^1 + 2A^2 + 2A^3 - \theta(A^2 - A^3) + \theta A^5 \\ &= 3A^1 + (2 - \theta)A^2 + (2 + \theta)A^3 + \theta A^5. \end{aligned}$$

In the first case, the point $(3 - \theta, 2, 2 - \theta, \theta, 0)$ is a feasible solution iff $0 \leq \theta \leq 2$, and the value of the objective function is $5 - \theta$. In the second case, the point $(3, 2 - \theta, 2 + \theta, 0, \theta)$ is a feasible solution iff $0 \leq \theta \leq 2$, and the value of the objective function is also $5 - \theta$. Since we must have $\theta \geq 0$ to have a feasible solution, there is no way to increase the objective function. In this situation, it turns out that we have reached an optimal solution, in our case $u_2 = (3, 2, 2, 0, 0)$, with the maximum of the objective function equal to 5.

We could also have applied the simplex algorithm to the vertex $u_0 = (0, 0, 1, 3, 2)$ and to the vector $(0, \theta, 1 - \theta, 3, 2 - \theta, 1)$, which is a feasible solution iff $0 \leq \theta \leq 1$, with new value of the objective function θ. By picking $\theta = 1$, we obtain the feasible solution $(0, 1, 0, 3, 1)$, corresponding to the basis (A^2, A^4, A^5), which is indeed a vertex. The new value of the objective function is 1. Then we express A^1 and A^3 in terms the basis (A^2, A^4, A^5) obtaining

$$A^1 = A^4 - A^3$$
$$A^3 = A^2 - A^5,$$

and repeat the process with $(0, 1, 0, 3, 1)$ and the basis (A^2, A^4, A^5). After three more steps we will reach the optimal solution $u_2 = (3, 2, 2, 0, 0)$.

Let us go back to the linear program of Example 26.1 with objective function x_2 and where the matrix A and the vector b are given by

$$A = \begin{pmatrix} -1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 2 \end{pmatrix}.$$

Recall that $u_0 = (0, 0, 0, 2)$ is a degenerate basic feasible solution, and the objective function has the value 0. See Figure 26.2 for a planar picture of the $\mathcal{H}$-polyhedron associated with Example 26.1.

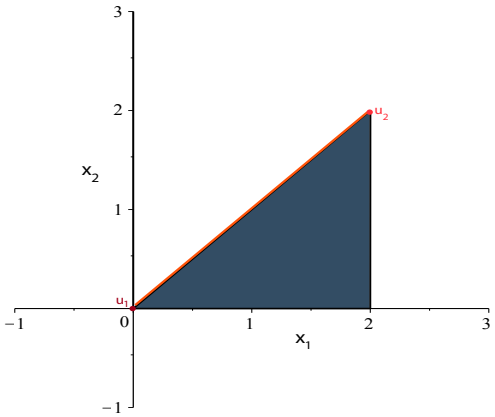


Figure 26.2: The planar $\mathcal{H}$-polyhedron associated with Example 26.1. The initial basic feasible solution is the origin. The simplex algorithm moves along the slanted orange line to the apex of the triangle.

Pick the basis (A^3, A^4). Then we have

$$\begin{aligned} A^1 &= -A^3 + A^4 \\ A^2 &= A^3, \end{aligned}$$

and we get

$$\begin{aligned} b &= 2A^4 - \theta A^1 + \theta A^1 \\ &= 2A^4 - \theta(-A^3 + A^4) + \theta A^1 \\ &= \theta A^1 + \theta A^3 + (2-\theta)A^4, \end{aligned}$$

and

$$\begin{aligned} b &= 2A^4 - \theta A^2 + \theta A^2 \\ &= 2A^4 - \theta A^3 + \theta A^2 \\ &= \theta A^2 - \theta A^3 + 2A^4. \end{aligned}$$

In the first case, the point $(\theta, 0, \theta, 2-\theta)$ is a feasible solution iff $0 \leq \theta \leq 2$, and the value of the objective function is 0, and in the second case the point $(0, \theta, -\theta, 2)$ is a feasible solution iff $\theta = 0$, and the value of the objective function is θ. However, since we must have $\theta = 0$ in the second case, there is no way to increase the objective function either.

It turns out that in order to make the cases considered by the simplex algorithm as mutually exclusive as possible, since in the second case the coefficient of θ in the value of the objective function is nonzero, namely 1, we should choose the second case. We must

pick $\theta = 0$, but we can swap the vectors A^3 and A^2 (because A^2 is coming in and A^3 has the coefficient $-\theta$, which is the reason why θ must be zero), and we obtain the basic feasible solution $u_1 = (0, 0, 0, 2)$ with the new basis (A^2, A^4). Note that this basic feasible solution corresponds to the same vertex $(0, 0, 0, 2)$ as before, but the basis has changed. The vectors A^1 and A^3 can be expressed in terms of the basis (A^2, A^4) as

$$\begin{aligned} A^1 &= -A^2 + A^4 \\ A^3 &= A^2. \end{aligned}$$

We now repeat the procedure with $u_1 = (0, 0, 0, 2)$ and the basis (A^2, A^4), and we get

$$\begin{aligned} b &= 2A^4 - \theta A^1 + \theta A^1 \\ &= 2A^4 - \theta(-A^2 + A^4) + \theta A^1 \\ &= \theta A^1 + \theta A^2 + (2 - \theta)A^4, \end{aligned}$$

and

$$\begin{aligned} b &= 2A^4 - \theta A^3 + \theta A^3 \\ &= 2A^4 - \theta A^2 + \theta A^3 \\ &= -\theta A^2 + \theta A^3 + 2A^4. \end{aligned}$$

In the first case, the point $(\theta, \theta, 0, 2-\theta)$ is a feasible solution iff $0 \leq \theta \leq 2$ and the value of the objective function is θ, and in the second case the point $(0, -\theta, \theta, 2)$ is a feasible solution iff $\theta = 0$ and the value of the objective function is θ. In order to increase the objective function we must choose the first case and pick $\theta = 2$. We obtain the feasible solution $u_2 = (2, 2, 0, 0)$ whose corresponding basis is (A^1, A^2) and the value of the objective function is 2.

The vectors A^3 and A^4 are expressed in terms of the basis (A^1, A^2) as

$$\begin{aligned} A^3 &= A^2 \\ A^4 &= A^1 + A^3, \end{aligned}$$

and we repeat the procedure with $u_2 = (2, 2, 0, 0)$ and the basis (A^1, A^2). We get

$$\begin{aligned} b &= 2A^1 + 2A^2 - \theta A^3 + \theta A^3 \\ &= 2A^1 + 2A^2 - \theta A^2 + \theta A^3 \\ &= 2A^1 + (2 - \theta)A^2 + \theta A^3, \end{aligned}$$

and

$$\begin{aligned} b &= 2A^1 + 2A^2 - \theta A^4 + \theta A^4 \\ &= 2A^1 + 2A^2 - \theta(A^1 + A^3) + \theta A^4 \\ &= (2 - \theta)A^1 + 2A^2 - \theta A^3 + \theta A^4. \end{aligned}$$

In the first case, the point $(2, 2-\theta, 0, \theta)$ is a feasible solution iff $0 \leq \theta \leq 2$ and the value of the objective function is $2-\theta$, and in the second case, the point $(2-\theta, 2, -\theta, \theta)$ is a feasible solution iff $\theta = 0$ and the value of the objective function is 2. This time there is no way to improve the objective function and we have reached an optimal solution $u_2 = (2, 2, 0, 0)$ with the maximum of the objective function equal to 2.

Let us now consider an example of an unbounded linear program.

Example 26.3. Let (P) be the following linear program in standard form.

$$\begin{aligned} &\text{maximize} \quad x_1 \\ &\text{subject to} \\ &\qquad x_1 - x_2 + x_3 = 1 \\ &\qquad -x_1 + x_2 + x_4 = 2 \\ &\qquad x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0,\ x_4 \geq 0. \end{aligned}$$

The matrix A and the vector b are given by

$$A = \begin{pmatrix} 1 & -1 & 1 & 0 \\ -1 & 1 & 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

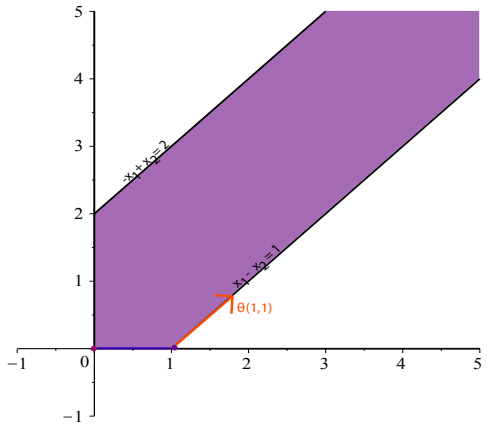


Figure 26.3: The planar $\mathcal{H}$-polyhedron associated with Example 26.3. The initial basic feasible solution is the origin. The simplex algorithm first moves along the horizontal indigo line to basic feasible solution at vertex $(1, 0)$. Any optimal feasible solution occurs by moving along the boundary line parameterized by the orange arrow $\theta(1, 1)$.

The vector $u_0 = (0, 0, 1, 2)$ corresponding to the basis $K = \{3, 4\}$ is a basic feasible solution, and the corresponding value of the objective function is 0. The vectors A^1 and A^2

are expressed in terms of the basis (A^3, A^4) by

$$\begin{aligned} A^1 &= A^3 - A^4 \\ A^2 &= -A^3 + A^4. \end{aligned}$$

Starting with $u_0 = (0, 0, 1, 2)$, we get

$$\begin{aligned} b &= A^3 + 2A^4 - \theta A^1 + \theta A^1 \\ &= A^3 + 2A^4 - \theta(A^3 - A^4) + \theta A^1 \\ &= \theta A^1 + (1 - \theta)A^3 + (2 + \theta)A^4, \end{aligned}$$

and

$$\begin{aligned} b &= A^3 + 2A^4 - \theta A^2 + \theta A^2 \\ &= A^3 + 2A^4 - \theta(-A^3 + A^4) + \theta A^2 \\ &= \theta A^2 + (1 + \theta)A^3 + (2 - \theta)A^4. \end{aligned}$$

In the first case, the point $(\theta, 0, 1 - \theta, 2 + \theta)$ is a feasible solution iff $0 \leq \theta \leq 1$ and the value of the objective function is θ, and in the second case, the point $(0, \theta, 1 + \theta, 2 - \theta)$ is a feasible solution iff $0 \leq \theta \leq 2$ and the value of the objective function is 0. In order to increase the objective function we must choose the first case, and we pick $\theta = 1$. We get the feasible solution $u_1 = (1, 0, 0, 3)$ corresponding to the basis (A^1, A^4), so it is a basic feasible solution, and the value of the objective function is 1.

The vectors A^2 and A^3 are given in terms of the basis (A^1, A^4) by

$$\begin{aligned} A^2 &= -A^1 \\ A^3 &= A^1 + A^4. \end{aligned}$$

Repeating the process with $u_1 = (1, 0, 0, 3)$, we get

$$\begin{aligned} b &= A^1 + 3A^4 - \theta A^2 + \theta A^2 \\ &= A^1 + 3A^4 - \theta(-A^1) + \theta A^2 \\ &= (1 + \theta)A^1 + \theta A^2 + 3A^4, \end{aligned}$$

and

$$\begin{aligned} b &= A^1 + 3A^4 - \theta A^3 + \theta A^3 \\ &= A^1 + 3A^4 - \theta(A^1 + A^4) + \theta A^3 \\ &= (1 - \theta)A^1 + \theta A^3 + (3 - \theta)A^4. \end{aligned}$$

In the first case, the point $(1 + \theta, \theta, 0, 3)$ is a feasible solution for all $\theta \geq 0$ and the value of the objective function if $1 + \theta$, and in the second case, the point $(1 - \theta, 0, \theta, 3 - \theta)$ is a

feasible solution iff $0 \leq \theta \leq 1$ and the value of the objective function is $1 - \theta$. This time, we are in the situation where the points

$$(1 + \theta, \theta, 0, 3) = (1, 0, 0, 3) + \theta(1, 1, 0, 0), \quad \theta \geq 0$$

form an infinite ray in the set of feasible solutions, and the objective function $1 + \theta$ is unbounded from above on this ray. This indicates that our linear program, although feasible, is unbounded.

Let us now describe a step of the simplex algorithm in general.

26.2 The Simplex Algorithm in General

We assume that we already have an initial vertex u_0 to start from. This vertex corresponds to a basic feasible solution with basis K_0. We will show later that it is always possible to find a basic feasible solution of a linear program (P) is standard form, or to detect that (P) has no feasible solution.

The idea behind the simplex algorithm is this: Given a pair (u, K) consisting of a basic feasible solution u and a basis K for u, find another pair (u^+, K^+) consisting of another basic feasible solution u^+ and a basis K^+ for u^+, such that K^+ is obtained from K by deleting some basic index $k^- \in K$ and adding some nonbasic index $j^+ \notin K$, in such a way that the value of the objective function increases (preferably strictly). The step which consists in swapping the vectors A^{k^-} and A^{j^+} is called a *pivoting step*.

Let u be a given vertex corresponds to a basic feasible solution with basis K. Since the m vectors A^k corresponding to indices $k \in K$ are linearly independent, they form a basis, so for every nonbasic $j \notin K$, we write

$$A^j = \sum_{k \in K} \gamma_k^j A^k. \tag{$*$}$$

We let $\gamma_K^j \in \mathbb{R}^m$ be the vector given by $\gamma_K^j = (\gamma_k^j)_{k \in K}$. Actually, since the vector γ_K^j depends on K, to be very precise we should denote its components by $(\gamma_K^j)_k$, but to simplify notation we usually write γ_k^j instead of $(\gamma_K^j)_k$ (unless confusion arises). We will explain later how the coefficients γ_k^j can be computed efficiently.

Since u is a feasible solution we have $u \geq 0$ and $Au = b$, that is,

$$\sum_{k \in K} u_k A^k = b. \tag{$**$}$$

For every nonbasic $j \notin K$, a candidate for entering the basis K, we try to find a new vertex $u(\theta)$ that improves the objective function, and for this we add $-\theta A^j + \theta A^j = 0$ to b in

the equation $(**)$ and then replace the occurrence of A^j in $-\theta A^j$ by the right hand side of equation $(*)$ to obtain

$$\begin{aligned} b &= \sum_{k\in K} u_k A^k - \theta A^j + \theta A^j \\ &= \sum_{k\in K} u_k A^k - \theta\Big(\sum_{k\in K} \gamma_k^j A^k\Big) + \theta A^j \\ &= \sum_{k\in K} \Big(u_k - \theta\gamma_k^j\Big) A^k + \theta A^j. \end{aligned}$$

Consequently, the vector $u(\theta)$ appearing on the right-hand side of the above equation given by

$$u(\theta)_i = \begin{cases} u_i - \theta\gamma_i^j & \text{if } i \in K \\ \theta & \text{if } i = j \\ 0 & \text{if } i \notin K \cup \{j\} \end{cases}$$

automatically satisfies the constraints $Au(\theta) = b$, and this vector is a feasible solution iff

$$\theta \geq 0 \quad \text{and} \quad u_k \geq \theta\gamma_k^j \quad \text{for all } k \in K.$$

Obviously $\theta = 0$ is a solution, and if

$$\theta^j = \min\left\{ \frac{u_k}{\gamma_k^j} \;\middle|\; \gamma_k^j > 0,\ k \in K \right\} > 0,$$

then we have a range of feasible solutions for $0 \leq \theta \leq \theta^j$. The value of the objective function for $u(\theta)$ is

$$cu(\theta) = \sum_{k\in K} c_k(u_k - \theta\gamma_k^j) + \theta c_j = cu + \theta\Big(c_j - \sum_{k\in K} \gamma_k^j c_k\Big).$$

Since the potential change in the objective function is

$$\theta\Big(c_j - \sum_{k\in K} \gamma_k^j c_k\Big)$$

and $\theta \geq 0$, if $c_j - \sum_{k\in K} \gamma_k^j c_k \leq 0$ then the objective function can't be increased.

However, if $c_{j^+} - \sum_{k\in K} \gamma_k^{j^+} c_k > 0$ for some $j^+ \notin K$, and if $\theta^{j^+} > 0$, then the objective function can be strictly increased by choosing any $\theta > 0$ such that $\theta \leq \theta^{j^+}$, so it is natural to zero at least one coefficient of $u(\theta)$ by picking $\theta = \theta^{j^+}$, which also maximizes the increase of the objective function. In this case (Case below (B2)), we obtain a new feasible solution $u^+ = u(\theta^{j^+})$.

Now, if $\theta^{j^+} > 0$, then there is some index $k \in K$ such $u_k > 0$, $\gamma_k^{j^+} > 0$, and $\theta^{j^+} = u_k/\gamma_k^{j^+}$, so we can pick such an index k^- for the vector A^{k^-} leaving the basis K. We claim that

$K^+ = (K - \{k^-\}) \cup \{j^+\}$ is a basis. This is because the coefficient $\gamma_{k^-}^{j^+}$ associated with the column A^{k^-} is nonzero (in fact, $\gamma_{k^-}^{j^+} > 0$), so equation $(*)$, namely

$$A^{j^+} = \gamma_{k^-}^{j^+} A^{k^-} + \sum_{k \in K - \{k^-\}} \gamma_k^{j^+} A^k,$$

yields the equation

$$A^{k^-} = (\gamma_{k^-}^{j^+})^{-1} A^{j^+} - \sum_{k \in K - \{k^-\}} (\gamma_{k^-}^{j^+})^{-1} \gamma_k^{j^+} A^k,$$

and these equations imply that the subspaces spanned by the vectors $(A^k)_{k \in K}$ and the vectors $(A^k)_{k \in K^+}$ are identical. However, K is a basis of dimension m so this subspace has dimension m, and since K^+ also has m elements, it must be a basis. Therefore, $u^+ = u(\theta^{j^+})$ is a basic feasible solution.

The above case is the most common one, but other situations may arise. In what follows, we discuss all eventualities.

Case (A).

We have $c_j - \sum_{k \in K} \gamma_k^j c_k \leq 0$ for all $j \notin K$. Then it turns out that u is an *optimal solution*. Otherwise, we are in Case (B).

Case (B).

We have $c_j - \sum_{k \in K} \gamma_k^j c_k > 0$ for some $j \notin K$ (not necessarily unique). There are three subcases.

Case (B1).

If for some $j \notin K$ as above we also have $\gamma_k^j \leq 0$ for all $k \in K$, since $u_k \geq 0$ for all $k \in K$, this places no restriction on θ, and the objective function is *unbounded above*.

Case (B2).

There is some index $j^+ \notin K$ such that simultaneously

(1) $c_{j^+} - \sum_{k \in K} \gamma_k^{j^+} c_k > 0$, which means that the objective function can potentially be increased;

(2) There is some $k \in K$ such that $\gamma_k^{j^+} > 0$, and for every $k \in K$, if $\gamma_k^{j^+} > 0$ then $u_k > 0$, which implies that $\theta^{j^+} > 0$.

If we pick $\theta = \theta^{j^+}$ where

$$\theta^{j^+} = \min\left\{ \frac{u_k}{\gamma_k^{j^+}} \;\middle|\; \gamma_k^{j^+} > 0,\ k \in K \right\} > 0,$$

then the feasible solution u^+ given by

$$u_i^+ = \begin{cases} u_i - \theta^{j^+}\gamma_i^{j^+} & \text{if } i \in K \\ \theta^{j^+} & \text{if } i = j^+ \\ 0 & \text{if } i \notin K \cup \{j^+\} \end{cases}$$

is a vertex of $\mathcal{P}(A, b)$. If we pick any index $k^- \in K$ such that $\theta^{j^+} = u_{k^-}/\gamma_{k^-}^{j+}$, then $K^+ = (K - \{k^-\}) \cup \{j^+\}$ is a basis for u^+. The vector A^{j^+} enters the new basis K^+, and the vector A^{k^-} leaves the old basis K. This is a *pivoting step*. The objective function increases strictly.

Case (B3).

There is some index $j \notin K$ such that $c_j - \sum_{k \in K} \gamma_k^j c_k > 0$, and for each of the indices $j \notin K$ satisfying the above property we have simultaneously

(1) $c_j - \sum_{k \in K} \gamma_k^j c_k > 0$, which means that the objective function can potentially be increased;

(2) There is some $k \in K$ such that $\gamma_k^j > 0$, and $u_k = 0$, which implies that $\theta^j = 0$.

Consequently, the objective function *does not change*. In this case, u is a degenerate basic feasible solution.

We can associate to $u^+ = u$ a new basis K^+ as follows: Pick any index $j^+ \notin K$ such that

$$c_{j^+} - \sum_{k \in K} \gamma_k^{j^+} c_k > 0,$$

and any index $k^- \in K$ such that

$$\gamma_{k^-}^{j^+} > 0,$$

and let $K^+ = (K - \{k^-\}) \cup \{j^+\}$. As in Case (B2), The vector A^{j^+} enters the new basis K^+, and the vector A^{k^-} leaves the old basis K. This is a *pivoting step*. However, the objective function *does not change* since $\theta^{j+} = 0$.

It is easy to prove that in Case (A) the basic feasible solution u is an optimal solution, and that in Case (B1) the linear program is unbounded. We already proved that in Case (B2) the vector u^+ and its basis K^+ constitutes a basic feasible solution, and the proof in Case (B3) is similar. For details, see Ciarlet [30] (Chapter 10).

It is convenient to reinterpret the various cases considered by introducing the followings sets:

$$B_1 = \Big\{ j \notin K \mid c_j - \sum_{k \in K} \gamma_k^j c_k > 0,\ \max_{k \in K} \gamma_k^j \leq 0 \Big\}$$

$$B_2 = \Big\{ j \notin K \mid c_j - \sum_{k \in K} \gamma_k^j c_k > 0,\ \max_{k \in K} \gamma_k^j > 0,\ \min\Big\{ \frac{u_k}{\gamma_k^j} \;\Big|\; k \in K, \gamma_k^j > 0 \Big\} > 0 \Big\}$$

$$B_3 = \Big\{ j \notin K \mid c_j - \sum_{k \in K} \gamma_k^j c_k > 0,\ \max_{k \in K} \gamma_k^j > 0,\ \min\Big\{ \frac{u_k}{\gamma_k^j} \;\Big|\; k \in K, \gamma_k^j > 0 \Big\} = 0 \Big\},$$

and

$$B = B_1 \cup B_2 \cup B_3 = \Big\{ j \notin K \mid c_j - \sum_{k \in K} \gamma_k^j c_k > 0 \Big\}.$$

Then it is easy to see that the following equivalences hold:

$$\begin{aligned} &\text{Case (A)} \Longleftrightarrow B = \emptyset, \qquad \text{Case (B)} \Longleftrightarrow B \neq \emptyset \\ &\text{Case (B1)} \Longleftrightarrow B_1 \neq \emptyset \\ &\text{Case (B2)} \Longleftrightarrow B_2 \neq \emptyset \\ &\text{Case (B3)} \Longleftrightarrow B_3 \neq \emptyset. \end{aligned}$$

Furthermore, (A) and (B), (B1) and (B3), (B2) and (B3) are mutually exclusive, while (B1) and (B2) are not.

If Case (B1) and Case (B2) arise simultaneously, we opt for Case (B1) which says that the linear program (P) is unbounded and terminate the algorithm.

Here are a few remarks about the method.

In Case (B2), which is the path followed by the algorithm most frequently, various choices have to be made for the index $j^+ \notin K$ for which $\theta^{j^+} > 0$ (the new index in K^+). Similarly, various choices have to be made for the index $k^- \in K$ leaving K, but such choices are typically less important.

Similarly in Case (B3), various choices have to be made for the new index $j^+ \notin K$ going into K^+. In Cases (B2) and (B3), criteria for making such choices are called *pivot rules*.

Case (B3) only arises when u is a degenerate vertex. But even if u is degenerate, Case (B2) may arise if $u_k > 0$ whenever $\gamma_k^j > 0$. It may also happen that u is nondegenerate but as a result of Case (B2), the new vertex u^+ is degenerate because at least two components $u_{k_1} - \theta^{j^+}\gamma_{k_1}^{j^+}$ and $u_{k_2} - \theta^{j^+}\gamma_{k_2}^{j^+}$ vanish for some distinct $k_1, k_2 \in K$.

Cases (A) and (B1) correspond to situations where the algorithm terminates, and Case (B2) can only arise a finite number of times during execution of the simplex algorithm, since the objective function is strictly increased from vertex to vertex and there are only finitely many vertices. Therefore, if the simplex algorithm is started on any initial basic feasible solution u_0, then one of three mutually exclusive situations may arise:

(1) There is a finite sequence of occurrences of Case (B2) and/or Case (B3) ending with an occurrence of Case (A). Then the last vertex produced by the algorithm is an optimal solution.

(2) There is a finite sequence of occurrences of Case (B2) and/or Case (B3) ending with an occurrence of Case (B1). We conclude that the problem is unbounded, and thus has no solution.

(3) There is a finite sequence of occurrences of Case (B2) and/or Case (B3), followed by an infinite sequence of Case (B3). If this occurs, the algorithm visits the some basis twice. This a phenomenon known as *cycling*. In this eventually the algorithm fails to come to a conclusion.

There are examples for which cycling occur, although this is rare in practice. Such an example is given in Chvatal [29]; see Chapter 3, pages 31-32, for an example with seven variables and three equations that cycles after six iterations under a certain pivot rule.

The third possibility can be avoided by the choice of a suitable pivot rule. Two of these rules are *Bland's rule* and the *lexicographic rule*; see Chvatal [29] (Chapter 3, pages 34-38).

Bland's rule says: choose the smallest of the elligible incoming indices $j^+ \notin K$, and similarly choose the smallest of the elligible outgoing indices $k^- \in K$.

It can be proved that cycling cannot occur if Bland's rule is chosen as the pivot rule. The proof is very technical; see Chvatal [29] (Chapter 3, pages 37-38), Matousek and Gardner [72] (Chapter 5, Theorem 5.8.1), and Papadimitriou and Steiglitz [79] (Section 2.7). Therefore, assuming that some initial basic feasible solution is provided, and using a suitable pivot rule (such as Bland's rule), the simplex algorithm always terminates and either yields an optimal solution or reports that the linear program is unbounded. Unfortunately Bland's rules is one of the slowest pivot rules.

The choice of a pivot rule affects greatly the number of pivoting steps that the simplex algorithms goes through. It is not our intention here to explain the various pivot rules. We simply mention the following rules, referring the reader to Matousek and Gardner [72] (Chapter 5, Section 5.7) or to the texts cited in Section 24.1.

1. Largest coefficient.
2. Largest increase.
3. Steepest edge.
4. Bland's Rule.
5. Random edge.

The steepest edge rule is one of the most popular. The idea is to maximize the ratio

$$\frac{c(u^+ - u)}{\|u^+ - u\|}.$$

The random edge rule picks the index $j^+ \notin K$ of the entering basis vector uniformly at random among all elligible indices.

Let us now return to the issue of the initialization of the simplex algorithm. We use the linear program $(\widehat{P})$ introduced during the proof of Theorem 25.7.

Consider a linear program $(P2)$

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax = b \text{ and } x \geq 0, \end{aligned}$$

in standard form where A is an $m \times n$ matrix of rank m.

First, observe that since the constraints are equations, we can ensure that $b \geq 0$, because evey equation $a_i x = b_i$ where $b_i < 0$ can be replaced by $-a_i x = -b_i$. The next step is to introduce the linear program $(\widehat{P})$ in standard form

$$\begin{aligned} &\text{maximize} \quad -(x_{n+1} + \cdots + x_{n+m}) \\ &\text{subject to} \quad \widehat{A}\,\widehat{x} = b \text{ and } \widehat{x} \geq 0, \end{aligned}$$

where $\widehat{A}$ and $\widehat{x}$ are given by

$$\widehat{A} = \begin{pmatrix} A & I_m \end{pmatrix}, \quad \widehat{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_{n+m} \end{pmatrix}.$$

Since we assumed that $b \geq 0$, the vector $\widehat{x} = (0_n, b)$ is a feasible solution of $(\widehat{P})$, in fact a basic feasible solutions since the matrix associated with the indices $n+1, \ldots, n+m$ is the identity matrix I_m. Furthermore, since $x_i \geq 0$ for all i, the objective function $-(x_{n+1} + \cdots + x_{n+m})$ is bounded above by 0.

If we execute the simplex algorithm with a pivot rule that prevents cycling, starting with the basic feasible solution $(0_n, d)$, since the objective function is bounded by 0, the simplex algorithm terminates with an optimal solution given by some basic feasible solution, say (u^*, w^*), with $u^* \in \mathbb{R}^n$ and $w^* \in \mathbb{R}^m$.

As in the proof of Theorem 25.7, for every feasible solution $u \in \mathcal{P}(A, b)$ the vector $(u, 0_m)$ is an optimal solution of $(\widehat{P})$. Therefore, if $w^* \neq 0$, then $\mathcal{P}(A, b) = \emptyset$, since otherwise for every feasible solution $u \in \mathcal{P}(A, b)$ the vector $(u, 0_m)$ would yield a value of the objective function $-(x_{n+1} + \cdots + x_{n+m})$ equal to 0, but (u^*, w^*) yields a strictly negative value since $w^* \neq 0$.

Otherwise, $w^* = 0$, and u^* is a feasible solution of (P). Since $(u^*, 0_m)$ is a basic feasible solution of $(\widehat{P})$ the columns corresponding to nonzero components of u^* are linearly independent. Some of the coordinates of u^* could be equal to 0, but since A has rank m we can add columns of A to obtain a basis K^* associated with u^*, and u^* is indeed a basic feasible solution of (P).

Running the simplex algorithm on the linear program $\widehat{P}$ to obtain an initial feasible solution (u_0, K_0) of the linear program $(P2)$ is called *Phase I* of the simplex algorithm. Running the simplex algorithm on the linear program $(P2)$ with some initial feasible solution

(u_0, K_0) is called *Phase II* of the simplex algorithm. If a feasible solution of the linear program $(P2)$ is readily available then Phase I is skipped. Sometimes, at the end of Phase I, an optimal solution of $(P2)$ is already obtained.

In summary, we proved the following fact worth recording.

Proposition 26.1. *For any linear program* (P_2)

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax = b \text{ and } x \geq 0, \end{aligned}$$

in standard form, where A *is an* $m \times n$ *matrix of rank* m *and* $b \geq 0$*, consider the linear program* $(\widehat{P})$ *in standard form*

$$\begin{aligned} &\text{maximize} \quad -(x_{n+1} + \cdots + x_{n+m}) \\ &\text{subject to} \quad \widehat{A}\,\widehat{x} = b \text{ and } \widehat{x} \geq 0. \end{aligned}$$

The simplex algorithm with a pivot rule that prevents cycling started on the basic feasible solution $\widehat{x} = (0_n, b)$ *of* $(\widehat{P})$ *terminates with an optimal solution* (u^*, w^*)*.*

(1) If $w^* \neq 0$*, then* $\mathcal{P}(A, B) = \emptyset$*, that is, the linear program* (P) *has no feasible solution.*

(2) If $w^* = 0$*, then* $\mathcal{P}(A, B) \neq \emptyset$*, and* u^* *is a basic feasible solution of* (P) *associated with some basis* K*.*

Proposition 26.1 shows that determining whether the polyhedron $\mathcal{P}(A, b)$ defined by a system of equations $Ax = b$ and inequalities $x \geq 0$ is nonempty is decidable. This decision procedure uses a fail-safe version of the simplex algorithm (that prevents cycling), and the proof that it always terminates and returns an answer is nontrivial.

26.3 How to Perform a Pivoting Step Efficiently

We now discuss briefly how to perform the computation of (u^+, K^+) from a basic feasible solution (u, K).

In order to avoid applying permutation matrices it is preferable to allow a basis K to be a sequence of indices, possibly out of order. Thus, for any $m \times n$ matrix A (with $m \leq n$) and any sequence $K = (k_1, k_2, \cdots, k_m)$ of m elements with $k_i \in \{1, \ldots, n\}$, the matrix A_K denotes the $m \times m$ matrix whose ith column is the k_ith column of A, and similarly for any vector $u \in \mathbb{R}^n$ (resp. any linear form $c \in (\mathbb{R}^n)^*$) the vector $u_K \in \mathbb{R}^m$ (the linear form $c_K \in (\mathbb{R}^m)^*$) is the vector whose ith entry is the k_ith entry in u (resp. the linear whose ith entry is the k_ith entry in c).

For each nonbasic $j \notin K$, we have

$$A^j = \gamma^j_{k_1} A^{k_1} + \cdots + \gamma^j_{k_m} A^{k_m} = A_K \gamma^j_K,$$

so the vector γ_K^j is given by $\gamma_K^j = A_K^{-1}A^j$, that is, by solving the system

$$A_K\gamma_K^j = A^j. \tag{$*_\gamma$}$$

To be very precise, since the vector γ_K^j depends on K its components should be denoted by $(\gamma_K^j)_{k_i}$, but as we said before, to simplify notation we write $\gamma_{k_i}^j$ instead of $(\gamma_K^j)_{k_i}$.

In order to decide which case applies ((A), (B1), (B2), (B3)), we need to compute the numbers $c_j - \sum_{k\in K}\gamma_k^j c_k$ for all $j \notin K$. For this, observe that

$$c_j - \sum_{k\in K}\gamma_k^j c_k = c_j - c_K\gamma_K^j = c_j - c_K A_K^{-1}A^j.$$

If we write $\beta_K = c_K A_K^{-1}$, then

$$c_j - \sum_{k\in K}\gamma_k^j c_k = c_j - \beta_K A^j.$$

and we see that $\beta_K^\top \in \mathbb{R}^m$ is the solution of the system $\beta_K^\top = (A_K^{-1})^\top c_k^\top$, which means that $\beta_K^\top$ is the solution of the system

$$A_K^\top\beta_K^\top = c_K^\top. \tag{$*_\beta$}$$

Remark: Observe that since u is a basis feasible solution of (P), we have $u_j = 0$ for all $j \notin K$, so u is the solution of the equation $A_K u_K = b$. As a consequence, the value of the objective function for u is $cu = c_K u_K = c_K A_K^{-1} b$. This fact will play a crucial role in Section 27.2 to show that when the simplex algorithm terminates with an optimal solution of the linear program (P), then it also produces an optimal solution of the dual linear program (D).

Assume that we have a basic feasible solution u, a basis K for u, and that we also have the matrix A_K as well its inverse A_K^{-1} (perhaps implicitly) and also the inverse $(A_K^\top)^{-1}$ of $A_K^\top$ (perhaps implicitly). Here is a description of an iteration step of the simplex algorithm, following almost exactly Chvatal (Chvatal [29], Chapter 7, Box 7.1).

An Iteration Step of the (Revised) Simplex Method

Step 1. Compute the numbers $c_j - \sum_{k\in K}\gamma_k^j c_k = c_j - \beta_K A^j$ for all $j \notin K$, and for this, compute $\beta_K^\top$ as the solution of the system

$$A_K^\top\beta_K^\top = c_K^\top.$$

If $c_j - \beta_K A^j \leq 0$ for all $j \notin K$, stop and return the optimal solution u (Case (A)).

Step 2. If Case (B) arises, use a pivot rule to determine which index $j^+ \notin K$ should enter the new basis K^+ (the condition $c_{j^+} - \beta_K A^{j^+} > 0$ should hold).

Step 3. Compute $\max_{k\in K}\gamma_k^{j^+}$. For this, solve the linear system

$$A_K\gamma_K^{j^+} = A^{j^+}.$$

Step 4. If $\max_{k\in K} \gamma_k^{j^+} \leq 0$, then stop and report that the linear program (P) is unbounded (Case (B1)).

Step 5. If $\max_{k\in K} \gamma_k^{j^+} > 0$, use the ratios $u_k/\gamma_k^{j^+}$ for all $k \in K$ such that $\gamma_k^{j^+} > 0$ to compute θ^{j^+}, and use a pivot rule to determine which index $k^- \in K$ such that $\theta^{j^+} = u_{k^-}/\gamma_{k^-}^{j+}$ should leave K (Case (B2)).

If $\max_{k\in K} \gamma_k^{j^+} = 0$, then use a pivot rule to determine which index k^- for which $\gamma_{k^-}^{j^+} > 0$ should leave the basis K (Case (B3)).

Step 6. Update u, K, and A_K, to u^+ and K^+, and A_{K^+}. During this step, given the basis K specified by the sequence $K = (k_1, \ldots, k_\ell, \ldots, k_m)$, with $k^- = k_\ell$, then K^+ is the sequence obtained by replacing k_ℓ by the incoming index j^+, so $K^+ = (k_1, \ldots, j^+, \ldots, k_m)$ with j^+ in the ℓth slot.

The vector u is easily updated. To compute A_{K^+} from A_K we take advantage that A_K and A_{K^+} only differ by a *single column*, namely the ℓth column A^{j^+}, which is given by the linear combination

$$A^{j^+} = A_K \gamma_K^{j^+}.$$

To simplify notation, denote $\gamma_K^{j^+}$ by γ, and recall that $k^- = k_\ell$. If $K = (k_1, \ldots, k_m)$, then $A_K = [A^{k_1} \cdots A^{k^-} \cdots A^{i_m}]$, and since A_{K^+} is the result of replacing the ℓth column A^{k^-} of A_K by the column A^{j^+}, we have

$$A_{K^+} = [A^{k_1} \cdots A^{j^+} \cdots A^{i_m}] = [A^{k_1} \cdots A_K\gamma \cdots A^{i_m}] = A_K E(\gamma),$$

where $E(\gamma)$ is the following invertible matrix obtained from the identity matrix I_m by replacing its ℓth column by γ:

$$E(\gamma) = \begin{pmatrix} 1 & & & \gamma_1 & & \\ & \ddots & & \vdots & & \\ & & 1 & \gamma_{\ell-1} & & \\ & & & \gamma_\ell & & \\ & & & \gamma_{\ell+1} & 1 & \\ & & & \vdots & & \ddots \\ & & & \gamma_m & & 1 \end{pmatrix}.$$

Since $\gamma_\ell = \gamma_{k^-}^{j^+} > 0$, the matrix $E(\gamma)$ is invertible, and it is easy to check that its inverse is given by

$$E(\gamma)^{-1} = \begin{pmatrix} 1 & & & -\gamma_\ell^{-1}\gamma_1 & & \\ & \ddots & & \vdots & & \\ & & 1 & -\gamma_\ell^{-1}\gamma_{\ell-1} & & \\ & & & \gamma_\ell^{-1} & & \\ & & & -\gamma_\ell^{-1}\gamma_{\ell+1} & 1 & \\ & & & \vdots & & \ddots \\ & & & -\gamma_\ell^{-1}\gamma_m & & 1 \end{pmatrix},$$

which is very cheap to compute. We also have

$$A_{K^+}^{-1} = E(\gamma)^{-1} A_K^{-1}.$$

Consequently, if A_K and A_K^{-1} are available, then A_{K^+} and $A_{K^+}^{-1}$ can be computed cheaply in terms of A_K and A_K^{-1} and matrices of the form $E(\gamma)$. Then the systems $(*_\gamma)$ to find the vectors γ_K^j can be solved cheaply.

Since

$$A_{K^+}^\top = E(\gamma)^\top A_K^\top$$

and

$$(A_{K^+}^\top)^{-1} = (A_K^\top)^{-1}(E(\gamma)^\top)^{-1},$$

the matrices $A_{K^+}^\top$ and $(A_{K^+}^\top)^{-1}$ can also be computed cheaply from $A_K^\top$, $(A_K^\top)^{-1}$, and matrices of the form $E(\gamma)^\top$. Thus the systems $(*_\beta)$ to find the linear forms β_K can also be solved cheaply.

A matrix of the form $E(\gamma)$ is called an *eta matrix*; see Chvatal [29] (Chapter 7). We showed that the matrix A_{K^s} obtained after s steps of the simplex algorithm can be written as

$$A_{K^s} = A_{K^{s-1}} E_s$$

for some eta matrix E_s, so A_{k^s} can be written as the product

$$A_{K^s} = E_1 E_2 \cdots E_s$$

of s beta matrices. Such a factorization is called an *eta factorization*. The eta factorization can be used to either invert A_{K^s} or to solve a system of the form $A_{K_s}\gamma = A^{j^+}$ iteratively. Which method is more efficient depends on the sparsity of the E_i.

In summary, there are cheap methods for finding the next basic feasible solution (u^+, K^+) from (u, K). We simply wanted to give the reader a flavor of these techniques. We refer the reader to texts on linear programming for detailed presentations of methods for implementing efficiently the simplex method. In particular, the *revised simplex method* is presented in Chvatal [29], Papadimitriou and Steiglitz [79], Bertsimas and Tsitsiklis [17], and Vanderbei [109].

26.4 The Simplex Algorithm Using Tableaux

We now describe a formalism for presenting the simplex algorithm, namely *(full) tableaux*. This is the traditional formalism used in all books, modulo minor variations. A particularly nice feature of the tableau formalism is that the update of a tableau can be performed using elementary row operations identical to the operations used during the reduction of a matrix to row reduced echelon form (rref). What differs is the criterion for the choice of the pivot.

Since the quantities $c_j - c_K\gamma_K^j$ play a crucial role in determining which column A^j should come into the basis, the notation $\overline{c}_j$ is used to denote $c_j - c_K\gamma_K^j$, which is called the *reduced cost* of the variable x_j. The reduced costs actually depend on K so to be very precise we should denote them by $(\overline{c}_K)_j$, but to simplify notation we write $\overline{c}_j$ instead of $(\overline{c}_K)_j$. We will see shortly how $(\overline{c}_{K^+})_i$ is computed in terms of in terms of $(\overline{c}_K)_i$.

Observe that the data needed to execute the next step of the simplex algorithm are

(1) The current basic solution u_K and its basis $K = (k_1, \ldots, k_m)$.

(2) The reduced costs $\overline{c}_j = c_j - c_K A_K^{-1} A^j = c_j - c_K\gamma_K^j$, for all $j \notin K$.

(3) The vectors $\gamma_K^j = (\gamma_{k_i}^j)_{i=1}^m$ for all $j \notin K$, that allow us to express each A^j as $A_K\gamma_K^j$.

All this information can be packed into a $(m+1) \times (n+1)$ matrix called a *(full) tableau* organized as follows:

$c_K u_K$	$\overline{c}_1$	$\cdots$	$\overline{c}_j$	$\cdots$	$\overline{c}_n$
u_{k_1}	γ_1^1	$\cdots$	γ_1^j	$\cdots$	γ_1^n
$\vdots$	$\vdots$		$\vdots$		$\vdots$
u_{k_m}	γ_m^1	$\cdots$	γ_m^j	$\cdots$	γ_m^n

It is convenient to think as the first row as row 0, and of the first column as column 0. Row 0 contains the current value of the objective function and the reduced costs, column 0 except for its top entry contains the components of the current basic solution u_K, and the remaining columns except for their top entry contain the vectors γ_K^j. Observe that the γ_K^j corresponding to indices j in K constitute a permutation of the identity matrix I_m. The entry $\gamma_{k^-}^{j^+}$ is called the *pivot* element. A tableau together with the new basis $K^+ = (K - \{k^-\}) \cup \{j^+\}$ contains all the data needed to compute the new u_{K^+}, the new $\gamma_{K^+}^j$, and the new reduced costs $(\overline{c}_{K^+})_j$.

If we define the $m \times n$ matrix Γ as the matrix $\Gamma = [\gamma_K^1 \ \cdots \ \gamma_K^n]$ whose jth column is γ_K^j, and $\overline{c}$ as the row vector $\overline{c} = (\overline{c}_1 \ \cdots \ \overline{c}_n)$, then the above tableau is denoted concisely by

$c_K u_K$	$\overline{c}$
u_K	Γ

We now show that the update of a tableau can be performed using elementary row operations identical to the operations used during the reduction of a matrix to row reduced echelon form (rref).

If $K = (k_1, \ldots, k_m)$, j^+ is the index of the incoming basis vector, $k^- = k_\ell$ is the index of the column leaving the basis, and if $K^+ = (k_1, \ldots, k_{\ell-1}, j^+, k_{\ell+1}, \ldots, k_m)$, since $A_{K^+} = A_K E(\gamma_K^{j^+})$, the new columns $\gamma_{K^+}^j$ are computed in terms of the old columns γ_K^j using the equations

$$\gamma_{K^+}^j = A_{K^+}^{-1} A^j = E(\gamma_K^{j^+})^{-1} A_K^{-1} A^j = E(\gamma_K^{j^+})^{-1}\gamma_K^j.$$

Consequently the matrix Γ^+ is given in terms of Γ by

$$\Gamma^+ = E(\gamma_K^{j^+})^{-1}\Gamma.$$

But the matrix $E(\gamma_K^{j^+})^{-1}$ is of the form

$$E(\gamma)^{-1} = \begin{pmatrix} 1 & & & -(\gamma_{k^-}^{j^+})^{-1}\gamma_{k_1}^{j^+} & & \\ & \ddots & & \vdots & & \\ & & 1 & -(\gamma_{k^-}^{j^+})^{-1}\gamma_{k_{\ell-1}}^{j^+} & & \\ & & & (\gamma_{k^-}^{j^+})^{-1} & & \\ & & & -(\gamma_{k^-}^{j^+})^{-1}\gamma_{k_{\ell+1}}^{j^+} & 1 & \\ & & & \vdots & & \ddots \\ & & & -(\gamma_{k^-}^{j^+})^{-1}\gamma_{k_m}^{j^+} & & 1 \end{pmatrix},$$

with the column involving the γs in the ℓth column, and this matrix is the product of the following elementary row operations:

1. Multiply row ℓ by $1/\gamma_{k^-}^{j^+}$ (the inverse of the pivot) to make the entry on row ℓ and column j^+ equal to 1.

2. subtract $\gamma_{k_i}^{j^+}$ $\times$ (the normalized) row ℓ from row i, for $i = 1, \ldots, \ell-1, \ell+1, \ldots, m$.

These are exactly the elementary row operations that reduce the ℓth column $\gamma_K^{j^+}$ of Γ to the ℓth column of the identity matrix I_m. Thus, this step is identical to the sequence of steps that the procedure to convert a matrix to row reduced echelon from executes on the ℓth column of the matrix. The only difference is the criterion for the choice of the pivot.

Since the new basic solution u_{K^+} is given by $u_{K^+} = A_{K^+}^{-1}b$, we have

$$u_{K^+} = E(\gamma_K^{j^+})^{-1}A_K^{-1}b = E(\gamma_K^{j^+})^{-1}u_K.$$

This means that u_+ is obtained from u_K by applying exactly the same elementary row operations that were applied to Γ. Consequently, just as in the procedure for reducing a matrix to rref, we can apply elementary row operations to the matrix $[u_k \ \Gamma]$, which consists of rows $1, \ldots, m$ of the tableau.

Once the new matrix Γ^+ is obtained, the new reduced costs are given by the following proposition.

Proposition 26.2. *Given any linear program $(P2)$ is standard form*

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax = b \text{ and } x \geq 0, \end{aligned}$$

where A is an $m \times n$ matrix of rank m, if (u, K) is a basic (not feasible) solution of $(P2)$ and if $K^+ = (K - \{k^-\}) \cup \{j^+\}$, with $K = (k_1, \ldots, k_m)$ and $k^- = k_\ell$, then for $i = 1, \ldots, n$ we have

$$c_i - c_{K^+}\gamma^i_{K^+} = c_i - c_K\gamma^i_K - \frac{\gamma^i_{k^-}}{\gamma^{j^+}_{k^-}}(c_{j^+} - c_K\gamma^{j^+}_K).$$

Using the reduced cost notation, the above equation is

$$(\overline{c}_{K^+})_i = (\overline{c}_K)_i - \frac{\gamma^i_{k^-}}{\gamma^{j^+}_{k^-}}(\overline{c}_K)_{j^+}.$$

Proof. Without any loss of generality and to simplify notation assume that $K = (1, \ldots, m)$ and write j for j^+ and ℓ for k_m. Since $\gamma^i_K = A_K^{-1}A^i$, $\gamma^i_{K^+} = A^{-1}_{K^+}A^i$, and $A_{K^+} = A_K E(\gamma^j_K)$, we have

$$c_i - c_{K^+}\gamma^i_{K^+} = c_i - c_{K^+}A^{-1}_{K^+}A^i = c_i - c_{K^+}E(\gamma^j_K)^{-1}A_K^{-1}A^i = c_i - c_{K^+}E(\gamma^j_K)^{-1}\gamma^i_K,$$

where

$$E(\gamma^j_K)^{-1} = \begin{pmatrix} 1 & & & -(\gamma^j_\ell)^{-1}\gamma^j_1 & & \\ & \ddots & & \vdots & & \\ & & 1 & -(\gamma^j_\ell)^{-1}\gamma^j_{\ell-1} & & \\ & & & (\gamma^j_\ell)^{-1} & & \\ & & & -(\gamma^j_\ell)^{-1}\gamma^j_{\ell+1} & 1 & \\ & & & \vdots & & \ddots \\ & & & -(\gamma^j_\ell)^{-1}\gamma^j_m & & 1 \end{pmatrix}$$

where the ℓth column contains the γs. Since $c_{K^+} = (c_1, \ldots, c_{\ell-1}, c_j, c_{\ell+1}, \ldots, c_m)$, we have

$$c_{K^+}E(\gamma^j_K)^{-1} = \left(c_1, \ldots, c_{\ell-1}, \frac{c_j}{\gamma^j_\ell} - \sum_{k=1, k\neq\ell}^{m} c_k\frac{\gamma^j_k}{\gamma^j_\ell}, c_{\ell+1}, \ldots, c_m\right),$$

and

$$c_{K^+}E(\gamma_K^j)^{-1}\gamma_K^i = \begin{pmatrix} c_1 & \dots & c_{\ell-1} & \dfrac{c_j}{\gamma_\ell^j} - \displaystyle\sum_{k=1,k\neq\ell}^{m} c_k\frac{\gamma_k^j}{\gamma_\ell^j} & c_{\ell+1} & \dots & c_m \end{pmatrix} \begin{pmatrix} \gamma_1^i \\ \vdots \\ \gamma_{\ell-1}^i \\ \gamma_\ell^i \\ \gamma_{\ell+1}^i \\ \vdots \\ \gamma_m^i \end{pmatrix}$$

$$\begin{aligned} &= \sum_{k=1,k\neq\ell}^{m} c_k\gamma_k^i + \frac{\gamma_\ell^i}{\gamma_\ell^j}\left(c_j - \sum_{k=1,k\neq\ell}^{m} c_k\gamma_k^j\right) \\ &= \sum_{k=1,k\neq\ell}^{m} c_k\gamma_k^i + \frac{\gamma_\ell^i}{\gamma_\ell^j}\left(c_j + c_\ell\gamma_\ell^j - \sum_{k=1}^{m} c_k\gamma_k^j\right) \\ &= \sum_{k=1}^{m} c_k\gamma_k^i + \frac{\gamma_\ell^i}{\gamma_\ell^j}\left(c_j - \sum_{k=1}^{m} c_k\gamma_k^j\right) \\ &= c_K\gamma_K^i + \frac{\gamma_\ell^i}{\gamma_\ell^j}(c_j - c_K\gamma_K^j), \end{aligned}$$

and thus

$$c_i - c_{K^+}\gamma_{K^+}^i = c_i - c_{K^+}E(\gamma_K^j)^{-1}\gamma_K^i = c_i - c_K\gamma_K^i - \frac{\gamma_\ell^i}{\gamma_\ell^j}(c_j - c_K\gamma_K^j),$$

as claimed. □

Since $(\gamma_{k^-}^1, \dots, \gamma_{k^-}^n)$ is the ℓth row of Γ, we see that Proposition 26.2 shows that

$$\overline{c}_{K^+} = \overline{c}_K - \frac{(\overline{c}_K)_{j^+}}{\gamma_{k^-}^{j^+}}\,\Gamma_\ell, \tag{$\dagger$}$$

where Γ_ℓ denotes the ℓ-th row of Γ and $\gamma_{k^-}^{j^+}$ is the pivot. This means that $\overline{c}_{K^+}$ is obtained by the elementary row operations which consist first normalizing the ℓth row by dividing it by the pivot $\gamma_{k^-}^{j^+}$, and then subtracting $(\overline{c}_K)_{j^+}\times$ the normalized row ℓ from $\overline{c}_K$. These are exactly the row operations that make the reduced cost $(\overline{c}_K)_{j^+}$ zero.

Remark: It easy easy to show that we also have

$$\overline{c}_{K^+} = c - c_{K^+}\Gamma^+.$$

We saw in section 26.2 that the change in the objective function after a pivoting step during which column j^+ comes in and column k^- leaves is given by

$$\theta^{j^+}\left(c_{j^+} - \sum_{k\in K} \gamma_k^{j^+} c_k\right) = \theta^{j^+}(\overline{c}_K)_{j^+},$$

where

$$\theta^{j^+} = \frac{u_{k^-}}{\gamma_{k^-}^{j^+}}.$$

If we denote the value of the objective function $c_K u_K$ by z_K, then we see that

$$z_{K^+} = z_K + \frac{(\overline{c}_K)_{j^+}}{\gamma_{k^-}^{j^+}}\, u_{k^-}.$$

This means that the new value z_{K^+} of the objective function is obtained by first normalizing the ℓth row by dividing it by the pivot $\gamma_{k^-}^{j^+}$, and then adding $(\overline{c}_K)_{j^+}\times$ the zeroth entry of the normalized ℓth line by $(\overline{c}_K)_{j^+}$ to the zeroth entry of line 0.

In updating the reduced costs, we subtract rather than add $(\overline{c}_K)_{j^+}\times$ the normalized row ℓ from row 0. This suggests storing $-z_K$ as the zeroth entry on line 0 rather than z_K, because then all the entries row 0 are updated by the same elementary row operations. Therefore, from now on, we use tableau of the form

$-c_K u_K$	$\overline{c}_1$	$\cdots$	$\overline{c}_j$	$\cdots$	$\overline{c}_n$
u_{k_1}	γ_1^1	$\cdots$	γ_1^j	$\cdots$	γ_1^n
$\vdots$	$\vdots$		$\vdots$		$\vdots$
u_{k_m}	γ_m^1	$\cdots$	γ_m^j	$\cdots$	γ_m^n

The simplex algorithm first chooses the incoming column j^+ by picking some column for which $\overline{c}_j > 0$, and then chooses the outgoing column k^- by considering the ratios $u_k/\gamma_k^{j^+}$ for which $\gamma_k^{j^+} > 0$ (along column j^+), and picking k^- to achieve the minimum of these ratios.

Here is an illustration of the simplex algorithm using elementary row operations on an example from Papadimitriou and Steiglitz [79] (Section 2.9).

Example 26.4. Consider the linear program

$$\begin{aligned}
&\text{maximize} && -2x_2 - x_4 - 5x_7\\
&\text{subject to}\\
&&x_1 + x_2 + x_3 + x_4 &= 4\\
&&x_1 + x_5 &= 2\\
&&x_3 + x_6 &= 3\\
&&3x_2 + x_3 + x_7 &= 6\\
&&x_1, x_2, x_3, x_4, x_5, x_6, x_7 &\geq 0.
\end{aligned}$$

We have the basic feasible solution $u = (0, 0, 0, 4, 2, 3, 6)$, with $K = (4, 5, 6, 7)$. Since $c_K = (-1, 0, 0, -5)$ and $c = (0, -2, 0, -1, 0, 0 - 5)$ the first tableau is

34	1	14	6	0	0	0	0
$u_4 = 4$	1	1	1	1	0	0	0
$u_5 = 2$	(1)	0	0	0	1	0	0
$u_6 = 3$	0	0	1	0	0	1	0
$u_7 = 6$	0	3	1	0	0	0	1

Row 0 is obtained by subtracting $-1\times$ (row 1) and $-5\times$ (row 4) from $c = (0, -2, 0, -1, 0, 0, -5)$. Let us pick column $j^+ = 1$ as the incoming column. We have the ratios (for positive entries on column 1)

$$4/1,\ 2/1,$$

and since the minimum is 2, we pick the outgoing column to be column $k^- = 5$. The pivot 1 is indicated in red. The new basis is $K = (4, 1, 6, 7)$. Next we apply row operations to reduce column 1 to the second vector of the identity matrix I_4. For this, we subtract row 2 from row 1. We get the tableau

34	1	14	6	0	0	0	0
$u_4 = 2$	0	1	1	1	-1	0	0
$u_1 = 2$	(1)	0	0	0	1	0	0
$u_6 = 3$	0	0	1	0	0	1	0
$u_7 = 6$	0	3	1	0	0	0	1

To compute the new reduced costs, we want to set $\bar{c}_1$ to 0 so we subtract row 2 from row 0, and we get the tableau

32	0	14	6	0	-1	0	0
$u_4 = 2$	0	1	(1)	1	-1	0	0
$u_1 = 2$	1	0	0	0	1	0	0
$u_6 = 3$	0	0	1	0	0	1	0
$u_7 = 6$	0	3	1	0	0	0	1

Next, pick column $j^+ = 3$ as the incoming column. We have the ratios (for positive entries on column 3)

$$2/1,\ 3/1,\ 6/1,$$

and since the minimum is 2, we pick the outgoing column to be column $k^- = 4$. The pivot 1 is indicated in red and the new basis is $K = (3, 1, 6, 7)$. Next we apply row operations to reduce column 3 to the first vector of the identity matrix I_4. For this, we subtract row 1 from row 3 and from row 4, to obtain the tableau:

32	0	14	6	0	−1	0	0
$u_3 = 2$	0	1	(1)	1	−1	0	0
$u_1 = 2$	1	0	0	0	1	0	0
$u_6 = 1$	0	−1	0	−1	1	1	0
$u_7 = 4$	0	2	0	−1	1	0	1

To compute the new reduced costs, we want to set $\bar{c}_3$ to 0 so we subtract 6× row 1 from row 0, and we get the tableau

20	0	8	0	−6	5	0	0
$u_3 = 2$	0	(1)	1	1	−1	0	0
$u_1 = 2$	1	0	0	0	1	0	0
$u_6 = 1$	0	−1	0	−1	1	1	0
$u_7 = 4$	0	2	0	−1	1	0	1

Next we pick $j^+ = 2$ as the incoming column. We have the ratios (for positive entries on column 2)

$$2/1,\ 4/2,$$

and since the minimum is 2, we pick the outgoing column to be column $k^- = 3$. The pivot 1 is indicated in red and the new basis is $K = (2, 1, 6, 7)$. Next we apply row operations to reduce column 2 to the first vector of the identity matrix I_4. For this, we add row 1 to row 3 and subtract 2× row 1 from row 4 to obtain the tableau:

20	0	8	0	−6	5	0	0
$u_2 = 2$	0	(1)	1	1	−1	0	0
$u_1 = 2$	1	0	0	0	1	0	0
$u_6 = 3$	0	0	1	0	0	1	0
$u_7 = 0$	0	0	−2	−3	3	0	1

To compute the new reduced costs, we want to set $\bar{c}_2$ to 0 so we subtract 8× row 1 from row 0 and we get the tableau

4	0	0	−8	−14	13	0	0
$u_2 = 2$	0	1	1	1	−1	0	0
$u_1 = 2$	1	0	0	0	1	0	0
$u_6 = 3$	0	0	1	0	0	1	0
$u_7 = 0$	0	0	−2	−3	(3)	0	1

The only possible incoming column corresponds to $j^+ = 5$. We have the ratios (for positive entries on column 5)

$$2/1,\ 0/3,$$

and since the minimum is 0, we pick the outgoing column to be column $k^- = 7$. The pivot 3 is indicated in red and the new basis is $K = (2, 1, 6, 5)$. Since the minimum is 0, the basis $K = (2, 1, 6, 5)$ is degenerate (indeed, the component corresponding to the index 5 is 0). Next we apply row operations to reduce column 5 to the fourth vector of the identity matrix I_4. For this, we multiply row 4 by 1/3, and then add the normalized row 4 to row 1 and subtract the normalized row 4 from row 2, and to obtain the tableau:

4	0	0	-8	-14	13	0	0
$u_2 = 2$	0	1	1/3	0	0	0	1/3
$u_1 = 2$	1	0	2/3	1	0	0	$-1/3$
$u_6 = 3$	0	0	1	0	0	1	0
$u_5 = 0$	0	0	$-2/3$	-1	(1)	0	1/3

To compute the new reduced costs, we want to set $\overline{c}_5$ to 0 so we subtract 13× row 4 from row 0 and we get the tableau

4	0	0	2/3	-1	0	0	$-13/3$
$u_2 = 2$	0	1	1/3	0	0	0	1/3
$u_1 = 2$	1	0	(2/3)	1	0	0	$-1/3$
$u_6 = 3$	0	0	1	0	0	1	0
$u_5 = 0$	0	0	$-2/3$	-1	1	0	1/3

The only possible incoming column corresponds to $j^+ = 3$. We have the ratios (for positive entries on column 3)

$$2/(1/3) = 6,\ 2/(2/3) = 3,\ 3/1 = 3,$$

and since the minimum is 3, we pick the outgoing column to be column $k^- = 1$. The pivot 2/3 is indicated in red and the new basis is $K = (2, 3, 6, 5)$. Next we apply row operations to reduce column 3 to the second vector of the identity matrix I_4. For this, we multiply row 2 by 2/3, subtract (1/3)× (normalized row 2) from row 1, and subtract normalized row 2 from row 3, add add row (2/3)× (normalized row 2) to row 4, to obtain the tableau:

4	0	0	2/3	-1	0	0	$-13/3$
$u_2 = 1$	$-1/2$	1	0	$-1/2$	0	0	1/2
$u_3 = 3$	3/2	0	(1)	3/2	0	0	$-1/2$
$u_6 = 0$	$-3/2$	0	0	$-3/2$	0	1	1/2
$u_5 = 2$	1	0	0	0	1	0	0

To compute the new reduced costs, we want to set $\overline{c}_3$ to 0 so we subtract (2/3)× row 2 from row 0 and we get the tableau

2	-1	0	0	-2	0	0	-4
$u_2 = 1$	$-1/2$	1	0	$-1/2$	0	0	$1/2$
$u_3 = 3$	$3/2$	0	1	$3/2$	0	0	$-1/2$
$u_6 = 0$	$-3/2$	0	0	$-3/2$	0	1	$1/2$
$u_5 = 2$	1	0	0	0	1	0	0

Since all the reduced cost are ≤ 0, we have reached an optimal solution, namely $(0, 1, 3, 0, 2, 0, 0, 0)$, with optimal value -2.

The progression of the simplex algorithm from one basic feasible solution to another corresponds to the visit of vertices of the polyhedron $\mathcal{P}$ associated with the constraints of the linear program illustrated in Figure 26.4.

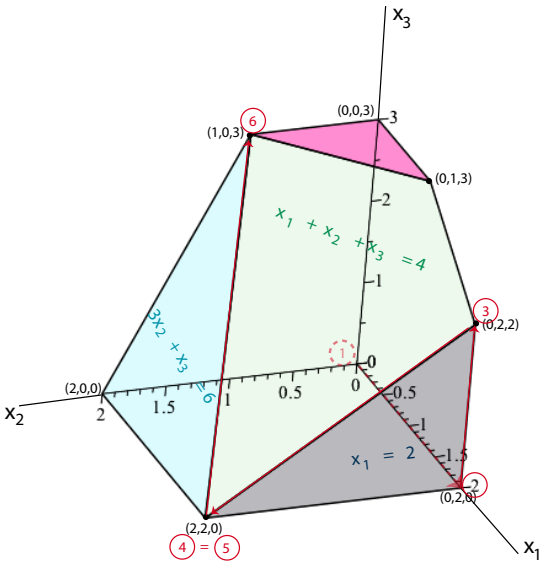


Figure 26.4: The polytope $\mathcal{P}$ associated with the linear program optimized by the tableau method. The red arrowed path traces the progression of the simplex method from the origin to the vertex $(0, 1, 3)$.

As a final comment, if it is necessary to run Phase I of the simplex algorithm, in the event that the simplex algorithm terminates with an optimal solution $(u^*, 0_m)$ and a basis K^* such that some $u_i = 0$, then the basis K^* contains indices of basic columns A^j corresponding to slack variables that need to be *driven out* of the basis. This is easy to achieve by performing a pivoting step involving some other column j^+ corresponding to one of the original variables (not a slack variable) for which $(\gamma_{K^*})_i^{j^+} \neq 0$. In such a step, it doesn't matter whether $(\gamma_{K^*})_i^{j^+} < 0$ or $(\overline{c}_{K^*})_{j^+} \leq 0$. If the original matrix A has no redundant equations, such a step

is always possible. Otherwise, $(\gamma_{K^*})_i^j = 0$ for all non-slack variables, so we detected that the ith equation is redundant and we can delete it.

Other presentations of the tableau method can be found in Bertsimas and Tsitsiklis [17] and Papadimitriou and Steiglitz [79].

26.5 Computational Efficiency of the Simplex Method

Let us conclude with a few comments about the efficiency of the simplex algorithm. In *practice*, it was observed by Dantzig that for linear programs with $m < 50$ and $m+n < 200$, the simplex algorithms typically requires less than $3m/2$ iterations, but at most $3m$ iterations. This fact agrees with more recent empirical experiments with much larger programs that show that the number iterations is bounded by $3m$. Thus, it was somewhat of a shock in 1972 when Klee and Minty found a linear program with n variables and n equations for which the simplex algorithm with Dantzig's pivot rule requires requires $2^n - 1$ iterations. This program (taken from Chvatal [29], page 47) is reproduced below:

$$\begin{aligned}
&\text{maximize} \quad \sum_{j=1}^{n} 10^{n-j} x_j \\
&\text{subject to} \\
&\left(2 \sum_{j=1}^{i-1} 10^{i-j} x_j\right) + x_i \leq 100^{i-1} \\
&x_j \geq 0,
\end{aligned}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, n$.

If $p = \max(m, n)$, then , in terms of worse case behavior, for all currently known pivot rules, the simplex algorithm has exponential complexity in p. However, as we said earlier, in practice, nasty examples such as the Klee–Minty example seem to be rare, and the number of iterations appears to be linear in m.

Whether or not a pivot rule (a clairvoyant rule) for which the simplex algorithms runs in polynomial time in terms of m is still an *open problem*.

The *Hirsch conjecture* claims that there is some pivot rule such that the simplex algorithm finds an optimal solution in $O(p)$ steps. The best bound known so far due to Kalai and Kleitman is $m^{1+\ln n} = (2n)^{\ln m}$. For more on this topic, see Matousek and Gardner [72] (Section 5.9) and Bertsimas and Tsitsiklis [17] (Section 3.7).

Researchers have investigated the problem of finding upper bounds on the expected number of pivoting steps if a randomized pivot rule is used. Bounds better than 2^m (but of course, not polynomial) have been found.

Understanding the complexity of linear programing, in particular of the simplex algorithm, is still ongoing. The interested reader is referred to Matousek and Gardner [72] (Chapter 5, Section 5.9) for some pointers.

In the next section we consider important theoretical criteria for determining whether a set of constraints $Ax \leq b$ and $x \geq 0$ has a solution or not.

Chapter 27

Linear Programming and Duality

27.1 Variants of the Farkas Lemma

If A is an $m \times n$ matrix and if $b \in \mathbb{R}^m$ is a vector, it is known from linear algebra that the linear system $Ax = b$ has no solution iff there is some linear form $y \in (\mathbb{R}^m)^*$ such that $yA = 0$ and $yb \neq 0$. This means that the linear from y vanishes on the columns $A^1, \ldots, A^n$ of A but does not vanish on b. Since the linear form y defines the linear hyperplane H of equation $yz = 0$ (with $z \in \mathbb{R}^m$), geometrically the equation $Ax = b$ has no solution iff there is a linear hyperplane H containing $A^1, \ldots, A^n$ and not containing b. This is a kind of separation theorem that says that the vectors $A^1, \ldots, A^n$ and b can be separated by some linear hyperplane H.

What we would like to do is to generalize this kind of criterion, first to a system $Ax = b$ subject to the constraints $x \geq 0$, and next to sets of inequality constraints $Ax \leq b$ and $x \geq 0$. There are indeed such criteria going under the name of *Farkas lemma*.

The key is a separation result involving polyhedral cones known as the Farkas–Minkowski proposition. We have the following fundamental separation lemma.

Proposition 27.1. *Let $C \subseteq \mathbb{R}^n$ be a closed nonempty cone. For any point $a \in \mathbb{R}^n$, if $a \notin C$, then there is a linear hyperplane H (through 0) such that*

1. *C lies in one of the two half-spaces determined by H.*
2. *$a \notin H$*
3. *a lies in the other half-space determined by H.*

We say that H strictly separates C and a.

Proposition 27.1 is an easy consequence of another separation theorem that asserts that given any two nonempty closed convex sets A and B with A compact, there is a hyperplane H strictly separating A and B (which means that $A \cap H = \emptyset$, $B \cap H = \emptyset$, that A lies in one

of the two half-spaces determined by H, and B lies in the other half-space determined by H); see Gallier [44] (Chapter 7, Corollary 7.4 and Proposition 7.3). This proof is nontrivial and involves a geometric version of the Hahn–Banach theorem.

The Farkas–Minkowski proposition is Proposition 27.1 applied to a polyhedral cone

$$C = \{\lambda_1 a_1 + \cdots + \lambda_n a_n \mid \lambda_i \geq 0,\ i = 1, \ldots, n\}$$

where $\{a_1, \ldots, a_n\}$ is a *finite* number of vectors $a_i \in \mathbb{R}^n$. By Proposition 24.2, any polyhedral cone is closed, so Proposition 27.1 applies and we obtain the following separation lemma.

Proposition 27.2. *(Farkas–Minkowski) Let $C \subseteq \mathbb{R}^n$ be a nonempty polyhedral cone $C = \text{cone}(\{a_1, \ldots, a_n\})$. For any point $b \in \mathbb{R}^n$, if $b \notin C$, then there is a linear hyperplane H (through 0) such that*

1. *C lies in one of the two half-spaces determined by H.*
2. *$a \notin H$*
3. *a lies in the other half-space determined by H.*

Equivalently, there is a nonzero linear form $y \in (\mathbb{R}^n)^*$ such that

1. *$ya_i \geq 0$ for $i = 1, \ldots, n$.*
2. *$yb < 0$.*

A direct proof of the Farkas–Minkowski proposition not involving Proposition 27.1 is given at the end of this section.

Remark: There is a generalization of the Farkas–Minkowski proposition applying to infinite dimensional real Hilbert spaces; see Theorem 28.11 (or Ciarlet [30], Chapter 9).

Proposition 27.2 implies our first version of Farkas' lemma.

Proposition 27.3. *(Farkas Lemma, Version I) Let A be an $m \times n$ matrix and let $b \in \mathbb{R}^m$ be any vector. The linear system $Ax = b$ has no solution $x \geq 0$ iff there is some nonzero linear form $y \in (\mathbb{R}^m)^*$ such that $yA \geq 0_n^\top$ and $yb < 0$.*

Proof. First, assume that there is some nonzero linear form $y \in (\mathbb{R}^m)^*$ such that $yA \geq 0$ and $yb < 0$. If $x \geq 0$ is a solution of $Ax = b$, then we get

$$yAx = yb,$$

but if $yA \geq 0$ and $x \geq 0$, then $yAx \geq 0$, and yet by hypothesis $yb < 0$, a contradiction.

Next assume that $Ax = b$ has no solution $x \geq 0$. This means that b does not belong to the polyhedral cone $C = \text{cone}(\{A^1, \ldots, A^n\})$ spanned by the columns of A. By Proposition 27.2, there is a nonzero linear form $y \in (\mathbb{R}^m)^*$ such that

1. $yA^j \geq 0$ for $j = 1, \ldots, n$.

2. $yb < 0$,

which says that $yA \geq 0_n^\top$ and $yb < 0$. □

Next consider the solvability of a system of inequalities of the form $Ax \leq b$ and $x \geq 0$.

Proposition 27.4. *(Farkas Lemma, Version II) Let A be an $m \times n$ matrix and let $b \in \mathbb{R}^m$ be any vector. The system of inequalities $Ax \leq b$ has no solution $x \geq 0$ iff there is some nonzero linear form $y \in (\mathbb{R}^m)^*$ such that $y \geq 0_m^\top$, $yA \geq 0_n^\top$, and $yb < 0$.*

Proof. We use the trick of linear programming which consists of adding "slack variables" z_i to convert inequalities $a_i x \leq b_i$ into equations $a_i x + z_i = b_i$ with $z_i \geq 0$ already discussed just before Definition 24.5. If we let $z = (z_1, \ldots, z_m)$, it is obvious that the system $Ax \leq b$ has a solution $x \geq 0$ iff the equation

$$\begin{pmatrix} A & I_m \end{pmatrix} \begin{pmatrix} x \\ z \end{pmatrix} = b$$

has a solution $\begin{pmatrix} x \\ z \end{pmatrix}$ with $x \geq 0$ and $z \geq 0$. Now by Farkas I, the above system has no solution with with $x \geq 0$ and $z \geq 0$ iff there is some nonzero linear form $y \in (\mathbb{R}^m)^*$ such that

$$y \begin{pmatrix} A & I_m \end{pmatrix} \geq 0_{n+m}^\top$$

and $yb < 0$, that is, $yA \geq 0_n^\top$, $y \geq 0_m^\top$, and $yb < 0$. □

In the next section we use Farkas II to prove the duality theorem in linear programming. Observe that by taking the negation of the equivalence in Farkas II we obtain a criterion of solvability, namely:

The system of inequalities $Ax \leq b$ has a solution $x \geq 0$ iff for every nonzero linear form $y \in (\mathbb{R}^m)^*$ such that $y \geq 0_m^\top$, if $yA \geq 0_n^\top$, then $yb \geq 0$.

We now prove the Farkas–Minkowski proposition without using Proposition 27.1. This approach uses a basic property of the distance function from a point to a closed set.

Let $X \subseteq \mathbb{R}^n$ be any nonempty set and let $a \in \mathbb{R}^n$ be any point. The *distance* $d(a, X)$ from a to X is defined as

$$d(a, X) = \inf_{x \in X} \|a - x\|.$$

Here, $\| \ \|$ denotes the Euclidean norm.

Proposition 27.5. *Let $X \subseteq \mathbb{R}^n$ be any nonempty set and let $a \in \mathbb{R}^n$ be any point. If X is closed, then there is some $z \in X$ such that $\|a - z\| = d(a, X)$.*

Proof. Since X is nonempty, pick any $x_0 \in X$, and let $r = \|a - x_0\|$. If $B_r(a)$ is the closed ball $B_r(a) = \{x \in \mathbb{R}^n \mid \|x - a\| \leq r\}$, then clearly

$$d(a, X) = \inf_{x \in X} \|a - x\| = \inf_{x \in X \cap B_r(a)} \|a - x\| .$$

Since $B_r(a)$ is compact and X is closed, $K = X \cap B_r(a)$ is also compact. But the function $x \mapsto \|a - x\|$ defined on the compact set K is continuous, and the image of a compact set by a continuous function is compact, so by Heine–Borel it has a minimum that is achieved by some $z \in K \subseteq X$. □

Remark: If U is a nonempty, closed and convex subset of a Hilbert space V, a standard result of Hilbert space theory (the projection theorem) asserts that for any $v \in V$ there is a *unique* $p \in U$ such that

$$\|v - p\| = \inf_{u \in U} \|v - u\| = d(v, U),$$

and

$$\langle p - v, u - p \rangle \geq 0 \quad \text{for all } u \in U.$$

Here $\|w\| = \sqrt{\langle w, w \rangle}$, where $\langle -, - \rangle$ is the inner product of the Hilbert space V.

We can now give a proof of the Farkas–Minkowski proposition (Proposition 27.2).

Proof of the Farkas–Minkowski proposition. Let $C = \text{cone}(\{a_1, \ldots, a_m\})$ be a polyhedral cone (nonempty) and assume that $b \notin C$. By Proposition 24.2, the polyhedral cone is closed, and by Proposition 27.5 there is some $z \in C$ such that $d(b, C) = \|b - z\|$; that is, z is a point of C closest to b. Since $b \notin C$ and $z \in C$ we have $u = z - b \neq 0$, and we claim that the linear hyperplane H orthogonal to u does the job, as illustrated in Figure 27.1.

First let us show that

$$\langle u, z \rangle = \langle z - b, z \rangle = 0. \tag{$*_1$}$$

This is trivial if $z = 0$, so assume $z \neq 0$. If $\langle u, z \rangle \neq 0$, then either $\langle u, z \rangle > 0$ or $\langle u, z \rangle < 0$. In either case we show that we can find some point $z' \in C$ closer to b than z is, a contradiction.

Case 1: $\langle u, z \rangle > 0$.

Let $z' = (1 - \alpha)z$ for any α such that $0 < \alpha < 1$. Then $z' \in C$ and since $u = z - b$

$$z' - b = (1 - \alpha)z - (z - u) = u - \alpha z,$$

so

$$\|z' - b\|^2 = \|u - \alpha z\|^2 = \|u\|^2 - 2\alpha\langle u, z \rangle + \alpha^2 \|z\|^2 .$$

If we pick $\alpha > 0$ such that $\alpha < 2\langle u, z \rangle / \|z\|^2$, then $-2\alpha\langle u, z \rangle + \alpha^2 \|z\|^2 < 0$, so $\|z' - b\|^2 < \|u\|^2 = \|z - b\|^2$, contradicting the fact that z is a point of C closest to b.

Case 2: $\langle u, z \rangle < 0$.

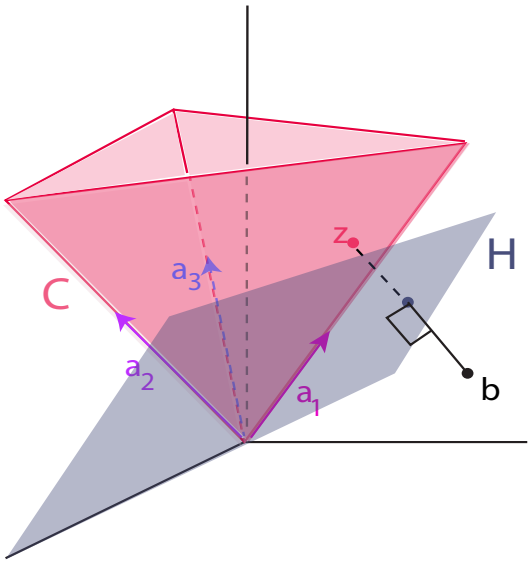


Figure 27.1: The hyperplane H, perpendicular to $z - b$, separates the point b from $C = \text{cone}(\{a_1, a_2, a_3\})$.

Let $z' = (1 + \alpha)z$ for any α such that $\alpha \geq -1$. Then $z' \in C$ and since $u = z - b$ we have $z' - b = (1 + \alpha)z - (z - u) = u + \alpha z$ so

$$\|z' - b\|^2 = \|u + \alpha z\|^2 = \|u\|^2 + 2\alpha\langle u, z\rangle + \alpha^2 \|z\|^2,$$

and if

$$0 < \alpha < -2\langle u, z\rangle / \|z\|^2,$$

then $2\alpha\langle u, z\rangle + \alpha^2 \|z\|^2 < 0$, so $\|z' - b\|^2 < \|u\|^2 = \|z - b\|^2$, a contradiction as above.

Therefore $\langle u, z\rangle = 0$. We have

$$\langle u, u\rangle = \langle u, z - b\rangle = \langle u, z\rangle - \langle u, b\rangle = -\langle u, b\rangle,$$

and since $u \neq 0$, we have $\langle u, u\rangle > 0$, so $\langle u, u\rangle = -\langle u, b\rangle$ implies that

$$\langle u, b\rangle < 0. \tag{$*_2$}$$

It remains to prove that $\langle u, a_i\rangle \geq 0$ for $i = 1, \ldots, m$. Pick any $x \in C$ such that $x \neq z$. We claim that

$$\langle b - z, x - z\rangle \leq 0. \tag{$*_3$}$$

Otherwise $\langle b - z, x - z\rangle > 0$, that is, $\langle z - b, x - z\rangle < 0$, and we show that we can find some point $z' \in C$ on the line segment $[z, x]$ closer to b than z is.

For any α such that $0 \leq \alpha \leq 1$, we have $z' = (1-\alpha)z + \alpha x = z + \alpha(x - z) \in C$, and since $z' - b = z - b + \alpha(x - z)$ we have

$$\|z' - b\|^2 = \|z - b + \alpha(x - z)\|^2 = \|z - b\|^2 + 2\alpha\langle z - b, x - z\rangle + \alpha^2 \|x - z\|^2,$$

so for any $\alpha > 0$ such that

$$\alpha < -2\langle z - b, x - z\rangle / \|x - z\|^2,$$

we have $2\alpha\langle z - b, x - z\rangle + \alpha^2 \|x - z\|^2 < 0$, which implies that $\|z' - b\|^2 < \|z - b\|^2$, contradicting that z is a point of C closest to b.

Since $\langle b - z, x - z\rangle \leq 0$, $u = z - b$, and by $(*_1)$ $\langle u, z\rangle = 0$, we have

$$0 \geq \langle b - z, x - z\rangle = \langle -u, x - z\rangle = -\langle u, x\rangle + \langle u, z\rangle = -\langle u, x\rangle,$$

which means that

$$\langle u, x\rangle \geq 0 \quad \text{for all } x \in C, \tag{$*_3$}$$

as claimed. In particular,

$$\langle u, a_i\rangle \geq 0 \quad \text{for } i = 1, \ldots, m. \tag{$*_4$}$$

Then, by $(*_2)$ and $(*_4)$, the linear form defined by $y = u^\top$ satisfies the properties $yb < 0$ and $ya_i \geq 0$ for $i = 1, \ldots, m$, which proves the Farkas–Minkowski proposition. □

There are other ways of proving the Farkas–Minkowski proposition, for instance using minimally infeasible systems or Fourier–Motzkin elimination; see Matousek and Gardner [72] (Chapter 6, Sections 6.6 and 6.7).

27.2 The Duality Theorem in Linear Programming

Let (P) be the linear program

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax \leq b \text{ and } x \geq 0, \end{aligned}$$

with A a $m \times n$ matrix, and assume that (P) has a feasible solution and is bounded above. Since by hypothesis the objective function $x \mapsto cx$ is bounded on $\mathcal{P}(A, b)$, it might be useful to deduce an *upper bound* for cx from the inequalities $Ax \leq b$, for any $x \in \mathcal{P}(A, b)$. We can do this as follows: for every inequality

$$a_i x \leq b_i \quad 1 \leq i \leq m,$$

pick a nonnegative scalar y_i, multiply both sides of the above inequality by y_i obtaining

$$y_i a_i x \leq y_i b_i \quad 1 \leq i \leq m,$$

(the direction of the inequality is preserved since $y_i \geq 0$), and then add up these m equations, which yields

$$(y_1 a_1 + \cdots + y_m a_m)x \leq y_1 b_1 + \cdots + y_m b_m.$$

If we can pick the $y_i \geq 0$ such that

$$c \leq y_1 a_1 + \cdots + y_m a_m,$$

then since $x_j \geq 0$ we have

$$cx \leq (y_1 a_1 + \cdots + y_m a_m)x \leq y_1 b_1 + \cdots + y_m b_m,$$

namely we found an upper bound of the value cx of the objective function of (P) for any feasible solution $x \in \mathcal{P}(A, b)$. If we let y be the linear form $y = (y_1, \ldots, y_m)$, then since

$$A = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix}$$

$y_1 a_1 + \cdots + y_m a_m = yA$, and $y_1 b_1 + \cdots + y_m b_m = yb$, what we did was to look for some $y \in (\mathbb{R}^m)^*$ such that

$$c \leq yA, \quad y \geq 0,$$

so that we have

$$cx \leq yb \quad \text{for all } x \in \mathcal{P}(A, b). \tag{$*$}$$

Then it is natural to look for a "best" value of yb, namely a minimum value, which leads to the definition of the *dual* of the linear program (P), a notion due to John von Neumann.

Definition 27.1. Given any linear program (P)

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax \leq b \text{ and } x \geq 0, \end{aligned}$$

with A a $m \times n$ matrix, the *dual* (D) of (P) is the following optimization problem:

$$\begin{aligned} &\text{minimize} \quad yb \\ &\text{subject to} \quad yA \geq c \text{ and } y \geq 0, \end{aligned}$$

where $y \in (\mathbb{R}^m)^*$. The original linear program (P) is called the *primal* linear program.

Here is an explicit example of a linear program and its dual.

Example 27.1. Consider the linear program illustrated by Figure 27.3

$$\begin{aligned} &\text{maximize} && 2x_1 + 3x_2 \\ &\text{subject to} && \\ & && 4x_1 + 8x_2 \leq 12 \\ & && 2x_1 + x_2 \leq 3 \\ & && 3x_1 + 2x_2 \leq 4 \\ & && x_1 \geq 0,\ x_2 \geq 0. \end{aligned}$$

Its dual linear program is illustrated in Figure27.2

$$\begin{aligned} &\text{minimize} && 12y_1 + 3y_2 + 4y_3 \\ &\text{subject to} && \\ & && 4y_1 + 2y_2 + 3y_3 \geq 2 \\ & && 8y_1 + y_2 + 2y_3 \geq 3 \\ & && y_1 \geq 0,\ y_2 \geq 0,\ y_3 \geq 0. \end{aligned}$$

It can be checked that $(x_1, x_2) = (1/2, 5/4)$ is an optimal solution of the primal linear program, with the maximum value of the objective function $2x_1 + 3x_2$ equal to $19/4$, and that $(y_1, y_2, y_3) = (5/16, 0, 1/4)$ is an optimal solution of the dual linear program, with the minimum value of the objective function $12y_1 + 3y_2 + 4y_3$ also equal to $19/4$.

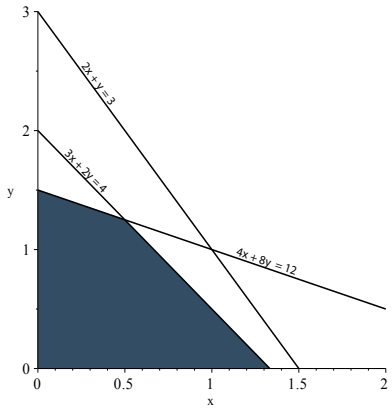


Figure 27.2: The $\mathcal{H}$-polytope for the linear program of Example 27.1. Note $x_1 \to x$ and $x_2 \to y$.

Observe that in the primal linear program (P), we are looking for a *vector* $x \in \mathbb{R}^n$ maximizing the form cx, and that the constraints are determined by the action of the *rows* of the matrix A on x. On the other hand, in the dual linear program (D), we are looking

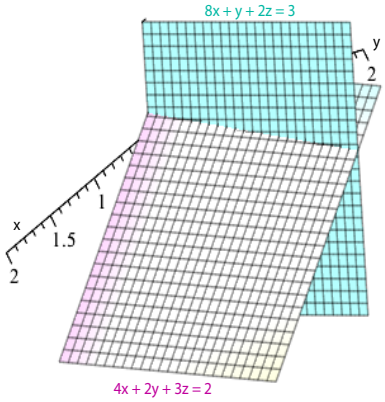


Figure 27.3: The $\mathcal{H}$-polyhedron for the dual linear program of Example 27.1 is the spacial region "above" the pink plane and in "front" of the blue plane. Note $y_1 \to x$, $y_2 \to y$, and $y_3 \to z$.

for a *linear form* $y \in (\mathbb{R}^*)^m$ minimizing the form yb, and the constraints are determined by the action of y on the *columns* of A. This is the sense in which (D) is the *dual* (P). In most presentations, the fact that (P) and (D) perform a search for a solution in spaces that are dual to each other is obscured by excessive use of transposition.

To convert the dual program (D) to a standard maximization problem we change the objective function yb to $-b^\top y^\top$ and the inequality $yA \geq c$ to $-A^\top y^\top \leq -c^\top$. The dual linear program (D) is now stated as (D')

$$\begin{aligned} &\text{maximize} && -b^\top y^\top \\ &\text{subject to} && -A^\top y^\top \leq -c^\top \text{ and } y^\top \geq 0, \end{aligned}$$

where $y \in (\mathbb{R}^m)^*$. Observe that the dual in maximization form (D'') of the dual program (D') gives back the primal program (P).

The above discussion established the following inequality known as *weak duality*.

Proposition 27.6. *(Weak Duality) Given any linear program* (P)

$$\begin{aligned} &\text{maximize} && cx \\ &\text{subject to} && Ax \leq b \text{ and } x \geq 0, \end{aligned}$$

with A *a* $m \times n$ *matrix, for any feasible solution* $x \in \mathbb{R}^n$ *of the primal problem* (P) *and every feasible solution* $y \in (\mathbb{R}^m)^*$ *of the dual problem* (D)*, we have*

$$cx \leq yb.$$

We say that the dual linear program (D) is *bounded below* if $\{yb \mid y^\top \in \mathcal{P}(-A^\top, -c^\top)\}$ is bounded below.

What happens if x^* is an optimal solution of (P) and if y^* is an optimal solution of (D)? We have $cx^* \leq y^*b$, but is there a "duality gap," that is, is it possible that $cx^* < y^*b$?

The answer is **no**, this is the *strong duality theorem*. Actually, the strong duality theorem asserts more than this.

Theorem 27.7. *(Strong Duality for Linear Programming) Let (P) be any linear program*

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax \leq b \text{ and } x \geq 0, \end{aligned}$$

with A a $m \times n$ matrix. The primal problem (P) has a feasible solution and is bounded above iff the dual problem (D) has a feasible solution and is bounded below. Furthermore, if (P) has a feasible solution and is bounded above, then for every optimal solution x^* of (P) and every optimal solution y^* of (D), we have

$$cx^* = y^*b.$$

Proof. If (P) has a feasible solution and is bounded above then we know from Proposition 25.1 that (P) has some optimal solution. Let x^* be any optimal solution of (P). First we will show that (D) has a feasible solution v.

Let $\mu = cx^*$ be the maximum of the objective function $x \mapsto cx$. Then for any $\epsilon > 0$, the system of inequalties

$$Ax \leq b, \quad x \geq 0, \quad cx \geq \mu + \epsilon$$

has no solution, since otherwise μ would not be the maximum value of the objective function cx. We would like to apply Farkas II, so first we transform the above system of inequalities into the system

$$\begin{pmatrix} A \\ -c \end{pmatrix} x \leq \begin{pmatrix} b \\ -(\mu + \epsilon) \end{pmatrix}.$$

By Proposition 27.3 (Farkas II), there is some linear form $(\lambda, z) \in (\mathbb{R}^{m+1})^*$ such that $\lambda \geq 0$, $z \geq 0$,

$$\begin{pmatrix} \lambda & z \end{pmatrix} \begin{pmatrix} A \\ -c \end{pmatrix} \geq 0_m^\top,$$

and

$$\begin{pmatrix} \lambda & z \end{pmatrix} \begin{pmatrix} b \\ -(\mu + \epsilon) \end{pmatrix} < 0,$$

which means that

$$\lambda A - zc \geq 0_m^\top, \quad \lambda b - z(\mu + \epsilon) < 0,$$

that is,

$$\begin{aligned} \lambda A &\geq zc \\ \lambda b &< z(\mu + \epsilon) \\ \lambda \geq 0,&\ \ z \geq 0. \end{aligned}$$

On the other hand, since $x^* \geq 0$ is an optimal solution of the system $Ax \leq b$, by Farkas II again (by taking the negation of the equivalence), since $\lambda A \geq 0$ (for the same λ as before), we must have

$$\lambda b \geq 0. \tag{$*_1$}$$

We claim that $z > 0$. Otherwize, since $z \geq 0$, we must have $z = 0$, but then

$$\lambda b < z(\mu + \epsilon)$$

implies

$$\lambda b < 0, \tag{$*_2$}$$

and since $\lambda b \geq 0$ by $(*_1)$, we have a contradiction. Consequently, we can divide by $z > 0$ without changing the direction of inequalities, and we obtain

$$\begin{aligned} \frac{\lambda}{z} A &\geq c \\ \frac{\lambda}{z} b &< \mu + \epsilon \\ \frac{\lambda}{z} &\geq 0, \end{aligned}$$

which shows that $v = \lambda/z$ is a feasible solution of the dual problem (D). However, weak duality (Proposition 27.6) implies that $cx^* = \mu \leq yb$ for any feasible solution $y \geq 0$ of the dual program (D), so (D) is bounded below and by Proposition 25.1 applied to the version of (D) written as a maximization problem, we conclude that (D) has some optimal solution. For any optimal solution y^* of (D), since v is a feasible solution of (D) such that $vb < \mu + \epsilon$, we must have

$$\mu \leq y^* b < \mu + \epsilon,$$

and since our reasoning is valid for *any* $\epsilon > 0$, we conclude that $cx^* = \mu = y^* b$.

If we assume that the dual program (D) has a feasible solution and is bounded below, since the dual of (D) is (P), we conclude that (P) is also feasible and bounded above. □

The strong duality theorem can also be proved by the simplex method, because when it terminates with an optimal solution of (P), the final tableau also produces an optimal solution y of (D) that can be read off the reduced costs of columns $n + 1, \ldots, n + m$ by flipping their signs. We follow the proof in Ciarlet [30] (Chapter 10).

Theorem 27.8. *Consider the linear program (P),*

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax \leq b \text{ and } x \geq 0, \end{aligned}$$

its equivalent version (P_2) *in standard form,*

$$\begin{aligned} &\text{maximize} \quad \widehat{c}\,\widehat{x} \\ &\text{subject to} \quad \widehat{A}\widehat{x} = b \text{ and } \widehat{x} \geq 0, \end{aligned}$$

where $\widehat{A}$ *is an* $m \times (n+m)$ *matrix,* $\widehat{c}$ *is a linear form in* $(\mathbb{R}^{n+m})^*$*, and* $\widehat{x} \in \mathbb{R}^{n+m}$*, given by*

$$\widehat{A} = \begin{pmatrix} A & I_m \end{pmatrix}, \quad \widehat{c} = \begin{pmatrix} c & 0_m^\top \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \overline{x} = \begin{pmatrix} x_{n+1} \\ \vdots \\ x_{n+m} \end{pmatrix}, \quad \widehat{x} = \begin{pmatrix} x \\ \overline{x} \end{pmatrix},$$

and the dual (D) *of* (P) *given by*

$$\begin{aligned} &\text{minimize} \quad yb \\ &\text{subject to} \quad yA \geq c \text{ and } y \geq 0, \end{aligned}$$

where $y \in (\mathbb{R}^m)^*$*. If the simplex algorithm applied to the linear program* $(P2)$ *terminates with an optimal solution* $(\widehat{u}^*, K^*)$*, where* $\widehat{u}^*$ *is a basic feasible solution and* K^* *is a basis for* $\widehat{u}^*$*, then* $y^* = \widehat{c}_{K^*}\widehat{A}_{K^*}^{-1}$ *is an optimal solution for* (D) *such that* $\widehat{c}\,\widehat{u}^* = y^*b$*. Furthermore,* y^* *is given in terms of the reduced costs by* $y^* = -((\overline{c}_{K^*})_{n+1} \ldots (\overline{c}_{K^*})_{n+m})$*.*

Proof. We know that K^* is a subset of $\{1, \ldots, n+m\}$ consisting of m indices such that the corresponding columns of $\widehat{A}$ are linearly independent. Let $N^* = \{1, \ldots, n+m\} - K^*$. The simplex methods terminates with an optimal solution in Case (A), namely when

$$\widehat{c}_j - \sum_{k \in k} \gamma_k^j \widehat{c}_k \leq 0 \quad \text{for all } j \in N^*,$$

where $\widehat{A}^j = \sum_{k \in K^*} \gamma_k^j \widehat{A}^k$, or using the notations of Section 26.3,

$$\widehat{c}_j - \widehat{c}_{K^*}\widehat{A}_{K^*}^{-1}\widehat{A}^j \leq 0 \quad \text{for all } j \in N^*.$$

The above inequalities can be written as

$$\widehat{c}_{N^*} - \widehat{c}_{K^*}\widehat{A}_{K^*}^{-1}\widehat{A}_{N^*} \leq 0_n^\top,$$

or equivalently as

$$\widehat{c}_{K^*}\widehat{A}_{K^*}^{-1}\widehat{A}_{N^*} \geq \widehat{c}_{N^*}. \tag{$*_1$}$$

The value of the objective function for the optimal solution $\widehat{u}^*$ is $\widehat{c}\,\widehat{u}^* = \widehat{c}_{K^*}\widehat{u}^*_{K^*}$, and since $\widehat{u}^*_{K^*}$ satisfies the equation $\widehat{A}_{K^*}\widehat{u}^*_{K^*} = b$, the value of the objective function is

$$\widehat{c}_{K^*}\,\widehat{u}^*_{K^*} = \widehat{c}_{K^*}\widehat{A}^{-1}_{K^*}b. \tag{$*_2$}$$

Then if we let $y^* = \widehat{c}_{K^*}\widehat{A}^{-1}_{K^*}$, obviously we have $y^*b = \widehat{c}_{K^*}\widehat{u}_{K^*}$, so if we can prove that y^* is a feasible solution of the dual linear program (D), by weak duality, y^* is an optimal solution of (D). We have

$$y^*\widehat{A}_{K^*} = \widehat{c}_{K^*}\widehat{A}^{-1}_{K^*}\widehat{A}_{K^*} = \widehat{c}_{K^*}, \tag{$*_3$}$$

and by $(*_1)$ we get

$$y^*\widehat{A}_{N^*} = \widehat{c}_{K^*}\widehat{A}^{-1}_{K^*}\widehat{A}_{N^*} \geq \widehat{c}_{N^*}. \tag{$*_4$}$$

Let P be the $(n+m)\times(n+m)$ permutation matrix defined so that

$$\widehat{A}\,P = \begin{pmatrix} A & I_m \end{pmatrix} P = \begin{pmatrix} \widehat{A}_{K^*} & \widehat{A}_{N^*} \end{pmatrix}.$$

Then we also have

$$\widehat{c}\,P = \begin{pmatrix} c & 0_m^\top \end{pmatrix} P = \begin{pmatrix} c_{K^*} & c_{N^*} \end{pmatrix}.$$

Using the equations $(*_3)$ and $(*_4)$ we obtain

$$y^*\begin{pmatrix} \widehat{A}_{K^*} & \widehat{A}_{N^*} \end{pmatrix} \geq \begin{pmatrix} c_{K^*} & c_{N^*} \end{pmatrix},$$

that is,

$$y^*\begin{pmatrix} A & I_m \end{pmatrix} P \geq \begin{pmatrix} c & 0_m^\top \end{pmatrix} P,$$

which is equivalent to

$$y^*\begin{pmatrix} A & I_m \end{pmatrix} \geq \begin{pmatrix} c & 0_m^\top \end{pmatrix},$$

that is

$$y^*A \geq c, \quad y \geq 0,$$

and these are exactly the conditions that say that y^* is a feasible solution of the dual program (D).

The reduced costs are given by $(\widehat{c}_{K^*})_i = \widehat{c}_i - \widehat{c}_{K^*}\widehat{A}^{-1}_{K^*}\widehat{A}^i$, for $i = 1,\ldots,n+m$. But for $i = n+1,\ldots,n+m$ each column $\widehat{A}^{n+j}$ is the jth vector of the identity matrix I_m, so

$$(\widehat{c}_{K^*})_{n+j} = -(\widehat{c}_{K^*}\widehat{A}^{-1}_{K^*})_j = -y^*_j \quad j = 1,\ldots,m,$$

as claimed. □

The fact that the above proof is fairly short is deceptive, because this proof relies on the fact that there are versions of the simplex algorithm using pivot rules that prevent cycling, but the proof that such pivot rules work correctly is quite lengthy. Other proofs are given

in Matousek and Gardner [72] (Chapter 6, Sections 6.3), Chvatal [29] (Chapter 5), and Papadimitriou and Steiglitz [79] (Section 2.7).

Observe that since the last m rows of the final tableau are actually obtained by multipling $[u\ \widehat{A}]$ by $\widehat{A}_{K^*}^{-1}$, the $m \times m$ matrix consisting of the last m columns and last m rows of the final tableau is $\widehat{A}_{K^*}^{-1}$ (basically, the simplex algorithm has performed the steps of a Gauss–Jordan reduction). This fact allows saving some steps in the primal dual method.

By combining weak duality and strong duality, we obtain the following theorem which shows that exactly four cases arise.

Theorem 27.9. *(Duality Theorem of Linear Programming) Let (P) be any linear program*

$$\begin{aligned} &\text{maximize} && cx \\ &\text{subject to} && Ax \leq b \text{ and } x \geq 0, \end{aligned}$$

and let (D) be its dual program

$$\begin{aligned} &\text{minimize} && yb \\ &\text{subject to} && yA \geq c \text{ and } y \geq 0, \end{aligned}$$

with A a $m \times n$ matrix. Then exactly one of the following possibilities occur:

(1) Neither (P) nor (D) has a feasible solution.

(2) (P) is unbounded and (D) has no feasible solution.

(3) (P) has no feasible solution and (D) is unbounded.

(4) Both (P) and (D) have a feasible solution. Then both have an optimal solution, and for every optimal solution x^* of (P) and every optimal solution y^* of (D), we have

$$cx^* = y^*b.$$

An interesting corollary of Theorem 27.9 is that there is a test to determine whether a linear program (P) has an optimal solution. Indeed, (P) has an optimal solution iff the following set of contraints is satisfiable:

$$\begin{aligned} Ax &\leq b \\ yA &\geq c \\ cx &\geq yb \\ x \geq 0,\ y &\geq 0_m^\top. \end{aligned}$$

In fact, for any feasible solution (x^*, y^*) of the above system, x^* is an optimal solution of (P) and y^* is an optimal solution of (D)

27.3 Complementary Slackness Conditions

Another useful corollary of the strong duality theorem is the following result known as the *equilibrium theorem*.

Theorem 27.10. *(Equilibrium Theorem) For any linear program (P) and its dual linear program (D) (with set of inequalities $Ax \leq b$ where A is an $m \times n$ matrix, and objective function $x \mapsto cx$), for any feasible solution x of (P) and any feasible solution y of (D), x and y are optimal solutions iff*

$$y_i = 0 \qquad \text{for all } i \text{ for which } \sum_{j=1}^{n} a_{ij}x_j < b_i \tag{$*_D$}$$

and

$$x_j = 0 \qquad \text{for all } j \text{ for which } \sum_{i=1}^{m} y_i a_{ij} > c_j. \tag{$*_P$}$$

Proof. First, assume that $(*_D)$ and $(*_P)$ hold. The equations in $(*_D)$ say that $y_i = 0$ unless $\sum_{j=1}^{n} a_{ij}x_j = b_i$, hence

$$yb = \sum_{i=1}^{m} y_i b_i = \sum_{i=1}^{m} y_i \sum_{j=1}^{n} a_{ij}x_j = \sum_{i=1}^{m}\sum_{j=1}^{n} y_i a_{ij} x_j.$$

Similarly, the equations in $(*_P)$ say that $x_j = 0$ unless $\sum_{i=1}^{m} y_i a_{ij} = c_j$, hence

$$cx = \sum_{j=1}^{n} c_j x_j = \sum_{j=1}^{n}\sum_{i=1}^{m} y_i a_{ij} x_j.$$

Consequently, we obtain

$$cx = yb.$$

By weak duality (Proposition 27.6), we have

$$cx \leq yb = cx$$

for all feasible solutions x of (P), so x is an optimal solution of (P). Similarly,

$$yb = cx \leq yb$$

for all feasible solutions y of (D), so y is an optimal solution of (D).

Let us now assume that x is an optimal solution of (P) and that y is an optimal solution of (D). Then, as in the proof of Proposition 27.6,

$$\sum_{j=1}^{n} c_j x_j \leq \sum_{i=1}^{m}\sum_{j=1}^{n} y_i a_{ij} x_j \leq \sum_{i=1}^{m} y_i b_i.$$

By strong duality, since x and y are optimal solutions the above inequalities are actually equalities, so in particular we have

$$\sum_{j=1}^{n}\Big(c_j - \sum_{i=1}^{m} y_i a_{ij}\Big)x_j = 0.$$

Since x and $y*$ are feasible, $x_i \geq 0$ and $y_j \geq 0$, so if $\sum_{i=1}^{m} y_i a_{ij} > c_j$, we must have $x_j = 0$. Similarly, we have

$$\sum_{i=1}^{m} y_i\Big(\sum_{j=1}^{m} a_{ij}x_j - b_i\Big) = 0,$$

so if $\sum_{j=1}^{m} a_{ij}x_j < b_i$, then $y_i = 0$. □

The equations in $(*_D)$ and $(*_P)$ are often called *complementary slackness conditions*. These conditions can be exploited to solve for an optimal solution of the primal problem with the help of the dual problem, and conversely. Indeed, if we guess a solution to one problem, then we may solve for a solution of the dual using the complementary slackness conditions, and then check that our guess was correct. This is the essence of the *primal-dual* methods. To present this method, first we need to take a closer look at the dual of a linear program already in standard form.

27.4 Duality for Linear Programs in Standard Form

Let (P) be a linear program in standard form, where $Ax = b$ for some $m \times n$ matrix of rank m and some objective function $x \mapsto cx$ (of course, $x \geq 0$). To obtain the dual of (P) we convert the equations $Ax = b$ to the following system of inequalities involving a $(2m) \times n$ matrix.

$$\begin{pmatrix} A \\ -A \end{pmatrix} x \leq \begin{pmatrix} b \\ -b \end{pmatrix}.$$

Then, if we denote the $2m$ dual variables by (y', y''), with $y', y'' \in (\mathbb{R}^m)^*$, the dual of the above program is

$$\begin{aligned} &\text{minimize} \quad y'b - y''b \\ &\text{subject to} \quad \begin{pmatrix} y' & y'' \end{pmatrix}\begin{pmatrix} A \\ -A \end{pmatrix} \geq c \text{ and } y', y'' \geq 0, \end{aligned}$$

where $y', y'' \in (\mathbb{R}^m)^*$, which is equivalent to

$$\begin{aligned} &\text{minimize} \quad (y' - y'')b \\ &\text{subject to} \quad (y' - y'')A \geq c \text{ and } y', y'' \geq 0, \end{aligned}$$

where $y', y'' \in (\mathbb{R}^m)^*$. If we write $y = y' - y''$, we find that the above linear program is equivalent to the following linear program (D):

$$\begin{aligned} &\text{minimize} && yb \\ &\text{subject to} && yA \geq c, \end{aligned}$$

where $y \in (\mathbb{R}^m)^*$. Observe that y is *not required* to be nonnegative; it is arbitrary.

Next, we would like to know what is the version of Theorem 27.8 for a linear program already in standard form. This is very simple.

Theorem 27.11. *Consider the linear program (P2) in standard form*

$$\begin{aligned} &\text{maximize} && cx \\ &\text{subject to} && Ax = b \text{ and } x \geq 0, \end{aligned}$$

and its dual (D) given by

$$\begin{aligned} &\text{minimize} && yb \\ &\text{subject to} && yA \geq c, \end{aligned}$$

where $y \in (\mathbb{R}^m)^*$. If the simplex algorithm applied to the linear program $(P2)$ terminates with an optimal solution (u^*, K^*), where u^* is a basic feasible solution and K^* is a basis for u^*, then $y^* = c_{K^*}A_{K^*}^{-1}$ is an optimal solution for (D) such that $cu^* = y^*b$. Furthermore, if we assume that the simplex algorithm is started with a basic feasible solution (u_0, K_0) where $K_0 = (n-m+1, \ldots, n)$ (the indices of the last m columns of A) and $A_{(n-m+1,\ldots,n)} = I_m$ (the last m columns of A constitute the identity matrix I_m), then the optimal solution $y^* = c_{K^*}A_{K^*}^{-1}$ for (D) is given in terms of the reduced costs by

$$y^* = c_{(n-m+1,\ldots,n)} - (\overline{c}_{K^*})_{(n-m+1,\ldots,n)},$$

and the $m \times m$ matrix consisting of last m columns and the last m rows of the final tableau is $A_{K^*}^{-1}$.

Proof. The proof of Theorem 27.8 applies with A instead of $\widehat{A}$ and we can show that

$$c_{K^*}A_{K^*}^{-1}A_{N^*} \geq c_{N^*},$$

and that $y^* = c_{K^*}A_{K^*}^{-1}$ satisfies, $cu^* = y^*b$, and

$$\begin{aligned} y^*A_{K^*} &= c_{K^*}A_{K^*}^{-1}A_{K^*} = c_{K^*}, \\ y^*A_{N^*} &= c_{K^*}A_{K^*}^{-1}A_{N^*} \geq c_{N^*}. \end{aligned}$$

Let P be the $n \times n$ permutation matrix defined so that

$$AP = \begin{pmatrix} A_{K^*} & A_{N^*} \end{pmatrix}.$$

Then we also have

$$cP = \begin{pmatrix} c_{K^*} & c_{N^*} \end{pmatrix},$$

and using the above equations and inequalities we obtain

$$y^* \begin{pmatrix} A_{K^*} & A_{N^*} \end{pmatrix} \geq \begin{pmatrix} c_{K^*} & c_{N^*} \end{pmatrix},$$

that is, $y^*AP \geq cP$, which is equivalent to

$$y^*A \geq c,$$

which shows that y^* is a feasible solution of (D) (remember, y^* is arbitrary so there is no need for the constraint $y^* \geq 0$).

The reduced costs are given by

$$(\overline{c}_{K^*})_i = c_i - c_{K^*}A_{K^*}^{-1}A^i,$$

and since for $j = n-m+1, \ldots, n$ the column A^j is the $(j+m-n)$th column of the identity matrix I_m, we have

$$(\overline{c}_{K^*})_j = c_j - (c_{K^*}A_{K^*})_{j+m-n} \quad j = n-m+1, \ldots, n,$$

that is,

$$y^* = c_{(n-m+1,\ldots,n)} - (\overline{c}_{K^*})_{(n-m+1,\ldots,n)},$$

as claimed. Since the last m rows of the final tableau is obtained by multiplying $[u_0 \;\; A]$ by $A_{K^*}^{-1}$, and the last m columns of A constitute I_m, the last m rows and the last m columns of the final tableau constitute $A_{K^*}^{-1}$. □

Let us now take a look at the complementary slackness conditions of Theorem 27.10. If we go back to the version of (P) given by

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad \begin{pmatrix} A \\ -A \end{pmatrix} x \leq \begin{pmatrix} b \\ -b \end{pmatrix} \text{ and } x \geq 0, \end{aligned}$$

and to the version of (D) given by

$$\begin{aligned} &\text{minimize} \quad y'b - y''b \\ &\text{subject to} \quad \begin{pmatrix} y' & y'' \end{pmatrix} \begin{pmatrix} A \\ -A \end{pmatrix} \geq c \text{ and } y', y'' \geq 0, \end{aligned}$$

where $y', y'' \in (\mathbb{R}^m)^*$, since the inequalities $Ax \leq b$ and $-Ax \leq -b$ together imply that $Ax = b$, we have equality for all these inequality constraints, and so the Conditions $(*_D)$ place no constraints at all on y' and y'', while the Conditions $(*_P)$ assert that

$$x_j = 0 \qquad \text{for all } j \text{ for which } \textstyle\sum_{i=1}^m (y'_i - y''_i)a_{ij} > c_j.$$

If we write $y = y' - y''$, the above conditions are equivalent to

$$x_j = 0 \qquad \text{for all } j \text{ for which } \sum_{i=1}^m y_i a_{ij} > c_j.$$

Thus we have the following version of Theorem 27.10.

Theorem 27.12. *(Equilibrium Theorem, Version 2) For any linear program $(P2)$ in standard form (with set of equalities $Ax \leq b$ where A is an $m \times n$ matrix, and objective function $x \mapsto cx$) and its dual linear program (D), for any feasible solution x of (P) and any feasible solution y of (D), x and y are optimal solutions iff*

$$x_j = 0 \qquad \textit{for all } j \textit{ for which } \sum_{i=1}^m y_i a_{ij} > c_j. \tag{$*_P$}$$

Therefore, the slackness conditions applied to a linear program $(P2)$ in standard form and to its dual (D) *only impose slackness conditions on the variables x_j of the primal problem.*

The above fact plays a crucial role in the primal-dual method.

27.5 The Dual Simplex Algorithm

Given a linear program $(P2)$ in standard form

$$\begin{aligned} &\text{maximize} && cx \\ &\text{subject to} && Ax = b \text{ and } x \geq 0, \end{aligned}$$

where A is an $m \times n$ matrix of rank m, if no obvious feasible solution is available but if $c \leq 0$, then rather than using the method for finding a feasible solution described in Section 26.2 we may use a method known as the dual simplex algorithm. This method uses basic solutions (u, K) where $Au = b$ and $u_j = 0$ for all $u_j \notin K$, but does not require $u \geq 0$, so u may not be feasible. However, $y = c_K A_K^{-1}$ is required to be feasible for the dual program

$$\begin{aligned} &\text{minimize} && yb \\ &\text{subject to} && yA \geq c, \end{aligned}$$

where $y \in (\mathbb{R}^*)^m$. Since $c \leq 0$, observe that $y = 0_m^\top$ is a feasible solution of the dual.

If a basic solution u of $(P2)$ is found such that $u \geq 0$, then $cu = yb$ for $y = c_K A_K^{-1}$, and we have found an optimal solution u for $(P2)$ and y for (D). The dual simplex method makes progress by attempting to make negative components of u zero and by decreasing the objective function of the dual program.

The dual simplex method starts with a basic solution (u, K) of $Ax = b$ which is not feasible but for which $y = c_K A_K^{-1}$ is dual feasible. In many cases, the original linear program is specified by a set of inequalities $Ax \leq b$ with some $b_i < 0$, so by adding slack variables it is

easy to find such basic solution u, and if in addition $c \leq 0$, then because the cost associated with slack variables is 0, we see that $y = 0$ is a feasible solution of the dual.

Given a basic solution (u, K) of $Ax = b$ (feasible or not), $y = c_K A_K^{-1}$ is dual feasible iff $c_K A_K^{-1} A \geq c$, and since $c_K A_K^{-1} A_K = c_K$, the inequality $c_K A_K^{-1} A \geq c$ is equivalent to $c_K A_K^{-1} A_N \geq c_N$, that is,

$$c_N - c_K A_K^{-1} A_N \leq 0, \tag{$*_1$}$$

where $N = \{1, \ldots, n\} - K$. Equation $(*_1)$ is equivalent to

$$c_j - c_K \gamma_K^j \leq 0 \quad \text{for all } j \in N, \tag{$*_2$}$$

where $\gamma_K^j = A_K^{-1} A^j$. Recall that the notation $\overline{c}_j$ is used to denote $c_j - c_K \gamma_K^j$, which is called the *reduced cost* of the variable x_j.

As in the simplex algorithm we need to decide which column A^k leaves the basis K and which column A^j enters the new basis K^+, in such a way that $y^+ = c_{K^+} A_{K^+}^{-1}$ is a feasible solution of (D), that is, $c_{N^+} - c_{K^+} A_{K^+}^{-1} A_{N^+} \leq 0$, where $N^+ = \{1, \ldots, n\} - K^+$. We use Proposition 26.2 to decide wich column k^- should leave the basis.

Suppose (u, K) is a solution of $Ax = b$ for which $y = c_K A_K^{-1}$ is dual feasible.

Case (A). If $u \geq 0$, then u is an optimal solution of $(P2)$.

Case (B). There is some $k \in K$ such that $u_k < 0$. In this case, pick some $k^- \in K$ such that $u_{k^-} < 0$ (according to some pivot rule).

Case (B1). Suppose that $\gamma_{k^-}^j \geq 0$ for all $j \notin K$ (in fact, for all j, since $\gamma_{k^-}^j \in \{0, 1\}$ for all $j \in K$). If so, we we claim that $(P2)$ is not feasible.

Indeed, let v be some basic feasible solution. We have $v \geq 0$ and $Av = b$, that is,

$$\sum_{j=1}^{n} v_j A^j = b,$$

so by multiplying both sides by A_K^{-1} and using the fact that by definition $\gamma_K^j = A_K^{-1} A^j$, we obtain

$$\sum_{j=1}^{n} v_j \gamma_K^j = A_K^{-1} b = u_K.$$

But recall that by hypothesis $u_{k^-} < 0$, yet $v_j \geq 0$ and $\gamma_{k^-}^j \geq 0$ for all j, so the component of index k^- is zero or positive on the left, and negative on the right, a contradiction. Therefore, $(P2)$ is indeed not feasible.

Case (B2). We have $\gamma_{k^-}^j < 0$ for some j.

We pick the column A^j entering the basis among those for which $\gamma_{k^-}^j < 0$. Since we assumed that $c_j - c_K \gamma_K^j \leq 0$ for all $j \in N$ by $(*_2)$, consider

$$\mu^+ = \max\left\{ -\frac{c_j - c_K \gamma_K^j}{\gamma_{k^-}^j} \;\middle|\; \gamma_{k^-}^j < 0,\ j \in N \right\} = \max\left\{ -\frac{\overline{c}_j}{\gamma_{k^-}^j} \;\middle|\; \gamma_{k^-}^j < 0,\ j \in N \right\} \leq 0,$$

and the set

$$N(\mu^+) = \left\{ j \in N \;\middle|\; -\frac{\overline{c}_j}{\gamma^j_{k^-}} = \mu^+ \right\}.$$

We pick some index $j^+ \in N(\mu^+)$ as the index of the column entering the basis (using some pivot rule).

Recall that by hypothesis $c_i - c_K\gamma^i_K \leq 0$ for all $j \notin K$ and $c_i - c_K\gamma^i_K = 0$ for all $i \in K$. Since $\gamma^{j^+}_{k^-} < 0$, for any index i such that $\gamma^i_{k^-} \geq 0$, we have $-\gamma^i_{k^-}/\gamma^{j^+}_{k^-} \geq 0$, and since by Proposition 26.2

$$c_i - c_{K^+}\gamma^i_{K^+} = c_i - c_K\gamma^i_K - \frac{\gamma^i_{k^-}}{\gamma^{j^+}_{k^-}}(c_{j^+} - c_K\gamma^{j^+}_K),$$

we have $c_i - c_{K^+}\gamma^i_{K^+} \leq 0$. For any index i such that $\gamma^i_{k^-} < 0$, by the choice of $j^+ \in K^*$,

$$-\frac{c_i - c_K\gamma^i_K}{\gamma^i_{k^-}} \leq -\frac{c_{j^+} - c_K\gamma^{j^+}_K}{\gamma^{j^+}_{k^-}},$$

so

$$c_i - c_K\gamma^i_K - \frac{\gamma^i_{k^-}}{\gamma^{j^+}_{k^-}}(c_{j^+} - c_K\gamma^{j^+}_K) \leq 0,$$

and again, $c_i - c_{K^+}\gamma^i_{K^+} \leq 0$. Therefore, if we let $K^+ = (K - \{k^-\}) \cup \{j^+\}$, then $y^+ = c_{K^+}A^{-1}_{K^+}$ is dual feasible. As in the simplex algorithm, θ^+ is given by

$$\theta^+ = u_{k^-}/\gamma^{j^+}_{k^-} \geq 0,$$

and u^+ is also computed as in the simplex algorithm by

$$u^+_i = \begin{cases} u_i - \theta^{j^+}\gamma^{j^+}_i & \text{if } i \in K \\ \theta^{j^+} & \text{if } i = j^+ \\ 0 & \text{if } i \notin K \cup \{j^+\} \end{cases}.$$

The change in the objective function of the prime and dual program (which is the same, since $u_K = A^{-1}_K b$ and $y = c_K A^{-1}_K$ is chosen such that $cu = c_K u_K = yb$) is the same as in the simplex algorithm, namely

$$\theta^+\left(c^{j^+} - c_K\gamma^{j^+}_K\right).$$

We have $\theta^+ > 0$ and $c^{j^+} - c_K\gamma^{j^+}_K \leq 0$, so if $c^{j^+} - c_K\gamma^{j^+}_K < 0$, then the objective function of the dual program decreases strictly.

Case (B3). $\mu^+ = 0$.

The possibity that $\mu^+ = 0$, that is, $c^{j^+} - c_K\gamma^{j^+}_K = 0$, may arise. In this case, the objective function doesn't change. This is a case of degeneracy similar to the degeneracy that arises in the simplex algorithm. We still pick $j^+ \in N(\mu^+)$, but we need a pivot rule that prevents

cycling. Such rules exist; see Bertsimas and Tsitsiklis [17] (Section 4.5) and Papadimitriou and Steiglitz [79] (Section 3.6).

The reader surely noticed that the dual simplex algorithm is very similar to the simplex algorithm, except that the simplex algorithm preserves the property that (u, K) is (primal) feasible, whereas the dual simplex algorithm preserves the property that $y = c_K A_K^{-1}$ is dual feasible. One might then wonder whether the dual simplex algorithm is equivalent to the simplex algorithm applied to the dual problem. This is indeed the case, there is a one-to-one correspondence between the dual simplex algorithm and the simplex algorithm applied to the dual problem. This correspondence is described in Papadimitriou and Steiglitz [79] (Section 3.7).

The comparison between the simplex algorithm and the dual simplex algorithm is best illustrated if we use a description of these methods in terms of *(full) tableaux*.

Recall that a *(full) tableau* is an $(m+1) \times (n+1)$ matrix organized as follows:

$-c_K u_K$	$\overline{c}_1$	$\cdots$	$\overline{c}_j$	$\cdots$	$\overline{c}_n$
u_{k_1}	γ_1^1	$\cdots$	γ_1^j	$\cdots$	γ_1^n
$\vdots$	$\vdots$		$\vdots$		$\vdots$
u_{k_m}	γ_m^1	$\cdots$	γ_m^j	$\cdots$	γ_m^n

The top row contains the current value of the objective function and the reduced costs, the first column except for its top entry contain the components of the current basic solution u_K, and the remaining columns except for their top entry contain the vectors γ_K^j. Observe that the γ_K^j corresponding to indices j in K constitute a permutation of the identity matrix I_m. A tableau together with the new basis $K^+ = (K - \{k^-\}) \cup \{j^+\}$ contains all the data needed to compute the new u_{K^+}, the new $\gamma_{K^+}^j$, and the new reduced costs $\overline{c}_i - (\gamma_{k^-}^i / \gamma_{k^-}^{j^+})\overline{c}_{j^+}$.

When executing the simplex algorithm, we have $u_k \geq 0$ for all $k \in K$ (and $u_j = 0$ for all $j \notin K$), and the incoming column j^+ is determined by picking one of the column indices such that $\overline{c}_j > 0$. Then, the index k^- of the leaving column is determined by looking at the minimum of the ratios $u_k/\gamma_k^{j^+}$ for which $\gamma_k^{j^+} > 0$ (along column j^+).

On the other hand, when executing the dual simplex algorithm, we have $\overline{c}_j \leq 0$ for all $j \notin K$ (and $\overline{c}_k = 0$ for all $k \in K$), and the outgoing column k^- is determined by picking one of the row indices such that $u_k < 0$. The index j^+ of the incoming column is determined by looking at the maximum of the ratios $-\overline{c}_j/\gamma_{k^-}^j$ for which $\gamma_{k^-}^j < 0$ (along row k^-).

More details about the comparison between the simplex algorithm and the dual simplex algorithm can be found in Bertsimas and Tsitsiklis [17] and Papadimitriou and Steiglitz [79].

Here is an example of the the dual simplex method.

Example 27.2. Consider the following linear program in standard form:

Maximize $-4x_1 - 2x_2 - x_3$

$$\text{subject to} \quad \begin{pmatrix} -1 & -1 & 2 & 1 & 0 & 0 \\ -4 & -2 & 1 & 0 & 1 & 0 \\ 1 & 1 & -4 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix} = \begin{pmatrix} -3 \\ -4 \\ 2 \end{pmatrix} \quad \text{and} \quad (x_1, x_2, x_3, x_4, x_5, x_6) \geq 0.$$

We initialize the dual simplex procedure with (u, K) where $u = \begin{pmatrix} 0 \\ 0 \\ 0 \\ -3 \\ -4 \\ 1 \end{pmatrix}$ and $K = (4, 5, 6)$.

The initial tableau, before explicitly calculating the reduced cost, is

0	$\overline{c}_1$	$\overline{c}_2$	$\overline{c}_3$	$\overline{c}_4$	$\overline{c}_5$	$\overline{c}_6$
$u_4 = -3$	-1	-1	2	1	0	0
$u_5 = -4$	-4	-2	1	0	1	0
$u_6 = 2$	1	1	-4	0	0	1

.

Since u has negative coordinates, Case (B) applies, and we will set $k^- = 4$. We must now determine whether Case (B1) or Case (B2) applies. This determination is accomplished by scanning the first three columns in the tableau, and observing each column has a negative entry. Thus Case (B2) is applicable, and we need to determine the reduced costs. Observe that $c = (-4, -2, -1, 0, 0, 0)$, which in turn implies $c_{(4,5,6)} = (0, 0, 0)$. Equation $(*_2)$ implies that the nonzero reduced costs are

$$\overline{c}_1 = c_1 - c_{(4,5,6)} \begin{pmatrix} -1 \\ -4 \\ 1 \end{pmatrix} = -4$$

$$\overline{c}_2 = c_2 - c_{(4,5,6)} \begin{pmatrix} -1 \\ -2 \\ 1 \end{pmatrix} = -2$$

$$\overline{c}_3 = c_3 - c_{(4,5,6)} \begin{pmatrix} -2 \\ 1 \\ 4 \end{pmatrix} = -1,$$

and our tableau becomes

0	-4	-2	-1	0	0	0
$u_4 = -3$	-1	-1	2	1	0	0
$u_5 = -4$	-4	-2	1	0	1	0
$u_6 = 2$	1	1	-4	0	0	1

.

Since $k^- = 4$, our pivot row is the first row of the tableau. To determine candidates for j^+, we scan this row, locate negative entries and compute

$$\mu^+ = \max\left\{ -\frac{\overline{c}_j}{\gamma_4^j} \;\middle|\; \gamma_4^j < 0,\ j \in \{1,2,3\} \right\} = \max\left\{ \frac{-2}{1}, \frac{-4}{1} \right\} = -2.$$

Since μ^+ occurs when $j = 2$, we set $j^+ = 2$. Our new basis is $K^+ = (2,5,6)$. We must normalize the first row of the tableau, namely multiply by -1, then add twice this normalized row to the second row, and subtract the normalized row from the third row to obtain the updated tableau.

0	-4	-2	-1	0	0	0
$u_2 = 3$	1	(1)	-2	-1	0	0
$u_5 = 2$	-2	0	-3	-2	1	0
$u_6 = -1$	0	0	-2	1	0	1

It remains to update the reduced costs and the value of the objective function by adding twice the normalized row to the top row.

6	-2	0	-5	-2	0	0
$u_2 = 3$	1	1	-2	-1	0	0
$u_5 = 2$	-2	0	-3	-2	1	0
$u_6 = -1$	0	0	(−2)	1	0	1

We now repeat the procedure of Case (B2) and set $k^- = 6$ (since this is the only negative entry of u^+). Our pivot row is now the third row of the updated tableaux, and the new μ^+ becomes

$$\mu^+ = \max\left\{ -\frac{\overline{c}_j}{\gamma_6^j} \;\middle|\; \gamma_6^j < 0,\ j \in \{1,3,4\} \right\} = \max\left\{ \frac{-5}{2} \right\} = -\frac{5}{2},$$

which implies that $j^+ = 3$. Hence the new basis is $K^+ = (2,5,3)$, and we update the tableau by taking $-\frac{1}{2}$ of Row 3, adding twice the normalized Row 3 to Row 1, and adding three times the normalized Row 3 to Row 2.

6	-2	0	-5	-2	0	0
$u_2 = 4$	1	1	0	-2	0	-1
$u_5 = 7/2$	-2	0	0	$-7/2$	1	$-3/2$
$u_3 = 1/2$	0	0	(1)	$-1/2$	0	$-1/2$

It remains to update the objective function and the reduced costs by adding five times the normalized row to the top row.

17/2	-2	0	0	$-9/2$	0	$-5/2$
$u_2 = 4$	1	1	0	-2	0	-1
$u_5 = 7/2$	-2	0	0	$-\frac{7}{2}$	1	$-3/2$
$u_3 = 1/2$	0	0	(1)	$-1/2$	0	$-1/2$

Since u^+ has no negative entries, the dual simplex method terminates and objective function $4x_1 - 2x_2 - x_3$ is maximized with $-\frac{17}{2}$ at $(0, 4, \frac{1}{2})$.

27.6 The Primal-Dual Algorithm

Let $(P2)$ be a linear program in standard form

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax = b \text{ and } x \geq 0, \end{aligned}$$

where A is an $m \times n$ matrix of rank m, and (D) be its dual given by

$$\begin{aligned} &\text{minimize} \quad yb \\ &\text{subject to} \quad yA \geq c, \end{aligned}$$

where $y \in (\mathbb{R}^m)^*$.

First, we may assume that $b \geq 0$ by changing every equation $\sum_{j=1}^n a_{ij}x_j = b_i$ with $b_i < 0$ to $\sum_{j=1}^n -a_{ij}x_j = -b_i$. If we happen to have some feasible solution y of the dual program (D), we know from Theorem 27.12 that a feasible solution x of $(P2)$ is an optimal solution iff the equations in $(*_P)$ hold. If we denote by J the subset of $\{1, \ldots, n\}$ for which the equalities

$$yA^j = c_j$$

hold, then by Theorem 27.12 a feasible solution x of $(P2)$ is an optimal solution iff

$$x_j = 0 \quad \text{for all } j \notin J.$$

Let $|J| = p$ and $N = \{1, \ldots, n\} - J$. The above suggests looking for $x \in \mathbb{R}^n$ such that

$$\begin{aligned} \sum_{j \in J} x_j A^j &= b \\ x_j &\geq 0 \quad \text{for all } j \in J \\ x_j &= 0 \quad \text{for all } j \notin J, \end{aligned}$$

or equivalently

$$A_J x_J = b, \quad x_J \geq 0, \tag{$*_1$}$$

and

$$x_N = 0_{n-p}.$$

To search for such an x, and just need to look for a feasible x_J, and for this we can use the *restricted primal* linear program (RP) defined as follows:

$$\begin{aligned} &\text{maximize} \quad -(\xi_1 + \cdots + \xi_m) \\ &\text{subject to} \quad \begin{pmatrix} A_J & I_m \end{pmatrix} \begin{pmatrix} x_J \\ \xi \end{pmatrix} = b \text{ and } x, \xi \geq 0. \end{aligned}$$

Since by hypothesis $b \geq 0$ and the objective function is bounded above by 0, this linear program has an optimal solution (x_J^*, ξ^*).

If $\xi^* = 0$, then the vector $u^* \in \mathbb{R}^n$ given by $u_J^* = x_J^*$ and $u_N^* = 0_{n-p}$ is an optimal solution of (P).

Otherwise, $\xi^* > 0$ and we have failed to solve $(*_1)$. However we may try to use ξ^* to improve y. For this, consider the dual (DRP) of (RP):

$$\begin{array}{ll} \text{minimize} & zb \\ \text{subject to} & zA_J \geq 0 \\ & z \geq -\mathbf{1}_m^\top. \end{array}$$

Observe that the program (DRP) has the same objective function as the original dual program (D). We know by Theorem 27.11 that the optimal solution (x_J^*, ξ^*) of (RP) yields an optimal solution z^* of (DRP) such that

$$z^*b = -(\xi_1^* + \cdots + \xi_m^*) < 0.$$

In fact, if K^* is the basis associated with (x_J^*, ξ^*) and if we write

$$\widehat{A} = \begin{pmatrix} A_J & I_m \end{pmatrix}$$

and $\widehat{c} = [0_p^\top \quad -\mathbf{1}^\top]$, then by Theorem 27.11 we have

$$z^* = \widehat{c}_{K^*}\widehat{A}_{K^*}^{-1} = -\mathbf{1}_m^\top - (\overline{c}_{K^*})_{(p+1,\ldots,p+m)},$$

where $(\overline{c}_{K^*})_{(p+1,\ldots,p+m)}$ denotes the row vector of reduced costs in the final tableau corresponding to the last m columns.

If we write

$$y(\theta) = y + \theta z^*,$$

then the new value of the objective function of (D) is

$$y(\theta)b = yb + \theta z^*b, \tag{$*_2$}$$

and since $z^*b < 0$, we have a chance of improving the objective function of (D), that is, decreasing its value for $\theta > 0$ small enough if $y(\theta)$ is feasible for (D). This will be the case iff $y(\theta)A \geq c$ iff

$$yA + \theta z^*A \geq c. \tag{$*_3$}$$

Now since y is a feasible solution of (D) we have $yA \geq c$, so if $z^*A \geq 0$ then $(*_3)$ is satisfied and $y(\theta)$ is a solution of (D) for all $\theta > 0$, which means that (D) is unbounded. But this implies that (P) is not feasible.

Let us take a closer look at the inequalities $z^*A \geq 0$. For $j \in J$, Since z^* is an optimal solution of (DRP), we know that $z^*A_J \geq 0$, so if $z^*A^j \geq 0$ for all $j \in N$, then (P) is not feasible.

Otherwise, there is some $j \in N = \{1, \ldots, n\} - J$ such that

$$z^* A^j < 0,$$

and then since by the definition of J we have $yA^j > c_j$ for all $j \in N$, if we pick $\theta > 0$ such that

$$\theta \leq \frac{yA^j - c_j}{-z^* A^j} \quad j \in N,\ z^* A^j < 0,$$

then we decrease the objective function $y(\theta)b = yb + \theta z^* b$ of (D) (since $z^* b < 0$). Therefore we pick the best θ, namely

$$\theta^+ = \min\left\{ \frac{yA^j - c_j}{-z^* A^j} \;\middle|\; j \notin J,\ z^* A^j < 0 \right\} > 0. \tag{$*_4$}$$

Next, we update y to $y^+ = y(\theta^+) = y + \theta^+ z^*$, we create the new restricted primal with the new subset

$$J^+ = \{j \in \{1, \ldots, n\} \mid y^+ A^j = c_j\},$$

and repeat the process. Here are the steps of the primal-dual algorithm.

Step 1. Find some feasible solution y of the dual program (D). We will show later that this is always possible.

Step 2. Compute

$$J^+ = \{j \in \{1, \ldots, n\} \mid yA^j = c_j\}.$$

Step 3. Set $J = J^+$ and solve the problem (RP) using the simplex algorithm, starting from the optimal solution determined during the previous round, obtaining the optimal solution (x_J^*, ξ^*) with the basis K^*.

Step 4.

If $\xi^* = 0$, then stop with an optimal solution u^* for (P) such that $u_J^* = x_J^*$ and the other components of u^* are zero.

Else let

$$z^* = -\mathbf{1}_m^\top - (\overline{c}_{K^*})_{(p+1,\ldots,p+m)},$$

be the optimal solution of (DRP) corresponding to (x_J^*, ξ^*) and the basis K^*.

If $z^* A^j \geq 0$ for all $j \notin J$, then stop; the program (P) has no feasible solution.

Else compute

$$\theta^+ = \min\left\{ -\frac{yA^j - c_j}{z^* A^j} \;\middle|\; j \notin J,\ z^* A^j < 0 \right\}, \qquad y^+ = y + \theta^+ z^*,$$

and

$$J^+ = \{j \in \{1, \ldots, n\} \mid y^+ A^j = c_j\}.$$

Go back to Step 3.

The following proposition shows that at each iteration we can start the program (RP) with the optimal solution obtained at the previous iteration.

Proposition 27.13. *Every $j \in J$ such that A^j is in the basis of the optimal solution ξ^* belongs to the next index set J^+.*

Proof. Such an index $j \in J$ correspond to a variable ξ_j such that $\xi_j > 0$, so by complementary slackness, the constraint $z^* A^j \geq 0$ of the dual program (DRP) must be an equality, that is, $z^* A^j = 0$. But then, we have

$$y^+ A^j = y A^j + \theta^+ z^* A^j = c_j,$$

which shows that $j \in J^+$. □

If (u^*, ξ^*) with the basis K^* is the optimal solution of the program (RP), Proposition 27.13 together with the last property of Theorem 27.11 allows us to restart the (RP) in Step 3 with $(u^*, \xi^*)_{K^*}$ as initial solution (with basis K^*). For every $j \in J - J^+$, column j is deleted, and for every $j \in J^+ - J$, the new column A^j is computed by multiplying $\widehat{A}_{K^*}^{-1}$ and A^j, but $\widehat{A}_{K^*}^{-1}$ is the matrix $\Gamma^*[1{:}m; p+1{:}p+m]$ consisting of the last m columns of Γ^* in the final tableau, and the new reduced $\overline{c}_j$ is given by $c_j - z^* A^j$. Reusing the optimal solution of the previous (RP) may improve efficiency significantly.

Another crucial observation is that for any index $j_0 \in N$ such that
$\theta^+ = (y A^{j_0} - c_{j_0})/(-z^* A^{j_0})$, we have

$$y^+ A_{j_0} = y A_{j_0} + \theta^+ z^* A^{j_0} = c_{j_0},$$

and so $j_0 \in J^+$. This fact that be used to ensure that the primal-dual algorithm terminates in a finite number of steps (using a pivot rule that prevents cycling); see Papadimitriou and Steiglitz [79] (Theorem 5.4).

It remains to discuss how to pick some initial feasible solution y of the dual program (D). If $c_j \leq 0$ for $j = 1, \ldots, n$, then we can pick $y = 0$.

We should note that in many applications, the natural primal optimization problem is actually the *minimization* some objective function $cx = c_1 x_1 + \cdots + c_n x_n$, rather its maximization. For example, many of the optimization problems considered in Papadimitriou and Steiglitz [79] are minimization problems.

Of course, minimizing cx is equivalent to maximizing $-cx$, so our presentation covers minimization too. But if we are dealing with a minimization problem, the weight c_j are often nonnegative, so from the point of view of maximization we will have $-c_j \leq 0$ for all j, and we will be able to use $y = 0$ as a starting point.

Going back to our primal problem in maximization form and its dual in minimization form, we still need to deal with the situation where $c_j > 0$ for some j, in which case there may not be any obvious y feasible for (D). Preferably we would like to find such a y very cheaply.

There is a trick to deal with this situation. We pick some very large positive number M and add to the set of equations $Ax = b$ the new equation

$$x_1 + \cdots + x_n + x_{n+1} = M,$$

with the new variable x_{n+1} constrained to be nonnegative. If the program (P) has a feasible solution, such an M exists. In fact, it can shown that for any basic feasible solution $u = (u_1, \ldots, u_n)$, each $|u_i|$ is bounded by some expression depending only on A and b; see Papadimitriou and Steiglitz [79] (Lemma 2.1). The proof is not difficult and relies on the fact that the inverse of a matrix can be expressed in terms of certain determinants (the adjugates). Unfortunately, this bound contains $m!$ as a factor, which makes it quite impractical.

Having added the new equation above, we obtain the new set of equations

$$\begin{pmatrix} A & 0_n \\ \mathbf{1}_n^\top & 1 \end{pmatrix} \begin{pmatrix} x \\ x_{n+1} \end{pmatrix} = \begin{pmatrix} b \\ M \end{pmatrix},$$

with $x \geq 0, x_{n+1} \geq 0$, and the new objective function given by

$$\begin{pmatrix} c & 0 \end{pmatrix} \begin{pmatrix} x \\ x_{n+1} \end{pmatrix} = cx.$$

The dual of the above linear program is

$$\begin{array}{ll} \text{minimize} & yb + y_{m+1}M \\ \text{subject to} & yA^j + y_{m+1} \geq c_j \quad j = 1, \ldots, n \\ & y_{m+1} \geq 0. \end{array}$$

If $c_j > 0$ for some j, observe that the linear form $\widetilde{y}$ given by

$$\widetilde{y}_i = \begin{cases} 0 & \text{if } 1 \leq i \leq m \\ \max_{1 \leq j \leq n}\{c_j\} > 0 & \end{cases}$$

is a feasible solution of the new dual program. In practice, we can choose M to be a number close to the largest integer representable on the computer being used.

Here is an example of the primal-dual algorithm given in the Math 588 class notes of T. Molla.

Example 27.3. Consider the following linear program in standard form:

$$\begin{array}{ll} \text{Maximize} & -x_1 - 3x_2 - 3x_3 - x_4 \\ \text{subject to} & \begin{pmatrix} 3 & 4 & -3 & 1 \\ 3 & -2 & 6 & -1 \\ 6 & 4 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 4 \end{pmatrix} \text{ and } x_1, x_2, x_3, x_4 \geq 0. \end{array}$$

The associated dual program (D) is

$$\begin{aligned} &\text{Minimize} && 2y_1 + y_2 + 4y_3 \\ &\text{subject to} && \begin{pmatrix} y_1 & y_2 & y_3 \end{pmatrix} \begin{pmatrix} 3 & 4 & -3 & 1 \\ 3 & -2 & 6 & -1 \\ 6 & 4 & 0 & 1 \end{pmatrix} \geq \begin{pmatrix} -1 \\ -3 \\ -3 \\ -1 \end{pmatrix}. \end{aligned}$$

We initialize the primal-dual algorithm with the dual feasible point $y = (-1/3 \;\; 0 \;\; 0)$. Observe that only the first inequality of (D) is actually an equality, and hence $J = \{1\}$. We form the restricted primal program $(RP1)$

$$\begin{aligned} &\text{Maximize} && -(\xi_1 + \xi_2 + \xi_3) \\ &\text{subject to} && \begin{pmatrix} 3 & 1 & 0 & 0 \\ 3 & 0 & 1 & 0 \\ 6 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 4 \end{pmatrix} \text{ and } x_1, \xi_1, \xi_2, \xi_3 \geq 0. \end{aligned}$$

We now solve $(RP1)$ via the simplex algorithm. The initial tableau with $K = (2, 3, 4)$ and $J = \{1\}$ is

	x_1	ξ_1	ξ_2	ξ_3
7	12	0	0	0
$\xi_1 = 2$	3	1	0	0
$\xi_2 = 1$	(3)	0	1	0
$\xi_3 = 4$	6	0	0	1

.

For $(RP1)$, $c = (0, -1, -1, -1)$, $(x_1, \xi_1, \xi_2, \xi_3) = (0, 2, 1, 4)$, and the nonzero reduced cost is given by

$$0 - (-1 \;\; -1 \;\; -1) \begin{pmatrix} 3 \\ 3 \\ 6 \end{pmatrix} = 12.$$

Since there is only one nonzero reduced cost, we must set $j^+ = 1$. Since $\min\{\xi_1/3, \xi_2/3, \xi_3/6\} = 1/3$, we see that $k^- = 3$ and $K = (2, 1, 4)$. Hence we pivot through the red circled 3 (namely we divide row 2 by 3, and then subtract $3\times$ (row 2) from row 1, $6\times$ (row 2) from row 3, and $12\times$ (row 2) from row 0), to obtain the tableau

	x_1	ξ_1	ξ_2	ξ_3
3	0	0	-4	0
$\xi_1 = 1$	0	1	-1	0
$x_1 = 1/3$	1	0	$1/3$	0
$\xi_3 = 2$	0	0	-2	1

.

At this stage the simplex algorithm for $(RP1)$ terminates since there are no positive reduced costs. Since the upper left corner of the final tableau is not zero, we proceed with Step 4 of

the primal dual algorithm and compute

$$z^* = (-1 \ \ -1 \ \ -1) - (0 \ \ -4 \ \ 0) = (-1 \ \ 3 \ \ -1),$$

$$(-1/3 \ \ 0 \ \ 0)\begin{pmatrix} 4 \\ -2 \\ 4 \end{pmatrix} + 3 = \frac{5}{3}, \qquad -(-1 \ \ 3 \ \ -1)\begin{pmatrix} 4 \\ -2 \\ 4 \end{pmatrix} = 14,$$

$$(-1/3 \ \ 0 \ \ 0)\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} + 1 = \frac{2}{3}, \qquad -(-1 \ \ 3 \ \ -1)\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} = 5,$$

so

$$\theta^+ = \min\left\{\frac{5}{42}, \frac{2}{15}\right\} = \frac{5}{42},$$

and we conclude that the new feasible solution for (D) is

$$y^+ = (-1/3 \ \ 0 \ \ 0) + \frac{5}{42}(-1 \ \ 3 \ \ -1) = (-19/42 \ \ 5/14 \ \ -5/42).$$

When we substitute y^+ into (D), we discover that the first two constraints are equalities, and that the new J is $J = \{1, 2\}$. The new reduced primal $(RP2)$ is

$$\begin{array}{ll} \text{Maximize} & -(\xi_1 + \xi_2 + \xi_3) \\ \text{subject to} & \begin{pmatrix} 3 & 4 & 1 & 0 & 0 \\ 3 & -2 & 0 & 1 & 0 \\ 6 & 4 & 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 4 \end{pmatrix} \ \text{ and } \ x_1, x_2, \xi_1, \xi_2, \xi_3 \geq 0. \end{array}$$

Once again, we solve $(RP2)$ via the simplex algorithm, where $c = (0, 0, -1, -1, -1)$, $(x_1, x_2, \xi_1, \xi_2, \xi_3) = (1/3, 0, 1, 0, 2)$ and $K = (3, 1, 5)$. The initial tableau is obtained from the final tableau of the previous $(RP1)$ by adding a column corresponding the the variable x_2, namely

$$\widehat{A}_K^{-1} A^2 = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1/3 & 0 \\ 0 & -2 & 1 \end{pmatrix}\begin{pmatrix} 4 \\ -2 \\ 4 \end{pmatrix} = \begin{pmatrix} 6 \\ -2/3 \\ 8 \end{pmatrix},$$

with

$$\overline{c}_2 = c_2 - z^* A^2 = 0 - \begin{pmatrix} -1 & 3 & -1 \end{pmatrix}\begin{pmatrix} 4 \\ -2 \\ 4 \end{pmatrix} = 14,$$

and we get

	x_1	x_2	ξ_1	ξ_2	ξ_3
3	0	14	0	-4	0
$\xi_1 = 1$	0	(6)	1	-1	0
$x_1 = 1/3$	1	$-2/3$	0	$1/3$	0
$\xi_3 = 2$	0	8	0	-2	1

.

Note that $j^+ = 2$ since the only positive reduced cost occurs in column 2. Also observe that since $\min\{\xi_1/6, \xi_3/8\} = \xi_1/6 = 1/6$, we set $k^- = 3$, $K = (2, 1, 5)$ and pivot along the red 6 to obtain the tableau

	x_1	x_2	ξ_1	ξ_2	ξ_3
2/3	0	0	$-7/3$	$-5/3$	0
$x_2 = 1/6$	0	1	1/6	$-1/6$	0
$x_1 = 4/9$	1	0	1/9	2/9	0
$\xi_3 = 2/3$	0	0	$-4/3$	$-2/3$	1

.

Since the reduced costs are either zero or negative the simplex algorithm terminates, and we compute

$$z^* = (-1 \;\; -1 \;\; -1) - (-7/3 \;\; -5/3 \;\; 0) = (4/3 \;\; 2/3 \;\; -1),$$

$$(-19/42 \;\; 5/14 \;\; -5/42) \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} + 1 = 1/14, \qquad -(4/3 \;\; 2/3 \;\; -1) \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} = 1/3,$$

so

$$\theta^+ = \frac{3}{14},$$

$$y^+ = (-19/42 \;\; 5/14 \;\; -5/42) + \frac{5}{14}(4/3 \;\; 2/3 \;\; -1) = (-1/6 \;\; 1/2 \;\; -1/3).$$

When we plug y^+ into (D), we discover that the first, second, and fourth constraints are equalities, which implies $J = \{1, 2, 4\}$. Hence the new restricted primal $(RP3)$ is

$$\text{Maximize} \quad -(\xi_1 + \xi_2 + \xi_3)$$

$$\text{subject to} \quad \begin{pmatrix} 3 & 4 & 1 & 1 & 0 & 0 \\ 3 & -2 & -1 & 0 & 1 & 0 \\ 6 & 4 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_4 \\ \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 4 \end{pmatrix} \quad \text{and} \quad x_1, x_2, x_4, \xi_1, \xi_2, \xi_3 \geq 0.$$

The initial tableau for $(RP3)$, with $c = (0, 0, 0, -1, -1, -1)$, $(x_1, x_2, x_4, \xi_1, \xi_2, \xi_3) = (4/9, 1/6, 0, 0, 0, 2/3)$ and $K = (2, 1, 6)$, is obtained from the final tableau of the previous $(RP2)$ by adding a column corresponding the the variable x_4, namely

$$\widehat{A}_K^{-1} A^4 = \begin{pmatrix} 1/6 & -1/6 & 0 \\ 1/9 & 2/9 & 0 \\ -4/3 & -2/3 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1/3 \\ -1/9 \\ 1/3 \end{pmatrix},$$

with

$$\overline{c}_4 = c_4 - z^* A^4 = 0 - \begin{pmatrix} 4/3 & 2/3 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} = 1/3,$$

and we get

	x_1	x_2	x_4	ξ_1	ξ_2	ξ_3
$2/3$	0	0	$1/3$	$-7/3$	$-5/3$	0
$x_2 = 1/6$	0	1	$1/3$	$1/6$	$-1/6$	0
$x_1 = 4/9$	1	0	$-1/9$	$1/9$	$2/9$	0
$\xi_3 = 2/3$	0	0	$1/3$	$-4/3$	$-2/3$	1

.

Since the only positive reduced cost occurs in column 3, we set $j^+ = 3$. Furthermore since $\min\{x_2/(1/3), \xi_3/(1/3)\} = x_2/(1/3) = 1/2$, we let $k^- = 2$, $K = (3, 1, 6)$, and pivot around the red circled $1/3$ to obtain

	x_1	x_2	x_4	ξ_1	ξ_2	ξ_3
$1/2$	0	-1	0	$-5/2$	$-3/2$	0
$x_4 = 1/2$	0	3	1	$1/2$	$-1/2$	0
$x_1 = 1/2$	1	$1/3$	0	$1/6$	$1/6$	0
$\xi_3 = 1/2$	0	-1	0	$-3/2$	$-1/2$	1

.

At this stage, there are no positive reduced costs, and we must compute

$$z^* = (-1 \;\; -1 \;\; -1) - (-5/2 \;\; -3/2 \;\; 0) = (3/2 \;\; 1/2 \;\; -1),$$

$$(-1/6 \;\; 1/2 \;\; -1/3)\begin{pmatrix} -3 \\ 6 \\ 0 \end{pmatrix} + 3 = 13/2, \qquad -(3/2 \;\; 1/2 \;\; -1)\begin{pmatrix} -3 \\ 6 \\ 0 \end{pmatrix} = 3/2,$$

so

$$\theta^+ = \frac{13}{3},$$
$$y^+ = (-1/6 \;\; 1/2 \;\; -1/3) + \frac{13}{3}(3/2 \;\; 1/2 \;\; -1) = (19/3 \;\; 8/3 \;\; -14/3).$$

We plug y^+ into (D) and discover that the first, third, and fourth constraints are equalities. Thus, $J = \{1, 3, 4\}$ and the restricted primal $(RP4)$ is

$$\text{Maximize} \quad -(\xi_1 + \xi_2 + \xi_3)$$

$$\text{subject to} \quad \begin{pmatrix} 3 & -3 & 1 & 1 & 0 & 0 \\ 3 & 6 & -1 & 0 & 1 & 0 \\ 6 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_3 \\ x_4 \\ \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 4 \end{pmatrix} \quad \text{and} \quad x_1, x_3, x_4, \xi_1, \xi_2, \xi_3 \geq 0.$$

The initial tableau for $(RP4)$, with $c = (0, 0, 0, -1, -1, -1)$, $(x_1, x_3, x_4, \xi_1, \xi_2, \xi_3) = (1/2, 0, 1/2, 0, 0, 1/2)$ and $K = (3, 1, 6)$ is obtained from the final tableau of the previous $(RP3)$ by replacing the column corresponding to the variable x_2 by a column corresponding to the variable x_3, namely

$$\widehat{A}_K^{-1} A^3 = \begin{pmatrix} 1/2 & -1/2 & 0 \\ 1/6 & 1/6 & 0 \\ -3/2 & -1/2 & 1 \end{pmatrix} \begin{pmatrix} -3 \\ 6 \\ 0 \end{pmatrix} = \begin{pmatrix} -9/2 \\ 1/2 \\ 3/2 \end{pmatrix},$$

with

$$\overline{c}_3 = c_3 - z^* A^3 = 0 - \begin{pmatrix} 3/2 & 1/2 & -1 \end{pmatrix} \begin{pmatrix} -3 \\ 6 \\ 0 \end{pmatrix} = 3/2,$$

and we get

	x_1	x_3	x_4	ξ_1	ξ_2	ξ_3
$1/2$	0	$3/2$	0	$-5/2$	$-3/2$	0
$x_4 = 1/2$	0	$-9/2$	1	$1/2$	$-1/2$	0
$x_1 = 1/2$	1	$1/2$	0	$1/6$	$1/6$	0
$\xi_3 = 1/2$	0	$3/2$	0	$-3/2$	$-1/2$	1

.

By analyzing the top row of reduced cost, we see that $j^+ = 2$. Furthermore, since $\min\{x_1/(1/2), \xi_3/(3/2)\} = \xi_3/(3/2) = 1/3$, we let $k^- = 6$, $K = (3, 1, 2)$, and pivot along the red circled $3/2$ to obtain

	x_1	x_3	x_4	ξ_1	ξ_2	ξ_3
0	0	0	0	-1	-1	-1
$x_4 = 2$	0	0	1	-4	-2	3
$x_1 = 1/3$	1	0	0	$2/3$	$1/3$	$-1/3$
$x_3 = 1/3$	0	1	0	-1	$-1/3$	$2/3$

.

Since the upper left corner of the final tableau is zero and the reduced costs are all ≤ 0, we are finally finished. Then $y = (19/3 \ 8/3 \ -14/3)$ is an optimal solution of (D), but more importantly $(x_1, x_2, x_3, x_4) = (1/3, 0, 1/3, 2)$ is an optimal solution for our original linear program and provides an optimal value of $-10/3$.

The primal-dual algorithm for linear programming doesn't seem to be the favorite method to solve linear programs nowadays. But it is important because its basic principle, to use a restricted (simpler) primal problem involving an objective function with fixed weights, namely 1, and the dual problem to provide feedback to the primal by improving the objective function of the dual, has led to a whole class of combinatorial algorithms (often approximation algorithms) based on the primal-dual paradigm. The reader will get a taste of this kind of algorithm by consulting Papadimitriou and Steiglitz [79], where it is explained

how classical algorithms such as Dijkstra's algorithm for the shortest path problem, and Ford and Fulkerson's algorithm for max flow can be derived from the primal-dual paradigm.

Part IV

NonLinear Optimization

Chapter 28

Basics of Hilbert Spaces

Most of the "deep" results about the existence of minima of real-valued functions proven in Chapter 29 rely on two fundamental results of Hilbert space theory:

(1) The projection lemma, which is a result about nonempty, closed, convex subsets of a Hilbert space V.

(2) The Riesz representation theorem, which allows us to express a continuous linear form on a Hilbert space V in terms of a vector in V and the inner product on V.

The correctness of the Karush–Kuhn–Tucker conditions appearing in Lagrangian duality follows from a version of the Farkas–Minkowski proposition, which also follows from the projection lemma.

Thus we feel that it is indispensible to review some basic results of Hilbert space theory, although in most applications considered here the Hilbert space in question will be finite-dimensional. However, in optimization theory, there are many problems where we seek to find a *function* minimizing some type of energy functional (often given by a bilinear form), in which case we are dealing with an infinite dimensional Hilbert space, so it necessary to develop tools to deal with the more general situation of infinite-dimensional Hilbert spaces.

28.1 The Projection Lemma, Duality

Given a Hermitian space $\langle E, \varphi \rangle$, we showed in Section 11.1 that the function $\| \| \colon E \to \mathbb{R}$ defined such that $\|u\| = \sqrt{\varphi(u,u)}$, is a norm on E. Thus, E is a normed vector space. If E is also complete, then it is a very interesting space.

Recall that completeness has to do with the convergence of Cauchy sequences. A normed vector space $\langle E, \| \| \rangle$ is automatically a metric space under the metric d defined such that $d(u,v) = \|v - u\|$ (see Chapter 18 for the definition of a normed vector space and of a metric space, or Lang [64, 65], or Dixmier [35]). Given a metric space E with metric d, a sequence

$(a_n)_{n\geq 1}$ of elements $a_n \in E$ is a *Cauchy sequence* iff for every $\epsilon > 0$, there is some $N \geq 1$ such that

$$d(a_m, a_n) < \epsilon \quad \text{for all} \quad m, n \geq N.$$

We say that E is *complete* iff every Cauchy sequence converges to a limit (which is unique, since a metric space is Hausdorff).

Every finite dimensional vector space over $\mathbb{R}$ or $\mathbb{C}$ is complete. For example, one can show by induction that given any basis $(e_1, \ldots, e_n)$ of E, the linear map $h\colon \mathbb{C}^n \to E$ defined such that

$$h((z_1, \ldots, z_n)) = z_1 e_1 + \cdots + z_n e_n$$

is a homeomorphism (using the *sup*-norm on $\mathbb{C}^n$). One can also use the fact that any two norms on a finite dimensional vector space over $\mathbb{R}$ or $\mathbb{C}$ are equivalent (see Chapter 6, or Lang [65], Dixmier [35], Schwartz [90]).

However, if E has infinite dimension, it may not be complete. When a Hermitian space is complete, a number of the properties that hold for finite dimensional Hermitian spaces also hold for infinite dimensional spaces. For example, any closed subspace has an orthogonal complement, and in particular, a finite dimensional subspace has an orthogonal complement. Hermitian spaces that are also complete play an important role in analysis. Since they were first studied by Hilbert, they are called Hilbert spaces.

Definition 28.1. A (complex) Hermitian space $\langle E, \varphi\rangle$ which is a complete normed vector space under the norm $\|\ \|$ induced by φ is called a *Hilbert space*. A real Euclidean space $\langle E, \varphi\rangle$ which is complete under the norm $\|\ \|$ induced by φ is called a *real Hilbert space*.

All the results in this section hold for complex Hilbert spaces as well as for real Hilbert spaces. We state all results for the complex case only, since they also apply to the real case, and since the proofs in the complex case need a little more care.

Example 28.1. The space l^2 of all countably infinite sequences $x = (x_i)_{i\in\mathbb{N}}$ of complex numbers such that $\sum_{i=0}^{\infty} |x_i|^2 < \infty$ is a Hilbert space. It will be shown later that the map $\varphi\colon l^2 \times l^2 \to \mathbb{C}$ defined such that

$$\varphi\left((x_i)_{i\in\mathbb{N}}, (y_i)_{i\in\mathbb{N}}\right) = \sum_{i=0}^{\infty} x_i \overline{y_i}$$

is well defined, and that l^2 is a Hilbert space under φ. In fact, we will prove a more general result (Proposition 33.3).

Example 28.2. The set $\mathcal{C}^\infty[a, b]$ of smooth functions $f\colon [a, b] \to \mathbb{C}$ is a Hermitian space under the Hermitian form

$$\langle f, g\rangle = \int_a^b f(x)\overline{g(x)}dx,$$

but it is not a Hilbert space because it is not complete. It is possible to construct its completion $\mathrm{L}^2([a, b])$, which turns out to be the space of Lebesgue integrable functions on $[a, b]$.

Theorem 18.22 yields a quick proof of the fact that any Hermitian space E (with Hermitian product $\langle -, - \rangle$) can be embedded in a Hilbert space E_h.

Theorem 28.1. *Given a Hermitian space $(E, \langle -, - \rangle)$ (resp. Euclidean space), there is a Hilbert space $(E_h, \langle -, - \rangle_h)$ and a linear map $\varphi \colon E \to E_h$, such that*

$$\langle u, v \rangle = \langle \varphi(u), \varphi(v) \rangle_h$$

for all $u, v \in E$, and $\varphi(E)$ is dense in E_h. Furthermore, E_h is unique up to isomorphism.

Proof. Let $(\widehat{E}, \| \ \|_{\widehat{E}})$ be the Banach space, and let $\varphi \colon E \to \widehat{E}$ be the linear isometry, given by Theorem 18.22. Let $\|u\| = \sqrt{\langle u, u \rangle}$ and $E_h = \widehat{E}$. If E is a real vector space, we know from Section 9.1 that the inner product $\langle -, - \rangle$ can be expressed in terms of the norm $\|u\|$ by the polarity equation

$$\langle u, v \rangle = \frac{1}{2}(\|u + v\|^2 - \|u\|^2 - \|v\|^2),$$

and if E is a complex vector space, we know from Section 11.1 that we have the polarity equation

$$\langle u, v \rangle = \frac{1}{4}(\|u + v\|^2 - \|u - v\|^2 + i\|u + iv\|^2 - i\|u - iv\|^2).$$

By the Cauchy-Schwarz inequality, $|\langle u, v \rangle| \leq \|u\| \|v\|$, the map $\langle -, - \rangle \colon E \times E \to \mathbb{C}$ (resp. $\langle -, - \rangle \colon E \times E \to \mathbb{R}$) is continuous. However, it is not uniformly continuous, but we can get around this problem by using the polarity equations to extend it to a continuous map. By continuity, the polarity equations also hold in E_h, which shows that $\langle -, - \rangle$ extends to a positive definite Hermitian inner product (resp. Euclidean inner product) $\langle -, - \rangle_h$ on E_h induced by $\| \ \|_{\widehat{E}}$ extending $\langle -, - \rangle$. □

Proof. We followed the approach in Schwartz [89] (Chapter XXIII, Section 42. Theorem 2). For other approaches, see Munkres [77] (Chapter 7, Section 43), and Bourbaki [21].

One of the most important facts about finite-dimensional Hermitian (and Euclidean) spaces is that they have orthonormal bases. This implies that, up to isomorphism, every finite-dimensional Hermitian space is isomorphic to $\mathbb{C}^n$ (for some $n \in \mathbb{N}$) and that the inner product is given by

$$\langle (x_1, \ldots, x_n), (y_1, \ldots, y_n) \rangle = \sum_{i=1}^{n} x_i \overline{y_i}.$$

Furthermore, every subspace W has an orthogonal complement $W^\perp$, and the inner product induces a natural duality between E and E^* (actually, between $\overline{E}$ and E^*) where E^* is the space of linear forms on E.

When E is a Hilbert space, E may be infinite dimensional, often of uncountable dimension. Thus, we can't expect that E always have an orthonormal basis. However, if we modify the notion of basis so that a "Hilbert basis" is an orthogonal family that is also dense in E,

i.e., every $v \in E$ is the limit of a sequence of finite combinations of vectors from the Hilbert basis, then we can recover most of the "nice" properties of finite-dimensional Hermitian spaces. For instance, if $(u_k)_{k\in K}$ is a Hilbert basis, for every $v \in E$, we can define the Fourier coefficients $c_k = \langle v, u_k\rangle/\|u_k\|$, and then, v is the "sum" of its Fourier series $\sum_{k\in K} c_k u_k$. However, the cardinality of the index set K can be very large, and it is necessary to define what it means for a family of vectors indexed by K to be summable. We will do this in Section 33.1. It turns out that every Hilbert space is isomorphic to a space of the form $l^2(K)$, where $l^2(K)$ is a generalization of the space of Example 28.1 (see Theorem 33.8, usually called the Riesz-Fischer theorem).

Our first goal is to prove that a closed subspace of a Hilbert space has an orthogonal complement. We also show that duality holds if we redefine the dual E' of E to be the space of *continuous* linear maps on E. Our presentation closely follows Bourbaki [21]. We also were inspired by Rudin [82], Lang [64, 65], Schwartz [90, 89], and Dixmier [35]. In fact, we highly recommend Dixmier [35] as a clear and simple text on the basics of topology and analysis. We first prove the so-called projection lemma.

Recall that in a metric space E, a subset X of E is *closed* iff for every convergent sequence (x_n) of points $x_n \in X$, the limit $x = \lim_{n\to\infty} x_n$ also belongs to X. The *closure* $\overline{X}$ *of* X is the set of all limits of convergent sequences (x_n) of points $x_n \in X$. Obviously, $X \subseteq \overline{X}$. We say that the subset X of E is *dense in* E iff $E = \overline{X}$, the closure of X, which means that every $a \in E$ is the limit of some sequence (x_n) of points $x_n \in X$. Convex sets will again play a crucial role.

First, we state the following easy "parallelogram inequality", whose proof is left as an exercise.

Proposition 28.2. *If E is a Hermitian space, for any two vectors $u, v \in E$, we have*

$$\|u + v\|^2 + \|u - v\|^2 = 2(\|u\|^2 + \|v\|^2).$$

From the above, we get the following proposition:

Proposition 28.3. *If E is a Hermitian space, given any $d, \delta \in \mathbb{R}$ such that $0 \leq \delta < d$, let*

$$B = \{u \in E \mid \|u\| < d\} \quad \textit{and} \quad C = \{u \in E \mid \|u\| \leq d + \delta\}.$$

For any convex set such A that $A \subseteq C - B$, we have

$$\|v - u\| \leq \sqrt{12d\delta},$$

for all $u, v \in A$ (see Figure 28.1).

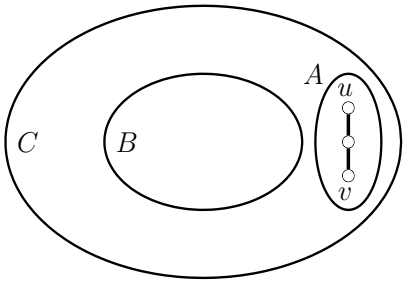


Figure 28.1: Inequality of Proposition 28.3

Proof. Since A is convex, $\frac{1}{2}(u+v) \in A$ if $u, v \in A$, and thus, $\|\frac{1}{2}(u+v)\| \geq d$. From the parallelogram inequality written in the form

$$\left\|\frac{1}{2}(u+v)\right\|^2 + \left\|\frac{1}{2}(u-v)\right\|^2 = \frac{1}{2}\left(\|u\|^2 + \|v\|^2\right),$$

since $\delta < d$, we get

$$\left\|\frac{1}{2}(u-v)\right\|^2 = \frac{1}{2}\left(\|u\|^2 + \|v\|^2\right) - \left\|\frac{1}{2}(u+v)\right\|^2 \leq (d+\delta)^2 - d^2 = 2d\delta + \delta^2 \leq 3d\delta,$$

from which

$$\|v-u\| \leq \sqrt{12d\delta}.$$

□

If X is a nonempty subset of a metric space (E, d), for any $a \in E$, recall that we define the *distance* $d(a, X)$ *of* a *to* X as

$$d(a, X) = \inf_{b \in X} d(a, b).$$

Also, the *diameter* $\delta(X)$ *of* X is defined by

$$\delta(X) = \sup\{d(a, b) \mid a, b \in X\}.$$

It is possible that $\delta(X) = \infty$. We leave the following standard two facts as an exercise (see Dixmier [35]):

Proposition 28.4. *Let* E *be a metric space.*

(1) For every subset $X \subseteq E$, $\delta(X) = \delta(\overline{X})$.

(2) If E *is a complete metric space, for every sequence* (F_n) *of closed nonempty subsets of* E *such that* $F_{n+1} \subseteq F_n$, *if* $\lim_{n\to\infty} \delta(F_n) = 0$, *then* $\bigcap_{n=1}^{\infty} F_n$ *consists of a single point.*

We are now ready to prove the crucial projection lemma.

Proposition 28.5. *(Projection lemma) Let E be a Hilbert space.*

(1) For any nonempty convex and closed subset $X \subseteq E$, for any $u \in E$, there is a unique vector $p_X(u) \in X$ such that

$$\|u - p_X(u)\| = \inf_{v \in X} \|u - v\| = d(u, X).$$

See Figure 28.2.

(2) The vector $p_X(u)$ is the unique vector $w \in E$ satisfying the following property (see Figure 28.3):

$$w \in X \quad \textit{and} \quad \Re \langle u - w, z - w \rangle \leq 0 \quad \textit{for all } z \in X. \tag{$*$}$$

(3) If X is a nonempty closed subspace of E then the vector $p_X(u)$ is the unique vector $w \in E$ satisfying the following property:

$$w \in X \quad \textit{and} \quad \langle u - w, z \rangle = 0 \quad \textit{for all } z \in X. \tag{$**$}$$

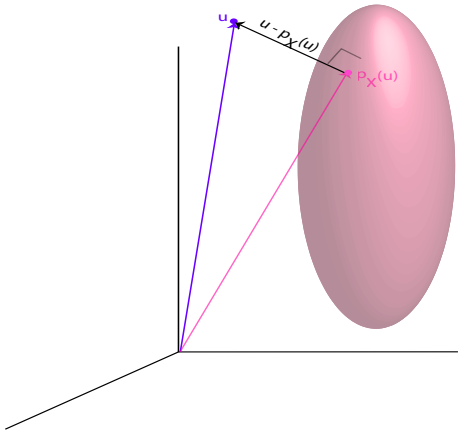


Figure 28.2: Let X be the solid pink ellipsoid. The projection of the purple point u onto X is the magenta point $p_X(u)$.

Proof. (1) Let $d = \inf_{v \in X} \|u - v\| = d(u, X)$. We define a sequence X_n of subsets of X as follows: for every $n \geq 1$,

$$X_n = \left\{ v \in X \mid \|u - v\| \leq d + \frac{1}{n} \right\}.$$

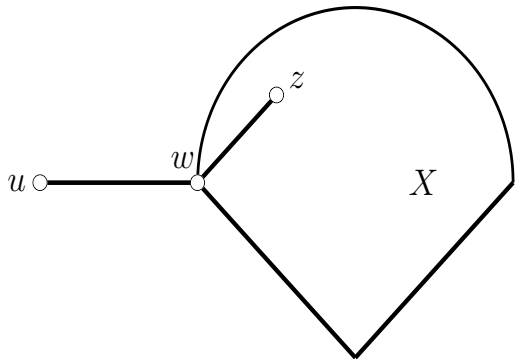


Figure 28.3: Inequality of Proposition 28.5

It is immediately verified that each X_n is nonempty (by definition of d), convex, and that $X_{n+1} \subseteq X_n$. Also, by Proposition 28.3, we have

$$\sup\{\|w - v\| \mid v, w \in X_n\} \leq \sqrt{12d/n},$$

and thus, $\bigcap_{n\geq 1} X_n$ contains at most one point. We will prove that $\bigcap_{n\geq 1} X_n$ contains exactly one point, namely, $p_X(u)$. For this, define a sequence $(w_n)_{n\geq 1}$ by picking some $w_n \in X_n$ for every $n \geq 1$. We claim that $(w_n)_{n\geq 1}$ is a Cauchy sequence. Given any $\epsilon > 0$, if we pick N such that

$$N > \frac{12d}{\epsilon^2},$$

since $(X_n)_{n\geq 1}$ is a monotonic decreasing sequence, which means that $X_{n+1} \subseteq X_n$ for all $n \geq 1$, for all $m, n \geq N$, we have

$$\|w_m - w_n\| \leq \sqrt{12d/N} < \epsilon,$$

as desired. Since E is complete, the sequence $(w_n)_{n\geq 1}$ has a limit w, and since $w_n \in X$ and X is closed, we must have $w \in X$. Also observe that

$$\|u - w\| \leq \|u - w_n\| + \|w_n - w\|,$$

and since w is the limit of $(w_n)_{n\geq 1}$ and

$$\|u - w_n\| \leq d + \frac{1}{n},$$

given any $\epsilon > 0$, there is some n large enough so that

$$\frac{1}{n} < \frac{\epsilon}{2} \quad \text{and} \quad \|w_n - w\| \leq \frac{\epsilon}{2},$$

and thus

$$\|u - w\| \leq d + \epsilon.$$

Since the above holds for every $\epsilon > 0$, we have $\|u - w\| = d$. Thus, $w \in X_n$ for all $n \geq 1$, which proves that $\bigcap_{n \geq 1} X_n = \{w\}$. Now, any $z \in X$ such that $\|u - z\| = d(u, X) = d$ also belongs to every X_n, and thus $z = w$, proving the uniqueness of w, which we denote as $p_X(u)$. See Figure 28.4.

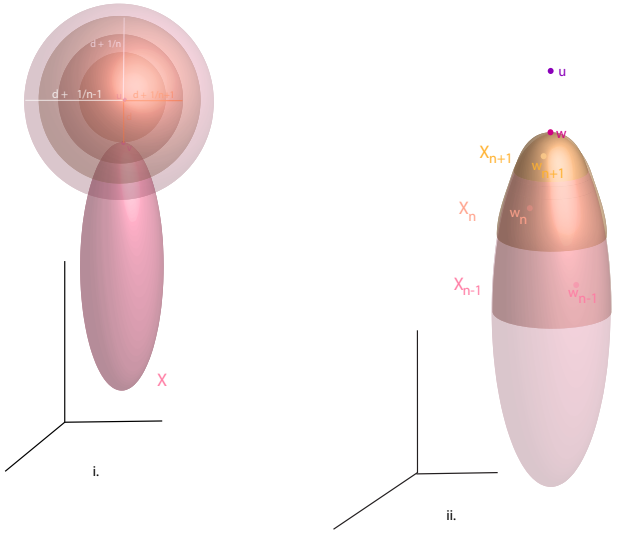


Figure 28.4: Let X be the solid pink ellipsoid with $p_X(u) = w$ at its apex. Each X_n is the intersection of X and a solid sphere centered at u with radius $d + 1/n$. These intersections are the colored "caps" of Figure ii. The Cauchy sequence $(w_n)_{n \geq 1}$ is obtained by selecting a point in each colored X_n.

(2) Let $z \in X$. Since X is convex, $w = (1 - \lambda)p_X(u) + \lambda z \in X$ for every λ, $0 \leq \lambda \leq 1$. Then, we have

$$\|u - w\| \geq \|u - p_X(u)\|$$

for all λ, $0 \leq \lambda \leq 1$, and since

$$\begin{aligned} \|u - w\|^2 &= \|u - p_X(u) - \lambda(z - p_X(u))\|^2 \\ &= \|u - p_X(u)\|^2 + \lambda^2 \|z - p_X(u)\|^2 - 2\lambda \Re \langle u - p_X(u), z - p_X(u) \rangle , \end{aligned}$$

for all λ, $0 < \lambda \leq 1$, we get

$$\Re \langle u - p_X(u), z - p_X(u) \rangle = \frac{1}{2\lambda} \left(\|u - p_X(u)\|^2 - \|u - w\|^2 \right) + \frac{\lambda}{2} \|z - p_X(u)\|^2,$$

and since this holds for every λ, $0 < \lambda \leq 1$ and

$$\|u - w\| \geq \|u - p_X(u)\|,$$

we have

$$\Re \langle u - p_X(u), z - p_X(u) \rangle \leq 0.$$

Conversely, assume that $w \in X$ satisfies the condition

$$\Re \langle u - w, z - w \rangle \leq 0$$

for all $z \in X$. For all $z \in X$, we have

$$\|u - z\|^2 = \|u - w\|^2 + \|z - w\|^2 - 2\Re \langle u - w, z - w \rangle \geq \|u - w\|^2,$$

which implies that $\|u - w\| = d(u, X) = d$, and from (1), that $w = p_X(u)$.

(3) If X is a subspace of E and $w \in X$, when z ranges over X the vector $z - w$ also ranges over the whole of X so Condition $(*)$ is equivalent to

$$w \in X \quad \text{and} \quad \Re\langle u - w, z\rangle \leq 0 \quad \text{for all } z \in X. \tag{$*_1$}$$

Since X is a subspace, if $z \in X$ then $-z \in X$, which implies that $(*_1)$ is equivalent to

$$w \in X \quad \text{and} \quad \Re\langle u - w, z\rangle = 0 \quad \text{for all } z \in X. \tag{$*_2$}$$

Finally, since X is a subspace if $z \in X$ then $iz \in X$, and this implies that

$$0 = \Re\langle u - w, iz\rangle = -i\Im\langle u - w, z\rangle,$$

so $\Im\langle u - w, z\rangle = 0$, but since we also have $\Re\langle u - w, z\rangle = 0$, we see that $(*_2)$ is equivalent to

$$w \in X \quad \text{and} \quad \langle u - w, z\rangle = 0 \quad \text{for all } z \in X, \tag{$**$}$$

as claimed. □

The vector $p_X(u)$ is called the *projection of u onto X*, and the map $p_X \colon E \to X$ is called the *projection of E onto X*. In the case of a real Hilbert space, there is an intuitive geometric interpretation of the condition

$$\langle u - p_X(u), z - p_X(u)\rangle \leq 0$$

for all $z \in X$. If we restate the condition as

$$\langle u - p_X(u), p_X(u) - z\rangle \geq 0$$

for all $z \in X$, this says that the absolute value of the measure of the angle between the vectors $u - p_X(u)$ and $p_X(u) - z$ is at most $\pi/2$. See Figure 28.5. This makes sense, since X is convex, and points in X must be on the side opposite to the "tangent space" to X at $p_X(u)$, which is orthogonal to $u - p_X(u)$. Of course, this is only an intuitive description, since the notion of tangent space has not been defined!

If X is a closed subspace of E, then Condition $(**)$ says that the vector $u - p_X(u)$ is orthogonal to X, in the sense that $u - p_X(u)$ is orthogonal to every vector $z \in X$.

The map $p_X \colon E \to X$ is continuous, as shown below.

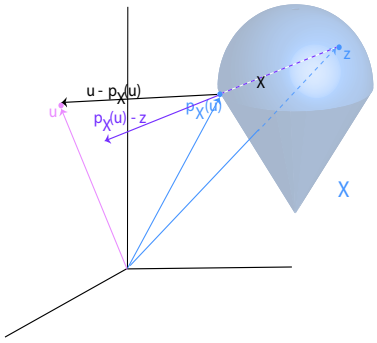


Figure 28.5: Let X be the solid blue ice cream cone. The acute angle between the black vector $u - p_X(u)$ and the purple vector $p_X(u) - z$ is less than $\pi/2$.

Proposition 28.6. *Let E be a Hilbert space. For any nonempty convex and closed subset $X \subseteq E$, the map $p_X \colon E \to X$ is continuous. In fact, p_X satisfies the Lipschitz condition*

$$\|p_X(v) - p_X(u)\| \leq \|v - u\| \quad \textit{for all } u, v \in E.$$

Proof. For any two vectors $u, v \in E$, let $x = p_X(u) - u$, $y = p_X(v) - p_X(u)$, and $z = v - p_X(v)$. Clearly, (as illustrated in Figure 28.6),

$$v - u = x + y + z,$$

and from Proposition 28.5 (2), we also have

$$\Re \langle x, y \rangle \geq 0 \quad \text{and} \quad \Re \langle z, y \rangle \geq 0,$$

from which we get

$$\begin{aligned} \|v - u\|^2 &= \|x + y + z\|^2 = \|x + z + y\|^2 \\ &= \|x + z\|^2 + \|y\|^2 + 2\Re \langle x, y \rangle + 2\Re \langle z, y \rangle \\ &\geq \|y\|^2 = \|p_X(v) - p_X(u)\|^2. \end{aligned}$$

However, $\|p_X(v) - p_X(u)\| \leq \|v - u\|$ obviously implies that p_X is continuous. □

We can now prove the following important proposition.

Proposition 28.7. *Let E be a Hilbert space.*

(1) For any closed subspace $V \subseteq E$, we have $E = V \oplus V^{\perp}$, and the map $p_V \colon E \to V$ is linear and continuous.

(2) For any $u \in E$, the projection $p_V(u)$ is the unique vector $w \in E$ such that

$$w \in V \quad \textit{and} \quad \langle u - w, z \rangle = 0 \quad \textit{for all } z \in V.$$

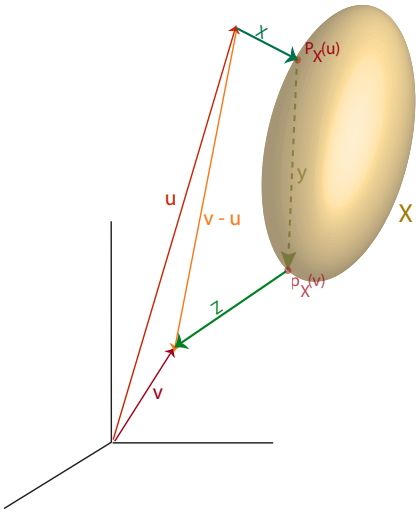


Figure 28.6: Let X be the solid gold ellipsoid. The vector $v-u$ is the sum of the three green vectors, each of which is determined by the appropriate projections.

Proof. (1) First, we prove that $u - p_V(u) \in V^\perp$ for all $u \in E$. For any $v \in V$, since V is a subspace, $z = p_V(u) + \lambda v \in V$ for all $\lambda \in \mathbb{C}$, and since V is convex and nonempty (since it is a subspace), and closed by hypothesis, by Proposition 28.5 (2), we have

$$\Re(\overline{\lambda}\,\langle u - p_V(u), v\rangle) = \Re(\langle u - p_V(u), \lambda v\rangle = \Re\,\langle u - p_V(u), z - p_V(u)\rangle \leq 0$$

for all $\lambda \in \mathbb{C}$. In particular, the above holds for $\lambda = \langle u - p_V(u), v\rangle$, which yields

$$|\,\langle u - p_V(u), v\rangle\,| \leq 0,$$

and thus, $\langle u - p_V(u), v\rangle = 0$. See Figure 28.7. As a consequence, $u - p_V(u) \in V^\perp$ for all $u \in E$. Since $u = p_V(u) + u - p_V(u)$ for every $u \in E$, we have $E = V + V^\perp$. On the other hand, since $\langle -, -\rangle$ is positive definite, $V \cap V^\perp = \{0\}$, and thus $E = V \oplus V^\perp$.

We already proved in Proposition 28.6 that $p_V \colon E \to V$ is continuous. Also, since

$$p_V(\lambda u + \mu v) - (\lambda p_V(u) + \mu p_V(v)) = p_V(\lambda u + \mu v) - (\lambda u + \mu v) + \lambda(u - p_V(u)) + \mu(v - p_V(v)),$$

for all $u, v \in E$, and since the left-hand side term belongs to V, and from what we just showed, the right-hand side term belongs to $V^\perp$, we have

$$p_V(\lambda u + \mu v) - (\lambda p_V(u) + \mu p_V(v)) = 0,$$

showing that p_V is linear.

(2) This is basically obvious from (1). We proved in (1) that $u - p_V(u) \in V^\perp$, which is exactly the condition

$$\langle u - p_V(u), z\rangle = 0$$

for all $z \in V$. Conversely, if $w \in V$ satisfies the condition

$$\langle u - w, z\rangle = 0$$

for all $z \in V$, since $w \in V$, every vector $z \in V$ is of the form $y - w$, with $y = z + w \in V$, and thus, we have

$$\langle u - w, y - w\rangle = 0$$

for all $y \in V$, which implies the condition of Proposition 28.5 (2):

$$\Re\langle u - w, y - w\rangle \leq 0$$

for all $y \in V$. By Proposition 28.5, $w = p_V(u)$ is the projection of u onto V. □

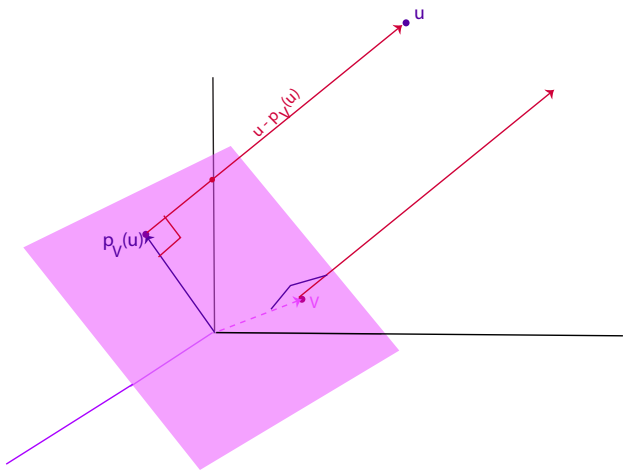


Figure 28.7: Let V be the pink plane. The vector $u - p_V(u)$ is perpendicular to any $v \in V$.

Remark: If $p_V : E \to V$ is linear, then V is a subspace of E. It follows that if V is a closed convex subset of E, then $p_V : E \to V$ is linear iff V is a subspace of E.

Let us illustrate the power of Proposition 28.7 on the following "least squares" problem. Given a real $m \times n$-matrix A and some vector $b \in \mathbb{R}^m$, we would like to solve the linear system

$$Ax = b$$

in the least-squares sense, which means that we would like to find some solution $x \in \mathbb{R}^n$ that minimizes the Euclidean norm $\|Ax - b\|$ of the error $Ax - b$. It is actually not clear that the problem has a solution, but it does! The problem can be restated as follows: Is there some $x \in \mathbb{R}^n$ such that

$$\|Ax - b\| = \inf_{y \in \mathbb{R}^n} \|Ay - b\|,$$

or equivalently, is there some $z \in \mathrm{Im}\,(A)$ such that

$$\|z - b\| = d(b, \mathrm{Im}\,(A)),$$

where $\mathrm{Im}\,(A) = \{Ay \in \mathbb{R}^m \mid y \in \mathbb{R}^n\}$, the image of the linear map induced by A. Since $\mathrm{Im}\,(A)$ is a closed subspace of $\mathbb{R}^m$, because we are in finite dimension, Proposition 28.7 tells us that there is a unique $z \in \mathrm{Im}\,(A)$ such that

$$\|z - b\| = \inf_{y \in \mathbb{R}^n} \|Ay - b\|,$$

and thus, the problem always has a solution since $z \in \mathrm{Im}\,(A)$, and since there is at least some $x \in \mathbb{R}^n$ such that $Ax = z$ (by definition of $\mathrm{Im}\,(A)$). Note that such an x is not necessarily unique. Furthermore, Proposition 28.7 also tells us that $z \in \mathrm{Im}\,(A)$ is the solution of the equation

$$\langle z - b, w\rangle = 0 \quad \text{for all } w \in \mathrm{Im}\,(A),$$

or equivalently, that $x \in \mathbb{R}^n$ is the solution of

$$\langle Ax - b, Ay\rangle = 0 \quad \text{for all } y \in \mathbb{R}^n,$$

which is equivalent to

$$\langle A^\top (Ax - b), y\rangle = 0 \quad \text{for all } y \in \mathbb{R}^n,$$

and thus, since the inner product is positive definite, to $A^\top(Ax - b) = 0$, i.e.,

$$A^\top Ax = A^\top b.$$

Therefore, the solutions of the original least-squares problem are precisely the solutions of the the so-called *normal equations*

$$A^\top Ax = A^\top b,$$

discovered by Gauss and Legendre around 1800. We also proved that the normal equations always have a solution.

Computationally, it is best not to solve the normal equations directly, and instead, to use methods such as the QR-decomposition (applied to A) or the SVD-decomposition (in the form of the pseudo-inverse). We will come back to this point later on.

As another corollary of Proposition 28.7, for any continuous nonnull linear map $h\colon E \to \mathbb{C}$, the null space

$$H = \mathrm{Ker}\, h = \{u \in E \mid h(u) = 0\} = h^{-1}(0)$$

is a closed hyperplane H, and thus, $H^\perp$ is a subspace of dimension one such that $E = H \oplus H^\perp$. This suggests defining the dual space of E as the set of all continuous maps $h\colon E \to \mathbb{C}$.

Remark: If $h\colon E \to \mathbb{C}$ is a linear map which is **not** continuous, then it can be shown that the hyperplane $H = \mathrm{Ker}\, h$ is dense in E! Thus, $H^\perp$ is reduced to the trivial subspace

$\{0\}$. This goes against our intuition of what a hyperplane in $\mathbb{R}^n$ (or $\mathbb{C}^n$) is, and warns us not to trust our "physical" intuition too much when dealing with infinite dimensions. As a consequence, the map $\flat\colon E \to E^*$ introduced in Section 11.2 (see just after Definition 28.2 below) is not surjective, since the linear forms of the form $u \mapsto \langle u, v\rangle$ (for some fixed vector $v \in E$) are continuous (the inner product is continuous).

We now show that by redefining the dual space of a Hilbert space as the set of continuous linear forms on E, we recover Theorem 11.5.

Definition 28.2. Given a Hilbert space E, we define the *dual space E' of E* as the vector space of all continuous linear forms $h\colon E \to \mathbb{C}$. Maps in E' are also called *bounded linear operators, bounded linear functionals, or simply, operators or functionals*.

As in Section 11.2, for all $u, v \in E$, we define the maps $\varphi_u^l\colon E \to \mathbb{C}$ and $\varphi_v^r\colon E \to \mathbb{C}$ such that

$$\varphi_u^l(v) = \overline{\langle u, v\rangle},$$

and

$$\varphi_v^r(u) = \langle u, v\rangle\,.$$

In fact, $\varphi_u^l = \varphi_u^r$, and because the inner product $\langle -, -\rangle$ is continuous, it is obvious that φ_v^r is continuous and linear, so that $\varphi_v^r \in E'$. To simplify notation, we write φ_v instead of φ_v^r.

Theorem 11.5 is generalized to Hilbert spaces as follows.

Proposition 28.8. *(Riesz representation theorem) Let E be a Hilbert space. Then, the map $\flat\colon E \to E'$ defined such that*

$$\flat(v) = \varphi_v,$$

is semilinear, continuous, and bijective. Furthermore, for any continuous linear map $\psi \in E'$, if $u \in E$ is the unique vector such that

$$\psi(v) = \langle v, u\rangle \quad \textit{for all } v \in E,$$

then we have $\|\psi\| = \|u\|$, where

$$\|\psi\| = \sup\left\{ \frac{|\psi(v)|}{\|v\|} \;\middle|\; v \in E,\ v \neq 0 \right\}.$$

Proof. The proof is basically identical to the proof of Theorem 11.5, except that a different argument is required for the surjectivity of $\flat\colon E \to E'$, since E may not be finite dimensional. For any nonnull linear operator $h \in E'$, the hyperplane $H = \operatorname{Ker} h = h^{-1}(0)$ is a closed subspace of E, and by Proposition 28.7, $H^\perp$ is a subspace of dimension one such that $E = H \oplus H^\perp$. Then, picking any nonnull vector $w \in H^\perp$, observe that H is also the kernel of the linear operator φ_w, with

$$\varphi_w(u) = \langle u, w\rangle\,,$$

and thus, since any two nonzero linear forms defining the same hyperplane must be proportional, there is some nonzero scalar $\lambda \in \mathbb{C}$ such that $h = \lambda\varphi_w$. But then, $h = \varphi_{\overline{\lambda}w}$, proving that $\flat\colon E \to E'$ is surjective.

By the Cauchy–Schwarz inequality we have

$$|\psi(v)| = |\langle v, u\rangle| \leq \|v\|\,\|u\|\,,$$

so by definition of $\|\psi\|$ we get

$$\|\psi\| \leq \|u\|\,.$$

Obviously $\psi = 0$ iff $u = 0$ so assume $u \neq 0$. We have

$$\|u\|^2 = \langle u, u\rangle = \psi(u) \leq \|\psi\|\,\|u\|\,,$$

which yields $\|u\| \leq \|\psi\|$, and therefore $\|\psi\| = \|u\|$, as claimed. □

Proposition 28.8 is known as the *Riesz representation theorem, or "Little Riesz Theorem."* It shows that the inner product on a Hilbert space induces a natural semilinear isomorphism between E and its dual E' (equivalently, a linear isomorphism between $\overline{E}$ and E'). This isomorphism is an isometry (it is preserves the norm).

Remark: Many books on quantum mechanics use the so-called Dirac notation to denote objects in the Hilbert space E and operators in its dual space E'. In the Dirac notation, an element of E is denoted as $|x\rangle$, and an element of E' is denoted as $\langle t|$. The scalar product is denoted as $\langle t| \cdot |x\rangle$. This uses the isomorphism between E and E', except that the inner product is assumed to be semi-linear on the left, rather than on the right.

Proposition 28.8 allows us to define the adjoint of a linear map, as in the Hermitian case (see Proposition 11.6). Actually, we can prove a slightly more general result which is used in optimization theory.

If $\varphi\colon E \times E \to \mathbb{C}$ is a sesquilinear map on a normed vector space $(E, \|\ \|)$, then Proposition 18.17 is immediately adapted to prove that φ is continuous iff there is some constant $k \geq 0$ such that

$$|\varphi(u, v)| \leq k\,\|u\|\,\|v\| \quad \text{for all } u, v \in E.$$

Thus we define $\|\varphi\|$ as in Definition 18.16 by

$$\|\varphi\| = \sup\{|\varphi(x, y)| \mid \|x\| \leq 1,\ \|y\| \leq 1,\ x, y \in E\}\,.$$

Proposition 28.9. *Given a Hilbert space E, for every continuous sesquilinear map $\varphi\colon E \times E \to \mathbb{C}$, there is a unique continuous linear map $f_\varphi\colon E \to E$, such that*

$$\varphi(u, v) = \langle u, f_\varphi(v)\rangle \quad \textit{for all } u, v \in E.$$

We also have $\|f_\varphi\| = \|\varphi\|$. If φ is Hermitian, then f_φ is self-adjoint, that is

$$\langle u, f_\varphi(v)\rangle = \langle f_\varphi(u), v\rangle \quad \textit{for all } u, v \in E.$$

Proof. The proof is adapted from Rudin [83] (Theorem 12.8). To define the function f_φ we proceed as follows. For any fixed $v \in E$ define the linear map φ_v by

$$\varphi_v(u) = \varphi(u, v) \quad \text{for all } u \in E.$$

Since φ is continuous φ_v is continuous so by Proposition 28.8, there is a unique vector in E that we denote $f_\varphi(v)$ such that

$$\varphi_v(u) = \langle u, f_\varphi(v)\rangle \quad \text{for all } u \in E,$$

and $\|f_\varphi(v)\| = \|\varphi_v\|$. Let us check that the map $v \mapsto f_\varphi(v)$ is linear.

We have

$$\begin{aligned}
\varphi(u, v_1 + v_2) &= \varphi(u, v_1) + \varphi(u, v_2) && \varphi \text{ is additive}\\
&= \langle u, f_\varphi(v_1)\rangle + \langle u, f_\varphi(v_2)\rangle && \text{by definition of } f_\varphi\\
&= \langle u, f_\varphi(v_1) + f_\varphi(v_2)\rangle && \langle -, -\rangle \text{ is additive}
\end{aligned}$$

for all $u \in E$, and since $f_\varphi(v_1+v_2)$ is the unique vector such that $\varphi(u, v_1+v_2) = \langle u, f_\varphi(v_1+v_2)\rangle$ for all $u \in E$, we must have

$$f_\varphi(v_1 + v_2) = f_\varphi(v_1) + f_\varphi(v_2).$$

For any $\lambda \in \mathbb{C}$ we have

$$\begin{aligned}
\varphi(u, \lambda v) &= \overline{\lambda}\varphi(u, v) && \varphi \text{ is sesquilinear}\\
&= \overline{\lambda}\langle u, f_\varphi(v)\rangle && \text{by definition of } f_\varphi\\
&= \langle u, \lambda f_\varphi(v)\rangle && \langle -, -\rangle \text{ is sesquilinear}
\end{aligned}$$

for all $u \in E$, and since $f_\varphi(\lambda v)$ is the unique vector such that $\varphi(u, \lambda v) = \langle u, f_\varphi(\lambda v)\rangle$ for all $u \in E$, we must have

$$f_\varphi(\lambda v) = \lambda f_\varphi(v).$$

Therefore f_φ is linear.

Then by definition of $\|\varphi\|$ we have

$$|\varphi_v(u)| = |\varphi(u, v)| \leq \|\varphi\|\, \|u\|\, \|v\|,$$

which shows that $\|\varphi_v\| \leq \|\varphi\|\, \|v\|$. Since $\|f_\varphi(v)\| = \|\varphi_v\|$, we have

$$\|f_\varphi(v)\| \leq \|\varphi\|\, \|v\|,$$

which shows that f_φ is continuous and that $\|f_\varphi\| \leq \|\varphi\|$. But by the Cauchy–Schwarz inequality we also have

$$|\varphi(u, v)| = |\langle u, f_\varphi(v)\rangle| \leq \|u\|\, \|f_\varphi(v)\| \leq \|u\|\, \|f_\varphi\|\, \|v\|,$$

so $\|\varphi\| \leq \|f_\varphi\|$, and thus

$$\|f_\varphi\| = \|\varphi\|.$$

If φ is Hermitian, $\varphi(v,u) = \overline{\varphi(u,v)}$, so

$$\langle f_\varphi(u), v\rangle = \overline{\langle v, f_\varphi(u)\rangle} = \overline{\varphi(v,u)} = \varphi(u,v) = \langle u, f_\varphi(v)\rangle,$$

which shows that f_φ is self-adjoint. □

Proposition 28.10. *Given a Hilbert space E, for every continuous linear map $f\colon E \to E$, there is a unique continuous linear map $f^*\colon E \to E$, such that*

$$\langle f(u), v\rangle = \langle u, f^*(v)\rangle \quad \textit{for all } u, v \in E,$$

and we have $\|f^*\| = \|f\|$. The map f^* is called the adjoint of f.

Proof. The proof is adapted from Rudin [83] (Section 12.9). By the Cauchy–Schwarz inequality

$$|\langle x, y\rangle| \leq \|x\|\,\|y\|$$

we see that the sesquilinear map $(x,y) \mapsto \langle x, y\rangle$ on $E \times E$ is continuous. Let $\varphi\colon E \times E \to \mathbb{C}$ be the sesquilinear map given by

$$\varphi(u,v) = \langle f(u), v\rangle \quad \text{for all } u, v \in E.$$

Since f is continuous and the inner product $\langle -, -\rangle$ is continuous, this is a continuous map. By Proposition 28.9 there is a unique linear map $f^*\colon E \to E$ such that

$$\langle f(u), v\rangle = \varphi(u,v) = \langle u, f^*(v)\rangle \quad \text{for all } u, v \in E,$$

with $\|f^*\| = \|\varphi\|$.

We can also prove that $\|\varphi\| = \|f\|$. First, by definition of $\|\varphi\|$ we have

$$\begin{aligned}
\|\varphi\| &= \sup\{|\varphi(x,y)| \mid \|x\| \leq 1,\ \|y\| \leq 1\} \\
&= \sup\{|\langle f(x), y\rangle| \mid \|x\| \leq 1,\ \|y\| \leq 1\} \\
&\leq \sup\{\|(f(x)\|\,\|y\| \mid \|x\| \leq 1, \|y\| \leq 1\} \\
&\leq \sup\{\|(f(x)\| \mid \|x\| \leq 1\} \\
&= \|f\|.
\end{aligned}$$

In the other direction we have

$$\|f(x)\|^2 = \langle f(x), f(x)\rangle = \varphi(x, f(x)) \leq \|\varphi\|\,\|x\|\,\|f(x)\|,$$

and if $f(x) \neq 0$ we get $\|f(x)\| \leq \|\varphi\|\,\|x\|$. This inequality holds trivially if $f(x) = 0$, so we conclude that $\|f\| \leq \|\varphi\|$. Therefore we have

$$\|\varphi\| = \|f\|,$$

as claimed, and consequently $\|f^*\| = \|\varphi\| = \|f\|$. □

It is easy to show that the adjoint satisfies the following properties:

$$(f+g)^* = f^* + g^*$$
$$(\lambda f)^* = \overline{\lambda} f^*$$
$$(f \circ g)^* = g^* \circ f^*$$
$$f^{**} = f.$$

One can also show that $\|f^* \circ f\| = \|f\|^2$ (see Rudin [83], Section 12.9).

As in the Hermitian case, given two Hilbert spaces E and F, the above results can be adapted to show that for any linear map $f\colon E \to F$, there is a unique linear map $f^*\colon F \to E$ such that

$$\langle f(u), v\rangle_2 = \langle u, f^*(v)\rangle_1$$

for all $u \in E$ and all $v \in F$. The linear map f^* is also called the adjoint of f.

28.2 Farkas–Minkowski Lemma in Hilbert Spaces

In this section, $(V, \langle -, -\rangle)$ is assumed to be a real Hilbert space. The projection lemma can be used to show an interesting version of the Farkas–Minkowski lemma in a Hilbert space.

Given a finite sequence of vectors $(a_1, \ldots, a_m)$ with $a_i \in V$, let C be the polyhedral cone

$$C = \text{cone}(a_1, \ldots, a_m) = \left\{ \sum_{i=1}^{m} \lambda_i a_i \mid \lambda_i \geq 0,\ i = 1, \ldots, m \right\}.$$

For any vector $b \in V$, the Farkas–Minkowski lemma gives a criterion for checking whether $b \in C$.

In Proposition 24.2 we proved that every polyhedral cone $\text{cone}(a_1, \ldots, a_m)$ with $a_i \in \mathbb{R}^n$ is closed. Close examination of the proof shows that it goes through if $a_i \in V$ where V is any vector space possibly of infinite dimension, because the important fact is that the number m of these vectors is finite, not their dimension.

Theorem 28.11. *(Farkas–Minkowski Lemma in Hilbert Spaces) Let $(V, \langle -, -\rangle)$ be a real Hilbert space. For any finite sequence of vectors $(a_1, \ldots, a_m)$ with $a_i \in V$, if C is the polyhedral cone $C = \text{cone}(a_1, \ldots, a_m)$, for any vector $b \in V$, we have $b \notin C$ iff there is a vector $u \in V$ such that*

$$\langle a_i, u\rangle \geq 0 \quad i = 1, \ldots, m, \quad \textit{and} \quad \langle b, u\rangle < 0.$$

Equivalently, $b \in C$ iff for all $u \in V$,

$$\textit{if } \langle a_i, u\rangle \geq 0 \quad i = 1, \ldots, m, \quad \textit{then} \quad \langle b, u\rangle \geq 0.$$

Proof. We follow Ciarlet [30] (Chapter 9, Theorem 9.1.1). We already established in Proposition 24.2 that the polyhedral cone $C = \text{cone}(a_1, \ldots, a_m)$ is closed. Next we claim the following:

Claim: If C is a nonempty, closed, convex subset of a Hilbert space V, and $b \in V$ is any vector such that $b \notin C$, then there exist some $u \in V$ and infinitely many scalars $\alpha \in \mathbb{R}$ such that

$$\begin{aligned} \langle v, u\rangle &> \alpha \quad \text{for every } v \in C \\ \langle b, u\rangle &< \alpha. \end{aligned}$$

We use the projection lemma (Proposition 28.5) which says that since $b \notin C$ there is some unique $c = p_C(b) \in C$ such that

$$\begin{aligned} \|b - c\| &= \inf_{v \in C} \|b - v\| > 0 \\ \langle b - c, v - c\rangle &\leq 0 \quad \text{for all } v \in C, \end{aligned}$$

or equivalently

$$\begin{aligned} \|b - c\| &= \inf_{v \in C} \|b - v\| > 0 \\ \langle v - c, c - b\rangle &\geq 0 \quad \text{for all } v \in C. \end{aligned}$$

As a consequence we have

$$\langle v, c - b\rangle \geq \langle c, c - b\rangle > \langle b, c - b\rangle,$$

and if we pick $u = c - b$ and any α such that

$$\langle c, c - b\rangle > \alpha > \langle b, c - b\rangle,$$

the claim is satisfied.

We now prove the Farkas–Minkowski Lemma. Asume that $b \notin C$. Since C is nonempty, convex, and closed, by the Claim there is some $u \in V$ and some $\alpha \in \mathbb{R}$ such that

$$\begin{aligned} \langle v, u\rangle &> \alpha \quad \text{for every } v \in C \\ \langle b, u\rangle &< \alpha. \end{aligned}$$

But C is a polyhedral cone containing 0 so we must have $\alpha < 0$. Then for every $v \in C$, since C a polyhedral cone if $v \in C$ then $\lambda v \in C$ for all $\lambda > 0$, so by the above

$$\langle v, u\rangle > \frac{\alpha}{\lambda} \quad \text{for every } \lambda > 0,$$

which implies that

$$\langle v, u\rangle \geq 0.$$

Since $a_i \in C$ for $i = 1, \ldots, m$, we proved that

$$\langle a_i, u\rangle \geq 0 \quad i = 1, \ldots, m \quad \text{and} \quad \langle b, u\rangle < \alpha < 0,$$

which proves Farkas Lemma. □

Observe that the claim established during the proof of Theorem 28.11 shows that the affine hyperplane $H_{u,\alpha}$ of equation $\langle v, u\rangle = \alpha$ for all $v \in V$ separates strictly C and $\{b\}$.

Chapter 29

General Results of Optimization Theory

29.1 Existence of Solutions of an Optimization Problem

The main goal of *optimization theory* is to construct *algorithms* to find solutions (often approximate) of problems of the form

$$\begin{aligned} &\text{find} \quad u \\ &\text{such that} \quad u \in U \ \text{ and } \ J(u) = \inf_{v \in U} J(v), \end{aligned}$$

where U is a given subset of a vector space V (possibly infinite dimensional) and $J\colon \Omega \to \mathbb{R}$ is a function defined on some open subset Ω of V such that $U \subseteq \Omega$.

To be very clear, $\inf_{v \in U} J(v)$ denotes the *greatest lower bound* of the set of real number $\{J(u) \mid u \in U\}$. To make sure that we are on firm grounds let us review the notions of greatest lower bound and least upper bound of a set of real numbers.

Let X be any nonempty subset of $\mathbb{R}$. The set $LB(X)$ of *lower bounds* of X is defined as

$$LB(X) = \{b \in \mathbb{R} \mid b \leq x \text{ for all } x \in X\}.$$

If the set X is not bounded below, which means that for every $r \in \mathbb{R}$ there is some $x \in X$ such that $x < r$, then $LB(X)$ is empty. Otherwise, if $LB(X)$ is nonempty, since it is bounded above by every element of X, by a fundamental property of the real numbers, the set $LB(X)$ has a greatest element denoted $\inf X$. The real number $\inf X$ is thus the *greatest lower bound* of X. In general, $\inf X$ does not belong to X, but if it does, then it is the least element of X.

If $LB(X) = \emptyset$, then X is *unbounded below* and $\inf X$ is undefined. In this case (with an abuse of notation), we write

$$\inf X = -\infty.$$

By convention, when $X = \emptyset$ we set

$$\inf \emptyset = +\infty.$$

Similarly the set $UB(X)$ of *upper bounds* of X is given by

$$UB(X) = \{u \in \mathbb{R} \mid x \leq u \text{ for all } x \in X\}.$$

If X is not bounded above, then $UB(X) = \emptyset$. Otherwise, if $UB(X) \neq \emptyset$, then it has least element denoted $\sup X$. Thus $\sup X$ is the *least upper bound* of X. If $\sup X \in X$, then it is the greatest element of X. If $UB(X) = \emptyset$, then

$$\sup X = +\infty.$$

By convention, when $X = \emptyset$ we set

$$\sup \emptyset = -\infty.$$

The element $\inf_{v \in U} J(v)$ is just $\inf\{J(v) \mid v \in U\}$. The notation J^* is often used to denote $\inf_{v \in U} J(v)$. If the function J is not bounded below, which means that for every $r \in \mathbb{R}$, there is some $u \in U$ such that $J(u) < r$, then

$$\inf_{v \in U} J(v) = -\infty,$$

and we say that our minimization problem has no solution, or that it is unbounded (below). For example, if $V = \Omega = \mathbb{R}$, $U = \{x \in \mathbb{R} \mid x \leq 0\}$, and $J(x) = -x$, then the function $J(x)$ is not bounded below and $\inf_{v \in U} J(v) = -\infty$.

The issue is that J^* may not belong to $\{J(u) \mid u \in U\}$, that is, it may not be achieved by some element $u \in U$, and solving the above problem consists in finding some $u \in U$ that achieves the value J^* in the sense that $J(u) = J^*$. If no such $u \in U$ exists, again we say that our minimization problem has no solution.

The minimization problem

$$\begin{aligned} &\text{find} \quad u \\ &\text{such that} \quad u \in U \text{ and } J(u) = \inf_{v \in U} J(v) \end{aligned}$$

is often presented in the following more informal way:

$$\begin{aligned} &\text{minimize} \quad J(v) \\ &\text{subject to} \quad v \in U. \end{aligned}$$

A vector $u \in U$ such that $J(u) = \inf_{v \in U} J(v)$ is often called a *minimizer* of J over U. Some authors denote the set of minimizers of J over U by $\arg\min_{v \in U} J(v)$ and write

$$u \in \arg\min_{v \in U} J(v)$$

to express that u is such a minimizer. When such a minimizer is unique, by abuse of notation, this unique minimizer u is denoted by

$$u = \arg\min_{v \in U} J(v).$$

We prefer not to use this notation, althoug it seems to have invaded the literature.

If we need to maximize rather than minimize a function, then we try to find some $u \in U$ such that

$$J(u) = \sup_{v \in U} J(v).$$

Here $\sup_{v \in U} J(v)$ is the least upper bound of the set $\{J(u) \mid u \in U\}$. Some authors denote the set of *maximizers* of J over U by $\arg\max_{v \in U} J(v)$.

Remark: Some authors define an *extended real-valued function* as a function $f \colon \Omega \to \mathbb{R}$ which is allowed to take the value $-\infty$ or even $+\infty$ for some of its arguments. Although this may be convenient to deal with situations where we need to consider $\inf_{v \in U} J(v)$ or $\sup_{v \in U} J(v)$, such "functions" are really partial functions and we prefer not to use the notion of extended real-valued function.

In most cases, U is defined as the set of solutions of a finite sets of *constraints*, either equality constraints $\varphi_i(v) = 0$, or inequality constraints $\varphi_i(v) \leq 0$, where the $\varphi_i \colon \Omega \to \mathbb{R}$ are some given functions. The function J is often called the *functional* of the optimization problem. This is a slightly odd terminology, but it is justified if V is a function space.

The following questions arise naturally:

(1) Results concerning the *existence and uniqueness* of a solution of the above problem. In the next section we state sufficient conditions either on the domain U or on the function J that ensure the existence of a solution.

(2) The *characterization* of the possible solutions of the above problem. These are conditions for any element $u \in U$ to be a solution of the problem. Such conditions usually involve the derivative dJ_u of J, and possibly the derivatives of the functions φ_i defininig U. Some of these conditions become sufficient when the functions φ_i are convex,

(3) The effective construction of *algorithms*, typically iterative algorihms that construct a sequence $(u_k)_{k \geq 1}$ of elements of U whose limit is a solution $u \in U$ of our problem. It is then necessary to understand when and how quickly such sequences converge. Gradient descent methods fall under this category. As a general rule, unconstrained problems (for which $U = \Omega = V$) are (much) easier to deal with than constrained problems (where $U \neq V$).

The material of this chapter is heavily inspired by Ciarlet [30]. In this chapter it is assumed that V is a real vector space with an inner product $\langle -, - \rangle$. If V is infinite dimensional, then we assume that it is a real Hilbert space (it is complete). As usual, we write

$\|u\| = \langle u, u\rangle^{1/2}$ for the norm associated with the inner product $\langle -, -\rangle$. The reader may want to review Section 28.1, especially the projection lemma and the Riesz representation theorem.

As a matter of terminology, if U is defined by inequality and equality constraints as

$$U = \{v \in \Omega \mid \varphi_i(v) \leq 0,\ i = 1, \ldots, m,\ \psi_j(v) = 0,\ j = 1, \ldots, p\},$$

if J and all the functions φ_i and ψ_j are affine, the problem is said to be *linear* (or a *linear program*), and otherwise *nonlinear*. If J is of the form

$$J(v) = \langle Av, v\rangle - \langle b, v\rangle$$

where A is a nonzero symmetric positive semidefinite matrix and the constraints are affine, the problem is called a *quadratic programming problem*.

We begin with the case where U is a closed but possibly unbounded subset of $\mathbb{R}^n$. In this case the following type of functions arise.

Definition 29.1. A real-valued function $J\colon V \to \mathbb{R}$ defined on a normed vector space V is *coercive* iff for any sequence $(v_k)_{k\geq 1}$ of vectors $v_k \in V$, if $\lim_{k\mapsto\infty} \|v_k\| = \infty$, then

$$\lim_{k\mapsto\infty} J(v_k) = +\infty.$$

For example, the function $f(x) = x^2 + 2x$ is coercive, but an affine function $f(x) = ax + b$ is not.

Proposition 29.1. *Let U be a nonempty, closed subset of $\mathbb{R}^n$, and let $J\colon \mathbb{R}^n \to \mathbb{R}$ be a continuous function which is coercive if U is unbounded. Then there is a least one element $u \in \mathbb{R}^n$ such that*

$$u \in U \quad \textit{and} \quad J(u) = \inf_{v\in U} J(v).$$

Proof. Since $U \neq \emptyset$, pick any $u_0 \in U$. Since J is coercive, there is some $r > 0$ such that for all $v \in V$, if $\|v\| > r$ then $J(u_0) < J(v)$. It follows that J is minimized over the set

$$U_0 = U \cap \{v \in \mathbb{R}^n \mid \|v\| \leq r\}.$$

Since U is closed and since the closed ball $\{v \in \mathbb{R}^n \mid \|v\| \leq r\}$ is compact, U_0 is compact, but we know that any continuous function on a compact set has a minimum which is achieved. □

The key point in the above proof is the fact that U_0 is compact. In order to generalize Proposition 29.1 to the case of an infinite dimensional vector space, we need some additional assumptions, and it turns out that the convexity of U and of the function J is sufficient. The key is that convex, closed and bounded subsets of a Hilbert space are "weakly compact."

Definition 29.2. Let V be a Hilbert space. A sequence $(u_k)_{k\geq 1}$ of vectors $u_k \in V$ *converges weakly* if there is some $u \in V$ such that

$$\lim_{k\mapsto\infty} \langle v, u_k\rangle = \langle v, u\rangle \quad \text{for every } v \in V.$$

Recall that a Hibert space is separable if it has a countable Hilbert basis (see Definition 33.4). Also, in a Euclidean space V the inner product induces an isomorphism between V and its dual V^*. In our case, we need the isomorphism $\sharp$ from V^* to V defined such that for every linear form $\omega \in V^*$, the vector $\omega^\sharp \in V$ is uniquely defined by the equation

$$\omega(v) = \langle v, \omega^\sharp\rangle \quad \text{for all } v \in V.$$

In a Hilbert space, the dual space V' is the set of all continuous linear forms $\omega\colon V \to \mathbb{R}$, and the existence of the isomorphism $\sharp$ between V' and V is given by the Riesz representation theorem; see Proposition 28.8. This theorem allows a generalization of the notion of gradient. Indeed, if $f\colon V \to \mathbb{R}$ is a function defined on the Hilbert space V and if f is differentiable at some point $u \in V$, then by definition, the derivative $df_u\colon V \to \mathbb{R}$ is a continuous linear form, so by the Riesz representation theorem (Proposition 28.8) there is a unique vector, denoted $\nabla f_u \in V$, such that

$$df_u(v) = \langle v, \nabla f_u\rangle \quad \text{for all } v \in V.$$

By definition, the vector ∇f_u is the gradient of f at u.

Similarly, since the second derivative $\mathrm{D}^2 f_u\colon V \to V'$ of f induces a continuous symmetric billinear form from $V \times V$ to $\mathbb{R}$, by Proposition 28.9, there is a unique continuous self-adjoint linear map $\nabla^2 f_u\colon V \to V$ such that

$$\mathrm{D}^2 f_u(v, w) = \langle \nabla^2 f_u(v), w\rangle \quad \text{for all } v, w \in V.$$

The map $\nabla^2 f_u$ is a generalization of the Hessian.

The next theorem is a rather general result about the existence of minima of convex functions defined on convex domains. The proof is quite involved and can be omitted upon first reading.

Theorem 29.2. *Let U be a nonempty, convex, closed subset of a separable Hilbert space V, and let $J\colon V \to \mathbb{R}$ be a convex, differentiable function which is coercive if U is unbounded. Then there is a least one element $u \in V$ such that*

$$u \in U \quad \textit{and} \quad J(u) = \inf_{v\in U} J(v).$$

Proof. As in the proof of Proposition 29.1, since the function J is coercive, we may assume that U is bounded and convex (however, if V infinite dimensional, then U is not compact in general). The proof proceeds in four steps.

Step 1. Consider a *minimizing sequence* $(u_k)_{k\geq 0}$, namely a sequence of elements $u_k \in V$ such that

$$u_k \in U \quad \text{for all } k \geq 0, \quad \lim_{k\mapsto\infty} J(u_k) = \inf_{v\in U} J(v).$$

At this stage, it is possible that $\inf_{v\in U} J(v) = -\infty$, but we will see that this is actually impossible. However, since U is bounded, the sequence $(u_k)_{k\geq 0}$ is bounded. Our goal is to prove that there is some subsequence of $(w_\ell)_{\ell\geq 0}$ of $(u_k)_{k\geq 0}$ that converges weakly.

Since the sequence $(u_k)_{k\geq 0}$ is bounded there is some constant $C > 0$ such that $\|u_k\| \leq C$ for all $k \geq 0$. Then, by the Cauchy–Schwarz inequality, for every $v \in V$ we have

$$|\langle v, u_k\rangle| \leq \|v\| \, \|u_k\| \leq C \, \|v\| \, ,$$

which shows that the sequence $(\langle v, u_k\rangle)_{k\geq 0}$ is bounded. Since V is a separable Hilbert space, there is a countable family $(v_k)_{k\geq 0}$ of vectors $v_k \in V$ which is dense in V. Since the sequence $(\langle v_1, u_k\rangle)_{k\geq 0}$ is bounded (in $\mathbb{R}$), we can find a convergent subsequence $(\langle v_1, u_{i_1(j)}\rangle)_{j\geq 0}$. Similarly, since the sequence $(\langle v_2, u_{i_1(j)}\rangle)_{j\geq 0}$ is bounded, we can find a convergent subsequence $(\langle v_2, u_{i_2(j)}\rangle)_{j\geq 0}$, and in general, since the sequence $(\langle v_k, u_{i_{k-1}(j)}\rangle)_{j\geq 0}$ is bounded, we can find a convergent subsequence $(\langle v_k, u_{i_k(j)}\rangle)_{j\geq 0}$.

We obtain the following infinite array:

$$\begin{pmatrix}
\langle v_1, u_{i_1(1)}\rangle & \langle v_2, u_{i_2(1)}\rangle & \cdots & \langle v_k, u_{i_k(1)}\rangle & \cdots \\
\langle v_1, u_{i_1(2)}\rangle & \langle v_2, u_{i_2(2)}\rangle & \cdots & \langle v_k, u_{i_k(2)}\rangle & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
\langle v_1, u_{i_1(k)}\rangle & \langle v_2, u_{i_2(k)}\rangle & \cdots & \langle v_k, u_{i_k(k)}\rangle & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots
\end{pmatrix}$$

Consider the "diagonal" sequence $(w_\ell)_{\ell\geq 0}$ defined by

$$w_\ell = u_{i_\ell(\ell)}, \quad \ell \geq 0.$$

We are going to prove that for every $v \in V$, the sequence $(\langle v, w_\ell\rangle)_{\ell\geq 0}$ has a limit.

By construction, for every $k \geq 0$, the sequence $(\langle v_k, w_\ell\rangle)_{\ell\geq 0}$ has a limit, which is the limit of the sequence $(\langle v_k, u_{i_k(j)}\rangle)_{j\geq 0}$, since the sequence $(i_\ell(\ell))_{\ell\geq 0}$ is a subsequence of every sequence $(i_\ell(j))_{j\geq 0}$ for every $\ell \geq 0$.

Pick any $v \in V$ and any $\epsilon > 0$. Since $(v_k)_{k\geq 0}$ is dense in V, there is some v_k such that

$$\|v - v_k\| \leq \epsilon/(4C).$$

Then we have

$$\begin{aligned}
|\langle v, w_\ell\rangle - \langle v, w_m\rangle| &= |\langle v, w_\ell - w_m\rangle| \\
&= |\langle v_k + v - v_k, w_\ell - w_m\rangle| \\
&= |\langle v_k, w_\ell - w_m\rangle + \langle v - v_k, w_\ell - w_m\rangle| \\
&\leq |\langle v_k, w_\ell\rangle - \langle v_k, w_m\rangle| + |\langle v - v_k, w_\ell - w_m\rangle|.
\end{aligned}$$

By Cauchy–Schwarz and since $\|w_\ell - w_m\| \leq \|w_\ell\| + \|w_m\| \leq C + C = 2C$,

$$|\langle v - v_k, w_\ell - w_m\rangle| \leq \|v - v_k\| \, \|w_\ell - w_m\| \leq (\epsilon/(4C))2C = \epsilon/2,$$

so

$$|\langle v, w_\ell\rangle - \langle v, w_m\rangle| \leq |\langle v_k, w_\ell - w_m\rangle| + \epsilon/2.$$

With the element v_k held fixed, by a previous argument the sequence $(\langle v_k, w_\ell\rangle)_{\ell\geq 0}$ converges, so it is a Cauchy sequence. Consequently there is some ℓ_0 (depending on ϵ and v_k) such that

$$|\langle v_k, w_\ell\rangle - \langle v_k, w_m\rangle| \leq \epsilon/2 \quad \text{for all } \ell, m \geq \ell_0,$$

so we get

$$|\langle v, w_\ell\rangle - \langle v, w_m\rangle| \leq \epsilon/2 + \epsilon/2 = \epsilon \quad \text{for all } \ell, m \geq \ell_0.$$

This proves that the sequence $(\langle v, w_\ell\rangle)_{\ell\geq 0}$ is a Cauchy sequence, and thus it converges.

Define the function $g\colon V \to \mathbb{R}$ by

$$g(v) = \lim_{\ell\mapsto\infty} \langle v, w_\ell\rangle, \quad \text{for all } v \in V.$$

Since

$$|\langle v, w_\ell\rangle| \leq \|v\| \, \|w_\ell\| \leq C \, \|v\| \quad \text{for all } \ell \geq 0,$$

we have

$$|g(v)| \leq C \, \|v\| \, ,$$

so g is a continuous linear map. By the Riesz representation theorem (Proposition 28.8), there is a unique $u \in V$ such that

$$g(v) = \langle v, u\rangle \quad \text{for all } v \in V,$$

which shows that

$$\lim_{\ell\mapsto\infty} \langle v, w_\ell\rangle = \langle v, u\rangle \quad \text{for all } v \in V,$$

namely the subsequence $(w_\ell)_{\ell\geq 0}$ of the sequence $(u_k)_{k\geq 0}$ converges weakly to $u \in V$.

Step 2. We prove that the "weak limit" u of the sequence $(w_\ell)_{\ell\geq 0}$ belongs to U.

Consider the projection $p_U(u)$ of $u \in V$ onto the closed convex set U. Since $w_\ell \in U$, by Proposition 28.5 we have

$$\langle p_U(u) - u, w_\ell - p_U(u)\rangle \geq 0 \quad \text{for all } \ell \geq 0.$$

The weak convergence of the sequence $(w_\ell)_{\ell\geq 0}$ to u implies that

$$\begin{aligned} 0 \leq \lim_{\ell\mapsto\infty} \langle p_U(u) - u, w_\ell - p_U(u)\rangle &= \langle p_U(u) - u, u - p_U(u)\rangle \\ &= - \, \|p_U(u) - u\| \leq 0, \end{aligned}$$

so $\|p_U(u) - u\| = 0$, which means that $p_U(u) = u$, and so $u \in U$.

Step 3. We prove that

$$J(v) \leq \liminf_{\ell \mapsto \infty} J(z_\ell)$$

for every sequence $(z_\ell)_{\ell \geq 0}$ converging weakly to some element $v \in V$.

Since J is assumed to be differentiable and convex, by Proposition 20.9 we have

$$J(v) + \langle \nabla J_v, z_\ell - v \rangle \leq J(z_\ell) \quad \text{for all } \ell \geq 0,$$

and by definition of weak convergence

$$\lim_{\ell \mapsto \infty} \langle \nabla J_v, z_\ell \rangle = \langle \nabla J_v, v \rangle,$$

so $\lim_{\ell \mapsto \infty} \langle \nabla J_v, z_\ell - v \rangle = 0$, and by definition of lim inf we get

$$J(v) \leq \liminf_{\ell \mapsto \infty} J(z_\ell)$$

for every sequence $(z_\ell)_{\ell \geq 0}$ converging weakly to some element $v \in V$.

Step 4. The weak limit $u \in U$ of the subsequence $(w_\ell)_{\ell \geq 0}$ extracted from the minimizing sequence $(u_k)_{k \geq 0}$ satisfies the equation

$$J(u) = \inf_{v \in U} J(v).$$

By Step (1) and Step (2) the subsequence $(w_\ell)_{\ell \geq 0}$ of the sequence $(u_k)_{k \geq 0}$ converges weakly to some element $u \in U$, so by Step (3) we have

$$J(u) \leq \liminf_{\ell \mapsto \infty} J(w_\ell).$$

On the other hand, by definition of $(w_\ell)_{\ell \geq 0}$ as a subsequence of $(u_k)_{k \geq 0}$, since the sequence $(J(u_k))_{k \geq 0}$ converges to $J(v)$, we have

$$J(u) \leq \liminf_{\ell \mapsto \infty} J(w_\ell) = \lim_{k \mapsto \infty} J(u_k) = \inf_{v \in U} J(v),$$

which proves that $u \in U$ achieves the minimum of J on U. □

Remark: Theorem 29.2 still holds if we only assume that J is convex and continuous. It also holds in a reflexive Banach space, of which Hilbert spaces are a special case; see Brezis [24], Corollary 3.23.

Theorem 29.2 is a rather general theorem whose proof is quite involved. For functions J of a certain type, we can obtain existence and uniqueness results that are easier to prove. This is true in particular for quadratic functionals.

Definition 29.3. Let V be a real Hilbert space. A function $J: V \to \mathbb{R}$ is called a *quadratic functional* if it is of the form

$$J(v) = \frac{1}{2}a(v, v) - h(v),$$

where $a: V \times V \to \mathbb{R}$ is a bilinear form which is symmetric and continuous, and $h: V \to \mathbb{R}$ is a continuous linear form.

Definition 29.3 is a natural extension of the notion of a quadratic functional on $\mathbb{R}^n$. Indeed, by Proposition 28.9, there is a unique continuous self-adjoint linear map $A: V \to V$ such that

$$a(u, v) = \langle Au, v\rangle \quad \text{for all } u, v \in V,$$

and by the Riesz representation theorem (Proposition 28.8), there is a unique $b \in V$ such that

$$h(v) = \langle b, v\rangle \quad \text{for all } v \in V.$$

Consequently, J can be written as

$$J(v) = \frac{1}{2}\langle Av, v\rangle - \langle b, v\rangle \quad \text{for all } v \in V.$$

Since a is bilinear and h is linear, observe that the derivative of J is given by

$$dJ_u(v) = a(u, v) - h(v) \quad \text{for all } v \in V,$$

or equivalently by

$$dJ_u(v) = \langle Au, v\rangle - \langle b, v\rangle = \langle Au - b, v\rangle, \quad \text{for all } v \in V.$$

Thus the gradient of J is given by

$$\nabla J_u = Au - b,$$

just as in the case of a quadratic function of the form $J(v) = (1/2)v^\top Av - b^\top v$, where A is a symmetric $n \times n$ matrix and $b \in \mathbb{R}^n$. To find the second derivative $\mathrm{D}^2 J_u$ of J at u we compute

$$dJ_{u+v}(w) - dJ_u(w) = a(u + v, w) - h(w) - (a(u, w) - h(w)) = a(v, w),$$

so

$$\mathrm{D}^2 J_u(v, w) = a(v, w) = \langle Av, w\rangle,$$

which yields

$$\nabla^2 J_u = A.$$

We will also make use of the following formula (if J is a quadratic functional):

$$J(u + \rho v) = \frac{\rho^2}{2}a(v, v) + \rho(a(u, v) - h(v)) + J(u).$$

Indeed, since a is symmetric bilinear and h is linear, we have

$$\begin{aligned} J(u+\rho v) &= \frac{1}{2}a(u+\rho v, u+\rho v) - h(u+\rho v) \\ &\quad \frac{\rho^2}{2}a(v,v) + \rho a(u,v) + \frac{1}{2}a(u,u) - h(u) - \rho h(v) \\ &= \frac{\rho^2}{2}a(v,v) + \rho(a(u,v) - h(v)) + J(u). \end{aligned}$$

Since $dJ_u(v) = a(u,v) - h(v) = \langle Au - b, v\rangle$ and $\nabla J_u = Au - b$, we can also write

$$J(u+\rho v) = \frac{\rho^2}{2}a(v,v) + \rho\langle \nabla J_u, v\rangle + J(u).$$

We have the following theorem about the existence and uniqueness of minima of quadratic functionals.

Theorem 29.3. *Given any Hilbert space V, let $J\colon V \to \mathbb{R}$ be a quadratic functional of the form*

$$J(v) = \frac{1}{2}a(v,v) - h(v).$$

Assume that there is some real number $\alpha > 0$ such that

$$a(v,v) \geq \alpha \|v\|^2 \quad \textit{for all } v \in V. \tag{$*_\alpha$}$$

If U is any nonempty, closed, convex subset of V, then there is a unique $u \in U$ such that

$$J(u) = \inf_{v\in U} J(v).$$

The element $u \in U$ satisfies the condition

$$a(u, v-u) \geq h(v-u) \quad \textit{for all } v \in U. \tag{$*$}$$

Conversely, if an element $u \in U$ satisfies $(*)$, then

$$J(u) = \inf_{v\in U} J(v).$$

If U is a subspace of V, then the above inequalities are replaced by the equations

$$a(u,v) = h(v) \quad \textit{for all } v \in U. \tag{$**$}$$

Proof. The key point is that the bilinear form a is actually an inner product in V. This is because it is positive definite, since $(*_\alpha)$ implies that

$$\sqrt{\alpha}\,\|v\| \leq (a(v,v))^{1/2},$$

and on the other hand the continuity of a implies that

$$a(v,v) \leq \|a\| \, \|v\|^2 ,$$

so we get

$$\sqrt{\alpha} \, \|v\| \leq (a(v,v))^{1/2} \leq \sqrt{\|a\|} \, \|v\| .$$

The above also shows that the norm $v \mapsto (a(v,v))^{1/2}$ induced by the inner product a is equivalent to the norm induced by the inner product $\langle -, - \rangle$ on V. Thus h is still continuous with respect to the norm $v \mapsto (a(v,v))^{1/2}$. Then by the Riesz representation theorem (Proposition 28.8), there is some unique $c \in V$ such that

$$h(v) = a(c,v) \quad \text{for all } v \in V.$$

Consequently, we can express $J(v)$ as

$$J(v) = \frac{1}{2} a(v,v) - a(c,v) = \frac{1}{2} a(v-c, v-c) - \frac{1}{2} a(c,c).$$

But then, minimizing $J(v)$ over U is equivalent to minimizing $(a(v-c,v-c))^{1/2}$ over $v \in U$, and by the projection lemma (Proposition 28.5) this is equivalent to finding the projection $p_U(c)$ of c on the closed convex set U with respect to the inner product a. Therefore, there is a unique $u = p_U(c) \in U$ such that

$$J(u) = \inf_{v \in U} J(v).$$

Also by Proposition 28.5, this unique element $u \in U$ is characterized by the condition

$$a(u-c, v-u) \geq 0 \quad \text{for all } v \in U.$$

Since

$$a(u-c, v-u) = a(u, v-u) - a(c, v-u) = a(u, v-u) - h(v-u),$$

the above inequality is equivalent to

$$a(u, v-u) \geq h(v-u) \quad \text{for all } v \in U. \tag{$*$}$$

If U is a subspace of V, then we have the condition

$$a(u-c, v) = 0 \quad \text{for all } v \in U,$$

which is equivalent to

$$a(u,v) = a(c,v) = h(v) \quad \text{for all } v \in U. \tag{$**$}$$

□

Note that the symmetry of the bilinear form a played a crucial role. Also, the inequalities

$$a(u, v-u) \geq h(v-u) \quad \text{for all } v \in U$$

are sometimes called *variational inequalities*.

A bilinear form $a\colon V \times V \to \mathbb{R}$ such that there is some real $\alpha > 0$ such that

$$a(v,v) \geq \alpha \|v\|^2 \quad \text{for all } v \in V$$

is said to be *coercive*.

Theorem 29.3 is the special case of Stampacchia's theorem, and the Lax–Milgram theorem when $U = V$, in the case where a is a symmetric bilinear form. To prove Stampacchia's theorem in general, we need to recall the *contraction mapping theorem*.

Definition 29.4. Let (E, d) be a metric space. A map $f\colon E \to E$ is a *contraction* (or a *contraction mapping*) if there is some real number k such that $0 \leq k < 1$ and

$$d(f(u), f(v)) \leq kd(u, v) \quad \text{for all } u, v \in E.$$

The number k is often called a *Lipschitz constant*.

The following theorem is proved in Section 18.8; see Theorem 18.23. A proof can be also found in Apostol [3], Dixmier [35], or Schwartz [90], among many sources. For the reader's convenience we restate this theorem.

Theorem 29.4. *(Contraction Mapping Theorem) Let (E, d) be a complete metric space. Every contraction $f\colon E \to E$ has a unique fixed point (that is, an element $u \in E$ such that $f(u) = u$).*

The contraction mapping theorem is also known as the *Banach fixed point theorem*.

Theorem 29.5. *(Lions–Stampacchia) Given a Hilbert space V, let $a\colon V \times V \to \mathbb{R}$ be a continuous bilinear form (not necessarily symmetric), let $h \in V'$ be a continuous linear form, and let J be given by*

$$J(v) = \frac{1}{2}\, a(v,v) - h(v), \quad v \in V.$$

If a is coercive, then for every nonempty, closed, convex subset U of V, there is a unique $u \in U$ such that

$$a(u, v-u) \geq h(v-u) \quad \textit{for all } v \in U. \tag{$*$}$$

If a is symmetric, then $u \in U$ is the unique element of U such that

$$J(u) = \inf_{v \in U} J(v).$$

Proof. As discussed just after Definition 29.3, by Proposition 28.9, there is a unique continuous linear map $A: V \to V$ such that

$$a(u, v) = \langle Au, v\rangle \quad \text{for all } u, v \in V,$$

with $\|A\| = \|a\| = C$, and by the Riesz representation theorem (Proposition 28.8), there is a unique $b \in V$ such that

$$h(v) = \langle b, v\rangle \quad \text{for all } v \in V.$$

Consequently, J can be written as

$$J(v) = \frac{1}{2}\langle Av, v\rangle - \langle b, v\rangle \quad \text{for all } v \in V. \tag{$*_1$}$$

Since $\|A\| = \|a\| = C$, we have $\|Av\| \leq \|A\| \, \|v\| = C \, \|v\|$ for all $v \in V$. Using $(*_1)$, the inequality $(*)$ is equivalent to finding u such that

$$\langle Au, v - u\rangle \geq \langle b, v - u\rangle \quad \text{for all } v \in V. \tag{$*_2$}$$

Let $\rho > 0$ be a constant to be determined later. Then $(*_2)$ is equivalent to

$$\langle \rho b - \rho Au + u - u, v - u\rangle \leq 0 \quad \text{for all } v \in V. \tag{$*_3$}$$

By the projection lemma (Proposition 28.5), $(*_3)$ is equivalent to finding $u \in U$ such that

$$u = p_U(\rho b - \rho Au + u). \tag{$*_4$}$$

We are led to finding a fixed point of the function $F: V \to V$ given by

$$F(v) = p_U(\rho b - \rho Av + v).$$

By Proposition 28.6, the projection map p_U does not increase distance, so

$$\|F(v_1) - F(v_2)\| \leq \|v_1 - v_2 - \rho(Av_1 - Av_2)\|.$$

Since a is coercive we have

$$a(v, v) \geq \alpha \, \|v\|^2,$$

since $a(v, v) = \langle Av, v\rangle$ we have

$$\langle Av, v\rangle \geq \alpha \, \|v\|^2 \quad \text{for all } v \in V, \tag{$*_5$}$$

and since

$$\|Av\| \leq C \, \|v\| \quad \text{for all } v \in V, \tag{$*_6$}$$

we get

$$\begin{aligned}\|F(v_1) - F(v_2)\|^2 &\leq \|v_1 - v_2\|^2 - 2\rho\langle Av_1 - Av_2, v_1 - v_2\rangle + \rho^2 \, \|Av_1 - Av_2\|^2 \\ &\leq \left(1 - 2\rho\alpha + \rho^2 C\right) \|v_1 - v_2\|^2.\end{aligned}$$

If we pick $\rho > 0$ such that $\rho < 2\alpha/C^2$, then

$$k^2 = 1 - 2\rho\alpha + \rho^2 C < 1,$$

and then

$$\|F(v_1) - F(v_2)\| \leq k \|v_1 - v_2\|, \tag{$*_7$}$$

with $0 \leq k < 1$, which shows that F is a contraction. By Theorem 29.4, the map F has a unique fixed point $u \in U$, which concludes the proof of the first statement. If a is also symmetric, then the second statement is just the first part of Proposition 29.3. □

Remark: Many physical problems can be expressed in terms of an unknown function u that satisfies some inequality

$$a(u, v - u) \geq h(v - u) \quad \text{for all } v \in U,$$

for some set U of "admissible" functions which is closed and convex. The bilinear form a and the linear form h are often given in terms of integrals. The above inequality is called a *variational inequality*.

In the special case where $U = V$ we obtain the Lax–Milgram theorem.

Theorem 29.6. *(Lax–Milgram's Theorem) Given a Hilbert space V, let $a\colon V \times V \to \mathbb{R}$ be a continuous bilinear form (not necessarily symmetric), let $h \in V'$ be a continuous linear form, and let J be given by*

$$J(v) = \frac{1}{2} a(v, v) - h(v), \quad v \in V.$$

If a is coercive, which means that there is some $\alpha > 0$ such that

$$a(v, v) \geq \alpha \|v\|^2 \quad \textit{for all } v \in V,$$

then there is a unique $u \in V$ such that

$$a(u, v) = h(v) \quad \textit{for all } v \in U.$$

If a is symmetric, then $u \in V$ is the unique element of V such that

$$J(u) = \inf_{v \in V} J(v).$$

The Lax–Milgram Theorem play an important role in solving linear elliptic partial differential equations; see Brezis [24].

We now consider various methods, known as gradient descents, to find minima of certain types of functionals.

29.2 Gradient Descent Methods for Unconstrained Problems

We begin by defining the notion of an elliptic functional which generalizes the notion of a quadratic function defined by a symmetric positive definite matrix. Elliptic functionals are well adapted to the types of iterative methods described in this section, and lend themselves well to an analysis of the convergence of these methods.

Definition 29.5. Given a Hilbert space V, a functional $J\colon V \to \mathbb{R}$ is said to be *elliptic* if it is continuously differentiable on V, and if there is some constant $\alpha > 0$ such that

$$\langle \nabla J_v - \nabla J_u, v - u\rangle \geq \alpha \|v - u\|^2 \quad \text{for all } u, v \in V.$$

The following proposition gathers properties of elliptic functionals that will be used later to analyze the convergence of various gradient descent methods.

Theorem 29.7. *Let V be a Hilbert space.*

(1) An elliptic functional $J\colon V \to \mathbb{R}$ is strictly convex and coercice. Furthermore, it satisfies the identity

$$J(v) - J(u) \geq \langle \nabla J_u, v - u\rangle + \frac{\alpha}{2}\|v - u\|^2 \quad \textit{for all } u, v \in V.$$

(2) If U is a nonempty, convex, closed subset of the Hilbert space V and if J is an elliptic functional, then the problem (P),

$$\begin{aligned} &\text{find} \quad u \\ &\text{such that} \quad u \in U \ \textit{ and } \ J(u) = \inf_{v \in U} J(v) \end{aligned}$$

has a unique solution.

(3) Suppose the set U is convex and that the functional J is elliptic. Then an element $u \in U$ is a solution of the problem (P) if and only if it satisfies the condition

$$\langle \nabla J_u, v - u\rangle \geq 0 \quad \textit{for every } v \in U$$

in the general case, or

$$\nabla J_u = 0 \quad \textit{if } U = V$$

(4) A functional J which is twice differentiable in V is elliptic if and only if

$$\langle \nabla^2 J_u(w), w\rangle \geq \alpha \|w\|^2 \quad \textit{for all } u, w \in V.$$

Proof. (1) Since J is a C^1-function, by Taylor's formula with integral remainder in the case $m = 0$ (Theorem 19.24), we get

$$\begin{aligned}
J(v) - J(u) &= \int_0^1 dJ_{u+t(v-u)}(v-u)dt \\
&= \int_0^1 \langle \nabla J_{u+t(v-u)}, v-u\rangle dt \\
&= \langle \nabla J_u, v-u\rangle + \int_0^1 \langle \nabla J_{u+t(v-u)} - \nabla J_u, v-u\rangle dt \\
&= \langle \nabla J_u, v-u\rangle + \int_0^1 \frac{\langle \nabla J_{u+t(v-u)} - \nabla J_u, t(v-u)\rangle}{t} dt \\
&\geq \langle \nabla J_u, v-u\rangle + \int_0^1 \alpha t \left\|v-u\right\|^2 dt \qquad \text{since } J \text{ is elliptic} \\
&= \langle \nabla J_u, v-u\rangle + \frac{\alpha}{2}\left\|v-u\right\|^2.
\end{aligned}$$

Using the inequality

$$J(v) - J(u) \geq \langle \nabla J_u, v-u\rangle + \frac{\alpha}{2}\left\|v-u\right\|^2 \quad \text{for all } u, v \in V,$$

by Proposition 20.9(2), since

$$J(v) > J(u) + \langle \nabla J_u, v-u\rangle \quad \text{for all } u, v \in V,\ v \neq u,$$

the function J is strictly convex. It is coercive because

$$\begin{aligned}
J(v) &\geq J(0) + \langle \nabla J_0, v\rangle + \frac{\alpha}{2}\left\|v\right\|^2 \\
&\geq J(0) - \left\|\nabla J_0\right\|\left\|v\right\| + \frac{\alpha}{2}\left\|v\right\|^2,
\end{aligned}$$

and the term $(-\left\|\nabla J_0\right\| + \frac{\alpha}{2}\left\|v\right\|)\left\|v\right\|$ goes to $+\infty$ when $\left\|v\right\|$ tends to $+\infty$.

(2) Since by (1) the functional J is coercive, by Theorem 29.2, problem (P) has a solution. Since J is strictly convex, by Theorem 20.11(2), it has a unique minimum.

(3) These are just the conditions of Theorem 20.11(3, 4).

(4) If J is twice differentiable, we showed in Section 19.4 that we have

$$\mathrm{D}^2 J_u(w,w) = \mathrm{D}_w(\mathrm{D}J)(u) = \lim_{\theta \mapsto 0} \frac{\mathrm{D}J_{u+\theta w}(w) - \mathrm{D}J_u(w)}{\theta},$$

and since

$$\begin{aligned}
\mathrm{D}^2 J_u(w,w) &= \langle \nabla^2 J_u(w), w\rangle \\
\mathrm{D}J_{u+\theta w}(w) &= \langle \nabla J_{u+\theta w}, w\rangle \\
\mathrm{D}J_u(w) &= \langle \nabla J_u, w\rangle,
\end{aligned}$$

and since J is elliptic, for all $u, w \in V$ we can write

$$\begin{aligned}\langle \nabla^2 J_u(w), w\rangle &= \lim_{\theta \mapsto 0} \frac{\langle \nabla J_{u+\theta w} - \nabla J_u, w\rangle}{\theta} \\ &= \lim_{\theta \mapsto 0} \frac{\langle \nabla J_{u+\theta w} - \nabla J_u, \theta w\rangle}{\theta^2} \\ &\geq \alpha \left\| w \right\|^2.\end{aligned}$$

Conversely, assume that the condition

$$\langle \nabla^2 J_u(w), w\rangle \geq \alpha \left\| w \right\|^2 \quad \text{for all } u, w \in V$$

holds. If we define the function $g\colon V \to \mathbb{R}$ by

$$g(w) = \langle \nabla J_w, v - u\rangle = dJ_w(v-u) = \mathrm{D}_{v-u}J(w),$$

where u and v are fixed vectors in V, then we have

$$dg_{u+\theta(v-u)}(v-u) = \mathrm{D}_{v-u}g(u+\theta(v-u)) = \mathrm{D}_{v-u}\mathrm{D}_{v-u}J(u+\theta(v-u)) = \mathrm{D}^2 J_{u+\theta(v-u)}(v-u, v-u)$$

and we can apply the Taylor–MacLaurin formula (Theorem 19.23 with $m = 0$) to g, and we get

$$\begin{aligned}\langle \nabla J_v - \nabla J_u, v - u\rangle &= g(v) - g(u) \\ &= dg_{u+\theta(v-u)}(v-u) \qquad (0 < \theta < 1) \\ &= \mathrm{D}^2 J_{u+\theta(v-u)}(v-u, v-u) \\ &= \langle \nabla^2 J_{u+\theta(v-u)}(v-u), v-u\rangle \\ &\geq \alpha \left\| v - u \right\|^2,\end{aligned}$$

which shows that J is elliptic. □

If $J\colon \mathbb{R}^n \to \mathbb{R}$ is a quadratic function given by

$$J(v) = \frac{1}{2}\langle Av, v\rangle - \langle b, v\rangle$$

(where A is a symmetric $n \times n$ matrix and $\langle -, -\rangle$ is the standard Eucidean inner product), then J is elliptic iff A is positive definite. This is because

$$\langle \nabla^2 J_u(w), w\rangle = \langle Aw, w\rangle \geq \lambda_1 \left\| w \right\|^2$$

where λ_1 is the smallest eigenvalue of A; see Proposition 13.23 (Rayleigh–Ritz). Note that by Proposition 13.23 (Rayleigh–Ritz) we also have

$$\langle \nabla^2 J_u(w), w\rangle \leq \lambda_n \left\| w \right\|^2$$

where λ_n is the largest eigenvalue of A; this fact will be useful later on.

Similarly, given a quadratic functional J defined on a Hilbert space V, where

$$J(v) = \frac{1}{2}a(v,v) - h(v),$$

by Theorem 29.7 (4), the functional J is elliptic iff there is some $\alpha > 0$ such that

$$\langle \nabla^2 J_u(v), v\rangle = a(v,v) \geq \alpha \|v\|^2 \quad \text{for all } v \in V.$$

This is precisely the hypothesis $(*_\alpha)$ used in Theorem 29.3.

We will now describe methods for solving unconstrained minimization problems, that is, finding the minimum (or minima) of a functions J over the whole space V. These methods are *iterative*, which means that given some *initial* vector u_0, we construct a sequence $(u_k)_{k\geq 0}$ that converges to a minimum u of the function J.

The key step is define u_{k+1} from u_k, and a first idea is to reduce the problem to a simpler problem, namely the minimization of a function of a *single (real) variable*. For this, we need two perform two steps:

(1) Find a *descent direction* at u_k, which is a some nonzero vector d_k which is usually determined from the gradient of J at various points.

(2) Find the minimum of the restriction of the function J along the line through u_k and parallel to the direction d_k. This means finding a real $\rho_k \in \mathbb{R}$ (depending on u_k and d_k) such that

$$J(u_k + \rho_k d_k) = \inf_{\rho\in\mathbb{R}} J(u_k + \rho d_k).$$

This problem only succeeds if ρ_k is unique, in which case we set

$$u_{k+1} = u_k + \rho_k d_k.$$

This step is often called a *line search* or *line minimization*, and ρ_k is called the *stepsize* parameter. See Figure 29.1.

If J is a quadratic elliptic functional of the form

$$J(v) = \frac{1}{2}a(v,v) - h(v),$$

then given d_k, there is a unique ρ_k solving the line search in Step (2). This is because, as we explained earlier, we have

$$J(u_k + \rho d_k) = \frac{\rho^2}{2}a(d_k,d_k) + \rho\langle \nabla J_{u_k}, d_k\rangle + J(u_k),$$

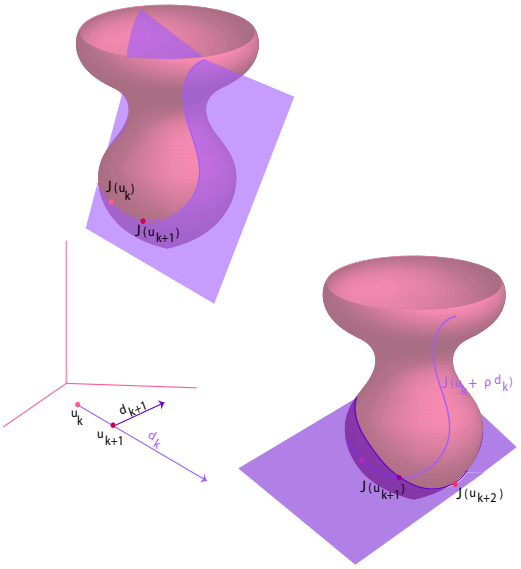


Figure 29.1: Let $J\colon \mathbb{R}^2 \to \mathbb{R}$ be the function whose graph is represented by the pink surface. Given a point u_k in the xy-plane, and a direction d_k, we calculate first u_{k+1} and then u_{k+2}.

and since $a(d_k, d_k) > 0$ (because J is elliptic), the above function of ρ has a unique minimum when its derivative is zero, namely

$$\rho\, a(d_k, d_k) + \langle \nabla J_{u_k}, d_k \rangle = 0.$$

We now consider one of the simplest methods for choosing the directions of descent in the case where $V = \mathbb{R}^n$, which is to pick the directions of the coordinate axes in a cyclic fashion. Such a method is called the *method of relaxation*.

If we write

$$u_k = (u_1^k, u_2^k, \ldots, u_n^k),$$

then the components u_i^{k+1} of u_{k+1} are computed in terms of u_k by solving from top down the following system of equations:

$$\begin{aligned}
J(\mathbf{u_1^{k+1}}, u_2^k, u_3^k, \ldots, u_n^k) &= \inf_{\lambda \in \mathbb{R}} J(\lambda, u_2^k, u_3^k, \ldots, u_n^k)\\
J(u_1^{k+1}, \mathbf{u_2^{k+1}}, u_3^k, \ldots, u_n^k) &= \inf_{\lambda \in \mathbb{R}} J(u_1^{k+1}, \lambda, u_3^k, \ldots, u_n^k)\\
&\vdots\\
J(u_1^{k+1}, \ldots, u_{n-1}^{k+1}, \mathbf{u_n^{k+1}}) &= \inf_{\lambda \in \mathbb{R}} J(u_1^{k+1}, \ldots, u_{n-1}^{k+1}, \lambda).
\end{aligned}$$

Another and more informative way to write the above system is to define the vectors $u_{k;i}$

by

$$\begin{aligned}
u_{k;0} &= (u_1^k, u_2^k, \ldots, u_n^k)\\
u_{k;1} &= (u_1^{k+1}, u_2^k, \ldots, u_n^k)\\
&\vdots\\
u_{k;i} &= (u_1^{k+1}, \ldots, u_i^{k+1}, u_{i+1}^k, \ldots, u_n^k)\\
&\vdots\\
u_{k;n} &= (u_1^{k+1}, u_2^{k+1}, \ldots, u_n^{k+1}).
\end{aligned}$$

Note that $u_{k;0} = u_k$ and $u_{k;n} = u_{k+1}$. Then our minimization problem can be written as

$$\begin{aligned}
J(u_{k;1}) &= \inf_{\lambda\in\mathbb{R}} J(u_{k;0} + \lambda e_1)\\
&\vdots\\
J(u_{k;i}) &= \inf_{\lambda\in\mathbb{R}} J(u_{k;i-1} + \lambda e_i)\\
&\vdots\\
J(u_{k;n}) &= \inf_{\lambda\in\mathbb{R}} J(u_{k;n-1} + \lambda e_n),
\end{aligned}$$

where e_i denotes the ith canonical basis vector in $\mathbb{R}^n$. If J is differentiable, necessary conditions for a minimum, which are also sufficient if J is convex, is that the directional derivatives $dJ_v(e_i)$ be all zero, that is,

$$\langle \nabla J_v, e_i\rangle = 0 \quad i = 0, \ldots, n.$$

The following result regarding the convergence of the mehod of relaxation is proved in Ciarlet [30] (Chapter 8, Theorem 8.4.2).

Proposition 29.8. *If the functional $J\colon \mathbb{R}^n \to \mathbb{R}$ is elliptic, then the relaxation method converges.*

Remarks: The proof of Proposition 29.8 uses Theorem 29.7. The finite dimensionality of $\mathbb{R}^n$ also plays a crucial role. The differentiability of the function J is also crucial. Examples where the method loops forever if J is not differentiable can be given; see Ciarlet [30] (Chapter 8, Section 8.4). The proof of Proposition 29.8 yields an *a priori* bound on the error $\|u - u_k\|$. If J is a quadratic functional

$$J(v) = \frac{1}{2} v^\top A v - b^\top v,$$

where A is a symmetric positive definite matrix, then $\nabla J_v = Av - b$, so the above system to solve for u_{k+1} in terms of u_k becomes the *Gauss–Seidel method* for solving a linear system; see Section 7.3.

We now discuss gradient methods. The intuition behind these methods is that the convergence of an iterative method ought to be better if the difference $J(u_k) - J(u_{k+1})$ is as large as possible during every iteration step. To achieve this, it is natural to pick the descent direction to be the one *in the opposite direction of the gradient vector* ∇J_{u_k}. This choice is justified by the fact that we can write

$$J(u_k + w) = J(u_k) + \langle \nabla J_{u_k}, w \rangle + \epsilon(w) \|w\|, \quad \text{with } \lim_{w \mapsto 0} \epsilon(w) = 0.$$

If $\nabla J_{u_k} \neq 0$, the first-order part of the variation of the function J is bounded in absolute value by $\|\nabla J_{u_k}\| \|w\|$ (by the Cauchy–Schwarz inequality), with equality if ∇J_{u_k} and w are collinear.

Gradient descent methods pick the direction of descent to be $d_k = -\nabla_{u_k} J$, so that we have

$$u_{k+1} = u_k - \rho_k \nabla J_{u_k},$$

where we put a negative sign in front of of the variable ρ_k as a reminder that the descent direction is *opposite* to that of the gradient; a positive value is expected for the scalar ρ_k.

There are three standard methods to pick ρ_k:

(1) *Gradient method with fixed stepsize parameter*. This is the simplest and cheapest method which consists of using the same constant $\rho_k = \rho$ for all iterations.

(2) *Gradient method with variable stepsize parameter*. In this method, the parameter ρ_k is adjusted in the course of iterations according to various criteria.

(3) *Gradient method with optimal stepsize parameter*, also called *steepest descent method for the Euclidean norm*. This is a version of method 2 in which ρ_k is determined by the following line search:

$$J(u_k - \rho_k \nabla J_{u_k}) = \inf_{\rho \in \mathbb{R}} J(u_k - \rho \nabla J_{u_k}).$$

This optimization problem only succeeds if the above minimization problem has a unique solution.

We have the following useful result about the convergence of the gradient method with optimal parameter.

Proposition 29.9. *Let $J : \mathbb{R}^n \to \mathbb{R}$ be an elliptic functional. Then the gradient method with optimal stepsize parameter converges.*

Proof. Since J is elliptic, by Theorem 29.7, the functional J has a unique minimum u characterized by $\nabla J_u = 0$. Our goal is to prove that the sequence $(u_k)_{k \geq 0}$ constructed using the gradient method with optimal parameter converges to u, started from any initial vector u_0. Without loss of generality we may assume that $u_{k+1} \neq u_k$ and $\nabla J_{u_k} \neq 0$ for all k, since otherwise the method converges in a finite number of steps.

Step 1. Any two consecutive descent directions are orthogonal, and

$$J(u_k) - J(u_{k+1}) \geq \frac{\alpha}{2} \|u_k - u_{k+1}\|^2.$$

Let $\varphi_k \colon \mathbb{R} \to \mathbb{R}$ be the function given by

$$\varphi_k(\rho) = J(u_k - \rho \nabla J_{u_k}).$$

Since the function φ_k is strictly convex and coercive, it has a unique minimum ρ_k which is the unique solution of the equation $\varphi_k'(\rho) = 0$. By the chain rule

$$\begin{aligned}\varphi_k'(\rho) &= dJ_{u_k - \rho \nabla J_{u_k}}(-\nabla J_{u_k}) \\ &= -\langle \nabla J_{u_k - \rho \nabla J_{u_k}}, \nabla J_{u_k} \rangle,\end{aligned}$$

and since $u_{k+1} = u_k - \rho_k \nabla J_{u_k}$ we get

$$\langle \nabla J_{u_{k+1}}, \nabla J_{u_k} \rangle = 0,$$

which shows that two consecutive descent directions are orthogonal.

Since $u_{k+1} = u_k - \rho_k \nabla J_{u_k}$ and we assumed that that $u_{k+1} \neq u_k$, we have $\rho_k \neq 0$, and we also get

$$\langle \nabla J_{u_{k+1}}, u_{k+1} - u_k \rangle = 0.$$

By the inequality of Theorem 29.7(1) we have

$$J(u_k) - J(u_{k+1}) \geq \frac{\alpha}{2} \|u_k - u_{k+1}\|^2.$$

Step 2. $\lim_{k \mapsto \infty} \|u_k - u_{k+1}\| = 0$.

It follows from the inequality proved in Step 1 that the sequence $(J(u_k))_{k \geq 0}$ is decreasing and bounded below (by $J(u)$, where u is the minimum of J), so it converges and we conclude that

$$\lim_{k \mapsto \infty} (J(u_k) - J(u_{k+1})) = 0,$$

which combined with the preceding inequality shows that

$$\lim_{k \mapsto \infty} \|u_k - u_{k+1}\| = 0.$$

Step 3. $\|\nabla J_{u_k}\| \leq \left\|\nabla J_{u_k} - \nabla J_{u_{k+1}}\right\|$.

Using the orthogonality of consecutive descent directions, by Cauchy–Schwarz we have

$$\begin{aligned}\|\nabla J_{u_k}\|^2 &= \langle \nabla J_{u_k}, \nabla J_{u_k} - \nabla J_{u_{k+1}} \rangle \\ &\leq \|\nabla J_{u_k}\| \left\|\nabla J_{u_k} - \nabla J_{u_{k+1}}\right\|,\end{aligned}$$

so that

$$\|\nabla J_{u_k}\| \leq \left\|\nabla J_{u_k} - \nabla J_{u_{k+1}}\right\|.$$

Step 4. $\lim_{k\mapsto\infty} \|\nabla J_{u_k}\| = 0$.

Since the sequence $(J(u_k))_{k\geq 0}$ is decreasing and the functional J is coercive, the sequence $(u_k)_{k\geq 0}$ must be bounded. By hypothesis, the derivative dJ is of J is continuous, so it is uniformly continuous over compact subsets of $\mathbb{R}^n$; here, we are using the fact that $\mathbb{R}^n$ is finite dimensional. Hence, we deduce that for every $\epsilon > 0$, if $\|u_k - u_{k+1}\| < \epsilon$ then

$$\left\|dJ_{u_k} - dJ_{u_{k+1}}\right\|_2 < \epsilon.$$

But by definition of the operator norm and using the Cauchy–Schwarz inequality

$$\begin{aligned}\left\|dJ_{u_k} - dJ_{u_{k+1}}\right\|_2 &= \sup_{\|w\|\leq 1} |dJ_{u_k}(w) - dJ_{u_{k+1}}(w)| \\ &= \sup_{\|w\|\leq 1} |\langle \nabla J_{u_k} - \nabla J_{u_{k+1}}, w\rangle| \\ &\leq \left\|\nabla J_{u_k} - \nabla J_{u_{k+1}}\right\|.\end{aligned}$$

But we also have

$$\begin{aligned}\left\|\nabla J_{u_k} - \nabla J_{u_{k+1}}\right\|^2 &= \langle \nabla J_{u_k} - \nabla J_{u_{k+1}}, \nabla J_{u_k} - \nabla J_{u_{k+1}}\rangle \\ &= dJ_{u_k}(\nabla J_{u_k} - \nabla J_{u_{k+1}}) - dJ_{u_{k+1}}(\nabla J_{u_k} - \nabla J_{u_{k+1}}) \\ &\leq \left\|dJ_{u_k} - dJ_{u_{k+1}}\right\|_2^2,\end{aligned}$$

and so

$$\left\|dJ_{u_k} - dJ_{u_{k+1}}\right\|_2 = \left\|\nabla J_{u_k} - \nabla J_{u_{k+1}}\right\|.$$

It follows that if

$$\lim_{k\mapsto\infty} \|u_k - u_{k+1}\| = 0$$

then

$$\lim_{k\mapsto\infty} \left\|\nabla J_{u_k} - \nabla J_{u_{k+1}}\right\| = \lim_{k\mapsto\infty} \left\|dJ_{u_k} - dJ_{u_{k+1}}\right\|_2 = 0,$$

and using the fact that

$$\|\nabla J_{u_k}\| \leq \left\|\nabla J_{u_k} - \nabla J_{u_{k+1}}\right\|,$$

we obtain

$$\lim_{k\mapsto\infty} \|\nabla J_{u_k}\| = 0.$$

Step 5. Finally we can prove the convergence of the sequence $(u_k)_{k\geq 0}$.

Since J is elliptic and since $\nabla J_u = 0$ (since u is the minimum of J over $\mathbb{R}^n$), we have

$$\begin{aligned}\alpha \|u_k - u\|^2 &\leq \langle \nabla J_{u_k} - \nabla J_u, u_k - u\rangle \\ &= \langle \nabla J_{u_k}, u_k - u\rangle \\ &\leq \|\nabla J_{u_k}\| \, \|u_k - u\|.\end{aligned}$$

Hence, we obtain

$$\|u_k - u\| \leq \frac{1}{\alpha} \|\nabla J_{u_k}\|,$$

and since we showed that

$$\lim_{k \mapsto \infty} \|\nabla J_{u_k}\| = 0,$$

we see that the sequence $(u_k)_{k \geq 0}$ converges to the mininum u. □

Remarks: As with the previous proposition, the assumption of finite dimensionality is crucial. The proof provides an *a priori* bound on the error $\|u_k - u\|$.

If J is a an elliptic quadratic functional

$$J(v) = \frac{1}{2}\langle Av, v\rangle - \langle b, v\rangle,$$

we can use the orthogonality of the descent directions ∇J_{u_k} and $\nabla J_{u_{k+1}}$ to compute ρ_k. Indeed, we have $\nabla J_v = Av - b$, so

$$0 = \langle \nabla J_{u_{k+1}}, \nabla J_{u_k}\rangle = \langle A(u_k - \rho_k(Au_k - b)) - b, Au_k - b\rangle,$$

which yields

$$\rho_k = \frac{\|w_k\|^2}{\langle Aw_k, w_k\rangle}, \quad \text{with} \quad w_k = Au_k - b = \nabla J_{u_k}.$$

Consequently, a step of the iteration method takes the following form:

(1) Compute the vector

$$w_k = Au_k - b.$$

(2) Compute the scalar

$$\rho_k = \frac{\|w_k\|^2}{\langle Aw_k, w_k\rangle}.$$

(3) Compute the next vector u_{k+1} by

$$u_{k+1} = u_k - \rho_k w_k.$$

This method is of particular interest when the computation of Aw for a given vector w is cheap, which is the case if A is sparse.

For a particular illustration of this method, we turn to the example provided by Shewchuk, with $A = \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix}$ and $b = \begin{pmatrix} 2 \\ -8 \end{pmatrix}$, namely

$$\begin{aligned} J(x, y) &= \frac{1}{2} \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} 2 & -8 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \\ &= \frac{3}{2}x^2 + 2xy + 3y^2 - 2x + 8y. \end{aligned}$$

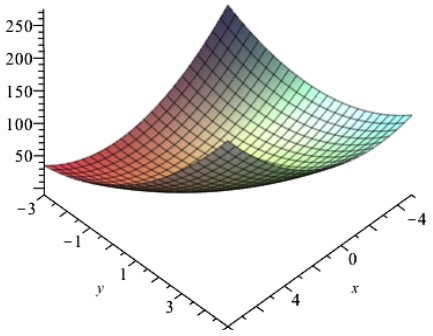


Figure 29.2: The ellipsoid $J(x, y) = \frac{3}{2}x^2 + 2xy + 3y^2 - 2x + 8y$.

This quadratic ellipsoid, which is illustrated in Figure 29.2, has a unique minimum at $(2, -2)$. In order to find this minimum via the gradient descent with optimal step size parameter, we pick a starting point, say $u_k = (-2, -2)$, and calculate the search direction $w_k = \nabla J(-2, -2) = (-12, -8)$. Note that

$$\nabla J(x, y) = (3x + 2y - 2, 2x + 6y + 8) = \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} 2 \\ -8 \end{pmatrix}$$

is perpendicular to the appropriate elliptical level curve; see Figure 29.3. We next perform

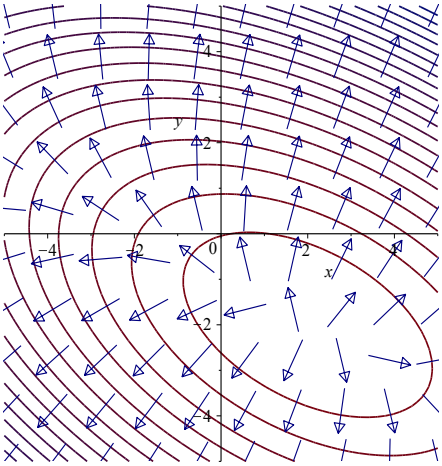


Figure 29.3: The level curves of $J(x, y) = \frac{3}{2}x^2 + 2xy + 3y^2 - 2x + 8y$ and the associated gradient vector field $\nabla J(x, y) = (3x + 2y - 2, 2x + 6y + 8)$.

the line search along the line given by the equation $-8x + 12y = -8$ and determine ρ_k. See Figures 29.4 and 29.5. In particular, we find that

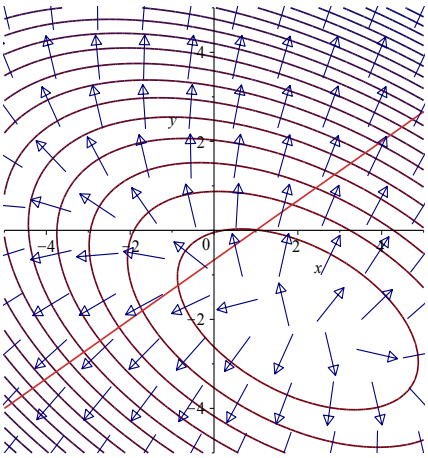


Figure 29.4: The level curves of $J(x,y) = \frac{3}{2}x^2 + 2xy + 3y^2 - 2x + 8y$ and the red search line with direction $\nabla J(-2,-2) = (-12,-8)$

$$\rho_k = \frac{\|w_k\|^2}{\langle Aw_k, w_k\rangle} = \frac{13}{75}.$$

This in turn gives us the new point

$$u_{k+1} = u_k - \frac{13}{75}w_k = (-2,-2) - \frac{13}{75}(-12,-8) = \left(\frac{2}{25}, -\frac{46}{75}\right),$$

and we continue the procedure by searching along the gradient direction $\nabla J(2/25, -46/75) = (-224/75, 112/25)$. Observe that $u_{k+1} = (\frac{2}{25}, -\frac{46}{75})$ has a gradient vector which is perpendicular to the search line with direction vector $w_k = \nabla J(-2,-2) = (-12\ -8)$; see Figure 29.5. Geometrically this procedure corresponds to intersecting the plane $-8x + 12y = -8$ with the ellipsoid $J(x,y) = \frac{3}{2}x^2 + 2xy + 3y^2 - 2x + 8y$ to form the parabolic curve $f(x) = 25/6x^2 - 2/3x - 4$ and then locating the x-coordinate of its apex which occurs when $f'(x) = 0$, i.e when $x = 2/25$; see Figure 29.6. After 31 iterations, this procedure stabilizes to point $(2,-2)$, which as we know, is the unique minimum of the quadratic ellipsoid $J(x,y) = \frac{3}{2}x^2 + 2xy + 3y^2 - 2x + 8y$.

We now give a sufficient condition for the gradient method with variable stepsize parameter to converge. In addition to requiring J to be an elliptic functional, we add a Lipschitz condition on the gradient of J. This time, the space V can be infinite dimensional.

Proposition 29.10. *Let $J\colon V \to \mathbb{R}$ be a continuously differentiable functional defined on a Hilbert space V. Suppose there exists two constants $\alpha > 0$ and $M > 0$ such that*

$$\langle \nabla J_v - \nabla J_u, v - u\rangle \geq \alpha \|v-u\|^2 \quad \textit{for all } u, v \in V,$$

and

$$\|\nabla J_v - \nabla J_u\| \leq M\|v-u\| \quad \textit{for all } u, v \in V.$$

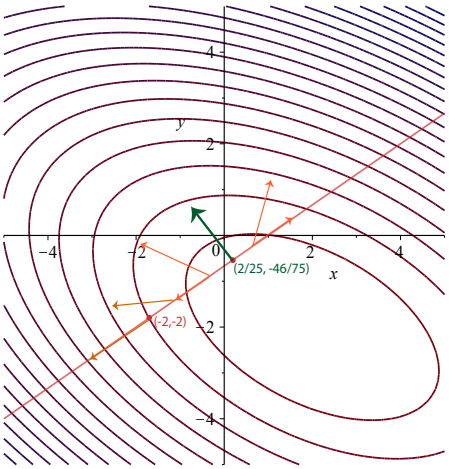


Figure 29.5: Let $u_k = (-2, -2)$. When traversing along the red search line, we look for the green perpendicular gradient vector. This gradient vector, which occurs at $u_{k+1} = (2/25, -46/75)$, provides a minimal ρ_k, since it has no nonzero projection on the search line.

If there exists two real numbers $a, b \in \mathbb{R}$ such that

$$0 < a \leq \rho_k \leq b \leq \frac{2\alpha}{M^2} \quad \textit{for all } k \geq 0,$$

then the gradient method with variable stepsize parameter converges. Furthermore, there is some constant $\beta > 0$ (depending on α, M, a, b) such that

$$\beta < 1 \quad \textit{and} \quad \|u_k - u\| \leq \beta^k \|u_0 - u\|,$$

where $u \in M$ is the unique minimum of J.

Proof. By hypothesis the functional J is elliptic, so by Theorem 29.7 it has a unique minimum u characterized by the fact that $\nabla J_u = 0$. Then since $u_{k+1} = u_k - \rho_k \nabla J_{u_k}$ we can write

$$u_{k+1} - u = (u_k - u) - \rho_k \langle \nabla J_{u_k} - \nabla J_u \rangle.$$

Using the inequalities

$$\langle \nabla J_{u_k} - \nabla J_u, u_k - u \rangle \geq \alpha \|u_k - u\|^2$$

and

$$\|\nabla J_{u_k} - \nabla J_u\| \leq M \|u_k - u\|,$$

and assuming that $\rho_k > 0$, it follows that

$$\begin{aligned} \|u_{k+1} - u\|^2 &= \|u_k - u\|^2 - 2\rho_k \langle \nabla J_{u_k} - \nabla J_u, u_k - u \rangle + \rho_k^2 \|\nabla J_{u_k} - \nabla J_u\|^2 \\ &\leq \left(1 - 2\alpha\rho_k + M^2\rho_k^2\right) \|u_k - u\|^2. \end{aligned}$$

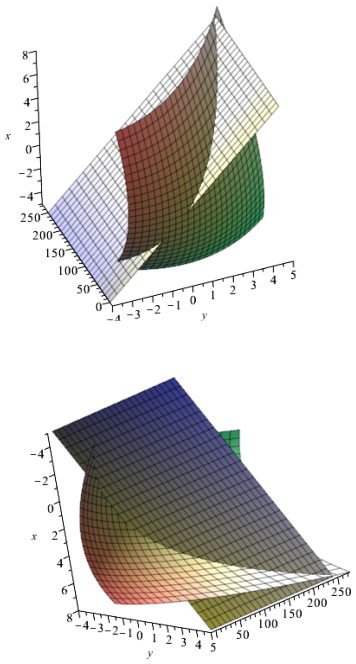


Figure 29.6: Two views of the intersection between the plane $-8x + 12y = -8$ and the ellipsoid $J(x, y) = \frac{3}{2}x^2 + 2xy + 3y^2 - 2x + 8y$. The point u_{k+1} is the minimum of the parabolic intersection.

Consider the function

$$T(\rho) = M^2\rho^2 - 2\alpha\rho + 1.$$

Its graph is a parabola intersecting the y-axis at $y = 1$ for $\rho = 0$, it has a minimum for $\rho = \alpha/M^2$, and it also has the value $y = 1$ for $\rho = 2\alpha/M^2$; see Figure 29.7. Therefore if we pick a, b and ρ_k such that

$$0 < a \leq \rho_k \leq b < \frac{2\alpha}{M^2},$$

we ensure that for $\rho \in [a, b]$ we have

$$T(\rho)^{1/2} = (M^2\rho^2 - 2\alpha\rho + 1)^{1/2} \leq (\max\{T(a), T(b)\})^{1/2} = \beta < 1.$$

Then by induction we get

$$\|u_{k+1} - u\| \leq \beta^{k+1} \|u_0 - u\|,$$

which proves convergence. □

Remarks: In the proof of Proposition 29.10, it is the fact that V is complete which plays a crucial role. If J is twice differentiable, the hypothesis

$$\|\nabla J_v - \nabla J_u\| \leq M \|v - u\| \quad \text{for all } u, v \in V$$

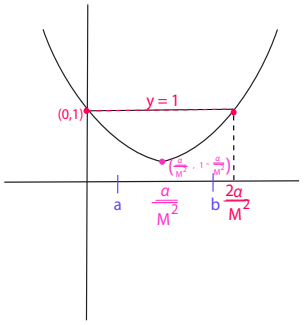


Figure 29.7: The parabola $T(\rho)$ used in the proof of Proposition 29.10.

can be expressed as

$$\sup_{v \in V} \left\| \nabla^2 J_v \right\| \leq M.$$

In the case of a quadratic elliptic functional defined over $\mathbb{R}^n$,

$$J(v) = \langle Av, v \rangle - \langle b, v \rangle,$$

the upper bound $2\alpha/M^2$ can be improved. In this case we have

$$\nabla J_v = Av - b,$$

and we know that we $\alpha = \lambda_1$ and $M = \lambda_n$ do the job, where λ_1 is the eigenvalue of A and λ_n is the largest eigenvalue of A. Hence we can pick a, b such that

$$0 < a \leq \rho_k \leq b < \frac{2\lambda_1}{\lambda_n^2}.$$

Since $u_{k+1} = u_k - \rho_k \nabla J_{u_k}$ and $\nabla J_{u_k} = Au_k - b$, we have

$$u_{k+1} - u = (u_k - u) - \rho_k(Au_k - u) = (I - \rho_k A)(u_k - u),$$

so we get

$$\|u_{k+1} - u\| \leq \|I - \rho_k A\|_2 \|u_k - u\|.$$

However, since $I - \rho_k A$ is a symmetric matrix, $\|I - \rho_k A\|_2$ is the largest absolute value of its eigenvalues, so

$$\|I - \rho_k A\|_2 \leq \max\{|1 - \rho_k \lambda_1|, |1 - \rho_k \lambda_n|\}.$$

The function

$$\mu(\rho) = \max\{|1 - \rho\lambda_1|, |1 - \rho\lambda_n|\}$$

is a piecewise affine function, and it is easy to see that if we pick a, b such that

$$0 < a \leq \rho_k \leq b \leq \frac{2}{\lambda_n},$$

then

$$\max_{\rho\in[a,b]} \mu(\rho) \leq \max\{\mu(a), \mu(b)\} < 1.$$

Therefore, the upper bound $2\lambda_1/\lambda_n^2$ can be replaced by $2/\lambda_n$, which is typically much larger. A "good" pick for ρ_k is $2/(\lambda_1 + \lambda_n)$ (as opposed to λ_1/λ_n^2 for the first version). In this case

$$|1 - \rho_k\lambda_1| = |1 - \rho_k\lambda_n| = \frac{\lambda_m - \lambda_1}{\lambda_m + \lambda_1},$$

so we get

$$\beta = \frac{\lambda_m - \lambda_1}{\lambda_m + \lambda_1} = \frac{\frac{\lambda_m}{\lambda_1} - 1}{\frac{\lambda_m}{\lambda_1} + 1} = \frac{\mathrm{cond}_2(A) - 1}{\mathrm{cond}_2(A) + 1},$$

where $\mathrm{cond}_2(A) = \lambda_m/\lambda_1$ is the condition number of the matrix A with respect to the spectral norm. Thus we see that the largest the condition number of A is, the slowest the convergence of the method will be. This is not surprising since we already know that linear systems involving ill-conditioned matrices (matrices with a large condition number) are problematic, and prone to numerical unstability. One way to deal with this problem is to use a method known as preconditioning.

We only described the most basic gradient descent methods. There are numerous variants, and we only mention a few of these methods.

The method of *scaling* consists in using $-\rho_k D_k \nabla J_{u_k}$ as descent direction, where D_k is some suitably chosen symmetric positive definite matrix.

In the *gradient method with extrapolation*, u_{k+1} is determined by

$$u_{k+1} = u_k - \rho_k \nabla J_{u_k} + \beta_k(u_k - u_{k-1}).$$

Another rule for choosing the stepsize is *Armijo's rule*.

These methods, and others, are discussed in detail in Berstekas [14]. Boyd and Vandenberghe discuss steepest descent methods for various types of norms besides the Euclidean norm; see Boyd and Vandenberghe [22] (Section 9.4).

Lax also discusses other methods in which the step ρ_k is chosen using roots of Chebyshev polynomials; see Lax [66], Chapter 17, Sections 2–4.

Contrary to intuition, the descent direction $d_k = -\nabla J_{u_k}$ given by the opposite of the gradient is not optimal. In the next section, we will see how a better direction can be picked; this is the method of *conjugate gradients*.

29.3 Conjugate Gradient Methods for Unconstrained Problems

The conjugate gradient method due to Hestenes and Stiefel (1952) is a gradient descent method that applies to an elliptic quadratic functional $J\colon \mathbb{R}^n \to \mathbb{R}$ given by

$$J(v) = \frac{1}{2}\langle Av, v\rangle - \langle b, v\rangle,$$

where A is an $n \times n$ symmetric positive definite matrix. Although it is presented as an iterative method, it terminates in at most n steps.

As usual, the conjugate gradient method starts with some arbitrary initial vector u_0 and proceeds through a sequence of iteration steps generating (better and better) approximations u_k of the optimal vector u minimizing J. During an iteration step, two vectors need to be determined:

(1) The descent direction d_k.

(2) The next approximation u_{k+1}. To find u_{k+1}, we need to find the stepsize $\rho_k > 0$ and then

$$u_{k+1} = u_k - \rho_k d_k.$$

Typically, ρ_k is found by performing a line search along the direction d_k, namely we find ρ_k as the real number such that the function $\rho \mapsto J(u_k - \rho d_k)$ is minimized.

We saw in Proposition 29.9 that during execution of the gradient method with optimal stepsize parameter that any two consecutive descent directions are orthogonal. The new twist with the conjugate gradient method is that given $u_0, u_1, \ldots, u_k$, the next approximation u_{k+1} is obtained as the solution of the problem which consists in minimizing J over the affine subspace $u_k + \mathcal{G}_k$, where $\mathcal{G}_k$ is the subspace of $\mathbb{R}^n$ spanned by the gradients

$$\nabla J_{u_0}, \nabla J_{u_1}, \ldots, \nabla J_{u_k}.$$

We may assume that $\nabla J_{u_\ell} \neq 0$ for $\ell = 0, \ldots, k$, since the method terminates as soon as $\nabla J_{u_k} = 0$. A priori the subspace $\mathcal{G}_k$ has dimension $\leq k+1$, but we will see that in fact it has dimension $k+1$. Then we have

$$u_k + \mathcal{G}_k = \left\{ u_k + \sum_{i=0}^{k} \alpha_i \nabla J_{u_i} \;\middle|\; \alpha_i \in \mathbb{R},\ 0 \leq i \leq k \right\},$$

and our minimization problem is to find u_{k+1} such that

$$u_{k+1} \in u_k + \mathcal{G}_k \quad \text{and} \quad J(u_{k+1}) = \inf_{v \in u_k + \mathcal{G}_k} J(v).$$

In the gradient method with optimal stepsize parameter the descent direction d_k is proportional to the gradient ∇J_{u_k}, but in the conjugate gradient method, d_k is equal to ∇J_{u_k} corrected by some multiple of d_{k-1}.

The conjugate gradient method is superior to the gradient method with optimal stepsize parameter for the following reasons proved correct later:

(a) The gradients ∇J_{u_i} and ∇J_{u_j} are orthogonal for all i, j with $0 \leq i < j \leq k$. This implies that if $\nabla J_{u_i} \neq 0$ for $i = 0, \ldots, k$, then the vectors ∇J_{u_i} are linearly independent, so the method stops in at most n steps.

(b) If we write $\Delta_\ell = u_{\ell+1} - u_\ell = -\rho_\ell d_\ell$, the second remarkable fact about the conjugate gradient method is that the vectors Δ_ℓ satisfy the following conditions:

$$\langle A\Delta_\ell, \Delta_i \rangle = 0 \quad 0 \leq i < \ell \leq k.$$

The vectors Δ_ℓ and Δ_i are said to be *conjugate* with respect to the matrix A (or A-*conjugate*). As a consequence, if $\Delta_\ell \neq 0$ for $\ell = 0, \ldots, k$, then the vectors Δ_ℓ are linearly independent.

(c) There is a simple formula to compute d_{k+1} from d_k, and to compute ρ_k.

We now prove the above facts. We begin with (a).

Proposition 29.11. *Assume that $\nabla J_{u_i} \neq 0$ for $i = 0, \ldots, k$. Then the minimization problem, find u_{k+1} such that*

$$u_{k+1} \in u_k + \mathcal{G}_k \quad \textit{and} \quad J(u_{k+1}) = \inf_{v \in u_k + \mathcal{G}_k} J(v),$$

has a unique solution, and the gradients ∇J_{u_i} and ∇J_{u_j} are orthogonal for all i, j with $0 \leq i < j \leq k$.

Proof. The affine space $u_\ell + \mathcal{G}_\ell$ is closed and convex, and since J is a quadratic elliptic functional it is coercise and strictly convex, so by Theorem 29.7(2) it has a unique minimum in $u_\ell + \mathcal{G}_\ell$. This minimum $u_{\ell+1}$ is also the minimum of the problem, find $u_{\ell+1}$ such that

$$u_{\ell+1} \in u_\ell + \mathcal{G}_\ell \quad \text{and} \quad J(u_{\ell+1}) = \inf_{v \in \mathcal{G}_\ell} J(u_\ell + v),$$

and since $\mathcal{G}_\ell$ is a vector space, by Theorem 20.8 we must have

$$dJ_{u_\ell}(w) = 0 \quad \text{for all } w \in \mathcal{G}_\ell,$$

that is

$$\langle \nabla J_{u_\ell}, w \rangle = 0 \quad \text{for all } w \in \mathcal{G}_\ell.$$

Since $\mathcal{G}_\ell$ is spanned by $(\nabla J_{u_0}, \nabla J_{u_1}, \ldots, \nabla J_{u_\ell})$, we obtain

$$\langle \nabla J_{u_\ell}, \nabla J_{u_j} \rangle = 0, \quad 0 \leq j < \ell,$$

and since this holds for $\ell = 0, \ldots, k$, we get

$$\langle \nabla J_{u_i}, \nabla J_{u_j} \rangle = 0, \quad 0 \leq i < j \leq k,$$

which shows the second part of the proposition. □

As a corollary of Proposition 29.11, if $\nabla J_{u_i} \neq 0$ for $i = 0, \ldots, k$, then the vectors ∇J_{u_i} are linearly independent and $\mathcal{G}_k$ has dimension $k+1$. Therefore, the conjugate gradient method terminates in at most n steps. Here is an example of a problem for which the gradient descent with optimal stepsize parameter does not converge in a finite number of steps.

Example 29.1. Let $J\colon \mathbb{R}^2 \to \mathbb{R}$ be the function given by

$$J(v_1, v_2) = \frac{1}{2}(\alpha_1 v_1^2 + \alpha_2 v_2^2),$$

where $0 < \alpha_1 < \alpha_2$. The minimum of J is attained at $(0, 0)$. Unless the initial vector $u_0 = (u_1^0, u_2^0)$ has the property that either $u_1^0 = 0$ or $u_2^0 = 0$, we claim that the gradient descent with optimal stepsize parameter does not converge in a finite number of steps. Observe that

$$\nabla J_{(v_1, v_2)} = \begin{pmatrix} \alpha_1 v_1 \\ \alpha_2 v_2 \end{pmatrix}.$$

As a consequence, given u_k, the line search for finding ρ_k and u_{k+1} yields $u_{k+1} = (0, 0)$ iff there is some $\rho \in \mathbb{R}$ such that

$$u_1^k = \rho \alpha_1 u_1^k \quad \text{and} \quad u_2^k = \rho \alpha_2 u_2^k.$$

Since $\alpha_1 \neq \alpha_2$, this is only possible if either $u_1^k = 0$ or $u_2^k = 0$. The formulae given just before Proposition 29.10 yield

$$u_1^{k+1} = \frac{\alpha_2^2(\alpha_2 - \alpha_1) u_1^k (u_2^k)^2}{\alpha_1^3 (u_1^k)^2 + \alpha_2^3 (u_2^k)^2}, \quad u_2^{k+1} = \frac{\alpha_1^2(\alpha_1 - \alpha_2) u_2^k (u_1^k)^2}{\alpha_1^3 (u_1^k)^2 + \alpha_2^3 (u_2^k)^2},$$

which implies that if $u_1^k \neq 0$ and $u_2^k \neq 0$, then $u_1^{k+1} \neq 0$ and $u_2^{k+1} \neq 0$, so the method runs forever from any initial vector $u_0 = (u_1^0, u_2^0)$ such that $u_1^0 \neq 0$ and, $u_2^0 \neq 0$.

We now prove (b).

Proposition 29.12. *Assume that $\nabla J_{u_i} \neq 0$ for $i = 0, \ldots, k$, and let $\Delta_\ell = u_{\ell+1} - u_\ell$, for $\ell = 0, \ldots, k$. Then $\Delta_\ell \neq 0$ for $\ell = 0, \ldots, k$, and*

$$\langle A\Delta_\ell, \Delta_i \rangle = 0, \quad 0 \leq i < \ell \leq k.$$

The vectors $\Delta_0, \ldots, \Delta_k$ are linearly independent.

Proof. Since J is a quadratic functional we have

$$\nabla J_{v+w} = A(v+w) - b = Av - b + Aw = \nabla J_v + Aw.$$

It follows that

$$\nabla J_{u_{\ell+1}} = \nabla J_{u_\ell + \Delta_\ell} = \nabla J_{u_\ell} + A\Delta_\ell, \quad 0 \leq \ell \leq k. \tag{$*_1$}$$

By Proposition 29.11, since

$$\langle \nabla J_{u_i}, \nabla J_{u_j} \rangle = 0, \quad 0 \leq i < j \leq k,$$

we get

$$0 = \langle \nabla J_{u_\ell+1}, \nabla J_{u_\ell} \rangle = \|\nabla J_{u_\ell}\|^2 + \langle A\Delta_\ell, \nabla J_{u_\ell} \rangle, \quad \ell = 0, \ldots, k,$$

and since by hypothesis $\nabla J_{u_i} \neq 0$ for $i = 0, \ldots, k$, we deduce that

$$\Delta_\ell \neq 0, \quad \ell = 0, \ldots, k.$$

If $k \geq 1$, for $i = 0, \ldots, \ell - 1$ and $\ell \leq k$ we also have

$$\begin{aligned} 0 = \langle \nabla J_{u_{\ell+1}}, \nabla J_{u_i} \rangle &= \langle \nabla J_{u_\ell}, \nabla J_{u_i} \rangle + \langle A\Delta_\ell, \nabla J_{u_i} \rangle \\ &= \langle A\Delta_\ell, \nabla J_{u_i} \rangle. \end{aligned}$$

Since $\Delta_j = u_{j+1} - u_j \in \mathcal{G}_j$ and $\mathcal{G}_j$ is spanned by $(\nabla J_{u_0}, \nabla J_{u_1}, \ldots, \nabla J_{u_j})$, we obtain

$$\langle A\Delta_\ell, \Delta_j \rangle = 0, \quad 0 \leq j < \ell \leq k.$$

For the last statement of the proposition, let $w_0, w_1, \ldots, w_k$ be any $k+1$ nonzero vectors such that

$$\langle Aw_i, w_j \rangle = 0, \quad 0 \leq i < j \leq k.$$

We claim that $w_0, w_1, \ldots, w_k$ are linearly independent.

If we have a linear dependence $\sum_{i=0}^k \lambda_i w_i = 0$, then we have

$$0 = \left\langle A\Big(\sum_{i=0}^k \lambda_i w_i\Big), w_j \right\rangle = \sum_{i=0}^k \lambda_i \langle Aw_i, w_j \rangle = \lambda_j \langle Aw_j, w_j \rangle.$$

Since A is symmetric positive definite (because J is a quadratic elliptic functional) and $w_j \neq 0$, we must have $\lambda_j = 0$ for $j = 0, \ldots, k$. Therefore the vectors $w_0, w_1, \ldots, w_k$ are linearly independent. □

Remarks:

(1) Since A is symmetric positive definite, the bilinear map $(u, v) \mapsto \langle Au, v \rangle$ is an inner product $\langle -, - \rangle_A$ on $\mathbb{R}^n$. Consequently, two vectors u, v are *conjugate* with respect to the matrix A (or A-*conjugate*), which means that $\langle Au, v \rangle = 0$, iff u and v are orthogonal with respect to the inner product $\langle -, - \rangle_A$.

(2) By picking the descent direction to be $-\nabla J_{u_k}$, the gradient descent method with optimal stepsize parameter treats the level sets $\{u \mid J(u) = J(u_k)\}$ as if they were spheres. The conjugate gradient method is more subtle, and takes the "geometry" of the level set $\{u \mid J(u) = J(u_k)\}$ into account, through the notion of conjugate directions.

(3) The notion of conjugate direction has its origins in the theory of projective conics and quadrics where A is a 2×2 or a 3×3 matrix and where u and v are conjugate iff $u^\top Av = 0$.

(4) The terminology conjugate gradient is somewhat misleading. It is not the gradients who are conjugate directions, but the descent directions.

By definition of the vectors $\Delta_\ell = u_{\ell+1} - u_\ell$, we can write

$$\Delta_\ell = \sum_{i=0}^{\ell} \delta_i^\ell \nabla J_{u_i}, \quad 0 \leq \ell \leq k. \tag{$*_2$}$$

In matrix form, we can write

$$\begin{pmatrix} \Delta_0 & \Delta_1 & \cdots & \Delta_k \end{pmatrix} = \begin{pmatrix} \nabla J_{u_0} & \nabla J_{u_1} & \cdots & \nabla J_{u_k} \end{pmatrix} \begin{pmatrix} \delta_0^0 & \delta_0^1 & \cdots & \delta_0^{k-1} & \delta_0^k \\ 0 & \delta_1^1 & \cdots & \delta_1^{k-1} & \delta_1^k \\ 0 & 0 & \cdots & \delta_2^{k-1} & \delta_2^k \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \delta_k^k \end{pmatrix},$$

which implies that $\delta_\ell^\ell \neq 0$ for $\ell = 0, \ldots, k$.

In view of the above fact, since Δ_ℓ and d_ℓ are collinear, it is convenient to write the descent direction d_ℓ as

$$d_\ell = \sum_{i=0}^{\ell-1} \lambda_i^\ell \nabla J_{u_i} + \nabla J_{u_\ell}, \quad 0 \leq \ell \leq k. \tag{$*_3$}$$

Our next goal is to compute u_{k+1}, assuming that the coefficients λ_i^k are known for $i = 0, \ldots, k$, and then to find simple formulae for the λ_i^k.

The problem reduces to finding ρ_k such that

$$J(u_k - \rho_k d_k) = \inf_{\rho \in \mathbb{R}} J(u_k - \rho d_k),$$

and then $u_{k+1} = u_k - \rho_k d_k$. In fact, by $(*_2)$, since

$$\Delta_k = \sum_{i=0}^{k} \delta_i^k \nabla J_{u_i} = \delta_k^k \left(\sum_{i=0}^{k-1} \frac{\delta_i^k}{\delta_k^k} \nabla J_{u_i} + \nabla J_{u_k} \right),$$

we must have

$$\Delta_k = \delta_k^k d_k \quad \text{and} \quad \rho_k = -\delta_k^k. \tag{$*_4$}$$

Remarkably, the coefficients λ_i^k and the descent directions d_k can be computed easily using the following formulae.

Proposition 29.13. *Assume that $\nabla J_{u_i} \neq 0$ for $i = 0, \ldots, k$. If we write*

$$d_\ell = \sum_{i=0}^{\ell-1} \lambda_i^\ell \nabla J_{u_i} + \nabla J_{u_\ell}, \quad 0 \leq \ell \leq k,$$

then we have

$$(\dagger) \quad \begin{cases} \lambda_i^k = \dfrac{\|\nabla J_{u_k}\|^2}{\|\nabla J_{u_i}\|^2}, \quad 0 \leq i \leq k-1, \\ d_0 = \nabla J_{u_0} \\ d_\ell = \nabla J_{u_\ell} + \dfrac{\|\nabla J_{u_\ell}\|^2}{\|\nabla J_{u_{\ell-1}}\|^2}\, d_{\ell-1}, \quad 1 \leq \ell \leq k. \end{cases}$$

Proof. Since by $(*_4)$ we have $\Delta_k = \delta_k^k d_k$, $\delta_k^k \neq 0$, (by Proposition 29.12) we have

$$\langle A\Delta_\ell, \Delta_i \rangle = 0, \quad 0 \leq i < \ell \leq k,$$

by $(*_1)$ we have $\nabla J_{u_{\ell+1}} = \nabla J_{u_\ell} + A\Delta_\ell$, and A is a symmetric matrix, we have

$$0 = \langle Ad_k, \Delta_\ell \rangle = \langle d_k, A\Delta_\ell \rangle = \langle d_k, \nabla J_{u_{\ell+1}} - \nabla J_{u_\ell} \rangle,$$

for $\ell = 0, \ldots, k-1$, and since

$$d_k = \sum_{i=0}^{k-1} \lambda_i^k \nabla J_{u_i} + \nabla J_{u_k},$$

we have

$$\left\langle \sum_{i=0}^{k-1} \lambda_i^k \nabla J_{u_i} + \nabla J_{u_k}, \nabla J_{u_{\ell+1}} - \nabla J_{u_\ell} \right\rangle = 0, \quad 0 \leq \ell \leq k-1.$$

Since by Proposition 29.11 the gradients ∇J_{u_i} are pairwise orthogonal, the above equations yield

$$\begin{aligned} -\lambda_{k-1}^k \left\|\nabla J_{u_{k-1}}\right\|^2 + \|\nabla J_k\|^2 &= 0 \\ -\lambda_\ell^k \|\nabla J_{u_\ell}\|^2 + \lambda_{\ell+1}^k \|\nabla J_{\ell+1}\|^2 &= 0, \quad 0 \leq \ell \leq k-2, \quad k \geq 2, \end{aligned}$$

and an easy induction yields

$$\lambda_i^k = \frac{\|\nabla J_{u_k}\|^2}{\|\nabla J_{u_i}\|^2}, \quad 0 \leq i \leq k-1.$$

Consequently, using $(*_3)$ we have

$$\begin{aligned} d_k &= \sum_{i=0}^{k-1} \frac{\|\nabla J_{u_k}\|^2}{\|\nabla J_{u_i}\|^2} \nabla J_{u_i} + \nabla J_{u_k} \\ &= \nabla J_{u_k} + \frac{\|\nabla J_{u_k}\|^2}{\left\|\nabla J_{u_{k-1}}\right\|^2} \left(\sum_{i=0}^{k-2} \frac{\left\|\nabla J_{u_{k-1}}\right\|^2}{\|\nabla J_{u_i}\|^2} \nabla J_{u_i} + \nabla J_{u_{k-1}} \right) \\ &= \nabla J_{u_k} + \frac{\|\nabla J_{u_k}\|^2}{\left\|\nabla J_{u_{k-1}}\right\|^2} d_{k-1}, \end{aligned}$$

which concludes the proof. □

It remains to compute ρ_k, which is the solution of the line search

$$J(u_k - \rho_k d_k) = \inf_{\rho \in \mathbb{R}} J(u_k - \rho d_k).$$

Since J is a quadratic functional, the function to be minimized is

$$\rho \mapsto \frac{\rho^2}{2} \langle A d_k, d_k \rangle - \rho \langle \nabla J_{u_k}, d_k \rangle + J(u_k),$$

whose mininum is obtained when its derivative is zero, that is,

$$\rho_k = \frac{\langle \nabla J_{u_k}, d_k \rangle}{\langle A d_k, d_k \rangle}. \tag{$*_5$}$$

In summary, the conjugate gradient method finds the minimum u of the elliptic quadratic functional

$$J(v) = \frac{1}{2} \langle Av, a \rangle - \langle b, v \rangle$$

by computing the sequence of vectors $u_1, d_1, \ldots, u_{k-1}, d_{k-1}, u_k$, starting from any vector u_0, with

$$d_0 = \nabla J_{u_0}.$$

If $\nabla J_{u_0} = 0$, then the algorithm terminates with $u = u_0$. Otherwise, for $k \geq 0$, assuming that $\nabla J_{u_i} \neq 0$ for $i = 1, \ldots, k$, compute

$$(*_6) \quad \begin{cases} \rho_k = \dfrac{\langle \nabla J_{u_k}, d_k \rangle}{\langle A d_k, d_k \rangle} \\ u_{k+1} = u_k - \rho_k d_k \\ d_{k+1} = \nabla J_{u_{k+1}} + \dfrac{\left\|\nabla J_{u_{k+1}}\right\|^2}{\|\nabla J_{u_k}\|^2} d_k. \end{cases}$$

If $\nabla J_{u_{k+1}} = 0$, then the algorihm terminate with $u = u_{k+1}$.

As we showed before, the algorithm terminates in at most n iterations.

Hestenes and Stiefel realized that the equations $(*_6)$ can be modified to make the computations more efficient, by having only one evaluation of the matrix A on a vector, namely d_k. The idea is to compute ∇_{u_k} inductively.

Since by $(*_1)$ and $(*_4)$ we have $\nabla J_{u_{\ell+1}} = \nabla J_{u_\ell} + A\Delta_\ell = \nabla J_{u_\ell} - \rho_k A d_k$, the gradient $\nabla J_{u_{\ell+1}}$ can be computed iteratively:

$$\begin{aligned}\nabla J_0 &= Au_0 - b\\ \nabla J_{u_{\ell+1}} &= \nabla J_{u_\ell} - \rho_k A d_k.\end{aligned}$$

Since by Proposition 29.13 we have

$$d_k = \nabla J_{u_k} + \frac{\|\nabla J_{u_k}\|^2}{\left\|\nabla J_{u_{k-1}}\right\|^2}\, d_{k-1}$$

and since d_{k-1} is a linear combination of the gradients ∇J_{u_i} for $i = 0, \ldots, k-1$, which are all orthogonal to ∇J_{u_k}, we have

$$\rho_k = \frac{\langle \nabla J_{u_k}, d_k\rangle}{\langle Ad_k, d_k\rangle} = \frac{\|\nabla J_{u_k}\|^2}{\langle Ad_k, d_k\rangle}.$$

It is customary to introduce the term r_k defined as

$$\nabla J_{u_k} = Au_k - b \tag{$*_7$}$$

and to call it the *residual*. Then the conjugate gradient method consists of the following steps. We inititialize the method starting from any vector u_0 and set

$$d_0 = r_0 = Au_0 - b.$$

The main iteration step is $(k \geq 0)$:

$$(*_8)\quad \begin{cases} \rho_k = \dfrac{\|r_k\|^2}{\langle Ad_k, d_k\rangle}\\ u_{k+1} = u_k - \rho_k d_k\\ r_{k+1} = r_k - \rho_k A d_k\\ \beta_{k+1} = \dfrac{\|r_{k+1}\|^2}{\|r_k\|^2}\\ d_{k+1} = r_{k+1} + \beta_{k+1}\, d_k. \end{cases}$$

Beware that some authors define the residual r_k as $r_k = b - Au_k$ and the descent direction d_k as $-d_k$. In this case, the second equation becomes

$$u_{k+1} = u_k + \rho_k d_k.$$

Since $d_0 = r_0$, the equations

$$r_{k+1} = r_k - \rho_k A d_k$$
$$d_{k+1} = r_{k+1} - \beta_{k+1} d_k$$

imply by induction that the subspace $\mathcal{G}_k$ spanned by $(r_0, r_1, \ldots, r_k)$ and $(d_0, d_1, \ldots, d_k)$ is the subspace spanned by

$$(r_0, Ar_0, A^2 r_0, \ldots, A^k r_0).$$

Such a subspace is called a *Krylov subspace*.

If we define the *error* e_k as $e_k = u_k - u$, then $e_0 = u_0 - u$ and $Ae_0 = Au_0 - Au = Au_0 - b = d_0 = r_0$, and then because

$$u_{k+1} = u_k - \rho_k d_k$$

we see that

$$e_{k+1} = e_k - \rho_k d_k.$$

Since d_k belongs to the subspace spanned by $(r_0, Ar_0, A^2 r_0, \ldots, A^k r_0)$ and $r_0 = Ae_0$, we see that d_k belongs to the subspace spanned by $(Ae_0, A^2 e_0, A^3 e_0, \ldots, A^{k+1} e_0)$, and then by induction we see that e_{k+1} belongs to the subspace spanned by $(e_0, Ae_0, A^2 e_0, A^3 e_0, \ldots, A^{k+1} e_0)$. This means that there is a polynomial P_k of degree $\leq k$ such that $P_k(0) = 1$ and

$$e_k = P_k(A) e_0.$$

This is an important fact because it allows an analysis of the convergence of the conjugate gradient method; see Trefethen and Bau [105] (Lecture 38). For this, since A is symmetric positive definite, we know that $\langle u, v\rangle_A = \langle Av, u\rangle$ is an inner product on $\mathbb{R}^n$ whose associated norm is denoted by $\|v\|_A$. Then observe that if $e(v) = v - u$, then

$$\begin{aligned}
\|e(v)\|_A^2 &= \langle Av - Au, v - u\rangle \\
&= \langle Av, v\rangle - 2\langle Au, v\rangle + \langle Au, u\rangle \\
&= \langle Av, v\rangle - 2\langle b, v\rangle + \langle b, u\rangle \\
&= 2J(v) + \langle b, u\rangle.
\end{aligned}$$

It follows that $v = u_k$ minimizes $\|e(v)\|_A$ on $u_{k-1} + \mathcal{G}_{k-1}$ since u_k minimizes J on $u_{k-1} + \mathcal{G}_{k-1}$. Since $e_k = P_k(A)e_0$ for some polynomial P_k of degree $\leq k$ such that $P_k(0) = 1$, if we let $\mathcal{P}_k$ be the set of polynomials $P(t)$ of degree $\leq k$ such that $P(0) = 1$, then we have

$$\|e_k\|_A = \inf_{P \in \mathcal{P}_k} \|P(A)e_0\|_A.$$

Since A is a symmetric positive definite matrix it has real positive eigenvalues $\lambda_1, \ldots, \lambda_n$ and there is an orthonomal basis of eigenvectors $h_1, \ldots, h_n$ so that if we write $e_0 = \sum_{j=1}^n a_j h_j$. then we have

$$\|e_0\|_A^2 = \langle Ae_0, e_0\rangle = \left\langle \sum_{i=1}^n a_i \lambda_i h_i, \sum_{j=1}^n a_j h_j \right\rangle = \sum_{j=1}^n a_j^2 \lambda_j$$

and

$$\|P(A)e_0\|_A^2 = \langle AP(A)e_0, P(A)e_0\rangle = \left\langle \sum_{i=1}^n a_i\lambda_i P(\lambda_i)h_i, \sum_{j=1}^n a_j P(\lambda_j)h_j \right\rangle = \sum_{j=1}^n a_j^2\lambda_j (P(\lambda_j))^2.$$

These equations imply that

$$\|e_k\|_A \leq \left(\inf_{P\in\mathcal{P}_k} \max_{1\leq i\leq n} |P(\lambda_i)|\right) \|e_0\|_A .$$

It can be shown that the conjugate gradient method requires of the order of

n^3 additions,

n^3 multiplications,

$2n$ divisions.

In theory, this is worse than the number of elementary operations required by the Cholesky method. Even though the conjugate gradient method does not seem to be the best method for *full* matrices, it usually outperforms other methods for *sparse* matrices. The reason is that the matrix A only appears in the computation of the vector Ad_k. If the matrix A is banded (for example, tridiagonal), computing Ad_k is very cheap and there is no need to store the entire matrix A, in which case the conjugate gradient method is fast. Also, although in theory, up to n iterations may be required, in practice, convergence may occur after a much smaller number of iterations.

Using the inequality

$$\|e_k\|_A \leq \left(\inf_{P\in\mathcal{P}_k} \max_{1\leq i\leq n} |P(\lambda_i)|\right) \|e_0\|_A ,$$

by choosing P to be shifted Chebyshev polynomial, it can be shown that

$$\|e_k\|_A \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k \|e_0\|_A ,$$

where $\kappa = \mathrm{cond}_2(A)$; see Trefethen and Bau [105] (Lecture 38, Theorem 38.5). Thus the rate of convergence of the conjugate gradient method is governed by the ratio

$$\frac{\sqrt{\mathrm{cond}_2(A)}-1}{\sqrt{\mathrm{cond}_2(A)}+1},$$

where $\mathrm{cond}_2(A) = \lambda_m/\lambda_1$ is the condition number of the matrix A. Since A is positive definite, λ_1 is its smallest eigenvalue and λ_m is its largest eigenvalue.

The above fact leads to the process of *preconditioning*, a method which consists in replacing the matrix of a linear system $Ax = b$ by an "equivalent" one for example $M^{-1}A$ (since

M is invertible, the system $Ax = b$ is equivalent to the system $M^{-1}Ax = M^{-1}b$), where M is chosen so that $M^{-1}A$ is still symmetric positive definite and has a smaller condition number than A; see Trefethen and Bau [105] (Lecture 40) and Demmel [33] (Section 6.6.5).

The method of conjugate gradients can be generalized to functionals that are not necessarily quadratic. The stepsize parameter ρ_k is still determined by a line search which consists in finding ρ_k such that

$$J(u_k - \rho_k d_k) = \inf_{\rho \in \mathbb{R}} J(u_k - \rho d_k).$$

This is more difficult than in the quadratic case and in general there is no guarantee that ρ_k is unique, so some criterion to pick ρ_k is needed. Then

$$u_{k+1} = u_k - \rho_k d_k,$$

and the next descent direction can be chosen in two ways:

(1) (*Polak–Ribière*)

$$d_k = \nabla J_{u_k} + \frac{\langle \nabla J_{u_k}, \nabla J_{u_k} - \nabla J_{u_{k-1}} \rangle}{\left\| \nabla J_{u_{k-1}} \right\|^2} d_{k-1},$$

(2) (*Fletcher–Reeves*)

$$d_k = \nabla J_{u_k} + \frac{\left\| \nabla J_{u_k} \right\|^2}{\left\| \nabla J_{u_{k-1}} \right\|^2} d_{k-1}.$$

Consecutive gradients are no longer orthogonal so these methods may run forever. There are various sufficient criteria for convergence. In practice, the Polak–Ribière method converges faster. There no longer any guarantee that these methods converge to a global minimum.

29.4 Gradient Projection Methods for Constrained Optimization

We now consider the problem of finding the minimum of a convex functional $J \colon V \to \mathbb{R}$ over a nonempty convex subset U of a Hilbert space V. By Theorem 20.11(3), the functional J has a minimum at $u \in U$ iff

$$dJ_u(v - u) \geq 0 \quad \text{for all } v \in U,$$

which can be expressed as

$$\langle \nabla J_u, v - u \rangle \geq 0 \quad \text{for all } v \in U.$$

On the other hand, by the projection lemma (Proposition 28.5), the condition for a vector $u \in U$ to be the projection of an element $w \in V$ onto U is

$$\langle u - w, v - u\rangle \geq 0 \quad \text{for all } v \in U.$$

These conditions are obviously analogous, and we can make this analogy more precise as follows. If $p_U \colon V \to U$ is the projection map onto U, we have the following chain of equivalences:

$$\begin{aligned}
&u \in U \quad \text{and} \quad J(u) = \inf_{v \in U} J(v) \quad \text{iff}\\
&u \in U \quad \text{and} \quad \langle \nabla J_u, v - u\rangle \geq 0 \quad \text{for every } v \in U, \text{ iff}\\
&u \in U \quad \text{and} \quad \langle u - (u - \rho \nabla J_u), v - u\rangle \geq 0 \quad \text{for every } v \in U \text{ and every } \rho > 0, \text{ iff}\\
&u = p_U(u - \rho \nabla J_u) \quad \text{for every } \rho > 0.
\end{aligned}$$

In other words, for every $\rho > 0$, $u \in V$ is a *fixed-point* of the function $g \colon V \to U$ given by

$$g(v) = p_U(v - \rho \nabla J_v).$$

The above suggests finding u by the method of successive approximations for finding the fixed-point of a contracting mapping, namely given any initial $u_0 \in V$, to define the sequence $(u_k)_{k \geq 0}$ such that

$$u_{k+1} = p_U(u_k - \rho_k \nabla J_{u_k}),$$

where the parameter $\rho_k > 0$ is chosen at each step. This method is called the *projected-gradient method with variable stepsize parameter*. Observe that if $U = V$, then this is just the gradient method with variable stepsize. We have the following result about the convergence of this method.

Proposition 29.14. *Let $J \colon V \to \mathbb{R}$ be a continuously differentiable functional defined on a Hilbert space V, and let U be nonempty, convex, closed subset of V. Suppose there exists two constants $\alpha > 0$ and $M > 0$ such that*

$$\langle \nabla J_v - \nabla J_u, v - u\rangle \geq \alpha \left\| v - u \right\|^2 \quad \textit{for all } u, v \in V,$$

and

$$\left\| \nabla J_v - \nabla J_u \right\| \leq M \left\| v - u \right\| \quad \textit{for all } u, v \in V.$$

If there exists two real nunbers $a, b \in \mathbb{R}$ such that

$$0 < a \leq \rho_k \leq b \leq \frac{2\alpha}{M^2} \quad \textit{for all } k \geq 0,$$

then the projected-gradient method with variable stepsize parameter converges. Furthermore, there is some constant $\beta > 0$ (depending on α, M, a, b) such that

$$\beta < 1 \quad \textit{and} \quad \left\| u_k - u \right\| \leq \beta^k \left\| u_0 - u \right\|,$$

where $u \in M$ is the unique minimum of J.

Proof. For every ≥ 0, define the function $g_k \colon V \to U$ by

$$g_k(v) = p_U(v - \rho_k \nabla J_v).$$

By Proposition 28.6, the projection map p_U has Lipschitz constant 1, so using the inequalities assumed to hold in the proposition, we have

$$\begin{aligned}
\|g_k(v_1) - g_k(v_2)\|^2 &= \|p_U(v_1 - \rho_k \nabla J_{v_1}) - p_U(v_2 - \rho_k \nabla J_{v_2})\|^2 \\
&\leq \|(v_1 - v_2) - \rho_k(\nabla J_{v_1} - \nabla J_{v_2})\|^2 \\
&= \|v_1 - v_2\|^2 - 2\rho_k \langle \nabla J_{v_1} - \nabla J_{v_2}, v_1 - v_2 \rangle + \rho_k^2 \|\nabla J_{v_1} - \nabla J_{v_2}\|^2 \\
&\leq \left(1 - 2\alpha\rho_k + M^2\rho_k^2\right) \|v_1 - v_2\|^2 .
\end{aligned}$$

As in the proof of Proposition 29.10, we know that if a and b satisfy the conditions $0 < a \leq \rho_k \leq b \leq \frac{2\alpha}{M^2}$, then there is some β such that

$$\left(1 - 2\alpha\rho_k + M^2\rho_k^2\right)^{1/2} \leq \beta < 1 \quad \text{for all } k \geq 0.$$

Since the minimizing point $u \in U$ is a fixed point of g_k for all k, by letting $v_1 = u_k$ and $v_2 = u$, we get

$$\|u_{k+1} - u\| = \|g_k(u_k) - g_k(u)\| \leq \beta \|u_k - u\| ,$$

which proves the convergence of the sequence $(u_k)_{k \geq 0}$. □

In the case of an elliptic quadratic functional

$$J(v) = \frac{1}{2}\langle Av, a\rangle - \langle b, v\rangle$$

defined on $\mathbb{R}^n$, the reasoning just after the proof of Proposition 29.10 can be immediately adapted to show that convergence takes pkace as long as a, b and ρ_k are chosen such that

$$0 < a \leq \rho_k \leq b \leq \frac{2}{\lambda_n}.$$

In theory, Proposition 29.14 gives a guarantee of the convergence of the projected-gradient method. Unfortunately, because computing the projection $p_U(v)$ effectively is generally impossible, the range of practical applications of Proposition 29.14 is rather limited. One exception is the case where U is a product $\prod_{i=1}^m [a_i, b_i]$ of closed intervals (where $a_i = -\infty$ or $b_i = +\infty$ is possible). In this case, it is not hard to show that

$$p_U(v)_i = \begin{cases} a_i & \text{if } w_i < a_i \\ w_i & \text{if } a_i \leq w_i \leq b_i \\ b_i & \text{if } b_i < w_i. \end{cases}$$

In particular, this is the case if

$$U = \mathbb{R}^n_+ = \{v \in \mathbb{R}^n \mid v \geq 0\}$$

and if

$$J(v) = \frac{1}{2}\langle Av, a\rangle - \langle b, v\rangle$$

is an elliptic quadratic functional on $\mathbb{R}^n$. Then the vector $u_{k+1} = (u_1^{k+1}, \ldots, u_n^{k+1})$ is given in terms of $u_k = (u_1^k, \ldots, u_n^k)$ by

$$u_i^{k+1} = \max\{u_i^k - \rho_k(Au_k - b)_i, 0\}, \quad 1 \leq i \leq n.$$

29.5 Penalty Methods for Constrained Optimization

In the case where $V = \mathbb{R}^n$, another method to deal with constrained optimization is to incorporate the domain U into the objective function J by adding a penalty function.

Definition 29.6. Given a nonempty closed convex subset U of $\mathbb{R}^n$, a function $\psi\colon \mathbb{R}^n \to \mathbb{R}$ is called a *penalty function* for U if ψ is convex and continuous and if the following conditions hold:

$$\psi(v) \geq 0 \quad \text{for all} \quad v \in \mathbb{R}^n, \quad \text{and} \quad \psi(v) = 0 \quad \text{iff} \quad v \in U.$$

The following proposition shows that the use of penalty functions reduces a constrained optimization problem to a sequence of unconstrained optimzation problems.

Proposition 29.15. *Let $J\colon \mathbb{R}^n \to \mathbb{R}$ be a continuous, coercive, strictly convex function, U be a nonempty, convex, closed subset of $\mathbb{R}^n$, $\psi\colon \mathbb{R}^n \to \mathbb{R}$ be a penalty function for U, and let $J_\epsilon\colon \mathbb{R}^n \to \mathbb{R}$ be the penalized objective function given by*

$$J_\epsilon(v) = J(v) + \frac{1}{\epsilon}\,\psi(v) \quad \textit{for all } v \in \mathbb{R}^n.$$

Then, for every $\epsilon > 0$, there exists a unique element $u_\epsilon \in \mathbb{R}^n$ such that

$$J_\epsilon(u_\epsilon) = \inf_{v \in \mathbb{R}^n} J_\epsilon(v).$$

Furthermore, if $u \in U$ is the unique minimizer of J over U, so that $J(u) = \inf_{v\in U} J(v)$, then

$$\lim_{\epsilon \mapsto 0} u_\epsilon = u.$$

Proof. Observe that since J is coercive, since $\psi(v) \geq 0$ for all $v \in \mathbb{R}^n$, and $J_\epsilon = J + (1/\epsilon)\psi$, we have $J_\epsilon(v) \geq J(v)$ for all $v \in \mathbb{R}^n$, so J_ϵ is also coercive. Since J is strictly convex and $(1/\epsilon)\psi$ is convex, it is immediately checked that $J_\epsilon = J + (1/\epsilon)\psi$ is also strictly convex. Then by Proposition 29.1 (and the fact that J and J_ϵ are strictly convex), J has a unique minimizer $u \in U$, and J_ϵ has a unique minimizer $u_\epsilon \in \mathbb{R}^n$.

Since $\psi(u) = 0$ iff $u \in U$, and $\psi(v) \geq 0$ for all $v \in \mathbb{R}^n$, we have $J_\epsilon(u) = J(u)$, and since u_ϵ is the minimizer of J_ϵ we have $J_\epsilon(u_\epsilon) \leq J_\epsilon(u)$, so we obtain

$$J(u_\epsilon) \leq J(u_\epsilon) + \frac{1}{\epsilon}\,\psi(u_\epsilon) = J_\epsilon(u_\epsilon) \leq J_\epsilon(u) = J(u),$$

that is,

$$J_\epsilon(u_\epsilon) \leq J(u). \tag{$*_1$}$$

Since J is coercive, the family $(u_\epsilon)_{\epsilon>0}$ is bounded. By compactness (since we are in $\mathbb{R}^n$), there exists a subsequence $(u_{\epsilon(i)})_{i\geq 0}$ with $\lim_{i\mapsto\infty} \epsilon(i) = 0$ and some element $u' \in \mathbb{R}^n$ such that

$$\lim_{i\mapsto\infty} u_{\epsilon(i)} = u'.$$

From the inequality $J(u_\epsilon) \leq J(u)$ proved in $(*_1)$ and the continuity of J, we deduce that

$$J(u') = \lim_{i\mapsto\infty} J(u_{\epsilon(i)}) \leq J(u). \tag{$*_2$}$$

By definition of $J_\epsilon(u_\epsilon)$ and $(*_1)$, we have

$$0 \leq \psi(u_{\epsilon(i)}) \leq \epsilon(i)(J(u) - J(u_{\epsilon(i)})),$$

and since the sequence $(u_{\epsilon(i)})_{i\geq 0}$ converges, the numbers $J(u) - J(u_{\epsilon(i)})$ are bounded independently of i. Consequently, since $\lim_{i\mapsto\infty} \epsilon(i) = 0$ and since the function ψ is continuous, we have

$$0 = \lim_{i\mapsto\infty} \psi(u_{\epsilon(i)}) = \psi(u'),$$

which shows that $u' \in U$. Since by $(*_2)$ we have $J(u') \leq J(u)$, and since both $u, u' \in U$ and u is the unique minimizer of J over U we must have $u' = u$. Therfore u' is the unique minimizer of J over U. But then the whole family $(u_\epsilon)_{\epsilon>0}$ converges to u since we can use the same argument as above for *every* subsequence of $(u_\epsilon)_{\epsilon>0}$. □

Note that a convex function $\psi\colon \mathbb{R}^n \to \mathbb{R}$ is automatically continuous, so the assumption of continuity is redundant.

As an application of Proposition 29.15, if U is given by

$$U = \{v \in \mathbb{R}^n \mid \varphi_i(v) \leq 0,\ i = 1, \ldots, m\},$$

where the functions $\varphi_i\colon \mathbb{R}^n \to \mathbb{R}$ are convex, we can take ψ to be the function given by

$$\psi(v) = \sum_{i=1}^{m} \max\{\varphi_i(v), 0\}.$$

In practice, the applicability of the penalty-function method is limited by the difficulty to construct effectively "good" functions ψ, for example, differentiable ones. Note that in the above example the function ψ is not diferentiable. A better penalty function is

$$\psi(v) = \sum_{i=1}^{m} (\max\{\varphi_i(v), 0\})^2.$$

Another way to deal with constrained optimization problems is to use *duality*. This approach is investigated in Chapter 30.

29.6 Summary

The main concepts and results of this chapter are listed below:

•

Chapter 30

Introduction to Nonlinear Optimization

In Chapter 20 we investigated the problem of determining when a function $J: \Omega \to \mathbb{R}$ defined on some open subset Ω of a normed vector space E has a local extremum in a subset U of Ω defined by equational constraints, namely

$$U = \{x \in \Omega \mid \varphi_i(x) = 0,\ 1 \leq i \leq m\},$$

where the functions $\varphi_i: \Omega \to \mathbb{R}$ are continuous (and usually, differentiable). Theorem 20.3 gives a necessary condition in terms of the Lagrange multipliers. In Section 20.3, we assume that U is a convex subset of Ω and Theorem 20.8 gives us a necessary condition for the function $J: \Omega \to \mathbb{R}$ to have a local minimum at u with respect to U if dJ_u exists, namely

$$dJ_u(v - u) \geq 0 \quad \text{for all } v \in U.$$

Our first goal is to find a necessary criterion for a function $J: \Omega \to \mathbb{R}$ to have a minimum on a subset U, even is this subset is not convex. This can be done by introducing a notion of "tangent cone" at a point $u \in U$.

Our approach is very much inspired by Ciarlet [30] because we find it one of the more direct, and it is general enough to accomodate Hilbert spaces. The field of nonlinear optimization and convex optimization is vast and there are many books on the subject. Among those we recommend (in alphabetic order) Bertsekas [13, 14, 15], Bertsekas, Nedić, and Ozdaglar [16], Boyd and Vandenberghe [22], Luenberger [68], and Luenberger and Ye [69].

30.1 The Cone of Feasible Directions

Let V be a normed vector space and let U be a nonempty subset of V. For any point $u \in U$, consider any converging sequence $(u_k)_{k \geq 0}$ of vectors $u_k \in U$ having u as their limit, with

$u_k \neq u$ for all $k \geq 0$, and look at the sequence of "unit chords,"

$$\frac{u_k - u}{\|u_k - u\|}.$$

This sequence could oscillate forever, or it could have a limit, some unit vector $\widehat{w} \in V$. In the second case, all nonzero vectors $\lambda\widehat{w}$ for all $\lambda > 0$, belong to the cone of feasible directions at u, which is defined as follows.

Definition 30.1. Let V be a normed vector space and let U be a nonempty subset of V. For any point $u \in U$, the *cone* $C(u)$ *of feasible directions at* u is the union of $\{0\}$ and the set of all nonzero vectors $w \in V$ for which there exists some convergent sequence $(u_k)_{k\geq 0}$ of vectors, such that

(1) $u_k \in U$ and $u_k \neq u$ for all $k \geq 0$, and $\lim_{k\mapsto\infty} = u$.

(2) $\lim_{k\mapsto\infty} \dfrac{u_k - u}{\|u_k - u\|} = \dfrac{w}{\|w\|}$, with $w \neq 0$.

Condition (2) can also be expressed as follows: there is a sequence $(\delta_k)_{k\geq 0}$ of vectors $\delta_k \in V$ such that

$$u_k = u + \|u_k - u\| \frac{w}{\|w\|} + \|u_k - u\|\, \delta_k, \quad \lim_{k\mapsto\infty} \delta_k = 0, \;\; w \neq 0.$$

Figure 30.1 illustrates the construction of w in $C(u)$.

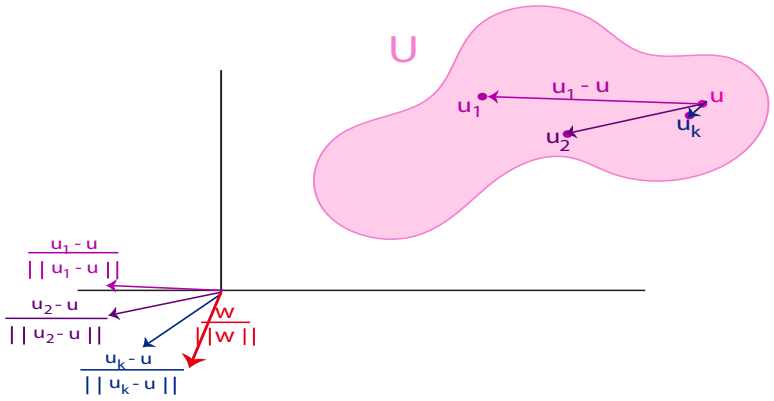


Figure 30.1: Let U be the pink region in $\mathbb{R}^2$ with fuchsia point $u \in U$. For any sequence $(u_k)_{k\geq 0}$ of points in U which converges to u, form the chords $u_k - u$ and take the limit to construct the red vector w.

The set $C(u)$ is a cone with apex 0, a notion defined as follows.

Definition 30.2. Given a vector space V, a nonempty subset $C \subseteq V$ is a *cone with apex* 0 (for short, a *cone*), if for any $v \in V$, if $v \in C$, then $\lambda v \in C$ for all $\lambda > 0$ ($\lambda \in \mathbb{R}$). For any $u \in V$, a *cone with apex* u is any nonempty subset of the form $u + C = \{u + v \mid v \in C\}$, where C is a cone with apex 0; see Figure 30.2.

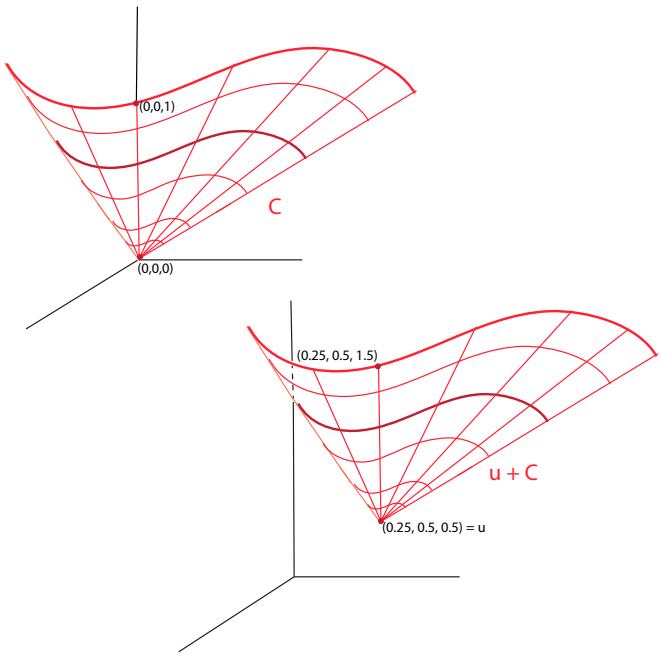


Figure 30.2: Let C be the cone determined by the bold orange curve through $(0, 0, 1)$ in the plane $z = 1$. Then $u+C$, where $u = (0.25, 0.5, 0.5)$, is the affine translate of C via the vector u.

Observe that a cone with apex 0 (or u) is not necessarily convex, and that 0 does not necessarily belong to C (resp. u does not necessarily belong to $u + C$), although in the case of the cone of feasible directions $C(u)$ we have $0 \in C(u)$ (and $u \in u + C(u)$). The condition for being a cone only asserts that if a nonzero vector v belongs to C, then the open ray $\{\lambda v \mid \lambda > 0\}$ (resp. the affine open ray $u + \{\lambda v \mid \lambda > 0\}$) also belongs to C.

Clearly, the cone $C(u)$ of feasible directions at u is a cone with apex 0, and $u + C(u)$ is a cone with apex u. Obviously, it would be desirable to have conditions on U that imply that $C(u)$ is a convex cone. Such conditions will be given later on.

Observe that the cone $C(u)$ of feasible directions at u contains the velocity vectors at u of all curves γ in U through u. If $\gamma\colon (-1, 1) \to U$ is such a curve with $\gamma(0) = u$, and if $\gamma'(u) \neq 0$ exists, then there is a sequence $(u_k)_{k\geq 0}$ of vectors in U converging to u as in Definition 30.1, with $u_k = \gamma(t_k)$ for some sequence $(t_k)_{k\geq 0}$ of reals $t_k > 0$ such that $\lim_{k\mapsto\infty} t_k = 0$, so that

$$u_k - u = t_k\gamma'(0) + t_k\epsilon_k, \quad \lim_{k\mapsto\infty} \epsilon_k = 0,$$

and we get

$$\lim_{k\mapsto\infty} \frac{u_k - u}{\|u_k - u\|} = \frac{\gamma'(0)}{\|\gamma'(0)\|}.$$

For an illustration of this paragraph in $\mathbb{R}^2$, see Figure 30.3.

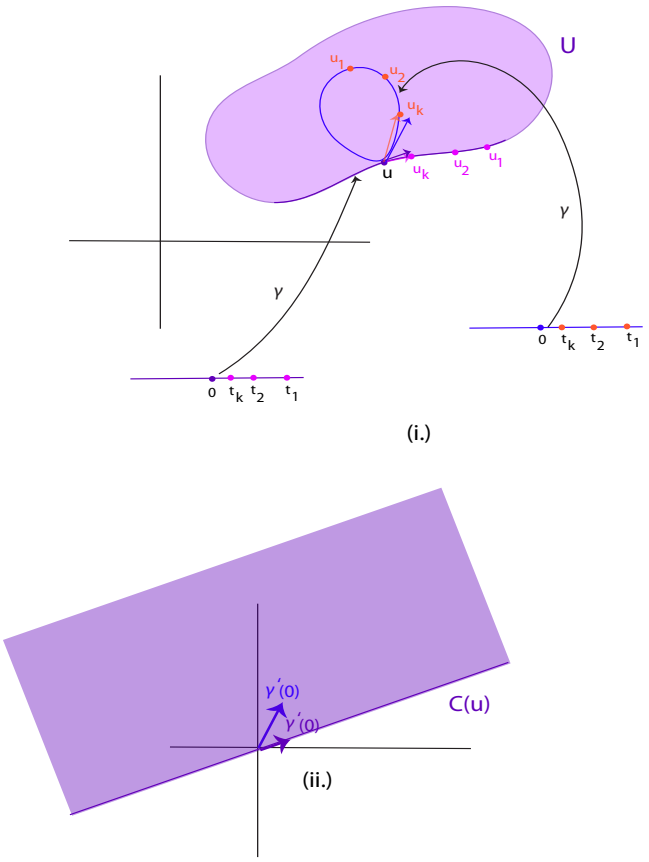


Figure 30.3: Let U be purple region in $\mathbb{R}^2$ and u be the designated point on the boundary of U. Figure (i.) illustrates two curves through u and two sequences $(u_k)_{k\geq 0}$ converging to u. The limit of the chords $u_k - u$ corresponds to the tangent vectors for the appropriate curve. Figure (ii.) illustrates the half plane $C(u)$ of feasible directions.

Example 30.1. In $V = \mathbb{R}^2$, let φ_1 and φ_2 be given by

$$\varphi_1(u_1, u_2) = -u_1 - u_2$$
$$\varphi_2(u_1, u_2) = u_1(u_1^2 + u_2^2) - (u_1^2 - u_2^2),$$

and let

$$U = \{(u_1, u_2) \in \mathbb{R}^2 \mid \varphi_1(u_1, u_2) \leq 0,\ \varphi_2(u_1, u_2) \leq 0\}.$$

The region U shown in Figure 30.4 is bounded by the curve given by the equation $\varphi_1(u_1, u_2) = 0$, that is, $-u_1 - u_2 = 0$, the line of slope -1 through the origin, and the curve given by the equation $u_1(u_1^2 + u_2^2) - (u_1^2 - u_2^2) = 0$, a nodal cubic through the origin. We obtain a parametric definition of this curve by letting $u_2 = tu_1$, and we find that

$$u_1(t) = \frac{1 - t^2}{1 + t^2}, \quad u_2(t) = \frac{t(1 - t^2)}{1 + t^2}.$$

The tangent vector at t is given by $(u_1'(t), u_2'(t))$ with

$$u_1'(t) = \frac{-2t(1+t^2) - (1-t^2)2t}{(1+t^2)^2} = \frac{-4t}{(1+t^2)^2}$$

and

$$u_2'(t) = \frac{(1-3t^2)(1+t^2) - (t-t^3)2t}{(1+t^2)^2} = \frac{1-2t^2-3t^4-2t^2+2t^4}{(1+t^2)^2} = \frac{1-4t^2-t^4}{(1+t^2)^2}.$$

The nodal cubic passes through the origin for $t = \pm 1$, and for $t = -1$ the tangent vector is $(1, -1)$, and for $t = 1$ the tangent vector is $(-1, -1)$. The cone of feasible directions $C(0)$ at the origin is given by

$$C(0) = \{(u_1, u_2) \in \mathbb{R}^2 \mid u_1 + u_2 \geq 0,\ |u_1| \geq |u_2|\}.$$

This is not a convex cone since it contains the sector delimited by the lines $u_2 = u_1$ and $u_2 = -u_1$, but also the ray supported by the vector $(-1, 1)$.

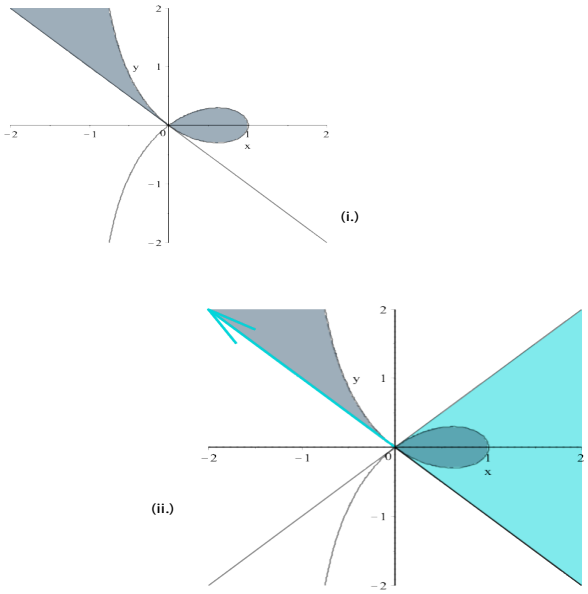


Figure 30.4: Figure (i.) illustrates U as the shaded gray region which lies between the line $y = -x$ and nodal cubic. Figure (ii.) shows the cone of feasible directions, $C(0)$, as the union of turquoise triangular cone and the turquoise the directional ray $(-1, 1)$.

The two crucial properties of the cone of feasible directions is shown in the following proposition.

Proposition 30.1. *Let U be any nonempty subset of a normed vector space V.*

(1) For any $u \in U$, the cone $C(u)$ of feasible directions at u is closed.

(2) Let $J\colon \Omega \to \mathbb{R}$ be a function defined on an open subset Ω containing U. If J has a local minimum with respect to the set U at a point $u \in U$, and if J'_u exists at u, then

$$J'_u(v-u) \geq 0 \quad \textit{for all } v \in u + C(u).$$

Proof. (1) Let $(w_n)_{n\geq 0}$ be a sequence of points $w_n \in C(u)$ converging to a limit $w \in V$. We may assume that $w \neq 0$, since $0 \in C(u)$ by definition, and thus we may also assume that $w_n \neq 0$ for all $n \geq 0$. By definition, for every $n \geq 0$, there is a sequence $(u_k^n)_{k\geq 0}$ of points in V and some $w_n \neq 0$ such that

(1) $u_k^n \in U$ and $u_k^n \neq u$ for all $k \geq 0$, and $\lim_{k\mapsto\infty} u_k^n = u$.

(2) There is a sequence $(\delta_k^n)_{k\geq 0}$ of vectors $\delta_k^n \in V$ such that

$$u_k^n = u + \|u_k^n - u\| \frac{w_n}{\|w_n\|} + \|u_k^n - u\| \delta_k^n, \quad \lim_{k\mapsto\infty} \delta_k^n = 0, \ \ w_n \neq 0.$$

Let $(\epsilon_n)_{n\geq 0}$ be a sequence of real numbers $\epsilon_n > 0$ such that $\lim_{n\mapsto\infty} \epsilon_n = 0$ (for example, $\epsilon_n = 1/(n+1)$). Due to the convergence of the sequences (u_k^n) and (δ_k^n) for every fixed n, there exist an integer $k(n)$ such that

$$\left\|u_{k(n)}^n - u\right\| \leq \epsilon_n, \quad \left\|\delta_{k(n)}^n\right\| \leq \epsilon_n.$$

Consider the sequence $(u_{k(n)}^n)_{n\geq 0}$. We have

$$u_{k(n)}^n \in U, \ \ u_{k(n)}^n \neq 0, \ \ \text{for all } n \geq 0, \ \ \lim_{n\mapsto\infty} u_{k(n)}^n = u,$$

and we can write

$$u_{k(n)}^n = u + \left\|u_{k(n)}^n - u\right\| \frac{w}{\|w\|} + \left\|u_{k(n)}^n - u\right\| \left(\delta_{k(n)}^n + \left(\frac{w_n}{\|w_n\|} - \frac{w}{\|w\|}\right)\right).$$

Since $\lim_{k\mapsto\infty}(w_n/\|w_n\|) = w/\|w\|$, we conclude that $w \in C(u)$. See Figure 30.5.

(2) Let $w = v - u$ be any nonzero vector in the cone $C(u)$, and let $(u_k)_{k\geq 0}$ be a sequence of points in $U - \{u\}$ such that

(1) $\lim_{k\mapsto\infty} u_k = u$.

(2) There is a sequence $(\delta_k)_{k\geq 0}$ of vectors $\delta_k \in V$ such that

$$u_k - u = \|u_k - u\| \frac{w}{\|w\|} + \|u_k - u\| \delta_k, \quad \lim_{k\mapsto\infty} \delta_k = 0, \ \ w \neq 0,$$

(3) $J(u) \leq J(u_k)$ for all $k \geq 0$.

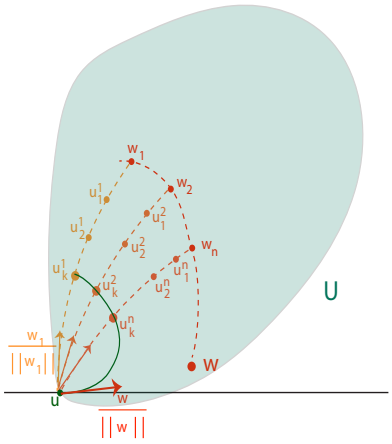


Figure 30.5: Let U be the mint green region in $\mathbb{R}^2$ with $u = (0, 0)$. Let $(w_n)_{n\geq 0}$ be a sequence of points along the upper dashed curve which converge to w. By following the dashed orange longitudinal curves, and selecting an appropriate point, we construct the dark green curve in U, which passes through u, and at u has tangent vector proportional to w.

Since J is differentiable at u, we have

$$0 \leq J(u_k) - J(u) = J'_u(u_k - u) + \|u_k - u\|\,\epsilon_k, \tag{$*$}$$

for some sequence $(\epsilon_k)_{k\geq 0}$ such that $\lim_{k\mapsto\infty} \epsilon_k = 0$. Since J'_u is linear and continuous, and

$$u_k - u = \|u_k - u\|\,\frac{w}{\|w\|} + \|u_k - u\|\,\delta_k, \quad \lim_{k\mapsto\infty} \delta_k = 0, \;\; w \neq 0,$$

$(*)$ implies that

$$0 \leq \frac{\|u_k - u\|}{\|w\|}\,(J'_u(w) + \eta_k),$$

with

$$\eta_k = \|w\|\,(J'_u(\delta_k) + \epsilon_k),$$

and since J'_u is continuous, we have $\lim_{k\mapsto\infty} \eta_k = 0$. But then, $J'_u(w) \geq 0$, since if $J'_u(w) < 0$, then for k large enough the expression $J'_u(w) + \eta_k$ would be negative, and since $u_k \neq u$, the expression
$(\|u_k - u\| / \|w\|)(J'_u(w) + \eta_k)$ would also be negative, a contradiction. □

From now on, we assume that U is defined by a set of inequalities, that is

$$U = \{x \in \Omega \mid \varphi_i(x) \leq 0, \;\; 1 \leq i \leq m\},$$

where the functions $\varphi_i \colon \Omega \to \mathbb{R}$ are continuous (and usually, differentiable). As we explained earlier, an equality constraint $\varphi_i(x) = 0$ is treated as the conjunction of the two inequalities

$\varphi_i(x) \leq 0$ and $-\varphi_i(x) \leq 0$. Later on, we will see that when the functions φ_i are convex, since $-\varphi_i$ is not necessarily convex, it is desirable to treat equality constraints separately, but for the time being we won't.

Our next goal is find sufficient conditions for the cone $C(u)$ to be convex, for any $u \in U$. For this, we assume that the functions φ_i are differentiable at u. It turns out that the constraints φ_i that matter are those for which $\varphi_i(u) = 0$, namely the constraints that are tight, or as we say, active.

Definition 30.3. Given m functions $\varphi_i \colon \Omega \to \mathbb{R}$ defined on some open subset Ω of some vector space V, let U be the set defined by

$$U = \{x \in \Omega \mid \varphi_i(x) \leq 0,\ \ 1 \leq i \leq m\}.$$

For any $u \in U$, a constraint φ_i is said to be *active* at u if $\varphi_i(u) = 0$, else *inactive* at u if $\varphi_i(u) < 0$.

If a constraint φ_i is active at u, this corresponds to u being on a piece of the boundary of U determined by some of the equations $\varphi_i(u) = 0$; see Figure 30.6.

Definition 30.4. For any $u \in U$, with

$$U = \{x \in \Omega \mid \varphi_i(x) \leq 0,\ \ 1 \leq i \leq m\},$$

we define $I(u)$ as the set of indices

$$I(u) = \{i \in \{1, \ldots, m\} \mid \varphi_i(u) = 0\}$$

where the constraints are active. Since each $(\varphi_i')_u$ is a linear form, the subset

$$C^*(u) = \{v \in V \mid (\varphi_i')_u(v) \leq 0,\ \ i \in I(u)\}$$

is the intersection of half spaces passing through the origin, so it is a convex set and obviously it is a cone. If $I(u) = \emptyset$, then $C^*(u) = V$.

The special kinds of $\mathcal{H}$-polyhedra of the form $C^*(u)$ cut out by hyperplanes through the origin are called $\mathcal{H}$-*cones*. It can be shown that every $\mathcal{H}$-cone is a polyhedral cone (also called a $\mathcal{V}$-*cone*), and conversely. The proof is nontrivial; see Gallier [45] and Ziegler [113].

We will prove shortly that we always have the inclusion

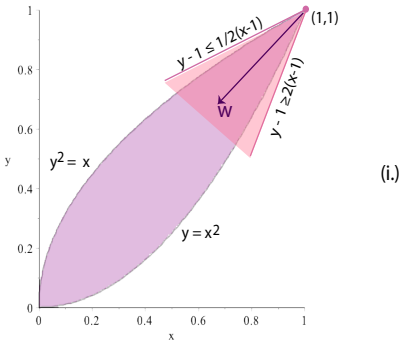
$$C(u) \subseteq C^*(u).$$

However, the inclusion can be strict, as in Example 30.1. Indeed for $u = (0, 0)$ we have $I(0, 0) = \{1, 2\}$ and since

$$(\varphi_1')_{(u_1,u_2)} = (-1 \ \ -1), \quad (\varphi_2')_{(u_1,u_2)} = (3u_1^2 + u_2^2 - 2u_1 \ \ 2u_1u_2 + 2u_2),$$

we have $(\varphi_2')_{(0,0)} = (0 \ \ 0)$, and thus $C^*(0) = \{(u_1, u_2) \in \mathbb{R}^2 \mid u_1 + u_2 \geq 0\}$ as illustrated in Figure 30.7.

The conditions stated in the following definition are sufficient conditions that imply that $C(u) = C^*(u)$, as we will prove next.



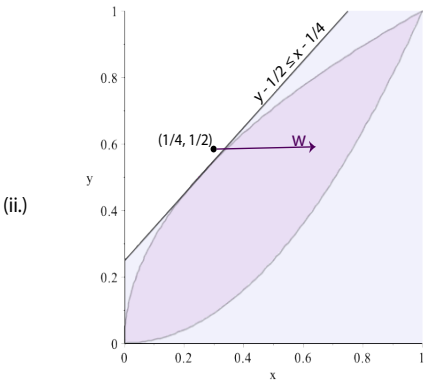


Figure 30.6: Let U be the light purple planar region which lies between the curves $y = x^2$ and $y^2 = x$. Figure (i.) illustrates the boundary point $(1, 1)$ given by the equalities $y - x^2 = 0$ and $y^2 - x = 0$. The affine translate of cone of feasible directions, $C(1, 1)$, is illustrated by the pink triangle whose sides are the tangent lines to the boundary curves. Figure (ii.) illustrates the boundary point $(1/4, 1/2)$ given by the equality $y^2 - x = 0$. The affine translate of $C(1/4, 1/2)$ is the lilac half space bounded by the tangent line to $y^2 = x$ through $(1/4, 1/2)$.

Definition 30.5. For any $u \in U$, with

$$U = \{x \in \Omega \mid \varphi_i(x) \leq 0, \;\; 1 \leq i \leq m\},$$

if the functions φ_i are differentiable at u (in fact, we only this for $i \in I(u)$), we say that the constraints are *qualified* at u if the following conditions hold:

(a) Either the constraints φ_i are affine for all $i \in I(u)$, or

(b) There is some nonzero vector $w \in V$ such that the following conditions hold for all $i \in I(u)$:

(i) $(\varphi_i')_u(w) \leq 0$.

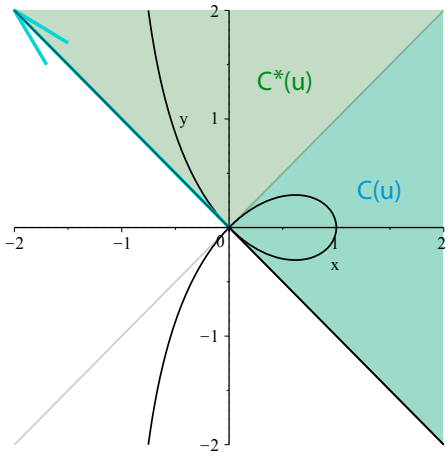


Figure 30.7: For $u = (0, 0)$, $C^*(u)$ is the sea green half space given by $u_1 + u_2 \geq 0$. This half space strictly contains $C(u)$, namely union the turquoise triangular cone and directional ray $(-1, 1)$.

(ii) If φ_i is not affine, then $(\varphi_i')_u(w) < 0$.

Condition (b)(ii) implies that u is not a critical point of φ_i for every $i \in I(u)$, so there is no singularity at u in the zero locus of φ_i. Intuitively, if the constraints are qualified at u then the boundary of U near u behaves "nicely."

The boundary points illustrated in Figure 30.6 are qualified. Observe that $U = \{x \in \mathbb{R}^2 \mid \varphi_1(x, y) = y^2 - x \leq 0,\ \varphi_2(x, y) = x^2 - y \leq 0\}$. For $u = (1, 1)$, $I(u) = \{1, 2\}$, $(\varphi_1')_{(1,1)} = (-1\ \ 2)$, $(\varphi_2')_{(1,1)} = (2\ \ -1)$, and $w = (-1, 1)$ ensures that $(\varphi_1')_{(1,1)}$ and $(\varphi_1')_{(1,1)}$ satisfy Condition (b) of Definition 30.5. For $u = (1/4, 1/2)$, $I(u) = \{1\}$, $(\varphi_1')_{(1,1)} = (-1\ \ 1)$, and $w = (-1, 0)$ will satisfy Condition (b).

In Example 30.1, the constraint $\varphi_2(u_1, u_2) = 0$ is not qualified at the origin because $(\varphi_2')_{(0,0)} = (0, 0)$; in fact, the origin is a self-intersection. In the example below, the origin is also a singular point, but for a different reason.

Example 30.2. Consider the region $U \subseteq \mathbb{R}^2$ determined by the two curves given by

$$\varphi_1(u_1, u_2) = u_2 - \max(0, u_1^3)$$
$$\varphi_2(u_1, u_2) = u_1^4 - u_2.$$

We have $I(0, 0) = \{1, 2\}$, and since $(\varphi_1)'_{(0,0)}(w_1, w_2) = (0\ 1)\binom{w_1}{w_2} = w_2$ and $(\varphi_2')_{(0,0)}(w_1, w_2) = (0\ \ -1)\binom{w_1}{w_2} = -w_2$, we have $C^*(0) = \{(u_1, u_2) \in \mathbb{R}^2 \mid u_2 = 0\}$, but the constraints are not qualified at $(0, 0)$ since it is impossible to have simultaneously $(\varphi_1')_{(0,0)}(w_1, w_2) < 0$ and $(\varphi_2')_{(0,0)}(w_1, w_2) < 0$, so in fact $C(0) = \{(u_1, u_2) \in \mathbb{R}^2 \mid u_1 \geq 0,\ u_2 = 0\}$ is strictly contained in $C^*(0)$; see Figure 30.8.

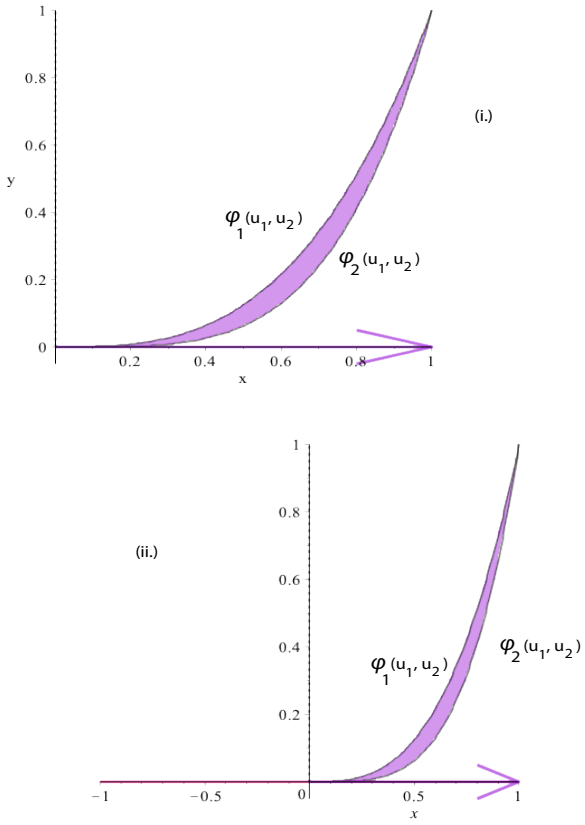


Figure 30.8: Figures (i.) and (ii.) illustrate the purple moon shaped region associated with Example 30.2. Figure (i.) also illustrates $C(0)$, the cone of feasible directions, while Figure (ii.) illustrates the strict containment of $C(0)$ in $C^*(0)$.

Proposition 30.2. *Let u be any point of the set*

$$U = \{x \in \Omega \mid \varphi_i(x) \leq 0,\ \ 1 \leq i \leq m\},$$

where Ω is an open subset of the normed vector space V, and assume that the functions φ_i are differentiable at u (in fact, we only this for $i \in I(u)$). Then the following facts hold:

(1) The cone $C(u)$ of feasible directions at u is contained in the convex cone $C^*(u)$; that is,

$$C(u) \subseteq C^*(u) = \{v \in V \mid (\varphi_i')_u(v) \leq 0,\ \ i \in I(u)\}.$$

(2) If the constraints are qualified at u (and the functions φ_i are continuous at u for all $i \notin I(u)$ if we only assume φ_i differentiable at u for all $i \in I(u)$), then

$$C(u) = C^*(u).$$

Proof. (1) For every $i \in I(u)$, since $\varphi_i(v) \leq 0$ for all $v \in U$ and $\varphi_i(u) = 0$, the function $-\varphi_i$ has a local minimum at u with respect to U, so by Proposition 30.1, we have

$$(-\varphi_i')_u(v) \geq 0 \quad \text{for all } v \in C(u),$$

which is equivalent to $(\varphi_i')_u(v) \leq 0$ for all $v \in C(u)$ and for all $i \in I(u)$, that is, $u \in C^*(u)$.

(2)(a) First, let us assume that φ_i is affine for every $i \in I(u)$. Recall that φ_i must be given by $\varphi_i(v) = h_i(v) + c_i$ for all $v \in V$, where h_i is a linear form and $c_i \in \mathbb{R}$. Since the derivative of a linear map at any point is itself,

$$(\varphi_i')_u(v) = h_i(v) \quad \text{for all } v \in V.$$

Pick any nonzero $w \in C^*(u)$, which means that $(\varphi_i')_u(w) \leq 0$ for all $i \in I(u)$. For any sequence $(\epsilon_k)_{k\geq 0}$ of reals $\epsilon_k > 0$ such that $\lim_{k\mapsto\infty} \epsilon_k = 0$, let $(u_k)_{k\geq 0}$ be the sequence of vectors in V given by

$$u_k = u + \epsilon_k w.$$

We have $u_k - u = \epsilon_k w \neq 0$ for all $k \geq 0$ and $\lim_{k\mapsto\infty} u_k = u$. Furthermore, since the functions φ_i are continuous for all $i \notin I$, we have

$$0 > \varphi_i(u) = \lim_{k\mapsto\infty} \varphi_i(u_k),$$

and since φ_i is affine and $\varphi_i(u) = 0$ for all $i \in I$, we have $\varphi_i(u) = h_i(u) + c_i = 0$, so

$$\varphi_i(u_k) = h_i(u_k) + c_i = h_i(u_k) - h_i(u) = h_i(u_k - u) = (\varphi_i')_u(u_k - u) = \epsilon_k(\varphi_i')_u(w) \leq 0,$$

which implies that $u_k \in U$ for all k large enough. Since

$$\frac{u_k - u}{\|u_k - u\|} = \frac{w}{\|w\|} \quad \text{for all } k \geq 0,$$

we conclude that $w \in C(u)$. See Figure 30.9.

(2)(b) Let us now consider the case where some function φ_i is not affine for some $i \in I(u)$. Let $w \neq 0$ be some vector in V such that Condition (b) of Definition 30.5 holds, namely: for all $i \in I(u)$, we have

(i) $(\varphi_i')_u(w) \leq 0$.

(ii) If φ_i is not affine, then $(\varphi_i')_u(w) < 0$.

Pick any nonzero vector $v \in C^*(u)$, which means that $(\varphi_i')_u(v) \leq 0$ for all $i \in I(u)$, and let $\delta > 0$ be any positive real number such that $v + \delta w \neq 0$. For any sequence $(\epsilon_k)_{k\geq 0}$ of reals $\epsilon_k > 0$ such that $\lim_{k\mapsto\infty} \epsilon_k = 0$, let $(u_k)_{k\geq 0}$ be the sequence of vectors in V given by

$$u_k = u + \epsilon_k(v + \delta w).$$

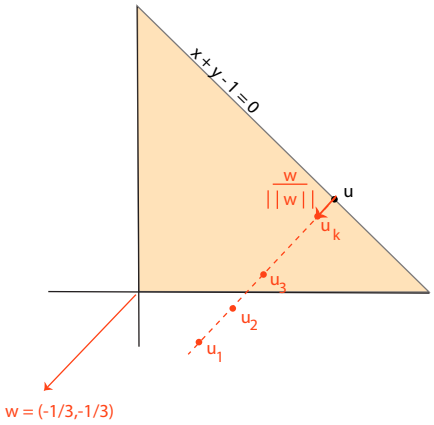


Figure 30.9: Let U be the peach triangle bounded by the lines $y = 0$, $x = 0$, and $y = -x+1$. Let u satisfy the affine constraint $\varphi(x, y) = y + x - 1$. Since $\varphi'_{(x,y)} = (1\ 1)$, set $w = (-1, -1)$ and approach u along the line $u + tw$.

We have $u_k - u = \epsilon_k(v + \delta w) \neq 0$ for all $k \geq 0$ and $\lim_{k\mapsto\infty} u_k = u$. Furthermore, since the functions φ_i are continuous for all $i \notin I(u)$, we have

$$0 > \varphi_i(u) = \lim_{k\mapsto\infty} \varphi_i(u_k) \quad \text{for all } i \notin I(u), \tag{$*_1$}$$

and as in the previous case, for all $i \in I(u)$ such that φ_i is affine, since $(\varphi'_i)_u(v) \leq 0$, $(\varphi'_i)_u(w) \leq 0$, and $\epsilon_k, \delta > 0$, we have

$$\varphi_i(u_k) = \epsilon_k((\varphi'_i)_u(v) + \delta(\varphi'_i)_u(w)) \leq 0 \quad \text{for all } i \in I(u) \text{ and } \varphi_i \text{ affine}, \tag{$*_2$}$$

and since φ_i is differentiable and $\varphi_i(u) = 0$ for all $i \in I(u)$, if φ_i is not affine we have

$$\varphi_i(u_k) = \epsilon_k((\varphi'_i)_u(v) + \delta(\varphi'_i)_u(w)) + \epsilon_k \|u_k - u\| \eta_k(u_k - u)$$

with $\lim_{\|u_k-u\|\mapsto 0} \eta_k(u_k - u) = 0$, so if we write $\alpha_k = \|u_k - u\| \eta_k(u_k - u)$, we have

$$\varphi_i(u_k) = \epsilon_k((\varphi'_i)_u(v) + \delta(\varphi'_i)_u(w) + \alpha_k)$$

with $\lim_{k\mapsto\infty} \alpha_k = 0$, and since $(\varphi'_i)_u(v) \leq 0$, we obtain

$$\varphi_i(u_k) \leq \epsilon_k(\delta(\varphi'_i)_u(w) + \alpha_k) \quad \text{for all } i \in I(u) \text{ and } \varphi_i \text{ not affine}. \tag{$*_3$}$$

The Equations $(*_1), (*_2), (*_3)$ show that $u_k \in U$ for k sufficiently large, where in $(*_3)$, since $(\varphi'_i)_u(w) < 0$ and $\delta > 0$, even if $\alpha_k > 0$, when $\lim_{k\mapsto\infty} \alpha_k = 0$, we will have $\delta(\varphi'_i)_u(w) + \alpha_k < 0$ for k large enough, and thus $\epsilon_k(\delta(\varphi'_i)_u(w) + \alpha_k) < 0$ for k large enough.

Since

$$\frac{u_k - u}{\|u_k - u\|} = \frac{v + \delta w}{\|v + \delta w\|}$$

for all $k \geq 0$, we conclude that $v + \delta w \in C(u)$ for $\delta > 0$ small enough. But now the sequence $(v_n)_{n \geq 0}$ given by

$$v_n = v + \epsilon_n w$$

converges to v, and for n large enough $v_n \in C(u)$. Since by Proposition 30.1, the cone $C(u)$ is closed, we conclude that $v \in C(u)$. See Figure 30.10.

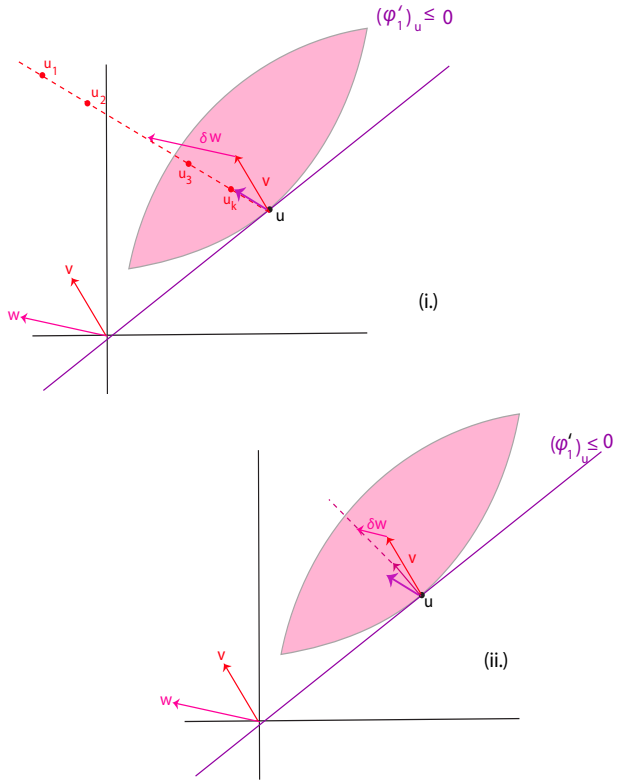


Figure 30.10: Let U be the pink lounge in $\mathbb{R}^2$. Let u satisfy the non-affine constraint $\varphi_1(u)$. Choose vectors v and w in the half space $(\varphi_1')_u \leq 0$. Figure (i.) approaches u along the line $u + t(\delta w + v)$ and shows that $v + \delta w \in C(u)$ for fixed δ. Figure (ii.) varies δ in order that the purple vectors approach v as $\delta \to \infty$.

In all cases, we proved that $C^*(u) \subseteq C(u)$, as claimed. □

In the case of m affine constraints $a_i x \leq b_i$, for some linear forms a_i and some $b_i \in \mathbb{R}$, for any point $u \in \mathbb{R}^n$ such that $a_i u = b_i$ for all $i \in I(u)$, the cone $C(u)$ consists of all $v \in \mathbb{R}^n$ such that $a_i v \leq 0$, so $u + C(u)$ consists of all points $u + v$ such that

$$a_i(u + v) \leq b_i \quad \text{for all } i \in I(u),$$

which is the cone cut out by the hyperplanes determining some face of the polyhedron defined by the m constraints $a_i x \leq b_i$.

We are now ready to prove one of the most important results of nonlinear optimization.

30.2 The Karush–Kuhn–Tucker Conditions

If the domain U is defined by inequality constraints satisfying mild differentiability conditions and if the constraints at u are qualified, then there is a necessary condition for the function J to have a local minimum at $u \in U$ involving generalized Lagrange multipliers. The proof uses a version of Farkas Lemma. In fact, the necessary condition stated next holds for infinite-dimensional vector spaces because there a version of Farkas Lemma holding for real Hilbert spaces, but we will content ourselves with the version holding for finite dimensional normed vector spaces. For the more general version, see Theorem 28.11 (or Ciarlet [30], Chapter 9).

We will be using the following version of Farkas Lemma.

Proposition 30.3. *(Farkas Lemma, Version I) Let A be an $m \times n$ matrix and let $b \in \mathbb{R}^m$ be any vector. The linear system $Ax = b$ has no solution $x \geq 0$ iff there is some nonzero linear form $y \in (\mathbb{R}^m)^*$ such that $yA \geq 0_n^\top$ and $yb < 0$.*

We will use the version of Farkas Lemma obtained by taking a contrapositive, namely: *if $yA \geq 0_n^\top$ implies $yb \geq 0$ for all linear forms $y \in (\mathbb{R}^m)^*$, then linear system $Ax = b$ some solution $x \geq 0$.*

Actually, it is more convenient to use a version of Farkas Lemma applying to a Euclidean vector space (with an inner product denoted $\langle -, - \rangle$). This version also applies to an infinite dimensional real Hilbert space; see Theorem 28.11. Recall that in a Euclidean space V the inner product induces an isomorphism between V and its dual V^*. In our case, we need the isomorphism $\sharp$ from V^* to V defined such that for every linear form $\omega \in V^*$, the vector $\omega^\sharp \in V$ is uniquely defined by the equation

$$\omega(v) = \langle v, \omega^\sharp \rangle \quad \text{for all } v \in V.$$

In $\mathbb{R}^n$, the isomorphism between $\mathbb{R}^n$ and $(\mathbb{R}^n)^*$ amounts to transposition: if $y \in (\mathbb{R}^n)^*$ is a linear form and $v \in \mathbb{R}^n$ is a vector, then

$$yv = v^\top y^\top.$$

The version of the Farkas–Minskowski lemma in term of an inner product is as follows.

Proposition 30.4. *(Farkas–Minkowski) Let V be a Euclidean space of finite dimension with inner product $\langle -, - \rangle$ (more generally, a Hilbert space). For any finite family $(a_1, \ldots, a_m)$ of m vectors $a_i \in V$ and any vector $b \in V$, for any $v \in V$,*

$$\text{if } \langle a_i, v \rangle \geq 0 \text{ for } i = 1, \ldots, m \text{ implies that } \langle b, v \rangle \geq 0,$$

then there exist $\lambda_1, \ldots, \lambda_m \in \mathbb{R}$ *such that*

$$\lambda_i \geq 0 \text{ for } i = 1, \ldots, m, \text{ and } b = \sum_{i=1}^{m} \lambda_i a_i,$$

that is, b *belong to the polyhedral cone* $\mathrm{cone}(a_1, \ldots, a_m)$.

Proposition 30.4 is the special case of Theorem 28.11 which holds for real Hilbert spaces.

We can now prove the following theorem.

Theorem 30.5. *Let* $\varphi_i \colon \Omega \to \mathbb{R}$ *be* m *constraints defined on some open subset* Ω *of a finite-dimensional Euclidean vector space* V *(more generally, a real Hilbert space* V*), let* $J \colon \Omega \to \mathbb{R}$ *be some function, and let* U *be given by*

$$U = \{x \in \Omega \mid \varphi_i(x) \leq 0, \ \ 1 \leq i \leq m\}.$$

For any $u \in U$*, let*

$$I(u) = \{i \in \{1, \ldots, m\} \mid \varphi_i(u) = 0\},$$

and assume that the functions φ_i *are differentiable at* u *for all* $i \in I(u)$ *and continuous at* u *for all* $i \notin I(u)$*. If* J *is differentiable at* u*, has a local minimum at* u *with respect to* U*, and if the constraints are qualified at* u*, then there exist some scalars* $\lambda_i(u) \in \mathbb{R}$ *for all* $i \in I(u)$*, such that*

$$J'_u + \sum_{i \in I(u)} \lambda_i(u)(\varphi'_i)_u = 0, \quad \text{and} \quad \lambda_i(u) \geq 0 \text{ for all } i \in I(u).$$

The above conditions are called the Karush–Kuhn–Tucker optimality conditions. Equivalently, in terms of gradients, the above conditions are expressed as

$$\nabla J_u + \sum_{i \in I(u)} \lambda_i(u) \nabla(\varphi_i)_u = 0, \quad \text{and} \quad \lambda_i(u) \geq 0 \text{ for all } i \in I(u).$$

Proof. By Proposition 30.1, we have

$$J'_u(w) \geq 0 \quad \text{for all } w \in C(u), \tag{$*_1$}$$

and by Proposition 30.2, we have $C(u) = C^*(u)$, where

$$C^*(u) = \{v \in V \mid (\varphi'_i)_u(v) \leq 0, \ \ i \in I(u)\}, \tag{$*_2$}$$

so $(*_1)$ can be expressed as: for all $w \in V$,

$$\text{if } w \in C^*(u) \text{ then } J'_u(w) \geq 0,$$

or

$$\text{if } -(\varphi'_i)_u(w) \geq 0 \text{ for all } i \in I(u) \text{ then } J'_u(w) \geq 0. \tag{$*_3$}$$

Under the isomorphism $\sharp$, the vector $(J'_u)^\sharp$ is the gradient ∇J_u, so that

$$J'_u(w) = \langle w, \nabla J_u \rangle, \tag{$*_4$}$$

and the vector $((\varphi'_i)_u)^\sharp$ is the gradient $\nabla(\varphi_i)_u$, so that

$$(\varphi'_i)_u(w) = \langle w, \nabla(\varphi_i)_u \rangle. \tag{$*_5$}$$

Using the Equations $(*_4)$ and $(*_5)$, the Equation $(*_3)$ can be written as: for all $w \in V$,

$$\text{if } \langle w, -\nabla(\varphi_i)_u \rangle \geq 0 \text{ for all } i \in I(u) \text{ then } \langle w, \nabla J_u \rangle \geq 0. \tag{$*_6$}$$

By the Farkas–Minkowski proposition (Proposition 30.4), there exist some sacalars $\lambda_i(u)$ for all $i \in I(u)$, such that $\lambda_i(u) \geq 0$ and

$$\nabla J_u = \sum_{i \in I(u)} \lambda_i(u)(-\nabla(\varphi_i)_u),$$

that is

$$\nabla J_u + \sum_{i \in I(u)} \lambda_i(u) \nabla(\varphi_i)_u = 0,$$

and using the inverse of the isomorphism $\sharp$ (which is linear), we get

$$J'_u + \sum_{i \in I(u)} \lambda_i(u)(\varphi'_i)_u = 0,$$

as claimed. □

Since the constraints are inequalities of the form $\varphi_i(x) \leq 0$, there is a way of expressing the Karush–Kuhn–Tucker optimality conditions, often abbreviated as *KKT conditions*, in a way that does not refer explicitly to the index set $I(u)$:

$$J'_u + \sum_{i=1}^{m} \lambda_i(u)(\varphi'_i)_u = 0, \tag{KKT$_1$}$$

and

$$\sum_{i=1}^{m} \lambda_i(u)\varphi_i(u) = 0, \quad \lambda_i(u) \geq 0, \quad i = 1, \ldots, m. \tag{KKT$_2$}$$

Indeed, if we have the strict inequality $\varphi_i(u) < 0$ (the constraint φ_i is inactive at u), since all the terms $\lambda_i(u)\varphi_i(u)$ are nonpositive, we must have $\lambda_i(u) = 0$; that is, we only need to consider the $\lambda_i(u)$ for all $i \in I(u)$. Yet another way to express the conditions in (KKT$_2$) is

$$\lambda_i(u)\varphi_i(u) = 0, \quad \lambda_i(u) \geq 0, \quad i = 1, \ldots, m. \tag{KKT$'_2$}$$

In other words, for any $i \in \{1, \ldots, m\}$, if $\varphi_i(u) < 0$, then $\lambda_i(u) = 0$; that is, if the constraint φ_i is inactive at u, then $\lambda_i(u) = 0$. By contrapositive, if $\lambda_i(u) \neq 0$, then $\varphi_i(u) = 0$; that is, if $\lambda_i(u) \neq 0$, then the constraint φ_i is active at u. The conditions in (KKT'_2) are referred to as *complementary slackness* conditions.

The scalars $\lambda_i(u)$ are often called *generalized Lagrange multipliers*. If $V = \mathbb{R}^n$, the necessary conditions of Theorem 30.5 are expressed as the following system of equations and inequalities in the unknowns $(u_1, \ldots, u_n) \in \mathbb{R}^n$ and $(\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^m_+$:

$$\begin{aligned}
\frac{\partial J}{\partial x_1}(u) + \lambda_1 \frac{\partial \varphi_1}{\partial x_1}(u) + \cdots + \lambda_m \frac{\partial \varphi_m}{\partial x_1}(u) &= 0 \\
\vdots \qquad\qquad & \vdots \\
\frac{\partial J}{\partial x_n}(u) + \lambda_1 \frac{\partial \varphi_n}{\partial x_1}(u) + \cdots + \lambda_m \frac{\partial \varphi_m}{\partial x_n}(u) &= 0 \\
\lambda_1 \varphi_1(u) + \cdots + \lambda_m \varphi_m(u) &= 0 \\
\varphi_1(u) &\leq 0 \\
\vdots \qquad & \vdots \\
\varphi_m(u) &\leq 0 \\
\lambda_1, \ldots, \lambda_m &\geq 0.
\end{aligned}$$

Example 30.3. Let J, φ_1 and φ_2 be the functions defined on $\mathbb{R}$ by

$$\begin{aligned}
J(x) &= x \\
\varphi_1(x) &= -x \\
\varphi_2(x) &= x - 1.
\end{aligned}$$

In this case

$$U = \{x \in \mathbb{R} \mid -x \leq 0,\ x - 1 \leq 0\} = [0, 1].$$

Since the constraints are affine, they are automatically qualified for any $u \in [0, 1]$. The system of equations and inequalities shown above becomes

$$\begin{aligned}
1 - \lambda_1 + \lambda_2 &= 0 \\
-\lambda_1 x + \lambda_2(x - 1) &= 0 \\
-x &\leq 0 \\
x - 1 &\leq 0 \\
\lambda_1, \lambda_2 &\geq 0.
\end{aligned}$$

The last four equations imply that either $x = 0$ or $x = 1$.

If $x = 0$, by the second equation we get $\lambda_2 = 0$, so $\lambda_1 = 1 \geq 0$. Indeed $x = 0$ is the minimun of $J(x) = x$ over $[0, 1]$.

If $x = 1$, by the second equation we get $\lambda_1 = 0$, so $\lambda_2 = -1$, a contradiction. Indeed, 1 is a maximum, and not a minimum of $J(x) = x$ over $[0, 1]$.

Remark: Unless the linear forms $(\varphi_i')_u$ for $i \in I(u)$ are linearly independent, the $\lambda_i(u)$ are generally not unique. Also, if $I(u) = \emptyset$, then the KKT conditions reduce to $J'_u = 0$. This is not surprising because in this case u belongs to the relative interior of U.

If the constraints are all affine equality constraints, then the KKT conditions are a bit simpler. We will consider this case shortly.

The conditions for the qualification of nonaffine constraints are hard (if not impossible) to use in practice, because they depend on $u \in U$ and on the derivatives $(\varphi_i')_u$. Thus it is desirable to find simpler conditions. Fortunately, this is possible if the nonaffine functions φ_i are convex.

Definition 30.6. Let $U \subseteq \Omega \subseteq V$ be given by

$$U = \{x \in \Omega \mid \varphi_i(x) \leq 0,\ \ 1 \leq i \leq m\},$$

where Ω is an open subset of the Euclidean vector space V. If the functions $\varphi_i \colon \Omega \to \mathbb{R}$ are convex, we say that the constraints are *qualified* if the following conditions hold:

(a) Either the constraints φ_i are affine for all $i = 1, \ldots, m$ and $U \neq \emptyset$, or

(b) There is some vector $v \in \Omega$ such that the following conditions hold for $i = 1, \ldots, m$:

(i) $\varphi_i(v) \leq 0$.

(ii) If φ_i is not affine, then $\varphi_i(v) < 0$.

The above qualification conditions are known as *Slater's conditions*.

Condition (b)(i) also implies that U has nonempty relative interior. If Ω is convex, then U is also convex. This is because for all $u, v \in \Omega$, if $u \in U$ and $v \in U$, that is $\varphi_i(u) \leq 0$ and $\varphi_i(v) \leq 0$ for $i = 1, \ldots, m$, since the functions φ_i are convex, for all $\theta \in [0, 1]$ we have

$$\begin{aligned} \varphi_i((1-\theta)u + \theta v) &\leq (1-\theta)\varphi_i(u) + \theta\varphi_i(v) && \text{since } \varphi_i \text{ is convex} \\ &\leq 0 && \text{since } 1-\theta \geq 0, \theta \geq 0, \varphi_i(u) \leq 0, \varphi_i(v) \leq 0, \end{aligned}$$

and any intersection of convex sets is convex.

It is important to observe that a *nonaffine equality constraint* $\varphi_i(u) = 0$ is *never* qualified.

Indeed, $\varphi_i(u) = 0$ is equivalent to $\varphi_i(u) \leq 0$ and $-\varphi_i(u) \leq 0$, so if these constraints are qualified and if φ_i is not affine then there is some nonzero vector $v \in \Omega$ such that both $\varphi_i(v) < 0$ and $-\varphi_i(v) < 0$, which is impossible. For this reason, equality constraints are often assumed to be affine.

The following theorem yields a more flexible version of Theorem 30.5 for constraints given by convex functions. If in addition, the function J is also convex, then the KKT conditions are also a sufficient condition for a local minimum.

Theorem 30.6. *Let $\varphi_i \colon \Omega \to \mathbb{R}$ be m convex constraints defined on some open convex subset Ω of a finite-dimensional Euclidean vector space V (more generally, a real Hilbert space V), let $J \colon \Omega \to \mathbb{R}$ be some function, let U be given by*

$$U = \{x \in \Omega \mid \varphi_i(x) \leq 0,\ \ 1 \leq i \leq m\},$$

and let $u \in U$ be any point such that the functions φ_i and J are differentiable at u.

(1) If J has a local minimum at u with respect to U, and if the constraints are qualified, then there exist some scalars $\lambda_i(u) \in \mathbb{R}$, such that the KKT condition hold:

$$J'_u + \sum_{i=1}^{m} \lambda_i(u)(\varphi'_i)_u = 0$$

and

$$\sum_{i=1}^{m} \lambda_i(u)\varphi_i(u) = 0, \quad \lambda_i(u) \geq 0, \quad i = 1, \ldots, m.$$

Equivalently, in terms of gradients, the above conditions are expressed as

$$\nabla J_u + \sum_{i=1}^{m} \lambda_i(u)\nabla(\varphi_i)_u = 0,$$

and

$$\sum_{i=1}^{m} \lambda_i(u)\varphi_i(u) = 0, \quad \lambda_i(u) \geq 0, \quad i = 1, \ldots, m.$$

(2) Conversely, if the restriction of J to U is convex and if there exist scalars $(\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^m_+$ such that the KKT conditions hold, then the function J has a (global) minimum at u with respect to U.

Proof. (1) It suffices to prove that if the convex constraints are qualified according to Definition 30.6, then they are qualified according to Definition 30.5, since in this case we can apply Theorem 30.5.

If $v \in \Omega$ is a vector such that Condition (b) of Definition 30.6 holds and if $v \neq u$, for any $i \in I(u)$, since $\varphi_i(u) = 0$ and since φ_i is convex, by Proposition 20.9,

$$\varphi_i(v) \geq \varphi_i(u) + (\varphi'_i)_u(v - u) = (\varphi'_i)_u(v - u),$$

so if we let $w = v - u$ then

$$(\varphi'_i)_u(w) \leq \varphi_i(v),$$

which shows that the nonaffine constraints φ_i for $i \in I(u)$ are qualified according to Definition 30.5, by Condition (b) of Definition 30.6.

If $v = u$, then the constraints φ_i for which $\varphi_i(u) = 0$ must be affine (otherwise, Condition (b)(ii) of Definition 30.6 would be false), and in this case we can pick $w = 0$.

(2) Let v be any arbitrary point in the convex subset U. Since $\varphi_i(v) \leq 0$ and $\lambda_i \geq 0$ for $i = 1, \ldots, m$, we have $\sum_{i=1}^m \lambda_i \varphi_i(v) \leq 0$, and using the fact that

$$\sum_{i=1}^m \lambda_i(u)\varphi_i(u) = 0, \quad \lambda_i(u) \geq 0, \quad i = 1, \ldots, m,$$

we have $\lambda_i = 0$ if $i \notin I(u)$ and $\varphi_i(u) = 0$ if $i \in I(u)$, so we have

$$\begin{aligned}
J(u) &\leq J(u) - \sum_{i=1}^m \lambda_i \varphi_i(v) \\
&\leq J(u) - \sum_{i \in I(u)} \lambda_i(\varphi_i(v) - \varphi_i(u)) && \lambda_i = 0 \text{ if } i \notin I(u),\ \varphi_i(u) = 0 \text{ if } i \in I(u) \\
&\leq J(u) - \sum_{i \in I(u)} \lambda_i (\varphi_i')_u(v - u) && \text{(by Proposition 20.9)} \\
&\leq J(u) + J'_u(v - u) && \text{(by the KKT conditions)} \\
&\leq J(v) && \text{(by Proposition 20.9),}
\end{aligned}$$

and this shows that u is indeed a (global) minimum of J over U. □

It is important to note that when *both* the constraints, the domain of definition Ω, *and* the objective function J are *convex*, if the KKT conditions hold for some $u \in U$ and some $\lambda \in \mathbb{R}^m_+$, then Theorem 30.6 implies that J has a (global) minimum at u with respect to U, independently of any assumption on the qualification of the constraints.

The above theorem suggests introducing the function $L\colon \Omega \times \mathbb{R}^m_+ \to \mathbb{R}$ given by

$$L(v, \lambda) = J(v) + \sum_{i=1}^m \lambda_i \varphi_i(v),$$

with $\lambda = (\lambda_1, \ldots, \lambda_m)$. The function L is called the *Lagrangian* of the *minimization problem* (P):

$$\begin{aligned}
&\text{minimize} && J(v) \\
&\text{subject to} && \varphi_i(v) \leq 0, \quad i = 1, \ldots, m.
\end{aligned}$$

The KKT conditions of Theorem 30.6 imply that for any $u \in U$, if the vector $\lambda = (\lambda_1, \ldots, \lambda_m)$ is known and if u is a minimum of J on U, then

$$\begin{aligned}
\frac{\partial L}{\partial u}(u) &= 0 \\
J(u) &= L(u, \lambda).
\end{aligned}$$

The Lagrangian technique "absorbs" the constraints into the new objective function L and reduces the problem of finding a constrained minimum of the function J, to the problem of finding an unconstrained minimum of the function $L(v, \lambda)$. This is the main point of Lagrangian duality which will be treated in the next section.

A case that arises often in practice is the case where the constraints φ_i are affine. If so, the m constraints $a_i x \leq b_i$ can be expressed in matrix form as $Ax \leq b$, where A is an $m \times n$ matrix whose ith row is the row vector a_i. The KKT conditions of Theorem 30.6 yield the following corollary.

Proposition 30.7. *If U is given by*

$$U = \{x \in \Omega \mid Ax \leq b\},$$

where Ω is an open convex subset of $\mathbb{R}^n$ and A is an $m \times n$ matrix, and if J is differentiable at u and J has a local minimum at u, then there exist some vector $\lambda \in \mathbb{R}^m$, such that

$$\begin{aligned} &\nabla J_u + A^\top \lambda = 0 \\ &\lambda_i \geq 0 \ \textit{and} \quad \textit{if } a_i u < b_i, \textit{ then } \lambda_i = 0, \ i = 1, \ldots, m. \end{aligned}$$

If the function J is convex, then the above conditions are also sufficient for J to have a minimum at $u \in U$.

Another case of interest is the generalization of the minimization problem involving the affine constraints of a linear program in standard form, that is, equality constraints $Ax = b$ with $x \geq 0$, where A is an $m \times n$ matrix. In our formalism, this corresponds to the $2m + n$ constraints

$$\begin{aligned} a_i x - b_i &\leq 0, \qquad i = 1, \ldots, m \\ -a_i x + b_i &\leq 0, \qquad i = 1, \ldots, m \\ -x_j &\leq 0, \qquad i = 1, \ldots, n. \end{aligned}$$

In matrix form, they can be expressed as

$$\begin{pmatrix} A \\ -A \\ -I_n \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \leq \begin{pmatrix} b \\ -b \\ 0_n \end{pmatrix}.$$

If we introduce the generalized Lagrange multipliers λ_i^+ and λ_i^- for $i = 1, \ldots, m$ and μ_j for $j = 1, \ldots, n$, then the KKT conditions are

$$\nabla J_u + \begin{pmatrix} A^\top & -A^\top & -I_n \end{pmatrix} \begin{pmatrix} \lambda^+ \\ \lambda^- \\ \mu \end{pmatrix} = 0_n,$$

that is,

$$\nabla J_u + A^\top \lambda^+ - A^\top \lambda^- - \mu = 0,$$

and $\lambda^+, \lambda^-, \mu \geq 0$, and if $a_i u < b_i$ then $\lambda_i^+ = 0$, if $-a_i u < -b_i$ then $\lambda_i^- = 0$, and if $-u_j < 0$, then $\mu_j = 0$. But the constraints $a_i u = b_i$ hold for $i = 1, \ldots, m$, so this places no restriction on the λ_i^+ and λ_i^-, and if we write $\lambda_i = \lambda_i^+ - \lambda_i^-$, then we have

$$\nabla J_u + A^\top \lambda = \mu,$$

with $\mu_j \geq 0$, and if $u_j > 0$ then $\mu_j = 0$, for $j = 1, \ldots, n$.

Thus we proved the following proposition (which is slight generalization of Proposition 8.7.2 in Matousek and Gardner [72]).

Proposition 30.8. *If U is given by*

$$U = \{x \in \Omega \mid Ax = b,\ x \geq 0\},$$

where where Ω is an open convex subset of $\mathbb{R}^n$ and A is an $m \times n$ matrix, and if J is differentiable at u and J has a local minimum at u, then there exist two vectors $\lambda \in \mathbb{R}^m$ $\mu \in \mathbb{R}^n$, such that

$$\nabla J_u + A^\top \lambda = \mu,$$

with $\mu_j \geq 0$, and if $u_j > 0$ then $\mu_j = 0$, for $j = 1, \ldots, n$. Equivalently, there exists a vector $\lambda \in \mathbb{R}^m$ such that

$$(\nabla J_u)_j + (A^j)^\top \lambda \quad \begin{cases} = 0 & \text{if } u_j > 0 \\ \geq 0 & \text{if } u_j = 0, \end{cases}$$

where A^j is the jth column of A. If the function J is convex, then the above conditions are also sufficient for J to have a minimum at $u \in U$.

Yet another special case that arises frequently in practice is the minimization problem involving the affine equality constraints $Ax = b$, where A is an $m \times n$ matrix, with no restriction on x. Reviewing the proof of Proposition 30.8, we obtain the following proposition.

Proposition 30.9. *If U is given by*

$$U = \{x \in \Omega \mid Ax = b\},$$

where Ω is an open convex subset of $\mathbb{R}^n$ and A is an $m \times n$ matrix, and if J is differentiable at u and J has a local minimum at u, then there exist some vector $\lambda \in \mathbb{R}^m$ such that

$$\nabla J_u + A^\top \lambda = 0.$$

Equivalently, there exists a vector $\lambda \in \mathbb{R}^m$ such that

$$(\nabla J_u)_j + (A^j)^\top \lambda = 0,$$

where A^j is the jth column of A. If the function J is convex, then the above conditions are also sufficient for J to have a minimum at $u \in U$.

Observe that in Proposition 30.9, the λ_i are just standard Lagrange multipliers, with no restriction of positivity. Thus, Proposition 30.9 is a slight generalization of Theorem 20.3 that requires A to have rank m, but in the case of equational affine constraints, this assumption is uncessary.

Here is an application of Proposition 30.9 to the *interior point method* in linear programming.

Example 30.4. In linear programming, the interior point method using a central path uses a logarithmic barrier function to keep the solutions $x \in \mathbb{R}^n$ of the equation $Ax = b$ away from boundaries by forcing $x > 0$, which means that $x_i > 0$ for all i; see Matousek and Gardner [72] (Section 7.2). Write

$$\mathbb{R}^n_{++} = \{x \in \mathbb{R}^n \mid x_i > 0,\ i = 1, \ldots, n\}.$$

Observe that $\mathbb{R}^n_{++}$ is open and convex. For any $\mu > 0$, we define the function f_μ defined on $\mathbb{R}^n_{++}$ by

$$f_\mu(x) = c^\top x + \mu \sum_{i=1}^{n} \ln x_i,$$

where $c \in \mathbb{R}^n$.

We would like to find necessary condition for f_μ to have a maximum on

$$U = \{x \in \mathbb{R}^n_{++} \mid Ax = b\},$$

or equivalently to solve the following problem:

$$\begin{array}{ll} \text{maximize} & f_\mu(x) \\ \text{subject to} & \\ & Ax = b \\ & x > 0. \end{array}$$

By Proposition 30.9 if x is an optimal of the above problem then there is some $y \in \mathbb{R}^m$ such that

$$\nabla f_\mu(x) + A^\top y = 0.$$

Since

$$\nabla f_\mu(x) = \begin{pmatrix} c_1 + \frac{\mu}{x_1} \\ \vdots \\ c_n + \frac{\mu}{x_n} \end{pmatrix},$$

we obtain the equation

$$c + \mu \begin{pmatrix} \frac{1}{x_1} \\ \vdots \\ \frac{1}{x_n} \end{pmatrix} = -A^\top y.$$

To obtain a more convenient formulation, we define $s \in \mathbb{R}^n_{++}$ such that

$$s = \mu \begin{pmatrix} \frac{1}{x_1} \\ \vdots \\ \frac{1}{x_n} \end{pmatrix}$$

which implies that

$$\begin{pmatrix} s_1 x_1 & \cdots & s_n x_n \end{pmatrix} = \mu \mathbf{1}_n^\top,$$

we rename $-y$ as y (which does not make any difference since $y \in \mathbb{R}^m$), and we obtain the following necessary conditions for f_μ to have a maximum:

$$\begin{aligned} Ax &= b \\ A^\top y - s &= c \\ \begin{pmatrix} s_1 x_1 & \cdots & s_n x_n \end{pmatrix} &= \mu \mathbf{1}_n^\top \\ s, x &> 0. \end{aligned}$$

It is not hard to show that if the primal linear program with objective function $c^\top x$ and equational constraints $Ax = b$ and the dual program with objective function $b^\top y$ and inequality constraints $A^\top y \geq c$ have interior feasible points x and y, which means that $x > 0$ and $s > 0$ (where $s = A^\top y - c$), then the above system of equations has a unique solution such that x is the unique maximizer of f_μ on U; see Matousek and Gardner [72] (Section 7.2, Lemma 7.2.1).

We now give an example illustrating Proposition 30.7, the *Support Vector Machine* (abbreviated as *SVM*).

30.3 Hard Margin Support Vector Machine; Version I

In this section we describe the following *classification problem*, or perhaps more accurately, *separation problem* (into two classes). Suppose we have two nonempty disjoint finite sets of p *blue* points $\{u_i\}_{i=1}^p$ and q *red* points $\{v_j\}_{j=1}^q$ in $\mathbb{R}^n$ (for simplicity, you may assume that these points are in the plane, that is, $n = 2$). Our goal is to find a hyperplane H of equation $w^\top x - b = 0$ (where $w \in \mathbb{R}^n$ is a nonzero vector and $b \in \mathbb{R}$), such that all the blue points u_i are in one of the two open half-spaces determined by H, and all the red points v_j are in the other open half-space determined by H; see Figure 30.11.

Without loss of generality, we may assume that

$$\begin{aligned} w^\top u_i - b &> 0 \qquad \text{for } i = 1, \ldots, p \\ w^\top v_j - b &< 0 \qquad \text{for } j = 1, \ldots, q. \end{aligned}$$

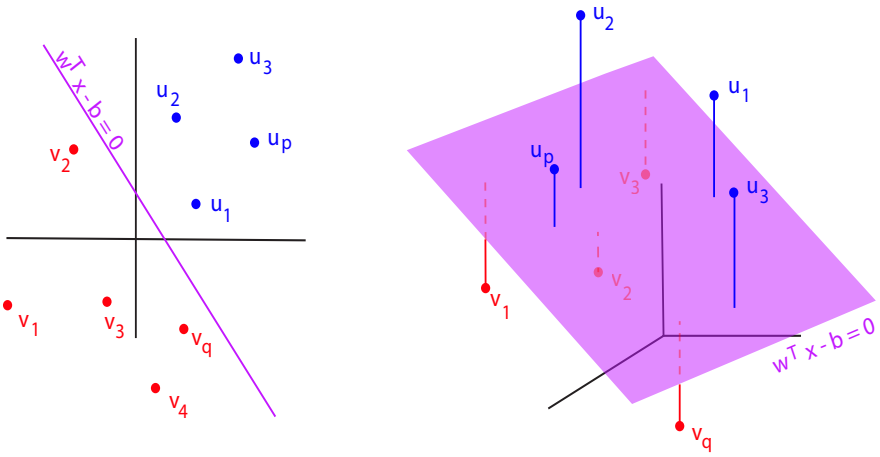


Figure 30.11: Two examples of the SVM separation problem. The left figure is SVM in $\mathbb{R}^2$, while the right figure is SVM in $\mathbb{R}^3$.

Of course, separating the blue and the red points may be impossible, as we see in Figure 30.12 for four points where the line segments (u_1, u_2) and (v_1, v_2) intersect. If a hyperplane separating the two subsets of blue and red points exists, we say that they are *linearly separable*.

Remark: Write $m = p + q$. The reader should be aware that in machine learning the classification problem is usually defined as follows. We assign m so-called *class labels* $y_k = \pm 1$ to the data points in such a way that $y_i = +1$ for each blue point u_i, and $y_{p+j} = -1$ for each red point v_j, and we denote the m points by x_k, where $x_k = u_k$ for $k = 1, \ldots, p$ and $x_k = v_{k-p}$ for $k = p+1, \ldots, p+q$. Then the classification constraints can be written as

$$y_k(w^\top x_k - b) > 0 \quad \text{for } k = 1, \ldots, m.$$

The set of pairs $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ is called a set of *training data* (or *training set*).

In the sequel, we will not use the above method, and we will stick to our two subsets of p *blue* points $\{u_i\}_{i=1}^p$ and q *red* points $\{v_j\}_{j=1}^q$.

Since there are infinitely many hyperplanes separating the two subsets (if indeed the two subsets are linearly separable), we would like to come up with a "good" criterion for choosing such a hyperplane.

The idea that was advocated by Vapnik (see Vapnik [110]) is to consider the distances $d(u_i, H)$ and $d(v_j, H)$ from *all* the points to the hyperplane H, and to pick a hyperplane H that maximizes the smallest of these distances. In machine learning this strategy is called finding a *maximal margin hyperplane*, or *hard margin support vector machine*, which definitely sounds more impressive.

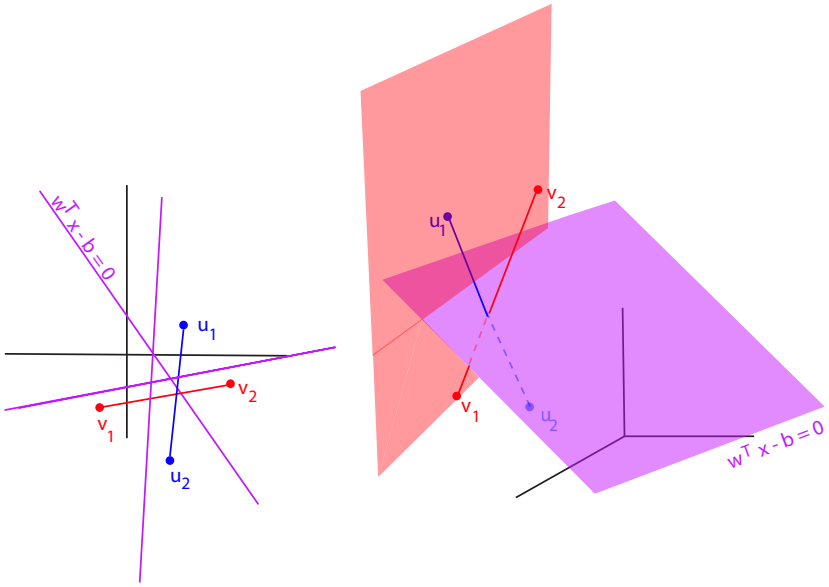


Figure 30.12: Two examples in which it is impossible to find purple hyperplanes which separate the red and blue points.

Since the distance from a point x to the hyperplane H of equation $w^\top x - b = 0$ is

$$d(x, H) = \frac{|w^\top x - b|}{\|w\|},$$

(where $\|w\| = \sqrt{w^\top w}$ is the Euclidean norm of w), it is convenient to temporarily assume that $\|w\| = 1$, so that

$$d(x, H) = |w^\top x - b|.$$

See Figure 30.13. Then with our sign convention, we have

$$\begin{aligned} d(u_i, H) &= w^\top u_i - b && i = 1, \ldots, p \\ d(v_j, H) &= -w^\top v_j + b && j = 1, \ldots, q. \end{aligned}$$

If we let

$$\delta = \min\{d(u_i, H),\ d(v_j, H) \mid 1 \leq i \leq p,\ 1 \leq j \leq q\},$$

then the hyperplane H should chosen so that

$$\begin{aligned} w^\top u_i - b &\geq \delta && i = 1, \ldots, p \\ -w^\top v_j + b &\geq \delta && j = 1, \ldots, q, \end{aligned}$$

and such that $\delta > 0$ is maximal. The distance δ is called the *margin* associated with the hyperplane H. This is indeed one way of formulating the two-class separation problem as an

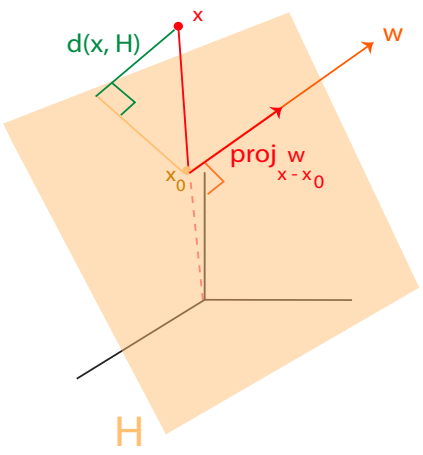


Figure 30.13: In $\mathbb{R}^3$, the distance from a point to the plane $w^\top x - b = 0$ is given by the projection onto the normal w.

optimization problem with a linear objective function $J(\delta, w, b) = \delta$, and affine and quadratic constraints (SVM_{h1}):

$$\begin{aligned}
&\text{maximize} \quad \delta \\
&\text{subject to} \\
&\qquad w^\top u_i - b \geq \delta \qquad i = 1, \ldots, p \\
&\qquad -w^\top v_j + b \geq \delta \qquad j = 1, \ldots, q \\
&\qquad \|w\| \leq 1.
\end{aligned}$$

Observe that the Problem (SVM_{h1}) has an optimal solution $\delta > 0$ iff the two subsets are linearly separable. We used the constraint $\|w\| \leq 1$ rather than $\|w\| = 1$ because the former is qualified, whereas the latter is not.

Actually, if (w, b, δ) is an optimal solution of Problem (SVM_{h1}), so in particular $\delta > 0$, then we claim that we must have $\|w\| = 1$. First, if $w = 0$, then we get the two inequalities

$$-b \geq \delta, \quad b \geq \delta,$$

which imply that $b \leq -\delta$ and $b \geq \delta$ for some positive δ, which is impossible. But then, if $w \neq 0$ and $\|w\| < 1$, by dividing both sides of the inequalities by $\|w\| < 1$ we would obtain the better solution $(w/\|w\|, b/\|w\|, \delta/\|w\|)$, since $\|w\| < 1$ implies that $\delta/\|w\| > \delta$.

We now prove that if the two subsets are linearly separable, then Problem (SVM_{h1}) has a unique optimal solution.

Theorem 30.10. *If two disjoint subsets of p blue points $\{u_i\}_{i=1}^p$ and q red points $\{v_j\}_{j=1}^q$ are linearly separable, then Problem* (SVM_{h1}) *has a unique optimal solution consisting of a*

hyperplane of equation $w^\top x - b = 0$ *separating the two subsets with maximum margin* δ. *Furthermore, if we define* $c_1(w)$ *and* $c_2(w)$ *by*

$$c_1(w) = \min_{1 \leq i \leq p} w^\top u_i$$
$$c_2(w) = \max_{1 \leq j \leq q} w^\top v_j,$$

then w *is the unique maximum of the function*

$$\rho(w) = \frac{c_1(w) - c_2(w)}{2}$$

over the convex subset U *of* $\mathbb{R}^n$ *given by the inequalities*

$$\begin{aligned} w^\top u_i - b &\geq \delta \qquad i = 1, \ldots, p \\ -w^\top v_j + b &\geq \delta \qquad j = 1, \ldots, q \\ \|w\| &\leq 1, \end{aligned}$$

and

$$b = \frac{c_1(w) + c_2(w)}{2}.$$

Proof. Our proof is adapted from Vapnik [110] (Chapter 10, Theorem 10.1). For any separating hyperplane H, since

$$\begin{aligned} d(u_i, H) &= w^\top u_i - b && i = 1, \ldots, p \\ d(v_j, H) &= -w^\top v_j + b && j = 1, \ldots, q, \end{aligned}$$

and since the smallest distance to H is

$$\begin{aligned} \delta &= \min\{d(u_i, H),\ d(v_j, H) \mid 1 \leq i \leq p,\ 1 \leq j \leq q\} \\ &= \min\{w^\top u_i - b,\ -w^\top v_j + b \mid 1 \leq i \leq p,\ 1 \leq j \leq q\} \\ &= \min\{\min\{w^\top u_i - b \mid 1 \leq i \leq p\}, \min\{-w^\top v_j + b \mid 1 \leq j \leq q\}\} \\ &= \min\{\min\{w^\top u_i \mid 1 \leq i \leq p\} - b\}, \min\{-w^\top v_j \mid 1 \leq j \leq q\} + b\} \\ &= \min\{\min\{w^\top u_i \mid 1 \leq i \leq p\} - b\}, -\max\{w^\top v_j \mid 1 \leq j \leq q\} + b\} \\ &= \min\{c_1(w) - b, -c_2(w) + b\}, \end{aligned}$$

in order for δ to be maximal we must have

$$c_1(w) - b = -c_2(w) + b,$$

which yields

$$b = \frac{c_1(w) + c_2(w)}{2}.$$

In this case,

$$c_1(w) - b = \frac{c_1(w) - c_2(w)}{2} = -c_2(w) + b,$$

so the maximum margin δ is indeed obtained when $\rho(w) = (c_1(w) - c_2(w))/2$ is maximal over U. Conversely, it is easy to see that any hyperplane of equation $w^\top x - b = 0$ associated with a w maximizing ρ over U and $b = (c_1(w) + c_2(w))/2$ is an optimal solution.

It remains to show that an optimal separating hyperplane exists and is unique. Since the unit ball is compact, U is compact, and since the function $w \mapsto \rho(w)$ is continuous, it achieves its maximum for some w_0 such that $\|w_0\| \leq 1$. Actually, we must have $\|w_0\| = 1$, since otherwise, by a familiar reasoning $w_0/\|w_0\|$ would be an even better solution. Therefore, w_0 is on the boundary of U. But ρ is a concave function (as an infimum of affine functions), so if it had two distinct maxima w_0 and w_0' with $\|w_0\| = \|w_0'\| = 1$, these would be global maxima since U is also convex, so we would have $\rho(w_0) = \rho(w_0')$ and then ρ would also have the same value along the segment (w_0, w_0') and in particular at $(w_0 + w_0')/2$, an interior point of U, a contradiction. □

We can proceed with the above formulation (SVM_{h1}) but there is a way to reformulate the problem so that the constraints are all affine, which might be preferable since they will be automatically qualified.

30.4 Hard Margin Support Vector Machine; Version II

Since $\delta > 0$ (otherwise the data would not be separable into two disjoint sets), we can divide the affine constraints by δ to obtain

$$\begin{aligned} w'^\top u_i - b' &\geq 1 && i = 1, \ldots, p \\ -w'^\top v_j + b' &\geq 1 && j = 1, \ldots, q, \end{aligned}$$

except that now, w' is not necessarily a unit vector. To obtain the distances to the hyperplane H, we need to divide by $\|w'\|$ and then we have

$$\begin{aligned} \frac{w'^\top u_i - b'}{\|w'\|} &\geq \frac{1}{\|w'\|} && i = 1, \ldots, p \\ \frac{-w'^\top v_j + b'}{\|w'\|} &\geq \frac{1}{\|w'\|} && j = 1, \ldots, q, \end{aligned}$$

which means that the shortest distance from the data points to the hyperplane is $1/\|w'\|$. Therefore, we wish to maximize $1/\|w'\|$, that is, to minimize $\|w'\|$, so we obtain the following optimization problem (SVM_{h2}):

Hard margin SVM (SVM_{h2}):

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\|w\|^2 \\
&\text{subject to} \\
&\qquad w^\top u_i - b \geq 1 \qquad i = 1, \ldots, p \\
&\qquad -w^\top v_j + b \geq 1 \qquad j = 1, \ldots, q.
\end{aligned}$$

The objective function $J(w) = 1/2\,\|w\|^2$ is convex, so Proposition 30.7 applies and gives us a necessary and sufficient condition for having a minimum in terms of the KKT conditions. First observe that the trivial solution $w = 0$ is impossible, because the blue constraints would be

$$-b \geq 1,$$

that is $b \leq -1$, and the red constraints would be

$$b \geq 1,$$

but these are contradictory. Our goal is to find w and b, and optionally, δ. We proceed in four steps first demonstrated on the following example.

Suppose that $p = q = n = 2$, so that we have two blue points

$$u_1^\top = (u_{11}, u_{12}) \qquad u_2^\top = (u_{21}, u_{22}),$$

two red points

$$v_1^\top = (v_{11}, v_{12}) \qquad v_2^\top = (v_{21}, v_{22}),$$

and

$$w^\top = (w_1, w_2).$$

Step 1: Write the constraints in matrix form. Let

$$C = \begin{pmatrix} -u_{11} & -u_{12} & 1 \\ -u_{21} & -u_{22} & 1 \\ v_{11} & v_{12} & -1 \\ v_{21} & v_{22} & -1 \end{pmatrix} \qquad d = \begin{pmatrix} -1 \\ -1 \\ -1 \\ -1 \end{pmatrix}. \tag{M}$$

The constraints become

$$C \begin{pmatrix} w \\ b \end{pmatrix} = \begin{pmatrix} -u_{11} & -u_{12} & 1 \\ -u_{21} & -u_{22} & 1 \\ v_{11} & v_{12} & -1 \\ v_{21} & v_{22} & -1 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ b \end{pmatrix} \leq \begin{pmatrix} -1 \\ -1 \\ -1 \\ -1 \end{pmatrix}. \tag{C}$$

Step 2: Write the objective function in matrix form.

$$J(w_1, w_2, b) = \frac{1}{2}\begin{pmatrix} w_1 & w_2 & b \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} w_1 \\ w_2 \\ b \end{pmatrix}. \tag{O}$$

Step 3: Apply Proposition 30.7 to solve for w in terms of λ and μ. We obtain

$$\begin{pmatrix} w_1 \\ w_2 \\ 0 \end{pmatrix} + \begin{pmatrix} -u_{11} & -u_{21} & v_{11} & v_{21} \\ -u_{12} & -u_{22} & v_{12} & v_{22} \\ 1 & 1 & -1 & -1 \end{pmatrix}\begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

i.e.

$$\nabla J_{(w,b)} + C^\top \begin{pmatrix} \lambda \\ \mu \end{pmatrix} = 0_{n+1}.$$

Then

$$\begin{pmatrix} w_1 \\ w_2 \\ 0 \end{pmatrix} = \begin{pmatrix} u_{11} & u_{21} & -v_{11} & -v_{21} \\ u_{12} & u_{22} & -v_{12} & -v_{22} \\ -1 & -1 & 1 & 1 \end{pmatrix}\begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \mu_1 \\ \mu_2 \end{pmatrix},$$

which implies

$$w = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \lambda_1 \begin{pmatrix} u_{11} \\ u_{12} \end{pmatrix} + \lambda_2 \begin{pmatrix} u_{21} \\ u_{22} \end{pmatrix} - \mu_1 \begin{pmatrix} v_{11} \\ v_{12} \end{pmatrix} - \mu_2 \begin{pmatrix} u_{21} \\ u_{22} \end{pmatrix} \tag{$*_1$}$$

with respect to

$$\mu_1 + \mu_2 - \lambda_1 - \lambda_2 = 0. \tag{$*_2$}$$

Step 4: Rewrite the constraints at (C) using $(*_1)$. In particular $C\begin{pmatrix} w \\ b \end{pmatrix} \leq d$ becomes

$$\begin{pmatrix} -u_{11} & -u_{12} & 1 \\ -u_{21} & -u_{22} & 1 \\ v_{11} & v_{12} & -1 \\ v_{21} & v_{22} & -1 \end{pmatrix}\begin{pmatrix} u_{11} & u_{21} & -v_{11} & -v_{21} & 0 \\ u_{12} & u_{22} & -v_{21} & -v_{22} & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \mu_1 \\ \mu_2 \\ b \end{pmatrix} \leq \begin{pmatrix} -1 \\ -1 \\ -1 \\ -1 \end{pmatrix}.$$

Rewriting the previous equation in "block" format gives us

$$-\begin{pmatrix} -u_{11} & -u_{12} \\ -u_{21} & -u_{22} \\ v_{11} & v_{12} \\ v_{21} & v_{22} \end{pmatrix}\begin{pmatrix} -u_{11} & -u_{21} & v_{11} & v_{21} \\ -u_{12} & -u_{22} & v_{21} & v_{22} \end{pmatrix}\begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \mu_1 \\ \mu_2 \end{pmatrix} + b\begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \leq \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix},$$

which with the definition

$$X = \begin{pmatrix} -u_{11} & -u_{21} & v_{11} & v_{21} \\ -u_{12} & -u_{22} & v_{21} & v_{22} \end{pmatrix}$$

yields

$$-X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + b \begin{pmatrix} \mathbf{1}_p \\ -\mathbf{1}_q \end{pmatrix} + \mathbf{1}_{p+q} \leq 0_{p+q}. \tag{$*_3$}$$

Let us now consider the general case.

Step 1: Write the constraints in matrix form. First we rewrite the constraints as

$$\begin{aligned} -u_i^\top w + b &\leq -1 && i = 1, \ldots, p \\ v_j^\top w - b &\leq -1 && j = 1, \ldots, q, \end{aligned}$$

and we get the $(p+q) \times (n+1)$ matrix C and the vector $d \in \mathbb{R}^{p+q}$ given by

$$C = \begin{pmatrix} -u_1^\top & 1 \\ \vdots & \vdots \\ -u_p^\top & 1 \\ v_1^\top & -1 \\ \vdots & \vdots \\ v_q^\top & -1 \end{pmatrix}, \quad d = \begin{pmatrix} -1 \\ \vdots \\ -1 \end{pmatrix},$$

so the set of inequality constraints is

$$C \begin{pmatrix} w \\ b \end{pmatrix} \leq d.$$

Step 2: The objective function in matrix form is given by

$$J(w, b) = \frac{1}{2} \begin{pmatrix} w^\top & b \end{pmatrix} \begin{pmatrix} I_n & 0_n \\ 0_n^\top & 0 \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix}.$$

Note that the corresponding matrix is symmetric positive semidefinite, but it is *not* invertible. Thus the function J is convex but not strictly convex. This will cause some minor trouble in finding the dual function of the problem.

Step 3: If we introduce the generalized Lagrange multipliers $\lambda \in \mathbb{R}^p$ and $\mu \in \mathbb{R}^q$, according to Proposition 30.7, the first KKT condition is

$$\nabla J_{(w,b)} + C^\top \begin{pmatrix} \lambda \\ \mu \end{pmatrix} = 0_{n+1},$$

with $\lambda \geq 0, \mu \geq 0$. By the result of Example 19.4,

$$\nabla J_{(w,b)} = \begin{pmatrix} I_n & 0_n \\ 0_n^\top & 0 \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} = \begin{pmatrix} w \\ 0 \end{pmatrix},$$

so we get

$$\begin{pmatrix} w \\ 0 \end{pmatrix} = -C^\top \begin{pmatrix} \lambda \\ \mu \end{pmatrix},$$

that is,

$$\begin{pmatrix} w \\ 0 \end{pmatrix} = \begin{pmatrix} u_1 & \cdots & u_p & -v_1 & \cdots & -v_q \\ -1 & \cdots & -1 & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \lambda \\ \mu \end{pmatrix}.$$

Consequently,

$$w = \sum_{i=1}^{p} \lambda_i u_i - \sum_{j=1}^{q} \mu_j v_j, \tag{$*_1$}$$

and

$$\sum_{j=1}^{q} \mu_j - \sum_{i=1}^{p} \lambda_i = 0. \tag{$*_2$}$$

Step 4: Rewrite the constrainst using $(*_1)$. Plugging the above expression for w into the constraints $C \begin{pmatrix} w \\ b \end{pmatrix} \leq d$ we get

$$\begin{pmatrix} -u_1^\top & 1 \\ \vdots & \vdots \\ -u_p^\top & 1 \\ v_1^\top & -1 \\ \vdots & \vdots \\ v_q^\top & -1 \end{pmatrix} \begin{pmatrix} u_1 & \cdots & u_p & -v_1 & \cdots & -v_q & 0_n \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 1 \end{pmatrix} \begin{pmatrix} \lambda \\ \mu \\ b \end{pmatrix} \leq \begin{pmatrix} -1 \\ \vdots \\ -1 \end{pmatrix},$$

so if let X be the $n \times (p+q)$ matrix given by

$$X = \begin{pmatrix} -u_1 & \cdots & -u_p & v_1 & \cdots & v_q \end{pmatrix},$$

we obtain

$$w = -X \begin{pmatrix} \lambda \\ \mu \end{pmatrix}, \tag{$*_1'$}$$

and the above inequalities are written in matrix form as

$$\begin{pmatrix} X^\top & \mathbf{1}_p \\ & -\mathbf{1}_q \end{pmatrix} \begin{pmatrix} -X & 0_n \\ 0_{p+q}^\top & 1 \end{pmatrix} \begin{pmatrix} \lambda \\ \mu \\ b \end{pmatrix} \leq -\mathbf{1}_{p+q};$$

that is,

$$-X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + b \begin{pmatrix} \mathbf{1}_p \\ -\mathbf{1}_q \end{pmatrix} + \mathbf{1}_{p+q} \leq 0_{p+q}. \tag{$*_3$}$$

Equivalently, the ith inequality is

$$-\sum_{j=1}^{p} u_i^\top u_j \lambda_j + \sum_{k=1}^{q} u_i^\top v_k \mu_k + b + 1 \leq 0 \qquad i = 1, \ldots, p,$$

and the $(p+j)$th inequality is

$$\sum_{i=1}^{p} v_j^\top u_i \lambda_i - \sum_{k=1}^{q} v_j^\top v_k \mu_k - b + 1 \leq 0 \qquad j = 1, \ldots, q.$$

We also have $\lambda \geq 0, \mu \geq 0$. Furthermore, if the ith inequality is inactive then $\lambda_i = 0$, and if the $(p+j)$th inequality is inactive then $\mu_j = 0$. Since the constraints are affine and since J is convex, if we can find $\lambda \geq 0$, $\mu \geq 0$, and b such that the inequalities in $(*_3)$ are satisfied, and $\lambda_i = 0$ and $\mu_j = 0$ when the corresponding constraint is inactive, then by Proposition 30.7 we have an optimum solution.

Remark: The second KKT condition can be written as

$$\begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \left(-X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + b \begin{pmatrix} \mathbf{1}_p \\ -\mathbf{1}_q \end{pmatrix} + \mathbf{1}_{p+q} \right) = 0;$$

that is,

$$-\begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + b \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \begin{pmatrix} \mathbf{1}_p \\ -\mathbf{1}_q \end{pmatrix} + \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \mathbf{1}_{p+q} = 0.$$

Since $(*_2)$ says that $\sum_{i=1}^{p} \lambda_i = \sum_{j=1}^{q} \mu_j$, the second term is zero, and by $(*_1')$ we get

$$w^\top w = \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} = \sum_{i=1}^{p} \lambda_i + \sum_{j=1}^{q} \mu_j.$$

Thus we obtain a simple expression for $\|w\|^2$ in terms of λ and μ.

The vectors u_i and v_j for which the i-th inequality is active and the $(p+j)$th inequality is active are called *support vectors*. For every vector u_i or v_j that is not a support vector, the corresponding inequality is inactive so $\lambda_i = 0$ and $\mu_j = 0$. Thus we see that only the support vectors contribute to a solution. If we can guess which vectors u_i and v_j are support vectors, namely, those for which $\lambda_i \neq 0$ and $\mu_j \neq 0$, then for each support vector u_i we have an equation

$$-\sum_{j=1}^{p} u_i^\top u_j \lambda_j + \sum_{k=1}^{q} u_i^\top v_k \mu_k + b + 1 = 0,$$

and for each support vector v_j we have an equation

$$\sum_{i=1}^{p} v_j^\top u_i \lambda_i - \sum_{k=1}^{q} v_j^\top v_k \mu_k - b + 1 = 0,$$

with $\lambda_i = 0$ and $\mu_j = 0$ for all non-support vectors, so together with the Equation $(*_2)$ we have a linear system with an equal number of equations and variables, which is solvable if our separation problem has a solutio. Thus,in principle we can find λ, μ, and b by solving a linear system.

Remark: We can first solve for λ and μ (by eliminating b), and by $(*_1)$ and since $w \neq 0$, there is a least some nonzero λ_{i_0} and thus some nonzero μ_{j_0}, so the corresponding inequalities are equations

$$-\sum_{j=1}^{p} u_{i_0}^\top u_j \lambda_j + \sum_{k=1}^{q} u_{i_0}^\top v_k \mu_k + b + 1 = 0$$
$$\sum_{i=1}^{p} v_{j_0}^\top u_i \lambda_i - \sum_{k=1}^{q} v_{j_0}^\top v_k \mu_k - b + 1 = 0,$$

so b is given in terms of λ and μ by

$$b = \frac{1}{2}(u_{i_0}^\top + v_{j_0}^\top)\left(\sum_{i=1}^{p} \lambda_i u_i - \sum_{j=1}^{p} \mu_j v_j\right).$$

Using the dual of the Lagrangian, we can solve for λ and μ, but typically b is not determined, so we use the above method to find b.

The above nondeterministic procedure in which we guess which vectors are support vectors is not practical. We will see later that a practical method for solving for λ and μ consists in maximizing the dual of the Lagrangian.

If w is an optimal solution, then $\delta = 1/\left\|w\right\|$ is the shortest distance from the support vectors to the separating hyperplane $H_{w,b}$ of equation $w^\top x - b = 0$. If we consider the two hyperplanes $H_{w,b+1}$ and $H_{w,b-1}$ of equations

$$w^\top x - b - 1 = 0 \quad \text{and} \quad w^\top x - b + 1 = 0,$$

then $H_{w,b+1}$ and $H_{w,b-1}$ are two hyperplanes parallel to the hyperplane $H_{w,b}$ and the distance between them is 2δ. Furthermore, $H_{w,b+1}$ contains the support vectors u_i, $H_{w,b-1}$ contains the support vectors v_j, and there are no data points u_i or v_j in the open region between these two hyperplanes containing the separating hyperplane $H_{w,b}$ (called a "slab" by Boyd and Vandenberghe; see [22], Section 8.6). This situation is illustrated in Figure 30.14.

Even if $p = 1$ and $q = 2$, a solution is not obvious. In the plane, there are four possibilities:

(1) If u_1 is on the segment (v_1, v_2), there is no solution.

(2) If the projection h of u_1 onto the line determined by v_1 and v_2 is between v_1 and v_2, that is $h = (1-\alpha)v_1 + \alpha_2 v_2$ with $0 \leq \alpha \leq 1$, then it is the line parallel to $v_2 - v_1$ and equidistant to u and both v_1 and v_2, as illustrated in Figure 30.15.

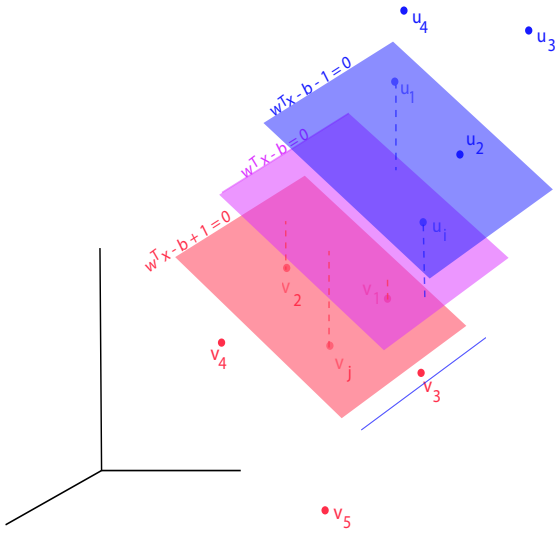


Figure 30.14: In $\mathbb{R}^3$, the solution to the hard margin SVM is the purple plane sandwiched between the red plane $w^\top x - b + 1 = 0$ and the blue plane $w^\top x - b - 1 = 0$, each of which contains the appropriate support vectors u_i and v_j.

(3) If the projection h of u_1 onto the line determined by v_1 and v_2 is to the right of v_2, that is $h = (1-\alpha)v_1 + \alpha_2 v_2$ with $\alpha > 1$, then it is the bisector of the line segment (u_1, v_2).

(4) If the projection h of u_1 onto the line determined by v_1 and v_2 is to the left of v_1, that is $h = (1-\alpha)v_1 + \alpha_2 v_2$ with $\alpha < 0$, then it is the bisector of the line segment (u_1, v_1).

If $p = q = 1$, we can find a solution explicitly. Then $(*_2)$ yields

$$\lambda = \mu,$$

and if we guess that the constraints are active, the corresponding equality constraints are

$$\begin{aligned} -u^\top u\lambda + u^\top v\mu + b + 1 &= 0 \\ u^\top v\lambda - v^\top v\mu - b + 1 &= 0, \end{aligned}$$

so we get

$$\begin{aligned} (-u^\top u + u^\top v)\lambda + b + 1 &= 0 \\ (u^\top v - v^\top v)\lambda - b + 1 &= 0, \end{aligned}$$

Adding up the two equations we find

$$(2u^\top v - u^\top u - v^\top v)\lambda + 2 = 0,$$

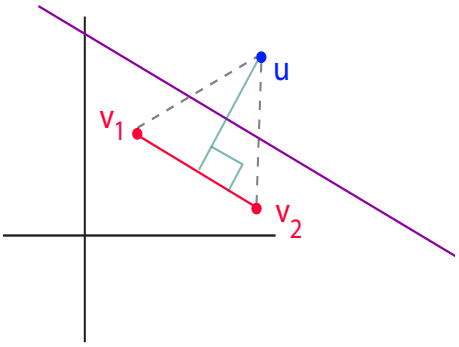


Figure 30.15: The purple line, which is the bisector of the altitude of the isosceles triangle, separates the two red points from the blue point in a manner which satisfies the hard margin SVM.

that is

$$\lambda = \frac{2}{(u-v)^\top (u-v)}.$$

By subtracting the first equation from the second, we find

$$(u^\top u - v^\top v)\lambda - 2b = 0,$$

which yields

$$b = \lambda \frac{(u^\top u - v^\top v)}{2} = \frac{u^\top u - v^\top v}{(u-v)^\top (u-v)}.$$

Then by $(*_1)$ we obtain

$$w = \frac{2(u-v)}{(u-v)^\top (u-v)}.$$

We verify easily that

$$2(u_1 - v_1)x_1 + \cdots + 2(u_n - v_n)x_n = (u_1^2 + \cdots + u_n^2) - (v_1^2 + \cdots + v_n^2)$$

is the equation of the bissector hyperplane between u and v; see Figure 30.16.

In the next section we will derive the dual of the optimization problem discussed in this section. We will also consider a more flexible solution invovling a *soft margin*.

30.5 Lagrangian Duality and Saddle Points

In this section we investigate methods to solve the *minimization problem* (P):

$$\begin{aligned} &\text{minimize} && J(v) \\ &\text{subject to} && \varphi_i(v) \leq 0, \quad i = 1, \ldots, m. \end{aligned}$$

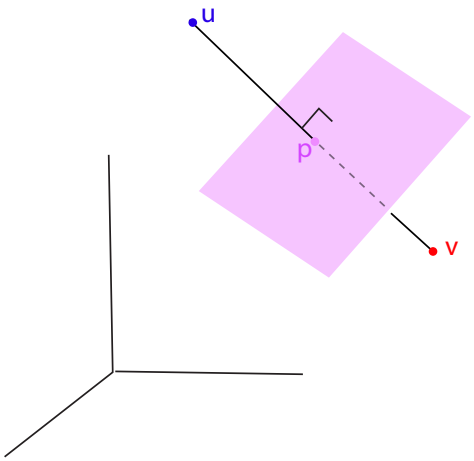


Figure 30.16: In $\mathbb{R}^3$, the solution to the hard margin SVM for the points u and v is the purple perpendicular planar bisector of $u - v$.

It turns out that under certain conditions the original problem (P), called *primal problem*, can be solved in two stages with the help another problem (D), called the *dual problem*. The dual problem (D) is a *maximization problem* involving a function G, called the *Lagrangian dual*, and it is obtained by *minimizing* the *Lagrangian* $L(v, \mu)$ of Problem (P) over the variable $v \in \mathbb{R}^n$, holding μ fixed, where $L\colon \Omega \times \mathbb{R}^m_+ \to \mathbb{R}$ is given by

$$L(v, \mu) = J(v) + \sum_{i=1}^{m} \mu_i \varphi_i(v),$$

with $\mu \in \mathbb{R}^m_+$.

The two steps of the method are:

(1) Find the dual function $\mu \mapsto G(\mu)$ explictly by solving the minimization problem of finding the minimum of $L(v, \mu)$ with respect to $v \in \Omega$, holding μ fixed. This is an unconstrained minimization problem (with $v \in \Omega$). If we are lucky, a unique minimizer u_μ such that $G(\mu) = L(u_\mu, \mu)$ can be found. We will address the issue of uniqueness later on.

(2) Solve the maximization problem of finding the maximum of the function $\mu \mapsto G(\mu)$ over all $\mu \in \mathbb{R}^m_+$. This is basically an unconstrained problem, except for the fact that $\mu \in \mathbb{R}^m_+$.

If steps (1) and (2) are successful, under some suitable conditions on the function J and the constraints φ_i (for example, if they are convex), for any solution $\lambda \in \mathbb{R}^m_+$ obtained in

step (2), the vector u_λ obtained in step (1) is an optimal solution of Problem (P). This is proved in Theorem 30.14.

In order to prove Theorem 30.14, which is our main result, we need two intermediate technical results of independent interest involving the notion of saddle point.

The local minima of a function $J\colon \Omega \to \mathbb{R}$ over a domain U defined by inequality constraints are saddle points of the Lagrangian $L(u,\mu)$ associated with J and the constraints φ_i. Then, under some mild hypotheses, the set of solutions of the minimization problem (P)

$$\begin{aligned} &\text{minimize} \quad J(v) \\ &\text{subject to} \quad \varphi_i(v) \leq 0, \quad i = 1, \ldots, m \end{aligned}$$

coincides with the set of first arguments of the saddle points of the Lagrangian

$$L(v,\mu) = J(v) + \sum_{i=1}^{m} \mu_i \varphi_i(v).$$

This is proved in Theorem 30.12. To prove Theorem 30.14, we also need Proposition 30.11, a basic property of saddle points.

Definition 30.7. Let $L\colon \Omega \times M \to \mathbb{R}$ be a function defined on a set of the form $\Omega \times M$. A point $(u,\lambda) \in \Omega \times M$ is a *saddle point* of L if u is a minimum of the function $L(-,\lambda)\colon \Omega \to \mathbb{R}$ given by $v \mapsto L(v,\lambda)$ for all $v \in \Omega$ and λ fixed, and λ is a maximum of the function $L(u,-)\colon M \to \mathbb{R}$ given by $\mu \mapsto L(u,\mu)$ for all $\mu \in M$ and u fixed; equivalently,

$$\sup_{\mu \in M} L(u,\mu) = L(u,\lambda) = \inf_{v \in \Omega} L(v,\lambda).$$

Note that the order of the arguments u and λ is important. The second set M will be the set of generalized *m*ultipliers, and this is why we use the symbol M

A saddle point is often depicted as a mountain pass, which explains the terminology; see Figure 30.17. However, this is a bit misleading since other situations are possible; see Figure 30.18.

Proposition 30.11. *If (u,λ) is a saddle point of a function $L\colon \Omega \times M \to \mathbb{R}$, then*

$$\sup_{\mu \in M} \inf_{v \in \Omega} L(v,\mu) = L(u,\lambda) = \inf_{v \in \Omega} \sup_{\mu \in M} L(v,\mu).$$

Proof. First we prove that the following inequality always holds:

$$\sup_{\mu \in M} \inf_{v \in \Omega} L(v,\mu) \leq \inf_{v \in \Omega} \sup_{\mu \in M} L(v,\mu). \tag{$*_1$}$$

Pick any $w \in \Omega$ and any $\rho \in M$. By definition of inf (the greatest lower bound) and sup (the least upper bound), we have

$$\inf_{v \in \Omega} L(v,\rho) \leq L(w,\rho) \leq \sup_{\mu \in M} L(w,\mu).$$

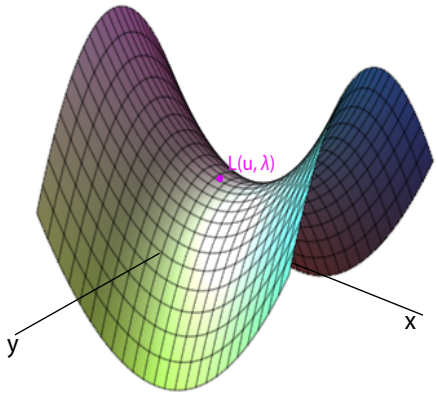


Figure 30.17: A three-dimensional rendition of a saddle point $L(u, \lambda)$ for the function $L(u, \lambda) = u^2 - \lambda^2$. The plane $x = u$ provides a maximum as the apex of a downward opening parabola, while the plane $y = \lambda$ provides a minimum as the apex of an upward opening parabola.

The cases where $\inf_{v\in\Omega} L(v, \rho) = -\infty$ or where $\sup_{\mu\in M} L(w, \mu) = +\infty$ may arise, but this is not a problem. Since

$$\inf_{v\in\Omega} L(v, \rho) \leq \sup_{\mu\in M} L(w, \mu)$$

and the right-hand side is independent of ρ, it is an upper bound of the left-hand side for all ρ, so

$$\sup_{\mu\in M} \inf_{v\in\Omega} L(v, \mu) \leq \sup_{\mu\in M} L(w, \mu).$$

Since the left-hand side is independent of w, it is a lower bound for the right-hand side for all w, so we obtain $(*_1)$:

$$\sup_{\mu\in M} \inf_{v\in\Omega} L(v, \mu) \leq \inf_{v\in\Omega} \sup_{\mu\in M} L(v, \mu).$$

To obtain the reverse inequality, we use the fact that (λ, μ) is a saddle point, so

$$\inf_{v\in\Omega} \sup_{\mu\in M} L(v, \mu) \leq \sup_{\mu\in M} L(u, \mu) = L(u, \lambda)$$

and

$$L(u, \lambda) = \inf_{v\in\Omega} L(v, \lambda) \leq \sup_{\mu\in M} \inf_{v\in\Omega} L(v, \mu),$$

and these imply that

$$\inf_{v\in\Omega} \sup_{\mu\in M} L(v, \mu) \leq \sup_{\mu\in M} \inf_{v\in\Omega} L(v, \mu), \qquad (*_2)$$

as desired. □

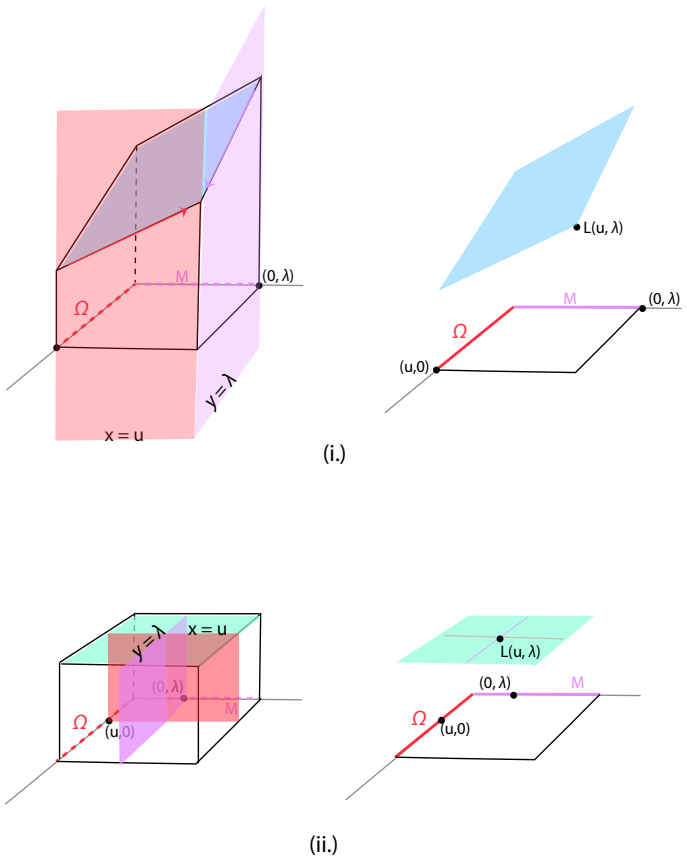


Figure 30.18: Let $\Omega = \{[t, 0, 0] \mid 0 \leq t \leq 1\}$ and $M = \{[0, t, 0] \mid 0 \leq t \leq 1\}$. In Figure $(i.)$, $L(u, \lambda)$ is the blue slanted quadrilateral whose forward vertex is a saddle point. In Figure $(ii.)$, $L(u, \lambda)$ is the planar green rectangle composed entirely of saddle points.

We now return to our main minimization problem (P):

$$\begin{aligned} &\text{minimize} \quad J(v) \\ &\text{subject to} \quad \varphi_i(v) \leq 0, \quad i = 1, \ldots, m, \end{aligned}$$

where $J\colon \Omega \to \mathbb{R}$ and the constraints $\varphi_i\colon \Omega \to \mathbb{R}$ are some functions defined on some open subset Ω of some finite-dimensional Euclidean vector space V (more generally, a real Hilbert space V).

Definition 30.8. The *Lagrangian* of the minimization problem (P) defined above is the function $L\colon \Omega \times \mathbb{R}^m_+ \to \mathbb{R}$ given by

$$L(v, \mu) = J(v) + \sum_{i=1}^{m} \mu_i \varphi_i(v),$$

with $\mu = (\mu_1, \ldots, \mu_m)$. The numbers μ_i are called *generalized Lagrange multipliers*.

The following theorem shows that under some suitable conditions, every solution u of the Problem (P) is the first argument of a saddle point (u, λ) of the Lagrangian L, and conversely, if (u, λ) is a saddle point of the Lagrangian L, then u is a solution of the Problem (P).

Theorem 30.12. *Consider Problem (P) defined above where $J\colon \Omega \to \mathbb{R}$ and the constraints $\varphi_i\colon \Omega \to \mathbb{R}$ are some functions defined on some open subset Ω of some finite-dimensional Euclidean vector space V (more generally, a real Hilbert space V). The following facts hold.*

(1) If $(u, \lambda) \in \Omega \times \mathbb{R}^m_+$ is a saddle point of the Lagrangian L associated with Problem (P), then $u \in U$, u is a solution of Problem (P), and $J(u) = L(u, \lambda)$.

(2) If Ω is convex (open), if the functions φ_i $(1 \leq i \leq m)$ and J are convex and differentiable at the point $u \in U$, if the constraints are qualified, and if $u \in U$ is a minimum of Problem (P), then there exists some vector $\lambda \in \mathbb{R}^m_+$ such that the pair $(u, \lambda) \in \Omega \times \mathbb{R}^m_+$ is a saddle point of the Lagrangian L.

Proof. (1) Since (u, λ) is a saddle point of L we have $\sup_{\mu \in M} L(u, \mu) = L(u, \lambda)$ which implies that $L(u, \mu) \leq L(u, \lambda)$ for all $\mu \in \mathbb{R}^m_+$, which means that

$$J(u) + \sum_{i=1}^m \mu_i \varphi_i(u) \leq J(u) + \sum_{i=1}^m \lambda_i \varphi_i(u),$$

that is,

$$\sum_{i=1}^m (\mu_i - \lambda_i)\varphi_i(u) \leq 0 \quad \text{for all } \mu \in \mathbb{R}^m_+.$$

If we let each μ_i be large enough, then $\mu_i - \lambda_i > 0$, and if we had $\varphi_i(u) > 0$ then the term $(\mu_i - \lambda_i)\varphi_i(u)$ could be made arbitrarily large and positive, so we conclude that $\varphi_i(u) \leq 0$ for $i = 1, \ldots, m$, and consequently, $u \in U$. For $\mu = 0$, we conclude that $\sum_{i=1}^m \lambda_i \varphi_i(u) \geq 0$, while since $\lambda_i \geq 0$ and $\varphi_i(u) \leq 0$ we have $\sum_{i=1}^m \lambda_i \varphi_i(u) \leq 0$, so we must have $u \in U$ and

$$\sum_{i=1}^m \lambda_i \varphi_i(u) = 0. \tag{$*_1$}$$

This shows that $J(u) = L(u, \lambda)$. Since the inequality $L(u, \lambda) \leq L(v, \lambda)$ is

$$J(u) + \sum_{i=1}^m \lambda_i \varphi_i(u) \leq J(v) + \sum_{i=1}^m \lambda_i \varphi_i(v),$$

by $(*_1)$ we obtain

$$\begin{aligned} J(u) &\leq J(v) + \sum_{i=1}^m \lambda_i \varphi_i(v) && \text{for all } v \in \Omega \\ &\leq J(v) && \text{for all } v \in U \text{ (since } \varphi_i(v) \leq 0 \text{ and } \lambda_i \geq 0), \end{aligned}$$

which shows that u is a minimum of J on U.

(2) The hypotheses required to apply Theorem 30.6(1) are satisfied. Consequently if $u \in U$ is a solution of Problem (P), then there exists some vector $\lambda \in \mathbb{R}^m_+$ such that the KKT conditions hold:

$$J'(u) + \sum_{i=1}^{m} \lambda_i (\varphi_i')_u = 0 \quad \text{and} \quad \sum_{i=1}^{m} \lambda_i \varphi_i(u) = 0.$$

The second equation yields

$$L(u, \mu) = J(u) + \sum_{i=1}^{m} \mu_i \varphi_i(u) \leq J(u) = J(u) + \sum_{i=1}^{m} \lambda_i \varphi_i(u) = L(u, \lambda),$$

that is,

$$L(u, \mu) \leq L(u, \lambda) \quad \text{for all } \mu \in \mathbb{R}^m_+ \tag{$*_2$}$$

(since $\varphi_i(u) \leq 0$ as $u \in U$), and since the function $v \mapsto J(v) + \sum_{i=1} \lambda_i \varphi_i(v) = L(v, \lambda)$ is convex as a sum of convex functions, by Theorem 20.11(4), the first equation is a sufficient condition for the existence of minimum. Consequently,

$$L(u, \lambda) \leq L(v, \lambda) \quad \text{for all } v \in \Omega, \tag{$*_3$}$$

and $(*_2)$ and $(*_3)$ show that (u, λ) is a saddle point of L. □

To recap what we just proved, under some mild hypotheses, the set of solutions of the minimization Problem (P)

$$\begin{aligned} &\text{minimize} \quad J(v) \\ &\text{subject to} \quad \varphi_i(v) \leq 0, \quad i = 1, \ldots, m \end{aligned}$$

coincides with the set of first arguments of the saddle points of the Lagrangian

$$L(v, \mu) = J(v) + \sum_{i=1}^{m} \mu_i \varphi_i(v),$$

and for any optimum $u \in U$ of Problem (P) we have $J(u) = L(u, \lambda)$.

Therefore, if we knew some particular second argument λ of these saddle points, then the *constrained* problem (P) would be replaced by the *unconstrained* problem (P_λ):

$$\begin{aligned} &\text{find } u_\lambda \in \Omega \text{ such that} \\ &L(u_\lambda, \lambda) = \inf_{v \in \Omega} L(v, \lambda). \end{aligned}$$

How do we find such an element $\lambda \in \mathbb{R}^m_+$?

For this, remember that for a saddle point (u_λ, λ), by Proposition 30.11, we have

$$L(u_\lambda, \lambda) = \inf_{v \in \Omega} L(v, \lambda) = \sup_{\mu \in \mathbb{R}^m_+} \inf_{v \in \Omega} L(v, \mu),$$

so we are naturally led to introduce the function $G\colon \mathbb{R}^m_+ \to \mathbb{R}$ given by

$$G(\mu) = \inf_{v \in \Omega} L(v, \mu) \quad \mu \in \mathbb{R}^m_+,$$

and then λ will be a solution of the problem

$$\begin{aligned} &\text{find } \lambda \in \mathbb{R}^m_+ \text{ such that} \\ &G(\lambda) = \sup_{\mu \in \mathbb{R}^m_+} G(\mu), \end{aligned}$$

which is equivalent to the maximization problem (D):

$$\begin{aligned} &\text{maximize} \quad G(\mu) \\ &\text{subject to} \quad \mu \in \mathbb{R}^m_+. \end{aligned}$$

Definition 30.9. Given the minimization problem (P)

$$\begin{aligned} &\text{minimize} \quad J(v) \\ &\text{subject to} \quad \varphi_i(v) \leq 0, \quad i = 1, \ldots, m, \end{aligned}$$

where $J\colon \Omega \to \mathbb{R}$ and the constraints $\varphi_i\colon \Omega \to \mathbb{R}$ are some functions defined on some open subset Ω of some finite-dimensional Euclidean vector space V (more generally, a real Hilbert space V), the function $G\colon \mathbb{R}^m_+ \to \mathbb{R}$ given by

$$G(\mu) = \inf_{v \in \Omega} L(v, \mu) \quad \mu \in \mathbb{R}^m_+,$$

is called the *Lagrange dual function* (or simply *dual function*). The problem (D)

$$\begin{aligned} &\text{maximize} \quad G(\mu) \\ &\text{subject to} \quad \mu \in \mathbb{R}^m_+ \end{aligned}$$

is called the *Lagrange dual problem*. The problem (P) is often called the *primal problem*, and (D) is the *dual problem*. The variable μ is called the *dual variable*. The variable $\mu \in \mathbb{R}^m_+$ is said to be *dual feasible* if $G(\mu)$ is defined (not $-\infty$). If $\lambda \in \mathbb{R}^m_+$ is a maximum of G, then we call it a *dual optimal* or an *optimal Lagrange multiplier*.

Since

$$L(v, \mu) = J(v) + \sum_{i=1}^m \mu_i \varphi_i(v),$$

the function $G(\mu) = \inf_{v\in\Omega} L(v,\mu)$ is the pointwise infimum of some affine functions of μ, so it it *concave*, even if the φ_i are not convex. One of the main advantages of the dual problem over the primal problem is that it is a *convex optimization problem*, since we wish to maximize a concave objective function G (thus minimize $-G$, a convex function), and the constraints $\mu \geq 0$ are convex. In a number of practical situations the dual function G can indeed be computed.

To be perfectly rigorous we should mention that the dual function G is actually a *partial function*, because it takes the value $-\infty$ when the map $v \mapsto L(v,\mu)$ is unbounded below.

Example 30.5. Consider the linear program (P)

$$\begin{aligned} &\text{minimize} \quad c^\top x \\ &\text{subject to} \quad Ax \leq b, \;\; x \geq 0, \end{aligned}$$

where A is an $m \times n$ matrix. The constraints $x \geq 0$ are rewritten as $-x_i \leq 0$, so we introduce Lagrange multipliers $\mu \in \mathbb{R}^m_+$ and $\nu \in \mathbb{R}^n_+$, and we have the Lagrangian

$$\begin{aligned} L(v,\mu,\nu) &= c^\top v + \mu^\top(Av - b) - \nu^\top v \\ &= -b^\top\mu + (c + A^\top\mu - \nu)^\top v. \end{aligned}$$

The linear function $v \mapsto (c + A^\top\mu - \nu)^\top v$ is unbounded below unless $c + A^\top\mu - \nu = 0$, so the dual function $G(\mu,\nu) = \inf_{v\in\mathbb{R}^n} L(v,\mu,\nu)$ is given for all $\mu \geq 0$ and $\nu \geq 0$ by

$$G(\mu,\nu) = \begin{cases} -b^\top\mu & \text{if } A^\top\mu - \nu + c = 0, \\ -\infty & \text{otherwise.} \end{cases}$$

The domain if G is a proper subset of $\mathbb{R}^m_+ \times \mathbb{R}^n_+$.

Observe that the value $G(\mu,\nu)$ of the function G, when it is defined, is independent of the second argument ν. Since we are interested in maximizing G, this suggests introducing at the function $\widehat{G}$ of the single argument μ given by

$$\widehat{G}(\mu) = -b^\top\mu,$$

which is defined for all $\mu \in \mathbb{R}^m_+$.

Of course, $\sup_{\mu\in\mathbb{R}^m_+} \widehat{G}(\mu)$ and $\sup_{(\mu,\nu)\in\mathbb{R}^m_+\times\mathbb{R}^n_+} G(\mu\,\nu)$ are generally different, but note that $\widehat{G}(\mu) = G(\mu,\nu)$ iff there is some $\nu \in \mathbb{R}^n_+$ such that $A^\top\mu - \nu + c = 0$ iff $A^\top\mu + c \geq 0$. Therefore, finding $\sup_{(\mu,\nu)\in\mathbb{R}^m_+\times\mathbb{R}^n_+} G(\mu,\nu)$ is equivalent to the constrained problem (D_1)

$$\begin{aligned} &\text{maximize} \quad -b^\top\mu \\ &\text{subject to} \quad A^\top\mu \geq -c, \;\; \mu \geq 0. \end{aligned}$$

The above problem is the dual of the linear program (P).

In summary, the dual function G of a primary problem (P) often contains hidden inequality constraints that define its domain, and sometimes it is possible to make these domain constraints $\psi_1(\mu) \leq 0, \ldots, \psi_p(\mu) \leq 0$ explicit, to define a new function $\widehat{G}$ that depends only on $q < m$ of the variables μ_i and is defined for all values $\mu_i \geq 0$ of these variables, and to replace the maximization problem (D), find $\sup_{\mu \in \mathbb{R}^m_+} G(\mu)$, by the constrained problem (D_1)

$$\begin{array}{ll} \text{maximize} & \widehat{G}(\mu) \\ \text{subject to} & \psi_i(\mu) \leq 0, \quad i = 1, \ldots, p. \end{array}$$

Problem (D_1) is different from the dual program (D), but it is equivalent to (D) as a maximization problem.

Another important property of the dual function G is that it provides a *lower bound* on the value of the objective function J. Indeed, we have

$$G(\mu) \leq L(u, \mu) \leq J(u) \quad \text{for all } u \in U \text{ and all } \mu \in \mathbb{R}^m_+, \tag{$\dagger$}$$

since $\mu \geq 0$ and $\varphi_i(u) \leq 0$ for $i = 1, \ldots, m$, so

$$G(\mu) = \inf_{v \in \Omega} L(v, \mu) \leq L(u, \mu) = J(u) + \sum_{i=1}^{m} \mu_i \varphi_i(u) \leq J(u).$$

If the primal problem (P) has a minimum denoted p^* and the dual problem (D) has a maximum denoted d^*, then the above inequality implies that

$$d^* \leq p^* \tag{$\dagger_w$}$$

known as *weak duality*. Equivalently, for every optimal solution λ^* of the dual problem and every optimal solution u^* of the primal problem, we have

$$G(\lambda^*) \leq J(u^*). \tag{$\dagger_{w'}$}$$

In particular, if $p^* = -\infty$, which means that the primal problem is unbounded below, then the dual problem is unfeasible. Conversely, if $d^* = +\infty$, which means that the dual problem is unbounded above, then the primal problem is unfeasible.

The difference $p^* - d^* \geq 0$ is called the *optimal duality gap*. If the duality gap is zero, that is, $p^* = d^*$, then we say that *strong duality* holds. Even when the duality gap is strictly positive, the inequality $(\dagger_w)$ can be helpful to find a lower bound on the optimal value of a primal problem that is difficult to solve, since the dual problem is always convex.

If the primal problem and the dual problem are feasible and if the optimal values p^* and d^* are finite and $p^* = d^*$ (no duality gap), then the complementary slackness conditions hold for the inequality constraints.

Proposition 30.13. *(Complementary Slackness) Given the minimization problem (P)*

$$\begin{aligned} &\text{minimize} && J(v) \\ &\text{subject to} && \varphi_i(v) \leq 0, \quad i = 1, \ldots, m, \end{aligned}$$

and its dual problem (D)

$$\begin{aligned} &\text{maximize} && G(\mu) \\ &\text{subject to} && \mu \in \mathbb{R}^m_+, \end{aligned}$$

if both (P) and (D) are feasible, $u \in U$ is an optimal solution of (P), $\lambda \in \mathbb{R}^m_+$ is an optimal solution of (D), and $J(u) = G(\lambda)$, then

$$\sum_{i=1}^m \lambda_i \varphi_i(u) = 0.$$

In other words, if the constraint φ_i is inactive at u, then $\lambda_i = 0$.

Proof. Since $J(u) = G(\lambda)$ we have

$$\begin{aligned} J(u) &= G(\lambda) \\ &= \inf_{v \in \Omega} \left(J(v) + \sum_{i=1}^m \lambda_i \varphi_i(v) \right) && \text{by definition of } G \\ &\leq J(u) + \sum_{i=1}^m \lambda_i \varphi_i(u) && \text{the greatest lower bound is a lower bound} \\ &\leq J(u) && \text{since } \lambda_i \geq 0, \varphi_i(u) \leq 0. \end{aligned}$$

which implies that $\sum_{i=1}^m \lambda_i \varphi_i(u) = 0$. □

Going back to Example 30.5, we see that weak duality says that for any feasible solution u of the primal problem (P), that is, some $u \in \mathbb{R}^n$ such that

$$Au \leq b, \;\; u \geq 0,$$

and for any feasible solution $\mu \in \mathbb{R}^m$ of the dual problem (D_1), that is,

$$A^\top \mu \geq -c, \;\; \mu \geq 0,$$

we have

$$-b^\top \mu \leq c^\top u.$$

Actually, if u and λ are optimal, then we know that strong duality holds, namely $-b^\top \mu = c^\top u$, but the proof of this fact is nontrivial.

The following theorem establishes a link between the solutions of the primal problem (P) and those of the dual problem (D). It also gives sufficient conditions for the duality gap to be zero.

Theorem 30.14. *Consider the minimization problem (P):*

$$\begin{aligned} &\text{minimize} && J(v) \\ &\text{subject to} && \varphi_i(v) \leq 0, \quad i = 1, \ldots, m, \end{aligned}$$

where the functions J and φ_i are defined on some open subset Ω of a finite-dimensional Euclidean vector space V (more generally, a real Hilbert space V).

(1) Suppose the functions $\varphi_i \colon \Omega \to \mathbb{R}$ are continuous, and that for every $\mu \in \mathbb{R}^m_+$, the problem (P_μ):

$$\begin{aligned} &\text{minimize} && L(v, \mu) \\ &\text{subject to} && v \in \Omega, \end{aligned}$$

has a unique solution u_μ, so that

$$L(u_\mu, \mu) = \inf_{v \in \Omega} L(v, \mu) = G(\mu),$$

and the function $\mu \mapsto u_\mu$ is continuous (on $\mathbb{R}^m_+$). If λ is any solution of problem (D):

$$\begin{aligned} &\text{maximize} && G(\mu) \\ &\text{subject to} && \mu \in \mathbb{R}^m_+, \end{aligned}$$

then the solution u_λ of the corresponding problem (P_λ) is a solution of Problem (P).

(2) Assume Problem (P) has some solution $u \in U$, and that Ω is convex (open), the functions φ_i ($1 \leq i \leq m$) and J are convex and differentiable at u, and that the constraints are qualified. Then Problem (D) has a solution $\lambda \in \mathbb{R}^m_+$, and $J(u) = G(\lambda)$; that is, the duality gap is zero.

Proof. (1) Our goal is to prove that for any solution λ of Problem (D), the pair (u_λ, λ) is a saddle point of L. By Theorem 30.12(1), the point $u_\lambda \in U$ is a solution of Problem (P).

Since $\lambda \in \mathbb{R}^m_+$ is a solution of Problem (D), by definition of $G(\lambda)$ and since u_λ satisfies Problem (P_λ), we have

$$G(\lambda) = \inf_{v \in \Omega} L(v, \lambda) = L(u_\lambda, \lambda),$$

which is one of the two equations characterizing a saddle point. In order to prove the second equation characterizing a saddle point,

$$\sup_{\mu \in \mathbb{R}^m_+} L(u_\mu, \mu) = L(u_\lambda, \lambda),$$

we will begin by proving that the function G is differentiable for any $\mu \in \mathbb{R}^m_+$, in order to be able to apply Theorem 20.8 to conclude that since G has a maximum at λ, that is, $-G$ has minimum at λ, then $-G'_\lambda(\mu - \lambda) \geq 0$ for all $\mu \in \mathbb{R}^m_+$. In fact, we prove that

$$G'_\mu(\xi) = \sum_{i=1}^m \xi_i \varphi_i(u_\mu) \quad \text{for all } \xi \in \mathbb{R}^m.$$

Consider any two points μ and $\mu+\xi$ in $\mathbb{R}^m_+$. By definition of u_μ we have

$$L(u_\mu, \mu) \leq L(u_{\mu+\xi}, \mu),$$

which means that

$$J(u_\mu) + \sum_{i=1}^m \mu_i \varphi_i(u_\mu) \leq J(u_{\mu+\xi}) + \sum_{i=1}^m \mu_i \varphi_i(u_{\mu+\xi}), \tag{$*_1$}$$

and since $G(\mu) = L(u_\mu, \mu) = J(u_\mu) + \sum_{i=1}^m \mu_i \varphi_i(u_\mu)$ and $G(\mu+\xi) = L(u_{\mu+\xi}, \mu+\xi) = J(u_{\mu+\xi}) + \sum_{i=1}^m (\mu_i + \xi_i)\varphi_i(u_{\mu+\xi})$, we have

$$G(\mu+\xi) - G(\mu) = J(u_{\mu+\xi}) - J(u_\mu) + \sum_{i=1}^m (\mu_i + \xi_i)\varphi_i(u_{\mu+\xi}) - \sum_{i=1}^m \mu_i \varphi_i(u_\mu), \tag{$*_2$}$$

and since $(*_1)$ can be written as

$$0 \leq J(u_{\mu+\xi}) - J(u_\mu) + \sum_{i=1}^m \mu_i \varphi_i(u_{\mu+\xi}) - \sum_{i=1}^m \mu_i \varphi_i(u_\mu),$$

by adding $\sum_{i=1}^m \xi_i \varphi_i(u_{\mu+\xi})$ to both sides of the above inequality and using $(*_2)$ we get

$$\sum_{i=1}^m \xi_i \varphi_i(u_{\mu+\xi}) \leq G(\mu+\xi) - G(\mu). \tag{$*_3$}$$

By definition of $u_{\mu+\xi}$ we have

$$L(u_{\mu+\xi}, \mu+\xi) \leq L(u_\mu, \mu+\xi),$$

which means that

$$J(u_{\mu+\xi}) + \sum_{i=1}^m (\mu_i + \xi_i)\varphi_i(u_{\mu+\xi}) \leq J(u_\mu) + \sum_{i=1}^m (\mu_i + \xi_i)\varphi_i(u_\mu), \tag{$*_4$}$$

which can be written as

$$J(u_{\mu+\xi}) - J(u_\mu) + \sum_{i=1}^m (\mu_i + \xi_i)\varphi_i(u_{\mu+\xi}) - \sum_{i=1}^m (\mu_i + \xi_i)\varphi_i(u_\mu) \leq 0,$$

and by adding $\sum_{i=1}^m \xi_i \varphi_i(u_\mu)$ to both sides of the above inequality and using $(*_2)$ we get

$$G(\mu+\xi) - G(\mu) \leq \sum_{i=1}^m \xi_i \varphi_i(u_\mu). \tag{$*_5$}$$

Putting $(*_3)$ and $(*_5)$ together we obtain

$$\sum_{i=1}^{m} \xi_i \varphi_i(u_{\mu+\xi}) \leq G(\mu+\xi) - G(\mu) \leq \sum_{i=1}^{m} \xi_i \varphi_i(u_\mu). \tag{$*_6$}$$

Consequently there is some $\theta \in [0,1]$ such that

$$\begin{aligned} G(\mu+\xi) - G(\mu) &= (1-\theta)\Big(\sum_{i=1}^{m} \xi_i \varphi_i(u_\mu)\Big) + \theta\Big(\sum_{i=1}^{m} \xi_i \varphi_i(u_{\mu+\xi})\Big) \\ &= \sum_{i=1}^{m} \xi_i \varphi_i(u_\mu) + \theta\Big(\sum_{i=1}^{m} \xi_i (\varphi_i(u_{\mu+\xi}) - \varphi_i(u_\mu))\Big). \end{aligned}$$

Since by hypothesis the functions $\mu \mapsto u_\mu$ (from $\mathbb{R}^m_+$ to Ω) and $\varphi_i \colon \Omega \to \mathbb{R}$ are continuous, for any $\mu \in \mathbb{R}^m_+$ we can write

$$G(\mu+\xi) - G(\mu) = \sum_{i=1}^{m} \xi_i \varphi_i(u_\mu) + \|\xi\|\, \epsilon(\xi), \quad \text{with } \lim_{\xi \mapsto 0} \epsilon(\xi) = 0, \tag{$*_7$}$$

for any $\|\ \|$ norm on $\mathbb{R}^m$. Equation $(*)_7$ show that G is differentiable for any $\mu \in \mathbb{R}^m_+$, and that

$$G'_\mu(\xi) = \sum_{i=1}^{m} \xi_i \varphi_i(u_\mu) \quad \text{for all } \xi \in \mathbb{R}^m. \tag{$*_8$}$$

Actually there is a small problem, namely that the notion of derivative was defined for a function defined on an *open* set, but $\mathbb{R}^m_+$ is not open. The difficulty only arises to ensure that the derivative is unique, but in our case we have a unique expression for the derivative so there is no problem as far as defining the derivative. There is still a potential problem, which is that we would like to apply Theorem 20.8 to conclude that since G has a maximum at λ, that is, $-G$ has minimum at λ, then

$$-G'_\lambda(\mu - \lambda) \geq 0 \quad \text{for all } \mu \in \mathbb{R}^m_+, \tag{$*_9$}$$

but the hypotheses of Theorem 20.8 require the domain of the function to be open. Fortunately, close examination of the proof of Theorem 20.8 shows that the proof still holds with $U = \mathbb{R}^m_+$. Therefore, $(*_8)$ holds, equivalently

$$G'_\lambda(\mu - \lambda) \leq 0 \quad \text{for all } \mu \in \mathbb{R}^m_+, \tag{$*_{10}$}$$

which, using the expression for G'_λ given in $(*_8)$ gives

$$\sum_{i=1}^{m} \mu_i \varphi_i(u_\lambda) \leq \sum_{i=1}^{m} \lambda_i \varphi_i(u_\lambda), \quad \text{for all } \mu \in \mathbb{R}^m_+. \tag{$*_{11}$}$$

As a consequence of $(*_{11})$, we obtain

$$\begin{aligned} L(u_\lambda, \mu) &= J(u_\lambda) + \sum_{i=1}^{m} \mu_i \varphi_i(u_\lambda) \\ &\leq J(u_\lambda) + \sum_{i=1}^{m} \lambda_i \varphi_i(u_\lambda) = L(u_\lambda, \lambda), \end{aligned}$$

for all $\mu \in \mathbb{R}^m_+$, that is,

$$L(u_\lambda, \mu) \leq L(u_\lambda, \lambda), \quad \text{for all } \mu \in \mathbb{R}^m_+, \tag{$*_{12}$}$$

which implies the second inequality

$$\sup_{\mu \in \mathbb{R}^m_+} L(u_\mu, \mu) = L(u_\lambda, \lambda)$$

stating that (u_λ, λ) is a saddle point. Therefore, (u_λ, λ) is a saddle point of L, as claimed.

(2) The hypotheses are exactly those required by Theorem 30.12(2), thus there is some $\lambda \in \mathbb{R}^m_+$ such that (u, λ) is a saddle point of the Lagrangian L, and by Theorem 30.12(1) we have $J(u) = L(u, \lambda)$. By Proposition 30.11, we have

$$J(u) = L(u, \lambda) = \inf_{v \in \Omega} L(v, \lambda) = \sup_{\mu \in \mathbb{R}^m_+} \inf_{v \in \Omega} L(v, \mu),$$

which can be rewritten as

$$J(u) = G(\lambda) = \sup_{\mu \in \mathbb{R}^m_+} G(\mu),$$

in other words, Problem (D) has a solution, and $J(u) = G(\lambda)$. □

Remark: If (u, λ) is a saddle point of the Lagrangian L (defined on $\Omega \times \mathbb{R}^m_+$), then by Proposition 30.11 the vector λ is a solution of Problem (D). Conversely, under the hypotheses of Part (1) of Theorem 30.14, if λ is a solution of Problem (D), then (u_λ, λ) is a saddle point of L. Consequently, under the above hypotheses, the set of solutions of the dual problem (D) coincide with the set of second arguments λ of the saddle points (u, λ) of L. In some sense, this result is the "dual" of the result stated in Theorem 30.12, namely that the set of solutions of Problem (P) coincides with set of first arguments u of the saddle points (u, λ) of L.

Informally, in Theorem 30.14(1), the hypotheses say that if $G(\mu)$ can be "computed nicely," in the sense that there is a unique minimizer u_μ of $L(v, \mu)$ (with $v \in \Omega$) such that $G(\mu) = L(u_\mu, \mu)$, and if a maximizer λ of $G(\mu)$ (with $\mu \in \mathbb{R}^m_+$) can be determined, then u_λ yields the minimum value of J, that is, $p^* = J(u_\lambda)$. If the constraints are qualified and if the functions J and φ_i are convex and differentiable, then since the KKT conditions hold, the duality gap is zero; that is,

$$G(\lambda) = L(u_\lambda, \lambda) = J(u_\lambda).$$

Example 30.6. Going back to Example 30.5 where we considered the linear program (P)

$$\begin{aligned} &\text{minimize} \quad c^\top x \\ &\text{subject to} \quad Ax \leq b, \;\; x \geq 0, \end{aligned}$$

with A an $m \times n$ matrix, the Lagrangian $L(\mu, \nu)$ is given by

$$L(v, \mu, \nu) = -b^\top \mu + (c + A^\top \mu - \nu)^\top v,$$

and we found that the dual function $G(\mu, \nu) = \inf_{v \in \mathbb{R}^n} L(v, \mu, \nu)$ is given for all $\mu \geq 0$ and $\nu \geq 0$ by

$$G(\mu, \nu) = \begin{cases} -b^\top \mu & \text{if } A^\top \mu - \nu + c = 0, \\ -\infty & \text{otherwise.} \end{cases}$$

The hypotheses of Theorem 30.14(1) certainly fail since there are infinitely $u_{\mu,\nu} \in \mathbb{R}^n$ such that $G(\mu, \nu) = \inf_{v \in \mathbb{R}^n} L(v, \mu, \nu) = L(u_{\mu,\nu}, \mu, \nu)$. Therefore, the dual function G is no help in finding a solution of the primal (P). As we saw earlier, if we consider the modified dual Problem (D_1) then strong duality holds, but this does not follow from Theorem 30.14, and a different proof is required.

Thus we have the somewhat counter-intuitive situation that the *general* theory of Lagrange duality does not apply, at least directly, to linear programming, a fact that is not sufficiently emphasized in many expositions. A separate treatment of duality if required.

Unlike the case of linear programming, which needs a separate treatment, Theorem 30.14 applies to the optimization problem involving a convex quadratic objective function and a set of affine inequality constraints. So in some sense, convex quadratic programming is simpler than linear programming!

Example 30.7. Consider the quadratic objective function

$$J(v) = \frac{1}{2} v^\top A v - v^\top b,$$

where A is an $n \times n$ matrix which is symmetric positive definite, $b \in \mathbb{R}^n$, and the constraints are affine inequality constraints of the form

$$Cx \leq d,$$

where C is an $m \times n$ matrix and $d \in \mathbb{R}^m$. For the time being, we do not assume that C has rank m. Since A is symmetric positive definite, J is strictly convex, as implied by Proposition 20.9 (see Example 20.1). The Lagrangian of this quadratic optimization problem is given by

$$\begin{aligned} L(v, \mu) &= \frac{1}{2} v^\top A v - v^\top b + (Cv - d)^\top \mu \\ &= \frac{1}{2} v^\top A v - v^\top (b - C^\top \mu) - \mu^\top d. \end{aligned}$$

Since A is symmetric positive definite, by Proposition 22.2, the function $v \mapsto L(v, \mu)$ has a unique minimum obtained for the solution u_μ of the linear system

$$Av = b - C^\top \mu;$$

that is,

$$u_\mu = A^{-1}(b - C^\top \mu).$$

This shows that the Problem (P_μ) has a unique solution which depends continuously on μ. Then for any solution λ of the dual problem, $u_\lambda = A^{-1}(b - C^\top \lambda)$ is an optimal solution of the primal problem.

We compute $G(\mu)$ as follows:

$$\begin{aligned} G(\mu) = L(u_\mu, \mu) &= \frac{1}{2} u_\mu^\top A u_\mu - u_\mu^\top (b - C^\top \mu) - \mu^\top d \\ &= \frac{1}{2} u_\mu^\top (b - C^\top \mu) - u_\mu^\top (b - C^\top \mu) - \mu^\top d \\ &= -\frac{1}{2} u_\mu^\top (b - C^\top \mu) - \mu^\top d \\ &= -\frac{1}{2} (b - C^\top \mu)^\top A^{-1} (b - C^\top \mu) - \mu^\top d \\ &= -\frac{1}{2} \mu^\top C A^{-1} C^\top \mu + \mu^\top (C A^{-1} b - d) - \frac{1}{2} b^\top A^{-1} b. \end{aligned}$$

Since A is symmetric positive definite, the matrix $CA^{-1}C^\top$ is symmetric positive semidefinite. It is invertible iff $C^\top \mu = 0$ implies $\mu = 0$, that is, $\operatorname{Ker} C^\top = (0)$, which is equivalent to $\operatorname{Im}(C) = \mathbb{R}^m$, namely if C has rank m (in which case, $m \leq n$).

It can be shown that the primal problem always has a solution, in fact unique. As a consequence by Theorem 30.14(2), the function $-G(\mu)$ always has a minimum, which is unique if C has rank m. We also verify easily that the gradient of G is given by

$$\nabla G_\mu = C u_\mu - d = -C A^{-1} C^\top \mu + C A^{-1} b - d.$$

Observe that since $CA^{-1}C^\top$ is symmetric positive semidefinite, $-G(\mu)$ is convex.

Therefore, if C has rank m, a solution of Problem (P) is obtained by finding the unique solution λ of the equation

$$-C A^{-1} C^\top \mu + C A^{-1} b - d = 0,$$

and then the minimum u_λ of Problem (P) is given by

$$u_\lambda = A^{-1}(b - C^\top \lambda).$$

If C has rank $< m$, then we can find $\lambda \geq 0$ by finding a feasible solution of the linear program whose set of constraints is given by

$$-C A^{-1} C^\top \mu + C A^{-1} b - d = 0,$$

using the standard method of adding nonnegative slack variables $\xi_1, \ldots, \xi_m$ and maximizing $-(\xi_1 + \cdots + \xi_m)$.

30.6 Handling Equality Constraints Explicitly

Sometimes it is desirable to handle equality constraints explicitly (for instance, this is what Boyd and Vandenberghe do, see [22]). The only difference is that the Lagrange multipliers associated with equality constraints are not required to be nonnegative, as we now show.

Consider the optimization problem (P')

$$\begin{array}{ll} \text{minimize} & J(v) \\ \text{subject to} & \varphi_i(v) \leq 0, \quad i = 1, \ldots, m \\ & \psi_j(v) = 0, \quad j = 1, \ldots, p. \end{array}$$

We treat each equality constraint $\psi_j(u) = 0$ as the conjunction of the inequalities $\psi_j(u) \leq 0$ and $-\psi_j(u) \leq 0$, and we associate Lagrange multipliers $\lambda \in \mathbb{R}^m_+$, and $\nu^+, \nu^- \in \mathbb{R}^p_+$. The KKT conditions are

$$J'_u + \sum_{i=1}^m \lambda_i (\varphi'_i)_u + \sum_{j=1}^p \nu_j^+ (\psi'_j)_u - \sum_{j=1}^p \nu_j^- (\psi'_j)_u = 0,$$

and

$$\sum_{i=1}^m \lambda_i \varphi_i(u) + \sum_{j=1}^p \nu_j^+ \psi_j(u) - \sum_{j=1}^p \nu_j^- \psi_j(u) = 0,$$

with $\lambda \geq 0, \nu^+ \geq 0, \nu^- \geq 0$. Since $\psi_j(u) = 0$ for $j = 1, \ldots, p$, these equations can be rewritten as

$$J'_u + \sum_{i=1}^m \lambda_i (\varphi'_i)_u + \sum_{j=1}^p (\nu_j^+ - \nu_j^-)(\psi'_j)_u = 0,$$

and

$$\sum_{i=1}^m \lambda_i \varphi_i(u) = 0$$

with $\lambda \geq 0, \nu^+ \geq 0, \nu^- \geq 0$, and if we introduce $\nu_j = \nu_j^+ - \nu_j^-$ we obtain the following KKT conditions for programs with explicit equality constraints:

$$J'_u + \sum_{i=1}^m \lambda_i (\varphi'_i)_u + \sum_{j=1}^p \nu_j (\psi'_j)_u = 0,$$

and

$$\sum_{i=1}^m \lambda_i \varphi_i(u) = 0$$

with $\lambda \geq 0$ and $\nu \in \mathbb{R}^p$ arbitrary.

Let us now assume that the functions φ_i and ψ_j are convex. As we explained just after Definition 30.6, nonaffine equality constraints are never qualified. Thus, in order to generalize Theorem 30.6 to explicit equality constraints, we assume that the equality constraints ψ_j are affine.

Theorem 30.15. *Let $\varphi_i \colon \Omega \to \mathbb{R}$ be m convex inequality constraints and $\psi_j \colon \Omega \to \mathbb{R}$ be p affine equality constraints defined on some open convex subset Ω of a finite-dimensional Euclidean vector space V (more generally, a real Hilbert space V), let $J \colon \Omega \to \mathbb{R}$ be some function, let U be given by*

$$U = \{x \in \Omega \mid \varphi_i(x) \leq 0,\ \psi_j(x) = 0,\ 1 \leq i \leq m,\ 1 \leq j \leq p\},$$

and let $u \in U$ be any point such that the functions φ_i and J are differentiable at u, and the functions ψ_j are affine.

(1) If J has a local minimum at u with respect to U, and if the constraints are qualified, then there exist some vectors $\lambda \in \mathbb{R}^m_+$ and $\nu \in \mathbb{R}^p$, such that the KKT condition hold:

$$J'_u + \sum_{i=1}^{m} \lambda_i(u)(\varphi'_i)_u + \sum_{j=1}^{p} \nu_j(\psi'_j)_u = 0,$$

and

$$\sum_{i=1}^{m} \lambda_i(u)\varphi_i(u) = 0, \quad \lambda_i \geq 0, \quad i = 1, \ldots, m.$$

Equivalently, in terms of gradients, the above conditions are expressed as

$$\nabla J_u + \sum_{i=1}^{m} \lambda_i \nabla(\varphi_i)_u + \sum_{j=1}^{p} \nu_j \nabla(\psi_j)_u = 0$$

and

$$\sum_{i=1}^{m} \lambda_i(u)\varphi_i(u) = 0, \quad \lambda_i \geq 0, \quad i = 1, \ldots, m.$$

(2) Conversely, if the restriction of J to U is convex and if there exist vectors $\lambda \in \mathbb{R}^m_+$ and $\nu \in \mathbb{R}^p$ such that the KKT conditions hold, then the function J has a (global) minimum at u with respect to U.

The Lagrangian $L(v, \lambda, \nu)$ of Problem (P') is defined as

$$L(v, \mu, \nu) = J(v) + \sum_{i=1}^{m} \mu_i \varphi_i(v) + \sum_{j=1}^{p} \nu_i \psi_j(v),$$

where $v \in \Omega$, $\mu \in \mathbb{R}^m_+$, and $\nu \in \mathbb{R}^p$.

The function $G\colon \mathbb{R}^m_+ \times \mathbb{R}^p \to \mathbb{R}$ given by

$$G(\mu, \nu) = \inf_{v \in \Omega} L(v, \mu, \nu) \quad \mu \in \mathbb{R}^m_+,\ \nu \in \mathbb{R}^p$$

is called the *Lagrange dual function* (or *dual function*), and the *dual problem* (D') is

$$\begin{aligned} &\text{maximize} && G(\mu, \nu) \\ &\text{subject to} && \mu \in \mathbb{R}^m_+,\ \nu \in \mathbb{R}^p. \end{aligned}$$

Observe that the Lagrange multipliers ν are not restricted to be nonnnegative.

Theorem 30.12 and Theorem 30.14 are immediately generalized to Problem (P'). We only state the new version of 30.14, leaving the new version of Theorem 30.12 as an exercise.

Theorem 30.16. *Consider the minimization problem (P'):*

$$\begin{aligned} &\text{minimize} && J(v) \\ &\text{subject to} && \varphi_i(v) \leq 0, \quad i = 1, \ldots, m \\ &&& \psi_j(v) = 0, \quad j = 1, \ldots, p. \end{aligned}$$

where the functions J, φ_i are defined on some open subset Ω of a finite-dimensional Euclidean vector space V (more generally, a real Hilbert space V), and the functions ψ_j are affine.

(1) Suppose the functions $\varphi_i\colon \Omega \to \mathbb{R}$ are continuous, and that for every $\mu \in \mathbb{R}^m_+$ and every $\nu \in \mathbb{R}^p$, the problem $(P_{\mu,\nu})$:

$$\begin{aligned} &\text{minimize} && L(v, \mu, \nu) \\ &\text{subject to} && v \in \Omega, \end{aligned}$$

has a unique solution $u_{\mu,\nu}$, so that

$$L(u_{\mu,\nu}, \mu, \nu) = \inf_{v \in \Omega} L(v, \mu, \nu) = G(\mu, \nu),$$

and the function $(\mu, \nu) \mapsto u_{\mu,\nu}$ is continuous (on $\mathbb{R}^m_+ \times \mathbb{R}^p$). If (λ, η) is any solution of problem (D):

$$\begin{aligned} &\text{maximize} && G(\mu, \nu) \\ &\text{subject to} && \mu \in \mathbb{R}^m_+,\ \nu \in \mathbb{R}^p, \end{aligned}$$

then the solution $u_{\lambda,\eta}$ of the corresponding problem $(P_{\lambda,\eta})$ is a solution of Problem (P').

(2) Assume Problem (P') has some solution $u \in U$, and that Ω is convex (open), the functions φ_i $(1 \leq i \leq m)$ and J are convex, differentiable at u, and that the constraints are qualified. Then Problem (D') has a solution $(\lambda, \eta) \in \mathbb{R}^m_+ \times \mathbb{R}^p$, and $J(u) = G(\lambda, \eta)$; that is, the duality gap is zero.

In the next example we derive the dual function and the dual program of the optimization problem of Section 30.4 (Hard margin SVM), which involves both inequality and equality constraints. We also derive the KKT conditions associated with the dual program.

Example 30.8. Recall the **Hard margin SVM** problem (SVM_{h2}):

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\|w\|^2 \\ &\text{subject to} \\ &\qquad w^\top u_i - b \geq 1 \qquad i = 1, \ldots, p \\ &\qquad -w^\top v_j + b \geq 1 \qquad j = 1, \ldots, q. \end{aligned}$$

We proceed in six steps.

Step 1: Write the constraints in matrix form.

The inequality constraints are written as

$$C \begin{pmatrix} w \\ b \end{pmatrix} \leq d,$$

where C is a $(p+q) \times (n+1)$ matrix C and $d \in \mathbb{R}^{p+q}$ is the vector given by

$$C = \begin{pmatrix} -u_1^\top & 1 \\ \vdots & \vdots \\ -u_p^\top & 1 \\ v_1^\top & -1 \\ \vdots & \vdots \\ v_q^\top & -1 \end{pmatrix}, \quad d = \begin{pmatrix} -1 \\ \vdots \\ -1 \end{pmatrix} = -\mathbf{1}_{p+q}.$$

If let X be the $n \times (p+q)$ matrix given by

$$X = \begin{pmatrix} -u_1 & \cdots & -u_p & v_1 & \cdots & v_q \end{pmatrix},$$

then

$$C = \begin{pmatrix} X^\top & \mathbf{1}_p \\ & -\mathbf{1}_q \end{pmatrix}$$

and so

$$C^\top = \begin{pmatrix} X & \\ \mathbf{1}_p^\top & -\mathbf{1}_q^\top \end{pmatrix}.$$

Step 2: Write the objective function in matrix form.

The objective function is given by

$$J(w, b) = \frac{1}{2} \begin{pmatrix} w^\top & b \end{pmatrix} \begin{pmatrix} I_n & 0_n \\ 0_n^\top & 0 \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix}.$$

Note that the corresponding matrix is symmetric positive semidefinite, but it is *not* invertible. Thus the function J is convex but not strictly convex.

Step 3: Write the Lagrangian in matrix form.

As in Example 30.7, we obtain the Lagrangian

$$L(w,b,\lambda,\mu) = \frac{1}{2}\begin{pmatrix} w^\top & b\end{pmatrix}\begin{pmatrix} I_n & 0_n \\ 0_n^\top & 0\end{pmatrix}\begin{pmatrix} w \\ b\end{pmatrix} - \begin{pmatrix} w^\top & b\end{pmatrix}\left(0_{n+1} - C^\top\begin{pmatrix}\lambda \\ \mu\end{pmatrix}\right) + \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\mathbf{1}_{p+q},$$

that is,

$$L(w,b,\lambda,\mu) = \frac{1}{2}\begin{pmatrix} w^\top & b\end{pmatrix}\begin{pmatrix} I_n & 0_n \\ 0_n^\top & 0\end{pmatrix}\begin{pmatrix} w \\ b\end{pmatrix} + \begin{pmatrix} w^\top & b\end{pmatrix}\begin{pmatrix} X\begin{pmatrix}\lambda \\ \mu\end{pmatrix} \\ \mathbf{1}_p^\top\lambda & -\mathbf{1}_q^\top\mu\end{pmatrix} + \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\mathbf{1}_{p+q}.$$

Step 4: Find the dual function $G(\lambda,\mu)$.

In order to find the dual function $G(\lambda,\mu)$ we need to minimize $L(w,b,\lambda,\mu)$ with respect to w and b and for this, since the objective function J is convex and since $\mathbb{R}^{n+1}$ is convex and open, we can apply Theorem 20.11, which gives a necessary and sufficient condition for a minimum. The gradient of $L(w,b,\lambda,\mu)$ with respect to w and b is

$$\begin{aligned}\nabla L_{w,b} &= \begin{pmatrix} I_n & 0_n \\ 0_n^\top & 0\end{pmatrix}\begin{pmatrix} w \\ b\end{pmatrix} + \begin{pmatrix} X\begin{pmatrix}\lambda \\ \mu\end{pmatrix} \\ \mathbf{1}_p^\top\lambda & -\mathbf{1}_q^\top\mu\end{pmatrix} \\ &= \begin{pmatrix} w \\ 0\end{pmatrix} + \begin{pmatrix} X\begin{pmatrix}\lambda \\ \mu\end{pmatrix} \\ \mathbf{1}_p^\top\lambda & -\mathbf{1}_q^\top\mu\end{pmatrix}.\end{aligned}$$

The necessary and sufficient condition for a minimum is

$$\nabla L_{w,b} = 0,$$

which yields

$$w = -X\begin{pmatrix}\lambda \\ \mu\end{pmatrix} \tag{$*_1$}$$

and

$$\mathbf{1}_p^\top\lambda - \mathbf{1}_q^\top\mu = 0. \tag{$*_2$}$$

The second equation can be written as

$$\sum_{i=1}^{p}\lambda_i = \sum_{j=1}^{q}\mu_j. \tag{$*_3$}$$

Plugging back w from $(*_1)$ into the Lagrangian and using $(*_2)$ we get

$$G(\lambda,\mu) = -\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda\\ \mu\end{pmatrix} + \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\mathbf{1}_{p+q}; \tag{$*_4$}$$

of course, $\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\mathbf{1}_{p+q} = \sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j$. Actually, to be perfectly rigorous $G(\lambda,\mu)$ is only defined on the intersection of the hyperplane of equation $\sum_{i=1}^p \lambda_i = \sum_{j=1}^q \mu_j$ with the convex octant in $\mathbb{R}^{p+1}$ given by $\lambda \geq 0, \mu \geq 0$, so for all $\lambda \in \mathbb{R}^p_+$ and all $\mu \in \mathbb{R}^q_+$, we have

$$G(\lambda,\mu) = \begin{cases} -\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda\\ \mu\end{pmatrix} + \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\mathbf{1}_{p+q} & \text{if } \sum_{i=1}^p \lambda_i = \sum_{j=1}^q \mu_j \\ -\infty & \text{otherwise.}\end{cases}$$

Note that the condition

$$\sum_{i=1}^p \lambda_i = \sum_{j=1}^q \mu_j$$

is Condition $(*_2)$ of Example 30.4, which is not surprising.

Step 5: Write the dual program in matrix form.

Maximizing the dual function $G(\lambda,\mu)$ over its domain of definition is equivalent to maximizing

$$\widehat{G}(\lambda,\mu) = -\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda\\ \mu\end{pmatrix} + \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\mathbf{1}_{p+q}$$

subject to the constraint

$$\sum_{i=1}^p \lambda_i = \sum_{j=1}^q \mu_j,$$

so we formulate the dual program as,

$$\begin{aligned}&\text{maximize} \quad -\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda\\ \mu\end{pmatrix} + \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\mathbf{1}_{p+q}\\ &\text{subject to}\\ &\qquad \sum_{i=1}^p \lambda_i = \sum_{j=1}^q \mu_j\\ &\qquad \lambda \geq 0,\ \mu \geq 0,\end{aligned}$$

or equivalently,

$$\begin{aligned}\text{minimize} \quad & \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda\\ \mu\end{pmatrix} - \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} \mathbf{1}_{p+q}\\ \text{subject to} \quad & \\ & \sum_{i=1}^{p} \lambda_i = \sum_{j=1}^{q} \mu_j\\ & \lambda \geq 0,\ \mu \geq 0.\end{aligned}$$

The constraints of the dual program are a lot simpler than the constraints

$$\begin{pmatrix}X^\top & \mathbf{1}_p\\ & -\mathbf{1}_q\end{pmatrix}\begin{pmatrix}w\\ b\end{pmatrix} \leq -\mathbf{1}_{p+q}$$

of the primal program because these constraints have been "absorbed" by the objective function $\widehat{G}(\lambda, \nu)$ of the dual program which involves the matrix $X^\top X$. The matrix $X^\top X$ is symmetric positive semidefinite, but not invertible in general.

Step 6: Solve the dual program.

This step involves using numerical procedures typically based on gradient descent to find λ and μ. Once λ and μ are determined, w is determined by $(*_1)$ and b is determined as in Section 30.4 using the fact that there is at least some i_0 such that $\lambda_{i_0} > 0$ and some j_0 such that $\mu_{j_0} > 0$.

Remarks:

(1) Since the constraints are affine and the objective function is convex, by Theorem 30.16(2) the duality gap is zero, so for any minimum w of $J(w, b) = (1/2)w^\top w$ and any maximum (λ, μ) of G, we have

$$J(w, b) = \frac{1}{2} w^\top w = G(\lambda, \mu).$$

But by $(*_1)$

$$w = -X\begin{pmatrix}\lambda\\ \mu\end{pmatrix} = \sum_{i=1}^{p} \lambda_i u_i - \sum_{j=1}^{q} \mu_j v_j,$$

so

$$\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda\\ \mu\end{pmatrix} = w^\top w,$$

and we get

$$\frac{1}{2} w^\top w = -\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda\\ \mu\end{pmatrix} + \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} \mathbf{1}_{p+q} = -\frac{1}{2} w^\top w + \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} \mathbf{1}_{p+q},$$

so

$$w^\top w = \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} \mathbf{1}_{p+q} = \sum_{i=1}^{p} \lambda_i + \sum_{j=1}^{q} \mu_j,$$

which yields

$$G(\lambda, \mu) = \frac{1}{2}\left(\sum_{i=1}^{p} \lambda_i + \sum_{j=1}^{q} \mu_j\right).$$

The above formulae are stated in Vapnik [110] (Chapter 10, Section 1).

(2) It is instructive to compute the Lagrangian of the dual program and to derive the KKT conditions for this Lagrangian.

The conditions $\lambda \geq 0$ being equivalent to $-\lambda \leq 0$, and the conditions $\mu \geq 0$ being equivalent to $-\mu \leq 0$, we introduce Lagrange multipliers $\alpha \in \mathbb{R}^p_+$ and $\beta \in \mathbb{R}^q_+$ as well as a multiplier $\rho \in \mathbb{R}$ for the equational constraint, and we form the Lagrangian

$$\begin{aligned} L(\lambda, \mu, \alpha, \beta, \rho) = \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda \\ \mu\end{pmatrix} - \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} \mathbf{1}_{p+q} \\ - \sum_{i=1}^{p} \alpha_i \lambda_i - \sum_{j=1}^{q} \beta_j \mu_j + \rho\left(\sum_{j=1}^{q} \mu_j - \sum_{i=1}^{p} \lambda_i\right). \end{aligned}$$

It follows that the KKT conditions are

$$X^\top X \begin{pmatrix}\lambda \\ \mu\end{pmatrix} - \mathbf{1}_{p+q} - \begin{pmatrix}\alpha \\ \beta\end{pmatrix} + \rho \begin{pmatrix}-\mathbf{1}_p \\ \mathbf{1}_q\end{pmatrix} = 0_{p+q}, \tag{$*_4$}$$

and $\alpha_i \lambda_i = 0$ for $i = 1, \ldots, p$ and $\beta_j \mu_j = 0$ for $j = 1, \ldots, q$.

But $(*_4)$ is equivalent to

$$-X^\top X \begin{pmatrix}\lambda \\ \mu\end{pmatrix} + \rho \begin{pmatrix}\mathbf{1}_p \\ -\mathbf{1}_q\end{pmatrix} + \mathbf{1}_{p+q} + \begin{pmatrix}\alpha \\ \beta\end{pmatrix} = 0_{p+q},$$

which is precisely the result of adding $\alpha \geq 0$ and $\beta \geq 0$ as slack variables to the inequalities $(*_3)$ of Example 30.4, namely

$$-X^\top X \begin{pmatrix}\lambda \\ \mu\end{pmatrix} + b \begin{pmatrix}\mathbf{1}_p \\ -\mathbf{1}_q\end{pmatrix} + \mathbf{1}_{p+q} \leq 0_{p+q},$$

to make them equalities, where ρ plays the role of b.

When the constraints are affine, the dual function $G(\lambda, \nu)$ can be expressed in terms of the conjugate of the objective function J.

30.7 Conjugate Function and Legendre Dual Function

The notion of conjugate function goes back to Legendre and plays an important role in classical mechanics for converting a Lagrangian to a Hamiltonian; see Arnold [4] (Chapter 3, Sections 14 and 15).

Definition 30.10. Let $f\colon A \to \mathbb{R}$ be a function defined on some subset A of $\mathbb{R}^n$. The *conjugate* f^* of the function f is the partial function $f^*\colon \mathbb{R}^n \to \mathbb{R}$ defined by

$$f^*(y) = \sup_{x \in A}(y^\top x - f(x)), \quad y \in \mathbb{R}^n.$$

The conjugate of a function is also called the *Fenchel conjugate*, or *Legendre transform* when f is differentiable.

As the pointwise supremum of a family of affine functions in y, the conjugate function f^* is convex, even if the original function f is not convex.

The domain of f^* can be very small, even if the domain of f is big. For example, if $f\colon \mathbb{R} \to \mathbb{R}$ is the affine function given by $f(x) = ax + b$ (with $a, b \in \mathbb{R}$), then the function $x \mapsto yx - ax - b$ is unbounded above unless $y = a$, so

$$f^*(y) = \begin{cases} -b & \text{if } y = a \\ +\infty & \text{otherwise.} \end{cases}$$

The domain of f^* can also be bigger than the domain of f; see Example 30.9(3).

The conjugate of many functions that come up in optimization are derived in Boyd and Vandenberghe; see [22], Section 3.3. We mention a few that will be used in this chapter.

Example 30.9.

(1) *Negative logarithm*: $f(x) = -\log x$, with $\mathrm{dom}(f) = \{x \in \mathbb{R} \mid x > 0\}$. The function $x \mapsto yx + \log x$ is unbounded above if $y \geq 0$, and when $y < 0$, its maximum is obtained iff its derivative is zero, namely

$$y + \frac{1}{x} = 0.$$

Substituting for $x = -1/y$ in $yx + \log x$, we obtain $-1 + \log(-1/y) = -1 - \log(-y)$, so we have

$$f^*(y) = -\log(-y) - 1,$$

with $\mathrm{dom}(f^*) = \{y \in \mathbb{R} \mid y < 0\}$.

(2) *Exponential*: $f(x) = e^x$, with $\mathrm{dom}(f) = \mathbb{R}$. The function $x \mapsto yx - e^x$ is unbounded if $y < 0$. When $y > 0$, it reaches a maximum iff its derivative is zero, namely

$$y - e^x = 0.$$

Substituting for $x = \log y$ in $yx - e^x$, we obtain $y \log y - y$, so we have

$$f^*(y) = y \log y - y,$$

with $\mathrm{dom}(f^*) = \{y \in \mathbb{R} \mid y \geq 0\}$, with the convention that $0 \log 0 = 0$.

(3) *Negative Entropy*: $f(x) = x \log x$, with $\mathrm{dom}(f) = \{x \in \mathbb{R} \mid x \geq 0\}$, with the convention that $0 \log 0 = 0$. The function $x \mapsto yx - x \log x$ is bounded above for all $y > 0$, and it attains its maximum when its derivative is zero, namely

$$y - \log x - 1 = 0.$$

Substituting for $x = e^{y-1}$ in $yx - x \log x$, we obtain $ye^{y-1} - e^{y-1}(y-1) = e^{y-1}$, which yields

$$f^*(y) = e^{y-1},$$

with $\mathrm{dom}(f) = \mathbb{R}$.

(4) *Strictly convex quadratic function*: $f(x) = \frac{1}{2} x^\top A x$, where A is an $n \times n$ symmetric positive definite matrix, with $\mathrm{dom}(f) = \mathbb{R}^n$. The function $x \mapsto y^\top x - \frac{1}{2} x^\top A x$ has a unique minimum when is gradient is zero, namely

$$y = Ax.$$

Substituting for $x = A^{-1}y$ in $y^\top x - \frac{1}{2} x^\top A x$, we obtain

$$y^\top A^{-1} y - \frac{1}{2} y^\top A^{-1} y = -\frac{1}{2} y^\top A^{-1} y,$$

so

$$f^*(y) = -\frac{1}{2} y^\top A^{-1} y$$

with $\mathrm{dom}(f^*) = \mathbb{R}^n$.

(5) *Log-determinant*: $f(X) = \log \det(X^{-1})$, where X is an $n \times n$ symmetric positive definite matrix. Then

$$f(Y) = \log \det((-Y)^{-1}) - n,$$

where Y is an $n \times n$ symmetric negative definite matrix; see Boyd and Vandenberghe; see [22], Section 3.3.1, Example 3.23.

(6) *Norm on* $\mathbb{R}^n$: $f(x) = \|x\|$ for any norm $\| \ \|$ on $\mathbb{R}^n$, with $\mathrm{dom}(f) = \mathbb{R}^n$. Recall from Section 11.6 that the dual norm $\| \ \|^D$ of the norm $\| \ \|$ (with respect to the canonical inner product $x \cdot y = y^\top x$ on $\mathbb{R}^n$ is given by

$$\|y\|^D = \sup_{\|x\|=1} |y^\top x|,$$

and that

$$|y^\top x| \leq \|x\| \, \|y\|^D .$$

We have

$$\begin{aligned} f^*(y) &= \sup_{x \in \mathbb{R}^n} (y^\top x - \|x\|) \\ &= \sup_{x \in \mathbb{R}^n, x \neq 0} \left(y^\top \frac{x}{\|x\|} - 1 \right) \|x\| \\ &\leq \sup_{x \in \mathbb{R}^n, x \neq 0} (\|y\|^D - 1) \|x\| , \end{aligned}$$

so if $\|y\|^D > 1$ this last term goes to $+\infty$, but if $\|y\|^D \leq 1$, then its maximum is 0. Therefore,

$$f^*(y) = \|y\|^* = \begin{cases} 0 & \text{if } \|y\|^D \leq 1 \\ +\infty & \text{otherwise.} \end{cases}$$

(7) *Norm squared*: $f(x) = \frac{1}{2} \|x\|^2$ for any norm $\| \, \|$ on $\mathbb{R}^n$, with $\mathrm{dom}(f) = \mathbb{R}^n$. Since $|y^\top x| \leq \|x\| \, \|y\|^D$, we have

$$y^\top x - (1/2) \|x\|^2 \leq \|y\|^D \|x\| - (1/2) \|x\|^2 .$$

The right-hand side is a quadratic function of $\|x\|$ which achieves its maximum at $\|x\| = \|y\|^D$, with maximum value $(1/2)(\|y\|^D)^2$. Therefore

$$y^\top x - (1/2) \|x\|^2 \leq (1/2) \Big(\|y\|^D \Big)^2$$

for all x, which shows that

$$f^*(y) \leq (1/2) \Big(\|y\|^D \Big)^2 .$$

By definition of the dual norm and because the unit sphere is compact, for any $y \in \mathbb{R}^n$ there is some $x \in \mathbb{R}^n$ such that $\|x\| = 1$ and $y^\top x = \|y\|^D$, so multiplying both sides by $\|y\|^D$ we obtain

$$y^\top \|y\|^D \, x = \Big(\|y\|^D \Big)^2$$

and for $z = \|y\|^D \, x$, since $\|x\| = 1$ we have $\|z\| = \|y\|^D \, \|x\| = \|y\|^D$, so we get

$$y^\top z - (1/2)(\|z\|)^2 = \Big(\|y\|^D \Big)^2 - (1/2) \Big(\|y\|^D \Big)^2 = (1/2) \Big(\|y\|^D \Big)^2 ,$$

which shows that the upper bound $(1/2) \Big(\|y\|^D \Big)^2$ is achieved. Therefore,

$$f^*(y) = \frac{1}{2} \Big(\|y\|^D \Big)^2 ,$$

and $\mathrm{dom}(f^*) = \mathbb{R}^n$.

(8) *Log-sum-exp function*: $f(x) = \log\left(\sum_{i=1}^n e^{x_i}\right)$, where $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. To determine the values of $y \in \mathbb{R}^n$ for which the maximum of $g(x) = y^\top x - f(x)$ over $x \in \mathbb{R}^n$ is attained, we compute its gradient and we find

$$\nabla g_x = \begin{pmatrix} y_1 - \dfrac{e^{x_1}}{\sum_{i=1}^n e^{x_i}} \\ \vdots \\ y_n - \dfrac{e^{x_n}}{\sum_{i=1}^n e^{x_i}} \end{pmatrix}.$$

Therefore, $(y_1, \ldots, y_n)$ must satisfy the system of equations

$$y_j = \frac{e^{x_j}}{\sum_{i=1}^n e^{x_i}}, \quad j = 1, \ldots, n. \tag{$*$}$$

The condition $\sum_{i=1}^n y_i = 1$ is obviously necessary, as well as the conditions $y_i > 0$, for $i = 1, \ldots, n$. Conversely, if $\mathbf{1}^\top y = 1$ and $y > 0$, then $x_j = \log y_i$ for $i = 1, \ldots, n$ is a solution. Since $(*)$ implies that

$$x_i = \log y_i + \log\left(\sum_{i=1}^n e^{x_i}\right), \tag{$**$}$$

we get

$$\begin{aligned} y^\top x - f(x) &= \sum_{i=1}^n y_i x_i - \log\left(\sum_{i=1}^n e^{x_i}\right) \\ &= \sum_{i=1}^n y_i \log y_i + \sum_{i=1}^n y_i \log\left(\sum_{i=1}^n e^{x_i}\right) - \log\left(\sum_{i=1}^n e^{x_i}\right) \quad \text{by } (**) \\ &= \sum_{i=1}^n y_i \log y_i + \left(\sum_{i=1}^n y_i - 1\right) \log\left(\sum_{i=1}^n e^{x_i}\right) \\ &= \sum_{i=1}^n y_i \log y_i \qquad \text{since } \textstyle\sum_{i=1}^n y_i = 1. \end{aligned}$$

Consequently, if $f^*(y)$ is defined, then $f^*(y) = \sum_{i=1}^n y_i \log y_i$. If we agree that $0 \log 0 = 0$, then it is an easy exercise (or, see Boyd and Vandenberghe [22], Section 3.3, Example 3.25) to show that

$$f^*(y) = \begin{cases} \sum_{i=1}^n y_i \log y_i & \text{if } \mathbf{1}^\top y = 1 \text{ and } y \geq 0 \\ \infty & \text{otherwise.} \end{cases}$$

Thus we obtain the negative entropy restricted to the domain $\mathbf{1}^\top y = 1$ and $y \geq 0$.

By definition of f^* we have

$$f(x) + f^*(y) \geq x^\top y,$$

whenever the left-hand side is defined. The above is known as *Fenchel's inequality* (or *Young's inequality* if f is differentiable).

If $f\colon A \to \mathbb{R}$ is convex (so A is convex) and if $\mathbf{epi}(f)$ is closed, then it can be shown that $f^{**} = f$. In particular, this is true if $A = \mathbb{R}^n$.

If $f\colon \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable, then x^* maximizes $x^\top y - f(x)$ iff x^* minimizes $-x^\top y + f(x)$ iff

$$\nabla f_{x^*} = y,$$

and so

$$f^*(y) = (x^*)^\top \nabla f_{x^*} - f(x^*).$$

Consequently, if we can solve the equation

$$\nabla f_z = y$$

for z given y, then we obtain $f^*(y)$.

It can be shown that if f is twice differentiable, strictly convex, and surlinear, which means that

$$\lim_{\|y\| \mapsto +\infty} \frac{f(y)}{\|y\|} = +\infty,$$

then there is a unique x_y such that $\nabla f_{x_y} = y$, so that

$$f^*(y) = x_y^\top \nabla f_{x_y} - f(x_y),$$

and f^* is differentiable with

$$\nabla f_y^* = x_y.$$

We now return to our optimization problem. Consider the problem (P)

$$\begin{array}{ll} \text{minimize} & J(v) \\ \text{subject to} & Ax \leq b \\ & Cx = d \end{array}$$

with affine inequality and equality constraints (with A an $m \times n$ matrix, C an $p \times n$ matrix, $b \in \mathbb{R}^m$, $d \in \mathbb{R}^p$). We are going show that the dual function G can be expressed in terms of the conjugate J^* of J.

The Lagrangian associated with the above program is

$$\begin{aligned} L(v, \lambda, \nu) &= J(v) + (Av - b)^\top \lambda + (Cv - d)^\top \nu \\ &= -b^\top \lambda - d^\top \nu + J(v) + (A^\top \lambda + C^\top \nu)^\top v, \end{aligned}$$

with $\lambda \in \mathbb{R}^m_+$ and $\nu \in \mathbb{R}^p$. By definition

$$\begin{aligned} G(\lambda, \nu) &= -b^\top \lambda - d^\top \nu + \inf_{v \in \mathbb{R}^n} (J(v) + (A^\top \lambda + C^\top \nu)^\top v) \\ &= -b^\top \lambda - d^\top \nu - \sup_{v \in \mathbb{R}^n} (-(A^\top \lambda + C^\top \nu)^\top v - J(v)) \\ &= -b^\top \lambda - d^\top \nu - J^*(-A^\top \lambda - C^\top \nu). \end{aligned}$$

Therefore, for all $\lambda \in \mathbb{R}^m_+$ and all $\nu \in \mathbb{R}^p$, we have

$$G(\lambda, \nu) = \begin{cases} -b^\top \lambda - d^\top \nu - J^*(-A^\top \lambda - C^\top \nu) & \text{if } -A^\top \lambda - C^\top \nu \in \mathrm{dom}(J^*), \\ -\infty & \text{otherwise.} \end{cases}$$

As application of this result, consider the following example.

Example 30.10. Consider the following problem:

$$\begin{aligned} &\text{minimize} && \|v\| \\ &\text{subject to} && Av = b, \end{aligned}$$

where $\| \ \|$ is any norm on $\mathbb{R}^n$. Using the result of Example 30.9, we obtain

$$G(\nu) = -b^\top \nu - \left\| -A^\top \nu \right\|^*,$$

that is,

$$G(\nu) = \begin{cases} -b^\top \nu & \text{if } \left\| A^\top \nu \right\|^D \leq 1 \\ -\infty & \text{otherwise.} \end{cases}$$

In the special case where $\| \ \| = \| \ \|_2$, we also have $\| \ \|^D = \| \ \|_2$.

Another interesting application is to the entropy minimization problem.

Example 30.11. Consider the following problem known as *entropy minimization*:

$$\begin{aligned} &\text{minimize} && f(x) = \sum_{i=1}^n x_i \log x_i \\ &\text{subject to} && Ax \leq b \\ & && \mathbf{1}^\top x = 1, \end{aligned}$$

where $\mathrm{dom}(f) = \{x \in \mathbb{R}^n \mid x \geq 0\}$. By Example 30.9(3), the conjugate of the negative entropy function $u \log u$ is e^{v-1}, so we easily see that

$$f^*(y) = \sum_{i=1}^n e^{y_i - 1},$$

which is defined on $\mathbb{R}^n$. Using our above result, the dual function $G(\lambda, \mu)$ of the entropy minimization problem is given by

$$G(\lambda, \mu) = -b^\top \lambda - \mu - e^{-\mu-1} \sum_{i=1}^{n} e^{-(A^i)^\top \lambda},$$

for all $\lambda \in \mathbb{R}^n_+$ and all $\mu \in \mathbb{R}$, where A^i is the ith column of A. It follows that the dual program is:

$$\begin{aligned} \text{maximize} \quad & -b^\top \lambda - \mu - e^{-\mu-1} \sum_{i=1}^{n} e^{-(A^i)^\top \lambda} \\ \text{subject to} \quad & \lambda \geq 0. \end{aligned}$$

We can simplify this problem by maximizing over the variable $\mu \in \mathbb{R}$. For fixed λ, the objective function is maximized when the derivative is zero, that is,

$$-1 + e^{-\mu-1} \sum_{i=1}^{n} e^{-(A^i)^\top \lambda} = 0,$$

which yields

$$\mu = \log\left(\sum_{i=1}^{n} e^{-(A^i)^\top \lambda} \right) - 1.$$

Plugging the above value back into the objective function of the dual we obtain the following program:

$$\begin{aligned} \text{maximize} \quad & -b^\top \lambda - \log\left(\sum_{i=1}^{n} e^{-(A^i)^\top \lambda} \right) \\ \text{subject to} \quad & \lambda \geq 0. \end{aligned}$$

The entropy minimization problem is another problem for which Theorem 30.15 applies, and thus can be solved using the dual program. Indeed, the Lagrangian of the primal program is given by

$$L(x, \lambda, \mu) = \sum_{i-1}^{n} x_i \log x_i + \lambda^\top (Ax - b) + \mu(\mathbf{1}^\top x - 1).$$

Using the second derivative criterion for convexity, we see that $L(x, \lambda, \mu)$ is strictly convex for $x \in \mathbb{R}^n_+$ and is bounded below, so it has a unique minimum which is obtain by settting the Laplacian ∇L_x to zero. We have

$$\nabla L_x = \begin{pmatrix} \log x_1 + 1 + (A^1)^\top \lambda + \mu \\ \vdots \\ \log x_n + 1 + (A^n)^\top \lambda + \mu. \end{pmatrix}$$

so by setting ∇L_x to 0 we obtain

$$x_i = e^{-((A^n)^\top \lambda + \mu + 1)}, \quad i = 1, \ldots, n. \tag{$*$}$$

By Theorem 30.15, since the objective function is convex and the constraints are affine, if the primal has a solution then so does the dual, and λ and μ constitute an optimal solution of the dual, then $x = (x_1, \ldots, x_n)$ given by the equations in $(*)$ is an optimal solution of the primal.

Other examples are given in Boyd and Vandenberghe; see [22], Section 5.1.6.

The derivation of the dual function of Problem (SVM_{h1}) from Section 30.3 involves a similar type of reasoning.

Example 30.12. Consider the hard margin Problem (SVM_{h1}):

$$\begin{aligned}
&\text{maximize} \quad \delta \\
&\text{subject to} \\
&\qquad w^\top u_i - b \geq \delta \qquad i = 1, \ldots, p \\
&\qquad -w^\top v_j + b \geq \delta \qquad j = 1, \ldots, q \\
&\qquad \|w\| \leq 1,
\end{aligned}$$

which is converted to the following minimization problem:

$$\begin{aligned}
&\text{minimize} \quad -2\delta \\
&\text{subject to} \\
&\qquad w^\top u_i - b \geq \delta \qquad i = 1, \ldots, p \\
&\qquad -w^\top v_j + b \geq \delta \qquad j = 1, \ldots, q \\
&\qquad \|w\|_2 \leq 1,
\end{aligned}$$

We replaced δ by 2δ because this will make it easier to find a nice geometric interpretation. Recall from Section 30.3 that Problem (SVM_{h1}) has a an optimal solution iff $\delta > 0$, in which case $\|w\| = 1$.

The corresponding Lagrangian with $\lambda \in \mathbb{R}^p_+, \mu \in \mathbb{R}^q_+, \gamma \in \mathbb{R}^+$, is

$$\begin{aligned}
L(w, b, \delta, \lambda, \mu, \gamma) &= -2\delta + \sum_{i=1}^{p} \lambda_i (\delta + b - w^\top u_i) + \sum_{j=1}^{q} \mu_j (\delta - b + w^\top v_j) + \gamma(\|w\|_2 - 1) \\
&= w^\top \left(-\sum_{i=1}^{p} \lambda_i u_i + \sum_{j=1}^{q} \mu_j v_j \right) + \gamma \|w\|_2 + \left(\sum_{i=1}^{p} \lambda_i - \sum_{j=1}^{q} \mu_j \right) b \\
&\quad + \left(-2 + \sum_{i=1}^{p} \lambda_i + \sum_{j=1}^{q} \mu_j \right) \delta - \gamma.
\end{aligned}$$

Next to find the dual function $G(\lambda, \mu, \gamma)$ we need to minimize $L(w, b, \delta, \lambda, \mu, \gamma)$ with respect to w, b and δ, so its gradient with respect to w, b and δ, so its gradient with respect to w, b and δ must be zero. This implies that

$$\sum_{i=1}^{p} \lambda_i - \sum_{j=1}^{q} \mu_j = 0$$
$$-2 + \sum_{i=1}^{p} \lambda_i + \sum_{j=1}^{q} \mu_j = 0,$$

which yields

$$\sum_{i=1}^{p} \lambda_i = \sum_{j=1}^{q} \mu_j = 1.$$

Our minimization problem is reduced to: find

$$\begin{aligned}
&\inf_{w, \|w\| \leq 1} \left(w^\top \left(\sum_{j=1}^{q} \mu_j v_j - \sum_{i=1}^{p} \lambda_i u_i \right) + \gamma \|w\|_2 - \gamma \right) \\
&= -\gamma - \gamma \inf_{w, \|w\| \leq 1} \left(-w^\top \frac{1}{\gamma} \left(\sum_{j=1}^{q} \mu_j v_j - \sum_{i=1}^{p} \lambda_i u_i \right) + \|-w\|_2 \right) \\
&= \begin{cases} -\gamma & \text{if } \left\| \frac{1}{\gamma} \left(\sum_{j=1}^{q} \mu_j v_j - \sum_{i=1}^{p} \lambda_i u_i \right) \right\|_2^D \leq 1 \\ -\infty & \text{otherwise} \end{cases} \qquad \text{by definition of } \| \|_2^* \\
&= \begin{cases} -\gamma & \text{if } \left\| \sum_{j=1}^{q} \mu_j v_j - \sum_{i=1}^{p} \lambda_i u_i \right\|_2 \leq \gamma \\ -\infty & \text{otherwise.} \end{cases} \qquad \text{since } \| \|_2^D = \| \|_2 \text{ and } \gamma > 0
\end{aligned}$$

It is immediately verified that the above formula is still correct if $\gamma = 0$. Therefore

$$G(\lambda, \mu, \gamma) = \begin{cases} -\gamma & \text{if } \left\| \sum_{j=1}^{q} \mu_j v_j - \sum_{i=1}^{p} \lambda_i u_i \right\|_2 \leq \gamma \\ -\infty & \text{otherwise.} \end{cases}$$

Since $\left\| \sum_{j=1}^{q} \mu_j v_j - \sum_{i=1}^{p} \lambda_i u_i \right\|_2 \leq \gamma$ iff $-\gamma \leq -\left\| \sum_{j=1}^{q} \mu_j v_j - \sum_{i=1}^{p} \lambda_i u_i \right\|_2$, the dual pro-

gram, maximizing $G(\lambda, \mu, \gamma)$, is equivalent to

$$\begin{aligned} \text{maximize} \quad & -\left\| \sum_{j=1}^{q} \mu_j v_j - \sum_{i=1}^{p} \lambda_i u_i \right\|_2 \\ \text{subject to} \quad & \\ & \sum_{i=1}^{p} \lambda_i = 1, \;\; \lambda \geq 0 \\ & \sum_{j=1}^{q} \mu_j = 1, \;\; \mu \geq 0, \end{aligned}$$

equivalently

$$\begin{aligned} \text{minimize} \quad & \left\| \sum_{j=1}^{q} \mu_j v_j - \sum_{i=1}^{p} \lambda_i u_i \right\|_2 \\ \text{subject to} \quad & \\ & \sum_{i=1}^{p} \lambda_i = 1, \;\; \lambda \geq 0 \\ & \sum_{j=1}^{q} \mu_j = 1, \;\; \mu \geq 0. \end{aligned}$$

Geometrically, $\sum_{i=1}^{p} \lambda_i u_i$ with $\sum_{i=1}^{p} \lambda_i = 1$ and $\lambda \geq 0$ is a convex combinations of the u_is, and $\sum_{j=1}^{q} \mu_j v_j$ with $\sum_{j=1}^{q} \mu_j = 1$ and $\mu \geq 0$ is a convex combination of the v_js, so the dual program is to minimize the distance between the polyhedron $\text{conv}(u_1, \ldots, u_p)$ (the convex hull of the u_is) and the polyhedron $\text{conv}(v_1, \ldots, v_q)$ (the convex hull of the v_js). Since both polyhedra are compact, the shortest distance between then is achieved. In fact, there is some vertex u_i such that if $P(u_i)$ is its projection onto $\text{conv}(v_1, \ldots, v_q)$ (which exists by Hilbert space theory), then the length of the line segment $(u_i, P(u_i))$ is the shortest distance between the two polyhedra (and similarly there is some vertex v_j such that if $P(v_j)$ is its projection onto $\text{conv}(u_1, \ldots, u_p)$ then the length of the line segment $(v_j, P(v_j))$ is the shortest distance between the two polyhedra).

If the two subsets are separable, in which case Problem (SVM_{h1}) has an optimal solution $\delta > 0$, because the objective function is convex and the convex constraint $\|w\|_2 \leq 1$ is qualified since δ may be negative, by Theorem 30.14(2) the duality gap is zero, so δ is half of the minimum distance between the two convex polyhedra $\text{conv}(u_1, \ldots, u_p)$ and $\text{conv}(v_1, \ldots, v_q)$; see Figure 30.19.

It should be noted that the constraint $\|w\| \leq 1$ yields a formulation of the dual problem which has the advantage of having a nice geometric interpretation: finding the minimal

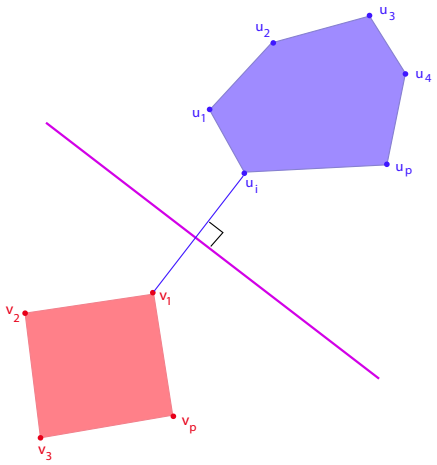


Figure 30.19: In $\mathbb{R}^2$ the convex hull of the u_is, namely the blue hexagon, is separated from the convex hull of the v_js, i.e. the red square, by the purple hyperplane (line) which is the perpendicular bisector to the blue line segment between u_i and v_1, where this blue line segment is the shortest distance between the two convex polygons.

distance between the convex polyhedra $\mathrm{conv}(u_1, \ldots, u_p)$ and $\mathrm{conv}(v_1, \ldots, v_q)$. Unfortunately this formulation is not useful for actually solving the problem. However, if the equivalent constraint $\|w\|^2 \, (= w^\top w) \leq 1$ is used, then the dual problem is much more useful as a solving tool.

In the next chapter we consider the case where the sets of points $\{u_1, \ldots, u_p\}$ and $\{v_1, \ldots, v_q\}$ are not linearly separable.

30.8 Some Techniques to Obtain a More Useful Dual Program

In some cases, it is advantageous to reformulate a primal optimization problem to obtain a more useful dual problem. Three different reformulations are proposed in Boyd and Vandenberghe; see [22], Section 5.7:

(1) Introducing new variables and associated equality constraints.

(2) Replacing the objective function with an increasing function of the the original function.

(3) Making explicit constraints implicit, that is, incorporating them into the domain of the objective function.

We only give illustrations of (1) and (2), and refer the reader to Boyd and Vandenberghe [22] (Section 5.7) for more examples of these techniques.

Consider the unconstrained program:

$$\text{minimize} \quad f(Ax + b),$$

where A is an $m \times n$ matrix and $b \in \mathbb{R}^m$. While the conditions for a zero duality gap are satisfied, the Lagrangian is

$$L(x) = f(Ax + b),$$

so the dual function G is the constant function whose value is

$$G = \inf_{x \in \mathbb{R}^n} f(Ax + b),$$

which is not useful at all.

Let us reformulate the problem as

$$\begin{aligned} &\text{minimize} && f(y) \\ &\text{subject to} \\ & && Ax + b = y, \end{aligned}$$

where we introduced the new variable $y \in \mathbb{R}^m$ and the equality constraint $Ax + b = y$. The two problems are obviously equivalent. The Lagrangian of the reformulated problem is

$$L(x, y, \mu) = f(y) + \mu^\top (Ax + b - y)$$

where $\mu \in \mathbb{R}^m$. To find the dual function $G(\mu)$ we minimize $L(x, y, \mu)$ over x and y. Minimizing over x we see that $G(\mu) = -\infty$ unless $A^\top \mu = 0$, in which case we are left with

$$G(\mu) = b^\top \mu + \inf_y (f(y) - \mu^\top y) = b^\top \mu - \inf_y (\mu^\top y - f(y)) = b^\top \mu - f^*(\mu),$$

where f^* is the conjugate of f. It follows that the dual program can be expressed as

$$\begin{aligned} &\text{maximize} && b^\top \mu - f^*(\mu) \\ &\text{subject to} \\ & && A^\top \mu = 0. \end{aligned}$$

This formulation of the dual is much more useful than the dual of the original program.

Example 30.13. As a concrete example, consider the following unconstrained program:

$$\text{minimize} \quad f(x) = \log\left(\sum_{i=1}^{n} e^{(a^i)^\top x + b_i}\right)$$

where a^i is a column vector in $\mathbb{R}^n$. We reformulate the problem by introducing new variables and equality constraints as follows:

$$\begin{aligned} &\text{minimize} \quad f(y) = \log\left(\sum_{i=1}^{n} e^{y_i}\right) \\ &\text{subject to} \\ &\qquad\qquad Ax + b = y, \end{aligned}$$

where A is the matrix whose columns are the vectors a^i and $b = (b_1, \ldots, b_n)$. Since by Example 30.9(8) the conjugate of the log-sum-exp function $f(y) = \log\left(\sum_{i=1}^{n} e^{y_i}\right)$ is

$$f^*(\mu) = \begin{cases} \sum_{i=1}^{n} \mu_i \log \mu_i & \text{if } \mathbf{1}^\top \mu = 1 \text{ and } \mu \geq 0 \\ \infty & \text{otherwise,} \end{cases}$$

the dual of the reformulated problem can be expressed as

$$\begin{aligned} &\text{maximize} \quad b^\top \mu - \log\left(\sum_{i=1}^{n} \mu_i \log \mu_i\right) \\ &\text{subject to} \\ &\qquad\qquad \mathbf{1}^\top \mu = 1 \\ &\qquad\qquad A^\top \mu = 0 \\ &\qquad\qquad\quad \mu \geq 0, \end{aligned}$$

an entropy maximization problem.

Example 30.14. Similarly the unconstrained norm minimization problem

$$\text{minimize} \quad \|Ax - b\|,$$

where $\| \|$ is any norm on $\mathbb{R}^m$, has a dual function which is a constant, and is not useful. This problem can be reformulated as

$$\begin{aligned} &\text{minimize} \quad \|y\| \\ &\text{subject to} \\ &\qquad\qquad Ax - b = y. \end{aligned}$$

By Example 30.9(6), the conjugate of the norm is given by

$$\|y\|^* = \begin{cases} 0 & \text{if } \|y\|^D \leq 1 \\ +\infty & \text{otherwise,} \end{cases}$$

so the dual of the reformulated program is:

$$\begin{aligned} &\text{maximize} && b^\top \mu \\ &\text{subject to} && \\ & && \|\mu\|^D \leq 1 \\ & && A^\top \mu = 0. \end{aligned}$$

Here is now an example of (2), replacing the objective function with an increasing function of the the original function.

Example 30.15. The norm minimization of Example 30.14 can be reformulated as

$$\begin{aligned} &\text{minimize} && \frac{1}{2}\|y\|^2 \\ &\text{subject to} && \\ & && Ax - b = y. \end{aligned}$$

This program is obviously equivalent to the original one. By Example 30.9(7), the conjugate of the square norm is given by

$$\frac{1}{2}\Big(\|y\|^D\Big)^2,$$

so the dual of the reformulated program is

$$\begin{aligned} &\text{maximize} && -\frac{1}{2}\Big(\|\mu\|^D\Big)^2 + b^\top \mu \\ &\text{subject to} && \\ & && A^\top \mu = 0. \end{aligned}$$

Note that this dual is different from the dual obtained in Example 30.14.

The objective function of the dual program in Example 30.14 is linear, but we have the nonlinear constraint $\|\mu\|^D \leq 1$. On the other hand, the objective function of the dual program of Example 30.15 is quadratic, whereas its constraints are affine. We have other examples of this trade-off with the Programs (SVM_{h2}) (quadratic objective function, affine constraints), and (SVM_{h1}) (linear objective function, one nonlinear constraint).

In the next example we revisit the problem of solving an overdetermined or underdetermined linear system $Ax = b$ considered in Section 16.1 from a different point of view.

Example 30.16. (Ridge regression)

The problem of solving an overdetermined or underdetermined linear system $Ax = y$ arises as a "learning problem" in which we observe a sequence of data $((a_1, y_1), \ldots, (a_m, y_m))$, where $a_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$, viewed as input-output pairs of some unknown function f that we are trying to infer. The simplest kind of function is a linear function $f(x) = x^\top w$,

where $w \in \mathbb{R}^n$ is a vector of coefficients usually called a *weight vector*. Since the problem is overdetermined and since our observations may be subject to errors, we can't solve for w exactly as the solution of the system $Aw = y$, so instead we solve the least-square problem of minimizing $\|Aw - y\|^2$.

In Section 16.1 we showed that this problem can be solved using the pseudo-inverse. We know that the minimizers w are solutions of the normal equations $A^\top Aw = A^\top y$, but when $A^\top A$ is not invertible, such a solution is not unique so some criterion has to be used to choose among these solutions.

The pseudo-inverse does so in a specific way that sets some of the components to 0. This is not always desirable and another way is to control the size of w by adding a regularization term to $\|Aw - y\|^2$, and a natural candidate is $\|w\|^2$. It is also customary to view each row of the matrix A as the transpose of an input vector $x_i \in \mathbb{R}^n$, and to define the $m \times n$ matrix X as

$$X = \begin{pmatrix} x_1^\top \\ \vdots \\ x_m^\top \end{pmatrix},$$

where the row vectors $x_i^\top$ are the rows of X, and thus the $x_i \in \mathbb{R}^n$ are column vectors. Our optimization problem, called *ridge regression*, is the problem ($RR1$):

$$\text{minimize} \quad \|y - Xw\|^2 + K\|w\|^2,$$

which by introducing the new variable $\xi = y - Xw$ can be rewritten as ($RR2$):

$$\begin{aligned} &\text{minimize} && \xi^\top \xi + Kw^\top w \\ &\text{subject to} && \\ & && y - Xw = \xi, \end{aligned}$$

where $K > 0$ is some constant determining the influence of the regularizing term $w^\top w$.

The objective function of the first version of our minimization problem can be expressed as

$$\begin{aligned} J(w) &= \|y - Xw\|^2 + K\|w\|^2 \\ &= (y - Xw)^\top (y - Xw) + Kw^\top w \\ &= y^\top y - 2w^\top X^\top y + w^\top X^\top Xw + Kw^\top w \\ &= w^\top (X^\top X + KI_n)w - 2w^\top X^\top y + y^\top y. \end{aligned}$$

The matrix $X^\top X$ is symmetric positive semidefinite and $K > 0$, so the matrix $X^\top X + KI_n$ is positive definite. It follows that

$$J(w) = w^\top (X^\top X + KI_n)w - 2w^\top X^\top y + y^\top y$$

is strictly convex, so it has a unique minimun iff $\nabla J_w = 0$. Since

$$\nabla J_w = 2(X^\top X + KI_n)w - 2X^\top y,$$

we deduce that

$$w = (X^\top X + KI_n)^{-1} X^\top y. \tag{$*_{wp}$}$$

The dual function of the first formulation of our problem is a constant function (with value the minimum of J) so it is not useful, but the second formulation of our problem yields an interesting dual problem. The Lagrangian is

$$\begin{aligned} L(\xi, w, \lambda) &= \xi^\top \xi + Kw^\top w + (y - Xw - \xi)^\top \lambda \\ &= \xi^\top \xi + Kw^\top w - w^\top X^\top \lambda - \xi^\top \lambda + \lambda^\top y. \end{aligned}$$

To derive the dual function $G(\lambda)$ we minimize $L(\xi, w, \lambda)$ with respect to ξ and w, and for this we set the gradient of $\nabla L_{\xi,w}$ to zero. Since

$$\nabla L_{\xi,w} = \begin{pmatrix} 2\xi - \lambda \\ 2Kw - X^\top \lambda \end{pmatrix},$$

we get

$$\begin{aligned} \lambda &= 2\xi \\ w &= \frac{1}{2K} X^\top \lambda = X^\top \frac{\xi}{K}. \end{aligned}$$

The above suggests defining the variable α so that $\xi = K\alpha$, so we have $\lambda = 2K\alpha$ and $w = X^\top \alpha$. Then we obtain the dual function as a function of α by substituting the above values of ξ, λ and w back in the Lagrangian and we get

$$\begin{aligned} G(\alpha) &= K^2 \alpha^\top \alpha + K\alpha^\top XX^\top \alpha - 2K\alpha^\top XX^\top \alpha - 2K^2 \alpha^\top \alpha + 2K\alpha^\top y \\ &= -K\alpha^\top (XX^\top + KI_m)\alpha + 2K\alpha^\top y. \end{aligned}$$

This is a strictly concave function so its maximum is achieved iff $\nabla G_\alpha = 0$, that is,

$$2K(XX^\top + KI_m)\alpha = 2Ky,$$

which yields

$$\alpha = (XX^\top + KI_m)^{-1} y.$$

Putting everything together we obtain

$$\begin{aligned} \alpha &= (XX^\top + KI_m)^{-1} y \\ w &= X^\top \alpha \\ \xi &= K\alpha, \end{aligned}$$

which yields

$$w = X^\top (XX^\top + KI_m)^{-1} y. \tag{$*_{wd}$}$$

Earlier in ($*_{wp}$) we found that

$$w = (X^\top X + KI_n)^{-1} X^\top y,$$

and it is easy to check that

$$(X^\top X + KI_n)^{-1} X^\top = X^\top (XX^\top + KI_m)^{-1}.$$

One interesting aspect of the dual is that it shows that the solution w being of the form $X^\top \alpha$, is linear combination

$$w = \sum_{i=1}^{m} \alpha_i x_i$$

of the data points x_i, with the coefficients α_i corresponding to the dual variable $\lambda = 2K\alpha$ of the dual function, and with

$$\alpha = (XX^\top + KI_m)^{-1} y.$$

If m is smaller than n, then it is more advantageous to solve for α. But what really makes the dual interesting is that with our definition of X as

$$X = \begin{pmatrix} x_1^\top \\ \vdots \\ x_m^\top \end{pmatrix},$$

the matrix $XX^\top$ consists of the inner products $x_i^\top x_j$, and similarly the function learned $f(x) = w^\top x$ can be expressed as

$$f(x) = \sum_{i=1}^{m} \alpha_i x_i^\top x,$$

namely that both w and $f(x)$ are given *in terms of the inner products* $x_i^\top x_j$ and $x_i^\top x$.

This fact is the key to a generalization to ridge regression in which the input space $\mathbb{R}^n$ is embedded in a larger (possibly infinite dimensional) Euclidean space F (with an inner product $\langle -, - \rangle$) usually called a *feature space*, using a function

$$\varphi \colon \mathbb{R}^n \to F.$$

The problem becomes (*kernel ridge regression*):

$$\begin{aligned} &\text{minimize} \quad \xi^\top \xi + K\langle w, w \rangle \\ &\text{subject to} \\ &\qquad y_i - \langle w, \varphi(x_i) \rangle = \xi_i, \quad i = 1, \ldots, m. \end{aligned}$$

This problem is discussed in Shawe–Taylor and Christianini [96] (Section 7.3).

We will show below that the solution is exactly the same:

$$\begin{aligned}\alpha &= (\mathbf{G} + KI_m)^{-1}y\\ w &= \sum_{i=1}^{m}\alpha_i\varphi(x_i)\\ \xi &= K\alpha,\end{aligned}$$

where $\mathbf{G}$ is the Gram matrix given by $\mathbf{G}_{ij} = \langle\varphi(x_i), \varphi(x_j)\rangle$.

In this framework, we have to be a little careful in using gradients since the inner product $\langle -, -\rangle$ on F is involved and F could be infinite dimensional, but this causes no problem because we can use derivatives, and by Proposition 19.5 we have

$$d\langle -, -\rangle_{(u,v)}(x, y) = \langle x, v\rangle + \langle u, y\rangle.$$

This implies that the derivative of the map $u \mapsto \langle u, u\rangle$ is

$$d\langle -, -\rangle_u(x) = 2\langle x, u\rangle.$$

Since the map $u \mapsto \langle u, v\rangle$ (with v fixed) is linear, its derivative is

$$d\langle -, v\rangle_u(x) = \langle x, v\rangle.$$

The derivative of the Lagrangian

$$L(\xi, w, \lambda) = \xi^\top\xi + K\langle w, w\rangle - \sum_{i=1}^{m}\lambda_i\langle\varphi(x_i), w\rangle - \xi^\top\lambda + \lambda^\top y$$

with respect to ξ and w is

$$dL_{\xi,w}\big(\widetilde{\xi}, \widetilde{w}\big) = 2(\widetilde{\xi})^\top\xi - (\widetilde{\xi})^\top\lambda + \left\langle 2Kw - \sum_{i=1}^{m}\lambda_i\varphi(x_i), \widetilde{w}\right\rangle.$$

We have $dL_{\xi,w}\big(\widetilde{\xi}, \widetilde{w}\big) = 0$ for all $\widetilde{\xi}$ and $\widetilde{w}$ iff

$$\begin{aligned}2Kw &= \sum_{i=1}^{m}\lambda_i\varphi(x_i)\\ \lambda &= 2\xi.\end{aligned}$$

Again we define $\xi = K\alpha$, so we have $\lambda = 2K\alpha$, and

$$w = \sum_{i=1}^{m}\alpha_i\varphi(x_i).$$

Plugging back into the Lagrangian we get

$$\begin{aligned} G(\alpha) &= K^2\alpha^\top\alpha + K\sum_{i,j=1}^m \alpha_i\alpha_j\langle\varphi(x_i),\varphi(x_j)\rangle - 2K\sum_{i,j=1}^m \alpha_i\alpha_j\langle\varphi(x_i),\varphi(x_j)\rangle \\ &\quad - 2K^2\alpha^\top\alpha + 2K\alpha^\top y \\ &= -K^2\alpha^\top\alpha - K\sum_{i,j=1}^m \alpha_i\alpha_j\langle\varphi(x_i),\varphi(x_j)\rangle + 2K\alpha^\top y. \end{aligned}$$

If $\mathbf{G}$ is the matrix given by $\mathbf{G}_{ij} = \langle\varphi(x_i),\varphi(x_j)\rangle$, then we have

$$G(\alpha) = -K\alpha^\top(\mathbf{G} + KI_m)\alpha + 2K\alpha^\top y.$$

The function G is strictly concave and has a maximum for

$$\alpha = (\mathbf{G} + KI_m)^{-1}y,$$

as claimed earlier.

It is easy to adapt the above method to learn an affine function $f(w) = x^\top w + b$ instead of a linear function $f(w) = x^\top w$, where $b \in \mathbb{R}$. We have the minimization problem:

$$\begin{aligned} &\text{minimize} \quad \xi^\top\xi + K\langle w, w\rangle + Kb^2 \\ &\text{subject to} \\ &\qquad y_i - \langle w, \varphi(x_i)\rangle - b = \xi_i, \quad i = 1,\ldots,m. \end{aligned}$$

We leave it as an exercise to show that

$$L(\xi, w, b, \lambda) = \xi^\top\xi + K\langle w, w\rangle + Kb^2 - \sum_{i=1}^m \lambda_i\langle\varphi(x_i), w\rangle - \xi^\top\lambda - b\mathbf{1}_m^\top\lambda + \lambda^\top y,$$

so $dL_{\xi,w,b} = 0$ yields

$$\begin{aligned} 2\xi &= \lambda \\ 2Kw &= \sum_{i=1}^m \lambda_i\varphi(x_i) \\ 2Kb &= \mathbf{1}_m^\top\lambda, \end{aligned}$$

so by setting $\xi = K\alpha$ the dual function G is given by

$$G(\alpha) = -K\alpha^\top(\mathbf{G} + \mathbf{1}_m\mathbf{1}_m^\top + KI_m)\alpha + 2K\alpha^\top y,$$

and we obtain

$$\begin{aligned} \alpha &= (\mathbf{G} + \mathbf{1}_m\mathbf{1}_m^\top + KI_m)^{-1}y \\ w &= \sum_{i=1}^m \alpha_i\varphi(x_i) \\ b &= \mathbf{1}_m^\top\alpha. \end{aligned}$$

It is easy to see that $\mathbf{G} + \mathbf{1}_m\mathbf{1}_m^\top$ is symmetric positive semidefinite, so $\mathbf{G} + \mathbf{1}_m\mathbf{1}_m^\top + KI_m$ is invertible.

Since the dimension of the feature space F may be very large, one might worry that computing the inner products $\langle\varphi(x_i), \varphi(x_j)\rangle$ might be very expensive. This is where kernel functions come to the rescue. A *kernel function* κ for an embedding $\varphi\colon \mathbb{R}^n \to F$ is a map $\kappa\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ with the property that

$$\kappa(u, v) = \langle\varphi(u), \varphi(v)\rangle \quad \text{for all } u, v \in \mathbb{R}^n.$$

If $\kappa(u, v)$ can be computed in a reasonably cheap way, and if $\varphi(u)$ can be computed cheaply, then the inner products $\langle\varphi(x_i), \varphi(x_j)\rangle$ (and $\langle\varphi(x_i), \varphi(x)\rangle$) can be computed cheaply. Fortunately there are good kernel functions. Two very good sources on kernel methods are Schölkopf and Smola [85] and Shawe–Taylor and Christianini [96]. We will investigate kernels in Chapter 31.

Sometimes, it is also helpful to replace a constraint by an increasing function of this constraint; for example, to use the constraint $\|w\|_2^2 (= w^\top w) \leq 1$ instead of $\|w\|_2 \leq 1$.

30.9 Uzawa's Method

Let us go back to our minimization problem

$$\begin{aligned} &\text{minimize} \quad J(v) \\ &\text{subject to} \quad \varphi_i(v) \leq 0, \quad i = 1, \ldots, m, \end{aligned}$$

where the functions J and φ_i are defined on some open subset Ω of a finite-dimensional Euclidean vector space V (more generally, a real Hilbert space V). As usual, let

$$U = \{v \in V \mid \varphi_i(v) \leq 0,\ 1 \leq i \leq m\}.$$

If the functional J satisfies the inequalities of Proposition 29.14 and if the functions φ_i are convex, in theory, the projected-gradient method converges to the unique minimizer of J over U. Unfortunately, it is usually impossible to compute the projection map $p_U\colon V \to U$.

On the other hand, the domain of the Lagrange dual function $G\colon \mathbb{R}_+^m \to \mathbb{R}$ given by

$$G(\mu) = \inf_{v \in \Omega} L(v, \mu) \quad \mu \in \mathbb{R}_+^m,$$

is $\mathbb{R}_+^m$, where

$$L(v, \mu) = J(v) + \sum_{i=1}^m \mu_i \varphi_i(v)$$

is the Lagrangian of our problem. Now the projection p_+ from $\mathbb{R}^m$ to $\mathbb{R}_+^m$ is very simple, namely

$$(p_+(\lambda))_i = \max\{\lambda_i, 0\}, \quad 1 \leq i \leq m.$$

It follows that the projection-gradient method should be applicable to the dual problem (D):

$$\begin{aligned} &\text{maximize} && G(\mu) \\ &\text{subject to} && \mu \in \mathbb{R}^m_+. \end{aligned}$$

If the hypotheses of Theorem 30.14 hold, then a solution λ of the dual program (D) yields a solution u_λ of the primal problem.

Uzawa's method is essentially the gradient method with fixed stepsize applied to the dual problem (D). However, it is designed to yield a solution of the primal problem.

Uzawa's method:

Given an arbitrary initial vectors $\lambda^0 \in \mathbb{R}^m_+$, two sequences $(\lambda^k)_{k\geq 0}$ and $(u^k)_{k\geq 0}$ are constructed, with $\lambda^k \in \mathbb{R}^m_+$ and $u^k \in V$.

Assuming that $\lambda^0, \lambda^1, \ldots, \lambda^k$ are known, u^k and λ^{k+1} are determined as follows:

u^k is the unique solution of the minimization problem, find $u^k \in V$ such that

$$\text{(UZ)} \quad \begin{cases} J(u^k) + \sum_{i=1}^m \lambda_i^k \varphi_i(u^k) = \inf_{v\in V}\Big(J(v) + \sum_{i=1}^m \lambda_i^k \varphi_i(v)\Big); \text{ and} \\ \lambda_i^{k+1} = \max\{\lambda_i^k + \rho\varphi_i(u^k), 0\}, \quad 1 \leq i \leq m, \end{cases}$$

where $\rho > 0$ is a suitably chosen parameter.

Recall that the proof of Theorem 30.14 shows that

$$G'_{\lambda^k}(\xi) = \langle \nabla G_{\lambda^k}, \xi\rangle = \sum_{i=1}^m \xi_i \varphi_i(u^k),$$

which means that $(\nabla G_{\lambda^k})_i = \varphi_i(u^k)$. Then the second equation in (UZ) corresponds to the gradient-projection step

$$\lambda^{k+1} = p_+(\lambda^k + \rho \nabla G_{\lambda^k}).$$

Note that because the problem is a maximization problem we use a positive sign instead of a negative sign. Uzawa's method is indeed a gradient method.

Basically, Uzawa's method replaces a constrained optimization problem by a sequence of unconstrained optimization problems involving the Lagrangian of the (primal) problem.

Interestingly, under certain hypotheses, it is possible to prove that the sequence of approximate solutions $(u_k)_{k\geq 0}$ converges to the minimizer u of J over U, even if the sequence $(\lambda^k)_{k\geq 0}$ does not converge. We prove such a result when the constraints φ_i are affine.

Theorem 30.17. *Suppose $J\colon \mathbb{R}^n \to \mathbb{R}$ is an elliptic functional, which means that J is continuously differentiable on $\mathbb{R}^n$, and there is some constant $\alpha > 0$ such that*

$$\langle \nabla J_v - \nabla J_u, v - u\rangle \geq \alpha \|v - u\|^2 \quad \textit{for all } u, v \in V,$$

and that U is a nonempty closed convex subset given by

$$U = \{v \in \mathbb{R}^n \mid Cv \leq d\},$$

where C is a real $m \times n$ matrix and $d \in \mathbb{R}^m$. If the scalar ρ satisfies the condition

$$0 < \rho < \frac{2\alpha}{\|C\|_2^2},$$

where $\|C\|_2$ is the spectral norm of C, then the sequence $(u^k)_{k\geq 0}$ computed by Uzawa's method converges to the unique minimizer $u \in U$ of J.

Furthermore, if C has rank m, then the sequence $(\lambda^k)_{k\geq 0}$ converges to the unique maximizer of the dual problem (D).

Proof.

Step 1. We establish algebraic conditions relating the unique minimizer $u \in U$ of J over U and some $\lambda \in \mathbb{R}^m_+$ such that (u, λ) is a saddle point.

Since J is elliptic and U is nonempty closed and convex, by Theorem 29.7, the functional J is strictly convex, so it has a unique minimizer $u \in U$. Since J is convex and the constraints are affine, by Theorem 30.14(2) the dual problem (D) has at least one solution. By Theorem 30.12(2), there is some $\lambda \in \mathbb{R}^m_+$ such that (u, λ) is a saddle point of the Lagrangian L.

If we define the affine function φ by

$$\varphi(v) = (\varphi_1(v), \ldots, \varphi_m(v)) = Cv - d,$$

then the Lagrangian $L(v, \mu)$ can be written as

$$L(v, \mu) = J(v) + \sum_{i=1}^{m} \mu_i \varphi_i(v) = J(v) + \langle C^\top \mu, v\rangle - \langle \mu, d\rangle.$$

Since

$$L(u, \lambda) = \inf_{v \in \mathbb{R}^n} L(v, \lambda),$$

by Theorem 20.11(4) we must have

$$\nabla J_u + C^\top \lambda = 0, \tag{$*_1$}$$

and since

$$G(\lambda) = L(u, \lambda) = \sup_{\mu \in \mathbb{R}^m_+} L(u, \mu),$$

by Theorem 20.11(3) (and since maximing a function g is equivalent to minimizing $-g$), we must have

$$G'_\lambda(\mu - \lambda) \leq 0 \quad \text{for all } \mu \in \mathbb{R}^m_+,$$

and since $\nabla G_\lambda = \varphi(u)$, we get

$$\langle \varphi(u), \mu - \lambda \rangle \leq 0 \quad \text{for all } \mu \in \mathbb{R}^m_+. \tag{$*_2$}$$

As in the proof of Proposition 29.14, $(*_2)$ can be expressed as follows for every $\rho > 0$:

$$\langle \lambda - (\lambda + \rho\varphi(u)), \mu - \lambda \rangle \geq 0 \quad \text{for all } \mu \in \mathbb{R}^m_+, \tag{$**_2$}$$

which shows that λ can be viewed as the projection onto $\mathbb{R}^m_+$ of the vector $\lambda + \rho\varphi(u)$. In summary we obtain the equations

$$(\dagger_1) \quad \begin{cases} \nabla J_u + C^\top \lambda = 0 \\ \lambda = p_+(\lambda + \rho\varphi(u)). \end{cases}$$

Step 2. We establish algebraic conditions relating the unique solution u_k of the minimization problem arising during an iteration of Uzawa's method in (UZ) and λ^k.

Observe that the Lagrangian $L(v, \mu)$ is strictly convex as a function of v (as the sum of a strictly convex function and an affine function). As in the proof of Theorem 29.7, we have

$$\begin{aligned} J(v) + \langle C^\top \mu, v \rangle &\geq J(0) + \langle \nabla J_0, v \rangle + \frac{\alpha}{2} \|v\|^2 + \langle C^\top \mu, v \rangle \\ &\geq J(0) - \|\nabla J_0\| \|v\| - \left\|C^\top \mu\right\| \|v\| + \frac{\alpha}{2} \|v\|^2, \end{aligned}$$

and the term $(-\|\nabla J_0\| - \left\|C^\top \mu\right\| \|v\| + \frac{\alpha}{2} \|v\|) \|v\|$ goes to $+\infty$ when $\|v\|$ tends to $+\infty$, so $L(v, \mu)$ is coercive as a function of v. Therefore, the minimization problem find u^k such that

$$J(u^k) + \sum_{i=1}^m \lambda_i^k \varphi_i(u^k) = \inf_{v \in \mathbb{R}^n} \left(J(v) + \sum_{i=1}^m \lambda_i^k \varphi_i(v) \right)$$

has a unique solution $u^k \in \mathbb{R}^n$. It follows from Theorem 20.11(4) that the vector u^k must satisfy the equation

$$\nabla J_{u^k} + C^\top \lambda^k = 0, \tag{$*_3$}$$

and since by definition of Uzawa's method

$$\lambda^{k+1} = p_+(\lambda^k + \rho\varphi(u^k)), \tag{$*_4$}$$

we obtain the equations

$$(\dagger_2) \quad \begin{cases} \nabla J_{u^k} + C^\top \lambda^k = 0 \\ \lambda^{k+1} = p_+(\lambda^k + \rho\varphi(u^k)). \end{cases}$$

Step 3. By subtracting the first of the two equations of $(\dagger_1)$ and $(\dagger_2)$ we obtain

$$\nabla J_{u^k} - \nabla J_u + C^\top(\lambda^k - \lambda) = 0,$$

and by subtracting the second of the two equations of ($\dagger_1$) and ($\dagger_2$) and using Proposition 28.6, we obtain

$$\left\|\lambda^{k+1} - \lambda\right\| \leq \left\|\lambda^k - \lambda + \rho C(u^k - u)\right\|.$$

In summary, we proved

$$(\dagger) \quad \begin{cases} \nabla J_{u^k} - \nabla J_u + C^\top(\lambda^k - \lambda) = 0 \\ \left\|\lambda^{k+1} - \lambda\right\| \leq \left\|\lambda^k - \lambda + \rho C(u^k - u)\right\|. \end{cases}$$

Step 4. Convergence of the sequence $(u^k)_{k\geq 0}$ to u.

Squaring both sides of the inequality in ($\dagger$) we obtain

$$\left\|\lambda^{k+1} - \lambda\right\|^2 \leq \left\|\lambda^k - \lambda\right\|^2 + 2\rho\langle C^\top(u^k - u), u_k - u\rangle + \rho^2 \left\|u^k - u\right\|^2.$$

Using the equation in ($\dagger$) and the inequality

$$\langle \nabla J_{u^k} - \nabla J_u, u^k - u\rangle \geq \alpha \left\|u^k - u\right\|^2,$$

we get

$$\begin{aligned} \left\|\lambda^{k+1} - \lambda\right\|^2 &\leq \left\|\lambda^k - \lambda\right\|^2 - 2\rho\langle \nabla J_{u^k} - \nabla J_u, u^k - u\rangle + \rho^2 \left\|u^k - u\right\|^2 \\ &\leq \left\|\lambda^k - \lambda\right\|^2 - \rho(2\alpha - \rho \left\|C\right\|_2^2) \left\|u^k - u\right\|^2. \end{aligned}$$

Consequently, if

$$0 \leq \rho \leq \frac{2\alpha}{\left\|C\right\|_2^2},$$

we have

$$\left\|\lambda^{k+1} - \lambda\right\| \leq \left\|\lambda^k - \lambda\right\|, \quad \text{for all } k \geq 0. \tag{$*_5$}$$

By ($*_5$), the sequence $(\left\|\lambda^k - \lambda\right\|)_{k\geq 0}$ is nonincreasing and bounded below by 0, so it converges, which implies that

$$\lim_{k\mapsto\infty}\Big(\left\|\lambda^{k+1} - \lambda\right\| - \left\|\lambda^k - \lambda\right\|\Big) = 0,$$

and since

$$\left\|\lambda^{k+1} - \lambda\right\|^2 \leq \left\|\lambda^k - \lambda\right\|^2 - \rho(2\alpha - \rho \left\|C\right\|_2^2) \left\|u^k - u\right\|^2,$$

we also have

$$\rho(2\alpha - \rho \left\|C\right\|_2^2) \left\|u^k - u\right\|^2 \leq \left\|\lambda^k - \lambda\right\|^2 - \left\|\lambda^{k+1} - \lambda\right\|^2,$$

so if

$$0 < \rho < \frac{2\alpha}{\left\|C\right\|_2^2},$$

then $\rho(2\alpha - \rho\,\|C\|_2^2) > 0$, and we conclude that

$$\lim_{k\mapsto\infty} \left\|u^k - u\right\| = 0,$$

that is, the sequence $(u^k)_{k\geq 0}$ converges to u.

Step 5. Convergence of the sequence $(\lambda^k)_{k\geq 0}$ to λ if C has rank m.

Since the sequence $(\left\|\lambda^k - \lambda\right\|)_{k\geq 0}$ is nonincreasing the sequence $(\lambda^k)_{k\geq 0}$ is bounded, and thus it has a convergent subsequence $(\lambda^{i(k)})_{i\geq 0}$ whose limit is some $\lambda' \in \mathbb{R}^m_+$. Since J' is continuous, by $(\dagger_2)$ we have

$$\nabla J_u + C^\top \lambda' = \lim_{i\mapsto\infty} (\nabla J_{u^{i(k)}} + C^\top \lambda^{i(k)}) = 0. \tag{$*_6$}$$

If C has rank m, then $\mathrm{Im}(C) = \mathbb{R}^m$, which is equivalent to $\mathrm{Ker}\,(C^\top) = (0)$, so $C^\top$ is injective and since by $(\dagger_1)$ we also have $\nabla J_u + C^\top \lambda = 0$, we conclude that $\lambda' = \lambda$. The above reasoning applies to any subsequence of $(\lambda^k)_{k\geq 0}$, so $(\lambda^k)_{k\geq 0}$ converges to λ. □

In the special case where J is an elliptic quadratic functional

$$J(v) = \frac{1}{2}\langle Av, v\rangle - \langle b, v\rangle,$$

where A is symmetric positive definite, an iteration of Uzawa's method gives

$$\begin{aligned} &Au^k - b + C^\top \lambda^k = 0\\ &\lambda_i^{k+1} = \max\{(\lambda^k + \rho(Cu^k - d))_i, 0\}, \quad 1 \leq i \leq m. \end{aligned}$$

Theorem 30.17 implies that Uzawa's method converges if

$$0 < \rho < \frac{2\lambda_1}{\|C\|_2^2},$$

where λ_1 is the smallest eigenvalue of A.

If we solve for u^k using the first equation, we get

$$\lambda^{k+1} = p_+(\lambda^k + \rho(-CA^{-1}C^\top \lambda^k + CA^{-1}b - d)). \tag{$*_7$}$$

In Example 30.7 we showed that the gradient of the dual function G is given by

$$\nabla G_\mu = Cu_\mu - d = -CA^{-1}C^\top \mu + CA^{-1}b - d,$$

so $(*_7)$ can be written as

$$\lambda^{k+1} = p_+(\lambda^k + \rho\nabla_{\lambda^k});$$

this shows that Uzawa's method is indeed the gradient method with fixed stepsize applied to the dual program.

30.10 Summary

The main concepts and results of this chapter are listed below:

- The cone of feasible directions.
- Cone with apex.
- Active and inactive constraints.
- Qualified constraint at u.
- Farkas lemma.
- Farkas–Minkowski lemma.
- Karush–Kuhn–Tucker optimality conditions (or KKT-conditions).
- Complementary slackness conditions.
- Generalized Lagrange multipliers.
- Qualified convex constraint.
- Lagrangian of a minimization problem.
- Hard margin support vector machine
- Training data
- Linearly separable sets of points.
- Maximal margin hyperplane.
- Support vectors
- Lagrangian duality.
- Saddle points.
- Lagrange dual function.
- Lagrange dual program.
- Duality gap.
- Weak duality.
- Strong Duality.

- Handling equality constaints in the Lagrangian.
- Dual of the Hard margin SVM (SVM$_{h2}$).
- Conjugate functions and Legendre dual functions.
- Dual of the Hard margin SVM (SVM$_{h1}$).
- Ridge regression.
- Kernel ridge regression.
- Kernel function.

Chapter 31

Positive Definite Kernels

31.1 Basic Properties of Positive Definite Kernels

Let X be a nonempty set. If the set X represents a set of highly nonlinear data, it may be advantageous to map X into a space H of much higher dimension called the *feature space*, using a function $\varphi\colon X \to H$ called a *feature map*. This idea is that φ "unwinds" the description of the objects in X, in an attempt to make it linear. The space H is usually a vector space equipped with an inner product $\langle -, - \rangle$. If H is infinite dimensional, then we assume that it is a Hilbert space.

Many algorithms to analyze or classify data make use of the inner products $\langle \varphi(x), \varphi(y) \rangle$, where $x, y \in X$. Thus it is natural to make the following definition.

Definition 31.1. Let X be a nonempty set, let H be a (complex) Hilbert space, and let $\varphi\colon X \to H$ be a function called a *feature map*. The function $\kappa\colon X \times X \to \mathbb{C}$ given by

$$\kappa(x, y) = \langle \varphi(x), \varphi(y) \rangle, \qquad x, y \in X,$$

is called a *kernel function*.

Remark: A *feature map* is often called a *feature embedding*, but this terminology is a bit misleading because it suggests that such a map is injective, which is not necessarily the case. Unfortunately, this terminology is used by most people.

Example 31.1. Suppose we have two feature maps $\varphi_1\colon X \to \mathbb{R}^{n_1}$ and $\varphi_2\colon X \to \mathbb{R}^{n_2}$, and let $\kappa_1(x, y) = \langle \varphi_1(x), \varphi_1(y) \rangle$ and $\kappa_2(x, y) = \langle \varphi_2(x), \varphi_2(y) \rangle$ be the corresponding kernel functions (where $\langle -, - \rangle$ is the standard inner product on $\mathbb{R}^n$). Define the feature map $\varphi\colon X \to \mathbb{R}^{n_1+n_2}$ by

$$\varphi(x) = (\varphi_1(x), \varphi_2(x)),$$

an $(n_1 + n_2)$-tuple. We have

$$\begin{aligned} \langle \varphi(x), \varphi(y) \rangle &= \langle (\varphi_1(x), \varphi_2(x)), (\varphi_1(y), \varphi_2(y)) \rangle = \langle \varphi_1(x), \varphi_1(y) \rangle + \langle \varphi_2(x), \varphi_2(y) \rangle \\ &= \kappa_1(x, y) + \kappa_2(x, y), \end{aligned}$$

which shows that the map κ given by

$$\kappa(x, y) = \kappa_1(x, y) + \kappa_2(x, y)$$

is the kernel function corresponding to the feature map $\varphi\colon X \to \mathbb{R}^{n_1+n_2}$.

Example 31.2. Let X be a subset of $\mathbb{R}^2$, and let $\varphi_1\colon X \to \mathbb{R}^3$ be the map given by

$$\varphi_1(x_1, x_2) = (x_1^2, x_2^2, \sqrt{2}x_1x_2).$$

Observe that linear relations in the feature space $H = \mathbb{R}^3$ correspond to quadratic relations in the input space (of data). We have

$$\begin{aligned}\langle \varphi_1(x), \varphi_1(y)\rangle &= \langle (x_1^2, x_2^2, \sqrt{2}x_1x_2), (y_1^2, y_2^2, \sqrt{2}y_1y_2)\rangle \\ &= x_1^2y_1^2 + x_2^2y_2^2 + 2x_1x_2y_1y_2 \\ &= (x_1y_1 + x_2y_2)^2 = \langle x, y\rangle^2,\end{aligned}$$

where $\langle x, y\rangle$ is the usual inner product on $\mathbb{R}^2$. Hence the function

$$\kappa(x, y) = \langle x, y\rangle^2$$

is a kernel function associated with the feature space $\mathbb{R}^3$.

If we now consider the map $\varphi_2\colon X \to \mathbb{R}^4$ given by

$$\varphi_2(x_1, x_2) = (x_1^2, x_2^2, x_1x_2, x_1x_2),$$

we check immediately that

$$\langle \varphi_2(x), \varphi_2(y)\rangle = \kappa(x, z) = \langle x, y\rangle^2,$$

which shows that the same kernel can arise from different maps into different feature spaces.

Example 31.3. Example 31.2 can be generalized as follows. Suppose we have a feature map $\varphi_1\colon X \to \mathbb{R}^n$ and let $\kappa_1(x, y) = \langle \varphi_1(x), \varphi_1(y)\rangle$ be the corresponding kernel function (where $\langle -, -\rangle$ is the standard inner product on $\mathbb{R}^n$). Define the feature map $\varphi\colon X \to \mathbb{R}^n \times \mathbb{R}^n$ by its n^2 components

$$\varphi(x)_{(i,j)} = (\varphi_1(x))_i(\varphi_1(x))_j, \qquad 1 \leq i, j \leq n,$$

with the inner product on $\mathbb{R}^n \times \mathbb{R}^n$ given by

$$\langle u, v\rangle = \sum_{i,j=1}^{n} u_{(i,j)}v_{(i,j)}.$$

Then we have

$$\begin{aligned}\langle \varphi(x), \varphi(y)\rangle &= \sum_{i,j=1}^{n} \varphi_{(i,j)}(x)\varphi_{(i,j)}(y)\\ &= \sum_{i,j=1}^{n} (\varphi_1(x))_i(\varphi_1(x))_j(\varphi_1(y))_i(\varphi_1(y))_j\\ &= \sum_{i=1}^{n}(\varphi_1(x))_i(\varphi_1(y))_i \sum_{j=1}^{n}(\varphi_1(x))_j(\varphi_1(y))_j\\ &= (\kappa_1(x,y))^2.\end{aligned}$$

Thus the map κ given by $\kappa(x,y) = (\kappa_1(x,y))^2$ is a kernel map associated with the feature map $\varphi\colon X \to \mathbb{R}^n \times \mathbb{R}^n$. The feature map φ is a direct generalization of the feature map φ_2 of Example 31.2.

The above argument is immediately adapted to show that if $\varphi_1\colon X \to \mathbb{R}^{n_1}$ and $\varphi_2\colon X \to \mathbb{R}^{n_2}$ are two feature maps and if $\kappa_1(x,y) = \langle \varphi_1(x), \varphi_1(y)\rangle$ and $\kappa_2(x,y) = \langle \varphi_2(x), \varphi_2(y)\rangle$ are the corresponding kernel functions, then the map defined by

$$\kappa(x,y) = \kappa_1(x,y)\kappa_2(x,y)$$

is a kernel function, for the feature space $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ and the feature map

$$\varphi(x)_{(i,j)} = (\varphi_1(x))_i(\varphi_2(x))_j, \qquad 1 \le i \le n_1,\ 1 \le j \le n_2.$$

Example 31.4. Note that the feature map $\varphi\colon X \to \mathbb{R}^n \times \mathbb{R}^n$ is not very economical because if $i \neq j$ then the components $\varphi_{(i,j)}(x)$ and $\varphi_{(j,i)}(x)$ are both equal to $(\varphi_1(x))_i(\varphi_1(x))_j$. Therefore we can define the more economical embedding $\varphi'\colon X \to \mathbb{R}^{\binom{n+1}{2}}$ given by

$$\varphi'(x)_{(i,j)} = \begin{cases}(\varphi_1(x))_i^2 & i = j,\\ \sqrt{2}(\varphi_1(x))_i(\varphi_1(x))_j & i < j,\end{cases}$$

where the pairs (i,j) with $1 \le i \le j \le n$ are ordered lexicographically. The feature map φ is a direct generalization of the feature map φ_1 of Example 31.2.

Observe that φ' can also be defined in the following way which makes it easier to come up with the generalization to any power:

$$\varphi'_{(i_1,\ldots,i_n)}(x) = \binom{2}{i_1 \cdots i_n}^{1/2} (\varphi_1(x))_1^{i_1}(\varphi_1(x))_1^{i_2}\cdots(\varphi_1(x))_1^{i_n}, \quad i_1 + i_2 + \cdots + i_n = 2,\ i_j \in \mathbb{N},$$

where the n-tuples $(i_1,\ldots,i_n)$ are ordered lexicographically. Recall that for any $m \ge 1$ and any $(i_1,\ldots,i_n) \in \mathbb{N}^m$ such that $i_1 + i_2 + \cdots + i_n = m$, we have

$$\binom{m}{i_1 \cdots i_n} = \frac{m!}{i_1! \cdots i_n!}.$$

More generally, for any $m \geq 2$, using the multinomial theorem, we can define a feature embedding $\varphi\colon X \to \mathbb{R}^{\binom{n+m-1}{m}}$ defining the kernel function κ given by $\kappa(x,y) = (\kappa_1(x,y))^m$, with φ given by

$$\varphi_{(i_1,\ldots,i_n)}(x) = \binom{m}{i_1 \cdots i_n}^{1/2} (\varphi_1(x))_1^{i_1} (\varphi_1(x))_1^{i_2} \cdots (\varphi_1(x))_1^{i_n}, \quad i_1 + i_2 + \cdots + i_n = m,\ i_j \in \mathbb{N},$$

where the n-tuples $(i_1, \ldots, i_n)$ are ordered lexicographically.

Example 31.5. For any positive real constant $R > 0$, the constant function $\kappa(x,y) = R$ is a kernel function corresponding to the feature map $\varphi\colon X \to \mathbb{R}$ given by $\varphi(x,y) = \sqrt{R}$.

By definition, the function $\kappa_1'\colon \mathbb{R}^n \to \mathbb{R}$ given by $\kappa_1'(x,y) = \langle x, y\rangle$ is a kernel function (the feature map is the identity map from $\mathbb{R}^n$ to itself). We just saw that for any positive real constant $R > 0$, the constant $\kappa_2'(x,y) = R$ is a kernel function. By Example 31.1, the function $\kappa_3'(x,y) = \kappa_1'(x,y) + \kappa_2'(x,y)$ is a kernel function, and for any integer $d \geq 1$, by Example 31.3, the function κ_d given by

$$\kappa_d(x,y) = (\kappa_3'(x,y))^d = (\langle x, y\rangle + R)^d,$$

is a kernel function on $\mathbb{R}^n$. By the binomial formula,

$$\kappa_d(x,y) = \sum_{m=0}^{d} R^{d-m} \langle x, y\rangle^m.$$

By Example 31.1, the feature map of this kernel function is the concatenation of the features of the $d+1$ kernel maps $R^{d-m}\langle x, y\rangle^m$. By Example 31.3, the components of the feature map of the kernel map $R^{d-m}\langle x, y\rangle^m$ are reweightings of the functions

$$\varphi_{(i_1,\ldots,i_n)}(x) = x_1^{i_1} x_2^{i_2} \cdots x_n^{i_n}, \quad i_1 + i_2 + \cdots + i_n = m,$$

with $(i_1, \ldots, i_n) \in \mathbb{N}^n$. Thus the components of the feature map of the kernel function κ_d are reweightings of the functions

$$\varphi_{(i_1,\ldots,i_n)}(x) = x_1^{i_1} x_2^{i_2} \cdots x_n^{i_n}, \quad i_1 + i_2 + \cdots + i_n \leq d,$$

with $(i_1, \ldots, i_n) \in \mathbb{N}^n$. It is easy to see that the dimension of this feature space is $\binom{m+d}{d}$.

There are a number of variations of the polynomial kernel κ_d; all-subsets embedding kernels, ANOVA kernels; see Shawe–Taylor and Christianini [96], Chapter III.

In the next example, the set X is not a vector space.

Example 31.6. Let D be a finite set and let $X = 2^D$ be its power set. If $|D| = n$, let $H = \mathbb{R}^X \cong \mathbb{R}^{2^n}$. We are assuming that the subsets of D are enumerated in some

fashion so that each coordinate of $\mathbb{R}^{2^n}$ corresponds to one of these subsets. For example, if $D = \{1, 2, 3, 4\}$, let

$$\begin{array}{llll} U_1 = \emptyset & U_2 = \{1\} & U_3 = \{2\} & U_4 = \{3\} \\ U_5 = \{4\} & U_6 = \{1, 2\} & U_7 = \{1, 3\} & U_8 = \{1, 4\} \\ U_9 = \{2, 3\} & U_{10} = \{2, 4\} & U_{11} = \{3, 4\} & U_{12} = \{1, 2, 3\} \\ U_{13} = \{1, 2, 4\} & U_{14} = \{1, 3, 4\} & U_{15} = \{2, 3, 4\} & U_{16} = \{1, 2, 3, 4\}. \end{array}$$

Let $\varphi\colon X \to H$ be the feature map defined as follows: for any subsets $A, U \in X$,

$$\varphi(A)_U = \begin{cases} 1 & \text{if } U \subseteq A \\ 0 & \text{otherwise.} \end{cases}$$

For example, if $A_1 = \{1, 2, 3\}$, we obtain the vector

$$\varphi(\{1, 2, 3\}) = (1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0),$$

and if $A_2 = \{2, 3, 4\}$, we obtain the vector

$$\varphi(\{2, 3, 4\}) = (1, 0, 1, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0).$$

For any two subsets A_1 and A_2 of D, it is easy to check that

$$\langle \varphi(A_1), \varphi(A_2) \rangle = 2^{|A_1 \cap A_2|},$$

the number of common subsets of A_1 and A_2. For example, $A_1 \cap A_2 = \{2, 3\}$, and

$$\langle \varphi(A_1), \varphi(A_2) \rangle = 4.$$

Therefore, the function $\kappa\colon X \times X \to \mathbb{R}$ given by

$$\kappa(A_1, A_2) = 2^{|A_1 \cap A_2|}, \qquad A_1, A_2 \subseteq D$$

is a kernel function.

Kernel functions have the following important property.

Proposition 31.1. *Let X be any nonempty set, let H be any (complex) Hilbert space, let $\varphi\colon X \to H$ be any function, and let $\kappa\colon X \times X \to \mathbb{C}$ be the kernel given by*

$$\kappa(x, y) = \langle \varphi(x), \varphi(y) \rangle, \qquad x, y \in X.$$

For any finite subset $S = \{x_1, \ldots, x_p\}$ of X, if K_S is the $p \times p$ matrix

$$K_S = (\kappa(x_j, x_i))_{1 \leq i,j \leq p} = (\langle \varphi(x_j), \varphi(x_i) \rangle)_{1 \leq i,j \leq p},$$

then we have

$$u^* K_S\, u \geq 0, \qquad \text{for all } u \in \mathbb{C}^p.$$

Proof. We have

$$\begin{aligned} u^* K_S\, u = u^\top K_S^\top\, \overline{u} &= \sum_{i,j=1}^{p} \kappa(x_i, x_j) u_i \overline{u_j} \\ &= \sum_{i,j=1}^{p} \langle \varphi(x), \varphi(y)\rangle u_i \overline{u_j} \\ &= \left\langle \sum_{i=1}^{p} u_i \varphi(x_i), \sum_{j=1}^{p} u_j \varphi(x_j) \right\rangle = \left\| \sum_{i=1}^{p} u_i \varphi(x_i) \right\|^2 \geq 0, \end{aligned}$$

as claimed. □

Proposition 31.1 suggests a second approach to kernel functions which does not assume that a feature space and a feature map are provided. We will see in Section 31.2 that the two approaches are equivalent. The second approach is useful in practice because it is often difficult to define a feature space and a feature map in a simple manner.

Definition 31.2. Let X be a nonempty set. A function $\kappa\colon X \times X \to \mathbb{C}$ is a *positive definite kernel* if for every finite subset $S = \{x_1, \ldots, x_p\}$ of X, if K_S is the $p \times p$ matrix

$$K_S = (\kappa(x_j, x_i))_{1 \leq i,j \leq p}$$

called a *Gram matrix*, then we have

$$u^* K_S\, u = \sum_{i,j=1}^{p} \kappa(x_i, x_j) u_i \overline{u_j} \geq 0, \qquad \text{for all } u \in \mathbb{C}^p.$$

Observe that Definition 31.2 does not require that $u^* K_S\, u > 0$ if $u \neq 0$, so the terminology *positive definite* is a bit abusive, and it would be more appropriate to use the terminology *positive semidefinite*. However, it seems customary to use the term *positive definite kernel*, or even *positive kernel*.

Proposition 31.2. *Let $\kappa\colon X \times X \to \mathbb{C}$ be a positive definite kernel. Then $\kappa(x, x) \geq 0$ for all $x \in X$, and for any finite subset $S = \{x_1, \ldots, x_p\}$ of X, the $p \times p$ matrix K_S given by*

$$K_S = (\kappa(x_j, x_i))_{1 \leq i,j \leq p}$$

is hermitian, that is, $K_S^* = K_S$.

Proof. The first property is obvious by choosing $S = \{x\}$. We have

$$(u + v)^* K_S (u + v) = u^* K_S u + u^* K_S v + v^* K_S u + v^* K_S v,$$

and since $(u+v)^* K_S(u+v), u^* K_S u, v^* K_S v \geq 0$, we deduce that

$$2A = u^* K_S v + v^* K_S u \tag{1}$$

must be real. By replacing u by iu, we see that

$$2B = -iu^* K_s v + iv^* K_S u \tag{2}$$

must also be real, By multiplying Equation (2) by i and adding it to Equation (1) we get

$$u^* K_S v = A + iB. \tag{3}$$

By subtracting Equation (3) from Equation (1) we get

$$v^* K_S u = A - iB.$$

Then

$$u^* K_S^* v = \overline{v^* K_S u} = \overline{A - iB} = A + iB = u^* K_S v,$$

for all $u, v \in \mathbb{C}^*$, which implies $K_S^* = K_S$. □

If the map $\kappa\colon X \times X \to \mathbb{R}$ is real-valued, then we have the following criterion for κ to be a positive definite kernel that only involves real vectors.

Proposition 31.3. *If $\kappa\colon X \times X \to \mathbb{R}$, then κ is a positive definite kernel iff for any finite subset $S = \{x_1, \ldots, x_p\}$ of X, the $p \times p$ real matrix K_S given by*

$$K_S = (\kappa(x_k, x_j))_{1 \leq j,k \leq p}$$

is symmetric, that is, $K_S^\top = K_S$, and

$$u^\top K_S\, u = \sum_{j,k=1}^{p} \kappa(x_j, x_k) u_j u_k \geq 0, \qquad \textit{for all } u \in \mathbb{R}^p.$$

Proof. If κ is a real-valued positive definite kernel, then the proposition is a trivial consequence of Proposition 31.2.

For the converse, assume that κ is symmetric and that it satisfies the second condition of the proposition. We need to show that κ is a positive definite kernel with respect to complex vectors. If we write $u_k = a_k + ib_k$, then

$$\begin{aligned} u^* K_S\, u &= \sum_{j,k=1}^{p} \kappa(x_j, x_k)(a_j + ib_j)(a_k - ib_k) \\ &= \sum_{j,k=1}^{p} (a_j a_k + b_j b_k)\kappa(x_j, x_k) + i \sum_{j,k=1}^{p} (b_j a_k - a_j b_k)\kappa(x_j, x_k) \\ &= \sum_{j,k=1}^{p} (a_j a_k + b_j b_k)\kappa(x_j, x_k) + i \sum_{1 \leq j < k \leq p} b_j a_k (\kappa(x_j, x_k) - \kappa(x_k, x_j)). \end{aligned}$$

Thus $u^* K_S u$ is real iff K_S is symmetric. □

Consequently we make the following definition.

Definition 31.3. Let X be a nonempty set. A function $\kappa\colon X \times X \to \mathbb{R}$ is a *(real) positive definite kernel* if $\kappa(x, y) = \kappa(y, x)$ for all $x, y \in X$, and for every finite subset $S = \{x_1, \ldots, x_p\}$ of X, if K_S is the $p \times p$ real symmetric matrix

$$K_S = (\kappa(x_i, x_j))_{1 \leq i,j \leq p},$$

then we have

$$u^\top K_S\, u = \sum_{i,j=1}^{p} \kappa(x_i, x_j) u_i u_j \geq 0, \qquad \text{for all } u \in \mathbb{R}^p.$$

Among other things, the next proposition shows that a positive definite kernel satisfies the Cauchy–Schwarz inequality.

Proposition 31.4. *A hermitian 2×2 matrix*

$$A = \begin{pmatrix} a & \overline{b} \\ b & d \end{pmatrix}$$

is positive semidefinite if and only if $a \geq 0$, $d \geq 0$, and $ad - |b|^2 \geq 0$.

Let $\kappa\colon X \times X \to \mathbb{C}$ be a positive definite kernel. For all $x, y \in X$, we have

$$|\kappa(x, y)|^2 \leq \kappa(x, x)\kappa(y, y).$$

Proof. For all $x, y \in \mathbb{C}$, we have

$$\begin{aligned} \begin{pmatrix} \overline{x} & \overline{y} \end{pmatrix} \begin{pmatrix} a & \overline{b} \\ b & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} &= \begin{pmatrix} \overline{x} & \overline{y} \end{pmatrix} \begin{pmatrix} ax + \overline{b}y \\ bx + dy \end{pmatrix} \\ &= a|x|^2 + bx\overline{y} + \overline{bx\overline{y}} + d|y|^2. \end{aligned}$$

If A is positive semidefinite, then we already know that $a \geq 0$ and $d \geq 0$. If $a = 0$, then we must have $b = 0$, since otherwise we can make $bx\overline{y} + \overline{bx\overline{y}}$, which is twice the real part of $bx\overline{y}$, as negative as we want. In this case, $ad - |b|^2 = 0$.

If $a > 0$, then

$$a|x|^2 + bx\overline{y} + \overline{bx\overline{y}} + d|y|^2 = a\left|x + \frac{\overline{b}}{a}y\right|^2 + \frac{|y|^2}{a}(ad - |b|^2).$$

If $ad - |b|^2 < 0$, we can pick $y \neq 0$ and $x = -(\overline{b}y)/a$, so that the above expression is negative. Therefore, $ad - |b|^2 \geq 0$. The converse is trivial.

If $x = y$, the inequality $|\kappa(x, y)|^2 \leq \kappa(x, x)\kappa(y, y)$ is trivial. If $x \neq y$, the inequality follows by applying the criterion for being positive semidefinite to the matrix

$$\begin{pmatrix} \kappa(x, x) & \overline{\kappa(x, y)} \\ \kappa(x, y) & \kappa(y, y) \end{pmatrix},$$

as claimed. □

The following property due to I. Schur (1911) shows that the pointwise product of two positive definite kernels is also a positive definite kernel.

Proposition 31.5. *(I. Schur) If $\kappa_1 \colon X \times X \to \mathbb{C}$ and $\kappa_2 \colon X \times X \to \mathbb{C}$ are two positive definite kernels, then the function $\kappa \colon X \times X \to \mathbb{C}$ given by $\kappa(x, y) = \kappa_1(x, y)\kappa_2(x, y)$ for all $x, y \in X$ is also a positive definite kernel.*

Proof. It suffices to prove that if $A = (a_{jk})$ and $B = (b_{jk})$ are two hermitian positive semidefinite $p \times p$ matrices, then so is their pointwise product $C = A \circ B = (a_{jk}b_{jk})$ (also known as Hadamard or Schur product). Recall that a hermitian positive semidefinite matrix A can be diagonalized as $A = U\Lambda U^*$, where Λ is a diagonal matrix with nonnegative entries and U is a unitary matrix. Let $\Lambda^{1/2}$ be the diagonal matrix consisting of the positive square roots of the diagonal entries in Λ. Then we have

$$A = U\Lambda U^* = U\Lambda^{1/2}\Lambda^{1/2}U^* = U\Lambda^{1/2}(U\Lambda^{1/2})^*.$$

Thus if we set $R = U\Lambda^{1/2}$, we have

$$A = RR^*,$$

which means that

$$a_{jk} = \sum_{h=1}^{p} r_{jh}\overline{r_{kh}}.$$

Then for any $u \in \mathbb{C}^p$, we have

$$\begin{aligned} u^*(A \circ B)u &= \sum_{j,k=1}^{p} a_{jk}b_{jk}u_j\overline{u_k} \\ &= \sum_{j,k=1}^{p}\sum_{h=1}^{p} r_{jh}\overline{r_{kh}}b_{jk}u_j\overline{u_k} \\ &= \sum_{h=1}^{p}\sum_{j,k=1}^{p} b_{jk}u_j r_{jh}\overline{u_k r_{kh}}. \end{aligned}$$

Since B is positive semidefinite, for each fixed h, we have

$$\sum_{j,k=1}^{p} b_{jk}u_j r_{jh}\overline{u_k r_{kh}} = \sum_{j,k=1}^{p} b_{jk}z_j\overline{z_k} \geq 0,$$

as we see by letting $z = (u_1 r_{1h}, \ldots, u_p r_{ph})$, □

In contrast, the ordinary product AB of two symmetric positive semidefinite matrices A and B may not be symmetric positive semidefinite; see Section 5.8 for an example.

Here are other ways of obtaining new positive definite kernels from old ones.

Proposition 31.6. *Let $\kappa_1 \colon X \times X \to \mathbb{C}$ and $\kappa_2 \colon X \times X \to \mathbb{C}$ be two positive definite kernels, $f \colon X \to \mathbb{C}$ be a function, $\psi \colon X \to \mathbb{R}^N$ be a function, $\kappa_3 \colon \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{C}$ be a positive definite kernel, and $a \in \mathbb{R}$ be any positive real. Then the following functions are positive definite kernels:*

(1) $\kappa(x, y) = \kappa_1(x, y) + \kappa_2(x, y)$.

(2) $\kappa(x, y) = a\kappa_1(x, y)$.

(3) $\kappa(x, y) = f(x)\overline{f(y)}$.

(4) $\kappa(x, y) = \kappa_3(\psi(x), \psi(y))$.

(5) If B is a symmetric positive semidefinite $n \times n$ matrix, then the map $\kappa \colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ given by

$$\kappa(x, y) = x^\top B y$$

is a positive definite kernel.

Proof. (1) For every finite subset $S = \{x_1, \ldots, x_p\}$ of X, if K_1 is the $p \times p$ matrix

$$K_1 = (\kappa_1(x_k, x_j))_{1 \leq j,k \leq p}$$

and if if K_2 is the $p \times p$ matrix

$$K_2 = (\kappa_2(x_k, x_j))_{1 \leq j,k \leq p},$$

then for any $u \in \mathbb{C}^p$, we have

$$u^*(K_1 + K_2)u = u^*K_1u + u^*K_2u \geq 0,$$

since $u^*K_1u \geq 0$ and $u^*K_2u \geq 0$ because κ_2 and κ_2 are positive definite kernels, which means that K_1 and K_2 are positive semidefinite.

(2) We have

$$u^*(aK_1)u = au^*K_1u \geq 0,$$

since $a > 0$ and $u^*K_1u \geq 0$.

(3) For every finite subset $S = \{x_1, \ldots, x_p\}$ of X, if K is the $p \times p$ matrix

$$K = (\kappa(x_k, x_j))_{1 \leq j,k \leq p} = (\overline{f(x_k)}f(x_j))_{1 \leq j,k \leq p}$$

then we have

$$u^*Ku = \sum_{j,k=1}^{p} \kappa(x_j, x_k)u_j\overline{u_k} = \sum_{j,k=1}^{p} u_j f(x_j)\overline{u_k f(x_k)} = \left| \sum_{j=1}^{p} u_j f(x_j) \right|^2 \geq 0.$$

(4) For every finite subset $S = \{x_1, \ldots, x_p\}$ of X, the $p \times p$ matrix K given by

$$K = (\kappa(x_k, x_j))_{1 \leq j,k \leq p} = (\kappa_3(\psi(x_k), \psi(x_j)))_{1 \leq j,k \leq p}$$

is symmetric positive semidefinite since κ_3 is a positive definite kernel.

(5) As in the proof of Proposition 31.5 (adapted to the real case) there is a matrix R such that

$$B = RR^\top,$$

so

$$\kappa(x, y) = x^\top By = x^\top RR^\top y = (R^\top x)^\top R^\top y = \langle R^\top x, R^\top y \rangle,$$

so κ is the kernel function given by the feature map $\varphi(x) = R^\top x$ from $\mathbb{R}^n$ to itself, and by Proposition 31.1, it is a symmetric positive definite kernel. □

Proposition 31.7. *Let $\kappa_1 \colon X \times X \to \mathbb{C}$ be a positive definite kernel, and let $p(z)$ be a polynomial with nonnegative coefficients. Then the following functions κ defined below are also positive definite kernels.*

(1) $\kappa(x, y) = p(\kappa_1(x, y))$.

(2) $\kappa(x, y) = e^{\kappa_1(x,y)}$.

(3) If X is real Hilbert space with inner product $\langle -, - \rangle_X$ and corresponding norm $\| \ \|_X$,

$$\kappa(x, y) = e^{-\frac{\|x-y\|_X^2}{2\sigma^2}}$$

for any $\sigma > 0$.

Proof. (1) If $p(z) = a_m z^m + \cdots + a_1 z + a_0$, then

$$p(\kappa_1(x, y)) = a_m \kappa_1(x, y)^m + \cdots + a_1 \kappa_1(x, y) + a_0.$$

Since $a_k \geq 0$ for $k = 0, \ldots, m$, by Proposition 31.5 and Proposition 31.6(2), each function $a_k \kappa_i(x, y)^k$ with $1 \leq k \leq m$ is a positive definite kernel, by Proposition 31.6(3) with $f(x) = \sqrt{a_0}$, the constant function a_0 is a positive definite kernel, and by Proposition 31.6(1), $p(\kappa_1(x, y))$ is a positive definite kernel.

(2) We have

$$e^{\kappa_1(x,y)} = \sum_{k=0}^{\infty} \frac{\kappa_1(x, y)^k}{k!}.$$

By (1), the partial sums

$$\sum_{k=0}^{m} \frac{\kappa_1(x, y)^k}{k!}$$

are positive definite kernels, and since $e^{\kappa_1(x,y)}$ is the (uniform) pointwise limit of positive definite kernels, it is also a positive definite kernel.

(3) By Proposition 31.6(2), since the map $(x, y) \mapsto \langle x, y\rangle_X$ is obviously a positive definite kernel (the feature map is the identity) and since $\sigma \neq 0$, the function $(x, y) \mapsto \langle x, y\rangle_X/\sigma^2$ is a positive definite kernel, so by (2),

$$\kappa_1(x, y) = e^{\frac{\langle x,y\rangle_X}{\sigma^2}}$$

is a positive definite kernel. Let $f\colon X \to \mathbb{R}$ be the function given by

$$f(x) = e^{-\frac{\|x\|^2}{2\sigma^2}}.$$

Then by Proposition 31.6(3),

$$\kappa_2(x, y) = f(x)f(y) = e^{-\frac{\|x\|^2}{2\sigma^2}} e^{-\frac{\|y\|^2}{2\sigma^2}} = e^{-\frac{\|x\|_X^2+\|y\|_X^2}{2\sigma^2}}$$

is a positive definite kernel. By Proposition 31.5, the function $\kappa_1\kappa_2$ is a positive definite kernel, that is

$$\kappa_1(x, y)\kappa_2(x, y) = e^{\frac{\langle x,y\rangle_X}{\sigma^2}} e^{-\frac{\|x\|_X^2+\|y\|_X^2}{2\sigma^2}} = e^{\frac{\langle x,y\rangle_X}{\sigma^2} - \frac{\|x\|_X^2+\|y\|_X^2}{2\sigma^2}} = e^{-\frac{\|x-y\|_X^2}{2\sigma^2}}$$

is a positive definite kernel. □

The positive definite kernel

$$\kappa(x, y) = e^{-\frac{\|x-y\|_X^2}{2\sigma^2}}$$

is called a *Gaussian kernel*. This kernel requires a feature map in an infinite-dimensional space because it is an infinite sum of distinct kernels.

Remark: If κ_1 is a positive definite kernel, the proof of Proposition 31.7(3) is immediately adapted to show that

$$\kappa(x, y) = e^{-\frac{\kappa_1(x,x)+\kappa_1(y,y)-2\kappa_1(x,y)}{2\sigma^2}}$$

is a positive definite kernel.

Next we prove that every positive definite kernel arises from a feature map in a Hilbert space which is a function space.

31.2 Hilbert Space Representation of a Positive Definite Kernel

The following result shows how to construct a so-called *reproducing kernel Hilbert space*, for short RKHS, from a positive definite kernel.

Theorem 31.8. *Let $\kappa\colon X \times X \to \mathbb{C}$ be a positive definite kernel on a nonempty set X. For every $x \in X$, let $\kappa_x\colon X \to \mathbb{C}$ be the function given by*

$$\kappa_x(y) = \kappa(x, y), \qquad y \in X.$$

Let H_0 be the subspace of the vector space $\mathbb{C}^X$ of functions from X to $\mathbb{C}$ spanned by the family of functions $(\kappa_x)_{\in X}$, and let $\varphi\colon X \to H_0$ be the map given by $\varphi(x) = \kappa_x$. There is a hermitian inner product $\langle -, - \rangle$ on H_0 such that

$$\kappa(x, y) = \langle \varphi(x), \varphi(y) \rangle, \qquad \textit{for all } x, y \in X.$$

The completion H of H_0 is a Hilbert space, and the map $\eta\colon H \to \mathbb{C}^X$ given by

$$\eta(f)(x) = \langle f, \kappa_x \rangle, \qquad x \in X,$$

is linear and injective, so H can be identified with a subspace of $\mathbb{C}^X$. We also have

$$\kappa(x, y) = \langle \varphi(x), \varphi(y) \rangle, \qquad \textit{for all } x, y \in X.$$

For all $f \in H_0$ and all $x \in X$,

$$\langle f, \kappa_x \rangle = f(x),$$

a property known as the **reproducing property**.

Proof. For any two linear combinations $f = \sum_{j=1}^{p} \alpha_j \kappa_{x_j}$ and $g = \sum_{k=1}^{q} \beta_k \kappa_{y_k}$ in H_0, with $x_j, y_k \in X$ and $\alpha_j, \beta_k \in \mathbb{C}$, define $\langle f, g \rangle$ by

$$\langle f, g \rangle = \sum_{j=1}^{p} \sum_{k=1}^{q} \alpha_j \overline{\beta_k} \kappa(x_j, y_k). \tag{$\dagger$}$$

At first glance, the above expression appears to depend on the expression of f and g as linear combinations, but since $\kappa(x_j, y_k) = \overline{\kappa(y_k, x_j)}$, observe that

$$\sum_{k=1}^{q} \overline{\beta_k} f(y_k) = \sum_{j=1}^{p} \sum_{k=1}^{q} \alpha_j \overline{\beta_k} \kappa(x_j, y_k) = \sum_{j=1}^{p} \alpha_j \overline{g(x_j)}, \tag{$*$}$$

and since the first and the third term are equal for all linear combinations representing f and g, we conclude that ($\dagger$) depends only on f and g and not on their representation as a linear combination.

Obviously ($\dagger$) defines a hermitian sequilinear form. For every $f \in H_0$, we have

$$\langle f, f \rangle = \sum_{j,k=1}^{p} \alpha_j \overline{\alpha_k} \kappa(x_j, x_k) \geq 0,$$

since κ is a positive definite kernel. For any finite subset $\{f_1, \ldots, f_n\}$ of H_0 and any $z \in \mathbb{C}^n$, we have

$$\sum_{j,k=1}^{n} \langle f_j, f_k \rangle z_j \overline{z_k} = \left\langle \sum_{j=1}^{n} z_j f_j, \sum_{j=1}^{n} z_j f_j \right\rangle \geq 0,$$

which shows that the map $(f, g) \mapsto \langle f, g \rangle$ from $H_0 \times H_0$ to $\mathbb{C}$ is a positive definite kernel.

Observe that for all $f \in H_0$ and all $x \in X$, (†) implies that

$$\langle f, \kappa_x \rangle = \sum_{j=1}^{k} \alpha_j \kappa(x_j, x) = f(x),$$

a property known as the *reproducing property*. The above implies that

$$\langle \kappa_x, \kappa_y \rangle = \kappa(x, y). \tag{$**$}$$

By Proposition 31.4 applied to the positive definite kernel $(f, g) \mapsto \langle f, g \rangle$, we have

$$|\langle f, \kappa_x \rangle|^2 \leq \langle f, f \rangle \langle \kappa_x, \kappa_x \rangle,$$

that is,

$$|f(x)|^2 \leq \langle f, f \rangle \kappa(x, x),$$

so $\langle f, f \rangle = 0$ implies that $f(x) = 0$ for all $x \in X$, which means that $\langle -, - \rangle$ as defined by (†) is positive definite. Therefore, $\langle -, - \rangle$ is a hermitian inner product on H_0, and by ($**$) and since $\varphi(x) = \kappa_x$, we have

$$\kappa(x, y) = \langle \varphi(x), \varphi(y) \rangle, \qquad \text{for all } x, y \in X.$$

Let H be the Hilbert space which is the completion of H_0, so that H_0 is dense in H. The map $\eta \colon H \to \mathbb{C}^X$ given by

$$\eta(f)(x) = \langle f, \kappa_x \rangle$$

is obviously linear, and it is injective because the family $(\kappa_x)_{x \in X}$ spans H_0 which is dense in H, thus it is also dense in H, so if $\langle f, \kappa_x \rangle = 0$ for all $x \in X$, then $f = 0$. □

If we identify a function $f \in H$ with the function $\eta(f)$, then we have the reproducing property

$$\langle f, \kappa_x \rangle = f(x), \qquad \text{for all } f \in H \text{ and all } x \in X.$$

If X is finite, then $\mathbb{C}^X$ is finite-dimensional. If X is a separable topological space and if κ is continuous, then it can be shown that H is a separable Hilbert space.

Also, if $\kappa \colon X \times X \to \mathbb{R}$ is a real symmetric positive definite kernel, then we see immediately that Theorem 31.8 holds with H_0 a real Euclidean space and H a real Hilbert space.

Remark: If $X = G$, where G is a locally compact group, then a function $p\colon G \to \mathbb{C}$ (not necessarily continuous) is *positive semidefinite* if for all $s_1, \ldots, s_n \in G$ and all $\xi_1, \ldots, \xi_n \in \mathbb{C}$, we have

$$\sum_{j,k=1}^{n} p(s_j^{-1}s_k)\xi_k\overline{\xi_j} \geq 0.$$

So if we define $\kappa\colon G \times G \to \mathbb{C}$ by

$$\kappa(s,t) = p(t^{-1}s),$$

then κ is a positive definite kernel on G. If p is continuous, then it is known that p arises from a unitary representation $U\colon G \to \mathbf{U}(H)$ of the group G in a Hilbert space H with inner product $\langle -, - \rangle$ (a homomorphism with a certain continuity property), in the sense that there is some vector $x_0 \in H$ such that

$$p(s) = \langle U(s)(x_0), x_0\rangle, \qquad \text{for all } s \in G.$$

Since the $U(s)$ are unitary operators on H,

$$\begin{aligned} p(t^{-1}s) &= \langle U(t^{-1}s)(x_0), x_0\rangle = \langle U(t^{-1})(U(s)(x_0)), x_0\rangle \\ &= \langle U(t)^*(U(s)(x_0)), x_0\rangle = \langle U(s)(x_0)), U(t)(x_0)\rangle, \end{aligned}$$

which shows that

$$\kappa(s,t) = \langle U(s)(x_0)), U(t)(x_0)\rangle,$$

so the map $\varphi\colon G \to H$ given by

$$\varphi(s) = U(s)(x_0)$$

is a feature map into the feature space H. This theorem is due to Gelfand and Raikov (1943).

The proof of Theorem 31.8 is essentially identical to part of Godement's proof of the above result about the correspondence between functions of positive type and unitary representations; see Helgason [54], Chapter IV, Theorem 1.5. Theorem 31.8 is a little more general since it does not assume that X is a group, but when G is a group, the feature map arises from a unitary representation.

Kernels on collections of sets can be defined in terms of measures.

Example 31.7. Let $(D, \mathcal{A})$ be a measurable space, where D is a nonempty set and $\mathcal{A}$ is a σ-algebra on D (the measurable sets). Let X be a subset of $\mathcal{A}$. If μ is a positive measure on $(D, \mathcal{A})$ and if μ is finite, which means that $\mu(D)$ is finite, then we can define the map $\kappa_1\colon X \times X \to \mathbb{R}$ given by

$$\kappa_1(A_1, A_2) = \mu(A_1 \cap A_2), \qquad A_1, A_2 \in X.$$

We can show that κ is a kernel function as follows. Let $H = \mathrm{L}^2_\mu(D, \mathcal{A}, \mathbb{R})$ be the Hilbert space of μ-square-integrable functions, with the inner product

$$\langle f, g\rangle = \int_D f(s)g(s)\, d\mu(s),$$

and let $\varphi\colon X \to H$ be the feature embedding given by

$$\varphi(A) = \chi_A, \qquad A \in X,$$

the characteristic function of A. Then we have

$$\begin{aligned}\kappa_1(A_1, A_2) = \mu(A_1 \cap A_2) &= \int_D \chi_{A_1 \cap A_2}(s)\, d\mu(s) \\ &= \int_D \chi_{A_1}(s)\chi_{A_2}(s)\, d\mu(s) = \langle \chi_{A_1}, \chi_{A_2} \rangle \\ &= \langle \varphi(A_1), \varphi(A_2) \rangle.\end{aligned}$$

The above kernel is called the *intersection kernel*. If we assume that μ is normalized so that $\mu(D) = 1$, then we also have the *union complement kernel*:

$$\kappa_2(A_1, A_2) = \mu(\overline{A_1} \cap \overline{A_2}) = 1 - \mu(A_1 \cup A_2).$$

The sum κ_3 of the kernels κ_1 and κ_2 is the *agreement kernel*:

$$\kappa_s(A_1, A_2) = 1 - \mu(A_1 - A_2) - \mu(A_2 - A_1).$$

Many other kinds of kernels can be designed, in particular, graph kernels. For comprehensive presentations of kernels, see Schölkopf and Smola [85] and Shawe–Taylor and Christianini [96].

31.3 Kernel PCA

As an application of kernel functions, we discuss a generalization of the method of principal component analysis (PCA). Suppose we have a set of data $S = \{x_1, \ldots, x_n\}$ in some input space $\mathcal{X}$, and pretend that we have an embedding $\varphi\colon \mathcal{X} \to F$ of $\mathcal{X}$ in a (real) feature space $(F, \langle -, - \rangle)$, but that we only have access to the kernel function $\kappa(x, y) = \langle \varphi(x), \varphi(y) \rangle$. We would like to do PCA analysis on the set $\varphi(S) = \{\varphi(x_1), \ldots, \varphi(x_n)\}$.

There are two obstacles:

(1) We need to center the data and compute the inner products of pairs of centered data. More precisely, if the centroid of $\varphi(S)$ is

$$\mu = \frac{1}{n}(\varphi(x_1) + \cdots + \varphi(x_n)),$$

then we need to compute the inner products $\langle \varphi(x) - \mu, \varphi(y) - \mu \rangle$.

(2) If we assume that $F = \mathbb{R}^d$ and that the data points $\varphi(x_i)$ are expressed as *row vectors* X_i of an $n \times d$ matrix X (as it is cutomary), then the inner products $\kappa(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle$ are given by the *kernel matrix* $\mathbf{K} = XX^\top$. Be aware that with this representation, $\varphi(x_i)$ is a d-dimensional column vector and that $\varphi(x_i) = X_i^\top$. However, the jth component $(Y_k)_j$ of the principal component Y_k (viewed as a n-dimensional column vector) is given by the projection of $\widehat{X}_j = X_j - \mu$ onto the direction u_k (viewing μ as a d-dimensional row vector), which is a unit eigenvector of the matrix $(X-\mu)^\top(X-\mu)$ (where $\widehat{X} = X - \mu$ is the matrix whose jth row is $\widehat{X}_j = X_j - \mu$), is given by the inner product

$$\langle X_j - \mu, u_k \rangle = (Y_k)_j;$$

see Definition 16.2 and Theorem 16.11. The problem is that we know what the matrix $(X - \mu)(X - \mu)^\top$ is from (1), because it can be expressed in terms of $\mathbf{K}$, but we don't know what $(X - \mu)^\top(X - \mu)$ is, because we don't have access to $\widehat{X} = X - \mu$.

Both difficulties are easily overcome. For (1), we have

$$\begin{aligned}\langle \varphi(x) - \mu, \varphi(y) - \mu \rangle &= \left\langle \varphi(x) - \frac{1}{n}\sum_{k=1}^{n} \varphi(x_k), \varphi(y) - \frac{1}{n}\sum_{k=1}^{n} \varphi(x_k) \right\rangle \\ &= \kappa(x, y) - \frac{1}{n}\sum_{i=1}^{n} \kappa(x, x_i) - \frac{1}{n}\sum_{j=1}^{n} \kappa(x_j, y) + \frac{1}{n^2}\sum_{i,j=1}^{n} \kappa(x_i, x_j).\end{aligned}$$

For (2), if $\mathbf{K}$ is the kernel matrix $\mathbf{K} = (\kappa(x_i, x_j))$, then the kernel matrix $\widehat{\mathbf{K}}$ corresponding to the kernel function $\widehat{\kappa}$ given by

$$\widehat{\kappa}(x, y) = \langle \varphi(x) - \mu, \varphi(y) - \mu \rangle$$

can be expressed in terms of $\mathbf{K}$. Let $\mathbf{1}$ be the column vector (of dimension n) whose entries are all 1. Then $\mathbf{1}\mathbf{1}^\top$ is the $n \times n$ matrix whose entries are all 1. If A is an $n \times n$ matrix, then $\mathbf{1}^\top A$ is the row vector consisting of the sums of the columns of A, $A\mathbf{1}$ is the column vector consisting of the sums of the rows of A, and $\mathbf{1}^\top A \mathbf{1}$ is the sum of all the entries in A. Then it is easy to see that the kernel matrix corresponding to the kernel function $\widehat{\kappa}$ is given by

$$\widehat{\mathbf{K}} = \mathbf{K} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\mathbf{K} - \frac{1}{n}\mathbf{K}\mathbf{1}\mathbf{1}^\top + \frac{1}{n^2}(\mathbf{1}^\top\mathbf{K}\mathbf{1})\mathbf{1}\mathbf{1}^\top.$$

Suppose $\widehat{X} = X - \mu$ has rank r. To overcome the second problem, note that if

$$\widehat{X} = VDU^\top$$

is an SVD for $\widehat{X}$, then

$$\widehat{X}^\top = UD^\top V^\top$$

is an SVD for $\widehat{X}^\top$, and the $r \times r$ submatrix of $D^\top$ consisting of the first r rows and r columns of $D^\top$ (and D), is the diagonal Σ^r matrix consisting of the singular values $\sigma_1 \geq \cdots \geq \sigma_r$ of

$\widehat{X}$, so we can express the matrix U_r consisting of the first r columns u_k of U in terms of the matrix V_r consisting of the first r columns v_k of V $(1 \leq k \leq r)$ as

$$U_r = \widehat{X}^\top V_r \Sigma_r^{-1}.$$

Furthermore, $\sigma_1^2 \geq \cdots \geq \sigma_r^2$ are the nonzero eigenvalues of $\widehat{\mathbf{K}} = \widehat{X}\widehat{X}^\top$, and the columns of V_r are corresponding unit eigenvectors of $\widehat{\mathbf{K}}$. From

$$U_r = \widehat{X}^\top V_r \Sigma_r^{-1}$$

the kth column u_k of U_r (which is a unit eigenvector of $\widehat{X}^\top \widehat{X}$ associated with the eigenvalue σ_k^2) is given by

$$u_k = \sum_{i=1}^n \sigma_k^{-1}(v_k)_i \widehat{X}_i^\top = \sum_{i=1}^n \sigma_k^{-1}(v_k)_i \widehat{\varphi(x_i)}, \quad 1 \leq k \leq r,$$

so the projection of $\widehat{\varphi(x)}$ onto u_k is given by

$$\begin{aligned}\langle \widehat{\varphi(x)}, u_k \rangle &= \left\langle \widehat{\varphi(x)}, \sum_{i=1}^n \sigma_k^{-1}(v_k)_i \widehat{\varphi(x_i)} \right\rangle \\ &= \sum_{i=1}^n \sigma_k^{-1}(v_k)_i \left\langle \widehat{\varphi(x)}, \widehat{\varphi(x_i)} \right\rangle = \sum_{i=1}^n \sigma_k^{-1}(v_k)_i \widehat{\kappa}(x, x_i).\end{aligned}$$

Therefore, the jth component of the principal component Y_k in the principal direction u_k is given by

$$(Y_k)_j = \langle X_j - \mu, u_k \rangle = \sum_{i=1}^n \sigma_k^{-1}(v_k)_i \widehat{\kappa}(x_j, x_i) = \sum_{i=1}^n \sigma_k^{-1}(v_k)_i \widehat{\mathbf{K}}_{ij}.$$

Since $\sigma_1^2 \geq \cdots \geq \sigma_r^2$ are the nonzero eigenvalues of $\widehat{\mathbf{K}}$ and $v_1, \ldots, v_r$ are corresponding unit eigenvectors, the above expression is completely determined in terms of the kernel matrix $\widehat{\mathbf{K}}$. Sometimes the notation

$$\alpha_k = \sigma_k^{-1} v_k$$

is used, where the α_k are called the *dual variables*, in which case,

$$(Y_k)_j = \sum_{i=1}^n (\alpha_k)_i \widehat{\mathbf{K}}_{ij}.$$

The column vector Y_k $(1 \leq k \leq r)$ defined by

$$Y_k = \left(\sum_{i=1}^n (\alpha_k)_i \widehat{\mathbf{K}}_{ij} \right)_{j=1}^n$$

is called the *kth kernel principal component* (for short *kth kernel PCA*) of the data set $S = \{x_1, \ldots, x_n\}$ in the direction $u_k = \sum_{i=1}^n \sigma_k^{-1}(v_k)_i \widehat{X}_i^\top$ (even though the matrix $\widehat{X}$ is not known).

In the next section, we give another illustration of the use of kernel functions in a generalization of ridge regression (see Example 30.16).

31.4 ν-SV Regression

Let $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ be a set of observed data usually called a set of *training data*, with $x_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$. Our goal is to learn an affine function f of the form $f(x) = w^\top x - b$ that fits the set of training data, but does not penalize errors below some given $\epsilon \geq 0$. Thus we try to fit a tube with radius ϵ to the data, but we also allow *errors*, in the sense that some data x_i may satisfy the equality $f(x_i) - y_i = \epsilon + \xi_i$ for some $\xi_i > 0$, or the equality $-(f(x_i) - y_i) = \epsilon + \xi_i'$ for some $\xi_i' > 0$. In this case, x_i lies outside of the tube with radius ϵ. The trade off between the size of ϵ and the size of the slack variables ξ_i and ξ_i' is achieved by using two constants $\nu \geq 0$ and $C > 0$. The method of *ν-support vector regression*, for short *ν-SV regression*, is specified by the following minimization problem:

ν-SV Regression:

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2} w^\top w + C\left(\nu\epsilon + \frac{1}{m}\sum_{i=1}^{m}(\xi_i + \xi_i')\right) \\
&\text{subject to} \\
&\qquad w^\top x_i - b - y_i \leq \epsilon + \xi_i, \quad \xi_i \geq 0 \qquad i = 1, \ldots, m \\
&\qquad -w^\top x_i + b + y_i \leq \epsilon + \xi_i', \quad \xi_i' \geq 0 \qquad i = 1, \ldots, m \\
&\qquad \epsilon \geq 0,
\end{aligned}$$

minimizing over the variables w, b, ϵ, ξ, and ξ'. The constraints are affine.

First, observe that the equations

$$\begin{aligned}
w^\top x_i - b - y_i &= \epsilon + \xi_i \\
-w^\top x_i + b + y_i &= \epsilon + \xi_i'
\end{aligned}$$

can only hold simultaneously if

$$\epsilon + \xi_i = -\epsilon - \xi',$$

that is,

$$2\epsilon + \xi_i + \xi_i' = 0,$$

and since $\epsilon, \xi_i, \xi_i' \geq 0$, this can happen only if $\epsilon = \xi_i = \xi_i' = 0$, and then

$$w^\top x_i - b = y_i.$$

In particular, if $\epsilon > 0$, then the equations

$$\begin{aligned}
w^\top x_i - b - y_i &= \epsilon + \xi_i \\
-w^\top x_i + b + y_i &= \epsilon + \xi_i'
\end{aligned}$$

cannot hold simultaneously. Also, since $-w^\top x_i + b + y_i = -(w^\top x_i - b - y_i)$, for an optimal solution, if $w^\top x_i - b - y_i \geq 0$, then $\xi_i' = 0$ since the inequality

$$-w^\top x_i + b + y_i \leq \epsilon + \xi_i'$$

is trivially satisfied (because $\epsilon, \xi_i' \geq 0$), and if $w^\top x_i - b - y_i \leq 0$, then similarly $\xi_i = 0$. Therefore, we have the equations

$$\xi_i \xi_i' = 0, \quad i = 1, \ldots, m. \tag{$\xi\xi'$}$$

Observe that if $\nu > 1$, then an optimal solution of the above program must yield $\epsilon = 0$. Indeed, if $\epsilon > 0$, we can reduce it by a small amount $\delta > 0$ and increase $\xi_i + \xi_i'$ by δ to still satisfy the constraints, but the objective function changes by the amount $-\nu\delta + \delta$, which is negative since $\nu > 1$, so $\epsilon > 0$ is not optimal.

Driving ϵ to zero is not the intented goal, because typically the data is not noise free so very few pairs (x_i, y_i) will satisfy the equation $w^\top x_i - b = y_i$, and then many pair (x_i, y_i) will correspond to an error ($\xi_i > 0$ or $\xi_i' > 0$). Thus, *typically we assume that* $0 < \nu \leq 1$.

To construct the Lagrangian, we assign Lagrange multipliers $\alpha_i \geq 0$ to the constraints $w^\top x_i - b - y_i \leq \epsilon + \xi_i$, Lagrange multipliers $\alpha_i' \geq 0$ to the constraints $-w^\top x_i + b + y_i \leq \epsilon + \xi_i'$, Lagrange multipliers $\eta_i \geq 0$ to the constraints $\xi_i \geq 0$, Lagrange multipliers $\eta_i' \geq 0$ to the constraints $\xi_i' \geq 0$, and the Lagrange multiplier $\beta \geq 0$ to the constraint $\epsilon \geq 0$. The Lagrangian is

$$\begin{aligned} L(w, b, \alpha, \alpha', \beta, \xi, \xi', \epsilon, \eta, \eta') = {} & \frac{1}{2} w^\top w + C\left(\nu\epsilon + \frac{1}{m}\sum_{i=1}^m (\xi_i + \xi_i')\right) \\ & - \beta\epsilon - \sum_{i=1}^m (\eta_i \xi_i + \eta_i' \xi_i') \\ & + \sum_{i=1}^m \alpha_i (w^\top x_i - b - y_i - \epsilon - \xi_i) \\ & + \sum_{i=1}^m \alpha_i' (-w^\top x_i + b + y_i - \epsilon - \xi_i'), \end{aligned}$$

The Lagrangian can also be written as

$$\begin{aligned} L(w, b, \alpha, \alpha', \beta, \xi, \xi', \epsilon, \eta, \eta') = {} & \frac{1}{2} w^\top w + w^\top \left(\sum_{i=1}^m (\alpha_i - \alpha_i') x_i\right) \\ & + \epsilon\left(C\nu - \beta - \sum_{i=1}^m (\alpha_i + \alpha_i')\right) \\ & + \sum_{i=1}^m \xi_i \left(\frac{C}{m} - \alpha_i - \eta_i\right) + \sum_{i=1}^m \xi_i' \left(\frac{C}{m} - \alpha_i' - \eta_i'\right) \\ & - b\left(\sum_{i=1}^m (\alpha_i - \alpha_i')\right) - \sum_{i=1}^m (\alpha_i - \alpha_i') y_i. \end{aligned}$$

To find the dual function $G(\alpha, \alpha', \eta, \eta', \beta)$, we minimize $L(w, b, \alpha, \alpha', \beta, \xi, \xi', \epsilon, \eta, \eta')$ with respect to the primal variables w, ϵ, b, ξ and ξ'. Observe that the Lagrangian is convex, and since $(w, \epsilon, \xi, \xi') \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^m \times \mathbb{R}^m$, a convex open set, by Theorem 20.11, the Lagrangian has a minimum iff $\nabla L_{w,\epsilon,b,\xi,\xi'} = 0$, so we compute the gradient $\nabla L_{w,\epsilon,b,\xi,\xi'}$. We obtain

$$\nabla L_{w,\epsilon,b,\xi,\xi'} = \begin{pmatrix} w + \sum_{i=1}^m (\alpha_i - \alpha_i')x_i \\ C\nu - \beta - \sum_{i=1}^m (\alpha_i + \alpha_i') \\ \sum_{i=1}^m (\alpha_i - \alpha_i') \\ \frac{C}{m} - \alpha - \eta \\ \frac{C}{m} - \alpha' - \eta' \end{pmatrix},$$

where

$$\left(\frac{C}{m} - \alpha - \eta\right)_i = \frac{C}{m} - \alpha_i - \eta_i, \quad \text{and} \quad \left(\frac{C}{m} - \alpha' - \eta'\right)_i = \frac{C}{m} - \alpha_i' - \eta_i'.$$

Consequently, if we set $\nabla L_{w,\epsilon,b,\xi,\xi'} = 0$, we obtain the equations

$$w = \sum_{i=1}^m (\alpha_i' - \alpha_i)x_i, \tag{$*_w$}$$

$$\begin{aligned} C\nu - \beta - \sum_{i=1}^m (\alpha_i + \alpha_i') &= 0 \\ \sum_{i=1}^m (\alpha_i - \alpha_i') &= 0 \\ \frac{C}{m} - \alpha - \eta = 0, \quad \frac{C}{m} - \alpha' - \eta' &= 0. \end{aligned}$$

Substituting the above equations in the second expression for the Lagrangian, we find that the dual function G is independent of the variables β, η, η' and is given by

$$G(\alpha, \alpha') = -\frac{1}{2}\sum_{i,j=1}^m (\alpha_i' - \alpha_i)(\alpha_j' - \alpha_j)x_i^\top x_j - \sum_{i=1}^m (\alpha_i - \alpha_i')y_i$$

if

$$\begin{aligned} C\nu - \beta - \sum_{i=1}^m (\alpha_i + \alpha_i') &= 0 \\ \sum_{i=1}^m (\alpha_i - \alpha_i') &= 0 \\ \frac{C}{m} - \alpha - \eta = 0, \quad \frac{C}{m} - \alpha' - \eta' &= 0, \end{aligned}$$

and $-\infty$ otherwise.

The dual program is obtained by maximizing $G(\alpha, \alpha')$ or equivalently by minimizing $-G(\alpha, \alpha')$, over $\alpha, \alpha' \in \mathbb{R}^m_+$. Taking into account the fact that $\eta, \eta' \geq 0$ and $\beta \geq 0$, we obtain the following dual program:

$$\begin{aligned}
\text{minimize} \quad & \frac{1}{2}\sum_{i,j=1}^{m}(\alpha'_i - \alpha_i)(\alpha'_j - \alpha_j)x_i^\top x_j + \sum_{i=1}^{m}(\alpha_i - \alpha'_i)y_i \\
\text{subject to} \quad & \\
& \sum_{i=1}^{m}(\alpha_i + \alpha'_i) \leq C\nu \\
& \sum_{i=1}^{m}(\alpha_i - \alpha'_i) = 0 \\
& 0 \leq \alpha_i \leq \frac{C}{m}, \quad 0 \leq \alpha'_i \leq \frac{C}{m}, \quad i = 1, \ldots, m.
\end{aligned}$$

The KKT conditions (for the primal program) are

$$\begin{aligned}
\alpha_i(w^\top x_i - b - y_i - \epsilon - \xi_i) &= 0, \quad i = 1, \ldots, m \\
\alpha'_i(-w^\top x_i + b + y_i - \epsilon - \xi'_i) &= 0, \quad i = 1, \ldots, m \\
\beta\epsilon &= 0 \\
\eta_i\xi_i &= 0, \quad i = 1, \ldots, m \\
\eta'_i\xi'_i &= 0, \quad i = 1, \ldots, m.
\end{aligned}$$

If $\epsilon > 0$, since the equations

$$\begin{aligned}
w^\top x_i - b - y_i &= \epsilon + \xi_i \\
-w^\top x_i + b + y_i &= \epsilon + \xi'_i
\end{aligned}$$

cannot hold simultaneously, we must have

$$\alpha_i\alpha'_i = 0, \quad i = 1, \ldots, m. \tag{$\alpha\alpha'$}$$

From the equations

$$\frac{C}{m} - \alpha_i - \eta_i = 0, \quad \frac{C}{m} - \alpha'_i - \eta'_i = 0, \quad \eta_i\xi_i = 0, \quad \eta'_i\xi'_i = 0,$$

we get the equations

$$\left(\frac{C}{m} - \alpha_i\right)\xi_i = 0, \quad \left(\frac{C}{m} - \alpha'_i\right)\xi'_i = 0, \quad i = 1, \ldots, m. \tag{$*$}$$

These equations show that if $\xi_i > 0$, then $\alpha_i = \frac{C}{m}$, so we have the active constraint

$$w^\top x_i - b - y_i = \epsilon + \xi_i$$

and x_i is an error, and similarly, if $\xi'_i > 0$, then $\alpha'_i = \frac{C}{m}$, so we have the active constraint

$$-w^\top x_i + b + y_i = \epsilon + \xi'_i$$

and x_i is an error.

If the primal has an optimal solution with $w \neq 0$ and $\epsilon > 0$, then by $(*_w)$ and since

$$\sum_{i=1}^{m}(\alpha_i - \alpha'_i) = 0 \quad \text{and} \quad \alpha_i\alpha'_i = 0,$$

there is there is some i_0 such that $\alpha_{i_0} > 0$ and some $j_0 \neq i_0$ such that $\alpha'_{j_0} > 0$. Under the mild hypothesis that there is some i_0 such that $0 < \alpha_{i_0} < \frac{C}{m}$ and there is some j_0 such that $0 < \alpha'_{j_0} < \frac{C}{m}$, then by $(*)$ we have $\xi_{i_0} = 0, \xi'_{j_0} = 0$, and we have the two equations

$$\begin{aligned} w^\top x_{i_0} - b - y_{i_0} &= \epsilon \\ -w^\top x_{j_0} + b + y_{j_0} &= \epsilon, \end{aligned}$$

so b and ϵ can be computed. In particular,

$$b = \frac{1}{2}\left(w^\top(x_{i_0} + x_{j_0}) - (y_{i_0} + y_{j_0})\right).$$

The function $f(x) = w^\top x - b$ (often called *regression estimate*) is given by

$$f(x) = \sum_{i=1}^{m}(\alpha'_i - \alpha_i)x_i^\top x_j - b.$$

The constraints

$$\begin{aligned} \sum_{i=1}^{m}(\alpha_i + \alpha'_i) &\leq C\nu \\ 0 \leq \alpha_i &\leq \frac{C}{m} \\ 0 \leq \alpha'_i &\leq \frac{C}{m} \end{aligned}$$

imply that at most a fraction ν of the data can have $\alpha_i = \frac{C}{m}$ or $\alpha'_i = \frac{C}{m}$. If follows that if $\epsilon > 0$ and $0 < \nu \leq 1$, then ν is an upper bound on the fraction of errors.

The KKT conditions imply that if $\epsilon > 0$, then $\beta = 0$, in which case

$$\sum_{i=1}^{m}(\alpha_i + \alpha_i') = C\nu.$$

Since $\alpha_i\alpha_i' = 0$, and since support vectors correspond to $0 < \alpha_i, \alpha_i' \leq \frac{C}{m}$, we see that ν is a lower bound on the fraction of support vectors.

Since the formulae for w, b, and $f(x)$,

$$\begin{aligned} w &= \sum_{i=1}^{m}(\alpha_i' - \alpha_i)x_i \\ b &= \frac{1}{2}\left(w^\top(x_{i_0} + x_{j_0}) - (y_{i_0} + y_{j_0})\right) \\ f(x) &= \sum_{i=1}^{m}(\alpha_i' - \alpha_i)x_i^\top x_j - b, \end{aligned}$$

only involve inner products among the data points x_i, and since the objective function $-G(\alpha, \alpha')$ of the dual program also only involves inner products among the data points x_i, we can kernelize the ν-SV regression method.

As in the previous section, we assume that our data points $\{x_1, \ldots, x_m\}$ belong to a set $\mathcal{X}$ and we pretend that we have feature space $(F, \langle -, - \rangle)$ and a feature embedding map $\varphi\colon \mathcal{X} \to F$, but we only have access to the kernel function $\kappa(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j)\rangle$. We wish to perform ν-SV regression in the feature space F on the data set $\{(\varphi(x_1), y_1), \ldots, (\varphi(x_m), y_m)\}$. Going over the previous computation, we see that the primal program is given by

kernel ν-SV Regression:

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\langle w, w\rangle + C\left(\nu\epsilon + \frac{1}{m}\sum_{i=1}^{m}(\xi_i + \xi_i')\right) \\ &\text{subject to} \\ &\qquad \langle w, \varphi(x_i)\rangle - b - y_i \leq \epsilon + \xi_i, \quad \xi_i \geq 0 \qquad i = 1, \ldots, m \\ &\qquad -\langle w, \varphi(x_i)\rangle + b + y_i \leq \epsilon + \xi_i', \quad \xi_i' \geq 0 \qquad i = 1, \ldots, m \\ &\qquad \epsilon \geq 0, \end{aligned}$$

minimizing over the variables w, ϵ, b, ξ, and ξ'. The Lagrangian is given by

$$\begin{aligned} L(w, b, \alpha, \alpha', \beta, \xi, \xi', \epsilon, \eta, \eta') = \frac{1}{2}\langle w, w\rangle + \left\langle w, \sum_{i=1}^{m}(\alpha_i - \alpha_i')\varphi(x_i)\right\rangle \\ + \epsilon\left(C\nu - \beta - \sum_{i=1}^{m}(\alpha_i + \alpha_i')\right) \\ + \sum_{i=1}^{m}\xi_i\left(\frac{C}{m} - \alpha_i - \eta_i\right) + \sum_{i=1}^{m}\xi_i'\left(\frac{C}{m} - \alpha_i' - \eta_i'\right) \\ - b\left(\sum_{i=1}^{m}(\alpha_i - \alpha_i')\right) - \sum_{i=1}^{m}(\alpha_i - \alpha_i')y_i. \end{aligned}$$

Setting the gradient $\nabla L_{w,\epsilon,b,\xi,\xi'}$ of the Lagrangian to zero, we also obtain the equations

$$w = \sum_{i=1}^{m}(\alpha_i' - \alpha_i)\varphi(x_i), \tag{$*_w$}$$

$$\begin{aligned} C\nu - \beta - \sum_{i=1}^{m}(\alpha_i + \alpha_i') = 0 \\ \sum_{i=1}^{m}(\alpha_i - \alpha_i') = 0 \\ \frac{C}{m} - \alpha - \eta = 0, \quad \frac{C}{m} - \alpha' - \eta' = 0. \end{aligned}$$

Using the above equations, we find that the dual function G is independent of the variables β, η, η', and we obtain the following dual program:

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\sum_{i,j=1}^{m}(\alpha_i' - \alpha_i)(\alpha_j' - \alpha_j)\kappa(x_i, x_j) + \sum_{i=1}^{m}(\alpha_i - \alpha_i')y_i \\ &\text{subject to} \\ &\sum_{i=1}^{m}(\alpha_i + \alpha_i') \leq C\nu \\ &\sum_{i=1}^{m}(\alpha_i - \alpha_i') = 0 \\ &0 \leq \alpha_i \leq \frac{C}{m}, \quad 0 \leq \alpha_i' \leq \frac{C}{m}, \quad i = 1, \ldots, m. \end{aligned}$$

Everything we said before also applies to the kernel ν-SV regression method, except that x_i is replaced by $\varphi(x_i)$ and that the inner product $\langle -, - \rangle$ must be used, and we have the formulae

$$\begin{aligned} w &= \sum_{i=1}^{m}(\alpha_i' - \alpha_i)\varphi(x_i) \\ b &= \frac{1}{2}\left(\sum_{i=1}^{m}(\alpha_i' - \alpha_i)(\kappa(x_i x_{i_0}) + \kappa(x_i, x_{j_0})) - (y_{i_0} + y_{j_0})\right) \\ f(x) &= \sum_{i=1}^{m}(\alpha_i' - \alpha_i)\kappa(x_i, x_j) - b, \end{aligned}$$

expressions that only involve κ.

Remark: There is a variant of ν-SV regression obtained by setting $\nu = 0$ and holding $\epsilon > 0$ fixed. This method is called *ϵ-SV regression* or *(linear) ϵ-insensitive SV regression*. The corresponding optimization program is

ϵ-SV Regression:

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}w^\top w + \frac{C}{m}\sum_{i=1}^{m}(\xi_i + \xi_i') \\ &\text{subject to} \\ &\qquad w^\top x_i - b - y_i \leq \epsilon + \xi_i, \quad \xi_i \geq 0 \qquad i = 1, \ldots, m \\ &\qquad -w^\top x_i + b + y_i \leq \epsilon + \xi_i', \quad \xi_i' \geq 0 \qquad i = 1, \ldots, m, \end{aligned}$$

minimizing over the variables w, b, ξ, and ξ'.

It is easy to see that the dual program is

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\sum_{i,j=1}^{m}(\alpha_i' - \alpha_i)(\alpha_j' - \alpha_j)x_i^\top x_j + \sum_{i=1}^{m}(\alpha_i - \alpha_i')y_i + \epsilon\sum_{i=1}^{m}(\alpha_i + \alpha_i') \\ &\text{subject to} \\ &\qquad \sum_{i=1}^{m}(\alpha_i - \alpha_i') = 0 \\ &\qquad 0 \leq \alpha_i \leq \frac{C}{m}, \quad 0 \leq \alpha_i' \leq \frac{C}{m}, \quad i = 1, \ldots, m. \end{aligned}$$

The constraint

$$\sum_{i=1}^{m}(\alpha_i + \alpha_i') \leq C\nu$$

is gone but the extra term $\epsilon \sum_{i=1}^{m}(\alpha_i + \alpha'_i)$ has been added to the dual function, to prevent α_i and α'_i from blowing up.

There is an obvious kernelized version of ϵ-SV regression. It is easy to show that ν-SV regression subsumes ϵ-SV regression, in the sense that if ν-SV regression succeeds and yields $w, b, \epsilon > 0$, then ϵ-SV regression with the same C and the same value of ϵ also succeeds and returns the same pair (w, b). For more details on these methods, see Schölkopf, Smola, Williamson, and Bartlett [87].

Remark: The linear penalty function $\sum_{i=1}^{m}(\xi_i + \xi'_i)$ can be repaced by the quadratic penalty function $\sum_{i=1}^{m}(\xi_i^2 + \xi_i'^2)$; see Shawe–Taylor and Christianini [96] (Chapter 7).

Yet another variant of ν-SV regression is to add the term $\frac{1}{2}b^2$ to the objective function. The new Lagrangian is

$$\begin{aligned} L(w, b, \alpha, \alpha', \beta, \xi, \xi', \epsilon, \eta, \eta') = {} & \frac{1}{2} w^\top w + w^\top \left(\sum_{i=1}^{m} (\alpha_i - \alpha'_i) x_i \right) \\ & + \epsilon \left(C\nu - \beta - \sum_{i=1}^{m} (\alpha_i + \alpha'_i) \right) \\ & + \sum_{i=1}^{m} \xi_i \left(\frac{C}{m} - \alpha_i - \eta_i \right) + \sum_{i=1}^{m} \xi'_i \left(\frac{C}{m} - \alpha'_i - \eta'_i \right) \\ & + \frac{1}{2} b^2 - b \left(\sum_{i=1}^{m} (\alpha_i - \alpha'_i) \right) - \sum_{i=1}^{m} (\alpha_i - \alpha'_i) y_i. \end{aligned}$$

We obtain the new equation

$$b = \sum_{i=1}^{m} (\alpha_i - \alpha'_i)$$

determining b, which replaces the equation

$$\sum_{i=1}^{m} (\alpha_i - \alpha'_i) = 0.$$

The new dual program is

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} \sum_{i,j=1}^{m} (\alpha'_i - \alpha_i)(\alpha'_j - \alpha_j)(x_i^\top x_j + 1) + \sum_{i=1}^{m} (\alpha_i - \alpha'_i) y_i \\ \text{subject to} \quad & \\ & \sum_{i=1}^{m} (\alpha_i + \alpha'_i) \leq C\nu \\ & 0 \leq \alpha_i \leq \frac{C}{m}, \quad 0 \leq \alpha'_i \leq \frac{C}{m}, \quad i = 1, \ldots, m. \end{aligned}$$

Chapter 32

Soft Margin Support Vector Machines

If the sets of points $\{u_1, \ldots, u_p\}$ and $\{v_1, \ldots, v_q\}$ are not linearly separable (with $u_i, v_j \in \mathbb{R}^n$), we can use a trick from linear programming, which is to introduce nonnegative "slack variables" $\epsilon = (\epsilon_1, \ldots, \epsilon_p) \in \mathbb{R}^p$ and $\xi = (\xi_1, \ldots, \xi_q) \in \mathbb{R}^q$ to relax the "hard" constraints

$$\begin{aligned} w^\top u_i - b &\geq \delta && i = 1, \ldots, p \\ -w^\top v_j + b &\geq \delta && j = 1, \ldots, q \end{aligned}$$

of Problem (SVM$_{h1}$) from Section 30.3 to the "soft" constraints

$$\begin{aligned} w^\top u_i - b &\geq \delta - \epsilon_i, \quad \epsilon_i \geq 0 \quad i = 1, \ldots, p \\ -w^\top v_j + b &\geq \delta - \xi_j, \quad \xi_j \geq 0 \quad j = 1, \ldots, q. \end{aligned}$$

Recall that $w \in \mathbb{R}^n$ and $b, \delta \in \mathbb{R}$.

If $\epsilon_i > 0$, the point u_i may be misclassified, in the sense that it can belong to the margin (the slab), or even to the wrong half-space classifying the negative (red) points. See Figures 32.1 (2) and (3). Similarly, if $\xi_j > 0$, the point v_j may be misclassified, in the sense that it can belong to the margin (the slab), or even to the wrong half-space classifying the positive (blue) points. We can think of ϵ_i as a measure of how much the constraint $w^\top u_i - b \geq \delta$ is violated, and similarly of ξ_j as a measure of how much the constraint $-w^\top v_j + b \geq \delta$ is violated. If $\epsilon = 0$ and $\xi = 0$, then we recover the original constraints. By making ϵ and ξ large enough, these constraints can always be satisfied. We add the constraint $w^\top w \leq 1$ and we minimize $-\delta$.

If instead of the constraints of Problem (SVM$_{h1}$) we use the hard constraints

$$\begin{aligned} w^\top u_i - b &\geq 1 && i = 1, \ldots, p \\ -w^\top v_j + b &\geq 1 && j = 1, \ldots, q \end{aligned}$$

of Problem (SVM$_{h2}$) (see Example 30.4), then we relax to the soft constraints

$$\begin{aligned} w^\top u_i - b &\geq 1 - \epsilon_i, \quad \epsilon_i \geq 0 \quad i = 1, \ldots, p \\ -w^\top v_j + b &\geq 1 - \xi_j, \quad \xi_j \geq 0 \quad j = 1, \ldots, q. \end{aligned}$$

In this case, there is no constraint on w, but we minimize $(1/2)w^\top w$.

Ideally we would like to find a separating hyperplane that *minimizes the number of misclassified points*, which means that the variables ϵ_i and ξ_j should be as small as possible, but there is a trade-off in maximizing the margin (the thickness of the slab), and minimizing the number of misclassified points. This is reflected in the choice of the objective function, and there are several options, depending on whether we minimize a linear function of the variables ϵ_i and ξ_j, or a quadratic functions of these variables, or whether we include the term $(1/2)b^2$ in the objective function. These methods are known as *support vector classification* algorithms (for short *SVC* algorithms).

SVC algorithms seek an "optimal" separating hyperplane H of equation $w^\top x - b = 0$. If some new data $x \in \mathbb{R}^n$ comes in, we can classify it by determining in which of the two half spaces determined by the hyperplane H they belong, by computing the sign of the quantity $w^\top x - b$. The function $\mathrm{sgn}\colon \mathbb{R} \to \{-1, 1\}$ is given by

$$\mathrm{sgn}(x) = \begin{cases} +1 & \text{if } x \geq 0 \\ -1 & \text{if } x < 0. \end{cases}$$

Then we define the *(binary) classification function* associated with the hyperplane H of equation $w^\top x - b = 0$ as

$$f(x) = \mathrm{sgn}(w^\top x - b).$$

Remarkably, all the known optimization problems for finding this hyperplane share the property that the weight vector w and the constant b are given by expressions that *only involves inner products of the input data points u_i and v_j*, and so does the classification function

$$f(x) = \mathrm{sgn}(w^\top x - b).$$

This is a key fact that allows a far reaching generalization of the support vector machine using the method of *kernels*.

The method of kernels consists in assuming that the input space $\mathbb{R}^n$ is embedded in a larger (possibly infinite dimensional) Euclidean space F (with an inner product $\langle -, - \rangle$) usually called a *feature space*, using a function

$$\varphi\colon \mathbb{R}^n \to F$$

called a *feature map*. The function $\kappa\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ given by

$$\kappa(x, y) = \langle \varphi(x), \varphi(y) \rangle$$

is the kernel function associated with the embedding φ; see Chapter 31. The idea is that the feature map φ "unwinds" the input data, making it somehow more linear in the higher dimensional space F. Now even if we don't know what the feature space F is and what the

embedding map φ is, we can pretend to solve our separation problem in F for the embedded data points $\varphi(u_i)$ and $\varphi(v_j)$. Thus we seek a hyperplane H of equation

$$\langle w, \zeta\rangle - b = 0, \quad \zeta \in F,$$

in the feature space F, to attempt to separate the points $\varphi(u_i)$ and the points $\varphi(v_j)$. As we said, it turns out that w and b are given by expression involving only the inner products $\kappa(u_i, u_j) = \langle\varphi(u_i), \varphi(u_j)\rangle, \kappa(u_i, v_j) = \langle\varphi(u_i), \varphi(v_j)\rangle$, and $\kappa(v_i, v_j) = \langle\varphi(v_i), \varphi(v_j)\rangle$, which form the symmetric $(p+q)\times(p+q)$ matrix $\mathbf{K}$ (a kernel matrix) given by

$$\mathbf{K}_{ij} = \begin{cases} \kappa(u_i, u_j) & 1 \le i \le p,\ 1 \le j \le q \\ -\kappa(u_i, v_{j-p}) & 1 \le i \le p,\ p+1 \le j \le p+q \\ -\kappa(v_{i-p}, u_j) & p+1 \le i \le p+q,\ 1 \le j \le p \\ \kappa(v_{i-p}, v_{j-q}) & p+1 \le i \le p+q,\ p+1 \le j \le p+q. \end{cases}$$

Then the classification function

$$f(x) = \mathrm{sgn}(\langle w, \varphi(x)\rangle - b)$$

for points in the original data space $\mathbb{R}^n$ is also expressed solely in terms of the matrix $\mathbf{K}$ and the inner products $\kappa(u_i, x) = \langle\varphi(u_i), \varphi(x)\rangle$ and $\kappa(v_j, x) = \langle\varphi(v_j), \varphi(x)\rangle$. As a consequence, in the original data space $\mathbb{R}^n$, the hypersurface

$$\mathcal{S} = \{x \in \mathbb{R}^n \mid \langle w, \varphi(x)\rangle - b = 0\}$$

separates the data points u_i and v_j, but it is not an affine subspace of $\mathbb{R}^n$. The classification function f tells us on which "side" of $\mathcal{S}$ is a new data point $x \in \mathbb{R}^n$. Thus, we managed to separate the data points u_i and v_j that are not separable by an affine hyperplane, by a *nonaffine hypersurface* $\mathcal{S}$, by assuming that an embdedding $\varphi\colon \mathbb{R}^n \to F$ exists, even though we don't know what it is, but having access to F through the kernel function $\kappa\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ given by the inner products $\kappa(x, y) = \langle\varphi(x), \varphi(y)\rangle$.

In practice, the art of using the kernel method is to choose the right kernel (as the knight says in Indiana Jones, to "choose wisely.").

The method of kernels is very flexible. It also applies to the soft margin versions of SVM, but also to regression problems, and to principal component analysis (PCA), and to other problems arising in machine learning.

Comprehensive presentations of the method of kernels are found in Schölkopf and Smola [85] and Shawe–Taylor and Christianini [96]. See also Bishop [18].

We first consider the soft margin SVM arising from Problem (SVM_{h1}).

32.1 Soft Margin Support Vector Machines; (SVM_{s1})

In this section we derive the dual function G associated with the following version of the soft margin SVM coming from Problem (SVM_{h1}), where the maximization of the margin δ has been replaced by the minimization of $-\delta$, and where we added a "regularizing term" $K\Big(\sum_{i=1}^{p}\epsilon_i + \sum_{j=1}^{q}\xi_j\Big)$ whose purpose is to make $\epsilon \in \mathbb{R}^p$ and $\xi \in \mathbb{R}^q$ *sparse* (that is, try to make ϵ_i and ξ_j have as many zeros as possible), where $K > 0$ is a fixed constant that can be adjusted to determine the influence of this regularizing term. If the primal problem (SVM_{s1}) has an optimal solution $(w, \delta, b, \epsilon, \xi)$, we attempt to use the dual function G to obtain it, but we will see that with this particular formulation of the problem, the constraint $w^\top w \leq 1$ causes troubles, even though it is convex.

Soft margin SVM (SVM_{s1}):

$$\begin{aligned}
\text{minimize} \quad & -\delta + K\Big(\sum_{i=1}^{p}\epsilon_i + \sum_{j=1}^{q}\xi_j\Big) \\
\text{subject to} \quad & \\
& w^\top u_i - b \geq \delta - \epsilon_i, \quad \epsilon_i \geq 0 \qquad i = 1, \ldots, p \\
& -w^\top v_j + b \geq \delta - \xi_j, \quad \xi_j \geq 0 \qquad j = 1, \ldots, q \\
& w^\top w \leq 1.
\end{aligned}$$

It is customary to write $\ell = p + q$.

For this problem, the primal problem may have an optimal solution $(w, \delta, b, \epsilon, \xi)$ with $\|w\| = 1$ and $\delta > 0$, but if the sets of points are not linearly separable then an optimal solution of the dual may not yield w.

The objective function of our problem is affine and the only nonaffine constraint $w^\top w \leq 1$ is convex. This constraint is qualified because for any $w \neq 0$ such that $w^\top w < 1$ and for any $\delta > 0$ and any b we can pick ϵ and ξ large enough so that the constaints are satisfied. Consequently, by Theorem 30.14(2) *if* the primal problem (SVM_{s1}) has an optimal solution, *then* the dual problem has a solution too, and the duality gap is zero.

Unfortunately this does not imply that an optimal solution of the dual yields an optimal solution of the primal because the hypotheses of Theorem 30.14(1) fail to hold. In general, there may not be a unique vector $(w, \epsilon, \xi, b, \delta)$ such that

$$\inf_{w,\epsilon,\xi,b,\delta} L(w, \epsilon, \xi, b, \delta, \lambda, \mu, \alpha, \beta, \gamma) = G(\lambda, \mu, \alpha, \beta, \gamma).$$

If the sets $\{u_i\}$ and $\{v_j\}$ are *not* linearly separable, then the dual problem may have a solution for which $\gamma = 0$,

$$\sum_{i=1}^{p}\lambda_i = \sum_{j=1}^{q}\mu_j = \frac{1}{2},$$

and

$$\sum_{i=1}^{p} \lambda_i u_i = \sum_{j=1}^{q} \mu_j v_j,$$

so that the dual function $G(\lambda, \mu, \alpha, \beta, \gamma)$, which is a *partial function*, is defined and has the value $G(\lambda, \mu, \alpha, \beta, 0) = 0$. Such a pair (λ, μ) corresponds to the coefficients of two convex combinations

$$\sum_{i=1}^{p} 2\lambda_i u_i = \sum_{j=1}^{q} 2\mu_j v_j$$

which correspond to the *same point* in the (nonempty) intersection of the convex hulls $\mathrm{conv}(u_1, \ldots, u_p)$ and $\mathrm{conv}(v_1, \ldots, v_q)$. It turns out that the only connection between w and the dual function is the equation

$$2\gamma w = \sum_{i=1}^{p} \lambda_i u_i - \sum_{j=1}^{q} \mu_j v_j,$$

and when $\gamma = 0$ this is equation is $0 = 0$, so the dual problem is useless to determine w. This point seems to have been missed in the literature (for example, in Shawe–Taylor and Christianini [96], Section 7.2). What the dual problem does show is that $\delta \geq 0$. However, if $\gamma \neq 0$, then w is determined by any solution (λ, μ) of the dual.

It still remains to compute δ and b, which can be done under a mild hypothesis that we call the **Standard Margin Hypothesis**.

If $(w, \delta, b, \epsilon, \xi)$ is an optimal solution of Problem (SVM_{s1}), then the points u_i and v_j are classified as follows:

(1) If $\epsilon_i = 0$, then the point u_i is correctly classified and is either on the blue margin (the hyperplane $H_{w,b+\eta}$ of equation $w^\top x = b + \eta$) or on the correct side of the blue margin (the blue side). Similarly, if $\xi_j = 0$, then the point v_j is correctly classified and is either on the red margin (the hyperplane $H_{w,b-\eta}$ of equation $w^\top x = b - \eta$) or on the correct side of the red margin (the red side).

(2) If $0 < \epsilon_i \leq \eta$, then the point u_i lies inside the margin (the slab), but on the correct side of the separating hyperplane (the blue side). If $\epsilon_i = \eta$, then u_i lies on the separating hyperplane. Similarly, if $0 < \xi_j \leq \eta$, then the point v_j lies inside the margin (the slab), but on the correct side of the separating hyperplane (the red side). If $\xi_j = \eta$, then v_j lies on the separating hyperplane.

(3) If $\epsilon_i > \eta$, then the point u_i lies on the wrong side of the separating hyperplane (the red side); it is misclassified. Similarly, if $\xi_j > \eta$, then the point v_j lies on the wrong side of the separating hyperplane (the blue side); it is misclassified.

Let $\lambda \in \mathbb{R}^p_+$ be the Lagrange multipliers associated with the inequalities $w^\top u_i - b \geq \delta - \epsilon_i$, let $\mu \in \mathbb{R}^q_+$ be the Lagrange multipliers are associated with the inequalities $-w^\top v_j + b \geq \delta - \xi_j$, let $\alpha \in \mathbb{R}^p_+$ be the Lagrange multipliers associated with the inequalities $\epsilon_i \geq 0$, $\beta \in \mathbb{R}^q_+$ be the Lagrange multipliers associated with the inequalities $\xi_j \geq 0$, and let $\gamma \in \mathbb{R}^+$ be the Lagrange multiplier associated with the inequality $w^\top w \leq 1$.

The linear constraints are given by the $2(p+q) \times (n+p+q+2)$ matrix given in block form by

$$C = \begin{pmatrix} X^\top & -I_{p+q} & \mathbf{1}_p & \mathbf{1}_{p+q} \\ & & -\mathbf{1}_q & \\ 0_{p+q,n} & -I_{p+q} & 0_{p+q} & 0_{p+q} \end{pmatrix},$$

where X is the $n \times (p+q)$ matrix

$$X = \begin{pmatrix} -u_1 & \cdots & -u_p & v_1 & \cdots & v_q \end{pmatrix},$$

and the linear constraints are expressed by

$$\begin{pmatrix} X^\top & -I_{p+q} & \mathbf{1}_p & \mathbf{1}_{p+q} \\ & & -\mathbf{1}_q & \\ 0_{p+q,n} & -I_{p+q} & 0_{p+q} & 0_{p+q} \end{pmatrix} \begin{pmatrix} w \\ \epsilon \\ \xi \\ b \\ \delta \end{pmatrix} \leq \begin{pmatrix} 0_{p+q} \\ 0_{p+q} \end{pmatrix}.$$

More explicitly, C is the following matrix:

$$C = \begin{pmatrix} -u_1^\top & -1 & \cdots & 0 & 0 & \cdots & 0 & 1 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ -u_p^\top & 0 & \cdots & -1 & 0 & \cdots & 0 & 1 & 1 \\ v_1^\top & 0 & \cdots & 0 & -1 & \cdots & 0 & -1 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ v_q^\top & 0 & \cdots & 0 & 0 & \cdots & -1 & -1 & 1 \\ 0 & -1 & \cdots & 0 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & \cdots & 0 & -1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & -1 & 0 & 0 \end{pmatrix}.$$

The objective function is given by

$$J(w, \epsilon, \xi, b, \delta) = -\delta + K \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} \mathbf{1}_{p+q}.$$

The Lagrangian $L(w, \epsilon, \xi, b, \delta, \lambda, \mu, \alpha, \beta, \gamma)$ with $\lambda, \alpha \in \mathbb{R}^p_+$, $\mu, \beta \in \mathbb{R}^q_+$, and $\gamma \in \mathbb{R}^+$ is given by

$$L(w, \epsilon, \xi, b, \delta, \lambda, \mu, \alpha, \beta, \gamma) = -\delta + K \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} \mathbf{1}_{p+q} + \begin{pmatrix} w^\top & \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} & b & \delta \end{pmatrix} C^\top \begin{pmatrix} \lambda \\ \mu \\ \alpha \\ \beta \end{pmatrix} + \gamma(w^\top w - 1).$$

Since

$$\begin{pmatrix} w^\top & \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} & b & \delta \end{pmatrix} C^\top \begin{pmatrix} \lambda \\ \mu \\ \alpha \\ \beta \end{pmatrix} = w^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} - \epsilon^\top(\lambda + \alpha) - \xi^\top(\mu + \beta) + b(\mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu) + \delta(\mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu),$$

the Lagrangian can be written as

$$\begin{aligned} L(w, \epsilon, \xi, b, \delta, \lambda, \mu, \alpha, \beta, \gamma) &= -\delta + K(\epsilon^\top \mathbf{1}_p + \xi^\top \mathbf{1}_q) + w^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + \gamma(w^\top w - 1) \\ &\quad - \epsilon^\top(\lambda + \alpha) - \xi^\top(\mu + \beta) + b(\mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu) + \delta(\mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu) \\ &= (\mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu - 1)\delta + w^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + \gamma(w^\top w - 1) \\ &\quad + \epsilon^\top(K\mathbf{1}_p - (\lambda + \alpha)) + \xi^\top(K\mathbf{1}_q - (\mu + \beta)) + b(\mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu). \end{aligned}$$

To find the dual function $G(\lambda, \mu, \alpha, \beta, \gamma)$ we minimize $L(w, \epsilon, \xi, b, \delta, \lambda, \mu, \alpha, \beta, \gamma)$ with respect to w, ϵ, ξ, b, and δ. Since the Lagrangian is convex and $(w, \epsilon, \xi, b, \delta) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R} \times \mathbb{R}$, a convex open set, by Theorem 20.11, the Lagrangian has a minimum in $(w, \epsilon, \xi, b, \delta)$ iff $\nabla L_{w,\epsilon,\xi,b,\delta} = 0$, so we compute the gradient with respect to $w, \epsilon, \xi, b, \delta$ and we get

$$\nabla L_{w,\epsilon,\xi,b,\delta} = \begin{pmatrix} X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + 2\gamma w \\ K\mathbf{1}_p - (\lambda + \alpha) \\ K\mathbf{1}_q - (\mu + \beta) \\ \mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu \\ \mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu - 1 \end{pmatrix}.$$

By setting $\nabla L_{w,\epsilon,\xi,b,\delta} = 0$ we get the equations

$$2\gamma w = -X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \tag{$*_w$}$$

and

$$\begin{aligned}\lambda + \alpha &= K\mathbf{1}_p\\ \mu + \beta &= K\mathbf{1}_q\\ \mathbf{1}_p^\top\lambda &= \mathbf{1}_q^\top\mu\\ \mathbf{1}_p^\top\lambda + \mathbf{1}_q^\top\mu &= 1.\end{aligned}$$

The second and third equations are equivalent to the inequalities

$$0 \leq \lambda_i, \mu_j \leq K, \quad i = 1, \ldots, p,\ j = 1, \ldots, q,$$

often called *box constraints*, and the fourth and fifth equations yield

$$\mathbf{1}_p^\top\lambda = \mathbf{1}_q^\top\mu = \frac{1}{2}.$$

First let us consider the singular case $\gamma = 0$. In this case, $(*_w)$ implies that

$$X\begin{pmatrix}\lambda\\ \mu\end{pmatrix} = 0,$$

and the term $\gamma(w^\top w - 1)$ is missing from the Lagrangian, which in view of the other four equations above reduces to

$$L(w, \epsilon, \xi, b, \delta, \lambda, \mu, \alpha, \beta, 0) = w^\top X\begin{pmatrix}\lambda\\ \mu\end{pmatrix} = 0.$$

In summary, we proved that if $\gamma = 0$, then

$$G(\lambda, \mu, \alpha, \beta, 0) = \begin{cases} 0 & \text{if } \begin{cases} \sum_{i=1}^p \lambda_i = \sum_{j=1}^q \mu_j = \frac{1}{2} \\ 0 \leq \lambda_i \leq K,\ i = 1, \ldots, p \\ 0 \leq \mu_j \leq K,\ j = 1, \ldots, q \end{cases} \\ -\infty & \text{otherwise} \\ & \text{and } \sum_{i=1}^p \lambda_i u_i - \sum_{j=1}^q \mu_j u_j = 0. \end{cases}$$

Geometrically, (λ, μ) corresponds to the coefficients of two convex combinations

$$\sum_{i=1}^p 2\lambda_i u_i = \sum_{j=1}^q 2\mu_j v_j$$

which correspond to the *same point* in the intersection of the convex hulls $\mathrm{conv}(u_1, \ldots, u_p)$ and $\mathrm{conv}(v_1, \ldots, v_q)$, iff the sets $\{u_i\}$ and $\{v_j\}$ are *not linearly separable*. If the sets $\{u_i\}$ and $\{v_j\}$ are *linearly separable*, then the convex hulls $\mathrm{conv}(u_1, \ldots, u_p)$ and $\mathrm{conv}(v_1, \ldots, v_q)$ are disjoint, which implies that $\gamma > 0$.

Let us now assume that $\gamma > 0$. Plugging back w from equation $(*_w)$ into the Lagrangian, after simplifications we get

$$\begin{aligned} G(\lambda, \mu, \alpha, \beta, \gamma) &= -\frac{1}{2\gamma} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + \frac{\gamma}{4\gamma^2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} - \gamma \\ &= -\frac{1}{4\gamma} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} - \gamma, \end{aligned}$$

so if $\gamma > 0$ the dual function is independent of α, β and is given by

$$G(\lambda, \mu, \alpha, \beta, \gamma) = \begin{cases} -\frac{1}{4\gamma} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} - \gamma & \text{if } \begin{cases} \sum_{i=1}^p \lambda_i = \sum_{j=1}^q \mu_j = \frac{1}{2} \\ 0 \leq \lambda_i \leq K,\ i = 1, \ldots, p \\ 0 \leq \mu_j \leq K,\ j = 1, \ldots, q \end{cases} \\ -\infty & \text{otherwise.} \end{cases}$$

Since $X^\top X$ is symmetric positive definite and $\gamma \geq 0$, obviously

$$G(\lambda, \mu, \alpha, \beta, \gamma) \leq 0$$

for all $\gamma > 0$.

The dual program is given by

$$\begin{aligned} \text{maximize} \quad & -\frac{1}{4\gamma} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} - \gamma \quad \text{if } \gamma > 0 \\ & 0 \quad \text{if} \quad \gamma = 0 \\ \text{subject to} \quad & \\ & \sum_{i=1}^p \lambda_i = \sum_{j=1}^q \mu_j = \frac{1}{2} \\ & 0 \leq \lambda_i \leq K, \quad i = 1, \ldots, p \\ & 0 \leq \mu_j \leq K, \quad j = 1, \ldots, q. \end{aligned}$$

Also, if $\gamma = 0$ then $X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} = 0$.

Maximizing with respect to $\gamma > 0$ yields

$$\gamma^2 = \frac{1}{4} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix},$$

so we obtain

$$G(\lambda, \mu) = -\left(\begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \right)^{1/2}.$$

Finally, since $G(\lambda, \mu) = 0$ and $X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} = 0$ if $\gamma = 0$, the dual program is equivalent to the following minimization program:

$$\begin{aligned} &\text{minimize} \quad \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \\ &\text{subject to} \\ &\qquad \sum_{i=1}^{p} \lambda_i = \sum_{j=1}^{q} \mu_j = \frac{1}{2} \\ &\qquad 0 \leq \lambda_i \leq K, \quad i = 1, \ldots, p \\ &\qquad 0 \leq \mu_j \leq K, \quad j = 1, \ldots, q. \end{aligned}$$

Observe that the constraints imply that K must be chosen so that

$$K \geq \max\left\{\frac{1}{2p}, \frac{1}{2q}\right\}.$$

The dual program is solved by making use of numerical procedures based on gradient descent. If the primal problem is solvable, this yields solutions for λ and μ.

If the optimal value is 0, then $\gamma = 0$ and $X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} = 0$, so in this case it is not possible to determine w. However, if the optimal value is > 0, then once a solution for λ and μ is obtained, by $(*_w)$, we have

$$\begin{aligned} \gamma &= \frac{1}{2}\left(\begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix}\right)^{1/2} \\ w &= \frac{1}{2\gamma}\left(\sum_{i=1}^{p} \lambda_i u_i - \sum_{j=1}^{q} \mu_j v_j\right), \end{aligned}$$

so we get

$$w = \frac{\displaystyle\sum_{i=1}^{p} \lambda_i u_i - \sum_{j=1}^{q} \mu_j v_j}{\left(\begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix}\right)^{1/2}},$$

which is the result of making $\sum_{i=1}^{p} \lambda_i u_i - \sum_{j=1}^{q} \mu_j v_j$ a unit vector, since

$$X = \begin{pmatrix} -u_1 & \cdots & -u_p & v_1 & \cdots & v_q \end{pmatrix}.$$

It remains to find b and δ, which are not given by the dual program.

The complementary slackness conditions yield a classification of the points in terms of the values of λ and μ. Indeed, we have $\epsilon_i\alpha_i = 0$ for $i = 1, \ldots, p$ and $\xi_j\beta_j = 0$ for $j = 1, \ldots, q$. Also, if $\lambda_i > 0$, then corresponding constraint is active, and similarly if $\mu_j > 0$. Since $\lambda_i + \alpha_i = K$, it follows that $\epsilon_i\alpha_i = 0$ iff $\epsilon_i(K - \lambda_i) = 0$, and since $\mu_j + \beta_j = K$, we have $\xi_j\beta_j = 0$ iff $\xi_j(K - \mu_j) = 0$. Thus if $\epsilon_i > 0$ then $\lambda_i = K$, and if $\xi_j > 0$, then $\mu_j = K$. Consequently, if $\lambda_i < K$ then $\epsilon_i = 0$ and u_i is correctly classified, and similarly if $\mu_j < K$ then $\xi_j = 0$ and v_j is correctly classified. We have the following classification:

(1) If $0 < \lambda_i < K$ then u_i is on the margin and is classified correctly. Similarly, if $0 < \mu_j < K$ then v_j is on the margin and is classified correctly.

(2) If $\lambda_i = K$, then if $\epsilon_i \leq \delta$ the point u_i may be classified correctly or it lies within the margin on the correct side, but if $\epsilon_i > \delta$ then it is misclassifed. Similarly, if $\mu_j = K$, then if $\xi_j \leq \delta$ the point v_j may be classified correctly or it lies within the margin on the correct side, but if $\xi_j > \delta$ then it is misclassifed.

(3) If $\lambda_i = 0$ then u_i is classified correctly. Similarly, if $\mu_j = 0$ then v_j is classified correctly.

The equations

$$\sum_{i=1}^{p} \lambda_i = \sum_{j=1}^{q} \mu_j = \frac{1}{2}$$

imply that there is some i_0 such that $\lambda_{i_0} > 0$ and some j_0 such that $\mu_{j_0} > 0$, but a priori, nothing prevents the situation where $\lambda_i = K$ for all nonzero λ_i or $\mu_j = K$ for all nonzero μ_j. If this happens, we can rerun the optimization method with a larger value of K. If the following mild hypothesis holds then b and δ can be found.

Standard Margin Hypothesis for (SVM_{s1}). There is some index i_0 such that $0 < \lambda_{i_0} < K$ and there is some index j_0 such that $0 < \mu_{j_0} < K$. This means that some u_{i_0} is correctly classified and on the blue margin, and some v_{j_0} is correctly classified and on the red margin.

If the **Standard Margin Hypothesis** for (SVM_{s1}) holds then $\epsilon_{i_0} = 0$ and $\mu_{j_0} = 0$, and then we have the active equations

$$w^\top u_{i_0} - b = \delta \quad \text{and} \quad -w^\top v_{j_0} + b = \delta,$$

and we obtain the value of b and δ as

$$b = \frac{1}{2}(w^\top u_{i_0} + w^\top v_{j_0})$$
$$\delta = \frac{1}{2}(w^\top u_{i_0} - w^\top v_{j_0}).$$

As we said earlier, the hypotheses of Theorem 30.14(2) hold, so *if* the primal problem (SVM_{s1}) has an optimal solution with $w \neq 0$, *then* the dual problem has a solution too, and the duality gap is zero. Therefore, for optimal solutions we have

$$L(w, \epsilon, \xi, b, \delta, \lambda, \mu, \alpha, \beta, \gamma) = G(\lambda, \mu, \alpha, \beta, \gamma),$$

which means that

$$-\delta + K\Big(\sum_{i=1}^{p}\epsilon_i + \sum_{j=1}^{q}\xi_j\Big) = -\left(\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda\\ \mu\end{pmatrix}\right)^{1/2},$$

so we get

$$\delta = K\Big(\sum_{i=1}^{p}\epsilon_i + \sum_{j=1}^{q}\xi_j\Big) + \left(\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda\\ \mu\end{pmatrix}\right)^{1/2}.$$

Therefore, we confirm that $\delta \geq 0$.

It is important to note that the objective function of the dual program

$$-G(\lambda,\mu) = \left(\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda\\ \mu\end{pmatrix}\right)^{1/2}$$

only involves the inner products of the u_i and the v_j through the matrix $X^\top X$, and similarly, the equation of the optimal hyperplane can be written as

$$\sum_{i=1}^{p}\lambda_i u_i^\top x - \sum_{j=1}^{q}\mu_j v_j^\top x - \left(\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda\\ \mu\end{pmatrix}\right)^{1/2} b = 0,$$

an expression that only involves inner products of x with the u_i and the v_j and inner products of the u_i and the v_j.

As explained at the beginning of this chapter, this is a key fact that allows ax generalization of the support vector machine using the method of *kernels*. We can define the following "kernelized" version of Problem (SVM_{s1}):

Soft margin kernel SVM (SVM_{s1}):

$$\begin{aligned}
&\text{minimize} \quad -\delta + K\Big(\sum_{i=1}^{p}\epsilon_i + \sum_{j=1}^{q}\xi_j\Big)\\
&\text{subject to}\\
&\qquad \langle w, \varphi(u_i)\rangle - b \geq \delta - \epsilon_i, \quad \epsilon_i \geq 0 \qquad i = 1,\ldots,p\\
&\qquad -\langle w, \varphi(v_j)\rangle + b \geq \delta - \xi_j, \quad \xi_j \geq 0 \qquad j = 1,\ldots,q\\
&\qquad \langle w, w\rangle \leq 1.
\end{aligned}$$

Tracing through the computation that led us to the dual program with u_i replaced by $\varphi(u_i)$ and v_j replaced by $\varphi(v_j)$, we find the following version of the dual program:

$$\begin{aligned}
&\text{minimize} \quad \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} \mathbf{K} \begin{pmatrix}\lambda\\ \mu\end{pmatrix}\\
&\text{subject to}\\
&\qquad \sum_{i=1}^{p} \lambda_i = \sum_{j=1}^{q} \mu_j = \frac{1}{2}\\
&\qquad 0 \leq \lambda_i \leq K, \quad i = 1, \ldots, p\\
&\qquad 0 \leq \mu_j \leq K, \quad j = 1, \ldots, q,
\end{aligned}$$

where $\mathbf{K}$ is the $\ell \times \ell$ kernel symmetric matrix (with $\ell = p + q$) given by

$$\mathbf{K}_{ij} = \begin{cases} \kappa(u_i, u_j) & 1 \leq i \leq p,\ 1 \leq j \leq q\\ -\kappa(u_i, v_{j-p}) & 1 \leq i \leq p,\ p+1 \leq j \leq p+q\\ -\kappa(v_{i-p}, u_j) & p+1 \leq i \leq p+q,\ 1 \leq j \leq p\\ \kappa(v_{i-p}, v_{j-q}) & p+1 \leq i \leq p+q,\ p+1 \leq j \leq p+q. \end{cases}$$

We also find that

$$w = \frac{\displaystyle\sum_{i=1}^{p} \lambda_i \varphi(u_i) - \sum_{j=1}^{q} \mu_j \varphi(v_j)}{\left(\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} K \begin{pmatrix}\lambda\\ \mu\end{pmatrix}\right)^{1/2}}.$$

Under the Standard Margin Hypothesis, there is some index i_0 such that $0 < \lambda_{i_0} < K$ and there is some index j_0 such that $0 < \mu_{j_0} < K$, and we obtain the value of b and δ as

$$\begin{aligned}
b &= \frac{1}{2}(\langle w, \varphi(u_{i_0}\rangle + \langle w, \varphi(v_{j_0})\rangle)\\
\delta &= \frac{1}{2}(\langle w, \varphi(u_{i_0})\rangle - \langle w, \varphi(v_{j_0})\rangle).
\end{aligned}$$

Using the above value for w, we obtain

$$b = \frac{\sum_{i=1}^{p} \lambda_i(\kappa(u_i, u_{i_0}) + \kappa(u_i, v_{j_0})) - \sum_{j=1}^{q} \mu_j(\kappa(v_j, u_{i_0}) + \kappa(v_j, v_{j_0}))}{2\left(\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} K \begin{pmatrix}\lambda\\ \mu\end{pmatrix}\right)^{1/2}}.$$

It follows that the classification function

$$f(x) = \mathrm{sgn}(\langle w, \varphi(x)\rangle - b)$$

is given by

$$f(x) = \operatorname{sgn}\Bigg(\sum_{i=1}^{p} \lambda_i(2\kappa(u_i, x) - \kappa(u_i, u_{i_0}) - \kappa(u_i, v_{j_0})) \\ - \sum_{j=1}^{q} \mu_j(2\kappa(v_j, x) - \kappa(v_j, u_{i_0}) - \kappa(v_j, v_{j_0}))\Bigg),$$

which is solely expressed in terms of the kernel κ.

Kernel methods for SVM are discussed in Schölkopf and Smola [85] and Shawe–Taylor and Christianini [96].

Since the constraint $w^\top w \leq 1$ causes troubles, we trade it for a different objective function in which $-\delta$ is replaced by $(1/2)\,\|w\|_2^2$. This way we are left with purely affine constraints. In the next section we discuss a generalization of Problem (SVM_{h2}) obtained by adding a linear regularizing term.

32.2 Soft Margin Support Vector Machines; (SVM_{s2})

In this section we consider the generalization of Problem (SVM_{h2}) where we minimize $(1/2)w^\top w$ by adding the "regularizing term" $K\Big(\sum_{i=1}^{p} \epsilon_i + \sum_{j=1}^{q} \xi_j,\Big)$ for some $K > 0$. Recall that the margin δ is given by $\delta = 1/\|w\|$.

Soft margin SVM (SVM_{s2}):

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}w^\top w + K\begin{pmatrix}\epsilon^\top & \xi^\top\end{pmatrix}\mathbf{1}_{p+q} \\ &\text{subject to} \\ &\qquad w^\top u_i - b \geq 1 - \epsilon_i, \quad \epsilon_i \geq 0 \qquad i = 1, \ldots, p \\ &\qquad -w^\top v_j + b \geq 1 - \xi_j, \quad \xi_j \geq 0 \qquad j = 1, \ldots, q. \end{aligned}$$

This is the classical problem discussed in all books on machine learning or pattern analysis, for instance Vapnik [110], Bishop [18], and Shawe–Taylor and Christianini [96]. The trivial solution where all variables are 0 is ruled out because of the presence of the 1 in the inequalities, but it is not clear that if (w, b, ϵ, ξ) is an optimal solution, then $w \neq 0$.

We prove that if the primal problem has an optimal solution (w, ϵ, ξ, b) with $w \neq 0$, then w is determined by any optimal solution (λ, μ) of the dual. We also prove that there is some i for which $\lambda_i > 0$ and some j for which $\mu_j > 0$. Under a mild hypothesis that we call the **Standard Margin Hypothesis**, b can be found.

If (w, ϵ, ξ, b) is an optimal solution of Problem (SVM_{s2}), then the points u_i and v_j are classified as follows:

(1) If $\epsilon_i = 0$, then the point u_i is correctly classified and is either on the margin or on the correct side of the margin (the blue side). Similarly, if $\xi_j = 0$, then the point v_j is correctly classified and is either on the margin or on the correct side of the margin (the red side). See Figure 32.1 (1).

(2) If $0 < \epsilon_i \leq 1$, then the point u_i lies inside the margin (the slab), but on the correct side of the separating hyperplane (the blue side). If $\epsilon_i = 1$, then u_i lies on the separating hyperplane. Similarly, if $0 < \xi_j \leq 1$, then the point v_j lies inside the margin (the slab), but on the correct side of the separating hyperplane (the red side). If $\xi_j = 1$, then v_j lies on the separating hyperplane. See Figure 32.1 (2).

(3) If $\epsilon_i > 1$, then the point u_i lies on the wrong side of the separating hyperplane (the red side); it is misclassified. Similarly, if $\xi_j > 1$, then the point v_j lies on the wrong side of the separating hyperplane (the blue side); it is misclassified. See Figure 32.1 (3).

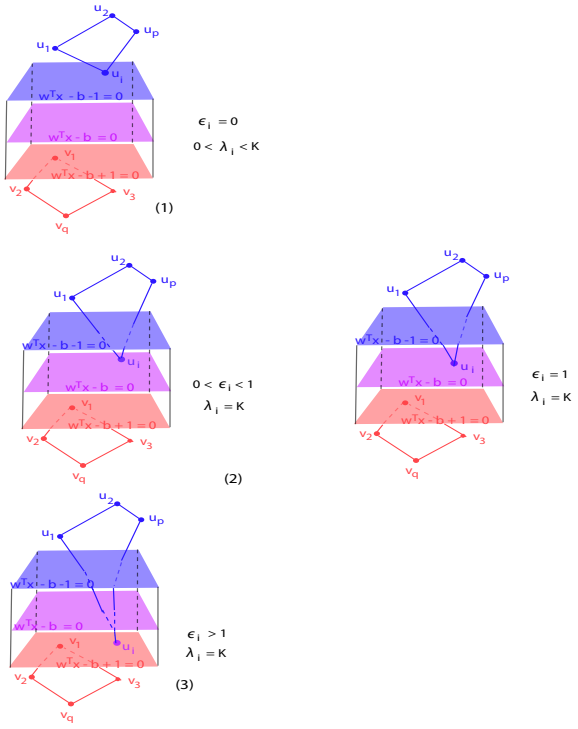


Figure 32.1: Figure (1) illustrates the case of u_i contained in the margin and occurs when $\epsilon_1 = 0$. The left illustration of Figure (2) is when u_i is inside the margin yet still on the correct side of the separating hyperplane $w^\top x - b = 0$; this occurs when $0 < \epsilon_1 < 1$. The right illustration depicts u_i on the separating hyperplane whenever $\epsilon_1 = 1$. Figure (3) illustrations a misclassification of u_i and occurs when $\epsilon_1 > 1$.

Points for which $\epsilon_i > 0$ (or $\xi_j > 0$) are called *margin-errors*; they either lie within the slab or they are misclassified.

Note that this framework is still somewhat sensitive to outliers because the penalty for misclassification is linear in ϵ and ξ.

First we write the constraints in matrix form. The $2(p+q) \times (n+p+q+1)$ matrix C is written in block form as

$$C = \begin{pmatrix} X^\top & -I_{p+q} & \mathbf{1}_p \\ & & -\mathbf{1}_q \\ 0_{p+q,n} & -I_{p+q} & 0_{p+q} \end{pmatrix},$$

and the constraints are expressed by

$$\begin{pmatrix} X^\top & -I_{p+q} & \mathbf{1}_p \\ & & -\mathbf{1}_q \\ 0_{p+q,n} & -I_{p+q} & 0_{p+q} \end{pmatrix} \begin{pmatrix} w \\ \epsilon \\ \xi \\ b \end{pmatrix} \leq \begin{pmatrix} -\mathbf{1}_{p+q} \\ 0_{p+q} \end{pmatrix}.$$

The objective function $J(w, \epsilon, \xi, b)$ is given by

$$J(w, \epsilon, \xi, b) = \frac{1}{2} w^\top w + K \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} \mathbf{1}_{p+q}.$$

The Lagrangian $L(w, \epsilon, \xi, b, \lambda, \mu, \alpha, \beta)$ with $\lambda, \alpha \in \mathbb{R}^p_+$ and with $\mu, \beta \in \mathbb{R}^q_+$ is given by

$$\begin{aligned} L(w, \epsilon, \xi, b, \lambda, \mu, \alpha, \beta) = {} & \frac{1}{2} w^\top w + K \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} \mathbf{1}_{p+q} \\ & + \begin{pmatrix} w^\top & \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} & b \end{pmatrix} C^\top \begin{pmatrix} \lambda \\ \mu \\ \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \mathbf{1}^\top_{p+q} & 0^\top_{p+q} \end{pmatrix} \begin{pmatrix} \lambda \\ \mu \\ \alpha \\ \beta \end{pmatrix}. \end{aligned}$$

Since

$$\begin{pmatrix} w^\top & \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} & b \end{pmatrix} C^\top \begin{pmatrix} \lambda \\ \mu \\ \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} w^\top & \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} & b \end{pmatrix} \begin{pmatrix} X & 0_{n,p+q} \\ -I_{p+q} & -I_{p+q} \\ \mathbf{1}^\top_p & -\mathbf{1}^\top_q \quad 0^\top_{p+q} \end{pmatrix} \begin{pmatrix} \lambda \\ \mu \\ \alpha \\ \beta \end{pmatrix}$$

we get

$$\begin{aligned} \begin{pmatrix} w^\top & \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} & b \end{pmatrix} C^\top \begin{pmatrix} \lambda \\ \mu \\ \alpha \\ \beta \end{pmatrix} &= \begin{pmatrix} w^\top & \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} & b \end{pmatrix} \begin{pmatrix} X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \\ -\begin{pmatrix} \lambda + \alpha \\ \mu + \beta \end{pmatrix} \\ \mathbf{1}^\top_p \lambda - \mathbf{1}^\top_q \mu \end{pmatrix} \\ &= w^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} - \epsilon^\top(\lambda + \alpha) - \xi^\top(\mu + \beta) + b(\mathbf{1}^\top_p \lambda - \mathbf{1}^\top_q \mu), \end{aligned}$$

and since

$$\begin{pmatrix}\mathbf{1}_{p+q}^\top & 0_{p+q}^\top\end{pmatrix}\begin{pmatrix}\lambda\\ \mu\\ \alpha\\ \beta\end{pmatrix} = \mathbf{1}_{p+q}^\top\begin{pmatrix}\lambda\\ \mu\end{pmatrix} = \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\mathbf{1}_{p+q},$$

the Lagrangian can be rewritten as

$$\begin{aligned}L(w,\epsilon,\xi,b,\lambda,\mu,\alpha,\beta) = \frac{1}{2}w^\top w + w^\top X\begin{pmatrix}\lambda\\ \mu\end{pmatrix} + \epsilon^\top(K\mathbf{1}_p - (\lambda+\alpha)) + \xi^\top(K\mathbf{1}_q - (\mu+\beta))\\ + b(\mathbf{1}_p^\top\lambda - \mathbf{1}_q^\top\mu) + \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\mathbf{1}_{p+q}.\end{aligned}$$

To find the dual function $G(\lambda,\mu,\alpha,\beta)$ we minimize $L(w,\epsilon,\xi,b,\lambda,\mu,\alpha,\beta)$ with respect to w,ϵ,ξ and b. Since the Lagrangian is convex and $(w,\epsilon,\xi,b)\in\mathbb{R}^n\times\mathbb{R}^p\times\mathbb{R}^q\times\mathbb{R}$, a convex open set, by Theorem 20.11, the Lagrangian has a minimum in (w,ϵ,ξ,b) iff $\nabla L_{w,\epsilon,\xi,b} = 0$, so we compute its gradient with respect to w,ϵ,ξ and b and we get

$$\nabla L_{w,\epsilon,\xi,b} = \begin{pmatrix} w + X\begin{pmatrix}\lambda\\ \mu\end{pmatrix}\\ K\mathbf{1}_p - (\lambda+\alpha)\\ K\mathbf{1}_q - (\mu+\beta)\\ \mathbf{1}_p^\top\lambda - \mathbf{1}_q^\top\mu\end{pmatrix}.$$

By setting $\nabla L_{w,\epsilon,\xi,b} = 0$ we get the equations

$$w = -X\begin{pmatrix}\lambda\\ \mu\end{pmatrix} \tag{$*_w$}$$

and

$$\begin{aligned}\lambda+\alpha &= K\mathbf{1}_p\\ \mu+\beta &= K\mathbf{1}_q\\ \mathbf{1}_p^\top\lambda &= \mathbf{1}_q^\top\mu.\end{aligned}$$

The first and the fourth equation are identical to the equations $(*_1)$ and $(*_2)$ that we obtained in Example 30.8. Since $\lambda,\mu,\alpha,\beta\geq 0$, the second and the third equation are equivalent to the box constraints

$$0\leq\lambda_i,\mu_j\leq K,\quad i=1,\ldots,p,\ j=1,\ldots,q.$$

Using the equations that we just derived, after simplifications we get

$$G(\lambda,\mu,\alpha,\beta) = -\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}X^\top X\begin{pmatrix}\lambda\\ \mu\end{pmatrix} + \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\mathbf{1}_{p+q},$$

which is independent of α and β and is identical to the dual function obtained in $(*_4)$ of Example 30.8. To be perfectly rigorous,

$$G(\lambda,\mu) = \begin{cases} -\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda\\ \mu\end{pmatrix} + \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\mathbf{1}_{p+q} & \text{if } \begin{cases} \sum_{i=1}^p \lambda_i = \sum_{j=1}^q \mu_j \\ 0 \le \lambda_i \le K,\ i = 1,\ldots,p \\ 0 \le \mu_j \le K,\ j = 1,\ldots,q \end{cases} \\ -\infty & \text{otherwise.} \end{cases}$$

As in Example 30.8, the the dual program can be formulated as

$$\begin{aligned} \text{maximize} \quad & -\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda\\ \mu\end{pmatrix} + \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\mathbf{1}_{p+q} \\ \text{subject to} \quad & \\ & \sum_{i=1}^p \lambda_i = \sum_{j=1}^q \mu_j \\ & 0 \le \lambda_i \le K, \quad i = 1,\ldots,p \\ & 0 \le \mu_j \le K, \quad j = 1,\ldots,q, \end{aligned}$$

or equivalently

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda\\ \mu\end{pmatrix} - \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\mathbf{1}_{p+q} \\ \text{subject to} \quad & \\ & \sum_{i=1}^p \lambda_i = \sum_{j=1}^q \mu_j \\ & 0 \le \lambda_i \le K, \quad i = 1,\ldots,p \\ & 0 \le \mu_j \le K, \quad j = 1,\ldots,q. \end{aligned}$$

The dual program is solved by making use of numerical procedures based on gradient descent. If the primal problem is solvable, this yields solutions for λ and μ.

Remark: The hard margin Problem (SVM_{h2}) corresponds to the special case of Problem (SVM_{s2}) in which $\epsilon = 0$, $\xi = 0$, and $K = +\infty$. Indeed, in Problem (SVM_{h2}) the terms involving ϵ and ξ are missing from the Lagrangian and the effect is that the box constraints are missing; we simply have $\lambda_i \ge 0$ and $\mu_j \ge 0$.

We can use the dual program to solve the primal. Once $\lambda \ge 0, \mu \ge 0$ have been found, w is given by

$$w = \sum_{i=1}^p \lambda_i u_i - \sum_{j=1}^q \mu_j v_j.$$

The complementary slackness conditions yield a classification of the points in terms of the values of λ and μ. Indeed, we have $\epsilon_i\alpha_i = 0$ for $i = 1, \ldots, p$ and $\xi_j\beta_j = 0$ for $j = 1, \ldots, q$. Also, if $\lambda_i > 0$, then corresponding constraint is active, and similarly if $\mu_j > 0$. Since $\lambda_i + \alpha_i = K$, it follows that $\epsilon_i\alpha_i = 0$ iff $\epsilon_i(K - \lambda_i) = 0$, and since $\mu_j + \beta_j = K$, we have $\xi_j\beta_j = 0$ iff $\xi_j(K - \mu_j) = 0$. Thus if $\epsilon_i > 0$ then $\lambda_i = K$, and if $\xi_j > 0$, then $\mu_j = K$. Consequently, if $\lambda_i < K$ then $\epsilon_i = 0$ and u_i is correctly classified, and similarly if $\mu_j < K$ then $\xi_j = 0$ and v_j is correctly classified. We have the following classification:

(1) If $0 < \lambda_i < K$ then u_i is on the margin and is classified correctly. Similarly, if $0 < \mu_j < K$ then v_j is on the margin and is classified correctly.

(2) If $\lambda_i = K$, then if $\epsilon_i \leq 1$ the point u_i may be classified correctly or it lies within the margin on the correct side, but if $\epsilon_i > 1$ then it is misclassifed. Similarly, if $\mu_j = K$, then if $\xi_j \leq 1$ the point v_j may be classified correctly or it lies within the margin on the correct side, but if $\xi_j > 1$ then it is misclassifed.

(3) If $\lambda_i = 0$ then u_i is classified correctly. Similarly, if $\mu_j = 0$ then v_j is classified correctly.

If the primal has a solution $w \neq 0$, then the equation

$$w = \sum_{i=1}^{p} \lambda_i u_i - \sum_{j=1}^{q} \mu_j v_j$$

implies that either there is some index i_0 such that $\lambda_{i_0} > 0$ or there is some index j_0 such that $\mu_{j_0} > 0$. The constraint

$$\sum_{i=1}^{p} \lambda_i = \sum_{j=1}^{q} \mu_j$$

implies that there is some index i_0 such that $\lambda_{i_0} > 0$ and there is some index j_0 such that $\mu_{j_0} > 0$. However, a priori, nothing prevents the situation where $\lambda_i = K$ for all nonzero λ_i or $\mu_j = K$ for all nonzero μ_j. If this happens, we can rerun the optimization method with a larger value of K. Observe that the equation

$$\sum_{i=1}^{p} \lambda_i = \sum_{j=1}^{q} \mu_j$$

implies that if there is some index i_0 such that $0 < \lambda_{i_0} < K$, then there is some index j_0 such that $0 < \mu_{j_0} < K$, and vice-versa. If the following mild hypothesis holds, then b can be found.

Standard Margin Hypothesis for (SVM_{s2}). There is some index i_0 such that $0 < \lambda_{i_0} < K$ and there is some index j_0 such that $0 < \mu_{j_0} < K$. This means that some u_{i_0} is correctly classified and on the blue margin, and some v_{j_0} is correctly classified and on the red margin.

If the **Standard Margin Hypothesis** for (SVM_{s2}) holds then $\epsilon_{i_0} = 0$ and $\mu_{j_0} = 0$, and then we have the active equations

$$w^\top u_{i_0} - b = 1 \quad \text{and} \quad -w^\top v_{j_0} + b = 1,$$

and we obtain

$$b = \frac{1}{2}(w^\top u_{i_0} + w^\top v_{j_0}).$$

Remark: There is a cheap version of Problem (SVM_{s2}) which consists in dropping the term $(1/2)w^\top w$ from the objective function:

Soft margin classifier (SVM_{s2l}):

$$\begin{aligned}
&\text{minimize} \quad \sum_{i=1}^{p} \epsilon_i + \sum_{j=1}^{q} \xi_j \\
&\text{subject to} \\
&\qquad w^\top u_i - b \geq 1 - \epsilon_i, \quad \epsilon_i \geq 0 \qquad i = 1, \ldots, p \\
&\qquad -w^\top v_j + b \geq 1 - \xi_j, \quad \xi_j \geq 0 \qquad j = 1, \ldots, q.
\end{aligned}$$

The above program is a linear program that minimizes the number of misclassified points but does not care about enforcing a minimum margin. An example of its use is given in Boyd and Vandenberghe; see [22], Section 8.6.1.

The "kernelized" version of Problem (SVM_{s2}) is the following:

Soft margin kernel SVM (SVM_{s2}):

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\langle w, w\rangle + K \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} \mathbf{1}_{p+q} \\
&\text{subject to} \\
&\qquad \langle w, \varphi(u_i)\rangle - b \geq 1 - \epsilon_i, \quad \epsilon_i \geq 0 \qquad i = 1, \ldots, p \\
&\qquad -\langle w, \varphi(v_j)\rangle + b \geq 1 - \xi_j, \quad \xi_j \geq 0 \qquad j = 1, \ldots, q.
\end{aligned}$$

Redoing the computation of the dual function, we find that the dual program is given by

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \mathbf{K} \begin{pmatrix} \lambda \\ \mu \end{pmatrix} - \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \mathbf{1}_{p+q} \\
&\text{subject to} \\
&\qquad \sum_{i=1}^{p} \lambda_i = \sum_{j=1}^{q} \mu_j \\
&\qquad 0 \leq \lambda_i \leq K, \quad i = 1, \ldots, p \\
&\qquad 0 \leq \mu_j \leq K, \quad j = 1, \ldots, q,
\end{aligned}$$

where $\mathbf{K}$ is the $\ell \times \ell$ kernel symmetric matrix (with $\ell = p + q$) given at the end of Section 32.1. We also find that

$$w = \sum_{i=1}^{p} \lambda_i \varphi(u_i) - \sum_{j=1}^{q} \mu_j \varphi(v_j),$$

so

$$b = \frac{1}{2}\left(\sum_{i=1}^{p} \lambda_i(\kappa(u_i, u_{i_0}) + \kappa(u_i, v_{j_0})) - \sum_{j=1}^{q} \mu_j(\kappa(v_j, u_{i_0}) + \kappa(v_j, v_{j_0}))\right),$$

and the classification function

$$f(x) = \mathrm{sgn}(\langle w, \varphi(x)\rangle - b)$$

is given by

$$f(x) = \mathrm{sgn}\Bigg(\sum_{i=1}^{p} \lambda_i(2\kappa(u_i, x) - \kappa(u_i, u_{i_0}) - \kappa(u_i, v_{j_0})) - \sum_{j=1}^{q} \mu_j(2\kappa(v_j, x) - \kappa(v_j, u_{i_0}) - \kappa(v_j, v_{j_0}))\Bigg).$$

32.3 Soft Margin Support Vector Machines; ($\mathrm{SVM}_{s2'}$)

In this section we consider a generalization of Problem (SVM_{s2}) for a version of the soft margin SVM coming from Problem (SVM_{h2}), by adding an extra degree of freedom, namely instead of the margin $\delta = 1/\|w\|$, we use the margin $\delta = \eta/\|w\|$ where η is some positive constant that we wish to maximize. To do so, we add a term $-K_m\eta$ to the objective function $(1/2)w^\top w$ as well as the "regularizing term" $K_s\left(\sum_{i=1}^{p} \epsilon_i + \sum_{j=1}^{q} \xi_j\right)$ whose purpose is to make ϵ and ξ sparse, where $K_m > 0$ and $K_s > 0$ are fixed constants that can be adjusted to determine the influence of η and the regularizing term.

Soft margin SVM ($\mathrm{SVM}_{s2'}$):

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}w^\top w - K_m\eta + K_s\begin{pmatrix}\epsilon^\top & \xi^\top\end{pmatrix}\mathbf{1}_{p+q}\\
&\text{subject to}\\
&\qquad w^\top u_i - b \geq \eta - \epsilon_i, \quad \epsilon_i \geq 0 \qquad i = 1, \ldots, p\\
&\qquad -w^\top v_j + b \geq \eta - \xi_j, \quad \xi_j \geq 0 \qquad j = 1, \ldots, q\\
&\qquad \eta \geq 0.
\end{aligned}$$

This version of the SVM problem was first discussed in Schölkopf, Smola, Williamson, and Bartlett [87] under the name of *ν-SVC* (or *ν-SVM*), and also used in Schölkopf, Platt,

Shawe–Taylor, and Smola [86]. The ν-SVC method is also presented in Schölkopf and Smola [85] (which contains much more). The difference between the ν-SVC method and the method presented in Section 32.2, sometimes called the C-SVM method, was thoroughly investigated by Chan and Lin [27].

For this problem, it is no longer clear that if $(w, \eta, b, \epsilon, \xi)$ is an optimal solution, then $w \neq 0$ and $\eta > 0$. In fact, if the sets of points are not linearly separable and if K_s is chosen too big, Problem ($\text{SVM}_{s2'}$) may fail to have an optimal solution.

We show that in order for the problem to have a solution we must pick K_m and K_s so that

$$K_m \leq \min\{2pK_s, 2qK_s\}.$$

If we define ν by

$$\nu = \frac{K_m}{(p+q)K_s},$$

then $K_m \leq \min\{2pK_s, 2qK_s\}$ is equivalent to

$$\nu \leq \min\left\{\frac{2p}{p+q}, \frac{2q}{p+q}\right\} \leq 1.$$

The reason for introducing ν is that $\nu(p+q)/2$ can be interpreted as a the maximum number of points failing to achieve the margin η. If the sets $\{u_i\}$ and $\{v_j\}$ are not linearly separable, then we must pick ν so that $\nu \geq 2/(p+q)$ for the method to have an optiomal solution. If $\nu < 3/(p+q)$ and at least three points are misclassified then we have some interesting guarantees; see Proposition 32.5 and Proposition 32.6.

The objective function of our problem is convex and the constraints are affine. Consequently, by Theorem 30.14(2) *if* the primal problem ($\text{SVM}_{s2'}$) has an optimal solution, *then* the dual problem has a solution too, and the duality gap is zero. This does not immediately imply that an optimal solution of the dual yields an optimal solution of the primal because the hypotheses of Theorem 30.14(1) fail to hold.

We show that if the primal problem has an optimal solution $(w, \eta, \epsilon, \xi, b)$ with $w \neq 0$, then any optimal solution of the dual problem determines λ and μ, which in turn determine w *via* the equation

$$w = -X\begin{pmatrix}\lambda\\ \mu\end{pmatrix} = \sum_{i=1}^{p}\lambda_i u_i - \sum_{j=1}^{q}\mu_j v_j, \tag{$*_w$}$$

and $\eta \geq 0$.

It remains to determine b, η, ϵ and ξ. The solution of the dual does not determine b, η, ϵ, ξ directly, and we are not aware of necessary and sufficient conditions that ensure that they can be determined. The best we can do is to use the KKT conditions.

The simplest sufficient condition is what we call the

Standard Margin Hypothesis for ($\mathrm{SVM}_{s2'}$): There is some i_0 such that $0 < \lambda_{i_0} < K_s$ and there is some μ_{j_0} such that $0 < \mu_{j_0} < K_s$. This means that some u_{i_0} is correctly classified and on the blue margin, and some v_{j_0} is correctly classified and on the red margin.

In this case, then by complementary slackness it can be shown that $\epsilon_{i_0} = 0$, $\xi_{i_0} = 0$, and the corresponding inequalities are active, that is we have the equations

$$w^\top u_{i_0} - b = \eta, \quad -w^\top v_{j_0} + b = \eta,$$

so we can solve for b and η. Then, since by complementary slackness if $\epsilon_i > 0$ then $\lambda_i = K_s$ and if $\xi_j > 0$ then $\mu_j = K_s$, all inequalities corresponding to such $\epsilon_i > 0$ and $\mu_j > 0$ are active, and we can solve for ϵ_i and ξ_j.

If $2/(p+q) \leq \nu < 3/(p+q)$ and at least three points are misclassified then we can guarantee that either there is some i_0 such that the constraint $w^\top u_{i_0} - b = \eta$ is active or there is some j_0 such that the constraint $-w^\top v_{j_0} + b = \eta$ is active.

If $(w, \eta, \epsilon, \xi, b)$ is an optimal solution of Problem ($\mathrm{SVM}_{s2'}$) with $w \neq 0$, then the points u_i and v_j are classified as follows:

(1) If $\epsilon_i = 0$, then the point u_i is correctly classified and is either on the blue margin (the hyperplane $H_{w,b+\eta}$ of equation $w^\top x = b + \eta$) or on the correct side of the blue margin (the blue side). Similarly, if $\xi_j = 0$, then the point v_j is correctly classified and is either on the red margin (the hyperplane $H_{w,b-\eta}$ of equation $w^\top x = b - \eta$) or on the correct side of the red margin (the red side).

(2) If $0 < \epsilon_i \leq \eta$, then the point u_i lies inside the margin (the slab), but on the correct side of the separating hyperplane (the blue side). If $\epsilon_i = \eta$, then u_i lies on the separating hyperplane. Similarly, if $0 < \xi_j \leq \eta$, then the point v_j lies inside the margin (the slab), but on the correct side of the separating hyperplane (the red side). If $\xi_j = \eta$, then v_j lies on the separating hyperplane.

(3) If $\epsilon_i > \eta$, then the point u_i lies on the wrong side of the separating hyperplane (the red side); it is misclassified. Similarly, if $\xi_j > \eta$, then the point v_j lies on the wrong side of the separating hyperplane (the blue side); it is misclassified.

Points for which $\epsilon_i > 0$ (or $\xi_j > 0$) are called *margin-errors*; they either lie within the slab or they are misclassified.

The linear constraints are given by the $(2(p+q)+1) \times (n+p+q+2)$ matrix given in block form by

$$C = \begin{pmatrix} X^\top & -I_{p+q} & \begin{matrix} \mathbf{1}_p \\ -\mathbf{1}_q \end{matrix} & \mathbf{1}_{p+q} \\ 0_{p+q,n} & -I_{p+q} & 0_{p+q} & 0_{p+q} \\ 0_n^\top & 0_{p+q}^\top & 0 & -1 \end{pmatrix},$$

and the linear constraints are expressed by

$$\begin{pmatrix} X^\top & -I_{p+q} & \mathbf{1}_p & \mathbf{1}_{p+q} \\ & & -\mathbf{1}_q & \\ 0_{p+q,n} & -I_{p+q} & 0_{p+q} & 0_{p+q} \\ 0_n^\top & 0_{p+q}^\top & 0 & -1 \end{pmatrix} \begin{pmatrix} w \\ \epsilon \\ \xi \\ b \\ \eta \end{pmatrix} \leq \begin{pmatrix} 0_{p+q} \\ 0_{p+q} \\ 0 \end{pmatrix}.$$

The objective function is given by

$$J(w, \epsilon, \xi, b, \eta) = \frac{1}{2} w^\top w - K_m \eta + K_s \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} \mathbf{1}_{p+q}.$$

The Lagrangian $L(w, \epsilon, \xi, b, \eta, \lambda, \mu, \alpha, \beta, \gamma)$ with $\lambda, \alpha \in \mathbb{R}^p_+$, $\mu, \beta \in \mathbb{R}^q_+$, and $\gamma \in \mathbb{R}_+$ is given by

$$\begin{aligned} L(w, \epsilon, \xi, b, \eta, \lambda, \mu, \alpha, \beta, \gamma) = \frac{1}{2} w^\top w - K_m \eta + K_s \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} \mathbf{1}_{p+q} \\ + \begin{pmatrix} w^\top & \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} & b & \eta \end{pmatrix} C^\top \begin{pmatrix} \lambda \\ \mu \\ \alpha \\ \beta \\ \gamma \end{pmatrix}. \end{aligned}$$

Since

$$\begin{aligned} \begin{pmatrix} w^\top & \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} & b & \eta \end{pmatrix} C^\top \begin{pmatrix} \lambda \\ \mu \\ \alpha \\ \beta \\ \gamma \end{pmatrix} = w^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} - \epsilon^\top(\lambda + \alpha) - \xi^\top(\mu + \beta) + b(\mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu) \\ + \eta(\mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu) - \gamma\eta, \end{aligned}$$

the Lagrangian can be written as

$$\begin{aligned} L(w, \epsilon, \xi, b, \eta, \lambda, \mu, \alpha, \beta, \gamma) &= \frac{1}{2} w^\top w - K_m \eta + K_s(\epsilon^\top \mathbf{1}_p + \xi^\top \mathbf{1}_q) + w^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} - \epsilon^\top(\lambda + \alpha) \\ &\quad - \xi^\top(\mu + \beta) + b(\mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu) + \eta(\mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu) - \gamma\eta, \\ &= \frac{1}{2} w^\top w + w^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + (\mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu - K_m - \gamma)\eta \\ &\quad + \epsilon^\top(K_s \mathbf{1}_p - (\lambda + \alpha)) + \xi^\top(K_s \mathbf{1}_q - (\mu + \beta)) + b(\mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu). \end{aligned}$$

To find the dual function $G(\lambda, \mu, \alpha, \beta, \gamma)$ we minimize $L(w, \epsilon, \xi, b, \eta, \lambda, \mu, \alpha, \beta, \gamma)$ with respect to w, ϵ, ξ, b, and η. Since the Lagrangian is convex and $(w, \epsilon, \xi, b, \eta) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^q \times$

$\mathbb{R} \times \mathbb{R}$, a convex open set, by Theorem 20.11, the Lagrangian has a minimum in $(w, \epsilon, \xi, b, \eta)$ iff $\nabla L_{w,\epsilon,\xi,b,\eta} = 0$, so we compute its gradient with respect to $w, \epsilon, \xi, b, \eta$ and we get

$$\nabla L_{w,\epsilon,\xi,b,\eta} = \begin{pmatrix} X\begin{pmatrix}\lambda\\ \mu\end{pmatrix} + w \\ K_s\mathbf{1}_p - (\lambda + \alpha) \\ K_s\mathbf{1}_q - (\mu + \beta) \\ \mathbf{1}_p^\top\lambda - \mathbf{1}_q^\top\mu \\ \mathbf{1}_p^\top\lambda + \mathbf{1}_q^\top\mu - K_m - \gamma \end{pmatrix}.$$

By setting $\nabla L_{w,\epsilon,\xi,b,\eta} = 0$ we get the equations

$$w = -X\begin{pmatrix}\lambda\\ \mu\end{pmatrix} \tag{$*_w$}$$

$$\begin{aligned} \lambda + \alpha &= K_s\mathbf{1}_p \\ \mu + \beta &= K_s\mathbf{1}_q \\ \mathbf{1}_p^\top\lambda &= \mathbf{1}_q^\top\mu, \end{aligned}$$

and

$$\mathbf{1}_p^\top\lambda + \mathbf{1}_q^\top\mu = K_m + \gamma. \tag{$*_\gamma$}$$

The second and third equations are equivalent to the box constraints

$$0 \leq \lambda_i, \mu_j \leq K_s, \quad i = 1, \ldots, p,\ j = 1, \ldots, q,$$

and since $\gamma \geq 0$ equation $(*_\gamma)$ is equivalent to

$$\mathbf{1}_p^\top\lambda + \mathbf{1}_q^\top\mu \geq K_m.$$

Plugging back w from $(*_w)$ into the Lagrangian, after simplifications we get

$$\begin{aligned} G(\lambda, \mu, \alpha, \beta) &= \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X\begin{pmatrix}\lambda\\ \mu\end{pmatrix} - \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X\begin{pmatrix}\lambda\\ \mu\end{pmatrix} \\ &= -\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X\begin{pmatrix}\lambda\\ \mu\end{pmatrix}, \end{aligned}$$

so the dual function is independent of α, β and is given by

$$G(\lambda, \mu) = -\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X\begin{pmatrix}\lambda\\ \mu\end{pmatrix}.$$

The dual program is given by

$$\begin{aligned}
\text{maximize} \quad & -\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda\\ \mu\end{pmatrix}\\
\text{subject to}\\
& \sum_{i=1}^{p}\lambda_i = \sum_{j=1}^{q}\mu_j\\
& \sum_{i=1}^{p}\lambda_i + \sum_{j=1}^{q}\mu_j \geq K_m\\
& 0 \leq \lambda_i \leq K_s, \quad i = 1,\ldots,p\\
& 0 \leq \mu_j \leq K_s, \quad j = 1,\ldots,q.
\end{aligned}$$

Finally, the dual program is equivalent to the following minimization program:

$$\begin{aligned}
\text{minimize} \quad & \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda\\ \mu\end{pmatrix}\\
\text{subject to}\\
& \sum_{i=1}^{p}\lambda_i = \sum_{j=1}^{q}\mu_j\\
& \sum_{i=1}^{p}\lambda_i + \sum_{j=1}^{q}\mu_j \geq K_m\\
& 0 \leq \lambda_i \leq K_s, \quad i = 1,\ldots,p\\
& 0 \leq \mu_j \leq K_s, \quad j = 1,\ldots,q.
\end{aligned}$$

The dual program is solved by making use of numerical procedures based on gradient descent. If the primal problem is solvable, this yields solutions for λ and μ. Once a solution for λ and μ is obtained, we have

$$w = -X\begin{pmatrix}\lambda\\ \mu\end{pmatrix} = \sum_{i=1}^{p}\lambda_i u_i - \sum_{j=1}^{q}\mu_j v_j.$$

As we said earlier, the hypotheses of Theorem 30.14(2) hold, so *if* the primal problem ($\text{SVM}_{s2'}$) has an optimal solution with $w \neq 0$, *then* the dual problem has a solution too, and the duality gap is zero. Therefore, for optimal solutions we have

$$L(w,\epsilon,\xi,b,\eta,\lambda,\mu,\alpha,\beta,\gamma) = G(\lambda,\mu,\alpha,\beta,\gamma),$$

which means that

$$\frac{1}{2}w^\top w - K_m\eta + K_s\Big(\sum_{i=1}^{p}\epsilon_i + \sum_{j=1}^{q}\xi_j\Big) = -\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda\\ \mu\end{pmatrix},$$

and since

$$w = -X\begin{pmatrix}\lambda\\ \mu\end{pmatrix}$$

we get

$$\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X\begin{pmatrix}\lambda\\ \mu\end{pmatrix} - K_m\eta + K_s\Big(\sum_{i=1}^{p}\epsilon_i + \sum_{j=1}^{q}\xi_j\Big) = -\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X\begin{pmatrix}\lambda\\ \mu\end{pmatrix},$$

which yields

$$\eta = \frac{K_s}{K_m}\Big(\sum_{i=1}^{p}\epsilon_i + \sum_{j=1}^{q}\xi_j\Big) + \frac{1}{K_m}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X\begin{pmatrix}\lambda\\ \mu\end{pmatrix}. \tag{$*$}$$

Therefore, $\eta \geq 0$.

Remarks:

(1) The objective function of Problem ($\mathrm{SVM}_{s2'}$) is half of the objective function of Problem (SVM_{s1}), but some of the constraints are different. However, the major advantage of Problem ($\mathrm{SVM}_{s2'}$) is that w is always determined.

(2) Since we proved that if the primal problem ($\mathrm{SVM}_{s2'}$) has an optimal solution with $w \neq 0$ then $\eta \geq 0$, one might wonder why the constraint $\eta \geq 0$ was included. If we delete this constraint, it is easy to see that the only difference is that instead of the equation

$$\mathbf{1}_p^\top\lambda + \mathbf{1}_q^\top\mu = K_m + \gamma$$

we obtain the equation

$$\mathbf{1}_p^\top\lambda + \mathbf{1}_q^\top\mu = K_m.$$

Since the equation

$$\mathbf{1}_p^\top\lambda = \mathbf{1}_q^\top\mu$$

holds, in the first case we obtain

$$\mathbf{1}_p^\top\lambda = \mathbf{1}_q^\top\mu = \frac{K_m}{2} + \frac{\gamma}{2} \tag{$*_1$}$$

and in the second case, we obtain

$$\mathbf{1}_p^\top\lambda = \mathbf{1}_q^\top\mu = \frac{K_m}{2}. \tag{$*_2$}$$

If $\eta > 0$, then by complementary slackness $\gamma = 0$, in which case ($*_1$) and ($*_2$) are equivalent. But if $\eta = 0$, then γ could be strictly positive.

It not clear that the option to include the constraint $\eta \geq 0$ in the primal is advantageous, except perhaps for the fact that in the dual program the equation and inequality

$$\begin{aligned} \mathbf{1}_p^\top \lambda &= \mathbf{1}_q^\top \mu \\ \mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu &\geq K_m \end{aligned}$$

are included rather than the equations

$$\mathbf{1}_p^\top \lambda = \mathbf{1}_q^\top \mu = \frac{K_m}{2}.$$

Perhaps the use of an inequality makes it easier to solve the dual. To settle this issue it seems that we need to run practical solvers on some test data.

Returning to Problem ($\text{SVM}_{s2'}$), the complementary slackness conditions yield a classification of the points in terms of the values of λ and μ. Indeed, we have $\epsilon_i \alpha_i = 0$ for $i = 1, \ldots, p$ and $\xi_j \beta_j = 0$ for $j = 1, \ldots, q$. Also, if $\lambda_i > 0$, then the corresponding constraint is active, and similarly if $\mu_j > 0$. Since $\lambda_i + \alpha_i = K_s$, it follows that $\epsilon_i \alpha_i = 0$ iff $\epsilon_i(K_s - \lambda_i) = 0$, and since $\mu_j + \beta_j = K_s$, we have $\xi_j \beta_j = 0$ iff $\xi_j(K_s - \mu_j) = 0$. Thus if $\epsilon_i > 0$ then $\lambda_i = K_s$, and if $\xi_j > 0$, then $\mu_j = K_s$. Consequently, if $\lambda_i < K_s$ then $\epsilon_i = 0$ and u_i is correctly classified, and similarly if $\mu_j < K_s$ then $\xi_j = 0$ and v_j is correctly classified.

In addition to the constraints

$$0 \leq \lambda_i \leq K_s, \quad 0 \leq \mu_j \leq K_s,$$

we also have the constraints

$$\begin{aligned} \sum_{i=1}^{p} \lambda_i &= \sum_{j=1}^{q} \mu_j \\ \sum_{i=1}^{p} \lambda_i + \sum_{j=1}^{q} \mu_j &\geq K_m \end{aligned}$$

which imply that

$$\sum_{i=1}^{p} \lambda_i \geq \frac{K_m}{2} \quad \text{and} \quad \sum_{j=1}^{q} \mu_j \geq \frac{K_m}{2}. \tag{$\dagger$}$$

Since λ, μ are all nonnegative, if $\lambda_i = K_s$ for all i and if $\mu_j = K_s$ for all j then

$$\frac{K_m}{2} \leq \sum_{i=1}^{p} \lambda_i \leq pK_s$$

and

$$\frac{K_m}{2} \leq \sum_{j=1}^{q} \mu_j \leq qK_s,$$

so these constraints are not satisfied unless $K_m \leq \min\{2pK_s, 2qK_s\}$, so we assume that $K_m \leq \min\{2pK_s, 2qK_s\}$. The equations in (†) also imply that there is some i_0 such that $\lambda_{i_0} > 0$ and some j_0 such that $\mu_{j_0} > 0$.

We have the following classification (recall that $\eta > 0$):

(1) If $0 < \lambda_i < K_s$ then u_i is on the margin and is classified correctly. Similarly, if $0 < \mu_j < K_s$ then v_j is on the margin and is classified correctly.

(2) If $\lambda_i = K_s$, then we can't say more without looking at ϵ_i. If $\epsilon_i = 0$ then the point u_i is on the margin and is classified correctly, and if $0 < \epsilon_i \leq \eta$, then u_i lies within the margin on the correct side, but if $\epsilon_i > \eta$ then it is misclassifed. Similarly, if $\mu_j = K_s$, then we can't say more without looking at ξ_j. If $\xi_j = 0$ then the point v_j is on the margin and is classified correctly, and if $0 < \xi_j \leq \eta$, then v_j lies within the margin on the correct side, but if $\xi_j > \eta$ then it is misclassifed.

(3) If $\lambda_i = 0$ then u_i is classified correctly. Similarly, if $\mu_j = 0$ then v_j is classified correctly. There is no way to tell whether u_i is on the margin or not, and similarly for v_j.

We find it convenient to define $\nu > 0$ such that

$$K_m = (p+q)K_s\,\nu,$$

that is

$$\nu = \frac{K_m}{(p+q)K_s},$$

so that the objective function $J(w, \epsilon, \xi, b, \eta)$ is given by

$$J(w, \epsilon, \xi, b, \eta) = \frac{1}{2}w^\top w + K\left(-\nu\eta + \frac{1}{p+q}\begin{pmatrix}\epsilon^\top & \xi^\top\end{pmatrix}\mathbf{1}_{p+q}\right),$$

with $K = (p+q)K_s$, and so $K_m = K\nu$ and $K_s = K/(p+q)$.

Observe that the condition $K_m \leq \min\{2pK_s, 2qK_s\}$ is equivalent to

$$\nu \leq \min\left\{\frac{2p}{p+q}, \frac{2q}{p+q}\right\} \leq 1,$$

and the condition $K_s \leq K_m/2$ is equivalent to

$$\frac{2}{p+q} \leq \nu.$$

Since we obtain an equivalent problem by rescaling by a common positive factor, it is convenient to normalize K_s as

$$K_s = \frac{1}{p+q},$$

in which case $K_m = \nu$. This method is called the *ν-support vector machine*.

Under the **Standard Margin Hypothesis** for ($\text{SVM}_{s2'}$), there is some i_0 such that $0 < \lambda_{i_0} < K_s$ and some j_0 such that $0 < \mu_{j_0} < K_s$, and by the complementary slackness conditions $\epsilon_{i_0} = 0$ and $\xi_{j_0} = 0$, so we have the two active constraints

$$w^\top u_{i_0} - b = \eta, \quad -w^\top v_{j_0} + b = \eta,$$

and we can solve for b and η and we get

$$b = \frac{w^\top u_{i_0} + w^\top v_{j_0}}{2}$$
$$\eta = \frac{w^\top u_{i_0} - w^\top v_{j_0}}{2}.$$

The equations (†) and the box inequalities

$$0 \leq \lambda_i \leq K_s, \quad 0 \leq \mu_j \leq K_s$$

also imply the following facts:

Proposition 32.1. *If Problem* ($\text{SVM}_{s2'}$) *has an optimal solution with $w \neq 0$ and $\eta > 0$, then the following facts hold:*

(1) At most $\nu(p+q)/2$ points u_i fail to achieve the margin η, and at most $\nu(p+q)/2$ points v_j fail to achieve the margin η.

(2) At least $\nu(p+q)/2$ points u_i have margin at most η, and at least $\nu(q+q)/2$ points have margin at most η.

Proof. (1) Recall that for an optimal solution with $w \neq 0$ and $\eta > 0$, we have $\gamma = 0$, so by ($*_\gamma$) we have the equations

$$\sum_{i=1}^{p} \lambda_i = \frac{K_m}{2} \quad \text{and} \quad \sum_{j=1}^{q} \mu_j = \frac{K_m}{2}.$$

If u_i fails to achieve the margin η, then $\epsilon_i > 0$, and by complementary slackness $\lambda_i = K_s = K_m/(\nu(p+q))$, so if there are p_f such points then

$$\frac{K_m}{2} = \sum_{i=1}^{p} \lambda_i \geq \frac{K_m p_f}{\nu(p+q)},$$

so

$$p_f \leq \frac{\nu(p+q)}{2}.$$

A similar reasoning applies if v_j fails to achieve the margin η with $\sum_{i=1}^{p} \lambda_i$ replaced by $\sum_{j=1}^{q} \mu_j$ (and where q_f is the number of points v_j that fail to achieve the margin η).

(2) A point u_i has margin at most η iff $\lambda_i > 0$. If

$$I_m = \{i \in \{1, \ldots, p\} \mid \lambda_i > 0\} \quad \text{and} \quad p_m = |I_m|,$$

then

$$\frac{K_m}{2} = \sum_{i=1}^{p} \lambda_i = \sum_{i \in I_m} \lambda_i,$$

and since $\lambda_i \leq K_s = K_m/(\nu(p+q))$, we have

$$\frac{K_m}{2} = \sum_{i \in I_m} \lambda_i \leq \frac{K_m p_m}{\nu(p+q)},$$

which yields

$$p_m \geq \frac{\nu(p+q)}{2}.$$

A similar reasoning applies if a point v_j has margin at most η. □

Note that if ν is chosen so that $\nu < 2/(p+q)$, then $\nu(p+q)/2 < 1$, which means that none of the data points are misclassified; in other words, the u_is and v_js are linearly separable. Thus again, we see that if the u_is and v_js are not linearly separable we must pick ν such that $2/(p+q) \leq \nu \leq \min\{2p/(p+q), 2q/(p+q)\}$ for the method to succeed.

The following proposition clarifies the role of the constant ν in establishing the trade-off between the width of the margin and the number of margin-error points. In particular, it shows that if Problem ($\mathrm{SVM}_{s2'}$) has an optimal solution with $w \neq 0$ and if $\nu < \min\{2p/(p+q), 2q/(p+q)\}$, then at least some u_i or some v_j is classified correctly. Obviously we have $2/(p+q) \leq \min\{2p/(p+q), 2q/(p+q)\}$.

Proposition 32.2. *Suppose $(w, b, \eta, \epsilon, \xi)$ is an optimal solution of Problem ($\mathrm{SVM}_{s2'}$) with $w \neq 0$ and $\eta > 0$, and let p_f be the number of points u_i that are misclassified ($\epsilon_i > 0$) and q_f be the number of points v_j that are misclassified ($\xi_j > 0$). If $p_f + q_f \geq 3$ and if $2/(p+q) \leq \nu < (p_f + q_f)/(p+q)$, then either there is some i such that $\epsilon_i = 0$ and the constraint $w^\top u_i - b = \eta$ is active, or there is some j such that $\xi_j = 0$ and the constraint $-w^\top v_j + b = \eta$ is active.*

Proof. (1) We may assume that $K_s = 1/(p+q)$. We proceed by contradiction. Thus we assume that for all $i \in \{1, \ldots, p\}$, if $\epsilon_i = 0$ then the constraint $w^\top u_i - b \geq \eta$ is not active, namely $w^\top u_i - b > \eta$, and for all $j \in \{1, \ldots, q\}$, if $\xi_j = 0$ then the constraint $-w^\top v_j + b \geq \eta$ is not active, namely $-w^\top v_j + b > \eta$.

Let $I = \{i \in \{1, \ldots, p\} \mid \epsilon_i > 0\}$, let $J = \{j \in \{1, \ldots, q\} \mid \xi_j > 0\}$, and let $p_f = |I|$ and $q_f = |J|$ (of course, $\eta > 0$).

Assume that $p_f + q_f \geq 3$. By complementary slackness all the constraints for which $i \in I$ and $j \in J$ are active, so our hypotheses are

$$\begin{array}{lll} w^\top u_i - b = \eta - \epsilon_i & \epsilon_i > 0 & i \in I \\ -w^\top v_j + b = \eta - \xi_j & \xi_j > 0 & j \in J \\ w^\top u_i - b > \eta & & i \notin I \\ -w^\top v_j + b > \eta & & j \notin J. \end{array}$$

For any $\theta > 0$ such that

$$\theta < \min\{\epsilon_i, \xi_j, \eta \mid i \in \{1, \ldots, p\}, j \in \{1, \ldots, q\}\},$$

we can write

$$\begin{array}{lll} w^\top u_i - b = \eta - \theta - (\epsilon_i - \theta) & \epsilon_i - \theta \geq 0 & i \in I \\ -w^\top v_j + b = \eta - \theta - (\xi_j - \theta) & \xi_j - \theta \geq 0 & j \in J \\ w^\top u_i - b > \eta - \theta & & i \notin I \\ -w^\top v_j + b > \eta - \theta & & j \notin J. \end{array}$$

The original value of the objective function is

$$\omega(0) = \frac{1}{2} w^\top w - \nu\eta + \frac{1}{p+q}\left(\sum_{i \in I} \epsilon_i + \sum_{j \in J} \xi_j\right),$$

and the new value is

$$\begin{aligned} \omega(\theta) &= \frac{1}{2} w^\top w - \nu(\eta - \theta) + \frac{1}{p+q}\left(\sum_{i \in I} (\epsilon_i - \theta) + \sum_{j \in J} (\xi_j - \theta)\right) \\ &= \frac{1}{2} w^\top w - \nu\eta + \frac{1}{p+q}\left(\sum_{i \in I} \epsilon_i + \sum_{j \in J} \xi_j\right) - \left(\frac{p_f + q_f}{p+q} - \nu\right)\theta. \end{aligned}$$

Since by hypothesis $p_f + q_f \geq 3$, if

$$\frac{2}{p+q} \leq \nu < \frac{p_f + q_f}{p+q},$$

then the term involving θ is negative so

$$\omega(\theta) < \omega(0),$$

and by the choice of θ we have $\eta - \theta > 0$, so $(w, b, \eta - \theta, \epsilon - \theta, \xi - \theta)$ is a feasible solution, contradicting the optimality of the solution $(w, b, \eta, \epsilon, \xi)$; here we write $\epsilon - \theta$ for the vector $(\epsilon_1 - \theta, \ldots, \epsilon_p - \theta)$, and similarly for $\xi - \theta$. □

Note that if $p_f + q_f = p + q$ and $\nu < \min\{2p/(p+q), 2q/(p+q)\} \leq 1$, then Proposition 32.5 yields a contradiction. Therefore $p_f + q_f < p + q$, that is, at least some u_i or some v_j is classified correctly

Remark: If the the sets $\{u_i\}$ and $\{v_j\}$ are linearly separable, then we know from Theorem 30.10 that some u_i is on the blue margin and some v_j is on the red margin.

We also have the following proposition that gives a sufficient condition implying that η and b can be found in terms of an optimal solution (λ, μ) of the dual.

Proposition 32.3. *If $(w, b, \eta, \epsilon, \xi)$ is an optimal solution of Problem ($\mathrm{SVM}_{s2'}$) with $w \neq 0$ and $\eta > 0$, and if $2/(p+q) \leq \nu < 4/(p+q)$ and $p_f, q_f \geq 2$, then η and b can always be determined from an optimal solution (λ, μ) of the dual.*

Proof. Since $p_f + q_f \geq 4$, by Proposition 32.5, either there is some i_0 such that $\epsilon_{i_0} = 0$ and the constraint $w^\top u_{i_0} - b = \eta$ is active, or there is some j_0 such that $\xi_{j_0} = 0$ and the constraint $-w^\top v_{j_0} + b = \eta$ is active As we already explained, Problem ($\mathrm{SVM}_{s2'}$) satisfies the conditions for having a zero duality gap. Therefore, for optimal solutions we have

$$L(w, \epsilon, \xi, b, \eta, \lambda, \mu, \alpha, \beta) = G(\lambda, \mu, \alpha, \beta),$$

which means that

$$\frac{1}{2} w^\top w - \nu\eta + \frac{1}{p+q}\left(\sum_{i=1}^{p} \epsilon_i + \sum_{j=1}^{q} \xi_j\right) = -\frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix},$$

and since

$$w = -X \begin{pmatrix} \lambda \\ \mu \end{pmatrix},$$

we get

$$\frac{1}{p+q}\left(\sum_{i=1}^{p} \epsilon_i + \sum_{j=1}^{q} \xi_j\right) = \nu\eta - \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix}. \tag{$*$}$$

Let $I = \{i \in \{1, \ldots, p\} \mid \epsilon_i > 0\}$ and $J = \{j \in \{1, \ldots, q\} \mid \xi_j > 0\}$. By hypothesis $|I| \geq 2$ and $|J| \geq 2$. We know that $\lambda_i = 1/(p+q)$ for all $i \in I$ and $\mu_j = 1/(p+q)$ for all $j \in J$, so the following equations are active:

$$\begin{aligned} w^\top u_i - b &= \eta - \epsilon_i && i \in I \\ -w^\top v_j + b &= \eta - \xi_j && j \in J. \end{aligned}$$

But $(*)$ can be written as

$$\frac{1}{p+q}\left(\sum_{i \in I} \epsilon_i + \sum_{j \in J} \xi_j\right) = \nu\eta - \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix}, \tag{$**$}$$

and since

$$\begin{aligned}\epsilon_i &= \eta - w^\top u_i + b && i \in I\\ \xi_j &= \eta + w^\top v_j - b && j \in J,\end{aligned}$$

by substituting in the equation $(**)$ we get

$$\left(\frac{|I|+|J|}{p+q} - \nu\right)\eta = \frac{|J|-|I|}{p+q}b + \frac{1}{p+q}w^\top\left(\sum_{i\in I} u_i - \sum_{j\in J} v_j\right) - \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda\\ \mu\end{pmatrix}.$$

We also know that either $w^\top u_{i_0} - b = \eta$ or $-w^\top v_{j_0} + b = \eta$. In the first case, $b = -\eta + w^\top u_{i_0}$, and by substituting b in the above equation we get an equation of the form

$$\left(\frac{|I|+|J|}{p+q} - \nu\right)\eta = -\frac{|J|-|I|}{p+q}\eta + T_1,$$

that is,

$$\left(\frac{2|J|}{p+q} - \nu\right)\eta = T_1.$$

In the second case $b = \eta + w^\top v_{j_0}$, and we get an equation of the form

$$\left(\frac{|I|+|J|}{p+q} - \nu\right)\eta = \frac{|J|-|I|}{p+q}\eta + T_2,$$

that is,

$$\left(\frac{2|I|}{p+q} - \nu\right)\eta = T_2.$$

We need to choose ν such that $2|I|/(p+q)| - \nu \neq 0$ and $2|J|/(p+q) - \nu \neq 0$. Since $|I| \geq 2$ and $|J| \geq 2$, this will be the case if $\nu < 4/(p+q)$. If this condition is satisfied we can solve for η, and then we find b from either $b = -\eta + w^\top u_{i_0}$ or $b = \eta + w^\top v_{j_0}$. □

Remark: If the the sets $\{u_i\}$ and $\{v_j\}$ are linearly separable, then we know from Theorem 30.10 that some u_i is on the blue margin and some v_j is on the red margin, so b and δ can be determined. Although we can ensure that some u_i is classified correctly or some v_j is classified correctly, it does not seem possible to prove that the corresponding constraints are active without additional hypotheses (such as $p_f + q_f \geq 3$).

Among its advantages, the support vector machinery is condusive to finding interesting statistical bounds in terms of the *VC dimension*, a notion invented by Vapnik and Chernovenkis. We will not go into this here and instead refer the reader to Vapnik [110] (especially, Chapter 4 and Chapters 9-13).

The "kernelized" version of Problem ($\text{SVM}_{s2'}$) is the following:

Soft margin kernel SVM ($\mathrm{SVM}_{s2'}$):

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\langle w, w\rangle - \nu\eta + \frac{1}{p+q}\begin{pmatrix}\epsilon^\top & \xi^\top\end{pmatrix}\mathbf{1}_{p+q}\\
&\text{subject to}\\
&\qquad \langle w, \varphi(u_i)\rangle - b \geq \eta - \epsilon_i, \quad \epsilon_i \geq 0 \qquad i = 1, \ldots, p\\
&\qquad -\langle w, \varphi(v_j)\rangle + b \geq \eta - \xi_j, \quad \xi_j \geq 0 \qquad j = 1, \ldots, q\\
&\qquad \eta \geq 0.
\end{aligned}$$

Tracing through the derivation of the dual program, we obtain

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\mathbf{K}\begin{pmatrix}\lambda\\ \mu\end{pmatrix}\\
&\text{subject to}\\
&\qquad \sum_{i=1}^{p}\lambda_i = \sum_{j=1}^{q}\mu_j\\
&\qquad \sum_{i=1}^{p}\lambda_i + \sum_{j=1}^{q}\mu_j \geq K_m\\
&\qquad 0 \leq \lambda_i \leq K_s, \quad i = 1, \ldots, p\\
&\qquad 0 \leq \mu_j \leq K_s, \quad j = 1, \ldots, q,
\end{aligned}$$

where $\mathbf{K}$ is the kernel matrix of Section 32.1.

As in Section 32.2, we obtain

$$w = \sum_{i=1}^{p}\lambda_i\varphi(u_i) - \sum_{j=1}^{q}\mu_j\varphi(v_j),$$

so

$$b = \frac{1}{2}\left(\sum_{i=1}^{p}\lambda_i(\kappa(u_i, u_{i_0}) + \kappa(u_i, v_{j_0})) - \sum_{j=1}^{q}\mu_j(\kappa(v_j, u_{i_0}) + \kappa(v_j, v_{j_0}))\right),$$

and the classification function

$$f(x) = \mathrm{sgn}(\langle w, \varphi(x)\rangle - b)$$

is given by

$$\begin{aligned}
f(x) = \mathrm{sgn}\Bigg(&\sum_{i=1}^{p}\lambda_i(2\kappa(u_i, x) - \kappa(u_i, u_{i_0}) - \kappa(u_i, v_{j_0}))\\
&- \sum_{j=1}^{q}\mu_j(2\kappa(v_j, x) - \kappa(v_j, u_{i_0}) - \kappa(v_j, v_{j_0}))\Bigg).
\end{aligned}$$

32.4 Soft Margin SVM; (SVM_{s3})

In this section we consider the version of Problem ($\mathrm{SVM}_{s2'}$) in which instead of using the function $K\left(\sum_{i=1}^{p}\epsilon_i + \sum_{j=1}^{q}\xi_j\right)$ as a regularizing function we use the quadratic function $K(\|\epsilon\|_2^2 + \|\xi\|_2^2)$.

Soft margin SVM (SVM_{s3}):

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}w^\top w - \nu\eta + K(\epsilon^\top\epsilon + \xi^\top\xi)\\
&\text{subject to}\\
&\qquad w^\top u_i - b \geq \eta - \epsilon_i, \qquad i = 1,\ldots,p\\
&\qquad -w^\top v_j + b \geq \eta - \xi_j, \qquad j = 1,\ldots,q\\
&\qquad \eta \geq 0,
\end{aligned}$$

where ν and K are two given positive constants. As we saw earlier, it is convenient to pick $K = 1/(p+q)$.

The new twist with this formulation of the problem is that if $\epsilon_i < 0$, then the corresponding inequality $w^\top u_i - b \geq \eta - \epsilon_i$ implies the inequality $w^\top u_i - b \geq \eta$ obtained by setting ϵ_i to zero while reducing the value of $\|\epsilon\|^2$, and similarly if $\xi_j < 0$, then the corresponding inequality $-w^\top v_j + b \geq \eta - \xi_j$ implies the inequality $-w^\top v_j + b \geq \eta$ obtained by setting ξ_j to zero while reducing the value of $\|\xi\|^2$. Therefore, if (w, b, ϵ, ξ) is an optimal solution of Problem (SVM_{s3}) it is not necessary to restrict the slack variables ϵ_i and ξ_j to the nonnegative, which simplifies matters a bit.

One of the advantages of this methods is that ϵ is determined by λ and ξ is determined by μ. We could also omit the constraint $\eta \geq 0$, because for an optimal solution it can be shown using duality that $\eta \geq 0$.

The Lagrangian is given by

$$\begin{aligned}
L(w,\epsilon,\xi,b,\eta,\lambda,\mu,\gamma) &= \frac{1}{2}w^\top w - \nu\eta + K(\epsilon^\top\epsilon + \xi^\top\xi) + w^\top X\begin{pmatrix}\lambda\\ \mu\end{pmatrix}\\
&\quad - \epsilon^\top\lambda - \xi^\top\mu + b(\mathbf{1}_p^\top\lambda - \mathbf{1}_q^\top\mu) + \eta(\mathbf{1}_p^\top\lambda + \mathbf{1}_q^\top\mu) - \gamma\eta\\
&= \frac{1}{2}w^\top w + w^\top X\begin{pmatrix}\lambda\\ \mu\end{pmatrix} + \eta(\mathbf{1}_p^\top\lambda + \mathbf{1}_q^\top\mu - \nu - \gamma)\\
&\quad + K(\epsilon^\top\epsilon + \xi^\top\xi) - \epsilon^\top\lambda - \xi^\top\mu + b(\mathbf{1}_p^\top\lambda - \mathbf{1}_q^\top\mu).
\end{aligned}$$

To find the dual function $G(\lambda,\mu,\gamma)$ we minimize $L(w,\epsilon,\xi,b,\eta,\lambda,\mu,\gamma)$ with respect to w, ϵ, ξ, b, and η. Since the Lagrangian is convex and $(w,\epsilon,\xi,b,\eta) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R} \times \mathbb{R}$, a convex open set, by Theorem 20.11, the Lagrangian has a minimum in (w,ϵ,ξ,b,η) iff $\nabla L_{w,\epsilon,\xi,b,\eta} = 0$,

so we compute $\nabla L_{w,\epsilon,\xi,b,\eta}$. The gradient $\nabla L_{w,\epsilon,\xi,b,\eta}$ is given by

$$\nabla L_{w,\epsilon,\xi,b,\eta} = \begin{pmatrix} w + X\begin{pmatrix}\lambda\\ \mu\end{pmatrix} \\ 2K\epsilon - \lambda \\ 2K\xi - \mu \\ \mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu \\ \mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu - \nu - \gamma \end{pmatrix}$$

By setting $\nabla L_{w,\epsilon,\xi,b,\eta} = 0$ we get the equations

$$w = -X\begin{pmatrix}\lambda\\ \mu\end{pmatrix} \tag{$*_w$}$$

and

$$\begin{aligned} 2K\epsilon &= \lambda \\ 2K\xi &= \mu \\ \mathbf{1}_p^\top \lambda &= \mathbf{1}_q^\top \mu \\ \mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu &= \nu + \gamma. \end{aligned}$$

The last two equations are identical to the last two equations obtained in Problem ($\mathrm{SVM}_{s2'}$). We can use the other equations to obtain the following expression for the dual function $G(\lambda, \mu, \gamma)$,

$$\begin{aligned} G(\lambda, \mu, \gamma) &= -\frac{1}{4K}(\lambda^\top \lambda + \mu^\top \mu) - \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda\\ \mu\end{pmatrix} \\ &= -\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(X^\top X + \frac{1}{2K} I_{p+q}\right)\begin{pmatrix}\lambda\\ \mu\end{pmatrix}. \end{aligned}$$

Consequently the dual program is equivalent to the minimization program

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(X^\top X + \frac{1}{2K} I_{p+q}\right)\begin{pmatrix}\lambda\\ \mu\end{pmatrix} \\ \text{subject to} \quad & \\ & \sum_{i=1}^{p} \lambda_i = \sum_{j=1}^{q} \mu_j \\ & \sum_{i=1}^{p} \lambda_i + \sum_{j=1}^{q} \mu_j \geq \nu \\ & \lambda_i \geq 0, \quad i = 1, \ldots, p \\ & \mu_j \geq 0, \quad j = 1, \ldots, q. \end{aligned}$$

The above program is similar to the program that was obtained for Problem (SVM$_{s2'}$) but the matrix $X^\top X$ is replaced by the matrix $X^\top X + (1/2K)I_{p+q}$, which is positive definite since $K > 0$, and also the inequalities $\lambda_i \leq K$ and $\mu_j \leq K$ no longer hold. However, the constraints imply that there is some i_0 such that $\lambda_{i_0} > 0$ and some j_0 such that $\mu_{j_0} > 0$.

The dual program is solved by making use of numerical procedures based on gradient descent. If the primal problem is solvable, this yields solutions for λ and μ. We obtain w from λ and μ, and γ, as in Problem (SVM$_{s2'}$); namely,

$$w = \sum_{i=1}^{p} \lambda_i u_i - \sum_{j=1}^{q} \mu_j v_j.$$

Since the variables ϵ_i and μ_j are not restricted to be nonnegative we no longer have complementary slackness conditions involving them, but we know that

$$\epsilon = \frac{\lambda}{2K}, \quad \xi = \frac{\mu}{2K}.$$

Also since the constraints

$$\sum_{i=1}^{p} \lambda_i \geq \frac{\nu}{2} \quad \text{and} \quad \sum_{j=1}^{q} \mu_j \geq \frac{\nu}{2}$$

imply that there is some i_0 such that $\lambda_{i_0} > 0$ and some j_0 such that $\mu_{j_0} > 0$, we have $\epsilon_{i_0} > 0$ and $\xi_{j_0} > 0$, which means that at least two points are misclassified, so Problem (SVM$_{s3}$) should only be used when the sets $\{u_i\}$ and $\{v_j\}$ are *not* linearly separable. We can solve for b and η using the active constraints corresponding to any i_0 such that $\lambda_{i_0} > 0$ and any j_0 such that $\mu_{j_0} > 0$ and we get

$$b = \frac{w^\top u_{i_0} + w^\top v_{j_0}}{2}$$
$$\eta = \frac{w^\top u_{i_0} - w^\top v_{j_0}}{2}.$$

We can also use the fact that the optimality gap is 0 to find η. We have

$$\frac{1}{2} w^\top w - \nu\eta + K(\epsilon^\top \epsilon + \xi^\top \xi) = -\frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \left(X^\top X + \frac{1}{2K} I_{p+q} \right) \begin{pmatrix} \lambda \\ \mu \end{pmatrix},$$

and since

$$w = -X \begin{pmatrix} \lambda \\ \mu \end{pmatrix}$$

we get

$$\nu\eta = K(\lambda^\top \lambda + \mu^\top \mu) + \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \left(X^\top X + \frac{1}{4K} I_{p+q} \right) \begin{pmatrix} \lambda \\ \mu \end{pmatrix}.$$

The above confirms that at optimality we have $\eta \geq 0$.

The "kernelized" version of Problem (SVM_{s3}) is the following:

Soft margin kernel SVM (SVM_{s3}):

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\langle w, w\rangle - \nu\eta + \frac{1}{p+q}(\epsilon^\top \epsilon + \xi^\top \xi)\\
&\text{subject to}\\
&\qquad \langle w, \varphi(u_i)\rangle - b \geq \eta - \epsilon_i, \qquad i = 1, \ldots, p\\
&\qquad -\langle w, \varphi(v_j)\rangle + b \geq \eta - \xi_j, \qquad j = 1, \ldots, q\\
&\qquad \eta \geq 0.
\end{aligned}$$

By going over the derivation of the dual program, we obtain

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(\mathbf{K} + \frac{p+q}{2} I_{p+q}\right)\begin{pmatrix}\lambda\\ \mu\end{pmatrix}\\
&\text{subject to}\\
&\qquad \sum_{i=1}^{p} \lambda_i = \sum_{j=1}^{q} \mu_j\\
&\qquad \sum_{i=1}^{p} \lambda_i + \sum_{j=1}^{q} \mu_j \geq \nu\\
&\qquad \lambda_i \geq 0, \quad i = 1, \ldots, p\\
&\qquad \mu_j \geq 0, \quad j = 1, \ldots, q,
\end{aligned}$$

where $\mathbf{K}$ is the kernel matrix of Section 32.1. Then w, b, and $f(x)$ are obtained exactly as in Section 32.3.

32.5 Soft Margin Support Vector Machines; (SVM_{s4})

In this section we consider a variation of Problem ($\mathrm{SVM}_{s2'}$) by adding the term $(1/2)b^2$ to the objective function. The result is that in minimizing the Lagrangian to find the dual function G, not just w but also b is determined. We also suppress the constraint $\eta \geq 0$ which turns out to be redundant.

Soft margin SVM (SVM_{s4}):

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}w^\top w + \frac{1}{2}b^2 + K\left(-\nu\eta + \frac{1}{p+q}\begin{pmatrix}\epsilon^\top & \xi^\top\end{pmatrix}\mathbf{1}_{p+q}\right)\\
&\text{subject to}\\
&\qquad w^\top u_i - b \geq \eta - \epsilon_i, \quad \epsilon_i \geq 0 \qquad i = 1, \ldots, p\\
&\qquad -w^\top v_j + b \geq \eta - \xi_j, \quad \xi_j \geq 0 \qquad j = 1, \ldots, q.
\end{aligned}$$

To simplify the presentation we assume that $K = 1$ and we write K_s for $1/(p+q)$.

The Lagrangian $L(w, \epsilon, \xi, b, \eta, \lambda, \mu, \alpha, \beta)$ with $\lambda, \alpha \in \mathbb{R}^p_+$, $\mu, \beta \in \mathbb{R}^q_+$ is given by

$$\begin{aligned} L(w, \epsilon, \xi, b, \eta, \lambda, \mu, \alpha, \beta) &= \frac{1}{2} w^\top w + w^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + \frac{b^2}{2} - \nu\eta + K_s(\epsilon^\top \mathbf{1}_p + \xi^\top \mathbf{1}_q) - \epsilon^\top(\lambda + \alpha) \\ &\quad - \xi^\top(\mu + \beta) + b(\mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu) + \eta(\mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu), \\ &= \frac{1}{2} w^\top w + w^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + \frac{b^2}{2} + b(\mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu) + \eta(\mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu - \nu) \\ &\quad + \epsilon^\top(K_s \mathbf{1}_p - (\lambda + \alpha)) + \xi^\top(K_s \mathbf{1}_q - (\mu + \beta)). \end{aligned}$$

To find the dual function $G(\lambda, \mu, \alpha, \beta)$, we minimize $L(w, \epsilon, \xi, b, \eta, \lambda, \mu, \alpha, \beta)$ with respect to w, ϵ, ξ, b, and η. Since the Lagrangian is convex and $(w, \epsilon, \xi, b, \eta) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R} \times \mathbb{R}$, a convex open set, by Theorem 20.11, the Lagrangian has a minimum in $(w, \epsilon, \xi, b, \eta)$ iff $\nabla L_{w,\epsilon,\xi,b,\eta} = 0$, so we compute its gradient with respect to $w, \epsilon, \xi, b, \eta$ and we get

$$\nabla L_{w,\epsilon,\xi,b,\eta} = \begin{pmatrix} X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + w \\ K_s \mathbf{1}_p - (\lambda + \alpha) \\ K_s \mathbf{1}_q - (\mu + \beta) \\ b + \mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu \\ \mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu - \nu \end{pmatrix}.$$

By setting $\nabla L_{w,\epsilon,\xi,b,\eta} = 0$ we get the equations

$$w = -X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \tag{$*_w$}$$

$$\begin{aligned} \lambda + \alpha &= K_s \mathbf{1}_p \\ \mu + \beta &= K_s \mathbf{1}_q \\ \mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu &= \nu, \end{aligned}$$

and

$$b = -(\mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu). \tag{$*_b$}$$

The second and third equations are equivalent to the box constraints

$$0 \leq \lambda_i, \mu_j \leq K_s, \quad i = 1, \ldots, p,\ j = 1, \ldots, q.$$

Since we assumed that the primal problem has an optimal solution with $w \neq 0$, we have

$$X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \neq 0.$$

Plugging back w from $(*_w)$ and b from $(*_b)$ into the Lagrangian, we get

$$\begin{aligned}
G(\lambda,\mu,\alpha,\beta) &= \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda\\ \mu\end{pmatrix} - \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda\\ \mu\end{pmatrix} + \frac{1}{2}b^2 - b^2\\
&= -\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda\\ \mu\end{pmatrix} - \frac{1}{2}b^2\\
&= -\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} \left(X^\top X + \begin{pmatrix}\mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top\\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top\end{pmatrix}\right)\begin{pmatrix}\lambda\\ \mu\end{pmatrix},
\end{aligned}$$

so the dual function is independent of α, β and is given by

$$G(\lambda,\mu) = -\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} \left(X^\top X + \begin{pmatrix}\mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top\\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top\end{pmatrix}\right)\begin{pmatrix}\lambda\\ \mu\end{pmatrix}.$$

The dual program is given by

$$\begin{aligned}
\text{maximize}\quad & -\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} \left(X^\top X + \begin{pmatrix}\mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top\\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top\end{pmatrix}\right)\begin{pmatrix}\lambda\\ \mu\end{pmatrix}\\
\text{subject to}\quad & \\
& \sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j = \nu\\
& 0 \leq \lambda_i \leq K_s, \quad i = 1,\ldots,p\\
& 0 \leq \mu_j \leq K_s, \quad j = 1,\ldots,q.
\end{aligned}$$

Finally, the dual program is equivalent to the following minimization program:

$$\begin{aligned}
\text{minimize}\quad & \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} \left(X^\top X + \begin{pmatrix}\mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top\\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top\end{pmatrix}\right)\begin{pmatrix}\lambda\\ \mu\end{pmatrix}\\
\text{subject to}\quad & \\
& \sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j = \nu\\
& 0 \leq \lambda_i \leq K_s, \quad i = 1,\ldots,p\\
& 0 \leq \mu_j \leq K_s, \quad j = 1,\ldots,q.
\end{aligned}$$

The dual program is solved by making use of numerical procedures based on gradient descent. If the primal problem is solvable, this yields solutions for λ and μ. Once a solution for λ and μ is obtained, we have

$$\begin{aligned}
w &= -X\begin{pmatrix}\lambda\\ \mu\end{pmatrix} = \sum_{i=1}^p \lambda_i u_i - \sum_{j=1}^q \mu_j v_j\\
b &= -\sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j.
\end{aligned}$$

As we said earlier, the hypotheses of Theorem 30.14(2) hold, so *if* the primal problem (SVM_{s4}) has an optimal solution with $w \neq 0$, *then* the dual problem has a solution too, and the duality gap is zero. Therefore, for optimal solutions we have

$$L(w, \epsilon, \xi, b, \eta, \lambda, \mu, \alpha, \beta) = G(\lambda, \mu, \alpha, \beta),$$

which means that

$$\frac{1}{2}w^\top w + \frac{b^2}{2} - \nu\eta + K_s\left(\sum_{i=1}^{p}\epsilon_i + \sum_{j=1}^{q}\xi_j\right) = -\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(X^\top X + \begin{pmatrix}\mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top \\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top\end{pmatrix}\right)\begin{pmatrix}\lambda \\ \mu\end{pmatrix}$$

and since

$$\frac{1}{2}w^\top w + \frac{b^2}{2} = \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(X^\top X + \begin{pmatrix}\mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top \\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top\end{pmatrix}\right)\begin{pmatrix}\lambda \\ \mu\end{pmatrix},$$

we get

$$\eta = \frac{K_s}{\nu}\left(\sum_{i=1}^{p}\epsilon_i + \sum_{j=1}^{q}\xi_j\right) + \frac{1}{\nu}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(X^\top X + \begin{pmatrix}\mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top \\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top\end{pmatrix}\right)\begin{pmatrix}\lambda \\ \mu\end{pmatrix}. \qquad (*)$$

Since

$$X^\top X + \begin{pmatrix}\mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top \\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top\end{pmatrix}$$

is positive semidefinite, so we confirm that $\eta \geq 0$.

Since $K_s = 1/(p+q)$, in order for the constraints

$$\sum_{i=1}^{p}\lambda_i + \sum_{j=1}^{q}\mu_j = \nu$$

and $0 \leq \lambda_i, \mu_j \leq 1/(p+q)$ to be satisfied we must have

$$\nu \leq 1.$$

The equation

$$\sum_{i=1}^{p}\lambda_i + \sum_{j=1}^{q}\mu_j = \nu$$

also implies that either there is some i_0 such that $\lambda_{i_0} > 0$ or there is some j_0 such that $\mu_{j_0} > 0$.

Under the **Standard Margin Hypothesis** for (SVM_{s4}), either there is some i_0 such that $0 < \lambda_{i_0} < K_s$ or there is some j_0 such that $0 < \mu_{j_0} < K_s$, and by the complementary slackness conditions $\epsilon_{i_0} = 0$ or $\xi_{j_0} = 0$, so we have

$$w^\top u_{i_0} - b = \eta, \quad \text{or} \quad -w^\top v_{j_0} + b = \eta,$$

and we can solve for η.

The equations (†) and the box inequalities

$$0 \leq \lambda_i \leq K_s, \quad 0 \leq \mu_j \leq K_s$$

also imply the following facts:

Proposition 32.4. *If Problem* (SVM_{s4}) *has an optimal solution with* $w \neq 0$ *and* $\eta > 0$ *then the following facts hold:*

(1) At most $\nu(p+q)$ *points* u_i *and* v_j *fail to achieve the margin* η.

(2) At least $\nu(p+q)$ *points* u_i *and* v_j *have margin at most* η.

Proof. (1) Recall that for an optimal solution with $w \neq 0$ and $\eta > 0$ we have the equation

$$\sum_{i=1}^{p} \lambda_i + \sum_{j=1}^{q} \mu_j = \nu.$$

If u_i fails to achieve the margin η, then $\epsilon_i > 0$, and by complementary slackness $\lambda_i = K_s = 1/(p+q)$. Similarly, if v_j fails to achieve the margin then $\xi_j > 0$, and by complementary slackness $\mu_j = K_s = 1/(p+q)$. Assume that p_f points u_i fail the margin and that q_f points v_j fail the margin. Then

$$\nu = \sum_{i=1}^{p} \lambda_i + \sum_{j=1}^{q} \mu_j \geq \frac{p_f + q_f}{p+q},$$

so

$$p_f + q_f \leq \nu(p+q).$$

(2) A point u_i has margin at most η iff $\lambda_i > 0$ and a point v_j has margin at most η iff $\mu_j > 0$. If

$$I_m = \{i \in \{1, \ldots, p\} \mid \lambda_i > 0\} \quad \text{and} \quad p_m = |I_m|$$

and

$$J_m = \{j \in \{1, \ldots, q\} \mid \mu_j > 0\} \quad \text{and} \quad q_m = |J_m|$$

then

$$\nu = \sum_{i=1}^{p} \lambda_i + \sum_{j=1}^{q} \mu_j = \sum_{i \in I_m} \lambda_i + \sum_{j \in J_m} \mu_j,$$

and since $\lambda_i, \mu_j \leq K_s = 1/(p+q)$, we have

$$\nu = \sum_{i \in I_m} \lambda_i + \sum_{j \in J_m} \mu_j \leq \frac{p_m + q_n}{p+q},$$

which yields

$$p_m + q_m \geq \nu(p+q).$$

□

Note that if ν is chosen so that $\nu < 1/(p+q)$, then $\nu(p+q) < 1$, which means that none of the data points are misclassified; in other words, the u_is and v_js are linearly separable. Thus we see that if the u_is and v_js are not linearly separable we must pick ν such that $1/(p+q) \leq \nu \leq 1$ for the method to succeed.

The following proposition clarifies the role of the constant ν in establishing the trade-off between the width of the margin and the number of margin-error points. In particular, it shows that if Problem (SVM_{s4}) has an optimal solution with $w \neq 0$ and $\eta > 0$, and if $\nu < 1$, then at least some u_i or some v_j is classified correctly. Obviously we have $1/(p+q) \leq 1$.

Proposition 32.5. *Suppose $(w, b, \eta, \epsilon, \xi)$ is an optimal solution of Problem (SVM_{s4}) with $w \neq 0$ and $\eta > 0$, and let p_f be the number of points u_i that are misclassified ($\epsilon_i > 0$) and q_f be the number of points v_j that are misclassified ($\xi_j > 0$). If $p_f + q_f \geq 2$ and if $1/(p+q) \leq \nu < (p_f + q_f)/(p+q)$, then either there is some i such that $\epsilon_i = 0$ and the constraint $w^\top u_i - b = \eta$ is active, or there is some j such that $\xi_j = 0$ and the constraint $-w^\top v_j + b = \eta$ is active.*

Proof. (1) We may assume that $K_s = 1/(p+q)$. We proceed by contradiction. Thus we assume that for all $i \in \{1, \ldots, p\}$, if $\epsilon_i = 0$ then the constraint $w^\top u_i - b \geq \eta$ is not active, namely $w^\top u_i - b > \eta$, and for all $j \in \{1, \ldots, q\}$, if $\xi_j = 0$ then the constraint $-w^\top v_j + b \geq \eta$ is not active, namely $-w^\top v_j + b > \eta$.

Let $I = \{i \in \{1, \ldots, p\} \mid \epsilon_i > 0\}$, let $J = \{j \in \{1, \ldots, q\} \mid \xi_j > 0\}$, and let $p_f = |I|$ and $q_f = |J|$ (of course, $\eta > 0$).

Assume that $p_f + q_f \geq 2$. By complementary slackness all the constraints for which $i \in I$ and $j \in J$ are active, so our hypotheses are

$$\begin{array}{lll} w^\top u_i - b = \eta - \epsilon_i & \epsilon_i > 0 & i \in I \\ -w^\top v_j + b = \eta - \xi_j & \xi_j > 0 & j \in J \\ w^\top u_i - b > \eta & & i \notin I \\ -w^\top v_j + b > \eta & & j \notin J. \end{array}$$

For any $\theta > 0$ such that

$$\theta < \min\{\epsilon_i, \xi_j, \eta \mid i \in \{1, \ldots, p\}, j \in \{1, \ldots, q\}\},$$

we can write

$$\begin{array}{lll} w^\top u_i - b = \eta - \theta - (\epsilon_i - \theta) & \epsilon_i - \theta \geq 0 & i \in I \\ -w^\top v_j + b = \eta - \theta - (\xi_j - \theta) & \xi_j - \theta \geq 0 & j \in J \\ w^\top u_i - b > \eta - \theta & & i \notin I \\ -w^\top v_j + b > \eta - \theta & & j \notin J. \end{array}$$

The original value of the objective function is

$$\omega(0) = \frac{1}{2}w^\top w - \nu\eta + \frac{1}{p+q}\left(\sum_{i\in I}\epsilon_i + \sum_{j\in J}\xi_j\right),$$

and the new value is

$$\begin{aligned}\omega(\theta) &= \frac{1}{2}w^\top w - \nu(\eta-\theta) + \frac{1}{p+q}\left(\sum_{i\in I}(\epsilon_i-\theta) + \sum_{j\in J}(\xi_j-\theta)\right)\\ &= \frac{1}{2}w^\top w - \nu\eta + \frac{1}{p+q}\left(\sum_{i\in I}\epsilon_i + \sum_{j\in J}\xi_j\right) - \left(\frac{p_f+q_f}{p+q} - \nu\right)\theta.\end{aligned}$$

Since by hypothesis $p_f + q_f \geq 2$, if

$$\frac{1}{p+1} \leq \nu < \frac{p_f+q_f}{p+q},$$

then the term involving θ is negative so

$$\omega(\theta) < \omega(0),$$

and by the choice of θ we have $\eta - \theta > 0$, so $(w, b, \eta-\theta, \epsilon-\theta, \xi-\theta)$ is a feasible solution, contradicting the optimality of the solution $(w, b, \eta, \epsilon, \xi)$; here we write $\epsilon - \theta$ for the vector $(\epsilon_1 - \theta, \ldots, \epsilon_p - \theta)$, and similarly for $\xi - \theta$. □

Note that if $p_f + q_f = p + q$ and $\nu < 1$, then Proposition 32.5 yields a contradiction. Therefore $p_f + q_f < p + q$, that is, at least some u_i or some v_j is classified correctly

Remark: If the the sets $\{u_i\}$ and $\{v_j\}$ are linearly separable, then we know from Theorem 30.10 that some u_i is on the blue margin and some v_j is on the red margin.

We also have the following proposition that gives a sufficient condition implying that η can be found in terms of an optimal solution (λ, μ) of the dual.

Proposition 32.6. *If $(w, b, \eta, \epsilon, \xi)$ is an optimal solution of Problem (SVM_{s4}) with $w \neq 0$ and $\eta > 0$, if $1/(p+q) \leq \nu < 2/(p+q)$ and $p_f + q_f \geq 2$, then η can always be determined from an optimal solution (λ, μ) of the dual.*

Proof. As we already explained, Problem (SVM_{s4}) satisfies the conditions for having a zero duality gap. Therefore, for optimal solutions we have

$$L(w, \epsilon, \xi, b, \eta, \lambda, \mu, \alpha, \beta) = G(\lambda, \mu, \alpha, \beta),$$

which means that

$$\nu\eta = \frac{1}{p+q}\left(\sum_{i=1}^{p}\epsilon_i + \sum_{j=1}^{q}\xi_j\right) + \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(X^\top X + \begin{pmatrix}\mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top\\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top\end{pmatrix}\right)\begin{pmatrix}\lambda\\ \mu\end{pmatrix}. \quad (*)$$

Let $I = \{i \in \{1, \ldots, p\} \mid \epsilon_i > 0\}$ and $J = \{j \in \{1, \ldots, q\} \mid \xi_j > 0\}$. If $I = J = \emptyset$, then

$$\eta = \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \left(X^\top X + \begin{pmatrix} \mathbf{1}_p \mathbf{1}_p^\top & -\mathbf{1}_p \mathbf{1}_q^\top \\ -\mathbf{1}_q \mathbf{1}_p^\top & \mathbf{1}_q \mathbf{1}_q^\top \end{pmatrix} \right) \begin{pmatrix} \lambda \\ \mu \end{pmatrix}.$$

Assume that $|I| + |J| \geq 2$. Then we know that $\lambda_i = 1/(p+q)$ for all $i \in I$ and $\mu_j = 1/(p+q)$ for all $j \in J$, so the following equations are active:

$$\begin{aligned} w^\top u_i - b &= \eta - \epsilon_i && i \in I \\ -w^\top v_j + b &= \eta - \xi_j && j \in J. \end{aligned}$$

But $(*)$ can be written as

$$\nu\eta = \frac{1}{p+q}\left(\sum_{i \in I} \epsilon_i + \sum_{j \in J} \xi_j \right) + \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \left(X^\top X + \begin{pmatrix} \mathbf{1}_p \mathbf{1}_p^\top & -\mathbf{1}_p \mathbf{1}_q^\top \\ -\mathbf{1}_q \mathbf{1}_p^\top & \mathbf{1}_q \mathbf{1}_q^\top \end{pmatrix} \right) \begin{pmatrix} \lambda \\ \mu \end{pmatrix}, \qquad (**)$$

and since

$$\begin{aligned} \epsilon_i &= \eta - w^\top u_i + b && i \in I \\ \xi_j &= \eta + w^\top v_j - b && j \in J, \end{aligned}$$

by substituting in the equation $(**)$ we get

$$\begin{aligned} \left(\frac{|I| + |J|}{p+q} - \nu \right) \eta = \frac{|J| - |I|}{p+q} b + \frac{1}{p+q} w^\top \left(\sum_{i \in I} u_i - \sum_{j \in J} v_j \right) \\ - \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \left(X^\top X + \begin{pmatrix} \mathbf{1}_p \mathbf{1}_p^\top & -\mathbf{1}_p \mathbf{1}_q^\top \\ -\mathbf{1}_q \mathbf{1}_p^\top & \mathbf{1}_q \mathbf{1}_q^\top \end{pmatrix} \right) \begin{pmatrix} \lambda \\ \mu \end{pmatrix}. \end{aligned}$$

We need to choose ν such that $(|I| + |J|)/(p+q)| - \nu \neq 0$ Since we are assuming that $|I| + |J| \geq 2$, this will be the case if $1/(p+q) \leq \nu < 2/(p+q)$. If this condition is satisfied we can solve for η. □

Remark: If the the sets $\{u_i\}$ and $\{v_j\}$ are linearly separable, then we know from Theorem 30.10 that some u_i is on the blue margin and some v_j is on the red margin, so b and δ can be determined. Although we can ensure that some u_i is classified correctly or some v_j is classified correctly, it does not seem possible to prove that the corresponding constraints are active without additional hypotheses (such as $p_f + q_f \geq 2$).

The "kernelized" version of Problem (SVM_{s4}) is the following:

Soft margin kernel SVM (SVM_{s4}):

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\langle w, w \rangle + \frac{1}{2} b^2 - \nu\eta + \frac{1}{p+q} \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} \mathbf{1}_{p+q} \\ &\text{subject to} \\ &\qquad \langle w, \varphi(u_i) \rangle - b \geq \eta - \epsilon_i, \quad \epsilon_i \geq 0 \qquad i = 1, \ldots, p \\ &\qquad -\langle w, \varphi(v_j) \rangle + b \geq \eta - \xi_j, \quad \xi_j \geq 0 \qquad j = 1, \ldots, q. \end{aligned}$$

Tracing through the derivation of the dual program, we obtain

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(\mathbf{K} + \begin{pmatrix}\mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top \\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top\end{pmatrix}\right)\begin{pmatrix}\lambda \\ \mu\end{pmatrix} \\
&\text{subject to} \\
&\qquad \sum_{i=1}^{p}\lambda_i + \sum_{j=1}^{q}\mu_j = \nu \\
&\qquad 0 \leq \lambda_i \leq K_s, \quad i = 1, \ldots, p \\
&\qquad 0 \leq \mu_j \leq K_s, \quad j = 1, \ldots, q,
\end{aligned}$$

where $\mathbf{K}$ is the kernel matrix of Section 32.1.

We obtain

$$\begin{aligned}
w &= \sum_{i=1}^{p}\lambda_i\varphi(u_i) - \sum_{j=1}^{q}\mu_j\varphi(v_j) \\
b &= -\sum_{i=1}^{p}\lambda_i + \sum_{j=1}^{q}\mu_j.
\end{aligned}$$

The classification function

$$f(x) = \mathrm{sgn}(\langle w, \varphi(x)\rangle - b)$$

is given by

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{p}\lambda_i(\kappa(u_i, x) + 1) - \sum_{j=1}^{q}\mu_j(\kappa(v_j, x) + 1)\right).$$

32.6 Soft Margin SVM; (SVM_{s5})

In this section we consider the version of Problem (SVM_{s3}) in which we add the term $(1/2)b^2$ to the objective function. We also drop the constraint $\eta \geq 0$ which is redundant.

Soft margin SVM (SVM_{s5}):

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}w^\top w + \frac{1}{2}b^2 - \nu\eta + K(\epsilon^\top\epsilon + \xi^\top\xi) \\
&\text{subject to} \\
&\qquad w^\top u_i - b \geq \eta - \epsilon_i, \qquad i = 1, \ldots, p \\
&\qquad -w^\top v_j + b \geq \eta - \xi_j, \qquad j = 1, \ldots, q,
\end{aligned}$$

where ν and K are two given positive constants. As we saw earlier, it is convenient to pick $K = 1/(p+q)$.

The Lagrangian is given by

$$\begin{aligned} L(w,\epsilon,\xi,b,\eta,\lambda,\mu) &= \frac{1}{2}w^\top w + \frac{1}{2}b^2 - \nu\eta + K(\epsilon^\top\epsilon + \xi^\top\xi) + w^\top X\begin{pmatrix}\lambda\\ \mu\end{pmatrix} \\ &\quad - \epsilon^\top\lambda - \xi^\top\mu + b(\mathbf{1}_p^\top\lambda - \mathbf{1}_q^\top\mu) + \eta(\mathbf{1}_p^\top\lambda + \mathbf{1}_q^\top\mu) \\ &= \frac{1}{2}w^\top w + w^\top X\begin{pmatrix}\lambda\\ \mu\end{pmatrix} + \eta(\mathbf{1}_p^\top\lambda + \mathbf{1}_q^\top\mu) \\ &\quad + K(\epsilon^\top\epsilon + \xi^\top\xi) - \epsilon^\top\lambda - \xi^\top\mu + b(\mathbf{1}_p^\top\lambda - \mathbf{1}_q^\top\mu) + \frac{1}{2}b^2. \end{aligned}$$

To find the dual function $G(\lambda,\mu)$ we minimize $L(w,\epsilon,\xi,b,\eta,\lambda,\mu)$ with respect to $w,\epsilon,\xi,\ b$, and η. Since the Lagrangian is convex and $(w,\epsilon,\xi,b,\eta) \in \mathbb{R}^n\times\mathbb{R}^p\times\mathbb{R}^q\times\mathbb{R}\times\mathbb{R}$, a convex open set, by Theorem 20.11, the Lagrangian has a minimum in (w,ϵ,ξ,b,η) iff $\nabla L_{w,\epsilon,\xi,b,\eta} = 0$, so we compute $\nabla L_{w,\epsilon,\xi,b,\eta}$. The gradient $\nabla L_{w,\epsilon,\xi,b,\eta}$ is given by

$$\nabla L_{w,\epsilon,\xi,b,\eta} = \begin{pmatrix} w + X\begin{pmatrix}\lambda\\ \mu\end{pmatrix} \\ 2K\epsilon - \lambda \\ 2K\xi - \mu \\ b + \mathbf{1}_p^\top\lambda - \mathbf{1}_q^\top\mu \\ \mathbf{1}_p^\top\lambda + \mathbf{1}_q^\top\mu - \nu \end{pmatrix}$$

By setting $\nabla L_{w,\epsilon,\xi,b,\eta} = 0$ we get the equations

$$w = -X\begin{pmatrix}\lambda\\ \mu\end{pmatrix} \tag{$*_w$}$$

and

$$\begin{aligned} 2K\epsilon &= \lambda \\ 2K\xi &= \mu \\ b &= -(\mathbf{1}_p^\top\lambda - \mathbf{1}_q^\top\mu) \\ \mathbf{1}_p^\top\lambda + \mathbf{1}_q^\top\mu &= \nu. \end{aligned}$$

The last two equations are identical to the last two equations obtained in Problem (SVM$_{s4}$). We can use the other equations to obtain the following expression for the dual function $G(\lambda,\mu,\gamma)$,

$$\begin{aligned} G(\lambda,\mu,\gamma) &= -\frac{1}{4K}(\lambda^\top\lambda + \mu^\top\mu) - \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X\begin{pmatrix}\lambda\\ \mu\end{pmatrix} - \frac{b^2}{2} \\ &= -\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(X^\top X + \begin{pmatrix}\mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top \\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top\end{pmatrix} + \frac{1}{2K}I_{p+q}\right)\begin{pmatrix}\lambda\\ \mu\end{pmatrix}. \end{aligned}$$

Consequently the dual program is equivalent to the minimization program

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(X^\top X + \begin{pmatrix}\mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top \\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top\end{pmatrix} + \frac{1}{2K}I_{p+q}\right)\begin{pmatrix}\lambda \\ \mu\end{pmatrix}\\
&\text{subject to}\\
&\sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j = \nu\\
&\lambda_i \geq 0, \quad i = 1, \ldots, p\\
&\mu_j \geq 0, \quad j = 1, \ldots, q.
\end{aligned}$$

The dual program is solved by making use of numerical procedures based on gradient descent. If the primal problem is solvable, this yields solutions for λ and μ.

The constraints imply that either there is some i_0 such that $\lambda_{i_0} > 0$ or there is some j_0 such that $\mu_{j_0} > 0$. We obtain w and b from λ and μ, as in Problem (SVM_{s4}); namely,

$$\begin{aligned}
w &= \sum_{i=1}^p \lambda_i u_i - \sum_{j=1}^q \mu_j v_j\\
b &= -\sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j.
\end{aligned}$$

Since the variables ϵ_i and μ_j are not restricted to be nonnegative we no longer have complementary slackness conditions involving them, but we know that

$$\epsilon = \frac{\lambda}{2K}, \quad \xi = \frac{\mu}{2K}.$$

Also since the constraint

$$\sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j = \nu$$

implies that either there is some i_0 such that $\lambda_{i_0} > 0$ or there is some j_0 such that $\mu_{j_0} > 0$, we have $\epsilon_{i_0} > 0$ or $\xi_{j_0} > 0$, which means that at least one point is misclassified, so Problem (SVM_{s5}) should only be used when the sets $\{u_i\}$ and $\{v_j\}$ are *not* linearly separable. We can solve for η using the active constraints corresponding to any i_0 such that $\lambda_{i_0} > 0$ or any j_0 such that $\mu_{j_0} > 0$.

We can also use the fact that the optimality gap is 0 to find η. We have

$$\frac{1}{2}w^\top w + \frac{b^2}{2} - \nu\eta + K(\epsilon^\top\epsilon + \xi^\top\xi) = -\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(X^\top X + \begin{pmatrix}\mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top \\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top\end{pmatrix} + \frac{1}{2K}I_{p+q}\right)\begin{pmatrix}\lambda \\ \mu\end{pmatrix},$$

so we get

$$\nu\eta = K(\lambda^\top\lambda + \mu^\top\mu) + \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(X^\top X\begin{pmatrix}\mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top \\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top\end{pmatrix} + \frac{1}{4K}I_{p+q}\right)\begin{pmatrix}\lambda \\ \mu\end{pmatrix}.$$

The above confirms that at optimality we have $\eta \geq 0$.

The "kernelized" version of Problem (SVM_{s5}) is the following:

Soft margin kernel SVM (SVM_{s5}):

$$\begin{aligned}
\text{minimize} \quad & \frac{1}{2}\langle w, w\rangle + \frac{1}{2}b^2 - \nu\eta + \frac{1}{p+q}(\epsilon^\top \epsilon + \xi^\top \xi) \\
\text{subject to} \quad & \\
& \langle w, \varphi(u_i)\rangle - b \geq \eta - \epsilon_i, \qquad i = 1, \ldots, p \\
& -\langle w, \varphi(v_j)\rangle + b \geq \eta - \xi_j, \qquad j = 1, \ldots, q.
\end{aligned}$$

Tracing through the derivation of the dual program, we obtain

$$\begin{aligned}
\text{minimize} \quad & \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(\mathbf{K} + \begin{pmatrix}\mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top \\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top\end{pmatrix} + \frac{p+q}{2}I_{p+q}\right)\begin{pmatrix}\lambda \\ \mu\end{pmatrix} \\
\text{subject to} \quad & \\
& \sum_{i=1}^{p}\lambda_i + \sum_{j=1}^{q}\mu_j = \nu \\
& \lambda_i \geq 0, \quad i = 1, \ldots, p \\
& \mu_j \geq 0, \quad j = 1, \ldots, q,
\end{aligned}$$

where $\mathbf{K}$ is the kernel matrix of Section 32.1. Then w, b, and $f(x)$ are obtained exactly as in Section 32.5.

32.7 Summary and Comparison of the SVM Methods

In this chapter we considered six variants for solving the soft margin binary classification problem for two sets of points $\{u_i\}_{i=1}^p$ and $\{v_j\}_{j=1}^q$ using support vector classification methods. The objective is to find a separating hyperplane $H_{w,b}$ of equation $w^\top x - b = 0$. We also try to find two "margin hyperplanes" $H_{w,b+\delta}$ of equation $w^\top x - b - \delta = 0$ and $H_{w,b-\delta}$ of equation $w^\top x - b + \delta = 0$ such that δ is as big as possible and yet the number of misclassified points is minimized, which is achieved by allowing an error $\epsilon_i \geq 0$ for every point u_i, in the sense that the constraint

$$w^\top u_i - b \geq \delta - \epsilon_i$$

should hold, and an error $\xi_j \geq 0$ for every point v_j, in the sense that the constraint

$$-w^\top v_j + b \geq \delta - \xi_j$$

should hold.

The goal is to design an objective function that minimizes ϵ and ξ and maximizes δ. The optimization problem should also solve for w and b, and for this some constraint has to be placed on w. Another goal is to try to use the dual program to solve the optimization problem, because the solutions involve inner products, and thus the problem is amenable to a generalization using kernel functions.

The first attempt, which is to use the objective function

$$J(w, \epsilon, \xi, b, \delta) = -\delta + K \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} \mathbf{1}_{p+q}$$

and the constraint $w^\top w \leq 1$ does not work very well, because this constraint needs to be guarded by a Lagrange multiplier $\gamma \geq 0$, and as a result, minimizing the Lagrangian L to find the dual function G gives an equation for solving w of the form

$$2\gamma w = -X^\top \begin{pmatrix} \lambda \\ \mu \end{pmatrix},$$

but if the sets $\{u_i\}_{i=1}^p$ and $\{v_j\}_{j=1}^q$ are not linearly separable, then an optimal solution may occurs for $\gamma = 0$, in which case it is impossible to determine w. This is Problem (SVM_{s1}) considered in Section 32.1.

Soft margin SVM (SVM_{s1}):

$$\begin{aligned} \text{minimize} \quad & -\delta + K\Big(\sum_{i=1}^p \epsilon_i + \sum_{j=1}^q \xi_j\Big) \\ \text{subject to} \quad & \\ & w^\top u_i - b \geq \delta - \epsilon_i, \quad \epsilon_i \geq 0 \qquad i = 1, \ldots, p \\ & -w^\top v_j + b \geq \delta - \xi_j, \quad \xi_j \geq 0 \qquad j = 1, \ldots, q \\ & w^\top w \leq 1. \end{aligned}$$

It is customary to write $\ell = p + q$.

It is shown in Section 32.1 that the dual program is equivalent to the following minimization program:

$$\begin{aligned} \text{minimize} \quad & \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \\ \text{subject to} \quad & \\ & \sum_{i=1}^p \lambda_i = \sum_{j=1}^q \mu_j = \frac{1}{2} \\ & 0 \leq \lambda_i \leq K, \quad i = 1, \ldots, p \\ & 0 \leq \mu_j \leq K, \quad j = 1, \ldots, q. \end{aligned}$$

Observe that the constraints imply that K must be chosen so that

$$K \geq \max\left\{\frac{1}{2p}, \frac{1}{2q}\right\}.$$

If the optimal value is 0, then $\gamma = 0$ and $X\begin{pmatrix}\lambda\\ \mu\end{pmatrix} = 0$, so in this case it is not possible to determine w. However, if the optimal value is > 0, then once a solution for λ and μ is obtained, we have

$$\gamma = \frac{1}{2}\left(\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda\\ \mu\end{pmatrix}\right)^{1/2}$$
$$w = \frac{1}{2\gamma}\left(\sum_{i=1}^{p}\lambda_i u_i - \sum_{j=1}^{q}\mu_j v_j\right),$$

so we get

$$w = \frac{\displaystyle\sum_{i=1}^{p}\lambda_i u_i - \sum_{j=1}^{q}\mu_j v_j}{\left(\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda\\ \mu\end{pmatrix}\right)^{1/2}},$$

If the following mild hypothesis holds then b and δ can be found.

Standard Margin Hypothesis for (SVM$_{s1}$). There is some index i_0 such that $0 < \lambda_{i_0} < K$ and there is some index j_0 such that $0 < \mu_{j_0} < K$. This means that some u_{i_0} is correctly classified and on the blue margin, and some v_{j_0} is correctly classified and on the red margin.

If the **Standard Margin Hypothesis** for (SVM$_{s1}$) holds then $\epsilon_{i_0} = 0$ and $\mu_{j_0} = 0$, and then we have the active equations

$$w^\top u_{i_0} - b = \delta \quad \text{and} \quad -w^\top v_{j_0} + b = \delta,$$

and we obtain the value of b and δ as

$$b = \frac{1}{2}(w^\top u_{i_0} + w^\top v_{j_0})$$
$$\delta = \frac{1}{2}(w^\top u_{i_0} - w^\top v_{j_0}).$$

The second more successful approach is to add the term $(1/2)w^\top w$ to the objective function and to drop the constraint $w^\top w \leq 1$. Then there are several variants of this method, depending on the choice of the regularizing term involving ϵ and ξ (linear or quadratic), how

the margin is delt with (implicitly with the term 1 or expliclity with a term η), and whether the term $(1/2)b^2$ is added to the objective function or not.

These methods all share the property that if the primal problem has an optimal solution with $w \neq 0$, then the dual problem always determines w, and then under mild conditions that we call standard margin hypotheses, b and η can be determined. Then ϵ and ξ can be determined using the constraints that are active. When $(1/2)b^2$ is added to the objective function, b is determined by the equation

$$b = -(\mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu).$$

All these problems are convex and the constraints are qualified, so the duality gap is zero, and if the primal has an optimal solution with $w \neq 0$, then it follows that $\eta \geq 0$.

We now consider five variants in more details.

(1) *Basic soft margin SVM*: (SVM_{s2}).

This is the optimization problem in which the regularization term $K \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} \mathbf{1}_{p+q}$ is linear and the margin δ is given by $\delta = 1/\|w\|$:

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2} w^\top w + K \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} \mathbf{1}_{p+q} \\ &\text{subject to} \\ &\qquad w^\top u_i - b \geq 1 - \epsilon_i, \quad \epsilon_i \geq 0 \qquad i = 1, \ldots, p \\ &\qquad -w^\top v_j + b \geq 1 - \xi_j, \quad \xi_j \geq 0 \qquad j = 1, \ldots, q. \end{aligned}$$

This problem is the classical one discussed in all books on machine learning or pattern analysis, for instance Vapnik [110], Bishop [18], and Shawe–Taylor and Christianini [96]. It is shown in Section 32.2 that the dual program is

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} - \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \mathbf{1}_{p+q} \\ &\text{subject to} \\ &\qquad \sum_{i=1}^{p} \lambda_i = \sum_{j=1}^{q} \mu_j \\ &\qquad 0 \leq \lambda_i \leq K, \quad i = 1, \ldots, p \\ &\qquad 0 \leq \mu_j \leq K, \quad j = 1, \ldots, q. \end{aligned}$$

We can use the dual program to solve the primal. Once $\lambda \geq 0, \mu \geq 0$ have been found, w is given by

$$w = -X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} = \sum_{i=1}^{p} \lambda_i u_i - \sum_{j=1}^{q} \mu_j v_j,$$

but b is not determined by the dual.

The complementary slackness conditions imply that if $\epsilon_i > 0$ then $\lambda_i = K$, and if $\xi_j > 0$, then $\mu_j = K$. Consequently, if $\lambda_i < K$ then $\epsilon_i = 0$ and u_i is correctly classified, and similarly if $\mu_j < K$ then $\xi_j = 0$ and v_j is correctly classified.

A priori nothing prevents the situation where $\lambda_i = K$ for all nonzero λ_i or $\mu_j = K$ for all nonzero μ_j. If this happens, we can rerun the optimization method with a larger value of K. If the following mild hypothesis holds then b can be found.

Standard Margin Hypothesis for (SVM_{s2}). There is some index i_0 such that $0 < \lambda_{i_0} < K$ and there is some index j_0 such that $0 < \mu_{j_0} < K$. This means that some u_{i_0} is correctly classified and on the blue margin, and some v_{j_0} is correctly classified and on the red margin.

If the **Standard Margin Hypothesis** for (SVM_{s2}) holds then $\epsilon_{i_0} = 0$ and $\mu_{j_0} = 0$, and then we have the active equations

$$w^\top u_{i_0} - b = 1 \quad \text{and} \quad -w^\top v_{j_0} + b = 1,$$

and we obtain

$$b = \frac{1}{2}(w^\top u_{i_0} + w^\top v_{j_0}).$$

(2) *Basic Soft margin ν-SVM Problem* ($\text{SVM}_{s2'}$).

This a generalization of Problem (SVM_{s2}) for a version of the soft margin SVM coming from Problem (SVM_{h2}), obtained by adding an extra degree of freedom, namely instead of the margin $\delta = 1/\|w\|$, we use the margin $\delta = \eta/\|w\|$ where η is some positive constant that we wish to maximize. To do so, we add a term $-K_m\eta$ to the objective function. We have the following optimization problem:

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}w^\top w - K_m\eta + K_s \begin{pmatrix}\epsilon^\top & \xi^\top\end{pmatrix} \mathbf{1}_{p+q} \\
&\text{subject to} \\
&\qquad w^\top u_i - b \geq \eta - \epsilon_i, \quad \epsilon_i \geq 0 \qquad i = 1, \ldots, p \\
&\qquad -w^\top v_j + b \geq \eta - \xi_j, \quad \xi_j \geq 0 \qquad j = 1, \ldots, q \\
&\qquad \eta \geq 0,
\end{aligned}$$

where $K_m > 0$ and $K_s > 0$ are fixed constants that can be adjusted to determine the influence of η and the regularizing term.

This version of the SVM problem was first discussed in Schölkopf, Smola, Williamson, and Bartlett [87] under the name of *ν-SVC*, and also used in Schölkopf, Platt, Shawe–Taylor, and Smola [86].

In order for the problem to have a solution we must pick K_m and K_s so that

$$K_m \leq \min\{2pK_s, 2qK_s\}.$$

It is shown in Section 32.3 that the dual program is

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda\\ \mu\end{pmatrix}\\
&\text{subject to}\\
&\qquad \sum_{i=1}^{p} \lambda_i = \sum_{j=1}^{q} \mu_j\\
&\qquad \sum_{i=1}^{p} \lambda_i + \sum_{j=1}^{q} \mu_j \geq K_m\\
&\qquad 0 \leq \lambda_i \leq K_s, \quad i = 1, \ldots, p\\
&\qquad 0 \leq \mu_j \leq K_s, \quad j = 1, \ldots, q.
\end{aligned}$$

If the primal problem has an optimal solution with $w \neq 0$, then using the fact that the duality gap is zero we can show that $\eta \geq 0$. Thus constraint $\eta \geq 0$ could be omitted. As in the previous case w is given by

$$w = -X \begin{pmatrix}\lambda\\ \mu\end{pmatrix} = \sum_{i=1}^{p} \lambda_i u_i - \sum_{j=1}^{q} \mu_j v_j,$$

but b and η are not determined by the dual.

If we drop the constraint $\eta \geq 0$, then the inequality

$$\sum_{i=1}^{p} \lambda_i + \sum_{j=1}^{q} \mu_j \geq K_m$$

is replaced by the equation

$$\sum_{i=1}^{p} \lambda_i + \sum_{j=1}^{q} \mu_j = K_m.$$

It convenient to define $\nu > 0$ such that

$$K_m = (p+q)K_s\,\nu,$$

that is

$$\nu = \frac{K_m}{(p+q)K_s},$$

so that the objective function $J(w, \epsilon, \xi, b, \eta)$ is given by

$$J(w, \epsilon, \xi, b, \eta) = \frac{1}{2} w^\top w + K \left(-\nu\eta + \frac{1}{p+q} \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} \mathbf{1}_{p+q} \right),$$

with $K = (p+q)K_s$, and so $K_m = K\nu$ and $K_s = K/(p+q)$.

Observe that the condition $K_m \leq \min\{2pK_s, 2qK_s\}$ is equivalent to

$$\nu \leq \min\left\{ \frac{2p}{p+q}, \frac{2q}{p+q} \right\} \leq 1.$$

Since we obtain an equivalent problem by rescaling by a common positive factor, it is convenient to normalize K_s as

$$K_s = \frac{1}{p+q},$$

in which case $K_m = \nu$. This method is called the *ν-support vector machine*.

Under the **Standard Margin Hypothesis** for ($\text{SVM}_{s2'}$), there is some i_0 such that $0 < \lambda_{i_0} < K_s$ and some j_0 such that $0 < \mu_{j_0} < K_s$, and by the complementary slackness conditions $\epsilon_{i_0} = 0$ and $\xi_{j_0} = 0$, so we have the two active constraints

$$w^\top u_{i_0} - b = \eta, \quad -w^\top v_{j_0} + b = \eta,$$

and we can solve for b and η and we get

$$b = \frac{w^\top u_{i_0} + w^\top v_{j_0}}{2} \qquad \eta = \frac{w^\top u_{i_0} - w^\top v_{j_0}}{2}.$$

Proposition 32.1 gives an upper bound on the number of points u_i and the number of points v_j that fail to achieve the margin, and that have margin at most η. As a consequence, if the u_is and v_js are not linearly separable we must pick ν such that $2/(p+q) \leq \nu \leq \min\{2p/(p+q), 2q/(p+q)\}$ for the method to succeed.

We also investigate conditions on ν that ensure that either some point u_i is correctly classified or some point v_i is correctly classified, and the corresponding constraint is active (so that u_i is on the margin, resp. v_j is on the margin). If there are p_f misclassified points u_i and q_f misclassified points v_j, then if $p_f + q_f \geq 3$ and $2/(p+q) < (p_f + q_f)/(p+q)$, then the above property holds; see Proposition 32.2. We also show that if $p_f, q_f \geq 2$ and if $2/(p+q) < 4/(p+q)$, then b and η can be found without reference to the standard margin hypothesis; see Proposition 32.3.

(3) *Basic Quadratic Soft margin ν-SVM Problem* (SVM_{s3}). This is the version of Problem ($\text{SVM}_{s2'}$) in which instead of using the linear function $K_s \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} \mathbf{1}_{p+q}$ as a regularizing

function we use the quadratic function $K(\|\epsilon\|_2^2 + \|\xi\|_2^2)$. The optimization problem is

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}w^\top w - \nu\eta + K(\epsilon^\top\epsilon + \xi^\top\xi)\\
&\text{subject to}\\
&\qquad w^\top u_i - b \geq \eta - \epsilon_i, \qquad i = 1, \ldots, p\\
&\qquad -w^\top v_j + b \geq \eta - \xi_j, \qquad j = 1, \ldots, q\\
&\qquad \eta \geq 0,
\end{aligned}$$

where ν and K are two given positive constants. As we saw earlier, it is convenient to pick $K = 1/(p+q)$.

In this method, it is no longer necessary to require $\epsilon \geq 0$ and $\xi \geq 0$, because an optimal solution satisfies these conditions. We can also omit the constraint $\eta \geq 0$, because for an optimal solution it can be shown using duality that $\eta \geq 0$. It is shown in Section 32.4 that the dual is given by

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(X^\top X + \frac{1}{2K}I_{p+q}\right)\begin{pmatrix}\lambda\\ \mu\end{pmatrix}\\
&\text{subject to}\\
&\qquad \sum_{i=1}^{p}\lambda_i = \sum_{j=1}^{q}\mu_j\\
&\qquad \sum_{i=1}^{p}\lambda_i + \sum_{j=1}^{q}\mu_j \geq \nu\\
&\qquad \lambda_i \geq 0, \quad i = 1, \ldots, p\\
&\qquad \mu_j \geq 0, \quad j = 1, \ldots, q.
\end{aligned}$$

The above program is similar to the program that was obtained for Problem (SVM$_{s2'}$) but the matrix $X^\top X$ is replaced by the matrix $X^\top X + (1/2K)I_{p+q}$, which is positive definite since $K > 0$, and also the inequalities $\lambda_i \leq K$ and $\mu_j \leq K$ no longer hold. However, the constraints imply that there is some i_0 such that $\lambda_{i_0} > 0$ and some j_0 such that $\mu_{j_0} > 0$. If the constraint $\eta \geq 0$ is dropped, then the inequality

$$\sum_{i=1}^{p}\lambda_i + \sum_{j=1}^{q}\mu_j \geq \nu$$

is replaced by the equation

$$\sum_{i=1}^{p}\lambda_i + \sum_{j=1}^{q}\mu_j = \nu.$$

We obtain w from λ and μ, and γ, as in Problem (SVM$_{s2'}$); namely,

$$w = -X\begin{pmatrix}\lambda\\ \mu\end{pmatrix} = \sum_{i=1}^{p}\lambda_i u_i - \sum_{j=1}^{q}\mu_j v_j,$$

but the dual does not determine b and η. However, ϵ and ξ are determined by

$$\epsilon = \frac{\lambda}{2K}, \quad \xi = \frac{\mu}{2K}.$$

Also since the constraints

$$\sum_{i=1}^{p}\lambda_i \geq \frac{\nu}{2} \quad \text{and} \quad \sum_{j=1}^{q}\mu_j \geq \frac{\nu}{2}$$

imply that there is some i_0 such that $\lambda_{i_0} > 0$ and some j_0 such that $\mu_{j_0} > 0$, we have $\epsilon_{i_0} > 0$ and $\xi_{j_0} > 0$, which means that at least two points are misclassified, so Problem (SVM$_{s3}$) should only be used when the sets $\{u_i\}$ and $\{v_j\}$ are *not* linearly separable. We can solve for b and η using the active constraints corresponding to any i_0 such that $\lambda_{i_0} > 0$ and any j_0 such that $\mu_{j_0} > 0$. With this method, there is no need for a standard margin hypothesis.

(4) *Soft margin ν-SVM Problem* (SVM$_{s4}$). This is the variation of Problem (SVM$_{s2'}$) obgtained by adding the term $(1/2)b^2$ to the objective function. The result is that in minimizing the Lagrangian to find the dual function G, not just w but also b is determined. We also suppress the constraint $\eta \geq 0$ which turns out to be redundant. The optimization problem is

$$\begin{aligned}&\text{minimize} \quad \frac{1}{2}w^\top w + \frac{1}{2}b^2 - \nu\eta + K_s\begin{pmatrix}\epsilon^\top & \xi^\top\end{pmatrix}\mathbf{1}_{p+q}\\ &\text{subject to}\\ &\qquad w^\top u_i - b \geq \eta - \epsilon_i, \quad \epsilon_i \geq 0 \qquad i = 1,\ldots,p\\ &\qquad -w^\top v_j + b \geq \eta - \xi_j, \quad \xi_j \geq 0 \qquad j = 1,\ldots,q,\end{aligned}$$

with $K_s = 1/(p+q)$.

It is shown in Section 32.5 that the dual is given by

$$\begin{aligned}&\text{minimize} \quad \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(X^\top X + \begin{pmatrix}\mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top\\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top\end{pmatrix}\right)\begin{pmatrix}\lambda\\ \mu\end{pmatrix}\\ &\text{subject to}\\ &\qquad \sum_{i=1}^{p}\lambda_i + \sum_{j=1}^{q}\mu_j = \nu\\ &\qquad 0 \leq \lambda_i \leq K_s, \quad i = 1,\ldots,p\\ &\qquad 0 \leq \mu_j \leq K_s, \quad j = 1,\ldots,q.\end{aligned}$$

Once a solution for λ and μ is obtained, we have

$$w = -X\begin{pmatrix}\lambda\\ \mu\end{pmatrix} = \sum_{i=1}^{p}\lambda_i u_i - \sum_{j=1}^{q}\mu_j v_j$$
$$b = -\sum_{i=1}^{p}\lambda_i + \sum_{j=1}^{q}\mu_j,$$

but η is not determined by the dual. Note that the constraint

$$\sum_{i=1}^{p}\lambda_i = \sum_{j=1}^{q}\mu_j$$

occurring in the dual of Program ($\mathrm{SVM}_{s2'}$) has been traded for the equation

$$b = -\sum_{i=1}^{p}\lambda_i + \sum_{j=1}^{q}\mu_j$$

determining b. This seems to be an advantage of Problem (SVM_{s4}).

It is also shown that if the primal problem (SVM_{s4}) has an optimal solution with $w \neq 0$, then $\eta \geq 0$. In order for the primal to have a solution we must have

$$\nu \leq 1.$$

Under the **Standard Margin Hypothesis** for (SVM_{s4}), either there is some i_0 such that $0 < \lambda_{i_0} < K_s$ or there is some j_0 such that $0 < \mu_{j_0} < K_s$, and by the complementary slackness conditions $\epsilon_{i_0} = 0$ or $\xi_{j_0} = 0$, so we have

$$w^\top u_{i_0} - b = \eta, \quad \text{or} \quad -w^\top v_{j_0} + b = \eta,$$

and we can solve for η.

Proposition 32.4 gives an upper bound on the number of points u_i and the number of points v_j that fail to achieve the margin, and that have margin at most η. As a consequence, if the u_is and v_js are not linearly separable we must pick ν such that $1/(p+q) \leq \nu \leq 1$ for the method to succeed.

We also investigate conditions on ν that ensure that either some point u_i is correctly classified or some point v_i is correctly classified, and the corresponding constraint is active (so that u_i is on the margin, resp. v_j is on the margin). If there are p_f misclassified points u_i and q_f misclassified points v_j, then if $p_f + q_f \geq 2$ and $1/(p+q) < (p_f + q_f)/(p+q)$, then the above property holds. See Proposition 32.5; this is a slight improvement over Proposition 32.2. We also show that if $p_f + q_f \geq 2$ and if $1/(p+q) < 3/(p+q)$, then η can be found without requiring the standard margin hypothesis; see Proposition 32.6. This is also a slight improvement over Proposition 32.3.

(5) *Quadratic Soft margin ν-SVM Problem* (SVM_{s5}). This is the variant of Problem (SVM_{s3}) in which we add the term $(1/2)b^2$ to the objective function. We also drop the constraint $\eta \geq 0$ which is redundant. We have the following optimization problem:

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}w^\top w + \frac{1}{2}b^2 - \nu\eta + K(\epsilon^\top \epsilon + \xi^\top \xi)\\
&\text{subject to}\\
&\qquad w^\top u_i - b \geq \eta - \epsilon_i, \qquad i = 1, \ldots, p\\
&\qquad -w^\top v_j + b \geq \eta - \xi_j, \qquad j = 1, \ldots, q,
\end{aligned}$$

where ν and K are two given positive constants. As we saw earlier, it is convenient to pick $K = 1/(p+q)$.

It is shown in Section 32.6 that the dual of Program (SVM_{s5}) is given by

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(X^\top X + \begin{pmatrix}\mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top\\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top\end{pmatrix} + \frac{1}{2K}I_{p+q}\right)\begin{pmatrix}\lambda\\ \mu\end{pmatrix}\\
&\text{subject to}\\
&\qquad \sum_{i=1}^{p}\lambda_i + \sum_{j=1}^{q}\mu_j = \nu\\
&\qquad \lambda_i \geq 0, \quad i = 1, \ldots, p\\
&\qquad \mu_j \geq 0, \quad j = 1, \ldots, q.
\end{aligned}$$

This time we obtain w, b, ϵ and ξ from λ and μ:

$$\begin{aligned}
w &= \sum_{i=1}^{p}\lambda_i u_i - \sum_{j=1}^{q}\mu_j v_j\\
b &= -\sum_{i=1}^{p}\lambda_i + \sum_{j=1}^{q}\mu_j\\
\epsilon &= \frac{\lambda}{2K}\\
\xi &= \frac{\mu}{2K}.
\end{aligned}$$

The constraint

$$\sum_{i=1}^{p}\lambda_i = \sum_{j=1}^{q}\mu_j$$

occurring in the dual of Program (SVM_{s3}) has been traded for the equation

$$b = -\sum_{i=1}^{p}\lambda_i + \sum_{j=1}^{q}\mu_j$$

determining b. This seems to be an advantage of Problem (SVM_{s5}).

The constraint

$$\sum_{i=1}^{p} \lambda_i + \sum_{j=1}^{q} \mu_j = \nu$$

implies that either there is some i_0 such that $\lambda_{i_0} > 0$ or there is some j_0 such that $\mu_{j_0} > 0$, we have $\epsilon_{i_0} > 0$ or $\xi_{j_0} > 0$, which means that at least one point is misclassified, so Problem (SVM_{s5}) should only be used when the sets $\{u_i\}$ and $\{v_j\}$ are *not* linearly separable. We can solve for η using the active constraints corresponding to any i_0 such that $\lambda_{i_0} > 0$ or any j_0 such that $\mu_{j_0} > 0$. Using duality, it can be shown that if the primal has an optimal solution with $w \neq 0$, then $\eta \geq 0$.

These methods all have a kernelized version.

In summary, from a theoretical point of view, Problems (SVM_{s4}) and (SVM_{s5}) seem to have more advantages than the others since they determine at least w and b, but this remains to be verified experimentally.

Chapter 33

Total Orthogonal Families in Hilbert Spaces

33.1 Total Orthogonal Families (Hilbert Bases), Fourier Coefficients

We conclude our quick tour of Hilbert spaces by showing that the notion of orthogonal basis can be generalized to Hilbert spaces. However, the useful notion is not the usual notion of a basis, but a notion which is an abstraction of the concept of Fourier series. Every element of a Hilbert space is the "sum" of its Fourier series.

Definition 33.1. Given a Hilbert space E, a family $(u_k)_{k\in K}$ of nonnull vectors is an *orthogonal family* iff the u_k are pairwise orthogonal, i.e., $\langle u_i, u_j\rangle = 0$ for all $i \neq j$ $(i, j \in K)$, and an *orthonormal family* iff $\langle u_i, u_j\rangle = \delta_{i,j}$, for all $i, j \in K$. A *total orthogonal family (or system)* or *Hilbert basis* is an orthogonal family that is dense in E. This means that for every $v \in E$, for every $\epsilon > 0$, there is some finite subset $I \subseteq K$ and some family $(\lambda_i)_{i\in I}$ of complex numbers, such that

$$\left\| v - \sum_{i\in I} \lambda_i u_i \right\| < \epsilon.$$

Given an orthogonal family $(u_k)_{k\in K}$, for every $v \in E$, for every $k \in K$, the scalar $c_k = \langle v, u_k\rangle / \|u_k\|^2$ is called the *k-th Fourier coefficient of v over $(u_k)_{k\in K}$*.

Remark: The terminology Hilbert basis is misleading, because a Hilbert basis $(u_k)_{k\in K}$ is not necessarily a basis in the algebraic sense. Indeed, in general, $(u_k)_{k\in K}$ does not span E. Intuitively, it takes linear combinations of the u_k's with infinitely many nonnull coefficients to span E. Technically, this is achieved in terms of limits. In order to avoid the confusion between bases in the algebraic sense and Hilbert bases, some authors refer to algebraic bases as *Hamel bases* and to total orthogonal families (or Hilbert bases) as *Schauder bases*.

Given an orthogonal family $(u_k)_{k\in K}$, for any finite subset I of K, we often call sums of the form $\sum_{i\in I}\lambda_i u_i$ *partial sums of Fourier series*, and if these partial sums converge to a limit denoted as $\sum_{k\in K} c_k u_k$, we call $\sum_{k\in K} c_k u_k$ a *Fourier series*.

However, we have to make sense of such sums! Indeed, when K is unordered or uncountable, the notion of limit or sum has not been defined. This can be done as follows (for more details, see Dixmier [35]):

Definition 33.2. Given a normed vector space E (say, a Hilbert space), for any nonempty index set K, we say that a family $(u_k)_{k\in K}$ of vectors in E is *summable with sum* $v \in E$ iff for every $\epsilon > 0$, there is some finite subset I of K, such that,

$$\Big\| v - \sum_{j\in J} u_j \Big\| < \epsilon$$

for every finite subset J with $I \subseteq J \subseteq K$. We say that the family $(u_k)_{k\in K}$ is *summable* iff there is some $v \in E$ such that $(u_k)_{k\in K}$ is summable with sum v. A family $(u_k)_{k\in K}$ is a *Cauchy family* iff for every $\epsilon > 0$, there is a finite subset I of K, such that,

$$\Big\| \sum_{j\in J} u_j \Big\| < \epsilon$$

for every finite subset J of K with $I \cap J = \emptyset$,

If $(u_k)_{k\in K}$ is summable with sum v, we usually denote v as $\sum_{k\in K} u_k$. The following technical proposition will be needed:

Proposition 33.1. *Let E be a complete normed vector space (say, a Hilbert space).*

(1) For any nonempty index set K, a family $(u_k)_{k\in K}$ is summable iff it is a Cauchy family.

(2) Given a family $(r_k)_{k\in K}$ of nonnegative reals $r_k \geq 0$, if there is some real number $B > 0$ such that $\sum_{i\in I} r_i < B$ for every finite subset I of K, then $(r_k)_{k\in K}$ is summable and $\sum_{k\in K} r_k = r$, where r is least upper bound of the set of finite sums $\sum_{i\in I} r_i$ $(I \subseteq K)$.

Proof. (1) If $(u_k)_{k\in K}$ is summable, for every finite subset I of K, let

$$u_I = \sum_{i\in I} u_i \quad \text{and} \quad u = \sum_{k\in K} u_k$$

For every $\epsilon > 0$, there is some finite subset I of K such that

$$\|u - u_L\| < \epsilon/2$$

for all finite subsets L such that $I \subseteq L \subseteq K$. For every finite subset J of K such that $I \cap J = \emptyset$, since $I \subseteq I \cup J \subseteq K$ and $I \cup J$ is finite, we have

$$\|u - u_{I\cup J}\| < \epsilon/2 \quad \text{and} \quad \|u - u_I\| < \epsilon/2,$$

and since

$$\|u_{I\cup J} - u_I\| \leq \|u_{I\cup J} - u\| + \|u - u_I\|$$

and $u_{I\cup J} - u_I = u_J$ since $I \cap J = \emptyset$, we get

$$\|u_J\| = \|u_{I\cup J} - u_I\| < \epsilon,$$

which is the condition for $(u_k)_{k\in K}$ to be a Cauchy family.

Conversely, assume that $(u_k)_{k\in K}$ is a Cauchy family. We define inductively a decreasing sequence (X_n) of subsets of E, each of diameter at most $1/n$, as follows: For $n = 1$, since $(u_k)_{k\in K}$ is a Cauchy family, there is some finite subset J_1 of K such that

$$\|u_J\| < 1/2$$

for every finite subset J of K with $J_1 \cap J = \emptyset$. We pick some finite subset J_1 with the above property, and we let $I_1 = J_1$ and

$$X_1 = \{u_I \mid I_1 \subseteq I \subseteq K,\ I \text{ finite}\}.$$

For $n \geq 1$, there is some finite subset J_{n+1} of K such that

$$\|u_J\| < 1/(2n+2)$$

for every finite subset J of K with $J_{n+1} \cap J = \emptyset$. We pick some finite subset J_{n+1} with the above property, and we let $I_{n+1} = I_n \cup J_{n+1}$ and

$$X_{n+1} = \{u_I \mid I_{n+1} \subseteq I \subseteq K,\ I \text{ finite}\}.$$

Since $I_n \subseteq I_{n+1}$, it is obvious that $X_{n+1} \subseteq X_n$ for all $n \geq 1$. We need to prove that each X_n has diameter at most $1/n$. Since J_n was chosen such that

$$\|u_J\| < 1/(2n)$$

for every finite subset J of K with $J_n \cap J = \emptyset$, and since $J_n \subseteq I_n$, it is also true that

$$\|u_J\| < 1/(2n)$$

for every finite subset J of K with $I_n \cap J = \emptyset$ (since $I_n \cap J = \emptyset$ and $J_n \subseteq I_n$ implies that $J_n \cap J = \emptyset$). Then, for every two finite subsets J, L such that $I_n \subseteq J, L \subseteq K$, we have

$$\|u_{J-I_n}\| < 1/(2n) \quad \text{and} \quad \|u_{L-I_n}\| < 1/(2n),$$

and since

$$\|u_J - u_L\| \leq \|u_J - u_{I_n}\| + \|u_{I_n} - u_L\| = \|u_{J-I_n}\| + \|u_{L-I_n}\|,$$

we get

$$\|u_J - u_L\| < 1/n,$$

which proves that $\delta(X_n) \leq 1/n$. Now, if we consider the sequence of closed sets $(\overline{X_n})$, we still have $\overline{X_{n+1}} \subseteq \overline{X_n}$, and by Proposition 28.4, $\delta(\overline{X_n}) = \delta(X_n) \leq 1/n$, which means that $\lim_{n\to\infty} \delta(\overline{X_n}) = 0$, and by Proposition 28.4, $\bigcap_{n=1}^{\infty} \overline{X_n}$ consists of a single element u. We claim that u is the sum of the family $(u_k)_{k\in K}$.

For every $\epsilon > 0$, there is some $n \geq 1$ such that $n > 2/\epsilon$, and since $u \in \overline{X_m}$ for all $m \geq 1$, there is some finite subset J_0 of K such that $I_n \subseteq J_0$ and

$$\|u - u_{J_0}\| < \epsilon/2,$$

where I_n is the finite subset of K involved in the definition of X_n. However, since $\delta(X_n) \leq 1/n$, for every finite subset J of K such that $I_n \subseteq J$, we have

$$\|u_J - u_{J_0}\| \leq 1/n < \epsilon/2,$$

and since

$$\|u - u_J\| \leq \|u - u_{J_0}\| + \|u_{J_0} - u_J\|,$$

we get

$$\|u - u_J\| < \epsilon$$

for every finite subset J of K with $I_n \subseteq J$, which proves that u is the sum of the family $(u_k)_{k\in K}$.

(2) Since every finite sum $\sum_{i\in I} r_i$ is bounded by the uniform bound B, the set of these finite sums has a least upper bound $r \leq B$. For every $\epsilon > 0$, since r is the least upper bound of the finite sums $\sum_{i\in I} r_i$ (where I finite, $I \subseteq K$), there is some finite $I \subseteq K$ such that

$$\left| r - \sum_{i\in I} r_i \right| < \epsilon,$$

and since $r_k \geq 0$ for all $k \in K$, we have

$$\sum_{i\in I} r_i \leq \sum_{j\in J} r_j$$

whenever $I \subseteq J$, which shows that

$$\left| r - \sum_{j\in J} r_j \right| \leq \left| r - \sum_{i\in I} r_i \right| < \epsilon$$

for every finite subset J such that $I \subseteq J \subseteq K$, proving that $(r_k)_{k\in K}$ is summable with sum $\sum_{k\in K} r_k = r$. □

Remark: The notion of summability implies that the sum of a family $(u_k)_{k\in K}$ is independent of any order on K. In this sense, it is a kind of "commutative summability". More precisely, it is easy to show that for every bijection $\varphi\colon K \to K$ (intuitively, a reordering of K), the family $(u_k)_{k\in K}$ is summable iff the family $(u_l)_{l\in\varphi(K)}$ is summable, and if so, they have the same sum.

The following proposition gives some of the main properties of Fourier coefficients. Among other things, at most countably many of the Fourier coefficient may be nonnull, and the partial sums of a Fourier series converge. Given an orthogonal family $(u_k)_{k\in K}$, we let $U_k = \mathbb{C}u_k$, and $p_{U_k}\colon E \to U_k$ is the projection of E onto U_k.

Proposition 33.2. *Let E be a Hilbert space, $(u_k)_{k\in K}$ an orthogonal family in E, and V the closure of the subspace generated by $(u_k)_{k\in K}$. The following properties hold:*

(1) For every $v \in E$, for every finite subset $I \subseteq K$, we have

$$\sum_{i\in I} |c_i|^2 \leq \|v\|^2,$$

where the c_k are the Fourier coefficients of v.

(2) For every vector $v \in E$, if $(c_k)_{k\in K}$ are the Fourier coefficients of v, the following conditions are equivalent:

(2a) $v \in V$

(2b) The family $(c_k u_k)_{k\in K}$ is summable and $v = \sum_{k\in K} c_k u_k$.

(2c) The family $(|c_k|^2)_{k\in K}$ is summable and $\|v\|^2 = \sum_{k\in K} |c_k|^2$;

(3) The family $(|c_k|^2)_{k\in K}$ is summable, and we have the Bessel inequality:

$$\sum_{k\in K} |c_k|^2 \leq \|v\|^2.$$

As a consequence, at most countably many of the c_k may be nonzero. The family $(c_k u_k)_{k\in K}$ forms a Cauchy family, and thus, the Fourier series $\sum_{k\in K} c_k u_k$ converges in E to some vector $u = \sum_{k\in K} c_k u_k$. Furthermore, $u = p_V(v)$.

Proof. (1) Let

$$u_I = \sum_{i\in I} c_i u_i$$

for any finite subset I of K. We claim that $v - u_I$ is orthogonal to u_i for every $i \in I$. Indeed,

$$\begin{aligned}
\langle v - u_I, u_i\rangle &= \left\langle v - \sum_{j\in I} c_j u_j, u_i \right\rangle \\
&= \langle v, u_i\rangle - \sum_{j\in I} c_j \langle u_j, u_i\rangle \\
&= \langle v, u_i\rangle - c_i\|u_i\|^2 \\
&= \langle v, u_i\rangle - \langle v, u_i\rangle = 0,
\end{aligned}$$

since $\langle u_j, u_i\rangle = 0$ for all $i \neq j$ and $c_i = \langle v, u_i\rangle / \|u_i\|^2$. As a consequence, we have

$$\begin{aligned}\|v\|^2 &= \Big\|v - \sum_{i\in I} c_i u_i + \sum_{i\in I} c_i u_i\Big\|^2 \\ &= \Big\|v - \sum_{i\in I} c_i u_i\Big\|^2 + \Big\|\sum_{i\in I} c_i u_i\Big\|^2 \\ &= \Big\|v - \sum_{i\in I} c_i u_i\Big\|^2 + \sum_{i\in I} |c_i|^2,\end{aligned}$$

since the u_i are pairwise orthogonal, that is,

$$\|v\|^2 = \Big\|v - \sum_{i\in I} c_i u_i\Big\|^2 + \sum_{i\in I} |c_i|^2.$$

Thus,

$$\sum_{i\in I} |c_i|^2 \leq \|v\|^2,$$

as claimed.

(2) We prove the chain of implications $(a) \Rightarrow (b) \Rightarrow (c) \Rightarrow (a)$.

$(a) \Rightarrow (b)$: If $v \in V$, since V is the closure of the subspace spanned by $(u_k)_{k\in K}$, for every $\epsilon > 0$, there is some finite subset I of K and some family $(\lambda_i)_{i\in I}$ of complex numbers, such that

$$\Big\|v - \sum_{i\in I} \lambda_i u_i\Big\| < \epsilon.$$

Now, for every finite subset J of K such that $I \subseteq J$, we have

$$\begin{aligned}\Big\|v - \sum_{i\in I} \lambda_i u_i\Big\|^2 &= \Big\|v - \sum_{j\in J} c_j u_j + \sum_{j\in J} c_j u_j - \sum_{i\in I} \lambda_i u_i\Big\|^2 \\ &= \Big\|v - \sum_{j\in J} c_j u_j\Big\|^2 + \Big\|\sum_{j\in J} c_j u_j - \sum_{i\in I} \lambda_i u_i\Big\|^2,\end{aligned}$$

since $I \subseteq J$ and the u_j (with $j \in J$) are orthogonal to $v - \sum_{j\in J} c_j u_j$ by the argument in (1), which shows that

$$\Big\|v - \sum_{j\in J} c_j u_j\Big\| \leq \Big\|v - \sum_{i\in I} \lambda_i u_i\Big\| < \epsilon,$$

and thus, that the family $(c_k u_k)_{k\in K}$ is summable with sum v, so that

$$v = \sum_{k\in K} c_k u_k.$$

$(b) \Rightarrow (c)$: If $v = \sum_{k \in K} c_k u_k$, then for every $\epsilon > 0$, there some finite subset I of K, such that

$$\left\| v - \sum_{j \in J} c_j u_j \right\| < \sqrt{\epsilon},$$

for every finite subset J of K such that $I \subseteq J$, and since we proved in (1) that

$$\|v\|^2 = \left\| v - \sum_{j \in J} c_j u_j \right\|^2 + \sum_{j \in J} |c_j|^2,$$

we get

$$\|v\|^2 - \sum_{j \in J} |c_j|^2 < \epsilon,$$

which proves that $(|c_k|^2)_{k \in K}$ is summable with sum $\|v\|^2$.

$(c) \Rightarrow (a)$: Finally, if $(|c_k|^2)_{k \in K}$ is summable with sum $\|v\|^2$, for every $\epsilon > 0$, there is some finite subset I of K such that

$$\|v\|^2 - \sum_{j \in J} |c_j|^2 < \epsilon^2$$

for every finite subset J of K such that $I \subseteq J$, and again, using the fact that

$$\|v\|^2 = \left\| v - \sum_{j \in J} c_j u_j \right\|^2 + \sum_{j \in J} |c_j|^2,$$

we get

$$\left\| v - \sum_{j \in J} c_j u_j \right\| < \epsilon,$$

which proves that $(c_k u_k)_{k \in K}$ is summable with sum $\sum_{k \in K} c_k u_k = v$, and $v \in V$.

(3) Since $\sum_{i \in I} |c_i|^2 \leq \|v\|^2$ for every finite subset I of K, by Proposition 33.1, the family $(|c_k|^2)_{k \in K}$ is summable. The Bessel inequality

$$\sum_{k \in K} |c_k|^2 \leq \|v\|^2$$

is an obvious consequence of the inequality $\sum_{i \in I} |c_i|^2 \leq \|v\|^2$ (for every finite $I \subseteq K$). Now, for every natural number $n \geq 1$, if K_n is the subset of K consisting of all c_k such that $|c_k| \geq 1/n$, the number of elements in K_n is at most

$$\sum_{k \in K_n} |nc_k|^2 \leq n^2 \sum_{k \in K} |c_k|^2 \leq n^2 \|v\|^2,$$

which is finite, and thus, at most a countable number of the c_k may be nonzero.

Since $(|c_k|^2)_{k\in K}$ is summable with sum c, for every $\epsilon > 0$, there is some finite subset I of K such that

$$\sum_{j\in J} |c_j|^2 < \epsilon^2$$

for every finite subset J of K such that $I \cap J = \emptyset$. Since

$$\Big\| \sum_{j\in J} c_j u_j \Big\|^2 = \sum_{j\in J} |c_j|^2,$$

we get

$$\Big\| \sum_{j\in J} c_j u_j \Big\| < \epsilon.$$

This proves that $(c_k u_k)_{k\in K}$ is a Cauchy family, which, by Proposition 33.1, implies that $(c_k u_k)_{k\in K}$ is summable, since E is complete. Thus, the Fourier series $\sum_{k\in K} c_k u_k$ is summable, with its sum denoted $u \in V$.

Since $\sum_{k\in K} c_k u_k$ is summable with sum u, for every $\epsilon > 0$, there is some finite subset I_1 of K such that

$$\Big\| u - \sum_{j\in J} c_j u_j \Big\| < \epsilon$$

for every finite subset J of K such that $I_1 \subseteq J$. By the triangle inequality, for every finite subset I of K,

$$\Big\| u - v \Big\| \leq \Big\| u - \sum_{i\in I} c_i u_i \Big\| + \Big\| \sum_{i\in I} c_i u_i - v \Big\|.$$

By (2), every $w \in V$ is the sum of its Fourier series $\sum_{k\in K} \lambda_k u_k$, and for every $\epsilon > 0$, there is some finite subset I_2 of K such that

$$\Big\| w - \sum_{j\in J} \lambda_j u_j \Big\| < \epsilon$$

for every finite subset J of K such that $I_2 \subseteq J$. By the triangle inequality, for every finite subset I of K,

$$\Big\| v - \sum_{i\in I} \lambda_i u_i \Big\| \leq \|v - w\| + \Big\| w - \sum_{i\in I} \lambda_i u_i \Big\|.$$

Letting $I = I_1 \cup I_2$, since we showed in (2) that

$$\Big\| v - \sum_{i\in I} c_i u_i \Big\| \leq \Big\| v - \sum_{i\in I} \lambda_i u_i \Big\|$$

for every finite subset I of K, we get

$$\begin{aligned}\|u-v\| &\leq \Big\|u-\sum_{i\in I} c_i u_i\Big\| + \Big\|\sum_{i\in I} c_i u_i - v\Big\| \\ &\leq \Big\|u-\sum_{i\in I} c_i u_i\Big\| + \Big\|\sum_{i\in I} \lambda_i u_i - v\Big\| \\ &\leq \Big\|u-\sum_{i\in I} c_i u_i\Big\| + \|v-w\| + \Big\|w-\sum_{i\in I}\lambda_i u_i\Big\|,\end{aligned}$$

and thus

$$\|u-v\| \leq \|v-w\| + 2\epsilon.$$

Since this holds for every $\epsilon > 0$, we have

$$\|u-v\| \leq \|v-w\|$$

for all $w \in V$, i.e. $\|v-u\| = d(v,V)$, with $u \in V$, which proves that $u = p_V(v)$. □

33.2 The Hilbert Space $l^2(K)$ and the Riesz-Fischer Theorem

Proposition 33.2 suggests looking at the space of sequences $(z_k)_{k\in K}$ (where $z_k \in \mathbb{C}$) such that $(|z_k|^2)_{k\in K}$ is summable. Indeed, such spaces are Hilbert spaces, and it turns out that every Hilbert space is isomorphic to one of those. Such spaces are the infinite-dimensional version of the spaces $\mathbb{C}^n$ under the usual Euclidean norm.

Definition 33.3. Given any nonempty index set K, the space $l^2(K)$ is the set of all sequences $(z_k)_{k\in K}$, where $z_k \in \mathbb{C}$, such that $(|z_k|^2)_{k\in K}$ is summable, i.e., $\sum_{k\in K} |z_k|^2 < \infty$.

Remarks:

(1) When K is a finite set of cardinality n, $l^2(K)$ is isomorphic to $\mathbb{C}^n$.

(2) When $K = \mathbb{N}$, the space $l^2(\mathbb{N})$ corresponds to the space l^2 of Example 2 in Section 11.1. In that example, we claimed that l^2 was a Hermitian space, and in fact, a Hilbert space. We now prove this fact for any index set K.

Proposition 33.3. *Given any nonempty index set K, the space $l^2(K)$ is a Hilbert space under the Hermitian product*

$$\langle (x_k)_{k\in K}, (y_k)_{k\in K}\rangle = \sum_{k\in K} x_k \overline{y_k}.$$

The subspace consisting of sequences $(z_k)_{k\in K}$ such that $z_k = 0$, except perhaps for finitely many k, is a dense subspace of $l^2(K)$.

Proof. First, we need to prove that $l^2(K)$ is a vector space. Assume that $(x_k)_{k\in K}$ and $(y_k)_{k\in K}$ are in $l^2(K)$. This means that $(|x_k|^2)_{k\in K}$ and $(|y_k|^2)_{k\in K}$ are summable, which, in view of Proposition 33.1, is equivalent to the existence of some positive bounds A and B such that $\sum_{i\in I}|x_i|^2 < A$ and $\sum_{i\in I}|y_i|^2 < B$, for every finite subset I of K. To prove that $(|x_k + y_k|^2)_{k\in K}$ is summable, it is sufficient to prove that there is some $C > 0$ such that $\sum_{i\in I}|x_i + y_i|^2 < C$ for every finite subset I of K. However, the parallelogram inequality implies that

$$\sum_{i\in I}|x_i + y_i|^2 \leq \sum_{i\in I} 2(|x_i|^2 + |y_i|^2) \leq 2(A + B),$$

for every finite subset I of K, and we conclude by Proposition 33.1. Similarly, for every $\lambda \in \mathbb{C}$,

$$\sum_{i\in I}|\lambda x_i|^2 \leq \sum_{i\in I}|\lambda|^2|x_i|^2 \leq |\lambda|^2 A,$$

and $(\lambda_k x_k)_{k\in K}$ is summable. Therefore, $l^2(K)$ is a vector space.

By the Cauchy-Schwarz inequality,

$$\sum_{i\in I}|x_i\overline{y_i}| \leq \sum_{i\in I}|x_i||y_i| \leq \big(\sum_{i\in I}|x_i|^2\big)^{1/2}\big(\sum_{i\in I}|y_i|^2\big)^{1/2} \leq \sum_{i\in I}(|x_i|^2 + |y_i|^2)/2 \leq (A + B)/2,$$

for every finite subset I of K. Here, we used the fact that

$$4CD \leq (C + D)^2,$$

which is equivalent to

$$(C - D)^2 \geq 0.$$

By Proposition 33.1, $(|x_k\overline{y_k}|)_{k\in K}$ is summable. The customary language is that $(x_k\overline{y_k})_{k\in K}$ is absolutely summable. However, it is a standard fact that this implies that $(x_k\overline{y_k})_{k\in K}$ is summable (For every $\epsilon > 0$, there is some finite subset I of K such that

$$\sum_{j\in J}|x_j\overline{y_j}| < \epsilon$$

for every finite subset J of K such that $I \cap J = \emptyset$, and thus

$$|\sum_{j\in J}x_j\overline{y_j}| \leq \sum_{i\in J}|x_j\overline{y_j}| < \epsilon,$$

proving that $(x_k\overline{y_k})_{k\in K}$ is a Cauchy family, and thus summable). We still have to prove that $l^2(K)$ is complete.

Consider a sequence $((\lambda_k^n)_{k\in K})_{n\geq 1}$ of sequences $(\lambda_k^n)_{k\in K} \in l^2(K)$, and assume that it is a Cauchy sequence. This means that for every $\epsilon > 0$, there is some $N \geq 1$ such that

$$\sum_{k\in K}|\lambda_k^m - \lambda_k^n|^2 < \epsilon^2$$

for all $m, n \geq N$. For every fixed $k \in K$, this implies that

$$|\lambda_k^m - \lambda_k^n| < \epsilon$$

for all $m, n \geq N$, which shows that $(\lambda_k^n)_{n\geq 1}$ is a Cauchy sequence in $\mathbb{C}$. Since $\mathbb{C}$ is complete, the sequence $(\lambda_k^n)_{n\geq 1}$ has a limit $\lambda_k \in \mathbb{C}$. We claim that $(\lambda_k)_{k\in K} \in l^2(K)$ and that this is the limit of $((\lambda_k^n)_{k\in K})_{n\geq 1}$.

Given any $\epsilon > 0$, the fact that $((\lambda_k^n)_{k\in K})_{n\geq 1}$ is a Cauchy sequence implies that there is some $N \geq 1$ such that for every finite subset I of K, we have

$$\sum_{i\in I} |\lambda_i^m - \lambda_i^n|^2 < \epsilon/4$$

for all $m, n \geq N$. Let $p = |I|$. Then,

$$|\lambda_i^m - \lambda_i^n| < \frac{\sqrt{\epsilon}}{2\sqrt{p}}$$

for every $i \in I$. Since λ_i is the limit of $(\lambda_i^n)_{n\geq 1}$, we can find some n large enough so that

$$|\lambda_i^n - \lambda_i| < \frac{\sqrt{\epsilon}}{2\sqrt{p}}$$

for every $i \in I$. Since

$$|\lambda_i^m - \lambda_i| \leq |\lambda_i^m - \lambda_i^n| + |\lambda_i^n - \lambda_i|,$$

we get

$$|\lambda_i^m - \lambda_i| < \frac{\sqrt{\epsilon}}{\sqrt{p}},$$

and thus,

$$\sum_{i\in I} |\lambda_i^m - \lambda_i|^2 < \epsilon,$$

for all $m \geq N$. Since the above holds for every finite subset I of K, by Proposition 33.1, we get

$$\sum_{k\in K} |\lambda_k^m - \lambda_k|^2 < \epsilon,$$

for all $m \geq N$. This proves that $(\lambda_k^m - \lambda_k)_{k\in K} \in l^2(K)$ for all $m \geq N$, and since $l^2(K)$ is a vector space and $(\lambda_k^m)_{k\in K} \in l^2(K)$ for all $m \geq 1$, we get $(\lambda_k)_{k\in K} \in l^2(K)$. However,

$$\sum_{k\in K} |\lambda_k^m - \lambda_k|^2 < \epsilon$$

for all $m \geq N$, means that the sequence $(\lambda_k^m)_{k\in K}$ converges to $(\lambda_k)_{k\in K} \in l^2(K)$. The fact that the subspace consisting of sequences $(z_k)_{k\in K}$ such that $z_k = 0$ except perhaps for finitely many k is a dense suspace of $l^2(K)$ is left as an easy exercise. □

Remark: The subspace consisting of all sequences $(z_k)_{k\in K}$ such that $z_k = 0$, except perhaps for finitely many k, provides an example of a subspace which is not closed in $l^2(K)$. Indeed, this space is strictly contained in $l^2(K)$, since there are countable sequences of nonnull elements in $l^2(K)$ (why?).

We just need two more propositions before being able to prove that every Hilbert space is isomorphic to some $l^2(K)$.

Proposition 33.4. *Let E be a Hilbert space, and $(u_k)_{k\in K}$ an orthogonal family in E. The following properties hold:*

(1) For every family $(\lambda_k)_{k\in K} \in l^2(K)$, the family $(\lambda_k u_k)_{k\in K}$ is summable. Furthermore, $v = \sum_{k\in K} \lambda_k u_k$ is the only vector such that $c_k = \lambda_k$ for all $k \in K$, where the c_k are the Fourier coefficients of v.

(2) For any two families $(\lambda_k)_{k\in K} \in l^2(K)$ and $(\mu_k)_{k\in K} \in l^2(K)$, if $v = \sum_{k\in K} \lambda_k u_k$ and $w = \sum_{k\in K} \mu_k u_k$, we have the following equation, also called Parseval identity:

$$\langle v, w\rangle = \sum_{k\in K} \lambda_k \overline{\mu_k}.$$

Proof. (1) The fact that $(\lambda_k)_{k\in K} \in l^2(K)$ means that $(|\lambda_k|^2)_{k\in K}$ is summable. The proof given in Proposition 33.2 (3) applies to the family $(|\lambda_k|^2)_{k\in K}$ (instead of $(|c_k|^2)_{k\in K}$), and yields the fact that $(\lambda_k u_k)_{k\in K}$ is summable. Letting $v = \sum_{k\in K} \lambda_k u_k$, recall that $c_k = \langle v, u_k\rangle / \|u_k\|^2$. Pick some $k \in K$. Since $\langle -, -\rangle$ is continuous, for every $\epsilon > 0$, there is some $\eta > 0$ such that

$$|\langle v, u_k\rangle - \langle w, u_k\rangle| < \epsilon \|u_k\|^2$$

whenever

$$\|v - w\| < \eta.$$

However, since for every $\eta > 0$, there is some finite subset I of K such that

$$\left\| v - \sum_{j\in J} \lambda_j u_j \right\| < \eta$$

for every finite subset J of K such that $I \subseteq J$, we can pick $J = I \cup \{k\}$, and letting $w = \sum_{j\in J} \lambda_j u_j$, we get

$$\left| \langle v, u_k\rangle - \left\langle \sum_{j\in J} \lambda_j u_j, u_k \right\rangle \right| < \epsilon \|u_k\|^2.$$

However,

$$\langle v, u_k\rangle = c_k \|u_k\|^2 \quad \text{and} \quad \left\langle \sum_{j\in J} \lambda_j u_j, u_k \right\rangle = \lambda_k \|u_k\|^2,$$

and thus, the above proves that $|c_k - \lambda_k| < \epsilon$ for every $\epsilon > 0$, and thus, that $c_k = \lambda_k$.

(2) Since $\langle -, - \rangle$ is continuous, for every $\epsilon > 0$, there are some $\eta_1 > 0$ and $\eta_2 > 0$, such that

$$| \langle x, y \rangle | < \epsilon$$

whenever $\|x\| < \eta_1$ and $\|y\| < \eta_2$. Since $v = \sum_{k \in K} \lambda_k u_k$ and $w = \sum_{k \in K} \mu_k u_k$, there is some finite subset I_1 of K such that

$$\left\| v - \sum_{j \in J} \lambda_j u_j \right\| < \eta_1$$

for every finite subset J of K such that $I_1 \subseteq J$, and there is some finite subset I_2 of K such that

$$\left\| w - \sum_{j \in J} \mu_j u_j \right\| < \eta_2$$

for every finite subset J of K such that $I_2 \subseteq J$. Letting $I = I_1 \cup I_2$, we get

$$\left| \left\langle v - \sum_{i \in I} \lambda_i u_i, w - \sum_{i \in I} \mu_i u_i \right\rangle \right| < \epsilon.$$

Furthermore,

$$\begin{aligned} \langle v, w \rangle &= \left\langle v - \sum_{i \in I} \lambda_i u_i + \sum_{i \in I} \lambda_i u_i, w - \sum_{i \in I} \mu_i u_i + \sum_{i \in I} \mu_i u_i \right\rangle \\ &= \left\langle v - \sum_{i \in I} \lambda_i u_i, w - \sum_{i \in I} \mu_i u_i \right\rangle + \sum_{i \in I} \lambda_i \overline{\mu_i}, \end{aligned}$$

since the u_i are orthogonal to $v - \sum_{i \in I} \lambda_i u_i$ and $w - \sum_{i \in I} \mu_i u_i$ for all $i \in I$. This proves that for every $\epsilon > 0$, there is some finite subset I of K such that

$$\left| \langle v, w \rangle - \sum_{i \in I} \lambda_i \overline{\mu_i} \right| < \epsilon.$$

We already know from Proposition 33.3 that $(\lambda_k \overline{\mu_k})_{k \in K}$ is summable, and since $\epsilon > 0$ is arbitrary, we get

$$\langle v, w \rangle = \sum_{k \in K} \lambda_k \overline{\mu_k}.$$

□

The next proposition states properties characterizing Hilbert bases (total orthogonal families).

Proposition 33.5. *Let E be a Hilbert space, and let $(u_k)_{k \in K}$ be an orthogonal family in E. The following properties are equivalent:*

(1) The family $(u_k)_{k\in K}$ is a total orthogonal family.

(2) For every vector $v \in E$, if $(c_k)_{k\in K}$ are the Fourier coefficients of v, then the family $(c_k u_k)_{k\in K}$ is summable and $v = \sum_{k\in K} c_k u_k$.

(3) For every vector $v \in E$, we have the Parseval identity:

$$\|v\|^2 = \sum_{k\in K} |c_k|^2.$$

(4) For every vector $u \in E$, if $\langle u, u_k\rangle = 0$ for all $k \in K$, then $u = 0$.

Proof. The equivalence of (1), (2), and (3), is an immediate consequence of Proposition 33.2 and Proposition 33.4.

(4) If $(u_k)_{k\in K}$ is a total orthogonal family and $\langle u, u_k\rangle = 0$ for all $k \in K$, since $u = \sum_{k\in K} c_k u_k$ where $c_k = \langle u, u_k\rangle/\|u_k\|^2$, we have $c_k = 0$ for all $k \in K$, and $u = 0$.

Conversely, assume that the closure V of $(u_k)_{k\in K}$ is different from E. Then, by Proposition 28.7, we have $E = V \oplus V^\perp$, where $V^\perp$ is the orthogonal complement of V, and $V^\perp$ is nontrivial since $V \neq E$. As a consequence, there is some nonnull vector $u \in V^\perp$. But then, u is orthogonal to every vector in V, and in particular,

$$\langle u, u_k\rangle = 0$$

for all $k \in K$, which, by assumption, implies that $u = 0$, contradicting the fact that $u \neq 0$. □

Remarks:

(1) If E is a Hilbert space and $(u_k)_{k\in K}$ is a total orthogonal family in E, there is a simpler argument to prove that $u = 0$ if $\langle u, u_k\rangle = 0$ for all $k \in K$, based on the continuity of $\langle -, -\rangle$. The argument is to prove that the assumption implies that $\langle v, u\rangle = 0$ for all $v \in E$. Since $\langle -, -\rangle$ is positive definite, this implies that $u = 0$. By continuity of $\langle -, -\rangle$, for every $\epsilon > 0$, there is some $\eta > 0$ such that for every finite subset I of K, for every family $(\lambda_i)_{i\in I}$, for every $v \in E$,

$$\left|\langle v, u\rangle - \left\langle \sum_{i\in I} \lambda_i u_i, u \right\rangle\right| < \epsilon$$

whenever

$$\left\|v - \sum_{i\in I} \lambda_i u_i\right\| < \eta.$$

Since $(u_k)_{k\in K}$ is dense in E, for every $v \in E$, there is some finite subset I of K and some family $(\lambda_i)_{i\in I}$ such that

$$\left\| v - \sum_{i\in I} \lambda_i u_i \right\| < \eta,$$

and since by assumption, $\left\langle \sum_{i\in I} \lambda_i u_i, u \right\rangle = 0$, we get

$$|\langle v, u\rangle| < \epsilon.$$

Since this holds for every $\epsilon > 0$, we must have $\langle v, u\rangle = 0$

(2) If V is any nonempty subset of E, the kind of argument used in the previous remark can be used to prove that $V^\perp$ is closed (even if V is not), and that $V^{\perp\perp}$ is the closure of V.

We will now prove that every Hilbert space has some Hilbert basis. This requires using a fundamental theorem from set theory known as *Zorn's Lemma*, which we quickly review.

Given any set X with a partial ordering $\leq$, recall that a nonempty subset C of X is a *chain* if it is totally ordered (i.e., for all $x, y \in C$, either $x \leq y$ or $y \leq x$). A nonempty subset Y of X is *bounded* iff there is some $b \in X$ such that $y \leq b$ for all $y \in Y$. Some $m \in X$ is *maximal* iff for every $x \in X$, $m \leq x$ implies that $x = m$. We can now state Zorn's Lemma. For more details, see Rudin [82], Lang [62], or Artin [6].

Proposition 33.6. *Given any nonempty partially ordered set X, if every (nonempty) chain in X is bounded, then X has some maximal element.*

We can now prove the existence of Hilbert bases. We define a partial order on families $(u_k)_{k\in K}$ as follows: For any two families $(u_k)_{k\in K_1}$ and $(v_k)_{k\in K_2}$, we say that

$$(u_k)_{k\in K_1} \leq (v_k)_{k\in K_2}$$

iff $K_1 \subseteq K_2$ and $u_k = v_k$ for all $k \in K_1$. This is clearly a partial order.

Proposition 33.7. *Let E be a Hilbert space. Given any orthogonal family $(u_k)_{k\in K}$ in E, there is a total orthogonal family $(u_l)_{l\in L}$ containing $(u_k)_{k\in K}$.*

Proof. Consider the set $\mathcal{S}$ of all orthogonal families greater than or equal to the family $B = (u_k)_{k\in K}$. We claim that every chain in $\mathcal{S}$ is bounded. Indeed, if $C = (C_l)_{l\in L}$ is a chain in $\mathcal{S}$, where $C_l = (u_{k,l})_{k\in K_l}$, the union family

$$(u_k)_{k\in \bigcup_{l\in L} K_l}, \text{ where } u_k = u_{k,l} \text{ whenever } k \in K_l,$$

is clearly an upper bound for C, and it is immediately verified that it is an orthogonal family. By Zorn's Lemma 33.6, there is a maximal family $(u_l)_{l\in L}$ containing $(u_k)_{k\in K}$. If $(u_l)_{l\in L}$ is not dense in E, then its closure V is strictly contained in E, and by Proposition 28.7, the

orthogonal complement $V^\perp$ of V is nontrivial since $V \neq E$. As a consequence, there is some nonnull vector $u \in V^\perp$. But then, u is orthogonal to every vector in $(u_l)_{l\in L}$, and we can form an orthogonal family strictly greater than $(u_l)_{l\in L}$ by adding u to this family, contradicting the maximality of $(u_l)_{l\in L}$. Therefore, $(u_l)_{l\in L}$ is dense in E, and thus, it is a Hilbert basis. □

Remark: It is possible to prove that all Hilbert bases for a Hilbert space E have index sets K of the same cardinality. For a proof, see Bourbaki [21].

Definition 33.4. A Hilbert space E is *separable* if its Hilbert bases are countable.

At last, we can prove that every Hilbert space is isomorphic to some Hilbert space $l^2(K)$ for some suitable K.

Theorem 33.8. *(Riesz-Fischer) For every Hilbert space E, there is some nonempty set K such that E is isomorphic to the Hilbert space $l^2(K)$. More specifically, for any Hilbert basis $(u_k)_{k\in K}$ of E, the maps $f\colon l^2(K) \to E$ and $g\colon E \to l^2(K)$ defined such that*

$$f\left((\lambda_k)_{k\in K}\right) = \sum_{k\in K} \lambda_k u_k \quad \textit{and} \quad g(u) = \left(\langle u, u_k\rangle / \|u_k\|^2\right)_{k\in K} = (c_k)_{k\in K},$$

are bijective linear isometries such that $g \circ f = \mathrm{id}$ and $f \circ g = \mathrm{id}$.

Proof. By Proposition 33.4 (1), the map f is well defined, and it it clearly linear. By Proposition 33.2 (3), the map g is well defined, and it is also clearly linear. By Proposition 33.2 (2b), we have

$$f(g(u)) = u = \sum_{k\in K} c_k u_k,$$

and by Proposition 33.4 (1), we have

$$g(f\left((\lambda_k)_{k\in K}\right)) = (\lambda_k)_{k\in K},$$

and thus $g \circ f = \mathrm{id}$ and $f \circ g = \mathrm{id}$. By Proposition 33.4 (2), the linear map g is an isometry. Therefore, f is a linear bijection and an isometry between $l^2(K)$ and E, with inverse g. □

Remark: The surjectivity of the map $g\colon E \to l^2(K)$ is known as the *Riesz-Fischer* theorem.

Having done all this hard work, we sketch how these results apply to Fourier series. Again, we refer the readers to Rudin [82] or Lang [64, 65] for a comprehensive exposition.

Let $\mathcal{C}(T)$ denote the set of all periodic continuous functions $f\colon [-\pi, \pi] \to \mathbb{C}$ with period 2π. There is a Hilbert space $\mathrm{L}^2(T)$ containing $\mathcal{C}(T)$ and such that $\mathcal{C}(T)$ is dense in $\mathrm{L}^2(T)$, whose inner product is given by

$$\langle f, g\rangle = \int_{-\pi}^{\pi} f(x)\overline{g(x)}dx.$$

The Hilbert space $\mathrm{L}^2(T)$ is the space of *Lebesgue square-integrable periodic functions* (of period 2π).

It turns out that the family $(e^{ikx})_{k\in\mathbb{Z}}$ is a total orthogonal family in $\mathrm{L}^2(T)$, because it is already dense in $\mathcal{C}(T)$ (for instance, see Rudin [82]). Then, the Riesz-Fischer theorem says that for every family $(c_k)_{k\in\mathbb{Z}}$ of complex numbers such that

$$\sum_{k\in\mathbb{Z}} |c_k|^2 < \infty,$$

there is a unique function $f \in \mathrm{L}^2(T)$ such that f is equal to its Fourier series

$$f(x) = \sum_{k\in\mathbb{Z}} c_k e^{ikx},$$

where the Fourier coefficients c_k of f are given by the formula

$$c_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)e^{-ikt}dt.$$

The Parseval theorem says that

$$\sum_{k=-\infty}^{+\infty} c_k\overline{d_k} = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)\overline{g(t)}dt$$

for all $f, g \in \mathrm{L}^2(T)$, where c_k and d_k are the Fourier coefficients of f and g.

Thus, there is an isomorphism between the two Hilbert spaces $\mathrm{L}^2(T)$ and $l^2(\mathbb{Z})$, which is the deep reason why the Fourier coefficients "work". Theorem 33.8 implies that the Fourier series $\sum_{k\in\mathbb{Z}} c_k e^{ikx}$ of a function $f \in \mathrm{L}^2(T)$ converges to f in the L^2-sense, i.e., in the mean-square sense. This does not necessarily imply that the Fourier series converges to f pointwise! This is a subtle issue, and for more on this subject, the reader is referred to Lang [64, 65] or Schwartz [92, 93].

We can also consider the set $\mathcal{C}([-1,1])$ of continuous functions $f\colon [-1,1] \to \mathbb{C}$. There is a Hilbert space $\mathrm{L}^2([-1,1])$ containing $\mathcal{C}([-1,1])$ and such that $\mathcal{C}([-1,1])$ is dense in $\mathrm{L}^2([-1,1])$, whose inner product is given by

$$\langle f, g\rangle = \int_{-1}^{1} f(x)\overline{g(x)}dx.$$

The Hilbert space $\mathrm{L}^2([-1,1])$ is the space of *Lebesgue square-integrable functions* over $[-1,1]$. The Legendre polynomials $P_n(x)$ defined in Example 5 of Section 9.2 (Chapter 9) form a Hilbert basis of $\mathrm{L}^2([-1,1])$. Recall that if we let f_n be the function

$$f_n(x) = (x^2-1)^n,$$

$P_n(x)$ is defined as follows:

$$P_0(x) = 1, \quad \text{and} \quad P_n(x) = \frac{1}{2^n n!} f_n^{(n)}(x),$$

where $f_n^{(n)}$ is the nth derivative of f_n. The reason for the leading coefficient is to get $P_n(1) = 1$. It can be shown with much efforts that

$$P_n(x) = \sum_{0 \le k \le n/2} (-1)^k \frac{(2(n-k))!}{2^n (n-k)! k! (n-2k)!} x^{n-2k}.$$

Bibliography

[1] Ralph Abraham and Jerrold E. Marsden. *Foundations of Mechanics*. Addison Wesley, second edition, 1978.

[2] George E. Andrews, Richard Askey, and Ranjan Roy. *Special Functions*. Cambridge University Press, first edition, 2000.

[3] Tom Apostol. *Analysis*. Addison Wesley, second edition, 1974.

[4] V.I. Arnold. *Mathematical Methods of Classical Mechanics*. GTM No. 102. Springer Verlag, second edition, 1989.

[5] Emil Artin. *Geometric Algebra*. Wiley Interscience, first edition, 1957.

[6] Michael Artin. *Algebra*. Prentice Hall, first edition, 1991.

[7] A. Avez. *Calcul Différentiel*. Masson, first edition, 1991.

[8] Sheldon Axler. *Linear Algebra Done Right*. Undergraduate Texts in Mathematics. Springer Verlag, second edition, 2004.

[9] Marcel Berger. *Géométrie 1*. Nathan, 1990. English edition: Geometry 1, Universitext, Springer Verlag.

[10] Marcel Berger. *Géométrie 2*. Nathan, 1990. English edition: Geometry 2, Universitext, Springer Verlag.

[11] Marcel Berger and Bernard Gostiaux. *Géométrie différentielle: variétés, courbes et surfaces*. Collection Mathématiques. Puf, second edition, 1992. English edition: Differential geometry, manifolds, curves, and surfaces, GTM No. 115, Springer Verlag.

[12] J.E. Bertin. *Algèbre linéaire et géométrie classique*. Masson, first edition, 1981.

[13] Dimitri P. Bertsekas. *Convex Optimization Theory*. Athena Scientific, first edition, 2009.

[14] Dimitri P. Bertsekas. *Convex Optimization Algorithms*. Athena Scientific, first edition, 2015.

[15] Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, third edition, 2016.

[16] Dimitri P. Bertsekas, Angelina Nedić, and Asuman E. Ozdaglar. *Convex Analysis and Optimization*. Athena Scientific, first edition, 2003.

[17] Dimitris Bertsimas and John N. Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, third edition, 1997.

[18] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, first edition, 2006.

[19] Nicolas Bourbaki. *Algèbre, Chapitres 1-3*. Eléments de Mathématiques. Hermann, 1970.

[20] Nicolas Bourbaki. *Algèbre, Chapitres 4-7*. Eléments de Mathématiques. Masson, 1981.

[21] Nicolas Bourbaki. *Espaces Vectoriels Topologiques*. Eléments de Mathématiques. Masson, 1981.

[22] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, first edition, 2004.

[23] Glen E Bredon. *Topology and Geometry*. GTM No. 139. Springer Verlag, first edition, 1993.

[24] Haim Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Universitext. Springer-Verlag, first edition, 2011.

[25] G. Cagnac, E. Ramis, and J. Commeau. *Mathématiques Spéciales, Vol. 3, Géométrie*. Masson, 1965.

[26] Henri Cartan. *Cours de Calcul Différentiel*. Collection Méthodes. Hermann, 1990.

[27] Chih-Chung Chang and Lin Chih-Jen. Training ν-support vector classifiers: Theory and algorithms. *Neural Computation*, 13:2119–2147, 2001.

[28] Yvonne Choquet-Bruhat, Cécile DeWitt-Morette, and Margaret Dillard-Bleick. *Analysis, Manifolds, and Physics, Part I: Basics*. North-Holland, first edition, 1982.

[29] Vasek Chvatal. *Linear Programming*. W.H. Freeman, first edition, 1983.

[30] P.G. Ciarlet. *Introduction to Numerical Matrix Analysis and Optimization*. Cambridge University Press, first edition, 1989. French edition: Masson, 1994.

[31] Timothée Cour and Jianbo Shi. Solving markov random fields with spectral relaxation. In Marita Meila and Xiaotong Shen, editors, *Artifical Intelligence and Statistics*. Society for Artificial Intelligence and Statistics, 2007.

[32] H.S.M. Coxeter. *Introduction to Geometry*. Wiley, second edition, 1989.

[33] James W. Demmel. *Applied Numerical Linear Algebra*. SIAM Publications, first edition, 1997.

[34] Jean Dieudonné. *Algèbre Linéaire et Géométrie Elémentaire*. Hermann, second edition, 1965.

[35] Jacques Dixmier. *General Topology*. UTM. Springer Verlag, first edition, 1984.

[36] Manfredo P. do Carmo. *Differential Geometry of Curves and Surfaces*. Prentice Hall, 1976.

[37] Manfredo P. do Carmo. *Riemannian Geometry*. Birkhäuser, second edition, 1992.

[38] David S. Dummit and Richard M. Foote. *Abstract Algebra*. Wiley, second edition, 1999.

[39] Charles L. Epstein. *Introduction to the Mathematics of Medical Imaging*. SIAM, second edition, 2007.

[40] Olivier Faugeras. *Three-Dimensional Computer Vision, A geometric Viewpoint*. the MIT Press, first edition, 1996.

[41] James Foley, Andries van Dam, Steven Feiner, and John Hughes. *Computer Graphics. Principles and Practice*. Addison-Wesley, second edition, 1993.

[42] David A. Forsyth and Jean Ponce. *Computer Vision: A Modern Approach*. Prentice Hall, first edition, 2002.

[43] Jean Fresnel. *Méthodes Modernes En Géométrie*. Hermann, first edition, 1998.

[44] Jean H. Gallier. *Geometric Methods and Applications, For Computer Science and Engineering*. TAM, Vol. 38. Springer, second edition, 2011.

[45] Jean H. Gallier. Notes on Convex Sets, Polytopes, Polyhedra, Combinatorial Topology, Voronoi Diagrams, and Delaunay Triangulations. Technical report, University of Pennsylvania, CIS Department, Philadelphia, PA 19104, 2016. Book in Preparation.

[46] Walter Gander, Gene H. Golub, and Urs von Matt. A constrained eigenvalue problem. *Linear Algebra and its Applications*, 114/115:815–839, 1989.

[47] Roger Godement. *Cours d'Algèbre*. Hermann, first edition, 1963.

[48] Gene H. Golub. Some modified eigenvalue problems. *SIAM Review*, 15(2):318–334, 1973.

[49] H. Golub, Gene and F. Van Loan, Charles. *Matrix Computations*. The Johns Hopkins University Press, third edition, 1996.

[50] A. Gray. *Modern Differential Geometry of Curves and Surfaces*. CRC Press, second edition, 1997.

[51] Jacques Hadamard. *Leçons de Géométrie Elémentaire. I Géométrie Plane*. Armand Colin, thirteenth edition, 1947.

[52] Jacques Hadamard. *Leçons de Géométrie Elémentaire. II Géométrie dans l'Espace*. Armand Colin, eighth edition, 1949.

[53] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, second edition, 2009.

[54] Sigurdur Helgason. *Groups and Geometric Analysis. Integral Geometry, Invariant Differential Operators and Spherical Functions*. MSM, Vol. 83. AMS, first edition, 2000.

[55] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, first edition, 1990.

[56] Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, first edition, 1994.

[57] Ramesh Jain, Rangachar Katsuri, and Brian G. Schunck. *Machine Vision*. McGraw-Hill, first edition, 1995.

[58] Hoffman Kenneth and Kunze Ray. *Linear Algebra*. Prentice Hall, second edition, 1971.

[59] D. Kincaid and W. Cheney. *Numerical Analysis*. Brooks/Cole Publishing, second edition, 1996.

[60] A.N. Kolmogorov and S.V. Fomin. *Introductory Real Analysis*. Dover, first edition, 1975.

[61] Erwin Kreyszig. *Differential Geometry*. Dover, first edition, 1991.

[62] Serge Lang. *Algebra*. Addison Wesley, third edition, 1993.

[63] Serge Lang. *Differential and Riemannian Manifolds*. GTM No. 160. Springer Verlag, third edition, 1995.

[64] Serge Lang. *Real and Functional Analysis*. GTM 142. Springer Verlag, third edition, 1996.

[65] Serge Lang. *Undergraduate Analysis*. UTM. Springer Verlag, second edition, 1997.

[66] Peter Lax. *Linear Algebra and Its Applications*. Wiley, second edition, 2007.

[67] N. N. Lebedev. *Special Functions and Their Applications*. Dover, first edition, 1972.

[68] David G. Luenberger. *Optimization by Vector Space Methods*. Wiley, first edition, 1997.

[69] David G. Luenberger and Yinyu Ye. *Linear and Nonlinear Programming*. Verlag, fourth edition, 2016.

[70] Saunders Mac Lane and Garrett Birkhoff. *Algebra*. Macmillan, first edition, 1967.

[71] Jerrold E. Marsden and J.R. Hughes, Thomas. *Mathematical Foundations of Elasticity*. Dover, first edition, 1994.

[72] Jiri Matousek and Bernd Gartner. *Understanding and Using Linear Programming*. Universitext. Springer Verlag, first edition, 2007.

[73] Dimitris N. Metaxas. *Physics-Based Deformable Models*. Kluwer Academic Publishers, first edition, 1997.

[74] Carl D. Meyer. *Matrix Analysis and Applied Linear Algebra*. SIAM, first edition, 2000.

[75] John W. Milnor. *Topology from the Differentiable Viewpoint*. The University Press of Virginia, second edition, 1969.

[76] James R. Munkres. *Analysis on Manifolds*. Addison Wesley, 1991.

[77] James R. Munkres. *Topology*. Prentice Hall, second edition, 2000.

[78] Joseph O'Rourke. *Computational Geometry in C*. Cambridge University Press, second edition, 1998.

[79] Christos H. Papadimitriou and Kenneth Steiglitz. *Combinatorial Optimization. Algorithms and Complexity*. Dover, first edition, 1998.

[80] Dan Pedoe. *Geometry, A comprehensive Course*. Dover, first edition, 1988.

[81] Eugène Rouché and Charles de Comberousse. *Traité de Géométrie*. Gauthier-Villars, seventh edition, 1900.

[82] Walter Rudin. *Real and Complex Analysis*. McGraw Hill, third edition, 1987.

[83] Walter Rudin. *Functional Analysis*. McGraw Hill, second edition, 1991.

[84] Giovanni Sansone. *Orthogonal Functions*. Dover, first edition, 1991.

[85] Berhnard Schölkopf and Alexander J. Smola. *Learning with Kernels*. MIT Press, first edition, 2002.

[86] Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, and Alex J. Smola. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13:1443–1471, 2001.

[87] Bernhard Schölkopf, Alex J. Smola, Robert C. Williamson, and Peter L. Bartlett. New support vector algorithms. *Neural Computation*, 12:1207–1245, 2000.

[88] Alexander Schrijver. *Theory of Linear and Integer Programming*. Wiley, first edition, 1999.

[89] Laurent Schwartz. *Topologie Générale et Analyse Fonctionnelle*. Collection Enseignement des Sciences. Hermann, 1980.

[90] Laurent Schwartz. *Analyse I. Théorie des Ensembles et Topologie*. Collection Enseignement des Sciences. Hermann, 1991.

[91] Laurent Schwartz. *Analyse II. Calcul Différentiel et Equations Différentielles*. Collection Enseignement des Sciences. Hermann, 1992.

[92] Laurent Schwartz. *Analyse III. Calcul Intégral*. Collection Enseignement des Sciences. Hermann, 1993.

[93] Laurent Schwartz. *Analyse IV. Applications à la Théorie de la Mesure*. Collection Enseignement des Sciences. Hermann, 1993.

[94] H. Seifert and W. Threlfall. *A Textbook of Topology*. Academic Press, first edition, 1980.

[95] Denis Serre. *Matrices, Theory and Applications*. GTM No. 216. Springer Verlag, second edition, 2010.

[96] John Shawe-Taylor and Nello Cristianini. *Kernel methods for Pattern Analysis*. Cambridge University Press, first edition, 2004.

[97] Ernst Snapper and Troyer Robert J. *Metric Affine Geometry*. Dover, first edition, 1989.

[98] G.W. Stewart. On the early history of the singular value decomposition. *SIAM review*, 35(4):551–566, 1993.

[99] J.J. Stoker. *Differential Geometry*. Wiley Classics. Wiley-Interscience, first edition, 1989.

[100] Eric J. Stollnitz, Tony D. DeRose, and David H. Salesin. *Wavelets for Computer Graphics Theory and Applications*. Morgan Kaufmann, first edition, 1996.

[101] Gilbert Strang. *Introduction to Applied Mathematics*. Wellesley-Cambridge Press, first edition, 1986.

[102] Gilbert Strang. *Linear Algebra and its Applications*. Saunders HBJ, third edition, 1988.

[103] Gilbert Strang and Nguyen Truong. *Wavelets and Filter Banks*. Wellesley-Cambridge Press, second edition, 1997.

[104] Claude Tisseron. *Géométries affines, projectives, et euclidiennes*. Hermann, first edition, 1994.

[105] L.N. Trefethen and D. Bau III. *Numerical Linear Algebra*. SIAM Publications, first edition, 1997.

[106] Emanuele Trucco and Alessandro Verri. *Introductory Techniques for 3D Computer Vision*. Prentice-Hall, first edition, 1998.

[107] B.L. Van Der Waerden. *Algebra, Vol. 1*. Ungar, seventh edition, 1973.

[108] J.H. van Lint and R.M. Wilson. *A Course in Combinatorics*. Cambridge University Press, second edition, 2001.

[109] Robert J. Vanderbei. *Linear Programming: Foundations and Extensions*. Springer, fourth edition, 2014.

[110] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, first edition, 1998.

[111] Frank Warner. *Foundations of Differentiable Manifolds and Lie Groups*. GTM No. 94. Springer Verlag, first edition, 1983.

[112] Stella X. Yu and Jianbo Shi. Grouping with bias. In Thomas G. Dietterich, Sue Becker, and Zoubin Ghahramani, editors, *Neural Information Processing Systems, Vancouver, Canada, 3-8 Dec. 2001*. MIT Press, 2001.

[113] Gunter Ziegler. *Lectures on Polytopes*. GTM No. 152. Springer Verlag, first edition, 1997.