Applications of Scientific Computation

EAS205, Some Notes

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

October 5, 2022

Contents

Chapter 1

Introduction to Vectors and Matrices

1.1 Vectors and Matrices; Some Motivations

Linear algebra provides a rich language to express problems expressible in terms of systems of (linear) equations, and a powerful set of tools to solve them. A valuable feature this language is that it is very effective at reducing the amount of bookkeeping:

variables and equations are neatly compressed using vectors and matrices.

But it is more than a convenient language. It is a way of thinking. If a problem can be linearized, or at least approximated by a linear system, then it has a better chance to be solved!

We begin by motivating the expressive power of the language of linear algebra on "the" typical linear problem: solving a system of linear equations.

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 some "column vectors." Let u, v, w, and b, be the *vectors* given by

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

and write our linear system

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

as

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

In writing the equation

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

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

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

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

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

For example

$$3u = 3 \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 6 \\ 3 \end{bmatrix}$$

and

$$u + v = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} + \begin{bmatrix} 2 \\ 1 \\ -2 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ -1 \end{bmatrix}.$$

We define $-z$ by

$$-z = - \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} -z_1 \\ -z_2 \\ -z_3 \end{bmatrix}.$$

Observe that

$$(-1)z = -z.$$

Also, note that

$$z + -z = -z + z = 0,$$

where 0 denotes the *zero vector*

$$0 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

If you don't like the fact that the symbol 0 is used both to denote the number 0 and the zero vector, you may denote the zero vector by $\mathbf{0}$.

More generally, you may feel more comfortable to denote vectors using boldface ($\mathbf{z}$ instead of z), but you will quickly get tired of that. You can also use the "arrow notation" $\overrightarrow{z}$, but nobody does that anymore!

Also observe that

$$0z = 0, \quad i.e. \quad 0z = \mathbf{0}.$$

Then,

$$\begin{aligned} x_1u + x_2v + x_3w &= x_1 \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} + x_2 \begin{bmatrix} 2 \\ 1 \\ -2 \end{bmatrix} + x_3 \begin{bmatrix} -1 \\ 1 \\ -2 \end{bmatrix} \\ &= \begin{bmatrix} x_1 \\ 2x_1 \\ x_1 \end{bmatrix} + \begin{bmatrix} 2x_2 \\ x_2 \\ -2x_2 \end{bmatrix} + \begin{bmatrix} -x_3 \\ x_3 \\ -2x_3 \end{bmatrix} \\ &= \begin{bmatrix} x_1 + 2x_2 - x_3 \\ 2x_1 + x_2 + x_3 \\ x_1 - 2x_2 - 2x_3 \end{bmatrix}. \end{aligned}$$

The set of all vectors with three components is denoted by $\mathrm{M}_{3,1}$ (some authors use $\mathbb{R}^{3\times 1}$). The reason for using the notation $\mathrm{M}_{3,1}$ rather than the more conventional notation $\mathbb{R}^3$ is that the elements of $\mathrm{M}_{3,1}$ are *column vectors*; they consist of three rows and a single column, which explains the subscript $3, 1$.

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*.

For the sake of clarity, in this introduction, we will denote the set of column vectors with n components by $\mathrm{M}_{n,1}$.

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*. If we let x_1, x_2, x_3 vary arbitrarily (keeping u, v, w fixed), we get a set of vectors that forms some kind of subspace of $\mathrm{M}_{3,1}$. 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

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

it can be shown that *every* vector in $\mathrm{M}_{3,1}$ can be written as a linear combination of u, v, w. In fact, in this case every vector $z \in \mathrm{M}_{3,1}$ can be written *in a unique way* as a linear combination

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

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,

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

The array

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

is a 3×3 matrix because it consists of 3 *columns* u, v, w also denoted by A^1, A^2, A^3, and 3 *rows* denoted by A_1, A_2, A_3, where

$$u = A^1 = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \qquad v = A^2 = \begin{bmatrix} 2 \\ 1 \\ -2 \end{bmatrix} \qquad w = A^3 = \begin{bmatrix} -1 \\ 1 \\ -2 \end{bmatrix}$$

and

$$A_1 = \begin{bmatrix} 1 & 2 & -1 \end{bmatrix} \qquad A_2 = \begin{bmatrix} 2 & 1 & 1 \end{bmatrix} \qquad A_3 = \begin{bmatrix} 1 & -2 & -2 \end{bmatrix}.$$

Given a matrix, our notation for the columns (use *superscripts*), and for the rows (use *subscripts*), is not universally used.

In `Matlab` a matrix is represented as a sequence of rows separated by semicolons, where the entries in each row are separed by blank spaces. For example, the matrix

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

is represented by

$$A = [1\ \ 2\ \ -1; 2\ \ 1\ \ 1; 1\ \ -2\ \ -2].$$

There is a convenient mechanism to denote the columns and the rows of the matrix A:

The jth column is denoted by $A(:, j)$. For example, the second column is $A(:, 2)$.

The ith row is denoted by $A(i, :)$. For example, the first row is $A(1, :)$.

We use the notation A^j for the jth column and A_i for the ith row.

If we form the vector of unknowns

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

and if we define the product Ax of the matrix A by the vector x by

$$\begin{bmatrix} 1 & 2 & -1 \\ 2 & 1 & 1 \\ 1 & -2 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = x_1 \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} + x_2 \begin{bmatrix} 2 \\ 1 \\ -2 \end{bmatrix} + x_3 \begin{bmatrix} -1 \\ 1 \\ -2 \end{bmatrix};$$

that is

$$Ax = x_1 u + x_2 v + x_3 w,$$

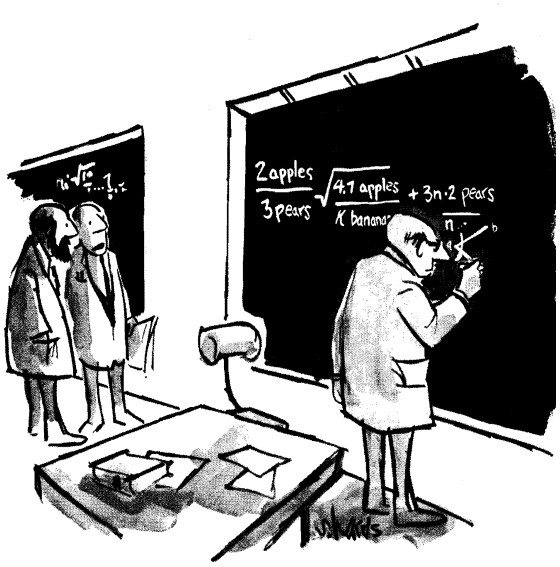

Figure 1.1: The power of abstraction

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

$$x_1u + x_2v + x_3w = Ax = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 1 & 1 \\ 1 & -2 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}.$$

So, our linear system is expressed by

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

or more concisely as

$$Ax = b.$$

Now, what if the vectors u, v, w are *linearly dependent*?

For example, if we consider the vectors

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

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

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

then

$$b = u + v,$$

since

$$u + v = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} + \begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 0 \end{bmatrix} = b,$$

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 system

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

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

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

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 $x \in \mathrm{M}_{3,1}$ and all $\lambda \in \mathbb{R}$, and that

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

for all $u, v \in \mathrm{M}_{3,1}$.

We can view the matrix A as a way of expressing a linear map from $\mathrm{M}_{3,1}$ to $\mathrm{M}_{3,1}$ and solving the system $Ax = b$ amounts to determining whether b belongs to the *image* (or *range*) 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

$$x_1 + 2x_2 - x_3 = 1$$
$$2x_1 + x_2 + x_3 = 2$$
$$x_1 - 2x_2 - 2x_3 = 3$$

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.

There is a *geometric interpretation of vectors* in terms of coordinate systems. For example, a vector u with two components

$$u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

is represented by an arrow whose source is the origin and whose tip is the point of coordinates (u_1, u_2); see Figure 1.2 for an example.

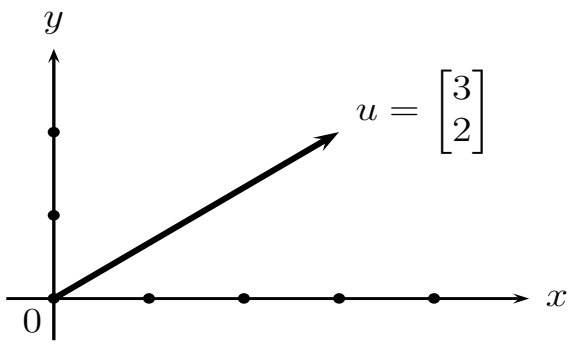

Figure 1.2: Geometric representation of a vector in $\mathbb{R}^2$

A vector u with three components

$$u = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}$$

is also represented by an arrow whose source is the origin and whose tip is the point of coordinates (u_1, u_2, u_3); see Figure 1.3 for an example.

Addition of vectors also has a geometric interpretation. To add two vectors u and v (both with two components), form the *parallelogram* having $0, u, v$ and $w = u + v$ as its corners, where $u+v$ is tip of the diagonal whose source is at the origin; see Figure 1.4 for an example.

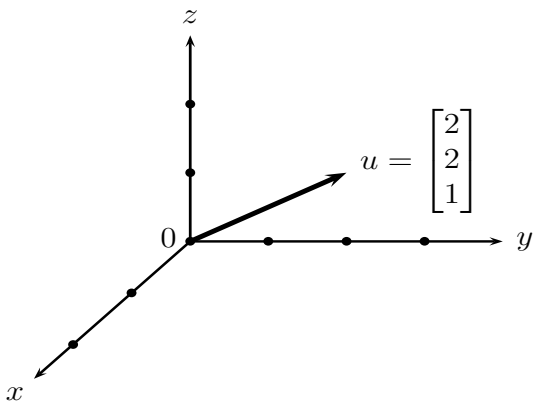

Figure 1.3: Geometric representation of a vector in $\mathbb{R}^3$

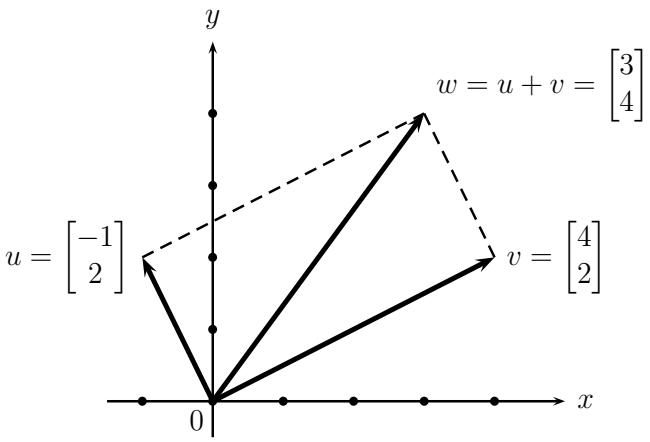

Figure 1.4: Geometric representation of vector addition in $\mathbb{R}^2$

A similar construction applies to vectors with three components, by performing this construction in the plane determined by u and v. (But what happens if u and v belong to the same line?)

The physical interpretation of the sum of vectors is the *resultant* of two forces. What about a geometric interpretation for vectors with four or more components?

It seems that humans have trouble visualizing spaces with more than three dimensions.

Vectors with four components

$$u = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix}$$

can be interpreted as 3D points in a space-time world, where the component u_4 corresponds to time, but this is not entirely satisfactory. What about vectors with five or more components?

Perhaps it is best to give up on a geometric interpretation for vectors with 4 or more components and take a more *algebraic view*.

It is still useful to keep as much as a *geometric point of view* when dealing with vectors and matrices, but one has to be cautious that our *intuition is often wrong in dimension greater than three*!

Still, we should consider seriously the advice given by the famous mathematician John Tate:

Think geometrically; Prove algebraically.

Before proceeding any further, let us recall that *sequences are not sets*.

In a set with n elements $\{u_1, \ldots, u_n\}$, the elements u_i are *all distinct* ($u_i \neq u_j$ for all $i \neq j$), and their *order does not matter*.

On the other hand, in a sequence with n elements $(u_1, \ldots, u_n)$, we may have $u_i = u_j$ for $i \neq j$, and the *order is important*. For example, the sequences $(1, 1, 2)$ and $(2, 1, 1)$ are different, and they are not sets. On the other hand, $\{1, 2\} = \{2, 1\}$ is a set.

If X is any set, the set of sequences with n-elements $(x_1, \ldots, x_n)$ with $x_i \in X$ (where $n \geq 1$) is denoted by X^n. Sequences in X^n are also called *n-tuples*. When $n = 1$, we identify the one-element sequence (x) with x, and thus we identify X^1 with X.

1.2 Linear Combinations, Linear Independence, Matrices

The set of n-tuples in $\mathbb{R}^n$ is an important example of a fundamental concept of linear algebra, a *vector space*.

Definition 1.1. The set $\mathbb{R}^n$ with the *addition operation* $+$ and the *scalar multiplication* $\cdot$ defined below is called a *vector space*:

$$(u_1, \ldots, u_n) + (v_1, \ldots, v_n) = (u_1 + v_1, \ldots, u_n + v_n)$$
$$\lambda \cdot (u_1, \ldots, u_n) = (\lambda u_1, \ldots, \lambda u_n),$$

for all $u_i, v_i \in \mathbb{R}$ and all $\lambda \in \mathbb{R}$.

The *zero vector* of $\mathbb{R}^n$ is the n-tuple $(0, \ldots, 0)$, denoted by $\mathbf{0}$ or even 0.

For any $u = (u_1, \ldots, u_n) \in \mathbb{R}^n$, we define $-u$ by

$$-u = -(u_1, \ldots, u_n) = (-u_1, \ldots, -u_n).$$

Remark: We say that $\mathbb{R}^n$ has *dimension n*. It is convenient to let $\mathbb{R}^0 = \{0\}$, the space consisting of the unique element 0.

For simplicity of notation, we usually drop the symbol for scalar multiplication (the dot) and write λu instead of $\lambda \cdot u$.

Important Notational Convention

Following Strang, we adopt the convention that writing a vector in $\mathbb{R}^n$ as a *horizontal* n-tuple

$$u = (u_1, \ldots, u_n)$$

is just a way of saving space, and that the *vertical* notation for the vector u, as a *column vector*, is

$$u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}.$$

The above is really a special matrix with a single column (an $n \times 1$ matrix). In particular, the n-tuple $(u_1, \ldots, u_n)$ should not be confused with the *row vector*

$$[u_1\ u_2\ \ldots\ u_n],$$

which is really a special matrix with a single row (an $1 \times n$ matrix).

Remark: Note that in a row matrix the entries are *not* separated by commas, but in an n-tuple, they are. Strang's notational trick is to use the brackets "[", "]" to denote matrices, and parentheses "(", ")" to denote tuples.

Beware that this notation is not universally accepted!

Many books use the parentheses "(", ")" to denote matrices. In this case, a row vector is denoted by

$$(u_1\ u_2\ \ldots\ u_n),$$

with no separating commas. Often, these books feel compelled to denote a column vector by

$$(u_1\, u_2\, \ldots\, u_n)^\top,$$

the transpose of a row vector, but we find this ugly (and unnecessary)!

It seems to us that Strang's convention is better. In fact, after a while, when we know what's going on, we can use parentheses to denote matrices without any risk of confusion.

For the time being, to be absolutely clear, let's stick to brackets to denote matrices.

Definition 1.2. Given a p-tuple $(u_1, \ldots, u_p)$ of vectors $u_i \in \mathbb{R}^n$, a *linear combination* of the u_i is a vector (in $\mathbb{R}^n$) of the form

$$\lambda_1 u_1 + \cdots + \lambda_p u_p,$$

with $\lambda_1, \ldots, \lambda_p$ some scalars in $\mathbb{R}$ (not necessarily distinct).

For example, if

$$u = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \qquad v = \begin{bmatrix} 4 \\ 2 \end{bmatrix} \qquad w = \begin{bmatrix} 2 \\ 4 \end{bmatrix} \qquad x = \begin{bmatrix} -2 \\ -1 \end{bmatrix}$$

then

$$2u + v - w + 2x$$

is a linear combination of (u, v, w, x). Actually,

$$2u + v - w + 2x = 0.$$

Observe that if we pick $\lambda_1 = \cdots = \lambda_p = 0$, then

$$0u_1 + \cdots + 0u_p = 0,$$

no matter what $u_1, \ldots, u_p$ are! We call this the *trivial linear combination*.

Given a fixed p-tuple $(u_1, \ldots, u_p)$ of vectors $u_i \in \mathbb{R}^n$, it is interesting to understand geometrically what the set of all linear combinations

$$\lambda_1 u_1 + \cdots + \lambda_p u_p$$

looks like when the $\lambda_1, \ldots, \lambda_p$ vary arbitrarily.

When $p = 1$, this is easy: we are dealing with rescaled versions λu_1 of the vector u_1.

If $u_1 = 0$, we get the space reduced to 0. Let us now assume that $u_1 \neq 0$.

In $\mathbb{R}^2$, this gives us a *line* through the origin. Similarly, in $\mathbb{R}^3$, we get a *line* through the origin. For $n \geq 4$, we still say that we get a *line* (through the origin).

Let us now consider the case $p = 2$.

1. If both $u_1 = 0$ and $u_2 = 0$, we get the space reduced to 0.

2. If $u_1 = 0$ or $u_2 = 0$ but not both, we get a line as in the case $p = 1$.

3. If $u_1 \neq 0$ and $u_2 \neq 0$, it is possible that u_1 and u_2 are not independent, which means that $u_2 = \lambda u_1$ for some λ (actually $\lambda \neq 0$). In this case, we get a line, as in the case $p = 1$.

4. The last possibility is that u_1 and u_2 are independent, which means that there is no λ such that $u_2 = \lambda u_1$, and there is no λ such that $u_1 = \lambda u_2$. It is easy to see that this implies that $u_1 \neq 0$ and $u_2 \neq 0$.

We claim that case (4) is equivalent to the following implication:

$$\textit{If } \lambda_1 u_1 + \lambda_2 u_2 = 0, \textit{ then } \lambda_1 = \lambda_2 = 0. \tag{$*$}$$

Proof. First, assume that case (4) holds and assume that

$$\lambda_1 u_1 + \lambda_2 u_2 = 0$$

for some λ_1, λ_2. We need to prove that $\lambda_1 = \lambda_2 = 0$. We proceed by contradiction.

First assume that $\lambda_1 \neq 0$. Since $\lambda_1 u_1 = -\lambda_2 u_2$, we get

$$u_1 = (-\lambda_2/\lambda_1)u_2,$$

contradicting the fact that u_1 is not a multiple of u_2.

Next, if $\lambda_2 \neq 0$, because $\lambda_2 u_2 = -\lambda_1 u_1$, we get

$$u_2 = (-\lambda_1/\lambda_2)u_1,$$

contradicting the fact that u_2 is not a multiple of u_1.

Let us now prove the converse, namely that condition $(*)$ implies that u_2 is not a multiple of u_1 and u_1 is not a multiple of u_2.

We proceed by contradiction and there are two cases.

1. If $u_1 = \lambda u_2$, then $\lambda_1 u_1 + \lambda_2 u_2 = 0$, with $\lambda_1 = 1$ and $\lambda_2 = -\lambda$, a contradiction.

2. If $u_2 = \lambda u_1$, then $\lambda_1 u_1 + \lambda_2 u_2 = 0$, with $\lambda_1 = \lambda$ and $\lambda_2 = -1$, a contradiction.

□

When u_1 and u_2 are independent (which implies that they are nonzero), the linear combinations $\lambda_1 u_1 + \lambda_2 u_2$ yield distinct vectors for distinct pairs (λ_1, λ_2), and these vectors form a *plane* through the origin.

This is easily confirmed when $n = 2, 3$, and for $n \geq 4$, we still say that we have a *plane* (through the origin).

If we now consider the case $p = 3$, we find that the case analysis is even more complicated and depends on the dependence or independence of the vectors u_1, u_2, u_3. This suggests taking a closer look at the notion of linear independence.

The key idea is that p vectors $(u_1, \ldots, u_p)$ are *linearly independent* if none of them can be expessed as a linear combination of the others.

This implies that $u_i \neq 0$ for all i, because the zero vector is equal to the linear combination of any sequence of arbitrary vectors if we pick all the scalars to be equal zero.

Equivalently, p vectors are *linearly dependent* if it is possible to form a nontrivial linear combination which yields the zero vector; that is, there exists p scalars, $\lambda_1, \ldots, \lambda_p$, *not all zero*, such that

$$\lambda_1 u_1 + \cdots + \lambda_p u_p = 0.$$

In this case, one of the u_i can be expressed as a linear combination of the others. For example, when $p = 3$, if

$$\lambda_1 u_1 + \lambda_2 u_2 + \lambda_3 u_3 = 0,$$

and say, $\lambda_2 \neq 0$, we can write

$$u_2 = (-\lambda_1/\lambda_2) u_1 + (-\lambda_3/\lambda_2) u_3.$$

Definition 1.3. We say that $p \geq 1$ vectors $(u_1, \ldots, u_p)$ with u_i in $\mathbb{R}^n$ are *linearly independent* if the equation

$$\lambda_1 u_1 + \cdots + \lambda_p u_p = 0$$

implies that $\lambda_1 = \cdots = \lambda_p = 0$.

We say that p vectors $(u_1, \ldots, u_p)$ with u_i in $\mathbb{R}^n$ are *linearly dependent* if they are not linearly independent; this is equivalent to the fact that there exist some scalars $\lambda_1, \ldots, \lambda_p$, with *some* $\lambda_i \neq 0$, such that

$$\lambda_1 u_1 + \cdots + \lambda_p u_p = 0.$$

Note that if $(u_1, \ldots, u_p)$ are linearly independent, then $u_i \neq 0$ for $i = 1, \ldots, p$, and the u_i are all distinct ($u_i \neq u_j$, whenever $i \neq j$). We must also have $n \geq 1$. Every tuple containing some zero vector is linearly dependent.

Going back to the case of 3 vectors (u_1, u_2, u_3) in $\mathbb{R}^3$, if they are linearly independent, then the set of their linear combinations fills $\mathbb{R}^3$.

We say that they form a *three-dimensional space*, and this is also the case in $\mathbb{R}^n$ for $n \geq 4$.

One of the major goals of linear algebra is to find "good" methods to check that some vectors are linearly independent.

Observe that p vectors $u_1, \ldots, u_p$, each in $\mathbb{R}^n$, can be arranged as a two-dimensional array A with n rows and p columns, where the j-th column of A is u_j.

$$A = \begin{bmatrix} & & & \\ u_1 & u_2 & \cdots & u_p \\ & & & \end{bmatrix}$$

Recall that that in the above, each u_i is viewed as a *column vector*. The above array is a *$n \times p$ matrix*. For example, the four vectors

$$u = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \qquad v = \begin{bmatrix} 4 \\ 2 \end{bmatrix} \qquad w = \begin{bmatrix} 2 \\ 4 \end{bmatrix} \qquad x = \begin{bmatrix} -2 \\ -1 \end{bmatrix}$$

in $\mathbb{R}^2$ can be used to form the 2×4 matrix

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

It is important to figure our the maximum number of linearly independent columns, and the maximum number of linearly independent rows, of a matrix.

We will prove later that this number is the same (a fundamental result of linear algebra)! It is called the *rank* of a matrix. We will present several algorithms for finding the rank of a matrix (Gaussian elimination, LU, QR, SVD).

We've talked about matrices without giving a definition. Here it is.

Definition 1.4. An *$m \times n$ matrix* $A = (a_{ij})$ is an array of $m \times n$ scalars $a_{ij} \in \mathbb{R}$, with $m \geq 1$ and $n \geq 1$, denoted by

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

We say that A has m *rows* and n *columns*. The index i is the index of the ith row, and the index j is the index of the jth column.

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

$$[a_{11} \cdots a_{1n}].$$

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

$$\begin{bmatrix} a_{11} \\ \vdots \\ a_{m1} \end{bmatrix}$$

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 $\mathrm{M}_{m,n}$. Some authors use the notation $\mathbb{R}^{m \times n}$. Matrices in $\mathrm{M}_{m,n}$ are also called *rectangular matrices*.

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 M_n.

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

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

The square matrix 0_n of dimension n containing 0 everywhere is called the *zero matrix*. For example, when $n = 3$,

$$0_3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Matrices can be added or rescaled (provided that they have the same dimensions m, n).

Definition 1.5. Given two $m \times n$ matrices $A = (a_{ij})$ and $B = (b_{ij})$, we define their *sum* $A + B$ as the matrix $C = (c_{ij})$ such that $c_{ij} = a_{ij} + b_{ij}$; that is,

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \dots & b_{mn} \end{bmatrix}$$

$$= \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \dots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \dots & a_{2n} + b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \dots & a_{mn} + b_{mn} \end{bmatrix}.$$

We define the matrix $-A$ as the matrix $(-a_{ij})$.

Given a scalar $\lambda \in \mathbb{R}$, we define the matrix λA as the matrix $C = (c_{ij})$ such that $c_{ij} = \lambda a_{ij}$; that is,

$$\lambda \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} = \begin{bmatrix} \lambda a_{11} & \lambda a_{12} & \dots & \lambda a_{1n} \\ \lambda a_{21} & \lambda a_{22} & \dots & \lambda a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda a_{m1} & \lambda a_{m2} & \dots & \lambda a_{mn} \end{bmatrix}.$$

For example

$$\begin{bmatrix} 1 & 2 & -1 & 3 \\ 0 & -1 & 4 & 2 \end{bmatrix} + \begin{bmatrix} 1 & -3 & 1 & 4 \\ -1 & 1 & -2 & 1 \end{bmatrix} = \begin{bmatrix} 2 & -1 & 0 & 7 \\ -1 & 0 & 2 & 3 \end{bmatrix}$$

and

$$3 \begin{bmatrix} 1 & 2 \\ 2 & -1 \\ 0 & -1 \\ 4 & 2 \end{bmatrix} = \begin{bmatrix} 3 & 6 \\ 6 & -3 \\ 0 & -3 \\ 12 & 6 \end{bmatrix}.$$

However,

$$\begin{bmatrix} 1 & 2 & -1 & 3 \\ 0 & -1 & 4 & 2 \end{bmatrix} + \begin{bmatrix} 1 & 2 \\ 2 & -1 \\ 0 & -1 \\ 4 & 2 \end{bmatrix}$$

does not make sense.

The analogy between the operations defined in Definition 1.1 on vectors in $\mathbb{R}^n$ and the operations in Definition 1.5 on matrices should not have escaped the reader. In some sense, these matrix operations are a two-dimensional generalization of the corresponding vector operations. In fact, they satisfy the same properties.

Proposition 1.1. *Matrices satisfy the following properties:*

$$\begin{aligned} A + (B + C) &= (A + B) + C \\ A + 0 &= 0 + A = A \\ A + -A &= -A + A = 0 \\ A + B &= B + A \\ \alpha(A + B) &= \alpha A + \alpha B \\ (\alpha + \beta)A &= \alpha A + \beta A \\ (\alpha\beta)A &= \alpha(\beta A) \\ 1A &= A, \end{aligned}$$

where A, B, C are $m \times n$ matrices (members of $\mathrm{M}_{m,n}$) and $\alpha, \beta \in \mathbb{R}$.

The first four properties are properties of addition.

1. The first property says that + is *associative*.
2. The second property says that 0 is an *identity element* (with respect to +).
3. The third property says that each matrix A has an *inverse* $-A$ (with respect to +).
4. The fourth property says that + is *commutative*.

These properties together make $\mathrm{M}_{m,n}$ into a *commutative group* under addition.

The last four properties are *distributivity properties* of addition and scalar multiplication; on the right, and on the left.

One can check that vectors in $\mathbb{R}^n$ and matrices in $\mathrm{M}_{m,n}$ satisfy all these properties. These properties characterize the structure known as *vector space*, the fundamental structures of linear algebra!

This suggests that each set of matrices $\mathrm{M}_{m,n}$ is a vector space. We will see later that this is indeed the case.

But first, we will discuss multiplication of matrices. For this, it is convenient to introduce the notion of *dot product*, another fundamental concept of linear algebra.

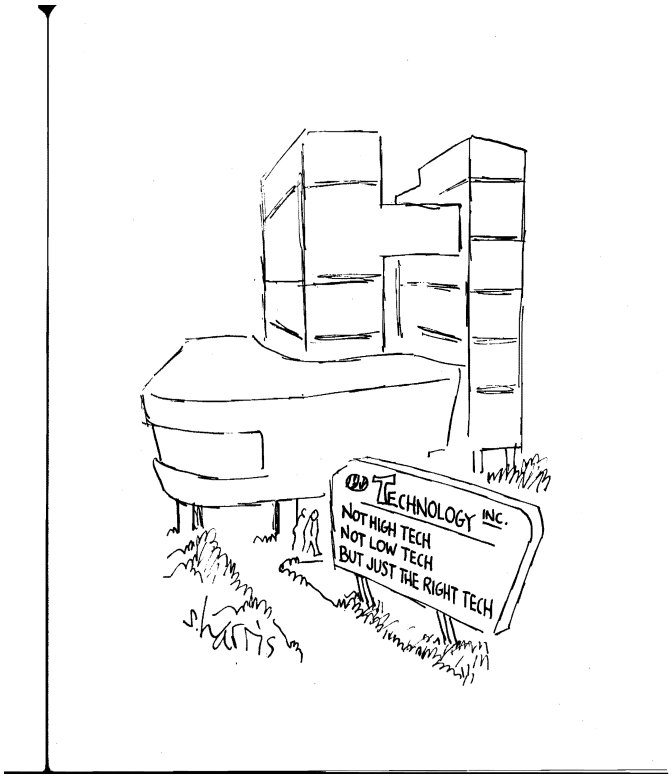

Figure 1.5: The right Tech

1.3 The Dot Product (also called Inner Product)

Recall that in Section 1.1 we defined the multiplication of a 3×3 matrix A by a vector x as the linear combination of the columns of A using the entries in the vector x as coefficients:

If

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \qquad A^1 = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \end{bmatrix} \qquad A^2 = \begin{bmatrix} a_{12} \\ a_{22} \\ a_{32} \end{bmatrix} \qquad A^3 = \begin{bmatrix} a_{13} \\ a_{23} \\ a_{33} \end{bmatrix},$$

and

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

then by definition

$$Ax = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = x_1 A^1 + x_2 A^2 + x_3 A^3.$$

A computation yields

$$x_1 A^1 + x_2 A^2 + x_3 A^3 = x_1 \begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \end{bmatrix} + x_2 \begin{bmatrix} a_{12} \\ a_{22} \\ a_{32} \end{bmatrix} + x_3 \begin{bmatrix} a_{13} \\ a_{23} \\ a_{33} \end{bmatrix}$$

$$= \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 \end{bmatrix}.$$

Therefore, we have

$$Ax = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 \end{bmatrix}.$$

Each entry in the rightmost matrix has the same structure involving the *rows* of A (rather than the columns of A). Indeed, if we let

$$\begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \end{bmatrix}$$

denote the first row of A, then we have

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = y_1x_1 + y_2x_2 + y_3x_3.$$

Similarly, if we let

$$\begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} = \begin{bmatrix} a_{21} & a_{22} & a_{23} \end{bmatrix}$$

denote the second row of A, then we have

$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = y_1x_1 + y_2x_2 + y_3x_3,$$

and if we let

$$\begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} = \begin{bmatrix} a_{31} & a_{32} & a_{33} \end{bmatrix}$$

denote the third row of A, then we have

$$a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = y_1x_1 + y_2x_2 + y_3x_3.$$

In all cases, the same expression

$$y_1x_1 + y_2x_2 + y_3x_3 = x_1y_1 + x_2y_2 + x_3y_3$$

shows up.

This expression is what is called the *dot product* or *inner product* of the vectors $x = (x_1, x_2, x_3)$ and $y = (y_1, y_2, y_3)$, viewed as column vectors.

We will denote the dot product of x and y by

$$x \cdot y, \quad \textit{or} \quad \langle x, y \rangle.$$

The notation $\langle x, y \rangle$ is preferable when the dot is already used to denote another operation. We can also define the inner product of two vectors $x = (x_1, y_1)$ and $y = (y_1, y_2)$ in $\mathbb{R}^2$ as

$$x \cdot y = x_1y_1 + x_2y_2.$$

The dot product of two one-dimensional vectors $x = (x_1)$ and $y = (y_1)$ is defined as

$$x \cdot y = x_1y_1.$$

In this very special case, it is just multiplication in $\mathbb{R}$.

Observe that the inner product of two vectors x and y is always a *number*, not a vector. Here is the general definition:

Definition 1.6. Given any two vectors $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ in $\mathbb{R}^n$, their *dot product* or *inner product* $x \cdot y$ (or $\langle x, y \rangle$) is the scalar (in $\mathbb{R}$) given by

$$x \cdot y = x_1y_1 + x_2y_2 + \cdots + x_ny_n.$$

Observe that

$$x \cdot y = y \cdot x.$$

For example,

$$\begin{bmatrix}4\\2\end{bmatrix} \cdot \begin{bmatrix}-1\\2\end{bmatrix} = 4 \times (-1) + 2 \times 2 = -4 + 4 = 0,$$

and

$$\begin{bmatrix}3\\2\\-1\end{bmatrix} \cdot \begin{bmatrix}-1\\2\\-1\end{bmatrix} = 3 \times (-1) + 2 \times 2 + (-1) \times (-1) = 2.$$

When the inner product of two vectors is zero, as in our first example, we say that the vectors are *orthogonal* (or *perpendicular*). This has to do with a geometric interpretation of the inner product in terms of angles that we will explore shortly.

The inner product can be interpreted as a *cost function*.

Say we have n products, where the price of product i is x_i, and say we want to buy (or sell) y_i units of product i. Then the total cost of the transaction is

$$x_1 y_1 + \cdots + x_n y_n.$$

We have defined the inner product on vectors, but it is the special case of the matrix multiplication of a row vector (a $1 \times n$ matrix) by a column vector (a $n \times 1$ matrix). Recall that

$$x \cdot y = \begin{bmatrix}x_1\\\vdots\\x_n\end{bmatrix} \cdot \begin{bmatrix}y_1\\\vdots\\y_n\end{bmatrix} = x_1 y_1 + \cdots + x_n y_n,$$

which suggests defining the product of the $1 \times n$ matrix (a row vector)

$$\begin{bmatrix}x_1 & \dots & x_n\end{bmatrix}$$

by the $n \times 1$ matrix (a column vector)

$$\begin{bmatrix}y_1\\\vdots\\y_n\end{bmatrix}$$

as

$$\begin{bmatrix}x_1 & \dots & x_n\end{bmatrix} \begin{bmatrix}y_1\\\vdots\\y_n\end{bmatrix} = x_1 y_1 + \cdots + x_n y_n.$$

This also suggests defining *transposition*, which converts a column vector to a row vector (and conversely). Given a column vector

$$x = \begin{bmatrix}x_1\\\vdots\\x_n\end{bmatrix},$$

the *transpose* $x^\top$ of x is the row vector

$$x^\top = \begin{bmatrix} x_1 & \dots & x_n \end{bmatrix},$$

and given a row vector

$$y = \begin{bmatrix} y_1 & \dots & y_n \end{bmatrix},$$

the *transpose* $y^\top$ of y is the column vector

$$y^\top = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

Then, the inner product of two vectors $x, y \in \mathbb{R}^n$ is also defined in terms of matrix multiplication and transposition by

$$x \cdot y = x^\top y.$$

The transposition operation actually applies to arbitrary $m \times n$ matrices.

Given an $m \times n$ matrix $A = (a_{ij})$, transposition forms a new $n \times m$ matrix $A^\top$ *whose columns are the rows of* A (and whose rows are the columns of A). Formally, the $n \times m$ matrix $A^\top$, the *transpose of* A is the matrix $(a_{ij}^\top)$, with

$$a_{ij}^\top = a_{ji},$$

for $i = 1, \dots, m$ and $j = 1, \dots, n$.

Observe that the original matrix A has m row and n columns, but the transpose matrix $A^\top$ has n rows and m columns. Also, `Matlab` uses the prime notation for transposition: A'.

Going back to the multiplication of a matrix A by a vector x, we we now have two ways of expressing Ax:

1. As a linear combination of the columns of A using the components of x as coefficients.

2. As a vector consisting of the inner products of the rows of A with x.

In the first case

$$Ax = \begin{bmatrix} A^1 & \cdots & A^n \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = x_1 A^1 + \cdots + x_n A^n,$$

where $A^1, \dots, A^n$ are the columns of A, and in the second case

$$Ax = \begin{bmatrix} \quad A_1 \quad \\ \vdots \\ \quad A_m \quad \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} A_1 x \\ \vdots \\ A_m x \end{bmatrix},$$

where $A_1, \ldots, A_m$ are the rows of A (as row vectors) and where $A_1x, .., A_mx$ are the inner products of $A_1, \ldots, A_m$ with x (in matrix form).

The inner product $x \cdot x$ of a vector x with itself has an important geometric interpretation.

For example, if $x = (2, 1)$, we have

$$x \cdot x = 2^1 + 1^2 = 4 + 1 = 5,$$

and for $x = (1, 2, 3)$, we have

$$x \cdot x = 1^2 + 2^2 + 3^2 = 14.$$

It turns out that $\sqrt{x \cdot x}$ is the *length* of the vector x.

In general, if $x = (x_1, \ldots, x_n)$, we have

$$x \cdot x = x_1^2 + x_2^2 + \cdots + x_n^2.$$

The following two properties of the inner product are immediately verified, but they are crucial:

Proposition 1.2. *The inner product on $\mathbb{R}^n$ satisfies the following properties:*

(a) For all $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, we have

$$x \cdot x \geq 0.$$

We say that the inner product is positive (semidefinite).

(b) For all $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, we have

$$x \cdot x = 0 \quad \textit{iff} \quad x = 0.$$

We say that the inner product is definite.

Properties (a) and (b) together say that the inner product is *positive definite*.

Property (a) is clear since if $x_i \in \mathbb{R}$, then
$x_1^2 + x_2^2 + \cdots + x_n^2 \geq 0$.

For property (b), we only need to check that if

$$x_1^2 + x_2^2 + \cdots + x_n^2 = 0,$$

then $x_1 = x_2 = \cdots = x_n = 0$. But this follows because all the x_i are real numbers, so $x_i^2 \geq 0$, and if $x_j \neq 0$, then $x_j^2 > 0$, and so $x_1^2 + x_2^2 + \cdots + x_n^2 > 0$.

Remark: Property (b) fails for complex numbers! For example,

$$1^2 + i^2 = 0,$$

yet $(1, i) \neq (0, 0)$. Here, $i = \sqrt{-1}$.

Since $x \cdot x \geq 0$ (property (a)), the square root of $x \cdot x$ makes sense, and we define the *(Euclidean) length* or *(Euclidean) norm* of x as

$$\|x\| = \sqrt{x \cdot x} = (x_1^2 + x_2^2 + \cdots + x_n^2)^{1/2}.$$

Remark: The Euclidean norm is also denoted by $\|x\|_2$, and is sometimes called the *2-norm*.

In view of property (b),

$$\|x\| = 0 \quad \text{iff} \quad x = 0.$$

Another useful property of the Euclidean norm is this: For every scalar $\lambda \in \mathbb{R}$, for every vector $x \in \mathbb{R}^n$,

$$\|\lambda x\| = |\lambda| \, \|x\| .$$

The absolute value is needed because λ could be negative, but a norm is always nonnegative.

If a vector $x \in \mathbb{R}^n$ is nonzero, then we know that $\|x\| \neq 0$, so we can write

$$x = \|x\| \left(\frac{1}{\|x\|} \, x \right).$$

Then, using the property stated above, we see that

$$\|(1/\|x\|)x\| = 1.$$

We say that $(1/\|x\|)x$ is a *unit vector*. This vector is sometimes denoted by $\widehat{x}$, and we can write

$$x = \|x\| \, \widehat{x}.$$

A *unit vector* x is any vector such that $\|x\| = 1$ (equivalently, $x \cdot x = 1$).

We can now give a geometric interpretation of the inner product of two vectors in $\mathbb{R}^2$. First, observe that if $x = 0$ or $y = 0$, then $x \cdot y = 0$, so we may assume that $x \neq 0$ and $y \neq 0$. Also, we can check easily that

$$(\lambda x) \cdot y = \lambda (x \cdot y) = x \cdot (\lambda y).$$

It follows that if $x \neq 0$ and $y \neq 0$, then

$$x \cdot y = \|x\| \, \|y\| \, (\widehat{x} \cdot \widehat{y}).$$

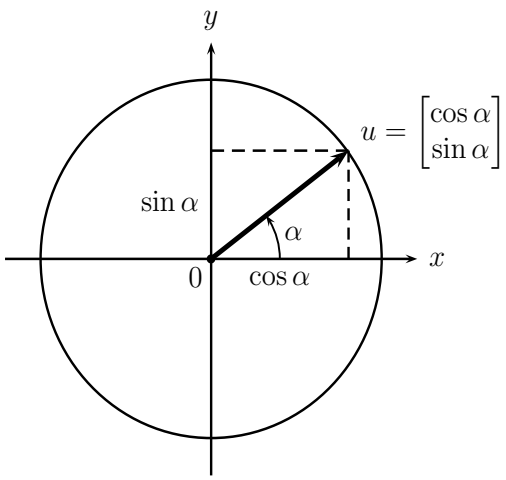

Figure 1.6: A unit vector in $\mathbb{R}^2$

Therefore, we just have to figure out what is the inner product of two unit vectors in the plane. However, a unit vector $\widehat{x}$ in the plane corresponds to a point on the unit circle.

Thus, its coordinates are of the form $(\cos\alpha, \sin\alpha)$, where α is the angle betwen the x-axis and the line supported by $\widehat{x}$. Similarly, the coordinates of the unit vector $\widehat{y}$ are of the form $(\cos\beta, \sin\beta)$; See Figures 1.6 and 1.7.

It follows that

$$\begin{aligned}\widehat{x}\cdot\widehat{y} &= (\cos\alpha, \sin\alpha)\cdot(\cos\beta, \sin\beta)\\ &= \cos\alpha\cos\beta + \sin\alpha\sin\beta\\ &= \cos(\beta-\alpha).\end{aligned}$$

Now, if we let $\theta = \beta - \alpha$, we see that θ is the angle between $\widehat{x}$ and $\widehat{y}$ (using a counterclockwise positive orientation).

Therefore, we proved that for two vectors x, y in the plane,

$$x\cdot y = \|x\|\,\|y\|\cos\theta,$$

where θ is the angle between x and y (we may assume $-\pi < \theta \leq \pi$).

Note that the formula applies even if $x = 0$ or $y = 0$, since in this case $\|x\|\,\|y\| = 0$.

A similar discussion applies to vectors in $\mathbb{R}^3$, except that this time, the sign of the angle θ depends on the orientation of the plane containing x and y.

In $\mathbb{R}^2$ or $\mathbb{R}^3$, the formula

$$x\cdot y = \|x\|\,\|y\|\cos\theta,$$

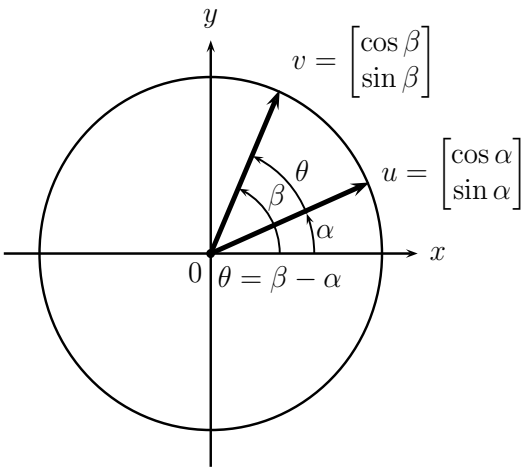

Figure 1.7: The inner product of two unit vectors in $\mathbb{R}^2$

explains why we say that x and y are orthogonal when $x \cdot y = 0$ (assuming $x \neq 0$ and $y \neq 0$). Indeed, we must have $\cos\theta = 0$, which implies that $\theta = \pm\pi/2$.

In the plane, the sign of $x \cdot y$ also tells us something about the magnitude of the angle θ: if $x \cdot y < 0$, then $|\theta| > \pi/2$.

What about vectors in $\mathbb{R}^n$ with $n \geq 4$?

Since $|\cos\theta| \leq 1$, we need to know that

$$|x \cdot y| \leq \|x\| \, \|y\|$$

always holds. We will prove this, but first we need to state a few properties of the inner product.

Proposition 1.3. *The inner product on $\mathbb{R}^n$ satisfies the following properties:*

$$\begin{aligned}
(x_1 + x_2) \cdot y &= (x_1 \cdot y) + (x_2 \cdot y)\\
(\lambda x) \cdot y &= \lambda(x \cdot y)\\
x \cdot (y_1 + y_2) &= (x \cdot y_1) + (x \cdot y_2)\\
x \cdot (\mu y) &= \mu(x \cdot y)\\
x \cdot y &= y \cdot x\\
x \cdot x &\geq 0\\
\textit{if} \quad x \neq 0, &\quad \textit{then} \quad x \cdot x > 0,
\end{aligned}$$

for all $x, x_1, x_2, y, y_1, y_2 \in \mathbb{R}^n$ and all $\lambda, \mu \in \mathbb{R}$.

The first four properties say that the inner product is linear in each argument; we say that it is *bilinear*.

The fifth property says that the inner product is *symmetric*.

The last two properties say that the inner product is *positive definite*.

Using the first five properties, we can show the following useful fact:

$$\|\lambda x + \mu y\|^2 = (\lambda x + \mu y) \cdot (\lambda x + \mu y) = \lambda^2 \|x\|^2 + 2\lambda\mu\, x \cdot y + \mu^2 \|y\|^2 .$$

In particular, for $\lambda = \mu = 1$ we get

$$\|x + y\|^2 = \|x\|^2 + 2\, x \cdot y + \|y\|^2 .$$

This shows that *Pythagoras Law* holds, namely

$$\|x + y\|^2 = \|x\|^2 + \|y\|^2 \quad \text{iff} \quad x \cdot y = 0,$$

that is iff x and y are orthogonal.

We are now ready to prove an important property of inner products, the Cauchy-Schwarz inequality. This property implies that the Euclidean norm satisfies what is known as the triangle inequality, another crucial property.

First, recall a property of quadratic equations. Consider the quadratic equation

$$ax^2 + bx + c = 0$$

and assume that $a > 0$. We know that this equation has two distinct real roots iff

$$b^2 - 4ac > 0.$$

Now, the curve of equation

$$y = ax^2 + bx + c$$

is a parabola, and because $a > 0$, this parabola does not go below the x-axis iff the equation $ax^2 + bx + c = 0$ does not have distinct real roots, which happens iff

$$b^2 - 4ac \leq 0.$$

Theorem 1.4. *For all $x, y \in \mathbb{R}^n$, we have the Cauchy-Schwarz inequality:*

$$|x \cdot y| \leq \|x\| \, \|y\| .$$

Furthermore, we also have the triangle inequality (also known as the Minkowski inequality):

$$\|x + y\| \leq \|x\| + \|y\| .$$

Proof. Let $t \in \mathbb{R}$ be any real number and consider the function F given by

$$F(t) = \|x + ty\|^2 .$$

From a previous calculation, we have

$$F(t) = t^2 \|y\|^2 + 2t(x \cdot y) + \|x\|^2 .$$

We also know that

$$F(t) = \|x + ty\|^2 \geq 0$$

for all t.

If $y = 0$, then $x \cdot y = 0$ and $\|y\| = 0$, in which case the inequality $|x \cdot y| \leq \|x\| \, \|y\|$ is trivial.

If $y \neq 0$, then $\|y\|^2 > 0$, and by our previous discussion, since $F(t)$ is nonnegative, the equation

$$t^2 \|y\|^2 + 2t(x \cdot y) + \|x\|^2 = 0$$

does not have distinct real roots, which implies that

$$4(x \cdot y)^2 - 4 \|x\|^2 \|y\|^2 \leq 0;$$

that is,

$$(x \cdot y)^2 \leq \|x\|^2 \|y\|^2 .$$

However, the above is equivalent to

$$|x \cdot y| \leq \|x\| \, \|y\| .$$

For the second inequality, recall that

$$\|x + y\|^2 = \|x\|^2 + 2\, x \cdot y + \|y\|^2 .$$

Also, since norms are positive, by squaring, we have

$$\|x + y\| \leq \|x\| + \|y\|$$

iff

$$\|x + y\|^2 \leq \|x\|^2 + \|y\|^2 + 2 \|x\| \, \|y\|$$

iff

$$\|x + y\|^2 - \|x\|^2 - \|y\|^2 \leq 2 \|x\| \, \|y\| . \tag{$*$}$$

Using the fact that

$$\|x + y\|^2 - \|x\|^2 - \|y\|^2 = 2\, x \cdot y,$$

the inequality $(*)$ is equivalent to

$$2x \cdot y \leq 2 \|x\| \, \|y\| .$$

The above is trivial if $x \cdot y < 0$, and otherwise follows from the Cauchy-Schwarz inequality. □

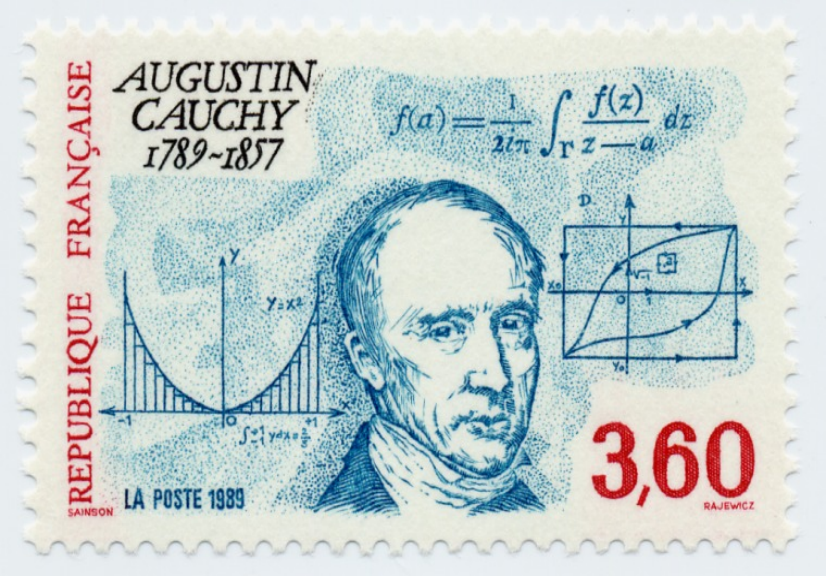

Figure 1.8: Augustin-Louis Cauchy, 1789–1857

Figure 1.9: Hermann Schwarz, 1843–1921

Figure 1.10: Hermann Minkowski, 1864–1909

It worth stating that the Euclidean norm satisfies the following three properties:

(1) $\|x\| \geq 0$ and $\|x\| = 0$ iff $x = 0$.

(2) $\|\lambda x\| = |\lambda| \, \|x\|$.

(3) $\|x + y\| \leq \|x\| + \|y\|$ (*triangle inequality*).

The Cauchy-Schwarz inequality

$$|x \cdot y| \leq \|x\| \, \|y\|$$

shows that when x and y are nonzero, we have

$$\left| \frac{x \cdot y}{\|x\| \, \|y\|} \right| \leq 1,$$

and so we can interpret this ratio as the cosine of the angle θ between the vectors x and y:

$$\cos \theta = \frac{x \cdot y}{\|x\| \, \|y\|}.$$

Remark: Unless we are in $\mathbb{R}^2$, the sign of this angle is not determined.

1.4 Matrix Multiplication

In Section 1.3, we defined the product of an $m \times n$ matrix A by a column vector $x \in \mathbb{R}^n$. The result is a vector $y \in \mathbb{R}^m$ such that

$$y = Ax.$$

It follows that the $m \times n$ matrix A defines a function from $\mathbb{R}^n$ to $\mathbb{R}^m$. On input $x \in \mathbb{R}^n$, this function produces the output $y = Ax \in \mathbb{R}^m$.

It turns out that such functions are very special: they are *linear*, but we will not discuss this right now.

Suppose now that we have two matrices A and B, where A is a $m \times n$ matrix and B is a $n \times p$ matrix. For every $x \in \mathbb{R}^p$, we get an output $Bx \in \mathbb{R}^n$, and for every $y \in \mathbb{R}^n$, we get an output $Ay \in \mathbb{R}^m$.

If we write $y = Bx$ and $z = Ay$, we should have

$$z = Ay = A(Bx).$$

The function that maps $x \in \mathbb{R}^p$ directly to $z \in \mathbb{R}^m$ should also be linear, and indeed it is given by a matrix AB, the *product* of A and B.

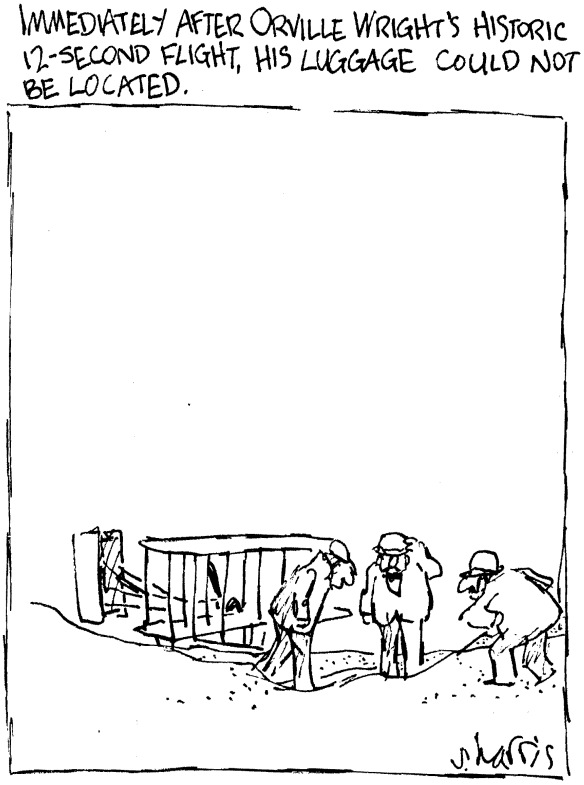

Figure 1.11: Early Traveling

The matrix AB is the $m \times p$ matrix whose jth column is the product AB^j of the matrix A by the jth column B^j of B; that is

$$AB = A \begin{bmatrix} B^1 & \cdots & B^p \end{bmatrix} = \begin{bmatrix} AB^1 & \cdots & AB^p \end{bmatrix}.$$

Observe that each AB^j is indeed a column vector in $\mathbb{R}^m$, since A is an $m \times n$ matrix and B^j is a vector in $\mathbb{R}^n$ (and B has p such columns, since it is an $n \times p$ matrix).

Going back to the definition of the product of a matrix A times a vector x in terms of the inner product of the rows of A with x, we see that if $C = AB$ and $C = (c_{ij})$, then

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} = (i\text{th row of } A) \cdot (j\text{th column of } B);$$

that is

$$c_{ij} = [a_{i\,1} \cdots a_{i\,n}] \begin{bmatrix} b_{1\,j} \\ \vdots \\ b_{n\,j} \end{bmatrix} = \sum_{k=1}^{n} a_{i\,k} b_{k\,j}.$$

Matrix multiplication has many of the properties similar to the multiplication of real numbers, but in general

$$AB \neq BA,$$

even for square matrices. For example

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

but

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

Here are some useful properties of matrix multiplication:

Proposition 1.5. *(1) Given any matrices $A \in \mathrm{M}_{m,n}$, $B \in \mathrm{M}_{n,p}$, and $C \in \mathrm{M}_{p,q}$, we have*

$$\begin{aligned} (AB)C &= A(BC) \\ I_m A &= A \\ A I_n &= A, \end{aligned}$$

that is, matrix multiplication is associative and has left and right identities.

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

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

We say that matrix multiplication $\cdot : \mathrm{M}_{m,n} \times \mathrm{M}_{n,p} \to \mathrm{M}_{m,p}$ is bilinear.

1.5 Inverse of a Matrix; Solving Linear Systems

Recall that a simple equation of the form

$$ax = b$$

(where a and b are real numbers) has a solution iff $a \neq 0$, in which case

$$x = \frac{b}{a} = a^{-1}b,$$

where $a^{-1} = 1/a$ is the *inverse* of a.

Since matrices correspond to linear maps, we should expect that only $n \times n$ matrices may have an inverse, since the corresponding functions should be bijective.

Given any $n \times n$ matrix A, we say that A *is invertible* iff there is some $n \times n$ matrix A^{-1} such that

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

A (square) matrix which is not invertible is said to be *singular*. It is important to observe that *not every matrix has an inverse*. For example, the matrix

$$\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$$

does not have any inverse. Why?

However, if a matrix A has an inverse, then it is *unique*. This is because if A has two inverses A' and A'', then

$$A'A = AA' = I \quad \text{and} \quad A''A = AA'' = I,$$

so we have

$$A'' = IA'' = (A'A)A'' = A'(AA'') = A'I = A'.$$

If a matrix A has an inverse, then the equation

$$Ax = b$$

has the *unique solution* $x = A^{-1}b$. Simply multiply both sides of $Ax = b$ by A^{-1} on the left.

A (square) matrix is invertible iff its determinant is nonzero (we will study determinants later). However, this is rarely a practical criterion.

A lot of efforts has been devoted to finding conditions that guarantee the invertibility of a matrix, and methods to compute the inverse of a matrix. This is generally a labor intensive process and it is rarely the best way to solve a linear system. Most methods solve a linear system without ever computing an inverse matrix.

The crucial point is that for any square matrix A, *the system $Ax = b$ has a solution for every $b \in \mathbb{R}^n$ iff the columns of A are linearly independent* (as vectors in $\mathbb{R}^n$).

There are various ways to prove this. One way to proceed it to first prove the following proposition which turns out to be a key ingredient in proving several fundamental results of linear algebra.

Proposition 1.6. *If A is any $n \times p$ matrix and if $p > n$ (there are more variables than equations), then the system $Ax = 0$ always has a nonzero solution; that is, there is some $x \in \mathbb{R}^p$, with $x \neq 0$, so that $Ax = 0$. Equivalently, any $p > n$ vectors in $\mathbb{R}^n$ must be linearly dependent.*

Proof. Let $u_1, \ldots, u_p$ be p vectors in $\mathbb{R}^n$. By definition, these vectors are linearly dependent iff there exist some scalars $x_1, \ldots, x_p$, not all zero, such that

$$x_1 u_1 + \cdots + x_p u_p = 0.$$

If we form the $n \times p$ matrix A having $u_1, \ldots, u_p$ as its columns, since

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

the system $Ax = 0$ has a nontrivial solution $x \neq 0$ iff $u_1, \ldots, u_p$ are linearly dependent.

We now prove by induction on $n \geq 1$ that the system $Ax = 0$ always has a nontrivial solution (if $p > n$).

If $n = 1$, there is a single equation of the form

$$a_1 x_1 + \cdots + a_p x_p = 0, \tag{$*$}$$

with $p \geq 2$.

If $a_j = 0$ for $j = 1, \ldots, p$, the equation is $0 = 0$, which is solved by any $x \in \mathbb{R}^p$.

Otherwise, there is some i such that $a_i \neq 0$, and pick the leftmost such i.

If $i > 1$, then x_1 does not appear in equation $(*)$, we have

$$a_i x_i + \cdots + a_p x_p = 0,$$

and $(x_1, 0, \ldots, 0)$ is a nontrivial solution for all $x_1 \neq 0$.

If $i = 1$, since $p \geq 2$, we can write

$$x_1 = (-a_2/a_1)x_2 + \cdots + (-a_p/a_1)x_p,$$

so we can assign arbitrary values to $x_2, \ldots, x_p$ and then solve for x_1. Thus, in all cases, equation $(*)$ has a nontrivial solution.

Let us now consider the induction step $n \geq 2$. The idea is to convert the system $Ax = 0$ to an equivalent system $A_2 x = 0$ with $n - 1$ equations, by *eliminating* x_1.

If the first column of A is zero, then $x = (x_1, 0, \ldots, 0)$ is a nontrivial solution of $Ax = 0$ for all $x_1 \neq 0$.

Otherwise, some entry in the first column of A is nonzero, pick the top one, say a_{i1}.

If $i \neq 1$, then swap the first and the ith row of A. Clearly, this preserves the set of solutions. Let A_1 be this new matrix.

Now, the coefficient π_1 of x_1 in the first equation is nonzero. We call π_1 a *pivot*.

Subtract $(a_{k1}/\pi_1) \times (row\ 1)$ from row k, for $k = 2, \ldots, n$.

The result is that *x_1 does not appear* in these $n-1$ new equations. Thus, the matrix of this new system is of the form

$$A_2 = \begin{pmatrix} \pi_1 & u_2 & \cdots & u_p \\ 0 & * & \cdots & * \\ \vdots & * & \cdots & * \\ 0 & * & \cdots & * \end{pmatrix} = \begin{pmatrix} \pi_1 & u \\ 0 & B \end{pmatrix}$$

where B is a $(n-1) \times (p-1)$ matrix.

Observe that we get back to the original matrix A_1 by *adding $(a_{k1}/\pi_1) \times (row\ 1)$ to row k*, for $k = 2, \ldots, n$.

It follows that the systems $A_1 x = 0$ and $A_2 x = 0$ have *the same set of solutions*. Now, since $p > n$ and $n \geq 2$, we have $p - 1 > n - 1 \geq 1$. By the induction hypothesis, the system

$$B \begin{pmatrix} x_2 \\ \vdots \\ x_p \end{pmatrix} = 0$$

which has $n-1$ equations and $p-1$ variables with $n-1 < p-1$ has a nonzero solution. Then, we can solve for x_1 using the first equation

$$\pi_1 x_1 + u_2 x_2 + \cdots + u_p x_p = 0,$$

and we obtain a nonzero solution $x = (x_1, x_2, \ldots, x_p)$ of the system $A_2 x = 0$.

Since the systems $A_1 x = 0$ and $A_2 x = 0$ have the same solutions, x is also a nontrivial solution of $A_1 x = 0$. Since the systems $Ax = 0$ and $A_1 x = 0$ also have the same solutions, this concludes the induction hypothesis. □

Proposition 1.6 has several important corollaries. Here is the first one.

Proposition 1.7. *If n vectors $u_1, \ldots, u_n$ with $u_i \in \mathbb{R}^n$ are linearly independent, then they span $\mathbb{R}^n$; that is, every vector in $\mathbb{R}^n$ is a linear combination of $u_1, \ldots, u_n$.*

Proof. We proceed by contradiction. If $u_1, \ldots, u_n$ do not span $\mathbb{R}^n$, then there is some nonzero vector $v \in \mathbb{R}^n$ which is *not* a linear combination of $u_1, \ldots, u_n$.

I claim that $u_1, \ldots, u_n, v$ are linearly independent. Assume that

$$\lambda_1 u_1 + \cdots + \lambda_n u_n + \mu v = 0,$$

for some scalars $\lambda_1, \ldots, \lambda_n, \mu \in \mathbb{R}$. We must have $\mu = 0$, because otherwise

$$v = (-\lambda_1/\mu)u_1 + \cdots + (-\lambda_n/\mu)u_n$$

contradicting the fact that v is not a linear combination of the u_is. But then,

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

and since $u_1, \ldots, u_n$ are linearly independent, we must have $\lambda_1 = \cdots = \lambda_n$. Therefore $u_1, \ldots, u_n, v$ are linearly independent. Now, $u_1, \ldots, u_n, v$ are $n+1$ linearly independent vectors in $\mathbb{R}^n$, contradicting Proposition 1.6. Therefore, $u_1, \ldots, u_n$ span $\mathbb{R}^n$. □

Here is a second important corollary of Proposition 1.6.

Proposition 1.8. *Let $u_1, \ldots, u_p$ and $v_1, \ldots, v_q$ be any vectors in $\mathbb{R}^n$. If $u_1, \ldots, u_p$ are linearly independent and if each u_j is a linear combination of the v_k, then $p \leq q$.*

Proof. Since each u_i is a linear combination of the v_j, we can write

$$u_j = \begin{bmatrix} v_1 & \cdots & v_q \end{bmatrix} a^j,$$

for some vector $a^j \in \mathbb{R}^q$, so if we form the $q \times p$ matrix $A = [a^1 \cdots a^p]$, we have

$$\begin{bmatrix} u_1 & \cdots & u_p \end{bmatrix} = \begin{bmatrix} v_1 & \cdots & v_q \end{bmatrix} A.$$

If $p > q$, then the matrix A has more columns than rows, so Proposition 1.6 implies that the system $Ax = 0$ has a nontrival solution $x \neq 0$. But then,

$$\begin{bmatrix} u_1 & \cdots & u_p \end{bmatrix} x = \begin{bmatrix} v_1 & \cdots & v_q \end{bmatrix} Ax = 0,$$

and since $x \neq 0$, we get a nontrivial linear dependence among the u_i's, a contradiction. Therefore, we must have $p \leq q$. □

Proposition 1.8 implies the following fact:

Proposition 1.9. *If $w_1, \ldots, w_n$ span $\mathbb{R}^n$, then they are linearly independent.*

Proof. If $w_1, \ldots, w_n$ are not linearly independent, since they span $\mathbb{R}^n$, there is a proper subset $\{v_1, \ldots, v_q\}$ of $\{w_1, \ldots, w_n\}$ that spans $\mathbb{R}^n$. If we pick $u_1, \ldots, u_n$, to be the canonical basis vectors with $u_i = (0, \ldots, 0, 1, 0, \ldots, 0)$, then $u_1, \ldots, u_n$ are linearly independent and they are linear combinations of $v_1, \ldots, v_q$ with $q < n$, contradicting Proposition 1.8. □

Proposition 1.8 has other important corollaries.

(1) If $(u_1, \ldots, u_p)$ and $(v_1, \ldots, v_q)$ are both linearly independent and span the same space, then $p = q$. This leads to the notions of *basis* and *dimension*, but we will postpone this topic to a later chapter.

(2) If p vectors in $\mathbb{R}^n$ are linearly independent and $p < n$, then they don't span $\mathbb{R}^n$.

Finally, we obtain the result we were seeking:

Theorem 1.10. *Let A be any square $n \times n$ matrix. The system $Ax = b$ has a solution for every $b \in \mathbb{R}^n$ iff the columns of A are linearly independent. In this case, for every b, the system $Ax = b$ has a unique solution.*

Proof. Assume that the n columns $A^1, \ldots, A^n$ of A are linearly independent. Then, by Proposition 1.7, these columns span $\mathbb{R}^n$, and consequently the system $Ax = b$ has a solution for every b (since Ax is a linear combination of the A^js).

Conversely, assume that the system $Ax = b$ has a solution for every $b \in \mathbb{R}^n$. This implies that $A^1, \ldots, A^n$ span $\mathbb{R}^n$, any by Proposition 1.9, they are linearly independent.

The uniqueness of the solution is a consequence of the linear independence of the columns of A. □

We can also give a criterion to decide when a square matrix A is invertible.

Theorem 1.11. *Let A be any square $n \times n$ matrix. The matrix A is invertible iff its columns $A^1, \ldots, A^n$ are linearly independent.*

Proof. First, assume that A is invertible. If

$$x_1 A^1 + \cdots + x_n A^n = 0,$$

which is equivalent to $Ax = 0$, applying A^{-1} to both sides of the equation $Ax = 0$, we get $A^{-1}Ax = Ix = x = 0$, so $A^1, \ldots, A^n$ are linearly independent.

Conversely, assume that $A^1, \ldots, A^n$ are linearly independent. By Proposition 1.7, $A^1, \ldots$, A^n span $\mathbb{R}^n$. Thus, for every $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)$, there is some $b_i \in \mathbb{R}^n$ such that

$$Ab_i = e_i.$$

If $B = [b_1 \ \cdots \ b_n]$ is the matrix whose jth column if b_j, then we have

$$AB = A[b_1 \ \cdots \ b_n] = [Ab_1 \ \cdots \ Ab_n] = [e_1 \ \cdots \ e_n] = I,$$

which shows that B is a right inverse of A.

We still have to prove that B is also a left inverse of A. Since $AB = I$, the columns $B^1, \ldots, B^n$ are linearly independent, because if we have a linear dependency $Bx = 0$, then $ABx = 0$, that is, $Ix = x = 0$.

By applying the same reasoning as above, $B^1, \ldots, B^n$ span $\mathbb{R}^n$, and thus B has some right inverse C, so that

$$BC = I.$$

However, we have

$$A = AI = A(BC) = (AB)C = IC = C,$$

which shows that

$$BA = I,$$

and B is a two-sided inverse of A, which proves that A is invertible □

The techniques used in the proof of Proposition 1.11 can be used to prove the following facts:

(1) If a square matrix A has a *left inverse* B, that is, a matrix such that $BA = I$, then A is invertible and $A^{-1} = B$.

(2) If a square matrix A has a *right inverse* C, that is, a matrix such that $AC = I$, then A is invertible and $A^{-1} = C$.

Before moving on to methods for solving linear systems, observe that if two $n \times n$ matrices A and B are invertible, then

$$(AB)^{-1} = B^{-1}A^{-1}.$$

Also if A is invertible, then so is its transpose $A^\top$ and

$$(A^\top)^{-1} = (A^{-1})^\top.$$

To prove the above, you may want to prove first that

$$(AB)^\top = B^\top A^\top,$$

even for rectangular matrices.

A 2×2 matrix

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

is invertible iff $ad - bc \neq 0$, in which case

$$A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

An $n \times n$ matrix $A = (a_{ij})$ is an *upper triangular matrix* iff $a_{ij} = 0$ whenever $i > j$. This means that all the entries below the diagonal are zero.

Here is an example of a triangular matrix:

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

The above matrix shows up in dealing with *cubic Bézier curves* and the *Bernstein polynomials*.

We will show later that an upper triangular matrix is invertible iff its diagonal entries are all nonzero.

Upper triangular matrices are important because linear systems $Ux = b$ can be solved easily by *backward substitution* (when U is upper triangular). For example, the inverse of the matrix A shown above is

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

It can be shown that the inverse of an upper triangular matrix is also upper triangular. There are useful matrix factorizations methods involving upper triangular matrices: LU, QR.

Figure 1.12: Hitting Power

Chapter 2

Gaussian Elimination, LU-Factorization, Cholesky Factorization, Reduced Row Echelon Form

2.1 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 2.1, 2.2, 2.3 illustrating various configuations.

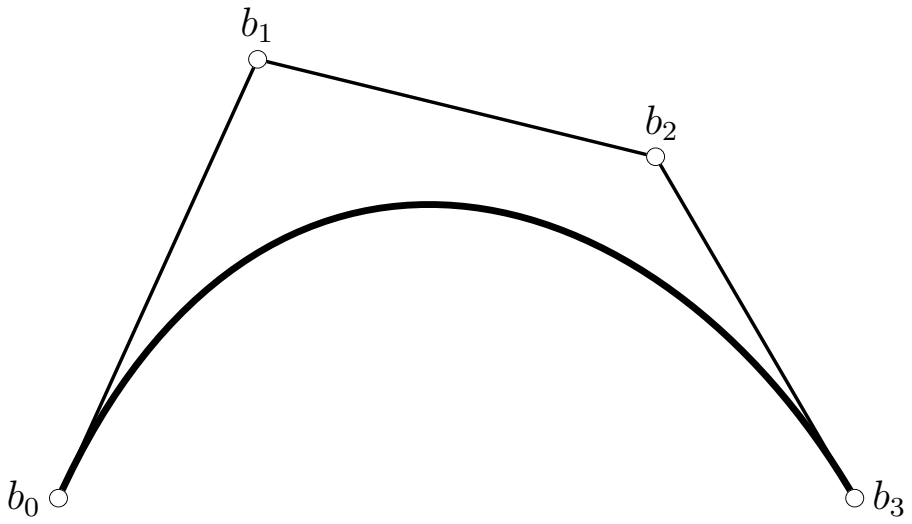

Figure 2.1: A "standard" Bézier curve

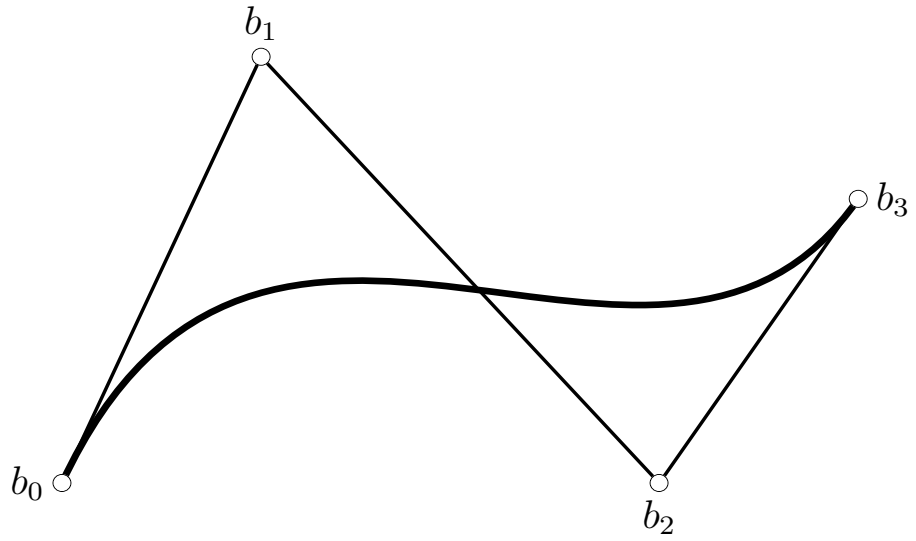

Figure 2.2: A Bézier curve with an inflexion point

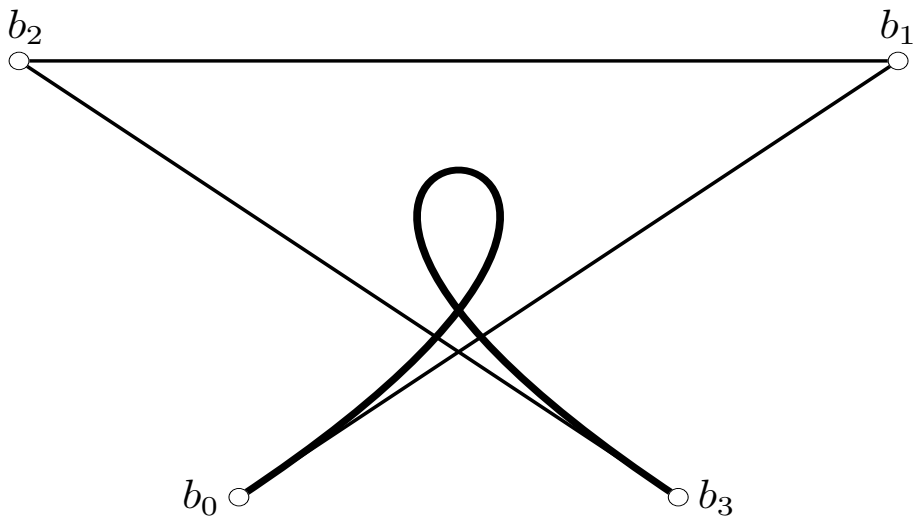

Figure 2.3: A self-intersecting Bézier curve

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 2.4 illustrates an interpolation problem involving $N+1 = 7+1 = 8$ data points. The control points d_0 and d_7 were chosen arbitrarily.

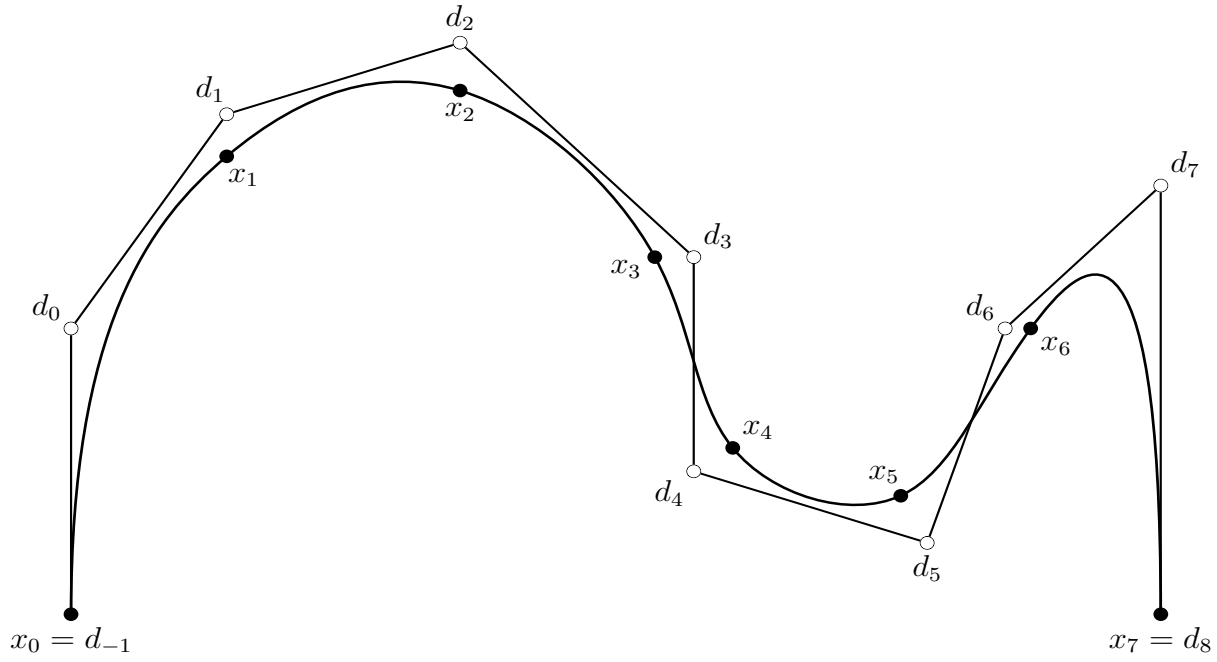

Figure 2.4: A C^2 cubic interpolation spline curve passing through the points $x_0, x_1, x_2, x_3, x_4, x_5, x_6, x_7$

It can be shown that $d_1, \ldots, d_{N-1}$ are given by the linear system

$$\begin{pmatrix} \frac{7}{2} & 1 & & & \\ 1 & 4 & 1 & & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & & 1 & 4 & 1 \\ & & & 1 & \frac{7}{2} \end{pmatrix} \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_{N-2} \\ d_{N-1} \end{pmatrix} = \begin{pmatrix} 6x_1 - \frac{3}{2}d_0 \\ 6x_2 \\ \vdots \\ 6x_{N-2} \\ 6x_{N-1} - \frac{3}{2}d_N \end{pmatrix}.$$

It can be shown 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.

2.2 Gaussian Elimination and LU-Factorization

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 $A_k = (a_{ij}^k)$ 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

$$a_{i\,j}^k = a_{i\,j}^i$$

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}.

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 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 2.1. *(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 $a_{ik}^k = 0$, 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

$$M = E_{n-1} P_{n-1} \cdots E_2 P_2 E_1 P_1,$$

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 expansive 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 $D_j x^{(j)} = M_j e_j$, 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).

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 $L_{ii} = 1$ 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 2.2. *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

$$A = \begin{pmatrix} A[1..k, 1..k] & A_2 \\ A_3 & A_4 \end{pmatrix} = \begin{pmatrix} L_1 & 0 \\ P & L_4 \end{pmatrix} \begin{pmatrix} U_1 & Q \\ 0 & U_4 \end{pmatrix} = \begin{pmatrix} L_1 U_1 & L_1 Q \\ P U_1 & PQ + L_4 U_4 \end{pmatrix},$$

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 $A[1..1, 1..1] = (a_{11})$, so, $a_{11} \neq 0$. 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

$$\begin{pmatrix} L_1 & 0 \\ P & L_4 \end{pmatrix} \begin{pmatrix} A[1..k, 1..k] & A_2 \\ A_3 & A_4 \end{pmatrix} = \begin{pmatrix} U_1 & B_2 \\ 0 & B_4 \end{pmatrix},$$

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 2.2 can be avoided if we use the fact that a triangular matrix is invertible iff all its diagonal entries are nonzero.

Corollary 2.3. *(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 2.2 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 2.2. 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.

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 5.2, 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 5.2 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 2.4. *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; so, 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_1A 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 $P_2P_1A[1..k, 1..k]$ 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_1P_2A is also invertible, and we are done. □

Remark: One can also prove Proposition 2.4 using a clever reordering of the Gaussian elimination steps suggested by Trefethen and Bau [56] (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

$$A_n = E_{n-1}P_{n-1} \cdots E_2P_2E_1P_1A,$$

where $U = A_n$ is upper-triangular. For example, when $n = 4$, we have $E_3P_3E_2P_2E_1P_1A = U$. We can define new matrices E_1', E_2', E_3' which are still products of elementary matrices so that we have

$$E_3'E_2'E_1'P_3P_2P_1A = U.$$

Indeed, if we let $E_3' = E_3$, $E_2' = P_3E_2P_3^{-1}$, and $E_1' = P_3P_2E_1P_2^{-1}P_3^{-1}$, we easily verify that each E_k' is a product of elementary matrices and that

$$E_3'E_2'E_1'P_3P_2P_1 = E_3(P_3E_2P_3^{-1})(P_3P_2E_1P_2^{-1}P_3^{-1})P_3P_2P_1 = E_3P_3E_2P_2E_1P_1.$$

It can also be proved that E_1', E_2', E_3' are lower triangular (see Theorem 2.5).

In general, we let

$$E_k' = P_{n-1} \cdots P_{k+1}E_kP_{k+1}^{-1} \cdots P_{n-1}^{-1},$$

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 2.5).

Using the above idea, we can prove the theorem below which also shows how to compute P, L and U using a simple adaptation of Gaussian elimination. We are not aware of a detailed proof of Theorem 2.5 in the standard texts. Although Golub and Van Loan [26] state a version of this theorem as their Theorem 3.1.4, they say that "The proof is a messy subscripting argument." Meyer [42] 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 2.5. *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 $E_{n-1}P_{n-1}\cdots E_1 P_1 A = U$ is the result of Gaussian elimination with some pivoting, write $A_k = E_{k-1}P_{k-1}\cdots E_1 P_1 A$, and define E_j^k, with $1 \leq j \leq n-1$ and $j \leq k \leq n-1$, such that, for $j = 1, \ldots, n-2$,

$$\begin{aligned} E_j^j &= E_j \\ E_j^k &= P_k E_j^{k-1} P_k, \quad \text{for } k = j+1, \ldots, n-1, \end{aligned}$$

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

$$P = P_{n-1} \cdots P_1$$
$$L = (E_1^{n-1})^{-1} \cdots (E_{n-1}^{n-1})^{-1},$$

then

$$PA = LU. \tag{$\dagger_1$}$$

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 rows 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 $n \times n$ 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

$$\Lambda'_k = P_k \Lambda_{k-1}, \tag{$\dagger_2$}$$

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 proven 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)

Step 1. Prove ($\dagger_1$).

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.

Step 2. Prove that the matrices $(E_j^k)^{-1}$ are lower-triangular. To achieve this, we prove that the matrices $\mathcal{E}_j^k$ are strictly lower triangular matrices of a very special form.

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 or $P_k = I$, so $P_k^2 = I$, and we get

$$(E_j^k)^{-1} = P_k (E_j^{k-1})^{-1} P_k = P_k (I + \mathcal{E}_j^{k-1}) P_k = P_k^2 + P_k \, \mathcal{E}_j^{k-1} \, P_k = I + P_k \, \mathcal{E}_j^{k-1} \, P_k.$$

Therefore, 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,

$$(E_j^k)^{-1} = I + P(k,i)\,\mathcal{E}_j^{k-1}.$$

But since

$$(E_j^k)^{-1} = I + \mathcal{E}_j^k,$$

we deduce that

$$\mathcal{E}_j^k = P(k,i)\,\mathcal{E}_j^{k-1}.$$

We also know that multiplying $\mathcal{E}_j^{k-1}$ 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.

Step 3. Express L as $L = I + \Lambda_{n-1}$, with $\Lambda_{n-1} = \mathcal{E}_1^1 + \cdots + \mathcal{E}_{n-1}^{n-1}$.

From Step 1 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 also 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 find that we get $\Lambda_1 = \mathcal{E}_1^1$, and the base step holds.

Since $(E_j^k)^{-1} = I + \mathcal{E}_j^k$ with

$$\mathcal{E}_j^k = \begin{pmatrix} 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & \ell_{j+1j}^{(k)} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \ell_{nj}^{(k)} & 0 & \cdots & 0 \end{pmatrix}$$

and $\mathcal{E}_i^k \mathcal{E}_j^k = 0$ if $i < j$, 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 (†), 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 (†), we obtain the desired equation: $L = I + \Lambda_{n-1}$. □

Part (3) of Theorem 2.5 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 L (really Λ_k, since the one's are not altered) 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 2.5(1), we can show the following result.

Proposition 2.6. *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

$$LDU = A = A^\top = U^\top DL^\top,$$

with $U^\top$ lower-triangular and $DL^\top$ upper-triangular. By the uniqueness of LU-factorization (part (1) of Theorem 2.5), 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.

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

$$\begin{array}{rcrcl} 10^{-4}x & + & y & = & 1 \\ x & + & y & = & 2. \end{array}$$

Since 10^{-4} is nonzero, it can be taken as pivot, and we get

$$\begin{array}{rcrcl} 10^{-4}x & + & y & = & 1 \\ & & (1-10^4)y & = & 2-10^4. \end{array}$$

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

$$\begin{array}{rcrcl} x & + & y & = & 2 \\ & & (1-10^{-4})y & = & 1-2\times 10^{-4}. \end{array}$$

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 $a^k_{i\,k}$ such that

$$|a^k_{i\,k}| = \max_{k \leq p \leq n} |a^k_{p\,k}|.$$

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 [56], or Question 2.19 in Demmel [14]).

Another strategy, called *complete pivoting*, consists in choosing some entry $a^k_{i\,j}$, where $k \leq i, j \leq n$, such that

$$|a^k_{i\,j}| = \max_{k \leq p,q \leq n} |a^k_{p\,q}|.$$

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.

2.3 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 2.7. *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 2.8. *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 2.5 (and Proposition 2.2), 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). 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.

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*.

2.4 SPD Matrices and the Cholesky Decomposition

First, we note that a symmetric positive definite matrix satisfies the condition of Proposition 2.2.

Proposition 2.9. *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 2.9 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 Cholseky factorization.

Let A be a symmetric positive definite matrix and write

$$A = \begin{pmatrix} a_{11} & W^\top \\ W & C \end{pmatrix}.$$

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.

Theorem 2.10. *(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 k. For $k = 1$, we must have $a_{11} > 0$, and if we let $\alpha = \sqrt{a_{11}}$ and $B = (\alpha)$, the theorem holds trivially. If $k \geq 2$, 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 $C - WW^\top/a_{11}$, 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 proof of Theorem 2.10 immediately yields an algorithm to compute B from A. 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 $i = j+1, \ldots, n$,

$$b_{i\,j} = \left(a_{i\,j} - \sum_{k=1}^{j-1} b_{i\,k} b_{j\,k} \right) / b_{j\,j}.$$

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.

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.

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

$$x^\top A x = x^\top B B^\top x = (B^\top x)^\top B^\top x,$$

it is clear that $x^\top A x > 0$ if $x \neq 0$. □

We now give three more criteria for a symmetric matrix to be positive definite.

Proposition 2.11. *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 2.9, 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 2.2, 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

$$Q = BB^\top,$$

with B lower-triangular and invertible. By the remark before Proposition 2.11, 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 [14], Trefethen and Bau [56], Ciarlet [11], Golub and Van Loan [26], Meyer [42], Strang [52, 53], and Kincaid and Cheney [34].

2.5 Reduced Row Echelon Form

Gaussian elimination described in Section 2.2 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 $k+1$.

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*. 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 2.12. *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 $Ax = b$. □

Before stating the next proposition, which states another important fact, we need the notion of rank.

Definition 2.1. Given any $m \times n$ matrix A, the *rank* (or *column rank*) of A is the maximum number of linearly independent columns of A. The *row rank* of A is the maximum number of linearly independent rows of A (equivalently, the maximum number of linearly independent columns of $A^\top$).

It turns out that the column rank and the row rank of a matrix are always equal. This is a fundamental result of linear algebra which is best proved using orthogonal complements. We postpone the proof until we discuss Euclidean inner products in more details.

Proposition 2.13. *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 2.12 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.

We will see that the reduction to row echelon form provides of proof of the fact that the row rank is equal to the column rank. Let us now define precisely what is a reduced row echelon matrix.

Definition 2.2. 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.

The following proposition shows that every matrix can be converted to a reduced row echelon form using row operations.

Proposition 2.14. *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 2.13 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 2.15. *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 2.14 and Theorem 2.15 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.

If we have a *homogeneous system* $Ax = 0$, which means that $b = 0$, of course $x = 0$ is always a solution, but Theorem 2.15 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 2.16. *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 2.15 we may assign arbitrary values to $n - r > 0$ nonpivot variables and we get nontrivial solutions. □

Theorem 2.15 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 2.17. *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 2.16, 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 2.17 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 2.17 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 2.18. *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 $Ax = b$. First, we have the following simple result.

Proposition 2.19. *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_n 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_n 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

$$\begin{pmatrix} I_r & F \\ 0_{m-r,r} & 0_{m-r,n-r} \end{pmatrix} \begin{pmatrix} d \\ 0_{n-r} \end{pmatrix} = \begin{pmatrix} d \\ 0_{m-r} \end{pmatrix},$$

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. 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

$$a_{22} + a_{33} + a_{44} - a_{12} - a_{13} - a_{14} = 0$$
$$a_{41} + a_{32} + a_{23} - a_{11} - a_{12} - a_{13} = 0,$$

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, and for $n = 4$, it is 34. 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!

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.

2.6 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 elmination step; *pivoting*.

- *Transposition matrix*; *elementary matrix*.
- The *Gaussian elimination theorem* (Theorem 2.1).
- *Gauss-Jordan factorization*.
- *LU-factorization*; Necessary and sufficient condition for the existence of an LU-factorization (Proposition 2.2).
- "$PA = LU$ theorem" (Theorem 2.5).
- 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 2.10).
- *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* $Ax = 0$.

Chapter 3

Vector Spaces, Bases, Linear Maps

3.1 Vector Spaces, Subspaces

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.

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 3.1. 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

$$a \cdot b = b \cdot a$$

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 3.1.

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 with real coefficients 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 $RR^\top = I_n$ 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 3.2. 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 3.2, 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

[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.

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 $u = 0$.

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 3.2.

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 $\mathbb{R}$, and $\mathbb{C}^n$ is a vector space over $\mathbb{C}$.

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 $\mathbb{C}$.

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 $\mathbb{C}$.

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 $f :]a, b[\to \mathbb{R}$ is a vector space over $\mathbb{R}$.

Let E be a vector space. We would like to define the important notions of linear combination and linear independence. These notions can be defined for sets of vectors in E, but it will turn out to be more convenient to define them for families $(v_i)_{i \in I}$, where I is any arbitrary index set.

3.2 Linear Independence, Subspaces

In this section, we revisit linear independence and introduce the notion of a *basis*. 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*). Let us review the notion of an indexed family.

Given a set A, a *family* $(a_i)_{i \in I}$ of elements of A is simply a function $a\colon I \to A$.

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.

We can deal with an arbitrary set X by viewing it as the family $(X_x)_{x \in X}$ corresponding to the identity function $\mathrm{id}\colon X \to X$. 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.

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 3.3. 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 $v = 0$. 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.

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

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

The above 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 $k \neq i, j$.

Example 3.3.

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 3.4. 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 3.4.

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 3.1. *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.

3.3 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. Subspaces having such a "generating family" play an important role, and motivate the following definition.

Definition 3.5. 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 3.5.

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* $\binom{n}{k}(1-X)^k X^{n-k}$ for $k = 0, \ldots, n$, also form a basis of that space. These polynomials play a major role in the theory of *spline curves*.

It is a standard result of linear algebra that every vector space E has a basis, and that for any two bases $(u_i)_{i\in I}$ and $(v_j)_{j\in J}$, 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. We begin with a crucial lemma.

Lemma 3.2. *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 3.3. *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, 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 3.2, 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 3.3 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.

The following proposition giving useful properties characterizing a basis is an immediate consequence of Theorem 3.3.

Proposition 3.4. *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.

Our next goal is to prove that any two bases in a finitely generated vector space have the same number of elements. We could use the *replacement lemma* due to Steinitz but this also follows from Proposition 1.8, which holds for any vector space (not just $\mathbb{R}^n$). Since this is an important fact, we repeat its statement and its proof.

Proposition 3.5. *For any vector space E, let $u_1, \ldots, u_p$ and $v_1, \ldots, v_q$ be any vectors in E. If $u_1, \ldots, u_p$ are linearly independent and if each u_j is a linear combination of the v_k, then $p \leq q$.*

Proof. Since each u_i is a linear combination of the v_j, we can write

$$u_j = \begin{bmatrix} v_1 & \cdots & v_q \end{bmatrix} a^j,$$

for some vector $a^j \in \mathbb{R}^q$, so if we form the $q \times p$ matrix $A = [a^1 \cdots a^p]$, we have

$$\begin{bmatrix} u_1 & \cdots & u_p \end{bmatrix} = \begin{bmatrix} v_1 & \cdots & v_q \end{bmatrix} A.$$

If $p > q$, then the matrix A has more columns than rows, so Proposition 1.6 implies that the system $Ax = 0$ has a nontrival solution $x \neq 0$. But then,

$$\begin{bmatrix} u_1 & \cdots & u_p \end{bmatrix} x = \begin{bmatrix} v_1 & \cdots & v_q \end{bmatrix} Ax = 0,$$

and since $x \neq 0$, we get a nontrivial linear dependence among the u_i's, a contradiction. Therefore, we must have $p \leq q$. □

Putting Theorem 3.3 and Proposition 3.5 together, we obtain the following fundamental theorem.

Theorem 3.6. *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. 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 3.3 with $L = \emptyset$ and $S = (u_i)_{i \in I}$. 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 3.5 implies that $|I| \leq |J|$. A symmetric argument yields $|J| \leq |I|$. □

Remark: Theorem 3.6 also holds for vector spaces that are not finitely generated.

When 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 $\dim(\{0\}) = 0$.

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

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 3.6. 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.

3.4 Linear Maps

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 3.6. 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 3.6.

Example 3.7.

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

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

   is a linear map.

2. For any vector space E, the *identity map* $\mathrm{id}: E \to E$ 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 $D: \mathbb{R}[X] \to \mathbb{R}[X]$ defined such that

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

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

Definition 3.7. Given a linear map $f\colon E \to F$, we define its *image (or range)* $\operatorname{Im} f = f(E)$, as the set

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

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

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

Proposition 3.8. *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 3.8, 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 $\operatorname{Im} f$.

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 3.9. *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 $v \in E$ is a linear combination $v = \sum_{i \in I} \lambda_i u_i$ of $(u_i)_{i \in I}$. If

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

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 $v = 0$. 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 3.9, an injective linear map $f\colon E \to F$ 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 $f\colon E \to F$ is injective, then $(f(u_i))_{i \in I}$ is a basis of F (by Proposition 3.4).

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

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

Such a map g 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 3.9 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 3.9 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 $f^{-1}\colon F \to E$ is also a linear map, and thus f is an isomorphism.

Another useful corollary of Proposition 3.9 is this:

Proposition 3.10. *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 3.9, 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 3.6). 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 $f(g(y)) = y$). Let $(u_1, \ldots, u_n)$ be any basis of E. By Proposition 3.9, 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 3.6). 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. □

We have the following important result relating the dimension of the kernel (nullspace) and the dimension of the image of a linear map.

Theorem 3.11. *Let $f\colon E \to F$ be a linear map. For any choice of a basis $(v_1, \ldots, v_r)$ of the image $\mathrm{Im}\, f$ of f, let $(u_1, \ldots, u_r)$ be any vectors in E such that $v_i = f(u_i)$, for $i = 1, \ldots, r$, and let $(h_1, \ldots, h_s)$ be a basis of the kernel (nullspace) $\mathrm{Ker}\, f$ of f. Then, $(u_1, \ldots, u_r, h_1, \ldots, h_s)$ is a basis of E, so that $n = \dim(E) = s + r$, and thus*

$$\dim(E) = \dim(\mathrm{Ker}\, f) + \dim(\mathrm{Im}\, f) = \dim(\mathrm{Ker}\, f) + \mathrm{rk}(f).$$

Proof. We claim that every vector $w \in E$ can be written as

$$w = u + h,$$

where u is a linear combination of $(u_1, \ldots, u_r)$ and h is a linear combination of $(h_1, \ldots, h_s)$. Consider $f(w) \in \mathrm{Im}\, f$. Since $(v_1, \ldots, v_r)$ is a basis of $\mathrm{Im}\, f$, we can write

$$f(w) = \lambda_1 v_1 + \cdots + \lambda_r v_r,$$

for some unique scalars $\lambda_i \in \mathbb{R}$. Let

$$u = \lambda_1 u_1 + \cdots + \lambda_r u_r,$$

then using linearity and the fact that $v_i = f(u_i)$, we have

$$\begin{aligned} f(w - u) &= f(w) - f(u) \\ &= \lambda_1 v_1 + \cdots + \lambda_r v_r - f(\lambda_1 u_1 + \cdots + \lambda_r u_r) \\ &= \lambda_1 v_1 + \cdots + \lambda_r v_r - (\lambda_1 f(u_1) + \cdots + \lambda_r f(u_r)) \\ &= \lambda_1 v_1 + \cdots + \lambda_r v_r - (\lambda_1 v_1 + \cdots + \lambda_r v_r) \\ &= 0. \end{aligned}$$

Consequently $w - u \in \mathrm{Ker}\, f$, which means that $w - u = h$ for some $h \in \mathrm{Ker}\, f$, that is

$$w = u + h,$$

with $u = \lambda_1 u_1 + \cdots + \lambda_r u_r$ and $h = \mu_1 h_1 + \cdots + \mu_s h_s$ for some $\mu_j \in \mathbb{R}$, since $(h_1, \ldots, h_s)$ is a basis of $\mathrm{Ker}\, f$. This proves that $(u_1, \ldots, u_r, h_1, \ldots, h_s)$ span E.

Let us now prove that $(u_1, \ldots, u_r, h_1, \ldots, h_s)$ are linearly independent.

Assume that

$$\lambda_1 u_1 + \cdots + \lambda_r u_r + \mu_1 h_1 + \cdots + \mu_s h_s = 0.$$

If we apply f, since $h_1, \ldots, h_s \in \operatorname{Ker} f$ we have $f(h_j) = 0$ for $j = 1, \ldots, s$, and since $v_i = f(u_i)$, we get

$$\lambda_1 v_1 + \cdots + \lambda_r v_r = 0.$$

However, $(v_1, \ldots, v_r)$ are linearly independent (a basis of $\operatorname{Im} f$) so $\lambda_1 = \cdots = \lambda_r = 0$, and we are left with

$$\mu_1 h_1 + \cdots + \mu_s h_s = 0.$$

But, $(h_1, \ldots, h_s)$ are also linearly independent (a basis of $\operatorname{Ker} f$), therefore, $\mu_1 = \cdots = \mu_s = 0$, which proves that $(u_1, \ldots, u_r, h_1, \ldots, h_s)$ are linearly independent.

In summary, we proved that $(u_1, \ldots, u_r, h_1, \ldots, h_s)$ is a basis of E, with $(h_1, \ldots, h_s)$ a basis of $\operatorname{Ker} f$, and $(u_1, \ldots, u_r)$ a basis of a subspace of E isomorphic to $\operatorname{Im} f$ (since $(v_1, \ldots, v_r)$ is a basis of $\operatorname{Im} f$). Then, it is immediate that

$$\dim(E) = \dim(\operatorname{Ker} f) + \dim(\operatorname{Im} f) = \dim(\operatorname{Ker} f) + \operatorname{rk}(f),$$

which completes the proof of our theorem. □

The set of all linear maps between two vector spaces E and F is denoted by $\operatorname{Hom}(E, F)$ 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 $\operatorname{Hom}_K(E, F)$.

The set $\operatorname{Hom}(E, F)$ is a vector space under the operations defined in Example 3.6, 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 $\operatorname{Hom}(E, F)$ also has finite dimension, as we shall see shortly. When $E = F$, a linear map $f \colon E \to E$ is also called an *endomorphism*. The space $\operatorname{Hom}(E, E)$ is also denoted by $\operatorname{End}(E)$.

It is also important to note that composition confers to $\operatorname{Hom}(E, E)$ a ring structure. Indeed, composition is an operation $\circ \colon \operatorname{Hom}(E, E) \times \operatorname{Hom}(E, E) \to \operatorname{Hom}(E, E)$, 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 $f: E \to E$ is a group under composition. Bijective linear maps 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 $\mathbf{GL}(n)$.

3.5 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 3.3).
- A family B of vectors is a basis iff it is a maximal linearly independent family iff it is a minimal generating family (Proposition 3.4).
- Proposition 3.5.
- Any two bases in a finitely generated vector space E have the *same number of elements*; this is the *dimension* of E (Theorem 3.6).
- *Hyperlanes*.
- Every vector has a *unique representation* over a basis (in terms of its coordinates).
- 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 3.8).

- The *unique homomorphic extension property* of linear maps with respect to bases (Proposition 3.9).
- The representation of linear maps by *matrices*.
- The vector space of linear maps $\mathrm{Hom}_K(E, F)$.

Chapter 4

Matrices, Linear Maps, and Affine Maps

4.1 Matrices

Proposition 3.9 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 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 $a_{i\,j}$ $(1 \leq i \leq m,\ 1 \leq j \leq n)$.

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, $1 \leq j \leq n$. 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,

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

and 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

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

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, $1 \leq j \leq p$. 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) suggests defining a multiplication operation on matrices, and we proceed to do so. We have the following definitions.

Definition 4.1. 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 $\{1, 2, \dots, m\} \times \{1, 2, \dots, n\}$ 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 $\{1, 2, \dots, m\}$ and $\{1, 2, \dots, n\}$, and to represent A as an array according to this ordering of the rows and columns.

We also define some operations on matrices as follows.

Definition 4.2. Given two $m \times n$ matrices $A = (a_{ij})$ and $B = (b_{ij})$, we define their *sum* $A + B$ as the matrix $C = (c_{ij})$ such that $c_{ij} = a_{ij} + b_{ij}$; that is,

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \dots & b_{mn} \end{pmatrix}$$

$$= \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \dots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \dots & a_{2n} + b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \dots & a_{mn} + b_{mn} \end{pmatrix}.$$

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

$$\lambda \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} = \begin{pmatrix} \lambda a_{11} & \lambda a_{12} & \dots & \lambda a_{1n} \\ \lambda a_{21} & \lambda a_{22} & \dots & \lambda a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda a_{m1} & \lambda a_{m2} & \dots & \lambda a_{mn} \end{pmatrix}.$$

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

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj},$$

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

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1p} \\ b_{21} & b_{22} & \dots & b_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \dots & b_{np} \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1p} \\ c_{21} & c_{22} & \dots & c_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \dots & c_{mp} \end{pmatrix}$$

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:

$$(a_{i1} \cdots a_{in}) \begin{pmatrix} b_{1j} \\ \vdots \\ b_{nj} \end{pmatrix} = \sum_{k=1}^{n} a_{ik} b_{kj}.$$

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

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

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 $AI_n = I_n A = A$. 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.

Proposition 3.10 shows 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 $C = A^{-1}$.

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 4.1 and Definition 4.2 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 3.2. 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.

Square matrices provide a natural example of a noncommutative ring with zero divisors.

Example 4.1. 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

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

We now formalize the representation of linear maps by matrices.

Definition 4.3. 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 $x = x_1 u_1 + \cdots + x_n u_n$ 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

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

for every j, $1 \leq j \leq n$. Explicitly, we have

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

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 4.1, 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 $(v_1, \ldots, v_m)$.

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

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

Recall that

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

Sometimes, it is necessary to incoporate the bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_m)$ 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: write $\mathcal{U} = (u_1, \ldots, u_n)$ and $\mathcal{V} = (v_1, \ldots, v_m)$ 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 $M(x) = (x_1, \ldots, x_n)$ of $x \in E$ w.r.t. the basis $\mathcal{U}$ and write $y_{\mathcal{V}}$ for the coordinates $M(y) = (y_1, \ldots, y_m)$ 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 [42] use the notation $[f]_{\mathcal{U},\mathcal{V}}$, and others such as Dummit and Foote [17] 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 [35]), 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,

$$\begin{aligned} d(P+Q) &= dP + dQ \\ d(\lambda P) &= \lambda dP, \end{aligned}$$

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

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

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 4.1. *(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 4.1 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 4.2. *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 $f\colon F \to G$:*

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

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 4.1), 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 4.1 is used to show that M_n is a ring). □

In view of Proposition 4.2, 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.

It is important to observe that the isomorphism $M\colon \mathrm{Hom}(E, F) \to \mathrm{M}_{n,p}$ given by Proposition 4.2 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 4.3. *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 3.9, 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 4.2, 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 4.3 suggests the following definition.

Definition 4.4. 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 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 $P = (a_{i\,j})$ 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 $x = x_1 u_1 + \cdots + x_n u_n$ over the basis $(u_1, \ldots, u_n)$ and $x = x'_1 v_1 + \cdots + x'_n v_n$ over the basis $(v_1, \ldots, v_n)$, from Proposition 4.2, 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 [42] uses the notation $[I]_{\mathcal{V},\mathcal{U}}$).

We prefer to use an abbreviation for $M_{\mathcal{V},\mathcal{U}}(\mathrm{id})$ and we will use the notation

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

for the *change of basis matrix from* $\mathcal{U}$ *to* $\mathcal{V}$. 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 [2]) 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 4.2. 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 4.3. 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.

4.2 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 4.1, we compute averages and

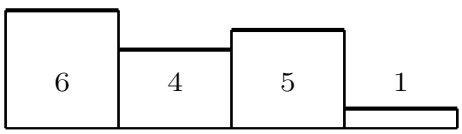

Figure 4.1: The original signal v

half differences obtaining Figure 4.2. We get the coefficients $c_1 = 4$ and $c_2 = 1$. Again, the signal on the left of Figure 4.2 can be reconstructed from the two signals in Figure 4.3.

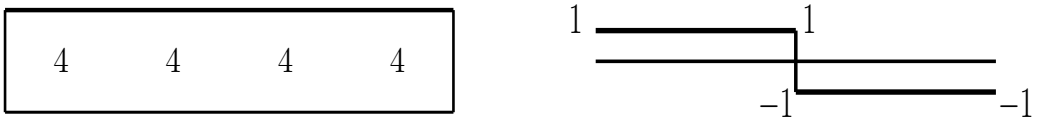

Figure 4.3: Second averages and second half differences

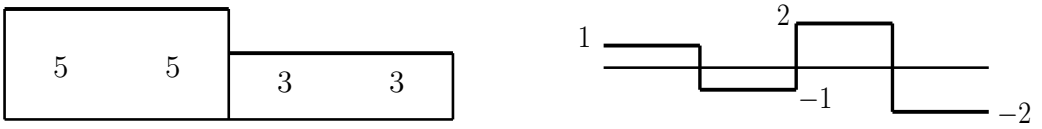

Figure 4.2: First averages and first half differences

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 begining 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

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

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. In this case, the Haar basis is denoted by

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

and

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

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

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 $[0, 1)$. We define the function $\mathrm{plf}(u)$ 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 $[(m-1)/m, 1)$. 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 4.4.

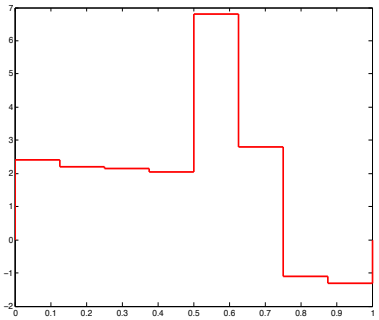

Figure 4.4: The piecewise linear function $\mathrm{plf}(u)$

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 4.5. Then, it is easy to see that ψ_k^j is given by the simple

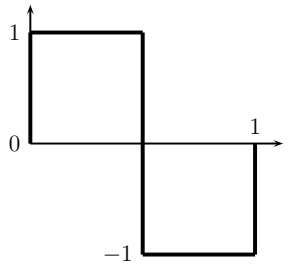

Figure 4.5: The Haar wavelet ψ

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. 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 ψ_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.

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 and 2^{n-j-1}. 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 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 $n=3$:

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

Then, we form the $2^n \times 2^{n-2}$ matrix obtained by "doubling" each column of odd index, which means replacing each such column by a column in which the block of 1 is doubled and the block of -1 is doubled. In general, given a current matrix of dimension $2^n \times 2^j$, we form a $2^n \times 2^{j-1}$ matrix by doubling each column of odd index, which means that we replace each such column by a column in which the block of 1 is doubled and the block of -1 is doubled. We repeat this process $n-1$ times until we get the vector

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

The first vector is the averaging vector $\underbrace{(1,\ldots,1)}_{2^n}$. This process is illustrated below for $n=3$:

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

Adding $\underbrace{(1,\ldots,1,1,\ldots,1)}_{2^n}$ as the first column, we obtain

$$W_3 = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0\\ 1 & 1 & 1 & 0 & -1 & 0 & 0 & 0\\ 1 & 1 & -1 & 0 & 0 & 1 & 0 & 0\\ 1 & 1 & -1 & 0 & 0 & -1 & 0 & 0\\ 1 & -1 & 0 & 1 & 0 & 0 & 1 & 0\\ 1 & -1 & 0 & 1 & 0 & 0 & -1 & 0\\ 1 & -1 & 0 & -1 & 0 & 0 & 0 & 1\\ 1 & -1 & 0 & -1 & 0 & 0 & 0 & -1 \end{pmatrix}.$$

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 4.6. Since some of

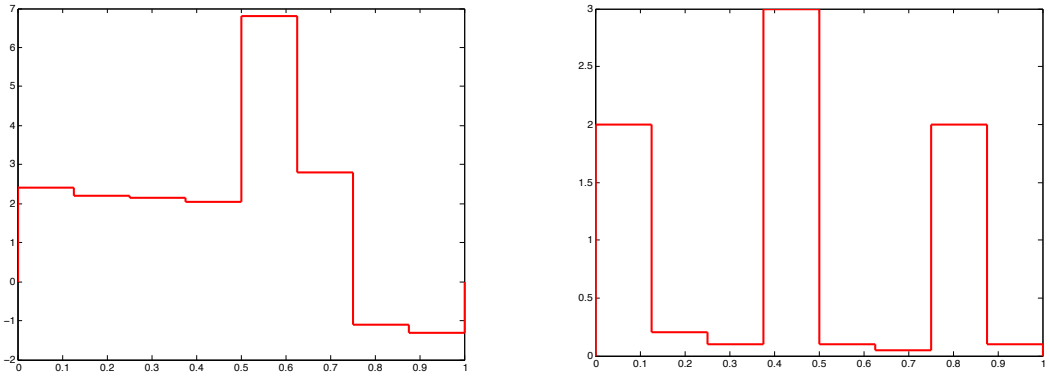

Figure 4.6: A signal and its Haar transform

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 4.7.

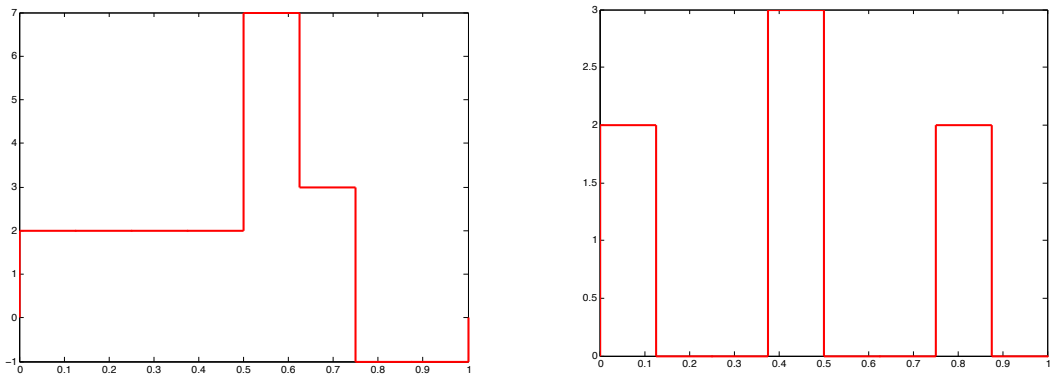

Figure 4.7: A compressed signal and its compressed Haar transform

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 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 4.8. Then, run a program that sets all

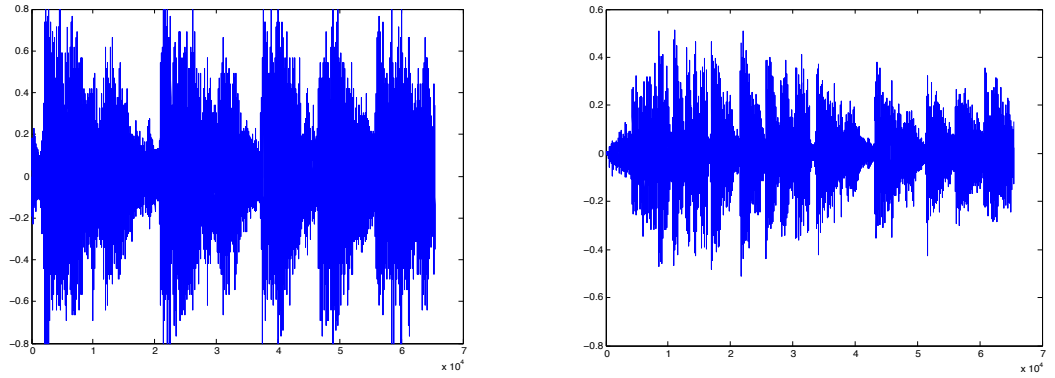

Figure 4.8: The signal "handel" and its Haar transform

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 4.9. When you type `sound(y2)`, you find that the music 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

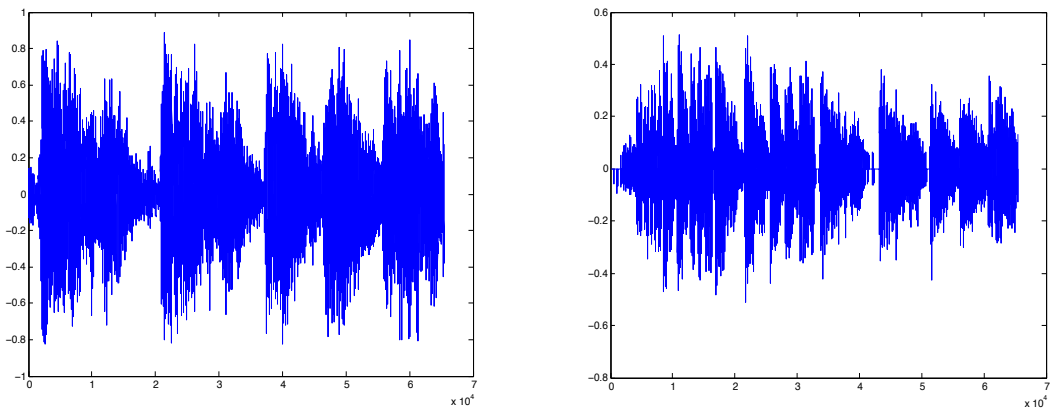

Figure 4.9: The compressed signal "handel" and its Haar transform

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 $D_n = \mathrm{diag}\Big(2^{-n}, \underbrace{2^{-n}}_{2^0}, \underbrace{2^{-(n-1)}, 2^{-(n-1)}}_{2^1}, \underbrace{2^{-(n-2)}, \ldots, 2^{-(n-2)}}_{2^2}, \ldots, \underbrace{2^{-1}, \ldots, 2^{-1}}_{2^{n-1}}\Big)$.

Therefore, we define the orthogonal matrix

$$H_n = W_n D_n^{\frac{1}{2}}$$

whose columns are the normalized Haar basis vectors, with

$$D_n^{\frac{1}{2}} = \mathrm{diag}\Big(2^{-\frac{n}{2}}, \underbrace{2^{-\frac{n}{2}}}_{2^0}, \underbrace{2^{-\frac{n-1}{2}}, 2^{-\frac{n-1}{2}}}_{2^1}, \underbrace{2^{-\frac{n-2}{2}}, \ldots, 2^{-\frac{n-2}{2}}}_{2^2}, \ldots, \underbrace{2^{-\frac{1}{2}}, \ldots, 2^{-\frac{1}{2}}}_{2^{n-1}}\Big).$$

We call H_n the *normalized Haar transform matrix*. Because H_n is orthogonal, $H_n^{-1} = H_n^\top$. Given a vector (signal) u, we call $c = H_n^\top u$ the *normalized Haar coefficients* of u. 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 [51]) 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 a 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)`, 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 4.10.

The compressed 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

$$C = H_m^\top A H_n$$
$$A = H_m C H_n^\top.$$

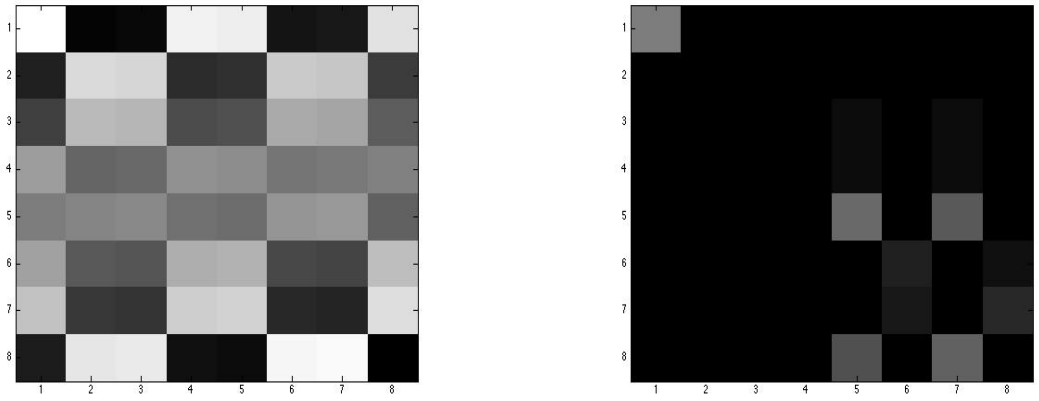

Figure 4.10: An image and its Haar transform

The compressed images corresponding to A_2 and C_2 are shown in Figure 4.11.

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.

Oberve 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.* This process is illustrated on the image shown in Figure 4.12.

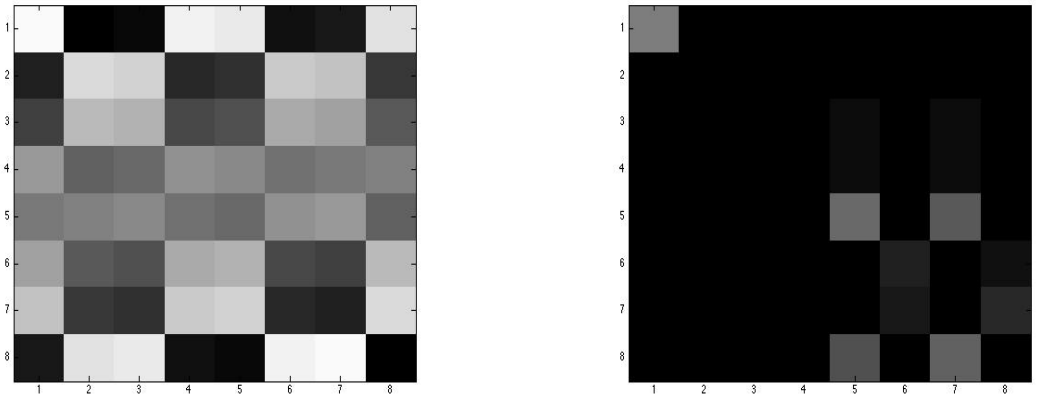

Figure 4.11: Compressed image and its Haar transform

Figure 4.12: Original drawing by Durer

The result of performing one round, two rounds, three rounds, and nine rounds of averaging is shown in Figure 4.13. 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.

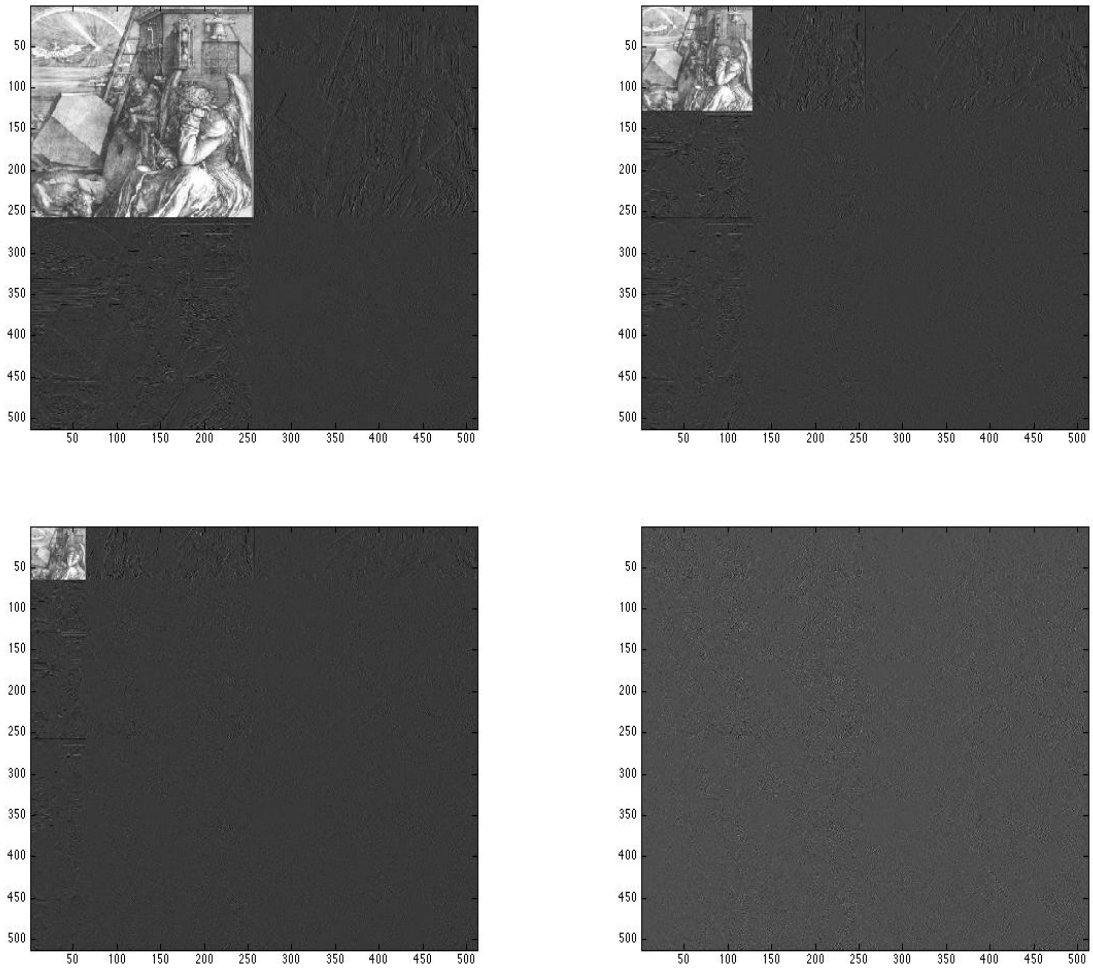

Figure 4.13: Haar tranforms after one, two, three, and nine rounds of averaging

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 [51]. A very detailed account is given in Strang and and Nguyen [54], 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} 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_j 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} for the 2D Haar transform, in the sense that for any matrix A, its matrix C of Haar coefficients is given by

$$C = \sum_{i=1}^{2^n} \sum_{j=1}^{2^n} a_{ij} h_{ij}.$$

If $W^{-1} = (w_{ij}^{-1})$, then

$$h_{ij} = w_i^{-1} (w_j^{-1})^\top.$$

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`.

4.3 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 4.4. *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 4.2, we have $M'(f) = Q^{-1} M(f) P$. □

As a corollary, we get the following result.

Corollary 4.5. *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 4.4. 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. Therefore, 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 4.5. 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 than we will not discuss in these notes (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.

4.4 Affine Maps

We showed in Section 3.4 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 3.4 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 is always an affine combination, but an affine combination is a linear combination, *with the restriction that the scalars λ_i must add up to* 1. 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 4.6. *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 4.6 also holds.

Proposition 4.7. *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 4.7 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 4.6 and 4.7, it is natural to make the following definition.

Definition 4.6. 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

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

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*, but 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 $(u_1, \ldots, u_n)$, we write

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

and over the basis $(v_1, \ldots, v_m)$, 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

$$y = Ax + d, \quad x \in \mathbb{R}^n.$$

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 4.14 illustrates this process for $n = 2$.

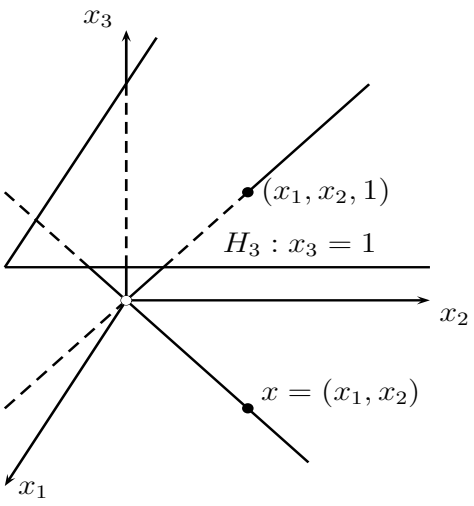

Figure 4.14: Viewing $\mathbb{R}^n$ as a hyperplane in $\mathbb{R}^{n+1}$ $(n = 2)$

For example, the map

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} 1 & 1 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 3 \\ 0 \end{pmatrix}$$

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 be used to define *projective maps*.

4.5 Summary

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

- 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.
- *Column vectors*, *row vectors*.
- *Matrix operations*: addition, scalar multiplication, multiplication.

- The *matrix representation mapping* $M\colon \mathrm{Hom}(E, F) \to \mathrm{M}_{n,p}$ and the representation isomorphism (Proposition 4.2).
- *Change of basis matrix* and Proposition 4.4.
- Haar basis vectors and a glimpse at *Haar wavelets*.
- *Affine maps* and their representations in terms of matrices.

Chapter 5

Determinants

5.1 Definition Using Expansion by Minors

Every square matrix A has a number associated to it and called its *determinant*, denoted by $\det(A)$. One of the most important properties of a determinant is that it gives us a criterion to decide whether the matrix is invertible:

A matrix A is invertible iff $\det(A) \neq 0$.

It is possible to define determinants in terms of a fairly complicated formula involving $n!$ terms (assuming A is a $n \times n$ matrix) but this way to proceed makes it more difficult to prove properties of determinants.

Consequently, we follow a more algorithmic approach due to Mike Artin. This approach avoids dealing with the sign of permutations (at least, not until we need an explicit formula for the determinant).

We will view the determinant as *a function of the rows of an $n \times n$ matrix*. Formally, this means that

$$\det : (\mathbb{R}^n)^n \to \mathbb{R}.$$

We will define the determinant *recursively* using a process called *expansion by minors*. Then, we will derive properties of the determinant and prove that there is a unique function satisfying these properties. As a consequence, we will have an axiomatic definition of the determinant.

For a 1×1 matrix $A = (a)$, we have

$$\det(A) = \det(a) = a.$$

For a 2×2 matrix,

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

it will turn out that

$$\det(A) = ad - bc.$$

The determinant has a geometric interpretation as a signed area, in higher dimension as a signed volume.

In order to describe the recursive process to define a determinant we need the notion of a minor.

Definition 5.1. 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*:

$$A_{ij} = \begin{bmatrix} & & & & \times & & \\ & & & & \times & & \\ \times & \times & \times & \times & \times & \times & \times \\ & & & & \times & & \\ & & & & \times & & \\ & & & & \times & & \\ & & & & \times & & \end{bmatrix}$$

For example, if

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

then

$$A_{23} = \begin{bmatrix} 2 & -1 & 0 & 0 \\ 0 & -1 & -1 & 0 \\ 0 & 0 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}.$$

We can now proceed with the definition of determinants.

Definition 5.2. Given any $n \times n$ matrix $A = (a_{ij})$, if $n = 1$, then

$$\det(A) = a_{11},$$

else

$$\det(A) = a_{11}\det(A_{11}) + \cdots + (-1)^{i+1}a_{i1}\det(A_{i1}) + \cdots + (-1)^{n+1}a_{n1}\det(A_{n1}), \qquad (*)$$

the expansion by minors on the first column.

When $n = 2$, we have

$$\det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = a_{11} \det[a_{22}] - a_{21} \det[a_{12}] = a_{11}a_{22} - a_{21}a_{12},$$

which confirms the formula claimed earlier. When $n = 3$, we get

$$\det \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = a_{11} \det \begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix} - a_{21} \det \begin{bmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{bmatrix} + a_{31} \det \begin{bmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{bmatrix},$$

and using the formula for a 2×2 determinant, we get

$$\det \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = a_{11}(a_{22}a_{33} - a_{32}a_{23}) - a_{21}(a_{12}a_{33} - a_{32}a_{13}) + a_{31}(a_{12}a_{23} - a_{22}a_{13}).$$

As we can see, the formula is already quite complicated!

Given a $n \times n$-matrix $A = (a_{i\,j})$, its determinant $\det(A)$ is also denoted by

$$\det(A) = \begin{vmatrix} a_{1\,1} & a_{1\,2} & \dots & a_{1\,n} \\ a_{2\,1} & a_{2\,2} & \dots & a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n\,1} & a_{n\,2} & \dots & a_{n\,n} \end{vmatrix}.$$

We now derive some important and useful properties of the determinant. Recall that we view the determinant $\det(A)$ as a function of the rows of the matrix A, so we can write

$$\det(A) = \det(A_1, \dots, A_n),$$

where $A_1, \dots, A_n$ are the rows of A.

Proposition 5.1. *The determinant function* $\det\colon (\mathbb{R}^n)^n \to \mathbb{R}$ *satisfies the following properties:*

(1) $\det(I) = 1$, *where* I *is the identity matrix.*

(2) The determinant is linear in each of its rows; this means that

$$\begin{aligned} \det(A_1, \dots, A_{i-1}, B + C, A_{i+1}, \dots, A_n) = {} & \det(A_1, \dots, A_{i-1}, B, A_{i+1}, \dots, A_n) \\ & + \det(A_1, \dots, A_{i-1}, C, A_{i+1}, \dots, A_n) \end{aligned}$$

and

$$\det(A_1, \dots, A_{i-1}, \lambda A_i, A_{i+1}, \dots, A_n) = \lambda \det(A_1, \dots, A_{i-1}, A_i, A_{i+1}, \dots, A_n).$$

(3) If two adjacent rows of A are equal, then $\det(A) = 0$. *This means that*

$$\det(A_1, \ldots, A_i, A_i, \ldots, A_n) = 0.$$

Property (2) says that det is a *multilinear map*, and property (3) says that det is an *alternating map*.

Proposition 5.1 is not hard to prove by direct computations and by induction. Let us verify property (3).

Proof of Property (3). Suppose that row j and $j+1$ are equal. Then, the $(n-1) \times (n-1)$ matrices A_{i1} also have two identical rows, except when $i = j$ or $i = j+1$. By the induction hypothesis, if $i \neq j, j+1$, then

$$\det(A_{i\,1}) = 0.$$

Consequently, $(*)$ yields

$$\det(A) = (-1)^{j+1} a_{j\,1} \det(A_{j1}) + (-1)^{j+2} a_{j+1\,1} \det(A_{j+11}).$$

But since the rows A_j and A_{j+1} are equal, we must have $A_{j\,1} = A_{j+1\,1}$ and $a_{j\,1} = a_{j+1\,1}$, so

$$\det(A) = (-1)^{j+1} a_{j\,1} \det(A_{j1}) - (-1)^{j+1} a_{j\,1} \det(A_{j1}) = 0,$$

as claimed. □

We now derive more useful properties from Proposition 5.1.

Proposition 5.2. *The determinant function* $\det\colon (\mathbb{R}^n)^n \to \mathbb{R}$ *satisfies the following properties:*

(4) If two adjacent rows are interchanged, then the determinant is multiplied by -1*; thus,*

$$\det(A_1, \ldots, A_{i+1}, A_i, \ldots, A_n) = -\det(A_1, \ldots, A_i, A_{i+1}, \ldots, A_n).$$

(5) If two rows are identical then the determinant is zero; that is,

$$\det(A_1, \ldots, A_i, \ldots, A_i, \ldots, A_n) = 0.$$

(6) If any two distinct rows of A are are interchanged, then the determinant is multiplied by -1*; thus,*

$$\det(A_1, \ldots, A_j, \ldots, A_i, \ldots, A_n) = -\det(A_1, \ldots, A_i, \ldots, A_j, \ldots, A_n).$$

(7) If a multiple of a row is added to another row, the determinant is unchanged; that is,

$$\det(A_1, \ldots, A_i + \lambda A_j, \ldots, A_n) = \det(A_1, \ldots, A_i, \ldots, A_n).$$

(8) If any row of A is zero, then $\det(A) = 0$.

Proof. (4) Observe that by linearity and property (3) by have

$$\begin{aligned}
0 &= \det(A_1, \ldots, A_i + A_{i+1}, A_i + A_{i+1}, \ldots, A_n) \\
&= \det(A_1, \ldots, A_i, A_i + A_{i+1}, \ldots, A_n) + \det(A_1, \ldots, A_{i+1}, A_i + A_{i+1}, \ldots, A_n) \\
&= \det(A_1, \ldots, A_i, A_i, \ldots, A_n) + \det(A_1, \ldots, A_i, A_{i+1}, \ldots, A_n) \\
&\quad + \det(A_1, \ldots, A_{i+1}, A_i, \ldots, A_n) + \det(A_1, \ldots, A_{i+1}, A_{i+1}, \ldots, A_n) \\
&= \det(A_1, \ldots, A_i, A_{i+1}, \ldots, A_n) + \det(A_1, \ldots, A_{i+1}, A_i, \ldots, A_n),
\end{aligned}$$

which shows that

$$\det(A_1, \ldots, A_{i+1}, A_i, \ldots, A_n) = -\det(A_1, \ldots, A_i, A_{i+1}, \ldots, A_n).$$

(5) If A_i and A_j are not adjacent we can interchange A_i and A_{i+1}, and then A_{i+1} and A_{i+2}, etc., until A_i and A_j become adjacent. By (4),

$$\det(A_1, \ldots, A_i, \ldots, A_j, \ldots, A_n) = \epsilon \det(A_1, \ldots, A_i, A_j, \ldots, A_n),$$

with $\epsilon = \pm 1$. However, if $A_i = A_j$, then by (3) $\det(A_1, \ldots, A_i, A_j, \ldots, A_n) = 0$, and so $\det(A_1, \ldots, A_i, \ldots, A_i, \ldots, A_n) = 0$.

(6) The proof is the same as in (4) using (5) instead of (3).

(7) Using linearity and (6) we have

$$\det(A_1, \ldots, A_i + \lambda A_j, \ldots, A_n) = \det(A_1, \ldots, A_i, \ldots, A_n) + \lambda \det(A_1, \ldots, A_j, \ldots, A_n),$$

but the matrix $[A_1 \cdots A_j \cdots A_n]$ contains the row A_j in two different positions, so $\det(A_1, \ldots, A_j, \ldots, A_n) = 0$, and (7) holds.

(8) This is an immediate consequence of linearity. □

Using property (6), it is easy to show that the expansion by minors formula $(*)$ can be adapted to any column. Indeed, we have

$$\det(A) = (-1)^{j+1} a_{1j} \det(A_{1j}) + \cdots + (-1)^{j+i} a_{ij} \det(A_{ij}) + \cdots + (-1)^{j+n} a_{nj} \det(A_{nj}). \quad (**)$$

The beauty of this approach is that properties (6) and (7) describe the effect of the elementary operations $P(i, j)$ and $E_{i,j,\lambda}$ on the determinant: Indeed, (6) says that

$$\det(P(i, j)A) = -\det(A), \quad \text{(a)}$$

and (7) says that

$$\det(E_{i,j;\lambda}A) = \det(A). \quad \text{(b)}$$

Furthermore, linearity (propery (2)) says that

$$\det(E_{i,\lambda}A) = \lambda \det(A). \quad \text{(c)}$$

Substituting the identity I for A in the above equations, since $\det(I) = 1$, we find the determinants of the elementary matrices:

(1) For any permutation matrix $P(i,j)$ $(i \neq j)$, we have

$$\det(P(i,j)) = -1.$$

(2) For any row operation $E_{i,j;\lambda}$ (adding λ times row j to row i), we have

$$\det(E_{i,j;\lambda}) = 1.$$

(3) For any row operation $E_{i,\lambda}$ (multiplying row i by λ), we have

$$\det(E_{i,\lambda}) = \lambda.$$

The above properties together with the equations (a), (b), (c) yield the following important proposition:

Proposition 5.3. *For every $n \times n$ matrix A and every elementary matrix E, we have*

$$\det(EA) = \det(E)\det(A).$$

We can now use Proposition 5.3 and the reduction to row echelon form to compute $\det(A)$. Indeed, recall that we showed (just before Proposition 2.17)) that every square matrix A can be reduced by elementary operations to a matrix A' which is either the identity or else whose last row is zero,

$$A' = E_k \cdots E_1 A.$$

If $A' = I$, then $\det(A') = 1$ by (1), else if A' has a zero row, then $\det(A') = 0$ by (8). Furthermore, by induction using Proposition 5.3 (see the proof of Proposition 5.7), we get

$$\det(A') = \det(E_k \cdots E_1 A) = \det(E_k) \cdots \det(E_1)\det(A).$$

Since all the determinants, $\det(E_k)$ of the elementary matrices E_i are known, we see that the formula

$$\det(A') = \det(E_k) \cdots \det(E_1)\det(A)$$

determines $\det(A)$. As a consequence, we have the following characterization of a determinant:

Theorem 5.4. *(Axiomatic Characterization of the Determinant) The determinant* det *is the unique function $f\colon (\mathbb{R}^n)^n \to \mathbb{R}$ satisfying properties (1), (2), and (3) of Proposition 5.1.*

Proof. We already proved that for every square matrix A, if $A' = E_k \cdots E_1 A$ is the reduced row echelon form of A, then

$$\det(A') = \det(E_k) \cdots \det(E_1)\det(A).$$

Now, observe that the proof that properties (4)–(8) follow from (1)–(3) applies to any function f satisfying properties (1)–(3). So for such a function, Proposition 5.3 also holds, and then we have

$$f(A') = f(E_k) \cdots f(E_1) f(A).$$

Furthermore, the same reasoning used for det shows that (a), (b), (c) hold, so $f(E) = \det(E)$ for all elementary matrices E, and $f(A') = 1$ iff $A' = I$ or $f(A') = 0$ iff A' has a zero row. It follows that $f(A) = \det(A)$ for all A. □

Instead of evaluating a determinant using expansion by minors on the columns, we can use expansion by minors on the rows. Indeed, define the function D given

$$D(A) = (-1)^{i+1} a_{i\,1} D(A_{i\,1}) + \cdots + (-1)^{i+n} a_{i\,n} D(A_{i\,n}), \tag{$\dagger$}$$

with $D([a]) = a$. Then, it is fairly easy to show that the properties of Proposition 5.1 hold for D, and thus, by Theorem 5.4, the function D also defines the determinant, that is,

$$D(A) = \det(A).$$

The definition of the determinant in terms of D can be used to prove that a matrix and its transpose have the same determinant. The proof is left as an exercise.

Proposition 5.5. *For any square matrix A, we have* $\det(A) = \det(A^\top)$.

We also obtain the important characterization of invertibility of a matrix in terms of its determinant.

Proposition 5.6. *A square matrix A is invertible iff* $\det(A) \neq 0$.

Proof. We know that for every square matrix A, if $A' = E_k \cdots E_1 A$ is the reduced row echelon form of A, then

$$\det(A') = \det(E_k) \cdots \det(E_1) \det(A).$$

Note that if any E_i is of the form $E_{k,\lambda}$, then $\lambda \neq 0$ since a row operation only rescales a pivot iff it is nonzero. Therefore $\det(E_i) \neq 0$ for $i = 1, \ldots, k$, and thus $\det(A) \neq 0$ iff $\det(A') \neq 0$. However, by Proposition 2.17, the matrix A is invertible iff $A' = I$, in which case $\det(A') = 1$. □

We can now prove one of the most useful properties of determinants.

Proposition 5.7. *Given any two $n \times n$ matrices A and B, we have*

$$\det(AB) = \det(A) \det(B).$$

Proof. First, assume that A is invertible. In this case, by Proposition 2.17, the matrix A is of product of elementary matrices

$$A = E_1 \cdots E_k.$$

We prove by induction on k that

$$\det(E_1 \cdots E_k B) = \det(E_1) \cdots \det(E_k) \det(B)$$

for any matrix B.

If $k = 1$, then by Proposition 5.3,

$$\det(E_1 B) = \det(E_1) \det(B).$$

For the induction step, by Proposition 5.3 again,

$$\det(E_1 \cdots E_k B) = \det(E_1) \det(E_2 \cdots E_k B),$$

and since by the induction hypothesis,

$$\det(E_2 \cdots E_k B) = \det(E_2) \cdots \det(E_k) \det(B),$$

we get

$$\det(E_1 \cdots E_k B) = \det(E_1) \cdots \det(E_k) \det(B).$$

Applying the above to $B = I$, we get

$$\det(A) = \det(E_1) \cdots \det(E_k),$$

so

$$\det(AB) = \det(E_1 \cdots E_k B) = \det(E_1) \cdots \det(E_k) \det(B) = \det(A) \det(B).$$

Let us now consider the case where A is singular. In this case, by Proposition 5.6, we have $\det(A) = 0$, and we just have to show that $\det(AB) = 0$. By Proposition 5.6 again, we know that A can be reduced to an echelon matrix

$$A' = E_k \cdots E_1 A$$

using some elementary matrices E_i, where the last row of A' is zero. Then, the last row of $A'B$ is also zero, and since

$$0 = \det(A'B) = \det(E_k \cdots E_1 AB) = \det(E_k) \cdots \det(E_1) \det(AB),$$

and $\det(E_i) \neq 0$ for $i = 1, \ldots, k$, we must have $\det(AB) = 0$. □

In order to give an explicit formula for the determinant, we need to discuss some properties of permutation matrices.

5.2 Permutations and Permutation Matrices

Let $[n] = \{1, 2 \ldots, n\}$, where $n \in \mathbb{N}$, and $n > 0$.

Definition 5.3. 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 have already encountered transpositions before, but represented by the matrices $P(i, j)$. We will also use the terminology product of permutations (or transpositions), as a synonym for composition of permutations.

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 $\mathfrak{S}_n$. It is easy to show by induction that the group $\mathfrak{S}_n$ has $n!$ elements.

There are various ways of denoting permutations. One way is to use a functional notation such as

$$\begin{pmatrix} 1 & 2 & \cdots & i & \cdots & n \\ \pi(1) & \pi(2) & \cdots & \pi(i) & \cdots & \pi(n) \end{pmatrix}.$$

For example the permutation $\pi\colon [4] \to [4]$ given by

$$\begin{aligned} \pi(1) &= 3 \\ \pi(2) &= 4 \\ \pi(3) &= 2 \\ \pi(4) &= 1 \end{aligned}$$

is represented by

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

The above representation has the advantage of being compact, but a matrix representation is also useful and has the advantage that composition of permutations corresponds to matrix multiplication.

A permutation can be viewed as an operation permuting the rows of a matrix. For example, the permutation

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 2 & 1 \end{pmatrix}$$

corresponds to the matrix

$$P_\pi = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Observe that the matrix P_π has a single 1 on every row and every column, all other entries being zero, and that if we multiply any 4×4 matrix A by P_π on the left, then the rows of $P_\pi A$ are permuted according to the permutation π; that is, the $\pi(i)$th row of $P_\pi A$ is the ith row of A. For example,

$$P_\pi A = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} a_{41} & a_{42} & a_{43} & a_{44} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \end{bmatrix}.$$

Equivalently, the ith row of $P_\pi A$ is the $\pi^{-1}(i)$th row of A. In order for the matrix P_π to move the ith row of A to the $\pi(i)$th row, the $\pi(i)$th row of P_π must have a 1 in column i and zeros everywhere else; this means that the ith column of P_π contains the basis vector $e_{\pi(i)}$, the vector that has a 1 in position $\pi(i)$ and zeros everywhere else.

This is the general situation and it leads to the following definition.

Definition 5.4. Given any permutation $\pi\colon [n] \to [n]$, the *permutation matrix* $P_\pi = (p_{ij})$ representing π is the matrix given by

$$p_{ij} = \begin{cases} 1 & \text{if } i = \pi(j) \\ 0 & \text{if } i \neq \pi(j); \end{cases}$$

equivalently, the jth column of P_π is the basis vector $e_{\pi(j)}$. A *permutation matrix* P is any matrix of the form P_π (where P is an $n \times n$ matrix, and $\pi\colon [n] \to [n]$ is a permutation, for some $n \geq 1$).

Remark: There is a confusing point about the notation for permutation matrices. A permutation matrix P acts on a matrix A by multiplication on the left by permuting the rows of A. As we said before, this means that the $\pi(i)$th row of $P_\pi A$ is the ith row of A, or equivalently that the ith row of $P_\pi A$ is the $\pi^{-1}(i)$th row of A. But then, observe that the row index of the entries of the ith row of PA is $\pi^{-1}(i)$, and not $\pi(i)$! See the following example:

$$\begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} a_{41} & a_{42} & a_{43} & a_{44} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \end{bmatrix},$$

where

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

The following proposition is easy to show and is left as an exercise.

Proposition 5.8. *The following properties hold:*

(1) Given any two permutations $\pi_1, \pi_2 \colon [n] \to [n]$, the permutation matrix $P_{\pi_2 \circ \pi_1}$ representing the composition of π_1 and π_2 is equal to the product $P_{\pi_2} P_{\pi_1}$ of the permutation matrices P_{π_1} and P_{π_2} representing π_1 and π_2; that is,

$$P_{\pi_2 \circ \pi_1} = P_{\pi_2} P_{\pi_1}.$$

(2) The matrix $P_{\pi_1^{-1}}$ representing the inverse of the permutation π_1 is the inverse $P_{\pi_1}^{-1}$ of the matrix P_{π_1} representing the permutation π_1; that is,

$$P_{\pi_1^{-1}} = P_{\pi_1}^{-1}.$$

Furthermore,

$$P_{\pi_1}^{-1} = (P_{\pi_1})^\top.$$

The following proposition shows the importance of transpositions.

Proposition 5.9. *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 5.9 shows that the transpositions generate the group of permutations $\mathfrak{S}_n$.

Since we already know that the determinant of a transposition matrix is -1, Proposition 5.9 implies that for every permutation matrix P, we have

$$\det(P) = \pm 1.$$

We can say more. Indeed if a given permutation π is factored into two different products of transpositions $\tau_p \circ \cdots \circ \tau_1$ and $\tau'_q \circ \cdots \circ \tau'_1$, because

$$det(P_\pi) = \det(P_{\tau_p}) \cdots \det(P_{\tau_1}) = \det(P_{\tau'_q}) \cdots \det(P_{\tau'_1}),$$

and $\det(P_{\tau_i}) = \det(P_{\tau'_j}) = -1$, the natural numbers p and q have the same parity.

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).

Definition 5.5. For every permutation σ of $[n]$, the parity $\epsilon(\sigma)$ of the number of transpositions involved in any decomposition of σ is called the *signature* of σ. We have $\epsilon(\sigma) = \det(P_\sigma)$.

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 $(-1)^0 = \epsilon(\mathrm{id}) = +1$.

It is also easy to see 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 are now ready to give an explicit formula for a determinant.

Given an $n \times n$ matrix A (with $n \geq 2$), we can view its first row A_1 as the sum of the n rows

$$[a_{11}\, 0 \cdots 0],\ [0\, a_{12}\, 0 \cdots 0],\ \ldots,\ [0 \cdots 0\, a_{1n}],$$

and we can expand A by linearity as

$$\det(A) = \det \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \end{bmatrix} + \det \begin{bmatrix} 0 & a_{12} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \end{bmatrix} + \cdots + \det \begin{bmatrix} \cdots & \cdots & 0 & a_{1n} \\ \cdots & \cdots & \vdots & \vdots \\ \vdots & \cdots & \vdots & \vdots \\ \cdots & \cdots & \vdots & \vdots \end{bmatrix}.$$

We can repeat this process on the second row, the third row, etc. At the end, we obtain a sum of determinants of matrices of the form

$$M = \begin{bmatrix} & & a_{1?} & & \\ & a_{2?} & & & \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ & & & & a_{n?} \end{bmatrix}$$

having a single entry left in each row, all the others being zero. Observe that all the determinants involving matrices having a zero column will be zero. Actually, the only determinants that survive are those that have a single entry a_{ij} in each row and each column. Such matrices are very similar to permutation matrices. In fact, they must be of the form $M_\pi = (m_{ij})$ for some permutation π of $[n]$, with

$$m_{ij} = \begin{cases} a_{ij} & \text{if } i = \pi(j) \\ 0 & \text{if } i \neq \pi(j). \end{cases}$$

Consequently, by multilinearity of determinants, we have

$$\begin{aligned} \det(A) &= \sum_{\pi \in \mathfrak{S}_n} a_{\pi(1)1} \cdots a_{\pi(n)n} \det(P_\pi) \\ &= \sum_{\pi \in \mathfrak{S}_n} \epsilon(\pi) a_{\pi(1)1} \cdots a_{\pi(n)n}. \end{aligned}$$

We summarize the above derivation as the following proposition which gives the *complete expansion* of the determinant.

Proposition 5.10. *For any $n \times n$ matrix $A = (a_{ij})$, we have*

$$\det(A) = \sum_{\pi \in \mathfrak{S}_n} \epsilon(\pi) a_{\pi(1)1} \cdots a_{\pi(n)n}.$$

Note that since $\det(A) = \det(A^\top)$, we also have

$$\det(A) = \sum_{\pi \in \mathfrak{S}_n} \epsilon(\pi) a_{1\pi(1)} \cdots a_{n\pi(n)}.$$

These formulae are more of theoretical than of practical importance. However, these formulae do show that the determinant is a polynomial function in the n^2 variables a_{ij}, and this has some importance consequences.

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 $\mathbb{R}^n$.

As we saw, the determinant of a matrix is a multilinear alternating map of its rows. This fact, combined with the fact that the determinant of a matrix is also 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 5.1. 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.

5.3 Inverse Matrices and Determinants

In the next two sections, K is a commutative ring and when needed, a field.

Definition 5.6. 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 5.11. *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 5.7, 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 $K = \mathbb{C}$).

5.4 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 5.12. *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 1.11 the matrix A is invertible, so by Proposition 5.6 we have $\det(A) \neq 0$. Finally, if $\det(A) \neq 0$, by Proposition 5.6 the matrix 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. Then, by Proposition 1.11, the matrix A is singular, and by Proposition 5.6, this 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.

5.5 Determinant of a Linear Map

In this section we define the determinant of a linear map $f\colon E \to E$.

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 4.5, the matrix of f with respect to the basis $(v_1, \ldots, v_n)$ is $P^{-1}M(f)P$. Now, by proposition 5.7, 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 5.7. 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 5.13. *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 $f: E \to E$ is invertible iff its matrix $M(f)$ in any basis is invertible (by Proposition 4.2), iff $\det(M(f)) \neq 0$, by Proposition 5.11. □

Given a vector space of finite dimension n, it is easily seen that the set of bijective linear maps $f: E \to E$ 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 $\mathbf{SL}(n)$.

5.6 The Cayley–Hamilton Theorem

We conclude this chapter with an interesting and important application of Proposition 5.11, 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 5.8. 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 5.14. *(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 5.11, 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

$$X^n D_0 + X^{n-1}D_1 + \cdots + D_n = (X^n + c_1X^{n-1} + \cdots + c_n)I$$

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

$$\begin{aligned} I &= B_0 \\ c_1 I &= B_1 - AB_0 \\ &\vdots \\ c_j I &= B_j - AB_{j-1} \\ &\vdots \\ c_{n-1} I &= B_{n-1} - AB_{n-2} \\ c_n I &= -AB_{n-1}, \end{aligned}$$

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

$$A^n + c_1 A^{n-1} + \cdots + c_n I = 0,$$

as claimed. □

As a concrete example, when $n = 2$, the matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

satisfies the equation

$$A^2 - (a+d)A + (ad - bc)I = 0.$$

5.7 Further Readings

Thorough expositions of the material covered in Chapter 1, 2, 3, 4, and Chapter 5 can be found in Strang [53, 52], Lax [38], Meyer [42], Artin [2], Lang [35], Mac Lane and Birkhoff [39], Hoffman and Kunze [33], Dummit and Foote [17], Bourbaki [6, 7], Van Der Waerden [58], Serre [48], Horn and Johnson [30], and Bertin [5]. These notions of linear algebra are nicely put to use in classical geometry, see Berger [3, 4], Tisseron [55] and Dieudonné [15].

Chapter 6

Euclidean Spaces

Rien n'est beau que le vrai.
—**Hermann Minkowski**

6.1 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. 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 using the Gram–Schmidt orthogonalization procedure. Using orthonormal bases, we show that every linear map has an adjoint. 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 12.

For a more detailed treatment of Euclidean geometry, see Berger [3, 4], Snapper and Troyer [49], or any other book on geometry, such as Pedoe [44], Coxeter [13], Fresnel [22], Tisseron [55], or Cagnac, Ramis, and Commeau [10]. Serious readers should consult Emil

Artin's famous book [1], 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 [27, 28] and Rouché and de Comberousse [45]. The first edition of [27] 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 6.1. 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 6.1. 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 6.2. 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 6.3. 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!

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. We will generalize polar forms to polynomials, and we will see that they play a very important role.

One of the very important properties of an inner product φ is that the map $u \mapsto \sqrt{\Phi(u)}$ is a norm.

Proposition 6.1. *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 the above quadratic equation has a double root λ_0, and we have $\Phi(u + \lambda_0 v) = 0$. If $\lambda_0 = 0$, then $\varphi(u,\, v) = 0$, and since $\Phi(v) > 0$, we must have $\Phi(u) = 0$, and thus $u = 0$. In this case, of course, $u = 0$ and v are linearly dependent. Finally, if $\lambda_0 \neq 0$, since $\Phi(u + \lambda_0 v) = 0$

implies that $u+\lambda_0 v = 0$, then 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.

At this stage, it is useful to define the notion of *norm* on a vector space. We let $\mathbb{R}_+$ denote the set of nonnegative real numbers,

$$\mathbb{R}_+ = \{\lambda \in \mathbb{R} \mid \lambda \geq 0\}.$$

Definition 6.2. 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) $\|\lambda x\| = |\lambda| \, \|x\|$. (scaling)

(N3) $\|x + y\| \leq \|x\| + \|y\|$. (triangle inequality)

A vector space E together with a norm $\| \|$ is called a *normed vector space*.

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 6.2, and since φ is bilinear and positive definite, it also satisfies conditions (N1) and (N2) of Definition 6.2, 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

$$\|u + v\| \leq \|u\| + \|v\|.$$

We now define orthogonality.

6.2 Orthogonality, Gram–Schmidt Procedure, Adjoint Maps

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. We prove that in a finite-dimensional Euclidean space, every basis can be orthormalized using the Gram–Schmidt orthonormalization procedure. Then, we show that every linear map has an adjoint.

Definition 6.3. 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 6.4. Going back to Example 6.3 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 $\sqrt{2\pi}$.

We leave the following simple two results as exercises.

Proposition 6.2. *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.*

Proposition 6.3. *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 $(e_i)_{i \in I}$.

A very important property of Euclidean spaces of finite dimension is that they possess orthonormal bases.

The existence of orthonormal bases can be shown 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 6.4. *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 6.4.

(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 6.4 also works for a countably infinite basis for E, producing a countably infinite orthonormal basis.

Example 6.5. 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:

$$\begin{aligned} P_0(x) &= 1, \\ P_1(x) &= x, \\ P_{n+1}(x) &= \frac{2n+1}{n+1}\, xP_n(x) - \frac{n}{n+1}\, P_{n-1}(x). \end{aligned}$$

The polynomials Q_n are related to the Legendre polynomials P_n as follows:

$$Q_n(x) = \sqrt{\frac{2n+1}{2}} P_n(x).$$

As a consequence of Proposition 6.4, 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

$$\|u\| = \|u_1 e_1 + \cdots + u_n e_n\| = \left(\sum_{i=1}^{n} u_i^2 \right)^{1/2}.$$

In matrix notation, if u and v are two column vectors of the same dimension n, we can write

$$u \cdot v = u^\top v = v^\top u.$$

We can also prove the following proposition regarding orthogonal spaces.

Proposition 6.5. *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 $E = F \oplus F^\perp$, which means that $F \cap F^\perp = (0)$, and that every $u \in E$ can be written as $u = v + w$, for some unique $v \in F$ and $w \in F^\perp$. Furthermore, we have $F^{\perp\perp} = F$.*

Proof. From Proposition 6.4, 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 6.4, 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 $w = w_1 u_1 + \cdots + w_n u_n \in E$ 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 $w = w_1 u_1 + \cdots + w_n u_n \in E$ 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 $F^{\perp\perp} = F$. □

Using orthonormal bases, it is easy to show that every linear map has an adjoint with respect to the inner product.

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 [52]).

Proposition 6.6. *Given a Euclidean space E of finite dimension, for every orthonormal basis $(e_1, \ldots, e_n)$ of E, for every linear map $f \colon E \to E$, if the matrix of f is A, then the linear map f^* whose matrix is the transpose $A^\top$ of A is the unique linear map such that*

$$f^*(u) \cdot v = u \cdot f(v), \quad \textit{for all } u, v \in E.$$

Proof. Assume that f^* exists, let $A = (a_{ij})$ be the matrix of f, and let $B = (b_{ij})$ be the matrix of f^*, with respect to the orthonormal basis $(e_1, \ldots, e_n)$. Since f^* must satisfy the condition

$$f^*(u) \cdot v = u \cdot f(v) \quad \text{for all } u, v \in E,$$

using the fact that if $w = w_1e_1 + \cdots + w_ne_n$ then $w_k = w \cdot e_k$ for all k, $1 \leq k \leq n$, if we let $u = e_i$ and $v = e_j$, since

$$f(e_j) = \sum_{i=1}^{n} a_{ij}e_i$$

and

$$f^*(e_i) = \sum_{j=1}^{n} b_{ji}e_j,$$

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, $B = A^\top$, which shows that if f^* exists, then it is unique. However, by the above computation, the linear map f^* whose matrix is $A^\top$ works, which establishes our proposition. □

The map f^* is called the *adjoint of f (w.r.t. to the inner product)*. 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).

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*.

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 6.6 can be adapted to show that there is a unique linear map $f^*\colon F \to E$ such that

$$\langle f(u), v\rangle_2 = \langle u, f^*(v)\rangle_1 \quad \text{for all } u \in E \text{ and all } v \in F.$$

The linear map f^* is also called the *adjoint of f*.

6.3 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 6.4. Given any two nontrivial Euclidean spaces E and F of the same finite dimension n, a function $f: E \to F$ 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 6.4 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 $f: E \to F$ (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 6.7. *Given any two nontrivial Euclidean 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) $\|f(v) - f(u)\| = \|v - u\|$, 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 $u = u_1e_1 + \cdots + u_ne_n$, 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. □

Remark: 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.

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

$$f(\tau + u) = f(\tau) + g(u)$$

for all $u \in E$. By Proposition 4.7, this shows that f is affine with associated linear map g. This fact is worth recording as the following proposition.

Proposition 6.8. *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 6.7, we will drop the word "linear" in "linear 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.

6.4 The Orthogonal Group, Orthogonal Matrices

In this section we explore some of the basic properties of the orthogonal group and of orthogonal matrices.

Proposition 6.9. *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 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 E, iff the rows of A form an orthonormal basis of E.

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) By Proposition 6.6, the condition

$$f \circ f^* = f^* \circ f = \mathrm{id}$$

is equivalent to

$$A\,A^\top = A^\top A = I_n.$$

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 6.9 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 6.9 also motivates the following definition.

Definition 6.5. 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 6.9 (2), the matrix P is orthogonal.

The proof of Proposition 6.7 (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*gon*al matrices are not called ortho*norm*al 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 6.6. 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.

6.5 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 6.10. *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 6.4 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, see Strang [52, 53], Golub and Van Loan [26], Trefethen and Bau [56], Demmel [14], Kincaid and Cheney [34], or Ciarlet [11].

Example 6.6. 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 6.7. 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 [52, 53], Golub and Van Loan [26], Demmel [14], Trefethen and Bau [56], or Serre [48].

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.

6.6 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 [43]).

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 9). 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 6.5, or better, using Householder matrices. 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 11). 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 [53, 52], Golub and Van Loan [26], Demmel [14], Serre [48], and Trefethen and Bau [56].

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 12.

6.7 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.
- The *Cauchy–Schwarz inequality*; the *Minkowski inequality*.
- *Orthogonality*, *orthogonal complement* $F^\perp$; *orthonormal family*.
- Existence of an orthonormal basis in a finite-dimensional Euclidean space (Proposition 6.4).

- The *Gram–Schmidt orthonormalization procedure* (Proposition 6.4).
- The *adjoint* of a linear map (with respect to an inner product).
- *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.

Chapter 7

Hermitian Spaces

7.1 Sesquilinear and Hermitian Forms, Pre-Hilbert Spaces and Hermitian Spaces

In this chapter we generalize the basic results of Euclidean geometry presented in Chapter 6 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.

We begin with a quick review of complex numbers. One of the main motivations for introducing the complex numbers is to ensure that every polynomial has a zero (or as we say, a *root*). Because the square λ^2 of a real number $\lambda \in \mathbb{R}$ is nonnegative, the equation

$$x^2 + 1 = 0,$$

which is equivalent to $x^2 = -1$, does not have any real root. Many other polynomials (with real coefficients) do not have any (real) roots.

In the eighteenth century, various mathematicians such as Euler and Gauss introduced and used complex numbers. The idea is to intoduce a new entity, the *pure imaginary number* i, which has the fundamental property that

$$i^2 = -1.$$

Now, as soon as we let i out of the bag, we realize that it is inevitable to add or multiply real numbers with i. Thus, we are immediately led to consider new numbers of the form

$$a + ib, \quad \text{with} \quad a, b \in \mathbb{R}.$$

The set of these "numbers" is the set of *complex numbers*, and it is denoted by $\mathbb{C}$. It is customary to denote complex numbers by the letter z.

If

$$z = a + ib$$

is a complex number, then a is the *real part* of z, denoted by $\mathrm{Re}(z)$ (or by $\Re z$), and b is the *imaginary part* of z, denoted by $\mathrm{Im}(z)$ (or by $\Im z$).

Complex numbers are added and multiplied as follows:

$$\begin{aligned}(a_1 + ib_1) + (a_2 + ib_2) &= (a_1 + a_2) + i(b_1 + b_2)\\ (a_1 + ib_1) \cdot (a_2 + ib_2) &= (a_1a_2 - b_1b_2) + i(a_1b_2 + a_2b_1).\end{aligned}$$

If $z = a + ib$, we define $-z$ as $-a - ib$.

One can check that addition of complex numbers is associative, commutative, and has 0 has an identity element. Every complex number z has an additive inverse $-z$. Therefore, under addition, $\mathbb{C}$ is a commutative group. Multiplication of complex numbers is also associative, commutative, and has 1 has an identity element. Furthermore, multiplication distributes with respect to addition.

The *conjugate* $\overline{z}$ of a complex number $z = a + ib$ is the complex number given by

$$\overline{z} = a - ib.$$

Observe that

$$\begin{aligned}\mathrm{Re}(z) &= \frac{z + \overline{z}}{2}\\ \mathrm{Im}(z) &= \frac{z - \overline{z}}{2i}.\end{aligned}$$

Also observe that

$$z\overline{z} = \overline{z}z = (a + ib)(a - ib) = a^2 - (ib)^2 = a^2 - i^2b^2 = a^2 + b^2.$$

We define $|z|$, the *modulus* (or *absolute value*) of $z = a + ib$ by

$$|z| = \sqrt{z\overline{z}} = \sqrt{a^2 + b^2}.$$

Complex numbers z with $|z| = 1$ are called *unit complex numbers*. The set of unit complex numbers is often denoted by $\mathbf{U}(1)$.

Clearly, $z = 0$ iff $|z| = 0$. Then, because

$$z\overline{z} = \overline{z}z = |z|^2,$$

if $z \neq 0$, we see that $\overline{z}/|z|^2$ is the *multiplicative inverse* of z. Therefore, $\mathbb{C} - \{0\} = \mathbb{C}^*$ is a commutative group under multiplication.

Complex numbers $z = a + ib \in \mathbb{C}$ have a geometric interpretation as points $(a, b) \in \mathbb{R}^2$ in the (real) plane. In fact, the mapping

$$z = a + ib \mapsto (a, b)$$

is a bijection from $\mathbb{C}$ to $\mathbb{R}^2$. For this reason, the set $\mathbb{C}$ of complex numbers is often called the *complex plane*, although this terminology is confusing.

The bijection between $\mathbb{C}$ and $\mathbb{R}^2$ is best revealed by the *polar form* of a complex number. Given any nonzero complex number $z = a + ib$, we can write

$$z = a + ib = \sqrt{a^2 + b^2}\left(\frac{a}{\sqrt{a^2 + b^2}} + i\frac{b}{\sqrt{a^2 + b^2}}\right).$$

But then, there is a unique angle $\theta \in [0, 2\pi)$ such that

$$\cos\theta = \frac{a}{\sqrt{a^2 + b^2}}$$
$$\sin\theta = \frac{b}{\sqrt{a^2 + b^2}},$$

so we can write

$$z = a + ib = \sqrt{a^2 + b^2}\,(\cos\theta + i\sin\theta) = |z|(\cos\theta + i\sin\theta).$$

We often write $r = |z|$, in which case

$$z = a + ib = r(\cos\theta + i\sin\theta),$$

where the expression $r(\cos\theta + i\sin\theta)$ is called the *polar form* of z. The angle θ is often called the *argument* of z. Observe that unit complex numbers correspond to points on the unit circle in $\mathbb{R}^2$.

It is also convenient to express $\cos\theta + i\sin\theta$ as the complex exponential

$$e^{i\theta} = \cos\theta + i\sin\theta.$$

The above equation known as *Euler's Formula* can be justified using power series.

One of the virtues of the polar reprentation of complex numbers is that multiplication has a geometric interpretation. Indeed, using some well-known trigonometric identities, we get

$$\begin{aligned}(\cos\alpha + i\sin\alpha)(\cos\beta + i\sin\beta) &= (\cos\alpha\cos\beta - \sin\alpha\sin\beta) + i(\cos\alpha\sin\beta + \sin\alpha\cos\beta)\\ &= \cos(\alpha + \beta) + i\sin(\alpha + \beta).\end{aligned}$$

Therefore,

$$r(\cos\alpha + i\sin\alpha)\cdot r'(\cos\beta + i\sin\beta) = rr'(\cos(\alpha + \beta) + i\sin(\alpha + \beta));$$

this means that we multiply the moduli and we add the angles.

In particular, if $r = r'$ and $\alpha = \beta = \theta$, we get

$$(r(\cos\theta + i\sin\theta))^2 = r^2(\cos 2\theta + i\sin 2\theta),$$

and by induction,

$$(r(\cos\theta + i\sin\theta))^n = r^n(\cos n\theta + i\sin n\theta),$$

for all $n \in \mathbb{N}$. In terms of the exponential notation, this is

$$(re^{i\theta})^n = r^n e^{in\theta}.$$

As a consequence, if we let $\omega = e^{i2\pi/n}$, observe that

$$(\omega^k)^n = 1, \quad k = 0, \ldots, n-1.$$

Therefore, $1, \omega, \omega^2, \ldots, \omega^{n-1}$ are the n distinct roots of the equation

$$z^n = 1.$$

For this reason, $1, \omega, \omega^2, \ldots, \omega^{n-1}$ are called *nth roots of* 1 *(or unity)*. The number z^k corresponds to the number on the unit circle whose angle is $k(2\pi/n)$. The number ω plays a crucial role in the *Discrete Fourier Transform*.

One should know that there is a neat way of realizing the complex numbers as 2×2 real matrices. Indeed, the function

$$a + ib \mapsto \begin{pmatrix} a & -b \\ b & a \end{pmatrix}$$

is obviously a bijection between the set of complex numbers $\mathbb{C}$ and the set of all real 2×2 matrices of the above form. Furthermore, it is easy to check that addition of complex numbers corresponds to addition of the corresponding matrices, and similarly multiplication of complex numbers corresponds to multiplication of the corresponding matrices. Also

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^2 = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} = -I_2,$$

which shows that the matrix

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

is a "real analog" of the imaginary complex number i. In the above correspondence, a unit complex number $z = \cos\theta + i\sin\theta$ corresponds to the rotation matrix

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix},$$

which represents the rotation around the origin by the angle θ.

In conclusion of this quick review of the complex numbers, let us state the reason why they are so important: Every polynomial

$$P(z) = a_0 z^n + a_1 z^{n-1} + \cdots + a_n$$

of degree $n \geq 1$ with real or complex coefficients $a_0, a_1, \ldots, a_n$ (with $a_0 \neq 0$) *always has n roots in* $\mathbb{C}$. This means that $P(z)$ can always be written as

$$P(z) = a_0(z - z_1)(z - z_2)\cdots(z - z_n),$$

for some sequence $(z_1, \ldots, z_n)$ of n complex numbers not necessarily distinct. This property is usually stated as: *the field* $\mathbb{C}$ *of complex numbers is algebraically closed*. This is one of the many results due to Gauss, who gave several proofs of this fundamental result. Every proof involves a little bit of analysis.

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 7.1. Given two vector spaces E and F over the complex field $\mathbb{C}$, a function $f\colon E \to F$ is *semilinear* if

$$f(u + v) = f(u) + f(v),$$
$$f(\lambda u) = \overline{\lambda} f(u),$$

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 7.2. 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 7.1. *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 7.1 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

$$\varphi((x_1,y_1),(x_2,y_2)) = x_1y_2 - x_2y_1$$

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 6.

As in the Euclidean case, Hermitian forms for which $\varphi(u,u) \geq 0$ play an important role.

Definition 7.3. 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 [36], define a pre-Hilbert space as what we define as a Hermitian space. We prefer following the terminology used in Schwartz [46] and Bourbaki [8]. 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 7.1. 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 7.2. 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 7.3. 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 7.4. 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.

The Cauchy–Schwarz inequality and the Minkowski inequalities extend to pre-Hilbert spaces and to Hermitian spaces.

Proposition 7.2. *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 the equation

$$\Phi(u) + 2t|\varphi(u,v)| + t^2\Phi(v) = 0$$

has a double root t_0, and thus

$$\Phi(u + t_0 e^{i\theta} v) = 0.$$

Since φ is positive definite, we must have

$$u + t_0 e^{i\theta} v = 0,$$

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 5 (or Lang [36, 37], Dixmier [16], or Schwartz [46, 47]). 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 $v \in E$.

We will now basically mirror the presentation of Euclidean geometry given in Chapter 6 rather quickly, leaving out most proofs, except when they need to be seriously amended.

7.2 Orthogonality, Gram–Schmidt Procedure, Adjoint Maps

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 6.3).

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 6.2 and 6.3 hold without any changes. It is easy to show that

$$\left\| \sum_{i=1}^{n} u_i \right\|^2 = \sum_{i=1}^{n} \|u_i\|^2 + \sum_{1 \leq i < j \leq n} 2\Re(u_i \cdot u_j).$$

The *Gram–Schmidt orthonormalization procedure* also applies to Hermitian spaces of finite dimension, without any changes from the Euclidean case!

Proposition 7.3. *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 6.4 also apply here, except that in the QR-decomposition, Q is a unitary matrix.

As a consequence of Proposition 7.3, 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

$$\|u\| = \|u_1 e_1 + \cdots + u_n e_n\| = \left(\sum_{i=1}^{n} |u_i|^2 \right)^{1/2}.$$

Proposition 6.5 also holds unchanged.

Proposition 7.4. *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 $F^{\perp\perp} = F$.*

As in the case of real Euclidean spaces, using orthonormal bases, it is easy to show that every linear map has an adjoint with respect to the Hermitian inner product. However, 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 7.4. Given a complex $m \times n$ matrix A, the *transpose* $A^\top$ *of* A is the $n \times m$ matrix $A^\top = \left(a^\top_{ij}\right)$ defined such that

$$a^\top_{ij} = a_{ji},$$

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 7.5. *Given a Hermitian space E of finite dimension, for every orthonormal basis $(e_1, \ldots, e_n)$ of E, for every linear map $f\colon E \to E$, if the matrix of f is A, then the linear map f^* whose matrix is the adjoint A^* of A is the unique linear map such that*

$$f^*(u) \cdot v = u \cdot f(v), \quad \textit{for all } u, v \in E.$$

Proof. Assume that f^* exists, let $A = (a_{ij})$ be the matrix of f, and let $B = (b_{ij})$ be the matrix of f^*, with respect to the orthonormal basis $(e_1, \ldots, e_n)$. Since f^* satisfies the condition

$$f^*(u) \cdot v = u \cdot f(v) \quad \text{for all } u, v \in E,$$

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, $1 \leq k \leq n$; if we let $u = e_i$ and $v = e_j$, we get

$$b_{ji} = f^*(e_i) \cdot e_j = e_i \cdot f(e_j) = \overline{f(e_j) \cdot e_i} = \overline{a_{ij}},$$

for all i, j, $1 \leq i, j \leq n$. Thus, $B = A^*$. However, by the above computation, the linear map f^* whose matrix is A^* works, which establishes our proposition. □

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 7.5 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.

7.3 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 6.3 extend to Hermitian spaces, except that orthogonal transformations are called unitary transformation, but Proposition 6.7 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 6.4.

Definition 7.5. 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 6.7 can be salvaged by strengthening condition (2).

Proposition 7.6. *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 6.7 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.

7.4 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.

Proposition 7.7. *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 $f\colon E \to E$ is an isometry iff

$$f \circ f^* = f^* \circ f = \mathrm{id}.$$

(2) For every orthonormal basis $(e_1, \ldots, e_n)$ of E, 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 E, iff the rows of A form an orthonormal basis of E.

Proof. (1) The proof is identical to that of Proposition 6.9 (1).

(2) By Proposition 7.5, the condition

$$f \circ f^* = f^* \circ f = \mathrm{id}$$

is equivalent to the condition

$$A\,A^* = A^*A = I_n.$$

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 6.9 shows that the inverse of an isometry f is its adjoint f^*. Proposition 6.9 also motivates the following definition.

Definition 7.6. 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 7.6 (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 unitary 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 7.7. 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 [52, 54].

Now that we have the definition of a unitary matrix, we can explain how the Gram–Schmidt orthonormalization procedure immediately yields the QR-decompo-sition for matrices.

Proposition 7.8. *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!

7.5 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 *positibe definite* Hermitian forms; *pre-Hilbert spaces*, *Hermitian spaces*.
- The *Cauchy–Schwarz inequality* and the *Minkowski inequality*.
- *Hermitian inner product*, *Hermitian norm*.
- Existence of orthonormal bases in a Hermitian space (Proposition 7.3).
- *Gram–Schmidt orthonormalization procedure*.
- The *adjoint* of a linear map (with respect to a Hermitian inner product).

- *Linear isometries* (*unitary transformatios*).
- The *unitary group*, *unitary matrices*.
- The *unitary group* $\mathbf{U}(n)$; The *special unitary group* $\mathbf{SU}(n)$.
- QR-Decomposition for invertible matrices.

Chapter 8

Eigenvectors and Eigenvalues

8.1 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*. As we will see in Section 8.2, 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 I denote the identity map on E.

Definition 8.1. Given any vector space E 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 $\mathrm{Ker}\,(\lambda I - f)$ is nontrivial (i.e., $\mathrm{Ker}\,(\lambda I - f) \neq \{0\}$). 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 $\mathrm{Ker}\,(\lambda I - f)$ 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: We emphasize that 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.

Let us now assume that E is of finite dimension n. The next proposition shows that the eigenvalues of a linear map $f: E \to E$ are the roots of a polynomial associated with f.

Proposition 8.1. *Let E be any vector space of finite dimension n and let f be any linear map $f: E \to E$. The eigenvalues of f are the roots (in K) of the polynomial*

$$\det(\lambda I - f).$$

Proof. A scalar $\lambda \in K$ is an eigenvalue of f iff there is some nonzero vector $u \neq 0$ in E such that

$$f(u) = \lambda u$$

iff

$$(\lambda I - f)(u) = 0$$

iff $(\lambda I - f)$ is not invertible iff, by Proposition 5.13,

$$\det(\lambda I - f) = 0. \qquad \square$$

In view of the importance of the polynomial $\det(\lambda I - f)$, we have the following definition.

Definition 8.2. Given any vector space E of dimension n, for any linear map $f: E \to E$, the polynomial $P_f(X) = \chi_f(X) = \det(XI - f)$ 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 5.6; see Definition 5.8.

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 $\chi_f(X) = \det(XI - f)$ of f by expanding the following determinant:

$$\det(XI - A) = \begin{vmatrix} X - a_{11} & -a_{12} & \ldots & -a_{1n} \\ -a_{21} & X - a_{22} & \ldots & -a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -a_{n1} & -a_{n2} & \ldots & X - a_{nn} \end{vmatrix}.$$

If we expand this determinant, we find that

$$\chi_A(X) = \det(XI - A) = X^n - (a_{11} + \cdots + a_{nn})X^{n-1} + \cdots + (-1)^n \det(A).$$

The sum $\mathrm{tr}(A) = a_{11} + \cdots + a_{nn}$ of the diagonal elements of A is called the *trace of A*. Since we proved in Section 5.6 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 $\det(f - XI)$. Since

$$\det(f - XI) = (-1)^n \det(XI - f),$$

this makes essentially no difference but the version $\det(XI - f)$ 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 $h = 1, \ldots, k$. 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 symmetric function* of the λ_i's. From this, it clear that

$$\sigma_k(\lambda) = \tau_k(A)$$

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

$$c_0 X^m + c_1 X^{m-1} + \cdots + c_m = c(X - \lambda_1) \cdots (X - \lambda_m)$$

for some complex numbers $\lambda_1, \ldots, \lambda_m \in \mathbb{C}$, not necessarily distinct, and some $c \in \mathbb{C}$ with $c \neq 0$. But then, since $c \neq 0$,

$$c_0 f^m(u) + c_1 f^{m-1}(u) + \cdots + c_m u = 0$$

is equivalent to

$$(f - \lambda_1 I) \circ \cdots \circ (f - \lambda_m I)(u) = 0.$$

If all the linear maps $f - \lambda_i I$ were injective, then $(f - \lambda_1 I) \circ \cdots \circ (f - \lambda_m I)$ would be injective, contradicting the fact that $u \neq 0$. Therefore, some linear map $f - \lambda_i I$ 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 $\det(XI - f)$.

Definition 8.3. 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 8.2. *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 obtained by picking a basis of E_{λ_i} and completing this basis. 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 8.3. *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

$$\mu_1 u_{i_1} + \cdots + \mu_k v_{i_k} = 0,$$

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 linear dependency among a proper subfamily of $(u_{i_1}, \ldots, u_{i_k})$, a contradiction. □

Thus, from Proposition 8.3, 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}. This means that $E_{\lambda_1} \oplus \cdots \oplus E_{\lambda_m}$ is the space of all vectors $u \in E$ that can be written as

$$u = u_1 + \cdots + u_m, \quad u_i \in E_{\lambda_i}, \quad i = 1, \ldots, m,$$

in a unique way. 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.

8.2 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,

$$\begin{pmatrix} a_{1\,1} & a_{1\,2} & a_{1\,3} & \dots & a_{1\,n-1} & a_{1\,n} \\ 0 & a_{2\,2} & a_{2\,3} & \dots & a_{2\,n-1} & a_{2\,n} \\ 0 & 0 & a_{3\,3} & \dots & a_{3\,n-1} & a_{3\,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & a_{n-1\,n-1} & a_{n-1\,n} \\ 0 & 0 & 0 & \dots & 0 & a_{n\,n} \end{pmatrix},$$

i.e., $a_{i\,j} = 0$ whenever $j < i$, $1 \leq i, j \leq n$.

Theorem 8.4. *Given any finite dimensional vector space over a field K, for any linear map $f\colon E \to E$, there is a basis $(u_1, \dots, u_n)$ 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 $(u_1, \dots, u_n)$ 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 $(v_2, \dots, v_n)$ such that $(u_1, v_2, \dots, v_n)$ is a basis of E, and let F be the subspace of dimension $n - 1$ spanned by $(v_2, \dots, v_n)$. In the basis $(u_1, v_2 \dots, v_n)$, the matrix of f is of the form

$$U = \begin{pmatrix} \lambda_1 & a_{1\,2} & \dots & a_{1\,n} \\ 0 & a_{2\,2} & \dots & a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n\,2} & \dots & a_{n\,n} \end{pmatrix},$$

since its first column contains the coordinates of $\lambda_1 u_1$ over the basis $(u_1, v_2, \dots, v_n)$. 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 $(v_2, \dots, v_n)$. 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 8.4 implies that every complex $n \times n$ matrix can be triangularized.

If E is a Hermitian space, the proof of Theorem 8.4 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 8.5. *(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. If E is a real Euclidean spaace and if the eigenvalues of f are all real, the proof also goes through with real matrices. □

Using, Theorem 8.5, we can derive the fact that if A is a Hermitian matrix, then there is a unitary matrix U and a real diagonal matrix D such that $A = UDU^*$. Indeed, since $A^* = A$, 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 real diagonal matrix. In fact, applying 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 8.5 again, we conclude that $A = QDQ^\top$, where Q is orthogonal and D is a real diagonal matrix. We will also prove this in Chapter 9.

When A has complex eigenvalues, there is a version of Theorem 8.5 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 8.5 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!

8.3 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 8.4. 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 8.6. *(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

$$|a_{i\,i}| > \sum_{j=1,\, j\neq i}^{n} |a_{i\,j}|, \quad \textit{for } i = 1, \ldots, n,$$

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 8.6 implies that if a symmetric matrix is strictly row diagonally dominant and has strictly positive diagonal entries, then it is positive definite. Theorem 8.6 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 8.6 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 8.7. *For any complex $n \times n$ matrix A, all the eigenvalues of A belong to the intersection of the Gershgorin discs, $G(A) \cap G(A^\top)$. 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 $a_{i\,i} > 0$ 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 [30], 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.

8.4 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 8.3).
- 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 8.6 and 8.7.

Chapter 9

Spectral Theorems in Euclidean and Hermitian Spaces

9.1 Introduction

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.

9.2 The Spectral Theorem for Self-Adjoint Maps; The Hermitian Case

Recall that if E is a finite-dimensional complex vector space with a Hermitian inner product $\langle -, - \rangle$, a linear map $f : E \to E$ is *self-adjoint* if $f = f^*$.

The first important fact about a self-adjoint linear map is that its eigenvalues are real.

Proposition 9.1. *Given a Hermitian space E, all the eigenvalues of any self-adjoint linear map $f : E \to E$ 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. □

The second important fact about a self-adjoint linear map is that eigenvectors associated with distinct eigenvalues are orthogonal.

Proposition 9.2. *Given a Hermitian space E, for any self-adjoint linear map $f\colon E \to E$, if u and v are eigenvectors of f associated with the eigenvalues λ and μ (in $\mathbb{R}$) where $\lambda \neq \mu$, then $\langle u, v\rangle = 0$.*

Proof. Let us compute $\langle f(u), v\rangle$ in two different ways. Proposition 9.1 tells us that λ and μ are real. We have

$$\langle f(u), v\rangle = \langle \lambda u, v\rangle = \lambda\langle u, v\rangle$$

and

$$\langle f(u), v\rangle = \langle u, f^*(v)\rangle = \langle u, \mu v\rangle = \mu\langle u, v\rangle,$$

where the last identity holds because $\overline{\mu} = \mu$ since μ is real 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 $\lambda \neq \mu$. □

Given any subspace W of a Hermitian 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 7.4 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).

Theorem 9.3. *(Spectral theorem for self-adjoint linear maps on a Hermitian space) Given a Hermitian 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. From Proposition 9.1, all the eigenvalues of f are real. If $n = 1$, the result is trivial. Assume now that $n \geq 2$. 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 $\langle v, w\rangle = 0$ (since λ is real, $\overline{\lambda} = \lambda$). 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 $n - 1$). □

9.3 The Spectral Theorem for Self-Adjoint Maps; The Euclidean Case

Proposition 9.1 also holds in the Euclidean case.

Proposition 9.4. *Given a Euclidean space E, if $f\colon E \to E$ is any self-adjoint linear map, then every eigenvalue of f is real.*

Proof. The problem is that we can't apply directly Proposition 9.1 because in that theorem, the eigenvector u could be complex. Instead, we can proceed as follows. Pick some orthonormal basis $(u_1, \ldots, u_n)$ of E, and let A be the matrix representing f on this basis. Since f is self-adjoint, A is symmetric. Then, consider the linear map $f_{\mathbb{C}}\colon \mathbb{C}^n \to \mathbb{C}^n$ defined by the matrix A (viewed as a complex matrix). Since A is real and symmetric, $A = A^\top = A^*$, so $f_{\mathbb{C}}$ is self-adjoint with respect to the standard Hermitian inner product on $\mathbb{C}^n$. By Proposition 9.1 applied to $f_{\mathbb{C}}$ (and the Hermitian space $\mathbb{C}^n$), all the eigenvalues of $f_{\mathbb{C}}$ are real. Now, the characteristic polynomials of both f and $f_{\mathbb{C}}$ are equal to $\det(zI - A)$, a polynomial with real coefficients, and we just proved that all its roots are real. Therefore, the eigenvalues of f are all real. □

Proposition 9.2 also holds in the Euclidean case and the proof is exactly the same.

Proposition 9.5. *Given a Euclidean space E, for any self-adjoint linear map $f\colon E \to E$, if u and v are eigenvectors of f associated with the eigenvalues λ and μ (in $\mathbb{R}$) where $\lambda \neq \mu$, then $\langle u, v\rangle = 0$.*

Given any subspace W of a Euclidean space E, 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 6.5 that $E = W \oplus W^{\perp}$ Then, we have a version of Theorem 9.3 for Euclidean spaces. The proof is the same, except that it uses Proposition 9.4.

Theorem 9.6. *(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 $\lambda_i \in \mathbb{R}$.

The theorems of this section and of the previous section can be immediately applied to matrices.

9.4 Normal and Other Special Matrices

First, we consider real matrices. Recall the following definitions.

Definition 9.1. Given a real $m \times n$ matrix A, the *transpose* $A^{\top}$ *of* A is the $n \times m$ matrix $A^{\top} = (a_{ij}^{\top})$ defined such that

$$a_{ij}^{\top} = a_{ji}$$

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 6.9 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.

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, Theorem 9.6 can be restated as follows.

Theorem 9.7. *For every symmetric matrix A there is an orthogonal matrix P and a diagonal matrix D such that $A = P D\, P^\top$, where D is of the form*

$$D = \begin{pmatrix} \lambda_1 & & \cdots & \\ & \lambda_2 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \cdots & \lambda_n \end{pmatrix},$$

where $\lambda_i \in \mathbb{R}$.

We now consider complex matrices.

Definition 9.2. 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 7.7 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.

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

$$B = P^*AP.$$

Theorem 9.3 can be restated in terms of matrices as follows.

Theorem 9.8. *For every complex Hermitian matrix A there is a unitary matrix U and a diagonal matrix D with real entries such that $A = UDU^*$.*

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.

9.5 Summary

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

- *self-adjoint* linear maps
- 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*.
- Eigenvectors of a self-adjoint map associated to distinct eigenvalues are orthogonal.
- Every self-adjoint linear map on a Hermitian space has an orthonormal basis of eigenvectors.
- Every self-adjoint linear map on a Euclidean space has an orthonormal basis of eigenvectors.
- The spectral theorem for symmetric matrices.
- The spectral theorem for Hermitian matrices.

Chapter 10

Variational Approximation of Boundary-Value Problems; Introduction to the Finite Elements Method

10.1 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 10.1.

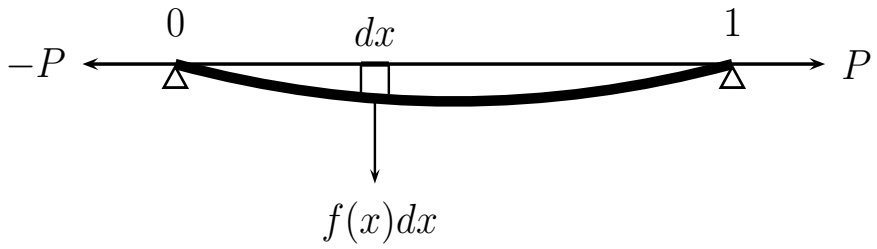

Figure 10.1: Vertical deflection of a beam

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

$$-u''(x)v(x) + c(x)u(x)v(x) = f(x)v(x), \qquad (*)$$

and integrate this equation! We get

$$-\int_0^1 u''(x)v(x)dx + \int_0^1 c(x)u(x)v(x)dx = \int_0^1 f(x)v(x)dx. \qquad (\dagger)$$

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 (†) 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 10.1. *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 10.1 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 10.1 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 10.2. *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 [11].

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 10.2.

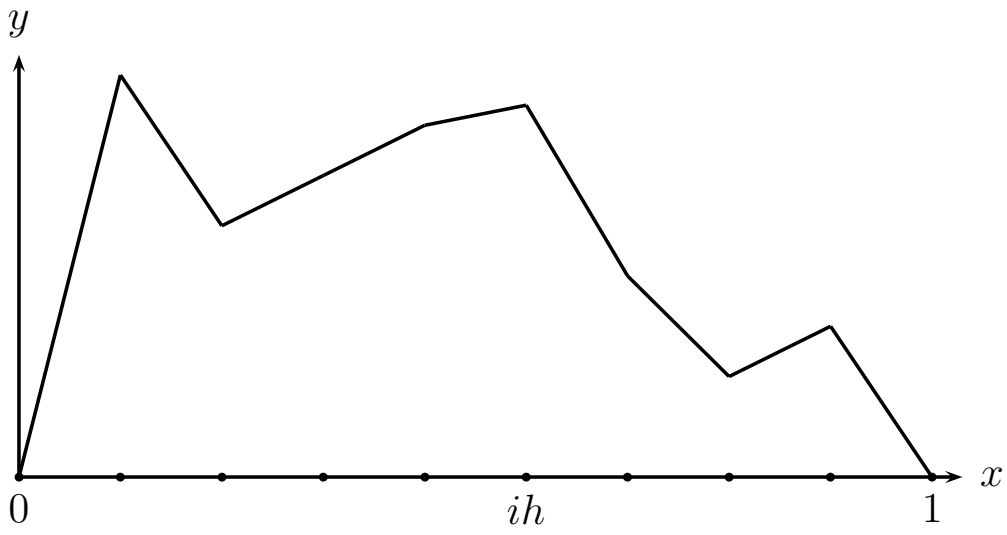

Figure 10.2: A piecewise affine function

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 10.3.

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):

$$v(x) = \sum_{i=1}^{N} v(ih) w_i(x), \quad x \in [0,1].$$

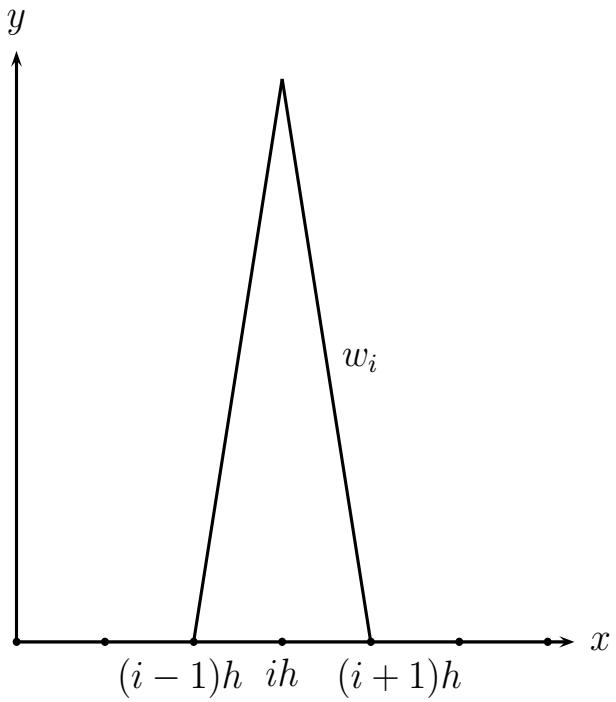

Figure 10.3: A basis "hat function"

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 10.4. 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.

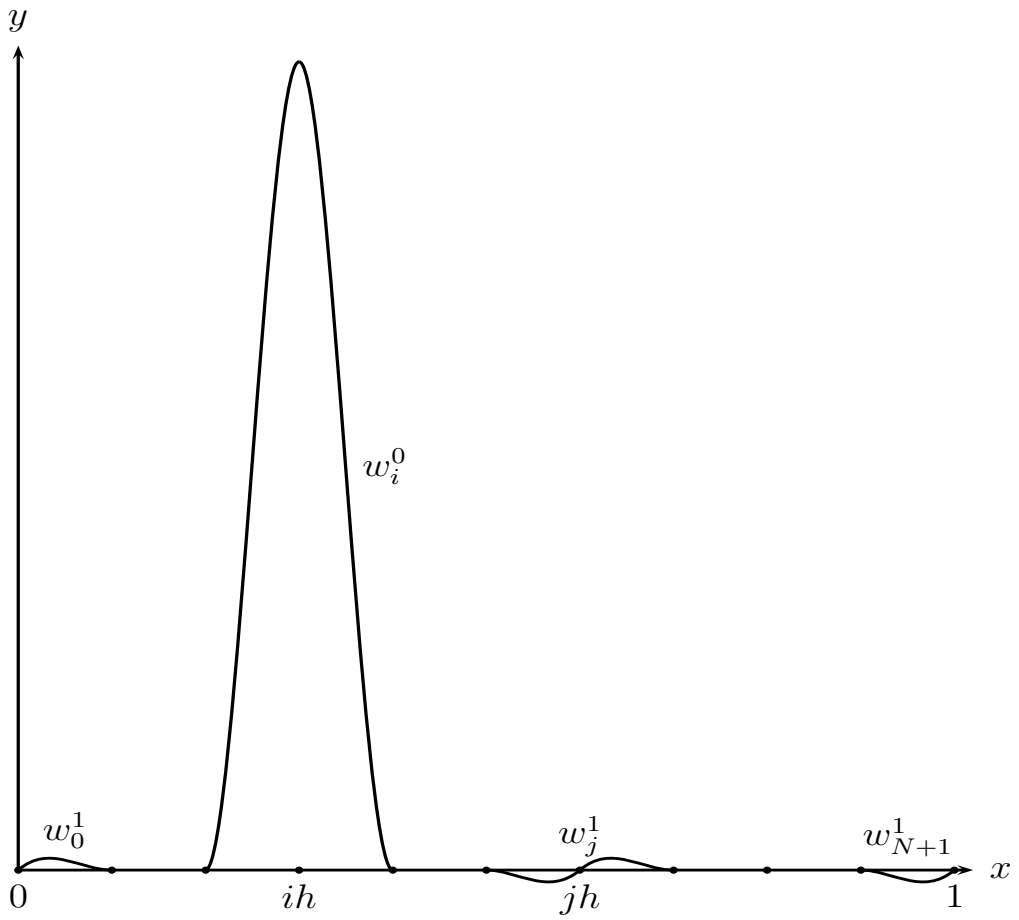

Figure 10.4: The basis functions w_i^0 and w_j^1

10.2 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 10.5. 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

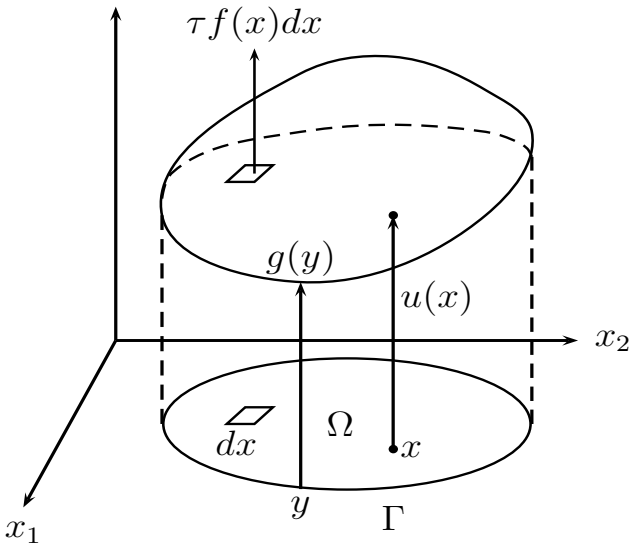

Figure 10.5: An elastic membrane

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 10.2, 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.

10.3 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 $x = L$.

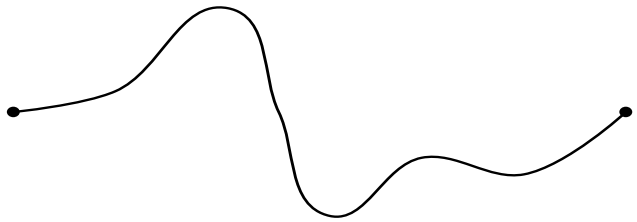

Figure 10.6: A vibrating string

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 2.10). 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 2.3 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 10.2. 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 11

Singular Value Decomposition and Polar Form

11.1 The Four Fundamental Subspaces

In this section we assume that we are dealing with a real Euclidean space E. 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 situation is even better, since we will show shortly that $f^* \circ f$ and $f \circ f^*$ have the same 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

$$(-\lambda)^m \det(g \circ f - \lambda\, I_n) = (-\lambda)^n \det(f \circ g - \lambda\, I_m),$$

and thus $g \circ f$ and $f \circ g$ always have the same nonzero eigenvalues!

Definition 11.1. 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 11.2. 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.

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 $(v_1, \ldots, v_n)$ 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 3.11)

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

and that (Propositions 6.5 and 7.4) for every subspace W of E,

$$\dim(W) + \dim(W^{\perp}) = \dim(E).$$

Proposition 11.1. *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. □

Since the matrix representing f^* with respect to an orthonormal basis is the transpose of the matrix representing f (see Proposition 6.6), Proposition 11.1 gives a short proof of the fundamental fact that a matrix A and its transpose $A^\top$ have the same rank.

Proposition 11.1 reveals that the four spaces

$$\operatorname{Im} f,\ \operatorname{Im} f^*,\ \operatorname{Ker} f,\ \operatorname{Ker} f^*$$

play a special role. They are often called the *fundamental subspaces* associated with f. These spaces are related in an intimate manner, since Proposition 11.1 shows that

$$\operatorname{Ker} f = (\operatorname{Im} f^*)^\perp$$
$$\operatorname{Ker} f^* = (\operatorname{Im} f)^\perp,$$

and that

$$\operatorname{rk}(f) = \operatorname{rk}(f^*).$$

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 an orthonormal basis $(u_1, \ldots, u_n)$ of E and an orthonormal 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^* is represented by the $n \times m$ matrix $A^\top$. Consequently, the four fundamental spaces

$$\operatorname{Im} f,\ \operatorname{Im} f^*,\ \operatorname{Ker} f,\ \operatorname{Ker} f^*$$

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 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 image $\operatorname{Im} f^*$ of f^*.

(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 [53]).

The two statements

$$\operatorname{Ker} f = (\operatorname{Im} f^*)^\perp$$
$$\operatorname{Ker} f^* = (\operatorname{Im} f)^\perp$$

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 [53]).

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, that 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 $y^\top b \neq 0$.

11.2 Singular Value Decomposition for Square Matrices

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 [23].

The early history of the singular value decomposition is described in a fascinating paper by Stewart [50]. 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 11.2. *(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

$$D = \begin{pmatrix} \sigma_1 & & \cdots & \\ & \sigma_2 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \cdots & \sigma_n \end{pmatrix},$$

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 $\sigma_{r+1} = \cdots = \sigma_n = 0$. 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

$$A^\top A = U D^2 U^\top,$$

with $D = \mathrm{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0)$, where $\sigma_1^2, \ldots, \sigma_r^2$ are the nonzero eigenvalues of $A^\top A$, and where r is the rank of A; that is, $\sigma_1, \ldots, \sigma_r$ are the singular values of A. It follows that

$$U^\top A^\top A U = (AU)^\top AU = D^2,$$

and if we let f_j 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 $(v_1, \ldots, v_r, v_{r+1}, \ldots, v_n)$ (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

$$V^\top AU = D,$$

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 11.2 suggests the following definition.

Definition 11.3. 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 11.2 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 1, 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 [53] 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 [26], Demmel [14], and Trefethen and Bau [56], 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 11.4. A pair (R, S) such that $A = RS$ with R orthogonal and S symmetric positive semidefinite is called a *polar decomposition of* A.

Theorem 11.2 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 11.3. *(Weyl's inequalities, 1949) For any complex $n \times n$ matrix, A, if $\lambda_1, \dots, \lambda_n \in \mathbb{C}$ are the eigenvalues of A and $\sigma_1, \dots, \sigma_n \in \mathbb{R}_+$ 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*

$$\begin{aligned} &|\lambda_1| \cdots |\lambda_n| = \sigma_1 \cdots \sigma_n \quad \textit{and} \\ &|\lambda_1| \cdots |\lambda_k| \leq \sigma_1 \cdots \sigma_k, \quad \textit{for} \quad k = 1, \dots, n-1. \end{aligned}$$

A proof of Theorem 11.3 can be found in Horn and Johnson [31], Chapter 3, Section 3.3, where more inequalities relating the eigenvalues and the singular values of a matrix are given.

Theorem 11.2 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 [53] Golub and Van Loan [26], Demmel [14], and Trefethen and Bau [56]).

11.3 Singular Value Decomposition for Rectangular Matrices

Here is the generalization of Theorem 11.2 to rectangular matrices.

Theorem 11.4. *(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*

$$D = \begin{pmatrix} \sigma_1 & & \dots & \\ & \sigma_2 & \dots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \dots & \sigma_n \\ 0 & \vdots & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \vdots & \dots & 0 \end{pmatrix} \quad \text{or } D = \begin{pmatrix} \sigma_1 & & \dots & & 0 & \dots & 0 \\ & \sigma_2 & \dots & & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & 0 & \vdots & 0 \\ & & \dots & \sigma_m & 0 & \dots & 0 \end{pmatrix},$$

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 $\sigma_{r+1} = \dots = \sigma_p = 0$, 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 11.2, 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 $\Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_r, 0, \dots, 0)$, where $\sigma_1^2, \dots, \sigma_r^2$ are the nonzero eigenvalues of $A^\top A$, and where r is the rank of A. Observe that $r \leq \min\{m, n\}$, and AU is an $m \times n$ matrix. It follows that

$$U^\top A^\top A U = (AU)^\top AU = \Sigma^2,$$

and if we let $f_j \in \mathbb{R}^m$ 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 $(v_1, \ldots, v_r)$ 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

$$V^\top A U = D,$$

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 11.4 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 12. The reader may also consult Strang [53], Demmel [14], Trefethen and Bau [56], and Golub and Van Loan [26].

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 [26], Chapter 8, Demmel [14], Trefethen and Bau [56], Lecture 26, or Ciarlet [11]. 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 [26], Chapter 8, Demmel [14], and Trefethen and Bau [56], 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 [40]). 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 12). Applications of the material of this chapter can be found in Strang [53, 52]; Ciarlet [11]; Golub and Van Loan [26], which contains many other references; Demmel [14]; and Trefethen and Bau [56].

11.4 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 [31] (Section 3.4) we can make the following definitions:

Definition 11.5. 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 5, 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 [31] (Section 3.4, Corollary 3.4.4 and Problem 3).

11.5 Summary

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

- For any linear map $f: E \to E$ 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 $\operatorname{Im} f$, $\operatorname{Ker} f$, $\operatorname{Im} f^*$, and $\operatorname{Ker} f^*$.
- The *singular value decomposition theorem* for square matrices (Theorem 11.2).
- The *SVD* of matrix.
- The *polar decomposition* of a matrix.
- The *Weyl inequalities*.
- The *singular value decomposition theorem* for $m \times n$ matrices (Theorem 11.4).
- Ky Fan k-norms, Ky Fan p-k-norms, Schatten p-norms.

Chapter 12

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*

12.1 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 12.1. *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 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 12.1. Given any $m \times n$ matrix A, if $A = VDU^\top$ is an SVD of A with

$$D = \mathrm{diag}(\lambda_1, \ldots, \lambda_r, 0, \ldots, 0),$$

where D is an $m \times n$ matrix and $\lambda_i > 0$, if we let

$$D^+ = \mathrm{diag}(1/\lambda_1, \ldots, 1/\lambda_r, 0, \ldots, 0),$$

an $n \times m$ matrix, the *pseudo-inverse* of A is defined by

$$A^+ = UD^+V^\top.$$

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 12.2. *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 12.2 and Theorem 12.1, A^+b is uniquely defined by every b, and thus A^+ depends only on A.

Let $A = U\Sigma V^\top$ be an SVD for 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,

$$AA^+ = U\Sigma V^\top V\Sigma^+U^\top = U\Sigma\Sigma^+U^\top = U\begin{pmatrix} I_r & 0 \\ 0 & 0_{n-r} \end{pmatrix}U^\top$$

and

$$A^+A = V\Sigma^+U^\top U\Sigma V^\top = V\Sigma^+\Sigma V^\top = V\begin{pmatrix} I_r & 0 \\ 0 & 0_{n-r} \end{pmatrix}V^\top.$$

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). *We claim that 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* $A^\top$.

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 Hermitian, $\mathrm{range}(A^+A) = \mathrm{Ker}(A^+A)^\perp = \mathrm{Ker}(A)^\perp$, as claimed.

It will also be useful to see that $\mathrm{range}(A) = \mathrm{range}(AA^+)$ consists of all vectors $y \in \mathbb{R}^n$ such that

$$U^\top y = \begin{pmatrix} z \\ 0 \end{pmatrix},$$

with $z \in \mathbb{R}^r$.

Indeed, if $y = Ax$, then

$$U^\top y = U^\top Ax = U^\top U\Sigma V^\top x = \Sigma V^\top x = \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0_{n-r} \end{pmatrix}V^\top x = \begin{pmatrix} z \\ 0 \end{pmatrix},$$

where Σ_r is the $r \times r$ diagonal matrix $\mathrm{diag}(\sigma_1, \ldots, \sigma_r)$. Conversely, if $U^\top y = \binom{z}{0}$, then $y = U\binom{z}{0}$, and

$$\begin{aligned} AA^+ y &= U \begin{pmatrix} I_r & 0 \\ 0 & 0_{n-r} \end{pmatrix} U^\top y \\ &= U \begin{pmatrix} I_r & 0 \\ 0 & 0_{n-r} \end{pmatrix} U^\top U \begin{pmatrix} z \\ 0 \end{pmatrix} \\ &= U \begin{pmatrix} I_r & 0 \\ 0 & 0_{n-r} \end{pmatrix} \begin{pmatrix} z \\ 0 \end{pmatrix} \\ &= U \begin{pmatrix} z \\ 0 \end{pmatrix} = y, \end{aligned}$$

which shows that y belongs to the range of A.

Similarly, we claim that $\mathrm{range}(A^+A) = \mathrm{Ker}(A)^\perp$ consists of all vectors $y \in \mathbb{R}^n$ such that

$$V^\top y = \begin{pmatrix} z \\ 0 \end{pmatrix},$$

with $z \in \mathbb{R}^r$.

If $y = A^+Au$, then

$$y = A^+Au = V \begin{pmatrix} I_r & 0 \\ 0 & 0_{n-r} \end{pmatrix} V^\top u = V \begin{pmatrix} z \\ 0 \end{pmatrix},$$

for some $z \in \mathbb{R}^r$. Conversely, if $V^\top y = \binom{z}{0}$, then $y = V\binom{z}{0}$, and so

$$\begin{aligned} A^+AV \begin{pmatrix} z \\ 0 \end{pmatrix} &= V \begin{pmatrix} I_r & 0 \\ 0 & 0_{n-r} \end{pmatrix} V^\top V \begin{pmatrix} z \\ 0 \end{pmatrix} \\ &= V \begin{pmatrix} I_r & 0 \\ 0 & 0_{n-r} \end{pmatrix} \begin{pmatrix} z \\ 0 \end{pmatrix} \\ &= V \begin{pmatrix} z \\ 0 \end{pmatrix} = y, \end{aligned}$$

which shows that $y \in \mathrm{range}(A^+A)$.

If A is a symmetric matrix, then in general, there is no SVD $U\Sigma V^\top$ of A with $U = V$. 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

$$A = U\Sigma U^\top.$$

Analogous results hold for complex matrices, but in this case, U and V 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 it can be shown 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 12.3. *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 [34].

Proposition 12.4. *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 [53], Golub and Van Loan [26], Demmel [14], and Trefethen and Bau [56], where extensions and applications of least squares (such as weighted least squares and recursive least squares) are described. Golub and Van Loan [26] also contains a very extensive bibliography, including a list of books on least squares.

12.2 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 5 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 12.5. *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*

$$A_k = \sum_{i=1}^{k} \sigma_i v_i u_i^\top = V \mathrm{diag}(\sigma_1, \ldots, \sigma_k) U^\top$$

and $\|A - A_k\|_2 = \sigma_{k+1}$.

Proof. By construction, A_k has rank k, and we have

$$\|A - A_k\|_2 = \Big\| \sum_{i=k+1}^{p} \sigma_i v_i u_i^\top \Big\|_2 = \big\| V \mathrm{diag}(0, \ldots, 0, \sigma_{k+1}, \ldots, \sigma_p) U^\top \big\|_2 = \sigma_{k+1}.$$

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 $p - k$. The subspace V_{k+1} spanned by $(v_1, \ldots, v_{k+1})$ has dimension $k + 1$, and because the sum of the dimensions of the kernel of B and of V_{k+1} is $(p - k) + k + 1 = p + 1$, these two subspaces must intersect in a subspace of dimension at least 1. Pick any unit vector h in $\mathrm{Ker}(B) \cap V_{k+1}$. Then since $Bh = 0$, we have

$$\|A - B\|_2^2 \geq \|(A - B)h\|_2^2 = \|Ah\|_2^2 = \big\|VDU^\top h\big\|_2^2 \geq \sigma_{k+1}^2 \big\|U^\top h\big\|_2^2 = \sigma_{k+1}^2,$$

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 12.5 in image compression is given in Demmel [14], 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 [14], which contains an overview of most known methods and an extensive list of references.

12.3 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 [18] (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 12.2. 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 12.6. *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 12.6 is often known as part of the *Rayleigh–Ritz theorem*.

Proposition 12.6 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 12.7. *(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 12.6, 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 12.6, 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.

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.

12.4 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 12.8. *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 [21] (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, [29] (Chapter 14, Section 14.5.1).

12.5 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 12.1).
- The *pseudo-inverse* A^+ of a matrix A.
- The least squares solution of smallest norm is given by the pseudo-inverse (Theorem 12.2)
- 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 12.6).
- The main theorem: *SVD yields PCA* (Theorem 12.7).
- Best affine approximation.
- SVD yields a best affine approximation (Theorem 12.8).
- Face recognition, eigenfaces.

Chapter 13

Quadratic Optimization Problems

13.1 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
$$f(x) = \frac{1}{2}x^\top Ax + x^\top b$$
over all $x \in \mathbb{R}^n$, or subject to linear or affine constraints.

2. Minimizing
$$f(x) = \frac{1}{2}x^\top Ax + x^\top b$$
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
$$P(x) = x^\top Ax - x^\top b,$$

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

$$P(x) = \frac{1}{2} x^\top A x - x^\top b.$$

The question is, under what conditions (on A) does $P(x)$ 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 $P(x)$ has a unique global minimum precisely when
$$Ax = b.$$
2. In Section 13.2, 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 11.2.

Definition 13.1. 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 13.1. *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

$$\langle x, f(x)\rangle > 0$$

for all $x \in E$ with $x \neq 0$.

(2) Every self-adjoint linear map $f : E \to E$ is positive semidefinite iff

$$\langle x, f(x)\rangle \geq 0$$

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

$$\langle x, f(x)\rangle = \Big\langle \sum_{i=1}^n x_i e_i, f\Big(\sum_{i=1}^n x_i e_i\Big)\Big\rangle = \Big\langle \sum_{i=1}^n x_i e_i, \sum_{i=1}^n \lambda_i x_i e_i\Big\rangle = \sum_{i=1}^n \lambda_i x_i^2,$$

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

$$\langle x, f(x)\rangle > 0$$

for all $x \neq 0$. Then for $x = e_i$, we get

$$\langle e_i, f(e_i)\rangle = \langle e_i, \lambda_i e_i\rangle = \lambda_i,$$

and thus $\lambda_i > 0$ for all $i = 1, \ldots, n$.

(2) As in (1), we have

$$\langle x, f(x)\rangle = \sum_{i=1}^{n} \lambda_i x_i^2,$$

and since $\lambda_i \geq 0$ for $i = 1, \ldots, n$ because f is positive semidefinite, we have $\langle x, f(x)\rangle \geq 0$, as claimed. The converse is as in (1) except that we get only $\lambda_i \geq 0$ since $\langle e_i, f(e_i)\rangle \geq 0$. □

Some special notation is customary (especially in the field of convex optinization) to express that a symmetric matrix is positive definite or positive semidefinite.

Definition 13.2. 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 [9], 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 $m \geq n$), then $A^\top CA$ is symmetric positive definite.

We can now prove that

$$P(x) = \frac{1}{2} x^\top Ax - x^\top b$$

has a global minimum when A is symmetric positive definite.

Proposition 13.2. *Given a quadratic function*

$$P(x) = \frac{1}{2} x^\top Ax - x^\top b,$$

if A is symmetric positive definite, then $P(x)$ has a unique global minimum for the solution of the linear system $Ax = b$. The minimum value of $P(x)$ is

$$P(A^{-1}b) = -\frac{1}{2} b^\top A^{-1} b.$$

Proof. Since A is positive definite, it is invertible, since its eigenvalues are all strictly positive. Let $x = A^{-1}b$, and compute $P(y) - P(x)$ for any $y \in \mathbb{R}^n$. Since $Ax = b$, we get

$$\begin{aligned} P(y) - P(x) &= \frac{1}{2}y^\top Ay - y^\top b - \frac{1}{2}x^\top Ax + x^\top b \\ &= \frac{1}{2}y^\top Ay - y^\top Ax + \frac{1}{2}x^\top Ax \\ &= \frac{1}{2}(y-x)^\top A(y-x). \end{aligned}$$

Since A is positive definite, the last expression is nonnegative, and thus

$$P(y) \geq P(x)$$

for all $y \in \mathbb{R}^n$, which proves that $x = A^{-1}b$ is a global minimum of $P(x)$. A simple computation yields

$$P(A^{-1}b) = -\frac{1}{2}b^\top A^{-1}b.$$

□

Remarks:

(1) The quadratic function $P(x)$ is also given by

$$P(x) = \frac{1}{2}x^\top Ax - b^\top x,$$

but the definition using $x^\top b$ is more convenient for the proof of Proposition 13.2.

(2) If $P(x)$ contains a constant term $c \in \mathbb{R}$, so that

$$P(x) = \frac{1}{2}x^\top Ax - x^\top b + c,$$

the proof of Proposition 13.2 still shows that $P(x)$ has a unique global minimum for $x = A^{-1}b$, but the minimal value is

$$P(A^{-1}b) = -\frac{1}{2}b^\top A^{-1}b + c.$$

Thus, when the energy function $P(x)$ of a system is given by a quadratic function

$$P(x) = \frac{1}{2}x^\top Ax - x^\top b,$$

where A is symmetric positive definite, finding the global minimum of $P(x)$ 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

$$Q(y_1, y_2) = \frac{1}{2}\left(y_1^2 + y_2^2\right)$$

subject to the constraint

$$2y_1 - y_2 = 5.$$

The solution for which $Q(y_1, y_2)$ is minimum is no longer $(y_1, y_2) = (0, 0)$, but instead, $(y_1, y_2) = (2, -1)$, as will be shown later.

Geometrically, the graph of the function defined by $z = Q(y_1, y_2)$ in $\mathbb{R}^3$ is a paraboloid of revolution P with axis of revolution Oz. The constraint

$$2y_1 - y_2 = 5$$

corresponds to the vertical plane H parallel to the z-axis and containing the line of equation $2y_1 - y_2 = 5$ 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*. But first, let us define precisely what kind of minimization problems we intend to solve.

Definition 13.3. The *quadratic constrained minimization problem* consists in minimizing a quadratic function

$$Q(y) = \frac{1}{2} y^\top C^{-1} y - b^\top y$$

subject to the linear constraints

$$A^\top y = f,$$

where C^{-1} is an $m \times m$ symmetric positive definite matrix, A is an $m \times n$ matrix of rank n (so that $m \geq n$), and where $b, y \in \mathbb{R}^m$ (viewed as column vectors), and $f \in \mathbb{R}^n$ (viewed as a column vector).

The reason for using C^{-1} instead of C is that the constrained minimization problem has an interpretation as a set of equilibrium equations in which the matrix that arises naturally is C (see Strang [52]). Since C and C^{-1} are both symmetric positive definite, this doesn't make any difference, but it seems preferable to stick to Strang's notation.

The method of Lagrange consists in incorporating the n constraints $A^\top y = f$ into the quadratic function $Q(y)$, by introducing extra variables $\lambda = (\lambda_1, \ldots, \lambda_n)$ called *Lagrange multipliers*, one for each constraint. We form the *Lagrangian*

$$L(y, \lambda) = Q(y) + \lambda^\top (A^\top y - f) = \frac{1}{2} y^\top C^{-1} y - (b - A\lambda)^\top y - \lambda^\top f.$$

We shall prove that our constrained minimization problem has a unique solution given by the system of linear equations

$$\begin{aligned} C^{-1}y + A\lambda &= b, \\ A^\top y &= f, \end{aligned}$$

which can be written in matrix form as

$$\begin{pmatrix} C^{-1} & A \\ A^\top & 0 \end{pmatrix} \begin{pmatrix} y \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ f \end{pmatrix}.$$

Note that the matrix of this system is symmetric. Eliminating y from the first equation

$$C^{-1}y + A\lambda = b,$$

we get

$$y = C(b - A\lambda),$$

and substituting into the second equation, we get

$$A^\top C(b - A\lambda) = f,$$

that is,

$$A^\top CA\lambda = A^\top Cb - f.$$

However, by a previous remark, since C is symmetric positive definite and the columns of A are linearly independent, $A^\top CA$ is symmetric positive definite, and thus invertible. Note that this way of solving the system requires solving for the Lagrange multipliers first.

Letting $e = b - A\lambda$, we also note that the system

$$\begin{pmatrix} C^{-1} & A \\ A^\top & 0 \end{pmatrix} \begin{pmatrix} y \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ f \end{pmatrix}$$

is equivalent to the system

$$\begin{aligned} e &= b - A\lambda, \\ y &= Ce, \\ A^\top y &= f. \end{aligned}$$

The latter system is called the *equilibrium equations* by Strang [52]. 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, y, e, b, C, λ, f, and $K = A^\top CA$ have a physical

interpretation. The matrix $K = A^\top CA$ is usually called the *stiffness matrix*. Again, the reader is referred to Strang [52].

In order to prove that our constrained minimization problem has a unique solution, we proceed to prove that the constrained minimization of $Q(y)$ subject to $A^\top y = f$ is equivalent to the unconstrained maximization of another function $-P(\lambda)$. We get $P(\lambda)$ by minimizing the Lagrangian $L(y, \lambda)$ treated as a function of y alone. Since C^{-1} is symmetric positive definite and

$$L(y, \lambda) = \frac{1}{2} y^\top C^{-1} y - (b - A\lambda)^\top y - \lambda^\top f,$$

by Proposition 13.2 the global minimum (with respect to y) of $L(y, \lambda)$ is obtained for the solution y of

$$C^{-1} y = b - A\lambda,$$

that is, when

$$y = C(b - A\lambda),$$

and the minimum of $L(y, \lambda)$ is

$$\min_y L(y, \lambda) = -\frac{1}{2}(A\lambda - b)^\top C(A\lambda - b) - \lambda^\top f.$$

Letting

$$P(\lambda) = \frac{1}{2}(A\lambda - b)^\top C(A\lambda - b) + \lambda^\top f,$$

we claim that the solution of the constrained minimization of $Q(y)$ subject to $A^\top y = f$ is equivalent to the unconstrained maximization of $-P(\lambda)$. Of course, since we minimized $L(y, \lambda)$ with respect to y, we have

$$L(y, \lambda) \geq -P(\lambda)$$

for all y and all λ. However, when the constraint $A^\top y = f$ holds, $L(y, \lambda) = Q(y)$, and thus for any admissible y, which means that $A^\top y = f$, we have

$$\min_y Q(y) \geq \max_\lambda -P(\lambda).$$

In order to prove that the unique minimum of the constrained problem $Q(y)$ subject to $A^\top y = f$ is the unique maximum of $-P(\lambda)$, we compute $Q(y) + P(\lambda)$.

Proposition 13.3. *The quadratic constrained minimization problem of Definition 13.3 has a unique solution (y, λ) given by the system*

$$\begin{pmatrix} C^{-1} & A \\ A^\top & 0 \end{pmatrix} \begin{pmatrix} y \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ f \end{pmatrix}.$$

Furthermore, the component λ of the above solution is the unique value for which $-P(\lambda)$ is maximum.

Proof. As we suggested earlier, let us compute $Q(y) + P(\lambda)$, assuming that the constraint $A^\top y = f$ holds. Eliminating f, since $b^\top y = y^\top b$ and $\lambda^\top A^\top y = y^\top A\lambda$, we get

$$\begin{aligned} Q(y) + P(\lambda) &= \frac{1}{2} y^\top C^{-1} y - b^\top y + \frac{1}{2}(A\lambda - b)^\top C(A\lambda - b) + \lambda^\top f \\ &= \frac{1}{2}(C^{-1}y + A\lambda - b)^\top C(C^{-1}y + A\lambda - b). \end{aligned}$$

Since C is positive definite, the last expression is nonnegative. In fact, it is null iff

$$C^{-1}y + A\lambda - b = 0,$$

that is,

$$C^{-1}y + A\lambda = b.$$

But then the unique constrained minimum of $Q(y)$ subject to $A^\top y = f$ is equal to the unique maximum of $-P(\lambda)$ exactly when $A^\top y = f$ and $C^{-1}y + A\lambda = b$, which proves the proposition. □

Remarks:

(1) There is a form of duality going on in this situation. The constrained minimization of $Q(y)$ subject to $A^\top y = f$ is called the *primal problem*, and the unconstrained maximization of $-P(\lambda)$ is called the *dual problem*. Duality is the fact stated slightly loosely as

$$\min_y Q(y) = \max_\lambda -P(\lambda).$$

Recalling that $e = b - A\lambda$, since

$$P(\lambda) = \frac{1}{2}(A\lambda - b)^\top C(A\lambda - b) + \lambda^\top f,$$

we can also write

$$P(\lambda) = \frac{1}{2} e^\top C e + \lambda^\top f.$$

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 $-P(\lambda)$).

(2) It is immediately verified that the equations of Proposition 13.3 are equivalent to the equations stating that the partial derivatives of the Lagrangian $L(y, \lambda)$ are null:

$$\begin{aligned} \frac{\partial L}{\partial y_i} &= 0, \quad i = 1, \ldots, m, \\ \frac{\partial L}{\partial \lambda_j} &= 0, \quad j = 1, \ldots, n. \end{aligned}$$

Thus, the constrained minimum of $Q(y)$ subject to $A^\top y = f$ is an extremum of the Lagrangian $L(y, \lambda)$. As we showed in Proposition 13.3, this extremum corresponds to simultaneously minimizing $L(y, \lambda)$ with respect to y and maximizing $L(y, \lambda)$ with respect to λ. Geometrically, such a point is a *saddle point* for $L(y, \lambda)$.

(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 [52].

Going back to the constrained minimization of $Q(y_1, y_2) = \frac{1}{2}(y_1^2 + y_2^2)$ subject to

$$2y_1 - y_2 = 5,$$

the Lagrangian is

$$L(y_1, y_2, \lambda) = \frac{1}{2}\big(y_1^2 + y_2^2\big) + \lambda(2y_1 - y_2 - 5),$$

and the equations stating that the Lagrangian has a saddle point are

$$\begin{aligned} y_1 + 2\lambda &= 0, \\ y_2 - \lambda &= 0, \\ 2y_1 - y_2 - 5 &= 0. \end{aligned}$$

We obtain the solution $(y_1, y_2, \lambda) = (2, -1, -1)$.

Much more should be said about the use of Lagrange multipliers in optimization or variational problems. This is a vast topic. Least squares methods and Lagrange multipliers are used to tackle many problems in computer graphics and computer vision; see Trucco and Verri [57], Metaxas [41], Jain, Katsuri, and Schunck [32], Faugeras [19], and Foley, van Dam, Feiner, and Hughes [20]. For a lucid introduction to optimization methods, see Ciarlet [11].

13.2 Quadratic Optimization: The General Case

In this section, we complete the study initiated in Section 13.1 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 13.4. *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 $A \succeq 0$. In this case, since $(x + A^{-1}b)^\top A(x + A^{-1}b) \geq 0$, 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 13.5. *If A is a 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 $A = U^\top \Sigma U$ is an SVD of A, 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 13.4, 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 Section 12.1 that b is in the range of A (here, U is $U^\top$), 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 13.4, 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, 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

$$x^* = -U^\top \begin{pmatrix} \Sigma_r^{-1} c \\ 0 \end{pmatrix} = -U^\top \begin{pmatrix} \Sigma_r^{-1} c & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} c \\ 0 \end{pmatrix} = -U^\top \begin{pmatrix} \Sigma_r^{-1} c & 0 \\ 0 & 0 \end{pmatrix} Ub = -A^+ b$$

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 $z \in \mathbb{R}^{n-r}$, our previous calculations show that $f(x) = -\frac{1}{2} b^\top A^+ b$. □

The case in which we add either linear constraints of the form $C^\top x = 0$ or affine constraints of the form $C^\top x = t$ (where $t \neq 0$) can be reduced to the unconstrained case using a QR-decomposition of C or N. 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, we may assume that

$$C = Q^\top \begin{pmatrix} R & S \\ 0 & 0 \end{pmatrix} \Pi,$$

where R is an $r \times r$ invertible upper triangular matrix and S is an $r \times (m-r)$ 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

$$\Pi^\top \begin{pmatrix} R & 0 \\ S & 0 \end{pmatrix} Qx = \Pi^\top \begin{pmatrix} R^\top & 0 \\ S^\top & 0 \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} = 0,$$

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

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}(y^\top, z^\top) QAQ^\top \begin{pmatrix} y \\ z \end{pmatrix} + (y^\top, z^\top) Qb \\ \text{subject to} & y = 0,\ y \in \mathbb{R}^r,\ z \in \mathbb{R}^{n-r}. \end{array}$$

Thus, the constraint $C^\top x = 0$ has been eliminated, and if we write

$$QAQ^\top = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix}$$

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 13.5.

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 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 $P^\top A P$; the details are left as an exercise.

13.3 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

$$\begin{array}{ll} \text{maximize} & x^\top A x \\ \text{subject to} & x^\top x = 1,\ x \in \mathbb{R}^n. \end{array}$$

In view of Proposition 12.6, 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 12.6, 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.

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 [25] (1973). The problem is the following: Given an $n \times n$ real symmetric matrix A and an $n \times p$ matrix C,

$$\begin{array}{ll} \text{minimize} & x^\top Ax \\ \text{subject to} & x^\top x = 1,\ C^\top x = 0,\ x \in \mathbb{R}^n. \end{array}$$

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 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 eliminated, 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. 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 $C^\top$. 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 [12] and implicitly by Yu and Shi [59].

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 [24] (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 $\left\|(N^\top)^\dagger t\right\| < 1$ (where $(N^\top)^\dagger$ 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 $\left\|(N^\top)^\dagger t\right\| < 1$ 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 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

$$P^\top A P = \begin{pmatrix} B & \Gamma^\top \\ \Gamma & C \end{pmatrix}$$

and

$$P^\top x = \begin{pmatrix} y \\ z \end{pmatrix},$$

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 12.6 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 [24].

13.4 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 $t \neq 0$.

Bibliography

[1] Emil Artin. *Geometric Algebra*. Wiley Interscience, first edition, 1957.

[2] Michael Artin. *Algebra*. Prentice Hall, first edition, 1991.

[3] Marcel Berger. *Géométrie 1*. Nathan, 1990. English edition: Geometry 1, Universitext, Springer Verlag.

[4] Marcel Berger. *Géométrie 2*. Nathan, 1990. English edition: Geometry 2, Universitext, Springer Verlag.

[5] J.E. Bertin. *Algèbre linéaire et géométrie classique*. Masson, first edition, 1981.

[6] Nicolas Bourbaki. *Algèbre, Chapitres 1-3*. Eléments de Mathématiques. Hermann, 1970.

[7] Nicolas Bourbaki. *Algèbre, Chapitres 4-7*. Eléments de Mathématiques. Masson, 1981.

[8] Nicolas Bourbaki. *Espaces Vectoriels Topologiques*. Eléments de Mathématiques. Masson, 1981.

[9] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, first edition, 2004.

[10] G. Cagnac, E. Ramis, and J. Commeau. *Mathématiques Spéciales, Vol. 3, Géométrie*. Masson, 1965.

[11] P.G. Ciarlet. *Introduction to Numerical Matrix Analysis and Optimization*. Cambridge University Press, first edition, 1989. French edition: Masson, 1994.

[12] 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.

[13] H.S.M. Coxeter. *Introduction to Geometry*. Wiley, second edition, 1989.

[14] James W. Demmel. *Applied Numerical Linear Algebra*. SIAM Publications, first edition, 1997.

[15] Jean Dieudonné. *Algèbre Linéaire et Géométrie Elémentaire*. Hermann, second edition, 1965.

[16] Jacques Dixmier. *General Topology*. UTM. Springer Verlag, first edition, 1984.

[17] David S. Dummit and Richard M. Foote. *Abstract Algebra*. Wiley, second edition, 1999.

[18] Charles L. Epstein. *Introduction to the Mathematics of Medical Imaging*. SIAM, second edition, 2007.

[19] Olivier Faugeras. *Three-Dimensional Computer Vision, A geometric Viewpoint*. the MIT Press, first edition, 1996.

[20] James Foley, Andries van Dam, Steven Feiner, and John Hughes. *Computer Graphics. Principles and Practice*. Addison-Wesley, second edition, 1993.

[21] David A. Forsyth and Jean Ponce. *Computer Vision: A Modern Approach*. Prentice Hall, first edition, 2002.

[22] Jean Fresnel. *Méthodes Modernes En Géométrie*. Hermann, first edition, 1998.

[23] Jean H. Gallier. *Geometric Methods and Applications, For Computer Science and Engineering*. TAM, Vol. 38. Springer, second edition, 2011.

[24] Walter Gander, Gene H. Golub, and Urs von Matt. A constrained eigenvalue problem. *Linear Algebra and its Applications*, 114/115:815–839, 1989.

[25] Gene H. Golub. Some modified eigenvalue problems. *SIAM Review*, 15(2):318–334, 1973.

[26] H. Golub, Gene and F. Van Loan, Charles. *Matrix Computations*. The Johns Hopkins University Press, third edition, 1996.

[27] Jacques Hadamard. *Leçons de Géométrie Elémentaire. I Géométrie Plane*. Armand Colin, thirteenth edition, 1947.

[28] Jacques Hadamard. *Leçons de Géométrie Elémentaire. II Géométrie dans l'Espace*. Armand Colin, eighth edition, 1949.

[29] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, second edition, 2009.

[30] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, first edition, 1990.

[31] Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, first edition, 1994.

[32] Ramesh Jain, Rangachar Katsuri, and Brian G. Schunck. *Machine Vision*. McGraw-Hill, first edition, 1995.

[33] Hoffman Kenneth and Kunze Ray. *Linear Algebra*. Prentice Hall, second edition, 1971.

[34] D. Kincaid and W. Cheney. *Numerical Analysis*. Brooks/Cole Publishing, second edition, 1996.

[35] Serge Lang. *Algebra*. Addison Wesley, third edition, 1993.

[36] Serge Lang. *Real and Functional Analysis*. GTM 142. Springer Verlag, third edition, 1996.

[37] Serge Lang. *Undergraduate Analysis*. UTM. Springer Verlag, second edition, 1997.

[38] Peter Lax. *Linear Algebra and Its Applications*. Wiley, second edition, 2007.

[39] Saunders Mac Lane and Garrett Birkhoff. *Algebra*. Macmillan, first edition, 1967.

[40] Jerrold E. Marsden and J.R. Hughes, Thomas. *Mathematical Foundations of Elasticity*. Dover, first edition, 1994.

[41] Dimitris N. Metaxas. *Physics-Based Deformable Models*. Kluwer Academic Publishers, first edition, 1997.

[42] Carl D. Meyer. *Matrix Analysis and Applied Linear Algebra*. SIAM, first edition, 2000.

[43] Joseph O'Rourke. *Computational Geometry in C*. Cambridge University Press, second edition, 1998.

[44] Dan Pedoe. *Geometry, A comprehensive Course*. Dover, first edition, 1988.

[45] Eugène Rouché and Charles de Comberousse. *Traité de Géométrie*. Gauthier-Villars, seventh edition, 1900.

[46] Laurent Schwartz. *Analyse I. Théorie des Ensembles et Topologie*. Collection Enseignement des Sciences. Hermann, 1991.

[47] Laurent Schwartz. *Analyse II. Calcul Différentiel et Equations Différentielles*. Collection Enseignement des Sciences. Hermann, 1992.

[48] Denis Serre. *Matrices, Theory and Applications*. GTM No. 216. Springer Verlag, second edition, 2010.

[49] Ernst Snapper and Troyer Robert J. *Metric Affine Geometry*. Dover, first edition, 1989.

[50] G.W. Stewart. On the early history of the singular value decomposition. *SIAM review*, 35(4):551–566, 1993.

[51] Eric J. Stollnitz, Tony D. DeRose, and David H. Salesin. *Wavelets for Computer Graphics Theory and Applications*. Morgan Kaufmann, first edition, 1996.

[52] Gilbert Strang. *Introduction to Applied Mathematics*. Wellesley-Cambridge Press, first edition, 1986.

[53] Gilbert Strang. *Linear Algebra and its Applications*. Saunders HBJ, third edition, 1988.

[54] Gilbert Strang and Nguyen Truong. *Wavelets and Filter Banks*. Wellesley-Cambridge Press, second edition, 1997.

[55] Claude Tisseron. *Géométries affines, projectives, et euclidiennes*. Hermann, first edition, 1994.

[56] L.N. Trefethen and D. Bau III. *Numerical Linear Algebra*. SIAM Publications, first edition, 1997.

[57] Emanuele Trucco and Alessandro Verri. *Introductory Techniques for 3D Computer Vision*. Prentice-Hall, first edition, 1998.

[58] B.L. Van Der Waerden. *Algebra, Vol. 1*. Ungar, seventh edition, 1973.

[59] 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.