Prerequisites for studies at advanced level in Image Science at Linköping University

Klas Nordberg

Computer Vision Laboratory
Department of Electrical Engineering
Linköping University

Version: 0.31 – August 4, 2015

Contents

Preface

This compendium presents a number of basic concept and results that are used as a basis for more advanced topics in Image Science. As such, they should be well-know to most readers, and they are collected here for easy reference and to allow you to, once again, put them at the top your head. In some cases, these results are complemented with additional material that may not be well-known, but at least is well-connected to the basic stuff. Since it is assumed that you have seen most or all of this before, there are little or no proofs for the statements made here. Readers interested in proofs are referred to basic textbooks on the subjects, mainly linear algebra and calculus.

Several people have been involved in the discussion, organization, and proof reading of the resulting compendium. In particular, I would like to thank Michael Felsberg, Per-Erik Forssén, John Hedborg, Jan-Åke Larsson, and Rudolf Mester for participating in this process.

Chapter 1

Sets and operations on sets

Sets are simplest, and most general, type of of mathematical structure, and the concept appears frequently throughout this and other presentations. This notwithstanding, the definitions of sets and some related concepts are not always provided during undergraduate studies. In order to have a clear notion of sets, which can be used when we want to talk about more advanced concepts, this chapter presents sets, operations on sets, and some related issues. Unless you are particularly interesting in the topic, you may skip parts of this chapter and only return if unfamiliar concepts appear later on.

1.1 Sets

A set consists of nothing but a rule that can be applied to virtually *any* object to determine if it belongs to the set or not. This rule is called a *membership function*, and it returns a binary value, e.g., YES and NO: a YES if the object belongs to the set, and a NO if it does not belong to the set. All objects that belong to the set are referred to as the elements of the set, and we can also say that the set consists of its elements. If a is an element of S, we write this as $a \in S$.

The definition of a set does not imply any ordering of its elements, i.e., there is no "first" or "last" element of a set, or any type of enumeration of its elements. Furthermore, all we can say about an object is whether or not it is an element of a specific set. There is no meaning to the statement that a set consists of two copies of the same element, at least on the sense that this set is distinct from a set that has only one copy of the element. The concept of collections, described in Section 1.1.3, define a group of elements that can be both ordered and include multiple copies of the same element.

Subsets

A set A is a *subset* of the set B if every element in A is also an element of B. We denote this as $A \subset B$. Trivially, this means that any set is a subset of itself. In the case that there exists at least one element of B that is not an element of A, we refer to A as a *proper subset* of B.

Some common sets

There are some sets that appear more frequently than others in this presentation, and have a specific notation.

- The set of natural numbers $\{0, 1, 2, \ldots\}$ is denoted $\mathbb{N}$.
- The set of integers is denoted $\mathbb{Z}$.
- The set of rational numbers is denoted $\mathbb{Q}$.
- The set of real numbers is denoted $\mathbb{R}$.
- The set of complex numbers is denoted $\mathbb{C}$.
- The empty set, that has no elements, is denoted $\varnothing$.

Universe set

The fact that the membership function can be applied to "anything" is what makes sets a very general concept. In most practical cases, however, the membership function is not applied to "anything". Instead it is restricted to a *universe set*, the set of objets that are relevant for the problem at hand. For example, if we want to talk about even numbers, this discussion is typically made in the context of the integers as the universe. There is no need to include also apples and pears, and to conclude that they are not even numbers. Instead, we define the set of even numbers as a subset of the integers, more precisely those that are even. The membership function is in this case restricted to only integers.

Equality of sets

A very basic operation on sets it to determine if two sets are equal. Since a set is completely determined by its membership function, two sets are equal exactly when their memberships functions are equal, i.e., they always give the same answer to the question of membership. A more practical consequence of this idea is to say that two sets, A and B, are equal if any element of A is also an element of B *and* any element of B is also an element of A. In this case we are allowed to write $A = B$. If we can find one single element that belongs to either of the two sets but not to both, the two sets are distinct: $A \neq B$.

Notation

If a set S only consists of a few elements, e.g., a, b, c, we can list them within a pair of braces:

$$S = \{a, b, c\}. \tag{1.1}$$

Since there is no ordering of the elements of a set, unless explicitly stated otherwise, the same set S can also be written as

$$S = \{c, a, b\}, \tag{1.2}$$

or as any other permutation of its elements. Many interesting sets have infinitely many elements, and then it is not possible to list its elements, as in Equation (1.1). Instead, we can write

$$S = \{x \in U \, : \text{ the membership function of } S\}, \tag{1.3}$$

explicitly specifying the universe set U and the membership function. For example, the set of all positive real numbers can be written as

$$\{x \in \mathbb{R} \, : \, x > 0\}. \tag{1.4}$$

Since the universe set U often is implied by the context in which a set appears, it is common to not mention U, other than implicitly by providing the general appearance of the elements of the set, i.e., how they are constructed from simpler sets, together with the membership function. Examples of this notation is provided below.

1.1.1 Cartesian Product and Tuples

The *Cartesian product* of two sets, A and B, consists of the set of all ordered pairs (a, b), where $a \in A$ and $b \in B$. The fact that (a, b) is an ordered pair means that it is a list rather than a set, it has a first element, here a, and a second element, here b. Furthermore, this list always contains exactly two elements. The Cartesian product can also be written as

$$A \times B = \{(a, b), \, a \in A, \, b \in B\} \tag{1.5}$$

Here, we omit the universe set in the construction of the Cartesian product since it is implied. An order pair is also referred to as a 2-tuple.

The idea of ordered pairs and 2-tuples can straight-forwardly be generalized to n-tuples as ordered lists of n elements: $(a_1, a_2, \ldots, a_n)$. Each element a_k in the n-tuple is itself an element from the set A_k. In this way, we define the Cartesian product of the n sets $A_1, \ldots, A_n$ as

$$A_1 \times \ldots \times A_n = \{(a_1, \ldots, a_n), \, a_1 \in A_1, \ldots, a_n \in A_n\} \tag{1.6}$$

Notice the distinction between a set and an n-tuple. The latter has always a fixed number of elements, n, while a set can have arbitrary many elements, even zero or infinitely many. Furthermore, the elements of the n-tuple are ordered and it is meaningful, for example, to talk about the first and second element of a 2-tuple, while an ordering does not exist for a set in general. Finally, for an n-tuple we always expect the k-th element to lie in the set A_k, while a general set does not have such specific requirement. Although it is possible to see an n-tuple as special type of set, with additional structures in terms of the orderings imposed by the Cartesian product, it is often better to see a tuple as a separate class of objects that are distinct from sets. For example, the operations defined on sets in Section 1.1.4, cannot be applied to n-tuples in a straight-forward manner.

1.1.2 Functions

A *function* from the set X to the set Y is a subset $f \subset A \times B$ which satisfy the requirements: (1) for each $x \in X$ there exists some $y \in Y$ such that $(x,y) \in f$, and (2) if $(x,y_1) \in f$ and $(x,y_2) \in f$ then $y_1 = y_2$. In a simpler language: for each $x \in X$ there can only be one single $y \in Y$ such that $(x,y) \in f$. This means that we can see f as a mechanism that assigns a unique $y \in Y$ to each $x \in X$, something we write as $f(x) = y$. In this case, we say that y is the *image* of x under the function f. For example, the membership function of a set is a function from the universe set U to the binary set $\{YES,NO\}$.

We use $f : X \to Y$ to denote that f is a function from X to Y. The set X is referred to as the *domain* of f, and Y is its *co-domain*. The smallest subset of Y that contains the images of all points in X is the *image* of X, sometimes denoted $f(X)$. In general, not every element in Y needs to be the image of some $x \in X$, i.e., it may be the case that $Y \neq f(X)$. An element x in the domain of f is often referred to as the *argument* or *variable* of f.

In terms of notation, we formally make a distinction between f, that denotes the function, and $f(x)$ that denotes the value of the function when applied to x. This distinction is sometimes important since $f \subset X \times Y$ while $f(x) \in Y$. This distinction, however, is sometimes impractical when we want to describe how f depends on its argument, for example, when we want to talk about the two functions $f(x)$ and $f(y)$ where y is, itself, depends on x.

It is common to use *mapping* as a synonym for a function. The set of all functions which maps from X to Y is denoted as maps$(X \to Y)$.

Onto

A function $f : X \to Y$ maps *onto* Y (or *surjective*) if $Y = f(X)$. This means that every $y \in Y$ is the image of at least one $x \in X$.

One-to-one

A function $f : X \to Y$ is *one-to-one* (or *injective*) if $f(x_1) = f(x_2)$ implies that $x_1 = x_2$. This means that two distinct $x_1, x_2 \in X$ cannot be mapped to the same $y \in Y$. This means that every $y \in Y$ is either not an image of any $x \in X$ or the image of a exactly one $x \in X$.

Invertible functions

A function $f : X \to Y$ is *invertible* (or *bijective*) if it maps one-to-one onto Y. This means that each $y \in Y$ is the image of exactly one $x \in X$, and implies that the inverse of f, often denoted f^{-1}, is a well-defined function that satisfies $y = f(x) \Rightarrow x = f^{-1}(y)$, for all $x \in X$ and $y \in Y$. If f is invertible, then f^{-1} has Y as its domain and X as its codomain.

Multi-variable functions

Let X_1 and X_2 be two sets and consider a function $f : X_1 \times X_2 \to Y$. This function maps an ordered pair of elements $(x_1,x_2) \in X_2 \times X_2$ to the co-domain Y. This implies that in order to determine the function value of f applied to (x_1,x_2), denoted $f(x_1,x_2)$, we need to know both x_1 and x_2. In a natural way this defines f as a two-variable function.

This idea can be generalized to functions of an arbitrary, but fixed, number of variables. Given n sets $A_1,\ldots,A_n$, we define an n-variable function on these sets as a function from the Cartesian product $A' = A_1 \times \ldots \times A_n$ to some co-domain Y. This means that if f is an n-variable function, we can choose to see it a function from n variables,

each from its specific set A_k, or a one-variable function where the variable comes from A'. Both views are useful, but perhaps in different contexts. For example, in order make the notation more compact, we may sometimes choose to see f as a one-variable function from A', and in order to define what we mean by the partial derivative of f with respect to one of its variables, we can see the same function as multi-variable function.

1.1.3 Collections

In its simplest form a collection is finite sequence of n elements, each from some common set A. This type of collection can be denoted as

$$C = \{a_1, a_2, \ldots, a_n\}, \quad \text{or} \quad C = \{a_i, i = 1, \ldots, n\}. \tag{1.7}$$

A collection is somewhat similar to the idea of a tuple, described in Section 1.1.1, the collection C above could be described as an n-tuple:

$$C \in \underbrace{A \times \ldots \times A}_{n}. \tag{1.8}$$

A more general description of a collection is as a function from I to A. Since I has a natural order, there is an order defined also for the elements of C. The set I is in this case referred to as the *index set* of the collection C, and the elements of C are then indexed by the set I.

This type of finite collection can be generalized in a straight-forward manner to infinite collections, which makes collections different from tuples. In this case, the collection is indexed either by the natural numbers, $\mathbb{N}$, or the integers, $\mathbb{Z}$. Again, the ordering of the index set induces an ordering also of any collection based on these index sets. We will sometimes use *sequence* as a synonym of collection.

Although sets and collections sometimes may have the same notation, in terms of brackets that enclose a list of elements, it should be clear from the definition if a set or a collection is constructed or, otherwise from the context in which they are used.

1.1.4 Basic operations on sets

With sets and functions established, we can now define a set of common operations on sets and relations between sets.

Set difference

Given two sets, A and B, we define the difference between the two sets as: all elements of A that are not also elements of B. The set difference is denoted as $A \backslash B$ and is formally defined as:

$$A \backslash B = \{a \in A, a \notin B\} \tag{1.9}$$

This means that $A \backslash B$ is always a subset of A.

Complement

Given a set S, we can sometimes be interested in the sets that contains everything except the elements of S. We refer to this as the *complement* of S, denoted $\mathsf{C}[S]$. This definition of the complement is a bit cumbersome since the "everything but" really means "everything but", including your grandmother if she is not already an element of S. To make the the complement of a set more manageable, we can define it relative to a universe set U, consisting of everything that are meaningful for the definition of S. In this case the complement $\mathsf{C}[S]$ becomes the set difference between the universe set and S:

$$\mathsf{C}[S] = U \backslash S. \tag{1.10}$$

Union

Given two sets, A and B, their *union* is the set that contains all elements in B *or* in B, including those that are in both A and B, denoted $A \cup B$. Although not necessary from a formal point of view, the definition of the union operation on sets often implies that A and B are subsets of a common universe set U. We can then write

$$A \cup B = \{x \in U, x \in A \text{ or } x \in B\} \tag{1.11}$$

Notice that both A and B are subsets of $A \cup B$.

Let C be a collection of sets, all having the same universe U, over some index set: $C = \{A_i, i = I\}$. We can then form the union of all sets in the collection as

$$\bigcup_{i \in I} A_i = \{a \in U, a \text{ is an element of at least one } A_i \in C\} \tag{1.12}$$

Intuitively, we can think of the union of some collection of sets as the "smallest" set that includes all elements in all the sets of the collection.

Cut

Given two sets, A and B, their *cut* is the set that contains all elements both of *SA and* of B, denoted $A \cap B$. In a similar way as for the union operation, we normally assume that both A and B are subsets of a common universe set U. We can then write

$$A \cap B = \{x \in U, x \in A \text{ and} x \in B\} \tag{1.13}$$

Notice that $A \cap B$ is a subset of both A and of B.

Let C be a collection of sets, all having the same universe U, over some index set: $C = \{A_i, i = I\}$. We can then form the cut of all sets in the collection as

$$\bigcap_{i \in I} A_i = \{a \in U, a \text{ is an element of every } A_i \in C\} \tag{1.14}$$

Intuitively, we can think of the cut of some collection of sets as the "largest" set of elements that belong to every set of the collection.

We sometime use *intersection* as a synonym to cut, in particular when referring to sets of geometric objects, such as points in 2D or 3D space.

Disjoint sets

Two sets, A and B, are disjoint if they have no elements in common, i.e., there is no element of one of the two sets that is also an element of the other. Yet another way to formulate the same thing is: A and B are disjoint if their cut is empty: $A \cap B = \varnothing$. A collection of sets is a collection of disjoint sets if every pair of sets in the collection are disjoint.

1.1.5 Binary functions on a set

A binary function on a set S is a function $S \times S : \rightarrow S$, i.e., it takes an ordered pair of elements from S and map it to an element in S.

There are many common examples of binary operations. For example, addition, subtraction, and multiplication are binary operations on $\mathbb{R}$. In principe, also division is a binary operation on $\mathbb{R}$, although we must exclude O as denominator. Another example of a binary operation is multiplication of square matrices.

A binary operation is sometimes characterized by the way it maps the two elements in S.

Commutative operation

A binary operation f is a *commutative operation* if $f(s_1, s_2) = f(s_2, s_1)$ for all $s_1, s_2 \in S$. Adding real numbers is a commutative binary operation, while subtracting real numbers is not.

Associative operations

A binary operation f is an *associative operation* if

$$f(s_1, f_(s_2, s_3)) = f(f(s_1, s_2), s_3) \quad \text{for all } s_1, s_2, s_3 \in S \tag{1.15}$$

Adding real numbers is an associative binary operation, while taking the vector cross product in $\mathbb{R}^3$ is not.

Identity element

A binary operation f has an *identity element* $e \in S$ if $f(s,e) = f(e,s) = s$ for all $s \in S$. 0 is an identity element when f is addition of real numbers, and 1 is an identity element when f is multiplication of real numbers. If f is the vector cross product in $\mathbb{R}^3$, then there is no identity element of f.

Inverse

An element $s \in S$ has an*inverse* s^{-1} relative to a binary operation f if $f(s,s^{-1}) = f(s^{-1},s) = e$ for each $s \in S$, where e is the neutral element.

Distributive operations

Let f_1 and f_2 be two binary operations on S. If

$$f_1(s_1, f_2(s_1, s_2)) = f_2(f_1(s_1, s_2), f_1(s_1, s_3)) \quad \text{for all } s_1, s_2, s_3 \in S \tag{1.16}$$

then f_1 is a distributive operation over f_2. Multiplication is a distributive operation over addition for real numbers.

1.2 Equivalence relations and equivalence classes

Given a set S, we consider a relation R on S, in terms of ordered pairs (s_1, s_2) where $s_1, s_2 \in S$. In the following, we use *sim* to denote the relation R, i.e., we write $s_1 \sim s_2$ as a shorthand for $(s_1, s_2) \in R$. In general, there exist many different relations on any specific S but, here, we are particularly interested in relations that satisfy the following properties:

1. For any $s \in S$: $s \sim s$ (reflexivity).
2. If $s_1 \sim s_2$ then $s_2 \sim s_1$ for any $s_1, s_2 \in S$ (symmetry).
3. If $s_1 \sim s_2$ and $s_2 \sim s_3$ then $s_1 \sim s_3$ for any $s_1, s_2, s_3 \in S$ (transitivity).

These three properties restrict $\sim$ to what is known as an *equivalence relation*.

Given an equivalence relation $\sim$, we may pick an arbitrary $s \in S$ and form the subset $E(s) \subset S$ as

$$E(s) = \{s' : s' \sim s\}. \tag{1.17}$$

This means that $E(s)$ is the subset of S that contains all $s' \in S$ that are equivalent to s. Now, consider $s, s' \in S$, and the corresponding subsets $E(s)$ and $E(s')$. Because the two subsets are constructed from an equivalence relation in Equation (1.17), it follows that:

$$\begin{cases} E(s) = E(s'), & s \sim s', \\ E(s) \cap E(s') = \varnothing, & \text{otherwise.} \end{cases} \tag{1.18}$$

The subsets constructed in this way are referred to as *equivalence classes*, and Equation (1.18) implies that two equivalence classes are either equal or disjoint. Another way to formulate the same thing is: each $s \in S$ belongs to *exactly one* equivalence class generated by $\sim$. Since there is exactly one equivalence class for each $s \in S$, it must be the case that the union of all equivalence classes constitutes S.

If E is an equivalence class of S generated by $\sim$, and $s \in E$, then s is referred to as a *representative* of E. In general, each equivalence class has many representatives, but it must be the case that if both s and s' are representatives of the same E, then $s \sim s'$.

We can summarize the properties of equivalence relations and equivalence classes as: the equivalence relation $\sim$ partitions S into disjoint equivalence classes that together fill out S. Furthermore, each $s \in S$ is a representative of exactly one of these equivalence classes.

1.3 Groups

A group is one of the simplest and also most common algebraic structure, although the fact that some well-known sets and operations on these sets form groups is perhaps not always made explicit. Formally, a group consists of a set of group elements, G, and a group operation on G, here denoted $\circ$. Together they satisfy the following properties, sometimes known as the *group axioms*:

1. The group operation is a binary operation on G that is *closed*, i.e., $\circ : G \times G \to G$.
2. The group operation is *associative*: $g_1 \circ (g_2 \circ g_3) = (g_1 \circ g_2) \circ g_3$, for all $g_1, g_2, g_3 \in G$.
3. There exists a *neutral element*, or *identity element* in $e \in G$ such that $e \circ g = g \circ e = g$ for all $g \in G$.
4. Each $g \in G$ has an *inverse*, here denoted $g^{-1} \in G$, such that $g \circ g^{-1} = g^{-1} \circ g = e$.

An important observation is that the group axioms alone lead to two additional properties of a group:

5. The neutral element e is unique.
6. For each $g \in G$, its inverse g^{-1} is unique.

The definition of a group implies that it consists of two parts: the set of group elements G and the group operation $\circ$, and this case we say that G forms a group under the operation $\circ$. This notwithstanding we often use only G to denote the group and rely on the context to define which group operation that is relevant for this set.

There is a extensive theory related to groups, obtained by successively introducing additional properties to the basic ones presented here. In this presentation, however, we will only have reasons to introduce two additional properties related to a group. First, an *Abelian group* is a group where the group operation is *commutative*, i.e., $g_1 \circ g_2 = g_2 \circ g_1$ for all $g_1, g_2 \in G$. A group that is not Abelian is referred to as a *non-Abelian group*. As you will see shortly, there are plenty of examples of both Abelian and non-Abelian group. Second, when G is a group, a subset $G' \subset G$ is a *subgroup* of G if G' itself forms a group with the same group operation, the same neutral element, and the same inverses as G.

Examples

The following is a short list of examples of groups that frequently appear in practical examples and theoretical analysis.

- The set of real numbers $\mathbb{R}$ and the addition operation form a group. The neutral element is 0, and the inverse of $x \in \mathbb{R}$ is $-x$. The set of complex numbers $\mathbb{C}$ also forms a group under the addition operation. Alternatively, we can restrict $\mathbb{R}$ to the set of rational numbers $\mathbb{Q}$, or to the set of integers $\mathbb{Z}$, that both form groups under the addition operation. The set of even integers form a group with the addition operation. All these groups are Abelian.
- We can extend this idea to the vector space $\mathbb{R}^n$ that forms a group with the operation of vector addition. The neutral element is the zero vector $\mathbf{0}$, and the inverse of vector $\mathbf{v} \in \mathbb{R}^n$ is $-\mathbf{v}$. This is an Abelian group.
- The set of real numbers excluding zero, here denoted $\mathbb{R}_{-0}$, and the multiplication operation form a group. The neutral element is 1, and the inverse of $x \in \mathbb{R}_{-0}$ is $1/x$. The extension to $\mathbb{C}_{-0}$ and the restriction to $\mathbb{Q}_{-0}$ also form groups with the multiplication operation. This is not the case for $\mathbb{Z}$ since the inverses, in general, then lie in $\mathbb{Q}$. However, the set $\{-1, 1\}$ and multiplication form a group. All these groups are Abelian.
- Linear non-singular transformations onto $\mathbb{R}^n$, as defined by $n \times n$ matrices of non-zero determinant, form a group together with the matrix product. This group is the *general linear group* on $\mathbb{R}^n$, denoted $GL(n)$. The neutral element is the identity matrix $\mathbf{I}$ and the inverse of matrix $\mathbf{M} \in GL(n)$ is its matrix inverse $\mathbf{M}^{-1}$. Since the matrix product in general is not commutative, this group is non-Abelian.
- There are several subgroups of $GL(n)$ that appear in the literature, and here we will only mention a few examples. The *special linear group* is the subgroup of $GL(n)$ consisting of matrices with determinant $= 1$. This group is denoted $SL(n)$. Again, this is a non-Abelian group.

- Another common subgroup of $GL(n)$ is the *orthogonal group*, i.e., matrices $\mathbf{R} \in GL(n)$ such that $\mathbf{R}^\top \mathbf{R} = \mathbf{I}$. This group is denoted $O(n)$. It follows that matrices in $O(n)$ have determinant $= \pm 1$. A subgroup of the orthogonal group is the *special orthogonal group*, consisting of orthogonal matrices with determinant $= 1$. This group consists of rotations in $\mathbb{R}^n$ and is denoted $SO(n)$ and is, in fact, a subgroup also of $SL(n)$. Both $O(n)$ and $SO(n)$ are non-Abelian groups, except when $n = 2$.

Not every set with an operation forms a group. For example, consider $\mathbb{R}^3$ and the vector cross product $\times$. The cross product is clearly a binary operation on $\mathbb{R}^3$ that is closed. It fails to be associative, however, and there is no neutral element or inverses for this operation. Similarly, if we extend the $GL(n)$ to include all $n \times n$ matrices, then inverses do not exist for all such matrices.

1.3.1 Semi-groups

By removing the requirements of having inverses, and a neutral element, but keeping the associative operation that is closed in the set, we have a semi-group. Although this concept is by far not as useful as a proper group, there are still has some applications where it appears.

Examples

Here are some common examples of semi-groups:

- The set of positive real numbers (with or without the number zero) is a semi-group together with the operation of addition.
- The set of positive functions form a semi-group under convolution.
- The set of $n \times n$ matrices form a semi-group under matrix multiplication.

The vector cross product in $\mathbb{R}^3$ is not a semi-group since the operation in not associative.

Chapter 2

Linear algebra

2.1 Vector spaces

The definition of a vector space implies that it consists of two sets, the set of vectors V, and the scalar field F, together with two operations: vector addition and scalar multiplication. A *field* is any set of "numbers", or scalars, for which there exist an addition and a multiplication operation that behave in the usual way. In this context, "the usual way" refers to how these operations work for the real numbers, for example in terms of associativity and distributivity. Hence, $F = \mathbb{R}$ defines a very common type of vector space, a *real vector space*. The most prototypical real vector space is $\mathbb{R}^n$ presented in Section 2.2. It is also not uncommon to consider the case when $F = \mathbb{C}$, a *complex vector space*, such as $\mathbb{C}^n$. More exotic possibilities for the scalar field includes quaternions. We will not consider these types of vector spaces here, but they may appear in more advanced topics. In this presentation, the scalars are always denoted by lower-case italicized letters, e.g., a or x.

The set of *vectors* together with the operation of vector addition must form an Abelian group. This means that there is a unique neutral element, the *zero vector*, denoted $\mathbf{0}$. In this presentation, vectors are always denoted by a lower-case bold face letter, e.g., $\mathbf{a}$ or $\mathbf{v}$, to distinguish them from scalars.

The operation of scalar multiplication, finally, maps a scalar and a vector to a vector. In the same way as vector addition is a distinct operation from adding scalars, scalar multiplication is a distinct operation from multiplication on the scalars field. In the case of a real or a complex vector space, however, we can in practice treat these distinct operations interchangeably when they are combined, in the sense that they must have the same properties as addition and multiplication have in a field, for example in terms of associativity and distributivity.

The above definition of a vector space is very general, and allows us to construct vector spaces that are not very useful other than as curiosities. To make practical use of a vector space, we need to introduce additional structures and properties. Instead of presenting these concepts in a general setting, they are introduced in Section 2.2 in the context of $\mathbb{R}^n$, the set of ordered n-tuples of real numbers, acting as a model vector space for which the additional concepts can described in a very concrete way. $\mathbb{R}^n$ is the principal vector space of this presentation, but we will also consider geometrical vector spaces in Section 2.7.1. Extensions to complex vector spaces appear in more advanced topics.

Subspace

A *subspace* S of a vector space V is any set of vectors in V that, by itself, form a vector space over the same scalar field and using the same vector addition and scalar multiplication as V does. We denote this as $S \subset V$ and refer to V as an *embedding space* of S. This implies that the vector sum of any two vectors in S, as well as the multiplication of any scalar onto any vector in S is, again, a vector in S. This definition means that V is a subspace of itself, and also that $\mathbf{0}$ is an element of all its subspaces. The *trivial subspace* of V consists of only $\mathbf{0}$, it is the smallest subspace of V. When there exists a vector in V not found in $S \subset V$, we refer to S as a *proper subspace* of V.

2.2 $\mathbb{R}^n$ as a vector space

$\mathbb{R}^n$ is the real vector space of ordered n-tuples of real numbers. As a graphical representation of an element in $\mathbb{R}^n$ we will often use a column of n real numbers, although this choice is arbitrary. Two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ are then represented as

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}, \quad \text{where} \quad a_k, b_k \in \mathbb{R}, \quad k = 1, \ldots, n. \tag{2.1}$$

$\mathbb{R}^n$ becomes a vector space by introducing an operation of vector addition that, for the case of $\mathbf{a}$ and $\mathbf{b}$ in Equation (2.1), is defined as

$$\mathbf{a} + \mathbf{b} = \begin{pmatrix} a_1 + b_1 \\ a_2 + b_2 \\ \vdots \\ a_n + b_n \end{pmatrix}. \tag{2.2}$$

Furthermore, with $c \in \mathbb{R}$ we define the scalar multiplication between c and, for example, $\mathbf{a}$ as the element in $\mathbb{R}^n$ given by

$$c\,\mathbf{a} = \begin{pmatrix} c\,a_1 \\ c\,a_2 \\ \vdots \\ c\,a_n \end{pmatrix}. \tag{2.3}$$

$\mathbb{R}^n$ is the principal vector space of this presentation, and unless explicitly stated otherwise, a *vector* refers to an element of the vector space $\mathbb{R}^n$. Also, a *scalar* refers to a real number.

In the rest of this section we present concepts that apply specifically to the vector space $\mathbb{R}^n$, but most of them can be generalized more or less straight-forward to more general vector spaces. In particular the concept of a scalar product, discussed in Section 2.2.3 can be given a more general definition.

2.2.1 Linear independence and span

Given a collection of m vectors $\mathbf{a}_1, \ldots, \mathbf{a}_m \in \mathbb{R}^n$ and m scalars $c_1, \ldots, c_m$ we define their *linear combination* as

$$c_1\mathbf{a}_1 + \ldots + c_m\mathbf{a}_m = \sum_{k=1}^{n} c_k\mathbf{a}_k. \tag{2.4}$$

If these m vectors are chosen such that we can find scalars $c_1, \ldots, c_m$, all not zero, such that the corresponding linear combination vanish

$$c_1\mathbf{a}_1 + \ldots + c_m\mathbf{a}_m = \mathbf{0}, \tag{2.5}$$

then these vectors are *linearly dependent*. Otherwise, they are *linearly independent*, i.e., Equation (2.5) is true only for coefficients $c_1 = \ldots = c_m = 0$. An equivalent definition is to say that a collection of vectors is linearly independent if it is not possible to write any of its vectors as a linear combination of the other.

Given a subset $S \subset \mathbb{R}^n$, and a collection C of m vectors $\mathbf{a}_1, \ldots, \mathbf{a}_m \in \mathbb{R}^n$, C is said to *span* S if any $\mathbf{s} \in S$ can be written as a linear combination of the vectors in C. More generally, the *linear span* of C, denoted $\text{span}(C)$, is defined as the set of all possible linear combinations formed between the vectors in C and all possible sets of m scalars:

$$\text{span}(C) = \{c_1\mathbf{a}_1 + \ldots + c_m\mathbf{a}_m \,:\, c_1, \ldots, c_m \in \mathbb{R}\}. \tag{2.6}$$

$\text{span}(C)$ is a subspace of $\mathbb{R}^n$, and an equivalent definition is to say that $\text{span}(C)$ is the smallest subspace of $\mathbb{R}^n$ that contains all vectors in C. When C spans a set S it follows that $S \subset \text{span}(C)$.

2.2.2 Bases and coordinates

A collection of n vectors $\mathbf{e}_1, \ldots, \mathbf{e}_n$ forms a *basis* of $\mathbb{R}^n$ if they span $\mathbb{R}^n$ and are linearly independent. A consequence of this definition is that any vector $\mathbf{a} \in \mathbb{R}^n$ can be written as a linear combination of the basis vectors:

$$\mathbf{a} = c_1\mathbf{e}_1 + \ldots + c_n\mathbf{e}_n, \tag{2.7}$$

for some set of scalars $c_1, \ldots, c_n$. Furthermore, these scalars are unique and we refer to them as the *coordinates* of $\mathbf{a}$ relative to the basis. The *dimensionality* of a vector space V, denoted $\dim(V)$, is the number of vectors in a basis of the space. Any basis of $\mathbb{R}^n$ consists of n basis vectors, and so $\mathbb{R}^n$ is an n-dimensional vector space.

An important observation we can make from the previous definition is that there are infinitely many choices of a basis for $\mathbb{R}^n$. As a vector space, $\mathbb{R}^n$ has no particular basis that is a better basis than any other. A vector $\mathbf{v}$ has a specific set of coordinates relative to a specific choice of basis, and they will changes if we instead choose another basis. Given the underlying structure of $\mathbb{R}^n$ in terms of n-tuples of real numbers, however, we can see that there is a *canonical basis*:

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \quad \ldots \quad \mathbf{e}_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}, \tag{2.8}$$

that has the unique property of producing coordinates of the vector $\mathbf{a}$ in Equation (2.1), that are equal to the elements of $\mathbf{a}$:

$$\mathbf{a} = a_1\mathbf{e}_1 + a_2\mathbf{e}_2 + \ldots + a_n\mathbf{e}_n. \tag{2.9}$$

This property means that the canonical basis of $\mathbb{R}^n$ in some cases is useful for the analysis of some problem, or for computing certain result. However, unless explicitly stated otherwise, a basis for $\mathbb{R}^n$ can be any collection of linearly independent vectors that also spans $\mathbb{R}^n$.

Subspace basis

Let S be a subspace of $\mathbb{R}^n$, and let B be a collection of vectors that forms a basis of S, i.e., they are linearly independent and span S. In this context, when S is a subspace of $\mathbb{R}^n$, we refer to B as a *subspace basis* to indicate that it is not necessarily a basis of the embedding space $\mathbb{R}^n$.

2.2.3 Scalar product and norm

$\mathbb{R}^n$ has an additional operation of a *scalar product*, defined as

$$\mathbf{a} \cdot \mathbf{b} = a_1b_1 + \ldots + a_nb_n = \sum_{k=1}^{n} a_kb_k, \tag{2.10}$$

for vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ described by Equation (2.1). This product is also known as an *inner product* or a *dot product*. Notice the distinction between the scalar multiplication operation, that combines a scalar and a vector to form a vector, and the scalar product operation, that combines two vectors to form a scalar. The scalar product is symmetric:

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a}, \tag{2.11}$$

and also bi-linear:

$$(\alpha_1\mathbf{a}_1 + \alpha_2\mathbf{a}_2) \cdot (\beta_1\mathbf{b}_1 + \beta_2\mathbf{b}_2) = \alpha_1\beta_1 (\mathbf{a}_1 \cdot \mathbf{b}_1) + \alpha_1\beta_2 (\mathbf{a}_1 \cdot \mathbf{b}_2) + \alpha_2\beta_1 (\mathbf{a}_2 \cdot \mathbf{b}_1) + \alpha_2\beta_2 (\mathbf{a}_2 \cdot \mathbf{b}_2). \tag{2.12}$$

The scalar product defines a *norm* that allows us to make an algebraic formulation of what we mean with the length of a vector $\mathbf{a} \in \mathbb{R}^n$. The norm is defined as

$$\|\mathbf{a}\| = (\mathbf{a} \cdot \mathbf{a})^{\frac{1}{2}} = \sqrt{a_1^2 + \ldots + a_n^2}. \tag{2.13}$$

and satisfies the *triangle inequality*:

$$\|\mathbf{a} + \mathbf{b}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\|. \tag{2.14}$$

In the following presentation, there will be several notions of normalization and normalized vectors but when used without any other specifiers, a vector $\mathbf{a}$ is referred to as *normalized* when $\|\mathbf{a}\| = 1$. Often, normalized vectors are denoted by a "hat" above the vector, as in $\hat{\mathbf{a}}$. The set of normalized vectors in $\mathbb{R}^n$ form the *unit sphere* in that vector space. The unit sphere in $\mathbb{R}^n$ is denoted S^{n-1}. All vectors $\mathbf{x}$ for which $\|\mathbf{x}\| \leq 1$ form the *unit ball* in $\mathbb{R}^3$. Hence, the difference between a sphere and a ball is that the sphere contains only the points on the "surface" of the sphere, while the ball also contains its interior points.

Both the scalar product and the norm in $\mathbb{R}^n$ can be given a more general definition than what is presented here. In particular, the norm can be given alternative definitions. The norm described in Equation (2.13) is referred to as the l^2-norm or simply 2-norm of $\mathbb{R}^n$, and to distinguish this norm from other norms, we can denote the 2-norm of $\mathbf{a} \in \mathbb{R}^n$ as $\|\mathbf{a}\|_2$. Alternative norms that appear frequently are

$$\|\mathbf{a}\|_1 = |a_1| + \ldots + |a_n| \qquad \text{: the 1-norm} \tag{2.15}$$

$$\|\mathbf{a}\|_p = \left(|a_1|^p + \ldots + |a_n|^p\right)^{1/p} \qquad \text{: the p-norm} \tag{2.16}$$

$$\|\mathbf{a}\|_\infty = \max(|a_1|, \ldots, |a_n|) \qquad \text{: the max-norm} \tag{2.17}$$

In the following presentation, the norm $\|\mathbf{a}\|$ without any specifier always refer to the 2-norm.

2.2.4 Angles and orthogonality

From the previous definitions follow that

$$-1 \leq \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \cdot \|\mathbf{b}\|} \leq 1, \tag{2.18}$$

for two non-zero vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$. We can describe angles in $\mathbb{R}^n$ that are consistent with the usual concept of angles in Euclidean spaces by defining cosine of an angle α between the two vectors as

$$\cos\alpha = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \cdot \|\mathbf{b}\|} \quad \text{and } 0 \leq \alpha \leq 180^\circ. \tag{2.19}$$

Two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ are *orthogonal* when their scalar product vanishes: $\mathbf{a} \cdot \mathbf{b} = 0$, i.e., the angle between them is 90°. Two subspaces $S_1, S_2 \subset \mathbb{R}^n$ are *orthogonal* if every vector in S_1 is orthogonal to any vector in S_2. We denote orthogonal vectors $\mathbf{a}$ and $\mathbf{b}$ as $\mathbf{a} \perp \mathbf{b}$, and orthogonal subspaces S_1 and S_2 as $S_1 \perp S_2$. An *orthogonal collection* of vectors C is a collection where any pair of distinct vectors in C are orthogonal. Consequently, an *orthogonal basis* is a basis that also forms an orthogonal collection of vectors. Furthermore, if the basis vectors in addition are normalized, the basis is *orthonormal*, or an *ON-basis* for short.

Orthonormal bases are special in the sense that *only for these bases* can we obtain the coordinates of a vector by means of scalar products with the basis vectors. For an ON-basis $\{\hat{\mathbf{e}}_k, k = 1, \ldots, n\}$ of $\mathbb{R}^n$, this means that any vector $\mathbf{a} \in \mathbb{R}^n$ can be expanded as

$$\mathbf{a} = c_1\hat{\mathbf{e}}_1 + \ldots + c_n\hat{\mathbf{e}}_n \quad \text{where} \quad c_k = \hat{\mathbf{e}}_k \cdot \mathbf{a}. \tag{2.20}$$

Given a subspace $S \subset \mathbb{R}^n$, its *orthogonal complement*, denoted $S_\perp$, is the set of all vectors in $\mathbb{R}^n$ that are orthogonal to S. It follows that also $S_\perp$ is a subspace of $\mathbb{R}^n$ and $S \perp S_\perp$. Furthermore, any $\mathbf{v} \in \mathbb{R}^n$ can be uniquely decomposed as $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_0$, where $\mathbf{v}_1 \in S$ and $\mathbf{v}_0 \in S_\perp$. In this case, the norm of $\mathbf{v}$ is given as

$$\|\mathbf{v}\|^2 = \|\mathbf{v}_1\|^2 + \|\mathbf{v}_0\|^2. \tag{2.21}$$

Trivially, the orthogonal complement of $S_\perp$ is S.

2.2.5 Linear transformations

A *linear transformation* or *linear mapping* from $\mathbb{R}^n$ to $\mathbb{R}^m$ is a function $f : \mathbb{R}^n \to \mathbb{R}^m$, such that

$$f(\lambda_1\mathbf{u}_1 + \lambda_2\mathbf{u}_2) = \lambda_1 f(\mathbf{u}_1) + \lambda_2 f(\mathbf{u}_2), \tag{2.22}$$

for all $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^n$ and scalars λ_1, λ_2.

The set of linear transformations $\mathbb{R}^n \to \mathbb{R}^m$ forms a vector space, here denoted $\mathbb{R}^{m\times n}$. For $f_1, f_2 \in \mathbb{R}^{m\times n}$, their vector sum is the mapping $f_1 + f_2 \in \mathbb{R}^{m\times n}$ defined by

$$(f_1 + f_2)(\mathbf{u}) = f_1(\mathbf{u}) + f_2(\mathbf{u}), \tag{2.23}$$

for any $\mathbf{u} \in \mathbb{R}^n$. Furthermore, the multiplication of $c \in \mathbb{R}$ by the linear transformation $f \in \mathbb{R}^{m\times n}$ is the mapping $(cf) \in \mathbb{R}^{m\times n}$ defined by

$$(cf)(\mathbf{u}) = c(f(\mathbf{u})). \tag{2.24}$$

We define the *range* of the linear transformation $f \in \mathbb{R}^{m\times n}$, as the image of its domain:

$$\text{Range}(f) = \{f(\mathbf{u}), \mathbf{u} \in \mathbb{R}^n\}. \tag{2.25}$$

It follows that $\text{Range}(f)$ is a subspace of $\mathbb{R}^m$, and $\dim(\text{Range}(f)) \leq n$. Furthermore, we define the *null space* of f as the set of vectors in $\mathbb{R}^n$ that vanish when f is applied to them:

$$\text{Null}(f) = \{\mathbf{u} \in \mathbb{R}^n : f(\mathbf{u}) = \mathbf{0}\}. \tag{2.26}$$

It follows that $\text{Null}(f)$ is a subspace of $\mathbb{R}^n$, and $\dim(\text{Null}(f)) \leq n$. A vector $\mathbf{n} \in \text{Null}(f)$ is called a *null vector* of f. $\text{Range}(f)$ and $\text{Null}(f)$ are subspaces of $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively, which in general are distinct. The two subspaces are, however, related by the *rank-nullity theorem*:

$$\dim(\text{Range}(f)) + \dim(\text{Null}(f)) = n. \tag{2.27}$$

Intuitively, this means that for a fixed n each dimension of $\text{Null}(f)$ removes a dimension of $\text{Range}(f)$, i.e., the range is as large as possible when the null space is the trivial subspace, consisting only of the zero vector $\mathbf{0} \in \mathbb{R}^n$. This relation between range and null space of a linear transformation f is relevant, e.g., when we want to determine an inverse of f.

A particular linear mapping $f \in \mathbb{R}^{m\times n}$ can be given an convenient representation in terms of a set of scalars in accordance with the following procedure. Let $\mathbf{e}_1, \ldots, \mathbf{e}_n$ be the canonical basis in $\mathbb{R}^n$. Applying f onto each of these basis vectors give a set of n vectors in V:

$$f(\mathbf{e}_1) \quad f(\mathbf{e}_2) \quad \ldots \quad f(\mathbf{e}_n). \tag{2.28}$$

For $\mathbf{a} \in \mathbb{R}^n$, with elements as in Equation (2.1), it can be expanded as

$$\mathbf{a} = \sum_{j=1}^{n} a_j \mathbf{e}_j, \tag{2.29}$$

in accordance with Equation (2.9). Applying f onto $\mathbf{a}$ produces a vector $\mathbf{b} = f(\mathbf{a})$:

$$\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} = f(\mathbf{a}) = f\left(\sum_{j=1}^{n} a_j \mathbf{e}_j\right) = /\text{due to linearity: Equation (2.22)}/ = \sum_{j=1}^{n} a_j f(\mathbf{e}_j). \tag{2.30}$$

Each of the n vectors in Equation (2.28) is an element of $\mathbb{R}^m$. If we use f_{ij} to denote the i-th element of $f(\mathbf{e}_j)$ this means that b_i, the i-th element of $\mathbf{b}$, can be written as

$$b_i = \sum_{j=1}^{n} f_{ij} a_j, \quad i = 1, \ldots, m. \tag{2.31}$$

The set of scalars that we have constructed in this way, f_{ij} for $i = 1, \ldots, m$ and $j = 1, \ldots, n$, is a complete representation of the linear mapping f, referring specifically to the canonical bases of $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. As we will see in the following section, however, this scalar set can be given a more practical interpretation as the elements of a matrix.

2.3 Matrices

A set of scalars, f_{ij}, for $i = 1,\ldots,m$ and $j = 1,\ldots,n$, can conveniently be arranged as a two-dimensional array of real numbers, a *matrix*. Graphically, a matrix is represented as a table of numbers, with $m \times n$ entries, for example as

$$\mathbf{F} = \begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1n} \\ f_{21} & f_{22} & \cdots & f_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ f_{m1} & f_{m2} & \cdots & f_{mn} \end{pmatrix}. \tag{2.32}$$

In this presentation, a matrix is always denoted by a upper-case bold letter, for example as $\mathbf{F}$. With $\mathbf{F}$ as the matrix defined by the last expression, we denote the element of $\mathbf{F}$ at row i and column j as $[\mathbf{F}]_{ij}$. Notice that the first index of $\mathbf{F}$ refers to the row and the second index to the column.

Matrices can be multiplied and there are, in fact, different ways to introduce a multiplication operation on matrices. In this presentation, we are interested in what is known as the *matrix product*. In terms of the matrix product, an $m_1 \times n$ matrix $\mathbf{A}$ can be multiplied with an $n \times m_2$ matrix $\mathbf{B}$, to form an $m_1 \times m_2$ matrix, denoted $\mathbf{A}\,\mathbf{B}$. The elements of this matrix are given by

$$[\mathbf{A}\,\mathbf{B}]_{ij} = \sum_{k=1}^{n} [\mathbf{A}]_{ik}[\mathbf{B}]_{kj}. \tag{2.33}$$

We notice that this expression implies that $[\mathbf{A}\,\mathbf{B}]_{ij}$ equals the scalar product of the i-th row of $\mathbf{A}$ with the j-th column of $\mathbf{B}$. The matrix product introduces a very useful operation on matrices, even if it is only defined for specific sizes of matrices, since it is not possible to multiply matrices of arbitrary sizes using the matrix product. The matrix product is associative: for matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ of suitable sizes it is the case that $\mathbf{A}\,(\mathbf{B}\,\mathbf{C}) = (\mathbf{A}\,\mathbf{B})\,\mathbf{C}$.

2.3.1 Matrices and linear transformations

In Section 2.2.5 we defined linear transformations between $\mathbb{R}^n$ and $\mathbb{R}^m$, and also introduced the canonical representation of such a transformation f in terms of a set of scalars f_{ij}. We will now make the formal connection between these linear transformations and matrices. By representing a vector $\mathbf{v} \in \mathbb{R}^n$ as an $n \times 1$ matrix, a *column vector*, in accordance with the graphical representation already established, it follows that a linear transformation f onto $\mathbf{v}$ is given as the matrix product $\mathbf{F}\,\mathbf{v}$, where $\mathbf{F}$ is the matrix defined in Equation (2.32). Consequently, any linear transformation in $\mathbb{R}^{m\times n}$ can be represented as an $m \times n$ matrix, with elements given by the canonical scalars described in Section 2.2.5. Also, and vice versa, any $m \times n$ matrix represents a linear transformation in $\mathbb{R}^{m\times n}$. Due to this correspondence, we can identify the set of linear transformations in $\mathbb{R}^{m\times n}$ with the set of $m \times n$ matrices. This means that when in the following we refer to a specific $m \times n$ matrix, we also refer to a specific linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$. Furthermore, we can use $\mathbb{R}^{m\times n}$ to denote also the set of $m \times n$ matrices. Once we have made this identification, it make sense to also identify vectors in $\mathbb{R}^n$ as $n \times 1$ matrices. This is not a strict identification, in the sense that we will sometime use a $1 \times n$ matrix, a *row vector*, to describe a vector in $\mathbb{R}^n$, especially in inline text.

The correspondence between linear transformations $\mathbb{R}^n \to \mathbb{R}^m$ and $m \times n$ matrices implies that we can define the range and null space of an $m \times n$ matrix $\mathbf{F}$ such that they correspond to the range and null space of the corresponding linear transformation in $\mathbb{R}^{m\times n}$. For example, with a matrix $\mathbf{F}$ that corresponds for a linear transformation f: $\text{Range}(\mathbf{F}) = \text{Range}(f)$ and $\text{Null}(\mathbf{F}) = \text{Null}(f)$. The range of $\mathbf{F}$ can equivalently be thought of as the subspace of $\mathbb{R}^m$ spanned by the columns of $\mathbf{F}$. Consequently, the range of $\mathbf{F}$) is sometimes also referred to as the *column space* of $\mathbf{F}$. Similarly, the *row space* of $\mathbf{F}$ is the subspace of $\mathbb{R}^n$ spanned by the rows of $\mathbf{F}$. In accordance with the previous discussion, the row space of $\mathbf{F}$ equals the orthogonal complement of the null space of $\mathbf{F}$.

Vectorization of matrices

Section 2.2.5 established the fact that the set of linear transformations in $\mathbb{R}^{m\times n}$ forms a vector spaces. Since we now have identified $\mathbb{R}^{m\times n}$ with the set of $m \times n$ matrices, this means that these matrices form a vector space. More trivially, the vector space character of $m \times n$ matrices follows directly if we *reshape* such a matrix into an mn element column vector in $\mathbb{R}^{mn}$. For example, this can be done by stacking each column of the matrix on top of the

next following column, from left to right. The exact implementation of the rearrangement is not important, as long as we stick to one and the same procedure each time. Once determined, such a rearrangement can also be inverted and allows us to map a vector in $\mathbb{R}^{mn}$ back to a matrix in $\mathbb{R}^{m\times n}$. With this correspondence between matrices in $\mathbb{R}^{m\times n}$ and vectors in $\mathbb{R}^{mn}$, we can use the ordinary algebraic structures defined for the latter vector space, described in Section 2.2, also for matrices.

2.3.2 Usual matrix forms

Here, we use the term *matrix form* to mean a set of matrices that have in common that specific elements are set to zero, while the other elements can have arbitrary values. The two most common matrix forms are described below.

Diagonal matrices

A special case of matrices are the *diagonal matrices* which have all elements set to zero, except along the diagonal where the elements can have arbitrary values. For a diagonal $m \times n$ matrix $\mathbf{D}$ this means that

$$[\mathbf{D}]_{ij} = \begin{cases} d_i & i = j \\ 0 & \text{otherwise} \end{cases}, \quad d_i \in \mathbb{R}, \quad i = 1, \ldots, \min(m,n). \tag{2.34}$$

This definition does not require $\mathbf{D}$ to be a square matrix, but rather than it has non-zero elements only in the diagonal. Square diagonal $n \times n$ matrices form a Abelian semi-group under matrix multiplication: the product of two such diagonal matrices is again a diagonal matrix, and they commute under matrix multiplication.

A common operation involving a diagonal matrix $\mathbf{D}$ is to multiply $\mathbf{D}$ from left or from right onto another matrix. The result of this operation ca be summarized as follows:

- $\mathbf{D}\,\mathbf{M}$ is equal to $\mathbf{M}$, except that each each *row* of $\mathbf{M}$ is multiplied by the corresponding diagonal element in $\mathbf{D}$.
- $\mathbf{M}\,\mathbf{D}$ is equal to $\mathbf{M}$, except that each each *column* of $\mathbf{M}$ is multiplied by the corresponding diagonal element in $\mathbf{D}$.

Triangular forms

A matrix $\mathbf{U}$ is an *upper triangular* matrix or *right triangular* matrix when it has all elements below the diagonal set to zero. In this case

$$[\mathbf{U}]_{ij} = 0, \quad i > j. \tag{2.35}$$

Similarly, a matrix $\mathbf{L}$ is a *lower triangular* matrix or a *left triangular* matrix when it has all elements above the diagonal set to zero. In this case

$$[\mathbf{L}]_{ij} = 0, \quad i < j. \tag{2.36}$$

Multiplying an upper triangular matrix with another upper triangular matrix of suitable size to make the matrix product well-defined, results again in an upper triangular matrix. This means that square $n \times n$ upper (or lower) triangular matrices is a semi-group.

2.3.3 Matrix transpose

Let $\mathbf{F}$ be an $m \times n$ matrix with elements given in accordance with Equation (2.32). The *transpose* of $\mathbf{F}$, denoted $\mathbf{F}^\top$, is then defined as the $n \times m$ matrix

$$\mathbf{F}^\top = \begin{pmatrix} f_{11} & f_{21} & \cdots & f_{m1} \\ f_{12} & f_{22} & \cdots & f_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ f_{1n} & f_{2n} & \cdots & f_{mn} \end{pmatrix}. \tag{2.37}$$

Intuitively, $\mathbf{F}^\top$ is obtained by "flipping" $\mathbf{F}$ along its diagonal.

As an immediate application of the transpose operation, we notice that two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$, both represented as column vectors, have a scalar product $\mathbf{a} \cdot \mathbf{b}$ that can be computed as a matrix product:

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^\top \mathbf{b} = \mathbf{b}^\top \mathbf{a} = \mathbf{b} \cdot \mathbf{a}. \tag{2.38}$$

In this context, it makes sense to also identify scalars with 1×1 matrices but it has to be done with some care. If, for example, we want to implement the multiplication of the scalar c onto a vector $\mathbf{u} \in \mathbb{R}^n$ in terms of a matrix product, it has to be done as $\mathbf{v}\, c$, where $\mathbf{v}$ represents an $n \times 1$ matrix and c is a 1×1 matrix. In this context, the expression $c\,\mathbf{v}$ is not compatible with the matrix product.

From the definition of the transpose operation follows immediately that if we take the transpose of $\mathbf{F}^\top$, we return again to $\mathbf{F}$:

$$(\mathbf{F}^\top)^\top = \mathbf{F}. \tag{2.39}$$

It is also a simple exercise to verify that for two matrices $\mathbf{A}$ and $\mathbf{B}$ of suitable sizes the transpose of their matrix product is given as

$$(\mathbf{A}\,\mathbf{B})^\top = \mathbf{B}^\top \mathbf{A}^\top. \tag{2.40}$$

Let $\mathbf{M}$ be an $m \times n$ matrix, i.e., its null space is a subspace of $\mathbb{R}^n$. In accordance with Section 2.2.5, Range$(\mathbf{M}^\top)$, too, is a subspace of $\mathbb{R}^n$. Any vector $\mathbf{u} \in$ Range$(\mathbf{M}^\top)$ can be written as: $\mathbf{u} = \mathbf{M}^\top \mathbf{v}$, with $\mathbf{v} \in \mathbb{R}^m$. Let $\mathbf{n} \in$ Null$(\mathbf{M})$, and consider the scalar product between $\mathbf{n}$ and $\mathbf{u}$:

$$\mathbf{u} \cdot \mathbf{n} = \mathbf{u}^\top \mathbf{n} = \mathbf{v}^\top \mathbf{M}\,\mathbf{n} = \mathbf{v}^\top \mathbf{0} = 0. \tag{2.41}$$

Consequently, any vector in Null$(\mathbf{M})$ is orthogonal to any vector in Range$(\mathbf{M}^\top)$, and the two subspaces of $\mathbb{R}^n$ are orthogonal. Hence, for any $m \times n$ matrix $\mathbf{M}$:

$$\text{Null}(\mathbf{M}) \perp \text{Range}(\mathbf{M}^\top). \tag{2.42}$$

This observation comes handy in certain discussions.

Adjoint operator

Although the transpose operation is straight-forward to implement, it is a slightly more subtle issue to describe what it means. Clearly $\mathbf{F}^\top$ is a linear transformation $\mathbb{R}^m \to \mathbb{R}^n$, i.e., it maps in the opposite direction relative to $\mathbf{F}$. In general, however, one is not the inverse of the other, so how are the two mappings $\mathbf{F}$ and $\mathbf{F}^\top$ related? Consider two vectors $\mathbf{u} \in \mathbb{R}^n$ and $\mathbf{v} \in \mathbb{R}^m$. Since the two vectors in general lie in distinct vector spaces they cannot be combined in terms of the scalar product. We can apply $\mathbf{F}$ onto $\mathbf{u}$, however, to obtain a vector in $\mathbb{R}^m$ and we can then compute the scalar product of that vector with $\mathbf{v}$. The result of this operation is

$$\mathbf{v} \cdot (\mathbf{F}\,\mathbf{u}) = \mathbf{v}^\top \mathbf{F}\,\mathbf{u}. \tag{2.43}$$

Alternatively, we can apply $\mathbf{F}^\top$ onto $\mathbf{v}$ and obtain a vector in $\mathbb{R}^n$, and then compute the scalar product of that vector with $\mathbf{u}$. This time the result is

$$\mathbf{u} \cdot (\mathbf{F}^\top \mathbf{v}) = \mathbf{u}^\top \mathbf{F}^\top \mathbf{v} = /\text{Using Equation (2.40)}/ = (\mathbf{F}\,\mathbf{u})^\top \mathbf{v} = (\mathbf{F}\,\mathbf{u}) \cdot \mathbf{v}. \tag{2.44}$$

Consequently, the scalar product in Equation (2.43) is the same as the one in Equation (2.44). In summary: we can map a vector $\mathbf{u}$ with a linear transformation $\mathbf{F}$ and then compute the scalar product of the result with some vector $\mathbf{v}$ in the image space, and this is the same as if we instead map $\mathbf{v}$ with $\mathbf{F}^\top$ and then compute the scalar product of the result with $\mathbf{u}$. Two transformations related in this way are called *adjoint operators*, and the transpose operation constructs $\mathbf{F}^\top$ as the adjoint operator of $\mathbf{F}$.

The operation of scalar product can be defined in a more general way than is illustrated in Section 2.2.3 for $\mathbb{R}^n$. Therefore, the concept of an adjoint operator is more general than the transpose operation for matrices discussed here.

2.3.4 Matrix rank

The number of linearly independent row and columns are equal for a general $m \times n$ matrix $\mathbf{M}$, and this number is referred to as the *rank* of $\mathbf{M}$, denoted $\text{rank}(\mathbf{M})$. Consequently, it must be the case that $\text{rank}(\mathbf{M}) \leq \min(m,n)$. In the case that $\text{rank}(\mathbf{M}) = \min(m,n)$ we say that $\mathbf{M}$ has *full rank* and otherwise it is *rank deficient*. If $\mathbf{M}$ has full rank and $m \geq n$ we say that $\mathbf{M}$ has *full column rank*, i.e., all columns are linearly independent. In this case, it is possible to find set of n rows that are linearly independent. If $\mathbf{M}$ has full rank and $m \leq n$, $\mathbf{M}$ it said to have *full row rank*, i.e., all rows are linearly independent. In this case, it is possible to find set of m columns that are linearly independent.

Let $\mathbf{A}$ and $\mathbf{B}$ be two matrices of sizes such that their product $\mathbf{A}\,\mathbf{B}$ is defined. Then

$$\text{rank}(\mathbf{A}\,\mathbf{B}) \leq \min(\text{rank}(\mathbf{A}), \text{rank}(\mathbf{B})), \tag{2.45}$$

with equality if and only if both $\mathbf{A}$ and $\mathbf{B}$ have full rank. For a matrix $\mathbf{F} \in \mathbb{R}^{m \times n}$, we have $\text{rank}(\mathbf{F}) = \dim(\text{Range}(\mathbf{F}))$. The rank-nullity theorem Equation (2.27) can then be reformulated as

$$\text{rank}(\mathbf{F}) + \dim(\text{Null}(\mathbf{F})) = n. \tag{2.46}$$

2.3.5 Outer product of vectors

The *outer product* of two vectors, $\mathbf{a} \in \mathbb{R}^m$ and $\mathbf{b} \in \mathbb{R}^n$, is the $m \times n$ matrix given as

$$\mathbf{a}\,\mathbf{b}^\top \tag{2.47}$$

Contrast to the inner product between two vectors, defined in Section 2.2.3, the result of the outer product is a matrix, rather than a scalar, and in general $\mathbf{a}\,\mathbf{b}^\top$ is not the same as $\mathbf{b}\,\mathbf{a}^\top$, i.e., the outer product is not commutative. Also, for the outer product the two vectors need not be of the same dimensionality.

The outer product $\mathbf{a}\,\mathbf{b}^\top$ is always of rank one, unless either of the two vectors is zero in which case the outer product is zero and has rank zero.

2.3.6 Submatrices

In some cases we are interested in describing parts of a matrix rather than the entire matrix. In particular, we may be interested in describing parts of matrix that, by themselves, form a 2-dimensional array of scalars. For a particular matrix $\mathbf{A}$ we can extract such a *submatrix* in a variety of ways. A common example is the matrix $\mathbf{A}_{ij}$ that is formed by removing row i and column j from $\mathbf{A}$. This means that if $\mathbf{A}$ is $m \times n$, then $\mathbf{A}_{ij}$ is a $(m-1) \times (n-1)$ matrix, and $i \leq m$ and $j \leq n$. Another example is $[\mathbf{A}]_{ij}$, which we can see as a 1×1 matrix consisting of the entry at row i and column j from $\mathbf{A}$. Similarly the j-th column of $\mathbf{A}$ represents a vector in $\mathbb{R}^m$ and the i-th row is a vector in $\mathbb{R}^n$.

We can also use matrices and vectors, of suitable sizes, to construct larger matrices through concatenation. For example, with $\mathbf{A}$ that is an $m \times n_1$ matrix and $\mathbf{B}$ that is an $m \times n_2$ matrix, we can form the $m \times (n_1 + n_2)$ matrix

$$\mathbf{C} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \end{pmatrix}. \tag{2.48}$$

Here, the columns of $\mathbf{A}$ are concatenated with the columns of $\mathbf{B}$, but for suitable sizes of the matrices, this concatenation can also be done over both rows and columns:

$$\mathbf{C} = \begin{pmatrix} \mathbf{A}_1 & \mathbf{B}_1 \\ \mathbf{A}_2 & \mathbf{B}_2 \end{pmatrix}. \tag{2.49}$$

Here, we assume that submatrices on the same row have the same number of rows, and submatrices on the same column have the same number of columns. This concatenation can also be done with vectors or combinations of matrices and vectors. In some cases we use a horizontal or vertical bar to separate the different submatrices to increase readability:

$$\mathbf{C} = \left(\begin{array}{c|c} \mathbf{A}_1 + \mathbf{I} & \mathbf{B}_1 \\ \hline \mathbf{A}_2 & \mathbf{B}_2 + \mathbf{I} \end{array}\right). \tag{2.50}$$

2.4 Square matrices

Now that we have discussed the general case of $m \times n$ matrices, we focus instead on the special case of square matrices, of size $n \times n$, representing linear transformations onto $\mathbb{R}^n$.

The identity matrix

The set of square matrices is not a group, since not every square matrix has an inverse, but in terms of the matrix product it has a neutral element, the *identity matrix* or *unit matrix*, denoted $\mathbf{I}$, a diagonal matrix that has all ones in the diagonal. Formally, there is one distinct identity matrix for each n, but we will use $\mathbf{I}$ to denote all of them, and rely on the context to determine which size it has. To make explicit that the identity matrix has a certain size, $n \times n$, the notation $\mathbf{I}_{n\times n}$ is used. The identity matrix is the unique neutral element of the matrix multiplication operation, i.e., it is the case that $\mathbf{I}_{m\times m}\,\mathbf{M} = \mathbf{M}\,\mathbf{I}_{n\times n} = \mathbf{M}$ for any $m \times n$ matrix $\mathbf{M}$.

An alternative form of the identity matrix is the *Kronecker delta function*, defined as

$$\delta_{ij} = \begin{cases} 1 & i = j, \\ 0 & i \neq j. \end{cases} \tag{2.51}$$

This function is well-defined for any domain, even an infinite domain, of integers i, j, although in most practical cases it is specified by the practical problem in which the function appears. The identity matrix and the Kronecker delta function are related as

$$[\mathbf{I}]_{ij} = \delta_{ij}. \tag{2.52}$$

2.4.1 Trace

The *trace* of an $n \times n$ matrix $\mathbf{M}$ is the sum of the diagonal elements:

$$\operatorname{trace}(\mathbf{M}) = \sum_{k=1}^{n} [\mathbf{M}]_{kk}. \tag{2.53}$$

The trace operation has a useful property in relation to the matrix product. For two matrices $\mathbf{A}$ and $\mathbf{B}$ such that $\mathbf{A}\,\mathbf{B}$ is a square matrix, it follows that $\mathbf{B}\,\mathbf{A}$, too, is a square matrix, and

$$\operatorname{trace}(\mathbf{A}\,\mathbf{B}) = \operatorname{trace}(\mathbf{B}\,\mathbf{A}). \tag{2.54}$$

This last relation may sometimes be mistaken to imply that we can permute matrices in an arbitrary way inside the trace operation, but this is incorrect for more than two matrices. If we take the trace of a general matrix product, we are allowed to take the first matrix in the product and put it last, and then take the trace of this new matrix product. What Equation (2.54) really mean is that the results of both these operations are equal, i.e., the trace of a matrix product has a *cyclic property* with respect to matrix multiplication. For example:

$$\operatorname{trace}(\mathbf{A}\,\mathbf{B}\,\mathbf{C}) = \operatorname{trace}(\mathbf{B}\,\mathbf{C}\,\mathbf{A}) = \operatorname{trace}(\mathbf{C}\,\mathbf{A}\,\mathbf{B}), \tag{2.55}$$

for matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$ of suitable sizes to make the expressions well-defined. In general, however, $\operatorname{trace}(\mathbf{A}\,\mathbf{B}\,\mathbf{C}) \neq \operatorname{trace}(\mathbf{A}\,\mathbf{C}\,\mathbf{B})$ even in the case that both matrix products are well-defined.

2.4.2 Determinant

The *determinant* of a square matrix is a scalar that characterizes how the corresponding linear transformation changes the volume of a parallelepiped in $\mathbb{R}^n$. The formal definition of the determinant for an $n \times n$ matrix $\mathbf{F}$, as a function of the elements of the matrix, amounts to an expression that relatively quickly becomes unmanageable for large n. For the simplest case, $n = 1$, the determinant is given as the single matrix element:

$$\det \mathbf{F} = [\mathbf{F}]_{11}. \tag{2.56}$$

For arbitrary $n > 1$, we can use a recursive formula:

$$\det \mathbf{F} = \sum_{k=1}^{n} (-1)^{k+1} [\mathbf{F}]_{1k} \det \mathbf{F}_{1k}, \tag{2.57}$$

where $\mathbf{F}_{1k}$ denotes the $(n-1)\times(n-1)$ matrix obtained from $\mathbf{F}$ by removing the first row and the k-th column. For example, for $n=2$ we get

$$\det\mathbf{F} = [\mathbf{F}]_{11}[\mathbf{F}]_{22} - [\mathbf{F}]_{12}[\mathbf{F}]_{21}, \tag{2.58}$$

and for $n=3$, the determinant is

$$\begin{aligned}\det\mathbf{F} = [\mathbf{F}]_{11}[\mathbf{F}]_{22}[\mathbf{F}]_{33} - [\mathbf{F}]_{11}[\mathbf{F}]_{23}[\mathbf{F}]_{32} - [\mathbf{F}]_{12}[\mathbf{F}]_{21}[\mathbf{F}]_{33} + \\ + [\mathbf{F}]_{12}[\mathbf{F}]_{23}[\mathbf{F}]_{31} + [\mathbf{F}]_{13}[\mathbf{F}]_{21}[\mathbf{F}]_{32} - [\mathbf{F}]_{13}[\mathbf{F}]_{22}[\mathbf{F}]_{31}.\end{aligned} \tag{2.59}$$

See also Section 2.8.3 for alternative but equivalent formulations of the determinant for the case of 3×3 matrices.

Notice that Equation (2.59) implies that each row and each column in $\mathbf{F}$ contributes with exactly one factor in each term of the determinant. It should be mentioned that the above expressions, and similar expressions formulated directly in the elements of $\mathbf{F}$, in practice may not be the most effective and numerically accurate method to compute the determinant. For example, in Section 2.4.6 it is shown how to compute the determinant of a diagonalizable matrix in terms of its eigenvalues, Equation (2.92).

Despite its somewhat complicated expression, the determinant is very useful since it describes how a linear transformation f changes the volume of a parallelepiped $P\subset\mathbb{R}^n$:

$$\text{volume of } f(P) = \det(\mathbf{F})\,\text{volume of } P \tag{2.60}$$

Here, $f(P)$ denotes the resulting parallelepiped after applying the linear transformation f on all points in $\mathbb{R}^n$, and $\mathbf{F}$ is the matrix representation of f. In particular this means that when the volume of P is >0 but the volume of $f(P)=0$, then $\det(\mathbf{F})=0$. This, in turn, implies that $\mathbf{F}$ is singular since at least some non-zero vector in $\mathbb{R}^n$ is a null vector of $\mathbf{F}$. We summarize this as:

$$\det\mathbf{F} = 0 \quad\Leftrightarrow\quad \mathbf{F}\text{ is singular.} \tag{2.61}$$

This means that we can use the determinant to determine whether or not a matrix is singular. In any practical situation, the left part of this relation must be interpreted as $\det(\mathbf{F})$ is very close to zero, where "very close" depends on the numerical resolution of the computations involved for computing the determinant. This is an important observation and is an issue in many applications that are based on solving linear equations.

The determinant has several useful properties in relation to other matrix operations:

$$\det(\mathbf{A}\,\mathbf{B}) = \det(\mathbf{A})\,\det(\mathbf{B}), \tag{2.62}$$

$$\det(\mathbf{A}^\top) = \det(\mathbf{A}), \tag{2.63}$$

$$\det(\mathbf{I}) = 1, \tag{2.64}$$

$$\det(c\,\mathbf{A}) = c^n\det(\mathbf{A}), \tag{2.65}$$

where $\mathbf{A}$ and $\mathbf{B}$ both are square $n\times n$ matrices, and c is a scalar.

Let $\mathbf{F}'$ be $\mathbf{F}$ where one of its rows or columns has been multiplied by the scalar c. This leads to $\det(\mathbf{F}') = c\,\det(\mathbf{F})$. For example, changing the sign of a column or row in $\mathbf{F}$ has the effect of changing the sign of the determinant.

Determinant of diagonal or triangular matrices

The square diagonal matrices exhibit an especially simple expression for their determinants. Given an $n\times n$ matrix $\mathbf{D}$ as in Equation (2.34), its determinant is just the product of the diagonal elements:

$$\det\mathbf{D} = d_1\cdot d_2\cdot\ldots\cdot d_n. \tag{2.66}$$

The same result is true also for square triangular forms. If $\mathbf{D}$ is either an upper or lower square triangular matrix, then its determinant is simply the product of its diagonal elements.

2.4.3 Matrix inverse

An $n \times n$ non-singular matrix $\mathbf{M}$ can be inverted, by forming its *matrix inverse* $\mathbf{M}^{-1}$, such that $\mathbf{M}^{-1}\mathbf{M} = \mathbf{M}\mathbf{M}^{-1} = \mathbf{I}$. Similar to the case of the determinant, explicit expressions for the matrix inverse in terms of the elements of $\mathbf{M}$ make little sense for large n. In the simplest case, for $n = 1$, the matrix inverse is given as the 1×1 matrix consisting of the reciprocal of the single element in $\mathbf{M}$:

$$\mathbf{M}^{-1} = \left(1/[\mathbf{M}]_{11}\right). \tag{2.67}$$

For arbitrary $n > 1$, we can use a recursive formula:

$$[\mathbf{M}^{-1}]_{ij} = (-1)^{i+j} \det \mathbf{M}_{ij} / \det \mathbf{M}, \tag{2.68}$$

where $\mathbf{M}_{ij}$ denotes the $(n-1) \times (n-1)$ matrix obtained from $\mathbf{M}$ by removing the i-th row and the j-th column. For example, for $n = 2$ we get

$$\mathbf{M}^{-1} = \frac{1}{\det \mathbf{M}} \begin{pmatrix} [\mathbf{M}]_{22} & -[\mathbf{M}]_{12} \\ -[\mathbf{M}]_{21} & [\mathbf{M}]_{11} \end{pmatrix}. \tag{2.69}$$

See also Section 2.8.4 for alternative but equivalent formulations of the matrix inverse for the case of 3×3 matrices.

It should be mentioned that the above expression, and similar expressions formulated directly in the elements of $\mathbf{M}$, in practice may not be the most effective and numerically accurate method to determine the matrix inverse. For example, in Section 2.4.6 it is shown how to compute the inverse of a diagonalizable matrix in terms of its eigensystem, Equation (2.93).

In relation to other matrix operations, the matrix inverse has a number of useful properties:

$$(\mathbf{A}^\top)^{-1} = (\mathbf{A}^{-1})^\top = /\text{more compact notation}/ = \mathbf{A}^{-T}, \tag{2.70}$$

$$\det(\mathbf{A}^{-1}) = 1/\det(\mathbf{A}), \tag{2.71}$$

$$(\mathbf{A}\,\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}, \tag{2.72}$$

$$(c\,\mathbf{A})^{-1} = \frac{1}{c}\mathbf{A}^{-1}, \tag{2.73}$$

for square matrices $\mathbf{A}, \mathbf{B}$ of equal size.

Square diagonal matrices are particularly simple to invert. With $\mathbf{D}$ an $n \times n$ diagonal and non-singular matrix, as described in Equation (2.34), its inverse $\mathbf{D}^{-1}$ is diagonal with elements given as

$$[\mathbf{D}^{-1}]_{ij} = \begin{cases} 1/d_i & i = j, \\ 0 & \text{otherwise.} \end{cases} \tag{2.74}$$

We remind again, that the matrix inverse defined here cannot be applied to matrices that are singular, and in practice this implies also matrices that are "close to singular" since the result may then be perturbed by numerical inaccuracies. It is, however, possible to define generalization of the matrix inverse both to non-square matrices and to singular matrices.

Implicit matrix inverse

As a linear transformation, $\mathbf{M}^{-1}$ can be interpreted as an explicit matrix inverse but also as a "placeholder" for the result of such an operation that can be combined with other matrices by means of the matrix product. For example, we can interpret the vector $\mathbf{c} = \mathbf{A}^{-1}\mathbf{b}$ as the result of first computing the matrix inverse of $\mathbf{A}$ and then multiply the result onto $\mathbf{b}$. In practice, however, it may be more efficient to see $\mathbf{c}$ as the result of solving the linear equation $\mathbf{A}\,\mathbf{c} = \mathbf{b}$ with respect to $\mathbf{c}$ which, for example, can be done using Gaussian elimination or other more advanced techniques for solving linear equations. This may be a more efficient way of implementing the computation of $\mathbf{c}$ than the first approach that involves an explicit matrix inverse.

2.4.4 Common types of square matrices

There are quite a few classes of matrices, and in particular of square matrices that appear frequently for analyzing problems and computing results. This section lists some common classes of square matrices that appear in this presentation.

Symmetric and anti-symmetric matrices

A square matrix $\mathbf{M}$ is *symmetric* if $\mathbf{M}^\top = \mathbf{M}$. This implies that $[\mathbf{M}]_{ij} = [\mathbf{M}]_{ji}$ for all elements in the matrix. Since the sum of two symmetric matrices is symmetric, it follows that the set of symmetric $n \times n$ matrices is a subspace of $\mathbb{R}^{n\times n}$. This subspace is here denoted $\mathrm{Sym}(n)$, and since there are $n(n+1)/2$ independent elements of a symmetric matrix, $\dim(\mathrm{Sym}(n)) = n(n+1)/2$.

A linear transformation $f : \mathbb{R}^n \to \mathbb{R}^n$ is *symmetric* if

$$\mathbf{u} \cdot f(\mathbf{v}) = \mathbf{v} \cdot f(\mathbf{u}), \tag{2.75}$$

for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$. A symmetric linear transformation f is represented by a symmetric matrix $\mathbf{F}$, and in this case it follows that

$$\mathbf{u}^\top \mathbf{F}\, \mathbf{v} = \mathbf{v}^\top \mathbf{F}\, \mathbf{u}, \tag{2.76}$$

for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$.

A square matrix $\mathbf{M}$ is *anti-symmetric* or *skew-symmetric* if $\mathbf{M}^\top = -\mathbf{M}$. Since the sum of two anti-symmetric matrices is anti-symmetric, it follows that the set of anti-symmetric $n \times n$ matrices is a subspace of $\mathbb{R}^{n\times n}$. This subspace is denoted $so(n)$, and since there are $n(n-1)/2$ independent elements of an anti-symmetric matrix, $\dim(so(n)) = n(n-1)/2$. If $\mathbf{M} \in so(n)$, then all elements in the diagonal of $\mathbf{M}$ vanish and, consequently, $\mathrm{trace}(\mathbf{M}) = 0$.

A linear transformation $f : \mathbb{R}^n \to \mathbb{R}^n$ is *anti-symmetric* if

$$\mathbf{u} \cdot f(\mathbf{v}) = -\mathbf{v} \cdot f(\mathbf{u}), \tag{2.77}$$

for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$. An anti-symmetric linear transformation f is represented by an anti-symmetric matrix $\mathbf{F}$, and in this case it follows that

$$\mathbf{u}^\top \mathbf{F}\, \mathbf{v} = -\mathbf{v}^\top \mathbf{F}\, \mathbf{u}, \tag{2.78}$$

for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$.

A general square matrix $\mathbf{M}$ can always be uniquely decomposed into a sum of a symmetric and an anti-symmetric matrix:

$$\mathbf{M} = \mathbf{S} + \mathbf{A}, \tag{2.79}$$

where $\mathbf{S} = \frac{1}{2}(\mathbf{M} + \mathbf{M}^\top)$ is symmetric and $\mathbf{A} = \frac{1}{2}(\mathbf{M} - \mathbf{M}^\top)$ is anti-symmetric.

General linear transformations

The set of square and non-singular matrices, i.e., matrices which inverses exist, in combination with matrix multiplication form the group of *general linear transformations*, denoted $GL(n)$.

The set of $n \times n$ matrices with determinant $+1$ is a group under matrix multiplication, denoted $SL(n)$. It is a subgroup of $GL(n)$.

Orthogonal matrices

An *orthogonal matrix* $\mathbf{M}$ satisfies $\mathbf{M}^\top \mathbf{M} = \mathbf{I}$. This implies that $\mathbf{M}^\top = \mathbf{M}^{-1}$, and $\mathbf{M}\mathbf{M}^\top = \mathbf{I}$, and also that $\det \mathbf{M} = \pm 1$. The set of $n \times n$ orthogonal matrices forms a group under matrix multiplication, he *orthogonal group*, denoted $O(n)$. $O(n)$ is a subgroup of both $GL(n)$ and consists of two separate components, one containing orthogonal matrices with determinant $= +1$, and one component containing orthogonal matrices with determinant $= -1$. They are separate in the sense that it is not possible to find a path in $\mathbb{R}^{n\times n}$ from one component to the other without leaving $O(n)$. The columns of $\mathbf{M} \in O(n)$ form a ON-basis of $\mathbb{R}^n$, this is the case also for the rows of $\mathbf{M}$.

An orthogonal transformation $f : \mathbb{R}^n \to \mathbb{R}^n$ is an orthogonal transformation if $\|f(\mathbf{v})\| = \|\mathbf{v}\|$ for all $\mathbf{v} \in \mathbb{R}^n$. An orthogonal transformation is represented by an orthogonal matrix.

A useful property that follows immediately for $\mathbf{M} \in O(n)$, as a consequence of $\mathbf{M}^\top \mathbf{M} = \mathbf{I}$, is that the magnitudes of the elements of $\mathbf{M}$ are restricted: $-1 \leq [\mathbf{M}]_{ij} \leq 1$, for $i, j = 1, \ldots, n$. Furthermore, if $\mathbf{M}_{ij} = \pm 1$, then all other elements in row i and column j vanish.

Rotations, SO(n)

A matrix $\mathbf{M} \in O(n)$ is a *rotation*[1] if $\det(\mathbf{M}) = 1$. The set of all rotation matrices in combination with matrix multiplication form the *special orthogonal group*, a subgroup of $O(n)$, denoted $SO(n)$. The other half of $O(n)$, where the determinant is negative, consists of rotations combined with reflections and do not form a group. $SO(n)$ is a subgroup also of $SL(n)$.

We have now introduced the subspace of $n \times n$ anti-symmetric matrices, denoted $so(n)$, and the group of rotations in $\mathbb{R}^n$, denoted $SO(n)$. The similarity in notation is not a coincidence, the two sets of matrices are indeed related although the connection belongs to a more advanced topic.

2.4.5 Basis matrix

In $\mathbb{R}^n$ we can select a basis and form an $n \times n$ matrix that holds the basis vectors in its columns. More generally, given an m-dimensional subspace $S \subset \mathbb{R}^n$ we can select a subspace basis for S, corresponding to an $n \times m$ matrix $\mathbf{E}$ that holds the basis vectors in its columns. In both cases is $\mathbf{E}$ of full column rank, it has linearly independent columns. Vice versa, if $\mathbf{E}$ its an $n \times m$ matrix of rank $m \leq n$, its columns form a basis for an m-dimensional subspace of $\mathbb{R}^n$. Consequently, in the following we will use the concept of a *basis matrix* to refer to an $n \times m$ matrix of full rank, with the specific intention of using its columns as a basis.

Let $\mathbf{E}$ be a basis matrix that holds the basis vectors $\mathbf{e}_k$ in its columns, and let $\mathbf{c} \in \mathbb{R}^m$ hold m coefficients c_k in its elements. The linear combination of the coefficients and the basis vectors is then given as

$$\mathbf{v} = c_1\mathbf{e}_1 + \ldots + c_m\mathbf{e}_m = \mathbf{E}\,\mathbf{c}. \tag{2.80}$$

Since we are discussing a basis, it follows that the elements of $\mathbf{c}$ are the coordinates of $\mathbf{v}$. Furthermore, in the case that $\mathbf{E}$ holds a basis for the entire of $\mathbb{R}^n$, it is $n \times n$, non-singular, and can be inverted. The coordinates of $\mathbf{v}$ are given as

$$\mathbf{c} = \mathbf{E}^{-1}\mathbf{v}. \tag{2.81}$$

In the case of an ON-basis of $\mathbb{R}^n$, then $\mathbf{E} \in O(n)$, i.e., $\mathbf{E}^{-1} = \mathbf{E}^\top$, and we get

$$\mathbf{c} = \mathbf{E}^\top\mathbf{v}. \tag{2.82}$$

Each element of $\mathbf{c}$, i.e., each coordinate of $\mathbf{v}$, is defined here as the scalar product between a row of $\mathbf{E}^\top$, i.e., a column of $\mathbf{E}$, i.e., a basis vector, and the vector $\mathbf{v}$. This is consistent with Equation (2.20).

2.4.6 Eigenvalue decomposition (EVD)

A square matrix $\mathbf{M}$ has an *eigenvalue* λ with a corresponding *eigenvector* $\mathbf{e}$ if

$$\mathbf{M}\,\mathbf{e} = \lambda\,\mathbf{e}. \tag{2.83}$$

Trivially this relation is satisfied for $\mathbf{e} = \mathbf{0}$, so the zero vector is excluded from the discussion about eigenvectors. We notice that if λ is an eigenvalue of $\mathbf{M}$ with a corresponding eigenvector $\mathbf{e}$, then $c\,\mathbf{e}$ is also an eigenvector with eigenvalue λ for any $c \in \mathbb{R}_{-0}$. Furthermore, if $\mathbf{e}_1$ and $\mathbf{e}_2$ are linearly independent eigenvectors with the same eigenvalue λ, then any linear combination of $\mathbf{e}_1$ and $\mathbf{e}_2$ is also an eigenvector with eigenvalue λ. Consequently, for each distinct eigenvalue λ there is a subspace of $\mathbb{R}^n$, an *eigenspace*, containing the eigenvectors corresponding to λ. The set of all eigenvalues together with their eigenspaces are referred to as the *eigensystem* of $\mathbf{M}$.

The eigenvalue and eigenvector equation Equation (2.83) can be rewritten as

$$(\mathbf{M} - \lambda\,\mathbf{I})\,\mathbf{e} = \mathbf{0}, \tag{2.84}$$

which means that $\mathbf{e}$ is a null vector of $\mathbf{M} - \lambda\,\mathbf{I}$. Consequently, any eigenvalue λ must be a root of $p_\mathbf{M}$, defined as

$$p_\mathbf{M}(\lambda) = \det(\mathbf{M} - \lambda\,\mathbf{I}). \tag{2.85}$$

[1] In some textbooks the concept of rotations applies to $O(n)$. Here, we use rotation to refer only to $SO(n)$.

The n-th order polynomial $p_{\mathbf{M}}$ is the *characteristic polynomial* of $\mathbf{M}$, and if we want to determine the eigensystem of $\mathbf{M}$, one way to start is to first determine its characteristic polynomial $p_{\mathbf{M}}$, then determine the roots of $p_{\mathbf{M}}$, and for each root determine the corresponding eigenspace.

The characteristic polynomial $p_{\mathbf{M}}$ has up to n distinct roots, and even if the coefficients of $p_{\mathbf{M}}$ are real, the roots can in general be complex, but in that case they come in complex conjugated pairs. In the case of a complex root, i.e., a complex eigenvalue, it cannot be the case that the corresponding eigenvector lies in $\mathbb{R}^n$, it must instead be an element of $\mathbb{C}^n$. We will not investigate the implications of this observation here, but return to it in more advanced topics. As a root of $p_{\mathbf{M}}$, each distinct eigenvalue λ_i of $\mathbf{M}$ has an *algebraic multiplicity* r_i such that we can write

$$p_{\mathbf{M}}(\lambda) = A\,(\lambda - \lambda_1)^{r_1} \ldots (\lambda - \lambda_m)^{r_m}. \tag{2.86}$$

Here, A is the coefficient of the n-th order term of $p_{\mathbf{M}}$, m is the number of distinct roots of $p_{\mathbf{M}}$, with $m \leq n$. With E_k denoting the eigenspace corresponding to eigenvalue λ_k, $\dim(E_k)$ is the *geometric multiplicity* of λ_k. The geometric multiplicity of an eigenvalue is related to its algebraic multiplicity, r_k, as $1 \leq \dim(E_k) \leq r_k$. While the algebraic multiplicities r_k always add to n, it may be the case that some eigenspaces do not have a dimensionality equal to r_k, and then their dimensions do not add to n, in which case $\mathbf{M}$ is said to be *defective*.

Although the concepts of algebraic and geometric multiplicities are important for eigenvalues, it is sometimes unpractical to characterize eigenvalues as being distinct and with some multiplicity. In the following presentation, we simply state that the n-th order characteristic polynomial $p_{\mathbf{M}}$ has exactly n roots, that may be distinct or not. Consequently, $\mathbf{M}$ has n eigenvalues that may be distinct or not, labeled λ_1 to λ_n. Normally we sort this sequence in descending order with respect to their values if they are real, or magnitudes if they are complex.

If $\mathbf{M}$ is not defective it is instead *diagonalizable* and this case has several interesting implications. First, if and only if $\mathbf{M}$ is diagonalizable is it possible to determine a basis for $\mathbb{R}^n$ that consists only of eigenvectors of $\mathbf{M}$. Second, let $\mathbf{E}$ be the corresponding basis matrix. It is then the case that

$$\mathbf{M}\,\mathbf{E} = \mathbf{E}\,\mathbf{D}, \tag{2.87}$$

where $\mathbf{D}$ is a diagonal $n \times n$ matrix that holds the corresponding eigenvalues in the same order as the eigenvectors in $\mathbf{E}$. This leads to

$$\mathbf{M} = \mathbf{E}\,\mathbf{D}\,\mathbf{E}^{-1}. \tag{2.88}$$

The matrix inverse of $\mathbf{E}$ is well-defined since $\mathbf{E}$ has full rank. The expression in the right-hand side of Equation (2.88) is a *eigenvalue decomposition* of $\mathbf{M}$, or *EVD* for short. It describes $\mathbf{M}$ as a product of three matrices, one related to its eigenvalues ($\mathbf{D}$) and two related to its eigenvectors ($\mathbf{E}$).

From Equation (2.88) follows immediately that

$$\mathbf{E}^{-1}\mathbf{M}\,\mathbf{E} = \mathbf{D}, \tag{2.89}$$

which means that $\mathbf{M}$ can be transformed into the diagonal matrix, *diagonalized*, by the linear transformation $\mathbf{E}$. In fact, applying the $n \times n$ diagonalizable matrix $\mathbf{M} = \mathbf{E}\,\mathbf{D}\,\mathbf{E}^{-1}$ onto $\mathbf{v} \in \mathbb{R}^n$ gives

$$\mathbf{M}\,\mathbf{v} = \mathbf{E}\,\mathbf{D}\,\mathbf{E}^{-1}\mathbf{v} \tag{2.90}$$

In the light of Section 2.4.5, we recognize $\mathbf{c} = \mathbf{E}^{-1}\mathbf{v}$ as the coordinates of $\mathbf{v}$ relative the eigenvector basis in $\mathbf{E}$. These coordinates are then transformed by the diagonal matrix $\mathbf{D}$, which is a trivial mapping: each coordinate c_k is multiplied with a corresponding eigenvalue in $\mathbf{D}$. Finally, the transformed coordinates are linearly combined with the basis vectors by the matrix multiplication with $\mathbf{E}$.

EVD and determinant

Let $\mathbf{M}$ be $n \times n$ and diagonalizable: $\mathbf{M} = \mathbf{E}\,\mathbf{D}\,\mathbf{E}^{-1}$, where $\mathbf{E}$ holds a basis of eigenvectors in its columns, and $\mathbf{D}$ is a diagonal matrix that holds the corresponding eigenvalues. Its determinant is given by

$$\det(\mathbf{M}) = \det(\mathbf{E})\,\det(\mathbf{D})\,\det(\mathbf{E}^{-1}) = \det(\mathbf{E})\,\det(\mathbf{D})\,1/\det(\mathbf{E}) = \det(\mathbf{D}). \tag{2.91}$$

We can then use Equation (2.66) to get

$$\det(\mathbf{M}) = \lambda_1 \cdot \ldots \cdot \lambda_n = \prod_{k=1}^{n} \lambda_k, \tag{2.92}$$

where $\lambda_k, k = 1, \ldots, n$, are the n eigenvalues of $\mathbf{M}$. As a consequence of this observation, we see that $\mathbf{M}$ is singular exactly when it has at least one vanishing eigenvalue.

EVD and matrix inverse

For the case that $\mathbf{M}$ is non-singular, i.e., it is diagonalizable and all its eigenvalues are non-zero and, therefore, its inverse is well-defined, we have

$$\mathbf{M}^{-1} = \mathbf{E}\,\mathbf{D}^{-1}\mathbf{E}^{-1}. \tag{2.93}$$

This implies that $\mathbf{M}$ and $\mathbf{M}^{-1}$ have all their eigenvectors in common, and if λ is an eigenvalue for some eigenvector $\mathbf{e}$ of $\mathbf{M}$, then $1/\lambda$ is the eigenvalue of $\mathbf{e}$ relative $\mathbf{M}^{-1}$.

2.4.7 The spectral theorem for symmetric matrices

If we consider eigenvalue decomposition for the special case when $\mathbf{M}$ is $n \times n$ and symmetric, the following result, known as the *spectral theorem*, can be derived:

1. It is possible to determine an orthonormal basis for $\mathbb{R}^n$, consisting of eigenvectors of $\mathbf{M}$.
2. All eigenvalues of $\mathbf{M}$ are real.

In terms of Equation (2.88), the first result implies that in we can choose $\mathbf{E}$ as an orthogonal matrix, i.e., $\mathbf{E}^{-1} = \mathbf{E}^\top$, and reformulate the equation as

$$\mathbf{M} = \mathbf{E}\,\mathbf{D}\,\mathbf{E}^\top, \tag{2.94}$$

where $\mathbf{E} \in O(n)$ holds an ON-basis of eigenvectors and $\mathbf{D}$ is real a diagonal matrix holding the corresponding eigenvalues of $\mathbf{M}$.

The formulation of the spectral theorem made here is not the only one possible. In more advanced topics it can be extended to anti-symmetric matrices, $so(n)$, and to rotation matrices, $SO(n)$.

Eigenvalues of $\mathbf{M}^\top\mathbf{M}$

Let $\mathbf{M}$ be an arbitrary $m \times n$ matrix, and consider the symmetric $n \times n$ matrix $\mathbf{M}^\top\mathbf{M}$ and its eigensystem. From the eigenvalue relation Equation (2.83) we get

$$\mathbf{M}^\top\mathbf{M}\,\mathbf{e} = \lambda\,\mathbf{e}, \quad \Rightarrow \quad \mathbf{e}^\top\mathbf{M}^\top\mathbf{M}\,\mathbf{e} = \lambda\,\mathbf{e}^\top\mathbf{e}, \quad \Rightarrow \quad \|\mathbf{M}\,\mathbf{e}\|^2 = \lambda\,\|\mathbf{e}\|^2, \tag{2.95}$$

From this follows that

$$\lambda \geq 0, \tag{2.96}$$

i.e., all eigenvalues of $\mathbf{M}^\top\mathbf{M}$ must be non-negative.

Eigensystem of the identity matrix

The $n \times n$ identity matrix $\mathbf{I}$ is a special case in terms of eigenvectors and eigenvalues: any non-zero vector $\mathbf{e} \in \mathbb{R}^n$ is an eigenvector with eigenvalue $= 1$ since $\mathbf{I}\,\mathbf{e} = 1 \cdot \mathbf{e}$. This observation leads to a couple of useful results.

First, let $\mathbf{M}$ be a square matrix with an eigenvector $\mathbf{e}$ and corresponding eigenvalue λ. Then, for any $\alpha \in \mathbb{R}$, $\mathbf{e}$ is also an eigenvector of $\mathbf{M} + \alpha\,\mathbf{I}$ with eigenvalue $\lambda + \alpha$. Consequently, adding $\alpha\,\mathbf{I}$ to a square matrix does not change its eigenvectors, it merely adds α to its eigenvalues.

Expansion of the identity matrix

It follows trivially that any vector $\mathbf{e} \in \mathbb{R}^n$ is an eigenvector of the $n \times n$ identity matrix $\mathbf{I}$, with corresponding eigenvalue 1. Let $\mathbf{E}$ be the basis matrix of an ON-basis for $\mathbb{R}^n$. It then follows that

$$\mathbf{I} = \mathbf{E}\,\mathbf{E}^\top = \hat{\mathbf{e}}_1\hat{\mathbf{e}}_1^\top + \ldots + \hat{\mathbf{e}}_n\hat{\mathbf{e}}_n^\top = \sum_{k=1}^{N} \hat{\mathbf{e}}_k\hat{\mathbf{e}}_k^\top, \tag{2.97}$$

where $\hat{\mathbf{e}}_k, k = 1, \ldots, n$, are the basis vectors of the ON-basis.

2.4.8 Quadratic forms

A symmetric $n \times n$ matrix $\mathbf{M}$ defines a function $f : \mathbb{R}^n \to \mathbb{R}$ in the following way

$$f(\mathbf{v}) = \mathbf{v}^\top \mathbf{M}\, \mathbf{v}, \tag{2.98}$$

which is referred to as a *quadratic form*, or 2-form, on $\mathbb{R}^n$. Since $\mathbf{M}$ is symmetric, it can be decomposed into the form described in Equation (2.94), leading to

$$f(\mathbf{v}) = \mathbf{v}^\top \mathbf{E}\, \mathbf{D}\, \mathbf{E}^\top \mathbf{v} = (\mathbf{E}^\top \mathbf{v})^\top \mathbf{D}\, (\mathbf{E}^\top \mathbf{v}). \tag{2.99}$$

This means that we can implement the mapping f as first applying the linear transformation $\mathbf{E}^\top$ and then computing the quadratic form defined by $\mathbf{D}$ on that resulting vector. This, in turn, means that much of the character of the quadratic form described by $\mathbf{M}$ is the same as the one described by $\mathbf{D}$, where the latter holds the eigenvalues of $\mathbf{M}$.

An important characterization of a quadratic form is its sign for general $\mathbf{v} \neq \mathbf{0}$. There are five distinct cases:

1. f is *positive definite* if $f(\mathbf{v}) > 0$ for $\mathbf{v} \neq \mathbf{0}$. This is equivalent to: all eigenvalues of $\mathbf{M}$ are positive.
2. f is *positive semi-definite* if $f(\mathbf{v}) \geq 0$ for $\mathbf{v} \neq \mathbf{0}$. This is equivalent to: all eigenvalues of $\mathbf{M}$ are non-negative.
3. f is *negative definite* if $f(\mathbf{v}) < 0$ for $\mathbf{v} \neq \mathbf{0}$. This is equivalent to: all eigenvalues of $\mathbf{M}$ are negative.
4. f is *negative semi-definite* if $f(\mathbf{v}) \leq 0$ for $\mathbf{v} \neq \mathbf{0}$. This is equivalent to: all eigenvalues of $\mathbf{M}$ are non-positive.
5. f is *indefinite* if $f(\mathbf{v})$ can be both positive and negative for $\mathbf{v} \neq \mathbf{0}$. This is equivalent to: $\mathbf{M}$ has both positive and negative eigenvalues.

Cases 1 and 3 are referred to as *elliptic* since the surface where f is constant forms an ellipsoid in $\mathbb{R}^n$. Similarly, cases 2 and 4 are referred to as *parabolic* since the surface forms a paraboloid. Case 5, finally, is referred to as *hyperbolic* since the surface forms a hyperboloid.

2.4.9 Projection operators

Consider two orthogonal subspaces S_1 and S_2 of $\mathbb{R}^n$: where $S_1 \perp S_2$, $\dim(S_1) = m_1$, $\dim(S_2) = m_2$, and $m_1 + m_2 = n$. This means that every $\mathbf{u} \in \mathbb{R}^n$ can be decomposed in a unique way as $\mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2$, where $\mathbf{u}_1 \in S_1$ and $\mathbf{u}_2 \in S_2$. Let $\mathbf{E}_1$ be an $n \times m_1$ basis matrix corresponding to an ON-basis of S_1 and $\mathbf{E}_2$ an $n \times m_2$ basis matrix corresponding to an ON-basis of S_2, leading to

$$\mathbf{u}_1 = \mathbf{E}_1 \mathbf{E}_1^\top \mathbf{u}, \quad \text{and} \quad \mathbf{u}_2 = \mathbf{E}_2 \mathbf{E}_2^\top \mathbf{u}. \tag{2.100}$$

This means that $\mathbf{P}_1 = \mathbf{E}_1 \mathbf{E}_1^\top$ acts a *projection operator* onto S_1. It takes an arbitrary vector $\mathbf{u} \in \mathbb{R}^n$ and performs an *orthogonal projection* of $\mathbf{u}$ onto S_1. Similarly, $\mathbf{P}_2 = \mathbf{E}_2 \mathbf{E}_2^\top$ is a projection operator onto S_2. Notice that a projection operator is a symmetric matrix. Furthermore, it is independent of the particular choice of ON-basis for the corresponding subspace, it only depends on the subspace itself.

Form the $n \times n$ matrix $\mathbf{E} = (\, \mathbf{E}_1\, \mathbf{E}_2\,)$ by concatenating the two basis matrices $\mathbf{E}_1$ and $\mathbf{E}_2$. $\mathbf{E}$ is then a basis matrix of an ON-basis for $\mathbb{R}^n$. It then follows that $\mathbf{P}_1 = \mathbf{E}\, \mathbf{D}_1 \mathbf{E}^\top$, where $\mathbf{D}_1$ is a diagonal matrix where the first m_1 elements are $= 1$ and the remaining m_2 elements are $= 0$. Consequently, $\mathbf{P}_1$ has m_1 eigenvalues that are $= 1$ and m_2 eigenvalues that vanish. Similarly, $\mathbf{P}_2$ has m_2 eigenvalues $= 1$ and m_1 eigenvalues that vanish.

If $\mathbf{P}$ is a projection operator onto some subspace S, then the projection operator onto $S_\perp$, the orthogonal complement of S, is given as

$$\mathbf{P}_\perp = \mathbf{I} - \mathbf{P}. \tag{2.101}$$

Based on these definitions of a projection operator, it follows that the projection operator corresponding to the one-dimensional subspace spanned by a non-zero vector $\mathbf{m}$ is given as

$$\mathbf{P} = \frac{\mathbf{m}\, \mathbf{m}^T}{\|\mathbf{m}\|^2}, \tag{2.102}$$

and the projection operator for the orthogonal complement is given as

$$\mathbf{P}_\perp = \mathbf{I} - \frac{\mathbf{m}\, \mathbf{m}^T}{\|\mathbf{m}\|^2}. \tag{2.103}$$

2.4.10 Commuting matrices

Let $\mathbf{A}$ and $\mathbf{B}$ be two $n \times n$ matrices. $\mathbf{A}$ and $\mathbf{B}$ are said to *commute* if $\mathbf{A}\,\mathbf{B} = \mathbf{B}\,\mathbf{A}$. As linear transformations, this means that it does not matter in which order $\mathbf{A}$ and $\mathbf{B}$ are applied to a vector in $\mathbb{R}^n$.

We now assume that both $\mathbf{A}$ and $\mathbf{B}$ are diagonalizable and that they share a common n-dimensional basis of eigenvectors. This means that we can write $\mathbf{A} = \mathbf{E}\,\mathbf{D}_1\mathbf{E}^{-1}$ and $\mathbf{B} = \mathbf{E}\,\mathbf{D}_2\mathbf{E}^{-1}$, where $\mathbf{E}$ holds the common basis of eigenvectors. From this follows:

$$\mathbf{A}\,\mathbf{B} = \mathbf{E}\,\mathbf{D}_1\mathbf{E}^{-1}\mathbf{E}\,\mathbf{D}_2\mathbf{E}^{-1} = \mathbf{E}\,\mathbf{D}_1\,\mathbf{D}_2\mathbf{E}^{-1} = \mathbf{E}\,\mathbf{D}_2\,\mathbf{D}_1\mathbf{E}^{-1} = \mathbf{E}\,\mathbf{D}_2\mathbf{E}^{-1}\mathbf{E}\,\mathbf{D}_1\mathbf{E}^{-1} = \mathbf{B}\,\mathbf{A} \tag{2.104}$$

which shows that sharing a basis of eigenvectors is sufficient for making $\mathbf{A}$ and $\mathbf{B}$ commute. It is, in fact, also a necessary condition for making $\mathbf{A}$ and $\mathbf{B}$ commute.

2.5 More on general matrices

Given that several useful concepts and operations are presented, we return to the case of matrices of general size to define additional features that are useful later on.

2.5.1 Frobenius scalar product and norm

In Section 2.3.1 it was concluded that the set of $m \times n$ matrices forms a real vector space which we can identify with $\mathbb{R}^{mn}$. This identification can intuitively be implemented by means of reshaping an $m \times n$ matrix to a column vector in $\mathbb{R}^{mn}$, and from there apply the usual machinery that is available for this vector space. For example, with $\mathbf{A}$ and $\mathbf{B}$ as two $m \times n$ matrices, we can compute their scalar product as

$$\mathbf{A} \cdot \mathbf{B} = \sum_{i=1}^{m}\sum_{j=1}^{n} [\mathbf{A}]_{ij}[\mathbf{B}]_{ij}. \tag{2.105}$$

Notice that this expression corresponds directly to the usual scalar product that is defined for vectors in $\mathbb{R}^{mn}$ in Section 2.2.3: as a sum of products between corresponding elements. This last expression can also be given more compact representations in terms of operations on matrices:

$$\mathbf{A} \cdot \mathbf{B} = \mathrm{trace}(\mathbf{A}^\top\mathbf{B}) = \mathrm{trace}(\mathbf{B}\,\mathbf{A}^\top) = \mathrm{trace}(\mathbf{B}^\top\mathbf{A}) = \mathrm{trace}(\mathbf{A}\,\mathbf{B}^\top) = \mathbf{B} \cdot \mathbf{A}. \tag{2.106}$$

The four different forms reflect that the scalar product on $\mathbb{R}^{mn}$ is symmetric: $\mathbf{A} \cdot \mathbf{B} = \mathbf{B} \cdot \mathbf{A}$. But they also reflect that the trace operation, when applied to a matrix product, allows a matrix at the end of a the product to be moved to the other end, see Section 2.4.1. These multiple and equivalent definitions are sometimes useful for simplifying derivations that includes scalar products of matrices. In the general case of rectangular $m \times n$ matrices, notice that the first two forms imply taking the trace of an $m \times m$ matrix, while the last two applies the trace onto an $n \times n$ matrix, and in practice the smaller size is the preferred. The scalar product defined in Equation (2.105) and Equation (2.106) is referred to as the *Frobenius scalar product* or *Frobenius inner product*.

Given the Frobenius scalar product, we can define also a norm for matrices:

$$\|\mathbf{A}\|_F = (\mathbf{A} \cdot \mathbf{A})^{1/2} = \sqrt{\mathrm{trace}(\mathbf{A}^\top\mathbf{A})}. \tag{2.107}$$

There are other types of norms for matrices that appear in the literature, and this one is referred to as the *Frobenius norm*.

A useful result in relation to the Frobenius norm is that it is invariant to orthogonal transformations:

$$\|\mathbf{A}\|_F = \|\mathbf{Q}_1\mathbf{A}\,\mathbf{Q}_2\|_F. \tag{2.108}$$

Here, $\mathbf{Q}_1 \in O(m)$ and $\mathbf{Q}_2 \in O(n)$. In the case that $\mathbf{A}$ is $n \times n$ and symmetric we can express $\mathbf{A}$ in terms of an eigenvalue decomposition: $\mathbf{A} = \mathbf{E}\,\mathbf{D}\,\mathbf{E}^\top$, where $\mathbf{E} \in O(n)$ and $\mathbf{D}$ is diagonal. This allows us to rewrite Equation (2.107) as

$$\|\mathbf{A}\|_F = \sqrt{\mathrm{trace}(\mathbf{E}\,\mathbf{D}^\top\,\mathbf{E}^\top\mathbf{E}\,\mathbf{D}\,\mathbf{E}^\top)} = \sqrt{\mathrm{trace}(\mathbf{D}^2)} = \sqrt{\lambda_1^2 + \ldots + \lambda_n^2}, \tag{2.109}$$

where $\lambda_k, k = 1, \ldots, n$, are the eigenvalues of $\mathbf{A}$. This result does not hold for a general $\mathbf{A}$ that is diagonalizable, but $\mathbf{E} \in O(n)$ is a sufficient condition for Equation (2.109) to be valid.

Example

Let $\mathbf{S} \in \mathrm{Sym}(n)$, a symmetric matrix, and $\mathbf{A} \in so(n)$, an anti-symmetric matrix. Given the Frobenius scalar product, these matrices are orthogonal:

$$\mathbf{S} \cdot \mathbf{A} = 0. \tag{2.110}$$

Consequently, the two subspaces $\mathrm{Sym}(n)$ and $so(n)$ are orthogonal in the vector space of $n \times n$ matrices. In fact, each of the two subspaces is the orthogonal complement of the other.

2.5.2 Matrix bases

Once $m \times n$ matrices have been put into the context of the vector space $\mathbb{R}^{mn}$, it is also possible to introduce a basis for this space and to determine the coordinates of a particular matrix relative to the basis. In this context, it may be the case that the matrix under consideration happens to be restricted to some subspace of matrices, e.g., symmetric or anti-symmetric matrices, and then we need a subspace basis rather than a basis for the entire embedding space. Furthermore, there are infinitely many ways to choose a basis for $\mathbb{R}^{mn}$, and from an algebraic point of view they are equally valid as bases. Depending on the problem at hand, however, one basis may be a better or more natural choice than others.

Assuming that we are dealing with general $m \times n$ matrices, representing linear transformations from $\mathbb{R}^n$ to $\mathbb{R}^m$, we can determine an arbitrary basis $\mathbf{e}_i, i = 1, \ldots, m$, for $\mathbb{R}^m$ and a basis $\mathbf{b}_j, j = 1, \ldots, n$, for $\mathbb{R}^n$. We can then form mn matrices $\mathbf{E}_{ij} \in \mathbb{R}^{m \times n}$ as

$$\mathbf{E}_{ij} = \mathbf{e}_i \mathbf{b}_j^\top, \quad i = 1, \ldots, m, \quad j = 1, \ldots, n. \tag{2.111}$$

In particular if both bases $\mathbf{e}_i$ and $\mathbf{b}_j$ are ON-bases, then $\mathbf{E}_{ij}$ forms an ON-basis of $\mathbb{R}^{m \times n}$. As a special case, we can choose the canonical basis of each space and obtain a canonical basis also for $\mathbb{R}^{m \times n}$. For example, if we consider the matrix space $\mathbb{R}^{2 \times 3}$, it has a canonical basis in terms of the 6 matrices

$$\begin{aligned} \mathbf{E}_{11} &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{E}_{12} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{E}_{13} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \\ \mathbf{E}_{21} &= \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad \mathbf{E}_{22} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \quad \mathbf{E}_{23} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \end{aligned} \tag{2.112}$$

2.6 Affine spaces

Although an affine space can be defined in a more formal manner, in terms of specific sets and operations on these sets, a more informal description of an affine space is given here. In fact, two slightly different but compatible descriptions are provided.

2.6.1 As a subset of a vector space

Given a vector space V, we have already defined a subset of $U \subset V$ as a subspace if U, itself, is a vector space over the same scalar field as V. As a consequence, it follows that U must intersect the origin of V. In various applications, however, we may be interested in describing subsets of V can be described as "displaced subspaces". This can be done by starting with a subspace $U \subset V$, and then construct the affine space A by adding a constant vector $\mathbf{v} \in V$ to each vector in U:

$$A = \{\mathbf{u} + \mathbf{v}, \mathbf{u} \in U\}. \tag{2.113}$$

The set A is a subset of V, but it is not a subspace unless $\mathbf{v} \in U$. Intuitively, we can see A as the subspace U that has been displaced by the vector $\mathbf{v}$. As an example of an affine space, take the vector space $V = \mathbb{R}^3$ and consider any 2-dimensional plane in V. Any such plane, regardless of whether it intersects the origin or not, is an affine space. This specific description of an affine does not define any internal structures, such as operations that can be applied to its elements, but still provides a useful characterization of an affine space as a subset of a vector space.

This way of constructing an affine space, by combining a subspace $U \subset V$ with a displacement vector $\mathbf{v} \in V$, does not produce a unique set A for every U and $\mathbf{v}$. Displacing U with $\mathbf{v}_1$ and $\mathbf{v}_2$ produces the same affine space A, as a subset of V, when $\mathbf{v}_1 - \mathbf{v}_2 \in U$. Given an affine space A that is produced by displacing a subspace $U \subset V$, the displacement of U can be done by any $\mathbf{a} \in A$.

Given the constriction of an affine space, described above in terms of a subset of a vector space, it should be clear that, in general, adding two elements from A, or multiplying an element of A with a scalar, does not produce an element in A. On the other hand, it should also be clear that the difference between two elements in A is always an element of the subspace U. This means that the operation of subtracting two elements in A can be given a well-defined interpretation as an element in A: as vectors in V their difference lies in U, but adding the stipulated displacement vector $\mathbf{v}$ to this difference moves it back to A.

2.6.2 As a vector space

Another possibility of introducing operations on an affine space is to construct a vector space from A. As already mentioned, A is in general not a subspace of V, so a vector space structure of A cannot be related directly to that of V, but it is possible to define a new set of operations for vector addition and scalar multiplication, indirectly from those in V. First we must choose a specific point $\mathbf{a}_0 \in A$, that is designated as the origin of the new vector space. In fact, we can refer to to this $\mathbf{a}_0$ as the displacement vector that generates A from subspace U.

Let $\mathbf{a}_1, \mathbf{a}_2 \in A$. Then the differences $\mathbf{a}_1 - \mathbf{a}_0$ and $\mathbf{a}_2 - \mathbf{a}_0$ both lie in U. Since U is a subspace, the sum of these two vectors also lies in U: $\mathbf{a}_1 + \mathbf{a}_2 - 2\,\mathbf{a}_0 \in U$. Finally, it can be transported back to A, by adding the displacement vector $\mathbf{a}_0$, producing $\mathbf{a}_1 + \mathbf{a}_2 - \mathbf{a}_0 \in A$. Similarly, to multiply $\mathbf{a} \in A$ by the scalar $s \in \mathbb{R}$, we first consider the difference $\mathbf{a} - \mathbf{a}_0 \in U$. Since this lies in the subspace U, the operation of multiplying it with the scalar s is a well-defined element of U: $s(\mathbf{a} - \mathbf{a}_0) \in U$. Finally, this element can be transported back to A, by adding the displacement vector $\mathbf{a}_0$: $s(\mathbf{a} - \mathbf{a}_0) + \mathbf{a}_0 \in A$. We can summarize these results as

$$\mathbf{a}_1 \bigoplus \mathbf{a}_2 = \mathbf{a}_1 + \mathbf{a}_2 - \mathbf{a}_0 \tag{2.114}$$

$$s \bigodot \mathbf{a} = s\,\mathbf{a} + (1-s)\,\mathbf{a}_0 \tag{2.115}$$

Here, we use $\bigoplus$ and $\bigodot$ to denote the new operations of adding elements in A, and of scalar multiplication in A, respectively. These two operations turn A into a vector space, but its vector space structure is distinct from that of V and must not be confused. The fact, that A together with $\bigoplus$ and $\bigodot$ is consistent with the properties of a vector space does not follow immediately from these definitions, but can be shown in a straight-forward way.

An important observation to make here is that $\mathbf{a}_0$ indeed corresponds to the origin of the vector space A and, consequently, the vector space structure that in defined in this way for A depends on the choice of $\mathbf{a}_0 \in A$. Two distinct choice of $\mathbf{a}_0$ generate two distinct vector spaces from A. Another observation that can be made here, it that the construction of a vector space from A only relies on the possibility of mapping differences between elements in A to some vector space U, and on the transporting elements of U back to A. It is, in fact, not necessary that U is a subspace of a larger vector space V, and that A is a subset of V. It suffices that the mapping of differences between elements in A to U, and the mapping from U back to A are consistent with the idea that both U and A could be subspaces of a vector space V. This possibility is exploited in the next section, where we consider Euclidean spaces and their relation to $\mathbb{R}^n$.

2.7 Euclidean spaces $\mathbb{E}^n$

Basic geometry is based on the idea that the world which we live in can be abstracted as a space of points that can be grouped into more complex objects, such as lines, planes, circles or general curves. On these objects we can define operations that extracts measurements of, for example, length or area, or compute a point of intersection, and determine properties such as parallel or perpendicular. Lines can be infinitely long, or we may consider a finite segment of a line that stretches between two distinct points.

We can even make the segment *directed* in the sense that it has designated start and end points. We can also transform points, or groups of points, by means of a translation specified by a certain direction and a certain distance in that direction. This transformation can be represented by a directed line segment, where the end point is the translation applied to the start point. Alternatively, we can rotate these points a certain angle about a certain point or line, or even combine translations and rotations. All these geometric constructions and operations should be well-known to the reader and can be summarized as *Euclidean geometry*, and the space in which it is defined is a *Euclidean space*.

A defining property of Euclidean geometry is that it can be implemented in terms of physical objects as they appear to our senses. We see the floor as a plane, the edge of our table as line, and the dot above the letter "i"

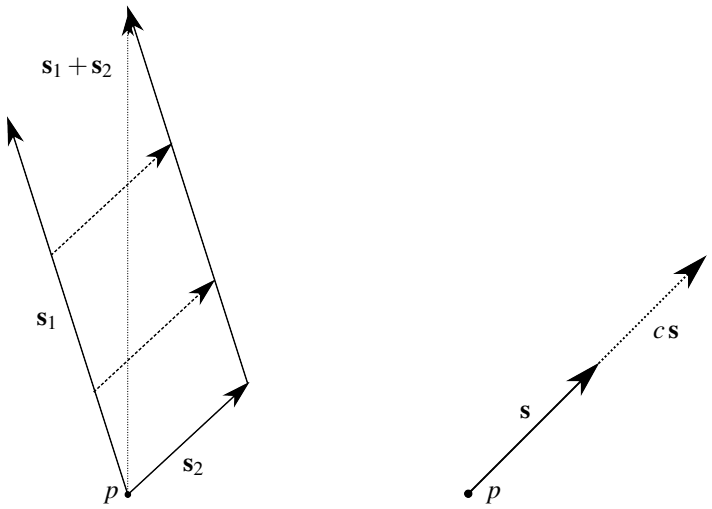

Figure 2.1: Left: Vector addition in $V_p(\mathbb{E}^n)$. Right: Scalar multiplication in $V_p(\mathbb{E}^n)$.

as a point. These objects and their geometrical relations can be described in terms of Euclidean geometry. This correspondence between our physical world and a Euclidean space is an approximation in the sense that every physical object, even a sub-atomic particle, has an extension and cannot be represented in terms of points, lines, or planes other than approximately. There is also an approximation involved in the sense that we cannot produce exact measurements of quantities such as distance or angle in the real world, they are always affected by measurement errors. Within the limits of these approximations, however, we can identify our three-dimensional physical world with the three-dimensional Euclidean space, denoted $\mathbb{E}^3$. By restricting $\mathbb{E}^3$ to a suitable plane of our choice, we obtain instead $\mathbb{E}^2$, the two-dimensional Euclidean space. In this presentation, we will only consider these two Euclidean spaces but, in principle, we can also consider Euclidean spaces of arbitrary dimension, where the usual geometrical concepts or points, lines, planes, angles, etc, have to be given an intuitive or algebraic generalization.

2.7.1 $\mathbb{E}^n$ as a vector space

A Euclidean space $\mathbb{E}^n$ is not a vector space, it is essentially just a set of point that has additional structures as defined by Euclidean geometry. We can try to think of the points in $\mathbb{E}^n$ as vectors of a vector space, but the problem is that $\mathbb{E}^n$ has no particular point that refers to the zero vector and, consequently, we cannot define in a meaningful way what it means to add a point with another point in $\mathbb{E}^n$. It is possible, however, to derive a real vector spaced from $\mathbb{E}^n$ by considering the set of directed line segments that have a common starting point $p \in \mathbb{E}^n$. This set of directed line segments is here denoted as $V_p(\mathbb{E}^n)$. The elements of $V_p(\mathbb{E}^n)$ form a vector space where vector addition of two segments, $\mathbf{s}_1$ and $\mathbf{s}_2$, is defined by translating the start and end points of $\mathbf{s}_1$ by means of $\mathbf{s}_2$, such that the starting point of the translated $\mathbf{s}_1$ coincides with the end point of $\mathbf{s}_2$. The resulting vector sum, $\mathbf{s}_1 + \mathbf{s}_2$, is then defined as the directed line segment that starts in p, and ends in the end point of $\mathbf{s}_1$ *after* it has been translated by $\mathbf{s}_2$. See Figure 2.1, left, that illustrates this vector addition. Similarly, the multiplication by $c \in \mathbb{R}$ onto a directed line segment $\mathbf{s} \in V_p(\mathbb{E}^n)$ implies moving only the end point of $\mathbf{s}$ along the corresponding line such that the distance between the start and end points of the resulting segment, $c\,\mathbf{s}$, is multiplied by c relative to the initial segment. In the case that $c < 0$, the end point is moved to the other side of the line relative the starting point, p, before the distance is scaled by $-c$. This scalar multiplication is illustrated in Figure 2.1, right. A consequence of this definition is that $V_p(\mathbb{E}^n)$ is a real n-dimensional vector space, where the zero vector is the degenerate line segment that start and ends at the same point: $p \in \mathbb{E}^n$.

Notice that the definition of $V_p(\mathbb{E}^n)$ is a purely geometric construction, we have not made use of concepts such as a basis or coordinates in this space. Also note that $V_p(\mathbb{E}^n)$ is distinct from the underlying Euclidean space $\mathbb{E}^n$: the elements of $V_p(\mathbb{E}^n)$ are directed line segments rather than points. For a particular choice of p, however, we can identify $V_p(\mathbb{E}^n)$ and $\mathbb{E}^n$ by equating the point $r \in \mathbb{E}^n$ with the vector in $V_p(\mathbb{E}^n)$ that starts at p and ends at r. This implies that for distinct choices of $p, q \in \mathbb{E}^n$ we obtain distinct derived vector spaces, $V_p(\mathbb{E}^n)$ and $V_q(\mathbb{E}^n)$.

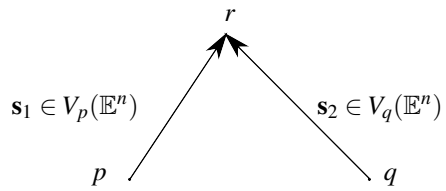

Figure 2.2: A point $q \in \mathbb{E}^n$ that is represented by two vectors $\mathbf{s}_1$ and $\mathbf{s}_2$ in two distinct vectors spaces $V_{p_1}(\mathbb{E}^n)$ and $V_{p_2}(\mathbb{E}^n)$, respectively.

Any vector in $V_p(\mathbb{E}^n)$ is a directed line segment that starts at point p and ends in some point $r \in \mathbb{E}^n$. By changing its starting point to q, we get instead a vector in $V_q(\mathbb{E}^n)$. Both vectors represent the same point $r \in \mathbb{E}^n$, given by the end points of each of the two vectors, but otherwise the two vectors are distinct as elements of distinct vector spaces. This is illustrated in Figure 2.2.

2.7.2 $\mathbb{E}^n$ and $\mathbb{R}^n$

We have seen that $\mathbb{E}^n$ can be identified with a geometrically constructed real vector space $V_p(\mathbb{E}^n)$, consisting of directed line segments originating at p. This identification is not unique since we can choose p as any point in $\mathbb{E}^n$. Despite being a vector space, $V_p(\mathbb{E}^n)$ is still a bit too abstract to be of practical use. In order to do numerical computations, it would be a considerable help if we can use the vector space $\mathbb{R}^n$ instead, where the geometrical operations that are defined in $\mathbb{E}^n$ can be given an algebraic representation that allows us to compute numerical values, for example, of distances or angles. In order to establish such a correspondence, we need to select a collection B of n elements in $V_p(\mathbb{E}^n)$, directed line segments all starting at p, that have two additional properties when seen as line segments in $\mathbb{E}^n$: they are mutually perpendicular and have unit length. The first property is completely determined from the notion of perpendicularity in $\mathbb{E}^n$. The second property, however, needs to be further specified by the notion of what exactly do we mean by *unit length*, which, in turn, depends on the length unit we are using in $\mathbb{E}^n$. For example, if we want to use meter or feet as the length unit, the vectors in B should be one meter or one feet long, respectively. Alternatively, we can select any n perpendicular vectors in $V_p(\mathbb{E}^n)$ that have the same length, and use this length as the definition of what we mean by one unit of length.

The collection B forms a basis of $V_p(\mathbb{E}^n)$, a *Cartesian basis* and the basis vectors are often referred to as *coordinate axes*. For example, in $\mathbb{E}^3$ we can choose a point of origin, p, that generates the vector space $V_p(\mathbb{E}^3)$, and in this vector space we select three vectors $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ to form a Cartesian basis B, such that they have unit length and are mutually perpendicular as line segments in $\mathbb{E}^3$. In the following, we refer to the point p together with the basis B in $V_p(\mathbb{E}^n)$ as a *frame of reference*, a *reference frame*, or a *coordinate system*.

Given a reference frame, any point $r \in \mathbb{E}^3$ can be represented by a specific vector $\mathbf{s} \in V_p(\mathbb{E}^3)$, a directed line segment in $\mathbb{E}^3$ that starts at p and ends at r, and from our intuitive notion of what $\mathbb{E}^3$ is all about, we maintain the idea that it is possible to determine unique scalars c_1, c_2, c_3 such that

$$\mathbf{s} = c_1\mathbf{e}_1 + c_2\mathbf{e}_2 + c_3\mathbf{e}_3. \tag{2.116}$$

Since B is a basis of $V_p(\mathbb{E}^3)$, the scalars c_1, c_2, c_3 are the coordinates of $\mathbf{s}$ relative to the basis vectors. These coordinates can conveniently be represented by the 3-tuple $(c_1, c_2, c_3) \in \mathbb{R}^3$. In the general case, where $\mathbb{E}^n$ is the underlying Euclidean space, any choice of a reference frame produces a set of coordinates $(c_1, \ldots, c_n)$ relative to the corresponding Cartesian basis B for any point $r \in \mathbb{E}^n$. We refer to these coordinates as the *Cartesian coordinates* of r relative to the reference frame.

Given a reference frame for $\mathbb{E}^n$, we can also map vectors in $\mathbb{R}^n$ back to points in $\mathbb{E}^n$. Any n-tuple, used as coordinates relative to the basis of the reference frame produces a vector in the corresponding vector space $V_p(\mathbb{E}^n)$, a directed line segment in $\mathbb{E}^n$, and the end point of that segment in $\mathbb{E}^n$ is the resulting point of this inverse mapping. From the construction of these two mappings, it follows that they are one-to-one, i.e., once the reference frame is determined, each point in $\mathbb{E}^n$ is represented by a unique vector in $\mathbb{R}^n$, and vice versa. Furthermore, it also follows that the operations of vector addition and scalar multiplication in $V_p(\mathbb{E}^n)$ correspond to the same operations in $\mathbb{R}^n$, and also that the n basis vectors in the reference frame are represented by the canonical basis in $\mathbb{R}^n$.

2.7.3 Concluding remarks

Why do we need to choose the basis vectors in the reference frame to be perpendicular and of unit length when seen as line segments? Can we not use any n vectors for B as long as they are linearly independent and span $V_p(\mathbb{E}^n)$? The answer is that we want the concepts of length and perpendicularity for line segments in $\mathbb{E}^n$ to correspond to the notions of vector norm and orthogonality of vectors in $\mathbb{R}^n$, as they are defined in Section 2.2.3 and Section 2.2.4. The distance between two points r_1 and r_2 in $\mathbb{E}^n$ should be equal to the vector norm of the difference between the two vectors in $\mathbb{R}^n$ that represent r_1 and r_2, respectively. And we also want two vectors in $V_p(\mathbb{E}^n)$ that are perpendicular as line segments in $\mathbb{E}^n$ to correspond to orthogonal vectors in $\mathbb{R}^n$. This correspondence between related concepts in $\mathbb{E}^n$ and $\mathbb{R}^n$ will only happen when the basis vectors of the reference frame are chosen as mutually perpendicular and are of unit length. Notice that although the concept of angles and perpendicularity is well-defined in $\mathbb{E}^2$ and $\mathbb{E}^3$, the notion of a scalar product between two vectors in $V_p(\mathbb{E}^n)$ is not really well-defined unless we first map the vectors to their correspondences in $\mathbb{R}^n$ and compute their scalar product there. This mapping and therefore also the resulting scalar product, however, depend on the choice of reference frame, so it is not possible to define a scalar product in $V_p(\mathbb{E}^n)$ that is invariant to the choice of reference frame.

In summary, for a Euclidean space $\mathbb{E}^n$ with given a reference frame it is possible to uniquely represent each point in $\mathbb{E}^n$ as a vector in $\mathbb{R}^n$. More precisely, the representation in $\mathbb{R}^n$ is in terms of coordinates of vectors in $V_p(\mathbb{E}^n)$ relative to the basis of the reference frame. This representation is heavily dependent on the choice of the reference frame. Changing either the point of origin, p, or the basis B implies that one and the same point in $\mathbb{E}^n$ is represented by another vector in $\mathbb{R}^n$. This means that when we discuss points in $\mathbb{E}^n$ as vectors in $\mathbb{R}^n$, this assumes a well-defined choice of reference frame. In most cases, we do not have to specify the reference frame, but merely assume that it exists and can be specified in its details if necessary.

For a specific Euclidean space $\mathbb{E}^n$ we can even introduce more than one reference frame, which implies that one and the same point in $\mathbb{E}^n$ has two or more representations as vectors in $\mathbb{R}^n$ depending on which reference frame the representation is derived from. Alternatively, we may think of a single reference frame for $\mathbb{E}^n$ that is not fixed. It could move, for example, along a smooth trajectory in $\mathbb{E}^n$ while at the same time its basis vectors in B vary, although always in such a way that they are mutually perpendicular and of unit length, e.g., by rotating the basis vectors in one way or another. Again, this variation of the reference frame causes the representation in $\mathbb{R}^n$ to vary as well, even if the original point in $\mathbb{E}^n$ is fixed.

2.8 What happens in $\mathbb{R}^3$ stays in $\mathbb{R}^3$

The last few sections provide us with the basis properties of vector spaces of type $\mathbb{R}^n$, for general integers $n \geq 1$. In this sections, some of these properties are given a more specific and sometimes also slightly simpler formulation for the case $n = 3$. There are also some operations presented here that are defined primarily for the case $n = 3$.

2.8.1 Handedness

When we define a reference frame for $\mathbb{E}^2$, they have in practice to be ordered such that we label one of the coordinate axes as "the first axis" and the other one as "the second axis". This implies that the reference frames of $\mathbb{E}^2$ come in two varieties: one where we can rotate the first axis 90° counter-clockwise to make it point in the same direction as the second axis, and one were this happens after a 90° clockwise rotation. Any two reference frames of the first type, two *counter-clockwise frames*, can be transformed one to the other by a suitable combination of a translation and a rotation, a *rigid transformation*. Similarly, any two reference frames of the second type, two *clockwise frames*, can be transformed one to the other by a suitable rigid transformation. A counter-clockwise reference frame can, however, not be transformed in this way to a clockwise reference frame, it is necessary to include a reflection to align one of the frames with the other. Another way to formulate the same thing: take an object O and its reflection O' relative to a line in $\mathbb{E}^2$. If O cannot be rigidly transformed into O' then the object has what it called *handedness* or *chirality*. Consequently, as soon as the coordinate axes are given a specific order, the frame has a handedness, in this case manifested as two types of "orientations" of the frames. The chirality aspect of a reference frame is not an issue until we start to discuss transformations in the space, in particular related to rotations. Depending on the orientation of the reference fame, one and the same rotation will have a slightly different representation, e.g., in terms of transformation matrices. In order to deal with these transformations in

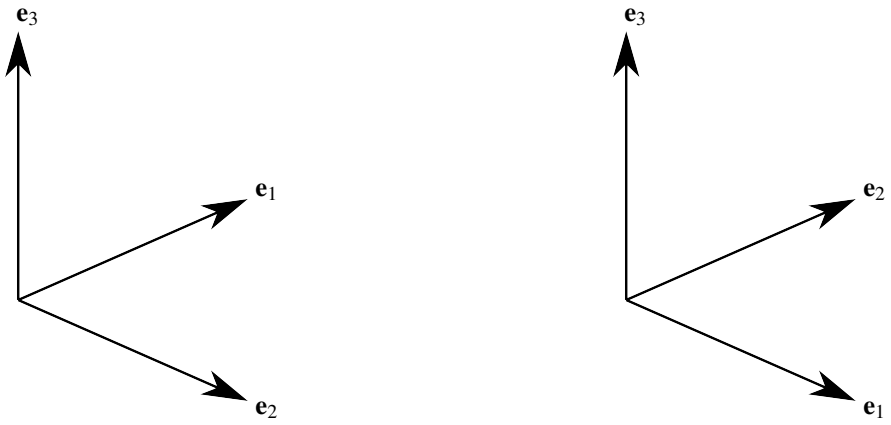

Figure 2.3: Left: a left-handed coordinate system. Right: a right-handed coordinate system.

practice, the orientation of the reference frame must be clear. Otherwise, it is possible that strange things start to happen, such as objects in $\mathbb{E}^2$ that rotate in the opposite direction compared to what you expected.

The same thing occurs in $\mathbb{E}^3$, where reflection is made relative to a plane instead of a line. Again, we get two types of orientations of a reference frame that usually are referred to as *right-handed* and *left-handed*. In a right-handed reference frame, you can place the right hand thumb pointing along the first coordinate axis, the index finger pointing along the second axis. The middle finger should then be pointing along the third axis. In a left-handed reference frame the same procedure applies but now to the fingers of the left hand. A right-handed frame cannot be rigidly transformed to a left-handed frame: we require a reflection to accomplish an alignment. The two types of reference systems are illustrated in Figure 2.3. As in $\mathbb{E}^2$, the handedness of a reference frame in $\mathbb{E}^3$ is not an issue until we start to rotate.

In the following presentation, and unless stated otherwise, we will use the "anti-clockwise" type of reference frames for $\mathbb{E}^2$, and the right-handed type for $\mathbb{E}^3$, since they are the standard types in mathematical literature. In general, pay attention when you are dealing with an externally defined reference system, it can be of any type.

2.8.2 Vector cross product in $\mathbb{R}^3$

Let $\mathbf{a}$ and $\mathbf{b}$ be two vectors in $\mathbb{R}^3$:

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}. \tag{2.117}$$

Given these two vectors, we define their *cross product*, denoted $\mathbf{a} \times \mathbf{b}$, as

$$\mathbf{a} \times \mathbf{b} = \begin{pmatrix} a_2 b_3 - a_3 b_2 \\ a_3 b_1 - a_1 b_3 \\ a_1 b_2 - a_2 b_1 \end{pmatrix}. \tag{2.118}$$

This operation on pairs of vectors in $\mathbb{R}^3$ has several useful properties, where those of immediate use in this presentation are listed below:

1. The cross product is anti-symmetric: $\mathbf{a} \times \mathbf{b} = -\mathbf{b} \times \mathbf{a}$.

2. The cross product $\mathbf{a} \times \mathbf{b}$ is always orthogonal to both $\mathbf{a}$ and $\mathbf{b}$: $(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{a} = (\mathbf{a} \times \mathbf{b}) \cdot \mathbf{b} = 0$.

3. The norm of the cross product scales with the norms of the two vectors and the sinus of the angle α between them: $\|\mathbf{a} \times \mathbf{b}\| = \|\mathbf{a}\| \cdot \|\mathbf{b}\| \cdot |\sin \alpha|$.

4. The cross product of a vector with itself vanishes: $\mathbf{a} \times \mathbf{a} = \mathbf{0}$.

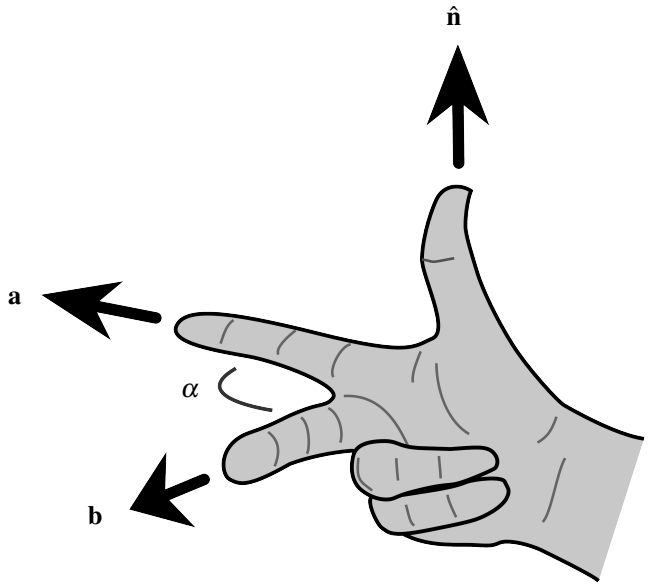

Figure 2.4: A geometric definition of the cross product in a right-handed coordinate system.

The cross product is here defined in a strictly algebraic way, Equation (2.118), in terms of how the elements of vectors **a** and **b** are mapped to the elements of the cross product $\mathbf{a} \times \mathbf{b}$. In practice it may be useful to also have a geometric interpretation of the cross product in $\mathbb{E}^3$. To do so, however, it is necessary to know the orientation of the coordinate system. In our case we have a right-handed coordinate system, and this implies that the cross product of **a** and **b** is given by

$$\mathbf{a} \times \mathbf{b} = \|\mathbf{a}\| \cdot \|\mathbf{b}\| \cdot \sin\alpha \cdot \hat{\mathbf{n}}, \tag{2.119}$$

where $\alpha \geq 0$ is the smallest positive angle between **a** and **b**, (which implies that $\sin\alpha \geq 0$), and $\hat{\mathbf{n}}$ is a normalized vector given by the *right-hand rule*, as illustrated in Figure 2.4. The right-hand rule implies that you can have your index and middle fingers point in the direction of **a** and **b**, respectively, and then your right-hand thumb points in the direction of $\hat{\mathbf{n}}$, perpendicular to **a** and **b**. In this case do **a** and **b** not have to be perpendicular, but the angle in between, α, should be the smallest possible. A consequence of this way of constructing $\hat{\mathbf{n}}$, it follows that $\{\mathbf{a}, \mathbf{b}, \hat{\mathbf{n}}\}$, in this order, form a right-handed basis.

The definition of the cross product applied to the vectors of a right-handed ON-basis in $\mathbb{R}^3$, $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$, and the right-hand rule, lead to

$$\mathbf{e}_1 \times \mathbf{e}_2 = \mathbf{e}_3, \qquad \mathbf{e}_2 \times \mathbf{e}_3 = \mathbf{e}_1, \qquad \mathbf{e}_3 \times \mathbf{e}_1 = \mathbf{e}_2. \tag{2.120}$$

2.8.3 The determinant of a 3×3 matrix

The expression of the determinant of a 3×3 matrix given in Equation (2.59) is correct but can be reformulated into a more practical form. Let $\mathbf{M}$ be a 3×3 matrix with columns $\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3$: $\mathbf{M} = (\, \mathbf{c}_1 \, \mathbf{c}_2 \, \mathbf{c}_3 \,)$. It then follows that

$$\det(\mathbf{M}) = (\mathbf{c}_1 \times \mathbf{c}_2) \cdot \mathbf{c}_3 = (\mathbf{c}_2 \times \mathbf{c}_3) \cdot \mathbf{c}_1 = (\mathbf{c}_3 \times \mathbf{c}_1) \cdot \mathbf{c}_2 \tag{2.121}$$

These expressions in the vectors $\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3$ are referred to as *scalar triple products*.

2.8.4 The inverse of a 3×3 matrix

The formulation of the inverse of a 2×2 matrix in Equation (2.69) can be extended to the 3×3 case without going over the recursive formula in Equation (2.68). Let $\mathbf{M}$ be a 3×3 matrix with columns $\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3$: $\mathbf{M} = (\, \mathbf{c}_1 \, \mathbf{c}_2 \, \mathbf{c}_3 \,)$. It then follows that

$$\mathbf{M}^{-1} = \frac{1}{\det(\mathbf{M})} (\, \mathbf{c}_2 \times \mathbf{c}_3 \,|\, \mathbf{c}_3 \times \mathbf{c}_1 \,|\, \mathbf{c}_1 \times \mathbf{c}_2 \,)^\top. \tag{2.122}$$

2.8.5 Rotation matrices in $\mathbb{R}^3$, $SO(3)$

A 3×3 rotation matrix $\mathbf{R} \in SO(3)$ satisfies the constraints

$$\mathbf{R}^\top \mathbf{R} = \mathbf{I}, \qquad \text{and} \qquad \det(\mathbf{R}) = 1. \tag{2.123}$$

The constraint $\mathbf{R}^\top \mathbf{R} = \mathbf{I}$ mean that the three columns in $\mathbf{R} = (\, \mathbf{r}_1 \, \mathbf{r}_2 \, \mathbf{r}_3 \,)$ form an ON-basis:

$$\mathbf{r}_1 \cdot \mathbf{r}_1 = \mathbf{r}_2 \cdot \mathbf{r}_2 = \mathbf{r}_3 \cdot \mathbf{r}_3 = 1, \quad \text{and} \quad \mathbf{r}_1 \cdot \mathbf{r}_2 = \mathbf{r}_2 \cdot \mathbf{r}_3 = \mathbf{r}_3 \cdot \mathbf{r}_1 = 0. \tag{2.124}$$

The constraint $\det(\mathbf{R}) = 1$ implies that the three vectors $\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3$ form a right-handed ON-basis. From this follows that each of the three vectors can be obtained as a cross product of the other two:

$$\mathbf{r}_1 \times \mathbf{r}_2 = \mathbf{r}_3, \quad \mathbf{r}_2 \times \mathbf{r}_3 = \mathbf{r}_1, \quad \mathbf{r}_3 \times \mathbf{r}_1 = \mathbf{r}_2. \tag{2.125}$$

Equations (2.124) and (2.125) are valid if we instead consider the rows of $\mathbf{R}$, rather than its columns. This implies that a rotation matrix is uniquely specified from only two of its columns or two of its rows.

2.9 Linear equations

Let $\mathbf{A}$ be an $m \times n$ non-singular matrix, $\mathbf{x} \in \mathbb{R}^n$, and $\mathbf{b} \in \mathbb{R}^m$ such that

$$\mathbf{A}\,\mathbf{x} = \mathbf{b}. \tag{2.126}$$

This type of relation between a matrix and two vectors appear frequently throughout this presentation, and may even be said to form a main theme. To make it more concrete we assume that $\mathbf{A}$ and $\mathbf{b}$ are known and we want to determine $\mathbf{x}$ such that Equation (2.126) is satisfied. Although this equation looks very simple to the eye, there are some issues that need to be considered in order to understand how to solve Equation (2.126). One issue is how many solutions Equation (2.126) has, and another is what they look like, i.e., how to compute them.

As for the first issue, it may be that there exists a unique solution for Equation (2.126), or it may have multiple solutions, or it may even have no solution at all. Which of the three cases that applies depends mainly on $\mathbf{A}$, but to some extent also on $\mathbf{b}$. As for the second issue, it makes sense to distinguish between the two cases when $\mathbf{b} \neq \mathbf{0}$ and $\mathbf{b} = \mathbf{0}$, respectively. This is because the set of solutions for the two cases have different character and the solution methods are also different. In the first case, when $\mathbf{b} \neq \mathbf{0}$, we refer to Equation (2.126) as an *inhomogeneous linear equation*. In the second case, when $\mathbf{b} = \mathbf{0}$, we refer to Equation (2.126) as a *homogeneous linear equation*. We will now analyze the two cases separately.

2.9.1 Inhomogeneous linear equations

We consider Equation (2.126) when $\mathbf{b} \neq \mathbf{0}$, i.e., we want to solve an inhomogeneous linear equation. This case can be further divided into several subcases that in a practical situation have to be treated differently. The first three subcases assume that $\mathbf{A}$ has full rank, while the last subcase deals with $\mathbf{A}$ being rank deficient.

- In the first case, we assume that $\mathbf{A}$ is square and non-singular, and we can then formulate $\mathbf{x}$ as

  $$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}. \tag{2.127}$$

  As was mentioned in Section 2.4.3, however, this expression shall not be taken as an explicit recipe for the computation of $\mathbf{x}$ as first finding the matrix inverse of $\mathbf{A}$ and then multiplying that inverse onto $\mathbf{b}$. It can also be taken as a description of $\mathbf{x}$ as the solution of solving Equation (2.126) by means of Gaussian elimination or other more advanced methods for solving a square non-singular inhomogeneous linear equation. We will leave the specific implementation of how to compute $\mathbf{x}$ in Equation (2.127) aside, and instead observe that as long as $\mathbf{A}$ is square and non-singular there exists a unique solution for $\mathbf{x}$.

- The second case to consider appears when $\mathbf{A}$ is $m \times n$ with full row rank. This means that $\mathbf{A}$ has a range that includes all of $\mathbb{R}^m$ and it is possible to determine at least one $\mathbf{x}$ that solves Equation (2.126). In accordance with Equation (2.46), it also follows that $\text{Null}(\mathbf{A})$ is of dimension $n - m$. Furthermore, if $\mathbf{x} = \mathbf{x}_0$ solves

Equation (2.126) then so will $\mathbf{x} = \mathbf{x}_0 + \mathbf{n}$ for any $\mathbf{n} \in \text{Null}(\mathbf{A})$. In summary, the set of solutions has $n - m$ "dimensions", but it is not a subspace of $\mathbb{R}^n$. Since $\mathbf{b} \neq \mathbf{0}$, the solution set cannot include $\mathbf{0}$, so it does not intersect with the origin. It is rather an $n - m$ dimensional subspace of $\mathbb{R}^n$ that has been displaced by $\mathbf{x}_0$, an affine space as it is defined in Section 2.6.1. With this observation in mind, we still need to determine an $\mathbf{x}_0$ that solves Equation (2.126). For example, we can choose

$$\mathbf{x}_0 = \mathbf{A}^\top (\mathbf{A}\,\mathbf{A}^\top)^{-1} \mathbf{b} = \mathbf{A}^+ \mathbf{b}, \tag{2.128}$$

where $\mathbf{A}^+$ is the pseudo-inverse of $\mathbf{A}$. This follows since $\mathbf{A}^+$ in this case is a right inverse of $\mathbf{A}$: $\mathbf{A}\mathbf{x} = \mathbf{A}\,\mathbf{A}^+\mathbf{b} = \mathbf{b}$. In summary, when $m \times n$ matrix $\mathbf{A}$ is of full row rank, Equation (2.126) is solved by

$$\mathbf{x} = \mathbf{x}_0 + \mathbf{n} = \mathbf{A}^+\mathbf{b} + \mathbf{n}, \tag{2.129}$$

where $\mathbf{n} \in \text{Null}(\mathbf{A})$. Before we leave this case, it should be noted that $\mathbf{x}_0$ described in Equation (2.128) is just one of infinitely many choices that can be made. Any point in the affine space described by Equation (2.129) can be used as $\mathbf{x}_0$. There is, however, one property of $\mathbf{x}_0$ in Equation (2.128) that makes it unique. Notice that this $\mathbf{x}_0 \in \text{Range}(\mathbf{A}^\top)$ and that $\mathbf{n}$ in Equation (2.129) lies in $\text{Null}(\mathbf{A})$. From Equation (2.42) follows that $\mathbf{x}_0$ and $\mathbf{n}$ are orthogonal and Equation (2.21), finally, gives

$$\|\mathbf{x}\|^2 = \|\mathbf{x}_0\|^2 + \|\mathbf{n}\|^2. \tag{2.130}$$

This implies that the norm of $\mathbf{x}$ is minimal when $\mathbf{n} = \mathbf{0}$, which means that $\mathbf{x}_0$ in Equation (2.128) is the vector of minimal norm that solves Equation (2.126).

- The third case appears when $\mathbf{A}$ is $m \times n$ with full column rank. Here, $\mathbf{A}$ represents a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$, but its range may not be the entirety of $\mathbb{R}^m$. This means that Equation (2.126) can only be solved in the case that $\mathbf{b} \in \text{Range}(\mathbf{A})$. If not, there is no $\mathbf{x}$ that solves the equation. As an alternative in this case, we can instead try to determine $\mathbf{x}$ that minimizes the difference between the left and right hand sides of Equation (2.126). We defer further analysis of this case until Section 2.10 where *least squares problems* are discussed more in detail, since that approach includes the case when $\mathbf{b} \in \text{Range}(\mathbf{A})$.

- The fourth case appears when $\mathbf{A}$ is rank deficient. This situation can be analyzed separately for different sizes of $\mathbf{A}$, but can be conveniently summarized as one consistent result based on the singular value decomposition, presented in more advanced topics.

2.9.2 Homogeneous linear equations

We now consider Equation (2.126) for $\mathbf{b} = \mathbf{0}$, when we have the homogeneous equation:

$$\mathbf{A}\,\mathbf{x} = \mathbf{0}. \tag{2.131}$$

A characteristic difference compared to the inhomogeneous case is that any solution $\mathbf{x}$ of Equation (2.131) can be multiplied by an arbitrary scalar and the result is, again, a solution of Equation (2.131). More general, if $\mathbf{x}_1$ and $\mathbf{x}_2$ are two distinct solutions of Equation (2.131), then any linear combination of the two is, again, a solution. Consequently, the solutions to Equation (2.131) form a subspace of $\mathbb{R}^n$, and trivially this subspace is $\text{Null}(\mathbf{A})$. Also trivially, $\mathbf{x} = \mathbf{0}$ always solves Equation (2.131), and normally this solution is therefore not interesting: we want to determine $\mathbf{x} \neq \mathbf{0}$ that solves Equation (2.131).

For a square $\mathbf{A}$, Equation (2.131) can be seen as a special case of the eigenvalue relation in Equation (2.83), where $\mathbf{e} = \mathbf{x}$ and $\lambda = 0$. Consequently, in this case we can solve Equation (2.131) by first determining the eigenvalues of $\mathbf{A}$. If there is an eigenvalue $= 0$, the corresponding eigenspace can be identified as $\text{Null}(\mathbf{A})$, and it contains all the solutions of Equation (2.131). If there is no eigenvalue of $\mathbf{A}$ that is zero, the null space of $\mathbf{A}$ is trivial: no solutions of Equation (2.131) exists other than $\mathbf{0}$.

This discussion applies also to the general case, when $\mathbf{A}$ is $m \times n$. We can identify the solutions of Equation (2.131) with $\text{Null}(\mathbf{A})$, although in this case it cannot be derived directly in terms of an eigensystem of $\mathbf{A}$, as this is not defined for this case. On the other hand, we notice that both sides of Equation (2.131) can be multiplied from left with $\mathbf{A}^\top$:

$$\mathbf{A}^\top \mathbf{A}\,\mathbf{x} = \mathbf{0}. \tag{2.132}$$

This means that an $\mathbf{x}$ that solves Equation (2.131) must be an eigenvector of $\mathbf{A}^\top\mathbf{A}$ corresponding to eigenvalue $= 0$. In fact, a complete set of solutions of Equation (2.131) is given by the eigenspace corresponding to an eigenvalue zero of $\mathbf{A}^\top\mathbf{A}$. Consequently, a practical implementation of finding the solutions of Equation (2.131) is to determine the eigenspace of $\mathbf{A}^\top\mathbf{A}$ that corresponds to an eigenvalue zero. This means that if there is no such eigenvalue, and Equation (2.131) has only the trivial solution $\mathbf{x} = \mathbf{0}$.

This last statement has to be taken with same care, since any practical numerical implementation that computes the eigenvalues of $\mathbf{A}^\top\mathbf{A}$ will be affected by some amount of inaccuracies. This means that a statement like "an eigenvalue equal to zero" must allow very small eigenvalues to be interpreted as "zero". On the other hand, the inaccuracies in the eigenvalues are also related to uncertainties in the elements of $\mathbf{A}$, usually because they are derived from quantities that are measured by means of a process that includes some measurement noise. Any practical implementation of solving Equation (2.131), using the eigenvalues of $\mathbf{A}^\top\mathbf{A}$ or any other method, must deal with this problem of interpreting what is meant by some quantity is "equal to zero" when that quantity has numerical inaccuracies.

The approach for solving Equation (2.131) that is discussed here, based on an eigenvalue decomposition of $\mathbf{A}^\top\mathbf{A}$, has the practical advantage of being straight-forward to implement, at least for moderate sizes of $\mathbf{A}$ and when software is available for the eigenvalue decomposition. However, it has numerical issues that makes it less attractive for solving Equation (2.131) in the general case. If m is very large, the result of computing $\mathbf{A}^\top\mathbf{A}$ may be affected by additional numerical inaccuracies since each element in this matrix is a sum of m terms where each term is a product of elements in $\mathbf{A}$, and each such product has a round-off error. This additional inaccuracy may introduce errors in the eigenvalues that can seriously perturb the solutions. Furthermore, if n is very large, we probably need to store $\mathbf{A}^\top\mathbf{A}$ as the very large $n \times n$ matrix in order to solve the eigenvalue decomposition. Furthermore, EVD of a large matrix has more numerical inaccuracies than of a smaller matrix.

In more advanced topics, we will consider an alternative approach for solving Equation (2.131) in the general case, based on a singular value decomposition of $\mathbf{A}$. In general, this is the preferred approach.

2.10 Least squares problems

In the discussion on how to solve the linear equation Equation (2.126) it was mentioned that a possible outcome is that there is no solution. This typically happens when $\mathbf{A}$ is $m \times n$ and of full column rank, but is also an possibility as soon as $\mathbf{A}$ is rank deficient. In all these cases, Range($\mathbf{A}$) is not the entirety of $\mathbb{R}^m$ and only if $\mathbf{b} \in$ Range($\mathbf{A}$) it is possible to find a solution $\mathbf{x}$. A practical approach to find a meaningful solution to the linear equation anyway, is to determine an $\mathbf{x}$ that minimizes the difference between the left and right hand sides of Equation (2.126). This difference is the *residual*, $\mathbf{r}$, defined as

$$\mathbf{r} = \mathbf{A}\,\mathbf{x} - \mathbf{b}, \quad \text{where} \quad \mathbf{r} = \begin{pmatrix} r_1 \\ \vdots \\ r_m \end{pmatrix} \in \mathbb{R}^m. \tag{2.133}$$

We want to find $\mathbf{x}$ that minimizes $\|\mathbf{r}\|$, but since it is easer to deal with squares of norms rather than the norms directly, and we get the same solution, we define a cost function ε as

$$\varepsilon = \|\mathbf{r}\|^2 = \mathbf{r}^\top\mathbf{r} = r_1^2 + \ldots + r_m^2, \tag{2.134}$$

and formulate the *least squares problems* related to Equation (2.126) as finding the $\mathbf{x}$ that minimizes ε.

To find this $\mathbf{x}$, we expand ε as a function of $\mathbf{x}$:

$$\varepsilon = \varepsilon(\mathbf{x}) = \|\mathbf{A}\,\mathbf{x} - \mathbf{b}\|^2 = (\mathbf{A}\,\mathbf{x} - \mathbf{b})^\top(\mathbf{A}\,\mathbf{x} - \mathbf{b}) = \mathbf{x}^\top\mathbf{A}^\top\mathbf{A}\,\mathbf{x} - 2\,\mathbf{x}^\top\mathbf{A}^\top\mathbf{b} + \mathbf{b}^\top\mathbf{b}. \tag{2.135}$$

With $\mathbf{y} = \mathbf{A}\,\mathbf{x}$, it follows that $\mathbf{y} \in S$, the subspace of $\mathbb{R}^m$ spanned by the columns of $\mathbf{A}$. If $\mathbf{y}$ also is given by the $\mathbf{x}$ that minimizes ε, it must be the case that $\mathbf{b} - \mathbf{y}^\top \perp S$. To see this, let $\mathbf{b} - \mathbf{y}^\top \perp S$ and consider $\mathbf{w} = \mathbf{y} + \mathbf{z}$ for some $\mathbf{z} \in S$, and minimize $\|\mathbf{b} - \mathbf{w}\|$ over different choices of $\mathbf{z}$. We get

$$\begin{aligned} \|\mathbf{b} - \mathbf{w}\|^2 &= (\mathbf{b} - \mathbf{w})^\top(\mathbf{b} - \mathbf{w}) = (\mathbf{b} - \mathbf{y} - \mathbf{z})^\top(\mathbf{b} - \mathbf{y} - \mathbf{z}) = \\ &= (\mathbf{b} - \mathbf{y})^\top\mathbf{b} + \mathbf{z}^\top\mathbf{z} - \underbrace{(\mathbf{b} - \mathbf{y})\,\mathbf{w}}_{=0} - \underbrace{\mathbf{z}^\top(\mathbf{b} - \mathbf{y})}_{=0} = (\mathbf{b} - \mathbf{y})^\top\mathbf{b} + \mathbf{z}^\top\mathbf{z} \end{aligned} \tag{2.136}$$

From this expression it follows immediately that $\|\mathbf{b}-\mathbf{w}\|$ is minimized when $\mathbf{z}=0$. Consequently, ε is minimized when $\mathbf{y}=\mathbf{A}\,\mathbf{x}$ is chosen such that $\mathbf{b}-\mathbf{y}\perp S$. This happens exactly when $\mathbf{A}^\top(\mathbf{b}-\mathbf{A}\,\mathbf{x})=\mathbf{0}$ or

$$\mathbf{A}^\top\mathbf{A}\,\mathbf{x}=\mathbf{A}^\top\mathbf{b}. \tag{2.137}$$

This is the *normal equation* of the least squares problem. Any $\mathbf{x}$ that minimizes the residual must solve the normal equation Equation (2.137).

Consequently, solving a least squares problem brings us back to solving a linear equation, Equation (2.137). This equation can be inhomogeneous or homogeneous, depending on whether $\mathbf{b}\neq\mathbf{0}$ or not. In particular when $\mathbf{A}$ is $m\times n$ of full column rank, it is not possible to solve the linear equation Equation (2.126) in the general case. We can then instead solve the corresponding least squares problem in terms of the normal equation Equation (2.137). The normal equation is either an inhomogeneous or homogeneous linear equation, and has to be solved accordingly.

In the case that $\mathbf{A}$ has full column rank, the normal equation Equation (2.137) is solved as

$$\mathbf{x}=(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{b}=\mathbf{A}^{+}\mathbf{b}. \tag{2.138}$$

2.10.1 Concluding remarks

We can summarize the results from the last two sections on solving the linear equation Equation (2.126) and least squares problems as: as long as $\mathbf{A}$ is of full rank we can determine solutions to Equation (2.126) in accordance with

- $\mathbf{A}$ is $n\times n$: $\mathbf{x}=\mathbf{A}^{-1}\mathbf{b}$.
- $\mathbf{A}$ has rank $m<n$: $\mathbf{x}=\mathbf{A}^{+}\mathbf{b}+\mathbf{n}$, where $\mathbf{n}\in\mathrm{Null}(\mathbf{A})$.
- $\mathbf{A}$ has rank $n>m$: the linear equation cannot be solved in general, but a least square solution that minimizes the residual is given as $\mathbf{x}=\mathbf{A}^{+}\mathbf{b}$.

All other cases can be analyzed as well, although this relies on more advanced techniques such as singular value decomposition that is not presented here.

Chapter 3

Calculus

Differentiation and integration of one-variable functions is assumed to be known already from secondary school. What is discussed here are certain concepts and properties required for understanding more advanced topics.

3.1 Functions on $\mathbb{R}$

The simplest class of functions studied in calculus is real-valued functions on $\mathbb{R}$, i.e., functions that map $\mathbb{R}$ onto itself.

3.1.1 Derivatives

The derivative of a one-variable function $f : \mathbb{R} \to \mathbb{R}$ should be a familiar concept to the reader and, here, we merely make the observation that it can be treated in two related, but slightly different, meanings.

Derivative as a linear mapping

In the first sense, the derivative of f is a new one-variable function $f' : \mathbb{R} \to \mathbb{R}$, with a value at point x given as the limit

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} = \lim_{x_0 \to x} \frac{f(x_0) - f(x)}{x_0 - x} \tag{3.1}$$

Whenever we use derivatives, it is assumed that this limit value is well-defined, at least for the values of x that are of interest for the particular problem at hand. This definition of the derivative f' can be seen as a linear transformation on the set of function, at least if we restrict our attention to functions f for which f' is a well-defined. This linear transformation, the *derivative operator*, is sometimes denoted D and we can write

$$f' = D\,f. \tag{3.2}$$

An important aspect of D is that this operator is independent of what we choose to call the function variable. If f is a function $\mathbb{R} \to \mathbb{R}$, then this means that it returns a real value for any real number that we apply it on. f can be applied to the value 0 or 42, or to x or y. These are just different values that f can be applied to, and independent of its value or what we choose to call the variable, $D\,f$ gives the same function f' as the derivative of f.

Higher order derivatives of f are obtained by applying the derivative operator D multiple times to f. For example, the second derivative of f is given as

$$f'' = D^2 f = D\,D\,f = D\,f' \tag{3.3}$$

Derivative with respect to a specific variable

As an alternative, we may refer specifically to the derivative of f with respect to some specific variable. In the case that this variable is x, the derivative of f with respect to x is a function that is denoted

$$\frac{df}{dx}. \tag{3.4}$$

In order to be well-defined, Equation (3.4) relies on x being established as the variable of f. For example, in the case that f is defined as $f(z) = 2 + z$, then the derivative in Equation (3.4) is not well-defined unless there is some connection between x and z. In many practical applications, f is defined as function of a variable, e.g., x, and then derivative in Equation (3.2) coincides with the derivative described in Equation (3.4). In general, however, it is not true that the derivative in Equation (3.2) is equal to the derivative in Equation (3.4). This is illustrated by the chain rule.

The chain rule

In some applications f may be defined as a function of a variable, e.g., y, and y, itself, is a function of x. For example:

$$f(y) = y^2, \quad \text{where} \quad y = \sin x. \tag{3.5}$$

The derivative of f is then given as

$$f'(y) = \frac{df}{dy} = 2y. \tag{3.6}$$

In some cases, we may instead be interested in the derivative of f with respect to x, rather than y. This former derivative is given by the *chain rule*:

$$\frac{df}{dx} = \frac{df}{dy}\frac{dy}{dx}. \tag{3.7}$$

Given the example above, the derivative of f with respect to x amounts to

$$\frac{df}{dx} = \frac{df}{dy}\frac{dy}{dx} = 2\,\mathbf{y}\cos x = 2\,\sin x\cos x = \sin 2x. \tag{3.8}$$

In practice, these observations imply that the concept "derivative of a function" should be used with some care. In particular, if there are several variables or parameters that appear in the definition of the function f, it is important to specify which one is the true function variable. The derivative in Equation (3.2) then refers to the derivative of this specific variable. If the derivative of f with respect to any other variable or parameters appears in some expression, we should instead use the chain rule to determine this derivative.

Stationary points

The value of $f'(x_0)$ represents the rate of change of the function f at x_0. If $f'(x_0) > 0$ the function increases, and if $f'(x_0) < 0$ the function decreases. If $f'(x_0) = 0$, the function has a *stationary point* at x_0. It is a necessary requirement for x_0 to be a stationary point in order for f to have a local minimum or maximum at x_0. The local properties of f in a region around x_0 are characterized also by the second order derivative f'', representing the rate of change in the first order derivative f'. The function f has a local minimum at x_0 if $f''(x_0) > 0$, or a local maximum if $f''(x_0) < 0$. If $f''(x_0) = 0$ at the stationary point x_0, the character of f at this point cannot be determined from the second order derivative along and, instead, we need to analyze higher order derivatives at x_0 in order to establish the local behavior of f at this point.

3.2 Functions on $\mathbb{R}^n$

In Chapter 2, we discussed linear transformations, a special class of functions on $\mathbb{R}^n$, and their representation in terms of matrices. In this section we look at general functions that map $\mathbb{R}^n$ to some set, the co-domain of f.

In the case that the co-domain is $\mathbb{R}$, we consider a function $f : \mathbb{R}^n \to \mathbb{R}$ and we use $f(\mathbf{v})$ to denote the result of applying f to a vector $\mathbf{v} \in \mathbb{R}^n$. In this case, we can use the fact that $\mathbb{R}^n$ can be expanded as the Cartesian product $\mathbb{R}^n = \mathbb{R} \times \mathbb{R} \times \ldots \times \mathbb{R}$. As a consequence, we can treat f either as a function of the single vector variable

$$\mathbf{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \in \mathbb{R}^n, \tag{3.9}$$

or as a multi-variable function of the n variables $v_1, \ldots, v_n \in \mathbb{R}$. Both views are applicable and it may even be useful to shift between the two depending on the situation, or what problem f refers to. For example, in some situations it may be convenient to use the compact notation $f(\mathbf{v})$ for the function value at $\mathbf{v}$, instead of $f(v_1, \ldots, v_n)$. In other cases, it may instead be useful to consider the derivatives f with respect to the different elements of $\mathbf{v}$, and then f as a function of n real variables may provide a better description.

Partial derivatives

When $f : \mathbb{R}^n \to \mathbb{R}$, we can study the derivative of f with respect to each of the n variables. These derivatives form what is referred to as the *n partial derivatives* of f, each of them describes the rate of change in f when one single variable changes and the other are kept fixed. If the n variables of f are denoted $v_1, \ldots, v_n$, the n partial derivatives of f are denoted

$$\frac{\partial f}{\partial v_k}, \quad k = 1, \ldots, n. \tag{3.10}$$

Notice the difference in notation relative to the derivative of a function of a single variable in Equation (3.4).

Gradient

The collection of all n partial derivatives of $f : \mathbb{R}^n \to \mathbb{R}$, seen as a vector in $\mathbb{R}^n$ is referred to as the *gradient* of f. The gradient of f, denoted ∇f, is defined as

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial v_1} \\ \frac{\partial f}{\partial v_2} \\ \vdots \\ \frac{\partial f}{\partial v_n} \end{pmatrix} \in \mathbb{R}^n. \tag{3.11}$$

Note that $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$.

Given a point $\mathbf{v} \in \mathbb{R}^n$ and a unit vector $\hat{\mathbf{n}} \in S^{n-1}$ we can define a one-variable function g as

$$g(x) = f(\mathbf{v} + \hat{\mathbf{n}}x), \tag{3.12}$$

which describes the variation of f in the direction defined by $\hat{\mathbf{n}}$ at the point $\mathbf{v}$. Since $\hat{\mathbf{n}}$ is normalized, x gives the distance from $\mathbf{v}$ in the direction of $\hat{\mathbf{n}}$. Using the chain rule, we get

$$g'(x) = \frac{dg}{dx} = \nabla f \cdot \hat{\mathbf{n}}. \tag{3.13}$$

This means that ∇f, as a vector in $\mathbb{R}^n$, represents the direction from $\mathbf{v}$ in which f has its highest positive rate of change. Similarly, $-\nabla f$ represents the direction from $\mathbf{v}$ in which f has its highest negative rate of change.

Stationary points

If $\nabla f(\mathbf{v}) = \mathbf{0}$ at some point $\mathbf{v} \in \mathbb{R}^n$, we say that $\mathbf{v}$ is a *stationary point* of f. If f has a local maximum or minimum at $\mathbf{v}$, then it must be the case that $\mathbf{v}$ is a stationary point, i.e., $\nabla f(\mathbf{v}) = \mathbf{0}$. The opposite implication, however, is not true: a stationary point of f must not correspond to a a local maximum or minimum. In order to establish the character of a stationary point $\mathbf{v}$, we also need to investigate the Hessian of f at $\mathbf{v}$.

Hessian

The local properties of f around a stationary point $\mathbf{v}$ are described by the *Hessian* of f at $\mathbf{v}$, consists of all possible combinations of second order derivatives of f, collected by the symmetric $n \times n$ matrix

$$\mathbf{H}\{f\} = \nabla\nabla^\top f = \begin{pmatrix} \frac{\partial^2 f}{\partial v_1^2} & \frac{\partial^2 f}{\partial v_1 \partial v_2} & \cdots & \frac{\partial^2 f}{\partial v_1 \partial v_n} \\ \frac{\partial^2 f}{\partial v_2 \partial v_1} & \frac{\partial^2 f}{\partial v_2^2} & \cdots & \frac{\partial^2 f}{\partial v_2 \partial v_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial v_n \partial v_1} & \frac{\partial^2 f}{\partial v_2 \partial v_n} & \cdots & \frac{\partial^2 f}{\partial v_n^2} \end{pmatrix}. \tag{3.14}$$

The character of the Hessian, as a quadratic form (see Section 2.4.8), determines the character of f in a local region of a stationary point $\mathbf{v}$ as follows:

1. If $\mathbf{H}\{f\}$ is positive definite: f has a local minimum at $\mathbf{v}$.
2. If $\mathbf{H}\{f\}$ is positive semi-definite: f may have a local minimum at $\mathbf{v}$, but this depends in higher order derivatives of f and cannot be determined from the Hessian alone.
3. If $\mathbf{H}\{f\}$ is negative definite: f has a local maximum at $\mathbf{v}$.
4. If $\mathbf{H}\{f\}$ is negative semi-definite: f may have a local maximum at $\mathbf{v}$, but this depends in higher order derivatives of f and cannot be determined from the Hessian alone.
5. If $\mathbf{H}\{f\}$ is indefinite: f has a *saddle point* at $\mathbf{v}$: it appears as local minimum in certain directions and a local maximum in other directions.

In practice, cases 1 and 3 that are of main interest since they allow us to immediately know something about the nature of a stationary point. Cases 2 and 4 can be further analyzed in terms of higher order derivatives but typically lead to non-trivial results, and will not be further discussed here. Case 5, finally, implies that f has as *saddle-point* at $\mathbf{v}$, a point from which f increases in some direction and decreases in other. This is an example of a stationary point that does not correspond to an optimum, neither as a minimum nor as a maximum.

Total derivative

In the case of an n-variable function $f : \mathbb{R}^n \to \mathbb{R}$, where each of the n variables, $v_k, k = 1, \ldots, n$, is a function of some variable t, then the derivative of f with respect to y is given by an extended version of the *chain rule*, as

$$\frac{df}{dt} = \frac{\partial f}{\partial v_1}\frac{dv_1}{dt} + \frac{\partial f}{\partial v_2}\frac{dv_1}{dt} + \ldots + \frac{\partial f}{\partial v_n}\frac{dv_1}{dt} = \nabla f \cdot \frac{d\mathbf{v}}{dt}, \tag{3.15}$$

where $\mathbf{v} = (v_1, \ldots, v_n)$. The left-hand side of Equation (3.15) is referred to as the *total derivative* of f, relative to the single variable t.

3.2.1 Taylor expansion

A one-variable function $f : \mathbb{R} \to \mathbb{R}$ can have a *Taylor expansion* at the point x_0 in terms of the power series

$$f(x_0 + h) = f(x_0) + f'(x_0)h + \frac{1}{2}f''(x_0)h^2 + \ldots = \sum_{k=0}^{\infty} \frac{1}{k!} f^{(k)}(x_0)h^k, \tag{3.16}$$

where $f^{(k)}(x_0)$ denotes the k-th order derivative of f evaluated at point x_0. The power series in the right hand side of Equation (3.16) is the *Taylor series* of f at the point x_0. This result implies that at, and in the vicinity of, the point x_0 we can approximate f by a low order polynomial in h, the displacement from x_0. For example, in a sufficiently small region around x_0 we can approximate f as

$$f(x_0 + h) \approx f(x_0) + f'(x_0)\,h + \frac{1}{2}f''(x_0)\,h^2, \tag{3.17}$$

or, more formally, as

$$f(x_0+h) = f(x_0) + f'(x_0)\,h + \frac{1}{2}f''(x_0)\,h^2 + O(h^3), \tag{3.18}$$

where $O(h^3)$ is a function of h that grows at least a fast as h^3 does for large h.

A necessary requirement on f to allow us to expand it as in Equation (3.16) is that its derivatives of all orders are continuous functions, but this is not sufficient in general. For a general function, even one with continuous derivatives of all orders, the Taylor series may or may not converge for different values of x_0 and Δ, and if it converges it may not converge to the expected function value of f. The set of functions for which the Taylor expansion is a correct representation of the function for all x_0 and Δ is referred to as *analytic functions*, and they are defined simply as those functions for which the Taylor series converges to the function itself. This means that the idea of Taylor series has to be used with some care, and should include an investigation of which x_0 and Δ that make the Taylor series converge to f. The functions that are considered for a Taylor expansion do not have to be analytic, it may be sufficient that the Taylor series converges to f in some interval to make it useful.

Taylor expansion of multi-variable functions

The concept of Taylor expansions can be extended to a function $f : \mathbb{R}^n \to \mathbb{R}$. Such a function can have a Taylor expansion at the point $\mathbf{x}_0 \in \mathbb{R}^n$ in accordance with

$$f(\mathbf{x}_0+\mathbf{h}) = f(\mathbf{x}_0) + \mathbf{h}^\top \underbrace{\nabla f(\mathbf{x}_0)}_{\substack{= \text{ gradient} \\ \text{of } f \text{ at } \mathbf{x}_0}} + \frac{1}{2}\mathbf{h}^\top \underbrace{\mathbf{H}\{f\}(\mathbf{x}_0)}_{\substack{= \text{ Hessian} \\ \text{of } f \text{ at } \mathbf{x}_0}} \mathbf{h} + O(\|\mathbf{h}\|^3) \tag{3.19}$$

In the same way as for the one-variable Taylor expansion, for general values of $\mathbf{x}_0$ and $\mathbf{h}$, the Taylor series in the right-hand side of Equation (3.19) may converge to the expected value of f or not. The power series defined by a Taylor series is unique: there cannot be two power series with distinct coefficients that converge to the same function f.

3.3 Optimizing functions on $\mathbb{R}^n$

Optimizing a function $f : \mathbb{R}^n \to \mathbb{R}$ implies finding its maximal or minimal value. The value of $\mathbf{v}$ for which f is minimal (or maximal) is the *minimizer* (or *maximizer*) of f. In general such a point $\mathbf{v}$ is also referred to a the *optimizer* or simply the *optimum*. In his case, we rely on the context to determine whether a particular optimization problem means to find a minimum or a maximum.

In terms of notation, we write

$$f_{\min} = \min_{\mathbf{v}} f(\mathbf{v}) \quad \text{and} \quad f_{\max} = \max_{\mathbf{v}} f(\mathbf{v}), \tag{3.20}$$

to denote that $f_{\min}$ is the minimal value of f, and that $f_{\max}$ is the maximal value of f, for all choices of $\mathbf{v}$ in the domain of f. Furthermore, we write

$$\mathbf{v}_{\min} = \operatorname{argmin} f(\mathbf{v}) \quad \text{and} \quad \mathbf{v}_{\max} = \operatorname{argmax} f(\mathbf{v}), \tag{3.21}$$

to denote that $\mathbf{v}_{\min}$ is the minimizer of f and that $\mathbf{v}_{\max}$ is the maximizer of f, i.e., $f(\mathbf{v}_{\min}) = f_{\min}$ and $f(\mathbf{v}_{\max}) = f_{\max}$. In some applications we are only interested in the optimizer of f, in some applications only in the optimal value of f, and in some applications in both. Notice that some functions do not have a minimizer or maximizer and also notice that a minimizer or maximizer need not be unique. In most practical applications, however, we consider functions f for which existence and uniqueness of the relevant optimizer can be assumed. In particular we normally need to assume that f can be differentiated with respect to its variables. It is often nor necessary that derivatives of arbitrary orders are well-defined, but at lest of first order and sometimes also of second order, i.e., the gradient and the Hessian of f.

In most practical cases, we often are interested in finding the *global optimum* of f, i.e., the maximal or minimal value of $f(\mathbf{v})$ for all $\mathbf{v} \in \mathbb{R}^n$. In the case that f is differentiable, the global optimum corresponds to a stationary point, and finding the stationary points of f is, therefore, sometimes an initial step in the search for the global optimum. This search, however, is complicated by two general observations.

First, the stationary points are determined from the condition $\nabla f = \mathbf{0}$. This condition corresponds to n equations in the n variables of f. By solving these equations for the unknown n-dimensional vector $\mathbf{v}$, the variable of f, we obtain the stationary points. In general, however, these equations can be very complicated so solve. For example, there may be no closed form expression for the solutions and, instead, they have to be determined numerically based on more or less heuristic methods. Only for special types of functions f, described in Section 3.3.3, can $\nabla f = \mathbf{0}$ be solved in a simple manner in terms of linear equations.

Second, in general f may have several stationary points. Given that the Hessian at a stationary point is either positive or negative definite, we can then at least be sure that the stationary points is a maximum or a minimum in a local region around the point: it is a *local optimum*. This means that a stationary point $\mathbf{v}$ can be optimal relative to a local neighborhood around $\mathbf{v}$, but it does not necessarily mean that $\mathbf{v}$ is a global optimum. In order to establish global optimality, therefore, we need to compare f for all local optima, and find which of them that gives the maximal or minimal value of f.

In summary, these two observations imply that finding the global optimum of f can be a computationally complex problem. Finding the stationary points can typically *not* be done by numerically solving $\nabla f = \mathbf{0}$ for $\mathbf{v}$. Instead we have to use iterative methods that start with some initial solution $\mathbf{v}_0$ and, in each iteration, moves from $\mathbf{v}_k$ to $\mathbf{v}_{k+1}$ such that $f(\mathbf{v}_{k+1}) < f(\mathbf{v}_k)$ when f is minimized (or $f(\mathbf{v}_{k+1}) > f(\mathbf{v}_k)$ when f is maximized). This iterative processing continues until, typically, the change in f is sufficiently small to allow us to believe that the final $\mathbf{v}_k$ is very close to a stationary point. If this is indeed the case, and if the stationary point is a global rather than a local optimum, have typically have to remain as hypotheses. Such iterative optimization methods are often referred to as non-linear optimization, and are not discussed here.

3.3.1 Constrained optimization and Lagrange's method

In several applications where we want to optimize a function $f : \mathbb{R}^n \to \mathbb{R}$ this is not done over all possible $\mathbf{v} \in \mathbb{R}^n$ but the search is instead restricted to a subset of $\mathbb{R}^n$ that can have a rather general shape. In the simplest case, such a subset can be identified as those points $\mathbf{v} \in \mathbb{R}^n$ that satisfy $c(\mathbf{v}) = 0$ for some function $c : \mathbb{R}^n \to \mathbb{R}$. Such a function is known as a *constraint* and if we optimize f over only $\mathbf{v}$ that satisfy the constraint, we are doing *constrained optimization*.

In constrained optimization we cannot use the general approach for optimization described above, since its stationary points may not satisfy the constraint. Instead we use *Lagrange's method* for constrained optimization where instead stationary points are defined as satisfying both

$$c(\mathbf{v}) = 0, \tag{3.22}$$

$$\nabla f = \lambda \nabla c. \tag{3.23}$$

Here, $\lambda \in \mathbb{R}$ is a scalar that can be determined if necessary, it is often referred to as a *Lagrange multiplier*. Notice that since we now have an additional equation in $c(\mathbf{v}) = 0$, it makes sense to introduce another unknown variable λ to the problem. In the same way as for unconstrained optimization, the stationarity of a point is merely a necessary condition but, in general, it is not sufficient for optimality. The stationary points produced by Lagrange's method need to be further analyzed in order to determine, e.g., if they represent a local minimum or maximum. This, however, will not be necessary for the investigations done in this presentation.

3.3.2 Gradient and Hessian of linear and quadratic forms

A particularly simple and practical case of optimization appears when f is a scalar product between the vector variable and some constant vector $\mathbf{a}$:

$$f_1(\mathbf{v}) = f_1(v_1, \ldots, v_n) = \mathbf{v} \cdot \mathbf{a} = \sum_{i=1}^{n} v_i [\mathbf{a}]_i. \tag{3.24}$$

or a quadratic form:

$$f_2(\mathbf{v}) = f_2(v_1, \ldots, v_n) = \mathbf{v}^\top \mathbf{M}\, \mathbf{v} = \sum_{i,j=1}^{n} v_i [\mathbf{M}]_{ij} v_j. \tag{3.25}$$

Both in the constrained and the unconstrained case are we interested in the gradient and the Hessian of f_1 or f_2. Element k of the gradient of f_1 is given as

$$[\nabla f_1]_k = \frac{\partial}{\partial v_k} f_1 = \frac{\partial}{\partial v_k} \sum_{i=1}^{n} v_i [\mathbf{a}]_i = \sum_{i=1}^{n} \frac{\partial v_i}{\partial v_k} [\mathbf{a}]_i = \sum_{i=1}^{n} \delta_{ik} [\mathbf{a}]_i = [\mathbf{a}]_k. \tag{3.26}$$

This means that this gradient can be expressed more compactly as

$$\nabla f_1 = \mathbf{a}. \tag{3.27}$$

In a similar way, element k of the gradient of f_2 is given as

$$[\nabla f]_k = \frac{\partial}{\partial v_k} f = \frac{d}{dv_k} \sum_{i,j=1}^{n} v_i [\mathbf{M}]_{ij} v_j = \sum_{i,j=1}^{n} \frac{dv_i}{dv_k} [\mathbf{M}]_{ij} v_j + \sum_{i,j=1}^{n} v_i [\mathbf{M}]_{ij} \frac{dv_j}{dv_k} = \tag{3.28}$$

$$= \sum_{i,j=1}^{n} \delta_{ik} [\mathbf{M}]_{ij} v_j + \sum_{i,j=1}^{n} v_i [\mathbf{M}]_{ij} \delta_{jk} = \sum_{j=1}^{n} [\mathbf{M}]_{kj} v_j + \sum_{i=1}^{n} v_i [\mathbf{M}]_{ik} = \tag{3.29}$$

$$= \Big/ \begin{matrix} \text{Change } j \text{ to } i \text{ in the first sum.} \\ \mathbf{M} \text{ is symmetric in second sum.} \end{matrix} \Big/ = \sum_{j=1}^{n} [\mathbf{M}]_{ki} v_i + \sum_{i=1}^{n} [\mathbf{M}]_{ki} v_i = 2 \sum_{i=1}^{n} [\mathbf{M}]_{ki} v_i. \tag{3.30}$$

This means that this gradient can be expressed more compactly as

$$\nabla f_2 = 2\,\mathbf{M}\,\mathbf{v}. \tag{3.31}$$

Since ∇f_1 is a vector that does not depend on $\mathbf{v}$, the Hessian of f_1 vanishes. The Hessian of f_2 is

$$\mathbf{H}\{f_2\} = 2\,\mathbf{M}. \tag{3.32}$$

With these results at hand, it is straight-forward to consider an expression like

$$f_3(\mathbf{v}) = \|\mathbf{A}\,\mathbf{v} - \mathbf{b}\|^2 = (\mathbf{A}\,\mathbf{v} - \mathbf{b})^\top (\mathbf{A}\,\mathbf{v} - \mathbf{b}) = \mathbf{v}^\top \mathbf{A}^\top \mathbf{A}\,\mathbf{v} - 2\,\mathbf{v}^\top \mathbf{A}\,\mathbf{b} + \mathbf{b}^\top \mathbf{b}. \tag{3.33}$$

As a function of v, it has been expanded into a sum of a quadratic form, a linear form, and a constant. The gradient is therefore given as

$$\nabla f_3 = 2(\mathbf{A}^\top \mathbf{A}\mathbf{v} - \mathbf{A}\,\mathbf{b}), \tag{3.34}$$

and the corresponding Hessian is

$$\mathbf{H}\{f_3\} = 2\,\mathbf{A}^\top \mathbf{A}. \tag{3.35}$$

Notice that the Hessian in this case is positive definite if $\mathbf{A}$ has full column rank, and positive semi-definite otherwise. In the first case, it means that f always has a minimum for any $\mathbf{v}$ that solves Equation (3.34).

3.3.3 Optimization of a second order function

A particularly simple case of optimization appears if f is function of $\mathbf{x} \in \mathbb{R}^n$ in terms of a second order polynomial:

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\,\mathbf{x} + \mathbf{x}^\top \mathbf{b} + c, \tag{3.36}$$

for $\mathbf{A} \in \mathrm{Sym}(n)$, $\mathbf{b} \in \mathbb{R}^n$ and $c \in \mathbb{R}$. Additionally, we assume that $\mathbf{A}$ is either positive definite or negative definite and, consequently, $\mathbf{A}$ is of full rank and has an inverse. The gradient of f is given as

$$\nabla f = 2\,\mathbf{A}\,\mathbf{x} + \mathbf{b}, \tag{3.37}$$

which means that any stationary point $\mathbf{x}_0$ must satisfy

$$2\,\mathbf{A}\,\mathbf{x}_0 + \mathbf{b} = \mathbf{0}. \tag{3.38}$$

The Hessian of f is $2\,\mathbf{A}$, and implies that this stationary point is the global maximum of f if $\mathbf{A}$ is negative definite and the global minimum of f if $\mathbf{A}$ is positive definite. This optimal value of f is reached at the unique point

$$\mathbf{x}_0 = -\frac{1}{2}\mathbf{A}^{-1}\mathbf{b}. \tag{3.39}$$

If $\mathbf{A}$ is indefinite and of full rank, $\mathbf{x}_0$ in Equation (3.39) is still a well-defined stationary point, but it represents a saddle-point.

If $\mathbf{A}$ is either positive or negative semi-definite, or indefinite but not of full rank, at least one of its eigenvalues is zero, and $\mathbf{A}$ does not have a well-defined inverse. In this case, the analysis depends also on $\mathbf{b}$ and the eigenspace E_0 consisting of all eigenvectors of $\mathbf{A}$ corresponding to eigenvalue 0. The conclusion is that we can determine a stationary point of f if and only if $\mathbf{b} \perp E_0$. Furthermore, if $\mathbf{x}_0$ is a stationary point, then so is $\mathbf{x}_0 + \mathbf{e}_0$, where $\mathbf{e}_0 \in E_0$. Such an affine space of stationary points represents the global maximum of f if $\mathbf{A}$ is negative semi-definite and the global minimum of f of $\mathbf{A}$ is positive semi-definite.

Index