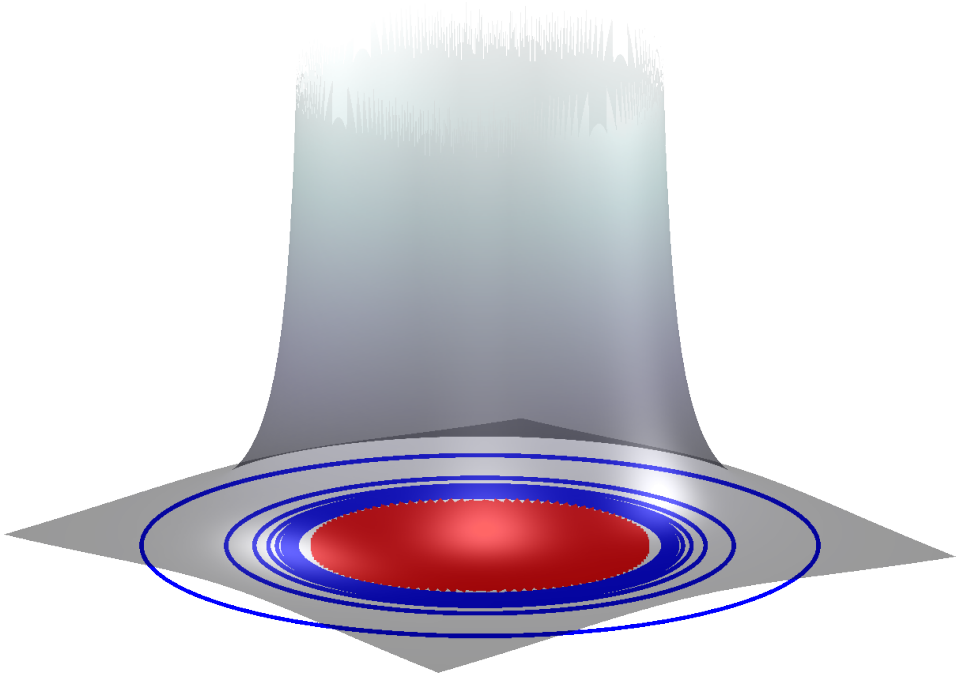
Notes on Linear Algebra and Functional Analysis

Bassam Bamieh

Department of Mechanical Engineering
University of California at Santa Barbara



Contents

****

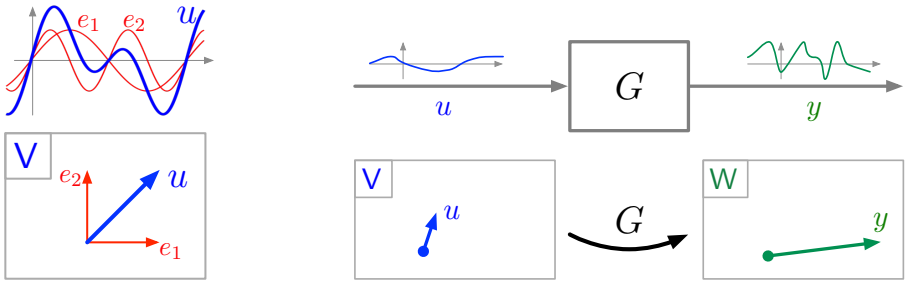
Preface

These notes are a summary of topics in linear algebra and functional analysis that are relevant to problems in Signals, Systems, and Controls. The term *Linear Algebra* is used in an expansive sense. The concepts behind vectors and matrices are generalizable to abstract vector spaces and linear operators on them. This is the subject of *Functional Analysis*, an incredibly useful and powerful mathematical tool in Systems and Controls. It is the study of the algebraic and geometric properties of abstract vector spaces, and mappings between them. In standard Euclidean space, vectors can be added, scaled, and their lengths and angles between them quantified with analytic geometry. Matrices can be interpreted as linear transformations of Euclidean space, and this geometric interpretation of matrix operations yields helpful intuition. It is this geometric view of linear algebra that generalizes naturally to abstract vector spaces and provides a powerful tool in Engineering and Science.

In Systems and Controls, the basic objects to study are the *signals*, and the *systems* that operate on signals to produce other signals. Signals can be viewed as "vectors" in abstract vector spaces, which are generalizations of the well known Euclidean space. Signals can be added or scaled in a similar manner as standard vectors. The "size" of a signal can be quantified using norms in a similar manner to lengths of vectors, and notions of orthogonality and angles between signals are also generalizable from the analogous notions on ordinary vectors. In this way, the formalism of abstract vector spaces provides a powerful framework for manipulating signals, and defining an underlying geometry of signals in the same manner as vectors in analytic geometry. Figure 1a illustrates a particular case where a (periodic) signal u can be viewed as the sum of two "mutually orthogonal" signals e_1 and e_2. The reader may already be familiar with this concept in the context of Fourier series, but the idea of treating signals as vectors in an abstract vector space has much wider applications.

Systems can be viewed as mappings between signal vector spaces. Systems that preserve the linear vector space structure as they map signals (i.e. satisfy the superposition property) are called *linear systems*. They can be thought of as generalizations of matrices to the concept of a *linear operator* on an abstract vector space. Figure 1b illustrates this point of view. Many concepts in matrix analysis, such as range and null spaces, eigenvalues and eigenvectors, diagonalization, and singular values can be defined for linear operators on abstract vector spaces. For example, transform analysis such as Fourier, Laplace or the $\mathbf{z}$-transform can be thought of as various forms of "diagonalizations" of the input-output system as a linear operator on properly defined vector spaces.

The emphasis in these notes is primarily on the geometric view of linear algebra discussed above, and its generalizations to function spaces using the notions of abstract vector spaces. The approach emphasizes the algebraic aspects of the subject while only dealing with analysis and convergence issues as needed. The reason for this is twofold. First, the algebraic aspects are the most easily generalized from the finite to the infinite vector space setting with a minimal amount of abstraction. The second reason is that the algebraic treatment is mostly constructive, and thus translates easily to computational algorithms. While these



(a) $u(t) = e_1(t) + e_2(t)$ for each t above. Thus $u = e_1 + e_2$ as elements of a signal vector space V where addition is defined pointwise. Here, e_1 and e_2 are depicted as mutually orthogonal in the geometry of V.

(b) A system G maps input signals u to output signals y. Thus $G : \mathsf{V} \longrightarrow \mathsf{W}$ is a mapping between signal vector spaces V and W. Here G amplifies the magnitude (and alters direction) of the input signal as measured by the lengths of the vectors u and y in their respective spaces.

Figure 1: (a) Signals can be viewed as elements of an abstract vector space where the addition of two signals is defined pointwise (i.e. at each time). (b) *Systems* map signals to signals, and they can therefore be viewed as mappings between vector spaces.

notes do not emphasize computational issues, considerations of computational algorithms often lead to more enlightening ways of thinking about the abstract mathematical concepts.

The first three chapters cover the basic fundamentals of linear algebra and functional analysis. The style is to emphasize as much as possible the commonalities between the finite and infinite-dimensional settings. The first chapter deals largely with the purely *algebraic* aspects of the theory of vector spaces including the concepts of null and image subspaces as well as isomorphisms. The second chapter introduces basic *geometric* notions of norms and inner products, i.e. measures of distances and angles in vector spaces. The third chapter covers the basic topological concepts of convergence and completeness, i.e. the *analysis* aspects of the subject. It is here that we begin to see differences between the finite and infinite-dimensional settings, but again an attempt is made to put those differences in a context where they are not as wide as they might initially seem. Some mathematicians take the view [1, 4, 2] that at a very basic level, mathematics can be broadly divided into three branches, namely algebra, geometry and analysis. In this sense the first three chapters are largely organized around these three loose grouping of concepts and techniques.

Another way to think about the organization of the first three chapters is that of overlaying new structures on top of existing ones, like a construction project. The vector space structures of addition, scaling and compatible (i.e. linear) operations is the most basic. On top of that one lays norms and inner products, these are additional structures, but to overlay them, one demands a certain compatibility with the vector space structure. The compatibility is captured by the translation invariance and scaling equivariance of norms. Finally, the third chapter overlays the topology, or equivalently, notions of convergence using the norms already defined.

The fourth chapter deals with the concept of duality. Given any vector space, the set of all linear *functionals* on it forms a vector space itself, called the *dual* space. Linear functionals can be thought of as generalizations of *row vectors*. A linear operator between vector spaces induces another linear operator between their dual spaces called the *adjoint* operator, which is a generalization of the transpose of a matrix. Many properties of vector spaces and mappings between them are best studied by going back and forth between the original and dual spaces, and between the original operator and its adjoint. Linear functionals have geometric interpretations in terms of hyperplanes, and geometric interpretations of adjoints can be given as mappings between hyperplanes. Duality also plays a prominent role in optimization problems. In particular, minimum distance-to-a-subspace problems

have useful characterizations in terms of their duals.

The fifth chapter cover some aspects of spectral theory emphasizing the role of the resolvent function and the concept of the pseudo-spectrum, which in the context of Systems and Controls may be even more important than the spectrum itself. The sixth chapter is about the kernel representation of linear operators, which is an intuitive and graphical way to visualize linear operators on function spaces as continuum analogs of matrices. The seventh chapter introduces matrix and operator partitions, which are useful algebraic tools for block decompositions of operators. Various block operations of LU, UL, and LDU decompositions lead to the frequently used Schur complements, as well as insights into Sylvester and Riccati equations.

Notation and Terminology

The notation $f(x)$ is often used to refer to a function f. This notation can cause confusion. When we refer to the function as a *whole object*, we say the function f, while the notation $f(x)$ refers to the *value of the function f at the point x*. This is like saying $f(1)$ or $f(3)$ to refer to the value of f at the points 1 or 3. Similarly, $f(x)$ means the value of f at the point x, even though x is a variable whose value is not specified. Some textbooks use the symbols $f(x)$ to refer to the whole function f. This is a notational convention, which is used to emphasize that f is not a number or a variable, but rather a function of another variable. As much as possible, we will try to avoid this confusing notation, and use f by itself to refer to the whole function as an object. Sometimes however, a slight abuse of notation may be called for, and we might use for example $U(s)$ to refer to the Laplace transform of a function of time $u(t)$. When such abuse of notation is used, it is simply to point out that the letter s is used to denote the frequency variable and that $U(s)$ is a Laplace transform of some function to distinguish it from $u(t)$ which is a function of the variable t.

Fonts

Sets and spaces	Sets are generally denoted by capital sans serif font, e.g. V as a vector space, P as a cone. Exceptions are for well-known font choices for sets such as $\mathbb{R}$, $\mathbb{C}$, etc.
Matrices/Operators	Matrices and operators are generally denoted by A, B, etc. Abstractly defined operators will sometimes be denoted with caligraphic fonts like $\mathcal{A}$, $\mathcal{B}$, etc.
Vectors	Vectors are generally denoted by small letters like x, or v. Individual components are denoted with subscripts, i.e. the i'th component of the vector x is denoted by x_i. We avoid the common space-saving notation $\begin{bmatrix} x_1^* & \cdots & x_n^* \end{bmatrix}^*$ for column vectors (or block partitioned matrices and operators), and instead use the n-tuple notation where needed for saving space $$(x_1, \ldots, x_n) \;=\; \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}.$$

Specific Notation

$\mathbb{R}$ ($\mathbb{C}$)	The real line (complex plane)
$\mathbb{R}^n$ ($\mathbb{C}^n$)	n-dimensional real (complex) space
$\mathbb{R}(s)$, $\mathbb{I}(s)$	The real and imaginary parts of a complex number s
$\mathbb{Z}$	The integers
$\mathbb{Z}^n$	The n-dimensional integer lattice
$\mathbb{N}$	The natural numbers $N := \{0, 1, 2, \ldots\}$
$\mathfrak{n}$	The set of numbers $\mathfrak{n} := \{1, \ldots, n\}$
$\mathbb{Z}^+$	The positive integers $\{1, 2, \ldots\}$
$\mathbb{Z}^-$	The negative integers $\{-1, -2, \ldots\}$
$\bar{\mathbb{Z}}^+$	The non-negative integers $\{0, 1, 2, \ldots\}$ (same as $\mathbb{N}$)
$\bar{\mathbb{Z}}^-$	The non-positive integers $\{0, -1, -2, \ldots\}$
$\mathbb{C}^-$	The open left half plane $\{s \in \mathbb{C};\ \mathbb{R}(s) < 0\}$
$\mathbb{C}^+$	The open right half plane $\{s \in \mathbb{C};\ \mathbb{R}(s) > 0\}$
$\bar{\mathbb{C}}^-$	The closed left half plane $\{s \in \mathbb{C};\ \mathbb{R}(s) \leq 0\}$
$\bar{\mathbb{C}}^+$	The closed right half plane $\{s \in \mathbb{C};\ \mathbb{R}(s) \geq 0\}$
$\mathbb{D}$	The open unit disk of the complex plane $\{s \in \mathbb{C};\ \lvert s\rvert < 1\}$
$\bar{\mathbb{D}}$	The closed unit disk of the complex plane $\{s \in \mathbb{C};\ \lvert s\rvert \leq 1\}$
$\ell^p_n(\Omega)$	The ℓ^p space of n-vector-valued sequences with index in $\Omega \subseteq \mathbb{Z}^d$ (2.11)
$\ell^p_{\mathsf{V}}(\Omega)$	The ℓ^p space of V-valued sequences (V a Banach space) (2.12)
$\mathsf{L}^{\mathsf{p}}_n(\Omega)$	The L^{p} space of n-vector-valued functions with domain in $\Omega \subseteq \mathbb{R}^d$. It is sometimes abbreviated simply as L^{p} when the dimension n is clear from context, or when n is irrelevant to the argument.
$\mathsf{L}^{\mathsf{p}}_{\mathsf{V}}(\Omega)$	The L^{p} space of V-valued functions (V a Banach space)
A^*	The complex-conjugate (Hermitian) transpose of a matrix A. Also the adjoint of the operator A
$A^\dagger$	The adjoint of a linear operator A. This notation is preferred in cases where the notation A^* could cause confusion.
$\boldsymbol{\lambda}(A)$	The spectrum of a linear operator A. The set of eigenvalues of a matrix A.
$\boldsymbol{\sigma}(A)$	The singular values of a matrix A. The spectrum of the operator AA^* (or A^*A).
$\mathbb{R}^{n\times m}$	The set of $n \times m$ matrices with real entries
$\mathbb{S}^n$	The set of symmetric $n \times n$ matrices with real entries
$\bar{\mathbb{P}}^n$	The set of symmetric positive $n \times n$ matrices $\{A \in \mathbb{S}^n;\ A \geq 0\}$, where $\geq$ is the Loewner order on matrices
$\mathbb{P}^n$	The set of symmetric, strictly-positive $n \times n$ matrices $\{A \in \mathbb{S}^n;\ A > 0\}$

Terminology

functional	any scalar-valued function, i.e. a function $f : \Omega \to \mathbb{R}$ ($\mathbb{C}$) from any set Ω to the scalars (either $\mathbb{R}$ or $\mathbb{C}$)

Chapter 1

Vector Spaces and Linear Operators

Abstract vector spaces are generalizations of the familiar notions of addition and scaling of vectors in two and three dimensional space. A prime example is a function space*, which is a set of functions on a common domain, and those functions can be added and scaled in an analogous manner to vectors in n-dimensional space. Thus, signals can be viewed and manipulated in the same way as vectors. Vector spaces can be built up from other vector spaces by taking direct sums, and can also be decomposed into direct sums of subspaces. Such decompositions are often useful in understanding operations on a vector space by examining the operation on subspaces, over which the operations have simpler structures.*

Mappings between vector spaces that preserve additions and scalings are said to be "linear" or satisfy the "superposition principle". They are generalizations of matrices acting on vectors via matrix-vector multiplication. A matrix is a "representation" of a linear operator in a particular basis, and a basis-free approach allows for a more unified view of linear operators. Many integral and differential operations on functions can be viewed as generalizations of matrix-vector multiplications. A linear mapping between two vector spaces that is also one-to-one and onto is called an isomorphism, and the two vector spaces are said to be isomorphic. Two isomorphic spaces are basically two "copies" of the same space, and thus many properties can be deduced by establishing isomorphisms between familiar and unfamiliar vector spaces.

This chapter is concerned primarily with the basic "algebraic" aspects of vector spaces.

Introduction

This chapter presents some of the basic concepts in linear algebra. The presentation is guided by two principles, (a) whenever possible, a geometric point of view is adopted, and (b) similarities between finite and infinite-dimensional vector spaces are emphasized. The first principle is motivated by the belief that geometric intuition always serves as a powerful reinforcer to the algebraic statements. In this view, linear algebra is fundamentally about the geometry of vector spaces, their subspaces, and the action of linear operators on them.

The second principle is adopted to make as many connections as possible between linear algebra (as normally understood to be about finite-dimensional vector spaces) and functional analysis, a topic normally thought of as dealing with infinite-dimensional vector spaces. We adopt an expansive view of linear algebra, and treat the algebraic aspects of finite and infinite-dimensional vector spaces as much as possible in the "same breath". This serves to introduce many of the powerful techniques of functional analysis (which are very useful for Signals and Systems) using familiar concepts from linear algebra. In particular, the purely "algebraic" aspects of vector spaces lend themselves to this approach. For the "analysis"

aspects, which involve notions of topology and convergence dealt with in later chapters, differences between finite and infinite-dimensional results are pointed out, although again, the emphasis will be on the commonalities, rather than the differences, between finite and infinite-dimensional results.

1.1 $\mathbb{R}^n$ and Abstract Vector Spaces

The space $\mathbb{R}^n$ is the set of n-tuples (n-vectors) of real numbers of the form

$$x \;:=\; (x_1, \ldots, x_n) \;=\; \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix},$$

where each $x_i \in \mathbb{R}$ is a real number. It will be useful to switch notation as convenient, and represent vectors either as an *n-tuple* $(x_1, \ldots, x_n)$, or as a *column vector* as shown above.

In analytic geometry, elements of $\mathbb{R}^n$ are visualized as directed line segments, or *vectors* in n dimensional space. There are two operations on n-tuples that have familiar geometric interpretations.

- n-tuples can be added component-wise. For $x := (x_1, \ldots, x_n)$ and $y := (y_1, \ldots, y_n)$ we define $x + y$ as

$$x + y \;:=\; (x_1 + y_1, \ldots, x_n + y_n)\,. \tag{1.1}$$

  This of course has the familiar geometric interpretation as vector addition when interpreting x and y as directed line segments.

- An n-tuple x can be multiplied by a *scalar* $\alpha \in \mathbb{R}$ by scaling each component by α

$$\alpha\, x \;:=\; (\alpha x_1, \ldots, \alpha x_2)\,. \tag{1.2}$$

  This has the geometric interpretations of scaling the length of the vector x by α, while keeping its direction unchanged if α is positive, or reversing its direction if α is negative. In either case, x and αx lie within the same line passing through the origin.

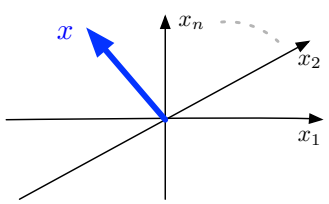
Note that in contrast to additions and scalings, there is "in general" no useful operation of vector multiplication, i.e. a product of two vectors that produces another vector[1]

It is possible to generalize the operations of vector addition and scaling to functions and more general objects. To start this generalization, we make the simple, yet powerful, observation that n-vectors are actually real-valued functions on the discrete set $\mathfrak{n} := \{1, 2, \ldots, n\}$. We can think of a vector in $\mathbb{R}^n$ as either an n-tuple of real numbers, or equivalently as a function from $\mathfrak{n}$ to $\mathbb{R}$

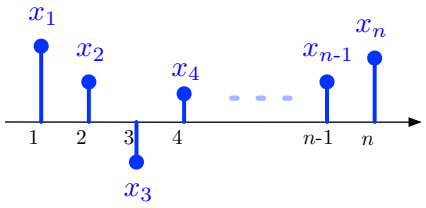
$$x = (x_1, \ldots, x_n) \qquad \longleftrightarrow \qquad x : \{1, \ldots, n\} \longrightarrow \mathbb{R}.$$

Thus the i'th component x_i of a vector $x \in \mathbb{R}^n$ can be viewed as the *value of the function* $x : \mathfrak{n} \longrightarrow \mathbb{R}$ at the index $i \in \mathfrak{n}$. This point of view is illustrated in Figure 1.1.

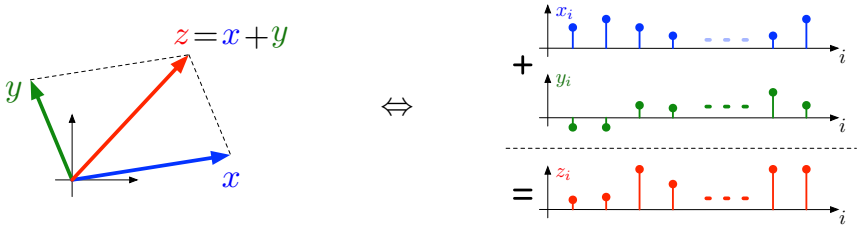
[1] There are very important exceptions, namely complex multiplication on $\mathbb{R}^2$, the cross product on $\mathbb{R}^3$, quaternion product on $\mathbb{R}^4$ and octonian product on $\mathbb{R}^8$. However, these are all special cases that do not generalize to $\mathbb{R}^n$ for all n. If on the other hand we attempt to imitate the definition of point-wise addition (1.1), and define the product operation as the point-wise multiplication of vectors, we run into the problem that division is not well defined if we divide by a vector that has at least one component which is zero.



(a) An n-tuple of real numbers $x := (x_1, \dots, x_n)$ is viewed as a a vector in n-dimensional space, with each scalar x_i being its component along the i'th axis in a Cartesian coordinate system.



(b) An n-tuple of real numbers $x := (x_1, \dots, x_n)$ is viewed as a real-valued function $x : \mathfrak{n} \to \mathbb{R}$ on the discrete index set $\mathfrak{n} := \{1, \dots, n\}$.



(c) Vector addition is defined component-wise, and function addition is defined point-wise. Thus when vectors are viewed as functions, vector addition and function addition are equivalent.

Figure 1.1: Two alternative views of an n-tuple of real numbers $x = (x_1, \dots, x_n)$. (a) As a directed line segment, i.e. a *vector* in n-dimensional space, and (b) as a *function* on the index set $\mathfrak{n} := \{1, \dots, n\}$. Vector addition corresponds exactly to function addition as depicted in (c).

A *function space* is simply a set of functions that have a common domain and range set. We will use the following compact notation to denote such spaces of functions

$$\mathsf{X}^{\Omega} \;:=\; \{u : \Omega \longrightarrow \mathsf{X}\},$$

i.e. the set of all mappings from Ω to X. For example if we view the function x as a signal, then x has Ω as its index set, and at each $i \in \Omega$, the function takes values in the set X, i.e. for each i, $x_i \in \mathsf{X}$. At first this notation might seem a little counter-intuitive since the domain is in the superscript in X^{Ω}, but it should become natural after examining a few examples.

For the vector example shown above, the set of n-vectors as real-valued functions on $\mathfrak{n}$ would be denoted by

$$\mathbb{R}^{\{1,\dots,n\}} \;=\; \mathbb{R}^{\mathfrak{n}} \;=\; \mathbb{R}^n.$$

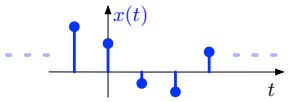
Note that by conventional abuse of notation we abbreviate $\mathbb{R}^{\mathfrak{n}}$ as $\mathbb{R}^n$. If the index set is countable, i.e. $\Omega = \{0, 1, 2, \dots\} = \mathbb{N}$, then $\mathsf{X}^{\{0,1,2,\dots\}}$ can be thought of as a countably-infinite number of ordered copies of X, and thus any element $x \in \mathsf{X}^{\mathbb{N}}$ looks like

$$\mathsf{X}^{\mathbb{N}} \;:=\; \big\{x = (x_0, x_1, x_2, \dots), \quad x_i \in \mathsf{X}\big\} \;=\; \mathsf{X} \times \mathsf{X} \times \mathsf{X} \times \cdots$$

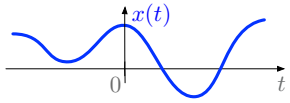
i.e. a sequence with each element in X, or equivalently a function $x : \mathbb{N} \to \mathsf{X}$. When Ω is uncountable (e.g. $\mathbb{R}$) then we can not think of a function like $x : \mathbb{R} \to \mathsf{X}$ as a sequence because the index is in a continuum, but we can still use this notation as in

$$\mathsf{X}^{\mathbb{R}} \;:=\; \big\{x : \mathbb{R} \to \mathsf{X}\big\}.$$

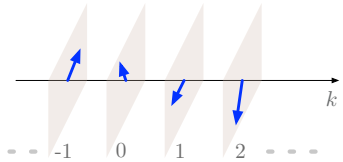
Figure 1.2 shows a few examples of this notation as applied to various signal spaces.



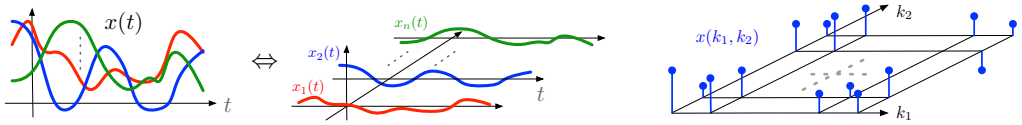
(a) A bilateral, scalar, discrete-time signal takes values in $\mathbb{R}$ for each value of $t \in \mathbb{Z}$. Thus $x : \mathbb{Z} \to \mathbb{R}$, and it is in the function space $\mathbb{R}^{\mathbb{Z}}$.



(b) A bilateral, scalar, continuous-time signal takes values in $\mathbb{R}$ for each value of t that runs from $-\infty$ to ∞. Thus $x : \mathbb{R} \to \mathbb{R}$, and it is in the function space $\mathbb{R}^{\mathbb{R}}$.



(c) A complex Fourier series is a complex-valued function of a discrete variable $k \in \mathbb{Z}$ depicted here as a vector in a complex plane at each k. Thus $x : \mathbb{Z} \to \mathbb{C}$, and therefore lies in the function space $\mathbb{C}^{\mathbb{Z}}$.



(d) A vector signal can be viewed in two equivalent ways. As a vector ($\mathbb{R}^{\mathsf{n}}$)-valued function on $\mathbb{R}$ (*left*), or as an $\mathbb{R}$-valued function on n copies of $\mathbb{R}$, i.e. on $\mathsf{n} \times \mathbb{R}$ (*right*). The corresponding signal spaces would then be $(\mathbb{R}^{\mathsf{n}})^{\mathbb{R}}$ (*left*) and $\mathbb{R}^{(\mathsf{n}\times\mathbb{R})}$ (*right*). The two signal spaces are in one-to-one correspondence.

(e) A grey-scale, $n \times n$ pixel image is an $\mathbb{R}$-valued function on the index set $\{1, \dots, n\} \times \{1, \dots, n\} =: \mathsf{n} \times \mathsf{n}$. Such an image is thus in the function space $\mathbb{R}^{\mathsf{n}\times\mathsf{n}}$.

Figure 1.2: The set of functions $x : \Omega \to \mathsf{X}$ from the set Ω to the set X is denoted by X^{Ω}. If this is thought of as a set of signals, then these signals are indexed by the domain set Ω, and take values in X. This way we can describe, discrete-time and continuous-time signals, as well as those that are scalar-, complex-, or vector-valued.

Function Space as a Vector Space

The two properties of addition and scaling are the fundamental properties of vectors. Real-valued functions can also be added and scaled just like vectors. Let f and g be two real-valued functions, then we typically define addition and scaling pointwise as follows

$$(f+g)(i) \;:=\; f(i)+g(i), \qquad (\alpha f)(i) \;:=\; \alpha f(i). \tag{1.3}$$

Compare those expressions to (1.1) and (1.2). We can define such operations not only on real-valued function, but in fact on any set of functions $f : \Omega \to \mathsf{V}$, where addition and scaling is defined on V itself, e.g. when V is a vector space. Note that the right hand side of the definitions (1.3) are in V. All the examples shown in Figure 1.2 are of this type.

We now return to our original aim of defining a vector space abstractly. The pattern of adding and scaling vectors (or functions) by component-wise (or point-wise) addition and scaling indicates that those two operations are the ones that define a vector space. Inspired by these examples, we define an abstract vector space as follows.

Definition 1.1. *A* vector space *over the reals $\mathbb{R}$ is a set V together with (a) an addition operation $+$, and (b) an operation of vector scaling (by scalars, i.e. elements of $\mathbb{R}$) such that V is closed under both operations*

$$u, v \in \mathsf{V} \quad \Rightarrow \quad u+v \in \mathsf{V}, \qquad\qquad \alpha \in \mathbb{R},\ v \in \mathsf{V}, \quad \Rightarrow \quad \alpha v \in \mathsf{V}.$$

These operations satisfy the following properties

1. *Commutativity and associativity of addition:*

$$u+v \;=\; v+u, \qquad (u+v)+w \;=\; u+(v+w).$$

2. *There exists an additive identity, denoted by* 0*, such that*

$$\forall v \in \mathsf{V}, \quad v + 0 \;=\; v,$$

and each $v \in \mathsf{V}$ *has an additive inverse, denoted by* $-v$*, such that*

$$(-v) + v \;=\; 0.$$

3. *Properties of scalings: For any* $\alpha, \beta \in \mathbb{R}$*, and any* $u, v \in \mathsf{V}$

$$\begin{aligned} (\alpha\beta)\, v &= \alpha\,(\beta v)\,, & (\alpha+\beta)\, v &= \alpha v + \beta v, \\ 1v &= v, & 0v &= 0, \\ \alpha\,(u+v) &= \alpha u + \alpha v. \end{aligned}$$

There are other equivalent ways of stating the above conditions. For example, we leave it as an exercise to show that being closed under addition and multiplication by scalars can be stated in four equivalent ways. Let $\alpha, \beta \in \mathbb{R}$ and $u, v, \in \mathsf{V}$ be arbitrarily elements, then

$$\begin{array}{ccccc} & & (u+v \in \mathsf{V}) \;\text{ and }\; (\alpha u \in \mathsf{V}) & & \\ & & \Updownarrow & & \\ \alpha u + v \in \mathsf{V} & \Leftrightarrow & \alpha u + \beta v \in \mathsf{V} & \Leftrightarrow & u + \beta v \in \mathsf{V} \end{array}$$

In addition, there are several consequences of the definitions above. For example, the scalings properties imply that every vector $v \in \mathsf{V}$ has an additive inverse, namely $(-1)v$ since

$$v + (-1)v \;=\; (1)v + (-1)v \;=\; (1-1)v \;=\; 0v \;=\; 0.$$

As an exercise, the reader should verify every equality in the preceding equation using the scalings properties listed in Definition 1.1.*3*.

Example 1.2. As examples of vector spaces, we have already seen $\mathbb{R}^n$. Now consider the set of all polynomials with real coefficients of degree at most n

$$\mathbb{P}_n \;:=\; \Big\{ p(x) = a_0 + a_1 x + \cdots + a_n x^n;\; a_k \in \mathbb{R},\; k = 0, 1, \cdots, n \Big\}.$$

Each polynomial in $\mathbb{P}_n$ is uniquely identified by its $n+1$ coefficients $(a_0, \ldots, a_n)$, and therefore there is a one-to-one correspondence between polynomials of degree n and vectors of dimension $n+1$. A natural question is whether this correspondence carries over to addition. Does the addition of two polynomials correspond to the addition of the two vectors? This is indeed the case since for any two polynomials p and q

$$\left.\begin{array}{l} p(x) = a_0 + a_1 x + \cdots + a_n x^n \\ q(x) = b_0 + b_1 x + \cdots + b_n x^n \end{array}\right\} \;\Rightarrow\; (p+q)\,(x) \;=\; (a_0{+}b_0){+}(a_1{+}b_1)\, x{+}\cdots{+}(a_n{+}b_n)\, x^n.$$

Similarly for scaling of a polynomial

$$(\alpha p)\,(x) \;=\; \alpha a_0 + \alpha a_1\; x + \cdots + \alpha a_n\; x^n.$$

Thus elements in $\mathbb{P}_n$ behave exactly like elements of $\mathbb{R}^{n+1}$ under additions and scaling. We say that these two vector spaces $\mathbb{P}_n$ and $\mathbb{R}^{n+1}$ are "isomorphic", a notion that will shortly be defined more precisely.

The function space examples we have seen in this section (e.g. in Figure 1.2) so far satisfy all the above requirements with pointwise addition and scaling. In fact, for any set Ω, the function space $\mathbb{R}^\Omega$ is a vector space with pointwise addition and scaling. We can take this a little further. Let V be any vector space, and Ω any other set (with no particular structure), then the set V^Ω is another vector space with pointwise addition and scaling.

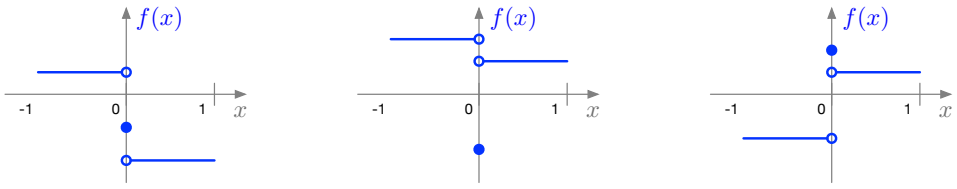


Figure 1.3: Examples of the vector space of functions over $[-1,1]$ that are piecewise constant over the intervals $[-1,0)$ and $(0,1]$. This is a vector space since additions and scalings of such functions maintain the piecewise-constant property. Each such function is completely specified by three real numbers, namely its values over $x=0$, and each of the two intervals respectively. Additions and scalings of these functions corresponds to additions and scalings of these numbers as a 3-tuple. Thus this space "looks like" (i.e. isomorphic to) $\mathbb{R}^3$.

Definition 1.3. *Let Ω be any set, and consider the collection of all vector-space-valued functions $f:\Omega\to\mathsf{V}$, where V is a vector space (e.g. $\mathbb{R}$ or $\mathbb{R}^n$)*

$$\mathsf{V}^\Omega \;:=\; \{f:\Omega\to\mathsf{V}\}\,.$$

This set is itself a vector space referred to as a function space *(the space of functions over the set Ω) with additions and scalings defined pointwise*

$$\begin{aligned}\big(f_1+f_2\big)(x) \;&:=\; f_1(x)\;+\;f_2(x)\\ \big(\alpha f\big)(x) \;&:=\; \alpha f(x).\end{aligned}$$

Note that both operations on the right hand side are performed in V, and thus we say that V^Ω "inherits" the vector space structure from V.

Function spaces as defined above are generally "too big" for many of the questions that arise in applications, so we typically impose additional conditions on functions and consider *subspaces* of a general function space. A subspace of a vector space is defined as a subset that is itself a vector space, i.e. closed under additions and scalings. The following examples are of subspaces of function spaces as defined above.

Example 1.4. Consider real-valued functions on the interval $[-1,1]$ that are constant on the subintervals $[-1,0)$ and $(0,1]$ as depicted in Figure 1.3. It is clear that additions and scalings of such functions maintain the property of being constant on the subintervals $[-1,0)$ and $(0,1]$, and thus the space of such functions is a vector space. In the notation of Definition 1.3 this space is denoted by $\mathbb{R}^\Omega$, where $\Omega:=\{[\text{-}1,0),0,(0,1]\}$, a three-element set where each element is a subset of $\mathbb{R}$. This space is clearly a subspace of the much larger function space $\mathbb{R}^{[\text{-}1,1]}$ (all real-valued functions on the interval $[\text{-}1,1]$).

Any function in $\mathbb{R}^{\{[\text{-}1,0),0,(0,1]\}}$ is completely described by a 3-tuple of numbers $(f_{\text{-}1},f_0,f_1)$ as follows

$$f(x) \;=\; \begin{cases} f_{\text{-}1}, & x\in[\text{-}1,0)\\ f_0, & x=0\\ f_1, & x\in(0,1]\end{cases}\quad. \tag{1.4}$$

Adding and scaling any two such functions corresponds to simply adding and scaling the 3-tuples that represent them

$$f(x)=\begin{cases} f_{\text{-}1}, & x\in[\text{-}1,0)\\ f_0, & x=0\\ f_1, & x\in(0,1]\end{cases},\qquad g(x)=\begin{cases} g_{\text{-}1}, & x\in[\text{-}1,0)\\ g_0, & x=0\\ g_1, & x\in(0,1]\end{cases}$$

$$\Rightarrow\qquad (f+g)(x)=\begin{cases} f_{\text{-}1}+g_{\text{-}1}, & x\in[\text{-}1,0)\\ f_0+g_0, & x=0\\ f_1+g_1, & x\in(0,1]\end{cases}.$$

This space of functions therefore "looks like" the space of 3-tuples of real numbers $\mathbb{R}^3$. The meaning of the phrase "looks like" is that there is a correspondence, or a one-to-one and onto mapping (indicated below by the double arrow $\leftrightarrow$) between $\mathbb{R}^{\{[-1,0),0,(0,1]\}}$ and $\mathbb{R}^3$

$$\begin{aligned} f &\quad\leftrightarrow\quad (f_{-1}, f_0, f_1) \\ g &\quad\leftrightarrow\quad (g_{-1}, g_0, g_1) \\ \alpha f + \beta g &\quad\leftrightarrow\quad \alpha(f_{-1}, f_0, f_1) + \beta(g_{-1}, g_0, g_1) = (\alpha f_{-1} + \beta g_{-1}, \alpha f_0 + \beta g_0, \alpha f_1 + \beta g_1). \end{aligned}$$

The last statement implies that vector space operations in $\mathbb{R}^{\{[-1,0),0,(0,1]\}}$ are mapped exactly to equivalent vector space operations in $\mathbb{R}^3$. Again, this notion will be formalized using the concept of isomorphism that will be described shortly.

Example 1.5. Let $\mathbb{R}^{n\times m}$ be the set of all $n \times m$ matrices with real coefficients. This set is a vector space with the usual operations of matrix addition and scaling which are defined "element-by-element", i.e. if we denote by a_{ij} the ij'th entry of a matrix $A = [a_{ij}]$, then for $n \times m$ matrices A, B and C

$$C \;=\; \alpha A + \beta B \qquad \Leftrightarrow \qquad c_{ij} \;=\; \alpha a_{ij} + \beta b_{ij}.$$

Let vec(A) be the operation of "vectorizing" a matrix, i.e. stacking all its columns into a single vector. It is clear that this operation is linear, one-to-one and onto from $n \times m$ matrices to vectors of size nm, i.e. an invertible linear mapping vec : $\mathbb{R}^{n\times m} \longrightarrow \mathbb{R}^{nm}$. Thus as a vector space, the space of $n\times m$ matrices behaves like (i.e. isomorphic) to the space of nm real vectors.

While the mapping vec is compatible with matrix addition and scalings, it does not make sense for matrix multiplication (since we can't multiply vectors). It does however have some uses in matrix equations we will encounter later.

Example 1.6. Continuous functions on a finite interval $[a, b] \subset \mathbb{R}$ form a vector space since sums and scalings of continuous functions are also continuous. This space is denoted by

$$\mathsf{C}[a, b] \;:=\; \Big\{ f : [a, b] \to \mathbb{R};\; f \text{ continuous} \Big\}.$$

Functions on $[a, b]$ with *continuous derivatives* also form a vector space denoted by $\mathsf{C}^1[a, b]$. More generally, the space of functions with n continuous derivatives is

$$\mathsf{C}^n[a, b] \;:=\; \Big\{ f : [a, b] \to \mathbb{R};\; f^{(n)} \text{ continuous} \Big\}, \qquad\qquad n \in \mathbb{N}.$$

Recall that if the n'th derivative $f^{(n)}$ of a function f is continuous, then the k'th derivative $f^{(k)}$ is also continuous for all $k = 0, 1, \ldots, n$. Thus we have a "nesting of vector spaces

$$\mathsf{C}^n[a, b] \;\subset\; \mathsf{C}^{n\text{-}1} \;\subset\; \cdots \;\subset\; \mathsf{C}^1[a, b] \;\subset\; \mathsf{C}[a, b].$$

Finally, $\mathsf{C}^\infty[a, b]$ denotes the vector space of functions on $[a, b]$ with all derivatives continuous. It can also be described as the intersection of all vector spaces $\mathsf{C}^k[a, b]$ for all k

$$\mathsf{C}^\infty[a, b] \;:=\; \bigcap_{k=0}^{\infty} \mathsf{C}^k[a, b].$$

All of the vector space just described are subspaces of the larger function space $\mathbb{R}^{[a,b]}$.

In contrast to the $\mathbb{R}^n$ and $\mathbb{P}_n$, each element of the space $\mathsf{C}[a, b]$ requires an infinite number of "parameters" to describe (e.g. one has to give the values of the function at the infinite number of points in $[a, b]$). This is an example of an "infinite-dimensional" vector space, a notion that we will make precise in Section 1.3.

Complex Vector Spaces

A *complex vector space* is defined exactly as in Definition 1.1 by replacing the set of scalars $\mathbb{R}$ with the complex scalars $\mathbb{C}$. More generally, let $\mathbb{F}$ be any field, then a vector space over the field $\mathbb{F}$ is defined the same way, but with scalars belonging to the field $\mathbb{F}$. For this to make sense, the scaling operation αx, with $\alpha \in \mathbb{F}$ and $x \in \mathsf{V}$ must be defined. In particular, for any index set Ω, the space $\mathbb{F}^\Omega$ is a well-defined vector space (e.g. $\mathbb{F}^n$ is the space of n-vectors with components in $\mathbb{F}$). Note that in Definition 1.3, V^Ω is a vector space over the same field that V is a vector space over.

In this book, we only consider vector spaces over $\mathbb{R}$ or $\mathbb{C}$.

1.2 Linear Operators

The reader is likely familiar with matrix-vector multiplication as an example of a linear operation on vectors. We will see that matrices are actually *representations* of linear operators in a particular basis. We will also see many useful examples of bases, and other representations of linear operators. However, we begin here with a "basis-free" way of defining and analyzing linear operators. In fact, the key idea in viewing matrices and general linear operators in the same framework is to work with them without any specific choice of bases.

Definition 1.7. *Let $A : \mathsf{V} \to \mathsf{W}$ be a mapping between two vector spaces V and W. This mapping is said to be a* homomorphism*, or equivalently a* linear operator *if*

$$\alpha, \beta \in \mathbb{R},\ v_1, v_2 \in \mathsf{V} \qquad \Rightarrow \qquad A(\alpha v_1 + \beta v_2) \ = \ \alpha A(v_1) + \beta A(v_2), \tag{1.5}$$

i.e. if it "respects the vector space structure" (also referred to as satisfying the "superposition property"). If in addition A is one-to-one and onto, then it is said to be an isomorphism*, and the two spaces V and W are said to be* isomorphic *(denoted as $\mathsf{V} \sim \mathsf{W}$).*

Thus a linear operator is a mapping that is "compatible" with the vector space structure of V and W. Note that for linear operators, we use the notation $A(v) = Av$ interchangeably[2]. The reader should verify that condition (1.5) is equivalent to the following condition

$$v_1, v_2 \in \mathsf{V} \ \Rightarrow \ A(v_1{+}v_2) \ = \ A(v_1){+}A(v_2), \qquad \text{and} \qquad \alpha \in \mathbb{R} \ \Rightarrow \ A(\alpha v) \ = \ \alpha\, A(v). \tag{1.6}$$

Remark 1.8. An isomorphism between two vector spaces is special. If two vector spaces are isomorphic, then they are essentially *two copies of the same space* since their addition and scaling structures are equivalent. Finding isomorphisms is usually a good way to understand an unfamiliar space, by establishing an isomorphism to a more familiar space. Example 1.2 was one such instance, where an isomorphism between polynomials of degree n, namely $\mathbb{P}_n$ and $n+1$ dimensional vectors in $\mathbb{R}^{n+1}$ was established. Also note that we have already used this concept (without calling it an isomorphism) when we presented the analogy between $\mathbb{R}^n$ (n-tuples of real numbers) and $\mathbb{R}^{\{1,\dots,n\}}$ (real-valued functions on the set $\{1,\dots,n\}$) in Figure 1.1. That correspondence was so obvious that we didn't have to formally justify it. It is clearly an isomorphism.

The following are examples of linear operators. They are generally not isomorphisms (i.e. not one-to-one and onto).

Example 1.9. *Matrices:* An $n \times m$ matrix A represents a mapping $A : \mathbb{R}^m \to \mathbb{R}^n$ by the usual matrix-vector product formula

$$y = Ax \qquad \Leftrightarrow \qquad y_k \ = \ \sum_{l=1}^{m} A_{kl}\, x_l, \tag{1.7}$$

[2]This comes from matrix-vector multiplication notation. If A is matrix and v is a column vector, the standard notation is Av for the matrix-vector product, which is also the action of A on v as a mapping.

where A_{kl} is the kl entry of the matrix A. The fact that this operation is linear follows from the distributive properties of multiplication and addition. This particular example is not done in a "basis-free" way. A matrix representation implies an implicit choice of basis as will be discussed in Section 1.3.

Example 1.10. *Shifts of functions:* Consider the function space $\mathbb{R}^{\mathbb{Z}}$ of real-valued functions on the integers (i.e. the space of two-sided sequences). The "shift operator" $\mathcal{S}$ shifts a function to the right by 1 step

$$[\mathcal{S}u](k) \;:=\; u(k-1).$$

This operator is clearly linear as seen from

$$\begin{aligned}\big[\mathcal{S}\big(\alpha u_1+\beta u_2\big)\big](k) \;&=\; [\alpha u_1+\beta u_2](k-1) \;=\; \alpha\; u_1(k-1) \;+\; \beta\; u_2(k-1)\\ &=\; \alpha\,[\mathcal{S}u_1](k) \;+\; \beta\,[\mathcal{S}u_2](k).\end{aligned}$$

Since this equality holds for each k in the domain $\mathbb{Z}$ of the functions, we write the conclusion equivalently as

$$\mathcal{S}\big(\alpha u_1+\beta u_2\big) \;=\; \alpha\;\mathcal{S}u_1 \;+\; \beta\;\mathcal{S}u_2.$$

Note that a basis is not needed to define the operator or to establish its linearity.

Example 1.11. *Multiplication operators:* Consider the vector space $\mathbb{R}^{\Omega}$ of real-valued functions[3] on some set Ω. Given a particular function $a:\Omega\to\mathbb{R}$ in this space, we can define the operator M_a of multiplication by a that acts on any other function $u:\Omega\to\mathbb{R}$ by

$$[M_a u](x) \;:=\; a(x)\;u(x), \qquad x\in\Omega. \tag{1.8}$$

Thus M_a is the operator of *pointwise multiplication* by the function a. The fact that this operator is linear follows from the distributive property of multiplication and addition of real numbers.

A familiar example of multiplication operators is given by the action of diagonal matrices on vectors. If we choose $\Omega=\{1,\dots,n\}$, and therefore $\mathbb{R}^{\Omega}=\mathbb{R}^n$, then given a vector $a\in\mathbb{R}^n$, it defines a multiplication operator M_a whose action on any other vector $u\in\mathbb{R}^n$ is given by

$$\begin{aligned} y \;&=\; M_a\;u\\ y_i \;&=\; (M_a u)_i = a_i u_i,\quad i\in\{1,\dots,n\}\end{aligned} \qquad\Leftrightarrow\qquad \begin{bmatrix} y_1\\ \vdots\\ y_n\end{bmatrix} = \begin{bmatrix} a_1 & & \\ & \ddots & \\ & & a_n\end{bmatrix}\begin{bmatrix} u_1\\ \vdots\\ u_n\end{bmatrix}.$$

Thus the operation $M_a u$ is represented by multiplying the column vector u by a diagonal matrix whose diagonal elements are made up of the components of the vector a.

The definition (1.8) of a multiplication operator on any function space should be thought of as a generalization of diagonal matrices. Diagonal matrices are the simplest, non-trivial matrices that can be studied. Similarly, multiplication operators are the simplest, non-trivial infinite-dimensional operators that can be studied. The concept of multiplication operators will be very useful in signal and system analysis later on. Signal and system transforms such as the Fourier, Laplace, Z-transform, etc. can be thought of as the infinite-dimensional analogue of diagonalizing matrices. Diagonalization of matrices, and more generally transforming linear operators to multiplication operators (whenever possible) is the most effective technique for uncovering properties of linear operators.

[3] All statements in this example apply equally to the vector space $\mathbb{C}^{\Omega}$ of complex-valued functions on Ω.

Example 1.12. *Integral operators:* Consider the space $\mathbb{R}^{[0,1]}$ of real-valued functions on the interval $[0,1]$. Given a function $a(.,.)$ of two variables, it defines an *integral operator* by

$$y = Au \qquad \Leftrightarrow \qquad y(x) = \int_0^1 a(x,\xi)\ u(\xi)\ d\xi, \qquad x \in [0,1]. \tag{1.9}$$

Since the integral may not converge for all functions u, this operator is not defined on all of $\mathbb{R}^{[0,1]}$, but rather on a subset of it. The exact specifications of the subset depends on properties of the function a, and will not be discusses here. The function $a(.,.)$ is called the *kernel function* of the operator A, and the integral equation (1.9) is called the *kernel representation* of the operator A. These kernel representations are important in understanding a large class of operators, and will be studied in detail in Chapter 6.

The reader should note the similarity between the matrix-vector product (1.7) and the integral (1.9), which can be thought of as a "continuum" version of a matrix-vector product. The integration variable ξ in (1.9) is analogous to the summation variable l in (1.7). Thus an integral, or kernel, representation of an operator on a function space can be thought of as a "continuum" version of a matrix representation. These useful analogies will be pursued further in Chapter 6.

Finally, the linearity of (1.9) follows from the linearity of integration, e.g.

$$\int_0^1 a(x,\xi)\ \Big(u_1(\xi) + u_2(\xi)\Big)\, d\xi = \int_0^1 a(x,\xi)\ u_1(\xi)\ d\xi + \int_0^1 a(x,\xi)\ u_2(\xi)\ d\xi. \qquad \blacksquare$$

Example 1.13. *Differential operators:* Differentiation is a linear operation

$$\tfrac{d}{dx}\Big(\alpha u_1(x)\ +\ \beta u_2(x)\Big)\ =\ \alpha\frac{du_1}{dx}(x)\ +\ \beta\frac{du_2}{dx}(x),$$

and so are higher order derivatives as well as partial derivatives. Thus, a large class of ordinary and partial differential equations, including those with varying coefficients, can be analyzed using the concepts of linear operators on function spaces. The ordinary differential operator (of a single variable) is defined formally as

$$(\mathbf{D}u)(x)\ :=\ \frac{du}{dx}(x).$$

The next question is on which vector spaces of functions does this operator act?

Recall the spaces $C[a,b]$ and $\mathsf{C}^{(n)}[a,b]$ defined in Example 1.6. The differential operator $\mathbf{D}$ acts on functions with n continuous derivatives to give a function with $n-1$ continuous derivatives. Therefore depending on the choice of the domain space, we can view $\mathbf{D}$ in any one of multiple ways

$$\mathbf{D} : \mathsf{C}^1[a,b] \longrightarrow \mathsf{C}[a,b], \qquad \mathbf{D} : \mathsf{C}^n[a,b] \longrightarrow \mathsf{C}^{n\text{-}1}[a,b], \qquad \mathbf{D} : \mathsf{C}^\infty[a,b] \longrightarrow \mathsf{C}^\infty[a,b].$$

Differential operators can also be defined on other (than C^n or C^∞) spaces. However, a little more care is needed in those cases since they have to operate on restricted classes of functions, namely those that have derivatives with certain other properties. These issues will be discussed in the context of so-called *unbounded operators*, of which differential operators are the prime example.

The Vector Space of Linear Operators

Recall that in Example 1.5 we showed how the set of matrices of a given size forms a vector space. This was a special case of the fact that the set of all linear operators between two vector spaces V and W is itself a vector space

$$\mathsf{L}(\mathsf{V},\mathsf{W})\ :=\ \Big\{A : \mathsf{V} \to \mathsf{W};\ A \text{ linear}\Big\}.$$

Addition and scalings of operators is defined "pointwise"

$$(A+B)\,v \;:=\; Av \;+\; Bv, \qquad (\alpha A)\,v \;:=\; \alpha\, Av.$$

The observant reader will realize that we do not need V to be a vector space for these definitions to make sense. Recall (Definition 1.3) that any function space Ω^{W} is a vector space provided the functions take values in a vector space W. The domain Ω of the functions need not have any structure. None the less, the set of linear operators between two vector spaces has some special structure which we will study later (Section 3.6), especially when the operators will be endowed with norms that are induced from the two vector space norms.

1.3 Bases and Dimension

For clarity of exposition, in this section (and later in Section 1.6) we adopt the notation that elements of a vector space are denoted by bold letters (e.g. $\boldsymbol{v}$), and scalars will be denoted in regular font (e.g. x_1). We will not use this notation in the remainder of the book, but here it serves as a useful visual aid when discussing bases and bases representations for the first time. To define bases, we need the concept of linear (in)dependence, and for that in turn we need the concept of the span of a collection of vectors.

Definition 1.14. *Let* $\mathsf{v} := \{\boldsymbol{v}_k;\ k = 1, \dots, n\}$ *be a finite set in a vector space* V. *The* span *of* v *is the set of a all linear combinations of its elements*

$$\mathsf{span}\{\mathsf{v}\} := \;\mathsf{span}\{\boldsymbol{v}_k\} := \;\left\{x_1\boldsymbol{v}_1 + \cdots + x_n\boldsymbol{v}_n;\ x_1, \dots, x_n \in \mathbb{R}\right\}.$$

It is a linear subspace of V. *We say that the set* v generates *(or* spans*)* $\mathsf{span}\{\mathsf{v}\}$.

Given a set of vectors, we want to generate its span without any "redundancy". For example, consider any two vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ and their span. We can add a third vector that is a linear combination of both, but then the three vectors generate the same span as the original two

$$\mathsf{span}\{\boldsymbol{v}_1, \boldsymbol{v}_2\} \;=\; \mathsf{span}\{\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_1 + \boldsymbol{v}_2\}.$$

To avoid such redundancies, we need the concept of linear (in)dependence. We first give the formal definition, and then explain why it captures the notion of no redundancy.

Definition 1.15. *Let* $\mathsf{v} := \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$ *be a finite set of elements of a vector space* V.

1. *The set* v *is said to be* linearly independent *if there is no non-trivial linear combination (i.e. not all coefficients are zero) of its elements that equals the zero vector, i.e.*

   $$x_k \in \mathbb{R}, \qquad x_1\boldsymbol{v}_1 + \cdots + x_n\boldsymbol{v}_n = 0 \qquad \Rightarrow \qquad x_1 = \cdots = x_n = 0.$$

   Otherwise, the set is said to be linearly dependent.

2. *A linearly independent set that generates all of* V *(i.e.* $\mathsf{span}\{\boldsymbol{v}_k\} = \mathsf{V}$*) is called a* basis.

A *linearly dependent* set has the property that one of the vectors can be written as a linear combination of the others. Given such a set, assume $x_l \neq 0$, then

$$x_1\boldsymbol{v}_1 + \cdots + x_n\boldsymbol{v}_n = 0 \qquad \Rightarrow \qquad \boldsymbol{v}_l \;=\; -\frac{1}{x_l}\sum_{k\neq l} x_k\boldsymbol{v}_k. \tag{1.10}$$

Thus a linearly dependent set has redundancy. If we remove the vector $\boldsymbol{v}_l$ above from the set, the smaller set will still generate the same span as the original set. Given a linearly dependent set, we can remove elements as above until "it becomes" linearly independent. This implies that a linearly independent set has a "minimality property" which can be stated as follows.

Lemma 1.16. *Let* $\mathsf{v} := \{\boldsymbol{v}_k;\ k = 1, \ldots, n\}$ *be a finite set that spans a vector space* V. v *is linearly independent iff there is no smaller set* $\mathsf{w} := \{\boldsymbol{w}_k;\ k = 1, \ldots, m\}$ *with* $|\mathsf{w}| = m < n = |\mathsf{v}|$ *that also spans* V.

Proof. The contrapositive is stated as

$$\mathsf{v} \text{ linearly dependent} \quad \Leftrightarrow \quad \exists \mathsf{w},\ |\mathsf{w}| < n,\ \mathsf{span}\{\mathsf{w}\} = \mathsf{span}\{\mathsf{v}\}\,.$$

The argument of Equation (1.10) showed that if v is linearly dependent, a smaller set with the same span is constructed by removing one element from v. This proves the $\Rightarrow$ direction. For the converse, assume we found a smaller set w with the same span as v. That means every element of v can be written as a linear combination of elements of w

$$\boldsymbol{v}_k \;=\; a_{k1}\boldsymbol{w}_1 + \cdots + a_{km}\boldsymbol{w}_m, \qquad k = 1, \ldots, n. \tag{1.11}$$

These relations can be written in "matrix-vector" form shown on the left below[4], and by a process of elimination with row operations (see Exercise 1.1), converted to the equations on the right

$$\begin{bmatrix} \boldsymbol{v}_1 \\ \vdots \\ \boldsymbol{v}_n \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_1 \\ \vdots \\ \boldsymbol{w}_m \end{bmatrix} \qquad \Leftrightarrow \qquad \begin{bmatrix} \boldsymbol{v}_{m+1} \\ \vdots \\ \boldsymbol{v}_n \end{bmatrix} = \begin{bmatrix} * & \cdots & * \\ \vdots & & \vdots \\ * & \cdots & * \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1 \\ \vdots \\ \boldsymbol{v}_m \end{bmatrix}. \tag{1.12}$$

Thus the vectors $\{\boldsymbol{v}_{m+1}, \ldots, \boldsymbol{v}_n\}$ can be written as a linear combination of the vectors $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m\}$ implying that the set v is linearly dependent. □

Another way to read the above lemma is that a set is linearly independent iff removal of any vector makes the span strictly smaller.

The preceding lemma implies that a basis is a generating set of *minimal size*. We can now formally define the notion of dimension of a vector space.

Definition 1.17. *If* $\mathsf{v} := \{\boldsymbol{v}_k;\ k = 1, \ldots, n\}$ *is a linearly independent set that spans a vector space* V *(i.e. a basis), we say that the* dimension of V is n *(*$\mathsf{dim}(\mathsf{V}) = n$*). Thus the dimension of a vector space is the minimal size of a linearly independent set that spans* V.
If there is no finite subset of V *that spans it, we say that* V *is* infinite dimensional.

Example 1.18. Consider $\mathbb{R}^n$ and the following set of vectors

$$\boldsymbol{e}_1 := (1, 0, \ldots, 0)\,, \qquad \cdots\,, \qquad \boldsymbol{e}_n := (0, \ldots, 0, 1)\,. \tag{1.13}$$

This set is linearly independent according to Definition 1.15, and any vector $\boldsymbol{x} \in \mathbb{R}^n$ can be written as a linear combination of its elements. Thus, it forms a basis for $\mathbb{R}^n$. We typically express the coefficients of the linear combination as a "column vector"

$$\boldsymbol{x} \;=\; x_1\boldsymbol{e}_1 + \cdots + x_n\boldsymbol{e}_n \;=\; x_1 \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \cdots + \; x_n \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \;=\; \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \;=\; (x_1, \ldots, x_n)\,. \tag{1.14}$$

The set (1.13) is called the *canonical basis* of $\mathbb{R}^n$. Thus, whenever a column vector or an n-tuple of numbers is written as on the right of (1.14), it is implicit that the vector components are the coefficients of the vector's representation in the canonical basis $\{\boldsymbol{e}_k\}$.

[4] This matrix-vector representation of the equations (1.11) is to be interpreted carefully. Each "component" $\boldsymbol{v}_k$ and $\boldsymbol{w}_k$ are themselves vectors in V. The matrix-vector form (1.12) is simply a compact representation of the n equations in (1.11).

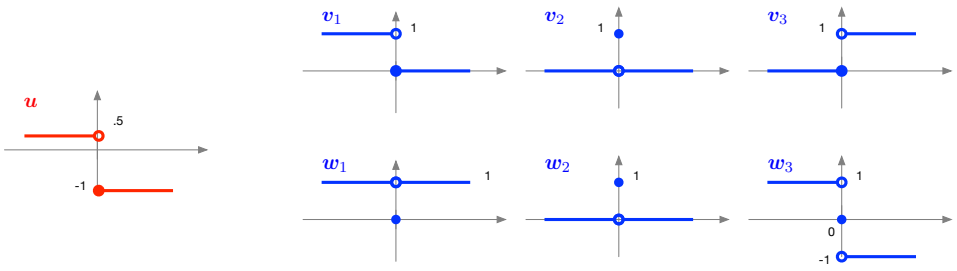


Figure 1.4: Example 1.19: Two different bases $\mathsf{v} := \{\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3\}$ and $\mathsf{w} := \{\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3\}$ of the vector space $\mathbb{R}^{\{[\text{-}1,0),0,(0,1]\}}$ of Example 1.4. The function $\boldsymbol{u}$ shown on the left can be expressed in terms of each of the bases sets as $\boldsymbol{u} = .5\boldsymbol{v}_1 - \boldsymbol{v}_2 - \boldsymbol{v}_3 = -.25\boldsymbol{w}_1 - \boldsymbol{w}_2 + .75\boldsymbol{w}_3$.

Example 1.19. Consider the vector space of Example 1.4. When we established the correspondence (1.4) with $\mathbb{R}^3$, we had implicitly chosen a basis. To make this point clear, consider the two sets of functions $\mathsf{v} := \{\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3\}$ and $\mathsf{w} := \{\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3\}$ shown in Figure 1.4. These two sets are linearly independent in the vector space $\mathbb{R}^{\{[\text{-}1,0),0,(0,1]\}}$. Each is also a basis. For example, the function $\boldsymbol{u}$ shown in Figure 1.4 can be written in each of the two bases as follows

$$\boldsymbol{u} \;=\; .5\boldsymbol{v}_1 \;-\; \boldsymbol{v}_2 \;-\; \boldsymbol{v}_3 \;=\; -.25\boldsymbol{w}_1 \;-\; \boldsymbol{w}_2 \;+\; .75\boldsymbol{w}_3.$$

The two bases sets are easily relatable by a set of equations similar to (1.12) as follows

$$\begin{bmatrix}\boldsymbol{w}_1\\ \boldsymbol{w}_2\\ \boldsymbol{w}_3\end{bmatrix} = \begin{bmatrix}1 & 0 & 1\\ 0 & 1 & 0\\ 1 & 0 & \text{-}1\end{bmatrix}\begin{bmatrix}\boldsymbol{v}_1\\ \boldsymbol{v}_2\\ \boldsymbol{v}_3\end{bmatrix} \quad\Leftrightarrow\quad \begin{bmatrix}\boldsymbol{v}_1\\ \boldsymbol{v}_2\\ \boldsymbol{v}_3\end{bmatrix} = \begin{bmatrix}1 & 0 & 1\\ 0 & 1 & 0\\ 1 & 0 & \text{-}1\end{bmatrix}^{\text{-}1}\begin{bmatrix}\boldsymbol{w}_1\\ \boldsymbol{w}_2\\ \boldsymbol{w}_3\end{bmatrix} = \begin{bmatrix}\frac{1}{2} & 0 & \frac{1}{2}\\ 0 & 1 & 0\\ \frac{1}{2} & 0 & \frac{\text{-}1}{2}\end{bmatrix}\begin{bmatrix}\boldsymbol{w}_1\\ \boldsymbol{w}_2\\ \boldsymbol{w}_3\end{bmatrix} \tag{1.15}$$

The first equation can be verified by inspection of Figure 1.4. The second equation follows by multiplying both sides by the inverse of the matrix of coefficients.

The representation (1.4) exactly corresponds to using the first basis set, i.e. for the function f as defined in (1.4) we can write

$$f \;=\; f_{\text{-}1}\boldsymbol{v}_1 \;+\; f_0\boldsymbol{v}_2 \;+\; f_1\boldsymbol{v}_3.$$

Once an expression for any function f is given in terms of the basis v, then the coefficients of its representation in the other basis w can be found from the relation (1.15) as follows

$$\begin{aligned} f &= f_{\text{-}1}\boldsymbol{v}_1 + f_0\boldsymbol{v}_2 + f_1\boldsymbol{v}_3\\ &= \hat{f}_{\text{-}1}\boldsymbol{w}_1 + \hat{f}_0\boldsymbol{w}_2 + \hat{f}_1\boldsymbol{w}_3\\ &= \hat{f}_{\text{-}1}(\boldsymbol{v}_1 + \boldsymbol{v}_3) + \hat{f}_0\boldsymbol{v}_2 + \hat{f}_1(\boldsymbol{v}_1 - \boldsymbol{v}_3)\\ &= \left(\hat{f}_{\text{-}1} + \hat{f}_1\right)\boldsymbol{v}_1 + \hat{f}_0\boldsymbol{v}_1 + \left(\hat{f}_{\text{-}1} - \hat{f}_1\right)\boldsymbol{v}_3. \end{aligned} \tag{1.16}$$

Note how (1.16) give two sets of coefficients for the representation of f in the v basis. Lemma 1.42 will shortly show that those two sets of coefficients must be equal, i.e. that

$$\begin{aligned} \hat{f}_{\text{-}1} + \hat{f}_1 &= f_{\text{-}1},\\ \hat{f}_{\text{-}1} - \hat{f}_1 &= f_1 \end{aligned} \quad\Leftrightarrow\quad \begin{bmatrix}f_{\text{-}1}\\ f_0\\ f_1\end{bmatrix} = \begin{bmatrix}1 & 0 & 1\\ 0 & 1 & 0\\ 1 & 0 & \text{-}1\end{bmatrix}\begin{bmatrix}\hat{f}_{\text{-}1}\\ \hat{f}_0\\ \hat{f}_1\end{bmatrix}. \tag{1.17}$$

it is clear that to obtain the coefficients $(\hat{f}_{\text{-}1}, \hat{f}_0, \hat{f}_1)$ from $(f_{\text{-}1}, f_0, f_1)$, we simply invert the matrix-vector relation above. We will see in Lemma 1.44 that this procedure generalizes to any finite-dimensional vector space.

The basis v might seem as the "natural" one for this vector space, but that is only one possible choice. Any other set of three linearly independent functions in $\mathbb{R}^{\{[-1,0),0,(0,1]\}}$ is equally valid as a basis.

The Infinite-dimensional Case

For infinite-dimensional vector spaces, the notion of basis will need to involve more than just algebraic properties in order for it to be useful. We will need to make sense of representations that involve infinite sums like

$$\boldsymbol{u} = \sum_{k=1}^{\infty} x_k \boldsymbol{v}_k. \tag{1.18}$$

This will require notions of convergence and topology which will be introduced in Chapter 3. There is however a definition of bases in infinite dimensions that is purely algebraic, and is referred to as a *Hamel basis*. Because it is purely algebraic, it turns out not to be too useful, but we describe it briefly in the next example for the sake of contrast with later definitions.

Definition 1.20. *Let* $\mathsf{v} := \{\boldsymbol{v}_k\}$ *be a (possibly infinite) set in a vector space* V.

1. *The* span *of* v *is the set of all* finite *linear combinations of its elements*

$$\mathsf{span}\{\mathsf{v}\} := \left\{ \sum_{k=1}^{n} x_k \boldsymbol{v}_k;\ \boldsymbol{v}_k \in \mathsf{v},\ x_k \in \mathbb{R},\ n \in \mathbb{N} \right\}.$$

2. *The set* v *is called* linearly independent *if no non-trivial* finite *linear combination of elements of* v *is zero.*

3. *A (possibly infinite) linearly independent set* $\mathsf{v} \subset \mathsf{V}$ *that spans a (possibly infinite-dimensional) vector space* V *is called a* Hamel basis.

The reason for using only *finite* linear combinations in the above definitions is that unless we have a notion of convergence, only finite sums are well defined. The next examples clarify this issue.

Example 1.21. Consider the vector space of real sequences $\mathbb{R}^{\mathbb{Z}}$, and consider the set

$$\boldsymbol{e}_k := (\ \cdots,\ 0,\ \underset{\uparrow\ k\text{'th entry}}{1},\ 0,\ \cdots\ ), \qquad k \in \mathbb{Z}. \tag{1.19}$$

This set is easily seen to be linearly independent. It is an infinite set. It does not span $\mathbb{R}^{\mathbb{Z}}$ however since every finite linear combination of such elements can only give a sequence with finitely many non-zero entries

$$\mathsf{span}\{\boldsymbol{e}_k\} \ = \ \text{sequences with finitely many non-zero entries} \ \subsetneq \ \mathbb{R}^{\mathbb{Z}}.$$

Therefore, $\mathsf{span}\{\boldsymbol{e}_k\}$ is an infinite-dimensional subspace of $\mathbb{R}^{\mathbb{Z}}$, and $\{\boldsymbol{e}_k\}$ is a Hamel basis for that subspace.

Example 1.22. Consider linear the set of all finite linear combinations of harmonic functions of arbitrary frequencies

$$\mathsf{V} := \left\{ f : \mathbb{R} \to \mathbb{C};\ f(t) = \alpha_1 e^{j\omega_1 t} + \cdots + \alpha_n e^{j\omega_n t},\ \alpha_i, \omega_i \in \mathbb{R},\ n \in \mathbb{N} \right\}.$$

This set is an infinite-dimensional vector space (over the scalars in $\mathbb{C}$). According to Definition 1.20, it is the span of the following set of functions indexed by $\mathbb{R}$

$$\mathsf{V} \;=\; \mathsf{span}\{\mathsf{v}\} \;:=\; \mathsf{span}\big\{e^{j\omega t};\ \omega \in \mathbb{R}\big\}\,.$$

The set v is linearly independent since no non-trivial linear combination of such elements can be the zero function provided the frequencies $\omega_1, \ldots, \omega_n$ are all distinct. Thus the set $\mathsf{v} := \big\{e^{j\omega t};\ \omega \in \mathbb{R}\big\}$ is a Hamel basis for V (almost by definition). Note that this is an *uncountable* basis since it has the same cardinality as $\mathbb{R}$. We will return to this example in later chapters as it forms the backbone of so-called "almost periodic functions".

A more useful concept introduced in later chapters involves the same basis (1.19). We will however define norms on subspaces of $\mathbb{R}^{\mathbb{Z}}$, and take the "closure" of $\mathsf{span}\{\boldsymbol{e}_k\}$ with respect to those norms. This will yield for example the $\ell^p(\mathbb{Z})$ spaces. By taking the closure, we can make sense of a series like (1.18), in that finite partial sums converge (in the defined norm) to the element $\boldsymbol{u}$. This discussion will have to wait until we define norms and convergence properties.

1.4 Subspaces, Direct Sums and Quotients

A subset $\mathsf{S} \subseteq \mathsf{V}$ of a vector space V that is itself a vector space is called a *subspace* of V. The set S has to therefore satisfy all the properties listed in Definition 1.1. Some of these properties are automatically inherited from V, namely, commutativity and associativity of vector addition, as well as the properties of scalings. Thus we only need to explicitly require the property of *closure* under linear combinations.

Definition 1.23. *A subset* $\mathsf{S} \subseteq \mathsf{V}$ *of a vector space is called a* subspace *if it is closed under linear combinations*

$$\forall \alpha, \beta \in \mathbb{R}, \quad x, y \in \mathsf{S} \quad \Rightarrow \quad \alpha x + \beta y \in \mathsf{S}.$$

In particular $0 \in \mathsf{S}$*, and* S *is itself a vector space.*

Example 1.24. Consider the set of "zero mean" vectors in $\mathbb{R}^n$

$$\mathsf{S} \;:=\; \left\{ x \in \mathbb{R}^n;\ \sum_{i=1}^{n} x_i = 0 \right\}.$$

Due to the linearity of sums, this subset is clearly closed under linear combinations. If $x, y \in \mathsf{S}$, then

$$\sum_{i=1}^{n} (\alpha x + \beta y)_i \;=\; \sum_{i=1}^{n} (\alpha x_i + \beta y_i) \;=\; \alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} y_i \;=\; 0 + 0.$$

Note that the zero mean condition can also be written as $\mathbb{1}^* x = 0$, where $\mathbb{1}^*$ is the transpose of $\mathbb{1}$, the vector whose entries are all 1. The reader should verify as an exercise that for any fixed vector $v \in \mathbb{R}^n$, the set $\{x \in \mathbb{R}^n;\ v^* x = 0\}$ is indeed a subspace. Thus the example above is a special case of this more general type of subspace.

Vector spaces can be "collected together" to form bigger vector spaces that contain them. This is the concept of the direct sum of vector spaces. There are two versions of the direct sum concept depending on whether we consider the vector spaces to be combined as unrelated, or if they are subspaces of some larger vector space. We begin with the former concept.

Definition 1.25. *Given two vector spaces* V_1 *and* V_2*, their* external direct sum *is the set of ordered pairs*

$$\mathsf{V}_1 \oplus_{\mathrm{e}} \mathsf{V}_2 \;:=\; \Big\{(v_1, v_2);\; v_1 \in \mathsf{V}_1, v_2 \in \mathsf{V}_2\Big\},$$

where the vector space operations are defined "componentwise", i.e.

$$(v_1, v_2) + (u_1, u_2) := (v_1 + u_1, v_2 + u_2), \qquad \alpha(v_1, v_2) := (\alpha v_1, \alpha v_2).$$

The basic example is that of $\mathbb{R}^2 = \mathbb{R} \oplus_{\mathrm{e}} \mathbb{R}$, which is ordered pairs of real numbers.

Each vector space V_1 and V_2 is embedded in $\mathsf{V}_1 \oplus_{\mathrm{e}} \mathsf{V}_2$ as a subspace since e.g.

$$\mathsf{S} \;:=\; \Big\{(v_1, 0) \in \mathsf{V}_1 \oplus_{\mathrm{e}} \mathsf{V}_2;\; v_1 \in \mathsf{V}_1\Big\}$$

is clearly a set closed under additions and scalings. S is itself a vector space which can be "identified" with V_1, i.e. S is isomorphic to V_1. Now let's look at another way we can combine *subspaces* of a given vector space.

Definition 1.26. *Given two subspaces* $\mathsf{S}_1, \mathsf{S}_2 \subseteq \mathsf{V}$ *of a vector space* V*, their* internal direct sum *is*

$$\mathsf{S}_1 \oplus_{\mathrm{i}} \mathsf{S}_2 \;:=\; \Big\{v \in \mathsf{V};\; v = v_1 + v_2,\; v_1 \in \mathsf{S}_1,\; v_2 \in \mathsf{S}_2\Big\}.$$

This is the set of all possible linear combinations of elements from S_1 *and* S_2*.*

Note the difference between this definition and the previous one. In the latter, one can make sense of the sum $v_1 + v_2$ since both vectors are in V, in which addition is well defined. The distinction in terminology between external and internal direct sums was made to emphasize this difference.

Consider for example the plane $\mathbb{R}^2$ and the subspace S which is the x-axis. The internal direct sum $\mathbb{R}^2 \oplus_{\mathrm{i}} \mathsf{S}$ is just all of $\mathbb{R}^2$ (adding a vector aligned with the x-axis to any other vector in the plane gives a vector in the plane). However, when we take an external direct sum, $\mathbb{R}^2$ and S are considered as vector spaces in their own right, and not as subspaces of another vector space. In this case therefore $\mathbb{R}^2 \oplus_{\mathrm{e}} \mathsf{S} = \mathbb{R}^3$ since it is the set of ordered 3-tuples (two coordinates come from $\mathbb{R}^2$, and the third coordinate is from S, which is just the set of real numbers).

The internal and external direct sums are equal, or more precisely they are isomorphic, if any vector can written as $v = v_1 + v_2$ in a unique way. This turns out to be equivalent to the two subspaces having only the trivial intersection.

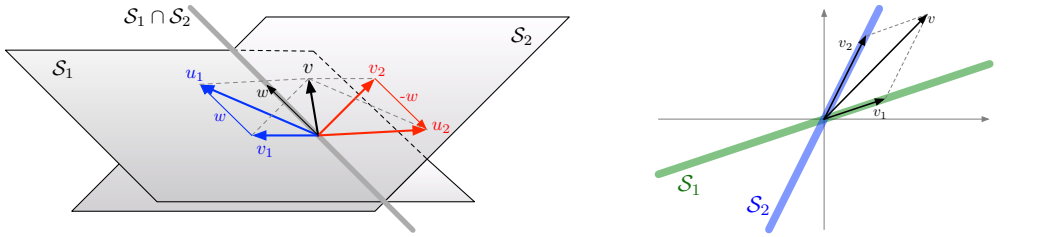
Lemma 1.27. *Let* $\mathsf{S}_1, \mathsf{S}_2 \subseteq \mathsf{V}$ *be subspaces of a vector space* V*.*

1. $\mathsf{S}_1 \cap \mathsf{S}_2 = 0$ *iff any vector* v *in the internal direct sum can be written* uniquely *as* $v = v_1 + v_2$ *with* $v_i \in \mathsf{S}_i$*.*
2. *If* $\mathsf{S}_1 \cap \mathsf{S}_2 = 0$*, then the internal sum* $\mathsf{S}_1 \oplus_{\mathrm{i}} \mathsf{S}_2$ *and the external sum* $\mathsf{S}_1 \oplus_{\mathrm{e}} \mathsf{S}_2$ *are isomorphic. In this case we simply write* $\mathsf{S}_1 \oplus \mathsf{S}_2$ *for either sum.*

Proof. 1. We show the contrapositive. Suppose $\mathsf{S}_1 \cap \mathsf{S}_2 \neq 0$, and select a non-zero vector w from it. Note that $w \in \mathsf{S}_1$ and $w \in \mathsf{S}_2$, and v can be alternatively decomposed as

$$v \;=\; v_1 + v_2 \;=\; (v_1 + w) + (v_2 - w), \qquad (v_1 + w) \in \mathsf{S}_1,\; (v_2 - w) \in \mathsf{S}_2,$$

which is another, distinct (since $w \neq 0$), representation of v as the sum of two vectors from S_1 and S_2 respectively. This geometry illustrated in Figure 1.5a.



(a) Two subspaces S_1 and S_2 with a non-trivial intersection $\mathsf{S}_1 \cap \mathsf{S}_2 \neq 0$ (here depicted as the thick grey line). Because the intersection is non-trivial, any vector of the form $v = v_1 + v_2$ (where $v_1 \in \mathsf{S}_1$ and $v_2 \in \mathsf{S}_2$) can be rewritten in an infinite number of other ways as the sum of two vectors from S_1 and S_2 respectively. Given any non-zero vector $w \in \mathsf{S}_1 \cap \mathsf{S}_2$, then $v = v_1 + v_2 = (v_1 + w) + (v_2 - w)$ is another distinct decomposition of v.

(b) A subspace S_2 is *complementary* to another subspace S_1 if $\mathsf{S}_1 \oplus \mathsf{S}_2 = \mathsf{V}$, and $\mathsf{S}_1 \cap \mathsf{S}_2 = 0$. This last condition insures that any vector $v \in \mathsf{V}$ can be written *uniquely* as $v = v_1 + v_2$ with $v_1 \in \mathsf{S}_1$ and $v_2 \in \mathsf{S}_2$.

Figure 1.5: Illustrations of the concepts of internal direct sum and complementary subspaces.

Conversely, if there exists two distinct representations

$$\begin{array}{l} v = v_1 + v_2 \\ \quad = \hat{v}_1 + \hat{v}_2 \end{array}, \quad \begin{array}{l} v_1, \hat{v}_1 \in \mathsf{S}_1 \\ v_2, \hat{v}_2 \in \mathsf{S}_2 \end{array}, \quad \begin{array}{l} v_1 \neq \hat{v}_1 \\ \text{or } v_2 \neq \hat{v}_2 \end{array} \quad \Rightarrow \quad \begin{cases} (v_1 + v_2) \text{-} (\hat{v}_1 + \hat{v}_2) = 0 \\ \Leftrightarrow \ (v_1 \text{-} \hat{v}_1) = (v_2 \text{-} \hat{v}_2) =: w \end{cases}$$

and this non-zero vector w belongs to both S_1 and S_2.

2. When $\mathsf{S}_1 \cap \mathsf{S}_2 = 0$, the unique decomposition $v = v_1 + v_2$ gives a mapping A between the external and internal sums

$$A : \mathsf{S}_1 \oplus_{\mathrm{e}} \mathsf{S}_2 \to \mathsf{S}_1 \oplus_{\mathrm{i}} \mathsf{S}_2, \qquad A(v_1, v_2) := v_1 + v_2.$$

The uniqueness of the decomposition implies that this map is one-to-one. It is clearly onto and linear, and therefore an isomorphism. □

It will be assumed going forward (unless otherwise stated) that when taking direct sums of subspaces, their intersection is 0. The next concept deals with decomposing a vector space into a direct sum of subspaces.

Definition 1.28. *Consider a vector space* V *with a subspace* $\mathsf{S}_1 \subset \mathsf{V}$*. A subspace* S_2 *is said to be* complementary *to* S_1 *if their intersection is zero, and their direct sum is all of* V

$$\mathsf{S}_2 \text{ complementary subspace to } \mathsf{S}_1 \quad \Leftrightarrow \quad \mathsf{S}_1 \cap \mathsf{S}_2 = 0, \text{ and } \mathsf{S}_1 \oplus \mathsf{S}_2 = \mathsf{V}.$$

This concept is illustrated in Figure 1.5b. Complementary subspaces are not unique. There is always an infinite number of choices of subspaces that are complementary to any given subspace[5].

Example 1.29. Consider the space $\mathbb{R}^{\mathbb{R}}$ of functions on the real line. We will give two different decompositions of it into complementary subspaces. First, recall that any function u on the real line can be written uniquely as the sum of its odd and even parts

$$u_{\mathrm{o}}(t) := u(t) - u(-t), \qquad u_{\mathrm{e}}(t) := u(t) + u(-t), \qquad \Rightarrow \qquad u(t) \ = \ (u_{\mathrm{o}}(t) + u_{\mathrm{e}}(t))/2.$$

[5]It is important to mention that a complementary subspace S_2 does not have to be "orthogonal" to S_1. The concept of orthogonal complements will be discussed later when inner products, and therefore the notion of orthogonality, are introduced. Orthogonal complements are however unique.

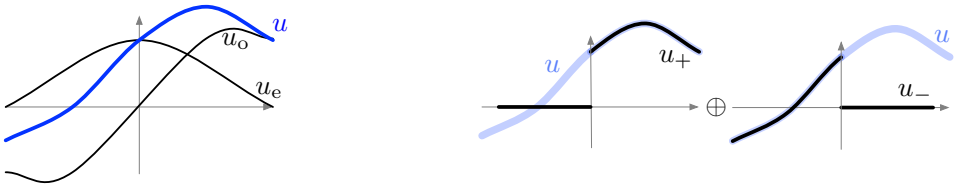


Figure 1.6: The decompositions of Example 1.29 of the space of functions on the real line $\mathbb{R}^{\mathbb{R}}$ into two complementary subspaces in two different ways. (*Left*) Any function can be written uniquely as $u(t) = u_{\mathrm{o}}(t) + u_{\mathrm{e}}(t)$, the sum of its odd and even parts respectively. (*Right*) Any function can be written uniquely as $u(t) = u_+(t) + u_-(t)$, the sum of functions supported on $[0, \infty)$ and $(-\infty, 0)$ respectively. Thus the vector space $\mathbb{R}^{\mathbb{R}}$ is isomorphic to the direct sum $\mathbb{R}^{[0,\infty)} \oplus \mathbb{R}^{(-\infty,0)}$.

Let $\left(\mathbb{R}^{\mathbb{R}}\right)_{\mathrm{o}}$ and $\left(\mathbb{R}^{\mathbb{R}}\right)_{\mathrm{e}}$ refer to the subspaces of odd and even functions on $\mathbb{R}$ respectively (convince yourself that they are indeed *subspaces*). Their intersection is zero since any function that is both even and odd must satisfy

$$u(-t) = u(t) = -u(t) \qquad \Rightarrow \qquad u(t) = 0.$$

Thus $\mathbb{R}^{\mathbb{R}} = \left(\mathbb{R}^{\mathbb{R}}\right)_{\mathrm{o}} \oplus \left(\mathbb{R}^{\mathbb{R}}\right)_{\mathrm{e}}$ is a decomposition into complementary subspaces. This decomposition is illustrated in Figure 1.6

Alternatively, consider the subspaces of $\mathbb{R}^{\mathbb{R}}$ of functions that are supported on the non-negative and negative real lines respectively

$$\left(\mathbb{R}^{\mathbb{R}}\right)_+ := \left\{u : \mathbb{R} \to \mathbb{R};\ u(t) = 0,\ \text{if } t < 0\right\}, \qquad \left(\mathbb{R}^{\mathbb{R}}\right)_- := \left\{u : \mathbb{R} \to \mathbb{R};\ u(t) = 0,\ \text{if } t \geq 0\right\}.$$

Any function can be decomposed uniquely into its corresponding positively and negatively supported parts

$$u_+(t) := \begin{cases} u(t), & t \geq 0, \\ 0, & t < 0, \end{cases} \qquad u_-(t) := \begin{cases} 0, & t \geq 0, \\ u(t), & t < 0, \end{cases} \qquad \Rightarrow \qquad u(t) = u_+(t) + u_-(t).$$

This decomposition is illustrated in Figure 1.6. The intersection of $\left(\mathbb{R}^{\mathbb{R}}\right)_+$ and $\left(\mathbb{R}^{\mathbb{R}}\right)_-$ is clearly zero, and therefore they are complementary subspaces of $\mathbb{R}^{\mathbb{R}}$.

Finally, we note that this last decomposition gives a useful illustration of the correspondence between internal and external direct sums. Consider the space $\mathbb{R}^{[0,\infty)}$ of functions defined on $[0, \infty)$. It is clearly isomorphic to the subspace $\left(\mathbb{R}^{\mathbb{R}}\right)_+$, but they are not equal. The former contains functions that are defined only on the interval $[0, \infty)$, while the latter has functions defined on all of $\mathbb{R}$, but constrained to be zero over $(-\infty, 0)$. Similarly for $\mathbb{R}^{(-\infty,0)}$. These isomorphisms can be summarized as follows

$$\begin{array}{rcl} \left(\mathbb{R}^{\mathbb{R}}\right)_+ & \sim & \mathbb{R}^{[0,\infty)} \\ \left(\mathbb{R}^{\mathbb{R}}\right)_- & \sim & \mathbb{R}^{(-\infty,0)} \end{array} \qquad \text{and} \qquad \mathbb{R}^{\mathbb{R}} = \left(\mathbb{R}^{\mathbb{R}}\right)_+ \oplus_{\mathrm{i}} \left(\mathbb{R}^{\mathbb{R}}\right)_- \qquad \Rightarrow \qquad \mathbb{R}^{\mathbb{R}} \sim \mathbb{R}^{[0,\infty)} \oplus_{\mathrm{e}} \mathbb{R}^{(-\infty,0)}.$$

Thus $\mathbb{R}^{\mathbb{R}}$ is equal to the internal direct sum of two complementary subspaces, while it is *isomorphic* to the external direct sum of two separately defined vector spaces. Once this notion is understood, we will not make a distinction in the sequel between internal and external direct sums (provided the intersection of the subspaces is zero), and we may simply write $\mathbb{R}^{\mathbb{R}} = \mathbb{R}^{[0,\infty)} \oplus \mathbb{R}^{(-\infty,0)}$ with the understanding that this is equality "up to isomorphisms".

Projections

Whenever a vector space V has a decomposition in terms of complementary subspaces S_1 and S_2, then *projection operators* onto those subspaces are defined as follows. Since every

$v \in \mathsf{V}$ has a unique decomposition as the sum of two vectors

$$v \;=\; v_1 + v_2, \qquad v_1 \in \mathsf{S}_1,\; v_2 \in \mathsf{S}_2,$$

this gives the well-defined mappings

$$\begin{array}{ll} \Pi_1 : \mathsf{V} \to \mathsf{S}_1 \\ \Pi_2 : \mathsf{V} \to \mathsf{S}_2 \end{array}, \qquad \begin{array}{lcl} \Pi_1 v & := & v_1 \\ \Pi_2 v & := & v_2 \end{array}.$$

These mappings are clearly linear. For example, for any two elements $v, w \in \mathsf{V}$, each has a unique decomposition $v = v_1 + v_2$, $w = w_1 + w_2$, and therefore $v + w = (v_1 + w_1) + (v_2 + w_2)$ is a unique decomposition, and consequently $\Pi_1(v + w) = v_1 + w_2 = \Pi_1 v + \Pi_1 w$.

Projections can be visualized as in Figure 1.5b, and are generally called *oblique projections*. In inner product spaces (to be introduced later), there is a notion of orthogonality, and when the complementary subspaces are orthogonal, then the projections Π_1 and Π_2 are called orthogonal projections.

The projections Π_i are linear operators, and they have a very special property. If we apply the projection twice to the same vector $\Pi^2 v = \Pi(\Pi v) = \Pi v_1 = v_1$ (since v_1 is already in S_1). This property is a defining property of projections as the following statement implies.

Lemma 1.30. *Let $\Pi : \mathsf{V} \to \mathsf{S}$ be a linear operator from a vector space V to a subspace $\mathsf{S} \subset \mathsf{V}$. If $\Pi^2 = \Pi$, then Π is a projection, and the linear operator*

$$\Pi_{\mathrm{c}} \;:=\; (I - \Pi),$$

is the complementary projection *which maps $\Pi_{\mathrm{c}} : \mathsf{V} \to \mathsf{S}_{\mathrm{c}}$, where $\mathsf{S}_c := \mathsf{Im}(\Pi_{\mathrm{c}})$ is a subspace complementary to S in V.*

Proof. First note that the two mappings Π and $(I - \Pi)$ give a decomposition of any vector

$$v \;=\; (\Pi + I - \Pi)\, v \;=\; \Pi v \;+\; (I - \Pi)v \;=\; v_1 + v_2, \qquad v_1 \in \mathsf{S},\; v_2 \in \mathsf{S}_{\mathrm{c}}.$$

To show that S and S_{c} are complementary, we need to show that their intersection is 0. Indeed, suppose a vector $v \in \mathsf{S}$ and $v \in \mathsf{S}_{\mathrm{c}}$

$$\Pi v = v, \text{ and } (I - \Pi)v = v \qquad \Rightarrow \qquad v = (I - \Pi)v = v - \Pi v = v - v = 0.$$

Thus only the 0 vector is in $\mathsf{S} \cap \mathsf{S}_{\mathrm{c}}$ and therefore the two subspaces are complementary. □

The lemma gives a technique for finding complementary subspaces. If we find an operator that is equal to its square $\Pi^2 = \Pi$, then the lemma guarantees that $\mathsf{Im}(\Pi)$ and $\mathsf{Im}(I - \Pi)$ are complementary subspaces, and the decomposition of any vector can be obtained by applying those two operators to the vector.

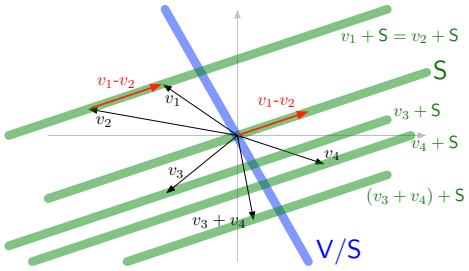
Example 1.31. Returning to Example 1.29, consider again the space of all functions on the real line $\mathsf{V} = \mathbb{R}^{\mathbb{R}}$. Define the operator Π which "zeros out" the part of a function on the negative real line $(-\infty, 0)$

$$(\Pi u)(t) \;:=\; \begin{cases} 0, & t < 0, \\ u(t), & t \geq 0. \end{cases}$$

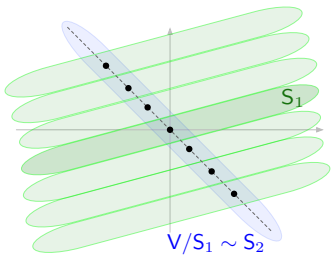
This operator clearly has the property $\Pi^2 = \Pi$. It is easy to see that its complementary projection $(I - \Pi)$ "zeros out" the part of a function supported on $[0, \infty)$

$$(I - \Pi)u \;=\; u \;-\; \Pi u \;=\; u \;-\; u_+ \;=\; u_- \;:=\; \begin{cases} u(t), & t < 0, \\ 0, & t \geq 0 \end{cases},$$

where we are using the notation of Example 1.29 for the positively u_+ and negatively u_- supported portions of u respectively.



(a) *Cosets* of a subspace S are sets of the form $v+S$ (depicted in green) for any vector v, called a *coset representative*. Representatives are not unique, any two vectors with $v_1 - v_2 \in S$ represent the same coset, i.e. $v_1 + S = v_2 + S$. At the top, the sum $(v_1 + S) + (v_2 + S) = (v_1 + v_2) + S$ is depicted. The *set of all cosets* of any subspace is itself a vector space with addition and scaling induced from the original space V. This set is denoted by V/S (depicted in blue), the *quotient space* of V by S.



(b) The set of all cosets V/S_1 can be identified with (i.e. is isomorphic to) any space S_2 that is complementary to S_1. Complementary subspaces S_1 and S_2 intersect only at 0, and it then follows that S_2 intersects each coset at exactly one point. Each such point is taken as the representative of its containing coset, and this gives the isomorphism $S_2 \sim V/S_1$.

Figure 1.7: The concepts of cosets of a subspace and the corresponding quotient space. The quotient of V by S_1, denoted V/S_1 is isomorphic to any subspace S_2 that is complementary to S_1.

1.4.1 Cosets and Quotient Spaces

The problem of finding complementary subspaces is a fundamental one that occurs often, and we will introduce various techniques for addressing it depending on the setting. At a more abstract level, there is a generic technique for constructing a complementary subspace using the notion of cosets of subspaces.

Definition 1.32. *Given a subspace $S \subset V$, a* coset *of S is a set of the form*

$$v + S \;:=\; \{v + x;\; x \in S\}\,,$$

where $v \in V$ is any vector, which is called a representative *of the coset $v + S$.*

Note that coset representatives are not unique. For example given any coset $v+S$, adding an element $u \in S$ to v gives the same coset

$$u \in S \qquad \Rightarrow \qquad (v+u)+S = \{(v+u)+x;\; x \in S\} = \{v+y;\; y \in S\} = v+S \quad (1.20)$$

Thus v and $v+u$ represent the same coset if $u \in S$. Also note that if the representative is actually a member of S, then its coset is just the "zero coset", which is the subspace itself

$$v \in S \qquad \Rightarrow \qquad v + S \;=\; 0 + S \;=\; S.$$

Cosets can be visualized as "affine shifts" of the subspace as illustrated in Figure 1.7a. The figure also illustrates another way to see (1.20). Two vectors v_1 and v_2 whose difference lies in S represent the same coset

$$(v_1 - v_2) \in S \qquad \Leftrightarrow \qquad v_1 + S \;=\; v_2 + S.$$

We now consider the *set of cosets* of a given subspace S. First observe that belonging to a coset is an *equivalence relation*, i.e. if v_1 and v_2 belong to a coset, v_2 and v_3 belong to the same coset, then clearly v_1 and v_3 belong to that coset. Therefore cosets of a given subspace partitions V into non-intersecting subsets whose union is all of V.

The set of cosets *is itself a vector space* with the natural definition of addition as

$$\begin{aligned}(v_1 + S) \;+\; (v_2 + S) \;&:=\; (v_1 + v_2) + S \\ \alpha\,(v + S) \;&:=\; \alpha v + S.\end{aligned}$$

These definitions are independent of the choice of cosets representatives, e.g.

$$\begin{aligned} v_1 + \mathsf{S} &= u_1 + \mathsf{S} \quad \Rightarrow \quad u_1 - v_1 \in \mathsf{S} \\ v_2 + \mathsf{S} &= u_2 + \mathsf{S} \quad \Rightarrow \quad u_2 - v_2 \in \mathsf{S} \\ \therefore \quad (v_1 + v_2) + \mathsf{S} &= (v_1 + v_2) + (u_1 - v_1) + (u_2 - v_2) + \mathsf{S} = (u_1 + u_2) + \mathsf{S}. \end{aligned}$$

Figure 1.7a depicts several cosets of a subspace S. These cosets are visualized as a "layered" collection of affine spaces, which can be added and scaled in the same manner as their coset representatives.

The set of all cosets of a subspace $\mathsf{S} \subset \mathsf{V}$ is a vector space called the *quotient space* V/S. Figure 1.7a illustrates that V/S can be thought of as a subspace complementary to S. We make this precise in the next statement.

Lemma 1.33. *Let $\mathsf{S}_1 \subset \mathsf{V}$ be subspace of a vector space V, and $\mathsf{S}_2 \subset \mathsf{V}$ be a complementary subspace. Then S_2 is isomorphic to the quotient space V/S_1. Thus for any subspace $\mathsf{S} \subset \mathsf{V}$, the vector space decomposes as $\mathsf{V} \sim \mathsf{S} \oplus \mathsf{V}/\mathsf{S}$.*

Note that this lemma also implies that all subspaces complementary to a given subspace are isomorphic to each other, and to V/S_1 in particular. The argument for this lemma is best illustrated by Figure 1.7b. Consider all cosets of S_1 (which partition the entire space V), and their intersection points with the complementary subspace S_2. Since S_1 and S_2 intersect at only the single point 0, then it follows that S_2 intersects each coset at exactly one point. Indeed, if $v_1 \neq v_2$ belong to the same coset (i.e. $v_1 - v_2 \in \mathsf{S}_1$), and they also belong to S_2, then $v_1 - v_2 \in \mathsf{S}_2$ (since S_2 is a subspace), and therefore $v_1 - v_2 \neq 0$ belongs to $\mathsf{S}_1 \cap \mathsf{S}_2$, which means that S_2 is not complementary to S_1.

Another perspective on this lemma is given by reexamination of Figure 1.7b. One can always choose one representative from each coset of S_1, and the collection of such representatives forms a set that is in one-to-one correspondence with V/S_1. However, in a vector space, those representatives can be chosen so that they themselves form a subspace, namely the subspace S_2. We note that this process of choosing representatives is not unique, as is the choice of complementary subspaces.

1.5 Image/Null Subspaces and Linear Equations Solvability

The theory of vector spaces and linear operators provides a powerful framework for treatment of the many of fundamental equations in science and engineering. These vary from matrix-vector equations to (ordinary or partial) differential equations as well as integral equations. If a vector space structure can be found in which the equations involve linear operators, then this theory is applicable. It is also often true that for nonlinear equations, analysis of the linear parts provides significant insight into properties of the solutions of the overall equations.

The first questions about equations are those of solvability, i.e. when do there exist solutions, and if they do, are they unique? If not, can one characterize all possible solutions? Regardless of the details of the equations (e.g. whether they involve matrices, differential or integral operators), the notions of image and null spaces are fundamental to answering solvability questions for linear equations. We first motivate the formal definitions.

When do there exist solutions to a linear equation? Consider a linear operator $A : \mathsf{V} \longrightarrow \mathsf{W}$ between vector spaces and the equation

$$Ax = b, \tag{1.21}$$

where $b \in \mathsf{W}$ is some given vector. It's almost a tautology to say that there exists a solution iff there exists a vector $\bar{x} \in \mathsf{V}$ such that $A\bar{x} = b$. We therefore are motivated to define the

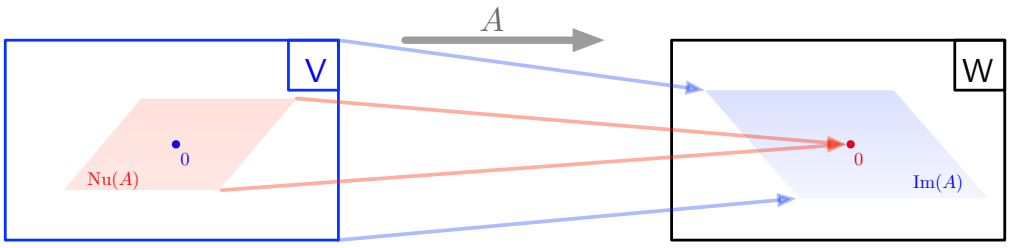


Figure 1.8: Illustration of the image $\mathsf{Im}(A)$ and null $\mathsf{Nu}(A)$ spaces of a linear operator A between two vector spaces V and W. The image space (depicted in light blue on the right) is elements of W that are images of any elements in V. It is a linear subspace of W. Clearly A is onto iff $\mathsf{Im}(A) = \mathsf{W}$ (i.e. the image space of A "fills up" all of W). The null space is the set of all elements that are mapped to 0 (depicted in light red on the left). It is a linear subspace of V. The null space always contains 0, but if it contains other elements, i.e. $\mathsf{Nu}(A) \neq 0$, then A is clearly not one-to-one (more than one element is mapped to the same element, namely zero). It is also true that $\mathsf{Nu}(A) = 0$ iff A is one-to-one.

set of all vectors in W for which there exists vectors in V mapped to them. This is the concept of the "image space".

Definition 1.34. *Given a linear operator $A : \mathsf{V} \to \mathsf{W}$ between two vector spaces V and W, the* image space $\mathsf{Im}(A) \subseteq \mathsf{W}$ *of A is*

$$\mathsf{Im}(A) := \{w \in \mathsf{W};\ \exists v \in \mathsf{V}, Av = w\}$$

As illustrated in Figure 1.8, the image space is the range of A as a function (i.e. all the elements of W that have at least one element of V mapped to them). The fact that it is a subspace (rather than an arbitrary set) is a consequence of the linearity of the mapping A

$$\begin{aligned} w_1 = Av_1,\ w_2 = Av_2 \quad &\Rightarrow \quad A(\alpha v_2 + \beta v_2) \ = \ \alpha\, Av_1 + \beta\, Av_2 \ = \ \alpha w_1 + \beta w_2, \\ \therefore\ \ w_1, w_2 \in \mathsf{Im}(A) \quad &\Rightarrow \quad (\alpha w_1 + \beta w_2) \in \mathsf{Im}(A)\,. \end{aligned}$$

We can now simply say that *there exists a solution to (1.21) iff* $b \in \mathsf{Im}(A)$. Suppose now we want to go further and find a criterion for solvability of (1.21) for *all possible* "right hand sides" $b \in \mathsf{W}$. It is immediate from Definition 1.34 that as a mapping A *is onto iff* $\mathsf{Im}(A) = \mathsf{W}$ (we can say that the image of A "fills up" all of W). We can now say that *for any $b \in \mathsf{W}$, there exists a solution to (1.21) iff* $\mathsf{Im}(A) = \mathsf{W}$.

Now for the other important questions of whether a solution is unique, and if not, how to characterize all possible solutions. The key is to consider the *homogenous* equation

$$Ax \ = \ 0. \tag{1.22}$$

Now suppose we have found one solution $\bar{x} \in \mathsf{V}$ to the original equation so that

$$A\bar{x} \ = \ b.$$

The linearity property then implies that any vector of the form $\bar{x} + \tilde{x}$ where $\tilde{x}$ solves the homogenous equation (1.22) (i.e. $A\tilde{x} = 0$) must also be a solution since

$$A(\bar{x} + \tilde{x}) \ = \ A\bar{x} \ + \ A\tilde{x} \ = \ b \ + \ 0 \ = \ b. \tag{1.23}$$

Therefore given one solution of the original equation, we can generate other solutions by simply adding to it any solution of the homogenous equation (1.22). Observe that the set of all solutions to the homogenous equation is actually a subspace since

$$Ax_1 = 0,\ Ax_2 = 0 \qquad \Rightarrow \qquad A(\alpha x_1 + \beta x_2) \ = \ \alpha\, Ax_1 + \beta\, Ax_2 \ = \ 0 + 0 \ = \ 0.$$

The set of all solutions to the homogenous equation (1.22) is called the "null space" of the operator A. We are therefore led to the following definition.

Definition 1.35. *Given a linear operator $A : \mathsf{V} \to \mathsf{W}$ between two vector spaces V and W, the* null space *of A is the subspace of vectors mapped to zero*

$$\mathsf{Nu}(A) := \{v \in \mathsf{V};\ Av = 0\} \subseteq \mathsf{V}.$$

Figure 1.8 illustrates the geometry where the null space is the set of all vectors in V mapped to the zero vector in W.

Recall that the image space characterizes when a map is onto. On the other hand, the null space characterizes when a linear map is one-to-one. Indeed, note that zero is always in the null space since zero is mapped to zero by any linear operator. If the null space of A contains more than just the zero element, then A is clearly not one-to-one since more than one element is mapped to zero. Conversely, suppose that A is not one-to-one, then there are two vectors $v_1 \neq v_2$ such that

$$Av_1 \ = \ Av_2 \qquad \Leftrightarrow \qquad Av_1 - Av_2 \ = \ 0 \qquad \Leftrightarrow \qquad A(v_1 - v_2) \ = \ 0,$$

i.e. we found a non-zero vector $v_1 - v_2 \neq 0$ that is in $\mathsf{Nu}(A)$. Note how the linearity of A was the key to this argument. We therefore have just argued the contrapositive of the following statement

$$A \text{ is one-to-one} \qquad \Leftrightarrow \qquad \mathsf{Nu}(A) = 0. \tag{1.24}$$

This criterion is exceedingly useful! Checking whether a mapping is one-to-one would require insuring that all elements of the domain set are each mapped to distinct elements. For linear mappings, it suffices to check the size of the null space. We now summarize all the previous arguments in the following lemma.

Lemma 1.36. (Linear Equations Solvability) *Let $A : \mathsf{V} \to \mathsf{W}$ be a linear operator between two vector spaces. Consider the abstract linear equation $Av = w$ where w is given, and v is the unknown.*

1. *For a fixed $\bar{w} \in \mathsf{W}$, the equation $Av = \bar{w}$ has a solution iff $\bar{w} \in \mathsf{Im}(A)$.*
2. *For each $w \in \mathsf{W}$, the equation $Av = w$ has a solution iff $\mathsf{Im}(A) = \mathsf{W}$ (i.e. A is onto).*
3. *Given $\bar{w} \in \mathsf{W}$, and one solution $A\bar{v} = \bar{w}$, this solution is unique iff $\mathsf{Nu}(A) = 0$.*

   If $\mathsf{Nu}(A) \neq 0$, then any other solution of $Av = \bar{w}$ is of the form $v = \bar{v} + \tilde{v}$, where $\tilde{v} \in \mathsf{Nu}(A)$.

We have already argued all the points above except for the last one. The argument (1.23) shows that if $\tilde{v} \in \mathsf{Nu}(A)$ then $\bar{v} + \tilde{v}$ is a solution. Conversely, if v any other solution with $Av = \bar{w}$, then consider $v - \bar{v}$

$$A\,(v - \bar{v}) \ = \ Av - A\bar{v} \ = \ b - b \ = \ 0 \qquad \Rightarrow \qquad \tilde{v} := v - \bar{v} \in \mathsf{Nu}(A)\,, \text{ and } v = \bar{v} + \tilde{v},$$

i.e. the solution v can be written as $\bar{v} + \tilde{v}$ where $\tilde{v}$ is in the null space. Thus the null space "parameterizes" *all solutions* of a linear equation.

In summary, Lemma 1.36 implies that to understand properties of solutions of linear equations, one must understand the null and image spaces of the underlying operator.

Example 1.37. For matrices, the image space is the so-called "column span" (the span of the all the columns viewed as vectors). This can be easily seen from the definition of matrix-vector products and partition notation as follows. Let $A : \mathbb{R}^m \to \mathbb{R}^n$ be an $n \times m$ matrix. It maps a vector $\boldsymbol{v} \in \mathbb{R}^m$ to a vector $\boldsymbol{w} = A\boldsymbol{v} \in \mathbb{R}^n$ by

$$\begin{bmatrix} \\ \boldsymbol{w} \\ \\ \end{bmatrix} = \left[\begin{array}{c:c:c:c} & & & \\ \boldsymbol{a}_1 & \cdots & \boldsymbol{a}_m & \\ & & & \end{array}\right] \begin{bmatrix} v_1 \\ \hdashline \vdots \\ \hdashline v_m \end{bmatrix} \ = \ \begin{bmatrix} \\ \boldsymbol{a}_1 \\ \\ \end{bmatrix} v_1 + \ \cdots \ + \begin{bmatrix} \\ \boldsymbol{a}_m \\ \\ \end{bmatrix} v_m,$$

where $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m\}$ are the columns of the matrix A. This way we can view $\boldsymbol{w}$ as a linear combination of the columns of A, formed with coefficients $\{v_1, \ldots, v_m\}$, which are the components of the vector $\boldsymbol{v}$. If we let $\boldsymbol{v}$ range over all possible vectors in $\mathbb{R}^m$, then the left hand side $\boldsymbol{w}$ will correspond to all possible linear combinations of the columns of A, i.e. to the *column span*. We therefore conclude that for a matrix A

$$\mathsf{Im}(A) \;=\; \mathrm{col.span}\big(A\big).$$

Example 1.38. Consider the two-variable equation

$$\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \tag{1.25}$$

To see if there exists a solution, check whether the vector $(1,1)$ is in the image space, or equivalently in the column span of the matrix. The two columns of the matrix are actually multiples of each other (column 2 is -1 times column 1) and therefore we conclude

$$\mathsf{Im}\left(\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\right) = \mathrm{span}\begin{bmatrix} 1 \\ -1 \end{bmatrix}, \qquad \begin{bmatrix} 1 \\ 1 \end{bmatrix} \notin \mathrm{span}\begin{bmatrix} 1 \\ -1 \end{bmatrix} \qquad \Rightarrow \qquad (1.25) \text{ has no solution.}$$

On the other hand, consider the equation with a different right hand side which is in the image space

$$\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}. \tag{1.26}$$

By inspection, one solution is the vector $(\bar{x}_1, \bar{x}_2) = (0,1)$. To characterize all solutions, we need to find the null space. By inspection again, one vector that is in the null space is $(1,1)$

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

Could the null space be any bigger? The answer is no, and the justification will be provided by the rank-nullity theorem of the next section[6], which in this case states that the dimension of the null space plus the dimension of the image space is exactly 2 (the dimension of the space in which x lives). We already showed that the image space is 1 dimensional, and therefore the null space must be one dimensional. Therefore all solutions of (1.26) are

$$\bar{x} \;+\; \tilde{x} \;=\; \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \alpha \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \alpha \\ 1+\alpha \end{bmatrix}, \qquad \alpha \in \mathbb{R}.$$

The reader should verify that any vector of this form satisfies (1.26).

Example 1.39. Consider the first order differential equation

$$\frac{df}{dt}(t) \;=\; \bar{w}(t), \qquad\qquad t \in [a,b], \tag{1.27}$$

where $\bar{w}(.)$ is a given function on the interval $[a,b]$. Recalling the derivative operator of Example 1.13, we can think of this differential equation as a linear equation in a function space of the form (1.21) as follows

$$\mathbf{D}f \;=\; \bar{w}, \qquad\qquad \mathbf{D} : \mathsf{C}^1[a,b] \longrightarrow \mathsf{C}[a,b]. \tag{1.28}$$

[6] In this simple example, this can also be justified directly. An vector x in the null space satisfies $x_1 - x_2 = 0$ and $-x_1 + x_2 = 0$. All such vectors have the property that $x_2 = -x_1$, and are therefore scalar multiples of the vector $(1,-1)$. The rank-nullity theorem is however useful for more complicated examples.

For a given right hand side $\bar{w}$, the equation has a solution iff $\bar{w} \in \mathsf{Im}(\mathbf{D})$. Examining the image space is relatively easy in this case since integration reverses $\mathbf{D}$ up to a constant, i.e. given a continuous function $g \in C[a,b]$, then for any constant $c \in \mathbb{R}$

$$\frac{d}{dt}\left(\int_a^t g(\tau)\, d\tau \;+\; c\right) \;=\; g(t).$$

The indefinite integral of $g \in \mathsf{C}[a,b]$ belongs to $\mathsf{C}^1[a,b]$, and therefore the mapping $\mathbf{D} : \mathsf{C}^1[a,b] \longrightarrow \mathsf{C}[a,b]$ is onto and $\mathsf{Im}(\mathbf{D}) = \mathsf{C}[a,b]$ in this case. Therefore, the equation (1.28) has a solution $f \in \mathsf{C}^1[a,b]$ for any $\bar{w} \in \mathsf{C}[a,b]$.

To find all solutions, we need to characterize the null space $\mathsf{Nu}(\mathbf{D})$

$$\mathbf{D}f \;=\; 0 \qquad \Leftrightarrow \qquad \frac{df}{dt}(t) \;=\; 0 \qquad \Leftrightarrow \qquad f(t) \;=\; c,\; c \in \mathbb{R},$$

i.e. it is the one-dimensional space of constant functions. Therefore, given any particular solution $\bar{f}$ of (1.27), all other solutions f are obtained by adding elements of the null space

$$f(t) \;=\; \bar{f}(t) \;+\; c, \qquad c \in \mathbb{R}.$$

1.5.1 The General Rank-Nullity Theorem

Some questions about linear operators or solvability of linear equations do not require a complete characterization of the image and null spaces, but only knowing their *dimensions*. This leads to the concepts of the "rank" and "nullity" of a linear operator.

Definition 1.40. *Given a linear operator $A : \mathsf{V} \to \mathsf{W}$ between two vector spaces, its* rank *is the dimension of its image space and its* nullity *is the dimension of its null space*

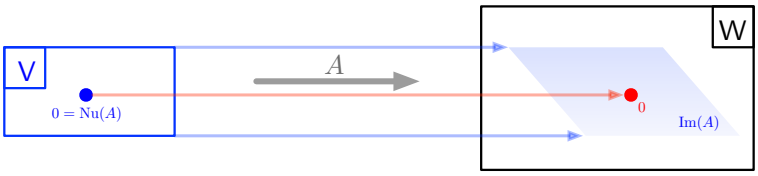
$$\mathbf{rk}(A) \;:=\; \dim(\mathsf{Im}(A))\,, \qquad\qquad \mathbf{nl}(A) \;:=\; \dim(\mathsf{Nu}(A))\,.$$

Both rank and nullity could be finite or infinite. Of course if W is finite dimensional, then necessarily the rank is finite, and similarly if V is finite dimensional then the nullity is finite. It is also possible that some operators between infinite dimensional spaces can have finite rank or finite nullity (but not both, see (1.29) below).

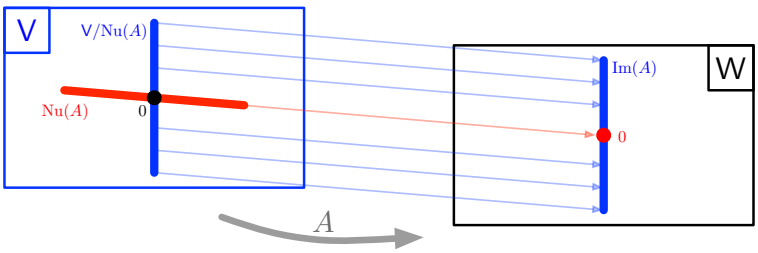
For matrices, we recall from Example 1.37 that the image space is the column span, which then implies that *the rank of a matrix is precisely the number of linearly independent columns*. This last statement is sometimes taken as the definition of the rank of a matrix. However, Definition 1.40 is preferable as a starting point since it is applicable to any linear operator and thus more general. The statement about the number of linearly independent columns of a matrix should be thought of as a result (justified in the arguments of Example 1.37) rather than a definition.

A related statement that the reader may be familiar with is that the number of linearly independent columns of a matrix is the same as the number of linearly independent rows (and is equal to the rank). While it is possible to prove this fact by algebraic manipulations, there are important geometric reasons for this fact that are difficult to appreciate at this point. A better geometric understanding can be achieved after discussing the fundamental concepts of duality and operator adjoints, and we therefore postpone a proof until then. In the meantime, we can still uncover interesting and fundamental relations between null, image and quotient spaces. This is the subject of the so-called *rank-nullity theorem*.

Let $A : \mathsf{V} \longrightarrow \mathsf{W}$ be a mapping between two vector spaces. Consider first the case when A has trivial null space (i.e. $\mathsf{Nu}(A) = 0$). Recall from (1.24) that this implies that A is one-to-one. Now consider $\mathsf{Im}(A) \subseteq \mathsf{W}$, which is a subspace of W (see Figure 1.9a for an



(a) When the null space $\mathsf{Nu}(A)$ is trivial (i.e. just the zero element), then the linear mapping A is one-to-one. This implies that V is mapped isomorphically onto $\mathsf{Im}(A) \subseteq \mathsf{W}$. Since $\mathsf{Im}(A)$ is itself a vector space, we have a vector space isomorphism $\mathsf{Im}(A) \sim \mathsf{V}$.



(b) The quotient space $\mathsf{V}/\mathsf{Nu}(A)$ can be viewed as a subspace complementary to $\mathsf{Nu}(A)$ in V. The null space is mapped to 0, but the complement $\mathsf{V}/\mathsf{Nu}(A)$ is mapped isomorphically (one-to-one and onto, depicted by the long blue arrows) to $\mathsf{Im}(A)$. Since $\mathsf{Nu}(A) \oplus \mathsf{V}/\mathsf{Nu}(A) \sim \mathsf{V}$, and $\mathsf{V}/\mathsf{Nu}(A)$ is isomorphic to $\mathsf{Im}(A)$, this implies the isomorphism $\mathsf{Nu}(A) \oplus \mathsf{Im}(A) \sim \mathsf{V}$. This isomorphism holds even though $\mathsf{Nu}(A)$ and $\mathsf{Im}(A)$ are in different spaces, and therefore $\oplus$ here is the external direct sum. The isomorphism implies the rank-nullity relation $\mathsf{dim}(\mathsf{V}) = \mathbf{nl}(A) + \mathbf{rk}(A)$.

Figure 1.9: Illustration of rank-nullity Lemma 1.41 in the simple case when $\mathsf{Nu}(A) = 0$ (top), and the general case $\mathsf{Nu}(A) \neq 0$ (bottom).

illustration). Since $\mathsf{Im}(A)$ is itself a vector space, we can think of A as a map $A : \mathsf{V} \to \mathsf{Im}(A)$, which is then onto by definition. Since this map is now one-to-one and onto, we can say that V *is isomorphic to* $\mathsf{Im}(A)$ if $\mathsf{Nu}(A) = 0$. This is true even if $\mathsf{Im}(A)$ doesn't "fill up" all of W, since now we consider A as mapping onto $\mathsf{Im}(A)$ rather than W.

What if $\mathsf{Nu}(A) \neq 0$? There is still an important statement we can make. Refer to Figure 1.9b and consider the quotient space $\mathsf{V}/\mathsf{Nu}(A)$. Recall (Lemma 1.33) that it can be viewed as a subspace (of V) complementary to $\mathsf{Nu}(A)$. The restriction of A to $\mathsf{V}/\mathsf{Nu}(A)$ has trivial null space (the intersection of $\mathsf{V}/\mathsf{Nu}(A)$ and $\mathsf{Nu}(A)$ is 0), thus restricting A to the subspace $\mathsf{V}/\mathsf{Nu}(A)$ (or equivalently to any subspace complementary to $\mathsf{Nu}(A)$) makes it into a one-to-one mapping onto $\mathsf{Im}(A)$. This implies that the quotient space $\mathsf{V}/\mathsf{Nu}(A)$ is isomorphic to $\mathsf{Im}(A)$. We summarize the above conclusions in the following Lemma.

Lemma 1.41. *Let $A : \mathsf{V} \to \mathsf{W}$ be linear operator between vector spaces.*

1. *If $\mathsf{Nu}(A) = 0$, then A is one-to-one and V is isomorphic to $\mathsf{Im}(A)$.*
2. *The* restriction of A to $\mathsf{V}/\mathsf{Nu}(A)$ *defined by*

$$A\left(v + \mathsf{Nu}(A)\right) \;:=\; Av,$$

   maps $\mathsf{V}/\mathsf{Nu}(A)$ isomorphically onto $\mathsf{Im}(A)$. Therefore $\mathsf{V}/\mathsf{Nu}(A) \sim \mathsf{Im}(A)$.

   Equivalently, any subspace of V complementary to $\mathsf{Nu}(A)$ is mapped by A isomorphically onto $\mathsf{Im}(A)$.
3. rank-nullity: $\mathsf{V} \;\sim\; \mathsf{Nu}(A) \oplus \mathsf{V}/\mathsf{Nu}(A) \;\sim\; \mathsf{Nu}(A) \oplus \mathsf{Im}(A)$. *In particular*

$$\mathsf{dim}(\mathsf{V}) \;=\; \mathsf{dim}(\mathsf{Nu}(A)) + \mathsf{dim}(\mathsf{Im}(A)) \;=\; \mathbf{nl}(A) + \mathbf{rk}(A)\,. \tag{1.29}$$

The third statement is illustrated in Figure 1.9b where V is the direct sum[7] of $\mathsf{Nu}(A)$ and $\mathsf{V}/\mathsf{Nu}(A)$, both viewed as subspace of V. Since $\mathsf{V}/\mathsf{Nu}(A)$ and $\mathsf{Im}(A)$ are isomorphic, we can then say that $\mathsf{V} \sim \mathsf{Nu}(A) \oplus \mathsf{Im}(A)$ even though $\mathsf{Im}(A)$ is a subspace of another vector space W.

Another consequence of (1.29) is that if V is infinite dimensional, then the rank and nullity cannot both be finite. A linear operator on an infinite dimensional space can possibly have finite-dimensional null space or a finite-dimensional image space, but not both.

1.6 Bases Representations and Change of Bases

We have shown several examples of finite-dimensional vector spaces that "look like" $\mathbb{R}^n$. In fact, for any finite-dimensional vector space, there are many ways to map it to $\mathbb{R}^n$ isomorphically by choosing different bases.

Lemma 1.42. *Let* $\mathsf{v} := \{\boldsymbol{v}_k;\ k = 1, \dots, n\} \subset \mathsf{V}$ *be a basis. Each element* $\boldsymbol{u} \in \mathsf{V}$ *can be written as a* unique *linear combination of the basis elements, i.e.*

$$\boldsymbol{u} \ = \ x_1\boldsymbol{v}_1 + \cdots + x_n\boldsymbol{v}_n \ = \ y_1\boldsymbol{v}_1 + \cdots + y_n\boldsymbol{v}_n \qquad \Rightarrow \qquad x_k = y_k,\ k = 1, \dots, n. \tag{1.30}$$

Thus a choice of basis v *induces a well-defined mapping* $\boldsymbol{u} \mapsto (x_1, \dots, x_n)$*, which is a vector space isomorphism from* V *to* $\mathbb{R}^n$*.*

The unique n *numbers* $(x_1, \cdots, x_n)$ *are called the* coordinates *or the* representation *of the vector* $\boldsymbol{u}$ *in the basis* v*.*

Proof. If for at least one index k, $x_k \neq y_k$, then subtract one representation from the other

$$0 \ = \ \boldsymbol{u} - \boldsymbol{u} \ = \ (x_1 - y_1)\ \boldsymbol{v}_1 + \cdots + (x_n - y_n)\ \boldsymbol{v}_n.$$

Since $(x_k - y_k) \neq 0$, we have found one non-trivial linear combinations of the basis elements that sums to zero, i.e. the set v is not linearly independent. This uniqueness property shows that the mapping $\boldsymbol{u} \mapsto (x_1, \dots, x_n)$ is well defined and one-to-one. It is also onto since any n-tuple of coefficients $(x_1, \dots, x_n)$ corresponds to a vector in V by taking the linear combination (1.30) (because v is a basis).

Finally, the mapping is linear since

$$\left.\begin{array}{l} \boldsymbol{u} = x_1\boldsymbol{v}_1 + \cdots + x_n\boldsymbol{v}_n \\ \boldsymbol{w} = y_1\boldsymbol{v}_1 + \cdots + y_n\boldsymbol{v}_n \end{array}\right\} \qquad \Rightarrow \qquad \boldsymbol{u} + \boldsymbol{w} = (x_1 + y_1)\ \boldsymbol{v}_1 + \cdots + (x_n + y_n)\ \boldsymbol{v}_n,$$

and recall that $(x_1, \dots, x_n) + (y_1, \dots, y_n) := (x_1 + y_1, \dots, x_n + y_n)$ is the definition of vector addition in $\mathbb{R}^n$. Thus the mapping $\boldsymbol{u} \mapsto (x_1, \dots, x_n)$ is an isomorphism. □

Lemma 1.42 implies that *every finite-dimensional (real*[8]*) vector space* V *is isomorphic to* $\mathbb{R}^n$*, where* n *is the dimension of* V. Every choice of basis $\mathsf{v} := \{\boldsymbol{v}_k;\ k = 1, \dots, n\} \subset \mathsf{V}$ induces an isomorphism between V and $\mathbb{R}^n$. We call such an isomorphism a *basis representation* of elements of V. To make this precise and avoid confusion, we adopt the following notation as needed for clarity

$$\boldsymbol{u} = \ x_1\boldsymbol{v}_1 + \cdots + x_n\boldsymbol{v}_n \qquad \Leftrightarrow \qquad [\boldsymbol{u}]_{\mathsf{v}} = \ (x_1, \dots, x_n). \tag{1.31}$$

Thus $[\boldsymbol{u}]_{\mathsf{v}}$ is the *vector of coefficients of* $\boldsymbol{u}$ *in the basis* $\{\boldsymbol{v}_k\}$. In a different basis, say $\{\boldsymbol{w}_k\}$, the same vector $\boldsymbol{u}$ will have a different set of basis coefficients $[\boldsymbol{u}]_{\mathsf{w}}$.

[7] Note that by definition of the quotient, $\mathsf{V} \sim \mathsf{U} \oplus \mathsf{V}/\mathsf{U}$ where U is any subspace of V. Here we apply this statement to the subspace $\mathsf{Nu}(A)$.

[8] Similar arguments imply that also any n-dimensional complex vector space is isomorphic to $\mathbb{C}^n$.

Recall that when we write a column vector $\boldsymbol{x} \in \mathbb{R}^n$, we are implicitly writing it using a basis expansion in the canonical basis $\mathsf{e} := \{\boldsymbol{e}_1, \dots, \boldsymbol{e}_n\}$

$$\boldsymbol{x} \;=\; \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \;=\; x_1 \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \cdots + \; x_n \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \;=\; x_1 \boldsymbol{e}_1 + \cdots + x_n \boldsymbol{e}_n.$$

In the notation of (1.31), $\boldsymbol{x} = [\boldsymbol{x}]_{\mathsf{e}}$, but since $\boldsymbol{x}$ is actually given in terms of the canonical basis to begin with, we sometimes simply write $\boldsymbol{x}$ rather than $[\boldsymbol{x}]_{\mathsf{e}}$.

Now given another basis $\mathsf{v} := \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$ of $\mathbb{R}^n$, how do we find the coefficients of any vector $\boldsymbol{x}$ (given initially in the canonical basis) in this new basis? We present now a method that gives a nice, compact formula for the new coefficients using matrix-vector notation.

Lemma 1.43. *Consider a basis $\mathsf{v} := \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$ of $\mathbb{R}^n$, and the representations $\{[\boldsymbol{v}_1]_{\mathsf{e}}, \dots, [\boldsymbol{v}_n]_{\mathsf{e}}\}$ of its vectors in the canonical basis. Given any vector $\boldsymbol{x} \in \mathbb{R}^n$, the relation between its representation in the canonical basis e and the new basis v is given by*

$$\begin{aligned} \boldsymbol{x} &= x_1 \boldsymbol{e}_1 + \cdots + x_n \boldsymbol{e}_n \\ &= \hat{x}_1 \boldsymbol{v}_1 + \cdots + \hat{x}_n \boldsymbol{v}_n \end{aligned} \quad \Leftrightarrow \quad [\boldsymbol{x}]_{\mathsf{v}} := \begin{bmatrix} \hat{x}_1 \\ \vdots \\ \hat{x}_n \end{bmatrix} = \left[\begin{array}{c:c:c} & & \\ [\boldsymbol{v}_1]_{\mathsf{e}} & \cdots & [\boldsymbol{v}_n]_{\mathsf{e}} \\ & & \end{array}\right]^{-1} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} =: V_{\mathsf{e}}^{-1} \, [\boldsymbol{x}]_{\mathsf{e}}, \quad (1.32)$$

where V_{e} is the matrix made up of the vectors $\{[\boldsymbol{v}_k]_{\mathsf{e}}\}$ as its columns.

Proof. Observe that the expansion of $\boldsymbol{x}$ in the new basis can be written compactly as the following matrix-vector product

$$\begin{array}{rcl} \boldsymbol{x} &=& \hat{x}_1 \boldsymbol{v}_1 + \cdots + \hat{x}_n \boldsymbol{v}_n \\ & \Updownarrow & \\ [\boldsymbol{x}]_{\mathsf{e}} &=& \hat{x}_1 \, [\boldsymbol{v}_1]_{\mathsf{e}} + \cdots + \hat{x}_n \, [\boldsymbol{v}_n]_{\mathsf{e}} \\ & \Updownarrow & \\ [\boldsymbol{x}]_{\mathsf{e}} \;=\; \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} &=& \left[\begin{array}{c:c:c} & & \\ [\boldsymbol{v}_1]_{\mathsf{e}} & \cdots & [\boldsymbol{v}_n]_{\mathsf{e}} \\ & & \end{array}\right] \begin{bmatrix} \hat{x}_1 \\ \vdots \\ \hat{x}_n \end{bmatrix} \end{array} \quad \Leftrightarrow \quad \begin{array}{rcl} [\boldsymbol{x}]_{\mathsf{e}} &=& V_{\mathsf{e}} \; [\boldsymbol{x}]_{\mathsf{v}} \\ & \Updownarrow & \\ [\boldsymbol{x}]_{\mathsf{v}} &=& V_{\mathsf{e}}^{-1} \; [\boldsymbol{x}]_{\mathsf{e}} \end{array},$$

where V_{e} is the matrix whose columns are the vectors $\{\boldsymbol{v}_k\}$ expressed in the canonical basis, and $[\boldsymbol{x}]_{\mathsf{v}}$ is a column vector containing the new coefficients. Since the columns of V_{e} are linearly independent, V_{e} is invertible. □

The matrix V defined above has a geometric interpretation. It maps each canonical basis vector to the respective new basis vector

$$V \; \boldsymbol{e}_k \;=\; \left[\begin{array}{c:c:c} & & \\ \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_n \\ & & \end{array}\right] \begin{bmatrix} 0 \\ \vdots \\ \mathrm{i} \\ \vdots \\ 0 \end{bmatrix} \swarrow k\text{'th entry} \;=\; \boldsymbol{v}_k,$$

where we have dropped the notation $[\boldsymbol{v}_k]_{\mathsf{e}}$ and V_{e} for simplicity. If we think of $\{\boldsymbol{e}_k\}$ as coordinate axes, and similarly consider $\{\boldsymbol{v}_k\}$ as new coordinate axes, then V is the linear transformation on $\mathbb{R}^n$ that transforms the old coordinate axes to the new ones. The coefficients of any vector however transform according to V^{-1} in (1.32). We say that the coefficients transform in a *contravariant* (i.e. in the "opposite") manner to the coordinate axes. Figure 1.10 illustrates this geometry with an example.

In Example 1.19 we saw how to relate two different bases and the corresponding coefficients in those bases. The arguments in (1.16)-(1.17) apply to bases of any size, and we state the conclusion next.

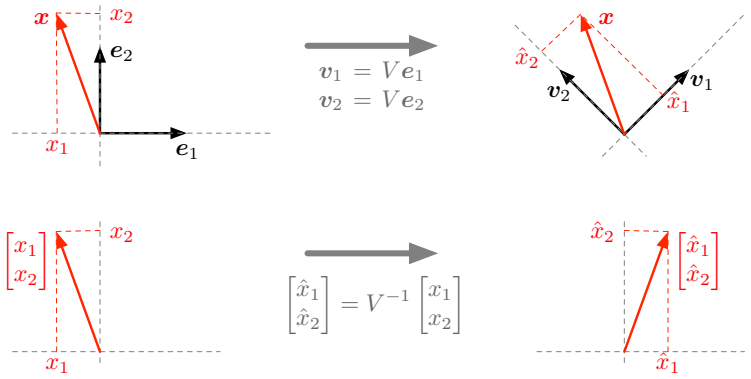


Figure 1.10: A change of basis can be viewed as a transformation of "coordinate axes". Here the matrix V transforms (top figure) the canonical basis vectors $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ to the vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, the two columns of the 2×2 matrix V. The transformation of bases is depicted as a 45° counter-clockwise rotation of the axes. On the other hand, the coefficients (x_1, x_2) and $(\hat{x}_1, \hat{x}_2)$ of any vector $\boldsymbol{x}$ in each of the two bases respectively transform with V^{-1}, i.e. in a manner "contravariant" to the transformation of the axes. The vector $(\hat{x}_1, \hat{x}_2)$ of coefficients (as distinct from the vector $\boldsymbol{x}$ itself) is depicted (bottom figure) as a 45° clockwise rotation of the original coefficents vector (x_1, x_2).

Lemma 1.44. *Let* $\mathsf{v} := \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ *and* $\mathsf{w} := \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n\}$ *be two bases of an* n*-dimensional vector space* V*. Let* $\{[\boldsymbol{v}_k]_{\mathsf{u}}\}$ *and* $\{[\boldsymbol{w}_k]_{\mathsf{u}}\}$ *be the vectors of coefficients of the sets* v *and* w *in any third basis* u*, and form the matrices*

$$V_{\mathsf{u}} := \left[\, [\boldsymbol{v}_1]_{\mathsf{u}} \,\middle|\, \cdots \,\middle|\, [\boldsymbol{v}_n]_{\mathsf{u}} \,\right], \qquad W_{\mathsf{u}} := \left[\, [\boldsymbol{w}_1]_{\mathsf{u}} \,\middle|\, \cdots \,\middle|\, [\boldsymbol{w}_n]_{\mathsf{u}} \,\right],$$

If $[\boldsymbol{u}]_{\mathsf{w}} = (x_1, \ldots, x_n)$ *and* $[\boldsymbol{u}]_{\mathsf{v}} = (\hat{x}_1, \ldots, \hat{x}_n)$ *are the respective basis coefficients of any vector* $\boldsymbol{u} \in \mathsf{V}$

$$\boldsymbol{u} \;=\; x_1\boldsymbol{w}_1 + \cdots + x_n\boldsymbol{w}_n \;=\; \hat{x}_1\boldsymbol{v}_1 + \cdots + \hat{x}_n\boldsymbol{v}_n, \tag{1.33}$$

then the two sets of coefficients are related by

$$[\boldsymbol{u}]_{\mathsf{v}} = \begin{bmatrix} \hat{x}_1 \\ \vdots \\ \hat{x}_n \end{bmatrix} = \begin{bmatrix} \\ A \\ \\ \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = A\,[\boldsymbol{u}]_{\mathsf{w}} \qquad \Leftrightarrow \qquad [\boldsymbol{u}]_{\mathsf{w}} = A^{-1}\,[\boldsymbol{u}]_{\mathsf{v}}\,, \tag{1.34}$$

where $A := V_{\mathsf{u}}^{-1} W_{\mathsf{u}}$*. The matrix* A *is independent of the choice of the third basis* u*.*

Proof. Since v and w are both bases of $\mathbb{R}^n$, each element of w can be written as a linear combination of elements of v. Denote the coefficients of these linear combinations as follows

$$\boldsymbol{w}_k \;=\; a_{1k}\boldsymbol{v}_1 + \cdots + a_{nk}\boldsymbol{v}_n, \qquad k = 1, .., n. \tag{1.35}$$

Note that these coefficients are determined by the sets v and w, and do not depend on any third bases u.

The equations (1.35) can each be expressed in the third basis u. First, each as an equation

with a matrix-vector product, and then all of them together as a *single matrix equation*

$$[\boldsymbol{w}_k]_{\mathsf{u}} = a_{1k}[\boldsymbol{v}_1]_{\mathsf{u}} + \cdots + a_{nk}[\boldsymbol{v}_n]_{\mathsf{u}} = \begin{bmatrix} [\boldsymbol{v}_1]_{\mathsf{u}} & \cdots & [\boldsymbol{v}_n]_{\mathsf{u}} \end{bmatrix} \begin{bmatrix} a_{1k} \\ \vdots \\ a_{nk} \end{bmatrix}, \qquad k = 1, .., n$$

$$\Leftrightarrow \qquad \begin{bmatrix} [\boldsymbol{w}_1]_{\mathsf{u}} & \cdots & [\boldsymbol{w}_n]_{\mathsf{u}} \end{bmatrix} = \begin{bmatrix} [\boldsymbol{v}_1]_{\mathsf{u}} & \cdots & [\boldsymbol{v}_n]_{\mathsf{u}} \end{bmatrix} \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}$$

$$\Rightarrow \qquad W_{\mathsf{u}} = V_{\mathsf{u}} A \qquad \Rightarrow \qquad A = V_{\mathsf{u}}^{-1} W_{\mathsf{u}},$$

where the matrix A is defined as above using all the coefficients $\{a_{ij}\}$. Note that the matrices V_{u} and W_{u} are invertible (since their columns are linearly independent respectively), and then so is A.

Now given any vector $\boldsymbol{u}$, we can write the expression (1.33) in the v basis as

$$\boldsymbol{u} = \hat{x}_1 \boldsymbol{v}_1 + \cdots + \hat{x}_n \boldsymbol{v}_n = x_1 \boldsymbol{w}_1 + \cdots + x_n \boldsymbol{w}_n$$

$$\Leftrightarrow \qquad [\boldsymbol{u}]_{\mathsf{v}} = \begin{bmatrix} \hat{x}_1 \\ \vdots \\ \hat{x}_n \end{bmatrix} = x_1 [\boldsymbol{w}_1]_{\mathsf{v}} + \cdots + x_n [\boldsymbol{w}_n]_{\mathsf{v}} = \begin{bmatrix} [\boldsymbol{w}_1]_{\mathsf{v}} & \cdots & [\boldsymbol{w}_n]_{\mathsf{v}} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

$$\Leftrightarrow \qquad [\boldsymbol{u}]_{\mathsf{v}} = W_{\mathsf{v}} [\boldsymbol{u}]_{\mathsf{w}} = A [\boldsymbol{u}]_{\mathsf{w}}.$$

The last equivalence follows from $A = V_{\mathsf{u}}^{-1} W_{\mathsf{u}}$ in any basis u. In the v basis $V_{\mathsf{v}} = I$, and therefore $A = V_{\mathsf{v}}^{-1} W_{\mathsf{v}} = W_{\mathsf{v}}$. □

Note how this lemma generalizes the previous Lemma 1.43. To obtain the previous lemma from the current one, set $\mathsf{w} = \mathsf{e}$, and use $\mathsf{u} = \mathsf{e}$. We then have that $W_{\mathsf{e}} = I$, and Lemma 1.44 says $A = V_{\mathsf{e}}^{-1} W_{\mathsf{e}} = V_{\mathsf{e}}^{-1}$, which the same statement as in Lemma 1.43.

1.6.1 Matrix Representations of Linear Operators

Consider a linear operator $\mathcal{A} : \mathsf{V} \to \mathsf{W}$ between two finite dimensional vector spaces. The operator $\mathcal{A}$ is given by some recipe or algorithm such that given any vector $\boldsymbol{f} \in \mathsf{V}$, the algorithm gives the vector $\mathcal{A}\boldsymbol{f} \in \mathsf{W}$.

Let $\mathsf{v} := \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m\}$ and $\mathsf{w} := \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n\}$ be bases in V and W respectively. We now ask the following question. If a vector $\boldsymbol{f}$ is mapped to a vector $\boldsymbol{g} = \mathcal{A}\boldsymbol{f}$, how is the basis representation of $\boldsymbol{f}$ mapped to that of $\boldsymbol{g}$?

Consider the basis representations of $\boldsymbol{f}$ and $\boldsymbol{g}$ (in the respective bases of V and W), and organize the coefficients into "column vectors" as follows

$$\begin{aligned} \boldsymbol{f} &= x_1 \boldsymbol{v}_1 + \cdots + x_m \boldsymbol{v}_m \\ \boldsymbol{g} &= y_1 \boldsymbol{w}_1 + \cdots + y_n \boldsymbol{w}_n \end{aligned}, \qquad \boldsymbol{x} := [\boldsymbol{f}]_{\mathsf{v}} = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}, \qquad \boldsymbol{y} := [\boldsymbol{g}]_{\mathsf{w}} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}. \tag{1.36}$$

Our goal is to find the matrix that relates those two coefficient vectors.

Each $\boldsymbol{v}_j$ is mapped to a vector $\mathcal{A}\boldsymbol{v}_j \in \mathsf{W}$. Since $\{\boldsymbol{w}_i\}$ is a basis in W, we can write $\mathcal{A}\boldsymbol{v}_j$ as a unique linear combination

$$\mathcal{A}\boldsymbol{v}_j = a_{1j}\, \boldsymbol{w}_1 + \cdots + a_{nj}\, \boldsymbol{w}_n, \qquad j = 1, \ldots, m. \tag{1.37}$$

where $\{a_{1j}, \ldots, a_{nj}\}$ are the coefficients of the representation of $A\boldsymbol{v}_j$ in the basis $\{\boldsymbol{w}_i\}$. This set of $n \times m$ numbers $\{a_{ij}\}$ is what we need to describe the relation between the column

vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in (1.36). Consider now the equation $\boldsymbol{g} = \mathcal{A}\boldsymbol{f}$ expressed using the basis expansions of $\boldsymbol{f}$ and $\boldsymbol{g}$

$$\begin{aligned}
\boldsymbol{g} \;=\; \mathcal{A}\boldsymbol{f} \;&=\; \mathcal{A}\left(\sum_{j=1}^{m} x_j \boldsymbol{v}_j\right) \;=\; \sum_{j=1}^{m} x_j\, \mathcal{A}(\boldsymbol{v}_j) && \text{(by linearity of } \mathcal{A}) \\
&= \sum_{j=1}^{m} x_j \left(\sum_{i=1}^{n} a_{ij}\boldsymbol{w}_i\right) = \sum_{j=1}^{m}\left(\sum_{i=1}^{n} a_{ij}x_j\right)\boldsymbol{w}_i && \text{(from (1.37) and rearranging sum)} \\
\Rightarrow \quad y_i \;&= \sum_{j=1}^{m} a_{ij}\; x_j.
\end{aligned}$$

Note that the last sum is the matrix-vector product between the matrix whose entries are $\{a_{ij}\}$ and the vector x. We summarize the above in the following statement.

Lemma 1.45. *Let $\mathcal{A} : \mathsf{V} \to \mathsf{W}$ be a linear operator between two finite dimensional vector spaces with bases $\mathsf{v} := \{\boldsymbol{v}_i;\ i = 1, \ldots, m\}$ and $\mathsf{w} := \{\boldsymbol{w}_i;\ i = 1, \ldots, n\}$ respectively. Let the array of numbers $\{a_{ij}\}$ be the coefficients of the vectors $\mathcal{A}\boldsymbol{v}_j$ in the basis w as*

$$\begin{aligned}
\mathcal{A}v_1 \;&=\; a_{11}\boldsymbol{w}_1 + \cdots + a_{n1}\boldsymbol{w}_n, \\
&\;\;\vdots \\
\mathcal{A}v_m \;&=\; a_{1m}\boldsymbol{w}_1 + \cdots + a_{nm}\boldsymbol{w}_n,
\end{aligned} \tag{1.38}$$

For any vectors $\boldsymbol{f} \in \mathsf{V}$ and $\boldsymbol{g} \in \mathsf{W}$ with $\boldsymbol{g} = \mathcal{A}\boldsymbol{f}$, their basis coefficients are related by

$$\begin{aligned}
&\boldsymbol{f} = x_1\boldsymbol{v}_1 + \cdots + x_m\boldsymbol{v}_m \\
&\boldsymbol{g} = y_1\boldsymbol{w}_1 + \cdots + y_n\boldsymbol{w}_n
\end{aligned} \qquad \Leftrightarrow \qquad \boldsymbol{y} := \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} a_{11} \cdots a_{1m} \\ \vdots \quad\;\; \vdots \\ a_{n1} \cdots a_{nm} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} =: A\;\boldsymbol{x} \tag{1.39}$$

Thus we say that $A := \left[\begin{smallmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nm} \end{smallmatrix}\right]$ is the *matrix representation* of the operator $\mathcal{A}$ in the bases $\{\boldsymbol{v}_j\}$ and $\{\boldsymbol{w}_i\}$. Note the arrangement of the coefficients $\{a_{ij}\}$ in (1.38) in comparison to that in (1.39). As arrays, they are "transposes" of each other.

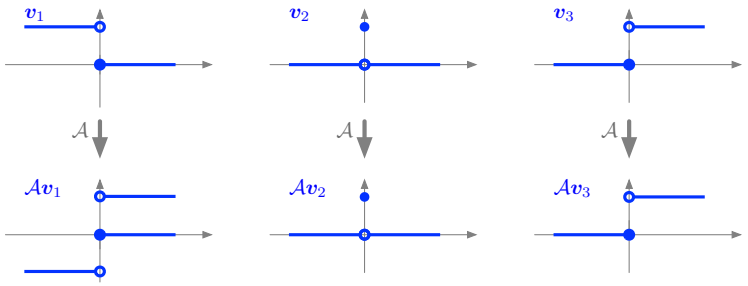
If the linear operator $\mathcal{A}$ is already described by a matrix representation (say with respect to the canonical basis of $\mathbb{R}^n$), then the lemma above describes how to change bases. This is worked out explicitly in Example 1.47 below. However, we begin here with a more abstract example where the operator $\mathcal{A}$ is first given in a "basis-free" manner.

Example 1.46. Recall the space $\mathbb{R}^{\{[-1,0),0,(0,1]\}}$ of Examples 1.4 and 1.19. Let $\mathcal{A}$ be an operator which acts on functions over $[-1, 1]$ in the following manner

$$(\mathcal{A}f)(x) \;:=\; \begin{cases} -f(x), & x \in [-1, 0), \\ f(x), & x = 0, \\ f(x) + f(-x), & x \in (0, 1]. \end{cases} \tag{1.40}$$

It's easy to verify that this operator is linear. If f is piece-wise constant on $[-1, 0)$ and $(0, 1]$, then so is $\mathcal{A}f$, and therefore $\mathcal{A}$ maps the vector space $\mathbb{R}^{\{[-1,0),0,(0,1]\}}$ to itself.

Now consider the basis $\mathsf{v} := \{\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3\}$ shown in Figure 1.4. What is the matrix representation of the operator $\mathcal{A}$ of (1.40) in this basis? To answer this, we simply repeat the procedure described earlier. In particular, we need to find the coefficients $\{a_{ij}\}$ of (1.37). Note that in this case, the two vector spaces V and W, and the two basis sets are the same respectively. The first step is to apply the operator $\mathcal{A}$ as described in (1.40) to each of the basis elements as shown here



The next step is to write out each $\mathcal{A}\boldsymbol{v}_i$ in terms of the basis, which will then yield the matrix representation coefficients $\{a_{ij}\}$ as follows

$$\begin{array}{rcl} \mathcal{A}\,\boldsymbol{v}_1 &=& -\boldsymbol{v}_1 + 0\boldsymbol{v}_2 + \boldsymbol{v}_3 \\ \mathcal{A}\,\boldsymbol{v}_2 &=& 0\boldsymbol{v}_1 + \boldsymbol{v}_2 + 0\boldsymbol{v}_3 \\ \mathcal{A}\,\boldsymbol{v}_3 &=& 0\boldsymbol{v}_1 + 0\boldsymbol{v}_2 + \boldsymbol{v}_3 \end{array} \quad \Rightarrow \quad A = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}.$$

Note again how the array of coefficients on the left is the transpose of the array of matrix entries on the right as per Lemma 1.45.

To take this example further, consider this same operator and $g = \mathcal{A}f$, but now we choose to represent f in the v basis and g in the w basis of Figure 1.4. What would its matrix representation be in this case? Following the procedure of Lemma 1.45 again, we calculate (from the above figure, and from the description of the basis w in Figure 1.4)

$$\begin{array}{rcl} \mathcal{A}\,\boldsymbol{v}_1 &=& 0\boldsymbol{w}_1 + 0\boldsymbol{w}_2 - \boldsymbol{w}_3 \\ \mathcal{A}\,\boldsymbol{v}_2 &=& 0\boldsymbol{w}_1 + \boldsymbol{w}_2 + 0\boldsymbol{w}_3 \\ \mathcal{A}\,\boldsymbol{v}_3 &=& \frac{1}{2}\boldsymbol{w}_1 + 0\boldsymbol{w}_2 - \frac{1}{2}\boldsymbol{w}_3 \end{array} \quad \Rightarrow \quad A = \begin{bmatrix} 0 & 0 & \frac{1}{2} \\ 0 & 1 & 0 \\ -1 & 0 & -\frac{1}{2} \end{bmatrix}.$$

Example 1.47. Recall that when we write a column vector $\boldsymbol{x} \in \mathbb{R}^n$, we are implicitly writing it using a basis expansion in the canonical basis

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = x_1 \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \cdots + x_n \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} = x_1\boldsymbol{e}_1 + \cdots + x_n\boldsymbol{e}_n.$$

Now given an $n \times m$ matrix A, it defines a linear operator $A : \mathbb{R}^m \to \mathbb{R}^n$ by the usual matrix-vector product. Comparing the matrix-vector product with (1.37)

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} \quad \Rightarrow \quad A\boldsymbol{e}_j = \begin{bmatrix} a_{1j} \\ \vdots \\ a_{nj} \end{bmatrix} = a_{1j}\boldsymbol{e}_1 + \cdots a_{nj}\boldsymbol{e}_n.$$

Thus the ij'th entry a_{ij} of the matrix A is the i'th coefficient of the expansion of the vector $A\boldsymbol{e}_j$ in the canonical basis $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$ of $\mathbb{R}^n$. In other words, when we write down a matrix, *it is the representation of a linear operator in the canonical basis*.

A natural question is what are the entries of the matrix representation of A if we choose different (other than the canonical) bases for $\mathbb{R}^m$ and $\mathbb{R}^n$? Lemma 1.45 gives the answer in general, and we will apply this lemma to give a compact formula using matrix notation as follows. Let $\mathsf{v} := \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m\}$ and $\mathsf{w} := \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n\}$ be bases for $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively. Each element of each basis can be written as a column vector, and those column vectors

can be "joined together" to form two matrices as follows

$$\boldsymbol{v}_j = \begin{bmatrix} v_{1j} \\ \vdots \\ v_{mj} \end{bmatrix},\ j = 1, .., m, \qquad V := \begin{bmatrix} \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_m \end{bmatrix} \qquad \Rightarrow \qquad V = \begin{bmatrix} v_{11} & \cdots & v_{1m} \\ \vdots & & \vdots \\ v_{m1} & \cdots & v_{mm} \end{bmatrix},$$

$$\boldsymbol{w}_j = \begin{bmatrix} w_{1j} \\ \vdots \\ w_{nj} \end{bmatrix},\ j = 1, .., n, \qquad W := \begin{bmatrix} \boldsymbol{w}_1 & \cdots & \boldsymbol{w}_n \end{bmatrix} \qquad \Rightarrow \qquad W = \begin{bmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & & \vdots \\ w_{n1} & \cdots & w_{nn} \end{bmatrix}.$$

Note that all these vectors are written in terms of the canonical basis, but we have now dropped notation like $[\boldsymbol{v}_j]_{\mathsf{e}}$ and V_{e} for simplicity.

Now let $\hat{A} = [\hat{a}_{ij}]$ be the matrix representation of A in the bases v and w of $\mathbb{R}^m$ and $\mathbb{R}^n$. By (1.38) (Lemma 1.45), the j'th column of $\hat{A}$ is given by the expansion coefficients

$$A\boldsymbol{v}_j \ = \ \hat{a}_{1j}\boldsymbol{w}_1 + \cdots + \hat{a}_{nj}\boldsymbol{w}_n \ = \ \begin{bmatrix} \boldsymbol{w}_1 & \cdots & \boldsymbol{w}_n \end{bmatrix} \begin{bmatrix} \hat{a}_{1j} \\ \vdots \\ \hat{a}_{nj} \end{bmatrix}, \qquad j = 1, .., m. \tag{1.41}$$

Combining all the vectors $A\boldsymbol{v}_j$ as columns of a matrix, we arrive at the *matrix equation*

$$\begin{bmatrix} \ & A & \ \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_m \end{bmatrix} \overset{1}{=} \begin{bmatrix} A\boldsymbol{v}_1 & \cdots & A\boldsymbol{v}_m \end{bmatrix} \overset{2}{=} \begin{bmatrix} \boldsymbol{w}_1 & \cdots & \boldsymbol{w}_n \end{bmatrix} \begin{bmatrix} \hat{a}_{11} & \cdots & \hat{a}_{1m} \\ \vdots & & \vdots \\ \hat{a}_{n1} & \cdots & \hat{a}_{nm} \end{bmatrix} \tag{1.42}$$

$$\Rightarrow \qquad A\ V \ = \ W\ \hat{A} \qquad \Rightarrow \qquad \hat{A} \ = \ W^{-1}AV. \tag{1.43}$$

Note that $\overset{1}{=}$ follows from the definition of the matrix-matrix product, while $\overset{2}{=}$ is simply the m equations (1.41) expressed as a single matrix equation.

The matrix formula (1.43) is undeniably elegant and compact. It involves the original matrix A (i.e. the representation of the linear transformation in the canonical basis), as well as the matrices V and W which contain all the bases vectors. This is an important enough result to summarize as a lemma.

Lemma 1.48. *Let $A : \mathbb{R}^m \to \mathbb{R}^n$ be an $n \times m$ matrix representing (in the canonical bases) a linear operator. Let $\mathsf{v} := \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m\}$ and $\mathsf{w} := \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n\}$ be vectors in $\mathbb{R}^m$ and $\mathbb{R}^n$ which are bases sets respectively. Then the matrix representation $\hat{A}$ of A in these bases is given by*

$$\boxed{\hat{A} \ = \ W^{-1}AV}, \qquad V := \begin{bmatrix} \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_m \end{bmatrix}, \quad W := \begin{bmatrix} \boldsymbol{w}_1 & \cdots & \boldsymbol{w}_n \end{bmatrix}.$$

In other words

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \ & A & \ \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} \qquad \textit{and} \qquad \begin{aligned} \boldsymbol{x} &= x_1\boldsymbol{e}_1 + \cdots + x_n\boldsymbol{e}_m = \hat{x}_1\boldsymbol{v}_1 + \cdots + \hat{x}_n\boldsymbol{v}_m \\ \boldsymbol{y} &= y_1\boldsymbol{e}_1 + \cdots + y_n\boldsymbol{e}_n = \hat{y}_1\boldsymbol{w}_1 + \cdots + \hat{y}_n\boldsymbol{w}_n \end{aligned}$$

$$\Downarrow$$

$$\begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} \ & W^{-1}AV & \ \end{bmatrix} \begin{bmatrix} \hat{x}_1 \\ \vdots \\ \hat{x}_m \end{bmatrix}$$

We note here that the arguments in (1.41) and (1.42) are done using *partitioned matrix notation*, which enables writing complicated sets of scalar equations as compact matrix equations. Much more will be done with partitioned matrix notation in Chapter 7.

Similarity Transformations as a Change of Basis Representations

An important special case of formula (1.43) is when the matrix is "square" $A : \mathbb{R}^n \to \mathbb{R}^n$, and the same new basis $\mathsf{v} := \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ is used for both the domain and range. The two bases v and w in the previous example are the same, the new matrix representation is

$$\hat{A} \;=\; V^{-1}\, A\, V, \tag{1.44}$$

where the columns of V are the new basis vectors.

A change of basis can be regarded as a transformation $V : \mathbb{R}^n \to \mathbb{R}^n$ which takes the canonical vectors $\boldsymbol{e}_i$ to the vectors $\boldsymbol{v}_i$ respectively

$$\boldsymbol{v}_i \;=\; V\boldsymbol{e}_i, \qquad i = 1, .., n.$$

The linearity property implies that this defines a transformation on all of $\mathbb{R}^n$ since $\{\boldsymbol{e}_i\}$ is a basis of $\mathbb{R}^n$. On the other hand, the relation (1.44) defines *a transformation on matrices*

$$A \;\longmapsto\; \hat{A} = V^{-1}AV.$$

A transformation of this form is called a *similarity transformation*. The term "similarity" is evocative. Both A and $\hat{A}$ are the same linear transformation, but expressed in two different bases. All the "basis-free" properties of a transformation (i.e. whether it is one-to-one, onto, its rank and nullity, and as we will see later, its eigenvalues) are exactly the same for A and $\hat{A}$. It is in this sense that A and $\hat{A}$ are similar.

Similarity transformations play a major role in linear algebra. Diagonalizing a matrix, or transforming it into Jordan normal form is done by finding a special basis (the choice depends on the given matrix) in which $\hat{A}$ in (1.44) has that form. Properties such as range and null spaces and eigenvalues can then be easily "read off" the special form of $\hat{A}$. Another very useful special form, namely the "Singular Value Decomposition" (SVD), is not a similarity transformation, but rather different bases in the domain and range are used, and that transformation is of the type (1.43).

Exercises

Exercise 1.1

Consider the proof of Lemma 1.16. Starting with the first equation in (1.12), at least one of the coefficients $a_{11}, \ldots, a_{1m}$ is non-zero. Assume without loss of generality that it is a_{11} (otherwise reindex the set w). Then

$$\boldsymbol{w}_1 \;=\; \frac{1}{a_{11}}\left(\boldsymbol{v}_1 + a_{12}\boldsymbol{w}_2 + \cdots + a_{1m}\boldsymbol{w}_m\right).$$

This can be substituted for $\boldsymbol{w}$ in (1.11), and the equations become

$$\begin{bmatrix} \boldsymbol{v}_2 \\ \vdots \\ \boldsymbol{v}_n \end{bmatrix} = \begin{bmatrix} * & \cdots & \cdots & * \\ \vdots & & & \vdots \\ * & \cdots & \cdots & * \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1 \\ \boldsymbol{w}_2 \\ \vdots \\ \boldsymbol{w}_m \end{bmatrix}.$$

Show that by repeating this process recursively, the form on the right in (1.12) is obtained.

Chapter 2

Norm and Inner Product Spaces

Additional structures can be layered on top of the additive vector space structure. The most basic is a metric *which measures distances between two points. If the metric is "compatible" with the vector space structure, i.e. translation invariant and scaling equivariant, then it becomes a* norm*, which is a measure of the length of a vector. A vector space can typically be endowed with many different norms, and the choice of the proper norm depends on the application. The geometry of a normed space is determined by the shape of its unit ball, and there is a one-to-one correspondence between convex sets with certain properties and norms. Normed spaces where the norms satisfy additional properties can be endowed with an* inner product*, which behaves similarly to the standard dot product in Euclidean space. Inner products give a notion of angles between abstract vectors, and induce a rather structured geometry which can be exploited in devising algorithms for construction and reconstruction of vectors and functions.*

This chapter is concerned primarily with the basic "geometric" aspects of vector spaces.

Introduction

On abstract sets, we can define geometrical notions such as distances, lengths and angles. The most basic notion is that of a distance between any two points, also called a *metric*. This makes a set into a so-called *metric space*. If in addition that set has a vector space structure, then a translation invariant and homogenous metric defines a norm (or length) of a vector. The distance between two points then becomes the length of the vector joining those two points, and those distances are unchanged by parallel translations of the two points, and also scale homogeneously as the vector is multiplied by a scalar. Vector spaces equipped with such vector norms are called *Normed Vector Spaces*.

If the vector norm satisfies further properties such as the parallelogram law, then we can define an inner product which has similar properties to the dot product in Euclidean space. The inner product gives a notion of angles and orthogonality akin to those in Euclidean geometry. The three types of overlaid structures are thus a metric space as the most general, then a normed vector space as a special case, and then an inner product space as the most special structure. The notion of distance in an inner product space is thus highly restricted, and has to obey several properties that hold in Euclidean geometry, but may not hold in more general geometries. This hierarchy of structures is illustrated in Figure 2.1.

Recall that in $\mathbb{R}^n$, the length of a vector v is traditionally defined as follows

$$\|v\|_2 \;:=\; \sqrt{v_1^2 + \cdots + v_n^2}. \tag{2.1}$$

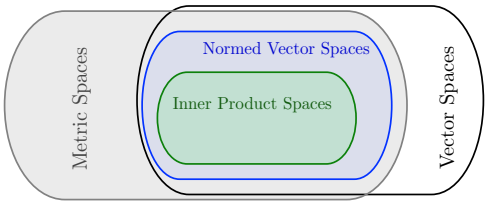


Figure 2.1: The hierarchy of structures on abstract spaces. One of the most basic is the *vector space* structure which allows for addition and scaling of elements, but has no notion of length or distances. A *metric space* has a notion of distances between two points, but may or may not be itself a vector space. If it is, and the metric is compatible with the vector space structure in the sense of being translation invariant and homogenous, then it is called a *normed vector space*. In such spaces, the "norm" is the length of a vector, and distances between points are given by the length of the vector joining them. In an *inner product space* the length of a vector is given by the inner product of a vector with itself. The inner product also characterizes angles between vectors, and therefore induces a notion of orthogonality.

This definition is motivated from the generalization of the Pythagorean theorem to more than 2 dimensions. This measure of vector length is a special case of a *norm* on a vector space, and is referred as the *Euclidean* norm on $\mathbb{R}^n$. It is also called the "2-norm", which explains the subscript in the notation $\|.\|_2$. It is a special case of more general norms that will be introduced shortly.

The vector length formula (2.1) above also gives a *metric* on $\mathbb{R}^n$, where the distance between two points is given by the length of the vector connecting the points

$$d(v,w) \;:=\; \|v-w\|_2 \;=\; \sqrt{(v_1-w_1)^2+\cdots+(v_n-w_n)^2}.$$

These familiar geometric notions in so-called Euclidean space can be abstracted and generalized to function spaces.

2.1 Metric Spaces

We start from the most basic structure of a *metric space*, which is just a set (not necessarily a vector space) with a notion of distance between its members.

Definition 2.1. *A* Metric Space *is a set* M *and a real-valued, non-negative distance function* $\mathsf{d}(.,.) : \mathsf{M}\times\mathsf{M} \longrightarrow \mathbb{R}$ *with the following properties*

- Symmetry: $\mathsf{d}(v,w) \;=\; \mathsf{d}(w,v)$.
  This is a natural requirement that the distance from v to w should be the same as the distance from w to v.

- Definiteness: $\mathsf{d}(v,w)=0 \quad\Leftrightarrow\quad v=w$.
  This means that the metric "separates distinct points", so that if two points v and w are distinct, then $\mathsf{d}(v,w)\neq 0$.

- Triangle Inequality: *For any three points u,v,w we have*

  $$\mathsf{d}(u,w) \;\leq\; \mathsf{d}(u,v)+\mathsf{d}(v,w). \tag{2.2}$$

  This means that there are "no short cuts", i.e. the distance from u to w cannot be made shorter by going through an intermediate point v, since that total traveled distance $\mathsf{d}(u,v)+\mathsf{d}(v,w)$ will be at least as large as the direct travel distance $\mathsf{d}(u,w)$.

Figure 2.2: The *2-sphere* in $\mathbb{R}^3$ is $\{x \in \mathbb{R}^3;\ \|x\|_2 = 1\} \subset \mathbb{R}^3$, the set of vectors of length 1. *(Left)* The 2-sphere is not a vector space (i.e. not a subspace of $\mathbb{R}^3$) since adding two vectors on the sphere produces a vector outside of it. *(Right)* It is however a metric space. The metric is given by the length of the *geodesic* (a path within the 2-sphere which is of minimum length) joining two points. Think of the 2-sphere as the surface of the earth. The geodesics are then the great circle arcs joining the two points.

The triangle inequality is illustrated in Figure 2.3. Although not stated explicitly in the requirements above, the distance function is always positive between distinct points. This follows from the three requirements above by observing that

$$\mathsf{d}(v,w) \overset{1}{=} \frac{1}{2}\left(\mathsf{d}(v,w) + \mathsf{d}(w,v)\right) \overset{2}{\geq} \frac{1}{2}\,\mathsf{d}(v,v) = 0,$$

where $\overset{1}{=}$ follows from symmetry, and $\overset{2}{\geq}$ from the triangle inequality. Finally, combining $\mathsf{d}(v,w) \geq 0$ with definiteness implies that $\mathsf{d}(v,w) > 0$ if $v \neq w$.

A metric space does not necessarily have to be a vector space. An example of such a space is the sphere (the shell of the unit ball) shown in Figure 2.2. It is clearly not a vector subspace of $\mathbb{R}^3$ since addition of vectors does not remain in the set. It is however a metric space when the metric is defined as the length of minimum-length path joining two points (called a *geodesic*). We will mostly deal with metrics on vector spaces, the most useful of which have additional structure that renders them into *normed* vector spaces.

2.2 Normed Vector Spaces

We begin with the formal definition of a norm, and then show how it induces a metric. The metric induced by a norm on a vector space has the additional important properties of translation invariance and homogeneity as well.

Definition 2.2. *A* Normed Space V *is a vector space with a* Norm *(a measure of the length of each vector), which is a real-valued functional* $\|.\| : \mathsf{V} \longrightarrow \mathbb{R}$ *with the following properties*

- Definitness: *For any vector* $v \in \mathsf{V}$*, its norm is zero iff it is the zero vector*

$$\|v\| = 0 \quad \Leftrightarrow \quad v = 0.$$

- Homogeneity: *If a vector* v *is scaled by a scalar* α*, then its norm is proportionally scaled*

$$\|\alpha v\| = |\alpha|\ \|v\|.$$

- Triangle Inequality: *For any three vectors* u, v, w *we have*

$$\|v+w\| \leq \|v\| + \|w\| \qquad \textit{or equivalently} \qquad \|v-w\| \leq \|v\| + \|w\|$$

Note that the equivalence of the two forms of the triangle inequality follows from simply substituting $-w$ for w in either of the forms, and using the homogeneity property which implies $\|-w\| = \|w\|$. This is depicted in the last panel of Figure 2.3.

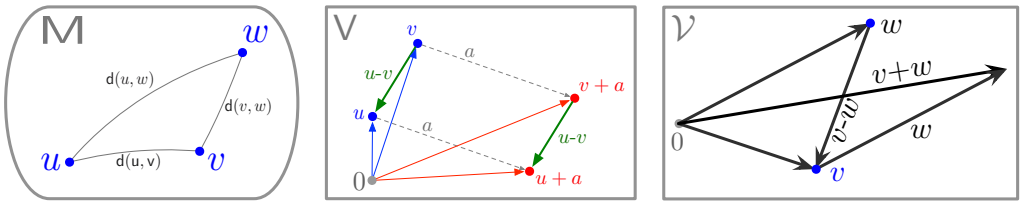


Figure 2.3: The requirements on metrics and norms in abstract spaces generalize the ordinary notions of distance and vector length in standard geometry. (*Left*) A metric space M is a set (not necessarily a vector space) with a distance function $\mathsf{d}(.,.)$ that satisfies the metric properties. Here the triangle inequality $\mathsf{d}(u,w) \leq \mathsf{d}(u,v)+\mathsf{d}(v,w)$ is depicted: one cannot decrease the total "travelled distance" between u and w by going through an intermediary point v. (*Middle*) In a normed vector space, the norm induces a metric where the distance between any two points u and v is the norm (aka vector length) $d(u,v) := \|u-v\|$ of the vector joining those two points. This metric is translation invariant $d(u+a, v+a) = \|u+a-(v+a)\| = d(u,v)$. (*Right*) Two equivalent forms of the triangle inequality $\|v-w\| \leq \|v\| + \|w\|$ and $\|v+w\| \leq \|v\| + \|w\|$ are depicted.

Given a norm, we can think about proposing a metric where the distance between two points is the length of the vector that joins them

$$d(x,y) \;:=\; \|x-y\|. \tag{2.3}$$

This satisfies all the properties of a metric as shown by

$$d(x,y) = \|x-y\| = \|y-x\| = d(y,x), \qquad \text{(Symmetry follows from } \|v\| = \|\text{-}v\|)$$
$$0 = d(x,y) = \|x-y\| \;\Leftrightarrow\; x-y=0 \;\Leftrightarrow\; x=y, \qquad \text{(Definiteness follows from definiteness of } \|.\|)$$
$$\mathsf{d}(u,w) = \|u-w\| = \|(u-v)+(v-w)\| \;\leq\; \|u-v\| + \|v-w\| = \mathsf{d}(u,v) + \mathsf{d}(v,w) \qquad \text{(Triangle inequality)}$$

The metric defined from a norm by (2.3) has some additional properties that not all metrics have. These properties can be understood as "compatibility properties" between the metric and the vector space structure. The first property is that the distance between two points remains the same if we translate those two points equally in a parallel manner

$$\mathsf{d}(v+a, w+a) \;=\; \|(v+a)-(w+a)\| \;=\; \|v-w\| \;=\; \mathsf{d}(v,w), \tag{2.4}$$

i.e. the metric is *translation invariant*. This property is depicted in Figure 2.3.

Another property comes from the fact that any point in a vector space can be scaled towards or away from the origin by multiplying it by a scalar. The distance between two points should scale in the same manner if we scale both points equally, i.e.

$$d(\alpha v, \alpha w) \;=\; \|\alpha v - \alpha w\| \;=\; \|\alpha\,(v-w)\| \;=\; |\alpha|\;\|v-w\| \;=\; |\alpha|\;d(v,w).$$

Note that this property and translation invariance (2.4) only make sense in a vector space. On a general metric space, the operations of addition $v+a$ and scaling αv are not necessarily defined.

We have so far seen that a metric induced by norm satisfies the two properties above. The converse is also true, if a metric posses those two properties, then it is a metric that is induced by a norm.

Theorem 2.3. *Let d be a metric on a vector space V with the following additional properties*

- Translation Invariance: *For any two vectors v and w, and any translation $a \in \mathsf{V}$*

$$\mathsf{d}(v,w) \;=\; \mathsf{d}(v+a, w+a). \tag{2.5}$$

- Homogeneity: *(or scale proportionality) For vectors v and w and any scalar α*

$$\mathsf{d}(\alpha v, \alpha w) \;=\; |\alpha|\; \mathsf{d}(v, w)$$

Then the metric d makes V into a normed vector space with the norm

$$\|v\| \;:=\; \mathsf{d}(0, v). \tag{2.6}$$

Another way to state this theorem is to say that *a metric is induced by a norm iff the metric is homogenous and translation invariant.*

Proof. Its immediate to show that the norm thus defined is definite and homogenous

$$\begin{aligned} 0 \;=\; \|v\| \;=\; \mathsf{d}(0, v) \quad &\Rightarrow \quad v \;=\; 0, && \text{(by definiteness of } d\text{)} \\ \mathsf{d}(0, \alpha v) \;=\; |\alpha|\; \mathsf{d}(0, v) \quad &\Rightarrow \quad \|\alpha v\| \;=\; |\alpha|\; \|v\|. \end{aligned}$$

In addition, the definition (2.6) together with translation invariance (2.5) imply that the distance between two points is the length of the vector joining them. This is because we can translate one point to the origin, and then measure the distance from zero to the other point by (2.6), which will be the norm of the difference

$$\mathsf{d}(u, w) \;=\; \mathsf{d}(u - u, w - u) \;=\; \mathsf{d}(0, w - u) \;=\; \|w - u\|.$$

The triangle inequality also follows from this

$$\|u - w\| \;=\; \mathsf{d}(u, w) \;\leq\; \mathsf{d}(u, v) + \mathsf{d}(v, w) \;=\; \|u - v\| + \|v - w\|.$$

Therefore, the definition (2.6) satisfies all the properties of a norm. □

It is rare that one would use a metric on a vector space that does not have the natural translation invariance and homogeneity properties. We therefore always work with normed vector spaces whenever a metric is needed.

An important property of the norm functional $\|.\| : \mathsf{V} \to \mathbb{R}$ is that it is a *convex functional*[1]. This can be easily verified as follows. Given $\alpha \in [0, 1]$

$$\|\alpha\; v_1 \;+\; (1 - \alpha)\; v_2\| \;\leq\; \alpha\; \|v_1\| \;+\; (1 - \alpha)\; \|v_2\|,$$

which follows from the triangle inequality and homogeneity of the norm. A particular sub-level set of the norm functional is the *unit ball*, namely the set of all vectors with norm less than one

$$\mathsf{B} \;:=\; \{v \in \mathbb{R}^n;\;\; \|v\| \leq 1\}\,.$$

The geometry of the unit ball of a normed vector space encodes many of the properties of a particular norm. Aside from being a convex set, it has other properties as well. In fact, any convex set that has certain other properties as outlined in Appendix 2.B induces a norm. The next few examples serve to illustrate some of those geometrical properties.

[1] The reader not familiar with the basics of convexity should now consult Appendix 2.A.

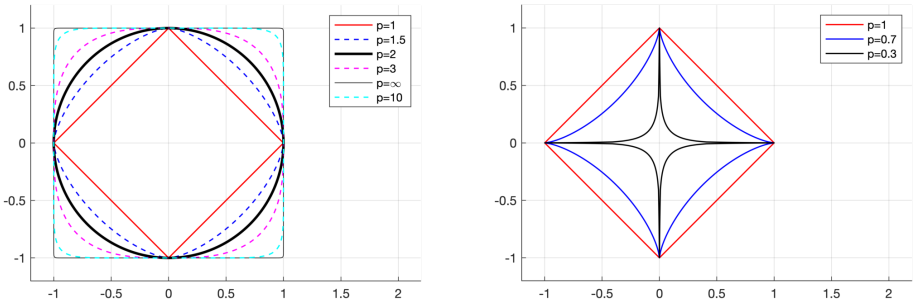


Figure 2.4: (*Left*) The boundaries of the unit balls of the $\|.\|_p$ norms for various values of $1 \leq p \leq \infty$. These curves represent points that are equidistant from the origin with $\|.\|_p$ as the distance measure. Note how the $p = 2$ norm is the only rotationally symmetric one (the unit ball is a perfect sphere), and how the $p = 10$ case is very close to the $p = \infty$ case. The boundary curves are smooth (infinitely differentiable) for the cases $1 < p < \infty$, but have corners for the cases $p = 1, \infty$. (*Right*) The boundaries of the sets $\|v\|_p \leq 1$ for $p < 1$. These are clearly not convex sets, implying that for $p < 1$, $\|.\|_p$ does not satisfy the triangle inequality, and is therefore not a norm.

2.2.1 Finite Dimensional Examples

We started this section with the Euclidean norm (2.1) in $\mathbb{R}^n$. There are many other possible norms on $\mathbb{R}^n$ as well. The most common are the so-called p-norms

$$\|v\|_p \;:=\; \Big(|v_1|^p + \cdots + |v_n|^p\Big)^{1/p}, \qquad 1 \leq p < \infty. \tag{2.7}$$

Note that $p = 2$ is the special case of the Euclidean norm. The $\|.\|_\infty$ norm is defined a little differently using

$$\|v\|_\infty \;:=\; \max\{|v_1|, \ldots, |v_n|\}. \tag{2.8}$$

It is possible to show that $\lim_{p\to\infty} \|v\|_p = \|v\|_\infty$, which explains the notation for $\|.\|_\infty$.

The unit balls of several representative p-norms are shown in Figure 2.4. The reader should note that the unit balls for $p \in [1, \infty]$ appear to be convex sets. Figure 2.4 also shows unit balls of the quantity $\|.\|_p$ for $p < 1$. It is important to note that in these latter cases, $\|.\|_p$ is not actually a norm since the unit balls of $\|.\|_p$ for $p < 1$ are clearly not convex. As already stated, this implies that the quantity $\|.\|_p$ does not satisfy the triangle inequality for $p < 1$.

Figure 2.4 also shows how $\lim_{p\to\infty} \|v\|_p = \|v\|_\infty$. Note how the unit ball for $p = 10$ is already almost identical to the unit ball for $p = \infty$. For $1 < p < \infty$, the curves are smooth (infinitely differentiable) implying that the norm function $\|.\|_p$ is smooth for these cases. For the extreme cases of $p = 1, \infty$, while the norm function is continuous, it is however not differentiable. This property plays an important role in some optimization problems.

The unit balls of norms serve as a nice geometrical illustration of the comparative properties of norms. Let's see what it means for one unit ball to be contained in another. Let $\|.\|_a$ and $\|.\|_b$ be two norms and B_a and B_b their respective unit balls, and note that

$$\|v\|_a \leq \|v\|_b \quad\Rightarrow\quad (\|v\|_b \leq 1 \;\Rightarrow\; \|v\|_a \leq 1) \quad\Rightarrow\quad (v \in \mathsf{B}_b \;\Rightarrow\; v \in \mathsf{B}_a) \quad\Rightarrow\quad \mathsf{B}_b \subseteq \mathsf{B}_a.$$

Conversely

$$\mathsf{B}_b \subseteq \mathsf{B}_a \quad\Rightarrow\quad (\|v\|_b = 1 \;\Rightarrow\; \|v\|_a \leq 1) \quad\Rightarrow\quad \|v\|_a \leq \|v\|_b. \tag{2.9}$$

Note how the smaller the norm is, the bigger is its unit ball (to achieve unit norm, the vector has to be longer). Thus we see that unit ball containment implies a bound on norms, but

in the opposite order

$$\mathsf{B}_{\mathrm{b}} \subseteq \mathsf{B}_{\mathrm{a}} \qquad \Leftrightarrow \qquad \|v\|_{\mathrm{a}} \leq \|v\|_{\mathrm{b}}. \tag{2.10}$$

Examining Figure 2.4 we observe the following containment of the p-norms for $p \in [1, \infty]$

$$\mathsf{B}_1 \subseteq \cdots \subseteq \mathsf{B}_2 \subseteq \cdots \mathsf{B}_\infty,$$

which implies the following inequalities (in reverse order of containment) between the norms

$$\|v\|_\infty \leq \cdots \|v\|_2 \leq \cdots \|v\|_1.$$

These, and other inequalities are discussed in further detail in Appendix 2.C.

It remains to show that the p-norms (2.7),(2.8) satisfy the properties of a norm as stated in Definition 2.2. Definiteness and homogeneity are easy to verify and the reader should do so as an exercise. Verifying that they satisfy the triangle inequality requires a little more work. We first do the simplest cases of the 1 and ∞ norms

$$\begin{aligned}
\|v+w\|_1 = \sum_{i=1}^n |v_i + w_i| &\leq \sum_{i=1}^n |v_i| + |w_i| = \sum_{i=1}^n |v_i| + \sum_{i=1}^n |w_i| = \|v\|_1 + \|w\|_1 \\
\|v+w\|_\infty = \max_{1\leq i\leq n} |v_i + w_i| &\leq \max_{1\leq i\leq n} \Big(|v_i| + |w_i|\Big) \leq \max_{1\leq i\leq n} |v_i| + \max_{1\leq i\leq n} |w_i| \\
&= \|v\|_\infty + \|w\|_\infty.
\end{aligned}$$

The triangle inequality for the other p-norms is the statement of the *Minkowski inequality* which simply states

$$\|v+w\|_p \leq \|v\|_p + \|w\|_p$$

for $p \in [1, \infty]$. We will revisit this inequality and other related inequalities such as the Hölder and Cauchy-Schwartz inequalities in Chapter 4. They are best understood using the concept of duality of normed vector spaces which also provides intuitive geometrical interpretations.

We close by giving a geometrical argument for the Minkowski inequality. We will show that the set $\{v \in \mathbb{R}^n;\ \|v\|_p \leq 1\}$ for $p \in [1, \infty)$ is convex. By Theorem 2.9 it would then follow that $\|.\|_p$ is a norm, and in particular, it satisfies the triangle inequality. First observe that the function $|x|^p$ (for a scalar x) is convex for $p \in [1, \infty)$, while it is not for $p < 1$. Now given two vectors $\|v\|_p \leq 1$ and $\|w\|_p \leq 1$, we take a convex combination

$$\begin{aligned}
\|\alpha v + (1-\alpha)w\|_p^p = \sum_{i=1}^n |\alpha v_i + (1-\alpha) w_i|^p &\leq \sum_{i=1}^n \alpha\,|v_i|^p + (1-\alpha)\,|w_i|^p \\
&= \alpha \|v\|_p^p + (1-\alpha)\|w\|_p^p \leq 1,
\end{aligned}$$

and note how we used the convexity of the scalar function $|x|^p$ in the inequality above. This argument shows why the set $\{v \in \mathbb{R}^n;\ \|v\|_p \leq 1\} = \left\{v \in \mathbb{R}^n;\ \|v\|_p^p \leq 1\right\}$ is convex for $p \in [0, \infty)$, while it is not for $p < 1$.

2.2.2 Function Space Examples

The typical function space examples are those of functions on some set Ω which is a subset of $\mathbb{R}^n$ or $\mathbb{C}^n$. These functions will typically (but not always) take values of real or complex numbers (we call these *scalar-valued* functions), or take values as n-vectors (real or complex, we call these *vector-valued* functions), or more generally take values in some vector space V. We will also view sequence spaces as function spaces since finite or infinite sequences are functions on some subset of $\mathbb{Z}$, and more generally on some subset of $\mathbb{Z}^d$.

The counter part of the p-norms in function space are the function space L^p and the sequence spaces ℓ^p. The sequence spaces are easier to deal with, so we start with them.

The ℓ^p Spaces

The ℓ^p spaces contain functions defined on a discrete set, typically a subset of the integers $\mathbb{Z}$ or the integer lattice $\mathbb{Z}^d$. For example

$$\begin{aligned}\ell^p(\mathbb{Z}) &:= \left\{ v : \mathbb{Z} \to \mathbb{R};\ \|v\|_p^p := \sum_{i\in\mathbb{Z}} |v_i|^p < \infty \right\}, && p \in [1,\infty) \\ &:= \left\{ v : \mathbb{Z} \to \mathbb{R};\ \|v\|_\infty := \sup_{i\in\mathbb{Z}} |v_i| < \infty \right\}, && p = \infty,\end{aligned}$$

which should be thought of as the space of all double-sided sequences of finite p-norm. We will use the notation $\ell^p(\Omega)$ where $\Omega \subseteq \mathbb{Z}$ (or $\Omega \subseteq \mathbb{Z}^d$) to specify the domain of the sequence index, e.g.

$$\ell^p(\mathbb{N}) := \left\{ v : \mathbb{N} \to \mathbb{R};\ \|v\|_p^p := \sum_{i=0}^{\infty} |v_i|^p < \infty \right\},$$

to denote a space of one-sided sequences with finite p-norm.

Recall Figure 1.1 and observe that we can identify $\mathbb{R}^n$ with the p-norm on the one hand with the function space

$$\ell^p(\mathfrak{n}) = \ell^p\Big(\{1,\ldots,N\}\Big) = \left\{ v : \mathfrak{n} \to \mathbb{R};\ \|v\|_p^p := \sum_{i=1}^{n} |v_i|^p \right\} = \mathbb{R}^n \text{ (with the } p\text{-norm)}$$

on the other. Note that the norm is always finite in this case, and there is no need to include the finiteness clause in the set definition. The sequence spaces ℓ^p are the closest to the finite dimensional $\mathbb{R}^n$ with p-norms. An element of $\ell^p(\mathbb{N})$ can be thought of as a one-sided sequence

$$v = (v_0, v_1, v_2, \ldots)$$

or a semi-infinite vector with components $\{v_i\}$. All of the arguments that we went through to verify that the $\|.\|_p$ norms in $\mathbb{R}^n$ are actually norms (i.e. definiteness, homogeneity and the triangle inequality) apply without change to the case of $\ell^p(\Omega)$ for any $\Omega \subseteq \mathbb{Z}^d$.

We will also have occasion to work with spaces of *vector-valued* sequences. The vector-valued ℓ^p spaces are defined similarly to the above. For any subset $\Omega \subseteq \mathbb{Z}^d$

$$\ell_n^p(\Omega) := \left\{ v : \Omega \to \mathbb{R}^n;\ \|v\|_p^p := \sum_{i\in\Omega\subseteq\mathbb{Z}^d} \|v_i\|_p^p < \infty \right\}. \tag{2.11}$$

The notation should be parsed carefully. There four integers, p, n, d and i which all play different roles. d is the dimension of underlying domain in $\mathbb{Z}^d$ (take it to be 1 for the sake of this explanation). At each $i \in \Omega \subseteq \mathbb{Z}^d$, v_i is an n-vector in $\mathbb{R}^n$. $\|v_i\|_p$ is the p-norm of that vector (if the signal were scalar-valued, i.e. $n = 1$, then we would simply have the absolute value $|v_i|^p$ in the expression above). The p-norm of the entire function v is then computed by summing all p-powers of the p-norm $\|v_i\|_p$ of those n-vectors at all points $i \in \Omega$ in the domain of the function.

We can extend the definition (2.11) a little further by generalizing from n-vector-valued sequences to sequences that take values in any normed vector space V. The definition is

$$\ell_{\mathsf{V}}^p(\Omega) := \left\{ v : \Omega \to \mathsf{V};\ \|v\|_p^p := \sum_{i\in\Omega\subseteq\mathbb{Z}^d} \|v_i\|_{\mathsf{V}}^p < \infty \right\}. \tag{2.12}$$

Note that $\|v_i\|_{\mathsf{V}}$ is the norm of the i'th element of the sequence in the normed space V. The reader should compare this with (2.11), where $\|v_i\|_p$ is the norm of the i'th sequence element (which is itself an n-vector) in $\mathbb{R}^n$ equipped with the p-norm.

We will often simplify the notation by writing ℓ^p instead of $\ell^p_{\mathsf{V}}(\Omega)$ or $\ell^p_n(\Omega)$ when the choice of domain and range spaces are clear from context. It turns out that this detail is immaterial to many of the algebraic relations and manipulations we use.

The L^p Spaces

Now we define the so-called Lebesgue spaces of functions of a continuous variable. Let $\Omega \subset \mathbb{R}^d$ be any subset. We define $\mathsf{L}^p_n(\Omega)$ as the space of all functions $f : \Omega \to \mathbb{R}^n$ that have finite p-norm-power integrals[2]

$$\begin{aligned}
\mathsf{L}^p_n(\Omega) &:= \left\{ v : \Omega \to \mathbb{R}^n; \quad \|v\|_p^p := \int_\Omega \|v(x)\|_p^p \, dx < \infty \right\}, \qquad p \in [1, \infty), \\
&:= \left\{ v : \Omega \to \mathbb{R}^n; \quad \|v\|_\infty := \sup_{x \in \Omega} \|v(x)\|_\infty < \infty \right\}, \qquad p = \infty.
\end{aligned}$$

It is instructive to compare this definition with (2.11). At each $x \in \Omega$, we take the p-norm of the n-vector $v(x)$, raise it to the p power, and then integrate (rather than sum) over the entire domain.

It can be shown that the norm $\|v\|_p$ defined here satisfies all the requirements of a norm, and thus this space is indeed a normed vector space closed under additions and scalings. The arguments are very similar to the ℓ^p case. In particular, the Minkowski inequalities for $\mathsf{L}^p(\Omega)$ follow by a similar argument as follows

$$\begin{aligned}
\|\alpha v + (1-\alpha) w\|_p^p = \int_\Omega \left| \alpha v(x) + (1-\alpha) w(x) \right|^p dx &\leq \int_\Omega \left(\alpha \, |v(x)|^p + (1-\alpha) \, |w(x)|^p \right) dx \\
&= \alpha \|v\|_p^p \; + \; (1-\alpha) \|w\|_p^p,
\end{aligned}$$

where the inequality follows from the convexity of the function $|.|^p$ for $p \in [1, \infty)$. This shows that the sets $\{\|v\|_p \leq 1\} = \left\{\|v\|_p^p \leq 1\right\}$ are convex, and therefore it follows from Theorem 2.9 (Appendix 2.B) that the L^p norms satisfy the triangle inequality.

The most commonly used L^p spaces are L^1, L^2 and L^∞. These different norms tend to weight different signal behaviors differently. An example is shown in Figure 2.5 which highlights one contrast between the L^1, L^2 and L^{20} norms (the latter is used as a sort of "approximation" to the L^∞ norm). The L^2 norm tends to emphasize the contribution of the peaks of signals more than the L^1 norm, and similarly the L^{20} norm tends to emphasize the peaks more than the L^2 norm. The extreme case of this situation is the limit $\lim_{p \to \infty} \|v\|_p = \|v\|_\infty$, which means that for large p, essentially only the peak of the signal values contributes to the norm. For comparison purposes, the signal in Figure 2.5 has been normalized so that its peak value is 1. This is without loss of generality since $\|\gamma v\|_p = |\gamma| \|v\|_p$, and therefore for comparison across different values of p, the factor γ is the same.

For both the ℓ^p and L^p sets of spaces, the case $p = 2$ is special. These are sets of square integrable signals, and it turns out that in addition to forming normed vector spaces, their norms have a very special property in that they come from an *inner product*. Such spaces have a much richer geometry which we examine next.

[2] For the case of $p = \infty$, the definition should be done with the "essential supremum" ess sup instead of the supremum sup, as the values of the function on sets of measure zero do not contribute to the norm. This technicality is not worth spending time on. For the classes of functions we deal with, sup and ess sup are the same.

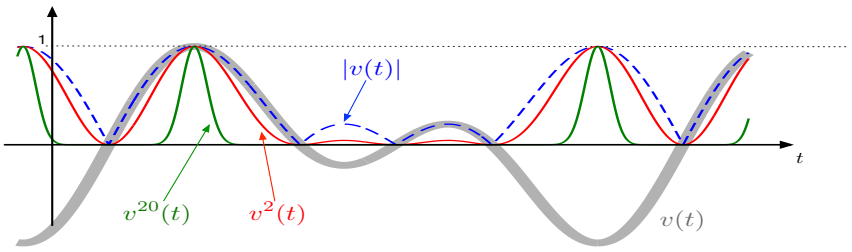


Figure 2.5: A graph of a signal v is shown as the solid gray curve. For comparison, its peak value has been normalized to 1. The other curves show its absolute value $|v(t)|$ (dashed blue), its square $v^2(t)$ (red), and $v^{20}(t)$ (green). The L^1, L^2 and L^{20} norms of v are given by the areas under the respective curves, and then taking the 1, 1/2 and 1/20 power of those quantities respectively. Notice how the values of the signal near its peaks contribute more to the L^{20} relative to the L^2 norm, as well as more to the L^2 relative to the L^1 norm of the signal. Thus as $p \to \infty$, the L^p norm tends to be dominated by the portions of the signal near where its peak value is achieved.

2.3 Inner Product Spaces

For any vector $v \in \mathbb{R}^n$ let v^* denote its transpose. For a complex number $\alpha \in \mathbb{C}$, let α^* denote its complex conjugate, and for a vector $v \in \mathbb{C}^n$, let v^* denote is complex conjugate transpose. The well-known "dot product" of vectors is

$$\langle v\,,\,w\rangle \;:=\; v^*w \;=\; \begin{bmatrix} v_1^* & \cdots & v_n^* \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix} \;=\; \sum_{k=1}^{n} v_k^* w_k. \tag{2.13}$$

Note that with this notational choice, the expressions are the same whether we are working with $\mathbb{R}^n$ or $\mathbb{C}^n$. The dot product is a special case of what is more generally referred to as an "inner product". The following definition turns out to capture all the important properties of the standard dot product that can be generalized to more abstract spaces.

Definition 2.4. *An* inner product *on a vector space* V *is a symmetric, positive definite, function* $\langle .\,,\,.\rangle : \mathsf{V} \times \mathsf{V} \to \mathbb{R}$ *(or* $\langle .\,,\,.\rangle : \mathsf{V} \times \mathsf{V} \to \mathbb{C}$ *if the underlying field of scalars is* $\mathbb{C}$*) which is bilinear, i.e.*

- Symmetry: *For any two vectors* $v, w \in \mathsf{V}$

$$\langle v, w\rangle \;=\; \langle w\,,\,v\rangle^*.$$

- Positive Definiteness: *For all vectors* v, $\langle v\,,\,0\rangle = 0$, *and for all non-zero vectors*

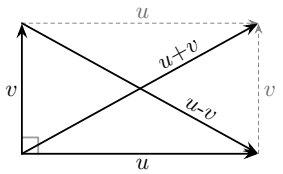
$$\langle v\,,\,v\rangle \in \mathbb{R}, \qquad \langle v\,,\,v\rangle > 0.$$

- Bilinearity: *For any vectors* $v, w, v_1, v_2, w_1, w_2 \in \mathsf{V}$ *and any scalars* $\alpha \in \mathbb{R}$ *(or* $\mathbb{C}$*)*

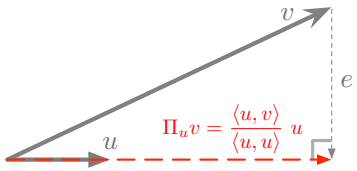
$$\begin{aligned} \langle v\,,\,w_1 + w_2\rangle &= \langle v\,,\,w_1\rangle + \langle v\,,\,w_2\rangle & \qquad \langle v\,,\,\alpha w\rangle &= \alpha\,\langle v\,,\,w\rangle \\ \langle v_1 + v_2\,,\,w\rangle &= \langle v_1\,,\,w\rangle + \langle v_2\,,\,w\rangle & \qquad \langle \alpha v\,,\,w\rangle &= \alpha^*\,\langle v\,,\,w\rangle \end{aligned} \tag{2.14}$$

The dot product on $\mathbb{R}^n$ clearly satisfies all these properties of an inner product, which is why it is traditionally referred to as the "Euclidean inner product" on $\mathbb{R}^n$. There are other possible inner products on $\mathbb{R}^n$ as we will see later in this section.

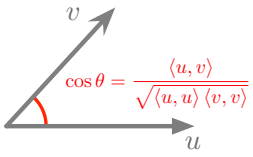
If the vector space is over the complex scalars, then the inner product can have complex values. Some references refer to the symmetry property $\langle u\,,\,v\rangle = \langle v\,,\,u\rangle^*$ as *conjugate symmetry*, and call an operation that satisfies $\langle u\,,\,\alpha v\rangle = \alpha\,\langle u\,,\,v\rangle$ and $\langle \alpha u\,,\,v\rangle = \alpha^*\,\langle u\,,\,v\rangle$



(a) The bilinearity property (2.14) of the inner product gives a version of Pythagoras' theorem. Orthogonality $\langle u\,,\,v\rangle = 0$ directly implies $\langle u\,,\,u\rangle + \langle v\,,\,v\rangle = \langle u+v\,,\,u+v\rangle = \langle u-v\,,\,u-v\rangle$.



(b) The *orthogonal projection* $\Pi_u v$ of one vector v onto another u is defined by the property that the difference $e := \Pi_u v - v$ is orthogonal to u, i.e. $\langle e\,,\,u\rangle = 0$. This condition gives the explicit formula above (in red) for the projection.



(c) The angle between two vectors in an abstract inner product space can be defined as above purely in terms of inner products. For the standard dot product, this formula gives the usual angles in Euclidean space.

Figure 2.6: The inner product induces a notion of angles. (a) Orthogonality implies a version of Pythagoras' theorem. (b) Orthogonal projections of one vector onto another can be defined, and an explicit formula obtained. The *Cauchy-Schwarz* inequality is obtained by applying Pythagoras to the right-angled triangle formed by the orthogonal projection. (c) The *normalized inner product* is always between -1 and 1 by the Cauchy-Schwarz inequality, and can therefore be used to define the angle between two vectors in an abstract inner product space.

sesquilinear. We will avoid this terminology and simply use the terms symmetry and bilinear with the understanding that they are defined as above for the complex case.

Good geometric intuition can be built up about inner products using notions of projections and angles. This generalizes the projection properties of the standard Euclidean inner product. Figure 2.6 illustrates the most basic two concepts. Two vectors u and v are said to be *orthogonal* if $\langle u\,,\,v\rangle = 0$. A version of Pythagoras' theorem holds for such orthogonal vectors, which we state using inner products

$$\begin{aligned}\langle u+v\,,\,u+v\rangle &= \langle u\,,\,u\rangle + \langle v\,,\,v\rangle + 2\cancelto{0}{\langle u\,,\,v\rangle} &&= \langle u\,,\,u\rangle + \langle v\,,\,v\rangle\\ \langle u-v\,,\,u-v\rangle &= \langle u\,,\,u\rangle + \langle v\,,\,v\rangle - 2\langle u\,,\,v\rangle &&= \langle u\,,\,u\rangle + \langle v\,,\,v\rangle\end{aligned} \tag{2.15}$$

and note that the only property used is the bilinearity of the inner product. Figure 2.6a illustrates this relation.

The projection of a vector v onto another vector u is a vector co-linear with u, i.e. the vector αu for some scalar α. The *orthogonal projection* $\Pi_u v$ of v onto u is defined by the additional property that the difference $e := \Pi_u v - v$ must be orthogonal to u (see Figure 2.6b). This requirement gives an explicit expression for α as follows

$$\langle e\,,\,u\rangle = 0 \quad\Leftrightarrow\quad \langle \alpha u - v\,,\,u\rangle = 0 \quad\Leftrightarrow\quad \langle \alpha u - v\,,\,u\rangle = 0 \quad\Leftrightarrow\quad \alpha\,\langle u\,,\,u\rangle - \langle v\,,\,u\rangle = 0.$$

Thus the orthogonal projection of v onto u is explicitly given using inner products as

$$\boxed{\Pi_u v \;=\; \frac{\langle u\,,\,v\rangle}{\langle u\,,\,u\rangle}\,u.} \tag{2.16}$$

Using the projection formula (2.16), we can derive another important property of inner products that characterizes relations between non-orthogonal vectors, and in particular a notion of the *angle* between them. Consider Figure 2.6b. The vectors v, $\Pi_u v$ and $e = \Pi_u v - v$ form a right angled triangle (in the sense that $\Pi_u v$ and $e = \Pi_u v - v$ are orthogonal, and $v = \Pi_u v - e$), and we can therefore apply Pythagoras (2.15)

$$\begin{aligned}\langle v\,,\,v\rangle &= \langle \Pi_u v\,,\,\Pi_u v\rangle + \langle e\,,\,e\rangle\\ \Rightarrow\quad \langle v\,,\,v\rangle &= \langle \alpha u\,,\,\alpha u\rangle + \langle e\,,\,e\rangle \;=\; \alpha^2\,\langle u\,,\,u\rangle + \langle e\,,\,e\rangle\\ \Rightarrow\quad \langle v\,,\,v\rangle - \frac{\langle u\,,\,v\rangle^2}{\langle u\,,\,u\rangle^2}\,\langle u\,,\,u\rangle &= \langle e\,,\,e\rangle \;\geq\; 0,\end{aligned}$$

where the inequality is strict (i.e. $\langle e\,,\,e\rangle > 0$) unless u and v are co-linear (for which then $e = 0$). This last inequality can be rewritten as follows

$$\frac{\langle u\,,\,v\rangle^2}{\langle u\,,\,u\rangle\,\langle v\,,\,v\rangle} \;\leq\; 1 \quad\Leftrightarrow\quad -1 \;\leq\; \frac{\langle u\,,\,v\rangle}{\sqrt{\langle u\,,\,u\rangle}\sqrt{\langle v\,,\,v\rangle}} \;\leq\; 1 \tag{2.17}$$

$$\Leftrightarrow\quad \boxed{|\langle u\,,\,v\rangle| \;\leq \sqrt{\langle u\,,\,u\rangle}\sqrt{\langle v\,,\,v\rangle}.} \tag{2.18}$$

The last expression is the *Cauchy-Schwartz* inequality in inner-product form.

Note that an inner product can be either positive or negative. We can think of the fraction in (2.17) as a sort of "normalized" inner product between u and v. The Cauchy-Schwarz inequality says that *the normalized inner product between any two vectors is always between* -1 *and* 1. This inequality motivates a definition of *angles between vectors* in an abstract inner product space. The normalized inner product between two vectors u and v is 1 if they're co-linear and have the same sense (i.e. the angle between them is $0°$), it is -1 if they're co-linear and have opposite sense (i.e. the angle between them is $180°$), and is zero if they're orthogonal. This quantity therefore seems to behave like a cosine. We thus adopt a definition of *the angle* θ *between any two vectors* u *and* v such that its cosine is the normalized inner product

$$\cos(\theta) \;:=\; \frac{\langle u\,,\,v\rangle}{\sqrt{\langle u\,,\,u\rangle}\sqrt{\langle v\,,\,v\rangle}}. \tag{2.19}$$

This corresponds to the standard angles between vectors in the Euclidean space $\mathbb{R}^2$ (with the inner product as the standard dot product). It also gives the standard angle in the Euclidean space $\mathbb{R}^n$ (since any two vectors are contained in a 2-dimensional subspace, which has the same geometry as $\mathbb{R}^2$). The formula however allows us to define angles in abstract inner product spaces. In particular, in $\mathbb{R}^n$ with an inner product that is different from the standard dot product, the formula (2.19) will give angles that are different from standard Euclidean geometry. In function space, this formula gives the angle and describes orthogonality between two functions.

2.3.1 The Norm Induced by an Inner Product

Recall that in Euclidean geometry, the length of a vector is the square root of the inner product of the vector with itself $\|v\|_2 = \sqrt{v^* v}$. We can try to generalize this statement by using any inner product to define a norm as follows

$$\|v\|^2 \;:=\; \langle v\,,\,v\rangle\,. \tag{2.20}$$

We must check that this definition satisfies the three properties of a norm in Definition 2.2. Definiteness is immediate to see, and homogeneity follows from the bilinearity of the inner product. Before checking the triangle inequality, observe that with the definition (2.20), the Cauchy-Schwarz inequality (2.18) can be rewritten as

$$|\langle u\,,\,v\rangle| \;\leq\; \|u\|\,\|v\|. \tag{2.21}$$

We now check the triangle inequality by first calculating

$$\|u+v\|^2 \;=\; \langle u+v\,,\,u+v\rangle = \;\langle u\,,\,u\rangle + \langle v\,,\,v\rangle + 2\,\langle u\,,\,v\rangle \tag{2.22}$$

$$\leq\; \|u\|^2 + \|v\|^2 + 2\|u\|\|v\| \;=\; (\|u\| + \|v\|)^2\,, \tag{2.23}$$

where the inequality above follows from the Cauchy-Schwarz inequality (2.21). Taking square roots of both sides in (2.23) shows that the triangle inequality is indeed satisfied. To summarize, we have just proved the following statement.

Theorem 2.5. *An inner product space is also a normed vector space with the norm*

$$\|v\|^2 \;:=\; \langle v\,,\, v\rangle\,. \tag{2.24}$$

This norm and the inner product are further related by the Cauchy-Schwarz inequality

$$\boxed{|\langle u\,,\, v\rangle| \;\leq\; \|u\|\;\|v\|.} \tag{2.25}$$

The canonical examples of an inner product space are $\mathbb{R}^n$ and $\mathbb{C}^n$ with the Euclidean inner product, and the function spaces ℓ^2 and L^2 (over any domain). The only difference between $\mathbb{R}^n$ and $\mathbb{C}^n$ on the one hand and ℓ^2 and L^2 on the other is that the finite sums in the inner product become infinite sums and integrals respectively. For example in $\ell^2(\mathbb{Z})$ and $L^2(\mathbb{R})$ the inner products are

$$\langle u\,,\, v\rangle \;:=\; \sum_{i\in\mathbb{Z}} u_i^*\; v_i, \qquad\qquad \langle u\,,\, v\rangle \;:=\; \int_{\mathbb{R}} u^*(x)\; v(x)\; dx \tag{2.26}$$

respectively[3]. It is a quick exercise to check that these definition satisfy all the requirements of an inner product laid out in Definition 2.4. The Cauchy-Schwartz inequality (2.21) for these inner product becomes

$$\left|\sum_{i\in\mathbb{Z}} u_i^*\; v_i\right| \;\leq\; \left(\sum_{i\in\mathbb{Z}} |u_i|^2\right)^{1/2} \left(\sum_{i\in\mathbb{Z}} |v_i|^2\right)^{1/2} \tag{2.27}$$

The definitions (2.26) can be generalized to $\ell^2_{\mathsf{V}}(\Omega)$ and $L^2_{\mathsf{V}}(\Omega)$ for spaces of functions over any domain Ω (in either $\mathbb{Z}^d$ or $\mathbb{R}^d$ respectively), and that take values in any inner-product space V

$$\big\langle u\,,\, v\big\rangle_{\ell^2_{\mathsf{V}}(\Omega)} \;:=\; \sum_{i\in\Omega} \big\langle u_i\,,\, v_i\big\rangle_{\mathsf{V}}\,, \qquad\qquad \big\langle u\,,\, v\big\rangle_{L^2_{\mathsf{V}}(\Omega)} \;:=\; \int_{\Omega} \big\langle u(x)\,,\, v(x)\big\rangle_{\mathsf{V}}\; dx. \tag{2.28}$$

The reader should note how that various inner products were carefully labeled in this case. For example, $\langle u(x)\,,\, v(x)\rangle_{\mathsf{V}}$ is the inner product in V since the function $u:\Omega\to\mathsf{V}$ takes its values in V for each $x\in\Omega$.

2.3.2 Other Inner Products in $\mathbb{R}^n$

There are many inner products on $\mathbb{R}^n$ other than the standard "dot product" (2.13). They are described here briefly for the sake of comparison with other norms in $\mathbb{R}^n$ that have been previously mentioned. Chapter **??** will provide a more thorough study of all the inner products on $\mathbb{R}^n$ since they are intimately related to positive definite matrices. The following material assumes some familiarity with positive definite matrices.

Let Q be any symmetric, positive definite matrix with ij'th entry denoted by q_{ij}. The *bilinear form* on $\mathbb{R}^n$

$$\langle u\,,\, v\rangle_Q \;:=\; u^*Qv \;=\; \sum_{1\leq i,j\leq n} q_{ij}\; u_i\; v_j \tag{2.29}$$

defines an inner product. It is easy to check that the properties of Definition 2.4 hold for this product. In particular, the definiteness property of $v\neq 0 \Rightarrow v^*Qv>0$ is the defining property of a positive definite matrix. The standard dot product corresponds to $Q=I$.

[3] If the functions are scalar valued, then the notation u_i^* stands for the complex conjugate of u_i, while if u is vector valued, then u_i^* is the complex conjugate transpose of u_i (and similarly for $u(x)$). This provides for a single notational convention that covers both scalar-valued and vector-valued functions.

If the same vector is used for each of the two vectors in the bilinear form (2.29), then the quantity $\langle x\,,\,x\rangle_Q$ is a called a *quadratic form*. It is instructive to "unpack" what this quantity is in terms of the vector components $x \in \mathbb{R}^n$

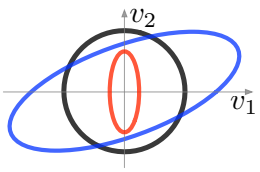
$$x^*Qx = \sum_{1\le i,j\le n} q_{ij}\; x_i x_j \;=\; \sum_{1\le i\le n} q_{ii}\; x_i^2 \;+\; 2\sum_{1\le i<j\le n} q_{ij}\; x_i x_j. \tag{2.30}$$

The first term corresponds to the diagonal entries of Q, while the second term represents the off-diagonal entries. Note that since Q is symmetric, then for off-diagonal terms $q_{ij} = q_{ji}$, and therefore we combined each pair of "cross terms" as $q_{ij}x_ix_j + q_{ji}x_jx_i = 2q_{ij}x_ix_j$. This is the most general form of a quadratic function of n variables, thus the term "quadratic form".

Some geometrical insight can be gained by studying the unit balls of the norm induced by this inner product

$$\mathsf{B}_Q \;:=\; \{v \in \mathbb{R}^n;\; v^*Qv \le 1\}\,.$$

The figure on the right shows examples of unit ball boundaries for three different choices of positive definite matrices Q, one of which is the identity. That latter unit ball B_I is just the unit disk which corresponds to the standard dot product. Note that the other balls look like ellipses, and in fact they are. It will be shown in Chapter **??** that all unit balls of inner products of the form (2.29) are indeed ellipsoids. Furthermore, their principal axes are determined by the eigenvectors and eigenvalues of the matrix Q. This fact distinguishes the geometry of inner product spaces from that of general normed spaces. The latter have unit balls that can be any arbitrary symmetric, convex sets (with the additional properties listed in Theorem 2.9). See e.g. Figure 2.10a. Inner product unit balls however can only be ellipsoids, and thus have a much more restricted geometry, which is parsimonously encoded in the properties of the matrix Q. This fact generalizes to infinite-dimensional inner product spaces where inner products are given by positive definite *operators* rather than matrices.



As just stated, the geometry of normed spaces is more general and richer than that of inner product spaces. However, the restrictive geometry induced by inner products enables much sharper results and algorithms when it comes to optimization and other applications. This is the typical tradeoff between generality versus structure that occurs throughout mathematics.

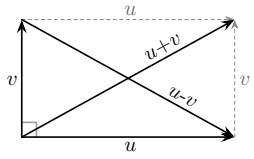
2.3.3 The Parallelogram Law and the Polarization Identity

We have seen how an inner product gives a norm by (2.24). A natural question to ask is whether a given norm comes from an inner product in this manner. If one is given only the norm, for example, the $\|.\|_p$ norms in $\mathbb{R}^n$, we ask whether there is an underlying inner product that gives rise to this norm?

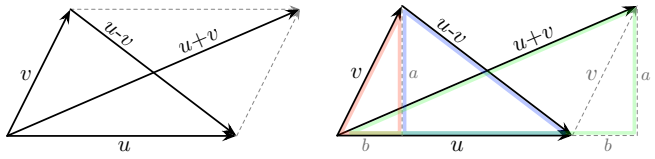
To answer the above question, we can look at properties that a norm induced by an inner product must satisfy. Pythagoras is one such property which states that for two orthogonal vectors u and v

$$\|u+v\|^2 \;=\; \|u\|^2 + \|v\|^2.$$

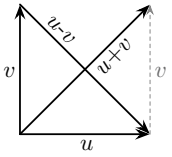
However, if we're only given the norm, we cannot check the orthogonality $\langle u\,,\,v\rangle = 0$ since that requires knowing the inner product! An equivalent statement to Pythagoras that



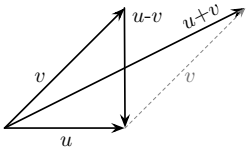
(a) Pythagoras' theorem states that for two *orthogonal* vectors u and v, we have $\|u+v\|^2 = \|u-v\|^2 = \|u\|^2 + \|v\|^2$. Statement of this theorem in an abstract vector space requires a norm, and a notion of *orthogonality*, i.e. an inner product.



(b) The parallelogram law states that for any two (not necessarily orthogonal) vectors, u and v, we have $\|u+v\|^2 + \|u-v\|^2 = 2\left(\|u\|^2 + \|v\|^2\right)$. It is valid in any inner product space, but its *statement* does not require knowing the inner product. Its proof follows from applying Pythagoras' theorem to the red, blue, and green right-angled triangles shown above, and then eliminating the lengths a and b from the resulting 3 equations.



(c) Failure of the parallelogram law for $\|.\|_\infty$ in $\mathbb{R}^2$. $\|u\|_\infty = \|(1,0)\|_\infty = 1$ and $\|v\|_\infty = \|(0,1)\|_\infty = 1$, and $\|u+v\|_\infty = \|(1,1)]\|_\infty = 1$, $\|u-v\|_\infty = \|(1,-1)\|_\infty = 1$. Thus $\|u+v\|_\infty + \|u-v\|_\infty = 2$, while $2\left(\|u\|_\infty + \|v\|_\infty\right) = 4$.



(d) Failure of the parallelogram law for $\|.\|_1$ in $\mathbb{R}^2$. $\|u\|_1 = \|(1,0)\|_1 = 1$ and $\|v\|_1 = \|(1,1)\|_1 = 2$, and $\|u+v\|_1 = \|(2,1)]\|_1 = 3$, $\|u-v\|_1 = \|(0,-1)\|_1 = 1$. Thus $\|u+v\|_1 + \|u-v\|_1 = 4$, while $2\left(\|u\|_1 + \|v\|_1\right) = 2(1+2) = 6$.

Figure 2.7: Pythagoras and the parallelogram law are equivalent statements valid in any inner product space. However, we need a notion of angles (or orthogonality) to state Pythagoras' theorem, while the parallelogram law requires only a notion of vector norms. The latter can thus be used to check whether a given norm arises out of an inner product. The bottom row figures give examples of how the parallelogram law is invalid for the norms $\|.\|_\infty$ and $\|.\|_1$ in $\mathbb{R}^n$, thus implying that these norms do not arise out of an inner product.

does not require knowing the inner product is the *parallelogram law* depicted in Figure 2.7. Applying Pythagoras' theorem to the three colored, right-angled triangles in Figure 2.7b

$$\begin{aligned} \|v\|^2 &= \|a\|^2 + \|b\|^2 \\ \|u-v\|^2 &= \|a\|^2 + (\|u\| - \|b\|)^2 \;=\; \|a\|^2 + \|u\|^2 + \|b\|^2 - 2\|u\|\|b\| \\ \|u+v\|^2 &= \|a\|^2 + (\|u\| + \|b\|)^2 \;=\; \|a\|^2 + \|u\|^2 + \|b\|^2 + 2\|u\|\|b\|. \end{aligned}$$

Adding the last two equations, and substituting for $\|a\|^2 + \|b\|^2$ from the first equation eliminates $\|a\|$ and $\|b\|$ to give the *parallelogram law*

$$\|u-v\|^2 + \|u+v\|^2 \;=\; 2\left(\|u\|^2 + \|v\|^2\right). \tag{2.31}$$

Several remarks are now in order. First, the parallelogram law is equivalent to Pythagoras. It was derived from Pythagoras, and conversely if u and v are orthogonal, then $\|u-v\| = \|u+v\|$, and (2.31) reduces to the statement $\|u+v\|^2 = \|u\|^2 + \|v\|^2$. The second observation is that checking whether (2.31) holds does not require knowing the inner product. Finally, recall the equation (2.22) and observe that we can use it to recover the inner product from the norm

$$\begin{aligned} \langle u\,,\, v\rangle &= \frac{1}{2}\left(\|u+v\|^2 - \left(\|u\|^2 + \|v\|^2\right)\right) \\ &= \frac{1}{2}\left(\left(\|u\|^2 + \|v\|^2\right) - \|u-v\|^2\right) \\ &= \frac{1}{4}\left(\|u+v\|^2 - \|u-v\|^2\right) \end{aligned} \tag{2.32}$$

The first equation is just (2.22), and the last two are different forms that follow from (2.31). We have thus shown that the norm derived from an inner product satisfies the parallelogram law. The converse is also true (Exercise 2.2), and we summarize the statement as follows.

Theorem 2.6. *A normed space where the norm satisfies the parallelogram law*

$$\|u+v\|^2 + \|u-v\|^2 = 2\left(\|u\|^2 + \|v\|^2\right) \tag{2.33}$$

is an inner product space where the inner product is given by the polarization identities (2.32).

Appendix

2.A Convexity

The reader is probably familiar with the notion of a convex scalar functional. This notion can be generalized to functionals of several variables. The only concept needed is that of taking convex combinations in the variables, and this is possible in any vector space.

Let $v_1, v_2 \in \mathsf{V}$ be vectors in any vector space V. A *convex combination* of v_1 and v_2 is the vector

$$v(\alpha) \;=\; \alpha\, v_1 \;+\; (1-\alpha)\, v_2, \qquad \alpha \in [0,1]. \tag{2.34}$$

This has a simple geometrical interpretation. The vector v lies on the straight line segment connecting v_1 and v_2 (see Figure 2.8a). In fact, the formula (2.34) is a *parametrization* of that line segment. As α changes from 0 to 1, the point v moves along the line segment from v_2 to v_1. To see that (2.34) parametrizes a straight line segment, take the derivative of $v(\alpha)$ with respect to the parameter α, and compute $dv/d\alpha = v_1 - v_2$. Thus the derivative is independent of α (i.e. constant velocity), and in the direction of the vector $v_1 - v_2$ which connects the two points v_1 and v_2.

It is worth noting that the convex combination (2.34) can be written in three different, but equivalent ways

$$\begin{aligned} v \;&=\; \alpha\, v_1 \;+\; (1-\alpha)\, v_2, && \alpha \in [0,1], \\ &=\; \gamma\, v_1 \;+\; \beta\, v_2, && \gamma+\beta = 1,\ \gamma,\beta \geq 0, \\ &=\; \tfrac{\alpha_1}{\alpha_1+\alpha_2}\, v_1 \;+\; \tfrac{\alpha_2}{\alpha_1+\alpha_2}\, v_2, && \alpha_1, \alpha_2 \geq 0. \end{aligned} \tag{2.35}$$

The reader should verify those equivalences as an exercise.

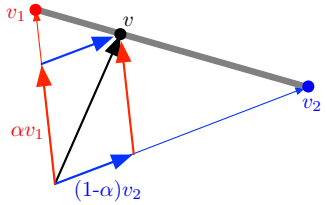
The notion of convex combination of points leads naturally to the notion of convex sets.

Definition 2.7. *A subset $\Omega \subset \mathsf{V}$ of a vector space is called* convex *if given any two points $v_1, v_2 \in \Omega$, all possible convex combinations of v_1 and v_2 (i.e. points on the entire straight line segment joining v_1 and v_2) belong to Ω.*

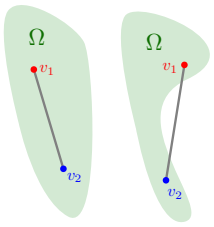
This concept is depicted in Figure 2.8b. The straight line segments joining any two points in Ω must lie entirely inside Ω. The figure also depicts an example of a non-convex set for comparison. For a set $\Omega \in \mathbb{R}^2$ with smooth boundaries, there is a physical interpretation. If one imagines a particle moving along the boundary of the set, then the acceleration vector of the particle is always pointing inwards into the set.

Clearly the entire vector space is a convex set. Other examples include the "unit balls" of normed spaces, e.g. the sets $\left\{v \in \mathbb{R}^2;\ \|v\|_p \leq 1\right\}$ for any $p \in [1,\infty]$ depicted in Figure 2.4. For $p < 1$, those sets are clearly not convex since e.g. the line segment connecting $(1,0)$ and $(0,1)$ lies outside the set (except for the end points of course).

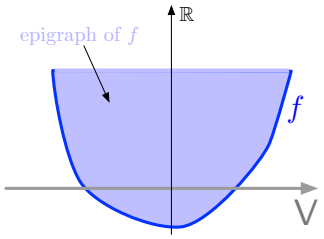
The next concept is that of a convex *functional*.



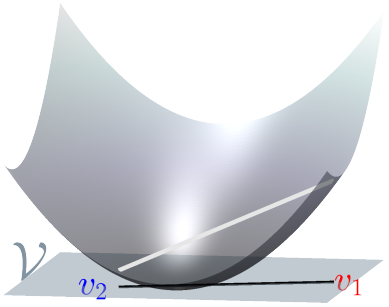
(a) Any *convex combination* $v = \alpha v_1 + (1\text{-}\alpha)v_2$ of two points v_1 and v_2 in a vector space lies on the straight line segment connecting those two points.



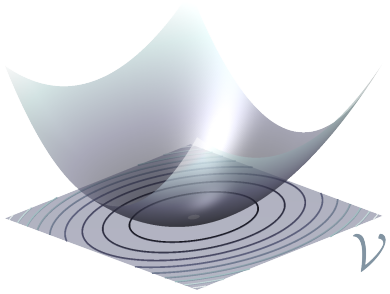
(b) (*Left*) A set Ω is *convex* if for every two points v_1 and v_2 in Ω, the straight line segment joining them is contained entirely inside Ω. (*Right*) An example of a non-convex set.



(c) A function f is convex iff its epigraph is a convex set. The epigraph is the set of points in $\mathsf{V} \times \mathbb{R}$ that are "above" the graph of f.



(d) A functional f on a vector space V is convex if given any convex combination of elements v_1, v_2 (a point on the black line segment in V), the value of the function f at that point is *smaller* than the convex combination of the *values* of $f(v_1)$ and $f(v_2)$ (those values depicted as the white line segment).



(e) The contours in V are the level sets of a function f. The contours are the boundaries of the *sub-level* sets of f. If f is convex, then its sub-level sets are convex subsets of V.

Figure 2.8: Illustrations of the concepts of (a) convex combination of two points, (b) convex sets, (c) convex epigraph, (d) convex functional, and (e) convex sub-level sets.

Definition 2.8. *A functional* $f : \mathsf{V} \longrightarrow \mathbb{R}$ *on a vector space* V *is called* convex *if for any* $v_1,\ v_2\ \in V$ *and* $\alpha \in [0, 1]$

$$f\big(\alpha v_1 + (1-\alpha)v_2\big)\ \leq\ \alpha f(v_1)\ +\ (1-\alpha)\ f(v_2). \tag{2.36}$$

In other words, if the value of f *at any convex combination of* v_1 *and* v_2 *is no larger than the same convex combination of the values* $f(v_1)$ *and* $f(v_2)$*. A function is called* strictly convex *if (2.36) holds with strict inequality for* $\alpha \in (0, 1)$*. It is called* concave *(*strictly concave*) if* $-f$ *is convex (strictly convex).*

This concept is depicted in Figure 2.8e. Convex functions can be thought of as "bowl shaped". For single-variable functionals $f : \mathbb{R} \to \mathbb{R}$, the reader may recall that convexity is equivalent to the second derivative being non-negative everywhere (i.e. the function never curves downwards). Similarly, if f is a twice-differentiable functional on $\mathbb{R}^n$, its convexity (strict convexity) is equivalent to its Hessian (which is the multivariate version of the second derivative, see Chapter **??**) being positive semi-definite (positive definite) everywhere.

Another characterization of convex functionals involves the concept of the epigraph of a function, which is defined as the subset of $\mathsf{V} \times \mathbb{R}$ such that

$$\text{epigraph}(f)\ :=\ \Big\{(v, r) \in \mathsf{V} \times \mathbb{R};\ r\ \geq\ f(v)\Big\}.$$

The epigraph is depicted in Figure 2.8c, it is the set of all points in $\mathsf{V} \times \mathbb{R}$ that include

the graph of f as well as all points "above it". It is a simple exercise to show that the criterion (2.36) implies that a function f is convex iff its epigraph is a convex set.

Let $f : \mathsf{V} \to \mathbb{R}$ be a convex functional, and consider its sub-level sets

$$\mathsf{S}_\gamma \;:=\; \Big\{ v \in \mathsf{V};\; f(v) \leq \gamma \Big\},$$

where the "level" $\gamma \in \mathbb{R}$ is any real number. An important property of convex functionals is that *their sub-level sets are convex sets*

$$\begin{aligned} f(v_1) \leq \gamma,\; f(v_2) \leq \gamma \qquad \Rightarrow \qquad f\big(\alpha v_1 + (1-\alpha) v_2\big) \;&\leq\; \alpha\, f(v_1) \;+\; (1-\alpha)\, f(v_2) \\ &\leq\; \alpha\, \gamma \;+\; (1-\alpha)\, \gamma \;=\; \gamma. \end{aligned}$$

Thus any convex combination of two elements v_1 and v_2 in S_γ is also in that sub-level set. This is depicted in Figure 2.8e. A particularly important convex functional is the norm functional $\|.\| : \mathsf{V} \to \mathbb{R}$ on a normed vector space V. Its convexity is a consequence of the triangle inequality property of the norm. An important sub-level set of the norm functional is the unit ball

$$\mathsf{B} \;:=\; \{ v \in \mathsf{V};\; \|v\| \leq 1 \}.$$

The preceding argument implies that the unit ball of any norm must be convex. This is a useful test to check whether a given set could be the unit ball of *some norm*. If that set is not convex (note that we don't need to know the norm to determine whether a set is convex or not, the definition only involves convex combinations and set membership), then it couldn't possibly be the unit ball of a norm. This criterion shows that the sets $\{\|v\|_p \leq 1\}$ for $p < 1$ in Figure 2.4 are not unit balls of a norm, i.e. the quantity $\|.\|_p$ does not define a norm on $\mathbb{R}^n$ if $p < 1$.

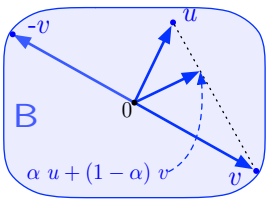
It is important to point out that there are functions whose sub-level sets are all convex, yet the function itself is not convex. Such functions are called *quasi-convex*, and they play an important role in optimization. However, this concept is not needed in the present chapter.

2.B Norms Induced by Convex Sets

As already mentioned, the unit ball of any norm is a convex set. This is a consequence of the triangle inequality. If $\|u\| \leq 1$ and $\|v\| \leq 1$, then

$$\|\alpha u + (1-\alpha) v\| \;\leq\; \|\alpha u\| + \|(1-\alpha) v\| \;=\; \alpha \|u\| + (1-\alpha) \|v\| \;\leq\; 1,$$

i.e. any convex combination of two vectors in the unit ball is also in the unit ball. This is illustrated in the figure on the right. Furthermore, one consequence of homogeneity is that if $\|v\| \leq 1$, then $\|-v\| \leq 1$, i.e. the unit ball is *symmetric* with respect to reflections about the origin.



Now consider the reverse question: *which types of convex sets are unit balls of norms?* In addition, *given such a convex set, how can we define the norm for which it is the unit ball?*

To answer the above questions, we investigate another important consequence of homogeneity, which describes how norms scale with vector scalings. First, observe that for any vector v, *normalizing* it by its length yields the vector $v/\|v\|$, which is a vector of unit length in the same direction as v. Thus scaling by the factor $1/\|v\|$ is such that the resulting vector just "touches" the edge of the

unit ball. This is illustrated in Figure 2.9a. We can therefore characterize the norm of a vector based on how much it has to be "scaled" before it is in or out of the unit ball, i.e.

$$\|v\| = \begin{cases} \inf\{\gamma \geq 0;\ \|v\| \leq \gamma\} = \inf\{\gamma \geq 0;\ \|v/\gamma\| \leq 1\} = \inf\left\{\gamma \geq 0;\ \frac{1}{\gamma}v \in \mathsf{B}\right\}, \\ \sup\{\gamma \geq 0;\ \gamma \leq \|v\|\} = \sup\{\gamma \geq 0;\ \|v/\gamma\| \geq 1\} = \sup\left\{\gamma \geq 0;\ \frac{1}{\gamma}v \notin \mathsf{B}\right\}. \end{cases} \tag{2.37}$$

The reader should now parse through Figure 2.9a for geometrical illustrations of these formulas as well as the following additional characterizations of the norm

$$\frac{1}{\|v\|} = \begin{cases} \sup\{\beta \geq 0;\ \beta v \in \mathsf{B}\}, \\ \inf\{\beta \geq 0;\ \beta v \notin \mathsf{B}\}. \end{cases} \tag{2.38}$$

We now observe that the quantities on the right in (2.37) and (2.38) are not written in terms of the norm $\|.\|$, but rather involve scalings and set membership in B. Thus given any convex set Ω in a vector space, we can try to define a norm such that the given set is its unit ball using those set membership conditions. We will also require additional properties on the convex set that guarantee that the quantities in (2.37) and (2.38) are finite and non-zero for non-zero vectors.

Theorem 2.9. *Let $\Omega \subset \mathsf{V}$ be a convex set containing the origin in a vector space V. If the convex set Ω is*

1. symmetric*: $v \in \Omega \Rightarrow -v \in \Omega$,*
2. absorbing*: any non-zero vector $v \in \mathsf{V}$ can be scaled so that it "enters" Ω, i.e.*

$$\inf\left\{\beta \geq 0;\ \beta v \notin \Omega\right\} > 0. \tag{2.39}$$

3. bounded*: any non-zero vector in Ω can be scaled so that it "exits" Ω*

$$\sup\left\{\beta \geq 0;\ \beta v \in \Omega\right\} < \infty, \tag{2.40}$$

then V becomes a normed vector space with the norm[4]

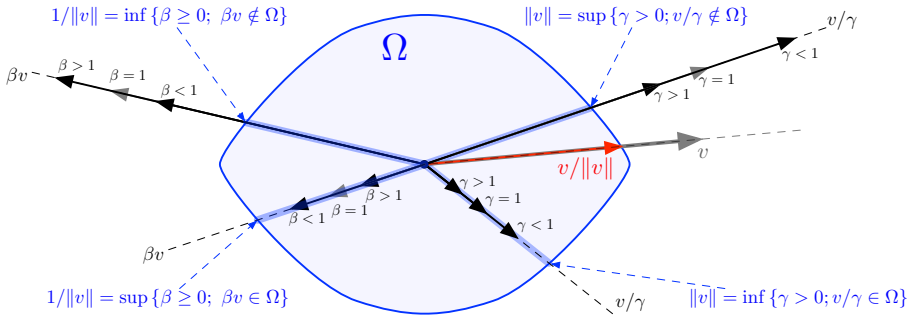
$$\begin{aligned} \|v\| &:= \inf\left\{\gamma \geq 0;\ v/\gamma \in \Omega\right\} = \sup\left\{\gamma \geq 0;\ v/\gamma \notin \Omega\right\} \\ &= \frac{1}{\sup\left\{\beta \geq 0;\ \beta v \in \Omega\right\}} = \frac{1}{\inf\left\{\beta \geq 0;\ \beta v \notin \Omega\right\}} \end{aligned} \tag{2.41}$$

Figure 2.9a illustrates the motivation for the norm definitions (2.41). Figures 2.9b and 2.9c explain why the "absorbing" and "boundedness" conditions are needed respectively. They illustrate why if the absorbing condition is not met, there exists vectors with infinite norms, while if the boundedness condition is not met, there exists non-zero vectors of zero norm.

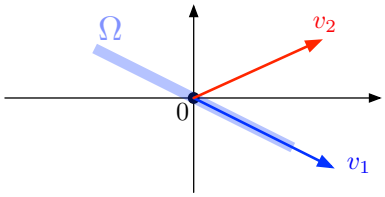
To prove Theorem 2.9, we need to show that the norm defined by (2.41) satisfies the three properties of a norm. Homogeneity is an immediate consequence of the definition

$$\begin{aligned} \|\bar{\gamma}v\| = \inf\left\{\gamma > 0;\ \bar{\gamma}v/\gamma \in \Omega\right\} &= \inf\left\{\bar{\gamma}r > 0;\ v/r \in \Omega\right\} \qquad \text{(substituting } \gamma/\bar{\gamma} = r\text{)} \\ &= \bar{\gamma}\inf\left\{r > 0;\ v/r \in \Omega\right\} = \bar{\gamma}\|v\|. \end{aligned} \tag{2.42}$$

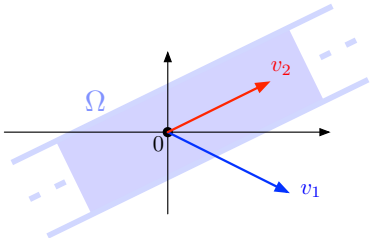
[4] In some literature, such functions defined in terms of a given convex set are called *Minkowski functionals*, though usually without the boundedness assumption, and therefore would only define *seminorms* rather than proper norms.



(a) Given any vector v in a normed space, its *normalization* $v/\|v\|$ (shown in red) is the vector in the same direction as v, and lies exactly on the boundary of the unit ball. This property can be used to define a norm from a convex set using *set membership*. The norm is given from the scalings v/γ or βv such that the scaled vectors lie on the boundary of the set Ω.



(b) A set Ω that is not "absorbing". Shown here as the light blue line segment. This set is one-dimensional in $\mathbb{R}^2$. The vector v_1 can be scaled so that it "enters" Ω, and is therefore of finite norm. However, there is no scaling of v_2 such that it enters Ω, and therefore v_2 has infinite norm as per (2.41) $\|v\| = \sup\{\gamma \geq 0; v/\gamma \notin \Omega\} = \infty$.



(c) A set Ω that is not "bounded". There is no scaling of the vector v_2 such that it "exits" this set, i.e. $\sup\{\beta \geq 0;\ \beta v_2 \in \Omega\} = \infty$. Therefore v_2 has zero norm $\|v_2\| = 1/\sup\{\beta \geq 0;\ \beta v_2 \in \Omega\} = 0$ as per (2.41), even though it is a non-zero vector.

Figure 2.9: A geometric interpretation of Theorem 2.9 where any convex set that is symmetric about origin with the absorbing and boundedness properties induces a norm.

We note that if $\bar{\gamma} < 0$, we substitute $\gamma/|\bar{\gamma}| = r$ and use the symmetry property of Ω.

Definiteness of the norm (2.41) is a consequence of boundedness property (2.40),(2.38) since for vectors inside Ω

$$\|v\| = 0 \qquad \Leftrightarrow \qquad \sup\left\{\beta \geq 0;\ \beta v \in \Omega\right\} = \infty \qquad \Leftrightarrow \qquad v = 0.$$

Definiteness also holds for vectors outside Ω since they can be scaled to be inside Ω. The fact that the norm is finite for any v follows from the property (2.39) since

$$\|v\| = \infty \qquad \Leftrightarrow \qquad \inf\left\{\beta \geq 0;\ \beta v \notin \Omega\right\} = 0.$$

Finally, the triangle inequality follows from convexity. Let u and v by any vectors. Use the definition (2.41) for $\|u\|$ and $\|v\|$, and normalize so that $u/\|u\|$ and $v/\|v\|$ are at the boundary of the set Ω (this follows by definition from (2.41)). Now the convexity of Ω implies that any convex combination of the normalized vectors will be inside of Ω, i.e.

$$\frac{\alpha_1}{\alpha_1+\alpha_2}\frac{u}{\|u\|} + \frac{\alpha_2}{\alpha_1+\alpha_2}\frac{v}{\|v\|} \in \Omega.$$

The particular choice of $\alpha_1 = \|u\|$ and $\alpha_2 = \|v\|$ says that

$$\frac{\|u\|}{\|u\|+\|v\|}\frac{u}{\|u\|} + \frac{\|v\|}{\|u\|+\|v\|}\frac{v}{\|v\|} = \frac{1}{\|u\|+\|v\|}\left(u+v\right) \in \Omega.$$

This last statement is indeed the triangle inequality since

$$\frac{1}{\|u\|+\|v\|}\,(u+v)\ \in \Omega \quad \Leftrightarrow \quad \left\|\frac{1}{\|u\|+\|v\|}\,(u+v)\right\| \le 1$$
$$\Leftrightarrow \quad \frac{1}{\|u\|+\|v\|}\,\|u+v\| \le 1 \quad \Leftrightarrow \quad \|u+v\| \le \|u\|+\|v\|.$$

Note that the only property of $\|.\|$ used in the above argument is the homogeneity property, which has already been established by (2.42).

Theorem 2.9 establishes that there is an infinite variety of norms that can be defined on $\mathbb{R}^n$, each corresponding to any symmetric, bounded convex set. We thus see that the p-norms depicted in Figure 2.4 are only a very special class of norms from amongst all the possible norms on $\mathbb{R}^n$.

2.C Equivalence of Norms in Finite Dimensions

Two norms $\|.\|_a$ and $\|.\|_b$ are said to be *equivalent* if each can be bounded by the other from above and below, i.e. if there exists constants $\underline{c}, \overline{c} > 0$ such that for all vectors v

$$\underline{c}\,\|v\|_b \ \le\ \|v\|_a \ \le\ \overline{c}\,\|v\|_b \qquad \Leftrightarrow \qquad \frac{1}{\overline{c}}\,\|v\|_a \ \le\ \|v\|_b \ \le\ \frac{1}{\underline{c}}\,\|v\|_a \qquad \Leftrightarrow \qquad \|.\|_a \ \sim\ \|.\|_b.$$

Note that the constants $\underline{c}, \overline{c}$ should not depend on the choice of vector v. We define the notation $\|.\|_a \sim \|.\|_b$ to mean that the two norms are equivalent. The equations above imply that this relation is *symmetric*, i.e. $\|.\|_a \sim \|.\|_b \ \Leftrightarrow\ \|.\|_b \sim \|.\|_a$. It is also easy to verify that this relation is *transitive*, meaning that

$$\Big(\|.\|_a \sim \|.\|_b\Big) \text{ and } \Big(\|.\|_b \sim \|.\|_c\Big) \qquad \Rightarrow \qquad \|.\|_a \sim \|.\|_c,$$

and therefore equivalent norms (no pun intended) form equivalence classes.

A little care is needed in interpreting the notion of equivalence above. Two norms do not have to be equal to be equivalent. If one tries to measure distances or optimize with respect to one norm, the answers will generally be quite different when done with another, but equivalent norm. The term "equivalent" above is to be understood with regard to convergence notions. Given two equivalent norms, a sequence is convergent in one norm iff it is convergent in another, equivalent norm.

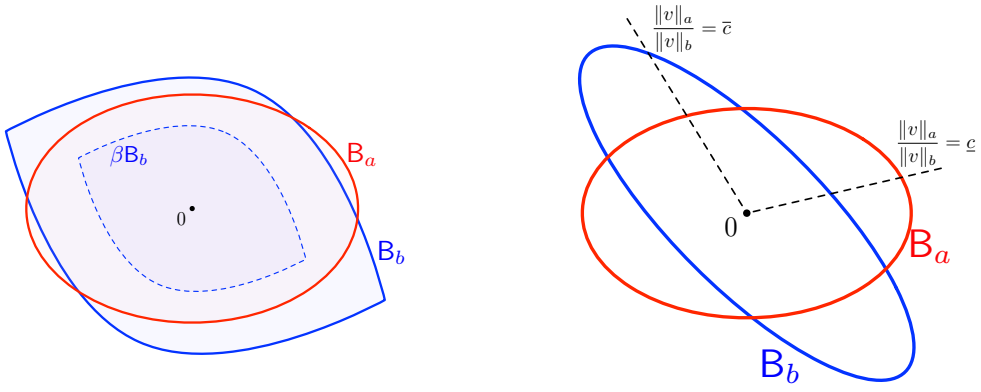
Examples of bounds on norms in $\mathbb{R}^n$ are easy to derive. The following are bounds for the 1, 2 and ∞ norms in $\mathbb{R}^n$, together with examples of when these bounds are tight

$$\begin{aligned} \|v\|_\infty \ &\le\ \|v\|_2 \ \le\ \|v\|_1 && \text{with equality for } v = (0,\ldots,0,1,0,\ldots,0)\,, \\ \|v\|_1 \ &\le\ \sqrt{n}\,\|v\|_2 \ \le\ n\,\|v\|_\infty && \text{with equality for } v = (1,\ldots,1). \end{aligned} \tag{2.43}$$

The inequalities involving $\|.\|_\infty$ are relatively straightforward to verify, and are left to the reader as an exercise. The inequality $\|v\|_2 \le \|v\|_1$ follows from the following calculation

$$\|v\|_2^2 = \sum_{i=1}^n v_i^2 = \sum_{i=1}^n |v_i|\ |v_i| \le \left(\sum_{i=1}^n |v_i|\right)\left(\max_{1\le i\le n} |v_i|\right) = \|v\|_1\|v\|_\infty \le \|v\|_1\|v\|_1 = \|v\|_1^2.$$

The inequality $\|v\|_1 \le \sqrt{n}\|v\|_2$ follows from the Cauchy-Schwartz inequality (Exercise 2.6) which will be introduced later. Note that the first set of inequalities in (2.43) can be visualized as in Figure 2.4 through the containment of their respective unit balls.



(a) If the unit ball B_b can be scaled to $\beta\mathsf{B}_b$ so that it is entirely contained in B_a, then $\|.\|_a \leq \|.\|_b/\beta$. Note that $\beta\mathsf{B}_b$ is the unit ball of the scaled norm $\|.\|_b/\beta$, and that unit ball containment and norm bounds are related by $(\beta\mathsf{B}_b \subseteq \mathsf{B}_a) \Leftrightarrow (\|.\|_a \leq \|.\|_b/\beta)$. In finite dimensions, the unit ball of any norm can be scaled as above so that it is entirely contained in another unit ball. The containment however is not "tight" in the sense that there could be unavoidable large gaps between $\beta\mathsf{B}_b$ and B_a, implying that the corresponding norm bounds are not tight. As the space dimension grows to infinity, these gaps can become arbitrarily large, and thus the norm bounds arbitrarily loose.

(b) Bounds between two norms $\|.\|_a$ and $\|.\|b$ are determined by the maxima and minima (2.46) of the ratio $\|v\|_a/\|v\|_b$ over all directions in space. Equivalently, they are the maxima and minima of $\|v\|_a$ as v ranges over the unit sphere $\|v\|_b = 1$. The maximal ($\bar{c}$) and minimal ($\underline{c}$) ratios and directions are shown above for an example of B_a and B_b. Note that for example when the boundary of B_b is inside B_a, then $\|v\|_a \leq \|v\|_b$, i.e. the ratio $\|v\|_a/\|v\|_b$ is less than 1.

Figure 2.10: Two graphical illustrations of the equivalence of any two norms in finite dimensions. B_a and B_b are the unit balls of two norms $\|.\|_a$ and $\|.\|_b$. The geometrical relationships between the two sets B_a and B_b determine the relative bounds between the norms $\|.\|_a$ and $\|.\|_b$. For example $(\mathsf{B}_b \subseteq \mathsf{B}_a) \Leftrightarrow (\|.\|_a \leq \|.\|_b)$, i.e. unit ball containment implies a norm bound in the reverse order.

The reader should note how the bounds in the inequalities above depend on the space dimension n. The second set of inequalities become progressively "looser" in higher dimensions. In fact, they are not valid in any infinite-dimensional ℓ^p spaces, while the first set of inequalities still hold in those spaces. This follows intuitively by observing that the first set are independent of n, and therefore one expects them to hold unchanged as $n \to \infty$. In addition, since the bounds become looser as $n \to \infty$, for large n, one would expect that an optimization problem for $\|.\|_p$ will yield very different answers from that for $\|.\|_q$ if $p \neq q$. These caveats should be kept in mind when interpreting the notion of "equivalence" of norms in finite dimensions that is stated in the next theorem.

Theorem 2.10. *Let $\|.\|_a$ and $\|.\|_b$ be any two norms on $\mathbb{R}^n$ (or $\mathbb{C}^n$) with n finite. The two norms are equivalent.*

Before we give the proof arguments, two geometric ideas provide helpful intuition. The first is illustrated in Figure 2.10a. If one of the unit balls (say B_b) can be scaled so that it is properly contained in the other, then a norm bound is obtained from the relation (2.10)

$$\beta\mathsf{B}_b \subseteq \mathsf{B}_a \qquad \Leftrightarrow \qquad \|v\|_a \;\leq\; \frac{1}{\beta}\|v\|_b,$$

since $\beta\mathsf{B}_b$ is the unit ball for the scaled norm $\frac{1}{\beta}\|.\|_b$. We can similarly scale B_a so that it is inside B_b and obtain the other bound. We note that such scalings are possible in finite dimensions, but it may not be possible in infinite dimensions to scale one unit ball so that it is contained in the other. Furthermore, if β is chosen so that $\beta\mathsf{B}_b$ "just fits" inside B_a, then we have the best possible bound. However, there will be directions in space where there are

potentially large gaps between the boundaries of the two balls. Along these directions, the bounds are loose and not very useful. For many norms, as the space dimension increases to infinity, these gaps can become arbitrarily large, and therefore the norm bounds along those directions can become arbitrarily loose.

Now to prove the theorem, we first observe that since norm equivalence is a transitive relation, we simply need to prove that any norm $\|.\|_a$ is equivalent to some conveniently chosen norm. In this case, it turns out that the $\|.\|_1$ norm or the $\|.\|_\infty$ norm will do nicely.

The bounds in one direction can be established by simple inequalities. Let $\{e_i\}$ be some basis of $\mathbb{R}^n$, and write a vector $v = \sum_{i=1}^n v_i e_i$ in that basis. Then

$$\|v\|_a \leq \left\| \sum_{i=1}^n v_i e_i \right\|_a \leq \sum_{i=1}^n |v_i| \, \|e_i\|_a \leq \left(\max_{1 \leq i \leq n} \|e_i\|_a \right) \left(\sum_{i=1}^n |v_i| \right) =: \bar{c} \, \|v\|_1, \tag{2.44}$$

where the last inequality is a version of the 1-∞ inequality (Exercise 2.5) . Note that the finite number $\bar{c}$ is a property of the vectors $\{e_i\}$, and is therefore independent of v. We can see how this argument might fail in infinite dimensions; the $\max_{1 \leq i \leq n} \|e_i\|_a$ term would instead be a supremum over an infinite index set, and therefore might itself be infinite.

The converse bound requires a different argument which also provides additional geometrical insight which comes from examining the maxima and minima of the ratio of norms $\|v\|_a / \|v\|_b$ over all vectors v. Given two such norms, consider the extrema of the ratio

$$\underline{c} := \inf_{v \neq 0} \frac{\|v\|_a}{\|v\|_b} \leq \frac{\|v\|_a}{\|v\|_b} \leq \sup_{v \neq 0} \frac{\|v\|_a}{\|v\|_b} =: \bar{c}. \tag{2.45}$$

If $\underline{c} > 0$ and $\bar{c} < \infty$, then we have our equivalence bounds since then

$$\underline{c} \, \|v\|_b \leq \|v\|_a \leq \bar{c} \, \|v\|_b,$$

There is a useful reformulation of the relations (2.45) which comes from observing that the ratio $\|v\|_a / \|v\|_b$ does not depend on the length of a vector v, but only its *direction* (norms are homogenous, therefore $\|\alpha v\|_a / \|\alpha v\|_b = \|v\|_a / \|v\|_b$). We can therefore restate (2.45) in the more useful form

$$\underline{c} := \inf_{\|v\|_b = 1} \|v\|_a = \inf_{v \neq 0} \frac{\|v\|_a}{\|v\|_b} \leq \frac{\|v\|_a}{\|v\|_b} \leq \sup_{v \neq 0} \frac{\|v\|_a}{\|v\|_b} = \sup_{\|v\|_b = 1} \|v\|_a =: \bar{c}. \tag{2.46}$$

This is illustrated in Figure 2.10b. We can think of $\|v\|_a$ as a functional $\|.\|_a : \mathbb{R}^n \to \mathbb{R}$ restricted to the unit sphere $\{\|v\|_b = 1\}$. The bounds $\underline{c}$ and $\bar{c}$ in (2.46) are the minimum and maximum respectively of the function $\|.\|_a$ over the set $\{\|v\|_b = 1\}$. We need to show that these are non-zero and finite respectively.

The *extreme value theorem* states that any continuous function on a compact set achieves its minimum and maximum in that set. For the present argument, $\|.\|_b = \|.\|_1$, i.e. the 1-norm, and the set $\{\|v\|_1 \leq 1\}$ is closed and bounded in $\mathbb{R}^n$, therefore compact. It remains to show that the function $\|.\|_a$ is continuous with respect to the 1-norm. This has effectively been shown by (2.44). Indeed, given any point $\bar{v}$ and $\epsilon > 0$, there exists $\delta > 0$ such that

$$\|v - \bar{v}\|_1 \leq \delta \qquad \Rightarrow \qquad \|v - \bar{v}\|_a \leq \epsilon.$$

In particular, the choice $\delta \leq \epsilon / \bar{c}$ (from (2.44)) provides this bound.

The extreme value theorem thus says that the constants $\underline{c}$ and $\bar{c}$ in

$$\underline{c} := \inf_{\|v\|_1 = 1} \|v\|_a \leq \sup_{\|v\|_1 = 1} \|v\|_a =: \bar{c} \tag{2.47}$$

are finite. To show that $\underline{c} > 0$, note that the function $\|.\|_a$ is strictly positive for all $\{\|v\|_1 = 1\}$ since all members of that set are non-zero vectors. Since $\|.\|_a$ achieves its minimum at some point in that set, that minimum must be strictly positive.

To recap, we have shown that all norms in finite dimensions are equivalent. In infinite dimensions, this statement is generally not true. As already stated, care should be taken in interpreting the statement even in finite dimensions. Although norms can be "equivalent" in the sense above, we have seen how the bounds become progressively looser as the space dimension increases, and many such bounds become arbitrarily loose as $n \to \infty$. When thought of this way, the large (but finite) dimension case is conceptually not all that different from the infinite dimensional case.

Exercises

Exercise 2.1

A *seminorm* on a vector space is a functional $|.| : \mathsf{V} \to \mathbb{R}$ which has all the properties of a norm listed in Definition 2.2, except for definiteness. Thus there could be non-zero vectors $v \in \mathsf{V}$ with zero norm $|v| = 0$. Consider the set of zero-norm vectors

$$\mathsf{Z} \;:=\; \big\{v \in \mathsf{V};\; |v| = 0\big\}\,.$$

1. Show that Z is a subspace.

2. Show that if the difference of two vectors is in Z, then they have the same seminorm

   $$(v_1 - v_2) \in \mathsf{Z} \qquad \Rightarrow \qquad |v_1| = |v_2|.$$

   Hint: Apply the triangle inequality to sums like $v_1 = (v_1 - v_2) + v_2$.

3. The previous part implies that if we consider a coset $x + \mathsf{Z}$ of Z, then the quantity

   $$\|x + \mathsf{Z}\| \;:=\; |x|,$$

   is well defined and independent of the coset representative x. Show that this defines a true norm (i.e. it is definite) on the quotient space V/Z.

Solution 2.1

1. Let v_1 and v_2 be in Z, i.e. $|v_1| = |v_2| = 0$, then

   $$|\alpha v_1 + \beta v_2| \;\le\; |\alpha v_1| + |\beta v_2| \;=\; |\alpha||v_1| + |\beta||v_2| \;=\; 0,$$

   where the first inequality is the triangle inequality for $|.|$.

2. As per the hint

   $$\begin{aligned} |v_1| &= |(v_1 - v_2) + v_2| \\ &\le |v_1 - v_2| + |v_2| \;=\; |v_2| \qquad \Big(\text{since } (v_1 - v_2) \in \mathsf{Z} \;\Leftrightarrow\; |v_1 - v_2| = 0\Big) \\ |v_2| &= |(v_2 - v_1) + v_1| \;\le\; |v_2 - v_1| + |v_1| \;=\; |v_1|. \end{aligned}$$

   Therefore $|v_1| = |v_2|$.

3. All properties of $\|.\|$ follow immediately from the properties of $|.|$. For example, the triangle inequality is validated by

$$\begin{aligned}\|\{x+\mathsf{Z}\}+\{y+\mathsf{Z}\}\| &= \|\{(x+y)+\mathsf{Z}\}\| = |x+y| \leq |x|+|y| \\ &= \|\{x+\mathsf{Z}\}\|+\|\{y+\mathsf{Z}\}\|.\end{aligned}$$

Definiteness is also immediate since if $\|\{x+\mathsf{Z}\}\|=0$, then $|x|=0$, and therefore $x\in\mathsf{Z}$, which then implies that the coset $x+\mathsf{Z}$ is the zero coset

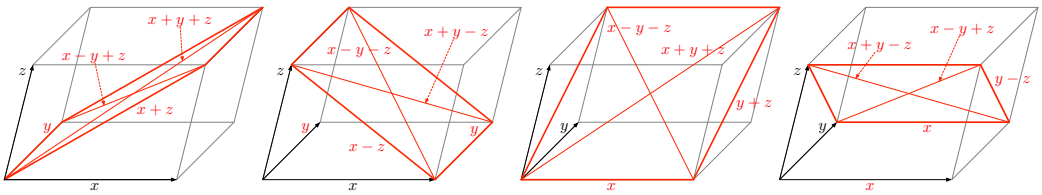
$$\{x+\mathsf{Z}\} = \{0+\mathsf{Z}\}.$$

Exercise 2.2

Given a norm that satisfies the parallelogram law (2.33), show that the form defined by the polarization identities (2.32) satisfies all the properties of an inner product. For example, utilizing the last equation in (2.32), showing $\langle x+y\,,\,z\rangle = \langle x\,,\,z\rangle + \langle y\,,\,z\rangle$ is equivalent to showing that

$$\|x+y+z\|^2-\|x+y-z\|^2 = \|x+z\|^2-\|x-z\|^2 + \|y+z\|-\|y-z\|^2. \tag{2.48}$$

The following diagram will be helpful. Note that the parallelograms in red are the ones that involve the six quantities above.



Solution 2.2

Definiteness and symmetry immediately follow from the definition of the inner product using the polarization identity. Additivity require more work though.

The first two parallelograms in the figure above involve the terms $x+y+z$ and $x+y-z$ as well as the terms $x+z$ and $x-z$. Writing the parallelogram law for each of them

$$\|x+y+z\|^2+\|x-y+z\|^2 = 2\left(\|x+z\|^2+\|y\|^2\right), \tag{2.49}$$

$$\|x-y-z\|^2+\|x+y-z\|^2 = 2\left(\|x-z\|^2+\|y\|^2\right). \tag{2.50}$$

The last two paralellograms also involve the terms $x+y+z$ and $x+y-z$, but the remaining terms $y+z$ and $y-z$ in (2.48) as well

$$\|x+y+z\|^2+\|x-y-z\|^2 = 2\left(\|y+z\|^2+\|x\|^2\right), \tag{2.51}$$

$$\|x+y-z\|^2+\|x-y+z\|^2 = 2\left(\|y-z\|^2+\|x\|^2\right). \tag{2.52}$$

We clearly need to get rid of the terms $\|y\|^2$ and $\|x\|^2$ since they don't occur in (2.48), so subtracting (2.50) from (2.49) and adding that to the difference between (2.51) and (2.52)

$$\begin{aligned}&\|x+y+z\|^2+\cancel{\|x-y+z\|^2}-\cancel{\|x-y-z\|^2}-\|x+y-z\|^2 \\ &+\|x+y+z\|^2+\cancel{\|x-y-z\|^2}-\|x+y-z\|^2-\cancel{\|x-y+z\|^2} \\ &\qquad = 2\left(\|x+z\|^2-\|x-z\|^2+\|y+z\|^2-\|y-z\|^2\right)\end{aligned}$$

This is precisely the relation (2.48).

Now the fact that additivity holds implies (by induction) that the following is valid

$$\langle \alpha x \,,\, y \rangle \;=\; \alpha \,\langle x \,,\, y \rangle \,, \qquad \alpha \in \mathbb{N}.$$

By replacing x with $-x$ we see that homogeneity furthermore is valid for any $\alpha \in \mathbb{Z}$. This fact also implies that homogeneity holds for any $\alpha = n/m \in \mathbb{Q}$ (where n and m are integers) due to the following implications

$$\left\langle \frac{n}{m} x \,,\, y \right\rangle = \frac{n}{m} \langle x \,,\, y \rangle \quad \Leftrightarrow \quad m \left\langle \frac{n}{m} x \,,\, y \right\rangle = n \langle x \,,\, y \rangle \quad \Leftrightarrow \quad \left\langle m \frac{n}{m} x \,,\, y \right\rangle = \langle nx \,,\, y \rangle \,.$$

Note that the only property used was homogeneity for $n, m \in \mathbb{Z}$.

Finally, homogeneity for any $\alpha \in \mathbb{R}$ follows by continuity. The function $\alpha \mapsto \langle \alpha x \,,\, y \rangle$

$$\langle \alpha x \,,\, y \rangle = \frac{1}{4} \left(\|\alpha x + y\|^2 - \|\alpha x - y\|^2 \right),$$

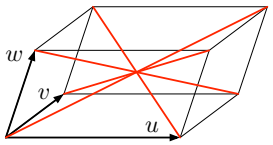
is continuous since $\alpha x + y$ and $\alpha x - y$ are a continuous functions of α, $\|.\|$ is a continuous function, and the sum of two continuous functions is continuous. Since two continuous functions $\alpha \langle x \,,\, y \rangle = \langle \alpha x \,,\, y \rangle$ are equal on a dense subset $\mathbb{Q} \subset \mathbb{R}$, they must be equal for all $\alpha \in \mathbb{R}$.

Exercise 2.3

The following identity is a corollary to the parallelogram law, but involving *three* vectors in an inner product space

$$\begin{aligned} &\|u+v+w\|^2 + \|u+v-w\|^2 + \|u-v+w\|^2 + \|u-v-w\|^2 \\ &\qquad\qquad = \; 4 \left(\|u\|^2 + \|v\|^2 + \|w\|^2 \right) \end{aligned} \tag{2.53}$$

Note the following similarity between (2.53) and (2.33). In both cases, the left hand side is the sum of all possible signed combinations of the vectors, which correspond to all possible diagonal lines in the parallelepiped formed by the vectors u, v, and w in 3D space.



Show that (2.53) follows from (2.33).

Solution 2.3

The identity (2.53) follows from the parallelogram law (2.33) by breaking down the sums on the left hand side of (2.53) into sums of pairs of vectors.

$$\begin{aligned} &\left(\|(u+v)+w\|^2 + \|(u+v)-w\|^2 \right) + \left(\|(u-v)+w\|^2 + \|(u-v)-w\|^2 \right) \\ &\qquad = \; 2 \left(\|u+v\|^2 + \|w\|^2 \right) + 2 \left(\|u-v\|^2 + \|w\|^2 \right) \\ &\qquad = \; 2 \left(\left(\|u+v\|^2 + \left(\|u-v\|^2 \right) + 2\|w\|^2 \right) \;=\; 2 \left(2 \left(\|u\|^2 + \|v\|^2 \right) + 2\|w\|^2 \right) \right). \end{aligned}$$

Exercise 2.4

Consider the set P' of all signals with "finite power" seminorm

$$\|u\| \;:=\; \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} u^2(t) \; dt \;<\; \infty.$$

This is a vector space, but the quantity $\|.\|$ is only a seminorm. For example, any signal that decays asymptotically (e.g. if it is in L^2) will have zero seminorm. Such signals have "finite energy" (if the L^2 norm is interpreted as "energy"), but over all time, their average power is zero.

Let N be the subspace of signals with zero power seminorm. This is clearly a subspace of P'. Now consider the quotient space $\mathsf{P} := \mathsf{P}'/\mathsf{N}$ with the bilinear product

$$\Big\langle \{u+\mathsf{N}\}\ ,\ \{v+\mathsf{N}\}\ \Big\rangle \ := \ \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} u(t)\ v(t)\ dt.$$

Show that this satisfies all the properties of an inner product on P'/N, including definiteness.

Solution 2.4

Symmetry and bilinearity follow immediately from the integral form. For definiteness, note that

$$0 \ = \ \Big\langle \{v+\mathsf{N}\}\ ,\ \{v+\mathsf{N}\}\ \Big\rangle \ = \ \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} v^2(t)\ dt,$$

implies that $v \in \mathsf{N}$. Therefore an element $\{v+\mathsf{Z}\}$ with zero inner product with itself is in the same coset as $\{0+\mathsf{N}\}$.

Exercise 2.5

One of the most basic inequalities used in many bounds involves the interplay between absolute sums and maxima. One could call it the 1-∞ *inequality*. For n-vectors v and u it states

$$\begin{aligned} |v^*u| \ &:= \ |v_1u_1+\cdots+v_nu_n| \ \le \ |u_1v_1|+\cdots+|u_nv_n| \ = \ |u_1|\,|v_1|+\cdots+|u_n|\,|v_n| \\ &\le \ \Big(\max_i |u_i|\Big)\,|v_1|+\cdots+\Big(\max_i |u_n|\Big)\,|v_n| \ = \ \Big(\max_i |u_i|\Big)\,\Big(|v_1|+\cdots+|v_n|\Big) \\ &= \ \|u\|_\infty\ \|v\|_1. \end{aligned}$$

This inequality can be used in many ways. Whenever one has a sum of several terms, the individual terms can be grouped into the entries of two different vectors in several different ways. Such arguments are useful e.g. in Exercise 2.6.

Show with a similar argument that for infinite sequences

$$\left|\sum_{i\in\mathbb{N}} u_iv_i\right| \ \le \ \Big(\sup_{i\in\mathbb{N}}|u_i|\Big)\left(\sum_{i\in\mathbb{N}}|v_i|\right),$$

and that for any function $f:\Omega\to\mathbb{R}$ defined on a subset $\Omega\subset\mathbb{R}^n$

$$\left|\int_\Omega f(x)\ g(x)\ dx\right| \ \le \ \Big(\sup_{x\in\Omega}|f(x)|\Big)\left(\int_\Omega |g(x)|\,dx\right).$$

Solution 2.5

For the discrete sum

$$\left|\sum_{i\in\mathbb{N}} u_iv_i\right| \ \le \ \sum_{i\in\mathbb{N}}|u_i|\,|v_i| \ \le \ \sum_{i\in\mathbb{N}}\Big(\sup_{i\in\mathbb{N}}|u_i|\Big)\,|v_i| \ = \Big(\sup_{i\in\mathbb{N}}|u_i|\Big)\left(\sum_{i\in\mathbb{N}}|v_i|\right) = \|u\|_\infty\|v\|_1.$$

For the integral

$$\left|\int_\Omega f(x)\ g(x)\ dx\right| \le \int_\Omega |f(x)|\,|g(x)|\ dx \le \int_\Omega \left(\sup_{x\in\Omega}|f(x)|\right)|g(x)|\ dx$$
$$= \left(\sup_{x\in\Omega}|f(x)|\right)\int_\Omega |g(x)|\ dx = \|u\|_\infty \|v\|_1.$$

Note: It is tempting to think of these bounds as parallels of the Cauchy-Schwartz inequality

$$\langle u\,,\,v\rangle \ \le\ \|u\|_\infty\|v\|_1.$$

It turns out that we can make this precise as we will see later. The quantity $\langle u\,,\,v\rangle$ is not to be interpreted as an inner product, but rather as a $u \in L^\infty$ "acting" on $v \in L^1$ as a linear functional. Unlike the case of an inner product space, here u and v are in different spaces, but one of them is the space of all bounded linear functionals (namely L^∞) on L^1. In fact, we will generalize the notation $\langle u\,,\,v\rangle$ to mean u acting on v as a linear functional.

Exercise 2.6

Use the Cauchy-Schwartz inequality

$$\left(\sum_{i=1}^n a_i b_i\right)^2 \ \le\ \left(\sum_{i=1}^n a_i^2\right)\left(\sum_{i=1}^n b_i^2\right)$$

to prove the norm bound $\|v\|_1 \le \sqrt{n}\|v\|_2$.
Hint: Rewrite $\|v\|_1 := \sum_{i=1}^n |v_i| \ = \ \sum_{i=1}^n v_i\ \mathrm{sign}(v_i)$*, and use Exercise 2.5.*

Solution 2.6

Examining the inequality $\|v\|_1 \le \sqrt{n}\|v\|_2$ that we need to prove

$$\sum_{i=1}^n |v_i| \ \le\ \sqrt{n}\left(\sum_{i=1}^n v_i^2\right)^{1/2} \qquad \Leftrightarrow \qquad \left(\sum_{i=1}^n |v_i|\right)^2 \ \le\ n\ \sum_{i=1}^n v_i^2.$$

Comparing with the Cauchy-Schwartz inequality, we see that if we apply it with $a = (v_1,\ldots,v_n)$ and $b = (\mathrm{sign}(v_1),\ldots,\mathrm{sign}(v_n))$ then

$$\left(\sum_{i=1}^n v_i\ \mathrm{sign}(v_i)\right)^2 \ \le\ \left(\sum_{i=1}^n v_i^2\right)\left(\sum_{i=1}^n (\mathrm{sign}(v_i))^2\right)$$
$$\left(\sum_{i=1}^n |v_i|\right)^2 \ \le\ \left(\sum_{i=1}^n v_i^2\right)\left(\sum_{i=1}^n 1\right) \ =\ \left(\sum_{i=1}^n v_i^2\right)\ n$$
$$\Leftrightarrow \qquad \|v\|_1 \ \le\ \sqrt{n}\ \|v\|_2^2.$$

Chapter 3

Completeness and Continuity: Banach and Hilbert Spaces

The metrics induced by norms and inner products imply a notion of convergence of sequences in the space. A space is called complete if every sequence that "should converge" does indeed converge in that space. Cauchy sequences are those that "should converge", and a space is complete if every Cauchy sequence converges in that space. Banach and Hilbert spaces are complete normed and inner product spaces respectively. In such complete spaces one can make sense of infinite bases, and the convergence of partial sums of basis expansions to any particular element.

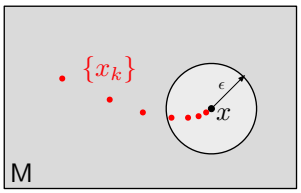
A notion of convergence induces a notion of continuity. A linear mapping between two vector spaces is a continuous mapping iff it is continuous at the origin. Norms on vector spaces induce a natural norm on linear operators called the "induced norm". The boundedness of the induced norm is equivalent to the continuity of the mapping at the origin, and therefore everywhere.

The set of all bounded linear operators on a Banach space is itself a Banach space, with the induced operator norm as the norm. The induced norm has the important property of "sub-multiplicativity", which makes the space of all bounded linear operators into a special type of algebra called a Banach Algebra.

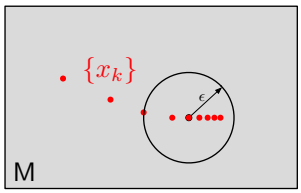
This chapter is concerned primarily with the basic "analysis" questions in vector spaces.

Introduction

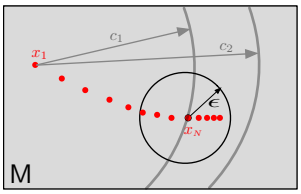
So far we have discussed the *algebraic* and *geometrical* properties of normed and inner product spaces. When describing limit processes or iterative algorithms in such spaces, we will also need a notion of convergence and topology. These are the *analysis* aspects of the subject. There are many ways to deal with these topological notions, but we will adopt here the least abstract setting where we use distances and norms to describe convergence. A metric can be used to define closed and open sets, and whether a vector space is complete or not. Intuitively, a set is closed or a space is complete if all sequences that "should converge" do converge in that space, i.e. the space has no holes or open boundaries in it. Complete normed spaces are called *Banach spaces*, while complete inner product spaces are called *Hilbert spaces*. We will see that once the algebraic and metric properties of a space are specified, it can always be "completed" (i.e. all the holes added to the space) so that it becomes a complete space.



(a) A sequence $\{x_k\}$ converges to a *limit point* x if given any $\epsilon > 0$ (no matter how small), an infinite "tail" of the sequence is inside an ϵ-neighborhood around x.



(b) A sequence $\{x_k\}$ is *Cauchy* if given any $\epsilon > 0$ (no matter how small), an infinite "tail" of the sequence is inside an ϵ ball, i.e. the tail "bunches up" even if there is no limit point.



(c) A Cauchy sequence is bounded since a tail is guaranteed to be inside an ϵ ball, and the remainder of the sequence (the "head") is finite, and therefore within a distance c_1 of the first element x_1.

Figure 3.1: The concepts of (a) convergent sequences, (b) Cauchy sequences, and (c) the fact that Cauchy sequences in a metric space M are bounded.

3.1 Convergence and Topology

The first concept to deal with is how to define convergence. If we are in a metric space, we can use the distance function to define convergence.

Definition 3.1. *Let* M *be a metric space with metric* $d(.,.)$*. We say a sequence* $\{x_k\} \subset \mathsf{M}$ *converges to a* limit point $x \in \mathsf{M}$ *(also written as* $\lim_{k\to\infty} x_k = x$*) if for any* $\epsilon > 0$ *there exists a number* N *such that*

$$\forall k \geq N, \quad \mathsf{d}(x_k, x) \leq \epsilon. \tag{3.1}$$

One way to parse this definition is to think of "tails" $\{x_k\}_{k\geq N}$ of the sequence. Each choice of N defines a tail of the sequence with an infinite number of elements in it. The definition says that x is a limit point, if for any distance ϵ, no matter how small, an entire tail of the sequence is within that small distance from the limit point. This is illustrated in Figure 3.1a.

To use the definition above to determine if a sequence is convergent requires knowing the limit x apriori. There is another way to define convergence that does not require knowing the limit. These are the sequences that "should converge".

Definition 3.2. *Let* M *be a metric space with metric* $d(.,.)$*. A sequence* $\{x_k\} \subset \mathsf{M}$ *is called* Cauchy *if for any* $\epsilon > 0$ *there exists a number* N *such that*

$$\forall k, l \geq N, \quad \mathsf{d}(x_k, x_l) \leq \epsilon. \tag{3.2}$$

This property will be equivalently stated in the abbreviated form

$$\lim_{k,l\to\infty} \mathsf{d}(x_k, x_l) \;=\; 0.$$

This means that given any distance ϵ, no matter how small, there is a tail of the sequence such that all elements of the tail are within ϵ-distance of each other (see Figure 3.1b). In contrast to the condition (3.1), the Cauchy condition (3.2) *does not require the existence (or knowing) of a limit point.* Now, the first property to establish is that a Cauchy sequence can not "stray too far".

Lemma 3.3. *Every Cauchy sequence in metric space is* bounded*.*

Proof. Choose some ϵ and find N such that all elements of the tail $\{x_k\}_{k\geq N}$ are within ϵ of each other. Now find a ball around the first element x_1 that contains all of the first N elements (see Figure 3.1c)

$$c \;:=\; \max_{1\leq k\leq N} \mathsf{d}(x_1, x_k),$$

and observe that c must be finite. Thus the first N elements are within distance c from x_1. In addition, *the entire sequence is within distance* $c+\epsilon$ *from* x_1 because (see also Figure 3.1c)

$$k \geq N \qquad \Rightarrow \qquad \mathsf{d}(x_1, x_k) \;\leq\; \mathsf{d}(x_1, x_N) + \mathsf{d}(x_N, x_k) \;\leq\; c + \epsilon$$

as follows from the triangle inequality. Thus the sequence is bounded. □

The Cauchy condition appears to be the right condition for our intuitive notion of when a sequence "should converge". This motivates the next definition.

Definition 3.4. *A metric space* M *is called* complete *if every Cauchy sequence is convergent, i.e. if it has a limit point in* M.

The classic example of an incomplete metric space is the set of rationals $\mathbb{Q}$ with the usual distance $d(x, y) := |x - y|$. Every decimal expansion of an irrational number (e.g. π or $\sqrt{2}$) defines a sequence of rationals (the truncation of that decimal expansion to a progressively larger, but finite, number of digits) that converges to an irrational.

Given any metric space that is not complete, there is a procedure for *completing* it by formally adding all the Cauchy sequences to it. The details are outlined in Appendix 3.A, and will not be elaborated here. For example, *the completion of* $\mathbb{Q}$ *is the real line* $\mathbb{R}$, and in fact, that is one way to formally construct the reals from the rationals. We will assume from now on that $\mathbb{R}$ is a complete metric space.

Open and Closed Sets

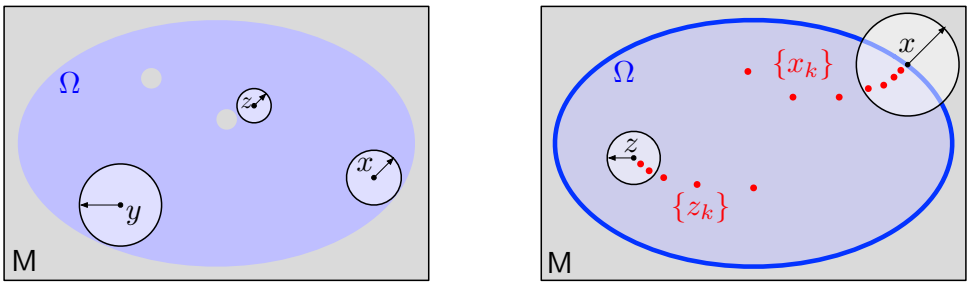
Now we introduce the concepts of closed and open sets. Consider the open and closed intervals $(0, 1)$ and $[0, 1]$ in $\mathbb{R}$. The interval $[0, 1]$ contains all its limit points, i.e. all Cauchy sequences in $[0, 1]$ have limits in $[0, 1]$. On the other hand, the open interval $(0, 1)$ does not since the sequence $1/k$ is a Cauchy sequence entirely in $(0, 1)$ with its limit $0 \notin (0, 1)$. The open interval does however have a property that the closed interval lacks. For every number $\bar{x} \in (0, 1)$, no matter how close to the boundary, we can find a "neighborhood" $\{x \in (0, 1);\ |x - \bar{x}| \leq \epsilon\}$ of it that is entirely contained within $(0, 1)$. These concepts generalize to metric spaces as follows.

Definition 3.5. *Let* M *be a complete metric space. A subset* $\Omega \subseteq \mathsf{M}$ *is called*

- closed*: if every Cauchy sequence* $\{x_k\} \subset \Omega$ *converges to a point* $x \in \Omega$*, i.e. if* Ω *is itself* complete*. For any arbitrary set* $\Omega \in \mathsf{M}$*, its* closure $\overline{\Omega}$ *is the smallest closed set containing it, namely the set obtained from* Ω*'s union with all of its limit points.*
- open*: if for each point* $\bar{x} \in \Omega$*, we can find a "neighborhood of* $\bar{x}$*", i.e. an* $\epsilon > 0$ *such that* $\mathsf{N}_\epsilon(\bar{\mathsf{x}}) := \{\mathsf{x} \in \Omega;\ \mathsf{d}(\mathsf{x}, \bar{\mathsf{x}}) \leq \epsilon\}$ *is entirely contained in* Ω*.*

Informally, a closed set is one which has no holes in it and contains all of of its boundaries. An open set is one without boundaries, i.e. there are no points at "edges of the set" where any neighborhood of those points, no matter how small, will contain points from outside the set. These concepts are depicted visually in Figure 3.2. As the terminology suggests, open intervals $(a, b) \subset \mathbb{R}$ are open sets, and closed intervals $[a, b]$ are closed set. It is not difficult to show from the definitions above that the set complement $\mathsf{M} - \Omega$ of an open set Ω is closed, and vice versa.

Note that a set can be open, closed, both open and closed, or neither. For example $\mathbb{R}$ is both open and closed (it is complete, therefore closed, and it contains all neighborhoods, therefore open). In fact, the entire complete metric space M in the definitions above is both open and closed. The set of rationals $\mathbb{Q} \subset \mathbb{R}$ is neither open nor closed. It is "full of holes", which are all the irrationals. Any irrational can be approximated by a Cauchy sequence of rationals. The real line $\mathbb{R}$ is actually the completion (or *closure*) of $\mathbb{Q}$ in $\mathbb{R}$, which we



(a) A set is open if for any point inside it (x, y or z above), there is a neighborhood of the point that is contained entirely in the set. Such sets cannot have boundaries, i.e. a point x (as in the diagram to the right) such that all neighborhoods contain contain points from outside the set.

(b) A set is closed if it contains all of its limit points. Such a set has no holes and must contain all of its boundaries as well. Here, Cauchy sequences $\{z_k\}$ and $\{x_k\}$ are depicted converging to a limits z and x inside and on the boundary respectively. The boundary must be part of the set.

Figure 3.2: Graphical depiction of open (a) and closed (b) sets Ω in a complete metric space M.

write as $\overline{\mathbb{Q}} = \mathbb{R}$. The set $\mathbb{Q}$ is also not open in $\mathbb{R}$ since any neighborhood around a rational contains infinitely many irrationals. Thus $\mathbb{Q}$ is neither open nor closed.

Continuous Mappings

Definition 3.6. *Let $f : \mathsf{M}_1 \to \mathsf{M}_2$ be a mapping between two metric spaces.*

1. *f is* continuous *if the inverse image of every open set is open, or equivalently, the inverse image of every closed set is closed.*

2. *f is* continuous at a point $\bar{x} \in \mathsf{M}_1$ *if given any $\epsilon > 0$, there exists a $\delta > 0$ such that*

$$\mathsf{d}(\bar{x}, \tilde{x}) \leq \delta \qquad \Rightarrow \qquad \mathsf{d}\big(f(\bar{x}), f(\tilde{x})\big) \leq \epsilon. \tag{3.3}$$

   f is called continuous *if it is continuous at each $x \in \mathsf{M}_1$.*

3. *f is called* sequentially continuous *if for every convergent sequence $x_k \overset{k\to\infty}{\longrightarrow} x$, the values of f also converge, i.e. $f(x_k) \overset{k\to\infty}{\longrightarrow} f(x)$.*

The first definition is the most general one and is valid in any topological space, where the topology is defined in terms of open sets and not necessarily in terms of a metric. Definition 2 requires a metric. Sequential continuity in general does not require a metric, but rather some notion of (sequential) convergence in the space. There is one important example of sequential convergence that does not initially arise from a metric, and that example is encountered in the theory of distributions (generalized functions). We leave this discussion for later when we study generalized functions.

It is not difficult to show that *1, 2* and *3* are equivalent in a metric space, and this is left as an exercise.

3.2 Banach and Hilbert Spaces

We now discuss the main issue in this chapter, which is complete *vector spaces* as opposed to metric spaces in general. Recall that an inner product space is also a normed space, and a normed space in turn is also a metric space, and therefore the completeness criterion in Definition 3.4 applies equally well to inner product and normed vector spaces. The next definition introduces the standard terminology for such spaces.

Definition 3.7. *A normed vector space that is complete is called a* Banach space. *A complete inner product space is called a* Hilbert space.

Given an incomplete normed space, we can always form its "completion" (by appending all the Cauchy sequences as described in Appendix 3.A) to obtain a complete normed space, i.e. a Banach space. Similarly, any incomplete inner product space[1] can be completed to a Hilbert space. The completed spaces contain the original ones as (incomplete, i.e. not closed) subspaces.

The spaces we will work most closely with are the L^p and ℓ^p spaces. We have already shown that they are normed spaces (and an inner product space in the case of $p = 2$). They are also complete spaces.

Theorem 3.8. *Let Ω be a subset of $\mathbb{R}^n$ or $\mathbb{Z}^n$. The spaces $\mathsf{L}^p(\Omega)$ or $\ell^p(\Omega)$ for $p \in [1, \infty]$ are complete, i.e. they are Banach spaces. In particular, ℓ^2 and L^2 are Hilbert spaces.*

The proof of the L^p case is somewhat technical, requiring concepts from measure theory and Lebesgue integration. It can be found in most textbooks on functional analysis and will therefore be omitted. The proof for the ℓ^p case is less technical and more instructive, so we illustrate it next.

Proof of Theorem 3.8 for the ℓ^p case. Let $\left\{v^{(k)}\right\}$ be a Cauchy sequence of elements in $\ell^p(\Omega)$, i.e. the sequence is such that given any $\epsilon > 0$, $\exists N$ such that for all $k, l \geq N$

$$\|v^{(k)} - v^{(l)}\|_p^p \;:=\; \textstyle\sum_{i\in\Omega} \left|v_i^{(k)} - v_i^{(l)}\right| \;\leq\; \epsilon.$$

Note that since Ω is discrete (i.e. countable), then the norm is given by a sum over Ω. Therefore, at any individual point $i \in \Omega$ we have

$$\left|v_i^{(k)} - v_i^{(l)}\right| \;\leq\; \textstyle\sum_{i\in\Omega} \left|v_i^{(k)} - v_i^{(l)}\right| \;\leq\; \epsilon.$$

This means that at any $i \in \Omega$, the sequence of real numbers $v_i^{(k)}$ is Cauchy, and therefore by completeness of $\mathbb{R}$, it must have a limit which we will call v_i. We have thus established that a sequence $\left\{v_i^{(n)}\right\}$ that is Cauchy in $\ell^p(\Omega)$ converges pointwise (i.e. at each $i \in \Omega$) to some function $v : \Omega \to \mathbb{R}$, i.e.

$$\text{for each } i \in \Omega, \qquad v_i \;=\; \lim_{k\to\infty} v_i^{(k)}.$$

It remains to prove that the function v has finite p-norm. Since Ω is countable, we can identify Ω with $\{1, 2, \ldots\}$ and examine the partial sums

$$\textstyle\sum_{i=1}^N \left|v_i^{(k)}\right|^p \;\leq\; \sum_{i=1}^\infty \left|v_i^{(k)}\right|^p \;\leq\; C,$$

where the last inequality is because $\left\{v^{(k)}\right\}$ is a Cauchy sequence in ℓ^p, and therefore is bounded[2], i.e. all their ℓ^p norms are bounded from above by some constant C. Since this bound is independent of N, and the finite-vector sequence $\left\{v_i^{(k)};\ 1 \leq i \leq N\right\}$ converges (in $\mathbb{R}^N$) to the finite vector $\{v_i;\ 1 \leq i \leq N\}$, then we also have the bound

$$\textstyle\sum_{i=1}^N |v_i|^p \;\leq\; C.$$

This bound is independent of N, and therefore is also a bound on the infinite sum

$$\textstyle\sum_{i=1}^\infty |v_i|^p \;\leq\; C,$$

and the limit function v is therefore in ℓ^p. □

[1] An incomplete inner product space is sometimes called a "pre-Hilbert" space.
[2] Recall that in a metric space, every Cauchy sequence must be bounded.

Example 3.9. Consider the subspace of $\ell^\infty(\mathbb{N})$ of asymptotically decaying sequences

$$\ell_0^\infty(\mathbb{N}) \;:=\; \left\{ v \in \ell^\infty(\mathbb{N});\;\; \lim_{k\to\infty} v_k = 0 \right\}.$$

This is clearly a subspace, but in addition it is a closed subspace in the ℓ^∞ norm. Therefore it is a Banach space in itself with the $\|.\|_\infty$ norm. For several reasons to be explained later, whenever the $\|.\|_\infty$ norm is needed in applications, the space ℓ_0^∞ is a better setting for a problem than ℓ^∞.

Example 3.10. Consider the space $\mathsf{C}[a,b]$ of continuous functions on the interval $[a,b]$ with the maximum norm

$$\mathsf{C}[a,b] \;:=\; \left\{ f : [a,b] \to \mathbb{R};\;\; f \text{ continuous} \right\}, \qquad \|f\|_\infty \;:=\; \max_{x\in[a,b]} |f(x)|.$$

Note that since f is continuous, then its supremum on a closed interval is achieved at some point in that interval (this is the "extreme value theorem" of Calculus), thus we write maximum instead of supremum. This is a subspace of $L^\infty[a,b]$ and therefore is a normed vector space. The question is whether it is complete? Let $\{f_k\}$ be a Cauchy sequence (in the max norm) of continuous functions. At each $x \in [a,b]$, $\{f_k(x)\}$ is a Cauchy sequence of real numbers because

$$\forall x \in [a,b], \quad |f_k(x) - f_l(x)| \;\leq\; \|f_k - f_l\|_\infty .$$

This means that at each x the sequence of real numbers $\{f_k(x)\}$ converges to a real number, which we can define as the value of the "limit function" f

$$f(x) \;:=\; \lim_{k\to\infty} f_k(x), \qquad x \in [a,b]$$

This function is the limit of the sequence $\{f_k\}$ in the max norm , i.e. $\lim_{k\to\infty} \|f_k - f\|_\infty = 0$. Note that convergence in the max norm is *uniform convergence*, and the *uniform convergence theorem* states that the uniform limit of continuous functions is continuous, thus the limit function f is itself a continuous function and therefore in $\mathsf{C}[a,b]$.

The space $\mathsf{C}[a,b]$ is therefore a Banach space, and it is a closed subspace of $L^\infty[a,b]$. This is only the case for bounded intervals. For example the vector space $C[0,\infty)$ is not a normed space with the max norm since continuous functions on an unbounded interval are not necessarily bounded.

Example 3.11. Consider the space $\mathsf{C}^1[a,b]$ of continuously differentiable functions on the interval $[a,b]$ equipped with the "max" norm

$$\mathsf{C}^1[a,b] \;:=\; \left\{ f : [a,b] \to \mathbb{R};\;\; f' \text{ continuous} \right\}, \qquad \|f\|_\infty \;:=\; \max_{x\in[a,b]} |f(x)|.$$

Note that since f is continuously differentiable, then its integral f is also continuous, and therefore the maximum above is achieved for some $x \in [a,b]$. This is clearly a normed vector space (it is closed under additions and scalings), which is also a subspace of $\mathsf{C}[a,b]$ (continuous functions on $[a,b]$ with the max norm). However, this subspace is not complete with respect to the max norm. It is actually a dense subspace in $\mathsf{C}[a,b]$ since any continuous function can be approximated arbitrarily closely by a continuously differentiable function in the max norm (see Figure 3.3). As we will see later, $\mathsf{C}^1[a,b]$ can be made into a Banach space by defining a different (Sobolev) norm on it.

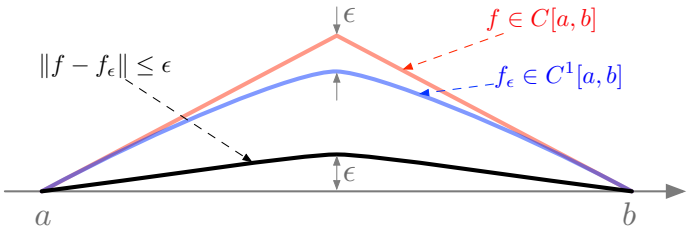


Figure 3.3: Any continuous function f can be approximated arbitrarily closely by a continuously differentiable function f_ϵ in the max norm $\|f-f_\epsilon\|_\infty := \max_{x\in[a,b]} |f(x)-f_\epsilon(x)|$. Thus the subspace of continuously differentiable functions $\mathsf{C}^1[a,b]$ is dense in the Banach space of continuous functions $C[a,b]$ with the max norm. While $\mathsf{C}^1[a,b]$ with this norm is itself a normed vector space, but it is not complete, and therefore not a Banach space itself.

3.3 Bases

There are several notions of bases that can be defined in infinite-dimensional spaces. We have already encountered Hamel bases (Definition 1.20), and recall the remark that that notion of basis is not particularly useful. The main application of bases in infinite-dimensional spaces is to enable approximations by finite-dimensional "truncations", and the approximation is to be measured using norms. The following definition which is sometimes referred to as a "Schauder basis" captures this idea.

Definition 3.12. *Let* $\mathsf{v} := \{\boldsymbol{v}_k\}_{k=0}^{\infty}$ *be an ordered, countable set in a Banach space* V*. This set is called a* basis *of* V *if every element* $\boldsymbol{u} \in \mathsf{V}$ *can be expressed uniquely as*

$$\boldsymbol{u} \;=\; \sum_{k=0}^{\infty} \alpha_k \; \boldsymbol{v}_k,$$

where the convergence is in the norm of V*. The unique sequence of numbers* $\{\alpha_k\}$ *are called the* coefficients *of* $\boldsymbol{u}$ *in the basis* v*.*

This definition implies that the closure of $\mathsf{span}\{\boldsymbol{v}_k\}$ is all of V, but it is a little stronger than that since the uniqueness of the coefficients is also required. We have already implicitly worked with basis for the ℓ^p spaces.

Example 3.13. Consider the spaces $\ell^p(\mathbb{N})$ for $p \in [1,\infty)$, and the set of vectors $\mathsf{e} := \{\boldsymbol{e}_k\}$

$$\boldsymbol{e}_k \;:=\; (0,\cdots,0,\underset{\uparrow\, k\text{'th entry}}{1},0,\cdots), \qquad\qquad k \in \mathbb{N}. \tag{3.4}$$

Now any element $\boldsymbol{u} \in \ell^p(\mathbb{N})$ can be written uniquely in this basis as

$$\boldsymbol{u} \;:=\; (u_0,u_1,\cdots) \;=\; \sum_{k=0}^{\infty} u_k \boldsymbol{e}_k. \tag{3.5}$$

The fact that $\boldsymbol{u} \in \ell^p(\mathbb{N})$ (for $p < \infty$) implies that the tails of the sequence $\{u_k\}$ decay $\sum_{k=n}^{\infty} |u_k|^p \overset{n\to\infty}{\longrightarrow} 0$, and therefore the partial sums of the series (3.5) converge in the ℓ^p norm

$$\left\| \sum_{k=0}^{\infty} u_k \boldsymbol{e}_k - \sum_{k=0}^{n-1} u_k \boldsymbol{e}_k \right\|_p^p \;=\; \left\| \sum_{k=n}^{\infty} u_k \boldsymbol{e}_k \right\|_p^p \;\leq\; \sum_{k=n}^{\infty} |u_k|^p \, \|\boldsymbol{e}_k\|_p^p \;=\; \sum_{k=n}^{\infty} |u_k|^p \overset{n\to\infty}{\longrightarrow} 0.$$

Thus the series (3.5) is convergent in the ℓ^p norm for any $\boldsymbol{u} \in \ell^p(\mathbb{N})$, $p \in [1,\infty)$. Finally note that the case $p = \infty$ is excluded in this example since the argument fails in that case. We will discuss this issue shortly in the context of the concept of separability.

Bases in Banach versus Hilbert Spaces

The main difficulty in working with bases in Banach spaces is that, except for very special cases (e.g. the canonical basis of $\mathbb{R}^n$, or the bases of Example 3.13), there is usually not a simple relationship between the vector norm and the basis coefficients. This problem is apparent even in finite dimensional Banach spaces such as $\mathbb{R}^n$ with the $\|.\|_p$ norms ($p \neq 2$). Consider for example the $\|.\|_1$ norm on $\mathbb{R}^2$. In the canonical basis, the expressions for this norm in terms of the vector components follow from the definition as relatively simple expressions

$$\|\boldsymbol{u}\|_1 \;=\; \left\| \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \right\|_1 \;=\; \|u_1\boldsymbol{e}_1 + u_2\boldsymbol{e}_2\|_1 \;=\; |u_1| + |u_2|.$$

On the other hand, if we choose a different basis, say $\boldsymbol{v}_1 := \boldsymbol{e}_1 + \boldsymbol{e}_2$ and $\boldsymbol{v}_2 := \boldsymbol{e}_1 - \boldsymbol{e}_2$, then

$$\begin{aligned} \|\boldsymbol{u}\|_1 \;&=\; \|u_1\boldsymbol{v}_1 + u_2\boldsymbol{v}_2\|_1 \;=\; \|u_1(\boldsymbol{e}_1 + \boldsymbol{e}_2) + u_2(\boldsymbol{e}_1 - \boldsymbol{e}_2)\|_1 \\ &=\; \|(u_1 + u_2)\boldsymbol{e}_1 + (u_1 - u_2)\boldsymbol{e}_2\|_1 \;=\; |u_1 + u_2| + |u_1 - u_2|. \end{aligned}$$

The expressions become even more complex for other choices of bases, and more so in higher dimensions. When the norms are given by a general convex set, it is generally not possible to give algebraic expressions in terms of the basis coefficients.

The situation above is to be contrasted with that in a Hilbert space. Consider $\mathbb{R}^n$ again with the Euclidean norm $\|.\|_2$. Suppose we choose a basis $\mathsf{v} := \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$. Recall from Lemma 1.43 that the relation between the coefficients in the new basis and the vector components (which are the coefficients in the canonical basis) are

$$\begin{aligned} \boldsymbol{u} \;&=\; x_1\boldsymbol{e}_1 + \cdots + x_n\boldsymbol{e}_n \\ &=\; \hat{x}_1\boldsymbol{v}_1 + \cdots + \hat{x}_n\boldsymbol{v}_n \end{aligned} \quad \Leftrightarrow \quad x := \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_n \end{bmatrix} \begin{bmatrix} \hat{x}_1 \\ \vdots \\ \hat{x}_n \end{bmatrix} =: V\hat{x}.$$

Now since the norm is given by the inner product, we can calculate the norm in terms of the new coefficients $\hat{x}_1, \ldots, \hat{x}_n$ by

$$\|\boldsymbol{u}\|_2^2 \;=\; x_1^2 + \cdots + x_n^2 \;=\; x^*x \;=\; \hat{x}^*V^*V\hat{x} \;=\; \hat{x}^*(V^*V)\,\hat{x}.$$

This is a modified inner product, and can be easily calculated with matrix-vector operations. It is for this reason that bases in Hilbert space are much easier to work with than in general Banach spaces.

Orthogonal Bases in Hilbert Space

Recall that inner product spaces have a much richer geometry than normed spaces due to the inner product which gives a notion of angles between vectors. In particular, mutually orthogonal vectors provide bases with particularly nice properties.

Definition 3.14. *A basis $\mathsf{v} := \{v_k\}$ of a Hilbert space with the properties*

$$\langle v_k \,,\, v_l \rangle \;=\; \begin{cases} 1, & k = l \\ 0, & k \neq l \end{cases}$$

is called an orthonormal basis.

Orthonormal bases play very nicely with norms and inner product as shown by the following famous identities.

Theorem 3.15. *Let* $\mathsf{v} := \{\boldsymbol{v}_k\}_{k=0}^{\infty}$ *be an orthonormal basis of a Hilbert space* V*. Then for any vectors* $\boldsymbol{u}, \boldsymbol{w} \in \mathsf{V}$ *with expansions* $\boldsymbol{u} = \sum_{k=0}^{\infty} \alpha_k \boldsymbol{v}_k$ *and* $\boldsymbol{u} = \sum_{k=0}^{\infty} \beta_k \boldsymbol{v}_k$

$$\langle \boldsymbol{u}\,,\, \boldsymbol{w} \rangle = \sum_{k=0}^{\infty} \alpha_k^* \beta_k \qquad \text{(Plancherel Identity)}$$

$$\|\boldsymbol{u}\|^2 = \sum_{k=0}^{\infty} |\alpha_k|^2 \qquad \text{(Parseval's Identity)}$$

Proof. First note that Parseval's identity follows from the Plancherel identity by choosing the two vectors equal[3]. Now compute

$$\begin{aligned}
\langle \boldsymbol{u}\,,\, \boldsymbol{w} \rangle &= \left\langle \sum_{k=0}^{\infty} \alpha_k \boldsymbol{v}_k \;,\; \sum_{l=0}^{\infty} \beta_l \boldsymbol{v}_l \right\rangle = \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} \langle \alpha_k \boldsymbol{v}_k \,,\, \beta_l \boldsymbol{v}_l \rangle \\
&= \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} \alpha_k^* \beta_l \; \langle \boldsymbol{v}_k \,,\, \boldsymbol{v}_l \rangle = \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} \alpha_k^* \beta_l \; \delta_{k\text{-}l} \qquad (\delta \text{ is the Kronecker delta}) \\
&= \sum_{k=0}^{\infty} \alpha_k^* \beta_k \qquad \square
\end{aligned}$$

Parseval's identity can be thought of as the infinite-dimensional version of the Pythagorean theorem. Another interpretation of the two identities is that any choice of orthonormal basis of a Hilbert space makes it "look like" $\ell^2(\mathbb{N})$. The one-to-one and onto correspondence is given by the mapping from a vector to its basis coefficients $\boldsymbol{u} \mapsto (\alpha_0, \alpha_1, \ldots)$. Parseval's identity implies that this is an isometric isomorphism between V and $\ell^2(\mathbb{N})$. Therefore *any choice of orthonormal basis of a Hilbert space induces an isometric isomorphism between it and* $\ell^2(\mathbb{N})$*.*

Riesz Bases in Hilbert Space

When it is not possible to use orthogonal bases in Hilbert space, the next best construction is as follows.

Definition 3.16. *A basis* $\mathsf{v} := \{v_k\}$ *of a Hilbert space* V *is called a* Riesz basis *if there are constants* $\underline{c}, \overline{c} > 0$ *such that for any vector* $\boldsymbol{u} \in \mathsf{V}$

$$\boldsymbol{u} = \sum_{k=0}^{\infty} \alpha_k \boldsymbol{v}_k \qquad \Rightarrow \qquad \underline{c} \sum_{k=0}^{\infty} |\alpha_k|^2 \;\leq\; \|\boldsymbol{u}\|^2 \;\leq\; \overline{c} \sum_{k=0}^{\infty} |\alpha_k|^2 \,. \tag{3.6}$$

To appreciate the need for such a definition, the reader should recall the concept of "equivalence of norms" of Section 2.C. The definition above says that the Hilbert space norm $\|\boldsymbol{u}\|$ and the ℓ^2 norm of the sequence of coefficients $\{\alpha_k\}$ are equivalent. Since each choice of basis induces an isomorphism from V to $\ell^2(\mathbb{N})$, the equivalence (3.6) implies that this isomorphism is at least *continuous* even if it is not an isometry (as would be the case with an orthonormal basis). Thus convergence arguments in V are equivalent to convergence arguments of the corresponding basis coefficients in $\ell^2(\mathbb{N})$.

[3] Since the inner product is also determined by the norm (via the polarization identity), then we can also say that Parseval's identity implies Plancherel's. For this reason, the two names are used interchangeably in the literature.

Separability

Recall the convergence argument of Example 3.13, and note that the argument fails in the case of ℓ^∞ since the norm of the tail

$$\left\| \sum_{k=n}^{\infty} u_k \boldsymbol{e}_k \right\|_\infty = \sup_{k \geq n} |u_k|,$$

and this quantity does not always decay as $n \to \infty$ (e.g. consider the element $\boldsymbol{u} := (1, 1, \cdots)$). However, the argument works in the space ℓ^∞_0 of Example 3.9, and therefore the set e is a basis for ℓ^∞_0.

The fact that the set (3.4) is not a basis for $\ell^\infty(\mathbb{N})$ is related to the important concept of separability.

Definition 3.17. *A vector space is called* separable *if it contains a countable dense subset.*

The importance of this definition is that if a space is not separable, then its elements cannot be approximated using finite arithmetic, e.g. on a computer. Therefore a non-separable space is usually not a useful setting for problems in applications.

If a Banach space V has a basis (in the sense of Definition 3.12), then it must be separable. indeed, let $\{\boldsymbol{v}_k\}_{k=0}^\infty$ be such a basis, and consider combinations of basis elements but with only *rational* coefficients

$$\sum_{k=0}^{\infty} \alpha_k \ \boldsymbol{v}_k, \qquad \alpha_k \in \mathbb{Q}.$$

The set of such elements is countable (its cardinality is the same as $\mathbb{N}^\mathbb{N}$), and also dense in V. The converse however is trickier. There exists separable Banach spaces that do not have a basis in the sense of Definition 3.12, but those are fairly esoteric and not of interest here.

It is possible to show (Exercise 3.1) that $\ell^\infty(\mathbb{N})$ is not separable, and since a space with a basis must be separable, then $\ell^\infty(\mathbb{N})$ cannot have a basis, and we need not search for alternatives to the basis $\{\boldsymbol{e}_k\}$ for ℓ^∞. On the other hand, its closed subspace $\ell^\infty_0(\mathbb{N})$ is separable since e (3.4) is indeed a basis for it.

Example 3.18. *(Finite-Power Signals and Almost-Periodic Functions)*
Consider the following bilinear form defined for functions on the real line

$$\langle u\,,\, w\rangle_{\mathsf{p}} := \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} u^*(t) w(t)\ dt, \qquad \|u\|^2_{\mathsf{p}} := \langle u\,,\, u\rangle_{\mathsf{p}} = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} |u(t)|^2\ dt. \quad (3.7)$$

In signal analysis, $\|.\|_{\mathsf{p}}$ is the Root Mean Square (RMS) value of a signal[4], and signals for which $\|.\|_{\mathsf{p}}$ is finite are called "finite power signals". Signals for which $\|.\|_{\mathsf{p}} \neq 0$ must be "persistent", i.e. not decay. For example, if $u \in \mathsf{L}^2(\mathbb{R})$, then $\|u\|_{\mathsf{p}} = 0$. Therefore $\langle .\,,\, .\rangle_{\mathsf{p}}$ is not quite an inner product since it is not definite. However, we can make it into a definite inner product by considering the space of equivalence classes with respect to this indefinite norm, so that $u \sim w \ \Leftrightarrow\ \|u - w\|_{\mathsf{p}} = 0$. More precisely, let

$$\mathsf{P}' := \left\{ u : \mathbb{R} \to \mathbb{R};\ \langle u\,,\, u\rangle_{\mathsf{p}} < \infty \right\}, \qquad \mathsf{N} := \left\{ u \in \mathsf{P}';\ \langle u\,,\, u\rangle_{\mathsf{p}} = 0 \right\}.$$

[4] In the definition, lim sup should be used instead of lim to guarantee the existence of a limit. For some exotic signals, the limit above with $T \to \infty$ may oscillate and not converge, even if it is bounded. We write lim here instead of lim sup for simplicity of notation.

Thus N contains all signals equivalent to the zero signal, and two signals are equivalent $u \sim w$ iff $(u - w) \in \mathsf{N}$. It is possible to show (recall Exercise 2.4) that N is a subspace of P′, and that the inner product $\langle \cdot , \cdot \rangle_{\mathsf{p}}$ is definite on the quotient space

$$\mathsf{P} \;:=\; \mathsf{P}'/\mathsf{N}.$$

Thus P is a Hilbert space[5] of *equivalence classes of finite-power signals*.

We will show that this Hilbert space is not separable! Before we do that, we explain the connection with what are called "almost-periodic functions". Consider signals of the form

$$\alpha_1 e^{j\omega_1 t} \;+\; \cdots \;+\; \alpha_n e^{j\omega_n t}, \tag{3.8}$$

where the frequencies $\omega_1, \ldots, \omega_n$ can be any real numbers, i.e. not necessarily commensurate, and therefore such signals are not necessarily periodic, but they are "almost-periodic" in a sense described in Exercise 3.2. The collection of all such signals is a vector space. Also by this definition, the uncountable set of functions $\left\{ e^{j\omega t};\ \omega \in \mathbb{R} \right\}$ is a Hamel basis for this vector space. It can be shown [3, page 109] that the completion of this vector space with the norm (3.7) is exactly the space P defined above. Furthermore, there is a Parseval-type relation as follows

$$u \in \mathsf{P} \quad \Leftrightarrow \quad u(t) = \sum_{k=0}^{\infty} \alpha_k e^{j\omega_k t}, \qquad \|u\|_{\mathsf{p}} := \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} |u(t)|^2 dt \;= \sum_{k=0}^{\infty} |a_k|^2 \;<\; \infty.$$

Thus for every $u \in \mathsf{P}$, there exists a countable set of "frequencies" $\{\omega_k\}$ such that u can be written as a trigonometric series. There are other classes of "almost-periodic functions" that can be defined by taking closures of signals of the form (3.8) in various norms (e.g. with the supremum norm or an "average" L^1 norm). However, the power norm appears to be the most useful one due to the above Parseval-type identity.

Now for separability. Observe that the Hamel basis described earlier is actually an uncountable *orthonormal* set since

$$\left\langle e^{j\omega t}, e^{j\gamma t} \right\rangle_{\mathsf{p}} \;= \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} e^{-j\omega t} e^{j\gamma t}\, dt \;= \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} e^{j(\omega-\gamma)t}\, dt \;=\; \begin{cases} 1, & \omega = \gamma \\ 0, & \omega \neq \gamma \end{cases}.$$

The existence of an uncountable, orthonormal set implies that the space is not separable. The argument is similar to that of Exercise 3.1, which can be informally described as follows: put an open ball of radius $< 1/2$ around each element of the orthonormal set (i.e. at the "tip" of each orthonormal vector). These balls do not intersect. Any dense subset must have at least one element in each ball. Since the number of balls is uncountable, any dense subset must also be uncountable.

3.4 Quotient Spaces and Minimum Distance Problems

In Section 1.4 we saw how to define quotient spaces in a general vector space. The construction was purely algebraic. Now that we have spaces equipped with norms, inner products and notions of completeness, we study quotient spaces with all those extra structures.

The first question is given a normed vector space V and a subspace $\mathsf{S} \subsetneq \mathsf{V}$, how does one define a norm on the quotient space V/S? Figure 3.4 gives a motivation for the definition to follow. Considering a coset $v + \mathsf{S}$, its "norm" $\|v + \mathsf{S}\|$ should be its "distance" from the zero

[5] In the mathematics literature, this construction with with the more general $p \in [1, \infty)$ in place of 2 in (3.7) is called a "Besicovitch space".

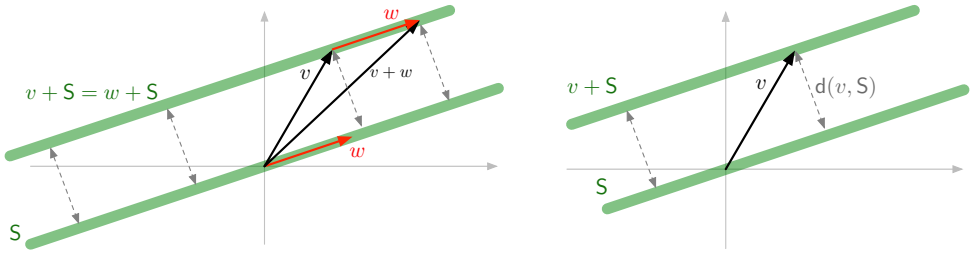


Figure 3.4: *(Left)* The definition of the norm of a cost $v+\mathsf{S}$ as the distance between the two cosets $v+\mathsf{S}$ and $0+\mathsf{S}$. The distance between two sets is defined as the infimum of the distance between all possible elements of the two sets respectively (3.9). *(Right)* This definition is independent of the choice of coset representative v, and is the same as the distance $\mathsf{d}(v,\mathsf{S})$ between v and the subspace S (3.10).

coset $0+\mathsf{S}$. This norm should of course be independent of the choice of coset representative v. We can therefore attempt a definition as follows

$$\|v+\mathsf{S}\| \;:=\; \inf_{w,u\in\mathsf{S}} \| \underbrace{v+w}_{\substack{\text{all possible}\\ \text{members}\\ \text{of } v+\mathsf{S}}} - \underbrace{u}_{\substack{\text{all possible}\\ \text{members}\\ \text{of } 0+\mathsf{S}}} \|. \tag{3.9}$$

This quantity captures the distance between the two cosets. The expression can be simplified a bit since we can combine $w,u\in\mathsf{S}$ as a single parameter $x=w-u$ and rewrite

$$\|v+\mathsf{S}\| \;:=\; \inf_{x\in\mathsf{S}} \|v+x\| \;=\; \inf_{x\in\mathsf{S}} \|v-x\| \;=\; \mathsf{d}(v,\mathsf{S}), \tag{3.10}$$

where $\mathsf{d}(v,\mathsf{S})$ is the minimum distance between the vector v and the subspace S. Note again that this distance is the same for any vector in the same coset that v is in. More precisely

$$v_1-v_2\in\mathsf{S} \qquad\Rightarrow\qquad \mathsf{d}(v_1,\mathsf{S}) \;=\; \mathsf{d}(v_2,\mathsf{S}).$$

Thus the definition (3.10) is independent of the choice of coset representative.

Now having defined the quantity (3.10), is it really a norm on the set of cosets? We need to check homogeneity, the triangle inequality and definiteness which we do as follows.

- *Homogeneity:* The coset $\alpha(v+\mathsf{S})$ is by definition $\alpha v+\mathsf{S}$ and therefore for $\alpha\neq 0$

$$\|\alpha(v+\mathsf{S})\| \;=\; \inf_{x\in\mathsf{S}} \|\alpha v+x\| \;=\; \inf_{x\in\mathsf{S}} |\alpha|\,\|v+x/\alpha\| \;\overset{1}{=}\; |\alpha| \inf_{y\in\mathsf{S}} \|v+y\| \;=\; |\alpha|\|(v+\mathsf{S})\|,$$

  where $\overset{1}{=}$ follows from S being a subspace, and therefore $y=x/\alpha\in\mathsf{S} \;\Leftrightarrow\; x\in\mathsf{S}$.

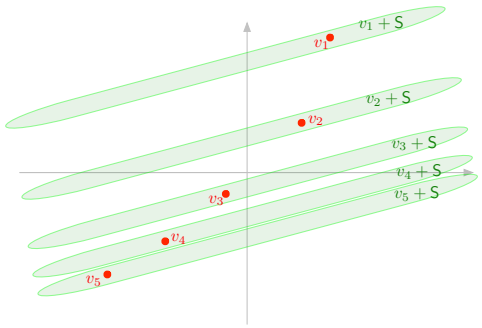
- *Triangle Inequality:* This follows from the triangle inequality for vectors

$$\begin{aligned}\|(v+\mathsf{S})+(w+\mathsf{S})\| \;&=\; \|(v+w)+\mathsf{S}\| \;=\; \inf_{x\in\mathsf{S}} \|v+w-x\| \;=\; \inf_{x,y\in\mathsf{S}} \|v+w-x-y\| \\ &\leq\; \inf_{x,y\in\mathsf{S}} (\|v-x\|+\|w-y\|) \;=\; \inf_{x\in\mathsf{S}} \|v-x\| + \inf_{y\in\mathsf{S}} \|w-y\|\,.\end{aligned}$$

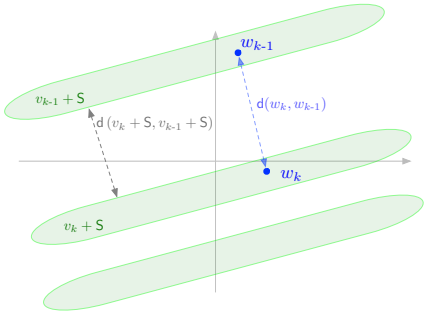
- *Definiteness:* Here we require an additional condition on the subspace S. If S is not closed, and $v\notin\mathsf{S}$ is in its closure $\overline{\mathsf{S}}$, then the minimum distance is zero

$$v\in\overline{\mathsf{S}} \qquad\Rightarrow\qquad \inf_{x\in\mathsf{S}} \|v-x\| \;=\; 0.$$

  Conversely, if the minimum distance $\|v+\mathsf{S}\|=0$, then v must be in the closure of S. Thus if S is closed, then the norm (3.10) on the quotient space V/S is definite. Note that in finite dimensions every subspace is closed, and in this case we do not require this additional assumption.



(a) A sequence $\{(v_k + \mathsf{S})\}_{k=0}^{\infty}$ of cosets can be Cauchy, while a sequence $\{v_k\}_{k=0}^{\infty}$ of their representatives is not.



(b) Constructing a Cauchy sequence of representatives for a Cauchy sequence of cosets. One can always choose $w_k \in v_k + \mathsf{S}$ and $w_{k\text{-}1} \in v_{k\text{-}1} + \mathsf{S}$ such that $\mathsf{d}(w_k, w_{k\text{-}1})$ is arbitrarily close to the distance between the two cosets as in (3.12).

Figure 3.5: Graphical depiction of the proof of completeness of the quotient space V/S of Lemma 3.19.

Thus with the norm defined in (3.10), the quotient space V/S becomes a normed space (provided S is closed). We can ask the next question which is about completeness.

Lemma 3.19. *Let S be a closed subspace of a Banach space V. Then the quotient space V/S endowed with the norm*

$$\|v + \mathsf{S}\| \;:=\; \inf_{x \in \mathsf{S}} \|v - x\|, \tag{3.11}$$

is itself a Banach space.

Proof. We have already shown that the norm (3.11) is independent of the choice v of coset representative, and satisfies the three requirements of a norm. It remains to show completeness of V/S. Take a Cauchy "sequence of cosets" $\{v_k + \mathsf{V}\}$, i.e.

$$\forall \epsilon,\ \exists N,\ \forall n, m \geq N, \quad \|(v_n + \mathsf{S}) - (v_m + \mathsf{S})\| \leq \epsilon.$$

If the sequence of coset representatives $\{v_k\}$ were Cauchy, then its limit can be used to find the limit coset. However, the difficulty illustrated in Figure 3.5a is that while the sequence of cosets is Cauchy, one can choose a sequence of representatives that are not themselves Cauchy. It is however easy to construct another sequence $\{w_k\}$ with distances no larger than the coset distances, and therefore $\{w_k\}$ will indeed be Cauchy. The construction is illustrated in Figure 3.5b

$$\begin{aligned} w_0 \;&:=\; v_0 \\ \text{choose } w_k \text{ s.t.} \quad \mathsf{d}(w_k, w_{k\text{-}1}) \;&\leq\; \mathsf{d}(v_k + \mathsf{S}, v_{k\text{-}1} + \mathsf{S}) + \gamma_k \qquad \text{(always possible)} \end{aligned} \tag{3.12}$$

where $\{\gamma_k\}$ is a sequence of positive numbers decaying to zero. Since the increments $\{\mathsf{d}(v_k + \mathsf{S}, v_{k\text{-}1} + \mathsf{S})\}$ and $\{\gamma_k\}$ are both Cauchy sequences, then so are the increments $\{\mathsf{d}(w_k, w_{k\text{-}1})\}$, and therefore $\{w_k\}$ is a Cauchy sequence. Thus $\lim_{k\to\infty} w_k =: \bar{w} \in \mathsf{V}$ exists since V is complete, and the coset $\bar{w} + \mathsf{S}$ is the limit of the sequence $\{v_k + \mathsf{S}\}$ in V/S. □

The infimization problem (3.11) deserves some attention. It is a "minimum distance problem" from a vector v to a subspace S. Such problems, and more general ones of finding minimum distances to affine spaces, are important in applications. They are fundamentally

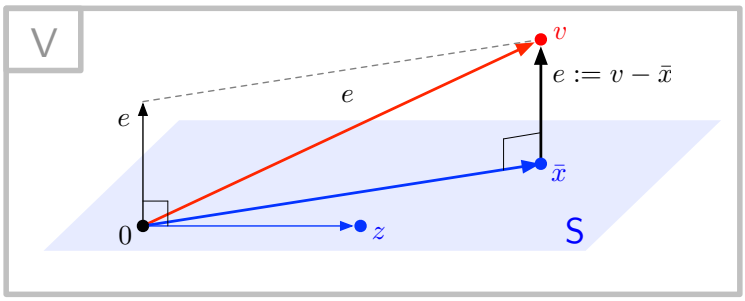


Figure 3.6: *The projection theorem in an inner product space.* The point $\bar{x}$ in a subspace S that is *closest* to a point b outside of it is called the *projection of* v *onto* S. This point is characterized by the "error vector" $e := v - \bar{x}$ being orthogonal to all other vectors $x \in \mathsf{S}$.

problems of approximation or error minimization. The reader should note that the proof of Lemma 3.19 does not provide an algorithm or technique for actually finding this minimum distance. We next consider minimum distance problems in inner product spaces, where the solution is given by an "orthogonal projection". Similar problems in Banach space are more subtle since there is no inner product. None the less, we will be able to address minimum distance problems using the concepts of duality and "orthogonal functionals" in Chapter 4.

3.4.1 The Projection Theorem in Hilbert Space

We begin with the projection theorem in an inner product space (though not necessarily a complete, i.e. Hilbert, space). This theorem states an orthogonality condition that any minimizer must satisfy, and thus it is a *necessary condition for optimality*. The proof is typical of most arguments for such necessary conditions by the contrapositive, namely, if the condition is not satisfied, then the objective can be improved by moving in a certain direction.

Theorem 3.20 (Minimum distance to a subspace of an inner product space)**.** *Let* S *be a subspace of a inner product space* V*, and consider the minimum distance problem between a vector* $v \in \mathsf{V}$ *and the subspace* S *(see Figure 3.6)*

$$\bar{J} \;:=\; \inf_{x \in \mathsf{S}} \|v - x\|. \tag{3.13}$$

If $\bar{x} \in \mathsf{S}$ *solves (3.13) (i.e.* $\|v - \bar{x}\| = \bar{J}$*), then it is unique, and the "optimal error" vector* $(v - \bar{x})$ *is orthogonal to* S

$$\forall x \in \mathsf{S}, \;\; \langle v - \bar{x}\,,\, x \rangle = 0. \tag{3.14}$$

Proof. We will show that if $\bar{x}$ does not satisfy the orthogonality condition, then we can find another point in S whose distance to v is smaller. If $\bar{x}$ doesn't satisfy (3.14), then there is some $\tilde{x} \in \mathsf{S}$ with $\langle b - \bar{x}\,,\, \tilde{x} \rangle \neq 0$. $\tilde{x}$ can be chosen[6] such that this number is positive, i.e. $\langle b - \bar{x}\,,\, \tilde{x} \rangle = c > 0$. Now "move" from $\bar{x}$ in the direction of $\tilde{x}$, and examine the distance to v (see Figure 3.7)

$$\begin{aligned} \|v - (\bar{x} + \epsilon\tilde{x})\|^2 \;&=\; \big\langle v - (\bar{x} + \epsilon\tilde{x})\,,\, v - (\bar{x} + \epsilon\tilde{x}) \big\rangle \;=\; \big\langle (v - \bar{x}) + \epsilon\tilde{x}\,,\, (v - \bar{x}) + \epsilon\tilde{x} \big\rangle \\ &=\; \big\langle v - \bar{x}\,,\, v - \bar{x} \big\rangle \;-\; 2\epsilon\, \langle v - \bar{x}\,,\, \tilde{x} \rangle \;+\; \epsilon^2\, \langle \tilde{x}\,,\, \tilde{x} \rangle \\ &=\; \|v - \bar{x}\|^2 \;-\; 2\epsilon\, c \;+\; \epsilon^2\, \|\tilde{x}\|^2. \end{aligned}$$

[6] If $\langle b - \bar{x}\,,\, \tilde{x} \rangle = c < 0$, then choose $-\tilde{x}$ instead of $\tilde{x}$.

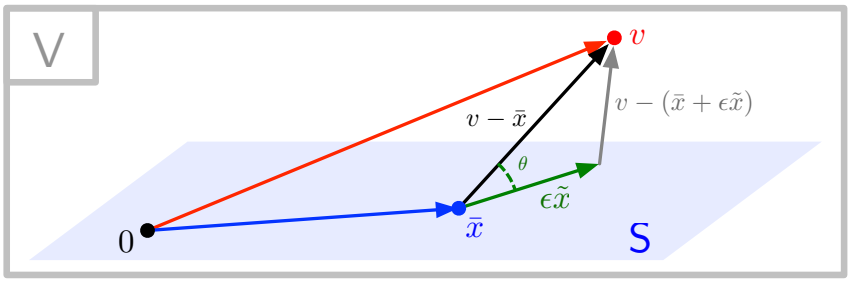


Figure 3.7: The proof of the projection theorem in an inner product space. For any $\bar{x} \in \mathsf{S}$, if $b - \bar{x}$ is not orthogonal to S, then we can select a $\tilde{x} \in \mathsf{S}$ such that $\langle b - \bar{x}\,,\, \tilde{x}\rangle = c > 0$ (i.e. the angle θ above is less than $90°$). We can then move in the direction of $\tilde{x}$ from $\bar{x}$, and choose ϵ such that $b - (\bar{x} + \epsilon\tilde{x})$ has smaller length than $b - \bar{x}$ as shown in (3.15).

The idea is that now we can choose ϵ sufficiently small so that the ϵ term dominates (in magnitude) the ϵ^2 term. Since $c > 0$, the right hand side can be made strictly smaller than the left hand side, i.e. $\exists \bar{\epsilon} > 0$ such that for all $\epsilon < \bar{\epsilon}$

$$\|v - (\bar{x} + \epsilon\tilde{x})\|^2 = \|v - \bar{x}\|^2 - 2c\,\epsilon + \|\tilde{x}\|^2\epsilon^2 < \|v - \bar{x}\|^2, \tag{3.15}$$

and therefore $\bar{x}$ is not optimal.

The argument that a minimizer must be unique is simple. Suppose $\bar{x}_1$ and $\bar{x}_2$ are both minimizers, then they each must satisfy (3.14), and therefore

$$\forall z \in \mathsf{S}, \quad 0 = \langle b - \bar{x}_1\,,\, z\rangle - \langle b - \bar{x}_2\,,\, z\rangle = \langle \bar{x}_2 - \bar{x}_1\,,\, z\rangle\,.$$

The last statement says that $\bar{x}_2 - \bar{x}_2$ must be orthogonal to S, but since $(\bar{x}_2 - \bar{x}_1) \in \mathsf{S}$, it must be zero, and $\bar{x}_2 = \bar{x}_1$. □

The previous theorem was stated in an inner product space because completeness plays no role in the necessary conditions for a minimum. However, the theorem does not say anything about the existence of a minimizer. We will show next that if S is a *closed* subspace of a Hilbert space V, then a minimizer always exists (and therefore must be unique by Theorem 3.20). While this is true in a Hilbert space, it is not true in a Banach space. This fact is intimately tied to the special structure of the norm in an inner product space, and is a consequence of the parallelogram law. Before stating and proving the minimizer existence result, we re-examine the parallelogram law and its implications for distances and the geometry in an inner product space. These observations are of interest in their own right.

Recall the parallelogram law which was stated earlier as

$$\|v + v\|^2 + \|u - v\|^2 = 2\|u\|^2 + 2\|v\|^2,$$

for any two vectors u and v in an inner product space. It can be restated as follows

$$\|u - v\|^2 = 2\|u\|^2 + 2\|v\|^2 - 4\|(u + v)/2\|^2, \tag{3.16}$$

which has an interesting geometric interpretation illustrated in Figure 3.8. The points u, v, and $(u+v)/2$ are *colinear*, with the segment $u-v$ bisected by $(u+v)/2$. The relation (3.16) implies that if the vectors u, v, and $(u + v)/2$ are of equal length, then $u - v$ must have length zero! It also implies that if they were approximately equal, say to within order ϵ of each other, then $u - v$ will have length of order ϵ as well. This is in sharp contrast with the geometry of a general normed space. Figure 3.8 shows an example from $\mathbb{R}^2$ with the

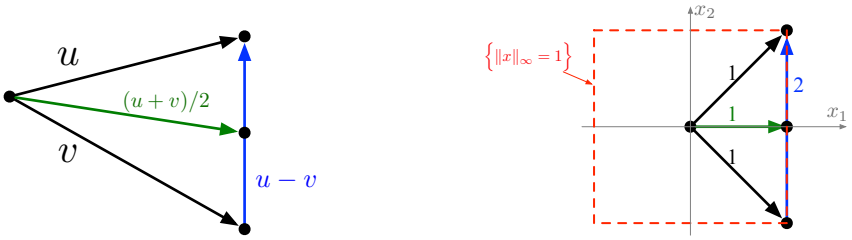


Figure 3.8: (*Left*) The restated parallelogram law (3.16) $\|u-v\|^2 = 2\|u\|^2 + 2\|v\|^2 - 4\|(u+v)/2\|^2$ constrains the length of $u-v$ in terms of the other three lengths. If u, v and $(u+v)/2$ have equal lengths, then $u-v$ must have zero length and the three points coalesce. Also, if u, v and $(u+v)/2$ have approximately equal lengths, then $u-v$ is forced to have small length of the order of the *differences* between the three lengths. (*Right*) This stands in sharp contrast to the geometry in a general normed space as this example in $\mathbb{R}^2$ with the ∞-norm illustrates.

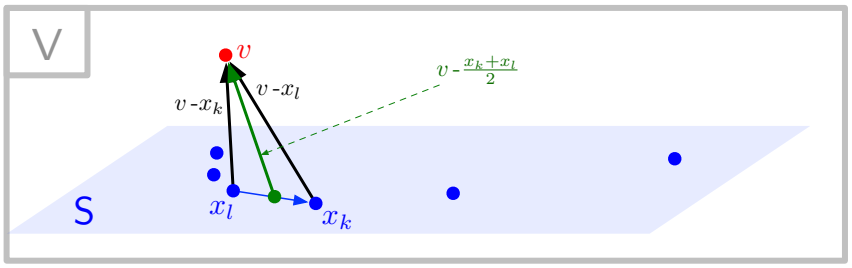


Figure 3.9: Due to the geometrical restrictions encoded in the parallelogram law in an inner product space, a sequence minimizing the distance to a subspace must be Cauchy. When the lengths of $v-x_k$ and $v-x_l$ are close to the optimum $\bar{J}$, and the length of the distance from v to the midpoint $(x_k+x_l)/2$ can't be smaller than $\bar{J}$, then the parallelogram law (3.19) implies that x_k must be close to x_l. Thus the sequence $\{x_n\}$ is Cauchy.

$\|.\|_\infty$ norm where three co-linear points of the same configuration have equal lengths, but the segment $u-v$ between them has twice their length! We will see in the next theorem that it is this geometry encoded in the parallelogram law that forces a minimizing sequence to be a Cauchy sequence.

Theorem 3.21 (Projection Theorem)**.** *Let* S *be a* closed *subspace of a Hilbert space* V*. For any element* $v \in \mathsf{V}$*, there exists a unique minimizer* $\bar{x}$ *of the distance between* v *and* S

$$\inf_{x\in\mathsf{S}} \|v-x\| \;=\; \|v-\bar{x}\|.$$

This minimizer satisfies the orthogonality condition

$$\forall x \in \mathsf{S},\;\; \langle v-\bar{x}\,,\, x\rangle = 0. \tag{3.17}$$

Proof. Uniqueness and the necessity of the orthogonality condition has already been shown in Theorem 3.20. To show existence, let $\{z_n\} \subset \mathsf{S}$ be a sequence that achieves the infimum

$$\lim_{n\to\infty} \|v-x_n\| \;=\; \inf_{x\in\mathsf{S}} \|v-z\| \;=:\; \bar{J}. \tag{3.18}$$

We will show that $\{x_n\}$ is a Cauchy sequence. It will then follow that since S is closed, this sequence must have a limit in S.

Now to show that $\{x_n\}$ is Cauchy, we apply version (3.16) of the parallelogram law to the vectors joining v to two points x_k and x_l in the sequence respectively (see Figure 3.9)

$$\|x_k-x_l\|^2 \;=\; 2\,\|v-x_k\|^2 + 2\,\|v-x_l\|^2 - 4\left\|v-\tfrac{x_k+x_l}{2}\right\|^2. \tag{3.19}$$

Now if those two points are chosen from the "tail" of the sequence, i.e. if $\|v - x_k\|$ and $\|v - x_l\|$ are within ϵ of the infimum $\bar{J}$, then $\|x_k - x_l\|$ is forced to be of order ϵ by (3.19). More precisely, given ϵ, there exists an N such that for all $k, l \geq N$

$$\left.\begin{array}{r} \|b - z_k\| \leq \bar{J} + \epsilon \\ \|b - z_l\| \leq \bar{J} + \epsilon \\ \left\|b - \frac{z_k + z_l}{2}\right\| \geq \bar{J} \end{array}\right\} \quad \Rightarrow \quad \left\{\begin{array}{ll} \|z_k - z_l\|^2 & \leq 2(\bar{J} + \epsilon)^2 + 2(\bar{J} + \epsilon)^2 - 4\bar{J}^2 \\ & = 8\bar{J}\ \epsilon + 4\ \epsilon^2 \end{array}\right.$$

The first two statements on the left follow from (3.18), and the third statement follows from $(x_k + x_l)/2$ being in S, and again by (3.18), the distance $\|v - (x_k + x_l)/2\|$ cannot be smaller than $\bar{J}$. Thus $\|x_k - x_l\|$ can be made as small as desired for all $k, l \geq N$, which implies that $\{x_n\}$ is a Cauchy sequence. □

The Orthogonal Complement and Orthogonal Projection

Definition 3.22. *Given any subspace $\mathsf{S} \subset \mathsf{V}$ of an inner product space V, its* orthogonal complement $\mathsf{S}^\perp$ *is*

$$\mathsf{S}^\perp \ := \ \{v \in \mathsf{V};\ \forall x \in \mathsf{S},\ \langle v\,,\, x\rangle = 0\}\,,$$

i.e. the set of vectors that are perpendicular to all vectors in S.

The fact that $\mathsf{S}^\perp$ is closed under additions and scalings, and therefore a subspace itself, is immediate from the definitions. In fact, in a Hilbert space $\mathsf{S}^\perp$ is a closed subspace even if S is not. Indeed, let $\{v_k\}$ be a Cauchy sequence in $\mathsf{S}^\perp \subset \mathsf{V}$, where V is a Hilbert space. The sequence has a limit $\lim_{k\to\infty} v_k = v \in \mathsf{V}$ (since V is complete), but this limit must also belong to $\mathsf{S}^\perp$ since

$$\forall x \in \mathsf{S},\ \ \langle v\,,\, x\rangle \ = \ \lim_{k\to\infty} \langle v_k\,,\, x\rangle \ = \ 0,$$

due to the continuity of the inner product.

The projection theorem 3.21 has the following important implication. Given a closed subspace $\mathsf{S} \subset \mathsf{V}$, any vector $v \in \mathsf{V}$ can be uniquely written as the sum

$$v \ = \ v_1 + v_2, \qquad v_1 \in \mathsf{S},\ v_2 \perp \mathsf{S}.$$

Indeed, v_1 is the unique minimizer $\bar{x} \in \mathsf{S}$ of the minimum distance problem between v and S, and v_2 is the optimal error vector $v_2 := v - \bar{x} = v - v_1$ in Theorem 3.21. Furthermore, orthogonality of v_1 and v_2 implies the Pythagorean identity $\|v\|^2 = \|v_1\|^2 + \|v_2\|^2$. This is stated formally next.

Lemma 3.23. *Let $\mathsf{S} \subset \mathsf{V}$ be a closed subspace of a Hilbert space. Then its orthogonal complement $\mathsf{S}^\perp$ is complementary to S in V, i.e. $\mathsf{V} = \mathsf{S} \oplus \mathsf{S}^\perp$ meaning every element $v \in \mathsf{V}$ can be written uniquely as*

$$v \ = \ v_1 + v_2, \qquad v_1 \in \mathsf{S},\ v_2 \in \mathsf{S}^\perp, \qquad \textit{with}\ \ \|v\|^2 = \|v_1\|^2 + \|v_2\|^2\,.$$

Recall that by Lemma 1.30, the decomposition $\mathsf{V} = \mathsf{S} \oplus \mathsf{S}^\perp$ implies that there are projection operators $\Pi : \mathsf{V} \to \mathsf{S}$ and $(I - \Pi) : \mathsf{V} \to \mathsf{S}^\perp$ onto S and $\mathsf{S}^\perp$ respectively. Since S and $\mathsf{S}^\perp$ are orthogonal complements, we call those projections *orthogonal projections*.

What happens if we take orthogonal complements twice? For any subspace $\mathsf{S} \subset \mathsf{V}$, denote $\left(\mathsf{S}^\perp\right)^\perp =: \mathsf{S}^{\perp\perp}$, and observe that any vector in S is orthogonal to all of $\mathsf{S}^\perp$

$$x \in \mathsf{S} \qquad \Rightarrow \qquad \forall w \in \mathsf{S}^\perp,\ \ \langle w\,,\, x\rangle = 0 \qquad \Rightarrow \qquad x \in \mathsf{S}^{\perp\perp}.$$

This means that $\mathsf{S} \subseteq \mathsf{S}^{\perp\perp}$. We can go one step further since orthogonal complements must be closed, and thus conclude that $\overline{\mathsf{S}} \subseteq \mathsf{S}^{\perp\perp}$. In fact, the two subspaces are equal.

Lemma 3.24. *Let $\mathsf{S} \subset \mathsf{V}$ be a (not necessarily closed) subspace of a Hilbert space V. Then $\mathsf{S}^{\perp\perp} = \overline{\mathsf{S}}$.*

Proof. It remains to show that $\mathsf{S}^{\perp\perp} \subseteq \overline{\mathsf{S}}$, which we do by the contrapositive. If $v \notin \overline{\mathsf{S}}$, then it must be a non-zero distance away from it, and by the projection theorem, the optimal error $v - \bar{x}$ is such that

$$\begin{aligned} 0 \;<\; \inf_{x\in\mathsf{S}} \|v-x\|^2 \;&=\; \|v-\bar{x}\|^2 \;=\; \langle v-\bar{x}\,,\, v-\bar{x}\rangle \;=\; \langle v-\bar{x}\,,\, v\rangle - \langle v-\bar{x}\,,\, x\rangle \\ &=\; \langle v-\bar{x}\,,\, v\rangle \qquad\qquad (x\in\mathsf{S} \;\Rightarrow\; \langle v-\bar{x}\,,\, x\rangle = 0) \end{aligned}$$

Thus we found one vector $\bar{w} := v - \bar{x} \in \mathsf{S}^{\perp}$ with $\langle \bar{w}\,,\, v\rangle \neq 0$. This implies $v \notin \left(\mathsf{S}^{\perp}\right)^{\perp}$. □

3.5 Continuity and Induced Norms of Linear Mappings

In previous chapters we dealt with linear operators in a purely algebraic manner. Now that we have notions of topology, distances and convergence on vector spaces, we can study further properties of operators. Those properties are induced from the norms on the vector spaces. It turns out that since we are dealing with *linear* mappings, the superposition property will have two important consequences. First that it suffices to check continuity at the origin, and second, that continuity puts an upper bound on how much the operator can "amplify" the norm of a vector. This last statement will be formalized in terms of the *induced norm* of the operator.

A linear operator $A : \mathsf{V} \to \mathsf{W}$ between two normed spaces V and W is a mapping, and we can define continuity of the mapping using the underlying norms in V and W. First we define a stronger type of continuity than what we've already encountered

Definition 3.25. *Let $f : \mathsf{M}_1 \to \mathsf{M}_2$ be a mapping between two metric spaces. f is called* uniformly Lipschitz continuous *(or simply* Lipschitz continuous[7]*) if there exists one constant $\bar{c}$ such that for all $x_1, x_2 \in \mathsf{M}_1$*

$$\mathsf{d}\big(f(x_1), f(x_2)\big) \;\leq\; \bar{c}\;\mathsf{d}(x_1, x_2). \tag{3.20}$$

Lipschitz continuity is stronger than notions of continuity in Definition 3.6. For example, it is easy to see that (3.20) implies (3.3), but the reverse may not hold in general. However, for *linear mappings*, the stronger Lipschitz continuity condition turns out to be equivalent to continuity in the sense of Definition 3.6. Thus in a normed *vector space*, all four definitions are equivalent. The proof of this equivalence is part of the next lemma.

Lemma 3.26. *Let $A : \mathsf{V} \to \mathsf{W}$ be a linear operator between two normed spaces.*

1. *A is a continuous mapping iff it is continuous at $0 \in \mathsf{V}$.*
2. *A is continuous iff the following ratio is "uniformly bounded"*

$$\|A\|_{\mathsf{i}} \;:=\; \sup_{0\neq v\in\mathsf{V}} \frac{\|Av\|_{\mathsf{W}}}{\|v\|_{\mathsf{V}}} \;=\; \bar{c} \;<\; \infty. \tag{3.21}$$

The quantity $\|A\|_{\mathsf{i}}$ is called the induced norm *of A.*

[7] The standard definition of Lipschitz continuity allows the constant $\bar{c}$ (the so-called Lipschitz constant) in (3.20) to vary with x_1 and x_2. We will not need this weaker concept of continuity here, and will simply refer to (3.20) as Lipschitz continuity.

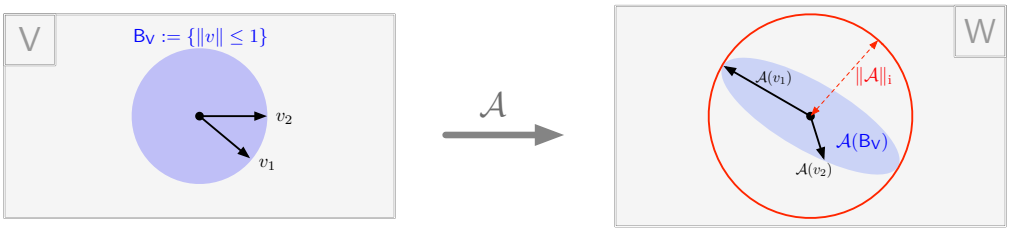


Figure 3.10: An illustration of the induced norm $\|\mathcal{A}\|_\mathsf{i}$ of an operator $\mathcal{A} : \mathsf{V} \to \mathsf{W}$ between normed vector spaces. The vectors v_1 and v_2 lie in the unit ball B_V of V, and they are mapped to the vectors $\mathcal{A}(v_1)$ and $\mathcal{A}(v_2)$ respectively. The operator $\mathcal{A}$ can amplify or shorten the length of these vectors depending on which direction they are in. The highest "amplification" is given by the induced norm $\|\mathcal{A}\|_\mathsf{i}$, which gives the tightest bound in the inequality $\|\mathcal{A}(v)\|_\mathsf{W} \le \|A\|_\mathsf{i}\|v\|_\mathsf{V}$. The bound implies that the image $\mathcal{A}(\mathsf{B}_\mathsf{V})$ of the unit ball is contained inside a ball of radius $\|\mathcal{A}\|_\mathsf{i}$ in W. In the diagram above, the vector v_1 achieves this bound with equality, thus it is the vector with the highest norm amplification, and achieves the supremum in (3.21).

Before proving the lemma, we examine the ratio (3.21). The homogeneity properties of both the norm and the linear operator A imply that this ratio can be written in several equivalent forms

$$\begin{aligned}\sup_{0\neq v\in\mathsf{V}} \frac{\|A(v)\|}{\|v\|} &= \sup_{0\neq v\in\mathsf{V}} \left\|\frac{1}{\|v\|}A(v)\right\| = \sup_{0\neq v\in\mathsf{V}} \left\|A\left(\frac{v}{\|v\|}\right)\right\| \overset{1}{=} \sup_{\|u\|=1} \|Au\| \\ &= \sup_{\|u\|\le 1} \|Au\| = \frac{1}{\alpha}\sup_{\|w\|\le\alpha} \|Aw\|. \end{aligned} \tag{3.22}$$

The equality $\overset{1}{=}$ follows by observing that $u := v/\|v\|$ is always a vector of unit norm. The remaining equalities follow from the the homogeneity of the norm and the operator A.

Another consequence of (3.21) is a bound relating the norm of a vector to that of its image under A

$$\begin{aligned}\sup_{0\neq v\in\mathsf{V}} \frac{\|Av\|_\mathsf{W}}{\|v\|_\mathsf{V}} =: \|A\|_\mathsf{i} \quad &\Rightarrow \quad \forall v \in \mathsf{V},\ \frac{\|Av\|_\mathsf{W}}{\|v\|_\mathsf{V}} \le \|A\|_\mathsf{i} \\ &\Leftrightarrow \quad \forall v \in \mathsf{V},\ \|Av\|_\mathsf{W} \le \|A\|_\mathsf{i}\,\|v\|_\mathsf{V}. \end{aligned} \tag{3.23}$$

Note that $Av \in \mathsf{W}$ is the image of $v \in \mathsf{V}$ under the operator A. The last inequality means that the induced operator norm $\|A\|_\mathsf{i}$ bounds how large the norm of any vector v can be "amplified" by the operator A. Note the careful labeling of the norms in (3.23) to indicate the respective spaces in which they are measured. Figure 3.10 gives a graphical illustration of the concept of the induced norm of an operator as a measure of how it amplifies or shrinks norms of vectors in different directions.

Proof of Lemma 3.26. The first clause is a consequence of the metric being translation invariant in a normed vector space. Since the metric is defined in terms of the norm, continuity at zero has the following implications for the norm

$$\begin{aligned}\Big(\mathsf{d}(0,\tilde{v}) := \|0-\tilde{v}\| = \|\tilde{v}\| \le \delta \quad &\Rightarrow \quad \mathsf{d}\big(A(0),A(\tilde{v})\big) := \|0-A(\tilde{v})\| = \|A(\tilde{v})\| \le \epsilon\Big) \\ \Leftrightarrow \quad \Big(\|\tilde{v}\| \le \delta \quad &\Rightarrow \quad \|A(\tilde{v})\| \le \epsilon\Big). \end{aligned} \tag{3.24}$$

On the other hand, continuity at any other point $\bar{v} \in \mathsf{V}$ means

$$\|\bar{v} - v\| \le \delta \quad \Rightarrow \quad \|A(\bar{v}) - A(v)\| \le \epsilon. \tag{3.25}$$

Since A is linear, we have $\|A(\bar{v}) - A(v)\| = \|A(\bar{v} - v)\|$, and therefore (3.24) implies (3.25) by choosing $\tilde{v} = \bar{v} - v$.

To prove the second clause, we first point out that the uniform boundedness condition is equivalent to uniform Lipschitz continuity. Indeed, since the metric is given by the norm, (3.20) becomes

$$\begin{aligned} \|A(v_1) - A(v_2)\| \;\leq\; \bar{c}\,\|v_1 - v_2\| \qquad &\Leftrightarrow \qquad \|A(v_1 - v_2)\| \;\leq\; \bar{c}\,\|v_1 - v_2\| \\ &\Leftrightarrow \qquad \|A(v)\| \;\leq\; \bar{c}\,\|v\| \qquad \Leftrightarrow \qquad \frac{\|Av\|}{\|v\|} \;\leq\; \bar{c}, \end{aligned}$$

which holds for any non-zero $v \in \mathsf{V}$. Thus the norm ratio is uniformly bounded by the Lipschitz constant $\bar{c}$. The converse also follows since the bound $\|A\|_{\mathrm{i}}$ in (3.21) gives the Lipschitz constant. Thus boundedness (3.21) implies Lipschitz continuity, which in turn implies continuity.

For the converse, assume A is continuous, its continuity at $v = 0$ means that given $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\Big(\forall \|v\| \leq \delta, \quad \|Av\| \;\leq\; \epsilon\Big) \qquad \Rightarrow \qquad \|A\|_{\mathrm{i}} \;=\; \frac{1}{\delta} \sup_{\|v\| \leq \delta} \|Av\| \;\leq\; \frac{\epsilon}{\delta},$$

where we have used the expression (3.22) for the induced norm. Therefore continuity at zero implies that A has bounded induced norm. □

For matrices, the supremum in (3.21) is always finite, thus linear mappings between finite-dimensional vector spaces are always Lipschitz continuous. Any given matrix will have different induced norms depending on the choice of vector norm in $\mathbb{R}^n$. We will shortly see several examples. On the other hand, not every operator on infinite-dimensional normed spaces will have a bounded induced norm. In fact, many operators of interest, such as differential operators, will typical have an unbounded supremum in (3.21). We will see examples of this as well.

Example 3.27. Let $A : \mathbb{R}^m \to \mathbb{R}^n$ be an $n \times m$ matrix. If $A = \mathrm{diag}(a_1, \ldots, a_n)$ is diagonal and square[8] (i.e. $n = m$), then its p-induced norm (written $\|A\|_{p\text{-i}}$) is the same for all $p \in [1, \infty]$, and is the maximum modulus of all the diagonal entries

$$\|A\|_{p\text{-i}} \;:=\; \sup_{\|v\|_p \leq 1} \|Av\|_p \;=\; \max_{1 \leq k \leq n} |a_k|\,, \qquad p \in [1, \infty]. \tag{3.26}$$

This is easy to show from the definition. We illustrate the argument here as it is typical of most induced norm calculations. First a bound is derived, and then one shows that the bound is tight by exhibiting a vector that achieves (or almost achieves) this bound. Let v be any vector and calculate that for a diagonal A

$$\begin{aligned} \|Av\|_p^p \;=\; \|(a_1 v_1, \ldots, a_n v_n)\|^p \;=\; \sum_{k=1}^n |a_k v_k|^p \;&\leq\; \Big(\max_{1 \leq k \leq n} |a_k|^p\Big) \sum_{k=1}^n |v_k|^p \\ &=\; \Big(\max_{1 \leq k \leq n} |a_k|\Big)^p \|v\|_p^p. \end{aligned} \tag{3.27}$$

By taking the p'th root of both sides we see that the quantity (3.26) bounds the p-induced norm of A. To find a vector that achieves this bound, let $\bar{k}$ be an index where the maximum

[8] If A is not square, but rather a 2×1 or a 1×2 block matrix, with a diagonal block and a zero block, then the same statement holds.

in (3.26) is achieved, and choose $\bar{v}$ to have all zero entries except for an entry of 1 at $\bar{k}$. Note that $\|\bar{v}\| = 1$ for any p, and calculate

$$\|A\bar{v}\|_p \;=\; \|(0,\ldots,0,a_{\bar{k}}1,0,\ldots,0)\|_p \;=\; |a_{\bar{k}}| \;:=\; \max_{1\le k\le n} |a_k|.$$

Thus the bound (3.27) is tight.

Example 3.28. The 1-induced and ∞-induced norm of A are given by the "maximum column sum" and "maximum row sum" respectively

$$\|A\|_{1\text{-i}} \;=\; \max_{1\le l\le m} \sum_{k=1}^{n} |a_{kl}|\,, \qquad\qquad \|A\|_{\infty\text{-i}} \;=\; \max_{1\le k\le m} \sum_{l=1}^{n} |a_{kl}|\,. \tag{3.28}$$

This is a consequence of the $1-\infty$ inequality of Exercise 2.5, and is itself left as an exercise. In contrast to the 1 and ∞ induced norms, there is not such a simple expression for the other p-induced norm in terms of the matrix entries.

Example 3.29. The 2-induced norm of A is its maximum singular value

$$\|A\|_{2\text{-i}} \;=\; \sigma_{\max}(A)\,.$$

We will prove this statement when we introduce the singular value decomposition later on. Note that in contrast to the direct computations required to obtain the 1 and ∞ induced norms (3.28), the 2-induced norm requires the more substantial calculation of the maximum sigular value.

Example 3.30. The bounds between p norms in $\mathbb{R}^n$ can be used to derive bounds between the respective p-induced norms. Upper and lower bounds between two vector norms can be used to bound the induced norms as follows

$$\underline{c}\,\|v\|_{\text{b}} \;\le\; \|v\|_{\text{a}} \;\le\; \bar{c}\,\|v\|_{\text{b}} \qquad\Leftrightarrow\qquad 1/\left(\bar{c}\,\|v\|_{\text{b}}\right) \;\le\; 1/\|v\|_{\text{a}} \;\le\; 1/\left(\underline{c}\,\|v\|_{\text{b}}\right)$$

$$\frac{\underline{c}}{\bar{c}}\,\|A\|_{\text{b-i}} := \frac{\underline{c}}{\bar{c}}\,\sup_{v\neq 0}\frac{\|Av\|_{\text{b}}}{\|v\|_{\text{b}}} \;\le\; \|A\|_{\text{a-i}} := \sup_{v\neq 0}\frac{\|Av\|_{\text{a}}}{\|v\|_{\text{a}}} \;\le\; \frac{\bar{c}}{\underline{c}}\,\sup_{v\neq 0}\frac{\|Av\|_{\text{b}}}{\|v\|_{\text{b}}} =: \frac{\bar{c}}{\underline{c}}\,\|A\|_{\text{b-i}}$$

Applying this to the 1, 2 and ∞ induced norms using the inequalities (2.43) gives

$$\left.\begin{array}{rcccl} \|v\|_\infty &\le& \|v\|_2 &\le& \|v\|_1 \\ \frac{1}{\sqrt{n}}\|v\|_1 &\le& \|v\|_2 &\le& \sqrt{n}\,\|v\|_\infty \end{array}\right\} \quad\Rightarrow\quad \left\{\begin{array}{rcccl} \frac{1}{\sqrt{n}}\,\|A\|_{\infty\text{-i}} &\le& \|A\|_{2\text{-i}} &\le& \sqrt{n}\,\|A\|_{\infty\text{-i}} \\ \frac{1}{\sqrt{n}}\,\|A\|_{1\text{-i}} &\le& \|A\|_{2\text{-i}} &\le& \sqrt{n}\,\|A\|_{1\text{-i}} \\ \frac{1}{n}\,\|A\|_{1\text{-i}} &\le& \|A\|_{\infty\text{-i}} &\le& n\,\|A\|_{1\text{-i}} \end{array}\right.$$

Example 3.31. Let $A : \mathsf{V} \to \mathsf{W}$ be an operator between two Banach spaces. If we have a bases $\mathsf{v} := \{v_k\}_{k=0}^{\infty} \subset \mathsf{V}$ and $\mathsf{v} := \{w_k\}_{k=0}^{\infty} \subset \mathsf{W}$, then the representation of A in those bases is a semi-infinite matrix. Indeed, for each basis element in the domain v_l, its image $Av_l \in \mathsf{W}$ has a basis expansion in the co-domain, so label that expansion as follows

$$Av_l \;=:\; \sum_{k=0}^{\infty} a_{kl}\; w_k. \tag{3.29}$$

Consider $y = Ax$, with $x \in \mathsf{V}$, $y \in \mathsf{W}$, and their respective basis expansions $x = \sum_{k=0}^{\infty} x_k v_k$, $y = \sum_{k=0}^{\infty} y_k w_k$. If x and y are represented by the semi-infinite vectors of their expansion coefficients, then the relation between them is the semi-infinite matrix with coefficients a_{kl}

$$\begin{bmatrix} y_0 \\ y_1 \\ \vdots \end{bmatrix} = \begin{bmatrix} a_{00} & a_{01} & \cdots \\ a_{10} & a_{11} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ \vdots \end{bmatrix}. \tag{3.30}$$

This is a simple consequence of the linearity of A. By the definition (3.29), the l'th column of the matrix contains the basis coefficients of the vector Av_l.

As specific examples, consider the spaces $\ell^p(\mathbb{N})$ with elements represented in the canonical basis as column vectors. Then every linear operator $A : \ell^{p_1}(\mathbb{N}) \to \ell^{p_2}(\mathbb{N})$ is represented by a semi-infinite matrix. For the cases of $\ell^1(\mathbb{N})$ and $\ell^\infty(\mathbb{N})$, it is possible to show (c.f. Examples 3.28 and 3.32) that the induced norms of an operator are obtained from its semi-infinite matrix representation as "sup-column-sum" and "sup-row-sum" respectively

$$\|A\|_{1\text{-i}} \;=\; \sup_{0\le l<\infty} \sum_{k=0}^{\infty} |a_{kl}|\,, \qquad\qquad \|A\|_{\infty\text{-i}} \;=\; \sup_{0\le k<\infty} \sum_{l=0}^{\infty} |a_{kl}|\,. \tag{3.31}$$

Note the similarity of these expressions with the finite dimensional case (3.28).

The attentive reader may have spotted a flaw in the previous argument. ℓ^∞ does not have a basis in the sense of Definition 3.12, so the bases argument leading to the semi-infinite representation (3.30) does not apply. Thus for the case of ℓ^∞, the statement in the previous example should be understood in the following sense. If the operator is representable by a semi-infinite matrix, then its induced norm is given by (3.31).

Example 3.32. Consider an operator $\mathcal{A} : \mathsf{L}^\infty(\mathbb{R}) \to \mathsf{L}^\infty(\mathbb{R})$ with an integral representation

$$v \;=\; \mathcal{A}u \qquad \Leftrightarrow \qquad v(x) \;=\; \int_{\mathbb{R}} A(x,\xi)\; u(\xi)\; d\xi. \tag{3.32}$$

The two variable function $A(.,.)$ is called the kernel of the operator $\mathcal{A}$, and the integral representation (3.32) is called the kernel representation of the operator. These representations are the continuum analogues of matrix representations and are studied further in Chapter 6.

Now calculate the L^∞-induced norm of $\mathcal{A}$ in terms of its kernel function $A(.,.)$ as follows (recall the 1-∞ inequality of Exercise 2.5)

$$\begin{aligned} |v(x)| \;&=\; \left|\int_{\mathbb{R}} A(x,\xi)\; u(\xi)\; d\xi\right| \;\le\; \left(\int_{\mathbb{R}} |A(x,\xi)|\; d\xi\right)\left(\sup_{\xi\in\mathbb{R}} |u(\xi)|\right) \\ \Rightarrow\; \|v\|_\infty \;&=\; \sup_{x\in\mathbb{R}} |v(x)| \;\le\; \left(\sup_{x\in\mathbb{R}} \int_{\mathbb{R}} |A(x,\xi)|\; d\xi\right)\; \|u\|_\infty. \end{aligned}$$

The upper bound obtained above is tight. Let $\bar{x}$ be such that the supremum in

$$\sup_{x\in\mathbb{R}} \int_{\mathbb{R}} |A(x,\xi)|\; d\xi \tag{3.33}$$

is almost achieved, then the function $\bar{u}(\xi) := \text{sign}\,(A(\bar{x},\xi))$ almost achieves the bound.

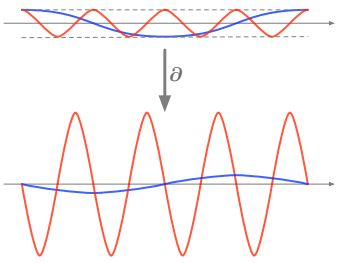
The expression (3.33) can be interpreted as the "sup-row-integral" of $A(.,.)$ by regarding ξ and x as "column" and "row" indices respectively. The integral representation (3.32) is a continuum analogue of a matrix representation. More on this in Chapter 6.

Example 3.33. Recall the space $\mathsf{C}^1[a,b]$ of continuously differentiable functions of Example 3.11

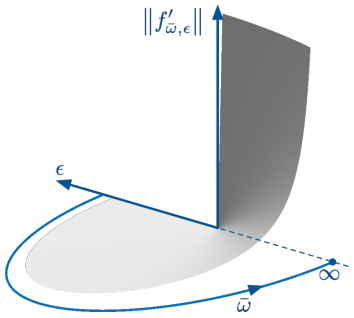
$$\mathsf{C}^1[a,b] \;:=\; \big\{f : [a,b] \to \mathbb{R};\; f' \text{ continuous}\big\}\,, \qquad \|f\|_\infty \;:=\; \max_{x\in[a,b]} |f(x)|.$$

Also recall that this is a normed, but incomplete space. Consider the differential operator

$$\big(\mathbf{D}f\big)(x) \;:=\; \frac{d}{dx} f(x), \qquad\qquad \mathbf{D} : \mathsf{C}^1[a,b] \to \mathsf{C}[a,b].$$



(a) The derivative operator $\mathbf{D}$ acting on functions of the same $\|.\|_\infty$ norm produces functions with very different $\|.\|_\infty$ norms. Here the functions $f_{\bar\omega}(x) := \cos(\bar\omega x)$ have norm $\|f_{\bar\omega}\|_\infty = 1$ for any $\bar\omega$, but their derivatives have norms $\|\mathbf{D}f_{\bar\omega}\|_\infty = \bar\omega$ that can become arbitrarily large as $\bar\omega \to \infty$. The derivative operator $\mathbf{D}$ is therefore an unbounded operator.



(b) A conceptual depiction of the discontinuity of the unbounded derivative operator $\mathbf{D}$. Around zero, $\mathbf{D}$ acts on the functions $f_{\bar\omega,\epsilon}(x) := \epsilon\cos(\bar\omega x)$ (which all have norm ϵ) to produce functions with norm $\|f'_{\bar\omega,\epsilon}\|_\infty = \epsilon\bar\omega$. We can think of $\bar\omega$ as a "direction" in C^1 of a vector $f_{\bar\omega,\epsilon}$ of length ϵ around zero. Regardless of how small ϵ is, as the direction $\bar\omega \to \infty$, the value of $\|\mathbf{D}f_{\bar\omega,\epsilon}\|_\infty$ becomes arbitrarily large.

Figure 3.11: The derivative operator $\mathbf{D} : \mathsf{C}^1[a,b] \to \mathsf{C}[a,b]$ of Example 3.33 is an unbounded operator on continuously differentiable functions. Its unboundedness is equivalent to discontinuity at 0.

We can see that this is an unbounded operator by acting on the function (see Figure 3.11a)

$$f_{\bar\omega}(x) := \cos(\bar\omega x) \quad \Rightarrow \quad f'_{\bar\omega}(x) = -\bar\omega \sin(\bar\omega x) \quad \Rightarrow \quad \frac{\|\mathbf{D}f_{\bar\omega}\|_\infty}{\|f_{\bar\omega}\|_\infty} = \frac{\bar\omega}{1}.$$

This ratio is unbounded as $\bar\omega \to \infty$.

We can interpret the unboundedness of the norm ratio as a discontinuity as follows. The rescaled function $f_{\bar\omega,\epsilon}(x) := \epsilon\cos(\bar\omega x)$ is a distance ϵ away from the zero vector in $\mathsf{C}^1[a,b]$ regardless of the frequency $\bar\omega$. However, its image under $\mathbf{D}$ is $f'_{\bar\omega,\epsilon}(x) = -\bar\omega\epsilon\sin(\bar\omega x)$, and can be made arbitrarily far from 0 in $\mathsf{C}[a,b]$ by choosing $\bar\omega$ sufficiently large. If we think of $\bar\omega$ as a "direction" in which the vector $f_{\bar\omega,\epsilon}$ is pointing, then the mapping $\mathbf{D}$ amplifies little in some directions, and amplifies unboundedly as directions are changed near zero. This function is thus discontinuous at zero as depicted in Figure 3.11b.

Inverses, Null and Image Spaces

Let $A : \mathsf{V} \to \mathsf{W}$ be a bounded (i.e. continuous) operator between Banach spaces. Its null space $\mathsf{Nu}(A) \subseteq \mathsf{V}$ is the inverse image of $0 \in \mathsf{W}$. The inverse image of every closed set under a continuous map is a closed set. The single point set $0 \in \mathsf{W}$ is closed, therefore $\mathsf{Nu}(A)$ must be closed in V. We therefore conclude that *the null space of any bounded operator is a closed subspace.*

On the other hand, the image space of a bounded operator may or may not be closed. Consider the following operator $A : \ell^1 \to \ell^1$

$$A\,(u_1, u_2, u_3, \ldots) \;:=\; \left(u_1, \frac{1}{2}u_2, \frac{1}{3}u_3, \ldots\right). \tag{3.34}$$

If we represent elements of ℓ^1 as semi-infinite vectors, then A is represented by the semi-infinite diagonal matrix $A = \mathsf{diag}\left(1, \frac{1}{2}, \frac{1}{3}, \cdots\right)$. From Example 3.31 its induced norm is the supremum of all diagonal elements, i.e. $\|A\| = 1$ and is therefore a bounded operator. The null space is exactly 0 since there is no other vector mapped to zero. What about the image space?

First note that the subspace of all finite sequences is mapped to the subspace of all finite sequences, and this subspace is dense in ℓ^1. Thus $\overline{\mathsf{Im}(A)} = \ell^1$. However there are elements in ℓ^1 that are not in the image. For example the absolutely summable sequence

$$\left(1, \frac{1}{2^2}, \ldots, \frac{1}{k^2}, \ldots\right) \notin \mathsf{Im}(A)\,,$$

for if it were, then $(1, 1/2, 1/3, \ldots)$ would be in ℓ^1, which it is not. Therefore $\mathsf{Im}(A) \neq \ell^1$, but dense in it, therefore it is not closed.

The operator $A = \mathsf{diag}\left(1, \frac{1}{2}, \frac{1}{3}, \cdots\right)$ is one-to-one, and if it has an inverse, it must be $A^{-1} = \mathsf{diag}(1, 2, 3, \ldots)$. This inverse however does not map all of ℓ^1 to ℓ^1, but it maps a dense subspace (namely $\mathsf{Im}(A)$) to ℓ^1. It is also unbounded on that subspace. A^{-1} is one example of the class of densely-defined, unbounded operators which we will study in Chapter **??**. The previous example also serves to highlight the next result whose proof is omitted.

Theorem 3.34 (Bounded Inverse Theorem)**.** *If a bounded operator $A : \mathsf{V} \to \mathsf{W}$ between Banach spaces has an inverse A^{-1} (equivalently if A is one-to-one and onto), then A^{-1} is a bounded operator.*

The operator (3.34), while one-to-one, is not onto, otherwise its inverse would have to be bounded by this theorem, and we calculated that its inverse is not bounded.

The Minimum Modulus

Recall the inequality (3.23) where the operator norm was interpreted as the "largest possible" amplification $\|Av\|/\|v\|$ of the norm of a vector v when acted on by A. Similarly, another useful notion is the "smallest possible" such amplification.

Definition 3.35. *Given an operator A, its* minimum modulus *is defined by*

$$\underline{\sigma}(A) \;:=\; \inf_{\|v\|=1} \|Av\| \;=\; \inf_{v\neq 0} \frac{\|Av\|}{\|v\|}$$

Note that the second equality follows from the homogeneity property of norms in a similar manner to the argument in (3.22).

If the operator is invertible, then the norm of its inverse and its minimum modulus are reciprocals

$$\|A^{-1}\| \;:=\; \sup_{v\neq 0} \frac{\|A^{-1}v\|}{\|v\|} \;=\; \frac{1}{\inf_{v\neq 0} \frac{\|v\|}{\|A^{-1}v\|}} \;=\; \frac{1}{\inf_{w\neq 0} \frac{\|A^{-1}w\|}{\|w\|}} \;=:\; \frac{1}{\underline{\sigma}(A^{-1})} \tag{3.35}$$

There are important examples where the operator inverse A^{-1} may exist, but is unbounded. Such cases are characterized by the minimum modulus being zero $\underline{\sigma}(A) = 0$ and equivalently $\left\|A^{-1}\right\| = \infty$. This is a special case of a more general fact that the minimum modulus is a measure of "how close" an operator is to a non-invertible operator (in this case the distance is zero since the operator itself is not boundedly invertible). This will be a consequence of the Neumann series of Theorem 3.45.

Recall the operator defined by (3.34). This operator has a diagonal matrix representation, and therefore its minimum modulus is simply the infimum of the diagonal elements

$$\underline{\sigma}\left(\mathsf{diag}\left(1, \frac{1}{2}, \frac{1}{3}, \ldots\right)\right) \;=\; \inf_{k\geq 1} \left|\frac{1}{k}\right| \;=\; 0.$$

Recall also that this operator's inverse was a densely-defined, but unbounded operator on ℓ^1, i.e. $\|A^{-1}\| = \infty$. This is consistent with (3.35).

3.6 Spaces of Linear Operators

Linear operators between two vector spaces themselves form a vector space by simply defining addition and scaling pointwise. The space of *bounded* linear operators is also equipped with the induced norm, which we will show is actually a norm, and therefore the space of all bounded linear operators between two Banach spaces is itself a Banach space. There are other important metrics on spaces of (possibly unbounded) linear operators that are not norms. Those are discussed in Chapter **??**.

Another important property of the induced norm is *submultiplicativity* which bounds the norm of the composition of two operators by the product of their respective norms. This property is very useful for analysis of linear operators that are made up of several other operators. Much of sensitivity analysis in Signals and Systems revolves around using this property. We will also see that the space of linear operators from a Banach space to itself has the additional structure of a so-called *Banach Algebra* to be discussed in Section 3.6.3.

3.6.1 The Space $\mathsf{L}(\mathsf{V},\mathsf{W})$ of Bounded Operators

The set of all linear operators between two vector spaces V and W is itself a vector space. The vector space operations are point-wise additions and scalings

$$\begin{aligned} A, B &: \mathsf{V} \to \mathsf{W} \\ \big(\alpha A + \beta B\big)(v) \;&:=\; \alpha\, A(v) \;+\; \beta\, B(v). \end{aligned}$$

It is an immediate exercise to show that the linear combination $\alpha A + \beta B$ is a linear operator from V to W.

When V and W are equipped with norms, then *bounded* operators between them have the naturally defined induced norm. The induced norm also satisfies the three properties of a norm. Homogeneity and definiteness of $\|.\|_{\mathrm{i}}$ follows from those same properties of the vector space norms[9]

$$\begin{aligned} \|\alpha A\|_{\mathrm{i}} \;&=\; \sup_{v\in\mathsf{V}} \frac{\|\alpha A v\|}{\|v\|} \;=\; \sup_{v\in\mathsf{V}} \frac{|\alpha|\|Av\|}{\|v\|} \;=\; |\alpha|\, \sup_{v\in\mathsf{V}} \frac{\|Av\|}{\|v\|} \;=\; |\alpha|\, \|A\|_{\mathrm{i}}, \\ \|A\|_{\mathrm{i}} = 0 \;\;&\Rightarrow\;\; \sup_{v\in\mathsf{V}} \frac{\|Av\|}{\|v\|} = 0 \;\;\Rightarrow\;\; \forall v\in\mathsf{V},\ \|Av\| = 0 \;\;\Rightarrow\;\; \forall v\in\mathsf{V},\ Av = 0 \;\;\Rightarrow\;\; A = 0. \end{aligned}$$

The induced norm $\|.\|_{\mathrm{i}}$ also satisfies the triangle inequality because the vector norms in V and W do

$$\begin{aligned} \|A+B\|_{\mathrm{i}} \;&:=\; \sup_{v\in\mathsf{V}} \frac{\big\|\big(A+B\big)(v)\big\|}{\|v\|} \;=\; \sup_{v\in\mathsf{V}} \frac{\|Av+Bv\|}{\|v\|} \;\le\; \sup_{v\in\mathsf{V}} \frac{\|Av\|+\|Bv\|}{\|v\|} \\ &=\; \sup_{v\in\mathsf{V}} \left(\frac{\|Av\|}{\|v\|} + \frac{\|Bv\|}{\|v\|} \right) \;\le\; \sup_{v\in\mathsf{V}} \frac{\|Av\|}{\|v\|} + \sup_{v\in\mathsf{V}} \frac{\|Bv\|}{\|v\|} \;=\; \|A\|_{\mathrm{i}} + \|B\|_{\mathrm{i}}. \end{aligned}$$

Thus the space of all bounded linear operators between two normed spaces V and W is itself and normed vector space with the induced norm. It turns out that this space is also complete if V and W are complete.

[9] For notational simplicity, from now on we drop the subscripts $\|.\|_{\mathsf{V}}$ on the vector space norm when no confusion can occur, and we will also rewrite $\sup_{0\neq v\in\mathsf{V}}$ simply as $\sup_{v\in\mathsf{V}}$ with the implicit assumption that $v \neq 0$ when used in such an expression.

Lemma 3.36. *Let* V *and* W *be Banach spaces. The set of all* bounded linear operators from V to W

$$\mathsf{L}(\mathsf{V},\mathsf{W}) \;:=\; \left\{ A:\mathsf{V}\to\mathsf{W};\; A \text{ linear, and } \|A\|_{\mathrm{i}} := \sup_{0\neq v\in\mathsf{V}} \frac{\|Av\|_{\mathsf{W}}}{\|v\|_{\mathsf{V}}} < \infty \right\}$$

is itself a Banach space with the induced norm $\|.\|_{\mathrm{i}}$.

Proof. It remains to show that $\mathsf{L}(\mathsf{V},\mathsf{W})$ is complete, i.e. to show that every Cauchy sequence $\{A_k\}$ (where the Cauchy property is measured by the induced norm) converges in $\mathsf{L}(\mathsf{V},\mathsf{W})$, i.e. converges to a *bounded* linear operator. To find the operator $\bar{A}$ which is the limit of such a sequence, we have to specify how it acts on each vector v. A reasonable guess is that the limit operator $\bar{A}$ should act like

$$\bar{A}v \;:=\; \lim_{k\to\infty} A_k v. \tag{3.36}$$

on each vector $v\in\mathsf{V}$. To show that this limit exists, note that $\{A_k\}$ being Cauchy means

$$\lim_{k,l\to\infty} \|A_k - A_l\|_{\mathrm{i}} \;=\; 0. \tag{3.37}$$

This shows that the sequence of vectors $\{A_k v\}$ is itself a Cauchy sequence in W since

$$\|A_k v - A_l v\| \;\leq\; \big\|(A_k - A_l)v\big\| \;\leq\; \|A_k - A_l\|_{\mathrm{i}}\,\|v\|,$$

and when combined with the limit (3.37), this implies that $\lim_{k,l\to\infty}\|A_k v - A_l v\| = 0$. Now, since $\{A_k v\}$ is a Cauchy sequence in a Banach space, it must have a unique limit and thus the mapping (3.36) is well defined.

Finally it remains to show that A defined by (3.36) is linear and bounded. Linearity follows from the linearity of each A_k

$$\begin{aligned} A(\alpha v_1+\beta v_2) = \lim_{k\to\infty} A_k(\alpha v_1+\beta v_2) &= \lim_{k\to\infty} \big(\alpha\, A_k v_1 \;+\; \beta\, A_k v_2\big) \\ &= \alpha \lim_{k\to\infty} A_k v_1 \;+\; \beta \lim_{k\to\infty} A_k v_2 = \alpha\, Av_1 + \beta\, Av_2. \end{aligned}$$

To establish the boundedness of A we first observe that the sequence $\{\|A_k\|\}$ is itself a Cauchy sequence of numbers. This follows from the triangle inequality

$$\Big|\, \|A_k\|_{\mathrm{i}} - \|A_l\|_{\mathrm{i}} \Big| \;\leq\; \|A_k - A_l\|_{\mathrm{i}} \;\overset{k,l\to\infty}{\longrightarrow}\; 0.$$

Since every Cauchy sequence is bounded, then we have an upper bound $\sup_k \|A_k\|_{\mathrm{i}} \leq c < \infty$ for some constant c. This can be used to bound the induced norm of A as follows

$$\|Av\| \;=\; \lim_{k\to\infty}\|A_k v\| \;\leq\; \lim_{k\to\infty}\|A_k\|_{\mathrm{i}}\,\|v\| \;\leq\; \sup_k \|A_k\|_{\mathrm{i}}\,\|v\| \;\leq\; c\,\|v\|$$

Therefore $\|A\|_{\mathrm{i}} \leq c < \infty$, and A is a bounded operator. □

Can $\mathsf{L}(\mathsf{V},\mathsf{W})$ Ever be a Hilbert Space?

Since Hilbert spaces are also Banach spaces, then the space of all bounded linear operators between two Hilbert spaces is a Banach space. Could it also be a Hilbert space itself? The answer is generally no, except for very simple cases.

First consider the space $\mathsf{L}(\mathbb{R},\mathsf{V})$ where V is a Hilbert space. This is the space of all linear operators from $\mathbb{R}$ to V, and each operator maps $\mathbb{R}$ to a one-dimensional subspace of V. Each

map can be uniquely identified by a single vector in V, namely the vector that the number 1 is mapped to, which means the induced norm of the map is the norm of that vector. Therefore this identification is an isometry between $\mathsf{L}(\mathbb{R}, \mathsf{V})$ and V, which is a Hilbert space.

Now consider the space $\mathsf{L}(\mathsf{V}, \mathbb{R})$. This is the space of all linear functionals on V, which is isometric to V itself by the Riesz representation theorem. Thus in this case $\mathsf{L}(\mathsf{V}, \mathbb{R})$ is a Hilbert space.

In general however, if $\dim \mathsf{V} > 1$ and $\dim \mathsf{W} > 1$, then the norm on $\mathsf{L}(\mathsf{V}, \mathsf{W})$ (the induced operator norm) can never arise from an inner product. A norm that comes from an inner product must satisfy the parallelogram law (2.33), and we now show that this does not hold if both $\dim \mathsf{V} \geq 2$ and $\dim \mathsf{W} \geq 2$. First consider the following two elements in $\mathsf{L}(\mathbb{R}^2, \mathbb{R}^2)$

$$\left.\begin{aligned} A_1 &:= \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \\ A_2 &:= \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \end{aligned}\right\} \quad \Rightarrow \quad \begin{aligned} \|A_1\| &= 1, \qquad \|A_1 + A_2\| = \left\| \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right\| = 1, \\ \|A_2\| &= 1, \qquad \|A_1 - A_2\| = \left\| \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \right\| = 1, \end{aligned} \tag{3.38}$$

where the matrix norm is the maximum singular value[10]. The parallelogram law

$$2 \;=\; \|A_1 + A_2\|^2 + \|A_1 - A_2\|^2 \;\neq\; 2\left(\|A_1\|^2 + \|A_2\|^2\right) \;=\; 4$$

is clearly violated in this case.

This counter-example can be generalized to any two Hilbert spaces that are not both one dimensional. Indeed, in this case we can find *unit* vectors $v_1 \perp v_2$ in V and $w_1 \perp w_2$ in W respectively. The two sets of vectors define 2-dimensional subspaces respectively, over which we can imitate the above argument as follows. Define the two mappings $A_1, A_2 : \mathsf{V} \to \mathsf{W}$

$$A_1 : \; v \;\mapsto \langle v_1 \,,\, v \rangle \;\; w_1 \qquad\qquad A_2 : \; v \;\mapsto \langle v_2 \,,\, v \rangle \;\; w_2. \tag{3.39}$$

Note that if we restrict those operators to $\mathsf{span}\{v_1, v_2\}$ and project onto $\mathsf{span}\{w_1, w_2\}$, then in the bases $\{v_1, v_2\}$ and $\{w_1, w_2\}$ the matrix representations $\hat{A}_1$ and $\hat{A}_2$ of these operators are given by

$$\hat{A}_1 \;=\; \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \end{bmatrix}, \qquad\qquad \hat{A}_1 \;=\; \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \end{bmatrix}.$$

The reader should observe here that the operation of multiplying by $[1 \;\; 0]$ in this basis is equivalent to $\langle v_1 \,,\, . \rangle$, and w_1 is represented by the vector $(1, 0)$. Similarly for $\hat{A}_2$. This correspondence guides how (3.38) was generalized to (3.39)

Since v_1, v_2, w_1, w_2 are all unit length vectors, we an easily calculate induced norms

$$\begin{aligned} \|A_1 v\| = \|\langle v_1 \,,\, v \rangle \;\; w_1\| \;\; &= \;\; |\langle v_1 \,,\, v \rangle| \, \|w_1\| \;\; \leq \;\; \|v\| && \Rightarrow \qquad \|A_1\| \leq 1 \\ \|A_1 v_1\| = |\, \langle v_1 \,,\, v_1 \rangle \,| = 1 \;\; &&& \Rightarrow \qquad \|A_1\| = 1 \\ \|(A_1 + A_2)\,(v)\|^2 \;\; &= \;\; \|A_1 v + A_2 v\|^2 = \|\langle v_1 \,,\, v \rangle \, w_1 + \langle v_2 \,,\, v \rangle \, w_2\|^2 && \\ &= \;\; |\, \langle v_1 \,,\, v \rangle \,|^2 + |\, \langle v_2 \,,\, v \rangle \,|^2 && (\text{since } w_1 \perp w_2) \\ &\leq \;\; \|v\|^2 && \Rightarrow \qquad \|A_1 + A_2\| \leq 1 \end{aligned}$$

and similarly $\|A_2\| = 1$ and $\|A_1 - A_2\| \leq 1$. The parallelogram law is violated since

$$\|A_1 + A_2\|^2 + \|A_1 - A_2\|^2 \;\leq\; 2 \;<\; 2\left(\|A_1\|^2 + \|A_2\|^2\right) \;=\; 4.$$

[10] Alternatively, the fact that all these induced norms are 1 can be quickly concluded from the basic definition of the induced norm in these simples cases.

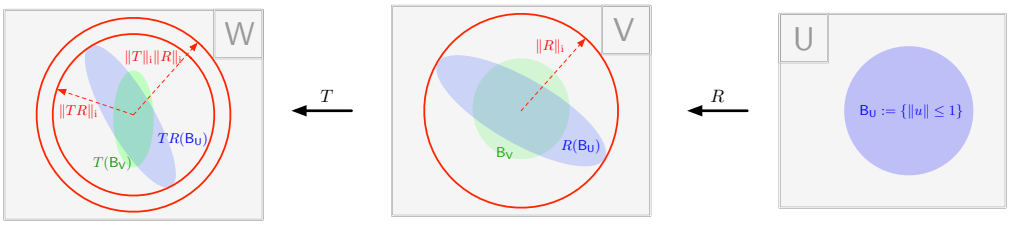


Figure 3.12: Illustration of the submultiplicativity property of the induced norm. B_U and B_V are the unit balls of U and V respectively. The radii of the smallest ball containing their images $R(\mathsf{B}_\mathsf{U})$ and $T(\mathsf{B}_\mathsf{V})$ are the induced norms $\|R\|_\mathrm{i}$ and $\|T\|_\mathrm{i}$ respectively. The induced norm $\|TR\|_\mathrm{i}$ of the composition is the radius of the smallest ball containing the image $TR(\mathsf{B}_\mathsf{U})$. This is always bounded from above by the product $\|R\|_\mathrm{i}\ \|T\|_\mathrm{i}$.

3.6.2 Submultiplicativity

Linear operators are mappings between sets, and mappings between sets can be composed together. The induced operator norm has useful property under compositions which is termed *submultiplicativity*.

Lemma 3.37. *Let $T : \mathsf{V} \to \mathsf{W}$ and $R : \mathsf{U} \to \mathsf{V}$ be linear operators between normed vector spaces. The induced norm of the composition TR is bounded by*

$$\|TR\|_\mathrm{i} \ \le\ \|T\|_\mathrm{i}\ \|R\|_\mathrm{i}. \tag{3.40}$$

Note that $\|R\|_\mathrm{i}$ is the norm induced as a mapping from U to V, while $\|T\|_\mathrm{i}$ is the norm induced from V to W. Regardless of the choices of norms on those vector spaces, the inequality holds when $\|T\|_\mathrm{i}$ and $\|R\|_\mathrm{i}$ are the corresponding induced norms. This lemma is illustrated in Figure 3.12.

Proof. This follows from the basic definitions. First observe that for any $u \in \mathsf{U}$

$$\frac{\|TR(u)\|}{\|u\|} \ =\ \frac{\|T(R(u))\|}{\|u\|} \ \le\ \frac{\|T(R(u))\|}{\|R(u)\|}\ \frac{\|R(u)\|}{\|u\|} \ \le\ \|T\|_\mathrm{i}\ \|R\|_\mathrm{i}.$$

$\|TR\|_\mathrm{i}$ is the supremum of the quantity on the left (over all $u \in \mathsf{U}$), and is therefore bounded by the quantity on the right (which is independent of u). □

Now we point out the special nature of a submultiplicative norm. First, we need to be able to make sense of a composition like TR, and if both are linear operators with a common space "in the middle", then the composition is well defined. Second, the norms on the operators need to be induced norms for (3.40) to hold. There are norms on operators which are not induced norms, and therefore may not be submultiplicative. We next present such an example for matrices.

Example 3.38. On the set of $n \times m$ matrices, we have seen several induced norms earlier, namely the $1-$, $2-$ and ∞-induced norms. All of those satisfy the submultiplicativity property. There are other norms that can be put on matrices. For example, consider the maximum absolute value of all entries

$$\|A\|_\mathrm{m} \ :=\ \max_{i,j} |a_{ij}|\,. \tag{3.41}$$

This is a norm on the space of matrices. Recall the operation vec : $\mathbb{R}^{n\times m} \to \mathbb{R}^{nm}$ introduced in Example 1.5 which takes a matrix and makes a vector out of it by stacking all the matrix

columns together. This is a vector space isomorphism, and also an *isometry* since the norm defined above is just the $\|.\|_\infty$ norm of the resulting vector, i.e. $\|A\|_{\mathrm{m}} = \|\mathrm{vec}(A)\|_\infty$. Thus the norm defined in (3.41) makes the space of $n \times m$ matrices into a normed vector space.

The norm (3.41) however is not submultplicative as the following example shows

$$2 = \left\| \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} \right\|_{\mathrm{m}} = \left\| \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \right\|_{\mathrm{m}} \not\leq \left\| \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \right\|_{\mathrm{m}} \left\| \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \right\|_{\mathrm{m}} = 1.$$

Since an induced norm must be submultiplicative, we conclude that the norm defined in (3.41) cannot be an induced norm of the matrix as linear operator between two normed vector spaces.

Example 3.39. The Frobenius norm on matrices is not an induced norm. One way to see this is that the identity operator should always have induced norm 1, but for an $n \times n$ identity matrix I_n

$$\|\mathsf{I}_n\|_{\mathsf{F}} = \sqrt{n}.$$

The Frobenius norm is none the less submultiplicative with respect to matrix products as the following argument (using partitioned matrix notation) shows

$$\begin{aligned} \|AB\|_{\mathsf{F}}^2 &= \left\| \begin{bmatrix} a_1^* \\ \vdots \\ a_n^* \end{bmatrix} \begin{bmatrix} b_1 & \cdots & b_q \end{bmatrix} \right\|_{\mathsf{F}}^2 = \left\| \begin{bmatrix} a_1^* b_1 & \cdots & a_1^* b_q \\ \vdots & & \vdots \\ a_n^* b_1 & \cdots & a_n^* b_q \end{bmatrix} \right\|_{\mathsf{F}}^2 \\ &= \sum_{i,j} (a_i^* b_j)^2 \leq \sum_{i,j} \|a_i\|_2^2 \, \|b_j\|_2^2 \qquad \text{(Cauchy-Schwarz: } \langle a_i \, , \, bj \rangle \leq \|a_i\|_2 \|b_j\|_2) \\ &= \left(\sum_i \|a_i\|_2^2 \right) \left(\sum_i \|b_j\|_2^2 \right) = \|A\|_{\mathsf{F}}^2 \, \|B\|_{\mathsf{F}}^2. \end{aligned}$$

Thus submultiplicativity is a necessary, but not sufficient condition for a norm to be an induced norm.

We finally note that although the statement (3.40) involves only two operators, by repeated applications it can be applied to any number of operator compositions

$$\begin{aligned} & \|A_1 A_2 A_3\| \leq \|A_1 A_2\| \, \|A_3\| \leq \|A_1\| \, \|A_2\| \, \|A_3\| \, , \\ \text{similarly} \quad & \|A_1 \, \cdots \, A_k\| \leq \|A_1\| \cdots \|A_k\| \end{aligned}$$

3.6.3 The Algebra of Bounded Operators

We now consider the space $\mathsf{L}(\mathsf{V}, \mathsf{V})$ of all bounded operators from a Banach space V to itself. $\mathsf{L}(\mathsf{V})$ equipped with the induced norm is a Banach space as well, but also has a "product operation", namely operator composition. The norm in $\mathsf{L}(\mathsf{V}, \mathsf{V})$ satisfies submultiplicativity with respect to this product operation. These properties represent a very special structure called a *Banach Algebra*. Before we define this structure, we define the preliminary structure of an *Algebra*.

Recall that the structure of a vector space is that of additions and scalings of vectors. If an operation of vector products is also defined, then we call such a space an algebra.

Definition 3.40. *An* algebra *is a vector space with a* product *operation that is associative, not necessarily commutative, and has the following additional properties of compatibility with the vector space structure*[11]

[11] Some references use the term "unital, associative algebra" for the definition given here since there exists useful algebras that are not associative (e.g. Lie Algebras), or without a unit element. The analysis of non-associative algebras is quite different from associative ones. Here we simply use the term "algebra" for "unital, associative algebra".

1. *Left and right distributivity over addition*

$$(u+v)w \;=\; uw+vw, \qquad w(u+v) \;=\; wu+wv.$$

2. *Compatibility with scalings* $(\alpha u)\ (\beta v) \;=\; (\alpha\beta)\ (uv)$.

3. *Existence of a "unit" element* I *such that* $\mathsf{I}u = u\mathsf{I} = u$ *for all elements* u *of the algebra.*

Note that although a product operation is defined, the existence of a multiplicative inverse is not required for all elements of the algebra. Familiar examples of algebras include the following.

- The space $\mathbb{R}^{n\times n}$ of square $n \times n$ matrices is an algebra with the matrix-matrix product. The unit element is the identity matrix. Some elements of this algebra have multiplicative inverses and some do not.
- The space $\mathbb{P}$ of polynomials of any order is an algebra with polynomial multiplication. Note that $\mathbb{P}_n$ (polynomials of degree n) is a vector space, but not an algebra since the product of two such polynomials may have degree larger than n. To form an algebra, we need to include polynomials of any (finite) degree. Thus $\mathbb{P}$ is an infinite-dimensional vector space which is also an algebra.
- Any function space $\Omega^{\mathcal{A}}$ where the range $\mathcal{A}$ is a itself an algebra.

Now we layer another structure on top of an algebra. If we start with a Banach space rather than just a vector space, we need the norm to "work nicely" with the product operation. This is where submultiplicativity comes in.

Definition 3.41. *An algebra that is also a Banach space is called a* Banach algebra *if the product operation satisfies the sub multiplicativity property*

$$\|AB\| \;\leq\; \|A\|\ \|B\|. \tag{3.42}$$

The concept of Banach algebras was developed to study the most important example stated next.

Example 3.42. Let V be a Banach space. The space $\mathsf{L}(\mathsf{V}) := \mathsf{L}(\mathsf{V},\mathsf{V})$ of all bounded linear operators from V to itself is an algebra since any two operators $A, B : \mathsf{V} \to \mathsf{V}$ can be composed $AB : \mathsf{V} \to \mathsf{V}$. The submultiplicativity property of the induced norm implies that AB is bounded if A and B are.

Example 3.43. Convolution of functions in $\mathsf{L}^1(\mathbb{R})$ is a product operation which is associative, distributive over additions, and compatible with scalings. The L^1 norm is also submultiplicative with convolution. Indeed, for any $f, g \in \mathsf{L}^1(\mathbb{R})$

$$\begin{aligned}\|f\star g\|_1 \;&=\; \int_{\mathbb{R}}\left|\left(\int_{\mathbb{R}} f(t-\tau)\ g(\tau)\ d\tau\right)\right| dt \;\leq\; \int_{\mathbb{R}}\int_{\mathbb{R}} |f(t-\tau)|\ \ |g(\tau)|\ \ d\tau\ dt\\ &=\; \int_{\mathbb{R}} |g(\tau)|\left(\int_{\mathbb{R}} |f(t-\tau)|\ \ dt\right) d\tau \;=\; \|f\|_1 \int_{\mathbb{R}} |g(\tau)|\, d\tau \;=\; \|f\|_1\|g\|_1.\end{aligned}$$

Thus $\mathsf{L}^1(\mathbb{R})$ with convolution meets all the requirements to be a Banach algebra except for the existence of a unit element. There is no function in $\mathsf{L}^1(\mathbb{R})$ with the unit property (with respect to convolution). The reader may suggest the Dirac delta function, but that is not an element of $\mathsf{L}^1(\mathbb{R})$. However, this is not a serious limitation as we now show.

Given an algebra $\bar{\mathcal{A}}$ without a unit, one can always formally "append" a unit element as follows. Define the vector space $\mathcal{A} := \mathbb{R} \oplus \bar{\mathcal{A}}$, and products on it as follows

$$\mathcal{A} := \left\{(\alpha, f);\ \alpha \in \mathbb{R},\ f \in \bar{\mathcal{A}}\right\}, \qquad (\alpha, a)\ (\beta, g)\ :=\ (\alpha\beta, \alpha g + \beta f + fg). \tag{3.43}$$

Note that $\alpha\beta$ is a product of real numbers, αg and βf are scalings of elements in $\bar{\mathcal{A}}$, while fg is a product in the algebra $\bar{\mathcal{A}}$. This new product on $\mathcal{A}$ satisfies all the properties of an algebra product. The unit in the algebra $\mathcal{A}$ is now $(1, 0)$ since

$$(1, 0)\ (\beta, g)\ =\ (1\beta, 1g + \beta 0 + 0g)\ =\ (\beta, g). \tag{3.44}$$

The construction above is essentially what one does by appending a Dirac delta function to $\mathsf{L}^1(\mathbb{R})$ when we formally write

$$f(t)\ =\ \alpha\delta(t)\ +\ \bar{f}(t), \qquad \bar{f} \in \mathsf{L}^1(\mathbb{R}), \tag{3.45}$$

for elements $f \in \mathbb{R} \oplus \mathsf{L}^1(\mathbb{R})$. The unit element is $\delta(.)$, which corresponds to $(1, 0)$ in (3.44). The reader should verify that convolutions of elements of the form (3.45) behave like the product defined in (3.43).

Thus the set $\mathbb{R} \oplus \mathsf{L}^1(\mathbb{R})$, which is $\mathsf{L}^1(\mathbb{R})$ with a unit element appended, is a Banach algebra with convolution as the product operation. Since convolution of scalar-valued functions is commutative, this is actually an example of a commutative Banach algebra.

Example 3.44. By arguments similar to those in the previous example, the space $\ell^1(\mathbb{Z})$ is a Banach algebra under convolutions. Since this space includes the Kronecker delta function as an element, it already comes equipped with a unit.

The Neumann Series

One of the most important consequences of submultiplicativity is the way it characterizes powers of a given operator $A : \mathsf{V} \to \mathsf{V}$

$$\|A^2\|\ =\ \|A\ A\|\ \leq\ \|A\|\ \|A\|\ =\ \|A\|^2.$$

This clearly can be carried further to any power of A

$$\|A^k\|\ =\ \|\underbrace{A\ \cdots\ A}_{k \text{ times}}\|\ \leq\ \underbrace{\|A\|\ \cdots\ \|A\|}_{k \text{ times}}\ =\ \|A\|^k.$$

Note that in particular if $\|A\| < 1$, then $\|A\|^k$ is a decaying geometric sequence, and therefore powers of A decay geometrically

$$\|A^k\|\ \leq\ \alpha^k, \qquad \alpha = \|A\| < 1.$$

This bound allows us to develop one of the most useful series in applications, the convergence of which is very simple to prove.

Theorem 3.45 (The Neumann Series)**.** *Let A be a bounded operator on a Banach space V. If $\sum_{k=0}^{\infty} \left\|A^k\right\| < 0$ (i.e. the series of norms is absolutely summable), then the inverse of $(I - A)$ exists as a bounded operator on V, and can be given by the following series*

$$(I - A)^{-1}\ =\ I + A + A^2 + \cdots\ =\ \sum_{k=0}^{\infty} A^k \tag{3.46}$$

which converges in $\mathsf{L}(\mathsf{V})$.

In the case when $\|A\| < 1$, the above series is absolutely summable and furthermore

$$\left\|(I - A)^{-1}\right\|\ \leq\ \frac{1}{1 - \|A\|}. \tag{3.47}$$

Proof. By Exercise 3.3, if a series in a Banach space is absolutely summable, then it is convergent in that Banach space ($\mathsf{L}(\mathsf{V})$ in this case). To show that the series indeed gives the inverse of $(I-A)$, look at the product of $(I-A)$ and the partial sums of the series

$$\begin{aligned}(I-A)\ \left(\textstyle\sum_{k=0}^{n} A^k\right) &= (I-A)\left(I+A+A^2+\cdots+A^n\right)\\ &= I+A+A^2+\cdots+A^n\\ &\quad -A-A^2-\cdots-A^n-A^{n+1} \ = \ I-A^{n+1}.\end{aligned}$$

The summability of the series implies that $\left\|A^{n+1}\right\| \overset{n\to\infty}{\longrightarrow} 0$, and therefore $A^{n+1} \overset{n\to\infty}{\longrightarrow} 0$, and we conclude that

$$\lim_{n\to\infty} \sum_{k=0}^{n} A^k \ = \ (I-A)^{-1}.$$

Now if $\|A\|<1$, we can make the stronger statement

$$\left\|\sum_{k=\infty}^{\infty} A^k\right\| \ \overset{1}{\leq} \ \sum_{k=0}^{\infty}\left\|A^k\right\| \ \overset{2}{\leq} \ \sum_{k=0}^{\infty}\|A\|^k \ \overset{3}{=} \ \frac{1}{1-\|A\|},$$

where $\overset{1}{\leq}$ follows from the triangle inequality, $\overset{2}{\leq}$ follows from submultiplicativity, and $\overset{3}{=}$ follows from $\sum_{k=0}^{\infty}\alpha^k = \frac{1}{1-|\alpha|}$, which holds for any number with $|\alpha|<1$. □

Remark 3.46. The condition $\|A\|<1$ is sufficient for the existence of $(I-A)^{-1}$ and the convergence of the series, but it is far from necessary for either. Thinking about a real or complex numbers α, the fraction $\frac{1}{1-\alpha}$ is finite for all $\alpha\neq 1$, so clearly the condition $|\alpha|<1$ is sufficient but not necessary. While $|\alpha|<1$ is necessary for the series of numbers $\sum_{k=0}^{\infty}\alpha^k$ to absolutely converge, the condition $\|A\|<1$ is not necessary when A is a matrix or an operator. For example, a nilpotent matrix has the property that $A^k=0$ for $k>n$ for some finite n, and thus clearly the Neumann series will converge if even if $\|A\|\geq 1$ for such a matrix. The next example is is an infinite-dimensional version of this phenomenon.

Example 3.47. Consider the so-called Volterra operator of indefinite integration

$$(\mathcal{V}f)(t) \ = \ \int_0^t f(\tau)\,d\tau.$$

This operator is well defined on a variety of function spaces. Here we can take for example $\mathcal{V}:\mathsf{C}[0,1]\to\mathsf{C}[0,1]$, and recall that $\mathsf{C}[0,1]$ is equipped with the maximum norm of functions. The induced norm of $\mathcal{V}$ is easy to calculate using the concept of kernel representations of operators (Chapter 6). Here we just give the answer that the induced norms (on $\mathsf{C}[0,1]$) of all powers of $\mathcal{V}$ are given by

$$\left\|\mathcal{V}^k\right\| \ = \ \frac{1}{k!}\,. \tag{3.48}$$

Thus although this operator is not nilpotent, the norms of its powers decay rapidly to zero, and the operator could therefore be thought of as *asymptotically nilpotent*. The bounds (3.48) certainly imply absolute summability $\sum_{k=0}^{\infty}\|\mathcal{V}^k\|<\infty$, and therefore the Neumann series is convergent to $(I-\mathcal{V})^{-1}$ which is a bounded operator on $\mathsf{C}[0,1]$

$$(I-\mathcal{V})^{-1} \ = \ \sum_{k=0}^{\infty}\mathcal{V}^k : \mathsf{C}[0,1]\to\mathsf{C}[0,1].$$

This abstract example has a very concrete interpretation as follows. Consider the ordinary differential equation over $[0,1]$

$$\dot{x}(t) \;=\; x(t), \qquad\qquad x(0) = \bar{x}. \tag{3.49}$$

Integrating both sides of the equation converts it to an integral equation, which can then be rewritten abstractly using the Volterra operator

$$\begin{aligned}
&\int_0^t \dot{x}(\tau)\, d\tau \;=\; \int_0^t x(\tau)\, d\tau \\
\Leftrightarrow \qquad & x(t) - \bar{x} \;=\; (\mathcal{V}x)(t) \qquad \Leftrightarrow \qquad x(t) \;-\; (\mathcal{V}x)(t) \;=\; \bar{x},\ t \in [0,1] \\
\Leftrightarrow \qquad & (I - \mathcal{V})\, x \;=\; \mathfrak{h}\bar{x},
\end{aligned}$$

where $\mathfrak{h}$ is the unit-step (Heaviside) function $\mathfrak{h}(t) = 1,\ t \in [0,1]$, i.e. $\mathfrak{h}\bar{x}$ is the constant function on $[0,1]$ with value $\bar{x}$. We can now use the Neumann series to give the solution as

$$\begin{aligned}
x \;&=\; (I - \mathcal{V})^{-1}\, \mathfrak{h}\ \bar{x} \;=\; \left(\sum_{k=0}^{\infty} \mathcal{V}^k\right) \mathfrak{h}\ \bar{x} \;=\; \left(\sum_{k=0}^{\infty} \mathcal{V}^k \mathfrak{h}\right) \bar{x} \\
\Rightarrow \qquad x(t) \;&=\; \left(\sum_{k=0}^{\infty} \frac{1}{k!} t^k\right) \bar{x}.
\end{aligned}$$

Note that each term $\mathcal{V}^k\mathfrak{h}$ is simply the k'th integral of the constant function 1, which gives $t^k/k!$. The reader should recognize that the last series is the definition of the exponential function e^t, which is the well-known solution of the differential equation (3.49).

The argument jut presented can be readily generalized to yield the matrix exponential, the Peano-Baker series, the Cauchy formula for repeated integration, as well as the so-called "variations of constants" formula. These seemingly distinct formulas can all be thought of as various manifestations of the Neumann series involving the Volterra operator. This development is detailed in Chapter **??**, where in addition, the Picard iteration for nonlinear differential equations is presented as a version of the Neumann series.

Remark 3.48. The Volterra operator example highlights the conservatism of the condition $\|A\| < 1$ in Theorem 3.45. Let α be any real scalar, then $\|\alpha\mathcal{V}\| = |\alpha|$, which can be made as large as desired. However, homogeneity of the norm and the bounds (3.48) imply that

$$\left\|(\alpha\mathcal{V})^k\right\| \;=\; |\alpha|^k \left\|\mathcal{V}^k\right\| \;\leq\; |\alpha|^k/k!$$

Thus even though $\|\alpha\mathcal{V}\|$ can be arbitrarily large, the Neumann series for $\alpha\mathcal{V}$ is still absolutely convergent.

We close this section with another application of the Neumann series to "operator perturbations", an important topic discussed in later chapters.

Lemma 3.49. *If A is an invertible element in a Banach algebra $\mathcal{A}$, then all element of $\mathcal{A}$ of the form*

$$A + \Delta, \qquad \|\Delta\| \;<\; \frac{1}{\|A^{-1}\|}$$

are also invertible in $\mathcal{A}$.

Proof. This follows immediately from the Neumann series and submultiplicativity

$$A+\Delta \;=\; A^{-1}\left(I+A^{-1}\Delta\right) \text{ invertible} \qquad \Leftarrow \qquad \begin{array}{c} \left\|A^{-1}\Delta\right\| \;\leq\; \left\|A^{-1}\right\|\|\Delta\| \;<\; 1 \\ \Uparrow \\ \|\Delta\| \;<\; \dfrac{1}{\|A^{-1}\|}. \end{array} \qquad \square$$

This lemma has several implications as well. The first is that the set of invertible elements in $\mathcal{A}$ is an open set since for any invertible A, all elements of a ball of radius at least $1/\|A^{-1}\|$ around it are invertible. In other words, a sufficiently small perturbation of an invertible element is also invertible.

Second, recall that for an operator $1/\|A^{-1}\| = \underline{\sigma}(A)$, so the minimum modulus gives a radius of a ball around an invertible operator made up of all invertible operators. It is possible to show for a large class of operators that this estimate is tight, i.e. that there exist an operator Δ with norm $\|\Delta\| = 1/\|A^{-1}\|$ such that $A+\Delta$ is not invertible. These issues will be discussed when we study perturbation problems for linear operators, and are also part of "robustness analysis" for dynamical systems.

3.6.4 Densely-Defined Operators

There are important applications where an operator can not be defined on an the entirety of a Hilbert or a Banach space, but rather on a *domain* which is a dense subspace. There are two types of such densely-defined operators. The first is when the operator norm has a bound on the dense subspace. In this case, the operators can be easily extended to be bounded operators on the whole space. In this section we show how this is done, and then use this procedure to define the Fourier transform on $\mathsf{L}^2(\mathbb{R})$.

The second case where the operator is unbounded is most commonly encountered with differential operators, either ordinary or partial. Such cases require more care, and are treated in Chapter **??**.

Suppose we have a linear operator $A : \mathsf{S} \to \mathsf{W}$ between a (not necessarily complete) normed vector space S and a Banach (i.e. complete) space W. If $\mathsf{S} \subset \mathsf{V}$ is a *dense* subspace of a Banach space V, and A is bounded, then we can extend the domain of A to all of V with the same bound as follows

$$v \in \mathsf{V} \;\Rightarrow\; \exists\, \{v_k\} \subset \mathsf{S},\; v_k \longrightarrow v \qquad \text{then define} \qquad Av := \lim_{k\to\infty} Av_k.$$

The fact that the limit exists in W is guaranteed by the boundedness of A since

$$\|Av_k - Av_l\| \;=\; \|A(v_k - v_l)\| \;\leq\; \|A\|_{\mathrm{i}}\,\|v_k - v_l\|\,.$$

This bound implies that $\{Av_k\}$ is a Cauchy sequence in W since $\{v_k\}$ is a Cauchy sequence in V. Furthermore, the induced norm of this extension of A on V is the same as that of A on S

$$\|Av\| \;=\; \left\|\lim_{k\to\infty} Av_k\right\| \;\overset{1}{=}\; \lim_{k\to\infty}\|Av_k\| \;\leq\; \lim_{k\to\infty}\|A\|_{\mathrm{i}}\,\|v_k\| \;=\; \|A\|_{\mathrm{i}}\lim_{k\to\infty}\|v_k\| \;\overset{2}{=}\; \|A\|_{\mathrm{i}}\,\|v\|\,.$$

Note that $\overset{1}{=}$ and $\overset{2}{=}$ are justified since $\{Av_k\}$ and $\{v_k\}$ are Cauchy in W and V respectively, and therefore the norm of the limit is equal to the limit of the norms.

One of the more useful applications of this technique is for defining the Fourier transform for square integrable functions on the real line.

Example 3.50. *The Fourier Transform on* $\mathsf{L}^2(\mathbb{R})$. Consider the Fourier transform for functions on the real line

$$(\mathcal{F}u)(\omega) \;=\; \hat{u}(\omega) \;:=\; \int_{-\infty}^{\infty} e^{-j\omega t} u(t)\; dt. \tag{3.50}$$

This is clearly a linear operator that maps functions on the real line to functions on the real line $u \mapsto \hat{u}$. For what class of functions is this well defined? Note that if u is absolutely integrable (i.e. in L^1)[12], then $\hat{u}$ is well-defined at each ω, and we can bound

$$|\hat{u}(\omega)| \;=\; \left|\int_{-\infty}^{\infty} e^{-j\omega t} u(t)\; dt\right| \leq \int_{-\infty}^{\infty} \left|e^{-j\omega t}\right| |u(t)|\; dt = \int_{-\infty}^{\infty} |u(t)|\; dt \;=\; \|u\|_{\mathsf{L}^1}$$
$$\Rightarrow \qquad \|\hat{u}\|_{\mathsf{L}^\infty} \;=\; \sup_{\omega\in\mathbb{R}} |\hat{u}(\omega)| \;\leq\; \|u\|_{\mathsf{L}^1}$$

Thus the Fourier transform is a bounded linear operator $\mathcal{F} : \mathsf{L}^1 \longrightarrow \mathsf{L}^\infty$ with induced norm of 1. In fact, with a little more care it is easy to show $\hat{u}$ is absolutely continuous if u is in L^1, so we can make the stronger statement that $\mathcal{F} : \mathsf{L}^1 \longrightarrow \mathsf{L}^\infty \cap \mathsf{C}$, and note that $\mathsf{L}^\infty \cap \mathsf{C}$ is the closed subspace (i.e. a Banach space) of L^∞ made up of continuous, bounded functions.

We would also like to define the Fourier transform for functions in L^2. However, the integral (3.50) is not guaranteed to converge if u is only square integrable but not absolutely integrable. Thus we can define the Fourier transform (3.50) only on the subspace $\mathsf{L}^1 \cap \mathsf{L}^2$. This subspace is however dense[13] in L^2, and if we can find an induced norm bound on the Fourier transform as a mapping on L^2, then the extension procedure described above extends the Fourier transform from $\mathsf{L}^1 \cap \mathsf{L}^2$ to all of L^2.

The bound we need is given by Parseval's theorem

$$\int_{-\infty}^{\infty} \hat{u}^2(\omega)\; d\omega \;=\; 2\pi \int_{-\infty}^{\infty} u^2(t)\; dt.$$

Thus the Fourier transform regarded as a mapping $\mathcal{F} : \mathsf{L}^1 \cap \mathsf{L}^2 \longrightarrow \mathsf{L}^\infty \cap \mathsf{L}^2$ has induced norm of 2π (with respect to L^2 norms on u and $\hat{u}$), and its domain can therefore be extended to all of L^2.

Parseval's theorem implies the even stronger conclusion that $\mathcal{F} : \mathsf{L}^2 \longrightarrow \mathsf{L}^2$ is actually an *isometry* (modulo the constant factor 2π), i.e. a norm-preserving isomorphism. For this reason, Fourier analysis is most profitable in the L^2 setting, but ironically, the definition (3.50) cannot be directly used on L^2 functions. The densely-defined-bounded-operator extension procedure provides the simplest resolution of this technicality.

Finally we note that all the arguments above apply just as easily to the Fourier transform for $\mathsf{L}^2(\mathbb{R}^n)$.

Appendix

3.A Completion using Cauchy Sequences

Definition 3.51. *Two Cauchy sequences* $\{x_k\}$, $\{y_l\}$ *in a metric space* M, *are said to be* equivalent *if*

$$\{x_k\} \sim \{y_l\} \qquad \Leftrightarrow \qquad \textit{given } \epsilon > 0,\ \exists N,\ \textit{such that } k,l \geq N \;\Rightarrow\; \mathsf{d}(x_k, y_l) \leq \epsilon, \tag{3.51}$$

i.e. the tails of the sequences become arbitrarily close together.

[12] In this example, the notation L^1, L^∞ and L^2 stand for $\mathsf{L}^1(\mathbb{R})$, $\mathsf{L}^\infty(\mathbb{R})$ and $\mathsf{L}^2(\mathbb{R})$. The domain $\mathbb{R}$ is dropped for notational simplicity.

[13] For example, continuous, compactly supported functions are in L^1 and L^2, and are dense in both spaces.

Intuitively, equivalent Cauchy sequences should be converging to the same point, and that can be used to define a completion of any incomplete metric space.

First we show that equivalent sequences form *equivalence classes* (pun intended). Symmetry is immediate since the metric $d(.,.)$ is itself symmetric. Transitivity follows from the triangle inequality; If $\{x_k\} \sim \{y_k\}$ and $\{y_k\} \sim \{z_k\}$, then for any $\epsilon > 0$, choose N_1 and N_2 such that for $k, l \geq N_1$ and $l, j \geq N_2$

$$\left.\begin{array}{l} \mathsf{d}(x_k, y_l) \leq \epsilon \\ \mathsf{d}(y_l, z_j) \leq \epsilon \end{array}\right\} \quad \Rightarrow \quad \mathsf{d}(x_k, z_j) \leq \mathsf{d}(x_k, y_l) + \mathsf{d}(y_l, z_j) \leq 2\epsilon,$$

and note that this holds for all $k, j \geq \max\{N_1, N_2\}$. Therefore $\{x_k\} \sim \{z_k\}$

Lemma 3.52. *Given any (not necessarily complete) metric space* M*, define its* completion $\overline{\mathsf{M}}$ *as the set of all equivalence classes of Cauchy sequences in* M*. Then*

1. $\mathsf{M} \subseteq \overline{\mathsf{M}}$ *by identifying* $x \in \mathsf{M}$ *with the "constant" Cauchy sequence* $x_k := x$*.*
2. *On* $\overline{\mathsf{M}}$*, the following defines a metric*

$$\bar{\mathsf{d}}\Big(\{x_k\}, \{y_k\}\Big) := \lim_{k\to\infty} \mathsf{d}(x_k, y_k), \tag{3.52}$$

   which coincides with d *on* $\mathsf{M} \subseteq \overline{\mathsf{M}}$*.*
3. $\overline{\mathsf{M}}$ *is a complete metric space with the metric* $\bar{d}$*.*

Proof. 1. Clearly the constant sequence $x_k := x$ is a Cauchy sequence. Any other Cauchy sequence that converges to x is in the same equivalence class as this constant sequence. This equivalence class then represents the point $x \in \mathsf{M}$ in the completion $\overline{\mathsf{M}}$.

2. We need to show two things. First, (a) that this metric is well defined, i.e. the limit in (3.52) exists, and its value is independent of the choice of equivalence class representative. Second, (b) that it satisfies all the properties of a metric.

   (a) To show that the limit exists, given two Cauchy sequences $\{x_k\}$ and $\{y_l\}$ in M, we show that the sequence of real numbers $\{\mathsf{d}(x_k, y_k)\}$ is itself a Cauchy sequence in $\mathbb{R}$. Indeed, given $\epsilon > 0$, choose N so that for $k, l \geq N$ we have $d(x_k, x_l) \leq \epsilon$ and $d(y_k, y_l) \leq \epsilon$. We then compare

$$\begin{array}{rrcl} & \mathsf{d}(x_k, y_k) & \leq & d(x_k, x_l) + \mathsf{d}(x_l, y_k) \;\leq\; d(x_k, x_l) + \mathsf{d}(x_l, y_l) + \mathsf{d}(y_l, y_k) \\ \Rightarrow & \mathsf{d}(x_k, y_k) & \leq & \mathsf{d}(x_l, y_l) + 2\epsilon \\ \text{similarly} & \mathsf{d}(x_l, y_l) & \leq & \mathsf{d}(x_k, y_k) + 2\epsilon \\ & \Rightarrow & & \big|\mathsf{d}(x_k, y_k) - \mathsf{d}(x_l, y_l)\big| \;\leq\; 2\epsilon. \end{array}$$

   Since $\{\mathsf{d}(x_k, y_k)\}$ is Cauchy sequence in $\mathbb{R}$, it has a limit since $\mathbb{R}$ is complete.
   A parallel argument can be used on a given pair of equivalent sequences $\{x_k\} \sim \{\bar{x}_k\}$ and $\{y_k\} \sim \{\bar{y}_k\}$

$$\begin{array}{rrcl} & \mathsf{d}(x_k, y_k) & \leq & d(x_k, \bar{x}_k) + d(\bar{x}_k, y_k) \;\leq\; \mathsf{d}(x_k, \bar{x}_l) + \mathsf{d}(\bar{x}_k, \bar{y}_k) + \mathsf{d}(\bar{y}_k, y_k) \\ \Rightarrow & \mathsf{d}(x_k, y_k) & \leq & d(\bar{x}_l, \bar{y}_k) + 2\epsilon \\ \text{similarly} & \mathsf{d}(\bar{x}_k, \bar{y}_k) & \leq & \mathsf{d}(x_k, y_k) + 2\epsilon \\ & \Rightarrow & & \big|\mathsf{d}(x_k, y_k) - \mathsf{d}(\bar{x}_k, \bar{y}_k)\big| \;\leq\; 2\epsilon. \end{array}$$

   Thus the two sequences of real numbers $\{\mathsf{d}(x_k, y_k)\}$ and $\{\mathsf{d}(\bar{x}_k, \bar{y}_k)\}$ converge to the same number.

(b) The three properties of a metric in Definition 2.1 hold for $\bar{\mathsf{d}}$ because they hold for the original metric d. Symmetry is clear. The triangle inequality also follows

$$\begin{aligned}\bar{\mathsf{d}}\big(\{x_k\},\{y_k\}\big) &= \lim_{k\to\infty}\mathsf{d}(x_k,y_k) \le \lim_{k\to\infty}\left(\mathsf{d}(x_k,z_k)+\mathsf{d}(z_k,y_k)\right)\\ &= \lim_{k\to\infty}\mathsf{d}(x_k,z_k)+\lim_{k\to\infty}\mathsf{d}(z_k,y_k)\\ &= \bar{\mathsf{d}}(\{x_k\},\{z_k\})+\bar{\mathsf{d}}(\{z_k\},\{y_k\}).\end{aligned}$$

The functional $\bar{\mathsf{d}}$ is clearly non-negative from its definition. Also from the definition (3.52), if $\bar{\mathsf{d}}\big(\{x_k\},\{y_k\}\big) = \lim_{k\to\infty}\mathsf{d}(x_k,y_k) = 0$, then by (3.51) the two sequences belong to the same equivalence class $\{x_k\}\sim\{y_k\}$, and therefore $\bar{\mathsf{d}}$ separates distinct equivalence classes.

3. To show that $\overline{\mathsf{M}}$ is complete, we must consider a Cauchy sequence in $\overline{\mathsf{M}}$ (i.e. a Cauchy sequence of Cauchy sequences from M), and show that its limit is in $\overline{\mathsf{M}}$. This is accomplished using a *diagonal sequence* argument as follows.

Let $\left\{\left\{x_k^{(n)}\right\};\ n\in\mathbb{N}\right\}$ be a sequence of Cauchy sequences in M indexed by the integer n (i.e. for each n, $\left\{x_k^{(n)}\right\}\subset\mathsf{M}$ is a Cauchy sequence in M). Order all the sequence elements in the following two-dimensional array

$$\begin{array}{cccc} x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & \overset{k}{\rightarrow}\\ x_1^{(2)} & x_2^{(2)} & x_3^{(2)} & \cdots\\ x_1^{(3)} & x_2^{(3)} & x_3^{(3)} & \cdots\\ {\scriptstyle n}\downarrow & \vdots & & \end{array}$$

and define the *diagonal sequence* $\bar{x}_k := x_k^{(k)}$. We claim that this sequence $\{\bar{x}_k\}$ is the limit in $\overline{\mathsf{M}}$ of the family of sequences.

The fact that the family $\left\{\left\{x_k^{(n)}\right\};\ n\in\mathbb{N}\right\}$ is a Cauchy sequence in $\overline{\mathsf{M}}$ means that given $\epsilon_1>0$, $\exists N_1$ such that $n,m\ge N_1$ implies

$$\bar{\mathsf{d}}\left(\left\{x_k^{(n)}\right\},\left\{x_k^{(m)}\right\}\right)\le\epsilon_1 \qquad\Rightarrow\qquad \lim_{k\to\infty}\mathsf{d}\left(x_k^{(n)},x_k^{(m)}\right)\le\epsilon_1.$$

The limit statement implies that given $\epsilon_2>0$, $\exists N_2$ such that $k\ge N_2$ implies

$$\mathsf{d}\left(x_k^{(n)},x_k^{(m)}\right)\le\epsilon_1+\epsilon_2.$$

Now comparing with the diagonal sequence we see that

$$k,m,n\ge\max\{N_1,N_2\} \qquad\Rightarrow\qquad \mathsf{d}\left(\bar{x}_k,x_k^{(n)}\right) = \mathsf{d}\left(x_k^{(k)},x_k^{(n)}\right)\le\epsilon_1+\epsilon_2,$$

We therefore conclude that

$$\lim_{n\to\infty}\bar{\mathsf{d}}\left(\{\bar{x}_k\},\left\{x_k^{(n)}\right\}\right) := \lim_{n,k\to\infty}\mathsf{d}\left(\bar{x}_k,x_k^{(n)}\right) = \lim_{n,k\to\infty}\mathsf{d}\left(x_k^{(k)},x_k^{(n)}\right) = 0,$$

and therefore $\{\bar{x}_k\}$ is indeed the limit (in $\overline{\mathsf{M}}$) of the Cauchy sequence of Cauchy sequences. □

Exercises

Exercise 3.1 ℓ^∞ *is not separabale*

Show that $\ell^\infty(\mathbb{N})$ is not separable using the following steps.

1. Consider the subset $\ell_{0,1}$ of $\ell^\infty(\mathbb{N})$ made up of sequences with entries of only 0 or 1. By comparing with decimal expansions of numbers, show that this subset is in one-to-one correspondence with the *uncountable* set of real numbers $(0, 1)$.
2. Show that a ball of radius $1/2$ around any element of $\ell_{0,1}$ cannot contain any other element of $\ell_{0,1}$. The number of these non-intersecting balls is equal to the cardinality of $(0, 1)$.
3. Any dense subset of $\ell^\infty(\mathbb{N})$ must have an element in each of the these balls, and must therefore be uncountable.

Exercise 3.2 *Almost-periodic functions*

Consider the sum of two oscillatory functions $u(t) = \alpha e^{j\omega_1 t} + \beta e^{j\omega_2 t}$.

1. Show that u is periodic iff the two frequencies ω_1 and ω_2 are commensurate, i.e. the ratio ω_1/ω_2 is rational. In this case show that the fundamental period of u is $\tau = 2\pi(m/\omega_1) = 2\pi(n/\omega_2)$, where $\omega_1/\omega_2 = m/n$ with n and m coprime.
2. If the frequencies are incommensurate, show that there exists an ϵ-period, i.e a number T such that

$$\forall t \in \mathbb{R}, \qquad |u(t) - u(t + \tau)| < \epsilon.$$

3. Show that the set $\mathsf{T} \subset \mathbb{R}$ of ϵ-periods is *relatively dense* in $\mathbb{R}$, i.e.

$$\inf \{\mathsf{d}(T, x);\ T \in \mathsf{T},\ x \in \mathbb{R}\} = d < \infty.$$

   In other words, the set T is "well dispersed" in $\mathbb{R}$. There exists a finite number d such that the distance between any real number and the set T is at most d. Compare this with the commensurate frequencies case, where $\mathsf{T} = \{k\tau;\ k \in \mathbb{Z}\}$.

Exercise 3.3

Using the fact that an absolutely summable series of real numbers is convergent, show that if a series in a Banach space V is absolutely summable, i.e.

$$\sum_{k=0}^{\infty} \|v_k\| < \infty,$$

then the partial sums sequence $\{\sum_{k=0}^{n} u_k\}_{n=0}^{\infty}$ is Cauchy, and therefore convergent in V. Note that submultiplicativity of the norm is not required. Only the triangle inequality.

Chapter 4

Duality and Adjoints

The concept of duality *plays an important role in many aspects of linear algebra and functional analysis. Linear functionals are scalar-valued linear operators, and they are considered as objects dual to vectors. In the most basic setting, multiplying a column vector from the left by a row vector is a linear operation that yields scalars, thus row vectors can be considered as dual objects to column vectors. The space of all continuous linear functionals is the dual space of a Banach space. Concepts of orthogonality and inner products can be generalized from within inner product spaces to be considered as relations between vectors and functionals instead of between vectors and vectors. This leads to generalizations of the projection theorem and to a duality theory for minimum distance problems. Weak and strong duality, and the Hahn-Banach theorem are part of this duality theory. The dual object to a linear operator between vector spaces is the adjoint, which acts between their respective duals. Understanding the interplay between an operator and its adjoint usually provides significant insight into the properties of that operator. The so-called fundamental theorem of linear algebra relates the image and null spaces of an operator and its adjoint. Questions about linear operators can be often more easily answered by understanding the interplay between the actions of the operator and its adjoint.*

Introduction

Recall that in Chapter 1 we defined $\mathbb{R}^n$ as the space of column vectors. We then generalized form column vectors to vectors in abstract vector spaces. What about *row vectors*? What role do they play, and what are the possible generalizations of row vectors?

To explore a bit, fix a particular real row vector $y = \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix}$ (note that this is not an n-tuple $(y_1, \ldots, y_n)$, but a row vector in the standard notation). Now consider the operation of multiplying column vectors by this particular row vector, and call that operation $\mathrm{y}(.)$, i.e.

$$\mathrm{y}(v) \;:=\; y\, v \;=\; \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix} \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} \;=\; \sum_{k=1}^{n} y_k v_k. \tag{4.1}$$

The reader should note the deliberately pedantic choice of fonts in the notation above. y is an operation on column vectors. This operation on any column vector v sums the components of v against the components of a specific row vector y. (4.1) is a particular representation of the operation y. We will have occasion to use different representations such as $\mathrm{y}(v) := y^* v$, where y is a column vector. Thus the careful distinction between an operation and its representations.

The operation (4.1) is a scalar-valued mapping, i.e. $\mathrm{y} : \mathbb{R}^n \to \mathbb{R}$. Since matrix and vector products are distributive over additions, this mapping is linear

$$\mathrm{y}\,(\alpha v_1 + \beta v_2) \;=\; y\;(\alpha v_1 + \beta v_2) \;=\; \alpha\; y v_1 + \beta\; y v_2 \;=\; \alpha\; \mathrm{y}(v_1) + \beta\; \mathrm{y}(v_2).$$

Thus y is a linear operator from $\mathbb{R}^n$ to $\mathbb{R}$. Scalar-valued linear operators are special, so they get their own name, they are called[1] linear functionals. It is not difficult to show (Lemma 4.1) that any linear functionals on $\mathbb{R}^n$ must be of the form (4.1) for some row vector y. Let's temporarily call the space of *row n-vectors* $\mathbb{R}^{n*}$

$$\mathbb{R}^{n*} \;:=\; \left\{ y := \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix};\; y_i \in \mathbb{R} \right\}.$$

Thus the space of all linear functionals on $\mathbb{R}^n$ is $\mathbb{R}^{n*}$. The space of all linear functionals on a vector space V is called the dual space of V, and denoted by V^*, thus the notation $\mathbb{R}^{n*}$ above. Note that $\mathbb{R}^{n*}$ is a vector space with row vector addition. This is true for any vector space V, its dual space V^* is also a vector space.

In the example of $\mathbb{R}^n$, its dual space $\mathbb{R}^{n*}$ is isomorphic to it. The isomorphism is given by the "transpose map" $()^* : \mathbb{R}^n \to \mathbb{R}^{n*}$, $w \mapsto w^*$, which takes a column vector w to a row vector w^*. We will see that this is true for any inner product space, where linear functionals can be generated from the vectors in that space by taking inner products, i.e. for any vector $w \in \mathsf{V}$ in an inner product space V, a linear functional w is defined by

$$\mathrm{w}(v) \;:=\; \langle w\,,\, v \rangle\,, \qquad\qquad v \in \mathsf{V}. \tag{4.2}$$

We will see that in a Hilbert space, all linear functionals are generated in this manner (this is the Riesz representation theorem). Thus a Hilbert space is isomorphic to its dual, and the isomorphism is given by the correspondence (4.2). This is the generalization to Hilbert space of the transpose map that takes column vectors to row vectors. In the absence of an inner product such as in a Banach space, the dual is typically different from the original space, and more care is needed in treating such problems.

What does duality say about operators? A linear operator acting on vectors induces in a natural way another linear operator, called the adjoint, acting on functionals. In the case of $\mathbb{R}^n$, if we act on a column vector v with a matrix A, and then apply a linear functional to the result by multiplying it by a row vector w^*, the result produced is a scalar $w^* A v$. We can now think about A acting on the row vector w^* rather than the column vector v. The mapping $w^* \mapsto w^* A$ takes row vectors to row vectors, so it is a mapping on the dual space $\mathbb{R}^{n*}$. This mapping is determined by the obvious condition

$$\forall v \in \mathbb{R}^n, \qquad \begin{bmatrix} \; w^* \; \end{bmatrix} \begin{bmatrix} \; A \; \end{bmatrix} \begin{bmatrix} v \end{bmatrix} \;=\; \begin{bmatrix} \; w^* A \; \end{bmatrix} \begin{bmatrix} v \end{bmatrix} \;=\; \begin{bmatrix} \; w^* \; \end{bmatrix} \begin{bmatrix} Av \end{bmatrix}. \tag{4.3}$$

The way to read this is to pretend for the moment that we don't know how to multiply a row vector w^* by a matrix, we only know how to multiply a column vector by a matrix. This is the situation in an abstract vector space, the operator on vectors is specified, but we have to do some work to figure out how it acts on functionals. Let's continue with the pretense that we don't know how to multiply row vectors and matrices. Formula (4.3) gives the recipe for finding $w^* A$ from w. Starting from w^* as a known functional on $\mathbb{R}^n$, we need to find the functional $w^* A$. If we know how this functional acts on all vectors v, then it is determined. The formula (4.3) says that $w^* A$ acts on v by first acting on v by Av, and then acting on the result by the known functional w^*. This serves to define the adjoint more abstractly as we now briefly outline.

[1] Any scalar-valued mapping, whether linear or not, is called a functional.

Let $A : \mathsf{V} \to \mathsf{W}$ be a linear operator between vector spaces. The adjoint $A^\dagger : \mathsf{W}^* \to \mathsf{V}^*$ maps their duals (i.e. maps linear functionals to linear functionals). The adjoint is determined by the following condition which generalizes (4.3)

$$\forall v \in \mathsf{V}, \qquad \left(A^\dagger \mathrm{w}\right)(v) \;:=\; \mathrm{w}\left(Av\right). \tag{4.4}$$

Let's parse this carefully. w is a functional on W, i.e. its in W^*. It is mapped by $A^\dagger$ to an element in V^*, i.e. a functional on V. The definition above specifies how $A^\dagger \mathrm{w}$ acts on every $v \in \mathsf{V}$, and therefore is a mapping $\mathrm{w} \mapsto A^\dagger \mathrm{w}$ from W^* to V^*.

The adjoint is a linear operator that is intimately related to the original operator. Concepts like row rank and left null space of a matrix are best understood as statements about the adjoint. Much of linear algebra and functional analysis involves the interplay between an operator and its adjoint. We now turn to developing these concepts more precisely.

4.1 Dual Vectors: The Dual Space

Let V be any vector space and let $\mathrm{w} : \mathsf{V} \to \mathbb{R}$ be a *linear functional*, i.e. a scalar-valued linear operator on V

$$\mathrm{w}\big(\alpha v_1 + \beta v_2\big) \;=\; \alpha\; \mathrm{w}\big(v_1\big) \;+\; \beta\; \mathrm{w}\big(v_2\big). \tag{4.5}$$

We denote the set of all linear functionals on V with the symbol V^*. This set inherits a vector space structure from V. Any two elements w_1 and w_2 in V^* can be scaled and added by the standard definition for functions (point-wise addition and scaling)

$$\big(a\mathrm{w}_1 + b\mathrm{w}_2\big)(v) \;:=\; a\; \mathrm{w}_1\big(v\big) \;+\; b\; \mathrm{w}_2\big(v\big). \tag{4.6}$$

It is immediate that $\mathrm{w}_1 + \mathrm{w}_2$ thus defined is also a linear functional

$$\begin{aligned}
\big(a\mathrm{w}_1 + b\mathrm{w}_2\big)\big(\alpha v_1 + \beta v_2\big) &= a\; \mathrm{w}_1\big(\alpha v_1 + \beta v_2\big) \;+\; b\; \mathrm{w}_2\big(\alpha v_1 + \beta v_2\big) && \text{(by (4.6))}\\
&= a\alpha\; \mathrm{w}_1(v_1) + a\beta\; \mathrm{w}_1(v_2) \;+\; b\alpha\; \mathrm{w}_2(v_1) + b\beta\; \mathrm{w}_2(v_2) && \text{(by (4.5))}\\
&= \alpha\; \big(a\mathrm{w}_1 + b\mathrm{w}_2\big)(v_1) \;+\; \beta\; \big(a\mathrm{w}_1 + b\mathrm{w}_2\big)(v_2) && \text{(by (4.6))}
\end{aligned}$$

Thus the set of all linear functionals V^* is itself a vector space.

In finite-dimensional vector spaces, any basis representation gives a representation of linear functionals as the product with row vectors as follows.

Lemma 4.1. *Let V be a finite-dimensional vector space. Given a particular basis $\mathsf{v} = \{\boldsymbol{v}_k\}_{k=1}^n$ of V, the action of a linear functional $\mathrm{w} : \mathsf{V} \to \mathbb{R}$ on any vector $\boldsymbol{u}$ can be written as a* dot product *of the vector $[\boldsymbol{u}]_\mathsf{v} =: (u_1, \ldots, u_n)$ of basis coefficients with a row vector w^*, which we call the* representation of w in the basis v

$$\mathrm{w}(\boldsymbol{u}) \;=\; w^*\,[\boldsymbol{u}]_\mathsf{v} \;=\; \begin{bmatrix} w_1 & \cdots & w_n \end{bmatrix} \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} \;=\; \sum_{k=1}^n w_k u_k,$$

where for each k, $w_k = \mathrm{w}\left(\boldsymbol{v}_k\right)$.

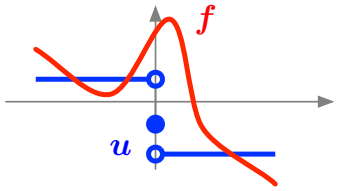
Proof. Consider the action of w on each of the basis elements

$$\mathrm{w}(\boldsymbol{v}_k) \;=:\; w_k, \qquad k = 1, \ldots, n.$$

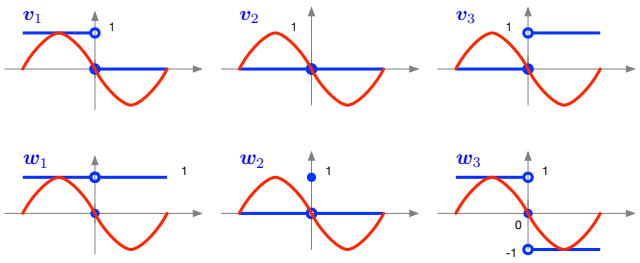
By linearity, the action of w on any vector is given by

$$\mathrm{w}(v) \;=\; \mathrm{w}\left(\sum_{k=1}^n u_k \boldsymbol{v}_k\right) \;=\; \sum_{k=1}^n u_k\; \mathrm{w}(\boldsymbol{v}_k) \;=\; \sum_{k=1}^n u_k\; w_k$$

□



(a) Any integrable function f defines a linear functional f on the space of piece-wise constant (on the interval [-1, 1]) functions by the integral $\mathrm{f}(u) := \int_{-1}^{1} f(t)u(t)dt$.



(b) Here $f(t) = -\sin((\pi t)$, and the basis representation of the functional f (in either of the v or the w bases) is given by integrating the function f against the respective bases elements.

Figure 4.1: Any (absolutely) integrable function f defines a linear functional f on a space of functions by $\mathrm{f}(u) := \int f(t)u(t)dt$ provided the integral converges. The basis representation of f in terms of any basis is obtained by acting with f on each of the individual basis elements.

Note that the lemma above says nothing about inner products or norms. It is a purely algebraic statement about vector spaces.

Example 4.2. Consider the vector space $\mathbb{R}^{\{[-1,0),0,(0,1]\}}$ of functions on $[-1, 1]$ that are constant on the sets $[-1, 0), 0, (0, 1]$ introduced in Example 1.4.

Given any (absolutely) integrable function $f : [-1, 1] \to \mathbb{R}$, we can use it to define a linear functional f by defining its action on any $u \in \mathbb{R}^{\{[-1,0),0,(0,1]\}}$ using the integral

$$\mathrm{f}(u) \;:=\; \int_{-1}^{1} f(t)\; u(t)\; dt. \tag{4.7}$$

This is illustrated in Figure 4.1a. We say that that function $f : [-1, 1] \to \mathbb{R}$ is the *kernel representation*[2] of the functional $\mathrm{f} : \mathbb{R}^{\{[-1,0),0,(0,1]\}} \to \mathbb{R}$.

Now recall the two bases v and w used earlier for this function space, and depicted again in Figure 4.1b. If we choose for example $f(t) := -\sin(\pi t)$, then as shown in the figure, the row vectors f^{v} and f^{w} representing f in the bases v and w respectively are

$$f^{\mathsf{v}} = \begin{bmatrix} f_1^{\mathsf{v}} & f_2^{\mathsf{v}} & f_3^{\mathsf{v}} \end{bmatrix} = \begin{bmatrix} \frac{2}{\pi} & 0 & -\frac{2}{\pi} \end{bmatrix}, \qquad f^{\mathsf{w}} = \begin{bmatrix} f_1^{\mathsf{w}} & f_2^{\mathsf{w}} & f_3^{\mathsf{w}} \end{bmatrix} = \begin{bmatrix} 0 & 0 & \frac{4}{\pi} \end{bmatrix}, \tag{4.8}$$

where each vector component is calculated by acting with f (4.7) on the respective basis element as in Lemma 4.1. For example

$$f_1^{\mathsf{v}} \;:=\; \mathrm{f}(\boldsymbol{v}_1) \;=\; \int_{-1}^{1} f(t)v(t)\; dt \;=\; -\int_{-1}^{0} \sin(\pi t)\; dt \;=\; \frac{1}{\pi}\cos(\pi t)\Big|_{-1}^{0} \;=\; \frac{2}{\pi}.$$

Remark 4.3. While for most function spaces, all linear functionals have a representation like (4.7), this is not so in the above example. This is due to the peculiarity of the function space $\mathbb{R}^{\{[-1,0),0,(0,1]\}}$ where the value $u(0)$ of an element at the single point $t = 0$ matters. Note that in (4.8), both f_2^{v} and f_2^{w} are zero. This is actually true for any (regular) function f that defines a linear functional by the integral (4.7)

$$f_2^{\mathsf{v}} \;:=\; \int_{-1}^{1} f(t)v_2(t)\; dt \;=\; \int_{0}^{0} f(t)\; dt \;=\; 0.$$

It turns out that while all functions f (provided they're integrable) define linear functionals by (4.7), not all linear functionals on this space are of that form. This problem is

[2] This is a special case of the *kernel representation of linear operators* discussed in Chapter 6.

easy to fix as follows. Define a linear functional by

$$\mathsf{f}(u) \;:=\; \int_{-1}^{1} f(t)\; u(t)\; dt \;+\; \bar{f}\; u(0), \tag{4.9}$$

where $\bar{f}$ is some scalar. Thus the functional f requires specifying two pieces, a function $f : [-1, 1] \to \mathbb{R}$, as well as a scalar $\bar{f}$. In this case the number $\bar{f}$ would also be equal to f_2^{v} and f_2^{w}. We can now conclude (e.g. by counting dimensions) that all linear functionals on $\mathbb{R}^{\{[-1,0),0,(0,1]\}}$ are of the form (4.9). ■

Formal Definition and Further Examples

We now give a formal definition of the dual space. The examples we've seen so far are for what is called the *algberaic dual* space since norms played no role in the discussion. However, when a vector space is equipped with a norm, then the dual space also has a natural norm given for each functional by its induced norm as a linear operator. This is sometimes called the *topological dual* space, but we will simply refer to it as the *dual space*.

Definition 4.4. *Let* V *be a Banach space. Its* dual space V^* *is the space of bounded (continuous) linear functionals*

$$\mathsf{V}^* \;:=\; \left\{ f : \mathsf{V} \to \mathbb{R};\; f \text{ is linear and, } \|f\| := \sup_{\|v\|=1} |f(v)| < \infty \right\},$$

i.e. $\mathsf{V}^* = \mathsf{L}(\mathsf{V}, \mathbb{R})$*, and therefore is itself a Banach space with the induced norm.*

Note that $\mathsf{V}^* = \mathsf{L}(\mathsf{V}, \mathbb{R})$, and that we've already shown in Section 3.6 that $\mathsf{L}(\mathsf{V}, \mathsf{W})$ is a Banach space (i.e. complete) with the induced norm whenever V and W are Banach spaces. $\mathbb{R}$ is a Banach space, and therefore $\mathsf{V}^* = \mathsf{L}(\mathsf{V}, \mathbb{R})$ is a Banach space.

Example 4.5. Consider the vector space $\mathbb{R}^n$. By Lemma 4.1 every linear functional w is of the form

$$\mathrm{w}(v) \;=\; w_1 v_1 + \cdots + w_n v_n \;=\; w^* v. \tag{4.10}$$

If we endow $\mathbb{R}^n$ with the Euclidean $\|.\|_2$ norm (call the space $\mathbb{R}^n_2$), what is the induced norm on w? First observe that by the Cauchy-Schwartz inequality

$$|\mathrm{w}(v)| \;=\; |w^* v| \;\leq\; \|w\|_2 \|v\|_2.$$

Thus the Euclidean norm $\|w\|_2$ of the vector w is an upper bound on the induced norm of the functional w. This upper bound is achieved by applying w to the vector w itself

$$\mathrm{w}(w) \;=\; w_1^2 + \cdots + w_n^2 \;=\; \|w\|_2^2.$$

We therefore conclude that the induced norm on $\mathbb{R}^n_2$ of the functional w in (4.10) is the Euclidean norm of the row vector w representing it. Therefore the dual of $\mathbb{R}^n_2$ is $\mathbb{R}^{n*}_2$. Since $\mathbb{R}^{n*}_2$ and $\mathbb{R}^n_2$ are isometrically isomorphic, we will often just say that the dual of $\mathbb{R}^n_2$ is $\mathbb{R}^n_2$ when the isomorphism is implicitly understood.

This example is a special case of a fact true in any Hilbert space. We say that the vector $w \in \mathsf{V}$ "represents" the functional $\mathrm{w} \in \mathsf{V}^*$ by the inner product as $\mathrm{w}(v) = \langle w\,,\, v \rangle$. The fact that every element of V^* in a Hilbert space is represented this way is the Riesz Representation Theorem 4.10. This will imply that the dual of a Hilbert space V is itself, or more precisely, isometrically isomorphic to V. More on this in the next subsection.

Example 4.6. Consider $\mathbb{R}^n_\infty$ with the $\|.\|_\infty$ norm and functionals acting by

$$\mathrm{w}(v) \;=\; w_1v_1 + \cdots + w_nv_n \;=\; w^*v. \tag{4.11}$$

The induced norm can be calculated using the 1-∞ inequality (Exercise 2.5)

$$\begin{aligned}\mathrm{w}(v) \;&=\; \sup_{\|v\|_\infty=1} |w^*v| \;=\; \sup_{\|v\|_\infty=1} |w_1v_1 + \cdots + w_nv_n| \\ &\le\; \sup_{\|v\|_\infty=1} |w_1v_1| + \cdots + |w_nv_n| \;\le\; \sup_{\|v\|_\infty=1} \left(\sum_{i=1}^n |w_i|\right)\left(\max_{1\le j\le n} |v_j|\right) \\ &=\; \|w\|_1,\end{aligned}$$

with equality achieved by using $v_i = \mathrm{sign}(w_i)$.

Thus the dual of $\mathbb{R}^n_\infty$ is $\mathbb{R}^n_1$ when the action of functionals is given by (4.11). We leave it as an exercise (Exercise 4.1) to show that the dual of $\mathbb{R}^n_1$ is $\mathbb{R}^n_\infty$, and more generally, the dual of $\mathbb{R}^n_p$ is $\mathbb{R}^n_q$ when $1/p + 1/q = 1$.

Example 4.7. If a Banach space V has a basis $\mathrm{v} = \{\boldsymbol{v}_k\}_{k=0}^\infty$, then every linear functional w must be of the form

$$\mathrm{w}(u) \;=\; \sum_{k=0}^\infty w_k u_k, \qquad \text{where} \qquad \boldsymbol{u} = \sum_{k=0}^\infty u_k \boldsymbol{v}_k, \qquad w_k := \mathrm{w}(\boldsymbol{v}_k). \tag{4.12}$$

Thus the sequence $\{w_k\}$ "represents" the functional w. The summability properties of the sequence $\{w_k\}$ will depend on both the basis set v as well as the norm in the space V. The statement (4.12) is a purely algebraic statement, and is the same as Lemma 4.1 irrespective of whether the space is finite or infinite dimensional. In finite dimensions, the finite set of numbers $\{w_k\}$ can be anything, while in infinite-dimensions, restrictions on the sequence have to imposed. Those restrictions depend on the norm in V as well as the particular choice of basis v.

Example 4.8. Let's calculate the dual of $\ell^1(\mathbb{R})$, but while keeping the example of $\mathbb{R}^n_1$ in mind since the calculations are analogous. Write any element of $\ell^1(\mathbb{R})$ in the canonical basis (i.e. $u = (u_0, u_1, \ldots)$, then by (4.12) every linear functional is of the form

$$\mathrm{w}(u) \;=\; \sum_{k=0}^\infty w_k u_k,$$

for some sequence $\{w_k\}_{k=0}^\infty$. Now let's see what the requirement (in Definition 4.4) that w be a *bounded* linear functional imply about the sequence $\{w_k\}$. A bound can be given using the 1-∞ inequality

$$\left|\sum_{k=0}^\infty w_k u_k\right| \;\le\; \sum_{k=0}^\infty |w_k|\,|u_k| \;\le\; \left(\sup_k |w_k|\right)\left(\sum_{k=0}^\infty |u_k|\right) \;=\; \|w\|_\infty\, \|u\|_1, \tag{4.13}$$

from which we conclude that the induced norm is bounded by

$$\sup_{u\neq 0} \frac{\mathrm{w}(u)}{\|u\|_1} \;\le\; \|w\|_\infty.$$

To show that this bound is tight, consider two separate cases. The first is if the sequence w achieves its supremum at some finite index $\bar{k}$, we then choose $\bar{u} = e_{\bar{k}}$. With this choice $\|\bar{u}\|_1 = \|e_{\bar{k}}\|_1 = 1$ and

$$\left|\sum_{k=0}^\infty w_k \bar{u}_k\right| \;=\; |w_{\bar{k}}| \;=\; \|w\|_\infty.$$

The other case is when the supremum of w is not achieved, but from the definition of the supremum we know that for any $\epsilon > 0$, $\exists \bar{k}$ such that $|w_{\bar{k}}| \leq \|w\|_\infty - \epsilon$. Choosing $\bar{u} = e_{\bar{k}}$ again

$$\left| \sum_{k=0}^{\infty} w_k \bar{u}_k \right| \;=\; |w_{\bar{k}}| \;=\; \|w\|_\infty - \epsilon.$$

Since ϵ can be made arbitrarily small, the bound is tight and $\|\mathrm{w}\| = \|w\|_\infty$, i.e. the norm of the functional w is given by the $\|.\|_\infty$ norm of its representing sequence $\{w_k\}$.

We therefore conclude that the dual of $\ell^1(\mathbb{N})$ is $\ell^\infty(\mathbb{N})$.

Example 4.9. Recall that the dual of $\mathbb{R}^n_1$ is $\mathbb{R}^n_\infty$ and vice versa. In light of the previous example of $\ell^1(\mathbb{N})$, we might suspect that the dual of $\ell^\infty(\mathbb{N})$ is $\ell^1(\mathbb{N})$. However, this is not quite right. Every element of ℓ^1 defines a bounded linear functional on ℓ^∞ by the same argument as (4.13), but there are other bounded functionals on ℓ^∞ that cannot be written in the form (4.12). This form was premised on the Banach space having a basis, and recall that $\ell^\infty(\mathbb{N})$ is not separable, and therefore cannot have a basis.

It is possible to show that the dual of $\ell^\infty(\mathbb{N})$ is strictly larger than $\ell^1(\mathbb{N})$, i.e. $\ell^1(\mathbb{N}) \subsetneq (\ell^\infty(\mathbb{N}))^*$ using the so-called Hahn-Banach theorem. However, the argument is non-constructive and one cannot exhibit those other elements (those not in $\ell^1(\mathbb{N})$) explicitly.

On the other hand, recall the closed subspace $\ell^\infty_{\mathrm{o}}(\mathbb{N})$ of $\ell^\infty(\mathbb{N})$ made up of sequences that decay to zero. This is a Banach space in itself, and in fact we can show that $(\ell^\infty_{\mathrm{o}}(\mathbb{N}))^* = \ell^1(\mathbb{N})$. The argument has essentially already been presented. Since $\{e_k\}$ is a basis for $\ell^\infty_{\mathrm{o}}(\mathbb{N})$, then any linear functional is represented by summing against a sequence, and the argument (4.12) says that the functionals induced norm (over ℓ^∞_{i}) is precisely the ℓ^1 norm of the sequence.

The take away from the above is that if the $\|.\|_\infty$ norm of sequences is needed (e.g. in an optimization problem), it is preferable whenever possible to set the problem up in ℓ^∞_{o} rather than ℓ^∞.

The Hilbert Dual

In a Hilbert space V every vector defines a functional by taking its inner product with other vectors. Let $w \in \mathsf{V}$ be any vector. Define the functional w from w by

$$\mathrm{w}(v) \;:=\; \langle w\,,\, v\rangle\,, \qquad\qquad v \in \mathsf{V}. \tag{4.14}$$

The functional w defined here is clearly linear. It is also bounded as follows from the Cauchy-Schwartz inequality $\|\mathrm{w}(v)\| = |\langle w\,,\, v\rangle| \leq \|w\|\|v\|$. Thus w(.) defined above is a bounded linear functional on V

Note that in (4.14) we are again using fonts to emphasize a distinction. w is a functional, i.e. an element in V^*, while w is a vector in V, so they are different (but obviously related) objects. Equation (4.14) says that every vector in V defines a functional in V^*. The question is whether every functional in V^* can be represented this way? In other words, given $\mathrm{w} \in \mathsf{V}^*$, does there exist a vector $v \in \mathsf{V}$ such that $\mathrm{w}(.) = \langle w\,,\, .\rangle$?

The answer is yes, and the construction is depicted geometrically in Figure 4.2. The key idea is that every linear functional is uniquely (up to scaling) determined by its null space. The reason is that a linear functional w is a mapping $\mathrm{w} : \mathsf{V} \to \mathbb{R}$. Its image $\mathsf{Im}(\mathrm{w}) = \mathsf{V}/\mathsf{Nu}(\mathrm{w})$ is isomorphic to $\mathbb{R}$, and therefore $\mathsf{Nu}(\mathrm{w})$ is a co-dimension 1 subspace. Such a subspace is uniquely determined by the *direction* of vectors orthogonal to it. To pick the "size" of the orthogonal vector, we need an appropriate normalization as follows. Take

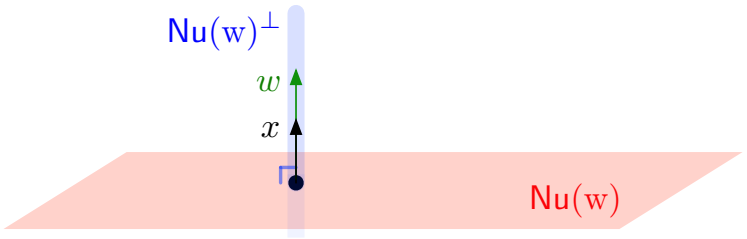


Figure 4.2: The construction of the Riesz representation theorem. Given an abstractly defined linear functional w, we need to find a vector w that represents w through the inner product, i.e. $\mathrm{w}(v) = \langle w\,,\, v\rangle$ for all $v \in \mathsf{V}$. The key is that the null space of w (which is a co-dimension 1 subspace) uniquely defines the direction of the vector w which must be orthogonal to it. Thus pick any vector x orthogonal to $\mathsf{Nu}(\mathrm{w})$, renormalize it appropriately by (4.15) to find the representative w.

any vector $x \in \mathsf{Nu}(\mathrm{w})^\perp$ and normalize it by (note that it is unique after normalization)

$$w \;:=\; \frac{\mathrm{w}(x)}{\|x\|^2}\,x \qquad \Rightarrow \qquad \|w\| \;=\; \frac{|\mathrm{w}(x)|}{\|x\|} \tag{4.15}$$

$$\mathrm{w}(w) \;=\; \mathrm{w}\left(\frac{\mathrm{w}(x)}{\|x\|^2}\,x\right) \;=\; \frac{\mathrm{w}(x)}{\|x\|^2}\,\mathrm{w}(x) \;=\; \|w\|^2 \;=\; \langle w\,,\, w\rangle\,. \tag{4.16}$$

Now since $\langle w\,,\, \mathsf{Nu}(\mathrm{w})\rangle = 0$, $\langle w\,,\, w\rangle = \mathrm{w}(w)$, and $\mathsf{V} = \mathsf{Nu}(\mathrm{w}) \oplus \mathsf{span}\{w\}$, then $\langle w\,,\, .\rangle = \mathrm{w}(.)$ on all of V, and therefore w is the representative of the functional w. We now state these conclusions formally.

Theorem 4.10 (Riesz Representation)**.** *If* V *is a Hilbert space, then every bounded functional* $\mathrm{w} \in \mathsf{V}^*$ *is* represented *by an inner product with a unique vector* $w \in \mathsf{V}$*, i.e.*

$$\forall v \in \mathsf{V},\quad \mathrm{w}(v) = \langle w\,,\, v\rangle\,. \qquad\qquad \textit{Furthermore} \qquad \|\mathrm{w}\| = \|w\|.$$

Note that $\|\mathrm{w}\|$ is the induced norm of w as a *functional* on V, while $\|w\|$ is the *vector norm* in V of its representative w. Their equality follows from the Cauchy-Schwartz inequality

$$|\mathrm{w}(v)| \;=\; |\langle w\,,\, v\rangle| \;\le\; \|w\|\,\|v\| \qquad \Rightarrow \qquad \sup_{v\neq 0}\frac{|\mathrm{w}(v)|}{\|v\|} \;=\; \sup_{v\neq 0}\frac{|\langle w\,,\, v\rangle|}{\|v\|} \;\le\; \|w\|,$$

and observing that this upper bound is achieved with $v = w$.

The Riesz representation theorem generalizes Example 4.5 which dealt with $\mathbb{R}^n$ when equipped with the standard Euclidean norm. The next example considers $\mathbb{R}^n$ equipped with a different inner product, so the above theorem still holds, but it needs to be interpreted carefully as we will demonstrate.

Example 4.11. Consider $\mathbb{R}^n_Q$, which is defined as $\mathbb{R}^n$ with a "weighted norm"

$$\|v\|_Q^2 \;=\; \langle v\,,\, v\rangle_Q \;:=\; v^*Qv \;=\; v^*Q^{\frac{1}{2}}Q^{\frac{1}{2}}v \;=:\; \left\langle Q^{\frac{1}{2}}v\,,\, Q^{\frac{1}{2}}v\right\rangle_2 \;=\; \|Q^{\frac{1}{2}}v\|_2^2,$$

where Q is a symmetric positive definite matrix, $\langle\,,\,\rangle_2$ is the Euclidean inner product and $\|.\|_2$ is the Euclidean norm.

The dual space is still isomorphic to $\mathbb{R}^n$, but what is the norm on the dual space? We have to be careful here with how we define linear functionals, because their resulting norms will depend on that definition. First, since $\mathbb{R}^n_Q$ is n-dimensional, we can take the canonical basis and Lemma 4.1 says (c.f. Example 4.5) that every linear functional w is represented by an n-vector w so that

$$\mathrm{w}(v) \;=\; w_1v_1 + \cdots + w_nv_n \;=\; w^*v, \qquad\qquad v \in \mathbb{R}^n_Q. \tag{4.17}$$

We can now calculate the induced norm as

$$\begin{aligned}\|\mathrm{w}(v)\|^2 &:= \sup_{v\neq 0}\frac{|w^*v|^2}{\|v\|_Q^2} = \sup_{v\neq 0}\frac{|w^*v|^2}{v^*\,Q^{1/2}Q^{\frac{1}{2}}\,v}\\ &= \sup_{u\neq 0}\frac{\left|w^*Q^{-1/2}\,u\right|^2}{u^*u} && \text{(substituting } u = Q^{1/2}\,v\text{)}\\ &= \sup_{u\neq 0}\frac{\left|\left(Q^{-1/2}\,w\right)^*u\right|^2}{u^*u} = \frac{\left\langle Q^{-1/2}w\,,\,u\right\rangle_2}{\langle u\,,\,u\rangle_2} = \left\|Q^{-1/2}\,w\right\|_2^2, && (4.18)\end{aligned}$$

Thus if the norm on $\mathbb{R}^n$ is given by $\|Q^{1/2}x\|_2$, and the functional action is given by $v \mapsto w^*v$, then the norm on this linear functional is given by $\left\|Q^{-1/2}w\right\|_2$. In other words, the dual of $\mathbb{R}^n_Q$ is $\mathbb{R}^n_{Q^{-1}}$.

At first glance this might seem to not be consistent with Thereom 4.10 which states that $\|\mathrm{w}\| = \|w\|$. However, there is no inconsistency if the theorem is interpreted correctly as follows. In the space $\mathbb{R}^n_Q$, the theorem says that any linear functional is given by

$$\mathrm{w}_1(v) = \langle w\,,\,v\rangle_Q = w^*Qv = (Qw)^*\,v. \tag{4.19}$$

This is a different functional w from w defined in (4.17)! This functional acts on a vector v by summing its components against components of the vector Qw rather than the vector w. According to the theorem, the induced norm of w_1 must be

$$\|\mathrm{w}_1\| = \|w\|_Q = \left\|Q^{1/2}w\right\|_2. \tag{4.20}$$

The two functionals w_1 and w can be related by defining another functional u

$$u := Qw, \qquad \mathrm{u}(v) := u^*v = \mathrm{w}_1(v)$$

The norm of u calculated according to (4.18) is the same as the norm of w_1 calculated according to (4.20)

$$\begin{aligned}\|\mathrm{u}\| &= \left\|Q^{-1/2}u\right\|_2 && \text{(by (4.18) since } \mathrm{u}(v) := u * v\text{)}\\ &= \left\|Q^{-1/2}Qw\right\|_2 = \left\|Q^{1/2}w\right\|_2 && \text{(since } u := Qw\text{)}\\ &= \|\mathrm{w}_1\| && \text{(by (4.20))}\end{aligned}$$

Therefore the calculations are consistent provided we apply the Riesz representation theorem correctly. This issue is a source of potential confusion whenever working with weighted norms in Hilbert space.

4.2 Duality and Orthogonality

In this section we will generalize the notion of orthogonality to Banach spaces using duality. It turns out that the best way to generalize the notion of orthogonality is to abandon the idea that orthogonality is between vectors in the same space. Instead, and more generally, orthogonality should be thought of as between *a functional and a vector*, which are objects that live in different spaces. It just happens that in Hilbert space, each functional is *represented* by taking an inner product with a particular vector. Thus we were lured into thinking of orthogonality as between vectors. A mental shift to a more general notion of

orthogonality will have great benefits. Many useful constructions in Hilbert space (such as the projection theorem) can generalize to a Banach space V, but now we have to think about V and its dual V^* simultaneously.

Before we begin, we make a note about a notational change. From this section on, linear functionals will be denoted with brackets like

$$v \in \mathsf{V},\ \mathrm{w} \in \mathsf{V}^*, \qquad \langle \mathrm{w}\,,\, v\rangle \ := \ \mathrm{w}(v),$$

which is the more traditional notation for functional action. To emphasize that this is not an inner product, but rather functional action, we will denote functionals (like w) with roman font, while vectors (like v) with italic font. In later chapters, after the reader has become accustomed to the distinctions between vectors and functionals, we will drop the distinction in fonts.

Orthogonality

In an inner product space, the two geometrical notions of orthogonality and alignment are defined in terms of the inner product. Those two notions generalize to normed spaces, but they are between vectors and functionals rather than between vectors and other vectors.

Definition 4.12. *A vector $v \in \mathsf{V}$ in a Banach space V and a functional $\mathrm{w} \in \mathsf{V}^*$ are said to be* orthogonal *if*

$$\langle \mathrm{w}\,,\, v\rangle \ = \ 0.$$

Given a subspace $\mathsf{S} \subset \mathsf{V}$, its orthogonal subspace[3] *$\mathsf{S}^\perp \subset \mathsf{V}^*$ is*

$$\mathsf{S}^\perp \ := \ \{\mathrm{w} \in \mathsf{V}^*;\ \forall v \in \mathsf{S},\ \ \langle \mathrm{w}\,,\, v\rangle = 0\}$$

The terminology and notation are suggestive, but should be parsed carefully. $\langle \mathrm{w}\,,\, v\rangle$ is the functional w acting on the vector v rather than an inner product. We use the term "orthogonal subspace" since $\mathsf{S}^\perp$ is in the dual space V, rather than "orthogonal complement" where $\mathsf{S}^\perp$ in a Hilbert space is a complementary subspace to S. None the less, in Banach spaces the orthogonal subspace $\mathsf{S}^\perp \subset \mathsf{V}^*$ plays a similar role to the orthogonal complement $\mathsf{S}^\perp \subset \mathsf{V}$ in Hilbert spaces. An easy, but important observation is that the orthogonal subspace is alway closed even if the original subspace is not.

Lemma 4.13. *If $\mathsf{S} \subset \mathsf{V}$ is any subspace of a Banach space, then its orthogonal subspace $\mathsf{S}^\perp \subset \mathsf{V}^*$ is a closed subspace in V^*.*

Proof. If the sequence $\{\mathrm{w}_k\} \subset \mathsf{S}^\perp$ has a limit $\lim_{k\to\infty} \mathrm{w}_k = \mathrm{w} \in \mathsf{V}^*$, then

$$\forall v \in \mathsf{S}, \qquad \langle \mathrm{w}\,,\, v\rangle \ = \ \left\langle \lim_{k\to\infty} \mathrm{w}_k\,,\, v \right\rangle \ = \ \lim_{k\to\infty} \langle \mathrm{w}_k\,,\, v\rangle \ = \ 0,$$

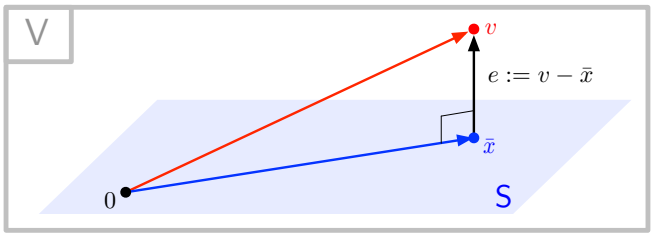
because the mapping $\langle .\,,\, v\rangle : \mathsf{V} \to \mathbb{R}$ is continuous. Thus $\mathrm{w} \in \mathsf{S}^\perp$. □

Definition 4.14. *In a Banach space V, a vector $v \in \mathsf{V}$ and a functional $\mathrm{w} \in \mathsf{V}^*$ are said to be* aligned *if*

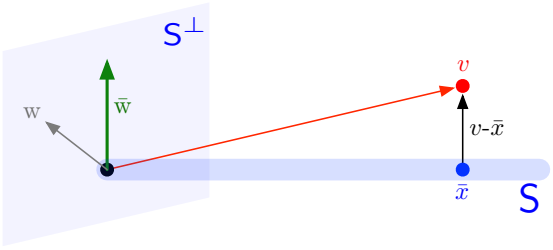
$$\langle \mathrm{w}\,,\, v\rangle \ = \ \|\mathrm{w}\|\ \|v\|.$$

This is very similar to the definition of alignment in an inner product space where two vector v, w are aligned if there is equality in the Cauchy-Schwartz inequality $\langle w\,,\, v\rangle = \|w\|\|v\|$. In a Banach space, we replace inner products with functional actions.

[3] This is alternatively termed "the annihilator" in many references.



(a) Depiction of a minimum distance problem in a Hilbert space. The point $\bar{x} \in \mathsf{S}$ is the point in S closest to v. The length of the optimal error vector $v - \bar{x}$ is the minimum distance between v S. The projection theorem says that the optimal error $v - \bar{x}$ must be *orthogonal* to the subspace S.



(b) Another geometrical interpretation of the projection theorem that is valid in either Hilbert or Banach space. We can't say that the optimal error vector $v - \bar{x}$ is orthogonal to S, since now orthogonality is between vectors and functionals, not between vectors and vectors. However, we can say that it must be "aligned" with a functional $\bar{\mathrm{w}} \in \mathsf{S}^\perp$ orthogonal to S, i.e. $\langle \bar{\mathrm{w}}, v - \bar{x} \rangle = \|\bar{\mathrm{w}}\| \|v - \bar{x}\|$. This special functional $\bar{\mathrm{w}}$ is characterized by a dual optimization problem in $\mathsf{S}^\perp \subset \mathsf{V}^*$.

Figure 4.3: A comparison of the projection theorem in a Hilbert space (top) with what might be its counterpart in a Banach space (bottom).

Now recall the problem of minimum distance between a vector and a subspace. In Hilbert space, such problems are addressed by the projection theorem. What would a counterpart of the projection theorem be like in Banach space? Figure 4.3 gives some geometrical intuition to help answer this question. The key point in the projection theorem is that an optimal error vector $v - \bar{x}$ (see Figure 4.3a) must be *orthogonal* to the subspace S. We can't make this statement in Banach space since orthogonality is between a functional and a vector. The statement $(v - \bar{x}) \perp \mathsf{S}$ does not make sense since $v - \bar{x}$ is a vector, not a functional. Now bring in the concept of alignment, and we can say that *the error vector $v - \bar{x}$ must be aligned with a functional* $\mathrm{w} \in \mathsf{S}^\perp$. This is how we can say that $v - \bar{x}$ is "orthogonal" to S.

Now the next question is which functional $\mathrm{w} \in \mathsf{S}^\perp$ is the error vector $v - \bar{x}$ aligned with? It turns out that we have to solve an optimization problem in $\mathsf{S}^\perp \subset \mathsf{V}^*$ (i.e. in the dual space) to find those special vectors. This is the subject of dual optimization problems to which we now turn.

Minimum Distance Problems: Weak Duality

We begin by establishing an easy (but very useful) inequality usually referred to as *weak duality*. First, let $v, x \in \mathsf{V}$ be vectors in a Banach space, and $\mathrm{w} \in \mathsf{V}^*$ be a functional, then by definition of $\|\mathrm{w}\|$

$$\|\mathrm{w}\| \, \|v - x\| \;\geq\; \langle \mathrm{w}, v - x \rangle$$

Now let $x \in \mathsf{S}$, a subspace of V, and $v \in \mathsf{V}$ a vector possibly outside of S. Furthermore, restrict $\|\mathrm{w}\| \leq 1$ and to be in the subspace orthogonal to S. Then

$$x \in \mathsf{S},\ \mathrm{w} \in \mathsf{S}^\perp,\ \|\mathrm{w}\| \leq 1 \qquad \Rightarrow \qquad \|v - x\| \;\geq\; \langle \mathrm{w}, v - x \rangle = \langle \mathrm{w}, v \rangle - \cancelto{0}{\langle \mathrm{w}, x \rangle} = \langle \mathrm{w}, v \rangle .$$

(4.21)

Note that the right side of the inequality is now independent of x, while the left side is independent of w. Now if we take the infimum on the left and supremum on the right, the inequality is preserved

$$\inf_{x\in\mathsf{S}} \|v-x\| \;\geq\; \sup_{\mathrm{w}\in\mathsf{S}^\perp,\ \|\mathrm{w}\|\leq 1} \langle \mathrm{w}\,,\, v\rangle\,. \tag{4.22}$$

This is the so-called "weak duality" statement which relates a minimization problem (here called the "primal problem") to a dual, maximization problem in the dual space V^*. The two problems may not always have equal optimal objectives, but the inequality above is always valid.

In many cases (such as Theorem 4.20 below), equality in (4.22) is achieved. In fact, the derivation above gives us a very useful criterion for equality of the two problems. Suppose we find a vector $\bar{x}$ and a functional $\bar{\mathrm{w}}$ such that equality in (4.21) is achieved, i.e.

$$\|\bar{\mathrm{w}}\|\,\|v-\bar{x}\| \;=\; \langle \bar{\mathrm{w}}\,,\, v\rangle\,.$$

Then $\bar{x}$ and $\bar{\mathrm{w}}$ must be the solutions to the two problems in (4.22) respectively! Note that this condition is an alignment condition (recall Definition 4.14), and can be very useful in explicit calculations. These conclusions, while simple to derive, are important enough to state precisely.

Theorem 4.15 (Weak Duality)**.** *Let $\mathsf{S}\subset\mathsf{V}$ be a subspace of a Banach space V, and $\mathsf{S}^\perp\subset\mathsf{V}^*$ its orthogonal subspace. Then the primal and dual optimization problems are related by*

$$\mathsf{d_p} \;:=\; \inf_{x\in\mathsf{S}} \|v-x\| \;\geq\; \sup_{\mathrm{w}\in\mathsf{S}^\perp,\ \|\mathrm{w}\|\leq 1} \langle \mathrm{w}\,,\, v\rangle \;=:\; \mathsf{d_d}. \tag{4.23}$$

If there exists $\bar{x}\in\mathsf{S}$ and $\bar{\mathrm{w}}\in\mathsf{S}^\perp$ such that the functional $\bar{\mathrm{w}}$ is "aligned" with the error $v-\bar{x}$

$$\|\bar{\mathrm{w}}\|\,\|v-\bar{x}\| \;=\; \langle \bar{\mathrm{w}}\,,\, v\rangle\,, \tag{4.24}$$

then $\bar{x}$ and $\bar{\mathrm{w}}$ are optimal for the primal and dual problems respectively, and $\mathsf{d_p}=\mathsf{d_d}$.

The dual problems (4.23) can be given a geometrical interpretation as shown in Figure 4.4. In Hilbert space, the figure can be considered as a reinterpretation of the projection theorem. $\langle \mathrm{w}\,,\, v\rangle$ is the projection of v onto a unit vector w in the orthogonal complement. This projection is maximized when w is aligned with the optimal error $v-\bar{x}$. This interpretation generalizes to Banach space by relabeling $\langle \mathrm{w}\,,\, v\rangle$ from "projection" to functional action, and $\mathsf{S}^\perp$ from orthogonal complement in V to orthogonal subspace in V^*.

The *duality gap* of (4.23) is defined as the difference $\mathsf{d_p}-\mathsf{d_d}$ (always positive) between the two optimal objectives. When they are equal, we say that the "duality gap is zero". There are many versions of duality theorems for various types of optimization problems. Conditions can be derived for when the duality gap is guaranteed to be zero even in cases where suprema and infima are not achieved (so optimal solutions do not exist). These conditions can be quite technical. However, for problems for which optimal solutions exist, the alignment condition (4.24) gives a much easier method to establish zero duality gap.

The result above was termed "weak duality", and the reader may suspect that therefore there must be a stronger version of the statement. Indeed, in Banach spaces, the dual problem always has a solution. A maximizing functional $\bar{\mathrm{w}}\in\mathsf{S}^\perp$ always exists even when the infimum in the primal problem is not achieved. This fact is a consequence of methods of constructing functionals that go by the name of Hahn-Banach theorems. This is the subject of the next section.

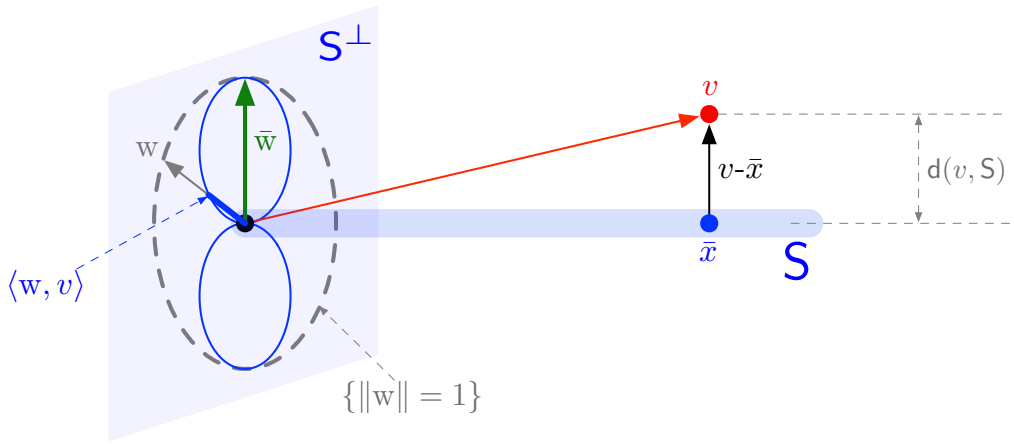


Figure 4.4: Illustration of the minimum-distance duality theorem 4.15. In a Hilbert space, we project v onto vectors w in $S^\perp$ of unit length (along the dashed grey circle). The length of the projection $\langle w\,,\, v\rangle$ is the length of thick blue line is shown above. The values of $\langle w\,,\, v\rangle$ for various directions $w \in S^\perp$ are depicted as the blue ellipses. The largest projection is achieved by a vector $\bar{w}$ which is aligned with the optimal "error" vector $v - \bar{x}$. Parallel interpretations are valid in a Banach space with "projection" replaced by functional action $\langle w\,,\, v\rangle$, and $S^\perp \subset V^*$ being the orthogonal subspace in V^* rather than the orthogonal complement.

4.3 Construction of Linear Functionals

We will be concerned with constructing various types of functionals. The constructions typically proceed by first defining a functional on a restricted subspace, and then "extending" it to the whole space. The extension process should guarantee certain properties of the functional, for example that the norm of the extension is not larger than the norm of the initially defined, restricted functional. This is the subject of the Hahn-Banach theorem to which we now turn.

Before formally stating the theorem, we motivate the important issues geometrically. Suppose we are given a functional $w : S \to \mathbb{R}$ defined on a proper subspace $S \subset V$ of a Banach space. How do we extend it to a functional $W : V \to \mathbb{R}$ defined on the entire space V? Extension here means that W is exactly w when restricted to the subspace, i.e. $W|_S = w$. You might imagine that if the norm of W is allowed to be larger than the norm of w, then this process is easy. We will require that the norm of the extension be no larger than the norm of the original restricted functional, i.e.

$$\|W\| \;:=\; \sup_{v\in V} \frac{\langle W\,,\, v\rangle}{\|v\|} \;=\; \|w\|_S \;:=\; \sup_{v\in S} \frac{\langle w\,,\, v\rangle}{\|v\|}$$

Let's examine how this extension process might work "one additional dimension at a time". Starting from the subspace S, select a vector v outside of it, and consider how to extend the functional to $\mathsf{span}\{S, v\}$, which is of one dimension larger than S. First, since any vector in $\mathsf{span}\{S, v\}$ can be written as $x + \alpha v$ with $x \in S$, linearity of the functional W implies

$$\langle W\,,\, x + \alpha v\rangle \;=\; \langle W\,,\, x\rangle + \alpha \langle W\,,\, v\rangle \;=\; \langle w\,,\, x\rangle + \alpha \langle W\,,\, v\rangle\,, \qquad x \in S,\ \alpha \in \mathbb{R}. \tag{4.25}$$

The extension W is now completely determined by the single number $\langle W\,,\, v\rangle$. This number needs to be chosen so that the norm of W is no larger than the norm of w.

At first, you might be tempted to select the trivial extension $\langle W\,,\, v\rangle = 0$ based on intuition. However, there is a subtlety here that should be appreciated. If v were orthogonal

to S (so this is only possible in a Hilbert space), then we can simply set $\langle \mathrm{W}\,,\, v\rangle = 0$, and orthogonality implies

$$\frac{\langle \mathrm{W}\,,\, x+\alpha v\rangle^2}{\|x+\alpha v\|^2} \;=\; \frac{\left(\langle \mathrm{w}\,,\, x\rangle + \alpha\,\langle \mathrm{W}\,,\, v\rangle\right)^2}{\|x+\alpha v\|^2} \;=\; \frac{\langle \mathrm{w}\,,\, x\rangle^2}{\|x\|^2+|\alpha|^2\|v\|^2} \;\leq\; \frac{\langle \mathrm{w}\,,\, x\rangle^2}{\|x\|^2}$$

$$\Rightarrow \qquad \|\mathrm{W}\|_{\mathsf{span}\{\mathsf{S},v\}} \;:=\; \sup_{x\in\mathsf{S},\ \alpha\in\mathbb{R}} \frac{\langle \mathrm{W}\,,\, x+\alpha v\rangle^2}{\|x+\alpha v\|^2} \;=\; \sup_{x\in\mathsf{S}} \frac{\langle \mathrm{w}\,,\, x\rangle^2}{\|x\|^2} \;=:\; \|\mathrm{w}\|_{\mathsf{S}}.$$

However, without orthogonality, we might have $\|x+\alpha v\| \leq \|x\|$ in the denominators, and then the trivial extension will actually have a larger norm than the original functional. Exercise 4.2 gives an example of such a situation. The point here is that a more elaborate and careful construction of the extension W needs to be done.

Theorem 4.16 (Hahn-Banach, One Dimensional Extension)**.** *Let $\mathsf{S}\subset\mathsf{V}$ be a subspace of a Banach space V. Let $\mathrm{w}:\mathsf{S}\to\mathbb{R}$ be a bounded linear functional on S, i.e.*

$$\|\mathrm{w}\|_{\mathsf{S}} \;:=\; \sup_{x\in\mathsf{S}} \frac{\langle \mathrm{w}\,,\, x\rangle}{\|x\|} \;=\; c \;<\; \infty.$$

Given any $v\in\mathsf{V}$, there exists an extension $\mathrm{W}:\mathsf{span}\{\mathsf{S},v\}\to\mathbb{R}$ of w (i.e. $\mathrm{W}|_{\mathsf{S}}=\mathrm{w}$) with the same norm $\|\mathrm{W}\|_{\mathsf{span}\{\mathsf{S},v\}}=c$.

Proof. Let's work backwards from the requirement $|\mathrm{W}(x+\alpha v)| \;\leq\; c\,\|x+\alpha v\|$, which we can rewrite (after using (4.25) and substituting $\mathsf{a}:=\mathrm{W}(v)$ for notational simplicity)

$$\forall\, x\in\mathsf{S},\ 0\neq\alpha\in\mathbb{R}, \qquad \begin{cases} \mathrm{w}(x)+\alpha\,\mathsf{a} \;\leq\; c\,\|x+\alpha v\| \\ -c\,\|x+\alpha v\| \;\leq\; \mathrm{w}(x)+\alpha\,\mathsf{a} \end{cases}, \tag{4.26}$$

(note that the case $\alpha=0$ is automatically satisfied so we exclude it). Rearranging and dividing through[4] by α gives a upper and a lower bound on the number a

$$\forall\, x\in\mathsf{S},\ 0\neq\alpha\in\mathbb{R}, \qquad \begin{cases} \mathsf{a} \;\leq\; c\,\|x/\alpha+v\| \text{ - } \mathrm{w}(x/\alpha) \\ \text{-}c\,\|x/\alpha+v\| \text{ - } \mathrm{w}(x/\alpha) \;\leq\; \mathsf{a} \end{cases} \tag{4.27}$$

Reparameterizing with $y:=x/\alpha\in\mathsf{S}$ gives slightly simpler conditions

$$\forall\, y\in\mathsf{S}, \qquad \begin{cases} \mathsf{a} \;\leq\; c\,\|y+v\| \;-\; \mathrm{w}(y) \\ -c\,\|y+v\| \;-\; \mathrm{w}(y) \;\leq\; \mathsf{a} \end{cases}$$

Now, there exists a real number a that satisfies both inequalities iff

$$\sup_{y\in\mathsf{S}}\Big(-c\|y+v\| \;-\; \mathrm{w}(y)\Big) \;\leq\; \inf_{z\in\mathsf{S}}\Big(c\|z+v\| \;-\; \mathrm{w}(z)\Big). \tag{4.28}$$

The linearity of w, and $x\in\mathsf{S} \;\Rightarrow\; \mathrm{w}(x)\leq c\|x\|$ gives a comparison of the two sides above

$$\begin{aligned} \big(-c\|y+v\|-\mathrm{w}(y)\big) \;-\; &\big(c\|z+v\|-\mathrm{w}(z)\big) \\ &= \;-c\left(\|z+v\|+\|y+v\|\right)+\mathrm{w}(z-y) \\ &\leq \;-c\|z+v-(y+v)\|+\mathrm{w}(z-y) \qquad \text{(triangle inequality)} \\ &\leq \;-c\|z-y\|+\mathrm{w}(z-y) \;\leq\; -c\|z-y\|+c\|z-y\| \;=\; 0 \\ \Rightarrow\quad \big(-c\|y+v\|-\mathrm{w}(y)\big) \;&\leq\; \big(c\|z+v\|-\mathrm{w}(z)\big). \end{aligned}$$

This last inequality implies (4.28), and therefore the existence of a real number a that satisfies (4.26). □

[4] The reader should check that dividing by a negative α maintains the equivalence. Indeed, the top inequality in (4.26) becomes the bottom inequality in (4.27) and vice versa.

Note that the proof is given for the case of a real Banach space. The proof for the complex case is similar, but somewhat messier, and is therefore omitted.

Theorem 4.17 (Hahn-Banach)**.** *Let* $\mathsf{S} \subset \mathsf{V}$ *be a subspace of a Banach space* V*. Let* $\mathrm{w} : \mathsf{S} \to \mathbb{R}$ *be a bounded linear functional, i.e.*

$$\|\mathrm{w}\|_{\mathsf{S}} \;:=\; \sup_{x\in\mathsf{S}} \frac{\langle \mathrm{w}\,,\, x\rangle}{\|x\|} \;=\; c \;<\; \infty.$$

Then there exists an extension $\mathrm{W} : \mathsf{V} \to \mathbb{R}$ *of* w *to all of* V *(i.e.* $\mathrm{W}|_{\mathsf{S}} = \mathrm{w}$*) with the same norm* $\|\mathrm{W}\| = c$*.*

Proof. The proof is given for a finite-dimensional or separable Banach space V.

First, we might as well assume that S is closed. If it were not, we can immediately extend w to the closure $\overline{\mathsf{S}}$ by the procedure for densely-defined bounded operators of Section 3.6.4. If the closure $\overline{\mathsf{S}} = \mathsf{V}$, then we're done. Thus from now on we assume that S is a proper, closed subspace of V. This implies the existence of an element $v_1 \notin \mathsf{S}$, and the one dimensional-extension Theorem 4.16 extends w to $\mathsf{S}_1 := \mathsf{span}\{\mathsf{S}, v_1\}$ without enlarging its norm.

The above procedure can be repeated to get a sequence of nested, proper subspaces

$$\mathsf{S} \;\subsetneq\; \mathsf{S}_1 \;\subsetneq\; \mathsf{S}_2 \;\cdots\,.$$

If V or V/S are finite dimensional, then this sequence terminates and extension procedure is complete. If neither are finite dimensional, we can take the union of these subspaces

$$\mathsf{S}_\infty \;=\; \bigcup_{k=0}^{\infty} \mathsf{S}_k,$$

and ask whether that is all of S. If V is separable, then we can choose $\{v_1, v_2, \ldots\}$ to be a total sequence in V/S. Since the span of this sequence is dense in V/S, then $\mathsf{S}_\infty = \mathsf{V}$.

If V is not separable, one must use a non-constructive existence statement like Zorn's lemma. This argument is omitted. □

The Hahn-Banach theorem has several immediate corollaries, useful in their own right.

Corollary 4.18. *Given any vector* $v \in \mathsf{V}$ *in a Banach space* V*, there exists a functional* $\bar{\mathrm{w}} \in \mathsf{V}^*$ *with* $\|\bar{\mathrm{w}}\| = 1$ *that solves the maximization problem*

$$\sup_{\|\mathrm{w}\|\le 1} \langle \mathrm{w}\,,\, v\rangle \;=\; \langle \bar{\mathrm{w}}\,,\, v\rangle \;=\; \|v\| \tag{4.29}$$

Consequently, given any bounded operator $A : \mathsf{V}_1 \to \mathsf{V}_2$ *then*

$$\|A\| \;:=\; \sup_{v\in\mathsf{V}_1,\ \|v\|\le 1} \|Av\|_{\mathsf{V}_2} \;=\; \sup_{v\in\mathsf{V}_1,\ \|v\|\le 1,\ \mathrm{w}\in\mathsf{V}_2^*,\ \|\mathrm{w}\|\le 1} \langle \mathrm{w}\,,\, Av\rangle\,. \tag{4.30}$$

Proof. Construct the functional first on the one dimensional subspace $\mathsf{span}\{v\}$ as

$$x \;=\; \alpha v \qquad \Rightarrow \qquad \langle \mathrm{w}\,,\, x\rangle \;:=\; \alpha\,\|v\|,$$

and therefore $\langle \mathrm{w}\,,\, v\rangle = \|v\|$. The norm of the functional on this subspace is

$$\sup_{x\in\mathsf{span}\{v\}} \frac{\langle \mathrm{w}\,,\, x\rangle}{\|x\|} \;=\; \sup_{\alpha\in\mathbb{R}} \frac{\langle \mathrm{w}\,,\, \alpha v\rangle}{\|\alpha v\|} \;=\; \sup_{\alpha\in\mathbb{R}} \frac{\alpha\|v\|}{|\alpha|\|v\|} \;=\; 1.$$

Now extend this functional to all of V while keeping its norm as 1. Finally, (4.30) follows immediately from (4.29). □

Note that the "flip side" of this corollary may not be true. That is, if we fix a functional w and optimize over the unit ball of the primal space

$$\sup_{v\in\mathsf{V},\ \|v\|\leq 1} \langle \mathrm{w}\,,\, v\rangle\,, \qquad\qquad \mathrm{w}\in\mathsf{V}^*,$$

there may not exist a vector $\bar{v}$ that achieves this supremum. There are cases however when this is true, and we'll discuss these when we cover "reflexive" spaces shortly.

Given a closed subspace S of a Hilbert space, one can always find a non-zero vector that is orthogonal to it. We simply pick any vector $v\notin\mathsf{S}$, and use the projection theorem to "drop a perpendicular" from v to the subspace S. Another corollary of the Hahn-Banach theorem is that we can do something similar in a Banach space, but we construct a *functional* that is orthogonal to S.

Corollary 4.19. *Let $\mathsf{S}\subset\mathsf{V}$ be a proper, closed subspace of a Banach space V. There exists a non-zero linear functional $\mathrm{w}\in\mathsf{V}^*$ such that $\mathsf{V}\subseteq\mathsf{Nu}(\mathrm{w})$.*

Proof. Pick any vector $v\notin\mathsf{S}$, and define the following functional w on $\mathsf{span}\{\mathsf{S},v\}$ by

$$\langle \mathrm{w}\,,\, x+\alpha v\rangle \;:=\; \alpha\,\boldsymbol{a}, \qquad\qquad x\in\mathsf{S},$$

where $\boldsymbol{a}$ is any non-zero number. Note that w is exactly zero on S. Since $\boldsymbol{a}\neq 0$, then $\|\mathrm{w}\|\neq 0$. Is this functional bounded on $\mathsf{span}\{\mathsf{S},v\}$? Let's check

$$\begin{aligned}
\sup_{x\in\mathsf{S},\ \alpha\in\mathbb{R}} \frac{\langle \bar{\mathrm{w}}\,,\, x+\alpha v\rangle}{\|x+\alpha v\|} \;&=\; \sup_{x\in\mathsf{S},\ \alpha\in\mathbb{R}} \frac{\alpha\,\boldsymbol{a}}{\|x+\alpha v\|} \;=\; \sup_{x\in\mathsf{S},\ \alpha\in\mathbb{R}} \frac{|\alpha|\,\boldsymbol{a}}{|\alpha|\ \|x/\alpha+v\|} \\
&=\; \sup_{y\in\mathsf{S}} \frac{\boldsymbol{a}}{\|y+v\|} \qquad\qquad (y:=x/\alpha\in\mathsf{S} \;\Leftrightarrow\; x\in\mathsf{S} \text{ if } \alpha\neq 0) \\
&=\; \frac{\boldsymbol{a}}{\inf_{y\in\mathsf{S}}\|y+v\|}. \qquad\qquad (4.31)
\end{aligned}$$

Now the quantity $\inf_{y\in\mathsf{S}}\|y+v\|$ is the distance between v and S. It must be positive since S is closed. Therefore the functional is bounded on $\mathsf{span}\{\mathsf{S},v\}$, and by Hahn-Banach can be extend to all of V. □

This corollary thus generates functionals in the orthogonal subspace $\mathsf{S}^\perp$. The preceding proof can be significantly strengthened by picking the number $\boldsymbol{a}$ judiciously as the distance between v and S. This allows us to show existence of a maximizing functional in the weak duality theorem as we show next.

Minimum Distance Problems: Existence of Dual Solutions

We will now use the Hahn-Banach theorem to strengthen the weak duality Theorem 4.15 and show that a solution to the dual problem always exists.

Theorem 4.20 (Minimum-Distance Duality)**.** *Let $\mathsf{S}\subset\mathsf{V}$ be a subspace of a Banach space V, and $\mathsf{S}^\perp\subset\mathsf{V}^*$ its orthogonal subspace. The minimum distance from any $v\in\mathsf{V}$ satisfies*

$$\inf_{x\in\mathsf{S}}\|v-x\| \;=\; \max_{\mathrm{w}\in\mathsf{S}^\perp,\ \|\mathrm{w}\|\leq 1} \langle \mathrm{w}\,,\, v\rangle \;=\; \langle \bar{\mathrm{w}}\,,\, v\rangle\,.$$

A vector $\bar{x}$ is a solution to the primal problem iff the optimal error $v-\bar{x}$ and the optimal functional $\bar{\mathrm{w}}$ are aligned

$$\langle \bar{\mathrm{w}}\,,\, v-\bar{x}\rangle \;=\; \|v-\bar{x}\|\ \|\bar{\mathrm{w}}\| \;=\; \|v-\bar{x}\|.$$

Proof. The proof is greatly aided by the diagram in Figure 4.4, and by recalling the alignment condition (4.24) of weak duality. For unit norm functional $\bar{\mathrm{w}}$ to be optimal for the dual problem it must (a) be orthogonal to S, and (b) $\langle \bar{\mathrm{w}}\,,\, v\rangle = \mathsf{d}(v,\mathsf{S})$. So construct a functional on S and v with those properties, and then extend it to all of V^* by Hahn-Banach.

Let $x+\alpha v$ with $x\in\mathsf{S}$ be any element of $\mathsf{span}\{\mathsf{S},v\}$, and define the functional $\bar{\mathrm{w}}$ by

$$\langle \bar{\mathrm{w}}\,,\, x+\alpha v\rangle \;:=\; \alpha\, \boldsymbol{d}, \qquad\qquad \boldsymbol{d} \;:=\; \mathsf{d}(v,\mathsf{S}) \;:=\; \inf_{x\in\mathsf{S}} \|v-x\|.$$

This is the same construction as in the proof of Corollary 4.19. Note that $\bar{\mathrm{w}}$ is exactly zero on S, and therefore it is in $\mathsf{S}^\perp$. When evaluated on only v it gives $\langle \bar{\mathrm{w}}\,,\, v\rangle = \boldsymbol{d}$, the distance from v to S. The reader should now reexamine Figure 4.4 with this in mind.

The last thing to show is that the norm of $\bar{\mathrm{w}}$ on $\mathsf{span}\{\mathsf{S},v\}$ is one. This calculation is exactly the same as (4.31) which here says

$$\sup_{x\in\mathsf{S},\ \alpha\in\mathbb{R}} \frac{\langle \bar{\mathrm{w}}\,,\, x+\alpha v\rangle}{\|x+\alpha v\|} \;=\; \frac{\boldsymbol{d}}{\inf_{y\in\mathsf{S}}\|y+v\|} \;=\; \frac{\boldsymbol{d}}{\boldsymbol{d}} \;=\; 1.$$

Finally use Hahn-Banach to extend this functional from $\mathsf{span}\{\mathsf{S},v\}$ to all of V. □

The Dual of the Dual: Reflexivity

Since the dual V^* of a vector space V is itself a vector space, one can ask about the dual of the dual space $(\mathsf{V}^*)^* =: \mathsf{V}^{**}$. Every element of $v\in\mathsf{V}$ can act on all of V^* as a linear functional as follows

$$\mathrm{v}(w) \;:=\; \mathrm{w}(v). \tag{4.32}$$

For example, if $v := (v_0, v_1, \ldots) \in \ell^1$ and $\mathrm{w} := (w_0, w_1, \ldots) \in \ell^\infty$, then

$$\mathrm{w}(v) \;:=\; \sum_{k=0}^{\infty} w_k v_k.$$

We can consider this sum as w acting on v, or v acting on w (and in this case we label it as $\mathrm{v}(w)$). Either way, the sum is well defined.

This action (4.32) is linear since

$$(\alpha \mathrm{v}_1 + \beta \mathrm{v}_2)\,(w) \;:=\; \mathrm{w}\,(\alpha v_1 + \beta v_2) \;=\; \alpha \mathrm{w}(v_1) + \beta \mathrm{w}(v_2) \;=:\; \alpha \mathrm{v}_1(\mathrm{w}) + \beta \mathrm{v}_2(\mathrm{w}).$$

Furthermore, the norm of $v\in\mathsf{V}$ gives a bound on the induced norm of $\mathrm{v}\in\mathsf{V}^{**}$

$$\|\mathrm{v}\| \;:=\; \sup_{0\neq \mathrm{w}\in\mathsf{V}^*} \frac{\|\mathrm{v}(\mathrm{w})\|}{\|\mathrm{w}\|} \;:=\; \sup_{0\neq \mathrm{w}\in\mathsf{V}^*} \frac{\|\mathrm{w}(v)\|}{\|\mathrm{w}\|} \;\leq\; \sup_{0\neq \mathrm{w}\in\mathsf{V}^*} \frac{\|\mathrm{w}\|\|v\|}{\|\mathrm{w}\|} \;=\; \|v\| \tag{4.33}$$

Therefore v defined by (4.32) is indeed a member of V^{**}. This together with the linearity of the action of v implies that the mapping $v\mapsto \mathrm{v}$ in (4.32) is vector space isomorphism from V to a subspace of V^{**}.

The mapping $v\mapsto \mathrm{v}$ would also be an *isometry* if the inequality in (4.33) is an equality. In a Hilbert space, we can simply choose $\mathrm{w}(v) = \langle v\,,\, v\rangle$, which achieves the equality. In a Banach space V, we need the statement that for every vector $v\in\mathsf{V}$ there exists a linear functional that achieves its norm, i.e.

$$\exists \mathrm{w}\in\mathsf{V}^*,\quad \mathrm{w}(v) \;=\; \|v\|,$$

but this is precisely Corollary 4.18 of the Hahn-Banach theorem. We can then conclude the following.

Lemma 4.21. *For any Banach space* V*, the mapping* $v \mapsto \mathrm{v}$ *from* V *to* V^{**} *defined by*

$$\mathrm{v}(\mathrm{w}) \;:=\; \mathrm{w}(v)$$

is an isometric isomorphism from V *to a subspace of* V^{**}*.*

Note that in a Hilbert space V, the dual V^* is isomorphic to V, and therefore so is the double dual, i.e. $\mathsf{V} \sim \mathsf{V}^* \sim \mathsf{V}^{**}$. In a Banach space, V and V^* are not generally isomorphic, but we just saw that V is isomorphic to a *subspace* of V^{**}. There are cases however where this subspace is actually all of V^{**}. Such spaces have special properties, so they're given a name.

Definition 4.22. *A Banach space* V *is called* reflexive *if* $\mathsf{V} \sim \mathsf{V}^{**}$*, i.e. if the isomorphic isometry (4.32) embedding* V *into* V^{**} *is onto.*

Any Hilbert space is reflexive, but some important Banach spaces are not. Most famously we have

$$(\ell^\infty_{\mathrm{o}})^* \;=\; \ell^1, \qquad \left(\ell^1\right)^* \;=\; \ell^\infty.$$

Recall that ℓ^∞_{o} is a closed, proper subspace of ℓ^∞, thus ℓ^∞_{o} is not reflexive. What about ℓ^1, if it were reflexive, then $(\ell^\infty)^* = \ell^1$. This is not possible since if a dual V^* is separable, then necessarily V is separable (Exercise 4.3). Thus $(\ell^\infty)^* = \ell^1$ would imply that ℓ^∞ is separable, which we know to be false.

Lemma 4.23. *In a reflexive space, every linear functional attains its max on the unit ball.*

Proof. If V is reflexive, we can regard V^* as the primal space and $\mathsf{V} \sim \mathsf{V}^{**}$ as the dual space. In this case, Corollary 4.18 says that for every element $\mathrm{w} \in \mathsf{V}^*$ (the primal space), there is an functional $\bar{\mathrm{v}} \in \mathsf{V}^{**} \sim \mathsf{V}$ that achieves its norm, i.e.

$$\sup_{\mathrm{v}\in\mathsf{V}^{**},\ \|\mathrm{v}\|\le 1} \mathrm{v}(\mathrm{w}) = \bar{\mathrm{v}}(\mathrm{w}) = \|\mathrm{w}\| \qquad \Leftrightarrow \qquad \sup_{v\in\mathsf{V},\ \|v\|\le 1} \langle \mathrm{w}\,,\, v\rangle = \langle \mathrm{w}\,,\, \bar{v}\rangle = \|\mathrm{w}\| \qquad \square$$

4.4 Dual Operators: The Adjoint

We have seen that the objects dual to vectors are linear functionals. What are the objects dual to linear operators between vector spaces? Given a linear operator $A : \mathsf{V}_1 \longrightarrow \mathsf{V}_2$ between Banach spaces, there is a natural way to define an operator between their dual spaces V_1^* and V_2^*. Consider the composition of the mappings

$$\mathsf{V}_1 \xrightarrow{A} \mathsf{V}_2 \xrightarrow{\mathrm{w}} \mathbb{R} \qquad \Leftrightarrow \qquad v \xmapsto{A} Av \xmapsto{\mathrm{w}} \langle \mathrm{w}\,,\, Av\rangle\,,$$

where $\mathrm{w} \in \mathsf{V}_2^*$ is any linear functional. Since A and w are linear, the composition $\mathrm{w} \circ A$ is also a linear mapping, and in this case from V_1 to $\mathbb{R}$, i.e. it is a linear functional on V_1. Therefore, there must be a $\mathrm{w}_1 \in \mathsf{V}_1^*$ that equals $\mathrm{w} \circ A$, i.e. w_1 should satisfy

$$\mathrm{w}_1 \;=\; \mathrm{w} \circ A \qquad \Leftrightarrow \qquad \forall v \in \mathsf{V}_1 \quad \langle \mathrm{w}_1\,,\, v\rangle = \langle \mathrm{w}\,,\, Av\rangle\,. \tag{4.34}$$

This defines a natural mapping $\mathrm{w} \mapsto \mathrm{w}_1$ from V_2^* to V_1^*. We call this mapping $A^\dagger$. Since this mapping applies to any $\mathrm{w} \in \mathsf{V}_2^*$, we rewrite the above relation as

$$\forall v \in \mathsf{V}_1,\ \forall \mathrm{w} \in \mathsf{V}_2^*, \qquad \left\langle A^\dagger \mathrm{w}\,,\, v\right\rangle = \langle \mathrm{w}\,,\, Av\rangle$$

These relations are illustrated in Figure 4.5.

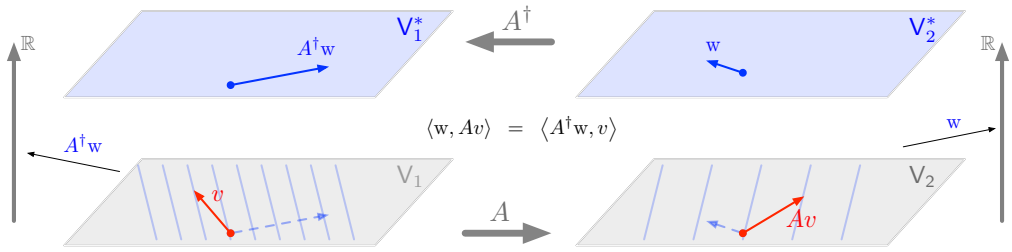


Figure 4.5: For any operator $A : \mathsf{V}_1 \to \mathsf{V}_2$, the adjoint operator $A^\dagger : \mathsf{V}_2^* \to \mathsf{V}_1^*$ maps linear functionals in V_2^* to linear functionals in V_1^*. For any $\mathrm{w} \in \mathsf{V}_2^*$, its image under the adjoint $A^\dagger \mathrm{w}$ is defined as the linear functional equal to the composition $(\mathrm{w} \circ A)(v) = \langle \mathrm{w}\,,\, Av\rangle$. Therefore the defining relation for the adjoint is $\langle A^\dagger \mathrm{w}\,,\, v\rangle = \langle \mathrm{w}\,,\, Av\rangle$.

The next question is whether the mapping $A^\dagger$ is well defined. Assume for now that A is a bounded operator, then the composition $\mathrm{w}_1 = \mathrm{w} \circ A$ is a bounded linear functional on V_1, and therefore there is a unique $\mathrm{w}_1 \in \mathsf{V}_1^*$ that satisfies (4.34). Furthermore, $A^\dagger$ must be linear. Indeed, pick any $\mathrm{w}, \mathrm{u} \in \mathsf{V}_2^*$, call their mappings under $A^\dagger$ as $\mathrm{w}_1 := A^\dagger \mathrm{w}$ and $\mathrm{u}_1 := A^\dagger \mathrm{u}$, then (4.34) implies

$$\forall v \in \mathsf{V}_1, \qquad \left.\begin{aligned} \langle \mathrm{w}_1\,,\, v\rangle &= \langle \mathrm{w}\,,\, Av\rangle \\ \langle \mathrm{u}_1\,,\, v\rangle &= \langle \mathrm{u}\,,\, Av\rangle \end{aligned}\right\} \quad \Rightarrow \quad \langle \alpha \mathrm{w}_1 + \beta \mathrm{u}_1\,,\, v\rangle = \langle \alpha \mathrm{w} + \beta \mathrm{u}\,,\, Av\rangle\,.$$

Finally $A^\dagger$ is a bounded operator if A is bounded as can be seen from

$$\begin{aligned} \left\|A^\dagger\right\| \;:=\; \sup_{\mathrm{w}\in\mathsf{V}_2^*,\ \|\mathrm{w}\|\le 1} \left\|A^\dagger \mathrm{w}\right\| \;&=\; \sup_{\mathrm{w}\in\mathsf{V}_2^*,\ \|\mathrm{w}\|\le 1,\ v\in\mathsf{V}_1,\ \|v\|\le 1} \langle A^\dagger \mathrm{w}\,,\, v\rangle \\ &=\; \sup_{\mathrm{w}\in\mathsf{V}_2^*,\ \|\mathrm{w}\|\le 1,\ v\in\mathsf{V}_1,\ \|v\|\le 1} \langle \mathrm{w}\,,\, Av\rangle \\ &=\; \sup_{v\in\mathsf{V}_1,\ \|v\|\le 1} \|Av\| \;=\; \|A\|\,, \end{aligned}$$

where we have also shown that the two induced operator norms are actually equal. We now summarize the conclusions from all the preceding arguments.

Lemma 4.24. *Let $A : \mathsf{V}_1 \to \mathsf{V}_2$ be a bounded operator between Banach spaces. Then there exists a linear operator mapping functionals $A^\dagger : \mathsf{V}_2^* \to \mathsf{V}_1^*$ called the* adjoint *of A, which is the unique operator satisfying the requirement*

$$\forall v \in \mathsf{V}_1,\ \mathrm{w} \in \mathsf{V}_2^*, \qquad \langle A^\dagger \mathrm{w}\,,\, v\rangle = \langle \mathrm{w}\,,\, Av\rangle \tag{4.35}$$

Furthermore, $\left\|A^\dagger\right\| = \|A\|$.

Remark 4.25. In calculations with adjoints, a typical step involves "moving" an operator from one side of functional action to the other by replacing it with its adjoint. For example, suppose two operators A and B can be composed as AB (i.e. the domains and co-domains allow for this), then

$$\langle \mathrm{w}\,,\, ABv\rangle \;=\; \langle \mathrm{w}\,,\, A(Bv)\rangle \;=\; \langle A^\dagger \mathrm{w}\,,\, Bv\rangle \;=\; \langle B^\dagger A^\dagger \mathrm{w}\,,\, v\rangle\,.$$

Since this holds for all functionals w and vectors v, this proves the identity $(AB)^\dagger = B^\dagger A^\dagger$.

Example 4.26. If a vector space is finite dimensional and we choose a basis, then each element of that vector space is identified with the column vector of its basis coefficients. Given two vector

spaces and choices of bases in each, every linear operator has a matrix representation with respect to those bases where the action of the linear operator is given by the matrix times the column vector. If we represent linear functionals with column vectors rather than row vectors, i.e. we represent $\mathrm{w}(v) := w^* v$ with w rather than w^*, then the adjoint acting on column vector w is the transpose of the matrix.

Indeed, let A be an $n \times m$ matrix. It is a linear mapping $A : \mathbb{C}^m \longrightarrow \mathbb{C}^n$, and functional action on $\mathbb{C}^n$ is given by a product $w^* v$. Now, equation (4.35) becomes

$$\mathrm{w}^* A v \;=\; \langle \mathrm{w}\,,\, Av \rangle \;\overset{1}{=}\; \langle A^\dagger \mathrm{w}\,,\, v \rangle \;=\; \left(A^\dagger \mathrm{w}\right)^* v \;=\; \mathrm{w}^* \left(A^\dagger\right)^* v$$

(proceeding outwards from the equality $\overset{1}{=}$). The fact that the equality $\mathrm{w}^* A v = \mathrm{w}^* \left(A^\dagger\right)^* v$ holds for all vectors v and w implies that $A \;=\; \left(A^\dagger\right)^*$ or equivalently that

$$A^\dagger \;=\; A^*.$$

In other words, if A is the matrix representation of a linear operator, then the matrix representation of its adjoint is the complex conjugate transpose of A.

Example 4.27. Let $\mathcal{A} : \mathsf{L}^2[a,b] \longrightarrow \mathsf{L}^2[a,b]$ be the bounded operator defined by a continuous kernel function $A(.,.)$. Specifically, $\mathcal{A} : f \mapsto g$ is given by

$$g(x) \;=\; \int_a^b A(x,\xi)\; f(\xi)\; d\xi.$$

Both f and g may be vector-valued, and in that case $A(.,.)$ would be a *matrix-valued* function. This is the kernel representation of a linear operator discussed in Chapter 6, where the kernel representation $A^\dagger$ of the adjoint $\mathcal{A}^\dagger$ is derived (Equation (6.8) as the nicely intuitive expression

$$\left(A^\dagger\right)(x,\xi) \;=\; A^*(\xi, x),$$

i.e. the kernel function of $\mathcal{A}^\dagger$ is obtained from that of $\mathcal{A}$ by "flipping" the arguments (x,ξ), and at each point taking a complex-conjugate transpose of the matrix value. Flipping the argument $(x,\xi) \mapsto (\xi, x)$ is akin to taking a transpose. This result is consistent with the interpretation of kernel functions as continuum analogues of matrices as emphasized in Chapter 6

Example 4.28. Given a vector function $f \in \mathsf{L}^\infty_n[0,\infty)$ (i.e. an n-vector where each component is an element of $\mathsf{L}^\infty[0,\infty)$), define the integral operator $F : \mathsf{L}^1[0,\infty) \to \mathbb{R}^n$

$$F(v) \;:=\; \int_0^\infty f(t)\; v(t)\; dt \;:=\; \int_0^\infty \begin{bmatrix} f_1(t) \\ \vdots \\ f_n(t) \end{bmatrix} v(t)\; dt \;:=\; \begin{bmatrix} \int_0^\infty f_1(t) v(t) dt \\ \vdots \\ \int_0^\infty f_n(t) v(t) dt \end{bmatrix}.$$

Note that F is really a "stacking" of n linear functionals on $\mathsf{L}^1[0,\infty)$ in an n-vector. The operator F takes a function on $[0,\infty)$ and returns a vector in $\mathbb{R}^n$. Its adjoint must then operate by taking a vector and returning a function on $[0,\infty)$. We can calculate its action from the requirement (4.35) as follows.

$$\begin{aligned} \langle F^\dagger \mathrm{w}\,,\, v \rangle \;&=\; \langle \mathrm{w}\,,\, Fv \rangle \\ \int_0^\infty \left(F^\dagger \mathrm{w}\right)(t)\; v(t)\; dt \;&=\; \mathrm{w}^* \int_0^\infty f(t)\; v(t)\; dt \end{aligned}$$

$\uparrow$ action of $F^\dagger \mathrm{w}$ as functional on v — $\uparrow$ action of w as a vector on the vector Fv

$$\int_0^\infty \left(F^\dagger \mathrm{w}\right)(t)\; v(t)\; dt \;=\; \int_0^\infty \left(f^*(t)\; \mathrm{w}\right)^*\; v(t)\; dt.$$

For this equality to hold for all $v \in \mathsf{L}^1[0,\infty)$, the following two functions must be equal

$$\big(F^\dagger \mathrm{w}\big)(t) \;=\; f^*(t)\,\mathrm{w} \;=\; \begin{bmatrix} f_1(t) & \cdots & f_n(t) \end{bmatrix} \begin{bmatrix} \mathrm{w}_1 \\ \vdots \\ \mathrm{w}_n \end{bmatrix} \;=\; \mathrm{w}_1 f_1(t) + \cdots + \mathrm{w}_n f_n(t).$$

Thus the operator $F^\dagger$ takes the vector w and makes a scalar-valued function by taking a linear combination of the n function $\{f_1, \ldots, f_n\}$ using the vector components $\{\mathrm{w}_k\}$ as coefficients for this linear combination.

It should be noted here that the setting of $v \in \mathsf{L}^1$ and $f \in \mathsf{L}^\infty_n$ is not special. The calculations above would be exactly the same if $v \in \mathsf{L}^p$ and $f \in \mathsf{L}^q_n$ as dual spaces, in particular for the case $p = q = 2$. We will return to this example again several times.

The following properties follow immediately from the definition of the adjoint and are left as an exercise.

Lemma 4.29. *Taking the adjoint of operators has the following properties.*

1. *For any two operators A and B on the same space $(\alpha A + \beta B)^\dagger = \alpha^* A^\dagger + \beta^* B^\dagger$.*
2. *If A^{-1} exists, then so does $\big(A^\dagger\big)^{-1}$ and it is equal to $\big(A^{-1}\big)^\dagger$. We use the notation $A^{-\dagger} := \big(A^{-1}\big)^\dagger = \big(A^\dagger\big)^{-1}$.*
3. *If the composition AB makes sense, then $(AB)^\dagger = B^\dagger A^\dagger$.*

Adjoints in Hilbert Space

The dual of a Hilbert space is itself, or more precisely isometrically isomorphic to itself in the sense that every vector $w \in \mathsf{V}$ defines a linear functional using the inner product $\langle w\,,\,.\rangle$. This isomorphism allows us to identify any Hilber space V with its dual V^*, and therefore the adjoint can now be defined as an operator on the Hilbert spaces themselves rather than their duals.

Definition 4.30. *Let $A : \mathsf{V} \to \mathsf{W}$ be a bounded operator between two Hilbert spaces. The* Hilbert adjoint *$A^\dagger : \mathsf{W} \to \mathsf{V}$ is the unique operator satisfying the requirement*

$$\forall v \in \mathsf{V},\ \forall w \in \mathsf{W}, \qquad \big\langle A^\dagger w\,,\, v\big\rangle_{\mathsf{V}} \;=\; \langle w\,,\, Av\rangle_{\mathsf{W}}\,. \tag{4.36}$$

The notation $\langle .\,,\,.\rangle_{\mathsf{V}}$, $\langle .\,,\,.\rangle_{\mathsf{W}}$ is used to emphasize the space where the inner product is taken.

Since the Banach adjoint of Lemma 4.24 makes sense for Hilbert spaces, the reader may wonder why a separate definition is made for the Hilbert adjoint above. As far as calculations are concerned, the two definitions are the same. The difference is conceptual. The Banach adjoint is between dual spaces rather than the original spaces. For Hilbert spaces, the Hilbert adjoint is the Banach adjoint if we identify the dual space with the original space via $\mathrm{w}(.) := \langle w\,,\,.\rangle$. The advantage of Definition 4.30 is that we can now compose an operator and its adjoint by $AA^\dagger$ or $A^\dagger A$. Note that this would not make sense for Banach adjoints. In addition, we can make sense of the concept of "self-adjoint" operators. This has far reaching implications as we will see below.

Another feature of the Hilbert adjoint is that taking the adjoint twice yields back the same operator $\big(A^\dagger\big)^\dagger = A$. Let $A : \mathsf{V} \to \mathsf{W}$, so $A^\dagger : \mathsf{W} \to \mathsf{V}$, and then $A^{\dagger\dagger} : \mathsf{V} \to \mathsf{W}$ with the requirement

$$\forall v \in \mathsf{V},\ \forall w \in \mathsf{W}, \qquad \big\langle A^\dagger w\,,\, v\big\rangle_{\mathsf{V}} \;=\; \langle w\,,\, Av\rangle_{\mathsf{W}} \qquad \Leftrightarrow \qquad \big\langle v\,,\, A^\dagger w\big\rangle_{\mathsf{V}} \;=\; \langle Av\,,\, w\rangle_{\mathsf{W}}\,,$$

which follows by symmetry of the inner product. The last equation says that A is the adjoint of $A^\dagger$ as claimed.

Remark 4.31. The fact that $A^{\dagger\dagger} = A$ means that in calculations, we can move operators freely to either side of the inner product by taking adjoints, e.g.

$$\langle Aw\,,\, v\rangle \;=\; \langle w\,,\, A^\dagger v\rangle\,.$$

This would not make sense in a Banach space since v is a vector, and $A^\dagger$ acts on functionals, so the action $A^\dagger v$ would not make sense.

Example 4.32. Define the right-shift operator $\mathcal{S}_{\rm r}$ on $\ell^2(\mathbb{N})$ by

$$\mathcal{S}_{\rm r}\,(u_0, u_1, \ldots) \;:=\; (0, u_0, u_1, \ldots)\,.$$

Now calculate its adjoint using the requirement (4.36)

$$\begin{aligned}
\langle \mathcal{S}_{\rm r}^\dagger v\,,\, u\rangle \;&=\; \langle v\,,\, \mathcal{S}_{\rm r} u\rangle \\
\Big\langle \mathcal{S}_{\rm r}^\dagger v\,,\, (u_0, u_1, \ldots)\Big\rangle \;&=\; \Big\langle (v_0, v_1, v_2, \ldots)\,,\, (0, u_0, u_1, \ldots)\Big\rangle \\
\Big\langle (v_1, v_2, \ldots)\,,\, (u_0, u_1, \ldots)\Big\rangle \;&=\; \Big\langle (v_0, v_1, v_2, \ldots)\,,\, (0, u_0, u_1, \ldots)\Big\rangle\,.
\end{aligned}$$

Therefore the adjoint of the right-shift operator is a left-shift operator $\mathcal{S}_{\rm l} = \mathcal{S}_{\rm r}^\dagger$ that drops the leftmost element of the sequence

$$\mathcal{S}_{\rm l}\,(u_0, u_1, \ldots) \;:=\; (u_1, u_2, \ldots)\,. \qquad \blacksquare$$

Example 4.33. Consider the following problem from linear systems theory where a linear system with an input u is given

$$\dot{x}(t) \;=\; Ax(t) + Bu(t), \qquad t \in [0, T],\; x(t) \in \mathbb{R}^n,\; u(t) \in \mathbb{R}^m,\;\; x(0) = 0.$$

The "variations-of-constants" formula says that the solution of the differential equations (with $x(0) = 0$) is given by

$$x(T) \;=\; \int_0^T e^{A(T-t)} B\; u(t)\; dt \;=:\; \mathcal{R}\,(u)\,, \tag{4.37}$$

where the integral defines the so-called reachability operator $\mathcal{R}$. If we choose $\mathsf{L}^2_m[0, T]$ as the Hilbert space for u, then $\mathcal{R} : \mathsf{L}^2[0, T] \to \mathbb{R}^n$ is a bounded operator which describes how a signal $\{u(t);\; t \in [0, T]\}$ is mapped to the state $x(T) \in \mathbb{R}^n$ at time T.

The adjoint is an operator $\mathcal{R}^\dagger : \mathbb{R}^n \to \mathsf{L}^2_m[0, T]$ which takes a vector to a function, and can be calculated as follows

$$\begin{aligned}
\langle v\,,\, \mathcal{R}(u)\rangle \;&=\; \langle \mathcal{R}^\dagger v\,,\, u\rangle \\
v^* \int_0^T e^{A(T-t)} B\; u(t)\; dt \;&=\; \int_0^T \big(\mathcal{R}^\dagger v\big)^*(t)\; u(t)\; dt \\
\int_0^T \Big(B^* e^{A^*(T-t)} v\Big)^*\; u(t)\; dt \;&=\; \int_0^T \big(\mathcal{R}^\dagger v\big)^*(t)\; u(t)\; dt.
\end{aligned}$$

Since this equality must hold for all $u \in \mathsf{L}^2_m[0, T]$, the two functions of t are equal

$$\big(\mathcal{R}^\dagger v\big)(t) \;=\; B^* e^{A^*(T-t)}\; v. \tag{4.38}$$

Thus the adjoint $\mathcal{R}^\dagger$ takes a vector $v \in \mathbb{R}^n$, and then produces a vector-valued function of t by multiplying v with the matrix-valued functions $B^* e^{A^*(T-t)}$ of t.

Example 4.34. Let $\{v_k\}$ be an orthonormal basis in a Hilbert space V. Every element $u \in \mathsf{V}$ has a unique expansion $u = \sum_{k=0}^{\infty} \alpha_k v_k$. Define the mapping $A : \mathsf{V} \to \ell^2(\mathbb{N})$ by $u \mapsto \alpha := (\alpha_0, \alpha_1, \ldots)$. This mapping takes a vector in V to the sequence of the coefficients of its basis representation. Parseval's theorem implies this is an isometric isomorphism from V to $\ell^2(\mathbb{N})$. In fact, calculating the adjoint in this case is basically the Plancherel identity (Theorem 3.15) since

$$\left.\begin{array}{l} u = \sum_{k=0}^{\infty} \alpha_k v_k \\ w = \sum_{k=0}^{\infty} \beta_k v_k \end{array}\right\} \quad \Rightarrow \quad \left\{\begin{array}{rl} & \left\langle A^\dagger \beta \,,\, u \right\rangle_{\mathsf{V}} = \langle \beta \,,\, Au \rangle_{\ell^2} \\ \Leftrightarrow & \left\langle A^\dagger \beta \,,\, u \right\rangle_{\mathsf{V}} = \langle \beta \,,\, \alpha \rangle_{\ell^2} = \sum_{k=0}^{\infty} \beta_k^* \alpha_k \\ \Leftrightarrow & \langle w \,,\, u \rangle_{\mathsf{V}} = \sum_{k=0}^{\infty} \beta_k^* \alpha_k \end{array}\right.$$

where the last statement is the Plancherel identity. Thus $A^\dagger \beta = w$, i.e. it takes a sequence in ℓ^2 and forms an element in V by using this sequence as coefficients of the basis expansion. Note that A does the opposite operation, but taking an element in V and finding its basis coefficients. Note that in this example

$$A^\dagger \;=\; A^{-1}, \qquad \text{and,} \qquad \|Au\|_{\ell^2} \;=\; \|u\|_{\mathsf{V}},$$

i.e. the operator A is an isometry.

Operators whose inverses are their adjoints have the following special property.

Lemma 4.35. *If the inverse of an operator A on a Hilbert space is equal to its adjoint, i.e. $A^\dagger A = AA^\dagger = I$, then A and $A^\dagger$ are isometries. Such operators are called* unitary.

The proof is the very simple calculation

$$\begin{array}{rcccccccc} \|Au\|^2 & = & \langle Au \,,\, Au \rangle & = & \left\langle AA^\dagger u \,,\, u \right\rangle & = & \langle u \,,\, u \rangle & = & \|u\|^2 \\ & & \text{or} & = & \left\langle u \,,\, A^\dagger Au \right\rangle & = & \langle u \,,\, u \rangle & = & \|u\|^2. \end{array}$$

Another very special class of operators is the following.

Definition 4.36. *An operator $A : \mathsf{V} \to \mathsf{V}$ on a Hilbert space* V *is called* self adjoint *if $A^\dagger = A$.*

One way to immediately generate self-adjoint operators is to compose any operator with its adjoint since

$$\left(AA^\dagger\right)^\dagger \;=\; \left(A^\dagger\right)^\dagger A^\dagger \;=\; AA^\dagger.$$

Thus for any operator A, the compositions $AA^\dagger$ and $A^\dagger A$ are both self adjoint. Self-adjoint operators have extremely useful properties, many of which are related to their eigenvectors. Those will be covered in Chapter 5. For now we list two other highlights.

Lemma 4.37. *If $A : \mathsf{V} \to \mathsf{V}$ is a self-adjoint operator on a Hilbert space, then*

1. *For any $v \in \mathsf{V}$, $\langle v \,,\, Av \rangle \in \mathbb{R}$.*
2. $\|A\| \;=\; \sup_{\|v\| \leq 1} \; \langle v \,,\, Av \rangle$.

Note that for a Hilbert space over complex scalars, the inner product $\langle v \,,\, Av \rangle$ can be a complex number. For self-adjoint operators however, it is guaranteed to be a real number. The second clause should be compared to formula (4.30) $\|A\| = \sup_{w,v} \langle w \,,\, Av \rangle$, which requires maximization over two parameters w and v. If A is self adjoint, then we have the simpler maximization over a single parameter v.

Proof. Of Lemma 4.37. The complex-conjugate symmetry of the inner product says $\langle w\,,\,v\rangle^* = \langle v\,,\,w\rangle$. Therefore

$$\langle v\,,\,Av\rangle^* \;=\; \langle Av\,,\,v\rangle \;=\; \left\langle v\,,\,A^\dagger v\right\rangle \;=\; \langle v\,,\,Av\rangle\,,$$

where the last equality is due to $A^\dagger = A$. Thus the number $\langle v\,,\,Av\rangle$ must be real.

For the second clause, we need to compare the two quantities

$$c \;:=\; \sup_{\|v\|\le 1} |\langle v\,,\,Av\rangle| \;\le\; \sup_{\|u\|\le 1,\|w\|\le 1} |\langle w\,,\,Au\rangle| \;=\; \|A\|. \tag{4.39}$$

Cleary $c \le \|A\|$ since c is a maximization of the same quantity over a smaller set. We need to show the opposite inequality. If we can express $\langle w\,,\,Av\rangle$ using terms of the form $\langle v\,,\,Av\rangle$, then we can make a comparison. Observe the following algebraic identity

$$\begin{aligned}\Big\langle (u+w)\,,\,A(u+w)\Big\rangle - \Big\langle (u-w)\,,\,A(u-w)\Big\rangle &= 2\,\langle w\,,\,Au\rangle + 2\,\langle u\,,\,Aw\rangle \\ &= 2\,\langle w\,,\,Au\rangle + 2\,\langle Aw\,,\,u\rangle \\ &= 2\,\langle w\,,\,Au\rangle + 2\,\langle w\,,\,Au\rangle^* = 4\,\mathbb{R}(\langle w\,,\,Au\rangle)\end{aligned}$$

The left hand side is in a form that can be bounded by the constant c

$$\begin{aligned}&\big|\big\langle (u+w)\,,\,A(u+w)\big\rangle - \big\langle (u-w)\,,\,A(u-w)\big\rangle\big| \\ &\le\; \big|\big\langle (u+w)\,,\,A(u+w)\big\rangle\big| + \big|\big\langle (u-w)\,,\,A(u-w)\big\rangle\big| \;\le\; c\left(\|u+w\|^2 + \|u-w\|^2\right) \\ &\qquad = \; 2c\left(\|u\|^2 + \|w\|^2\right),\end{aligned}$$

where the last equality is the parallelogram law. This together with the algebraic identity gives $\mathbb{R}(\langle w\,,\,Av\rangle) \;\le\; (c/2)\left(\|u\|^2 + \|w\|^2\right)$. It then follows (Exercise 4.4) that

$$|\langle w\,,\,Av\rangle| \;\le\; (c/2)\left(\|u\|^2 + \|w\|^2\right).$$

We can now compare the two maximization problems

$$\|A\| \;=\; \sup_{\|u\|\le 1,\|w\|\le 1} |\langle w\,,\,Au\rangle| \;\le\; \sup_{\|u\|\le 1,\|w\|\le 1} \frac{c}{2}\left(\|u\|^2 + \|w\|^2\right) \;\le\; c,$$

finally implying that $\|A\| = c := \sup_{\|v\|\le 1} \langle v\,,\,Av\rangle$. □

To appreciate that $\|A\| \;=\; \sup_{\|v\|\le 1} \;\langle v\,,\,Av\rangle$ may not be true in general, consider the 90° counterclockwise rotation matrix in $\mathbb{R}^2$

$$A \;=\; \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \qquad \Rightarrow \qquad \forall v \in \mathbb{R}^2,\;\; \langle v\,,\,Av\rangle = -v_1v_2 + v_2v_1 = 0,$$

since (by design) Av is orthogonal to v. However, A is an isometry, so $\|A\| = 1$, and we have a large gap in the inequality (4.39).

For any operator A, the composition $AA^\dagger$ is not only self adjoint, but many properties of the original operator A can be deduced from those of $AA^\dagger$ or $A^\dagger A$. The following is one example.

Lemma 4.38. *If $A : \mathsf{V} \to \mathsf{W}$ is a bounded operator on Hilbert spaces, then*

$$\|A\|^2 \;=\; \left\|AA^\dagger\right\| \;=\; \left\|A^\dagger A\right\|.$$

Proof. This is a quick application of the previous Lemma 4.37. Since $A^\dagger A$ is self adjoint, its norm can be calculated from the "single parameter" maximization problem

$$\left\|A^\dagger A\right\| \;=\; \sup_{\|v\|\leq 1} \left\langle v\,,\, A^\dagger A v\right\rangle \;=\; \sup_{\|v\|\leq 1} \left\langle Av\,,\, Av\right\rangle \;=\; \sup_{\|v\|\leq 1} \|Av\|^2 \;=\; \|A\|^2. \qquad \square$$

This fact can be immensely useful. Suppose $A : \mathsf{V} \to \mathbb{R}^n$, and V is an infinite-dimensional space. Calculating $\|A\|$ directly from the definition is generally difficult and one must resort to approximation techniques. However, note that $AA^\dagger : \mathbb{R}^n \to \mathbb{R}^n$, i.e. a matrix, which is a finite-dimensional object whose norm calculation is much easier than that of A. Thus the identity $\|A\|^2 = \left\|AA^\dagger\right\|$ allows us to convert some infinite-dimensional problems to finite-dimensional ones.

The lemma also highlights the special relationship between an operator and its adjoint. Suppose B is any other operator where the composition AB makes sense, and $\|B\| = \|A\|$. Submultiplicativity implies

$$\|AB\| \;\leq\; \|A\|\,\|B\| \;=\; \|A\|\,\|A\| \;=\; \|A\|^2.$$

Thus in general, we only have the inequality $\|AB\| \leq \|A\|^2$. Composition with the adjoint is very special since we have equality in this inequality. Intuitively, one can say that the norm of A is not "reduced" by composing it with its adjoint. Other, similarly special, properties of the composition with the adjoint will be explored in the next section.

4.5 The Four Fundamental Subspaces

An operator and its adjoint each have image and null spaces, and there are easily established, but fundamental, relations between them. Let $A : \mathsf{V}_1 \to \mathsf{V}_2$ and $A^\dagger : \mathsf{V}_2^* \to \mathsf{V}_1^*$ as shown in Figure 4.6 and examine each of their null and image spaces.

First begin with $\mathsf{Im}(A)$ and $\mathsf{Nu}\big(A^\dagger\big)$, which are illustrated on the right side of Figure 4.6

$$\begin{aligned} \mathrm{w} \in \mathsf{Im}(A)^\perp \quad &\Leftrightarrow \quad \forall v \in \mathsf{V}_1,\ \langle \mathrm{w}\,,\, Av\rangle = 0 \quad \Leftrightarrow \quad \forall v \in \mathsf{V}_1,\ \left\langle A^\dagger \mathrm{w}\,,\, v\right\rangle = 0 \\ &\Leftrightarrow \quad A^\dagger \mathrm{w} = 0, \quad \text{i.e. } \mathrm{w} \in \mathsf{Nu}\big(A^\dagger\big)\,, \end{aligned}$$

which means that $\mathsf{Im}(A)^\perp = \mathsf{Nu}\big(A^\dagger\big)$ as depicted in the figure. Now we examine the relation between $\mathsf{Nu}(A)$ and $\mathsf{Im}\big(A^\dagger\big)$ (left side of figure)

$$\begin{aligned} v \in \mathsf{Nu}(A)\,,\ \text{i.e. } Av = 0 \quad &\Leftrightarrow \quad \forall \mathrm{w} \in \mathsf{V}_2^*,\ \langle \mathrm{w}\,,\, Av\rangle = 0 \quad \Leftrightarrow \quad \forall \mathrm{w} \in \mathsf{V}_2^*,\ \left\langle A^\dagger \mathrm{w}\,,\, v\right\rangle = 0 \\ &\Leftrightarrow \quad v \perp \mathsf{Im}\big(A^\dagger\big)\,. \end{aligned}$$

The last statement means that every element of $\mathsf{Im}\big(A^\dagger\big)$ is orthogonal to all of $\mathsf{Nu}(A)$, i.e. $\mathsf{Im}\big(A^\dagger\big) \subseteq \mathsf{Nu}(A)^\perp$, but it does not necessarily imply the opposite containment. One can say a little bit more. Orthogonal subspaces like $\mathsf{Nu}(A)^\perp$ are always closed, so we at least have $\overline{\mathsf{Im}(A^\dagger)} \subseteq \mathsf{Nu}(A)^\perp$. However, for the two subspaces to be equal, we need an additional condition.

Theorem 4.39. *Let $A : \mathsf{V}_1 \to \mathsf{V}_2$ be a bounded operator between Banach spaces with $A^\dagger : \mathsf{V}_2^* \to \mathsf{V}_1^*$ its adjoint. Then their null and image spaces are related by*

$$\begin{aligned} &\mathsf{Im}(A)^\perp = \mathsf{Nu}\big(A^\dagger\big)\,, \\ &\mathsf{Nu}(A)^\perp \supseteq \overline{\mathsf{Im}(A^\dagger)}, \qquad\qquad \mathsf{Nu}(A)^\perp = \mathsf{Im}\big(A^\dagger\big) \ \textit{if } \mathsf{Im}(A) \textit{ is closed.} \end{aligned}$$

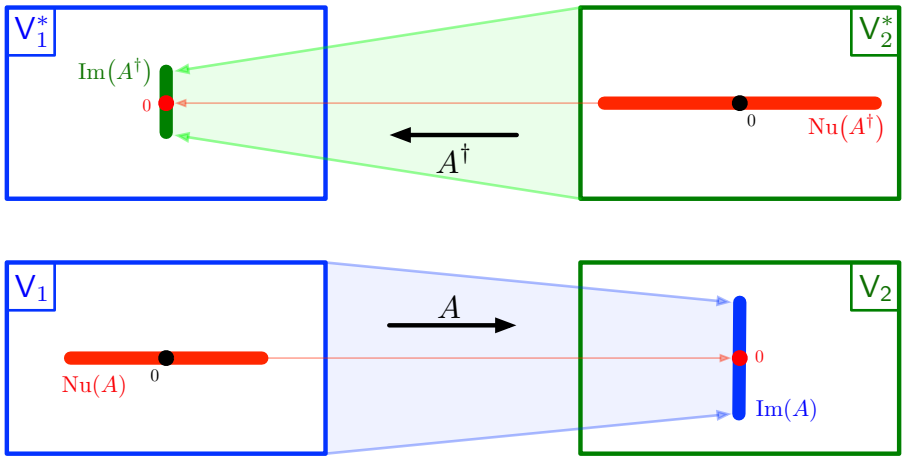


Figure 4.6: Depiction of the four fundamental subspaces. The null and image spaces of A and its adjoint $A^\dagger$ have special orthogonality relationships $\mathsf{Im}(A) \perp \mathsf{Nu}(A^\dagger)$ and $\mathsf{Im}(A^\dagger) \perp \mathsf{Nu}(A)$. For example, if $v \in \mathsf{Nu}(A)$, then $0 = \langle \mathrm{w}\,,\, Av\rangle = \langle A^\dagger \mathrm{w}\,,\, v\rangle$, which means $A^\dagger \mathrm{w}$, which is in the image of $A^\dagger$, is orthogonal to v.

We have already shown all but the last statement of this theorem. We will not give a proof[5] of it, but give an example to demonstrate why it is needed. Consider the example (3.34) of the operator $A := \mathsf{diag}(1, 1/2, 1/3, \ldots)$ on ℓ^1, whose image space was shown not to be closed. The null space of A is 0, and therefore all of $\left(\ell^1\right)^* = \ell^\infty$ is $\mathsf{Nu}(A)^\perp$. Now since A is diagonal, its adjoint is also $A^\dagger = \mathsf{diag}(1, 1/2, 1/3, \ldots)$ on ℓ^∞. Any element in $\mathsf{Im}\left(A^\dagger\right)$ must be a decaying sequence, i.e. in ℓ^∞_0, which is the proper, closed subspace of ℓ^∞ made up of sequences that decay to zero asymptotically. Thus $\overline{\mathsf{Im}(A^\dagger)} = \ell^\infty_0 \subsetneq \ell^\infty = \mathsf{Nu}(A)^\perp$.

If V_2 is finite dimensional, then the image of A is closed, and we have equality in the above theorem. An immediate application of Theorem 4.39 is to the concepts of column rank and row rank of a matrix.

Definition 4.40. *Let $A : \mathbb{C}^m \to \mathbb{C}^n$ be an $n \times m$ matrix. Its* column rank *is the number of linearly independent columns, or equivalently the dimension of $\mathsf{Im}(A)$. Its* row rank *is the number of linearly independent rows, or equivalently the dimension of $\mathsf{Im}(A^*)$.*

Now recall the Rank-Nullity Theorem 1.41 which in this case relates the dimensions of the null and image spaces of A and A^* to the dimensions of their domains

$$m \;=\; \mathsf{dim}(\mathbb{C}^m) \;=\; \mathsf{dim}(\mathsf{Nu}(A)) + \mathsf{dim}(\mathsf{Im}(A)) \;=\; \mathbf{nl}(A) + \mathbf{rk}(A)\,, \tag{4.40}$$

$$n \;=\; \mathsf{dim}(\mathbb{C}^n) \;=\; \mathsf{dim}(\mathsf{Nu}(A^*)) + \mathsf{dim}(\mathsf{Im}(A^*)) \;=\; \mathbf{nl}(A^*) + \mathbf{rk}(A^*)\,. \tag{4.41}$$

These relations, together with the orthogonality relations of Theorem 4.39, allow us to count dimensions and prove the following fact.

Lemma 4.41. *For any matrix, its column rank equals its row rank.*

Proof. As already mentioned, let's count dimensions. Theorem 4.39 says

$$\mathsf{Nu}(A)^\perp = \mathsf{Im}(A^*) \quad\Rightarrow\quad \mathbb{C}^m = \mathsf{Nu}(A) \oplus \mathsf{Im}(A^*) \quad\Rightarrow\quad m = \mathbf{nl}(A) + \mathbf{rk}(A^*) \tag{4.42}$$

Combining this with (4.40) gives $\mathbf{rk}(A) = \mathbf{rk}(A^*)$. □

[5] The proof requires the Bounded Inverse Theorem 3.34.

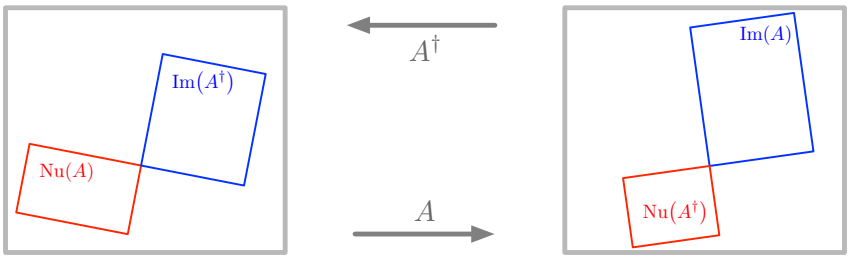


Figure 4.7: A graphical depiction of the four fundamental subspaces of an operator between Hilbert spaces. Each subspace is depicted as a square, and orthogonal complements are depicted as being at right angles to each other.

In addition, combining (4.42) with (4.41) gives a relation between the null spaces

$$\mathbf{nl}(A) - \mathbf{nl}(A^*) \;=\; m - n.$$

The null space of A^* is sometimes called the left null space of A. This relation says that for square matrices ($n = m$), the dimensions of the left and right null spaces are equal. For non-square matrices, the difference in dimensions of the null spaces is precisely the difference between the numbers of rows and columns.

The Four Subspaces in Hilbert Space

In Hilbert space, the picture of the four subspaces is somewhat simplified since we can now think about orthogonal complements in the same space.

Theorem 4.42. *Let $A : \mathsf{V}_1 \to \mathsf{V}_2$ be a bounded operator between Hilbert spaces with $A^\dagger : \mathsf{V}_2 \to \mathsf{V}_1$ its Hilbert adjoint. Then their null and image spaces are related by*

$$\mathsf{Im}(A)^\perp = \mathsf{Nu}\big(A^\dagger\big) \qquad \Leftrightarrow \qquad \mathsf{Im}\big(A^\dagger\big)^\perp = \mathsf{Nu}(A)\,, \tag{4.43}$$

$$\mathsf{Nu}(A)^\perp = \overline{\mathsf{Im}(A^\dagger)} \qquad \Leftrightarrow \qquad \mathsf{Nu}\big(A^\dagger\big)^\perp = \overline{\mathsf{Im}(A)}. \tag{4.44}$$

A good way to remember the closures above is that Null spaces of bounded operators are always closed, so $\mathsf{Nu}\big(A^\dagger\big) = \overline{\mathsf{Nu}(A^\dagger)}$. An orthogonal complement is also always closed, so $\mathsf{Nu}(A)^\perp$ must be a closed subspace, thus taking the closure $\overline{\mathsf{Im}(A^\dagger)}$.

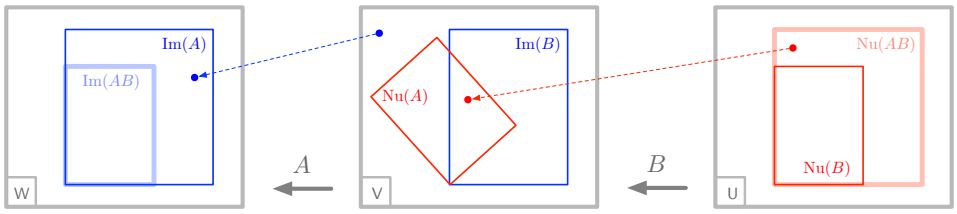
Proof. The equivalences in (4.43) and (4.44) follow by replacing A with $A^\dagger$ and recalling that $A^{\dagger\dagger} = A$, e.g.

$$\text{for any operator } A,\ \mathsf{Im}(A)^\perp = \mathsf{Nu}\big(A^\dagger\big) \qquad \Rightarrow \qquad \mathsf{Im}\big(A^\dagger\big)^\perp = \mathsf{Nu}\big(A^{\dagger\dagger}\big) = \mathsf{Nu}(A)\,.$$

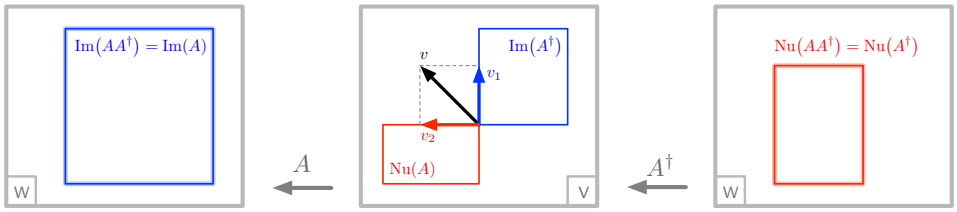
The first statement in (4.43) is the same as that in Theorem 4.39. The second statement (4.44) follows from the first by recalling that for any subspace $\mathsf{S} \subset \mathsf{V}$, we have $\mathsf{S}^{\perp\perp} = \overline{\mathsf{S}}$ (Lemma 3.24). Therefore

$$\mathsf{Nu}(A) = \mathsf{Im}\big(A^\dagger\big)^\perp \qquad \Rightarrow \qquad \mathsf{Nu}(A)^\perp = \mathsf{Im}\big(A^\dagger\big)^{\perp\perp} = \overline{\mathsf{Im}(A^\dagger)}. \qquad \square$$

The theorem is depicted graphically in Figure 4.7. These orthogonality relations have many consequences, and we now describe one of the important ones about the relationships between the image and null spaces of the compositions $AA^\dagger$ and $A^\dagger A$. To provide a larger



(a) When two operators A and B are composed as AB, the null space of B is contained in that of AB (shown in red). If a vector is nulled by B, it must also be nulled by AB, thus the containment $\mathsf{Nu}(B) \subseteq \mathsf{Nu}(AB)$. On the other hand, any element $w = ABu$ in the image space of AB must also be in the image space of A since $w = A(Bu)$, thus the containment $\mathsf{Im}(AB) \subseteq \mathsf{Im}(A)$. The containments may be strict. If an element is mapped to a non-zero element in $\mathsf{Im}(B) \cap \mathsf{Nu}(A)$ (shown as the dashed red arrow), then it is in $\mathsf{Nu}(AB)$, but not in $\mathsf{Nu}(B)$. Similarly, an element not in $\mathsf{Im}(B)$ is mapped to $\mathsf{Im}(A)$ but not in $\mathsf{Im}(AB)$ (shown as the dashed blue arrow), for otherwise it would have been in $\mathsf{Im}(B)$.



(b) When an operator is composed with its adjoint the strict containment situation described in part (a) above cannot occur since $\mathsf{Nu}(A) \perp \mathsf{Im}\left(A^{\perp}\right)$. Nothing in $\mathsf{Im}\left(A^{\dagger}\right)$ can be nulled by A.

Figure 4.8: Depiction of null and image spaces containments for operator compositions. When an operator is composed with its adjoint, the geometry becomes very special and we get the above equalities of spaces as described in Lemma 4.43.

context consider the composition AB of two operators $\mathsf{U} \xrightarrow{B} \mathsf{V} \xrightarrow{A} \mathsf{W}$ depicted in Figure 4.8a. The null spaces of B and AB are in U, while the image spaces of A and AB are in W. The subspaces obey the following containment relations

$$\begin{aligned} u \in \mathsf{Nu}(B) \quad &\Rightarrow \quad ABu = A0 = 0 \quad \Rightarrow \quad u \in \mathsf{Nu}(AB) \quad \Rightarrow \quad \boxed{\mathsf{Nu}(B) \subseteq \mathsf{Nu}(AB)} \\ w \in \mathsf{Im}(AB) \quad &\Rightarrow \quad \exists u,\ w = ABu \quad \Rightarrow \quad \exists v\ (v = Bu),\ w = Av \quad \Rightarrow \quad w \in \mathsf{Im}(A) \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \boxed{\Rightarrow \quad \mathsf{Im}(AB) \subseteq \mathsf{Im}(A)} \end{aligned}$$

Figure 4.8a illustrates these relations and gives an example of how these containments can be strict. A non-zero element $v = Bu$ which is in $\mathsf{Im}(B)$, but happens to be in $\mathsf{Nu}(A)$ will have $0 = Av = ABu$. Thus $u \notin \mathsf{Nu}(B)$, but $u \in \mathsf{Nu}(AB)$. An example is also shown for strict containment of the image spaces.

Now if we compose an operator with its adjoint, say $AA^{\dagger}$, the situation just described can not happen since $\mathsf{Im}\left(A^{\dagger}\right) \perp \mathsf{Nu}(A)$ (see Figure 4.8b), which can be interpreted as saying that nothing in the image space of $A^{\dagger}$ can be nulled by A, so we have an unobstructed "pass through" by A. The precise statement is as follows.

Lemma 4.43. *Let $A : \mathsf{V} \to \mathsf{W}$ be a bounded operator between Hilbert spaces. Then*

$$\begin{aligned} \mathsf{Im}\left(AA^{\dagger}\right) &= \mathsf{Im}(A) \quad &\Leftrightarrow \quad \mathsf{Im}\left(A^{\dagger}A\right) &= \mathsf{Im}\left(A^{\dagger}\right) \\ \mathsf{Nu}\left(AA^{\dagger}\right) &= \mathsf{Nu}\left(A^{\dagger}\right) \quad &\Leftrightarrow \quad \mathsf{Nu}\left(A^{\dagger}A\right) &= \mathsf{Nu}(A) \end{aligned}$$

Note that the equivalences follow by replacing A by $A^{\perp}$.

Proof. The proof is illustrated in Figure 4.8b. Consider $\mathsf{Im}(AA^\dagger) = \mathsf{Im}(A)$ first. Let $w = Av$ be in $\mathsf{Im}(A)$, and consider the orthogonal decomposition of $v \in \mathsf{V}$

$$v = v_1 + v_2, \qquad \begin{cases} v_1 \in \mathsf{Im}(A^\dagger) & \Rightarrow \quad \exists u,\ v_1 = A^\dagger u \\ v_2 \in \mathsf{Nu}(A) & \Rightarrow \quad Av_2 = 0 \end{cases}$$

$$\Rightarrow \qquad w = Av = Av_1 + Av_2 = Av_1 = AA^\dagger u \qquad \Rightarrow \qquad w \in \mathsf{Im}(AA^\dagger)\,.$$

For the null spaces, suppose $u \in \mathsf{W}$ is such that $AA^\dagger u = 0$. Consider $v := A^\dagger u$, then $Av = 0$ and therefore $v \in \mathsf{Im}(A^\dagger) \cap \mathsf{Nu}(A) = 0$. This means $\mathsf{Nu}(AA^\dagger) \subseteq \mathsf{Nu}(A^\dagger)$, and since $\mathsf{Nu}(A^\dagger) \subseteq \mathsf{Nu}(AA^\dagger)$ always, the two subspaces must be equal. □

The preceding lemma should be considered in concert with Lemma 4.38, which stated that $\|AA^\dagger\| = \|A\|\|A^\dagger\| = \|A\|^2$. This last statement was compared with the generally valid bound $\|AB\| \leq \|A\|\|B\|$, and the comment was made that composing with the adjoint does not "reduce" the norm. Lemma 4.43 is similar in spirit. When composing an operator with its adjoint, null and image spaces are not "distorted".

Operators like $AA^\dagger$ or $A^\dagger A$ are sometimes called "Grammians" and occur in many problems that involve solvability of linear equations exactly or approximately. The next set of examples highlight this. They can all be described as "least-squares problems" in Hilbert space.

Least-Squares Problems in Hilbert Space

Example 4.44. Given an overdetermined linear system of equations like $Ax = b$ where the dimension of b is higher (perhaps infinite) than that of x. In this case, exact solutions generally do not exist, and a typical approach is to try to find a solution that minimizes the equation error $\|Ax - b\|$, i.e.

$$\inf_x \|Ax - b\|.$$

If this problem is set up in Hilbert spaces with $A : \mathsf{V} \to \mathsf{W}$, then it is really a minimum distance to a subspace problem

$$\inf_x \|Ax - b\| \;=\; \inf_{y \in \mathsf{Im}(A)} \|y - b\| \;=\; \inf_{y \in \overline{\mathsf{Im}(A)}} \|y - b\|.$$

If $\mathsf{Im}(A)$ is closed, a solution exists and the projection theorem says that the optimal error $\bar{y} - b$ must be orthogonal to the subspace, i.e. $(\bar{y} - b) \in \mathsf{Im}(A)^\perp$. We know that $\mathsf{Im}(A)^\perp = \mathsf{Nu}(A^\dagger)$, therefore the optimal error satisfies

$$A^\dagger(\bar{y} - b) \;=\; 0$$

If $\mathsf{Im}(A)$ is closed, there exists $\bar{x}$ such that $A\bar{x} = \bar{y}$. Putting this all together gives

$$A^\dagger\,(A\bar{x} - b) = 0 \qquad \Rightarrow \qquad A^\dagger A\,\bar{x} \;=\; A^\dagger b \tag{4.45}$$

$$\Rightarrow \qquad \bar{x} = \left(A^\dagger A\right)^{-1} A^\dagger b. \qquad \text{(if } A^\dagger A \text{ invertible)}$$

The last expression is precisely the Moore-Penrose pseudo-inverse formula for the solution of overdetermined least squares problems. The equation (4.45) is a necessary condition for optimality of $\bar{x}$ regardless of whether $A^\dagger A$ is invertible or not.

Example 4.45. Now consider an operator $A : \mathsf{V} \to \mathsf{W}$ between Hilbert spaces, and the underdetermined system $Ax = b$ where the dimension of x is higher (possibly infinite) than that

of b. Such equations generally have many solutions parameterized by the null space of A. One approach is to seek a solution of minimum norm

$$\inf_{x,\ Ax=b} \|x\|.$$

Since the set $\{x;\ Ax = b\}$ is an affine space, this is minimum distance problem from zero to the an affine space. The projection theorem gives the answer that the optimal solution[6] $\bar{x}$ must be orthogonal to the affine space, i.e. orthogonal to $\mathsf{Nu}(A)$

$$\bar{x} \in \mathsf{Nu}(A)^{\perp} = \mathsf{Im}\big(A^{\dagger}\big) \qquad \Rightarrow \qquad \bar{x} \ = \ A^{\dagger} v,\ v \in \mathsf{W}. \tag{4.46}$$

If we substitute this last condition in the "constraint"

$$A\bar{x} = b \text{ and } \bar{x} = A^{\dagger} v \qquad \Rightarrow \qquad \begin{cases} AA^{\dagger} v \ = \ b \\ \bar{x} \ = \ A^{\dagger} v \end{cases} \tag{4.47}$$

$$\Rightarrow \qquad \bar{x} = A^{\dagger} \left(AA^{\dagger}\right)^{-1} b \qquad \text{(if } AA^{\dagger} \text{ invertible)}$$

Again, (4.47) is a necessary condition for optimality, while the last expression is the Moore-Penrose pseudo-inverse formula. We note that in some derivations, the vector v in (4.47) is introduced as a Lagrange multiplier.

Example 4.46. Consider again the linear system of Example 4.33

$$\dot{x}(t) \ = \ Ax(t) + Bu(t), \qquad t \in [0, T],\ x(t) \in \mathbb{R}^n,\ u(t) \in \mathbb{R}^m.$$

The "reachability problem" is the following. Given a "target state "$\bar{x} \in \mathbb{R}^n$, does there exist an input signal $u_{[0,T]} := \{u(t);\ t \in [0, T]\}$ such that the solution of the system starting from $x(0) = 0$ satisfies $x(T) = \bar{x}$? i.e. can the target state $\bar{x}$ be "reached" from the zero state by applying an input over the time interval $[0, T]$.

Recall the "reachability operator" $\mathcal{R} : \mathsf{L}^2_m[0, T] \to \mathbb{R}^n$ defined by the variations of constant formula (4.37)

$$x(T) \ = \ \int_0^T e^{A(T-t)} B\ u(t)\ dt \ =: \ \mathcal{R}\,(u)\,.$$

In terms of this operator, *the reachability problem is feasible iff* $\bar{x} \in \mathsf{Im}(\mathcal{R})$. This is an infinite-dimensional problem, but since $\mathsf{Im}(\mathcal{R}) = \mathsf{Im}\big(\mathcal{R}\mathcal{R}^{\dagger}\big)$, and the adjoint maps from a finite-dimensional space $\mathcal{R}^{\dagger} : \mathbb{R}^n \to \mathsf{L}^2_m[0, T]$, then the composition $\mathcal{R}\mathcal{R}^{\dagger} : \mathbb{R}^n \to \mathbb{R}^n$ is a matrix, and we can reduce this problem to a finite-dimensional one!

The adjoint $\mathcal{R}^{\dagger}$ has already been calculated in (4.38). To calculate the composition $\mathcal{R}\mathcal{R}^{\dagger}$, let it act on a vector v

$$\begin{aligned} \mathcal{R}\mathcal{R}^{\dagger} v \ = \ \mathcal{R}\left(\mathcal{R}^{\dagger} v\right) \ &= \int_0^T e^{A(T-t)} B \left(B^* e^{A^*(T-t)} v\right)\ dt \qquad \text{(from (4.38))} \\ &= \left(\int_0^T e^{A(T-t)} BB^* e^{A^*(T-t)}\ dt\right)\ v = \left(\int_0^T e^{A\tau} BB^* e^{A^*\tau}\ d\tau\right)\ v \\ &=: \ W_T\ v \end{aligned}$$

where the expression is simplified by the change of integration variable $\tau := T - t$.

The $n \times n$ matrix $W_T := \mathcal{R}\mathcal{R}^{\dagger}$ is known as the "reachability Grammian" (over the time horizon $[0, T]$). The fact that $\mathsf{Im}(\mathcal{R}) = \mathsf{Im}\big(\mathcal{R}\mathcal{R}^{\dagger}\big)$ says that any target states $\bar{x} \in \mathbb{R}^n$ is

[6] If A is a bounded operator, then the affine space $\{x;\ Ax = b\}$ is a translation of $\mathsf{Nu}(A)$. It is therefore closed, and a unique solution to the minimum distance problem exists.

reachable from zero iff the reachability Grammian $W_T := \mathcal{R}\mathcal{R}^\dagger$ is full rank. We can in fact say more. Note that this is an underdetermined linear problem in the sense of Example 4.45, and therefore the optimal (minimum L^2 norm) input $\bar{u}$ is given by

$$\begin{aligned} \bar{u} &= \mathcal{R}^\dagger \left(\mathcal{R}\mathcal{R}^\dagger\right)^{-1} \bar{x} \\ \Rightarrow \qquad \bar{u}(t) &= B^* e^{A^*(T-t)}\, W_T^{-1} \bar{x}, \qquad\qquad t \in [0, T]. \end{aligned} \tag{4.48}$$

Note that $W_T^{-1}\bar{x} \in \mathbb{R}^n$ is a vector, and therefore this formula gives the optimal input as linear combinations of the entries of the matrix-valued function $B^* e^{A^*(T-t)}$ of t.

The calculation (4.48) of optimal inputs is ultimately a consequence of the projection theorem. When adjoint calculations can be made, explicit expressions for optimal solutions like (4.48) can be obtained.

4.6 Geometric Interpretations of Adjoints

In a Hilbert space, any functional is represented by a vector in the same space and this enables geometric interpretations such as orthogonality and alignment in terms of vectors. In a Banach space, it is possible to give functionals a geometric interpretation in terms of objects in the primal, rather than the dual space. The key idea is that (after a normalization) each functional is associated with a *hyperplane*, i.e. a co-dimension 1 subspace. First, some terminology.

Definition 4.47. *Let* $\mathsf{S} \subset \mathsf{V}$ *be a subspace of a vector space* V. *Any coset* $v + \mathsf{S}$ *of* S *is called an* affine space. *The dimension of* V/S *is called the* co-dimension *of the affine space* $v + \mathsf{S}$. *An affine space of co-dimension* 1 *is called a* hyperplane.

Linear functionals and the notion of duality is more general than that for an inner product space, and can be defined for any vector space without an inner product. Inner products however make it easy to visualize linear functionals since any linear functional is expressed as the inner product with a particular vector. A natural question is whether there is a similar geometrical view of linear functionals without inner products. Figure 4.9 depicts such a visualization using *level sets*.

For any linear functional $\mathrm{w} : \mathsf{V} \to \mathbb{R}$ and any real number α, the α-level set $\mathsf{S}^{\mathrm{w}}_\alpha$ of w is

$$\mathsf{S}^{\mathrm{w}}_\alpha \;:=\; \{v \in \mathsf{V};\; \langle \mathrm{w}\,,\, v\rangle = \alpha\}\,.$$

Note that with this terminology, the null space is the 0-level set ($\mathsf{Nu}(\mathrm{w}) = \mathsf{S}^{\mathrm{w}}_0$). Level sets of different levels are disjoint, and each element of V belongs to one and only one level set. All level sets are co-sets of the Null space $\mathsf{S}^{\mathrm{w}}_0$, and therefore the set of all level sets is isomorphic to the quotient space $\mathsf{V}/\mathsf{Nu}(\mathrm{w})$. Recall (Figure 1.9 of Chapter 1) that the quotient space is isomorphic to the image $\mathsf{V}/\mathsf{Nu}(\mathrm{w}) \sim \mathsf{Im}(\mathrm{w}) = \mathbb{R}$, which in this case is just $\mathbb{R}$ (unless w is the zero functional). Therefore each level set of a functional is a co-dimension 1 affine space, i.e. a hyperplane.

Figure 4.9 depicts a visualization of functionals. The level sets of each functional form a family of parallel hyperplanes in V. The value of $\mathrm{w}(v)$ is determined by which level set the vector v is in. In an inner product space V, the functional w is represented by an inner product with a particular element $w \in \mathsf{V}$, i.e. $\mathrm{w}(v) = \langle w\,,\, v\rangle$, and thus the level-sets hyperplanes are all orthogonal to the vector w. In a vector space which is not an inner-product space, the functional w can not in general be represented by a vector in V. However, its level sets are in V, and can thus provide a visualization of w using objects (the level sets) that are in V.

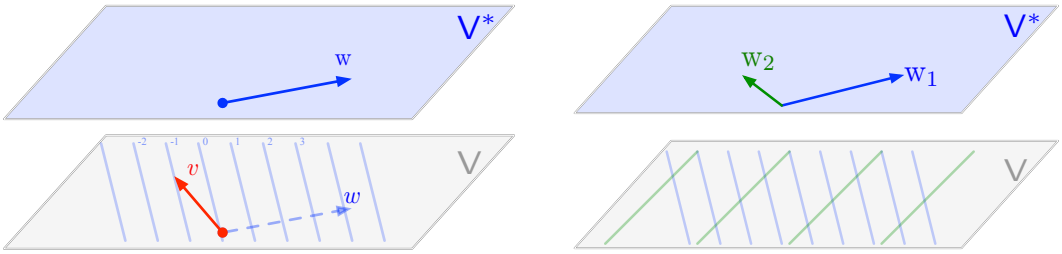


Figure 4.9: *(Left)* A linear functional $\mathrm{w} : \mathsf{V} \longrightarrow \mathbb{R}$ can be visualized in V using its level sets, shown here labeled with integer level values. The level sets are hyperplanes (translates of a co-dimension-1 subspace) in V. If V is an inner-product space, the functional w can be represented by a vector $w \in \mathsf{V}$ (shown as the dashed arrow) so that $\mathrm{w}(v) = \langle w\,,\, v\rangle$. *(Right)* Note that the longer a vector $\mathrm{w} \in \mathsf{V}^*$ is, the tighter is the spacing between its level sets.

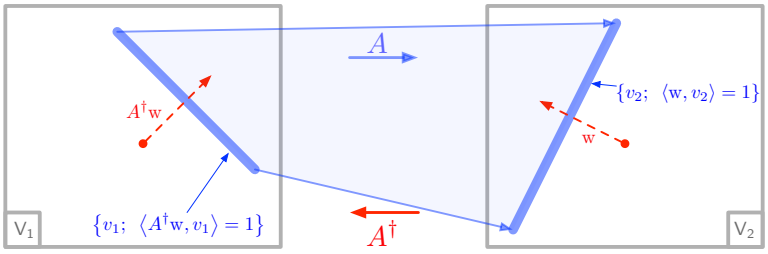


Figure 4.10: Geometric interpretation of the adjoint $A^\dagger$. Each linear functional is identified with its unit hyperplane. The inverse image of the hyperplane of $\mathrm{w} \in \mathsf{V}_2^*$ is precisely the hyperplane of $A^\dagger\mathrm{w} \in \mathsf{V}_1^*$. In other words, if A is viewed as mapping hyperplanes to hyperplanes (i.e. as sets), then $A^\dagger$ is the inverse of A viewed as mapping hyperplane sets (but not the inverse as mapping individual elements).

The interpretation of functionals as hyperplanes makes it possible to give a geometric interpretation to the operator adjoint as well. Let w be a functional on a Banach space V, and consider the hyperplane which is the 1-level set[7]

$$\mathsf{S}^{\mathrm{w}} \;:=\; \{v \in \mathsf{V};\;\; \langle \mathrm{w}\,,\, v\rangle = 1\}\,.$$

We will call S^{w} the unit hyperplane of w. Each functional in $\mathrm{w} \in \mathsf{V}^*$ is uniquely identified with its unit hyperplane. Since adjoints map functionals to functionals, we can give geometric interpretations to adjoints by understanding how they map hyperplanes to hyperplanes. This is illustrated in Figure 4.10. Let $A : \mathsf{V}_1 \to \mathsf{V}_2$ be a bounded operator, and let $\mathrm{w} \in \mathsf{V}_2^*$ be a linear functional on V_2. Its unit hyperplane is the set

$$\mathsf{S}^{\mathrm{w}} = \big\{v_2 \in \mathsf{V}_2;\;\; \langle \mathrm{w}\,,\, v\rangle = 1\big\}\,.$$

The inverse image under A of this set is

$$\big\{v_1 \in \mathsf{V}_1;\;\; \langle \mathrm{w}\,,\, Av_1\rangle = 1\big\} \;\;=\;\; \big\{v_1 \in \mathsf{V}_1;\;\; \big\langle A^\dagger\mathrm{w}\,,\, v_1\big\rangle = 1\big\} \;\;=\;\; \mathsf{S}^{A^\dagger\mathrm{w}},$$

i.e. the unit hyperplane of the functional $A^\dagger\mathrm{w} \in \mathsf{V}_1^*$.

This has an interesting interpretation as depicted in Figure 4.10 in the case when A is invertible. The operator $A : \mathsf{V}_1 \to \mathsf{V}_2$ maps hyperplanes in V_1 to hyperplanes in V_2. It can therefore be also viewed as mapping hyperplane sets to hyperplane sets. The adjoint $A^\dagger$ also maps hyperplane sets to hyperplane sets, and as such "set mappings", $A^\dagger$ is the inverse of A. They are however not inverses as mappings of individual elements of the vector spaces.

[7]The number 1 is chosen arbitrarily here. We could choose any other number α provided we use the same level for all functionals.

Appendix

4.A Riesz Lemma

In $\mathbb{R}^n$, any proper subspace has non-zero vectors orthogonal to it. In fact, any hyperplane passing through the origin (co-dimension 1 subspace) is uniquely determined by the direction of a vector perpendicular to it. Thus there is a one-to-one correspondence between hyperplanes and (unit-length) vectors, and this facilitates working with hyperplanes. Now let S be a proper, closed subspace of a Hilbert space V. How do we construct a vector $e \in \mathsf{V}$ orthogonal to S? Well, since S has an orthogonal complement, we can simply pick any vector in $\mathsf{S}^\perp$. However, we will instead use an argument illustrated in Figure 4.11 that generalizes to Banach spaces. We first note another equivalent characterization of orthogonality to a subspace in a Hilbert space.

Lemma 4.48. *Let $\mathsf{S} \subset \mathsf{V}$ be a subspace of an inner product space V. For any vector $e \in \mathsf{V}$*

$$\forall w \in \mathsf{S}, \quad \langle e, w \rangle = 0 \qquad (i.e.\ e \perp \mathsf{S}) \tag{4.49}$$

$$\Leftrightarrow \qquad \forall w \in \mathsf{S}, \quad \|e - w\| \geq \|e\|. \tag{4.50}$$

Proof. The direction ($\Rightarrow$) is immediate by Pythagoras. Since w and e are orthogonal, $\|e-w\|^2 = \|e\|^2 + \|w\|^2$, and therefore $\|e\|$ is a lower bound on $\|e-w\|$ for all choices of $w \in \mathsf{S}$. For the converse ($\Leftarrow$), note that $\|e-w\| \geq \|e\|$ implies that

$$\begin{aligned} \|e-w\|^2 &= \langle e-w, e-w \rangle = \|e\|^2 + \|w\|^2 - 2\langle e, w \rangle \geq \|e\|^2 \\ \Rightarrow \qquad \|w\|^2 &\geq 2\langle e, w \rangle . \end{aligned}$$

Now replace $w \in \mathsf{S}$ in the last expression by any scaling of it $\alpha w \in \mathsf{S}$ and observe

$$\|\alpha w\|^2 \geq 2\langle e, \alpha w \rangle \qquad \Leftrightarrow \qquad \alpha^2 \|w\|^2 \geq 2|\alpha| \langle e, w \rangle \qquad \Leftrightarrow \qquad |\alpha| \|w\|^2 \geq 2 \langle e, w \rangle .$$

The only way this inequality holds for all $\alpha \in \mathbb{R}$ is if $\langle e, w \rangle = 0$. Therefore $e \perp \mathsf{S}$. □

The interesting feature of the above lemma is that the criterion (4.49) is only applicable in an inner product space, but the criterion (4.50) is applicable in any normed space. We can therefore attempt to use the latter to characterize a type of "orthogonality" in a normed space without an inner product. This is the context of the Riesz lemma.

Lemma 4.49 (Riesz Lemma)**.** *Let $\mathsf{S} \subset \mathsf{V}$ be a closed proper subspace of a Banach space V. Then for any $\epsilon > 0$, there exists a vector e such that*

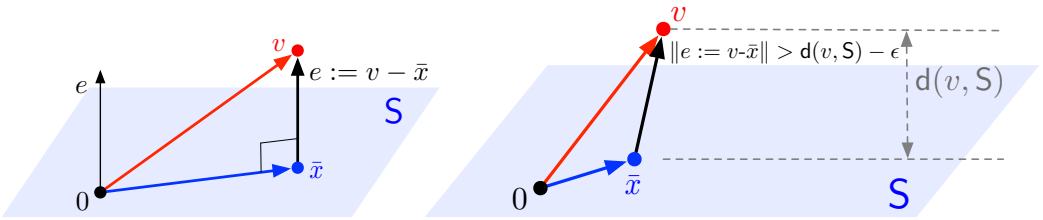
$$\forall y \in \mathsf{S}, \quad \|e - y\| > (1-\epsilon)\|e\|. \tag{4.51}$$

If V is reflexive, $\epsilon = 0$ can be taken in the above.

Before proving this theorem, we give some geometrical intuition. Since $\mathsf{S} \subset \mathsf{V}$ is a proper closed subspace, then there exists $v \notin \mathsf{S}$. In a Hilbert space, we can use the orthogonal projection $\bar{x}$ of v onto S (see Figure 4.11a), and recall that $\bar{x}$ solves the minimum distance problem

$$\inf_{x \in \mathsf{S}} \|v - x\| = \|v - \bar{x}\|.$$

The projection theorem states that $e := v - \bar{x}$ is orthogonal to all of S. To generalize this to Banach spaces, the key is not to use an orthogonal projection (which is not applicable), but rather the minimum distance problem itself. The infimum of this problem may not be achieved, but we can always construct "almost solutions" and those will provide "almost orthogonal" vectors in the sense of criterion (4.50).



(a) in a Hilbert space, we can take any vector $v \notin \mathsf{S}$, project v onto S, and obtain $e := v - \bar{x} := v - \Pi v$. The vector e is then orthogonal to the subspace S.

(b) In a Banach space, we can take any vector $v \notin \mathsf{S}$, but then for any ϵ, we can always find a vector $\bar{x}$ such that $\|e\| = \|v - \bar{x}\|$ is slightly larger than the distance $\mathsf{d}(v, \mathsf{S})$. This vector has the property that for any $w \in \mathsf{S}$, $\|e - w\| > (1 - \epsilon)\|e\|$, and this can be interpreted as "almost orthogonality".

Figure 4.11: Find a vector that is orthogonal in a Hilbert space to a closed, proper subspace S, as well as finding a vector that has an "almost orthogonal" property in a Banach space (Riesz Lemma 4.49).

Proof. Since S is a closed, proper subspace, there exists a vector $v \notin \mathsf{S}$. Consider the minimum distance problem to v, for which a minimum may not exist, but for any $\delta > 0$ we can find $\bar{x} \in \mathsf{S}$ such that (see Figure 4.11b)

$$\|v - \bar{x}\| \;<\; \mathsf{d}(v, \mathsf{S}) + \delta, \qquad \mathsf{d}(v, \mathsf{S}) \;:=\; \inf_{x \in \mathsf{S}} \|v - x\|. \tag{4.52}$$

The "error vector" $e := v - \bar{x}$ provides the answer since for any $y \in \mathsf{S}$

$$\frac{\|e - y\|}{\|e\|} \;=\; \frac{\|(v - \bar{x}) - y\|}{\|v - \bar{x}\|} \overset{1}{>} \frac{\|v - (\bar{x} + y)\|}{\mathsf{d}(v, \mathsf{S}) + \delta} \overset{2}{\geq} \frac{\mathsf{d}(v, \mathsf{S})}{\mathsf{d}(v, \mathsf{S}) + \delta},$$

where $\overset{1}{>}$ follows from (4.52), and $\overset{2}{\geq}$ follows from $(\bar{x} + y) \in \mathsf{S}$. Finally, given $\epsilon > 0$, choose $\delta > 0$ such that

$$\frac{\mathsf{d}(v, \mathsf{S})}{\mathsf{d}(v, \mathsf{S}) + \delta} \;\geq\; (1 - \epsilon) \qquad \Rightarrow \qquad \frac{\|e - y\|}{\|e\|} \;\geq\; (1 - \epsilon).$$

□

Exercises

Exercise 4.1

Show by a similar argument to that of Example 4.6 that the dual of $\mathbb{R}^n_1$ is $\mathbb{R}^n_\infty$. More generally, show by using the Minkowski inequality that for $1/p + 1/q = 1$, the dual of $\mathbb{R}^n_p$ is $\mathbb{R}^n_q$.

Exercise 4.2

Show by counterexample that setting $\mathrm{w}^e_1(x) = 0$ in the one step extension (4.25) in the proof of the Hahn-Banach theorem does not work. Take $\mathbb{R}^2$ with the $\|.\|_\infty$ norm. Consider the horizontal axis as S and $\mathrm{w}(x_1, 0) = x_1$, so the norm of w restricted to the horizontal axis is 1. If $x = (1, 1)$ is chosen, show that the extension with $\mathrm{w}^e_1(x) = 0$ will actually have $\|\mathrm{w}^e_1\| = 2$.

Exercise 4.3

Show that for a Banach space V, if V^* is separable, then so is V.
Hint: Take a countable dense set $\{\mathrm{w}_k\}$ *in the unit sphere of* V^*. *For each functional, find* $v_k \in \mathsf{V}$ *such that* $\mathrm{w}_k(v_k) \geq 1/2$. *Show that the span of* $\{v_k\}$ *is dense in* V.

Exercise 4.4

In a vector space over the complex scalars, suppose $[w, u]$ is any bilinear complex-valued form on two vectors w and u. Show that

$$\mathbb{R}([u,v]) \;\leq\; f\left(\|u\|, \|v\|\right) \qquad \Rightarrow \qquad \Big|\,[u,v]\Big| \;\leq\; f\left(\|u\|, \|v\|\right),$$

where f is any function of two real variables.

Hint: Try "rotating" the vector v to $e^{j\theta}v$.

Chapter 5

Eigenvectors, Invariant Subspaces and the Spectrum

Finding a subspace that is invariant to a matrix or an operator enables a decomposition of the operator into simpler pieces. The simplest invariant subspaces are spanned by eigenvectors, and those are one-dimensional invariant subspaces. A matrix or operator whose eigenvectors span the entire space are diagonalizable. When this is not possible, higher dimensional invariant subspaces lead to the Jordan form for matrices. For general operators, the concept of eigenvalues is generalized to that of the spectrum which is a subset of the complex plane. The resolvent of an operator is an operator-valued function on the complex plane. The behavior of this function encodes many of the properties of the operator.

Introduction

You have probably heard that an eigenvector v of a (square) matrix A is such that $Av = \lambda v$ for some (possibly) complex number λ, which we call its corresponding eigenvalue. Geometrically, this means that when A acts on the vector v, it does not rotate or change its direction, but simply scales it by λ.

It is important to notice that eigenvectors are not uniquely defined because if v is an eigenvector, then so is any non-zero scaling αv of it since $A(\alpha v) = \alpha Av = \alpha\ \lambda v\ =\ \lambda(\alpha v)$. Thus it maybe more accurate to speak of an "eigendirection" or to say that the one-dimensional subspace $\mathsf{span}\{v\}$ is A-invariant, i.e.

$$A\left(\mathsf{span}\{v\}\right)\ \subseteq\ \mathsf{span}\{v\}\,, \tag{5.1}$$

where $A\left(\mathsf{span}\{v\}\right)$ is the image of the one-dimensional subspace $\mathsf{span}\{v\}$ (as a set) when acted on by A. The image of a subspace is also a subspaces of dimension no larger than the original subspace, so $A\left(\mathsf{span}\{v\}\right)$ is either of dimension 1 (in which case $A\left(\mathsf{span}\{v\}\right) = \mathsf{span}\{v\}$) or 0.

The statement (5.1) is equivalent to $Av = \lambda v$, but can be generalized to higher dimensional spaces, where we say that a subspace S is A-invariant if

$$A\mathsf{S}\ \subseteq\ \mathsf{S}.$$

Eigenvectors (or more accurately their spans) are one-dimensional invariant subspaces. However, other types of (higher dimensional) invariant subspaces are useful because they allow for decomposing the matrix or operator into simpler forms as we will see in the next section.

Yet another way understand the relation $Av = \lambda v$ is to rewrite it equivalently as

$$(\lambda I - A)\, v \;=\; 0.$$

Thus for a matrix, λ is an eigenvalue iff $(\lambda I - A)$ is singular, and all the corresponding eigenvectors are the null space of $(\lambda I - A)$. We can therefore think of $R_A(\lambda) := (\lambda I - A)^{-1}$ as a function defined everywhere on the complex plane except at the eigenvalues of A. This function is called the resolvent of A and its behavior (as a function) says much more about the operator A than its eigenvalues. For operators, the spectrum is the set of points in the complex plane where $R_A(.)$ is not a bounded operator, and can be either isolated points or other types of sets.

5.1 Invariant Subspaces and Eigenvectors

Eigenvalues and invariant subspaces are defined only for square matrices, or more generally operators between a vector space and itself. The simplest square matrices to study are the diagonal matrices since they are basically a decoupled system of scalar relations

$$y = Ax \qquad \Leftrightarrow \qquad \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} a_1 & & \\ & \ddots & \\ & & a_n \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \qquad \Leftrightarrow \qquad \begin{array}{rcl} y_1 & = & a_1\, x_1, \\ & \vdots & \\ y_n & = & a_n\, x_n. \end{array}$$

There is another way to see how special diagonal matrices are. Let $A = \mathsf{diag}(a_1, \ldots, a_n)$ be a diagonal matrix, and consider the n subspaces $\mathsf{span}\{e_i\} =: \mathsf{E}_i \subset \mathbb{R}^n$ that are spanned by the canonical basis vectors e_i

$$e_i := (0, \cdots, 0, 1, 0, \cdots, 0) \quad \Rightarrow \quad Ae_i = a_i e_i \quad \Leftrightarrow \quad A\mathsf{E}_i \subseteq \mathsf{E}_i,$$
$$\uparrow i\text{'th position}$$

The statement $Ae_i = a_i e_i$ means that when A acts on e_i, it returns a vector in the same direction as e_i, but scaled by the factor a_i. In other words, A does not change the direction of e_i when it acts on it. Since $\mathsf{E}_i = \mathsf{span}\{e_i\}$ is a one-dimensional subspace, this statement is equivalent to saying that when A acts on the entire subspace E_i, all the resulting vectors remain in E_i. We write this as $A\mathsf{E}_i \subseteq \mathsf{E}_i$, and we say that E_i is an *invariant subspace* for A.

Now consider the "next best thing" to a diagonal matrix, namely a block-diagonal matrix. Consider for example the 3×3 matrix A and the canonical vectors

$$A \;:=\; \begin{bmatrix} 1 & 2 & 0 \\ 3 & 4 & 0 \\ 0 & 0 & 5 \end{bmatrix}, \qquad e_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \;\; e_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \;\; e_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

It is clear that we have $Ae_3 = 5e_3$, so E_3 is invariant, but $Ae_1 \notin \mathsf{E}_1$ and $Ae_2 \notin \mathsf{E}_2$. However, we do have $Ae_1 \in \mathsf{E}_1 \oplus \mathsf{E}_2$ and $Ae_2 \in \mathsf{E}_1 \oplus \mathsf{E}_2$ (all vectors with zero entries in the third component), and therefore $A\,(\mathsf{E}_1 \oplus \mathsf{E}_2) \subseteq \mathsf{E}_1 \oplus \mathsf{E}_2$. Thus $\mathsf{E}_1 \oplus \mathsf{E}_2$ is an invariant subspace, but its dimension is higher than 1.

The key insight to take from the previous paragraphs is to actually go in reverse. Given a not necessarily diagonal (or block-diagonal) matrix, if we can find invariant subspaces, then we can find a new basis so that the representation in that basis is diagonal (or block diagonal). This leads to the following fundamental concepts.

Definition 5.1. *Consider a linear operator $A : \mathsf{V} \to \mathsf{V}$ on a vector space V.*

1. *A subspace $\mathsf{S} \subseteq \mathsf{V}$ is called A-*invariant *if $A\mathsf{S} \subseteq \mathsf{S}$.*

2. *An invariant subspace is called* minimal *if it does not contain any other invariant subspaces (other than* 0 *and itself).*

3. *An* eigenvector *of A is a non-zero vector v such that $Av = \lambda v$ for some $\lambda \in \mathbb{C}$. The number λ is called the* eigenvalue *of A associated with the eigenvector v.*

   In particular, an eigenvector of A spans a 1-dimensional (and therefore minimal) A-invariant subspace.

   The set of all eigenvectors with eigenvalue λ is precisely $\mathsf{Nu}(\lambda I - A)$.

The ultimate goal in analysis of any linear operator is to find all of its minimal invariant subspaces. Once those are found, the operator can be "decomposed" into its simplest possible form. To illustrate, we begin with the case of square matrices that have a sufficient number of linearly independent eigenvectors.

Given an $n \times n$ matrix A, assume it has n linearly independent eigenvectors

$$A\, v_i \;=\; \lambda_i\, v_i, \qquad i = 1, \ldots, n. \tag{5.2}$$

It is an elementary, but powerful observation that these n matrix-vector relations can be rewritten as the following single matrix equation

$$\begin{bmatrix} Av_1 & \cdots & Av_n \end{bmatrix} = \begin{bmatrix} \lambda_1 v_1 & \cdots & \lambda_n v_n \end{bmatrix}$$

$$\Updownarrow$$

$$\begin{bmatrix} & A & \end{bmatrix} \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix} = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} \quad \Leftrightarrow \quad AV = V\Lambda, \tag{5.3}$$

where V is a matrix whose columns are the eigenvectors of A, and Λ is the diagonal matrix made up of the eigenvalues of A. Equation (5.3) states that V is the similarity transformation that diagonalizes A

$$\Lambda \;:=\; \mathsf{diag}(\lambda_1, \ldots, \lambda_n) \;=\; V^{-1}AV. \tag{5.4}$$

Note that V is non-singular since its columns were assumed to be linearly independent. We have therefore obtained the following criterion for a matrix to be diagonalizable.

Lemma 5.2. *Let A be $n \times n$ matrix. A is diagonalizable with a similarity transformation iff it has a* full set of eigenvectors*, i.e. n linearly independent vectors with $Av_i = \lambda_i v_i$, $i = 1, \ldots, n$.*

In this case, the diagonalizing similarity transformation (5.3) V is made up of the eigenvectors as its columns, and

$$V^{-1}AV \;=\; \mathsf{diag}(\lambda_1, \ldots, \lambda_n)\,.$$

Higher Dimensional Invariant Subspaces

We now consider invariant subspaces that are of dimension possibly larger than 1 (i.e., they don't correspond to eigenvectors). First we point out a very useful observation about the connection between invariant subspaces and "block-triangular" matrix forms. Let A be a linear operator on a vector space S. Suppose we find an A-invariant subspace $\mathsf{S} \subsetneq \mathsf{V}$,

and a complement of it S^c (S^c does not itself have to be A-invariant). The decomposition $\mathsf{V} = \mathsf{S} \oplus \mathsf{S}^c$ induced a 2×2 block partitioning of A as follows

$$A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} : \begin{matrix} \mathsf{S} \\ \oplus \\ \mathsf{S}^c \end{matrix} \longrightarrow \begin{matrix} \mathsf{S} \\ \oplus \\ \mathsf{S}^c \end{matrix}.$$

The $\{21\}$-block is $\Pi_{\mathsf{S}^c} A|_{\mathsf{S}}$, and must be zero because A maps any vector in S to a vector whose S^c component is zero (since $A(\mathsf{S}_1) \subseteq \mathsf{S}_1$). Thus having an invariant subspace is equivalent to finding a representation that is in block-triangular form.

Now we turn to the question of existence of eigenvectors. Not every matrix has a full set of eigenvectors. We have already seen one extreme example, namely diagonal matrices where every canonical basis vector e_i is an eigenvector. Another extreme example is any multiple αI of the identity, where every vector is an eigenvector! At the other extreme is the $n \times n$ "Jordan block"

$$J := \begin{bmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{bmatrix}, \qquad \begin{bmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \lambda \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

which has e_1 as an eigenvector with eigenvalue λ. This matrix however has no other eigenvectors (Exercise 5.2) regardless of its size n! It does however have higher-dimensional invariant subspaces as can be seen from the following

$$\left[\begin{array}{c:c} J_r & * \\ \hdashline 0 & J_{n-r} \end{array}\right] \begin{bmatrix} * \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} * \\ \vdots \\ 0 \end{bmatrix}, \tag{5.5}$$

where we have partitioned J_n into smaller pieces, and $*$ indicates possibly non-zero entries. Note that the block-diagonal portions are also Jordan block matrices of dimensions r and $n - r$ respectively, where $1 \leq r \leq n - 1$. The "block-upper-triangular" structure of (5.5) implies that for any such r

$$J(\mathsf{E}_1 \oplus \cdots \oplus \mathsf{E}_r) \subseteq \mathsf{E}_1 \oplus \cdots \oplus \mathsf{E}_r,$$

i.e. that $\mathsf{E}_1 \oplus \cdots \oplus \mathsf{E}_r$ is a nested sequence of invariant subspace, each of dimension r. Thus, even though a Jordan block has only one eigenvector, it does have invariant subspaces of higher dimension, all arranged in a nested sequence. In fact, having a nested sequence of invariant subspaces implies an "upper-triangular" form.

Theorem 5.3. *Let $\mathcal{A}$ be a linear operator on a n-dimensional vector space V.*

1. *$\mathcal{A}$ has at least one eigenvector.*
2. *If S is an $\mathcal{A}$-invariant subspace, then there exists at least one eigenvector of $\mathcal{A}$ in S.*
3. *There exists a basis $\{v_i\}$ of V such that the matrix representation of $\mathcal{A}$ in that basis is upper (or lower) triangular. The eigenvalues of $\mathcal{A}$ are the diagonal entries in either triangular form.*

We have already seen from the Jordan block example that a matrix can have only one eigenvector regardless of the dimension of the matrix. The fact that any matrix must have at least one eigenvector follows from the fact that any polynomial has (possibly complex)

roots, and the following construction[1] of a sequence of vectors from powers of $\mathcal{A}$. Start with any non-zero vector $v \in \mathsf{S}$ and consider the set of vectors

$$\left\{v, \mathcal{A}v, \mathcal{A}^2 v, \ldots, \mathcal{A}^n v\right\}.$$

This is a set of $n+1$ vectors in an n-dimensional vector space, so it must be a linearly dependent set, i.e. there is a non-trivial linear combination such that

$$0 \;=\; a_0 v + a_1 \mathcal{A}v + \ldots + a_n \mathcal{A}^n v \;=\; \left(a_0 I + a_1 \mathcal{A} + \ldots + a_n \mathcal{A}^n\right) v \;=:\; p(\mathcal{A})\; v.$$

Let $m \leq n$ be the integer of the highest nonzero coefficient above. Note that we can always find a v such that $m \geq 1$ (otherwise $Av = 0$ for all vector v). The polynomial p then has m roots, and can be factored as $p(x) = a_m(x - z_1)\cdots(x - z_m)$, where $\{z_i\}$ are the zeros of the polynomial. It then follows (Exercise 5.3) that $p(\mathcal{A})$ is also factored as

$$0 \;=\; p(\mathcal{A})\; v \;=\; a_m \left(\mathcal{A} - z_1 I\right) \cdots \left(\mathcal{A} - z_m I\right)\; v.$$

Thus the action of a sequence of matrix products on v produces zero, so that must happen for some index (call it r) and we have

$$0 \;=\; \left(\mathcal{A} - z_r I\right) \underbrace{\left(\mathcal{A} - z_{r+1} I\right) \cdots \left(\mathcal{A} - z_m I\right)\; v}_{w} = \;\left(\mathcal{A} - z_r I\right)\; w.$$

Therefore w is an eigenvector of $\mathcal{A}$ with eigenvalue z_r.

It is tempting to think about the preceding construction as a numerical algorithm for finding eigenvalues/vectors. However, the step of finding the roots of a polynomial given its coefficients has increasing (with polynomial order, thus with matrix size) sensitivity to small perturbations in the polynomial's coefficients, and therefore will not produce reliable results for large matrices.

The second clause of Theorem 5.3 follows from the first. If S is $\mathcal{A}$-invariant, choose a subspace S^{c} complementary to S, then the decomposition $\mathsf{V} = \mathsf{S} \oplus \mathsf{S}^{\mathrm{c}}$ induces an upper triangular block partitioning of the operator $\mathcal{A}$

$$\begin{bmatrix} \mathcal{A}_{11} & \mathcal{A}_{12} \\ 0 & \mathcal{A}_{22} \end{bmatrix}.$$

Now $\mathcal{A}_{11}$ is a linear operator on a finite dimensional vector space S, and therefore has at least one eigenvector in S.

The final clause of the theorem follows by repeated applications of the above decomposition. Given $\mathcal{A}$, we find one eigenvector, call it v_1 with eigenvalue λ_1, and complete it to a basis $\{v_1, v_2, \ldots, v_n\}$ of V. The fact that v_1 is an eigenvector means that it spans an invariant subspace, and the matrix representation A_n of $\mathcal{A}$ in that basis is block upper triangular

$$A_n \;=\; \left[\begin{array}{c:c} \lambda_1 & * \\ \hdashline 0 & A_{n\text{-}1} \end{array}\right],$$

where $A_{n\text{-}1}$ is an $(n\text{-}1) \times (n\text{-}1)$ matrix. Now $A_{n\text{-}1}$ has at least one eigenvector as it acts on the n-1-dimensional subspace $\mathsf{span}\{v_2, \ldots, v_n\}$, and therefore we can now find a different basis for that subspace in which $A_{n\text{-}1}$ is also block lower triangular. Clearly this process

[1] This is the construction of a so-called *Krylov subspace*, and is very common in many problems in linear algebra (numerical and theoretical) and functional analysis.

can be repeated until we finally find a basis in which the matrix representation A_1 of $\mathcal{A}$ is upper triangular

$$A_1 = \begin{bmatrix} \lambda_1 & & * \\ & \ddots & \\ & & \lambda_n \end{bmatrix}. \tag{5.6}$$

It is useful to contrast this conclusion with that of Lemma 5.2. The latter says that an $n \times n$ matrix is diagonalizable iff it has n linearly independent eigenvectors, which is not always the case for any matrix. However, Theorem 5.3 says that *any* $n \times n$ matrix can be brought into upper (or lower) triangular form. This very useful form, especially when the new basis is selected to be orthonormal, is referred to as the *Schur form* of a matrix.

Geometrically, the fact that any matrix can be brought into the upper triangular form (5.6) is equivalent to finding n *properly nested* $\mathcal{A}$*-invariant subspaces*

$$0 \subsetneq \mathsf{S}_1 \subsetneq \mathsf{S}_2 \subsetneq \cdots \subsetneq \mathsf{S}_n = \mathsf{V}, \qquad \mathcal{A}(\mathsf{S}_i) \subseteq \mathsf{S}_i. \tag{5.7}$$

In the basis that produces the form (5.6) those nested subspaces are simply the sum of the canonical subspaces

$$\mathsf{S}_r = \mathsf{E}_1 \oplus \cdots \oplus \mathsf{E}_r = \left\{ x \in \mathbb{R}^n;\ x = (x_1, \ldots, x_r, 0, \ldots, 0) \right\}.$$

Thus Theorem 5.3 is equivalent to the statement that any linear operator on an n-dimensional space has a sequence of properly nested invariant subspaces (5.7).

Example 5.4. This simple example is useful to summarize the results of this section

$$A = \left[\begin{array}{cc:c:cc} \lambda_1 & 0 & 0 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 & 0 \\ \hdashline 0 & 0 & \lambda_2 & 0 & 0 \\ \hdashline 0 & 0 & 0 & \lambda_3 & 1 \\ 0 & 0 & 0 & 0 & \lambda_3 \end{array}\right] \begin{array}{c} \mathsf{S}_1 \\ \oplus \\ \mathsf{S}_2 \\ \oplus \\ \mathsf{S}_3 \end{array} \longrightarrow \begin{array}{c} \mathsf{S}_1 \\ \oplus \\ \mathsf{S}_2, \\ \oplus \\ \mathsf{S}_3 \end{array}$$

where we assume that λ_1 is distinct from either λ_2 or λ_3.

S_2 is a 1-dimensional invariant subspace corresponding to the eigenvalue λ_2, and therefore there is a single eigenvector for λ_2. On the other hand S_3 is a 2-dimensional invariant subspace corresponding to the eigenvalue λ_3. In this case however, there is only a single eigenvector although the invariant subspace is 2-dimensional (this is a Jordan block of size 2). S_1 is a 2-dimensional invariant subspace corresponding to the eigenvalue λ_1. In this case, every vector in S_2 is actually an eigenvector. This happens because the $\{11\}$-block of A is $\lambda_1 I$, where I is the 2×2 identity matrix. Thus S_1 is an invariant subspace made up of vectors all of which are eigenvectors for the same eigenvalue. Therefore S_1 is *not a minimal invariant subspace* since it has lower dimensional subspaces that are also invariant. Note that if we chose another basis for S_1, the matrix form above would remain the same, i.e. we can choose any two linearly independent vectors in S_1 as a basis, and the matrix form remains as above.

5.2 The Spectrum of an Operator

In the finite dimensional case, a complex number λ is an eigenvalue of an operator A if $\mathsf{Nu}(\lambda I - A) \neq 0$. The fact that $\lambda I - A$ is not invertible is equivalent to $\mathsf{Nu}(\lambda I - A)$ having non-trivial elements, which are precisely the eigenvectors of A associated with the eigenvalue

λ. In the more general case, the key object will still be the operator $\lambda I - A$, but it will fail to be invertible in several different ways, each of which will characterize a different part of the *spectrum* of the operator. To begin with however, we will first formally define and establish some general properties of the spectrum.

Definition 5.5. *Let $A : \mathsf{V} \to \mathsf{V}$ be a (possibly unbounded) operator on a Banach space V.*

1. *The* spectrum $\boldsymbol{\lambda}(A)$ *of A is the set of all points $\lambda \in \mathbb{C}$ such that $\lambda I - A$ is not boundedly invertible.*

   The spectral radius $|\boldsymbol{\lambda}(A)|$ *is the supremum of the moduli of all points in the spectrum*

   $$|\boldsymbol{\lambda}(A)| \;:=\; \sup_{\lambda \in \boldsymbol{\lambda}(A)} |\lambda|.$$

2. *The* resolvent set $\boldsymbol{\rho}(A)$ *of A is the complement of the spectrum, i.e. all $\lambda \in \mathbb{C}$ such that $(\lambda I - A)^{-1}$ exists and is a bounded operator on V, thus $\boldsymbol{\rho}(A) = \mathbb{C} \setminus \boldsymbol{\lambda}(A)$.*

   The resolvent $\mathsf{R}_A(.)$ *of A is the function $\mathsf{R}_A(\lambda) := (\lambda I - A)^{-1}$, which is a function from the resolvent set $\boldsymbol{\rho}(A) \subset \mathbb{C}$ to the algebra $\mathsf{B}(\mathsf{V})$ of all bounded operators on V.*

5.2.1 Bounded Operators

In the case of bounded operators, there are some easily established properties of its spectrum which we briefly cover.

Lemma 5.6. *Let $A : \mathsf{V} \to \mathsf{V}$ be a bounded operator on a Banach space V. The spectrum of A is bounded in the complex plane by its norm. Equivalently, the spectral radius of an operator is bounded by its norm*

$$\boldsymbol{\lambda}(A) \;\subseteq\; \Big\{\lambda \in \mathbb{C};\; |\lambda| \leq \|A\|\Big\} \qquad \Leftrightarrow \qquad |\boldsymbol{\lambda}(A)| \;\leq\; \|A\|.$$

This means that the spectrum is confined within the disk of radius $\|A\|$ in the complex plane. This lemma is an easy consequence of the Neumann series. If $\lambda > \|A\|$, then $\|A/\lambda\| < 1$ and

$$(\lambda I - A) \;=\; \lambda\,(I - A/\lambda)$$

is boundedly invertible since its Neumann series converges in $\mathsf{B}(\mathsf{V})$.

The spectrum of any bounded operator on a Banach space V has certain properties that follow from understanding the structure of the Banach algebra $\mathsf{B}(\mathsf{V})$ of all bounded operators on V. The first important property is that the set of all invertible elements in $\mathsf{B}(\mathsf{V})$ is open. In other words, given A invertible, then $A + \Delta$ is invertible for all "perturbations" such that $\|\Delta\| < \epsilon$ for sufficiently small norm ϵ. This is again a consequence of the Neumann series formula which implies

$$\begin{aligned}
A \text{ invertible in } \mathsf{B}(\mathsf{V}) \qquad &\Rightarrow \qquad A + \Delta = A\left(I + A^{-1}\Delta\right) \\
\|\Delta\| \;<\; \frac{1}{\|A^{-1}\|} \qquad &\Rightarrow \qquad \left\|A^{-1}\Delta\right\| \;\leq\; \left\|A^{-1}\right\| \|\Delta\| \;<\; 1 \\
&\Rightarrow \qquad \left\|(A+\Delta)^{-1}\right\| = \left\|\left(I + A^{-1}\Delta\right)^{-1} A^{-1}\right\| \leq \frac{\left\|A^{-1}\right\|}{1 - \|A^{-1}\Delta\|}.
\end{aligned}$$

Thus for all $\Delta \in \mathsf{B}(\mathsf{V})$ with $\|\Delta\| < 1/\left\|A^{-1}\right\|$, the element $A + \Delta$ is invertible with the above bound on the norm of its inverse. We now state this together with other corollaries.

Lemma 5.7. *Let* $\mathsf{B}(\mathsf{V})$ *be the Banach algebra of all bounded linear operators on the Banach space* V*. Given any* $A \in \mathsf{B}(\mathsf{V})$

1. *If* A *is invertible in* $\mathsf{B}(\mathsf{V})$*, then for any* $\Delta \in \mathsf{B}(\mathsf{V})$ *with* $\|\Delta\| < 1/\|A^{-1}\|$*, the operator* $A + \Delta$ *is invertible in* $\mathsf{B}(\mathsf{V})$ *with norm bound* $\|(A+\Delta)^{-1}\| \leq 1/(1 - \|\Delta\|\|A^{-1}\|)$*.*

   Thus the set of invertible elements is open in $\mathsf{B}(\mathsf{V})$*.*

2. *The resolvent set* $\boldsymbol{\rho}(A)$ *of* A *is an open set in* $\mathbb{C}$*.*

3. *The spectrum* $\boldsymbol{\lambda}(A)$ *of* A *is closed in* $\mathbb{C}$*.*

The second clause follows from the first by observing that if $\bar{\lambda}$ is in the resolvent set, then $\bar{\lambda}I - A$ is an invertible element, and for $\lambda = \bar{\lambda} + \epsilon$

$$\lambda I - A \;=\; (\bar{\lambda} + \epsilon)I - A \;=\; \big(\bar{\lambda}I - A\big) \;+\; \epsilon I.$$

Since $\|\epsilon I\| = |\epsilon|$, then the first clause implies that $\lambda I - A$ is boundedly invertible for all $|\epsilon| < 1/\| \big(\bar{\lambda}I - A\big)^{-1} \|$. The last clause follows from the second since the spectrum is the complement of the resolvent set, and is therefore closed in $\mathbb{C}$.

The next characterization is the same as the matrix case. If we know the spectrum of an operator, we also know the spectrum of its powers and its inverse (if it exists).

Lemma 5.8. *Given a bounded operator* $A : \mathsf{V} \to \mathsf{V}$ *on a Banach space* V *with spectrum* $\boldsymbol{\lambda}(A) \subseteq \mathbb{C}$*. Then*

$$\boldsymbol{\lambda}(A^n) \;=\; \big(\boldsymbol{\lambda}(A)\big)^n, \qquad \textit{and} \qquad \boldsymbol{\lambda}\big(A^{-1}\big) \;=\; 1/\boldsymbol{\lambda}(A)\,,$$

if A *is invertible in* $\mathsf{B}(\mathsf{V})$*.*

The first statement is the subject of Exercise 5.5. For the second statement, if A is invertible, then $0 \notin \boldsymbol{\lambda}(A)$, and rewrite

$$(\lambda I - A) \;=\; \lambda \left(A^{-1} - \lambda^{-1} I\right) A.$$

Then $(\lambda I - A)$ is boundedly invertible iff $\left(\lambda^{-1} I - A^{-1}\right)$ is boundedly invertible. Thus $\lambda \in \boldsymbol{\lambda}(A)$ iff $\lambda^{-1} \in \boldsymbol{\lambda}\big(A^{-1}\big)$.

We note that this lemma is a special case of the so-called *spectral mapping theorem*, which says that $\boldsymbol{\lambda}(f(A)) = (\boldsymbol{\lambda}(A))$ for any function f analytic in a neighborhood of the spectrum of A. In the lemma above, the functions $f(x) := x^n$ are analytic everywhere, and thus the spectral mapping works for any operator. For the case of $f(x) = 1/x$, this function is analytic outside of $x = 0$, and we assumed that A was invertible, so the spectrum of A (a closed set) does not contain the point $\lambda = 0$.

5.2.2 The Components of the Spectrum

Before we give the formal definitions, it will be useful to categorize the different ways an operator can fail to be invertible. Given a bounded operator $A : \mathsf{V} \to \mathsf{V}$ on a Banach space V, we will be interested in the following three categories.

1. If $\mathsf{Nu}(A) \neq 0$, then the operator A is not one-to-one, and therefore not invertible.

2. If $\mathsf{Nu}(A) = 0$ but $\mathsf{Im}(A) \neq \mathsf{V}$, then this operator is also not invertible. This case can be further classified into two possible categories.

   (a) $\mathsf{Im}(A) \neq \mathsf{V}$ but $\mathsf{Im}(A)$ is dense in V. In this case, the inverse A^{-1} is defined only on a dense subspace of V, and A^{-1} is an unbounded operator.

(b) $\mathsf{Im}(A) \neq \mathsf{V}$ and $\overline{\mathsf{Im}(A)} \neq \mathsf{V}$. Thus the closure $\overline{\mathsf{Im}(A)}$ has a non-zero co-dimension in the space V. In this case, there is no way to define A^{-1}.

It is useful to contrast these possibilities with the finite-dimensional case, which is constrained by the rank-nullity theorem statement (1.29) $\mathsf{dim}(\mathsf{Nu}(A)) + \mathsf{dim}(\mathsf{Im}(A)) = \mathsf{dim}(\mathsf{V})$, from which we can state that

$$\mathsf{Nu}(A) \neq 0 \qquad \Leftrightarrow \qquad \mathsf{Im}(A) \neq \mathsf{V}.$$

Thus the three categories listed above collapse to a single category in finite dimensions.

Example 5.9. To understand the different possibilities in the general case, the following examples are useful. Consider the "bilateral" and "unilateral" shift operators on $\ell^2(\mathbb{Z})$ and $\ell^2(\mathbb{N})$ respectively

$$\begin{aligned} \mathcal{S}\,(\ldots, u_{\text{-}1}, u_o, u_1, \ldots) &:= (\ldots, u_{\text{-}2}, u_{\text{-}1}, u_0, \ldots)\,, \\ \mathcal{S}^{\text{-}1}(\ldots, u_{\text{-}1}, u_o, u_1, \ldots) &:= (\ldots, u_0, u_1, u_2, \ldots)\,, \\ \mathcal{S}_r\,(u_o, u_1, \ldots) &:= (0, u_0, u_1, \ldots)\,, \qquad \mathcal{S}_l\,(u_o, u_1, \ldots) := (u_1, u_2, \ldots)\,. \end{aligned}$$

On $\ell^2(\mathbb{Z})$, $\mathcal{S}$ and $\mathcal{S}^{\text{-}1}$ are the right and left shift operators respectively. They are clearly inverses of each other. On $\ell^2(\mathbb{N})$ on the other hand, $\mathcal{S}_r$ shifts a sequence to the right and "pads" with a zero, while $\mathcal{S}_l$ shifts to the left and discards[2] the first element u_0. The latter two are not inverses of each other, but we do have $\mathcal{S}_l\mathcal{S}_r = I$, i.e. $\mathcal{S}_r$ is a right inverse of $\mathcal{S}_l$. Now consider the following observations.

- Note that if $\mathcal{S}_l\,(u_0, u_1, \ldots) = (u_1, u_2, \ldots) = (0, 0, \ldots)$, then $u_k = 0$ for $k \geq 1$. Thus $\mathsf{Nu}(\mathcal{S}_l)$ is the one-dimensional subspace $\mathsf{span}\{(1, 0, \ldots)\}$.
- On the other hand $\mathsf{Nu}(\mathcal{S}_r) = 0$. Note that $\mathcal{S}_r$ is actually an isometry ($\|\mathcal{S}_r v\| = \|v\|$), so clearly its null space is trivial. The same statements hold for $\mathcal{S}$ and $\mathcal{S}^{\text{-}1}$ on $\ell^2(\mathbb{Z})$, i.e. they are both invertible and are isometries.
- Even though $\mathcal{S}_l$ has a non-trivial null space, it has a right inverse since $\mathcal{S}_l\mathcal{S}_r = I$, but not an actual inverse (again because its null space is not trivial). The fact that it has a right inverse means that it is onto $\mathsf{Im}(\mathcal{S}_l) = \mathsf{V} = \ell^2(\mathbb{N})$.
- $\mathsf{Im}(\mathcal{S}_r)$ is clearly missing the subspace $\mathsf{span}\{(1, 0, \ldots)\}$, which is of dimension one. Thus $\mathsf{Im}(\mathcal{S}_r) \neq \ell^2(\mathbb{N})$, and is in fact of co-dimension one in $\ell^2(\mathbb{N})$.

The above examples do not exhibit the case where $\mathsf{Im}(\mathsf{M})$ is not all of V, but rather dense in V. We will see this case once we consider $\lambda I - \mathcal{S}_r$. We are now ready to state the formal definitions.

Definition 5.10. *Let $A : \mathsf{V} \to \mathsf{V}$ be a bounded operator on a Banach space V. The spectrum can be divided into the following disjoint components*

1. *If $\mathsf{Nu}(\lambda I - A) \neq 0$, then λ is called an* eigenvalue *of A, and elements of $\mathsf{Nu}(\lambda I - A)$ are its associated* eigenvectors. *The set $\mathsf{eigs}(A)$ of eigenvalues of A is a subset of the spectrum $\boldsymbol{\lambda}(A)$. The set $\mathsf{eigs}(A)$ is also called the* point spectrum.
2. *If $\mathsf{Nu}(\lambda I - A) = 0$ and $\mathsf{Im}(\lambda I - A) \neq \mathsf{V}$, but is dense in V, then λ is said to belong to the* continuous spectrum *$\boldsymbol{\lambda}_c(A)$.*

[2] The operators $\mathcal{S}_r$ and $\mathcal{S}_l$ are sometimes called "ladder operators", or "creation" and "annihilation" operators respectively, since the right shift creates a new empty slot, while the left shift "annihilates" the left-most element. This terminology reflects the Physicists' flare for dramatic language and overbearing metaphors. A more precise interpretation is the given in Exercise 5.7.

3. *If* $\mathsf{Nu}(\lambda I - A) = 0$ *and* $\overline{\mathsf{Im}(\lambda I - A)} \neq \mathsf{V}$, *then* λ *is said to belong to the* residual spectrum $\boldsymbol{\lambda}_r(A)$.

Thus the spectrum is the union of the three disjoint sets $\boldsymbol{\lambda}(A) = \mathsf{eigs}(A) \cup \boldsymbol{\lambda}_c(A) \cup \boldsymbol{\lambda}_r(A)$.

The reader should be aware that the above classificationa are not the only possible ones. There are other categories, or subdivisions[3], of the spectrum that may be more relevant depending on the application. For example, another possible decomposition of the spectrum is that of the *discrete spectrum* (which roughly are isolated eigenvalues), and the remainder. For our purposes, the most important portions are the eigenvalues $\mathsf{eigs}(A)$ (the point spectrum), and the continuous spectrum. We begin with an example of calculating the points spectrum.

Example 5.11. Consider the left-shift operator $\mathcal{S}_l : \ell^2(\mathbb{N}) \to \ell^2(\mathbb{N})$. A vector u is in the null space of $\lambda I - \mathcal{S}_l$ iff

$$(\lambda I - \mathcal{S}_l)\,(u_0, u_1, \ldots) \;=\; (\lambda u_0 - u_1, \lambda u_1 - u_2, \ldots) \;=\; (0, 0, \ldots)\,.$$

Thus the components of u satisfy the recursion

$$u_{t+1} \;=\; \lambda\, u_t, \qquad t \geq 0.$$

The solution of this recursion is the sequence $u_t = u_0 \lambda^t$, which is in ℓ^2 iff $|\lambda| < 1$. Thus the open unit disk of the complex plane is the set of eigenvalues (the point spectrum)

$$\mathsf{eigs}(\mathcal{S}_l) \;=\; \{\lambda \in \mathbb{C};\ |\lambda| < 1\}\,.$$

A natural question now is what about the case when $|\lambda| = 1$. The vector $u_t = \lambda^t u_0$ is no longer in ℓ^2 (it does not decay), but it looks like it "almost" is an eigenvector. This notion of "almost eignvector/value" is actually how the continuous spectrum can be characterized. The precise statement is as follows.

Lemma 5.12. *A point* $\lambda \in \mathbb{C}$ *is in the continuous spectrum of an operator* $A : \mathsf{V} \to \mathsf{V}$ *iff it is not an eigenvalue, and the minimum modulus* $\underline{\sigma}(\lambda I - \mathcal{A}) = 0$. *The latter condition is equivalent to the existence of a sequence* $\left\{v^{(k)}\right\}$ *of vectors that satisfy either of the following two conditions*

$$\left\|v^{(k)}\right\| = c < \infty, \quad \textit{and} \quad \left\|(\lambda I - A)\; v^{(k)}\right\| \overset{k\to\infty}{\longrightarrow} 0 \tag{5.8}$$

$$\left\|v^{(k)}\right\| \overset{k\to\infty}{\longrightarrow} \infty, \quad \textit{and} \quad \left\|(\lambda I - A)\; v^{(k)}\right\| \leq c < \infty. \tag{5.9}$$

The first condition can be understood intuitively as follows. Although the sequence $\left\{v^{(k)}\right\}$ is not made up of eigenvectors, and may not even be convergent in V, it "wants to" limit to an eigenvector since $(\lambda I - A)\; v^{(k)}$ is limiting to zero (recall that $(\lambda I - A)v = 0$ is the condition for an eigenvector). The difficulty is that usually the sequence is "converging" to something that is not in V. Before we prove this lemma, we work through an example of how it can be applied to the left shift operator of Example 5.11,

Example 5.13. Consider the "almost eigenvectors" of Example 5.11, namely the functions λ^t for $|\lambda| = 1$, and their truncations

$$u_t^{(k)} \;=\; \begin{cases} \lambda^t, & t \leq k, \\ 0, & t > k. \end{cases} \tag{5.10}$$

[3] To mention just a few, there is the compression spectrum, the peripheral spectrum, and various species of the essential spectrum.

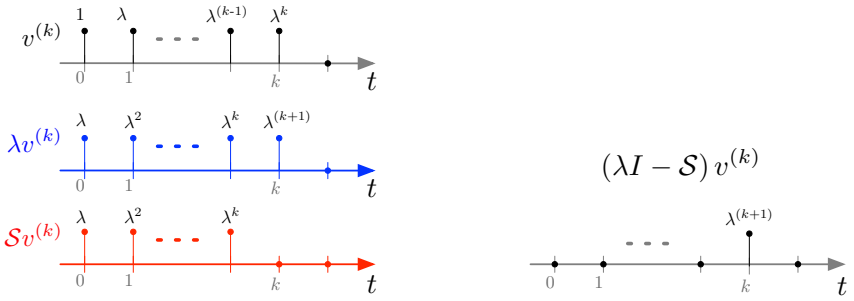


Figure 5.1: For $|\lambda| = 1$, the functions $\{\lambda^t\}$ are candidates as eigenvectors of the left-shift operator $\mathcal{S}_l$ on $\ell^2(\mathbb{N})$. However, the functions are not square summable, and thus not actual eigenvectors. Their truncations (5.10) over $[0, k]$ shown above as $v^{(k)}$ are indeed in $\ell^2(\mathbb{N})$, and satisfy the criterion (5.9) of Lemma 5.12 for the continuous spectrum. Shown here is $\lambda v^{(k)}$ (in blue), $\mathcal{S}_l v^{(k)}$ (in red), and their difference $(\lambda I - \mathcal{S}_l)\, v^{(k)}$. The norm of $v^{(k)}$ is $\sqrt{k}$, and thus grows unboundedly with k, while the norm of $(\lambda I - \mathcal{S}_l)\, v^{(k)}$ remains at 1 for all truncations k. This means that any $|\lambda| = 1$ is in the continuous spectrum.

Note that since $|\lambda| = 1$, the $\ell^2(\mathbb{N})$ norms of these function are $\left\|u^{(k)}\right\| = \sqrt{k}$. Now consider the action of $\lambda I - \mathcal{S}_l$ on $u^{(k)}$ (note that this λ is the same as that in (5.10))

$$\left((\lambda I - S_l)\, u^{(k)}\right)_t \;=\; \lambda\, u_t^{(k)} \;-\; u_{t+1}^{(k)} \;=\; \begin{cases} \lambda\lambda^t \;-\; \lambda^{t+1} \;=\; 0 & t \le k-1, \\ \lambda\lambda^k \;-\; 0 & t = k, \\ 0 & t > k. \end{cases}$$

The result is a function that is zero everywhere except at $t = k$, and the norm of this function is 1. See Figure 5.1.

We have thus found a sequence of functions $u^{(k)}$ of growing norms such that the action of $\lambda I - \mathcal{S}_l$ on them has bounded norms

$$\left\|u^{k)}\right\| \;=\; \sqrt{k} \;\overset{k\to\infty}{\longrightarrow}\; \infty, \qquad\qquad \left\|(\lambda I - S_l)\, u^{(k)}\right\| \;=\; 1.$$

This fulfills criterion (5.9) of Lemma 5.12, and thus every $|\lambda| = 1$ is in the continuous spectrum of $\mathcal{S}_l$. We have so far established that the point spectrum of $\mathcal{S}_l$ is the open unit disk, and the continuous spectrum contains the unit circle. Are there other portions of the spectrum that we missed? The answer is no. Recall that the spectrum is a closed set bounded by the operator norm, which in this case is $\|\mathcal{S}_l\| = 1$, thus the entire spectrum must be contained in the closed unit disk. We can finally conclude

$$\begin{array}{ccccc} \boldsymbol{\lambda}(\mathcal{S}_l) & = & \mathsf{eigs}(\mathcal{S}_l) & \cup & \boldsymbol{\lambda}_c(\mathcal{S}_l) \\ \| & & \| & & \| \\ \{|\lambda| \le 1\} & = & \{|\lambda| < 1\} & \cup & \{|\lambda| = 1\} \end{array}$$

We established that $\boldsymbol{\lambda}_c(\mathcal{S}_l) \subseteq \{|\lambda| = 1\}$, but those two sets are equal because the whole spectrum must be contained in the closed unit disk. Note that we have also shown that the residual spectrum of $\mathcal{S}_l$ must be empty.

Proof of Lemma 5.12. The two conditions (5.8) and (5.9) are equivalent. If (5.8) is satisfied then the sequence $w^{(k)} := v^{(k)} / \left\|(\lambda I - A)\, v^{(k)}\right\|$ satisfies (5.9). Conversely if (5.9) is satisfied, then the sequence $w^{(k)} := v^{(k)} / \left\|v^{(k)}\right\|$ satisfies (5.8).

The proof of Lemma 5.12 relies on the following observation. For $\lambda \in \boldsymbol{\lambda}_c(A)$, $\lambda I - A$ is one-to-one (since it has trivial kernel), and since its image is dense in V, then its inverse $(\lambda I - A)^{-1}$ is a densely defined operator. It is necessarily unbounded, for otherwise it can

be extended to a bounded operator on all of V (and then λ would be in the resolvent set). Since it is unbounded, there exists a sequence of vectors $\{w^{(k)}\}$ with

$$\|w^{(k)}\| = 1, \qquad \lim_{k\to\infty} \left\|(\lambda I - A)^{-1} w^{(k)}\right\| \;=\; \infty.$$

Now define $v^{(k)} := (\lambda I - A)^{-1} w^{(k)}$, which then implies $w^{(k)} = (\lambda I - A) v^{(k)}$. The sequence $\{v^{(k)}\}$ then satisfies condition (5.9). □

We can now generalize the truncation construction of the previous example to devise a method for calculating the continuous spectrum without having to examine truncations in every specific case. The important ingredients of this construction were (a) defining a sequence of truncations, (b) finding an eigenfunctions $\mathcal{A}v = \lambda v$, where v is not in the Banach space, but any truncation of it is, and (c) the operator $\mathcal{A}$ is such that the difference between truncating then acting with the operator compared to acting first and truncating second can be bounded. This can be formalized for the function spaces L^p as follows.

Definition 5.14. *Consider $\Omega \subseteq \mathbb{R}^n$ (or $\mathbb{Z}^n$), and the nested sequence of subsets*

$$\Omega_k \;:=\; \Big\{x \in \Omega;\; \|x\| \le k\Big\}.$$

Note that $\bigcup_{k=1}^{\infty} \Omega_k = \Omega$. The sequence of projections

$$\big(\Pi_k v\big)(x) \;:=\; \begin{cases} v(x) & x \in \Omega_k \\ 0 & x \notin \Omega_k \end{cases}$$

is called an increasing sequence of truncations.

Note that the difference $v - \Pi_k v$ is the "tail" of the function v, and in $L^p(\Omega)$ for $p \in [1,\infty)$, the norm of the tail converges to zero

$$\|v - \Pi_k v\|_{L^p(\Omega)} \;\overset{k\to\infty}{\longrightarrow}\; 0.$$

We also have that the truncation of any $L^\infty(\Omega)$ function is in $L^p(\Omega)$

$$\Pi_k\big(L^\infty(\Omega)\big) \;\subseteq\; L^p(\Omega) \qquad \text{for any } p \in [1,\infty].$$

Lemma 5.15. *Consider $L^p(\Omega)$ (where $\Omega \subseteq \mathbb{R}^n$ or $\mathbb{Z}^n$, and $p \in [1,\infty)$), together with a nested sequence of truncations $\{\Pi_k\}$ as in Definition 5.14. Consider also an operator $\mathcal{A} : L^p(\Omega) \to L^p(\Omega)$, which is also defined[4] on $L^\infty(\Omega)$, such that $\big\|\big(\Pi_k A - A\Pi_k\big)v\big\|_p$ are uniformly (in k) bounded as follows*

$$\forall k, \quad \big\|\big(\Pi_k A - A\Pi_k\big)v\big\|_p \;\le\; c\,\|v\|_\infty \;<\; \infty. \tag{5.11}$$

If there exists a vector $v \in L^\infty(\Omega)$ with $v \notin L^p(\Omega)$ such that $\mathcal{A}v = \lambda v$, then λ is in the continuous spectrum of $\mathcal{A}$.

This lemma formalizes the intuition that when we find bounded functions v with $\mathcal{A}v = \lambda v$, the corresponding λ should be part of the spectrum. The lemma states that such λ's are indeed in the continuous spectrum.

[4] More precisely, there is another operator $\mathcal{A}_\infty : L^\infty(\Omega) \to L^\infty(\Omega)$ such that $\mathcal{A} = \mathcal{A}_\infty$ on the intersection $L^p(\Omega) \cap L^\infty(\Omega)$.

Proof. Given such a function v, the truncations $v^{(k)} := \Pi_k v$ are candidates for the sequence that satisfies criterion (5.9) of Lemma 5.12. Indeed, the fact that $v \in L^\infty(\Omega)$ but not in $L^p(\Omega)$ implies that truncations of v have unbounded L^p norms

$$\|\Pi_k v\|_p \overset{k\to\infty}{\longrightarrow} \infty.$$

The fact that $\mathcal{A}v = \lambda v$ and the bound (5.11) together imply

$$\begin{aligned}\|(\lambda I - \mathcal{A})\ \Pi_k v\|_p &= \|\lambda\ \Pi_k v\ -\ \mathcal{A}\Pi_k v\|_p = \|\Pi_k\ \lambda v\ -\ \Pi_k\mathcal{A}v + \Pi_k\mathcal{A}v - \mathcal{A}\Pi_k v\|_p \\ &\le \big\|\Pi_k\big(\lambda v\ -\ \mathcal{A}v\big)\big\|_p + \big\|\big(\Pi_k\mathcal{A} - \mathcal{A}\Pi_k\big)v\big\|_p \ \le\ 0\ +\ c\ \|v\|_\infty.\end{aligned}$$

Note that since v is an eigenfunction $Av = \lambda v$, it can always be chosen such that $\|v\|_\infty = 1$. Therefore the truncations $\Pi_k v$ satisfy criterion (5.9) of Lemma 5.12. □

We now apply Lemma 5.15 to the bilateral shift operator, which will turn out to have only a continuous spectrum.

Example 5.16. Consider the bilateral shift operator on $\ell^2(\mathbb{Z})$. Since it is norm preserving, its induced norm is $\|\mathcal{S}\| = 1$, and therefore by Lemma 5.6 its spectrum must be inside the unit disk of the complex plane. Furthermore, $\mathcal{S}^{-1}$ is also an isometry and therefore its spectrum must be inside the unit disk. However, since $\boldsymbol{\lambda}(\mathcal{S}) = 1/\boldsymbol{\lambda}\big(\mathcal{S}^{-1}\big)$ we conclude

$$\left.\begin{aligned}\boldsymbol{\lambda}(\mathcal{S}) &\subseteq \Big\{|\lambda| \le 1\Big\} \\ 1/\boldsymbol{\lambda}(\mathcal{S}) = \boldsymbol{\lambda}\big(\mathcal{S}^{-1}\big) &\subseteq \Big\{|\lambda| \le 1\Big\}\end{aligned}\right\} \quad\Rightarrow\quad \boldsymbol{\lambda}(\mathcal{S}) \subseteq \Big\{|\lambda| = 1\Big\}. \tag{5.12}$$

We have thus established that the spectrum of $\mathcal{S}$ is confined to the unit circle. Now investigate the eigenvector equation for $\mathcal{S}$

$$(\lambda I - \mathcal{S})\, v = 0 \quad\Leftrightarrow\quad \lambda v_t - v_{t-1} = 0 \quad\Leftrightarrow\quad v_t = \lambda^{-1} v_{t-1} \quad\Leftrightarrow\quad v_t = \lambda^{-t} v_0.$$

For $|\lambda| = 1$, the function $\{\lambda^{-t}\}$ is not in $\ell^2(\mathbb{Z})$, and we therefore conclude that the point spectrum is empty (there are no eigenvectors).

For $|\lambda| = 1$, the function $\{\lambda^{-t}\}$ is in $\ell^\infty(\mathbb{Z})$, and therefore a candidate for application of Lemma 5.15. It remains to check the condition (5.11) for the shift operator $\mathcal{S}$. First note that the truncation Π_k can be expressed as point-wise multiplication by $\mathbf{1}_{[-k,k]}$, the indicator function of the set $[-k, k]$

$$\big(\Pi_k v\big)(t) = \mathbf{1}_{[-k,k]}(t)\ v(t).$$

Now This can be used to express $\Pi_k\mathcal{S} - \mathcal{S}\Pi_k$ as follows

$$\begin{aligned}\big(\,(\Pi_k\mathcal{S} - \mathcal{S}\Pi_k)v\big)(t) &= \big(\Pi_k\mathcal{S}v\big)(t) - \big(\mathcal{S}\Pi_k v\big)(t) = \mathbf{1}_{[-k,k]}(t)v(t\text{-}1) - \mathbf{1}_{[-k,k]}(t\text{-}1)v(t\text{-}1) \\ &= \Big(\mathbf{1}_{[-k,k]}(t) - \mathbf{1}_{[-k,k]}(t\text{-}1)\Big)\ v(t\text{-}1) = \Big(\delta(t+k) - \delta(t\text{-}k)\Big)\ v(t\text{-}1),\end{aligned}$$

where $\delta(.)$ is the Kronecker delta. Each of the two functions on the right hand side are non-zero only at a single point, and thus have ℓ^p norms bounded by the ℓ^∞ norm of v. This gives the bound

$$\|(\Pi_k\mathcal{S} - \mathcal{S}\Pi_k)v\|_p \ \le\ 2\ \|v\|_\infty,$$

which fulfills the requirements of Lemma 5.15.

The previous argument shows that the unit circle is contained in the continuous spectrum. Since the entire spectrum must be contained in the unit circle by (5.12), the two sets are equal, and we conclude that on $\ell^2(\mathbb{Z})$

$$\boldsymbol{\lambda}_c(\mathcal{S}) = \{\lambda \in \mathbb{C};\ |\lambda| = 1\},$$

and the remaining portions of the spectrum are empty, i.e. $\mathcal{S}$ does not have eigenvalues or a residual spectrum.

5.2.3 Adjoint Relations and the Residual Spectrum

As already seen, the points spectrum is usually the easiest to compute, and the continuous spectrum has to be deduced from further considerations. The third portion of the spectrum, namely the residual spectrum, would be even more difficult to compute directly since there isn't even a characterization like that of the approximate eigenvalues. However, the residual spectrum of an operator is easily related to the eigenvalues of its adjoint. These relations are summarized in the following statement.

Lemma 5.17. *Let $\mathcal{A}$ be an operator on a Banach space V, and denote its adjoint by $\mathcal{A}^*$. The full spectrum, and its point and residual subsets for both $\mathcal{A}$ and $\mathcal{A}^*$ are related by*

$$\begin{aligned} \boldsymbol{\lambda}(\mathcal{A}^*) &=^* \boldsymbol{\lambda}(\mathcal{A})\,, \\ \boldsymbol{\lambda}_r(\mathcal{A}) &\subseteq^* \mathsf{eigs}(\mathcal{A}^*)\,, \\ \mathsf{eigs}(\mathcal{A}) &\subseteq^* \boldsymbol{\lambda}_r(\mathcal{A}^*) \cup \mathsf{eigs}(\mathcal{A}^*)\,, \end{aligned}$$

where the notation $=^*$ and $\subseteq^*$ between sets means equality and subset after complex conjugation respectively.

Proof. The first relation is a consequence of the fact that the inverse of the adjoint (when it exists) is the adjoint of the inverse, and therefore

$$\Big((\lambda I - \mathcal{A})^{-1}\Big)^* \;=\; \big((\lambda I - \mathcal{A})^*\big)^{-1} \;=\; (\lambda^* I - \mathcal{A}^*)^{-1}\,.$$

This means that $\lambda I - \mathcal{A}$ is boundedly invertible iff $\lambda^* - \mathcal{A}^*$ is boundedly invertible.

The remaining two relations are a consequence of the fundamental theorem of linear algebra which states that for any bounded operator B

$$\mathsf{Im}(\mathsf{B})^\perp = \mathsf{Nu}(\mathsf{B}^*) \qquad \Leftrightarrow \qquad \mathsf{Im}(\lambda I - \mathcal{A})^\perp = \mathsf{Nu}(\lambda^* I - \mathcal{A}^*)\,.$$

Also recall that a subspace $\mathcal{S} \subseteq \mathsf{V}$ is dense iff $\mathcal{S}^\perp = 0$. Now recall the definition of the residual spectrum

$$\lambda \in \boldsymbol{\lambda}_r(A) \;\Rightarrow\; \overline{\mathsf{Im}(\lambda I - \mathcal{A})} \neq \mathsf{V} \;\Leftrightarrow\; \mathsf{Nu}(\lambda^* I - \mathcal{A}^*) \neq 0 \;\Leftrightarrow\; \lambda^* \in \mathsf{eigs}(\mathcal{A}^*)\,.$$

This gives the containment $\boldsymbol{\lambda}_r(\mathcal{A}) \subseteq^* \mathsf{eigs}(\mathcal{A}^*)$. Conversely we can attempt to reverse the implications above

$$\lambda \in \mathsf{eigs}(\mathcal{A}) \;\Leftrightarrow\; \mathsf{Nu}(\lambda I - \mathcal{A}) \neq 0 \;\Leftrightarrow\; \overline{\mathsf{Im}(\lambda^* I - \mathcal{A}^*)} \neq \mathsf{V} \;\Rightarrow\; \begin{cases} \lambda^* \in \boldsymbol{\lambda}_r(\mathcal{A}^*) \\ \text{or } \lambda^* \in \mathsf{eigs}(\mathcal{A}^*) \end{cases}$$

The reason for that last statement is that by definition, $\lambda^* \in \boldsymbol{\lambda}_r(\mathcal{A}^*)$ iff $\overline{\mathsf{Im}(\lambda^* I - \mathcal{A}^*)} \neq \mathsf{V}$ *and* $\mathsf{Nu}(\lambda^* I - \mathcal{A}^*) = 0$. If the latter statement is not true, then $\lambda^* \in \mathsf{eigs}(\mathcal{A}^*)$. □

Example 5.18. Consider the right shift operator $\mathcal{S}_r$ on unilateral sequences $\ell^2(\mathbb{N})$. The null space of $\lambda I - \mathcal{S}_r$

$$(\lambda I - \mathcal{S}_r)\,(u_0, u_1, \ldots) \;=\; (\lambda u_0, \lambda u_1 - u_0, \lambda u_2 - u_1, \ldots) \;=\; (0, 0, \ldots)\,.$$

If $\lambda = 0$, the 2nd component states that $u_0 = 0$, the 3rd component states that $u_1 = 0$, and so on. Thus the null space is trivial. If $\lambda \neq 0$, then the 1st component above states that $u_0 = 0$, which then implies from the 2nd component that $u_1 = 0$, and so on. Thus again, the null space is trivial. We therefore conclude that there is no $\lambda \in \mathbb{C}$ such that $\mathsf{Nu}(\lambda I - \mathcal{S}_r) \neq 0$. in other words, $\mathcal{S}_r$ has no eigenvalues and its point spectrum is empty.

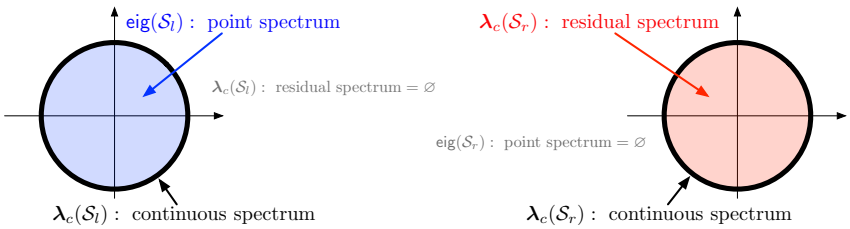


Figure 5.2: The decomposition of the spectra of the left and right shift operators $\mathcal{S}_l$ and $\mathcal{S}_r$ respectively on unilateral sequences in $\ell^2(\mathbb{N})$. Because they are adjoints $\mathcal{S}_l^* = \mathcal{S}_r$, their spectra are equal, and are equal to the closed unit disk. However, the decomposition of their spectra into their individual components of point, continuous, and residual spectra is different in each case. These decompositions are however related by (5.13), which is partially derived from the adjoint relations of Lemma 5.17.

It is easy to establish that $\mathcal{S}_r = \mathcal{S}_l^*$ on $\ell^2(\mathbb{N})$, i.e. the adjoint of the left-shift operator. We have already calculated parts of the spectrum of $\mathcal{S}_l$ in a previous example. Lemma 5.17 provides relationships between the various portions of the spectra of $\mathcal{S}_l$ and $\mathcal{S}_r = \mathcal{S}_l^*$. They are summarized in the following diagram

$$
\begin{array}{ccccccccc}
 & & & & \{|\lambda|<1\} & \cup & \varnothing & \cup & \{|\lambda|=1\} \\
 & & & & \| & & \| \text{①} & & \| \text{③} \\
\{|\lambda|\le 1\} & \supseteq & \boldsymbol{\lambda}(\mathcal{S}_l) & = & \mathsf{eigs}(\mathcal{S}_l) & \cup & \boldsymbol{\lambda}_r(\mathcal{S}_l) & \cup & \boldsymbol{\lambda}_c(\mathcal{S}_l) \\
 & & \overset{*}{\|} & & \overset{*}{\cup}\| & & \|\overset{*}{\cap} & & \\
\{|\lambda|\le 1\} & \supseteq & \boldsymbol{\lambda}(\mathcal{S}_r) & = & \boldsymbol{\lambda}_r(\mathcal{S}_r) & \cup & \mathsf{eigs}(\mathcal{S}_r) & \cup & \boldsymbol{\lambda}_c(\mathcal{S}_r) \\
 & & & & \| \text{②} & & \| & & \| \text{③} \\
 & & & & \{|\lambda|<1\} & \cup & \varnothing & \cup & \{|\lambda|=1\}
\end{array}
\tag{5.13}
$$

The relations in blue are those of Lemma 5.17, which are valid in general for any operator (e.g. $\boldsymbol{\lambda}(A) =^* \boldsymbol{\lambda}(A^*)$). The sets in red are the explicit calculations we did earlier in this example and in Example 5.11. The remaining relations are due to the following observations.

① $\mathcal{S}_r = \mathcal{S}_l^*$, and therefore $\boldsymbol{\lambda}_r(\mathcal{S}_l) \subseteq^* \mathsf{eigs}(\mathcal{S}_r)$. The latter is empty, then so is $\boldsymbol{\lambda}_r(\mathcal{S}_l)$.

② Again, $\mathcal{S}_r = \mathcal{S}_l^*$, and therefore $\mathsf{eigs}(\mathcal{S}_l) \subseteq \boldsymbol{\lambda}_r(\mathcal{S}_r) \cup \mathsf{eigs}(\mathcal{S}_r)$. However, we already know that $\mathsf{eigs}(\mathcal{S}_r)$ is empty, so we must have $\boldsymbol{\lambda}_r(\mathcal{S}_r) = \mathsf{eigs}(\mathcal{S}_l)$, which has been calculated to be the set $\{|\lambda| < 1\}$.

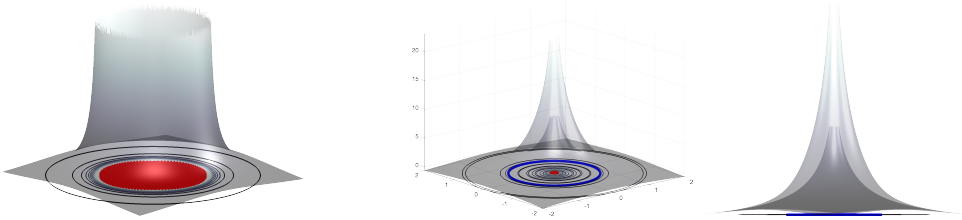
③ Now we know that $\mathsf{eigs}(.) \cup \boldsymbol{\lambda}_r(.)$ is the open unit disk for both operators. The entire spectrum is a closed set that must be contained inside the closed unit disk, and the difference between the closure and the open disk is simply the unit circle. This remainder must be the continuous spectrum $\boldsymbol{\lambda}_c(.)$

These relations are depicted in Figure 5.2.

5.3 The Resolvent and the Pseudospectrum

Example 5.19. Consider the right-shift operator $\mathcal{S}_r$ on $\ell^1(\mathbb{N})$. We want to compute the norm of its resolvent $\| (\lambda I - \mathcal{S}_r)^{-1} \|$ as a function of λ. The easiest way to do this is from the matrix representations

$$
\mathcal{S}_r = \begin{bmatrix} 0 & & \\ 1 & \ddots & \\ & \ddots & \end{bmatrix}, \qquad \lambda I\text{-}\mathcal{S}_r = \begin{bmatrix} \lambda & & \\ \text{-}1 & \ddots & \\ & \ddots & \end{bmatrix}, \qquad (\lambda I\text{-}\mathcal{S}_r)^{\text{-}1} = \begin{bmatrix} \lambda^{\text{-}1} & & \\ \lambda^{\text{-}2} & \ddots & \\ & \ddots & \end{bmatrix}.
$$



(a) The norm of the resolvent of the infinite shift operator. The resolvent has infinite norm on its spectrum (the unit disk, shown in red), while outside the unit disk, its value is the minimum distance to the unit disk.

(b) The logarithm of the norm of the resolvent of the truncated shift operator $\mathcal{S}_n$ for $n = 4$ and $n = 10$. In each case, the spectrum is the point $0 \in \mathbb{C}$ (red dot), but the resolvent grows exponentially (with n) inside the unit circle (shown in blue). Outside the unit circle the resolvent converges (as $n \to \infty$) to that of the infinite shift operator $\mathcal{S}$.

Figure 5.3: Surface and contour plots of the norm of the resolvents of the infinite shift operator $\mathcal{S}$, as well as its truncations $\mathcal{S}_n$, each of which is a Jordan block of size n. The spectrum of $\mathcal{S}$ is the entire unit disk, while the spectrum of $\mathcal{S}_n$ for any n is just the point $0 \in \mathbb{C}$. Thus, while the spectrum is "fragile" in the sense that $\boldsymbol{\lambda}(\mathcal{S}_n) \not\longrightarrow \boldsymbol{\lambda}(\mathcal{S})$, the resolvent is "robust" in the sense that for any $\lambda \in \mathbb{C}$, we have $\| (\lambda I - \mathcal{S}_n)^{-1} \| \overset{n\to\infty}{\longrightarrow} \| (\lambda I - \mathcal{S})^{-1} \|$.

The last expression can be verified directly by multiplying $(\lambda I\text{-}\mathcal{S}_r)$ with $(\lambda I\text{-}\mathcal{S}_r)^{-1}$ and showing that the product is the identity matrix. Note that due to the lower triangular structure, each entry in the product involves a finite sum, and thus issues of convergence do not arise. Alternatively a detailed calculation is presented in Exercise 5.6

Recall that the ℓ^1-induced norm is supremum of absolute sums of columns. For this operator, which is a lower triangular Toeplitz matrix, each column has the same sum. We therefore compute

$$\left\|(\lambda I - \mathcal{S})^{-1}\right\|_{1-i} = \sum_{k=0}^{\infty} |\lambda|^{-k-1} = |\lambda|^{-1} \sum_{k=0}^{\infty} |\lambda|^{-k} = \begin{cases} |\lambda|^{-1} \frac{1}{1-|\lambda|^{-1}} = \frac{1}{|\lambda|-1} & |\lambda| > 1, \\ \infty & |\lambda| \leq 1. \end{cases}$$

Therefore the resolvent's norm is infinite on the closed unit disk, which is to be expected since that set is the spectrum of $\mathcal{S}$. The norm is finite outside the unit disk and equal to $1/(|\lambda| - 1)$, i.e. the reciprocal of the distance between λ and the unit disk. This is plotted in Figure 5.3a.

It is insightful to repeat these calculations for the truncated shift operator $\mathcal{S}_n$, and then compare to those for $\mathcal{S}$.

Example 5.20. The truncated shift operator and its resolvent are given by

$$\mathcal{S}_n = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix}, \qquad (\lambda I - \mathcal{S}_n)^{-1} = \begin{bmatrix} \lambda^{-1} & \lambda^{-2} & \cdots & \lambda^{-n} \\ & \ddots & \ddots & \vdots \\ & & \ddots & \lambda^{-2} \\ & & & \lambda^{-1} \end{bmatrix}.$$

The $\|.\|_1$-induced norm is the maximum column sum, which in this case is the sum of the last column

$$\left\|(\lambda I - \mathcal{S}_n)^{-1}\right\|_{1-i} = \sum_{k=0}^{n-1} |\lambda|^{-k-1} = |\lambda|^{-1} \sum_{k=0}^{n-1} |\lambda|^{-k} = \begin{cases} |\lambda|^{-1} \frac{1-|\lambda|^{-n}}{1-|\lambda|^{-1}} = \frac{1-|\lambda|^{-n}}{|\lambda|-1}, & |\lambda| > 1, \\ n, & |\lambda| = 1, \\ \frac{1-|\lambda|^{-n}}{|\lambda|-1} = |\lambda|^{-n} \frac{|\lambda|^n - 1}{|\lambda|-1}, & |\lambda| < 1. \end{cases}$$

This resolvent norm is finite for all $\lambda \neq 0$. Around $\lambda = 0$, it behaves like $1/\lambda^n$, i.e. a pole of n'th order. That is consistent with $\lambda = 0$ being an eigenvalue of $\mathcal{S}_n$ with multiplicity n.

We analyze the asymptotic behavior (as $n \to \infty$) in each of the three regions

$$\left\|(\lambda I - \mathcal{S}_n)^{-1}\right\|_{1-i} \sim \begin{cases} \frac{1}{|\lambda|-1}, & |\lambda| > 1, \\ n, & |\lambda| = 1, \\ |\lambda|^{-n} \frac{|\lambda|^n - 1}{|\lambda|-1}, & |\lambda| < 1. \end{cases}$$

- For $|\lambda| = 1$, the norm of the resolvent grows linearly with n and is unbounded.
- For $|\lambda| > 1$, $|\lambda|^{-n} \to 0$, and therefore for large n, the norm of $\mathsf{R}_{\mathcal{S}_n}$ behaves just like the norm of the infinite case $\mathsf{R}_{\mathcal{S}}$.
- For $|\lambda| < 1$ the norm is

$$\left\|(\lambda I - \mathcal{S}_n)^{-1}\right\|_{1-i} \;=\; \left|\lambda^{-1}\right|^n \frac{1 - |\lambda|^n}{1 - |\lambda|}, \qquad |\lambda| < 1.$$

  Thus for each $|\lambda| < 1$, the norm grows exponentially in n. L'Hopital's rule shows that the limit as $|\lambda| \to 1$ is n, thus agreeing with the case $|\lambda = 1$ above.

Plots of the resolvent's norm for finite n are shown in Figure 5.3b.

Plots of the resolvents' norms for both $\mathcal{S}$ and $\mathcal{S}_n$ are shown in Figure 5.3 for comparison. These plots and the above calculations lead to the following observation, whose importance cannot be overemphasized. Intuitively, we want to think of $\mathcal{S}$ somehow as the limit of $\mathcal{S}_n$ (as $n \to \infty$). However, we are faced with the fact that the spectrum of $\mathcal{S}_n$ is $\lambda = 0$ for each n, while the spectrum of $\mathcal{S}$ is the entire unit disk. The spectra clearly don't limit, i.e.

$$\lim_{n\to\infty} \boldsymbol{\lambda}(\mathcal{S}_n) \;\neq\; \boldsymbol{\lambda}(\mathcal{S})\,.$$

This is often given as an example of a "discontinuity" at infinity (i.e. as $n \to \infty$), and that $\mathcal{S}_n$ cannot be considered as an approximation to $\mathcal{S}$ even for arbitrarily large n. The calculations above however, show that the resolvents' norms do limit correctly, i.e. for each $\lambda \in \mathbb{C}$

$$\lim_{n\to\infty} \left\|(\lambda I - \mathcal{S}_n)^{-1}\right\| \;=\; \left\|(\lambda I - \mathcal{S})^{-1}\right\|.$$

Here we make an important, but rather philosophical observation. One would always want to develop a theory where conclusions for large n truncations approximate the conclusions for n infinite. In other words, we want "continuity at infinity". The example above shows that the spectrum does not have this "continuity". However, the difficulty should not be viewed as a fundamental difference between infinite versus finite dimensions, but rather that the spectrum is a "fragile" object, i.e. if we change n from infinity to a large number, the spectrum abruptly changes. The resolvent norms do not appear to have this discontinuity at infinity, and therefore the resolvent can be regarded as a *robust* object, unlike the *fragile* spectrum.

The fragility of the spectrum is most dramatically exhibited by some non-normal operators. This fragility has had far reaching implications historically in many fields such as fluid turbulence, condensed matter physics, numerical analysis, and control theory. In the remainder of this section, we introduce the *pseudospectrum*, which captures the level sets of the resolvent's norm. The pseudospectrum provides one particular framework to understand fragility/robustness of a spectrum. In particular, robustness analysis in control theory generalizes the notion of the pseudospectrum using more detailed descriptions of operator perturbations than an additive, unstructured perturbation.

The Pseudospectrum

Definition 5.21. *Given an (possibly unbouded) operator $\mathcal{A} : \mathsf{V} \to \mathsf{V}$ on a Banach space V, its* ϵ-pseudospectrum *is a super-level set of its resolvent's norm*

$$\boldsymbol{\lambda_\epsilon}(\mathcal{A}) \;:=\; \left\{\lambda \in \mathbb{C};\;\; \left\|(\lambda I - \mathcal{A})^{-1}\right\| \geq \frac{1}{\epsilon}\right\} \;=\; \left\{\lambda \in \mathbb{C};\;\; \inf_{\|v\|=1} \|(\lambda I - \mathcal{A})\, v\| \leq \epsilon\right\}.$$

By definition, the actual spectrum is a subset of the ϵ-pseudospectrum for any $\epsilon > 0$. The second characterization follows from the first by the definition of the induced norm.

For matrices with the 2-induced norm, recall that a number λ is an eigenvalue iff the minimum singular value of $\lambda I - \mathcal{A}$ is zero. For this case, the second characterization above states that λ is in the ϵ-pseudospectrum iff the minimum singular value of $\lambda I - \mathcal{A}$ is less than ϵ. Thus if we think of the spectrum as the set of λ's in $\mathbb{C}$ such that $\lambda I - \mathcal{A}$ fails to be invertible, then the ϵ-pseudospectrum is the set such that $\lambda I - \mathcal{A}$ is within a "distance" ϵ of failing to be invertible. This is stated precisely in the following "robustness criterion".

Lemma 5.22. *A complex number is in the* ϵ-psuedospectrum *of an operator $\mathcal{A}$ iff it is in the spectrum of a "nearby" operator $\mathcal{A} + \Delta$, with $\|\Delta\| \leq \epsilon$*

$$\boldsymbol{\lambda_\epsilon}(\mathcal{A}) \;=\; \left\{\lambda \in \mathbb{C};\;\; \lambda \in \boldsymbol{\lambda}(\mathcal{A} + \Delta)\,,\;\; \|\Delta\| \leq \epsilon\right\}$$

For normal operators, the pseudospectrum does not have surprising behavior. To see this, let A be a normal matrix. Its eigenvectors $\{v_k\}_{k=1}^n$ form an orthonormal basis of $\mathbb{R}^n$. If we write A in terms of its dyadic decomposition, we can get a nice formula for the resolvent's norm as follows

$$A \;=\; \sum_{k=1}^{n} \lambda_k\, v_k v_k^* \qquad \Rightarrow \qquad \left\|(\lambda I - A)^{-1}\right\| = \left\|\sum_{k=1}^{n} \frac{1}{\lambda - \lambda_k}\; v_k v_k^*\right\| = \max_{1\leq k\leq n} \frac{1}{|\lambda - \lambda_k|},$$

where the last inequality follows from $\{v_k\}$ being a mutually orthonormal set. This quantity has a geometric interpretation

$$\left\|(\lambda I - A)^{-1}\right\| \;=\; \frac{1}{\min_{1\leq k\leq n} |\lambda - \lambda_k|},$$

which is the reciprocal of the distance from λ to the closest eigenvalue. Therefore the ϵ-pseudospectrum

$$\begin{aligned}\boldsymbol{\lambda_\epsilon}(\mathcal{A}) &:= \left\{\lambda \in \mathbb{C};\;\; \left\|(\lambda I - \mathcal{A})^{-1}\right\| \geq \frac{1}{\epsilon}\right\} = \left\{\lambda \in \mathbb{C};\;\; \min_{1\leq k\leq n} |\lambda - \lambda_k| \leq \epsilon\right\} \\ &= \left\{\lambda \in \mathbb{C};\;\; |\lambda - \lambda_k| \leq \epsilon,\; \text{for any } k = 1, .., n\right\}. \end{aligned} \tag{5.14}$$

This means that the ϵ-pseudospectrum is the union of disks of radius ϵ around each eigenvalue. This is illustrated in Figure 5.4 and Example 5.23 below.

When a matrix or operator is not normal, the expression (5.14) is no longer valid, and the pseudospectrum can be quite unpredictable from the spectrum itself. This can have many interpretations. In particular if the ϵ-pseudospectrum for very small ϵ is very different from the actual spectrum, it implies that the matrix or operator's eigenvalues are very sensitive to small perturbations in the matrix entries. In the shift operator example of the previous section, we argued that this sensitivity is due to the presence of large Jordan blocks, i.e. eigenvalues with high algebraic multiplicity and low geometric multiplicity. Another cause of sensitivity is when a matrix or operator has nearly aligned eigenvectors, even though there are no eigenvalue multiplicities. This is given in the next example.

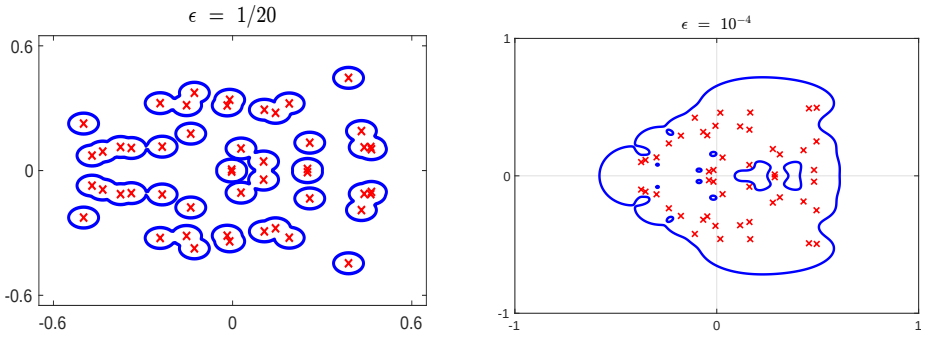


Figure 5.4: ϵ-pseudospectrum boundaries for the two matrices in Example 5.23. (*Left*) A normal matrix has pseudospectrum which is simply the union of disks around each eigenvalue of radius ϵ. (*Right*) For a non-normal matrix, the ϵ pseudospectrum is not easily predicable from the location of the eigenvalues.

Example 5.23. Figure 5.4 shows two examples of 50×50 matrices, their eigenvalues and examples of their pseudospectra. The first example in Figure 5.4 (*left*) is a normal matrix constructed with random eigenvalues and vectors such that the eigenvectors are mutually orthogonal. The fact that the pseudospectrum is a union of disks centered at the eigenvalues as predicted by (5.14) is clearly seen in this plot.

The second example in Figure 5.4 (*right*) is matrix constructed with random eigenvalues, and random eigenvectors but chosen to be nearly aligned. The behavior of the pseudospectrum is very different from the normal example. The pseudospectrum deviates significantly from the eigenvalues even for very small ϵ. Furthermore, the "shape" of the pseudospectrum is not very predictable from the location of the eigenvalues. Note in particular how it bulges far away from some eigenvalues, while it has "voids" very close to other eigenvalues.

Appendix

5.A Analyticity of the Resolvent

Many important calculations involving resolvents utilize the following (very useful) *resolvent formulas*

$$\begin{aligned} \mathsf{R}_A(\lambda_1) - \mathsf{R}_A(\lambda_2) &= (\lambda_1 I - A)^{-1} - (\lambda_2 I - A)^{-1} \\ &= (\lambda_1 I - A)^{-1} \Big((\lambda_2 I - A) - (\lambda_1 I - A) \Big) (\lambda_2 I - A)^{-1} \\ \boxed{(\lambda_1 I - A)^{-1} - (\lambda_2 I - A)^{-1} = (\lambda_1 I - A)^{-1} (\lambda_2 I - A)^{-1} \quad (\lambda_2 - \lambda_1)\,.} & \qquad (5.15) \\ \mathsf{R}_A(\lambda) - \mathsf{R}_B(\lambda) &= (\lambda I - A)^{-1} - (\lambda I - B)^{-1} \\ &= (\lambda I - A)^{-1} (A - B) (\lambda I - B)^{-1} \\ \boxed{(\lambda I - A)^{-1} - (\lambda I - B)^{-1} = (\lambda I - A)^{-1} (A - B) (\lambda I - B)^{-1}.} & \qquad (5.16) \end{aligned}$$

The first compares the resolvent of one operator at two different points in the complex plane, while the second compares the resolvent of two different operators at the same point in the complex plane. As an illustration of one use of these formulas, we investigate the concept of analyticity of a Banach-space-valued function of a complex variable. It turns out that much of the theory is the same as standard complex analytic functions, except that the complex absolute magnitude $|.|$ is replaced by the Banach space norm $\|.\|$.

Definition 5.24. *Let $\Omega \subseteq \mathbb{C}$ be an open set. A Banach-space-valued function $f : \Omega \to \mathsf{V}$ is called* analytic *(or* holomorphic*) if for each $z \in \Omega$ the limit*

$$\lim_{h \to 0} \frac{1}{h} \Big(f(z+h) - f(z) \Big)$$

exists in V. In other words, if at each $z \in \Omega$ there exists $f'(z) \in \mathsf{V}$ such that

$$\lim_{\epsilon \searrow 0} \sup_{h \in \mathbb{C},\ |h| \leq \epsilon} \left\| \frac{f(z+h) - f(z)}{h} - f'(z) \right\| \ = \ 0. \tag{5.17}$$

Note that analyticity is "complex differentiability", so it is important in the above definition that the limit be the same as $z + h \to z$ from all possible directions in $\mathbb{C}$.

Lemma 5.25. *Given a bounded operator A on a Banach space V, its resolvent $\mathsf{R}_A : \boldsymbol{\rho}(A) \to \mathsf{B}(\mathsf{V})$ is an analytic function over the resolvent set $\boldsymbol{\rho}(A) \subset \mathbb{C}$.*

Proof. We can first guess at what the derivative might be (in analogy with the matrix case), and then verify with the definition (5.17). If A were a matrix, then

$$\mathsf{R}'_A(z) \ := \ \frac{d}{dz} (zI - A)^{-1} \ = \ -(zI - A)^{-2}.$$

Now checking the condition (5.17)

$$\begin{aligned} \frac{\mathsf{R}_A(z+h) - \mathsf{R}_A(z)}{h} - \mathsf{R}'(z) &\overset{1}{=} \frac{\big((z+h)I - A\big)^{-1} (zI - A)^{-1} \ (-h)}{h} \ + \ (zI - A)^{-2} \\ &= \ -\big((z+h)I - A\big)^{-1} (zI - A)^{-1} \ + \ (zI - A)^{-2} \\ &= \ \Big(-\big((z+h)I - A\big)^{-1} \ + \ (zI - A)^{-1} \Big) (zI - A)^{-1} \\ &\overset{2}{=} \ h \ \big((z+h)I - A\big)^{-1} (zI - A)^{-1} (zI - A)^{-1}, \end{aligned}$$

where we have used the resolvent formula (5.15) in $\overset{1}{=}$ and again in $\overset{2}{=}$.

Finally, we can bound the norms by

$$\begin{aligned} \sup_{h \in \mathbb{C},\ |h| \leq \epsilon} \left\| \frac{\mathsf{R}_A(z+h) - \mathsf{R}_A(z)}{h} - \mathsf{R}'(z) \right\| &\leq \sup_{|h| \leq \epsilon} \Big\| h \ \big((z+h)I - A\big)^{-1} (zI - A)^{-2} \Big\| \\ &\leq \ \epsilon \ \Big\| (zI - A)^{-2} \Big\| \sup_{|h| \leq \epsilon} \Big\| \big((z+h)I - A\big)^{-1} \Big\|. \end{aligned}$$

The fact that the last quantity is finite for sufficiently small ϵ follows from the bound

$$\begin{aligned} \Big\| \big((z+h)I - A\big)^{-1} \Big\| = \Big\| \big((zI - A) + hI\big)^{-1} \Big\| &= \Big\| (zI - A)^{-1} \ \big(I + h(zI - A)^{-1}\big)^{-1} \Big\| \\ &\leq \ \big\| (zI - A)^{-1} \big\| \ \frac{1}{1 - |h| \, \|(zI - A)^{-1}\|}, \end{aligned}$$

where the last bound follows from the Neumann series provided $|h| < \big\|(zI - A)^{-1}\big\|$. Taking the limit as $\epsilon \searrow 0$ shows that the resolvent is complex differentiable with derivative $-(zI - A)^{-2}$. □

Exercises

Exercise 5.1

Show that for any upper-triangular (or lower-triangular) matrix, its eigenvalues are precisely the entries on the diagonal of the matrix.

Hint: Use the recursive formula for the determinant of $\lambda I - A$.

Exercise 5.2

Show that the $n \times n$ Jordan block J_n has no eigenvectors other than e_1.

Hint: Since J is upper triangular, any eigenvalue of J is a diagonal entry (Exercise 5.1), i.e. any eigenvalue of J is λ. Given the structure of J, the components $\{x_k\}$ of any eigenvector x must satisfy the recursion $\lambda x(k) + x(k+1) = \lambda x(k)$, $k = 1, \ldots, n-1$. Show that e_1 is the only possible solution to this recursion.

Exercise 5.3

Let A be an $n \times n$ square matrix, and $p(x) := x^m + a_{m\text{-}1}x^{m\text{-}1} + \cdots + a_0$ a polynomial with roots $z_1, \ldots, z_m$. Show that

$$p(A) \;:=\; A^m + a_{m\text{-}1}A^{m\text{-}1} + \cdots + a_0 I \;=\; (A - z_1 I)\cdots(A - z_m I)\,.$$

Exercise 5.4

Show that for any $n \times n$ matrix M

$$\left\|M^{-1}\right\| \;\leq\; c\|M\|^{n-1}\frac{1}{|\det M|},$$

where the constant c is independent of M. The formula $M^{-1} = \mathrm{Adj}(M)/\det M$ will be useful.

To put this result in context, recall that in general there need not be a relationship between $\sigma_{\max}(M)$ and the eigenvalues other than the bound $\rho(M) \leq \sigma_{\max}(M)$, which may be arbitrarily conservative. On the other hand, the result above does give a lower bound on $\sigma_{\min}(M) = 1/\|M^{-1}\| \geq \det M/c\|M\|^{n-1}$. Recalling that $\det M$ is the product of the eigenvalues of M, this can be thought of as a lower bound on $\sigma_{\min}(M)$ in terms of the eigenvalues.

Solution 5.4

First observe that for any $n \times n$ matrix M with entries m_{ij}, we have the following bounds between the norm of M and its entries

$$\max_{i,j}|m_{ij}| \;\leq\; \|M\| \;\leq\; n\,\max_{i,j}|m_{ij}|. \tag{5.18}$$

The first inequality is trivial, and the second one follows from standard matrix norm bounds.

Now starting from

$$\left\|M^{-1}\right\| \;=\; \|\mathrm{Adj}(M)/\det M\| \;=\; \|\mathrm{Adj}(M)\|\,\frac{1}{\det M},$$

we see that a bound for $\|\mathrm{Adj}(M)\|$ is required. Recall that the entries of $\mathrm{Adj}(M)$ are the $(n-1)\times(n-1)$ minors of M, and each of these entries is a polynomial, with each term a

homogenous, degree $n-1$ monomial of the entries of M. This fact gives the first inequality in

$$\left|\big(\text{Adj}\,(M)\,\big)_{ij}\right| \;\leq\; n\left(\max_{i,j}|m_{ij}|\right)^{n-1} \;\leq\; n\,\|M\|^{n-1},$$

while the second inequality follows from the lower bound in (5.18). We can now bound $\|\text{Adj}(M)\|$ using the upper bound in (5.18) and conclude

$$\left\|M^{-1}\right\| \;=\; \|\text{Adj}(M)\|\,\frac{1}{|\det M|} \;\leq\; n^2\|M\|^{n-1}\,\frac{1}{|\det M|}.$$

Exercise 5.5

Prove the first part of Lemma 5.8 which states that for any bounded operator $\mathcal{A}$

$$\boldsymbol{\lambda}(A^n) \;=\; \big(\boldsymbol{\lambda}(A)\,\big)^n.$$

The following fact will be useful. Let α be any complex number and consider the polynomial factorization

$$(x^n-\alpha^n) \;=\; (x-\alpha\;\rho_0)\;\cdots\;(x-\alpha\;\rho_{n\text{-}1}), \tag{5.19}$$

where $\{\rho_l\}$ are the n, n'th roots of unity ($\rho_l := e^{j\frac{2\pi}{n}l}$).

Solution 5.5

Substitute the operator A in the polynomial (5.19)

$$(A^n-\lambda I) \;=\; \left(A-\lambda^{1/n}\rho_0\;I\right)\;\cdots\;\left(A-\lambda^{1/n}\;\rho_{n\text{-}1}\;I\right),$$

where $\lambda^{1/n}$ is any n'th root of λ. Now this product is not boundedly invertible iff $\lambda^{1/n}\rho_l \in \boldsymbol{\lambda}(A)$ for some $l=0,\ldots,n-1$. Note that for any set C in the complex plane

$$\exists l\in\{0,1,\ldots,n-1\},\ \text{s.t. }\lambda^{1/n}\rho_l \;\in\; C \qquad\Leftrightarrow\qquad \lambda \;\in\; C^n.$$

We therefore conclude that $\lambda\in\boldsymbol{\lambda}(A^n)$ iff one of its n'th roots is in $\boldsymbol{\lambda}(A)$, which implies that $\boldsymbol{\lambda}(A^n)=(\boldsymbol{\lambda}(A))^n$.

Exercise 5.6

The resolvent equations for the right-shift operator are $v=\big(\lambda I-\mathcal{S}_r\big)^{-1}w \;\Leftrightarrow\; \big(\lambda I-\mathcal{S}_r\big)v=w$. Show that these equations are equivalent to the following recursion, and the solution shown on the second line.

$$\begin{array}{r} \lambda v_0 = w_0 \\ \lambda v_t - v_{t-1} = w_t \\ t\geq 1 \end{array} \quad\Leftrightarrow\quad \begin{array}{rcl} v_0 &=& \frac{1}{\lambda}w_0 \\ v_{t+1} &=& \frac{1}{\lambda}v_t+\frac{1}{\lambda}w_{t+1} \\ t &\geq& 0 \end{array} \quad\Leftrightarrow\quad \begin{array}{rcl} x_0 &=& 0 \\ x_{t+1} &=& \frac{1}{\lambda}x_t+\frac{1}{\lambda^2}w_t \\ v_t &=& x_t+\frac{1}{\lambda}w_t \end{array}$$

$$v_t \;=\; \sum_{l=0}^{t-1}\left(\tfrac{1}{\lambda}\right)^{t-l}\left(\tfrac{1}{\lambda^2}w_l\right)+\tfrac{1}{\lambda}w_t$$

Exercise 5.7

The right and left shift operators on $\ell^2(\mathbb{N})$ are sometimes referred to as "ladder operators", or "creation" and "annihilation" operators respectively. To see the justification for this terminology, consider two operators A and S on a Hilbert space H whose commutator is

$$[A, S] \;:=\; AS - SA \;=\; \alpha S,$$

where $\alpha \in \mathbb{C}$.

1. Show that if λ is an eigenvalue of A (resp. A^*), then so is $\lambda + \alpha$ (resp. $\lambda - \alpha^*$).

2. Now suppose $\alpha > 0$ is real, $A > 0$ is self-adjoint and is such that its countable eigenvectors $\{v_k\}$ span the Hilbert space H. Arrange the mutually orthonormal eigenvectors $\{v_k\}_{k=0}^{\infty}$ in ascending order of corresponding eigenvalues. Since they form a basis, there is an isometric isomorphism $V : \mathsf{H} \to \ell^2(\mathbb{N})$ which takes any vector in H to the sequence of its coefficients in the basis.

   Show that the right and left shift operators on $\ell^2(\mathbb{N})$ are the representations of S and S^* in that basis, i.e.

$$VSV^{-1} \;=\; \mathcal{S}_r, \qquad\qquad VS^*V^{-1} \;=\; \mathcal{S}_l.$$

Chapter 6

The Kernel Representation of Linear Operators

Linear operators on function spaces are abstractions of matrices. They are often defined abstractly or in terms of their action on functions. For a large class of linear operators there is also an integral representation, called the kernel representation *which often provides considerable insight into the structure of the operator. The kernel representation can be considered to be the continuum limit of a matrix representation. Thus, this representation is the natural generalization of matrix-vector and matrix-matrix multiplication. In the same way that certain matrix structures (e.g. symmetric, diagonal, rank-one, Toeplitz, etc.) provide useful insight, the structure of the operator kernel provides similar insights into operators on function spaces. Several operator norms, such as induced* L^1 *and* L^∞ *norms, as well as the Hilbert-Schmidt norm have simple expressions in terms of the kernel representation.*

6.1 Motivation: Kernels as Continuum Matrices

Recall the basic definition of matrix-vector multiplication. A matrix $A : \mathbb{R}^n \to \mathbb{R}^m$ acting on a vector u to produce a vector v operates as

$$v_i \;=\; \sum_{j=1}^{n} A_{ij} u_j, \quad i = 1, \ldots, m. \qquad \Leftrightarrow \qquad \begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix} \;=\; \begin{bmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & & \vdots \\ A_{m1} & \cdots & A_{mn} \end{bmatrix} \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} \tag{6.1}$$

The summation on the left is expressed graphically on the right as each v_i obtained from multiplying each element of the i'th row of A with the corresponding elements of u and then adding the result up. A matrix is a two dimensional array of numbers which defines a linear operator on $\mathbb{R}^n$ by the operation described above. Similarly we will see that a function of two variables $A(x, \xi)$ also defines a linear operator on a function space.

Recall that real vectors in $\mathbb{R}^n$ can be viewed as real-valued functions (Figure 1.1) on the set $\{1, \ldots, n\}$, i.e. as elements of the function space $\mathbb{R}^{\{1,\ldots,n\}}$. To generalize the operation (6.1) to a function space $\mathbb{R}^{\mathsf{I}}$ over any index set I, we need the ability to "sum" over this index. Assume for the moment that $\mathsf{I} \subseteq \mathbb{R}$ so we can integrate over it. The counterpart of (6.1) would be

$$v(x) \;=\; \int_{\mathsf{I}} A(x, \xi)\; u(\xi)\; d\xi, \tag{6.2}$$

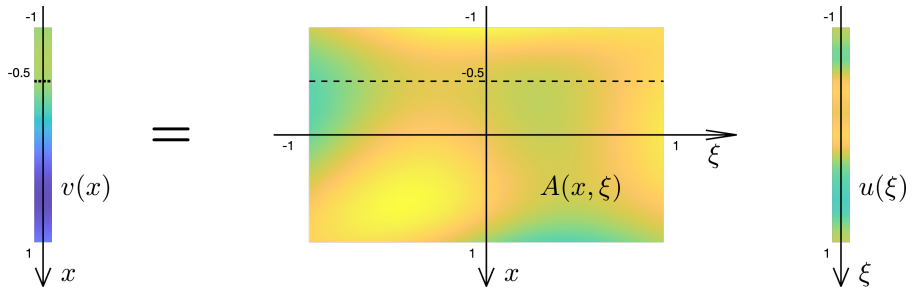


Figure 6.1: Graphical depiction of the integral operator (6.2) as an abstraction of matrix-vector multiplication. The two-variable kernel function $A(x,\xi)$ is the counterpart of matrix entries, with the coordinate x as "row index" and ξ as "column index". The operation $v(x) = \int_{-1}^{1} A(x,\xi)\, u(\xi)\, d\xi$ gives the value of $v(x)$ at any x (an instance above is depicted by the dashed lines at $x = -0.5$) as the multiply-then-integrate of the x'th row of $A(.,.)$ with the all the values of $u(\xi)$ viewed as a "column vector". The case shown above is for an integral operator on functions defined over the interval $[-1, 1]$. The unusual choice of the vertical axis positive direction as downwards is made to be in analogy with matrix rows being indexed from top to bottom.

where the integration variable ξ plays the same role as the column index j over which the summation in (6.1) is performed. The two variable function $A(.,.)$ is called the *kernel function* of the operator A, and the formula (6.2) is called the *kernel representation*[1] of A.

The operation in (6.2) can be regarded as a linear operator $A : u \mapsto v$ on some subspace of the function space $\mathbb{R}^{\mathsf{I}}$ provided the type of functions $u(.)$ and $A(.,.)$ ensure convergence of the integral. Linearity of the integral operator (6.2) follows immediately from the linearity of integration

$$\begin{aligned}\big(A\big(\alpha_1 u_1 + \alpha_2 u_2\big)\big)(x) &= \int_{\mathsf{I}} A(x,\xi)\, \big(\alpha_1 u_1(\xi) + \alpha_2 u_2(\xi)\big)\, d\xi \\ &= \alpha_1 \int_{\mathsf{I}} A(x,\xi)\, u_1(\xi)\, d\xi \;+\; \alpha_2 \int_{\mathsf{I}} A(x,\xi)\, u_2(\xi)\, d\xi \\ &= \alpha_1 \big(A\big(u_1\big)\big)(x) \;+\; \alpha_2 \big(A\big(u_2\big)\big)(x).\end{aligned}$$

The operation (6.2) is depicted in Figure 6.1. The one-variable functions $u(\xi)$ and $v(x)$ are analogous to "column vectors", while the two-variable kernel function $A(x,\xi)$ is analogous to a matrix, i.e. a two-dimensional array. For each x, the value of $v(x)$ is given by the operation of multiply-then-integrate of the corresponding "row" of $A(x,\xi)$ with the function $u(\xi)$ in an analogous manner to matrix-vector multiplication.

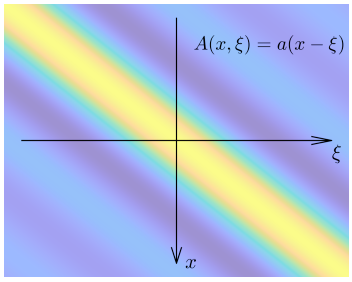
Just like certain matrix structures encode certain symmetries or properties of the linear operations they represent, the structure of a kernel encodes important properties of the operators they represent. Figure 6.2 illustrates the four examples we examine below.

- **Toeplitz Operators**

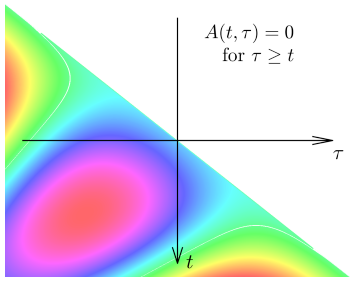
  A kernel defined from a single variable function a by $A(x,\xi) := a(x-\xi)$ is called *Toeplitz*. This is depicted in Figure 6.2a, where the kernel function appears as "constant along diagonals". Such a kernel defines a *convolution* operator as follows

$$v(x) \;=\; \int A(x,\xi)\, u(\xi)\, d\xi \;=\; \int a(x-\xi)\, u(\xi)\, d\xi.$$

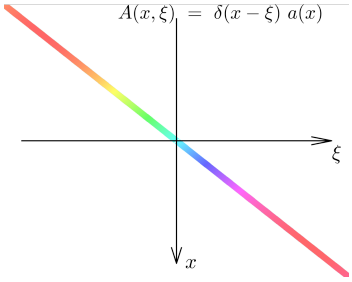
[1] The reader should be careful not to confuse this with the null space of the operator, which is sometimes referred to as the kernel of the operator. The two concepts are unrelated. To avoid confusion, we will always use the phrase null space instead of kernel space, and the word "kernel" will only be used to refer to the above kernel representation.



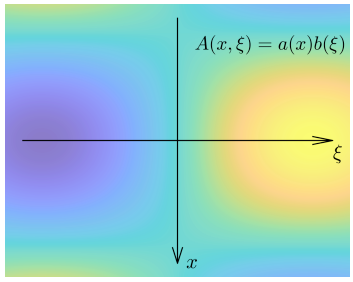
(a) A *Toeplitz kernel* $A(.,.)$ is generated from a function $a(.)$ of a single variable such that $A(x,\xi) = a(x-\xi)$. This kernel is "constant along diagonals" in a similar manner to a Toeplitz matrix. Integrating against this kernel amounts to convolution.



(b) A *lower-triangular kernel* is such that $A(t,\tau) = 0$ for $\tau \geq t$. If it operates on time signals, a lower-triangular kernel is a *causal system*, i.e. past values of the output do not depend on future values of the input.



(c) A *diagonal kernel* has a "modulated impulse sheet" $A(x,\xi) = \delta(x-\xi)a(x)$ along the diagonal. This kernel arises in the integral representation of a multiplication operator so that the expression (6.2) amounts to $v(x) = a(x)u(x)$.



(d) A *rank-1 kernel* (also called a *tensor product kernel*) is formed from the product of two functions $A(x,\xi) = a(x)b(\xi)$ (and thus is separable). It is akin to the outer product of two vectors, and is written as a tensor product $A = a \otimes b$ of the two functions.

Figure 6.2: Examples and visualizations of integral kernels (6.2) of various operators. A Toeplitz kernel represents a convolution operator. A lower triangular kernel arises when a time-varying, causal system acts on temporal signals. Diagonal kernels represent multiplication operators and generalize diagonal matrices. In particular, the identity operator $v(x) = u(x)$ has an impulse sheet $A(x,\xi) = \delta(x-\xi)$ of unit strength along the diagonal. Rank-1 kernels are formed from the product of two functions and are analogous to rank-1 matrices formed from the outer product of two vectors.

Thus Toeplitz operators are convolution operators. They have special symmetry properties where on certain domains they can be characterized by shift invariance.

- **Lower-Triangular Operators**

  Such operators arise when modeling time-varying causal systems. The lower-triangular property is illustrated in Figure 6.2b, where the kernel is restricted to be zero in the "upper triangular part" of the (τ, t) plane

$$A(t,\tau) \ = \ 0, \quad \text{for } \tau \geq t. \tag{6.3}$$

  If u and y are temporal signals over the entire real line, then the lower-triangular property of the kernel implies that the integral (6.2) has the following limits

$$y(t) \ = \ \int_{-\infty}^{\infty} A(t,\tau)\ u(\tau)\ d\tau \ = \ \int_{-\infty}^{t} A(t,\tau)\ u(\tau)\ d\tau. \tag{6.4}$$

  When t and τ are interpreted as time, then (6.4) is the description of a general time-varying system mapping u to y that has the causality property, i.e. for any given time

T, current and past values of the output $\{y(t);\ t \leq T\}$ do not depend on future values of the input $\{u(\tau);\ \tau > T\}$.

An alternative way of imposing the lower-triangular condition (6.3) is by using the unit *Heaviside step function* $\mathfrak{h}$ as follows. Given any kernel function $A(x,\xi)$, observe that the product $A(x,\xi)\mathfrak{h}(x-\xi)$ becomes a lower triangular kernel

$$\int_{\underline{\xi}}^{\overline{\xi}} \big(A(x,\xi)\mathfrak{h}(x-\xi)\big)\ u(\xi)\ d\xi \ = \ \int_{\underline{\xi}}^{x} A(x,\xi)\ u(\xi)\ d\xi,$$

since $\mathfrak{h}(x-\xi) = 0$ when $\xi > x$. The above holds regardless of the original upper and lower integration limits $\overline{\xi}$ and $\underline{\xi}$ respectively.

Operators with a lower triangular kernel are sometimes called *Volterra operators* if the kernel function is bounded. For Volterra operators acting on function spaces $\mathsf{L}^p(\Omega)$ where Ω is compact, these operators have the important property that the Neumann series converges even if the operator norm is greater than one (Exercise 6.2). This property gives an iterative scheme for solving certain integral equations involving Volterra operators.

- **Diagonal Operators**

  Operators with diagonal kernels are really multiplication operators. First we should observe that to implement the identity operator $v(x) = u(x)$ with the operation (6.2), we must allow the kernel function $A(.,.)$ to contain distributions (Dirac delta functions). Observe that

  $$v(x) \ = \ \int \delta(x-\xi)\ u(\xi)\ d\xi \ = \ u(x).$$

  Any distribution that is supported on the diagonal $x = \xi$ of the (ξ, x) plane represents a multiplication operator. Such kernels are depicted in Figure 6.2c, where

  $$v(x) \ = \ \int A(x,\xi)\ u(\xi)\ d\xi \ = \ \int \Big(\delta(x-\xi)\ a(x)\Big)\ u(\xi)\ d\xi \ = \ a(x)\ u(x).$$

  This kernel is visualized as a diagonal impulse sheet that is modulated by the function $a(.)$. This is a generalization of a diagonal matrix. Multiplication operators are discussed in detail in Chapter **??**.

- **Rank-1 (Tensor Product) Operators**

  Given any function space, and two functions a and b in that space, we can form an operator from the "separable product" of those functions as follows

  $$A(x,\xi) \ := \ a(x)\ b(\xi). \tag{6.5}$$

  This is akin to the outer product of two vectors. If a and b were vectors in $\mathbb{R}^n$, then the rank-1 matrix $A = ab^*$ has as its ij'th entry

  $$A_{ij} \ = \ a_i\ b_j.$$

  This expression should be compared with (6.5) where the role of the row index i is played by the variable x, while the column index j is analogous to ξ.

  If the function space is also an inner product space (with the usual inner product), then the action of this operator $v = Au$ can be expressed as follows

  $$v(x) = \int A(x,\xi)\ u(\xi)\ d\xi \quad \Leftrightarrow \quad v(x) = \int a(x)\ b(\xi)\ u(\xi)\ d\xi = a(x) \int b(\xi)\ u(\xi)\ d\xi$$
  $$\Leftrightarrow \quad v \ = \ a\ \langle b\,,\, u\rangle\,.$$

Thus the image space of A is the one-dimensional subspace spanned by the vector a. This justifies calling this a rank-1 kernel. We note here that in a general Hilbert space, we can define the *tensor product* of two vectors a and b as a linear operator on that same Hilbert space as follows

$$(a \otimes b)\, u \;:=\; a\ \langle b\,,\, u\rangle\,.$$

This is an abstract definition in terms of the inner product. For function spaces, this definition amounts to the expression (6.5) for the kernel of the operator.

Figure 6.2d shows an example of a rank-1 kernel. It is often not easy to visually discern from the shape of the kernel whether it is a rank-1 kernel or not.

Kernel representations of linear dynamical systems will be examined in some detail later in Chapter **??** where causality, time invariance, and time periodicity properties of such systems will be studied.

6.2 Basic Properties: Compositions and Adjoints

Let $\Omega_m \subseteq \mathbb{R}^m$ and $\Omega_n \subseteq \mathbb{R}^n$ be two domains over which function spaces are defined. For simplicity, assume the functions to be real valued, i.e. the function spaces are $\mathbb{R}^{\Omega_m}$ and $\mathbb{R}^{\Omega_n}$. Consider a linear operator $A : \mathbb{R}^{\Omega_m} \longrightarrow \mathbb{R}^{\Omega_n}$ defined in terms of its kernel representation

$$v \;=\; A\ u \qquad \Leftrightarrow \qquad v(x) \;=\; \int_{\Omega_m} A(x,\xi)\ u(\xi)\ d\xi, \quad x \in \Omega_n. \tag{6.6}$$

The type of functions $A(.,.)$ that produce well-defined operators will be discussed later as it depends on which class of functions u and v belong to. For now, we examine structural properties, and assume the function classes have been chosen so that all integrals are convergent and all manipulations are allowed.

Addition and Composition of Operators

Given two operators A and B in terms of their respective kernel functions, it is easy to see that the operator sum $C := A + B$ has as its kernel function $C(x,\xi) = A(x,\xi) + B(x,\xi)$

$$\begin{aligned} v = \ & \big(A+B\big)u \;=\; Au \;+\; Bu \\ v(x) = \ & \int A(x,\xi)\ u(\xi)\ d\xi \;+\; \int B(x,\xi)\ u(\xi)\ d\xi \;=\; \int \big(A(x,\xi)+B(x,\xi)\big)\ u(\xi)\ d\xi. \end{aligned}$$

Therefore, under addition, kernel functions behave just like matrix-matrix addition which is element-by-element.

Another intuitive property of kernel representations is that they can be composed in a manner similar to matrix-matrix multiplication. Let $A : u \mapsto v$ and $B : v \mapsto w$ be two operators with kernel representations

$$v(x) \;=\; \int A(x,\xi)\ u(\xi)\ d\xi, \qquad\qquad w(x) \;=\; \int B(x,\xi)\ v(\xi)\ d\xi.$$

Define a third operator as the composition $C := BA : u \mapsto w$, and calculate its kernel representation from those of A and B as follows

$$\begin{aligned} w(x) = \int B(x,\xi)\ v(\xi)\ d\xi &= \int B(x,\xi) \left(\int A(\xi,r)\ u(r)\ dr \right)\ d\xi \\ &= \int \left(\int B(x,\xi)\ A(\xi,r)\ d\xi \right)\ u(r)\ dr = \int C(x,r)\ u(r)\ dr. \end{aligned}$$

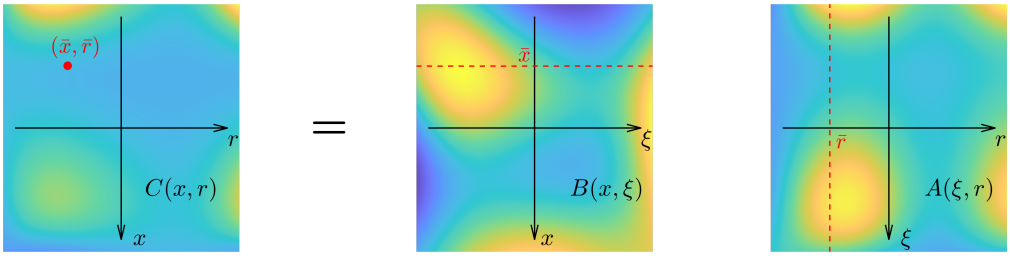


Figure 6.3: A graphical depiction of the composition of two operators $C = BA$ as the integral operation (6.7) on their respective kernels. This operation is akin to matrix-matrix multiplication as shown above. The value of the kernel C at a point $(\bar{x},\bar{r})$ is obtained from integrating the "row" $B(\bar{x},.)$ against the "column" $A(.,\bar{r})$.

Thus the kernel of the composition $C = BA$ is obtained from the formula

$$C(x,r) \;=\; \int B(x,\xi)\; A(\xi,r)\; d\xi, \tag{6.7}$$

which looks like matrix-matrix multiplication except for integration instead of summation. Each "row" $B(x,.)$ of the kernel of B is integrated against each "column" $A(.,r)$ of the kernel of A. The composition operation (6.7) is depicted graphically in Figure 6.3. The reader should compare this visually with the usual matrix-matrix multiplication.

Matrix-valued Kernels

The notation we are using applies without modification to the case of operators acting on vector-valued functions. Let $u : \Omega \to \mathbb{R}^n$ be a vector-valued function. The space of all such functions is $(\mathbb{R}^n)^\Omega$. An operator mapping m-vector-valued functions to n-vector-valued functions $A : (\mathbb{R}^m)^\Omega \to (\mathbb{R}^n)^\Omega$ (here we assumed the functions to have the same domain Ω so as not to clutter the notation) has a kernel representation

$$v(x) \;=\; \int_\Omega A(x,\xi)\; u(\xi)\; d\xi, \qquad x \in \Omega,$$

where for each (x,ξ), $A(x,\xi)$ is an $n \times m$ matrix. We call such a function $A(.,.)$ a *matrix-valued* function for the obvious reason. Notice that at each ξ, the n-vector $A(x,\xi)u(\xi)$ is obtained by multiplying the m-vector $u(\xi)$ with the $n \times m$ matrix $A(x,\xi)$. Thus, we can use the same notation for scalar-valued and vector-valued functions without explicitly indicating the vector dimensions. The reader should verify that the addition and composition properties verified in the previous paragraphs are valid for vector-valued functions and matrix-valued kernels provided all the dimensions are compatible.

Adjoints

The adjoint of an operator is the generalization of the concept of the matrix transpose. If the reader is not familiar with this notion, then the following two paragraphs should be read after becoming familiar with adjoints as described in Chapter 4. For readers with some familiarity with the concept, recall that the adjoint of a matrix is its transpose (or complex conjugate transpose in the case of complex matrices). If a linear operator A has kernel $A(x,\xi)$, we will show that the kernel of its adjoint $A^\dagger$ is simply $A^\dagger(x,\xi) = A^*(\xi,x)$. Thus, if the kernel is real-valued, then the kernel of the adjoint is obtained from the original kernel by "flipping" the two arguments x and ξ. This is analogous to transposing a matrix

by flipping the row and column index. If the kernel is matrix-valued, then in addition to flipping the arguments, we also take complex-conjugate transpose at each (x,ξ) as well. This is analogous to transposing block-structured matrices.

To demonstrate the previous statement, we start from the definition of the adjoint of an operator on a function space and work through the kernel representation. Starting from the definition $\stackrel{1}{=}$ below and working outwards

$$\int_{\Omega_n} v^*(x) \left(\int_{\Omega_m} A(x,\xi)\ u(\xi)\ d\xi \right) dx \ = \langle v\,,\, Au \rangle \ \stackrel{1}{=}\ \langle A^\dagger v\,,\, u \rangle = \int_{\Omega_m} \big(A^\dagger v\big)^*(x)\ u(x)\ dx$$

$$\|$$

$$\int_{\Omega_m} \left(\int_{\Omega_n} A^*(x,\xi)\ v(x)\ dx \right)^* u(\xi)\ d\xi$$

In order for the two expressions to be equal for all test functions u, we see that (after relabeling the integration variables) the action of $A^\dagger$ is given by

$$\big(A^\dagger v\big)(x) \ = \ \int A^*(\xi,x)\ v(\xi)\ d\xi \qquad \Rightarrow \qquad A^\dagger(x,\xi) \ = \ A^*(\xi,x). \tag{6.8}$$

Dyadic Decompositions

Recall that the diagonalization of a (diagonalizable) matrix A can be expressed as the "dyadic decomposition" (7.8)

$$Au \ = \ \sum_{i=1}^{n} \lambda_i \,\langle w_i\,,\, u \rangle\ v_i \qquad \Leftrightarrow \qquad A \ = \ \sum_{i=1}^{n} \lambda_i\ v_i w_i^* \tag{6.9}$$

where v and w are the eigenvectors of A and A^* respectively. The expression on the right is the same as that on the left, but expressed in terms of the outer products $v_i w_i^*$, each of which is a rank-1 square matrix. A dyadic decomposition therefore expresses a diagonalizable matrix as a linear combination of rank-1 matrices.

As already seen in the rank-1 kernel example (6.5), the counterpart of rank-1 operators are given by the tensor product of two vectors. Assume for simplicity that we are in the setting of a Hilbert space $\mathcal{H}$ which is also a function space. The tensor product of two elements $a, b \in \mathcal{H}$ is a bounded operator on $\mathcal{H}$ defined by

$$(a \otimes b)\, u \ := \ a\ \langle b\,,\, u \rangle \qquad \Leftrightarrow \qquad \big((a \otimes b)\, u\big)(x) \ := \ \left(\int b(\xi)\ u(\xi)\ d\xi \right) a(x).$$

Thus the kernel function of the rank-1 operator $a \otimes b$ is given by the "outer product" of the two functions

$$\big(a \otimes b\big)(x,\xi) \ = \ a(x)\ b(\xi). \tag{6.10}$$

Recall that an operator A on a Hilbert space with a purely discrete spectrum can be decomposed similarly to (6.9), but with a potentially infinite sum

$$A \ = \ \sum_{k=1}^{\infty} \lambda_k\ \big(v_k \otimes w_k\big), \tag{6.11}$$

where λ_k are the eigenvalues of A, and v_k and w_k are the eigenfunctions of A and A^* respectively. Thus the outer product of two eigenvectors in (6.9) is replaced by the tensor

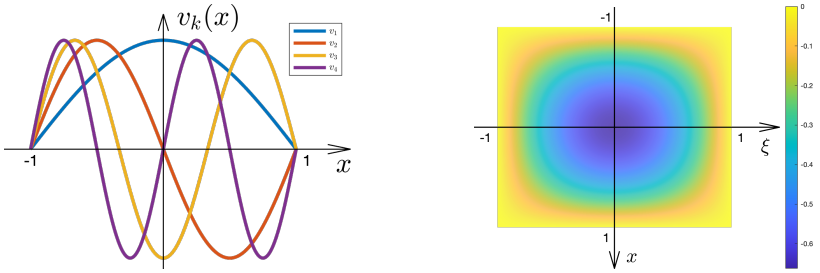


Figure 6.4: (*Left*) The first four eigenfunctions of the second derivative operator ∂_x^2 on $\mathsf{L}^2[-1,1]$ with zero Dirichlet boundary conditions ($\psi(\pm 1) = 0$). (*Right*) The kernel function of the *inverse* of ∂_x^2 (with those boundary conditions) generated from the dyadic expansion (6.13).

product of two eigenfunctions here. From the explicit expression (6.10) for a rank-1 kernel, we can write a summation formula for the kernel of A based on the expansion (6.11)

$$A(x,\xi) \;=\; \sum_{k=1}^{\infty} \lambda_k \; v_k(x) w_k(\xi). \tag{6.12}$$

Thus the kernel of A is a weighted sum of rank-1 kernels made up of the outer products of eigenfunctions of A and A^*.

As an example, consider the differential operator $A = \partial_x^2$ on $\mathsf{L}^2[-1,1]$ with zero Dirichlet boundary conditions. Its eigenfunctions and eigenvalues are known to be

$$v_k(x) = \sqrt{\frac{3}{2}}\frac{1}{\pi k} \sin\left(k\frac{\pi}{2}(x+1)\right), \qquad \lambda_k = -\frac{\pi^2}{4}k^2, \qquad k = 1,2,\ldots$$

It can be shown that this is a self-adjoint operator, and thus the eigenfunctions w of its adjoint are the same as those shown above, and the expression for its kernel would be

$$A(x,\xi) \;=\; \sum_{k=1}^{\infty} \lambda_k \; v_k(x) v_k(\xi).$$

However, this series is not convergent since λ_k is unbounded. This was to be expected since ∂_x^2 is an unbounded operator. It is also a differential operator and its kernel function contains distributions. We therefore would not expect this series expansion to be convergent in a standard sense.

On the other had, the inverse of this operator is indeed a bounded operator with eigenvalues of $1/\lambda_k$, and with the same eigenfunctions. Let $A^{-1}(x,\xi)$ be the kernel function of the inverse of ∂_x^2 with the given boundary conditions. The expansion for this kernel is

$$A^{-1}(x,\xi) = \sum_{k=1}^{\infty} \frac{1}{\lambda_k} \; v_k(x) v_k(\xi) = -\sum_{k=1}^{\infty} \tfrac{4}{\pi^2 k^2} \;\; \tfrac{3}{2\pi^2 k^2} \sin\!\left(\tfrac{\pi k}{2}(x+1)\right) \sin\!\left(\tfrac{\pi k}{2}(\xi+1)\right), \tag{6.13}$$

which is a convergent series. Figure 6.4 illustrates the shape of this kernel.

The above construction can be used to derive the kernel representation of many operators (and functions of those operators) for which the eigenfunctions can be calculated either analytically or numerically.

6.3 Boundedness and Operator Norms

Recall that certain matrix norms are easy to express in terms of the matrix entries. For example, the ℓ^1 and ℓ^∞ induced norms are the "max column sum" and "max row sum"

respectively, while the Frobenius norm is the sum of the squares of all entries. These norms stand in contrast with the ℓ^2 (Euclidean) induced norm. The latter is the maximum singular value and cannot be immediately calculated from the entries. Similarly, on function spaces, the L^1 and L^∞ induced norms can be calculated from the kernel as max column and row integrals respectively. The counterpart of the Frobenius norm is the so-called Hilbert-Schmidt norm, which is simply the squared integral of the kernel in analogy with the sum of the squares of matrix entries.

6.3.1 L^p-induced Norms

Consider again the setting (6.6) of function spaces over subsets of $\mathbb{R}^n$ and $\mathbb{R}^m$

$$v(x) \;=\; \int_{\Omega_m} A(x,\xi)\; u(\xi)\; d\xi, \qquad x \in \Omega_n. \tag{6.14}$$

where the functions u and v are in $\mathsf{L}^p(\Omega_m)$ and $\mathsf{L}^p(\Omega_n)$ respectively for $p \in [1,\infty]$. We want to find conditions on the kernel $A(.,.)$ to ensure that the operator has bounded L^p-induced norm. It is not difficult to show that if the domains Ω_m and Ω_n are compact, and if $A(.,.)$ is bounded, then the operation (6.14) defines a bounded operator on any of the L^p spaces. However, we would like a more detailed condition which gives the actual induced norms. For this, we first consider the two extreme cases of the L^∞-induced and L^1-induced norms. In the calculations below, the key step is the following bound for any two functions f and g

$$\begin{aligned}
&\int \Big|f(x)\Big|\,\Big|g(x)\Big|\,dx \le \int \Big|f(x)\Big| \left(\sup_x \Big|g(x)\Big|\right) dx \;=\; \int \Big|f(x)\Big|\,dx \left(\sup_x \Big|g(x)\Big|\right) \\
&\qquad \Rightarrow \quad \|fg\|_1 \le \|f\|_1 \|g\|_\infty,
\end{aligned}$$

and note that the roles of f and g can be reversed if the respective norms are finite.

In the case of L^∞ norms on u and v, we can calculate the following bound[2]

$$\begin{aligned}
|v(x)| = \;& \left|\int_{\Omega_m} A(x,\xi)\; u(\xi)\; d\xi\right| \;\le\; \int_{\Omega_m} |A(x,\xi)|\,|u(\xi)|\; d\xi \\
&\le \int_{\Omega_m} |A(x,\xi)| \left(\sup_{\xi\in\Omega_m} |u(\xi)|\right) d\xi \;=\; \left(\int_{\Omega_m} |A(x,\xi)|\; d\xi\right) \left(\sup_{\xi\in\Omega_m} |u(\xi)|\right) \\
\Rightarrow\; \|v\|_\infty = \;& \sup_{x\in\Omega_n} |v(x)| \;\le\; \left(\sup_{x\in\Omega_n} \int_{\Omega_m} |A(x,\xi)|\; d\xi\right) \|u\|_\infty.
\end{aligned} \tag{6.15}$$

This bound can be interpreted as the "max-row-integral" of $A(.,.)$ by regarding ξ and x as the "column" and "row" indices respectively. The fact that this bound is tight follows from a standard argument; the function $\bar{u}$ that almost achieves this upper bound is

$$\bar{u}(\xi) \;=\; \operatorname{sign}\left(A(\bar{x},\xi)\right),$$

with $\bar{x}$ chosen where the supremum in (6.15) is almost achieved.

The calculation for the L^1-induced norm proceeds as follows

$$\begin{aligned}
\|v\|_1 = \int_{\Omega_n} |v(x)|\,dx \;&=\; \int_{\Omega_n} \left|\int_{\Omega_m} A(x,\xi)\; u(\xi)\; d\xi\right| dx \\
&\le \int_{\Omega_n}\int_{\Omega_m} |A(x,\xi)|\,|u(\xi)|\; d\xi\; dx \;=\; \int_{\Omega_m} \left(\int_{\Omega_n} |A(x,\xi)|\,dx\right) |u(\xi)|\; d\xi \\
&\le \left(\sup_{\xi\in\Omega_m} \int_{\Omega_n} |A(x,\xi)|\,dx\right) \int_{\Omega_m} |u(\xi)|\; d\xi = \left(\sup_{\xi\in\Omega_m} \int_{\Omega_n} |A(x,\xi)|\,dx\right) \|u\|_1.
\end{aligned} \tag{6.16}$$

[2]The term "bounded" here means "essentially bounded", and "supremum" means "essential supremum".

This bound can be interpreted as "max-column-integral" of $A(.,.)$. The tightness of the bound can be ascertained as follows; the function $\bar{u}$ that almost achieves the bound is an approximation to the dirac delta function $\delta(\xi-\bar{\xi})$ centered at $\bar{\xi}$ where the supremum in (6.16) is almost achieved.

The above calculations show that the L^1- and L^∞-induced norms can be directly computed from the kernel representation of the operator. For other L^p-induced norms, such a direct calculation is not typically possible, but one can bound the L^p-induced norms using the two extreme cases of $p = 1$ and $p = \infty$. For this we need the so-called Riesz-Thorin Convexity Theorem which we state for the special case of $\Omega_m = \Omega_n = \Omega$ for simplicity.

Theorem 6.1 (Riesz-Thorin)**.** *Let Ω be a measure space and let A be a bounded linear operator on $\mathsf{L}^1(\Omega)$ and $\mathsf{L}^\infty(\Omega)$ with induced norms $\|A\|_{1\text{-}i}$ and $\|A\|_{\infty\text{-}i}$ respectively. Then A is bounded on $\mathsf{L}^p(\Omega)$ for all $p \in [1,\infty]$. Furthermore*

$$\|A\|_{p\text{-}i} \;\leq\; \|A\|_{1\text{-}i}^{1/p}\; \|A\|_{\infty\text{-}i}^{1-1/p}.$$

In particular

$$\|A\|_{2\text{-}i} \;\leq\; \sqrt{\|A\|_{1\text{-}i}\; \|A\|_{\infty\text{-}i}}.$$

Combining this theorem with the calculations above, we arrive at the following.

Theorem 6.2. *Consider the kernel representation (6.14) of an operator A defined on the function spaces $\mathsf{L}^p(\Omega)$.*

1. *The L^∞-induced norm of A is the "max-row-integral" of A:*

$$\|A\|_{\infty\text{-}i} \;=\; \sup_{x\in\Omega} \int_\Omega \Big|A(x,\xi)\Big|\; d\xi \;=:\; \|A\|_{\text{mri}}.$$

2. *The L^1-induced norm of A is the "max-column-integral" of A:*

$$\|A\|_{1\text{-}i} \;=\; \sup_{\xi\in\Omega} \int_\Omega \Big|A(x,\xi)\Big|\; dx \;=:\; \|A\|_{\text{mci}}.$$

3. *If the above two quantities are finite, then A is also bounded on L^p for all $p \in [1,\infty]$ with induced norm*

$$\|A\|_{p\text{-}i} \;\leq\; \|A\|_{1\text{-}i}^{1/p}\; \|A\|_{\infty\text{-}i}^{1-1/p}.$$

In particular, the above theorem gives a condition for L^2 boundedness. We refer the reader to Exercise 6.3 for an alternative condition for L^2 boundedness which uses a more direct argument for that particular case.

Finally we point out that all of the above is equally true for the sequence spaces $\ell^p(\Omega)$, where $\Omega \subseteq \mathbb{Z}^n$ is a subset of the integer lattice. All of the calculations above hold for this case by simply replacing dx and $d\xi$ with counting measure on $\Omega \subseteq \mathbb{Z}^n$, and the integrals become sums.

Application to LTI Systems (Toeplitz Kernels)

It is illuminating to see the implications of the above result to the kernel representation of an LTI system. Let $G(.,.)$ be the kernel function of a general linear time varying system. Theorem 6.2 gives conditions for the L^p boundedness (aka L^p-stability) of the system G.

Note that the standard definition of Bounded Input Bounded Output (BIBO) stability is exactly L^∞-stability.

In the special case that G is time invariant, the kernel $G(.,.)$ is Toeplitz. For Toeplitz kernels, the "max-row-integral" and "max-column-integral" are equal, and are in fact equal to the L^1 norm of the impulse response. More precisely, when G is Toeplitz, then

$$G(t,\tau) \;=\; g(t-\tau),$$

and we can easily show that

$$\begin{aligned}
\|G\|_{\infty\text{-}i} &= \sup_{-\infty<t<\infty} \int_{-\infty}^{\infty} \Big|G(t,\tau)\Big| \; d\tau \;=\; \sup_{-\infty<t<\infty} \int_{-\infty}^{\infty} \Big|g(t\text{-}\tau)\Big| \; d\tau \;=\; \int_{-\infty}^{\infty} \Big|g(\tau)\Big| \; d\tau \\
\|G\|_{1\text{-}i} &= \sup_{-\infty<\tau<\infty} \int_{-\infty}^{\infty} \Big|G(t,\tau)\Big| \; dt \;=\; \sup_{-\infty<\tau<\infty} \int_{-\infty}^{\infty} \Big|g(t\text{-}\tau)\Big| \; dt \;=\; \int_{-\infty}^{\infty} \Big|g(t)\Big| \; dt
\end{aligned}$$

The convexity theorem then implies that for LTI systems, the finiteness of the L^1 norm of the impulse response is a sufficient condition for L^p-stability for all $p \in [0,\infty]$. This is the often stated condition for BIBO (i.e. L^∞-) stability. The above calculations shows that BIBO stability is equivalent to L^1-stability, and is a sufficient (but not necessary) condition for L^p-stability for all $p \in [1,\infty]$.

The above calculation also gives a bound on the L^p-induced norm (for any $p \in [1,\infty]$) in terms of the L^1 norm of the impulse response

$$\|G\|_{p\text{-}i} \;\leq\; \|G\|_{1\text{-}i}^{1/p} \; \|G\|_{\infty\text{-}i}^{1-1/p} \;=\; \|g\|_1^{1/p} \; \|g\|_1^{1-1/p} \;=\; \|g\|_1. \tag{6.17}$$

At this point it is important for the reader not to confuse the induced norms of systems, the norms of signals, and the norms of the impulse response representations of systems. For example, $\|G\|_{\infty\text{-}i}$ above is the L^∞-induced norm of the system G. It turns out to be equal to the L^1 norm $\|g\|_1$ of its impulse response when regarded as a signal. Those two are also distinct from norms of signals that the system G operates on.

6.3.2 The Trace and the Hilbert-Schmidt Norm

The trace of a square matrix is easily computed from its entries as the sum of the diagonal entries. It is also the sum of the eigenvalues. Thus, while each individual eigenvalue may not be easily computable from the matrix entries, the sum of the eigenvalues is. Given a linear operator A on a function space $\mathbb{R}^\Omega$ and its kernel representation $A(.,.)$, we define the trace in an analogous manner to matrices by integrating the kernel along its "diagonal"

$$\mathrm{tr}(A) \;:=\; \int_\Omega A(x.x) \; dx. \tag{6.18}$$

Unlike the matrix trace however, this quantity may be infinite for some operators. An operator with finite trace is called a *trace class* operator. When the kernel is matrix valued, we adopt the following natural definition

$$\mathrm{tr}(A) \;:=\; \int_\Omega \mathrm{tr}\Big(A(x.x)\Big) \; dx, \tag{6.19}$$

where the trace of the integrand is the usual matrix trace.

An important property of the matrix trace is that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$. A similar formula holds for operators whose kernels can be integrated as below with convergent integrals

$$\begin{aligned}
\mathrm{tr}(AB) &= \int_\Omega \mathrm{tr}\Big(\big(AB\big)(x,x)\Big) \; dx \;=\; \int_\Omega \mathrm{tr}\bigg(\int_\Omega A(x,r)B(r,x) \; dr\bigg) \; dx, \\
\mathrm{tr}(BA) &= \int_\Omega \mathrm{tr}\Big(\big(BA\big)(x,x)\Big) \; dx \;=\; \int_\Omega \mathrm{tr}\bigg(\int_\Omega B(x,r)A(r,x) \; dr\bigg) \; dx.
\end{aligned}$$

These two quantities are equal as can be seen by interchanging the matrix trace with integration, and applying the matrix result $\text{tr}(A(x,r)B(x,r)) = \text{tr}(B(x,r)A(x,r))$.

To see that the operator trace is also the sum of the eigenvalues, assume for simplicity that an operator A has a discrete spectrum and the expansion (6.12). Applying the integration formula (6.18) to (6.12) we see that

$$\text{tr}(A) \;=\; \int_\Omega \sum_{k=1}^{\infty} \lambda_k \; v_k(x) w_k(x) \; dx \;=\; \sum_{k=1}^{\infty} \lambda_k \int_\Omega \; v_k(x) w_k(x) \; dx \;=\; \sum_{k=1}^{\infty} \lambda_k. \quad (6.20)$$

Note that the product $v_k w_k$ integrates to 1 since $\{v_k\}$ and $\{w_k\}$ are dual bases. Since the trace is the sum of the eigenvalues, we see that for some operators, the series (6.20) may not be summable, and this corresponds to when the kernel function is not integrable (6.19) along the diagonal.

The Hilber-Schmidt Norm

Now recall that the Frobenius norm (squared) of a matrix A is the sum of squares of its entries, which is also the trace of AA^*, which in turn is the sum of squares of the singular values of A

$$\|A\|_{\text{F}}^2 \;:=\; \sum_{ij} |A_{ij}| \;=\; \text{tr}(AA^*) \;=\; \sum_k \lambda_k \left(AA^*\right) \;=\; \sum_k \sigma_k^2 \left(A\right). \quad (6.21)$$

Consider an operator A on a function space $\mathbb{R}^\Omega$ with a kernel representation $A(.,.)$. We can similarly define the generalization of the Frobenius norm as the *Hilbert-Schmidt (HS) norm* of the kernel

$$\|A\|_{\text{HS}}^2 \;:=\; \int_\Omega \int_\Omega \|A(x,\xi)\|_{\text{F}}^2 \; dx \; d\xi \;=\; \int_\Omega \int_\Omega \text{tr}\Big(A(x,\xi)A^*(x,\xi)\Big) \; dx \; d\xi. \quad (6.22)$$

Note that this formula is written for the general case of a matrix-valued kernel, where the trace in the integrand is the matrix trace.

The HS norm is also equal to $\text{tr}\big(AA^\dagger\big)$ as can be seen from applying the composition formula for kernels

$$\begin{aligned}\text{tr}\big(AA^\dagger\big) = \int_\Omega \text{tr}\Big(\left(AA^\dagger\right)(x,x)\Big) \; dx &= \int_\Omega \text{tr}\bigg(\int_\Omega A(x,r)A^\dagger(r,x) \; dr\bigg) \; dx \\ &= \int_\Omega \int_\Omega \text{tr}\Big(A(x,r)A^*(x,r)\Big) \; dr \; dx = \|A\|_{\text{HS}}^2 .\end{aligned}$$

Note that the value of the kernel $\left(AA^\dagger\right)(x,x)$ at each point (x,x) on the diagonal is equal to the integral of the kernel A (squared) over the x'th row. Thus integrating $\left(AA^\dagger\right)(x,x)$ over x integrates the kernel A (squared) over all rows and columns. Alternatively, the value $\left(A^\dagger A\right)(x,x)$ is the integral of the kernel A (squared) over the x'th column, and then integrating that over x yields the HS norm. The reader should recall that $\text{tr}\big(AA^\dagger\big) = \text{tr}\big(A^\dagger A\big)$.

Since the operator trace is also the sum of the eigenvalues for an operator with a discrete spectrum, we can now make a similar conclusion to (6.21) for such operators

$$\|A\|_{\text{HS}}^2 \;:=\; \int_\Omega \int_\Omega \|A(x,\xi)\|_{\text{F}}^2 \; dx \; d\xi \;=\; \text{tr}\big(AA^\dagger\big) \;=\; \sum_k \lambda_k \left(AA^\dagger\right) \;=\; \sum_k \sigma_k^2 \left(A\right). \quad (6.23)$$

Thus we see that an operator has finite Hilbert-Schmidt norm iff its singular values (the square roots of the eigenvalues of $AA^\dagger$) form an ℓ^2 sequence, or equivalently, when the eigenvalue sequence of $AA^\dagger$ is an ℓ^1 sequence.

Finally we recall that the space of $n \times n$ matrices is an inner product space (of dimension n^2) with the inner product

$$\langle A\,,\, B\rangle \;:=\; \mathrm{tr}(A^* B)\,.$$

This inner product induces the Frobenius norm since $\|A\|_F^2 = \langle A\,,\, A\rangle$. Similarly, observe that the expression (6.22) is the square of the $\mathsf{L}^2(\Omega \times \Omega)$ norm[3] of the two-variable function $A(.,.)$. Thus we can identify the set of all HS operators with the function space $\mathsf{L}^2(\Omega \times \Omega)$ which has the inner product

$$\langle A\,,\, B\rangle_{\mathrm{HS}} \;:=\; \int_\Omega \int_\Omega \mathrm{tr}\Big(A^*(x,\xi)\; B(x,\xi)\Big)\; dx\; d\xi \;=\; \mathrm{tr}\big(A^\dagger B\big)\,.$$

This inner product induces the HS norm since $\langle A\,,\, A\rangle_{\mathrm{HS}} = \|A\|_{\mathrm{HS}}^2$. In addition, the Cauchy-Schwartz inequality for $\mathsf{L}^2(\Omega \times \Omega)$ implies the following inequality

$$\langle A\,,\, B\rangle_{\mathrm{HS}} \;=\; \mathrm{tr}\big(A^\dagger B\big) \;\leq\; \|A\|_{\mathrm{HS}}\, \|B\|_{\mathrm{HS}}\,.$$

Thus the composition $A^\dagger B$ of two HS operators $A^\dagger$ and B is a trace class operator, with its trace bounded by the product of the two HS norms. Note that the composition of two HS operators, while is trace class, is not necessarily HS. This is similar to the the product of two L^2 functions being in L^1, but not necessarily in L^2.

The reader should recall that the set of bounded operators on a Hilbert space (or any Banach space) is itself a Banach space with the induced norm. Hilbert-Schmidt operators are special in the sense that they can be endowed with an inner product, and thus be made into a Hilbert space, which has much more structure than a general Banach space. The Hilber-Schmidt norm however is not an induced operator norm, and therefore has limited utility especially when it comes to norm bounds and sensitivity and robustness calculations.

Exercises

Exercise 6.1

Consider the following two differential operators

$$\begin{aligned}(Au)(x) \;&:=\; a_2(x)\; u''(x) \;+\; a_1(x)\; u'(x) \;+\; a_0(x)\; u(x),\\ (Bu)(x) \;&:=\; \big(a_2(x)\; u(x)\big)'' \;+\; \big(a_1(x)\; u(x)\big)' \;+\; a_0(x)\; u(x).\end{aligned}$$

Show that their kernel representations are

$$\begin{aligned}A(x,\xi) \;&=\; a_2(x)\; \delta''(x-\xi) \;+\; a_1(x)\; \delta'(x-\xi) \;+\; a_0(x)\; \delta(x-\xi),\\ B(x,\xi) \;&=\; a_2(\xi)\; \delta''(x-\xi) \;+\; a_1(\xi)\; \delta'(x-\xi) \;+\; a_0(x)\; \delta(x-\xi).\end{aligned}$$

Exercise 6.2

Consider a Volterra (lower triangular) operator on $\mathsf{L}^p([a,b])$ of the form

$$y \;=\; Au \qquad \Leftrightarrow \qquad y(t) \;=\; \int_a^b A(t,\tau)\; u(\tau)\; d\tau, \qquad t,\tau \in [a,b],$$

where $A(.,.,)$ is lower triangular (i.e. $A(t,\tau) = 0$ for $\tau > t$), and uniformly bounded

$$\sup_{t,\tau\in[a,b]} |A(t,\tau)| \;=\; \bar{A} \;<\; \infty.$$

[3] More explicitly, we should write $L_{\mathbb{R}^{n\times n}}(\Omega\times\Omega)$ when the kernel $A(.,.)$ is matrix valued. This is suppressed for simplicity of notation.

1. Show that A is a bounded operator on $\mathsf{L}^1([a,b])$, on $\mathsf{L}^\infty([a,b])$ and on $\mathsf{L}^p([a,b])$ for all $p \in [1,\infty]$.

2. Let $A^k(.,.)$ be the kernel function of the operator A^k (A composed with itself k times). Show by induction (on k) that the following bound holds

$$\left|A^k(t,\tau)\right| \;\leq\; \bar{A}^k \frac{1}{(k-1)!}\;|t-\tau|^{k-1}\,, \qquad t,\tau \in [a,b].$$

3. Show that the L^p-induced norms of A^k (for $p \in [1,\infty]$) are all bounded by

$$\left\|A^k\right\|_{p-i} \;\leq\; \bar{A}\;\frac{\left(|b-a|\;\bar{A}\right)^{k-1}}{(k-1)!}$$

4. By recalling the fact that the sequence $\alpha^k/k!$ converges to zero for any number α, show that the Neumann series $(I-A)^{-1} = \sum_{k=0}^{\infty} A^k$ converges (in L^p-induced norm) for the Volterra operator described above.

5. Find a bound on $\left\|(I-A)^{-1}\right\|$.

   Hint: It will involve the exponential function.

Exercise 6.3

Provide a direct proof that when Ω is compact, the boundedness of the kernel function $A(.,.)$ on Ω is sufficient to ensure the boundedness of the operator A it represents on $\mathsf{L}^2(\Omega)$. This can be done with a similar set of bounds (though not tight) to those in the calculations for (6.15) and (6.16).

Solution 6.3

L^2-induced norm bounds in terms of the kernel function $A(.,.)$ can be derived as follows

$$\begin{aligned}
\|v\|^2 \;&=\; \int_\Omega |v(x)|^2 dx \;=\; \int_\Omega \left|\int_\Omega A(x,\xi)\;u(\xi)\;d\xi\right|^2 dx \\
&\leq\; \int_\Omega\int_\Omega |A(x,\xi)|^2|u(\xi)|^2\;d\xi\;dx \;\leq \int_\Omega \left(\sup_{\xi\in\Omega}|A(x,\xi)|^2\right)\left(\int_\Omega |u(\xi)|^2 d\xi\right) dx \\
&\leq\; |\Omega|\;\sup_{x,\xi\in\Omega}|A(x,\xi)|^2\;\|u\|^2, \qquad (6.24)
\end{aligned}$$

where $|\Omega|$ is the measure of Ω (which is finite by the compactness assumption). Note that the first inequality is a consequence of Jensen's inequality.

Note that in contrast to the L^∞ and L^1 cases, the above bound is not tight. The L^2-induced norm is the supremum of the singular values of A, which are difficult to compute directly from the function $A(.,.)$ in general. This is in complete analogy with the matrix case where the singular values can not be seen immediately from the entries of the matrix.

In the case that Ω is not compact, we can rework the last step before inequality (6.24) as follows

$$\|v\|^2 \;\leq\; \|u\|^2 \int_\Omega \left(\sup_{\xi\in\Omega}|A(x,\xi)|^2\right) dx.$$

Thus the integral on the right can be used as a bound in the non-compact Ω case. This is however not a good bound. To see that, assume the kernel $A(.,.)$ is Toeplitz, then

$$\sup_{\xi\in\Omega} |A(x,\xi)|^2 \;=\; \sup_{\xi\in\Omega} |a(x-\xi)|^2 \;=\; \|a\|_\infty^2 ,$$

and the integral of that constant over the non-compact set Ω will be divergent. Thus this bound is infinite for any Toeplitz operator over a non-compact set Ω, while there are many cases of such Toeplitz operators that have a finite $\mathsf{L}^2(\Omega)$-induced norms.

Chapter 7

Matrix/Operator Partitions

A vector space can be constructed as the direct sum of several vector spaces, each of which can then be viewed as subspace of the overall vector space. The structure of a vector space can be analyzed by decomposing it as the direct sum of several of its subspaces. Similarly, a linear operator can usually be understood in detail by examining how it acts on each of the individual subspaces. Various canonical decompositions of matrices and operators, such as the eigenvalue decomposition, the singular value decomposition and others can be described using the language of decomposition over certain subspaces. Other constructions, such as the Schur complement, can also be easily understood in this setting.

This geometric view is nicely complemented using the algebraic notion of matrix and operator partitions. These are conformable partitions of a matrix into "blocks" of matrices, each corresponding to the action of a matrix on a subspace. Thus, conformable partitioning of matrices and operators is a useful tool for synthesizing algebraic and geometric intuition in linear algebra. In certain instances, rather compact arguments can be provided using this technique.

Introduction: Matrix Partitions Notation

An example of conformably partitioned matrices is the following. Let H and G be matrices with dimensions such that the product HG makes sense. Suppose H is partitioned in 2×2 blocks and G in 2×1 blocks as

$$HG = \left[\begin{array}{c:c} H_{11} & H_{12} \\ \hdashline H_{21} & H_{22} \end{array}\right] \left[\begin{array}{c} G_1 \\ \hdashline G_2 \end{array}\right] = \left[\begin{array}{c} H_{11}G_1 + H_{12}G_2 \\ \hdashline H_{21}G_1 + H_{22}G_2 \end{array}\right].$$

For this to make sense, the relative partitions of H and G have to be so that all the products make sense, i.e. they should be *conformable partitions* (e.g. the number of rows of G_1 is equal to the number of columns of H_{11} and H_{21}). The utility of this is that we can multiply H and G as if they were a 2×2 and 2×1 matrices respectively. Care must be taken with the order of multiplication though since the elements of the block partitioned matrices do not commute (since they are matrices themselves).

Matrix-Vector Products

A matrix-vector product has at least two interpretations which can be arrived at by considering either column or row partitioning of the matrix. Begin with the product $x = Tz$

with T partitioned columnwise to observe

$$\left[\begin{array}{c} x \end{array}\right] = \left[\begin{array}{c:c:c} T_1 & \cdots & T_n \end{array}\right] \left[\begin{array}{c} z_1 \\ \hdashline \vdots \\ \hdashline z_n \end{array}\right] = \left[\begin{array}{c} T_1 \end{array}\right] z_1 + \cdots + \left[\begin{array}{c} T_n \end{array}\right] z_n, \tag{7.1}$$

which can be interpreted as writing the vector x as a linear combination of the *vectors* $T_1, \cdots, T_n$, each being multiplied by the *scalar* coefficients $z_1, \cdots, z_n$. This observation has two uses

- If each z_i ranges over all possible scalars, then the above simply states that all possible resulting vectors x are in span $\{T_1, \cdots, T_n\}$. This is another way of seeing that the image space of T is its column span.
- If T is square and non-singular, i.e. $\{T_1, \cdots, T_n\}$ is a full linearly independent set of vectors, then the product $x = Tz$ can be regarded as expanding the vector x in the basis made up of the columns of T. The corresponding z_i's are then the expansions coefficients, which can be obtained directly from $z = T^{-1}x$.

An example of the change of basis interpretation is for a linear differential equation $\dot{x}(t) = Ax(t)$. With the state transformation $x(t) =: Tz(t)$, the differential equation becomes $\dot{z}(t) = \left(T^{-1}AT\right) z(t)$. The new state variables $z_i(t)$ should be interpreted as the time-varying coefficients of the expansion of $x(t)$ in the basis $\{T_1, \cdots, T_n\}$

$$\left[\begin{array}{c} x(t) \end{array}\right] = \left[\begin{array}{c} T_1 \end{array}\right] z_1(t) + \cdots + \left[\begin{array}{c} T_n \end{array}\right] z_n(t),$$

The second interpretation of matrix-vector products arises from row partitioning of the matrix in a product like $w = Mv$ as follows

$$\left[\begin{array}{c} w_1 \\ \hdashline \vdots \\ \hdashline w_n \end{array}\right] = \left[\begin{array}{c} M_1^* \\ \hdashline \vdots \\ \hdashline M_n^* \end{array}\right] \left[\begin{array}{c} v \end{array}\right] = \left[\begin{array}{c} M_1^* v \\ \hdashline \vdots \\ \hdashline M_n^* v \end{array}\right],$$

where we have denoted the rows of M by M_i^* (equivalently M_i are the columns of M^*). Each entry of the vector w is then $w_i = M_i^* v$, namely the *inner product* of v with each of the columns of M^*.

Matrix Products as Inner and Outer Vector Products

In a similar manner to matrix-vector products, matrix-matrix products can be given multiple interpretations. Given any two matrices H and G with compatible dimensions, the product matrix HG can be thought of in at least two ways. The standard definition of matrix multiplication involves partitioning H into rows and G into columns, and defining the product as

$$HG = \left[\begin{array}{c} H_1^* \\ \hdashline \vdots \\ \hdashline H_n^* \end{array}\right] \left[\begin{array}{c:c:c} G_1 & \cdots & G_q \end{array}\right] = \left[\begin{array}{c:c:c} H_1^* G_1 & \cdots & H_1^* G_q \\ \hdashline \vdots & & \vdots \\ \hdashline H_n^* G_1 & \cdots & H_n^* G_q \end{array}\right],$$

i.e. the ij'th scalar element of HG is the *inner product* of the i'th column of H^* with the j'th column of G.

A second interpretation arises from partitioning H into columns and G into rows

$$HG = \left[\begin{array}{c:c:c} H_1 & \cdots & H_m \end{array}\right] \left[\begin{array}{c} G_1^* \\ \hdashline \vdots \\ \hdashline G_m^* \end{array}\right] = \left[H_1\right]\left[\begin{array}{c} G_1^* \end{array}\right] + \cdots + \left[H_m\right]\left[\begin{array}{c} G_m^* \end{array}\right]. \tag{7.2}$$

Viewed this way, HG is expressed as the sum of m outer products of corresponding columns of H and G^*. Each of those outer products is a rank-1 matrix. Such rank-1 matrices are simple objects whose properties can be completely characterized and understood[1]. This point of view is most useful when a given matrix A is written as the product of several matrices with special properties. We can then interpret this as a rank-1 decomposition of A. Several notions such as the diagonalization of a matrix and the Singular Value Decomposition (SVD) can be understood as various types of rank-1 decompositions. These notions are easily generalizable to operators on infinite dimensional spaces.

Eigenvalue/vector Decomposition

Partition notation is particularly useful to illustrate the connections between the concepts of eigenvalues/eigenvectors and that of diagonalization. Assume that an $n \times n$ matrix A has n eigenvalues $\{\lambda_i\}$ with corresponding eigenvectors $\{v_i\}$ that can be chosen to be linearly independent[2]. We thus have the following relations

$$Av_i \;=\; \lambda_i v_i, \qquad\qquad i = 1, \ldots, n,$$

with the set $\{v_i\}$ being linearly independent (note that the eigenvalues need not be distinct). It is an elementary, but powerful observation that these n matrix-vector relations can be rewritten as the following single matrix equation

$$\begin{aligned}
\left[\begin{array}{c:c:c} Av_1 & \cdots & Av_n \end{array}\right] &= \left[\begin{array}{c:c:c} \lambda_1 v_1 & \cdots & \lambda_n v_n \end{array}\right] \\
&\Updownarrow \\
\left[\begin{array}{c} A \end{array}\right]\left[\begin{array}{c:c:c} v_1 & \cdots & v_n \end{array}\right] &= \left[\begin{array}{c:c:c} v_1 & \cdots & v_n \end{array}\right]\left[\begin{array}{c:c:c} \lambda_1 & & \\ \hdashline & \ddots & \\ \hdashline & & \lambda_n \end{array}\right] \qquad (7.3) \\
&\Updownarrow \\
AV &= V\Lambda, \qquad (7.4)
\end{aligned}$$

where V is a matrix whose columns are the eigenvectors of A, and Λ is the diagonal matrix made up of the eigenvalues of A. Equation (5.3) states that V is the similarity transformation that diagonalizes A

$$A \;=\; V\Lambda V^{-1}. \tag{7.5}$$

This actually proves that a matrix is diagonalizable iff it has a full set of eigenvectors[3] (regardless of the eigenvalues). The diagonalizing similarity transformation V in (7.5) is the

[1] A rank-1 matrix can also be interpreted geometrically as a projection on a 1-dimensional subspace. This provides useful geometric intuition.

[2] This statement is equivalent to several other well-known characterizations. Amongst them is that (i) the matrix is diagonalizable, (ii) the Jordan form contains no non-trivial Jordan blocks, or (iii) the geometric multiplicity of each eigenvalue is equal to its algebraic multiplicity.

[3] A *full set of eigenvectors* means a set of linearly independent eigenvectors that span the whole space (in finite dimensions, this means that we have n linearly independent eigenvectors for an $n \times n$ matrix). The choice of eigenvectors will not be unique of there are eigenvalue multiplicities.

non-singular matrix made up of the eigenvectors of A as its columns. Thus diagonalizing a matrix or an operator is equivalent to finding all of its eigenvalues and eigenvectors.

Equation (7.5) can also be given an interpretation as a rank-1 decomposition of A as follows. Let $W^* := V^{-1}$ (or equivalently $W := V^{-*}$), and observe that (7.5) can be rewritten as

$$\begin{aligned} A &= \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} \begin{bmatrix} w_1^* \\ \vdots \\ w_n^* \end{bmatrix} \\ &= \lambda_1 \begin{bmatrix} v_1 \end{bmatrix} \begin{bmatrix} w_1^* \end{bmatrix} + \cdots + \lambda_n \begin{bmatrix} v_n \end{bmatrix} \begin{bmatrix} w_n^* \end{bmatrix} = \sum_{i=1}^{n} \lambda_i \, v_i w_i^* . \end{aligned} \tag{7.6}$$

This is a rank-1 decomposition of A, namely into n rank-1 matrices made up of outer products of the respective columns of V and V^{-1} scaled by the respective eigenvalues. A rank-1 matrix has a geometrical interpretation as a projection, and therefore this decomposition can be geometrically interpreted as decomposing A into n (possibly oblique) projections.

When the eigenvectors $\{v_i\}$ are not mutually orthogonal, there is no direct way to obtain V^{-1} from V. However, there are important relationships between the columns of V (the eigenvectors of A) and the columns $W := V^{-*}$ (which are the *rows* of V^{-1}) :

1. *The columns of W are the eigenvectors of A^*:* This can be seen from the following calculation

   $$AV = V\Lambda \;\Leftrightarrow\; V^*A^* = \Lambda^* V^* \;\Leftrightarrow\; WV^*A^*W = W\Lambda^* V^* W \;\Leftrightarrow\; A^*W = W\Lambda^*,$$

   where the last step follows from $W^*V = VW^* = I$.

2. *The sets $\{v_i\}$ and $\{w_i\}$ form reciprocal bases:* Reciprocal bases[4] have the property that $\langle v_i \, , \, w_j \rangle = v_i^* w_j = \delta_{i-j}$. This is easily seen to be true from the following partitioning of $W^*V = I$

   $$W^*V = \begin{bmatrix} w_1^* \\ \vdots \\ w_n^* \end{bmatrix} \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix} = \begin{bmatrix} w_1^* v_1 & \cdots & w_1^* v_n \\ \vdots & & \vdots \\ w_n^* v_1 & \cdots & w_n^* v_n \end{bmatrix} = I.$$

   The reciprocal basis is useful since it allows for writing any vector x in terms of a basis $\{v_i\}$ by observing

   $$x = VW^*x = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix} \begin{bmatrix} w_1^* \\ \vdots \\ w_n^* \end{bmatrix} \begin{bmatrix} x \end{bmatrix} = \sum_{i=1}^{n} v_i \, \langle w_i \, , \, x \rangle \, . \tag{7.7}$$

   Thus the coefficients of expansion of a vector x in a basis $\{v_i\}$ are the inner products $\langle w_i \, , \, x \rangle$ of the vector with the respective elements of the reciprocal basis $\{w_i\}$.

We can obtain yet another interpretation of the action of a diagonalizable matrix on a vector by comparing (7.7) with what equation (7.6) gives for Ax

$$Ax = \sum_{i=1}^{n} \lambda_i \, v_i \, \langle w_i \, , \, x \rangle \, . \tag{7.8}$$

Thus A acts on any vector x by scaling its components along each eigenvector v_i by the corresponding eigenvalue λ_i multiplied by the projection $\langle w_i \, , \, x \rangle$.

[4] Another common term is *dual bases*.

Symmetric (Hermitian) Matrices

The above decompositions are considerably simplified when the matrix A is Hermitian ($A = A^*$). In this case all its eigenvectors are mutually orthogonal (more precisely, one can choose a complete mutually orthonormal set from amongst the eigenvector of A). This implies that $V^{-1} = V^*$, i.e. V is unitary, and consequently, the above decompositions takes the particularly simple form

$$A \;=\; \lambda_1 \begin{bmatrix} v_1 \end{bmatrix} \begin{bmatrix} \; v_1^* \; \end{bmatrix} + \cdots + \lambda_n \begin{bmatrix} v_n \end{bmatrix} \begin{bmatrix} \; v_n^* \; \end{bmatrix}.$$

The action of A on any vector x in (7.8) becomes

$$Ax \;=\; \sum_{i=1}^{n} \lambda_i \; v_i \, \langle v_i \, , \, x \rangle \, .$$

and can be interpreted geometrically as scaling the *orthogonal projection* of x on v_i by the corresponding eigenvalue λ_i.

Singular Value Decompositions

Any matrix (i.e. not necessarily diagonalizable or Hermitian) has a Singular Value Decomposition (SVD) of the form

$$A \;=\; U \; \Sigma \; V^*, \tag{7.9}$$

where Σ is a diagonal matrix and both U and V are unitary matrices. In fact, the columns of U and V are actually the (mutually orthonormal) eigenvectors of AA^* and A^*A respectively. This can be seen from the calculation

$$\begin{aligned} AA^* \;&=\; U\Sigma V^* V \Sigma U^* \;=\; U\Sigma^2 U^*, \\ A^*A \;&=\; V\Sigma U^* U \Sigma V^* \;=\; V\Sigma^2 V^*. \end{aligned}$$

These relations provide one way to calculate the SVD, by finding the eigenvectors and eigenvalues of the two Hermitian matrices AA^* and A^*A. Note that the singular values of A are the square roots of the eigenvalues of AA^* (or A^*A, the non-zero ones are the same).

The SVD can be thought of as a rank-1 decomposition in a similar manner as the previous decompositions. Partition U and V into columns $\{u_i\}$ and $\{v_i\}$ respectively

$$\begin{aligned} A \;&=\; \left[\begin{array}{c:c:c} u_1 & \cdots & u_n \end{array}\right] \left[\begin{array}{c:c:c} \sigma_1 & & \\ \hdashline & \ddots & \\ \hdashline & & \sigma_n \end{array}\right] \left[\begin{array}{c} v_1^* \\ \hdashline \vdots \\ \hdashline v_n^* \end{array}\right] \\ &=\; \sigma_1 \begin{bmatrix} u_1 \end{bmatrix} \begin{bmatrix} \; v_1^* \; \end{bmatrix} + \cdots + \sigma_n \begin{bmatrix} u_n \end{bmatrix} \begin{bmatrix} \; v_n^* \; \end{bmatrix} \;=\; \sum_{i=1}^{n} \sigma_i \; u_i v_i^*. \end{aligned} \tag{7.10}$$

Using this, the action of A as a linear transformation on an arbitrary vector x can be

rewritten as

$$\begin{aligned} Ax &= \left[\begin{array}{c:c:c} u_1 & \cdots & u_n \end{array}\right] \left[\begin{array}{c:c:c} \sigma_1 & & \\ \hdashline & \ddots & \\ \hdashline & & \sigma_n \end{array}\right] \left[\begin{array}{c} v_1^* \\ \hdashline \vdots \\ \hdashline v_n^* \end{array}\right] \left[\begin{array}{c} x \end{array}\right] \\ &= \left[\begin{array}{c:c:c} u_1 & \cdots & u_n \end{array}\right] \left[\begin{array}{c:c:c} \sigma_1 & & \\ \hdashline & \ddots & \\ \hdashline & & \sigma_n \end{array}\right] \left[\begin{array}{c} v_1^* x \\ \hdashline \vdots \\ \hdashline v_n^* x \end{array}\right] = \left[\begin{array}{c:c:c} u_1 & \cdots & u_n \end{array}\right] \left[\begin{array}{c} \sigma_1 \langle v_1 \, , \, x \rangle \\ \hdashline \vdots \\ \hdashline \sigma_n \langle v_n \, , \, x \rangle \end{array}\right] \\ &= \sum_{i=1}^{n} \sigma_i \; u_i \langle v_i \, , \, x \rangle \, . \end{aligned}$$

The last equation can be interpreted as follows. The linear transformation A acts on x by first taking its projections onto each of the orthonormal vectors v_i, these numbers are then multiplied ("amplified") by σ_i respectively to produce a linear combination of the orthonormal vectors u_i. This interpretation immediately shows that the image space of A is the span of the vectors u_i corresponding to non-zero singular values, while the null space of A is the span of the vectors v_i corresponding to the zero singular values.

The SVD also clearly shows how a matrix "amplifies" vector lengths. For example, any vector x aligned with v_1 will produce a vector Ax aligned with u_1 but with length amplified by σ_1. The vectors v_i are called the *right singular vectors* of A, while u_i are called the *left singular vectors* of A.

7.1 Block LU, UL, and LDU Decompositions: Schur Complements

Recall that any square matrix M has a Lower-Diagonal-Upper (LDU) factorization (possibly after column or row permutations) of the form

$$M \; = \; LDU,$$

where L is a lower-triangular matrix, U an upper-triangular matrix, both with all ones on the diagonals, and D is a diagonal matrix. Such factorization are done by Gaussian elimination and are useful in solving linear equations by forwards or backwards substitution.

In this section, we look at "block factorizations" similar to the above, but with block-triangular and block-diagonal matrices. The so-called Schur complement and related results are concerned with matrices or operators with a 2×2 block decomposition and block-LDU factorizations of the form (see e.g. Equation (7.15))

$$M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} = \begin{bmatrix} I & 0 \\ M_{21}M_{11}^{-1} & I \end{bmatrix} \begin{bmatrix} M_{11} & 0 \\ 0 & M_{22} - M_{21}M_{11}^{-1}M_{12} \end{bmatrix} \begin{bmatrix} I & M_{11}^{-1}M_{12} \\ 0 & I \end{bmatrix}, \quad (7.11)$$

which is valid under the assumption that M_{11} is invertible. Notice the block-lower-triangular, block-diagonal, and block-upper-triangular structure of this factorization.

We can derive conditions on invertibility and definiteness of M in terms of the matrices occurring in the factorization above. For example, it is clear that when M_{11} is invertible, then M is invertible iff $M_{22} - M_{21}M_{11}^{-1}M_{12}$, which is called a Schur complement, is invertible. Similar statements can be made about definiteness when M is Hermitian.

The key idea is to find transformations (constructed from the submatrices M_{ij}) that will block-triangularize and block-diagonalize M. The transformations are in general not similarity transformations, as they may transform the domain of M differently from its

range. None the less, they are useful because they are simple to construct directly from M (e.g. no calculations of invariant subspaces are required), but yet yield valuable conditions on invertibility, definiteness and the like.

The geometric interpretation of the partitioning in (7.11) in general is that if $M : \mathsf{V} \to \mathsf{W}$ is a linear operator where the vector spaces have decompositions $\mathsf{V} = \mathsf{V}_1 \oplus \mathsf{V}_2$ and $\mathsf{W} = \mathsf{W}_1 \oplus \mathsf{W}_2$ respectively, then $M_{ij} : \mathsf{V}_j \to \mathsf{W}_i$ are the four possible restriction/projections of M with respect to those decompositions

$$\begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} = \begin{bmatrix} \Pi_{\mathsf{W}_1} M|_{\mathsf{V}_1} & \Pi_{\mathsf{W}_1} M|_{\mathsf{V}_2} \\ \Pi_{\mathsf{W}_2} M|_{\mathsf{V}_1} & \Pi_{\mathsf{W}_2} M|_{\mathsf{V}_2} \end{bmatrix} \quad \Leftrightarrow \quad \begin{bmatrix} \mathsf{W}_1 \\ \mathsf{W}_2 \end{bmatrix} = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \begin{bmatrix} \mathsf{V}_1 \\ \mathsf{V}_2 \end{bmatrix}. \tag{7.12}$$

The equality on the right should be thought of as useful notation to intuitively interpret the identity on the left.

Block-LU, UL and UDL Decompositions

We will need to assume that either M_{22} or M_{11} is invertible. We begin with the former case, and state two problems whose solutions below can be thought of as types of *Gaussian eliminations on block rows or block columns*.

1. Find an invertible transformation $R : \mathsf{W}_1 \oplus \mathsf{W}_2 \to \mathsf{W}_1 \oplus \mathsf{W}_2$ such that the composition $RM : \mathsf{V} \to \mathsf{W}$ is block-lower-triangular.

   This is easily done as follows. Write (7.12) in terms of vectors v and w, and observe that if $H := RM$ is to be block-lower-triangular, then we must have

   $$\begin{aligned} w_1 &= M_{11}v_1 + M_{12}v_2 \\ w_2 &= M_{21}v_1 + M_{22}v_2 \end{aligned}, \qquad \begin{aligned} y_1 &= H_{11}v_1 \\ y_2 &= H_{21}v_1 + H_{22}v_2 \end{aligned}.$$

   In order to arrive at an equation for y_1 that does not involve v_2, it looks like we need to eliminate v_2 from the equations for w as follows

   $$\begin{array}{rrcl} & w_1 & = & M_{11}v_1 + M_{12}v_2 \\ - & M_{12}M_{22}^{-1}w_2 & = & M_{12}M_{22}^{-1}M_{21}v_1 + M_{12}v_2 \\ \hline & w_1 - M_{12}M_{22}^{-1}w_2 & = & -M_{12}M_{22}^{-1}M_{21}v_1 + 0 \end{array}.$$

   Thus it seems like we should define y_1 to be $w_1 - M_{12}M_{22}^{-1}w_2$ in order to have an equation for y_1 that depends only on v_1. Since we have no requirements on what y_2 depends on, we leave it as $y_2 = w_2$, i.e.

   $$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} := \begin{bmatrix} I & M_{12}M_{22}^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix},$$

   Now we can easily verify that this transformation gives a block-UL decomposition of M

   $$\begin{bmatrix} I & -M_{12}M_{22}^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} = \begin{bmatrix} M_{11} - M_{12}M_{22}^{-1}M_{21} & 0 \\ M_{21} & M_{22} \end{bmatrix}$$

   $$\Rightarrow \quad \boxed{\begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} = \begin{bmatrix} I & M_{12}M_{22}^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} M_{11} - M_{12}M_{22}^{-1}M_{21} & 0 \\ M_{21} & M_{22} \end{bmatrix}}, \tag{7.13}$$

   where the last equation follows from the following easily verifiable fact

   $$\begin{bmatrix} I & Z \\ 0 & I \end{bmatrix}^{-1} = \begin{bmatrix} I & -Z \\ 0 & I \end{bmatrix}, \qquad \begin{bmatrix} I & 0 \\ Z & I \end{bmatrix}^{-1} = \begin{bmatrix} I & 0 \\ -Z & I \end{bmatrix}$$

   for any matrix Z of compatible dimensions.

   Notice that the transformation above leaves the second block-rows of M unchanged.

2. Find an invertible transformation $T : \mathsf{V}_1 \oplus \mathsf{V}_2 \to \mathsf{V}_1 \oplus \mathsf{V}_2$ such that the composition $MT : V \to V$ is block-upper-triangular.

   We again write (7.12) in terms of vectors v and w, and observe that if $G := MT$ is to be block-upper-triangular, then we must have

$$\begin{array}{rl} w_1 = & M_{11}v_1 + M_{12}v_2 \\ w_2 = & M_{21}v_1 + M_{22}v_2 \end{array}, \qquad \begin{array}{rl} w_1 = & G_{11}x_1 + G_{12}x_2 \\ w_2 = & G_{22}x_2 \end{array}.$$

   To discover what G_{22} should be, we simply manipulate the second v, w equation as

$$\begin{array}{rl} w_1 = & M_{11}v_1 + M_{12}v_2 \\ w_2 = & M_{22}\left(M_{22}^{-1}M_{21}v_1 + v_2\right) \end{array}, \qquad \begin{array}{rl} w_1 = & G_{11}x_1 + G_{12}x_2 \\ w_2 = & G_{22}x_2 \end{array}.$$

   Thus it seems like we should define x_2 to be $M_{22}^{-1}M_{21}v_1 + v_2$, and leave $x_1 = v_1$, i.e.

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} := \begin{bmatrix} I & 0 \\ M_{22}^{-1}M_{21} & I \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \qquad \Leftrightarrow \qquad \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} I & 0 \\ -M_{22}^{-1}M_{21} & I \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

   Again, it is now easy to see what this transformation converts M into block-upper-triangular form

$$\begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \begin{bmatrix} I & 0 \\ -M_{22}^{-1}M_{21} & I \end{bmatrix} = \begin{bmatrix} M_{11} - M_{12}M_{22}^{-1}M_{21} & M_{12} \\ 0 & M_{22} \end{bmatrix}$$

$$\Rightarrow \qquad \boxed{\begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} = \begin{bmatrix} I & 0 \\ M_{22}^{-1}M_{21} & I \end{bmatrix} \begin{bmatrix} M_{11} - M_{12}M_{22}^{-1}M_{21} & M_{12} \\ 0 & M_{22} \end{bmatrix}.} \tag{7.14}$$

   Notice that this transformation leaves the second block-columns of M unchanged.

If M_{11} rather M_{22} is invertible, then similar transformations can be derived as those above that involve only M_{11}^{-1} rather than M_{22}^{-1}. See Exercise 7.1.

The two transformations constructed in (7.13) and (7.14) will block-triangularize M by either post or premultiplying by a simple transformation matrix. (7.13) has the property that it leaves the 2nd block-row of M unchanged, while (7.14) leaves the 2nd block-column unchanged. Thus performing the two transformations successively leads to a form of M that is both block-upper-triangular and block-lower-triangular, i.e block diagonal

$$\begin{bmatrix} I & -M_{12}M_{22}^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \begin{bmatrix} I & 0 \\ -M_{22}^{-1}M_{21} & I \end{bmatrix} = \begin{bmatrix} M_{11} - M_{12}M_{22}^{-1}M_{21} & 0 \\ 0 & M_{22} \end{bmatrix},$$

which we rewrite as a block-UDL and block-LDU decompositions respectively

$$\boxed{\begin{aligned} \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} &= \begin{bmatrix} I & M_{12}M_{22}^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} M_{11} - M_{12}M_{22}^{-1}M_{21} & 0 \\ 0 & M_{22} \end{bmatrix} \begin{bmatrix} I & 0 \\ M_{22}^{-1}M_{21} & I \end{bmatrix} \\ &= \begin{bmatrix} I & 0 \\ M_{21}M_{11}^{-1} & I \end{bmatrix} \begin{bmatrix} M_{11} & 0 \\ 0 & M_{22} - M_{21}M_{11}^{-1}M_{12} \end{bmatrix} \begin{bmatrix} I & M_{11}^{-1}M_{12} \\ 0 & I \end{bmatrix} \end{aligned}} \tag{7.15}$$

It should be emphasized that the transformation of M shown here is in general *not a similarity transformation* since the two transformation matrices are not inverses of each other. None the less, several useful conclusions can be drawn from this transformation as described next.

7.1.1 Corollaries of Block-Decompositions

The first conclusion we can draw from (7.15) is about the invertibility of a partitioned matrix. Observe that the two transformation matrices in (7.15) are clearly invertible. Which leads to the following conclusion.

Lemma 7.1. *Consider the partitioned matrix* $M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}$.

1. *If* M_{22} *is invertible, and its Schur complement*

$$\Delta_{11} := M_{11} - M_{12}M_{22}^{-1}M_{21}$$

is invertible, then M *is invertible with inverse*

$$\begin{aligned} \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}^{-1} &= \begin{bmatrix} I & 0 \\ -M_{22}^{-1}M_{21} & I \end{bmatrix} \begin{bmatrix} \Delta_{11}^{-1} & 0 \\ 0 & M_{22}^{-1} \end{bmatrix} \begin{bmatrix} I & -M_{12}M_{22}^{-1} \\ 0 & I \end{bmatrix} \\ &= \begin{bmatrix} \Delta_{11}^{-1} & -\Delta_{11}^{-1}M_{12}M_{22}^{-1} \\ -M_{22}^{-1}M_{21}\Delta_{11}^{-1} & M_{22}^{-1} + M_{22}^{-1}M_{21}\Delta_{11}^{-1}M_{12}M_{22}^{-1} \end{bmatrix} \end{aligned} \tag{7.16}$$

2. *If* M_{11} *is invertible, and its Schur complement*

$$\Delta_{22} := M_{22} - M_{21}M_{11}^{-1}M_{12}$$

is invertible, then M *is invertible with inverse*

$$\begin{aligned} \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}^{-1} &= \begin{bmatrix} I & -M_{11}^{-1}M_{12} \\ 0 & I \end{bmatrix} \begin{bmatrix} M_{11}^{-1} & 0 \\ 0 & \Delta_{22}^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -M_{21}M_{11}^{-1} & I \end{bmatrix} \\ &= \begin{bmatrix} M_{11}^{-1} + M_{11}^{-1}M_{12}\Delta_{22}^{-1}M_{21}M_{11}^{-1} & -M_{11}^{-1}M_{12}^{-1}\Delta_{22}^{-1} \\ -\Delta_{22}^{-1}M_{21}M_{11}^{-1} & \Delta_{22}^{-1} \end{bmatrix} \end{aligned} \tag{7.17}$$

3. *Depending on whether* M_{11} *or* M_{22} *is invertible, we have*

$$\det(M) = \det(M_{11})\det\big(M_{22} - M_{21}M_{11}^{-1}M_{12}\big) = \det(M_{22})\det\big(M_{11} - M_{12}M_{22}^{-1}M_{21}\big)$$

The last statement follows from (7.15) by recalling that for any two matrices with compatible dimensions $\det(AB) = \det(A)\det(B)$, and for block-upper (or lower) matrices

$$\det\left(\begin{bmatrix} A & 0 \\ B & C \end{bmatrix}\right) = \det(A)\det(C) \qquad \Rightarrow \qquad \det\left(\begin{bmatrix} I & 0 \\ B & I \end{bmatrix}\right) = 1.$$

The lemma above basically says that the inversion of a block-2x2 matrix can be achieved by the inversion of one of its diagonal blocks and the corresponding Schur complement. There is another important identity hidden in (7.16) and (7.17). Equating the $(1,1)$ blocks in both of these expressions for M^{-1} gives

$$\begin{aligned} \Delta_{11}^{-1} &= M_{11}^{-1} + M_{11}^{-1}M_{12}\Delta_{22}^{-1}M_{21}M_{11}^{-1} \\ \big(M_{11} - M_{12}M_{22}^{-1}M_{21}\big)^{-1} &= M_{11}^{-1} + M_{11}^{-1}M_{12}\left(M_{22} - M_{21}M_{11}^{-1}M_{12}\right)^{-1}M_{21}M_{11}^{-1}. \end{aligned} \tag{7.18}$$

This last identity can be thought of as an identity involving four matrices M_{ij} with compatible dimensions. This is sometimes done without reference to the 2×2 block matrix from which they can be considered to have arisen. If we simply rename the matrices, this last identity can be stated as the famous *Matrix Inversion Lemma*, also sometimes known as the *Sherman-Morrison-Woodbury* formula.

Corollary 7.2. *Consider four matrices A, B, C, D of compatible dimensions. If A and D are invertible, then*

$$\begin{aligned} &D \text{ invertible} \quad \Rightarrow \quad \begin{bmatrix} A & B \\ C & D \end{bmatrix} \text{invertible} \quad \Leftrightarrow \quad \left(A - BD^{-1}C\right) \text{invertible} \\ &A \text{ invertible} \quad \Rightarrow \quad \begin{bmatrix} A & B \\ C & D \end{bmatrix} \text{invertible} \quad \Leftrightarrow \quad \left(D - CA^{-1}B\right) \text{invertible} \end{aligned} \tag{7.19}$$

If both A and D are invertible, and either of the two conditions above hold, then

$$\left(A - BD^{-1}C\right)^{-1} = A^{-1} + A^{-1}B\left(D - CA^{-1}B\right)^{-1}CA^{-1}. \tag{7.20}$$

This lemma is usually stated as (7.20) only, which makes no mention of a 2x2-block matrix. Rather, on the face of it, it appears to be a lemma about sums of products of matrices. It is however best understood and proved through comparison with Lemma 7.1 as follows. Given four matrices A, B, C, D of compatible dimensions, form the following matrix[5] and apply the transformation (7.15) to it

$$\begin{aligned} \begin{bmatrix} A & B \\ C & D \end{bmatrix} &= \begin{bmatrix} I & BD^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} A - BD^{-1}C & 0 \\ 0 & D \end{bmatrix} \begin{bmatrix} I & 0 \\ D^{-1}C & I \end{bmatrix} \\ &= \begin{bmatrix} I & 0 \\ CA^{-1} & I \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & D - CA^{-1}B \end{bmatrix} \begin{bmatrix} I & A^{-1}B \\ 0 & I \end{bmatrix}. \end{aligned} \tag{7.21}$$

The first statement in Lemma 7.2 follows immediately; The 2x2-block matrix is invertible iff $A - BD^{-1}C$ is invertible (from the first equality above), which holds iff $D - CA^{-1}B$ is invertible (second equality above). The formula (7.20) follows from (7.18), which we recall follows from simply equating terms in the two different block-diagonalizations (7.16) and (7.17).

There are many uses of the Matrix Inversion Lemma. Perhaps the most prominent is for producing a formula for the inverse of a matrix made up of two pieces, the first being a matrix whose inverse is known, and the second being a "low rank" addition to the matrix. To emphasize this point, redraw (7.20) to show dimensions in a typical usage

$$\begin{aligned} &\left(\begin{bmatrix} \\ A \\ \\ \end{bmatrix} + \begin{bmatrix} \\ B \\ \\ \end{bmatrix} [D] \begin{bmatrix} & C & \end{bmatrix}\right)^{-1} = \\ &\qquad \begin{bmatrix} \\ A^{-1} \\ \\ \end{bmatrix} - \begin{bmatrix} \\ A^{-1} \\ \\ \end{bmatrix} \begin{bmatrix} \\ B \\ \\ \end{bmatrix} \left([D^{-1}] + \begin{bmatrix} & C & \end{bmatrix} \begin{bmatrix} \\ A^{-1} \\ \\ \end{bmatrix} \begin{bmatrix} \\ B \\ \\ \end{bmatrix}\right)^{-1} \begin{bmatrix} & C & \end{bmatrix} \begin{bmatrix} \\ A^{-1} \\ \\ \end{bmatrix}, \end{aligned}$$

where D is a much smaller matrix than A. Note that we have replaced $-B$ with B and D^{-1} with D in this last formula compared with (7.20). If A^{-1} is known, then $(A + BDC)^{-1}$ given by this formula only requires inverting the much smaller matrix $D^{-1} + CA^{-1}B$. An extreme case of this situation is when BDC is of rank one. This is the so-called *rank-one update* $A + uv^*$ where A is square and u and v are vectors. The rank-1 matrix uv^* is called a rank-one update of A. The formula (7.20) applied to this case is sometimes referred to as the *Sherman-Morrison formula*

$$\left(A + uv^*\right)^{-1} = A^{-1} - \frac{1}{1 + v^*A^{-1}u}\left(A^{-1}u\right)\left(v^*A^{-1}\right). \tag{7.22}$$

[5] Observe that the dimension compatibility condition that allows $A - BD^{-1}C$ to be formed is exactly the same as that allows the matrix (7.21) to be formed.

Note that in this case $D = 1$, and $D + CA^{-1}B$ is a scalar, so its inverse is just the reciprocal scalar. Note also that $(A + uv^*)^{-1}$ is itself a rank-one update of A^{-1} since $\left(A^{-1}u\right)$ is a column vector and $\left(v^*A^{-1}\right)$ is a row vector.

The rank-one update formula is very useful in deriving *recursive algorithms* such as Recursive Least Squares. To obtain an estimate from a batch of data, a solution of some linear system of equations is performed. When new data comes in, one would like to "update" the solution with this new data without having to solve the entire system from scratch. The rank-one update provides such recursive algorithms.

Hermitian Matrices

We now examine what the block-UDL decomposition (7.15) implies about the definiteness of a Hermitian matrix M, for which it reads

$$\begin{bmatrix} M_{11} & M_o \\ M_o^* & M_{22} \end{bmatrix} = \begin{bmatrix} I & M_o M_{22}^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} M_{11} - M_o M_{22}^{-1} M_o^* & 0 \\ 0 & M_{22} \end{bmatrix} \begin{bmatrix} I & 0 \\ M_{22}^{-1} M_o^* & I \end{bmatrix}. \tag{7.23}$$

Note that M Hermitian implies that both M_{11} and M_{22} are Hermitian. Furthermore, in this case the transformation matrices are adjoints of each other

$$\begin{bmatrix} I & M_o M_{22}^{-1} \\ 0 & I \end{bmatrix}^* = \begin{bmatrix} I & 0 \\ M_{22}^{-1} M_o^* & I \end{bmatrix}.$$

Thus the transformation of M in (7.23) is a *congruence transformation*, and therefore preserves the definiteness of a matrix. The definiteness of a block-diagonal matrix is simply determined by the definiteness of its diagonal blocks. We summarize this in the following statement.

Corollary 7.3. *Consider the partitioned Hermitian matrix* $M = \begin{bmatrix} M_{11} & M_o \\ M_o^* & M_{22} \end{bmatrix}$. *If* $M_{22} > 0$ *(*$M_{11} > 0$*) is positive definite, then the definiteness of* M *is equivalent to the definiteness of its* Schur complement

$$\begin{aligned} M > 0 \quad &\Leftrightarrow \quad M_{11} - M_o M_{22}^{-1} M_o^* > 0 \\ \left(M > 0 \quad \right. &\left. \Leftrightarrow \quad M_{22} - M_o^* M_{11}^{-1} M_o > 0 \right) \end{aligned}. \tag{7.24}$$

The statement (7.24) also holds if $>$ *is replaced by* $\geq$.

Note that there is an analogous statement for negative (semi)-definiteness. The reader should write that one out as an exercise.

Completion of Squares

Corollary 7.3 has an interpretation as a "completion of squares" statement which gives additional insight into the Schur complement in the Hermitian case. Consider the quadratic form generated from a partitioned Hermitian matrix M

$$v^* M v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}^* \begin{bmatrix} M_{11} & M_o \\ M_o^* & M_{22} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = v_1^* M_{11} v_1 + v_2^* M_{22} v_2 + 2 v_1^* M_o v_2. \tag{7.25}$$

A necessary condition for $M > 0$ is that $M_{11} > 0$ and $M_{22} > 0$, thus the first two terms above are always positive for non-zero v_1 and v_2. However, the third term is not guaranteed to be positive, but the Schur complement condition insures that if it becomes negative, the

first two terms dominate so that the overall sum is positive. This is a consequence of the following completion of squares argument.

Assuming that M_{22} is invertible and using the decomposition (7.23), we can rewrite the quadratic form (7.25) as

$$\begin{aligned}
&\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}^* \begin{bmatrix} M_{11} & M_o \\ M_o^* & M_{22} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \\
&= \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}^* \begin{bmatrix} I & M_o M_{22}^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} M_{11} - M_o M_{22}^{-1} M_o^* & 0 \\ 0 & M_{22} \end{bmatrix} \begin{bmatrix} I & 0 \\ M_{22}^{-1} M_o^* & I \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \\
&= \begin{bmatrix} v_1 \\ w \end{bmatrix}^* \begin{bmatrix} M_{11} - M_o M_{22}^{-1} M_o^* & 0 \\ 0 & M_{22} \end{bmatrix} \begin{bmatrix} v_1 \\ w \end{bmatrix}, \quad \text{by defining } \begin{bmatrix} v_1 \\ w \end{bmatrix} := \begin{bmatrix} I & 0 \\ M_{22}^{-1} M_o^* & I \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}.
\end{aligned}$$

In other words, if we define the vector

$$w \;:=\; v_2 + M_{22}^{-1} M_o^* v_1,$$

then the quadratic form is rewritten as

$$v_1^* M_{11} v_1 + v_2^* M_{22} v_2 + 2 v_1^* M_o v_2 = v_1^* \left(M_{11} - M_o M_{22}^{-1} M_o^* \right) v_1 \;+\; w^* M_{22} w. \tag{7.26}$$

Note that the invertibility of the mapping $(v_1, v_2) \mapsto (v_1, w)$ implies that $(v_1, v_2) = 0 \quad \Leftrightarrow (v_1, w) = 0$. Therefore if M_{22} is invertible, then the positivity $M_{11} - M_o M_{22}^{-1} M_o^* \geq 0$ (or strict positivity $M_{11} - M_o M_{22}^{-1} M_o^* > 0$) of the Schur complement along with $M_{11} \geq 0$ (or $M_{11} > 0$) insures the positivity (or strict positivity) of the quadratic form (7.25).

The equality (7.26) is sometimes referred to as a matrix version of the classical completion of squares argument since it can be viewed as

$$\begin{aligned}
&v_1^* M_{11} v_1 + v_2^* M_{22} v_2 + 2 v_1^* M_o v_2 \\
&= v_1^* \left(M_{11} - M_o M_{22}^{-1} M_o^* \right) v_1 \;+\; \left(v_2 + M_{22}^{-1} M_o^* v_1 \right)^* M_{22} \left(v_2 + M_{22}^{-1} M_o^* v_1 \right) \\
&\qquad = v_1^* \left(M_{11} - M_o M_{22}^{-1} M_o^* \right) v_1 \;+\; v_2^* M_{22} v_2 \;+\; 2 v_1^* M_o v_2 \;+\; v_1^* M_o M_{22}^{-1} M_o^* v_1.
\end{aligned}$$

In other words, but adding and subtracting the term $v_1^* M_o M_{22}^{-1} M_o^* v_1$, the quadratic form can be turned into a perfect square.

Eigenvalues and Eigenvectors "Updates"

The matrix inversion Lemma 7.2 can be put to use for characterizing eigenvalues/vectors for matrices of the form $A + UV$, where it is assumed that the eigenvalues/vectors of A are known, and UV is a low-rank matrix.

If λ is not an eigenvalue of A, then Corollary 7.2 states

$$\underset{\text{invertible}}{\left(\lambda I - (A + UV)\right)} \quad \Leftrightarrow \quad \underset{\text{invertible}}{\left((\lambda I - A) - UV\right)} \quad \Leftrightarrow \quad \underset{\text{invertible}}{\left(I - V\left(\lambda I - A\right)^{-1} U\right)}.$$

In the case when UV has finite rank, the last statement gives a test for characterizing the eigenvalues of $A + UV$ as the zeros of a function.

Lemma 7.4. *Consider an operator update $A + UV$ where U and V have finite-dimensional domain and range respectively (i.e. UV is finite rank). Any number $\lambda \in \mathbb{C}$ which is not an eigenvalue of A is an eigenvalue of $A + UV$ iff it is the root of the* characteristic function

$$f(\lambda) \;:=\; \det\left(I - V\left(\lambda I - A\right)^{-1} U\right).$$

This test can be applied in the cases where the computation of the characteristic function is feasible such as in the next example.

Example 7.5. Consider the following matrix M which arises in a (2nd order) finite-difference discretization of the 2nd derivative operator with Homogenous Neumann boundary conditions. It can be written as the sum of a circulant matrix A and a rank one matrix as follows

$$M = \begin{bmatrix} -1 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -1 \end{bmatrix} = \begin{bmatrix} -2 & 1 & & & 1 \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ 1 & & & 1 & -2 \end{bmatrix} + \begin{bmatrix} 1 \\ \\ \\ -1 \end{bmatrix} \begin{bmatrix} 1 & & & -1 \end{bmatrix} =: C_{\mathrm{a}} + vv^*,$$

where C_{a} is the circulant matrix formed from the vector $\mathrm{a} := \{-2, 1, 0, \ldots, 0, 1\}$. The eigenvalues of M that are not eigenvalues of C_{a} are the zeros of the characteristic function

$$f(\lambda) := 1 - v^* \left(\lambda I - C_{\mathrm{a}}\right)^{-1} v$$

Since C_{a} is circulant, its eigenvalues/vectors are well known, and are precisely the values of the Discrete Fourier Transform (DFT) of the vector a (see Chapter **??**). The inverse of $\lambda I - C_{\mathrm{a}}$, as well as the inner product $v^* \left(\lambda I - C_{\mathrm{a}}\right)^{-1} v$ are easiest to compute with the DFT as follows.

Denote the DFTs of a and v above by $\hat{\mathrm{a}}$ and $\hat{v}$ respectively, i.e. $\hat{v} = Fv$, where F is the matrix representation of the DFT[6] which has the property $FF^* = nI$. Now transform the inner product as follows (using the fact $v = F^{-1}\hat{v} = (1/n)F^*\hat{v}$)

$$\begin{aligned} v^* \left(\lambda I - C_{\mathrm{a}}\right)^{-1} v &= \left(F^{-1}\hat{v}\right)^* \left(\lambda I - C_{\mathrm{a}}\right)^{-1} \left(F^{-1}\hat{v}\right) = \hat{v}^* \left(\lambda FF^* - FC_{\mathrm{a}}F^*\right)^{-1} \hat{v} \\ &= \hat{v}^* \Big(n\,\lambda I - n\,\mathsf{diag}(\hat{\mathrm{a}})\Big)^{-1} \hat{v} = \frac{1}{n} \sum_{l=0}^{n-1} \frac{|\hat{v}_l|^2}{\lambda - \hat{\mathrm{a}}_l}. \end{aligned} \tag{7.27}$$

Now $\hat{\mathrm{a}}$ and $\hat{v}$ are easy to compute as

$$\begin{aligned} \hat{\mathrm{a}}_l &= -2 + e^{-j\frac{2\pi}{n}l} + e^{j\frac{2\pi}{n}l} = -2 + 2\cos\left(\tfrac{2\pi}{n}l\right), \qquad l \in \mathbb{Z}_n \\ \hat{v}_l &= 1 - e^{j\frac{2\pi}{n}l} = 1 - \cos\left(\tfrac{2\pi}{n}l\right) - j\sin\left(\tfrac{2\pi}{n}l\right) \quad \Rightarrow \quad |\hat{v}_l|^2 = 2 - 2\cos\left(\tfrac{2\pi}{n}l\right) \end{aligned} \tag{7.28}$$

The expression (7.27) becomes

$$f(\lambda) := 1 - v^* \left(\lambda I - C_{\mathrm{a}}\right)^{-1} v = 1 - \frac{1}{n} \sum_{l=0}^{n-1} \frac{2 - 2\cos\left(\frac{2\pi}{n}l\right)}{\lambda + 2 - 2\cos\left(\frac{2\pi}{n}l\right)} \tag{7.29}$$

A plot of this function is shown in Figure 7.1 for an example with $n = 10$. The figure clearly shows that the zeros of f capture all the eigenvalues of M that are not shared with C_{a}.

The matrix inversion Lemma 7.2 can also be used to characterize *eigenvectors* as well. An eigenvector of a 2×2-block matrix with an eigenvalue λ is characterized by the null-space condition

$$\begin{bmatrix} \lambda I - A & -B \\ -C & \lambda I - D \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{7.30}$$

[6] The (non-unitary) DFT of a vector v of length n is defined by $\hat{v}_l := \sum_{k\in\mathbb{Z}_n} v_k e^{-j\frac{2\pi}{n}kl}$. We can write this as $\hat{v} = Fv$, where F has the property $F^{-1} = (1/n)F^*$. A circulant matrix is diagonalized by F so that $FC_{\mathrm{a}}F^{-1} = (1/n)FC_{\mathrm{a}}F^* = \mathsf{diag}(\hat{\mathrm{a}})$, where $\hat{\mathrm{a}}$ is the DFT of the vector a.

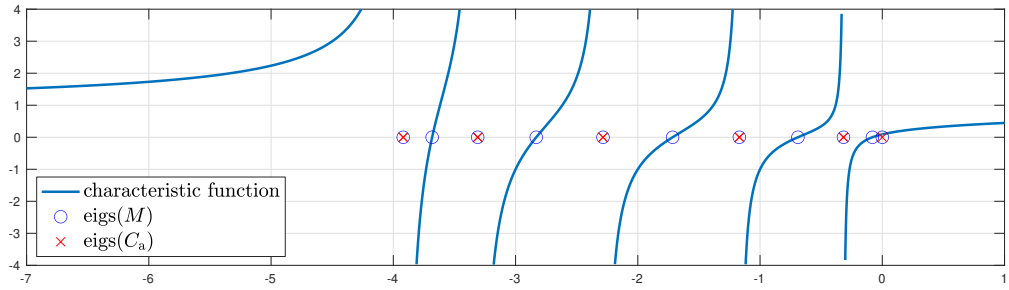


Figure 7.1: The characteristic function (7.29) $f(\lambda)$ of Example 7.5 for the matrix update $M = C_{\mathrm{a}} + vv^*$ (with matrix sizes of 10). The zeros of f are the eigenvalues of M that are *not* eigenvalues of C_{a}. As shown, the zeros of f capture all the eigenvalues of M that are not shared with C_{a}.

If λ is not an eigenvalue of either A or D, then we can use this relation to obtain v_1 from v_2 and vice versa

$$(\lambda I - A)\, v_1 - B v_2 = 0 \qquad \Rightarrow \qquad v_1 = (\lambda I - A)^{-1} B v_2, \tag{7.31}$$

$$-C v_1 + (\lambda I - D)\, v_2 = 0 \qquad \Rightarrow \qquad v_2 = (\lambda I - D)^{-1} C v_1. \tag{7.32}$$

We can combine the above four equations in two different ways

$$0 \;=\; (\lambda I\text{-}A)\, v_1 - B v_2 \;=\; \Big((\lambda I\text{-}A) - B\,(\lambda I\text{-}D)^{-1} C\Big)\, v_1 \qquad (\text{using } (7.32)) \tag{7.33}$$

$$0 \;=\; -C v_1 + (\lambda I\text{-}D)\, v_2 \;=\; \Big((\lambda I\text{-}D) - C\,(\lambda I\text{-}A)^{-1} B\Big)\, v_2 \qquad (\text{using } (7.31)) \tag{7.34}$$

These relations can be used as follows. Suppose for a given λ (not an eigenvalue of either A or D) we find a a vector v_2 that satisfies (7.34), then (7.31) can be used to construct a vector v_1 that satisfies (7.33) since

$$\begin{aligned}
\Big((\lambda I\text{-}A) - B\,(\lambda I\text{-}D)^{-1} C\Big)\, v_1 \;&=\; \Big((\lambda I\text{-}A) - B\,(\lambda I\text{-}D)^{-1} C\Big)\,(\lambda I - A)^{-1} B v_2 \\
&=\; \Big(B - B\,(\lambda I\text{-}D)^{-1} C\,(\lambda I - A)^{-1} B\Big)\, v_2 \\
&=\; B\,(\lambda I\text{-}D)^{-1}\Big((\lambda I\text{-}D) - C\,(\lambda I - A)^{-1} B\Big)\, v_2 \;=\; 0.
\end{aligned}$$

We can now use the above development to device a method for finding *eigenvector updates* similar to that for eigenvalue updates of Lemma 7.4

Lemma 7.6. *Consider an operator update $A + UV$ where U and V have finite-dimensional domain and range respectively (i.e. UV is finite rank). For any eigenvalue λ of $A + UV$ that is not an eigenvalue of A, any corresponding eigenvector v is given by*

$$v \;=\; (\lambda I - A)^{-1} U\, w, \qquad \textit{where} \qquad \Big(I - V\,(\lambda I - A)^{-1} U\Big)\, w \;=\; 0.$$

Proof. In the formulas (7.33)-(7.34), define $D := (\lambda - 1)I$, thus $\lambda I - D = I$. Now define $B := U$ and $C := V$, and the statement follows. □

This lemma is most useful when compositions like $V\,(\lambda - A)^{-1} U$ can be calculated and their eigenvectors computed. The case of circulant matrices, or more generally, Toeplitz operators are possible applications.

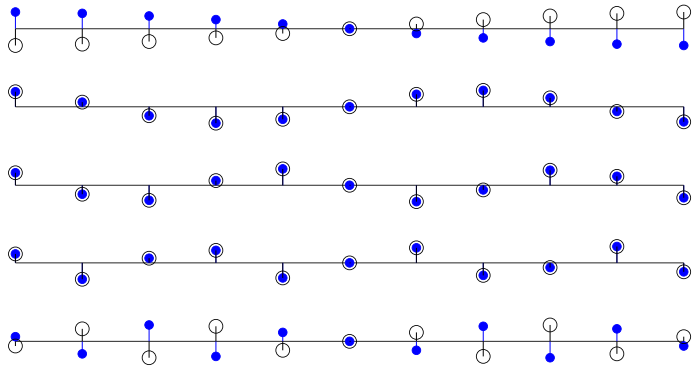


Figure 7.2: Eigenvectors of the matrix $M = C_{\rm a} + vv^*$ of Examples 7.5-7.7 for the case $n = 11$. Only five eigenvalues of M are distinct from those of $C_{\rm a}$, and Lemma 7.6 is applied to those. The solid blue circles depict the eigenvectors computed with the formula (7.35), while the hollow circles are the eigenvectors computed directly for M. Since eigenvectors are only determined up to scalar multiples, the two computations differ only up to a scalar multiple as can be seen above.

Example 7.7. We now revisit Example 7.5 to calculate some of the eigenvectors of the update using the above lemma. For the subset of eigenvalues of M that are not eigenvalues of $C_{\rm a}$ (see Figure 7.1), the eigenvectors v are calculated from the lemma by

$$\mathsf{v} \;=\; (\lambda I - C_{\rm a})^{-1}\, v\;\alpha, \qquad \text{where} \qquad \Big(1 - v^*\,(\lambda I - C_{\rm a})^{-1}\, v\Big)\,\alpha \;=\; 0. \tag{7.35}$$

Note that in this case α is any scalar, and therefore the eigenvectors are any scalar multiple of $(\lambda I - C_{\rm a})^{-1}\, v$ (where λ is the eigenvalue of M that is not an eigenvalue of $C_{\rm a}$). This circulant-matrix times vector product can be calculated using the DFT formulas (7.28), or directly with numerical computations. Figure 7.2 illustrates the computation of the eigenvectors with the formula (7.35) and with direct numerical computations using eigenvalue/vector routines.

7.2 Block Similarity Transformations: Sylvester and Riccati Equations

The block LU, UL and UDL decompositions described in the previous section can be thought of as block triangularization (for the LU and UL decompositions), and block diagonalization (in the UDL case). However, the transformations required are not similarity transformations as can be seen for example from the statements in Lemma 7.1. Those transformations are however useful in studying the invertibility of block-decomposed matrices. In the Hermitian case, the required transformations (7.23) are actually congruence transformations, which preserve the sign definiteness of Hermitian matrices, and this makes them useful in studying the definiteness of block-partitioned matrices. On the other hand, if one is interested in studying the spectra and invariant subspaces of block-partitioned matrices, it is useful to try to block-triangularize or block-diagnonalize using similarity transformations. These question will lead naturally to matrix Riccati and Sylvester equations as we now illustrate.

Observe that all transformations used in the previous section (e.g. (7.16) or (7.17)) are of the forms

$$U_X := \begin{bmatrix} I & X \\ 0 & I \end{bmatrix}, \qquad L_X := \begin{bmatrix} I & 0 \\ X & I \end{bmatrix}.$$

Such transformations are convenient to work with since their inverses are easily established

$$U_X^{-1} := \begin{bmatrix} I & -X \\ 0 & I \end{bmatrix}, \qquad L_X^{-1} := \begin{bmatrix} I & 0 \\ -X & I \end{bmatrix}.$$

Now we investigate whether we can use such transformations to block-diagonalize or block-triangularize a 2×2 partitioned matrix. Consider first a block upper-triangular matrix and a similarity transformation of the form U_X

$$\begin{bmatrix} I & -X \\ 0 & I \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} \begin{bmatrix} I & X \\ 0 & I \end{bmatrix} = \begin{bmatrix} A_{11} & A_{11}X - XA_{22} + A_{12} \\ 0 & A_{22} \end{bmatrix}.$$

Thus if we can find a matrix X such that the (1,2) block is zero, then we have found a similarity transformation that renders the given upper triangular matrix into a block-diagonal form. Finding such a matrix is equivalent to solving a matrix Sylvester equation

$$A_{11}X - XA_{22} = -A_{12} \qquad \Rightarrow \qquad \begin{bmatrix} I & -X \\ 0 & I \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} \begin{bmatrix} I & X \\ 0 & I \end{bmatrix} = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}. \quad (7.36)$$

The solvability of this equation is governed by the properties of the matrix-valued operator $\mathcal{S}(X) := AX + BX$. Note that in this case this operator is "square" since X and A_{12} have the same size. Recall that the eigenvalues of a Sylvester operator are given by

$$\mathcal{S}(X) := AX + XB \qquad \Rightarrow \qquad \mathsf{eigs}(\mathcal{S}) \;=\; \mathsf{eigs}(A) + \mathsf{eigs}(B)\,,$$

where $+$ stands for all possible sums of elements of the two sets. Thus $\mathcal{S}(X) := A_{11}X - XA_{22}$ is invertible iff no eigenvalue of A_{11} is equal to an eigenvalue of A_{22} (for otherwise the difference of those two eigenvalues would be zero). We can now conclude that there exists a block-triangularizing transformation of the form (7.36) if[7] there are no common eigenvalues between A_{11} and A_{22}.

Now consider a 2×2 block-partitioned matrix and a similarity transformation of the form

$$\begin{bmatrix} I & 0 \\ -X & I \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} I & 0 \\ X & I \end{bmatrix}. = \begin{bmatrix} A_{11} + A_{12}X & A_{12} \\ -XA_{11} - XA_{12}X + A_{21} + A_{22}X & -XA_{12} + A_{22} \end{bmatrix}$$

Thus if the matrix X solves the following matrix Algebraic Riccati Equation (ARE), then the similarity transformation above converts A to block upper-triangular form

$$A_{22}X - XA_{11} - XA_{12}X + A_{21} \;=\; 0 \quad (7.37)$$

$$\Rightarrow \qquad \begin{bmatrix} I & 0 \\ -X & I \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} I & 0 \\ X & I \end{bmatrix} = \begin{bmatrix} A_{11} + A_{12}X & A_{12} \\ 0 & A_{22} - XA_{12} \end{bmatrix}. \quad (7.38)$$

Unlike a Sylvester equation, the Riccati equation (7.37) always has multiple solutions. The fact that the transformation (7.38) is a similarity transformation aids greatly in understanding the set of solutions. For example, we know that

$$\begin{aligned} \mathsf{eigs}(A) \;=\; \mathsf{eigs}\left(\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}\right) \;&=\; \mathsf{eigs}\left(\begin{bmatrix} A_{11} + A_{12}X & A_{12} \\ 0 & A_{22} - XA_{12} \end{bmatrix}\right) \\ &=\; \mathsf{eigs}(A_{11} + A_{12}X) \;\cup\; \mathsf{eigs}(A_{22} - XA_{12})\,. \end{aligned}$$

[7] Note that for any given matrix, this condition is sufficient but may not be necessary. There could be cases where $\mathcal{S}$ is not invertible, but A_{12} is in its image space, and then there is an infinite number of solutions of the Sylvester equation.

The fact that the set $\mathsf{eigs}(A)$ is independent of X means that any solution of the ARE is such that the eigenvalues of A are "divided up" between the eigenvalues of $A_{11}+A_{12}X$ and $A_{22}-XA_{12}$.

We can go further by examining the eigenvectors after rearranging (7.38) into

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} I & 0 \\ X & I \end{bmatrix} = \begin{bmatrix} I & 0 \\ X & I \end{bmatrix} \begin{bmatrix} A_{11}+A_{12}X & A_{12} \\ 0 & A_{22}-XA_{12} \end{bmatrix}$$
$$\Rightarrow \quad \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} I \\ X \end{bmatrix} = \begin{bmatrix} I \\ X \end{bmatrix} (A_{11}+A_{12}X)\,.$$

This implies that $\mathsf{span}\begin{bmatrix} I \\ X \end{bmatrix}$ is an invariant subspace for A, and the eigenvalues of $A_{11}+A_{12}X$ are the eigenvalues of A restricted to that subspace. Under some conditions, we can actually go in reverse, that is, by finding invariant subspaces of A, we can construct solutions to the ARE as described next.

Suppose we are given an ARE of the following form

$$BX + XA - XMX + Q \;=\; 0,$$

where no assumptions are made on the matrices other than compatibility of dimensions. Let X be $n \times m$. Compatibility of dimensions implies that A must have m columns, B has n rows, M is $m \times n$ and Q is $n \times m$. This also implies that A and B are square. We can therefore arrange the four coefficient matrices A, B, M, Q into a 2×2 partitioned matrix as follows

$$H \;:=\; \begin{bmatrix} -A & M \\ Q & B \end{bmatrix}.$$

Now we look for invariant subspaces of H that can be represented as $\mathsf{span}\begin{bmatrix} I \\ X \end{bmatrix}$ for some matrix $n \times m$ matrix X, which must be of dimension m. Any subspace of dimension m can be parameterized as the image space of an $(n+m) \times m$ matrix $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$, where the partitioning is such that X_1 is an $m \times m$ square matrix. If X_1 is invertible, then

$$\mathsf{Im}\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \;=\; \mathsf{Im}\left(\begin{bmatrix} I \\ X_2X_1^{-1} \end{bmatrix} X_1\right) \;=\; \mathsf{Im}\begin{bmatrix} I \\ X \end{bmatrix}, \qquad X := X_2X_1^{-1}.$$

The condition that X_1 is invertible is equivalent to the condition that $\begin{bmatrix} X_1 & 0 \\ X_2 & I \end{bmatrix}$ is invertible, which can also expressed as the subspaces $\mathsf{Im}\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ and $\mathsf{Im}\begin{bmatrix} 0 \\ I \end{bmatrix}$ being complementary in $\mathbb{R}^{n+m}$. The conclusions above are summarized next.

Lemma 7.8. *Consider the following matrix Algebraic Riccati Equation (ARE)*

$$BX + XA - XMX + Q \;=\; 0,$$

where X is $n \times m$, and all other matrices are of compatible dimensions. Consider also the 2×2 block partitioned matrix

$$H \;:=\; \begin{bmatrix} -A & M \\ Q & B \end{bmatrix}.$$

1. *Every solution X of the ARE is such that* $\mathsf{Im}\begin{bmatrix} I \\ X \end{bmatrix}$ *is an m-dimensional invariant subspace of H, and* $\mathsf{eigs}(-A+MX)$ *are the m eigenvalues of H restricted to that invariant subspace.*

2. *Every m-dimensional invariant subspace* $\mathsf{Im}\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ *of H that is complementary to* $\mathsf{Im}\begin{bmatrix} 0 \\ I \end{bmatrix}$ *corresponds to a solution $X = X_2 X_1^{-1}$ of the ARE. Furthermore,* $\mathsf{eigs}(-A+MX)$ *are the eigenvalues of H restricted to that subspace.*

The lemma above gives a computational procedure for finding ARE solutions. First look for m-dimensional invariant subspaces of H that satisfy the complementarity condition, and then form X from a matrix that spans that subspace. In the simplest case that H has distinct eigenvalues, there are at most $\binom{n}{m}$ such subspaces. Another way to think about this algorithm is as follows. Since any solution of the ARE corresponds to a selection of a size m subset of the eigenvalues of H, first decide on which subset should be the eigenvalues of $-A+MX$. Second, check if that invariant subspace satisfies the complementarity condition, and if it does, construct $X = X_2 X_1^{-1}$ as above.

Exercises

Exercise 7.1

If M_{11} rather than M_{22} is invertible, then show using similar arguments to those leading to (7.13), (7.14) and (7.15) that the following holds

$$\begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} = \begin{bmatrix} I & 0 \\ M_{21}M_{11}^{-1} & I \end{bmatrix} \begin{bmatrix} M_{11} & M_{12} \\ 0 & M_{22} - M_{21}M_{11}^{-1}M_{12} \end{bmatrix} \tag{7.39}$$

$$= \begin{bmatrix} M_{11} & 0 \\ M_{21} & M_{22} - M_{21}M_{11}^{-1}M_{12} \end{bmatrix} \begin{bmatrix} I & M_{11}^{-1}M_{12} \\ 0 & I \end{bmatrix} \tag{7.40}$$

$$= \begin{bmatrix} I & 0 \\ M_{21}M_{11}^{-1} & I \end{bmatrix} \begin{bmatrix} M_{11} & 0 \\ 0 & M_{22} - M_{21}M_{11}^{-1}M_{12} \end{bmatrix} \begin{bmatrix} I & M_{11}^{-1}M_{12} \\ 0 & I \end{bmatrix} \tag{7.41}$$

Note that it is easy to simply verify the above equations by direct calculations. However, it is a useful exercise to imitate the arguments leading to (7.13) and (7.14) for the following reason. What if the above equations were not given to you? What if you only knew that M_{11} is invertible, and you needed to discover those transformations?

Bibliography

[1] H Behnke, F Bachmann, K Fladt, W Suss, and H Kunle. *Fundamentals of Mathematics, Volume I, Foundations of Mathematics/The Real Number System and Algebra*. MIT Press, Cambridge, MA, 1974.

[2] Heinrich Behnke, Friedrich Bachmann, Kuno Fladt, W Suss, H Kunle, and SH Gould. *Fundamentals of Mathematics, Volume III: Analysis*. The MIT Press, 1984.

[3] Abram Samoilovitch Besicovitch. *Almost periodic functions*, volume 4. Dover New York, 1954.

[4] Sydney Henry Gould, Heinrich Behnke, F Bachmann, and K Fladt. *Fundamentals of mathematics, Volume II: Geometry*. MIT Press, 1974.