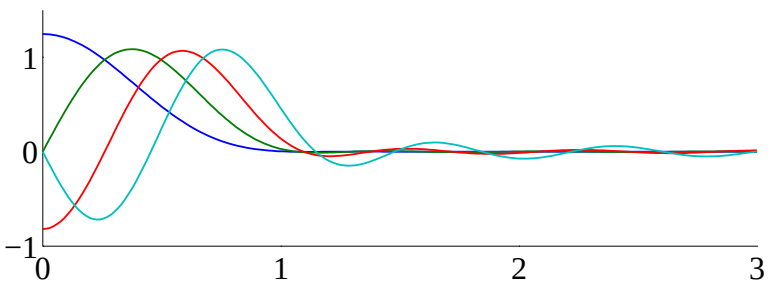
ECE 735 Course Notes

John A. Gubner

Department of Electrical and Computer Engineering
University of Wisconsin–Madison

January 11, 2020



Contents

CHAPTER 1
Introduction

1.1. Overview of Problems of Interest

Signal synthesis and recovery is all about the situation illustrated in Figure 1.1, when the system and the output are given, and the goal is to find a corresponding

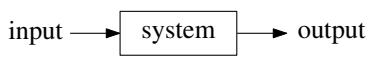


Figure 1.1. A typical system.

input. In the **signal synthesis problem**, the output is a design specification, and the goal is to find an input that causes the system to generate the desired output. In the **signal recovery problem**, the output is measurement data, and the goal is to find the input that generated it. In practice, there may be many inputs that can generate the same output; hence, additional constraints must be imposed on the input to select a particular solution.

We can pose the situation in Figure 1.1 somewhat more formally as shown in Figure 1.2, which suggests the equation

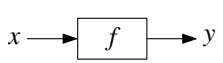


Figure 1.2. A mathematically defined system.

$$y = f(x). \tag{1.1}$$

Equation (1.1) immediately raises several mathematical questions. First, what kind of object is x? We answer this by requiring that $x \in X$, where X is some set of admissible system inputs; i.e., admissible arguments for f. Second, what kind of object is y? Certainly, y must be the same kind of object as $f(x)$ for any $x \in X$. In general, we require that for all $x \in X$, $f(x) \in Y$ for some fixed set Y. Note that it is *not* required that for all $y \in Y$, there exist an $x \in X$ with $f(x) = y$.

In many problems, we have a mathematical model in which a measurement $y_0 \in Y$ is equal to $f(x_0)$ for some $x_0 \in X$. However, due to noise or modeling errors, when x_0 is applied to the system, the output that is actually measured is

$$y_1 \neq y_0. \tag{1.2}$$

Somehow, based on the observation y_1, we want to find x_0. There are two situations to consider. First suppose there is an x_1 such that $y_1 = f(x_1)$. Then we would like to say something like, "if y_1 is close to y_0, then x_1 will be close to x_0." Second, suppose that there is no $x_1 \in X$ with $y_1 = f(x_1)$. In this case, we could consider the problem

$$\min_{x \in X} \text{distance}\big(y_1, f(x)\big). \tag{1.3}$$

What do "close" and "distance" mean? In general, since x and y may be very different kinds of objects, we may need different notions of closeness or distance. In order to examine these questions precisely, we must learn a bit about **metric spaces**.

In many signal processing applications, the sets X and Y are **vector spaces**. Among other things, this means that there is a notion of addition for objects in X and a notion of addition for objects in Y. In the vector-space context, we can employ additive noise models. For example, instead of (1.2), we can write $y_1 = y_0 + \Delta y$ for some nonzero Δy. Distance in vector spaces is often measured by a **norm**. In this case, every vector has a notion of size associated with it. This is usually denoted by $\|x\|$. The distance between two vectors x_0 and x_1 is then taken to be $\|x_0 - x_1\|$. Since different spaces have different norms, for emphasis we sometimes write $\|x\|_X$ for $x \in X$ and $\|y\|_Y$ for $y \in Y$.

Another advantage of having X and Y be vector spaces is that it makes sense to talk about linear functions (usually called **linear transformations** or **linear operators**). In this case, we often denote the function (transformation or operator) by A; we write $y = Ax$ instead of $y = f(x)$. When Y is a normed vector space, (1.3) becomes

$$\min_{x \in X} \|y_1 - Ax\|_Y. \tag{1.4}$$

As x runs over X, Ax runs over

$$\text{range}\, A := \{Ax : x \in X\}.$$

Hence, we are trying to find a point in range$A$ that is closest to y_1. This is a **projection problem**. Since A is linear, its range is a subspace. When y_1 is a point in the plane and the subspace is a line through the origin, the projection problem is straightforward, as shown in Figure 1.3. The point we need has the property that the error vector is perpendicular (**orthogonal**) to every vector in the subspace. How can we generalize this idea when y_1 is a waveform, e.g., a sine wave, instead of a point in two-dimensional space? This brings us to the topic of **inner-product spaces**. If Y is an inner-product space with inner product denoted by $\langle \cdot, \cdot \rangle_Y$, we show later that x_1 achieves the minimum in (1.4) if and only if

$$\langle y_1 - Ax_1, Ax \rangle_Y = 0, \quad \text{for all } x \in X.$$

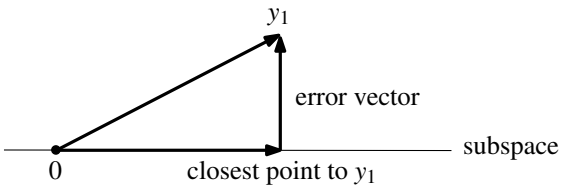


Figure 1.3. A projection problem in the plane.

If in addition X is an inner-product space with inner product denoted by $\langle \cdot, \cdot \rangle_X$, we show later that the above formula holds if and only if

$$\langle A^* y_1 - (A^* A) x_1, x \rangle_X = 0, \quad \text{for all } x \in X,$$

where A^* is the **adjoint** of A (defined later). Since x is arbitrary, it follows that x_1 satisfies the *linear* equation

$$(A^* A) x_1 = A^* y_1. \tag{1.5}$$

In many cases the solution of this equation can be found. In particular, if X is finite dimensional, then A^*A can be identified with a matrix, $A^* y_1$ is a column vector, and x_1 can be found using MATLAB.

If you have studied linear algebra, you may be familiar with the **diagonalization** of matrices and the **singular-value decomposition** (SVD) of matrices. These are fundamental tools for studying linear operators on finite-dimensional spaces. However, operators encountered in applications are often defined on infinite-dimensional spaces. Fortunately, the notions of diagonalization and SVD can be generalized to infinite-dimensional settings.

Formula (1.4) is an example of an optimization problem. Not all such problems can be solved so easily. To minimize a real-valued function of x, which we now call f, what should we do? If x is a real number or an element of $\mathbb{R}^d$, we can differentiate. What should we do if x is a waveform? How does one differentiate with respect to a waveform? Later we generalize the notion of derivative to functions defined on infinite-dimensional spaces by introducing the Fréchet and Gâteaux derivatives. Since setting these derivatives equal to zero means solving $f'(x) = 0$, we have a special case of (1.1). It's all about solving equations....

1.2. Examples and Some Basic Definitions

Recall that a linear, time-invariant system with input x and output y can be described by the equation

$$y(t) = \int_{-\infty}^{\infty} h(t - \tau) x(\tau)\, d\tau,$$

where h is the system **impulse response**. We consider the situation in which the output y is given, and our job is to find an input x that causes the system to produce the output y.

When y is a system design parameter, the determination of x is regarded as a **synthesis problem**. When y is given by measurement data (usually noisy), the determination of x is regarded as a **recovery problem**.

At first glance, it appears that the synthesis problem can be solved easily using Fourier transforms. If Y, H, and X denote the Fourier transforms of y, h, and x, then $Y(f) = H(f)X(f)$, and it follows that x is the inverse Fourier transform of $Y(f)/H(f)$. However, a little thought raises the following issues. It is possible that the the system blocks certain frequencies; i.e., there may be a range of frequencies f for which $H(f) = 0$. If $Y(f)$ is nonzero for these f, there is no solution of $Y(f) = H(f)X(f)$ for these f. For example, if y is a "bang-bang control" signal, then y will need to have jump discontinuities; however, such a signal cannot be generated by a bandlimited system.[1]

Other difficulties arise if we impose the following two conditions. First, we assume that $y(t)$ is given only for a finite range of times, say $0 \le t \le T_y$. Second, x is to be a finite-duration signal, say $0 \le t \le T_x$. Under these conditions, we have

$$y(t) = \int_0^{T_x} h(t-\tau)x(\tau)\,d\tau, \quad 0 \le t \le T_y.$$

See Sections 7.5.1 and 7.6 for details.

Notes

Note 1.1. If y is a finite-energy waveform that is bandlimited, say $Y(f) = 0$ for $|f| > W$, then

$$y(t) = \int_{-W}^{W} Y(f)e^{j2\pi ft}\,dt,$$

where Y is square integrable on $(-\infty,\infty)$ [34, the **Plancherel Theorem**]. It is then easy to show that y is a continuous function of t;[a] i.e., y cannot have jump discontinuities.

[a] The dual of this fact is proved in Problem 7.19.

CHAPTER 2
Vector Spaces

A **vector space** is a nonempty collection of objects that we can add and subtract, very much like the real or complex numbers that you are used to.[1] Also, there is a set of **scalars**, which for us will always be the real numbers, $\mathbb{R}$, or the complex numbers, $\mathbb{C}$. These scalars interact with vectors via an operation called **scalar multiplication** such that if a is a scalar and x is a vector in X, then ax is a vector in X.[2] When the set of scalars is $\mathbb{R}$, we call X a **real vector space**, and when the set of scalars is $\mathbb{C}$, we call X a **complex vector space**.

There are many examples of vector spaces. The ones you are most familiar with are the finite-dimensional Euclidean spaces $\mathbb{R}^d$ and $\mathbb{C}^d$. However, we are most interested in vector spaces of waveforms defined on some time interval. For example, we may consider the vector space of continuous functions defined on a given time interval.

2.1. Linear Combinations

If $x_1,\ldots,x_n$ are vectors, and $c_1,\ldots,c_n$ are scalars, we call

$$\sum_{k=1}^{n} c_k x_k$$

a **linear combination**. Notice that a linear combination is a sum with a *finite* number of terms. Infinite sums are *not* considered linear combinations.

As mentioned, we are most interested in vector spaces of waveforms. In many cases, the waveforms x_k are related to a common pulse v by time delays. For example, we may have $x_k(t) = v(t-\tau_k)$ for some delay τ_k. Suppose we are given scalars $c_1,\ldots,c_n$, and we wish to plot the linear combination of delayed waveforms

$$y(t) := \sum_{k=1}^{n} c_k x_k(t) = \sum_{k=1}^{n} c_k v(t-\tau_k)$$

in MATLAB.

Example 2.1. We use `lincmb`[3] to plot the linear combination

$$y(t) = 5e^{-(t-4)^2/2} - e^{-(t-5)^2/2} + 3e^{-(t-8)^2/2}, \quad 0 \le t \le 12.$$

The commands

```
v = @(t)exp(-t.^2/2);    % Define v(t) = exp(-t^2/2)
t = linspace(0,12,200);
tau = [4 5 8];
c = [5 -1 3];
y = lincmb(t,c,v,tau);
subplot(2,1,1)
plot(t,y);
```

generate the top graph in Figure 2.1. Notice that we defined `v` as an **anonymous function** using the `@` notation. If we had defined `v` in an M-file, then statement with `lincmb` should include single quotes around `v` as in `lincmb(t,c,'v',tau)`.

The function `lincmb` can also return the individual delayed pulses x_k so that they can be plotted. This is accomplished by replacing the last three lines of MATLAB code above with

```
[y,xmat] = lincmb(t,c,v,tau);
subplot(2,1,1)
plot(t,y);
subplot(2,1,2)
plot(t,xmat);
```

The results are shown in Figure 2.1.

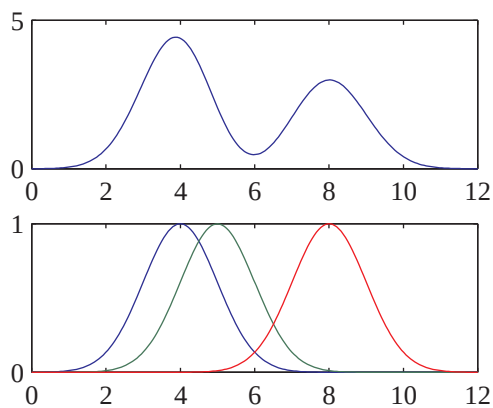


Figure 2.1. Linear combination y (top) and shifted pulses x_k (bottom) for Example 2.1.

More generally, we may include a scale factor s_k as well and have

$$x_k(t) = v(s_k(t - \tau_k)).$$

Suppose we are given scalars $c_1, \ldots, c_n$, and we wish to plot

$$y(t) := \sum_{k=1}^{n} c_k x_k(t) = \sum_{k=1}^{n} c_k v(s_k(t - \tau_k))$$

in MATLAB.

Example 2.2. Consider the common decaying exponential pulse,

$$v(t) := \begin{cases} e^{-t}, & t \geq 0, \\ 0, & t < 0. \end{cases}$$

We use `lincmb` to plot the linear combination of the five decaying exponentials,

$$y(t) = \sum_{k=1}^{5} (-1)^k v((t-k)/k), \quad 0 \leq t \leq 7,$$

as shown in the top graph in Figure 2.2. The bottom graph in the figure shows the

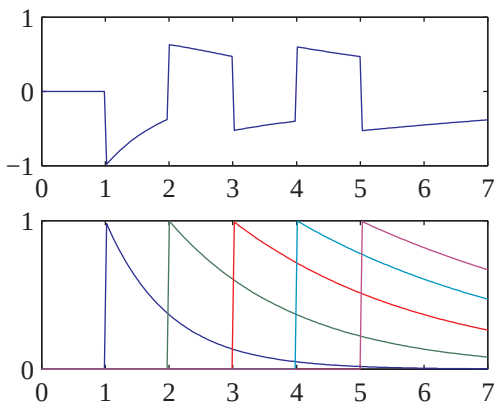


Figure 2.2. Linear combination y (top) and scaled and shifted pulses x_k (bottom) for Example 2.2.

scaled and shifted pulses x_k. The commands

```
v = @(t)exp(-t).*(t >= 0);
t = linspace(0,7,200);
tau = [1:5];
c = (-1).^tau;
s = 1./tau;
[y,xmat] = lincmb(t,c,v,tau,s);
```

```
subplot(2,1,1)
plot(t,y);
subplot(2,1,2)
plot(t,xmat);
```

generate Figure 2.2.

2.2. Linear Independence

Let G denote a nonempty subset of a vector space X. The set G may have finitely many elements or infinitely many elements. We say that G is **linearly independent** if whenever $x_1,\ldots,x_n$ is a finite collection of vectors from G and $c_1,\ldots,c_n$ are scalars,

$$\sum_{k=1}^{n} c_k x_k = 0 \text{ implies all the coefficients } c_k \text{ must be zero.}$$

If G is not linearly independent, we say G is **linearly dependent**.

Some care is needed to appreciate the equation

$$\sum_{k=1}^{n} c_k x_k = 0$$

when the x_k are waveforms. For example, if the x_k are waveforms defined on some time interval, say $[3,7]$, the above equation is understood as shorthand for

$$\sum_{k=1}^{n} c_k x_k(t) = 0, \quad \text{for all } t \in [3,7].$$

Example 2.3. Consider the waveforms $x(t) = 1/t$ and $y(t) = 1/t^2$ for $0 < t < 1$. We show that x and y are linearly independent. To do this, we let c_1 and c_2 be arbitrary scalars, and we assume that

$$\frac{c_1}{t} + \frac{c_2}{t^2} = 0, \quad t \in (0,1). \tag{2.1}$$

Since (2.1) holds for *all* $t \in (0,1)$, it must hold for any particular values of t, say $t = 1/4$ and $t = 1/2$. This leads to the system of equations

$$\begin{aligned} 4c_1 \;+\; 16c_2 \;&=\; 0, \\ 2c_1 \;+\; \;4c_2 \;&=\; 0, \end{aligned}$$

which can easily be solved to show that $c_2 = 0$ and then that $c_1 = 0$ as well. Consider, however, another approach. Multiply (2.1) by t^2 to get

$$c_1 t + c_2 = 0, \quad t \in (0,1). \tag{2.2}$$

Since equality holds for $t \in (0,1)$, equality holds in the limit as $t \to 0$. Taking the required limit shows that $c_2 = 0$. It then follows that $c_1 t = 0$ for $t \in (0,1)$. Specializing to $t = 1/2$ shows that $c_1 = 0$ as well. For a third approach, which avoids taking an explicit limit in (2.2), let us differentiate (2.2) instead. This yields $c_1 = 0$. Using this in (2.2) yields $c_2 = 0$ as well.

Remark. The preceding example shows that there are many approaches to proving that a collection of waveforms is linearly independent. In a specific case, one approach may be significantly easier to carry out than another.

Example 2.4 (Lagrange Interpolation). Given *distinct* times $\tau_1, \ldots, \tau_n$, we can define the **Lagrange fundamental interpolating polynomials**

$$\ell_j(t) := \frac{(t-\tau_1)\cdots(t-\tau_{j-1})(t-\tau_{j+1})\cdots(t-\tau_n)}{(\tau_j-\tau_1)\cdots(\tau_j-\tau_{j-1})(\tau_j-\tau_{j+1})\cdots(\tau_j-\tau_n)}$$

for $j = 1, \ldots, n$. Each ℓ_j is a polynomial of degree less than n. Note that $\ell_j(\tau_j) = 1$, and that for $i \neq j$, $\ell_j(\tau_i)$ contains the factor $(\tau_i - \tau_i)$ in the numerator and so must be zero. In other words,

$$\ell_j(\tau_i) = \begin{cases} 1, & j = i, \\ 0, & j \neq i. \end{cases}$$

To show that the ℓ_j are linearly independent, suppose that for some scalars $c_1, \ldots, c_n$,

$$\sum_{j=1}^{n} c_j \ell_j(t) = 0, \quad \text{for all } t.$$

Since this holds for all t, if we put $t = \tau_i$, then all the terms except the one with $j = i$ are zero, and we get $c_i = 0$. Thus, the Lagrange fundamental interpolating polynomials are linearly independent. Furthermore, if

$$p(t) = \sum_{j=1}^{n} c_j \ell_j(t),$$

then $p(\tau_i) = c_i$. Hence, p is the unique polynomial of degree less than n that takes the values c_i at $t = \tau_i$ for $i = 1, \ldots, n$.

The following lemma illustrates a method for proving linear independence that can sometimes be easy to use. First, however, we need a definition. A scalar-valued function f defined on a vector space X is said to be a **linear functional** if for *all* scalars a and b and *all* vectors x and y, we have $f(ax + by) = af(x) + bf(y)$. Note

that a linear functional has the property that $f(0) = 0$; i.e., applying f to the zero vector always yields the zero scalar. To see this, write

$$f(0) = f(0+0) = f(0) + f(0).$$

where the second equality follows because f is linear. Now subtract $f(0)$ from the left- and right-hand sides to get $0 = f(0)$.

Lemma 2.5. *Let vectors $x_1, \ldots, x_n$ be given. If one can find linear functionals, say $f_1, \ldots, f_n$, such that $f_i(x_j) = 1$ for $j = i$ and $f_i(x_j) = 0$ for $j \neq i$, then $x_1, \ldots, x_n$ are linearly independent.*

Proof. Let $c_1, \ldots, c_n$ be arbitrary scalars, and suppose

$$\sum_{j=1}^{n} c_j x_j = 0. \tag{2.3}$$

Then for each i, we can write

$$\begin{aligned} 0 &= f_i(0), && \text{since } f_i \text{ is linear,} \\ &= f_i\Big(\sum_{j=1}^{n} c_j x_j\Big), && \text{by (2.3),} \\ &= \sum_{j=1}^{n} c_j f_i(x_j), && \text{by linearity,} \\ &= c_i, \end{aligned}$$

where the last step uses the hypothesis about f_i. □

Example 2.6 (Linear Independence of the Power Functions). Let $\mathbb{P}_n$ denote the set of all polynomials of degree less than n. Every such polynomial $x(t)$ is a linear combination of the powers, $1, t, t^2, \ldots, t^{n-1}$. To establish linear independence of the powers, it is more convenient to write the typical element $x \in \mathbb{P}_n$ in the form

$$x = \sum_{r=0}^{n-1} c_r x_r, \tag{2.4}$$

where $x_r(t) := t^r/r!$, $r = 0, \ldots, n-1$. Now consider the linear functionals f_q defined by $f_q(x) := x^{(q)}(0)$, where $x^{(q)}$ is the qth derivative of the polynomial x, with $x^{(0)} := x$. Observe that for $q \leq r$,

$$x_r^{(q)}(t) = \frac{t^{r-q}}{(r-q)!}.$$

Hence, $f_q(x_r) = x_r^{(q)}(0)$ equals 0 for $q < r$, and equals 1 for $q = r$. Since $x_r^{(r)}$ is the constant polynomial equal to 1, $x_r^{(q)} = 0$ for $q > r$, and so $f_q(x_r) = 0$ in this case. By Lemma 2.5, the power functions x_r are linearly independent.

Example 2.7. We can combine some of the ideas from Example 2.3 and use Lemma 2.5 to show that x and y are linearly independent. Specifically, with w denoting any linear combination of x and y, put

$$f_1(w) := \frac{\partial}{\partial t}[t^2 w(t)]\Big|_{t=1/2}$$
$$f_2(w) := \lim_{t\to 0}[t^2 w(t)].$$

Then $f_1(x) = 1$, $f_1(y) = 0$, $f_2(x) = 0$, and $f_2(y) = 1$.

Example 2.8. If we review Example 2.4, we can see Lemma 2.5 at work again. Use the linear functionals $f_i(x) := x(\tau_i)$, where x is a polynomial. Then $f_i(\ell_j) = \ell_j(\tau_i)$ has the required properties.

2.3. Subspaces

This section summarizes some basic terminology and results about subspaces. On a first reading, it may be helpful to focus on the examples and statements of results, and to skim over the derivations rather than get bogged down in technical details.

Let X be a vector space, and let W be a *nonempty* subset of X. If W has the property that for every pair of vectors $w_1, w_2 \in W$ and every pair of scalars c_1, c_2, the linear combination $c_1 w_1 + c_2 w_2 \in W$, then we say that W is a **subspace** of X. Notice that taking $c_1 = c_2 = 0$ shows that a subspace always contains the zero vector. We also point out two special cases: the set consisting of only the zero vector is a subspace (called the **zero subspace** or the **trivial subspace**), and the whole space X is a subspace.

Example 2.9. It is easily checked that subsets of the Euclidean plane of the form $W = \{(x,y) : ax + by = 0\}$ for constants a and b satisfy the definition of a subspace. Geometrically, these subspaces are straight lines passing through the origin. In three-dimensional space, subsets of the form $W = \{(x,y,z) : ax + by + cz = 0\}$ also satisfy the definition of a subspace. Geometrically, these subspaces are planes passing through the origin as shown in Figure 2.3.

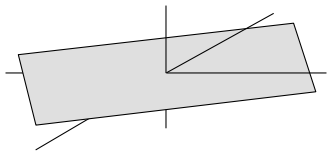


Figure 2.3. A plane in three-dimensional space.

Example 2.10. Let X denote the vector space of all complex-valued waveforms $x(t)$ defined for $-\infty < t < \infty$. Let W denote the subset of all causal waveforms. In symbols, $W = \{x \in X : x(t) = 0 \text{ for } t < 0\}$. Show that W is a subspace of X.

Solution. First note that W is nonempty since the zero waveform is causal. Next, fix any $w_1, w_2 \in W$ and any scalars c_1, c_2. We must show that $c_1 w_1 + c_2 w_2 \in W$. More explicitly, we must show that

$$(c_1 w_1 + c_2 w_2)(t) := c_1 w_1(t) + c_2 w_2(t)$$

is equal to zero for $t < 0$. Now, since $w_1, w_2 \in W$, for $t < 0$, $w_1(t) = w_2(t) = 0$, and we can write

$$(c_1 w_1 + c_2 w_2)(t) = c_1 \cdot 0 + c_2 \cdot 0 = 0, \quad t < 0.$$

This shows that $c_1 w_1 + c_2 w_2$ is causal and therefore in W.

Example 2.11 (The L^p Spaces). Let X denote the set of all real-valued or complex-valued waveforms defined on some time interval. If $1 \le p < \infty$, we say that $x \in X$ belongs to L^p if

$$\int |x(t)|^p \, dt < \infty,$$

where the integral is over the time interval under consideration. Show that L^p is a subspace of X.

Solution. We use the fact (shown below) that for real or complex numbers a and b,

$$|a+b|^p \le 2^p (|a|^p + |b|^p). \tag{2.5}$$

If w_1 and w_2 belong to L^p and c_1 and c_2 are real numbers, then

$$\begin{aligned} \int |c_1 w_1(t) + c_2 w_2(t)|^p \, dt &\le \int 2^p \big(|c_1 w_1(t)|^p + |c_2 w_2(t)|^p\big) \, dt \\ &= 2^p \left(|c_1|^p \int |w_1(t)|^p \, dt + |c_2|^p \int |w_2(t)|^p \, dt \right) \end{aligned}$$

is finite.

To establish (2.5), we start with the **triangle inequality**,[a] $|a+b| \le |a|+|b|$. If $|b| \le |a|$, then $|a+b| \le 2|a|$, from which it follows that

$$|a+b|^p \le 2^p |a|^p \le 2^p (|a|^p + |b|^p).$$

A similar argument works if $|a| < |b|$.

Closure under n-term Linear Combinations

A subspace has the property that every linear combination of one or more of its elements always lies in the subspace. We prove this by induction on the number n of vectors combined. (See Appendix B for background on mathematical induction.) By definition, the result is true for linear combinations of two vectors (or even one vector by setting $c_2 = 0$). Denote the subspace by W, and suppose the result is true from some $n \ge 2$. To show that $\sum_{k=1}^{n+1} c_k w_k \in W$ for scalars c_k and vectors $w_k \in W$, write

$$\sum_{k=1}^{n+1} c_k w_k = \sum_{k=1}^{n} c_k w_k + c_{n+1} w_{n+1}.$$

Now the sum on the right has only n terms and lies in W because we have assumed the result is true for linear combinations of n terms. Denote this sum by w_0. Then the left-hand side is equal to $w_0 + c_{n+1} w_{n+1}$, which lies in W since this is a linear combination of two elements of W.

Intersections of Subspaces Are Subspaces

Another important property of subspaces is that any intersection of subspaces is a subspace. To see this, for every α, let W_α be a subspace of X. Put

$$W := \bigcap_\alpha W_\alpha.$$

We must show that W is a subspace. Since each W_α is a subspace, the zero vector belongs to each one and to their intersection. Thus, W contains the zero vector. Given any $w_1, w_2 \in W$, this pair belongs to all the W_α, and so the linear combination $c_1 w_1 + c_2 w_2$ belongs to all the W_α and hence to their intersection; thus, the linear combination belongs to W.

[a] The triangle inequality is easy to verify for real numbers by checking the various combinations of signs for a and b. To establish the result for complex numbers, it is convenient to identify them with $\mathbb{R}^2$. Then the triangle inequality becomes a simple consequence of the Cauchy–Schwarz inequality; see Corollary 3.2.

Span

If G is a subset of X, then the **span** of G is defined to be the intersection of all subspaces of X that contain G. To see that this definition makes sense, observe that there is at least one subspace that contains G, namely the whole space X. Notice also that since the span is defined as the intersection of subspaces, we have from the preceding paragraph that the span is a subspace. The span of the empty set is easily seen to be the zero subspace.

In order to derive an alternative characterization of the span, we write the span of G symbolically as

$$\operatorname{span} G := \bigcap_{\substack{W: G \subset W \text{ and} \\ W \text{ is a subspace}}} W. \tag{2.6}$$

Proposition 2.12. *The span of a nonempty set is equal to the collection of all linear combinations of vectors in the set.*

Proof. Let G be a nonempty subset, and denote the set of all linear combinations of vectors in G by W_G. The first thing to check is that W_G is in fact a subspace. Consider two linear combinations from G, say $\sum_{k=1}^{n} \alpha_k g_k$ and $\sum_{k=1}^{n} \beta_k g_k$, where each $g_k \in G$. Then

$$c_1 \left(\sum_{k=1}^{n} \alpha_k g_k \right) + c_2 \left(\sum_{k=1}^{n} \beta_k g_k \right) = \sum_{k=1}^{n} (c_1 \alpha_k + c_2 \beta_k) g_k$$

is another element in W_G. Thus, W_G is a subspace. Since the one-term linear combination $1g = g \in G$, we see that W_G contains G. Since W_G is a subspace that contains G, W_G is one of the subspaces being intersected in (2.6). Using the identity $A \cap B \subset B$ for any sets A and B, it follows that

$$\bigcap_{\substack{W: G \subset W \text{ and} \\ W \text{ is a subspace}}} W \subset W_G. \tag{2.7}$$

To show the reverse containment, observe that if W is any subspace containing G, then W must contain linear combinations from G; i.e., $W_G \subset W$. In other words, W_G is a subset of every W in the intersection in (2.7). Hence,

$$W_G \subset \bigcap_{\substack{W: G \subset W \text{ and} \\ W \text{ is a subspace}}} W. \qquad \square$$

A subspace is said to be **finite dimensional** if it is equal to the span of a finite set. By Proposition 2.12 this means that if $W = \text{span}\{w_1, \ldots, w_n\}$ for some n, then every $w \in W$ can be expressed in the form

$$w = \sum_{i=1}^{n} c_i w_i$$

for some scalars c_i. The scalars corresponding to a given w are unique if and only if $w_1, \ldots, w_n$ are linearly independent, in which case we say that $w_1, \ldots, w_n$ constitute a **basis** for W; a basis is defined to be a spanning set that is linearly independent. Every finite-dimensional subspace W, except the zero subspace, has a basis,[4] and any two bases of a subspace have the same number of vectors;[5] this number is called the **dimension** of the subspace and is denoted by $\dim W$. If W is the zero subspace, $\dim W$ is taken to be zero. If W is a subspace of a finite-dimensional space X, then $\dim W \leq \dim X$, with equality if and only if $W = X$.[6] If a subspace W is such that there is no finite set whose span is equal to W, then W is said to be **infinite dimensional**.

Sums of Subspaces

If U and W are subspaces of X, then we define the **sum of subspaces**

$$U + W := \{u + v : u \in U \text{ and } w \in W\}.$$

It is easy to check that $U + W$ is a subspace (Problem 2.16). If the subspaces are such that every element of $U + W$ has a unique representation, then we say the sum is a **direct sum**, and we write $U \oplus W$. By a "unique representation," we mean that for $u, u' \in U$ and $w, w' \in W$,

$$u + w = u' + w' \text{ implies } u = u' \text{ and } w = w'.$$

Linear Varieties

If $x \in X$ and W is a subspace of X, we define

$$x + W := \{x + w : w \in W\}.$$

Such a set is called a **linear variety** or a **translated subspace**. For example, if we move the plane in Figure 2.3 down so that it no longer passes through the origin, we get the translated subspace shown in Figure 2.4.

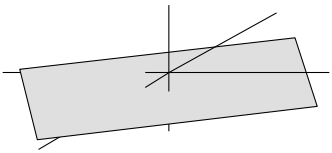


Figure 2.4. A translated subspace.

Example 2.13. The set of causal waveforms W is a subspace of the set of all waveforms X (Example 2.10). Put

$$x_1(t) := \begin{cases} 1, & t < 0, \\ 0, & t \geq 0, \end{cases}$$

and consider the translated subspace $x_1 + W$. This linear variety consists of all waveforms x such that $x(t) = 1$ for $t < 0$.

Although a given translated subspace can be represented in different ways, the next proposition shows that the subspace part is unique.

Proposition 2.14 (Subspace Uniqueness). *Suppose $x+W = y+U$, where U and W are subspaces and $x, y \in X$. Then $W = U$.*

Proof. First note that $x+0 \in x+W = y+U$ implies $x = y+u_0$ for some $u_0 \in U$. Now fix any $w \in W$. Then $x+w \in x+W = y+U$ implies $x+w = y+u$ for some $u \in U$. Replacing x with $y+u_0$ shows that $y+u_0+w = y+u$, which implies $w = u-u_0 \in U$. Hence $W \subset U$. By a similar argument, $U \subset W$, and it follows that $W = U$ as claimed. □

A linear variety $x+W$ is said to be finite dimensional if W is finite dimensional. In this case, the dimension of $x+W$ is defined to be the dimension of W. Otherwise, the linear variety is said to be infinite dimensional.

Hyperplanes §

Let X be a vector space of dimension n, and let $W \subset X$ be a subspace of dimension $n-1$. Now suppose V is a subspace with $W \subset V \subset X$. If $\dim V = n-1$, then $V = W$, while if $\dim V = n$, then $V = X$. In other words, there are no subspaces "between"

§This material is not used in the sequel.

an $(n-1)$-dimensional subspace and the whole space X having dimension n. In this sense, W is a ***maximal* proper subspace**. (Recall that if $A \subset B$, we say that A is a ***proper* subset** if $A \neq B$.)

Even in a vector space X that is not finite dimensional, we can define the notion of a maximal proper subspace. We say W is a **maximal proper subspace** if W is a proper subspace of X with the property that whenever V is a subspace satisfying $W \subset V \subset X$, we must have either $V = W$ or $V = X$; i.e., there is no subspace "between" W and the whole space X.

A subset H of a vector space X is called a **hyperplane** if H is the translation of a maximal proper subspace. In other words, H is a hyperplane if $H = x_0 + W$ for some vector x_0 and some maximal proper subspace W.

It can be shown[7] that *every* maximal proper subspace is of the form $\{x \in X : f(x) = 0\}$ for some nonzero linear functional f. It is then easy to verify that every hyperplane is of the form $\{x \in X : f(x) = c\}$ for some nonzero linear functional f and some scalar c. Sets of the form $\{x \in X : f(x) \leq c\}$ and $\{x \in X : f(x) \geq c\}$ are called **half-spaces**.

2.4. Affine Sets

This section summarizes some basic terminology and results about affine sets. On a first reading, it may be helpful to focus on the examples and statements of results, and to skim over the derivations rather than get bogged down in technical details.

Let X be a vector space, and let A be a subset of X. If A has the property that for every pair of vectors $a_1, a_2 \in A$ and every pair of scalars c_1, c_2 with $c_1 + c_2 = 1$, the linear combination $c_1 a_1 + c_2 a_2 \in A$, then we say A is **affine**.[b] It is easy to see that every translated subspace is affine. Let W be a subspace, and consider the translated subspace $x + W$; for $c_1 + c_2 = 1$ and $w_1, w_2 \in W$, we can write

$$c_1(x + w_1) + c_2(x + w_2) = \underbrace{(c_1 + c_2)}_{1} x + \underbrace{(c_1 w_1 + c_2 w_2)}_{\in W},$$

which is an element of $x + W$. In particular, every subspace, including the whole space X, is affine. At the end of this section, we show that every nonempty affine set is a translated subspace; hence, a nonempty affine set is said to be finite dimensional or infinite dimensional according to whether it is the translation of a finite or infinite dimensional subspace; the dimension of a nonempty affine set is defined to be the dimension of the subspace it is the translation of.

[b] This condition can also be expressed as $\lambda a_1 + (1-\lambda) a_2 \in A$ for all scalars λ.

Closure under n-term Affine Combinations

A linear combination of vectors whose coefficients sum to one is called an **affine combination**. An affine set has the property that every affine combination of one or more of its elements always lies in the set. We prove this by induction on the number n of vectors combined. By definition, the result is true for affine combinations of two vectors (or even one vector by setting $c_1 = 1$ and $c_2 = 0$). Denote the affine set by A, and suppose the result is true from some $n \geq 2$. To show that $\sum_{k=1}^{n+1} c_k a_k \in A$ for scalars c_k with $\sum_{k=1}^{n+1} c_k = 1$ and vectors $a_k \in A$, we proceed as follows. First, note that since $n \geq 2$, at least one of the $c_k \neq 1$ (otherwise $\sum_{k=1}^{n+1} c_k = n+1 > 1$). Suppose $c_i \neq 1$. Write

$$\sum_{k=1}^{n+1} c_k a_k = (1-c_i) \sum_{k \neq i} \frac{c_k}{1-c_i} a_k + c_i a_i.$$

The linear combination on the right has only n terms, and since its coefficients sum to one, the combination is affine; hence, it lies in A because we have assumed the result is true for n terms. Denote this affine combination by a_0. Then the left-hand side is equal to $(1-c_i)a_0 + c_i a_i$, which lies in A since this is an affine combination of two elements of A.

Intersections of Affine Sets Are Affine

Just as we showed any intersection of subspaces is a subspace, it can be shown that any intersection of affine sets is affine (Problem 2.19).

The Affine Hull

If G is a subset of X, then the **affine hull** of G is defined to be the intersection of all affine sets that contain G. To see that this definition makes sense, observe that there is at least one affine set that contains G, namely the whole space X. Notice also that since the affine hull is defined as the intersection of affine sets, we have from the preceding paragraph that the affine hull is an affine set. The affine hull of the empty set is easily seen to be the empty set.

In order to derive an alternative characterization of the affine hull, we write the affine hull of G symbolically as

$$\operatorname{aff} G := \bigcap_{\substack{A: G \subset A \text{ and} \\ A \text{ is affine}}} A.$$

Proposition 2.15. *The affine hull of a nonempty set is equal to the collection of all affine combinations of vectors in the set.*

Proof. Imitate the proof of Proposition 2.12 (Problem 2.20). □

The Only Nonempty Affine Sets are Translated Subspaces

We show that a nonempty affine set is a translated subspace. Let A be any nonempty affine set. Fix any $a_0 \in A$, and put $W := \{a - a_0 : a \in A\}$. It is easy to show that W is a subspace (Problem 2.22). Hence, given any $a \in A$, $a - a_0 \in W$ and so $a - a_0 = w$ for some $w \in W$. In other words, $a = a_0 + w$, which says that $a \in a_0 + W$. Conversely, given any $x \in a_0 + W$, $x = a_0 + w$ for some $w \in W$. However, from the definition of W, $w = a - a_0$ for some $a \in A$. Hence, $x = a_0 + (a - a_0) = a \in A$. We conclude that $A = a_0 + W$, where W is a subspace.

2.5. Convex Sets

Geometrically, a convex set is one for which the line segment joining any two points in the set lies entirely in the set. Figure 2.5 shows a convex set on the left and a nonconvex set on the right. Here is a more precise mathematical definition. Let X

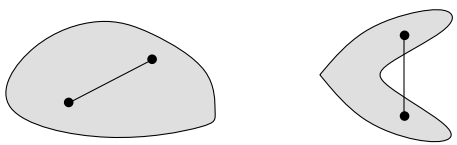


Figure 2.5. A convex set (left) and a nonconvex set (right).

be a vector space, and let C be a subset of X. If C has the property that for every pair of vectors $x_1, x_2 \in C$ and every pair of nonnegative scalars c_1, c_2 with $c_1 + c_2 = 1$, the linear combination $c_1 x_1 + c_2 x_2 \in C$, then we say C is **convex**.[c] It is easy to see that every affine set, and hence every subspace, is convex. Just as the empty set is affine, the empty set is convex.

Example 2.16. Again let X denote the set of all complex-valued waveforms defined on $(-\infty, \infty)$ as in Example 2.10. Let C denote the subset of X consisting of real-valued waveforms x with $x(t) \geq 0$ for $|t| \leq 1$. Show that C is convex.

Solution. First note that C is nonempty since the zero waveform belongs to C. Next, fix any $x_1, x_2 \in C$ and any $0 \leq \lambda \leq 1$. We must show that $\lambda x_1 + (1 - \lambda) x_2 \in C$. Fix t with $|t| \leq 1$, and use the fact that $x_1, x_2 \in C$ along with the fact that λ and $1 - \lambda$ are nonnegative to write

$$\lambda x_1(t) + (1 - \lambda) x_2(t) \geq \lambda \cdot 0 + (1 - \lambda) \cdot 0 = 0.$$

[c] This condition can also be expressed as $\lambda x_1 + (1 - \lambda) x_2 \in C$ for all $0 \leq \lambda \leq 1$.

This shows that $\lambda x_1 + (1-\lambda)x_2$ is nonnegative for $|t| \leq 1$ and therefore in C.

Closure under n-term Convex Combinations

A linear combination of vectors whose coefficients are nonnegative and sum to one is called an **convex combination**. A convex set has the property that every convex combination of one or more of its elements always lies in the set. This can be shown by trivial modification of the induction proof of the analogous result for affine sets (Problem 2.23).

Intersections of Convex Sets Are Convex

Just as we showed any intersection of subspaces is a subspace, it can be shown that any intersection of convex sets is convex (Problem 2.24).

The Convex Hull

If G is a subset of X, then the **convex hull** of G is defined to be the intersection of all convex sets that contain G. To see that this definition makes sense, observe that there is at least one convex set that contains G, namely the whole space X. Notice also that since the convex hull is defined as the intersection of convex sets, we have from the preceding paragraph that the convex hull is a convex set. The convex hull of the empty set is easily seen to be the empty set. The convex hull of a finite set is called a (convex) **polytope**.

In order to derive an alternative characterization of the convex hull, we write the convex hull of G symbolically as

$$\operatorname{co} G := \bigcap_{\substack{C: G \subset C \text{ and} \\ C \text{ is convex}}} C.$$

Proposition 2.17. *The convex hull of a nonempty set is equal to the collection of all convex combinations of vectors in the set.*

Proof. Imitate the proof of Proposition 2.12 (Problem 2.25). □

Carathéodory's Theorem §

Carathéodory has a number of theorems named after him. We are concerned with the one saying that if $C \subset \mathbb{R}^n$ is the convex hull of finitely many vectors (i.e., C is a polytope), then there is a subset of at most $n+1$ of these vectors such that C is the convex hull of the subset.

Carathéodory's Theorem is frequently used to simplify the characterization of the capacity region of multiuser channels in the study of information theory.

Theorem 2.18 (Carathéodory). *Let $x_1, \ldots, x_p$ be vectors in $\mathbb{R}^n$. Suppose $x = \sum_{i=1}^p \lambda_i x_i$, where $\lambda_i \geq 0$ and $\sum_{i=1}^p \lambda_i = 1$. If $p \geq n+1$, then there exist $\{\widetilde{x}_1, \ldots, \widetilde{x}_{n+1}\} \subset \{x_1, \ldots, x_p\}$, and there exist $\mu_i \geq 0$, $\sum_{i=1}^{n+1} \mu_i = 1$ with $x = \sum_{i=1}^{n+1} \mu_i \widetilde{x}_i$.*

Proof. The proof is based on [42, p. 41]. The proof is by induction on p. The case $p = n+1$ is trivial. Suppose $p > n+1$. Then the set $\{x_1 - x_p, \ldots, x_{p-1} - x_p\}$ contains more than n vectors, and is therefore linearly dependent. Hence, there exist scalars $c_1, \ldots, c_{p-1}$ not all zero such that

$$\sum_{i=1}^{p-1} c_i (x_i - x_p) \;=\; 0 \;=\; \sum_{i=1}^{p-1} c_i x_i + \left(- \sum_{i=1}^{p-1} c_i \right) x_p.$$

Put $a_i := c_i$ for $i = 1, \ldots, p-1$, and put $a_p := -\sum_{i=1}^{p-1} c_i$. Then

$$\sum_{i=1}^{p} a_i x_i \;=\; 0 \qquad \text{and} \qquad \sum_{i=1}^{p} a_i \;=\; 0.$$

Note also that since the c_i are not all zero, at least one of the a_i must be positive, and at least one of the a_i must be negative. Next observe that for any $\rho \in \mathbb{R}$,

$$\sum_{i=1}^{p} (\lambda_i - \rho a_i) x_i \;=\; x.$$

Put $\mu_i := \lambda_i - \rho a_i$. Note that $\sum_{i=1}^p \mu_i = \sum_{i=1}^p \lambda_i = 1$. Our goal is to choose ρ so that $\mu_i \geq 0$ *and* such that for some j, $\mu_j = 0$. If we can do this, then

$$x \;=\; \sum_{i=1}^{p} \mu_i x_i \;=\; \sum_{i=1, i \neq j}^{p} \mu_i x_i,$$

and we can apply the induction hypothesis to $\{x_i\}_{i=1, i\neq j}^p$, which contains $p-1$ vectors.

We claim that setting $\rho := \max_{a_i < 0} \lambda_i / a_i$ works. Since $\rho \leq 0$, for $a_i \geq 0$, $\mu_i = \lambda_i - \rho a_i \geq 0$. For $a_i < 0$, $\rho \geq \lambda_i / a_i$, or equivalently, $\rho a_i \leq \lambda_i$. Hence, for $a_i < 0$, we also have $\mu_i = \lambda_i - \rho a_i \geq 0$. Finally, since the maximum defining ρ is achieved, for some j, $\rho = \lambda_j / a_j$, and $\mu_j = \lambda_j - \rho a_j = 0$. □

§This material is not used in the sequel.

Extreme Points and Linear Programming §

A point x in a convex set C is said to be an **extreme point** of C if it is impossible to write x as a nontrivial convex combination of two distinct points of C; more precisely, it is not possible to write $x = \lambda y + (1-\lambda)z$ for $y, z \in C$, $0 < \lambda < 1$, and $y \neq z$. Equivalently, x is an extreme point of C if whenever $x = \lambda y + (1-\lambda)z$ for some $y, z \in C$ and $0 < \lambda < 1$, we must have $y = z$.

The standard form of a **linear programming** problem [28] is[d]

$$\min_{x \in \mathbb{R}^n} c^{\mathsf{T}} x \quad \text{subject to } Ax = b \text{ and } x \geq 0,$$

where A is a given $m \times n$ matrix of real numbers and $b \in \mathbb{R}^m$ and $c \in \mathbb{R}^n$ are given vectors. We call

$$F := \{x \in \mathbb{R}^n : Ax = b \text{ and } x \geq 0\}$$

the set of **feasible vectors**. It is easy to check that the set of feasible vectors is convex. Hence, the linear programming problem is to minimize the linear functional $c^{\mathsf{T}} x$ over the convex set of feasible vectors.

Let $a_1, \ldots, a_n$ denote the columns of A. Each $a_i \in \mathbb{R}^m$. For $x \in \mathbb{R}^n$, let $I(x) := \{i : x_i > 0\}$, and note that for $x \in F$, if $i \notin I(x)$, then $x_i = 0$.

Proposition 2.19. *Let $x \in F$. Then x is an extreme point of F if and only if $\{a_i : i \in I(x)\}$ are linearly independent.*

In other words, extreme points of F are those nonnegative solutions of $Ax = b$ with the property that the strictly positive components of x correspond to linearly independent columns of A.

The importance of the proposition is that the extreme points of F can be found by looking at linearly independent subsets of the columns of A, and there are only finitely many such subsets. In fact, we may restrict attention to linearly independent subsets whose cardinality is equal to the dimension of the subspace spanned by the columns of A (we will see in Section 4.2 that this is the **rank** of A; i.e., the dimension of the **range** of A).

Proof of Proposition 2.19. Suppose $x \in F$ and $\{a_i : i \in I(x)\}$ are linearly independent. We must show that x is an extreme point of F. So we suppose $x = \lambda y + (1-\lambda)z$ for some $y, z \in F$ and $0 < \lambda < 1$, and we must show that $y = z$. To begin, observe that $y_i = z_i = 0$ for $i \notin I(x)$. This follows because $0 < \lambda < 1$ and because $x, y, z \in F$ implies they have nonnegative components. Since we now have $x_i = y_i = z_i = 0$ for $i \notin I(x)$, and since $Ax = Ay = Az = b$, we can write

$$\sum_{i \in I(x)} x_i a_i = \sum_{i \in I(x)} y_i a_i = \sum_{i \in I(x)} z_i a_i.$$

§ This material is not used in the sequel.

[d] The superscript T denotes the **transpose**.

Since $\{a_i : i \in I(x)\}$ is linearly independent, we must have $x_i = y_i = z_i$ for $i \in I(x)$. Thus, $x = y = z$.

We now prove that if x is an extreme point of F, then $\{a_i : i \in I(x)\}$ is linearly independent. Suppose otherwise that $\{a_i : i \in I(x)\}$ is linearly dependent. Since $Ax = b$ with $x \geq 0$, we can write

$$\sum_{i \in I(x)} x_i a_i = b.$$

Linear dependence implies there exist $\{y_i \in \mathbb{R} : i \in I(x)\}$, not all zero, with

$$\sum_{i \in I(x)} y_i a_i = 0.$$

Taking $y_i := 0$ for $i \notin I(x)$ and then $y := [y_1, \ldots, y_n]^\mathsf{T}$ yields a y with $Ay = 0$. Thus, for all $\varepsilon > 0$, $A(x \pm \varepsilon y) = b$. Furthermore, for small $\varepsilon > 0$, we have $x \pm \varepsilon y \geq 0$. Writing $x = (x + \varepsilon y)/2 + (x - \varepsilon y)/2$ expresses x as the nontrivial convex combination of two distinct points from F (the points are different since at least one y_i is nonzero). □

Notes

Note 2.1. Here are the precise additivity properties of a vector space X.

- (i) **closure:** For all $x, y \in X$, $x + y \in X$.
- (ii) **commutative law:** For all $x, y \in X$, $x + y = y + x$.
- (iii) **associative law:** For all $x, y, z \in X$, $(x + y) + z = x + (y + z)$.
- (iv) **additive identity:** There exists an element of X, denoted by 0, such that for all $x \in X$, $x + 0 = x$.
- (v) **additive inverse:** For every $x \in X$, there exists a unique element of X, denoted by $-x$, such that $x + (-x) = 0$.

We can summarize these properties informally as follows. For a set X to be a vector space, the sum of two elements of X must be an element of X. The order in which elements are added does not affect the result. There is a zero element, and every element can be negated.

Note 2.2. Scalar multiplication has the following five properties.

- (i) **closure:** For each scalar a and vector $x \in X$, $ax \in X$.
- (ii) **associative law:** For all scalars a and b and any vector $x \in X$, $a(bx) = (ab)x$.
- (iii) **first distributive law:** For any scalar a and any two vectors $x, y \in X$, $a(x + y) = ax + ay$.

(*iv*) **second distributive law:** For any two scalars a and b and any vector $x \in X$, $(a+b)x = ax + bx$.

(*v*) For the scalar 1 and any vector $x \in X$, $1x = x$.

Note 2.3. Here is the MATLAB function `lincmb`.

```
function [y,xmat] = lincmb(t,c,xfun,m,varargin)
%
% Usage:  lincmb(t,c,xfun,m)
%     or  lincmb(t,c,xfun,m,s)
%
%     where if the optional argument s is
%     omitted, it is taken to be 1.
%
% Compute
%
%      y(t_i) = sum_j c_j xfun( s_j*(t_i-m_j) )
%
%   and
%
%      xmat = matrix with ij element xfun( s_j*(t_i-m_j) ).
%
% If you want to plot the functions
% xfun(s_j*(t-m_j)) themselves instead of
% their linear combination, you can plot(t,xmat)
%
% t = ARRAY of times at which the linear combination will
%     be evaluated.
% c = ROW vector of coefficients of the linear combination.
% xfun = either a STRING containing the name of the
%        underlying function to be used or the handle of an
%        anonymous function.
% m = ROW vector of shifts.
% s = OPTIONAL ROW vector of scale factors.
%     WARNING: If length(s)>1, then s and m are assumed to
%     have the same length; otherwise an ERROR may result.
%
% y = output (same dimension(s) as t)
% xmat = matrix (see below).
%
% The key idea is to observe that the formula
%
%    y(t_i) = sum_j c_j xfun(t_i-m_j)
%
```

```
% can be viewed as matrix-vector multiplication if we think
% of c and y as column vectors.
%
if nargin==4
   % Create matrix xfun(t_i-m_j)
   xmat = feval(xfun,bsxfun(@minus,t(:),m));
else % Assume nargin = 5
   % Create matrix xfun(s_j*(t_i-m_j))
   s = varargin{1};
   xmat=feval(xfun,bsxfun(@times,s,bsxfun(@minus,t(:),m)));
end
y = reshape(xmat*(c.'),size(t));
```

Note 2.4. To see that every finite-dimensional subspace, other than the zero subspace, has a basis, we start with the fact that by definition, to say that a subspace W is finite dimensional means that $W = \text{span}\{w_1, \ldots, w_n\}$ for some finite n. If each w_i is the zero vector, then W is the zero subspace and does not have a basis. If $w_1, \ldots, w_n$ are linearly independent, they form a basis. Otherwise, there exist scalars $c_1, \ldots, c_n$, not all zero, with

$$\sum_{i=1}^{n} c_i w_i = 0.$$

Suppose some $c_j \neq 0$. Then

$$w_j = -\frac{1}{c_j} \sum_{i \neq j} c_i w_i.$$

Using this formula, we can convert every linear combination of $w_1, \ldots, w_n$ into a linear combination of $\{w_i : i \neq j\}$. We have thus converted a spanning set of n vectors into a spanning set of $n-1$ vectors. We continue in this way until we are left with a spanning set that is linearly independent (the desired basis).

Note 2.5. Before showing that any two bases of a finite-dimensional subspace have the same number of vectors, we need the following result.

Lemma. *Let U and V be finite subsets of a vector space, and assume that $V \subset \text{span}\, U$. If V is linearly independent, then the number of vectors in V cannot exceed the number of vectors in U.*

Proof. Suppose otherwise that U contains n vectors and that V contains $m > n$ vectors. Pick any $v_1 \in V$. Since $v_1 \in \text{span}\, U$, v_1 can be expressed as a linear combination of elements of U in which at least one coefficient is not zero. Hence, the corresponding element of U can be expressed in terms of v_1 and the remaining elements of U. Let U_1 denote these remaining elements along with v_1. Then U_1 contains

n vectors, and $\operatorname{span} U_1 = \operatorname{span} U$. Now pick $v_2 \in V$. Then v_2 can be expressed as a linear combination of elements of U_1 in which at least one of the coefficients of one of the vectors other than v_1 is nonzero (otherwise v_2 is proportional to v_1). Remove the corresponding vector, and let U_2 denote the remaining vectors along with v_2. Thus, U_2 contains v_1, v_2, and $n-2$ elements of U and $\operatorname{span} U_2 = \operatorname{span} U$. Continuing in this way, after a total of n steps, $U_n = \{v_1, \ldots, v_n\}$ and $\operatorname{span} U_n = \operatorname{span} U$. Hence, the $m-n$ vectors in V that are not in U_n can be expressed in terms of $v_1, \ldots, v_n$, contradicting the assumption that V is linearly independent. □

It is now easy to see that any two bases of a finite-dimensional space must have the same number of elements. Suppose U and V are two different bases for the space. Then V is a linearly independent subset of $\operatorname{span} U$, and U is a linearly independent subset of $\operatorname{span} V$. Two applications of the lemma show that U have V must have the same number of elements.

Note 2.6. We show that if $\dim X < \infty$ and W is a subspace of X, then $\dim W \leq \dim X$. Furthermore, if $\dim W = \dim X$, then $W = X$. Let $n := \dim X$. If W is the zero subspace, the result is obviously true. If W is not the zero subspace, then there is some nonzero $v_1 \in W$. The set $V_1 := \{v_1\}$ is linearly independent. If $\operatorname{span} V_1 = W$, we have a basis for W. Otherwise, there is a $v_2 \in W$ such that $V_2 = \{v_1, v_2\}$ is linearly independent. If for some $k < n$, this procedure stops with $\operatorname{span} V_k = W$, we are finished. If the procedure continues until we obtain $\operatorname{span} V_n = W$, we still have to show that $W = X$. If $W \neq X$, there is an $x \in X$ that does not belong to $W = \operatorname{span} V_n$. But then $\{v_1, \ldots, v_n, x\}$ is a linearly independent subset of X that contains more vectors than those in the basis for X, contradicting the lemma of the above note.

Note 2.7. The following two lemmas show that every maximal proper subspace of a vector space X is of the form $\{x \in X : f(x) = 0\}$ for some nonzero linear functional f.

Lemma. *If f is a nonzero linear functional on a vector space X, then $\{x \in X : f(x) = 0\}$ is a maximal proper subspace of X.*

Proof. Put $W := \{x \in X : f(x) = 0\}$. It is easy to check that W is a subspace. Furthermore, W is a proper subspace because f is nonzero; i.e., since there is an $x_1 \in X$ with $f(x_1) \neq 0$, $x_1 \notin W$. It remains to show that W is maximal. Let V be a subspace with $W \subset V \subset X$. We must show that $V \neq W$ implies $V = X$.[e] Actually, since $V \subset X$, it suffices to prove $X \subset V$. So suppose $V \neq W$. Then there is an $x_0 \in V$ with $x_0 \notin W$. This implies $f(x_0) \neq 0$; without loss of generality, we may assume $f(x_0) = 1$. To prove $X \subset V$, take any $x \in X$, and observe that $f\big(x - f(x)x_0\big) =$

[e] We are making use of Example A.6 in the Appendix.

0. Hence, $x - f(x)x_0 \in W$, and we may write $x - f(x)x_0 = w$ for some $w \in W$. Rearranging, we have $x = w + f(x)x_0$, which implies $x \in V$ because both $w \in W \subset V$ and $x_0 \in V$. □

Lemma. *If W is a maximal proper subspace of a vector space X, then there exists a nonzero linear functional f such that $W = \{x \in X : f(x) = 0\}$.*

Proof. Let W be a maximal proper subspace of a vector space X. Since W is proper, there is some $x_0 \notin W$. Put $V := \text{span}(x_0 + W)$. The reader should verify that $x_0 \notin W$ together with maximality of W forces $V = X$. It now follows that every $x \in X$ can be written in the form $x = ax_0 + w$ for some unique scalar a and vector $w \in W$ (the reader should verify the uniqueness of a and w). Hence, the formula $f(x) := a$ is well defined and is a nonzero linear functional. To conclude the proof, observe that $\{x \in X : f(x) = 0\} = W$. □

Problems

1. **MATLAB.** Use `lincmb` to plot

$$y(t) = v(t) - v((t-1)/2) + v((t-2)/4), \quad 0 \le t \le 7,$$

if

$$v(t) := \begin{cases} t(1-t), & 0 \le t \le 1, \\ 0, & \text{otherwise.} \end{cases}$$

Also plot the scaled and shifted pulses. Show your code.

2. Let $x(t) = 1$ for $0 \le t \le 1$, and $x(t) = 0$ otherwise. Let $y(t) = 1$ for $0 \le t \le 2$, and $y(t) = 0$ otherwise. Show that x and y are linearly independent on $(-\infty, \infty)$.

3. Let x and y be as in the preceding problem, and let $z(t) = t$ for $0 \le t \le 2$, and $z(t) = 0$ otherwise. Consider the task of showing that x, y, and z are linearly independent. List all of the reasons that the following attempted solution is not correct.

   Suppose $ax(t) + by(t) + cz(t) = 0$. For $0 \le t \le 1$, we have

$$a + b + ct = 0,$$

   while for $1 < t \le 2$, we have

$$b + ct = 0.$$

   Subtracting the second equation from the first yields $a = 0$. In the second equation, taking $t = 0$ shows that $b = 0$; this leaves $ct = 0$. Taking $t = 1$ here yields $c = 0$ as well.

4. For $1 < t < 2$, put $x(t) := t$, $y(t) := t^2$, and $z(t) := e^t$. Determine whether or not they are linearly independent.

5. We showed in Example 2.6 that the power functions $1, t, \ldots, t^{n-1}$ are linearly independent. For $k = 0, \ldots, n-1$, let $x_k(t)$ be a polynomial of degree k. Show that the n polynomials x_k are linearly independent. *Hint:* Use the fact that in the polynomial $x_k(t)$, the coefficient of t^k is nonzero.

6. Let $p_1(t), \ldots, p_n(t)$ be linearly independent waveforms for $-\infty < t < \infty$. Suppose that the $p_k(t)$ are differentiable. Are the derivatives $p_k'(t)$ linearly independent? If "yes," prove it; if "no," give a counterexample of linearly independent p_k for which the p_k' are linearly dependent.

7. Given real numbers $\beta_1 < \cdots < \beta_n$, show that the functions $x_k(t) = e^{\beta_k t}$ defined on $[0, \infty)$ are linearly independent. *Hint:* Suppose that for some coefficients $c_1, \ldots, c_n$,

$$\sum_{k=1}^{n} c_k e^{\beta_k t} = 0, \quad \text{for all } t \geq 0.$$

Multiply the above equation by $e^{-\beta_n t}$ and then take the limit as $t \to \infty$.

8. Solve the preceding problem if the $x_k(t)$ are defined only on a finite interval (a, b), and if the β_k are allowed to be distinct, possibly complex, numbers.

9. A certain communication system transmits linearly independent waveforms x_k, $k = 1, \ldots, n$, over a linear, time-invariant channel with impulse response h. In the absence of noise, the receiver sees the corresponding waveforms $y_1, \ldots, y_n$, where

$$y_k(t) = \int_{-\infty}^{\infty} h(t - \tau) x_k(\tau)\, d\tau.$$

If $h(t) = e^{-t} u(t)$, where u is the unit-step function, determine whether or not the y_k are linearly independent.

10. Let $T_1, \ldots, T_n$ be distinct real numbers, and put $x_k(t) = \operatorname{sinc}(t - T_k)$, where

$$\operatorname{sinc}(t) := \begin{cases} \dfrac{\sin(\pi t)}{\pi t}, & t \neq 0, \\ 1, & t = 0. \end{cases}$$

Show that these functions are linearly independent on $(-\infty, \infty)$. *Hint:* The Fourier transform of $h(t) := \operatorname{sinc}(t)$ is

$$H(f) = \begin{cases} 1, & |f| \leq 1/2, \\ 0, & |f| > 1/2. \end{cases}$$

11. Let $\beta_1, \ldots, \beta_n$ be distinct, possibly complex numbers. Consider the waveforms $x_1, \ldots, x_n$ on the interval $[0, 2\pi]$ defined by

$$x_k(t) := \int_0^t \sin(\tau) e^{\beta_k \tau} \, d\tau, \quad 0 \le t \le 2\pi.$$

Determine whether or not $x_1, \ldots, x_n$ are linearly independent.

12. Determine whether or not the signals

$$x_k(t) := \int_{-\infty}^{t/k} e^{-|\tau|} \, d\tau, \quad -\infty < t < \infty, \quad k = 1, \ldots, n,$$

are linearly independent.

13. ***The ℓ^p Spaces.*** Let X denote the set of all infinite sequences of the form $x = (x_1, x_2, \ldots)$. If $1 \le p < \infty$, we say that $x \in X$ belongs to ℓ^p if

$$\sum_{k=1}^{\infty} |x_k|^p < \infty.$$

Show that ℓ^p is a subspace of X.

14. This problem addresses a subtlety in the proof of Proposition 2.12. Let G be a subset of a vector space X, and suppose G contains infinitely many vectors. Consider two linear combinations from G, say

$$\sum_{i=1}^{p} a_i x_i \quad \text{and} \quad \sum_{j=1}^{q} b_j y_j,$$

where a_i and b_j are scalars and x_i and y_j are vectors in G. Let $n := p + q$. Specify scalars $\alpha_1, \ldots, \alpha_n$ and $\beta_1, \ldots, \beta_n$ and specify vectors $g_1, \ldots, g_n$ in G such that

$$\sum_{i=1}^{p} a_i x_i = \sum_{k=1}^{n} \alpha_k g_k \quad \text{and} \quad \sum_{j=1}^{q} b_j y_j = \sum_{k=1}^{n} \beta_k g_k.$$

15. Proposition 2.12 applies only to nonempty sets. What is the span of the empty set?

16. If U and W are subspaces, show that $U + W := \{u + w : u \in U \text{ and } w \in W\}$ is a subspace.

17. Show that the sum of two subspaces, say $U + W$, is a direct sum if and only if their intersection $U \cap W$ is the zero subspace.

18. If a finite-dimensional vector space X can be written as the direct sum of two subspaces, say $X = U \oplus W$, show that $\dim X = \dim U + \dim W$.

19. Show that any intersection of affine sets is an affine set.

20. Show that the affine hull of a nonempty set is equal to the collection of all affine combinations of vectors in the set.

21. Show that $\operatorname{aff}\{a_0, \ldots, a_m\} = a_0 + \operatorname{span}\{a_1 - a_0, \ldots, a_m - a_0\}$.

22. Let A be a nonempty affine set, and fix any $a_0 \in A$. Show that $W := \{a - a_0 : a \in A\}$ is a subspace.

23. Let C be a convex set, and let $x_1, \ldots, x_n$ belong to C. If $c_1, \ldots, c_n$ are nonnegative and sum to one, show that $\sum_{k=1}^{n} c_k x_k \in C$.

24. Show that any intersection of convex sets is a convex set.

25. Show that the convex hull of a nonempty set is equal to the collection of all convex combinations of vectors in the set.

26. Recall that an $n \times n$ matrix is **stochastic** if its entries are nonnegative and the sum of each row is equal to one. Let S denote the set of $n \times n$ stochastic matrices. Determine whether or not S is convex.

CHAPTER 3
Inner-Product Spaces

You are probably familiar with the standard **dot product** on $\mathbb{R}^d$. Recall that if[a] $x = [x_1, \ldots, x_d]^\mathsf{T}$ and $y = [y_1, \ldots, y_d]^\mathsf{T}$, then the dot product of x and y is $\sum_{k=1}^d x_k y_k$. For $x, y \in \mathbb{C}^d$, the analogous quantity is

$$\sum_{k=1}^{d} x_k \overline{y_k},$$

where the overbar denotes the **complex conjugate**. For real-valued waveforms, the obvious generalization is

$$\int x(t) y(t)\, dt,$$

and for complex-valued waveforms it is[1]

$$\int x(t) \overline{y(t)}\, dt.$$

The properties of the foregoing expressions suggest the following formal definition. We say that $\langle \cdot, \cdot \rangle$ is an **inner product** on a vector space X if $\langle x, y \rangle$ is a scalar-valued function of $x, y \in X$ such that the following three conditions hold.

- (i) For all $x \in X$, $0 \leq \langle x, x \rangle < \infty$ and $\langle x, x \rangle = 0$ if and only if x is the zero vector.
- (ii) For all $x, y \in X$, $\langle x, y \rangle = \overline{\langle y, x \rangle}$.
- (iii) For all $x, y, z \in X$, and all scalars a and b, $\langle ax + by, z \rangle = a\langle x, z \rangle + b\langle y, z \rangle$.

In a real vector space, the complex conjugate in (ii) is omitted. In a complex vector space $\langle x, y \rangle$ may be a complex number, but $\langle x, x \rangle$ is always a real number and nonnegative as well.

Property (iii) says that the inner product as a function of its left-hand argument is linear; hence, for fixed y, $f(x) := \langle x, y \rangle$ defines a linear functional on X. Since we showed earlier that for a linear functional $f(0) = 0$, it follows that $\langle 0, y \rangle = 0$.

In complex spaces, the inner product is not linear in its right-hand argument, but it does satisfy

$$\langle z, ax + by \rangle = \overline{a}\langle z, x \rangle + \overline{b}\langle z, y \rangle.$$

A vector space on which an inner product is defined is called an **inner-product space**.

[a] The superscript T denotes the **transpose**.

Inner-product spaces are useful settings in which to analyze and design communication and control systems. This is most evident for continuous-time signals and systems when the inner product is given by an integral as above.[2] In this case, we have the obvious realization shown in Figure 3.1. The inner product $\langle x, y \rangle$ can also

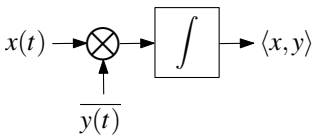


Figure 3.1. Realization of the inner product of waveforms.

be expressed as the sampled convolution of x with a filter "matched" to y. If $y(t)$ is nonzero only on $[0, T]$, put $h(\theta) := \overline{y(T-\theta)}$. Then

$$\begin{aligned}\left(\int_{-\infty}^{\infty} h(t-\tau)x(\tau)\,d\tau\right)\Bigg|_{t=T} &= \int_{-\infty}^{\infty} h(T-\tau)x(\tau)\,d\tau \\ &= \int_{-\infty}^{\infty} \overline{y(T-[T-\tau])}x(\tau)\,d\tau \\ &= \int_{-\infty}^{\infty} x(\tau)\overline{y(\tau)}\,d\tau = \langle x, y \rangle.\end{aligned}$$

Since h is causal, the inner product $\langle x, y \rangle$ can be realized by a physical system. The block diagram of such a system is shown in Figure 3.2, where $H(f)$ denotes the Fourier transform of $h(t)$.

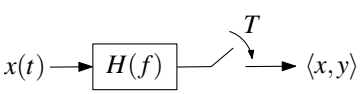


Figure 3.2. Sampled matched-filter realization of the inner product of x and y, where $H(f)$ is the Fourier transform of $h(t) := \overline{y(T-t)}$.

Remark. If in the preceding example $y(t)$ is not of finite duration, we could put $h(\theta) = \overline{y(-\theta)}$ and sample the convolution at $t = 0$. However, since h is not causal, the system would not be physically realizable.

When X is an inner-product space, we usually put $\|x\| := \langle x, x \rangle^{1/2}$ for $x \in X$. As we will see, this formula for $\|\cdot\|$ satisfies the properties of a **norm**, whose precise definition is given later in Section 6.2. For now, it is convenient to introduce the

following terminology. We call $\|x\|$ the **length** of x. When x is a waveform, we call $\|x\|^2$ the **energy** of x. We define the **distance** between two vectors x and y by $\|x-y\|$. We say x and y are **orthogonal** if $\langle x,y\rangle = 0$. They are **orthonormal** if they are orthogonal *and* $\|x\| = \|y\| = 1$. A vector with $\|x\| = 1$ is called a **unit vector**.

Proposition 3.1 (Cauchy–Schwarz Inequality). *For all x and y in an inner-product space,*

$$|\langle x,y\rangle| \le \|x\|\,\|y\|. \tag{3.1}$$

Furthermore, if $y \neq 0$, then equality holds if and only if $x = ay$ for some scalar a.

Proof. If $y = 0$, then both sides of (3.1) are equal to zero. If $y \neq 0$, put

$$t := \left\| x - \frac{\langle x,y\rangle}{\|y\|^2} y \right\|^2 = \left\langle x - \frac{\langle x,y\rangle}{\|y\|^2} y, x - \frac{\langle x,y\rangle}{\|y\|^2} y \right\rangle. \tag{3.2}$$

Expanding, we find that

$$\begin{aligned} t &= \|x\|^2 - \frac{\langle x,y\rangle\langle y,x\rangle}{\|y\|^2} - \frac{\langle x,y\rangle\langle y,x\rangle}{\|y\|^2} + \frac{|\langle x,y\rangle|^2}{\|y\|^4}\|y\|^2 \\ &= \|x\|^2 - \frac{|\langle x,y\rangle|^2}{\|y\|^2}. \end{aligned} \tag{3.3}$$

From (3.2), $t \ge 0$, and so rearranging (3.3) yields (3.1). Now, suppose (3.1) holds with equality. Then from (3.3), $t = 0$, and from (3.2), $x = ay$ with $a = \langle x,y\rangle/\|y\|^2$. Conversely, if $x = ay$ then both sides of (3.1) are equal to $|a|\,\|y\|^2$. □

Corollary 3.2 (Triangle Inequality). *Vectors x and y in an inner-product space satisfy $\|x+y\| \le \|x\| + \|y\|$.*

Proof. This is a simple consequence of the Cauchy–Schwarz inequality (3.1). Write

$$\begin{aligned} \|x+y\|^2 &= \langle x+y, x+y\rangle \\ &= \|x\|^2 + 2\,\mathrm{Re}\langle x,y\rangle + \|y\|^2 \\ &\le \|x\|^2 + 2|\,\mathrm{Re}\langle x,y\rangle| + \|y\|^2 \\ &\le \|x\|^2 + 2|\langle x,y\rangle| + \|y\|^2 \\ &\le \|x\|^2 + 2\|x\|\,\|y\| + \|y\|^2, \qquad \text{by (3.1)}, \\ &= (\|x\| + \|y\|)^2. \end{aligned}$$

□

3.1. Projections onto Subspaces

Consider a set X containing a subset W and a point x. The problem of finding a point in W that is as close as possible to x is called the **projection problem**. To quantify "closeness" requires a notion of distance. When X is an inner-product space, the distance between two vectors x and y is taken to be $\|x-y\|$. In some cases, the projection problem may have a unique solution, while in others there may be no solution or many solutions. This is illustrated in Figure 3.3. In most situations, the

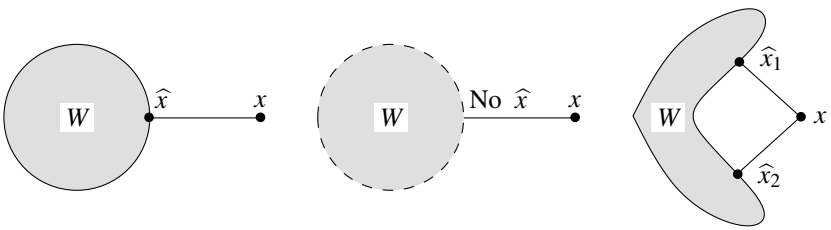


Figure 3.3. Examples for which a unique projection exists, no projection exists, and multiple projections exist.

projection problem is very hard. However, if X is an inner-product space and W is a subspace, the projection problem can often be solved quite easily.

3.1.1. The Orthogonality Principle

Theorem 3.3 (Orthogonality Principle). *Let X be an inner-product space, and let W be a subspace of X. Fix any $x \in X$. Then a vector $\widehat{x} \in W$ has the property that*

$$\|x-\widehat{x}\| \le \|x-w\| \qquad \text{for all } w \in W, \tag{3.4}$$

if and only if

$$\langle x-\widehat{x}, w\rangle = 0 \qquad \text{for all } w \in W. \tag{3.5}$$

Furthermore, there is at most one element $\widehat{x} \in W$ satisfying (3.4) *and* (3.5).

Remark. The theorem does not guarantee the existence of *any* projection $\widehat{x} \in W$ satisfying either (3.4) or (3.5). The theorem only says that if there exists an $\widehat{x} \in W$ satisfying either property, then the other property is also satisfied. The theorem also says that if such an $\widehat{x}$ exists, it is unique.

Before proving the Orthogonality Principle below, we give a simple example of its application.

Example 3.4. Let x_0 be a finite-energy waveform that may not be causal. Find the best causal approximation of x_0.

Solution. Intuitively,

$$\widehat{x}_0(t) := \begin{cases} x_0(t), & t \geq 0, \\ 0, & t < 0, \end{cases}$$

should be the best causal approximation of x_0. However, it is important to realize that the problem as stated is *not* well defined. We *choose* to interpret the question in a way that allows us to apply the Orthogonality Principle. First, let X denote the set of all finite-energy waveforms x defined on $(-\infty,\infty)$. More precisely, X is the set of all waveforms on $(-\infty,\infty)$ that satisfy $\int_{-\infty}^{\infty} |x(t)|^2\,dt < \infty$. We equip X with the inner product

$$\langle x,y\rangle := \int_{-\infty}^{\infty} x(t)\overline{y(t)}\,dt.$$

Let W denote the subset of waveforms in X that are also causal. Arguing as in Example 2.10, W is a subspace of X. We interpret the question as asking us to find a point $\widehat{x}_0 \in W$ that minimizes the distance $\|x-w\|$ over all $w \in W$. To show that the formula above for $\widehat{x}_0$ achieves the minimum distance to W, we use the Orthogonality Principle. We must show that $\langle x_0 - \widehat{x}_0, w\rangle = 0$ for all $w \in W$. Write

$$\begin{aligned} \langle x_0-\widehat{x}_0, w\rangle &= \int_{-\infty}^{\infty} [x_0(t)-\widehat{x}_0(t)]\overline{w(t)}\,dt \\ &= \int_{0}^{\infty} [x_0(t)-\widehat{x}_0(t)]\overline{w(t)}\,dt, \quad \text{since } w \text{ is causal,} \\ &= \int_{0}^{\infty} 0\cdot\overline{w(t)}\,dt, \quad \text{since } x_0(t)-\widehat{x}_0(t) = 0 \text{ for } t \geq 0, \\ &= 0. \end{aligned}$$

Proof of the Orthogonality Principle. We first show that (3.5) implies (3.4). Suppose that (3.5) holds for some $\widehat{x} \in W$. Then for all $w \in W$,

$$\begin{aligned} \|x-w\|^2 &= \|x-\widehat{x}+\widehat{x}-w\|^2 \\ &= \|x-\widehat{x}\|^2 + 2\,\mathrm{Re}\langle x-\widehat{x}, \underbrace{\widehat{x}-w}_{\in W}\rangle + \|\widehat{x}-w\|^2 \\ &= \|x-\widehat{x}\|^2 + \|\widehat{x}-w\|^2, \ \text{using (3.5)}, \\ &\geq \|x-\widehat{x}\|^2, \end{aligned} \tag{3.6}$$

and thus (3.4) holds.

To prove the converse result, suppose (3.4) holds. To obtain a contradiction, suppose there is a $w \in W$ such that $\langle x-\widehat{x}, w\rangle = c \neq 0$. Without loss of generality, we may assume that $\|w\| = 1$. (Otherwise, let $w' := w/\|w\|$, $c' := c/\|w\| \neq 0$, and note that $\langle x-\widehat{x}, w'\rangle = c' \neq 0$.) Write

$$\begin{aligned}
\|x - \underbrace{(\widehat{x}+cw)}_{\in W}\|^2 &= \|(x-\widehat{x}) - cw\|^2 \\
&= \|x-\widehat{x}\|^2 - \langle x-\widehat{x}, cw\rangle \\
&\quad - \langle cw, x-\widehat{x}\rangle + \|cw\|^2 \\
&= \|x-\widehat{x}\|^2 - \overline{c}c - c\overline{c} + |c|^2\|w\|^2 \\
&= \|x-\widehat{x}\|^2 - |c|^2 \\
&< \|x-\widehat{x}\|^2, \qquad \text{since } c \neq 0.
\end{aligned}$$

Since $\widehat{x}+cw \in W$, we have contradicted (3.4).

Finally, we show that if $\widehat{x}$ exists, it must be unique. Suppose that $\widehat{x} \in W$ satisfies (3.5), and suppose there exists a $\tilde{x} \in W$ satisfying

$$\langle x-\tilde{x}, w\rangle = 0 \qquad \text{for all } w \in W.$$

Then

$$\begin{aligned}
\|\widehat{x}-\tilde{x}\|^2 &= \langle \widehat{x}-\tilde{x}, \widehat{x}-\tilde{x}\rangle \\
&= \langle \widehat{x}-x+x-\tilde{x}, \widehat{x}-\tilde{x}\rangle \\
&= \langle x-\widehat{x}, \tilde{x}-\widehat{x}\rangle + \langle x-\tilde{x}, \widehat{x}-\tilde{x}\rangle \\
&= 0+0, \qquad \text{since } \widehat{x}-\tilde{x} \in W.
\end{aligned}$$

Thus, $\|\widehat{x}-\tilde{x}\| = 0$, and $\widehat{x} = \tilde{x}$. □

The Orthogonality Principle says that (3.4) and (3.5) are equivalent. Hence, if $\widehat{x}$ satisfies (3.5), we say that $\widehat{x}$ is the ***orthogonal*** **projection** of x onto W. See Figure 3.4.

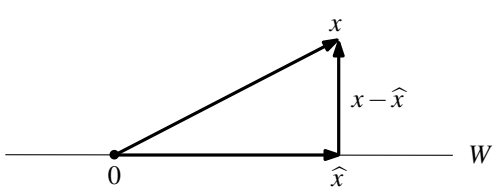


Figure 3.4. An orthogonal projection.

Figure 3.4 suggests that we introduce the following concepts. We say that a vector y is orthogonal to a subset G, denoted by $y \perp G$, if $\langle y, g \rangle = 0$ for all $g \in G$. The set of all such y is called the **orthogonal complement** of G and is denoted by $G^\perp$. For example, since the Orthogonality Principle tells us $\langle x - \widehat{x}, w \rangle = 0$ for all $w \in W$, we can write $x - \widehat{x} \perp W$ or $x - \widehat{x} \in W^\perp$.

We now derive some simple properties of orthogonal projections. Since (3.4) implies (3.5), we can take $w = 0$ in (3.6) and obtain

$$\|x\|^2 = \|x - \widehat{x}\|^2 + \|\widehat{x}\|^2,$$

from which we obtain the error formula,

$$\|x - \widehat{x}\|^2 = \|x\|^2 - \|\widehat{x}\|^2, \tag{3.7}$$

as well as the inequality,

$$\|x\| \geq \|\widehat{x}\|. \tag{3.8}$$

This inequality shows that the operation of projection onto a subspace does not increase the energy of the output signal.

Theorem 3.5 (Linearity of Projections). *Suppose x_1 and x_2 belong to an inner-product space X and have projections $\widehat{x}_1$ and $\widehat{x}_2$ onto a subspace W. Then for any scalars c_1 and c_2, the projection of $c_1 x_1 + c_2 x_2$ onto W is given by $c_1 \widehat{x}_1 + c_2 \widehat{x}_2$.*

Proof. Problem 3.6. □

3.1.2. The Projection Theorem for Finite-Dimensional Subspaces

Theorem 3.6 (Finite-Dimensional Projection Theorem). *If W is a finite-dimensional subspace of an inner-product space X (whose dimension may be finite or infinite), then the projection of any $x \in X$ onto W always exists; i.e., there exists a unique element of W, denoted by $\widehat{x}$, that satisfies* (3.4) *and* (3.5).

Remark. The theorem tells us that for finite-dimensional W, every $x \in X$ has the unique representation $x = \widehat{x} + (x - \widehat{x})$, where $\widehat{x} \in W$ and, by (3.5), $x - \widehat{x} \in W^\perp$. In other words, we have the direct sum, $X = W \oplus W^\perp$.

Proof. If we can establish the existence of at least one projection, uniqueness follows from the Orthogonality Principle. To say that W is finite dimensional means that there is a finite set of vectors, say $w_1, \ldots, w_n$, such that $W = \operatorname{span}\{w_1, \ldots, w_n\}$. To be explicit, every element of W is of the form

$$\sum_{i=1}^{n} c_i w_i \tag{3.9}$$

for some scalars c_i (Proposition 2.12). Without loss of generality, we may assume $w_1, \ldots, w_n$ are orthonormal (if not, we can apply the **Gram–Schmidt procedure**[3] to replace the w_i with orthonormal vectors). We claim that

$$\widehat{x} := \sum_{j=1}^{n} \langle x, w_j \rangle w_j \tag{3.10}$$

satisfies (3.5), and by the Orthogonality Principle, $\widehat{x}$ satisfies (3.4) as well. Since every $w \in W$ has the form (3.9), we see that

$$\langle x - \widehat{x}, w \rangle = \left\langle x - \widehat{x}, \sum_{i=1}^{n} c_i w_i \right\rangle = \sum_{i=1}^{n} \overline{c_i} \langle x - \widehat{x}, w_i \rangle.$$

This quantity will be zero if $\langle x - \widehat{x}, w_i \rangle = 0$ for all i. Using our proposed $\widehat{x}$, write

$$\begin{aligned} \langle x - \widehat{x}, w_i \rangle &= \left\langle x - \sum_{j=1}^{n} \langle x, w_j \rangle w_j, w_i \right\rangle \\ &= \langle x, w_i \rangle - \sum_{j=1}^{n} \langle x, w_j \rangle \langle w_j, w_i \rangle \\ &= \langle x, w_i \rangle - \langle x, w_i \rangle = 0. \end{aligned}$$

□

When $w_1, \ldots, w_n$ are orthonormal as in the foregoing proof, it is easy to check that (3.10) implies

$$\|\widehat{x}\|^2 = \sum_{i=1}^{n} |\langle x, w_i \rangle|^2. \tag{3.11}$$

Combining this with (3.8), we obtain **Bessel's inequality** for an orthonormal basis,

$$\sum_{i=1}^{n} |\langle x, w_i \rangle|^2 \leq \|x\|^2 < \infty. \tag{3.12}$$

Further, we note that since $x = \widehat{x}$ if and only if $x \in W$, (3.7) and (3.11) together imply that

$$\|x\|^2 = \sum_{i=1}^{n} |\langle x, w_i \rangle|^2, \quad x \in W. \tag{3.13}$$

3.1.3. Computing Projections with an Orthonormal Basis

When the subspace W has an orthonormal basis, the proof of the Finite-Dimensional Projection Theorem shows that the projection is given by (3.10).

In general, the vectors w_j in (3.10) can belong to any inner-product space. For example, the w_j could be waveforms. However, when the w_j are column vectors in $\mathbb{C}^m$, if we let $\mathtt{W}$ denote the $m \times n$ matrix whose columns are $w_1, \ldots, w_n$, then the right-hand side of (3.10) can be expressed as $\mathtt{W}\mathtt{W}^{\mathsf{H}}x$, where the superscript H denotes the **Hermitian** or complex-conjugate transpose. The corresponding MATLAB expression is `W*W'*x`.[b]

Example 3.7 (Work). How do we compute the **work** done by applying a force vector F over a distance $\ell > 0$ in the direction w (a unit vector)? The first step is to compute the component of the force in the direction w. In other words, we compute the projection of F onto the subspace spanned by w. By (3.10),

$$\widehat{F} = \langle F, w \rangle w.$$

The magnitude of the work is $\left\|\widehat{F}\right\| \ell = \left|\langle F, w \rangle\right| \ell = \left|\langle F, \ell w \rangle\right|$. The work itself is $\langle F, \ell w \rangle$.

3.1.4. Computing Projections without an Orthonormal Basis

When the subspace W is described as the span of vectors that are not orthonormal, the proof of the Finite-Dimensional Projection Theorem suggests that we apply the Gram–Schmidt procedure and the resulting orthonormal basis to compute the projection. Here we describe another approach.

Theorem 3.8. *If X is an inner-product space (whose dimension may be finite or infinite) and $W = \operatorname{span}\{w_1, \ldots, w_n\}$, then the projection of any $x \in X$ onto the subspace W is given by*

$$\widehat{x} = \sum_{j=1}^{n} c_j w_j,$$

where $c := [c_1, \ldots, c_n]^{\mathsf{T}}$ is any solution of the matrix-vector equation $Gc = b$, where G is the $n \times n$ **Gram matrix** *whose $i\,j$ entry is $\langle w_j, w_i \rangle$, and $b := [\langle x, w_1 \rangle, \ldots, \langle x, w_n \rangle]^{\mathsf{T}}$. The equation $Gc = b$ always has at least one solution, and the solution is unique if and only if $w_1, \ldots, w_n$ are linearly independent.*

[b] In MATLAB, `'` denotes the complex-conjugate transpose, while `.'` denotes the ordinary transpose without complex conjugation.

Remark. When $w_1, \ldots, w_n$ are linearly independent, the solution of $Gc = b$ can be computed in MATLAB with the command `c=G\b`. If the w_i are linearly dependent, or nearly so, MATLAB produces a warning that G is close to singular or badly scaled.

Proof. The key observation is that $\widehat{x} \in W$ satisfies $\langle x - \widehat{x}, w \rangle = 0$ for all $w \in W$ if and only if

$$\left\langle x - \widehat{x}, \sum_{i=1}^{n} c_i w_i \right\rangle = 0 \text{ for } \textit{all} \text{ choices of coefficients } c_i,$$

which happens if and only if $\langle x - \widehat{x}, w_i \rangle = 0$ for each i. Since $\widehat{x} \in W$, this happens if and only if there are scalars $c_1, \ldots, c_n$ such that

$$\left\langle x - \sum_{j=1}^{n} c_j w_j, w_i \right\rangle = 0, \quad i = 1, \ldots, n,$$

or

$$\sum_{j=1}^{n} \langle w_j, w_i \rangle c_j = \langle x, w_i \rangle, \quad i = 1, \ldots, n,$$

which we recognize as $Gc = b$. Hence, if c is any solution of $Gc = b$, then $\widehat{x} = \sum_{j=1}^{n} c_j w_j$ satisfies $\langle x - \widehat{x}, w \rangle = 0$ for all $w \in W$, and so must be the projection by the Orthogonality Principle. Furthermore, since the Finite-Dimensional Projection Theorem tells us that $\widehat{x} \in W$ exists and satisfies $\langle x - \widehat{x}, w \rangle = 0$ for all $w \in W$, we know that a solution of $Gc = b$ exists; the proof that the solution is unique if and only if $w_1, \ldots, w_n$ are linearly independent is left to Problem 3.9. □

Corollary 3.9. *In an inner-product space, if $\widehat{x}$ is the projection of x onto a finite-dimensional subspace $W = \text{span}\{w_1, \ldots, w_n\}$, then knowledge of $\widehat{x}$ is equivalent to knowledge of the column vector of inner products $b = [\langle x, w_1 \rangle, \ldots, \langle x, w_n \rangle]^{\mathsf{T}}$ in that each is a function of the other.*

Proof. First suppose we know the column vector b. Then the foregoing discussion shows there is a solution of $Gc = b$ and that $\widehat{x} = \sum_{j=1}^{n} c_j w_j$. Hence, knowledge of the column vector b of inner products is sufficient to compute $\widehat{x}$. Conversely, suppose we know $\widehat{x}$. Since $\langle x - \widehat{x}, w_i \rangle = 0$ implies $\langle x, w_i \rangle = \langle \widehat{x}, w_i \rangle$, we can compute the entries of the column vector b. □

The utility of Corollary 3.9 is most evident when the space X is a space of continuous-time waveforms. In this case, x, $\widehat{x}$, and $w_1, \ldots, w_n$ are waveforms, but b is a column vector of numbers. Hence, any signal processing on the waveform $\widehat{x}$ can be accomplished by operating on the column vector b of inner products.

Example 3.10. An example of the foregoing arises in the design of receivers for digital communication systems. Messages are transmitted by sending linear combinations of n different **signaling waveforms**, $w_1, \ldots, w_n$. Suppose that a signal

$$s := \sum_{j=1}^{n} c_j w_j$$

is transmitted. Due to noise, the receiver sees $y = s + z$, where z is a noise waveform. Since $s \in W = \text{span}\{w_1, \ldots, w_n\}$, the projection of $y = s + z$ onto W is just $s + \widehat{z}$. This operation has the virtue that $\|\widehat{z}\| \leq \|z\|$; i.e., the energy of the projected noise is no greater than that of the original noise. Of course, receivers do not actually compute the projection, they simply compute the inner products $\langle y, w_i \rangle$ for $i = 1, \ldots, n$ using a bank of matched filters.

Remark. Although the receiver in the above example does not lose any information about the signal by doing the projection, the reader should wonder if the receiver loses information about the noise that could be helpful. If the noise is white and Gaussian, it can be proved that nothing is lost. Otherwise, restricting attention only to inner products with the w_i can be suboptimal. Consider the following situation. Let w and v be orthonormal vectors, and put $z := w + v$. Suppose that the received signal $y = cw + nz$, where the scalar c represents a message to be decoded, and n is a scalar noise factor. The standard receiver would compute only

$$\langle y, w \rangle = \langle cw + nz, w \rangle = \langle cw + n(w + v), w \rangle = c + n.$$

However, if the receiver also computes

$$\langle y, v \rangle = \langle cw + nz, v \rangle = \langle cw + n(w + v), v \rangle = n,$$

the receiver can recover c exactly using the formula $\langle y, w \rangle - \langle y, v \rangle$.

3.1.5. The Euclidean Case

We noted in Theorem 3.8 that the projection onto the span of $w_1, \ldots, w_n$ can be computed in terms of any solution of $Gc = b$, where G is the Gram matrix. However, when X is m-dimensional Euclidean space ($m > n$), we can avoid computation of the Gram matrix. The key is to observe that

$$\left\| x - \sum_{j=1}^{n} c_j w_j \right\| = \|x - Ac\|,$$

where A denotes the $m \times n$ matrix whose columns are $w_1, \ldots, w_n$. In MATLAB, a vector c that minimizes this expression can be obtained with the command `c=A\x`, and the projection $\widehat{x}$ is `A*c`. If the w_i are linearly dependent, or nearly so, MATLAB produces a warning that the matrix A is rank deficient.

A little thought shows that the Gram matrix $G = A^{\mathsf{H}}A$ and $b = A^{\mathsf{H}}x$. Hence, $Gc = b$ is the same as $(A^{\mathsf{H}}A)c = A^{\mathsf{H}}x$. As long as the columns of A are linearly independent, $c = (A^{\mathsf{H}}A)^{-1}A^{\mathsf{H}}x$, and $Ac = A(A^{\mathsf{H}}A)^{-1}A^{\mathsf{H}}x$ is the projection of x onto the subspace spanned by the columns of A. These formulas for c and Ac are excellent theoretical tools, but they should be avoided for numerical work. To see why suppose your computer keeps only two significant digits. If $A = 12$, then $G = A^{\mathsf{H}}A = 12 \times 12 = 144$, but the computer has to round this to 140.

In many cases, A is a block matrix of the form $A = [B\ C]$ so that

$$(A^{\mathsf{H}}A)^{-1}A^{\mathsf{H}} = \begin{bmatrix} B^{\mathsf{H}}B & B^{\mathsf{H}}C \\ C^{\mathsf{H}}B & C^{\mathsf{H}}C \end{bmatrix}^{-1} \begin{bmatrix} B^{\mathsf{H}} \\ C^{\mathsf{H}} \end{bmatrix}.$$

Using **block matrix inversion formulas**,[4] it is easy to show that

$$(A^{\mathsf{H}}A)^{-1}A^{\mathsf{H}} = \begin{bmatrix} (B^{\mathsf{H}}P_C^{\perp}B)^{-1}B^{\mathsf{H}}P_C^{\perp} \\ (C^{\mathsf{H}}P_B^{\perp}C)^{-1}C^{\mathsf{H}}P_B^{\perp} \end{bmatrix}, \tag{3.14}$$

where $P_B := B(B^{\mathsf{H}}B)^{-1}B^{\mathsf{H}}$ and $P_B^{\perp} := I - P_B$; P_C and $P_C^{\perp}$ are defined similarly. Note also that since $P_BP_B = P_B$ and $P_B^{\mathsf{H}} = P_B$, the same is true for $P_B^{\perp}$. Hence, if we put $\widetilde{B} := P_C^{\perp}B$ and $\widetilde{C} := P_B^{\perp}C$, then

$$(A^{\mathsf{H}}A)^{-1}A^{\mathsf{H}} = \begin{bmatrix} (\widetilde{B}^{\mathsf{H}}\widetilde{B})^{-1}\widetilde{B}^{\mathsf{H}} \\ (\widetilde{C}^{\mathsf{H}}\widetilde{C})^{-1}\widetilde{C}^{\mathsf{H}} \end{bmatrix}.$$

3.1.6. Least-Squares Approximation of Waveforms

Consider the problem of approximating a waveform x in terms of a linear combination of given waveforms $w_1, \ldots, w_n$. There are many choices for these waveforms that are easy to work with. For example, we might use complex exponentials if we are approximating periodic waveforms. Or we might use the power functions if we are doing polynomial approximation. Here we consider finite-energy waveforms defined on a finite interval $[a,b]$, and we use the inner product

$$\langle x, y \rangle := \int_a^b x(t)\overline{y(t)}\,dt.$$

Our goal is to find the $w \in W := \operatorname{span}\{w_1, \ldots, w_n\}$ that minimizes

$$\|x - w\|^2 = \int_a^b |x(t) - w(t)|^2 \, dt. \tag{3.15}$$

In other words, we want to find the projection of x onto W. To do this using the Gram matrix and solving $Gc = b$, we have to be able to compute the entries of the column vector b, which means we have to compute the inner products

$$\langle x, w_i \rangle = \int_a^b x(t)\overline{w_i(t)} \, dt.$$

For theoretical problems in which $x(t)$ is given by a formula, we can compute these integrals and solve the projection problem (Problem 3.11). However, in practical problems, we often know only measured values $x(t)$ for a finite set of times, say $t_1, \ldots, t_m$. In this case, we cannot compute the required inner products to solve the projection problem. Instead of choosing $w \in W$ to minimize (3.15), we choose $w \in W$ to minimize

$$\sum_{k=1}^{m} |x(t_k) - w(t_k)|^2. \tag{3.16}$$

This seems possible since we know the measurement values $x(t_k)$. We show how to cast this problem as a projection problem in m-dimensional Euclidean space.

Observe that the column vectors

$$\mathbf{x} := \begin{bmatrix} x(t_1) \\ \vdots \\ x(t_m) \end{bmatrix} \quad \text{and} \quad \mathbf{w}_j := \begin{bmatrix} w_j(t_1) \\ \vdots \\ w_j(t_m) \end{bmatrix}$$

lie in m-dimensional Euclidean space, which we equip with the usual Euclidean inner product, $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{C}^m} := \mathbf{y}^{\mathsf{H}}\mathbf{x}$. The corresponding Euclidean norm is $\|\mathbf{x}\|_{\mathbb{C}^m} := \langle \mathbf{x}, \mathbf{x} \rangle_{\mathbb{C}^m}^{1/2}$. If $w(t) = \sum_{j=1}^n c_j w_j(t)$, then (3.16) becomes

$$\begin{aligned} \sum_{k=1}^{m} |x(t_k) - w(t_k)|^2 &= \sum_{k=1}^{m} \left| x(t_k) - \sum_{j=1}^{n} c_j w_j(t_k) \right|^2 \\ &= \left\| \mathbf{x} - \sum_{j=1}^{n} c_j \mathbf{w}_j \right\|_{\mathbb{C}^m}^2. \end{aligned}$$

Thus, minimizing (3.16) is equivalent to projecting the column vector $\mathbf{x}$ onto the span of $\mathbf{w}_1, \ldots, \mathbf{w}_n$, regarded as a subspace of m-dimensional Euclidean space. As noted above, to compute $c = [c_1, \ldots, c_n]^{\mathsf{T}}$ in MATLAB, we should use the command `c=A\x`, where `A` is the $m \times n$ matrix whose columns are $\mathbf{w}_1, \ldots, \mathbf{w}_n$.

If the w_j are scaled shifts, then `lincmb` can be used to compute `A`, and, once the c_j are found, `lincmb` can compute the approximation $w(t) = \sum_{j=1}^{n} c_j w_j(t)$.

Example 3.11. Consider the approximation of $x(t) = \cos(2\pi t/5)$ based on 31 uniformly spaced samples from $[-10, 10]$. The waveforms used for the approximation are $w_j(t) = v(t - \tau_j)$, where $v(t) = \exp(-t^2/2)$ is the same pulse used in Example 2.1, and the shifts τ_j are the 21 integers $-10, \ldots, 10$. The MATLAB script

```
v = @(t)exp(-t.^2/2);          % Define v(t) = exp(-t^2/2)
tvec = linspace(-10,10,30);    % Sample times
xvec = cos(2*pi*tvec/5).';     % Waveform samples
tau = [-10:10];                % Shifts
c = ones(size(tau));
[y,A] = lincmb(tvec,c,v,tau); % Ignore y
c = A\xvec;
t = linspace(-10,10,200);      % Plot approximation
w = lincmb(t,c.',v,tau);
subplot(2,1,1)
plot(tvec,xvec,'o',t,w);
```

generated the results shown in Figure 3.5, where the circles are the values $x(t_k)$ and the solid line is $\sum_{j=1}^{n} c_j w_j(t)$ using the optimal coefficients.

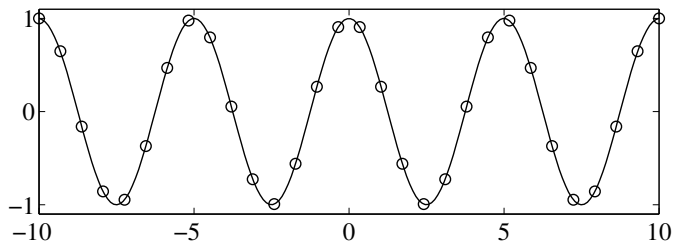


Figure 3.5. Waveform samples $x(t_k)$ (circles) and approximation (solid line).

Least-Squares Polynomial Approximation

MATLAB makes it very easy to compute polynomial least-squares approximations of waveforms. If $w_j(t) = t^{n-j}$, for $j = 1, \ldots, n$, then the vector of coefficients c_j can be computed with the command `c=polyfit(tvec,xvec,n-1)`, where `tvec` is the vector of sample times $t_1, \ldots, t_m$ and `xvec` is the vector of samples $x(t_1), \ldots, x(t_m)$. To evaluate the approximation $w(t) = c_1 t^{n-1} + \cdots + c_{n-1} t + c_n$, use the command `polyval(c,t)`.

3.2. Projections onto Convex Sets

We have seen that the Orthogonality Principle is a powerful tool for solving the problem of projecting onto subspaces. Here we modify the Orthogonality Principle to solve the problem of projecting onto convex sets.

Theorem 3.12. *Let C be a nonempty convex set of an inner-product space X. Fix any $x \in X$. Then $\widehat{x} \in C$ satisfies*

$$\|x-\widehat{x}\| \le \|x-y\| \quad \text{for all } y \in C \tag{3.17}$$

if and only if

$$\mathrm{Re}\langle x-\widehat{x}, y-\widehat{x}\rangle \le 0 \quad \text{for all } y \in C. \tag{3.18}$$

Furthermore, there is at most one $\widehat{x} \in C$ satisfying (3.17) *and* (3.18).

Proof. If (3.18) holds, then we can proceed as in the proof of the Orthogonality Principle and write

$$\|x-y\|^2 = \|(x-\widehat{x})-(y-\widehat{x})\|^2 = \|x-\widehat{x}\|^2 \underbrace{-2\mathrm{Re}\langle x-\widehat{x}, y-\widehat{x}\rangle}_{\ge 0} + \|y-\widehat{x}\|^2 \ge \|x-\widehat{x}\|^2.$$

Conversely, if (3.17) holds, then for any $y \in C$ and any $0 < \lambda \le 1$,

$$\|x-\widehat{x}\|^2 \le \|x-\{\underbrace{\widehat{x}+\lambda(y-\widehat{x})}_{\in C}\}\|^2 = \|(x-\widehat{x})-\lambda(y-\widehat{x})\|^2$$
$$= \|x-\widehat{x}\|^2 - 2\lambda\,\mathrm{Re}\langle x-\widehat{x}, y-\widehat{x}\rangle + \lambda^2\|y-\widehat{x}\|^2.$$

Rearranging and using the fact that $\lambda > 0$ yields

$$\mathrm{Re}\langle x-\widehat{x}, y-\widehat{x}\rangle \le \lambda\|y-\widehat{x}\|^2/2.$$

Since $\lambda > 0$ can be arbitrarily small, (3.18) must hold.

For uniqueness, suppose $\widehat{x}_1$ and $\widehat{x}_2$ are both elements of C that satisfy (3.18). Writing (3.18) for $\widehat{x}_1$ and putting $y = \widehat{x}_2$ yields

$$\mathrm{Re}\langle x-\widehat{x}_1, \widehat{x}_2-\widehat{x}_1\rangle \le 0. \tag{3.19}$$

Similarly, writing (3.18) for $\widehat{x}_2$ and putting $y = \widehat{x}_1$ yields $\mathrm{Re}\langle x-\widehat{x}_2, \widehat{x}_1-\widehat{x}_2\rangle \le 0$, which we can rewrite as

$$\mathrm{Re}\langle \widehat{x}_2-x, \widehat{x}_2-\widehat{x}_1\rangle \le 0.$$

Adding this to (3.19), we obtain

$$\mathrm{Re}\langle \widehat{x}_2 - \widehat{x}_1, \widehat{x}_2 - \widehat{x}_1 \rangle \leq 0.$$

This inner product is already real and equal to $\|\widehat{x}_2 - \widehat{x}_1\|^2$. Hence, $\|\widehat{x}_2 - \widehat{x}_1\|^2 \leq 0$, which implies $\widehat{x}_2 = \widehat{x}_1$. □

Example 3.13. Let X denote the set of all complex-valued, finite-energy waveforms defined on $(-\infty, \infty)$. Let C denote the subset of X consisting of real-valued waveforms x with $x(t) \geq 0$ for $|t| \leq 1$. The set C can be shown to be convex as in Example 2.16. Given a real-valued, finite-energy waveform x_0, find the waveform in C that is closest to x_0.

Solution. Intuitively,

$$\widehat{x}_0(t) := \begin{cases} x_0(t), & |t| > 1, \\ x_0(t), & |t| \leq 1 \text{ and } x_0(t) \geq 0, \\ 0, & |t| \leq 1 \text{ and } x_0(t) < 0, \end{cases}$$

should be the best approximation of x_0 from C. To apply Theorem 3.12, we must show that $\mathrm{Re}\langle x_0 - \widehat{x}_0, y - \widehat{x}_0 \rangle \leq 0$ for all $y \in C$. To begin, observe that since x_0, $\widehat{x}_0$, and y are real, and since $x_0(t) - \widehat{x}_0(t) = 0$ for $|t| > 1$,

$$\begin{aligned} \mathrm{Re}\langle x_0 - \widehat{x}, y - \widehat{x}_0 \rangle &:= \mathrm{Re} \int_{-\infty}^{\infty} [x_0(t) - \widehat{x}_0(t)] \overline{[y(t) - \widehat{x}_0(t)]}\, dt \\ &= \int_{-\infty}^{\infty} [x_0(t) - \widehat{x}_0(t)][y(t) - \widehat{x}_0(t)]\, dt \\ &= \int_{-1}^{1} [x_0(t) - \widehat{x}_0(t)][y(t) - \widehat{x}_0(t)]\, dt. \end{aligned}$$

Consider this last integrand for $|t| \leq 1$. For such t, either $x_0(t) \geq 0$, which implies $\widehat{x}_0(t) = x_0(t)$, making the integrand zero, or $x_0(t) < 0$, in which case $\widehat{x}_0(t) = 0$, and we see that

$$[x_0(t) - \widehat{x}_0(t)][y(t) - \widehat{x}_0(t)] = [x_0(t) - 0][y(t) - 0] = x_0(t) y(t) \leq 0,$$

since $y \in C$ implies $y(t) \geq 0$ for $|t| \leq 1$. Since the integrand is nonpositive, so is the integral. Hence, the condition of Theorem 3.12 is satisfied, and we have proved that $\widehat{x}_0$ as defined is the closest element to x_0 from C.

Notes

Note 3.1. The **Lebesgue integral** [6], [14], [33], [34] of any nonnegative (measurable) function is well defined, and the value ∞ is allowed. If $\int |x(t)|\,dt < \infty$, then the Lebesgue integral $\int x(t)\,dt$ exists and is a finite real or complex number. If $\int |x(t)|^2\,dt < \infty$ and $\int |y(t)|^2\,dt < \infty$, then $\int |x(t)y(t)|\,dt < \infty$ by Hölder's inequality (see Section 6.2.1), and so the inner product $\int x(t)y(t)\,dt$ exists as a finite real or complex number.

Note 3.2. When using inner products defined by an integral, we agree that any waveform x with $\int |x(t)|^2\,dt = 0$ is called the **zero waveform**. The reason for this is if we change the value of $x(t)$ at one time t, the value of the integral will not change. More generally, if $\int |x(t)-y(t)|^2\,dt = 0$, we consider x and y to be the same waveform even if they are not equal at all times in the interval of integration.

Note 3.3. The Gram–Schmidt Procedure. Let $w_1,\ldots,w_n$ be nonzero vectors in an inner-product space. Consider the sequence $v_1 := w_1$,

$$v_k := w_k - \sum_{i=1}^{k-1} \left\langle w_k, \frac{v_i}{\|v_i\|} \right\rangle \frac{v_i}{\|v_i\|}, \quad k = 2,\ldots,n. \tag{3.20}$$

Let V_n denote the nonzero vectors generated by this procedure. Then the elements of V_n are orthogonal, and $\operatorname{span} V_n = \operatorname{span}\{w_1,\ldots,w_n\}$. By replacing every $v \in V_n$ with $v/\|v\|$ we obtain an ortho*normal* basis. Under the assumption that $w_1,\ldots,w_n$ are linearly independent, we prove these claims by induction on n. Later we discuss the linearly dependent case. It is obvious that the results hold for $n = 1$. Suppose they hold for some n. Put

$$v_{n+1} = w_{n+1} - \sum_{i=1}^{n} \frac{\langle w_{n+1}, v_i\rangle}{\|v_i\|^2} v_i.$$

Using this equation along with the induction hypothesis, it is not hard to show that $\operatorname{span}\{v_1,\ldots,v_{n+1}\} = \operatorname{span}\{w_1,\ldots,w_{n+1}\}$. We claim that $v_{n+1} \neq 0$. Otherwise,

$$w_{n+1} = \sum_{i=1}^{n} \frac{\langle w_{n+1}, v_i\rangle}{\|v_i\|^2} v_i \in \operatorname{span}\{v_1,\ldots,v_n\} = \operatorname{span}\{w_1,\ldots,w_n\},$$

which contradicts the assumed linear independence of $w_1,\ldots,w_n,w_{n+1}$. Finally, we check orthogonality. For $1 \le j \le n$,

$$\begin{aligned}
\langle v_{n+1}, v_j\rangle &= \langle w_{n+1}, v_j\rangle - \sum_{i=1}^{n} \frac{\langle w_{n+1}, v_i\rangle}{\|v_i\|^2} \langle v_i, v_j\rangle \\
&= \langle w_{n+1}, v_j\rangle - \frac{\langle w_{n+1}, v_j\rangle}{\|v_j\|^2} \langle v_j, v_j\rangle \\
&= \langle w_{n+1}, v_j\rangle - \langle w_{n+1}, v_j\rangle = 0,
\end{aligned}$$

where in the second equation we have used the induction hypothesis that $\langle v_i, v_j\rangle = 0$ for $1 \le i, j \le n$.

Remarks. (i) The Gram–Schmidt procedure is *not* numerically stable.

(ii) Recall our argument that v_{n+1} could not be zero. This shows that if the w_i are not linearly independent, then the Gram–Schmidt procedure detects if w_{n+1} is linearly dependent on $w_1, \ldots, w_n$. If this happens, we simply discard w_{n+1}, and continue with w_{n+2} instead.

(iii) Recalling the proof of the Finite-Dimensional Projection Theorem, when we look at the Gram–Schmidt procedure, we see that the sum in (3.20) is the projection of w_k onto the span of the orthonormal vectors $v_1/\|v_1\|, \ldots, v_{k-1}/\|v_{k-1}\|$, which is the span of $w_1, \ldots, w_{k-1}$. Hence, v_k is the projection of w_k onto $(\text{span}\{w_1, \ldots, w_{k-1}\})^\perp$.

(iv) Assume $w_1, \ldots, w_n$ are linearly independent so that no v_i from the Gram–Schmidt procedure is the zero vector. Then rewrite (3.20) as $w_k = \sum_{i=1}^k \widehat{R}_{ik} q_i$, where $q_i := v_i/\|v_i\|$ and $\widehat{R}_{ik} := \langle w_k, q_i\rangle$ for $i = 1, \ldots, k-1$, $\widehat{R}_{kk} := 1$, and $\widehat{R}_{ik} := 0$ for $i > k$. When the w_i are length-$m \ge n$ column vectors, consider the $m \times n$ matrices $\widehat{Q} := [q_1|\cdots|q_n]$ and $A := [w_1|\cdots|w_n]$. Then we can write $A = \widehat{Q}\widehat{R}$, which is the ***thin*** or ***reduced*** **QR decomposition** of A in which the columns of $\widehat{Q}$ are orthonormal and R is an upper-triangular matrix. If $n < m$, we can also write the full **QR decomposition**,

$$A = \underbrace{\begin{bmatrix} \widehat{Q} & \widetilde{Q} \end{bmatrix}}_{=:Q} \underbrace{\begin{bmatrix} \widehat{R} \\ 0 \end{bmatrix}}_{=:R},$$

where the $m-n$ columns of $\widetilde{Q}$ are chosen to form an orthonormal basis for $(\text{range}\, Q)^\perp$, and 0 denotes an $(m-n) \times n$ matrix of zeros. Notice that Q is unitary since $Q^{\mathsf{H}}Q = I$.

(v) The QR decomposition provides and easy derivation of the **Hadamard inequality**, $|\det A| \le \prod_{i=1}^n \|\mathbf{a}_i\|$, where $\mathbf{a}_i$ is the ith column of A. If the columns of A are linearly dependent, $\det A = 0$ and there is nothing to prove. So assume the columns of A are linearly independent and apply the QR decomposition to write $A = QR$, which implies $\det A = \det Q \det R$. Since $Q^{\mathsf{H}}Q = I$ and $\det Q^{\mathsf{H}} = \det Q$, we see that $1 = \det I = \det Q^{\mathsf{H}}Q = (\det Q)^2$ implies $|\det Q| = 1$. Since R is upper triangular, $\det R = R_{11} \cdots R_{nn}$. Now observe that $\|\mathbf{a}_i\|^2 = (A^{\mathsf{H}}A)_{ii}$ and that from $A = QR$, we have $A^{\mathsf{H}}A = R^{\mathsf{H}}R$. Hence,

$$\begin{aligned} |\det A| = |\det R| = \prod_{i=1}^n |R_{ii}| &\le \prod_{i=1}^n (\text{norm of column } i \text{ of } R) \\ &= \prod_{i=1}^n (R^{\mathsf{H}}R)_{ii}^{1/2} = \prod_{i=1}^n (A^{\mathsf{H}}A)_{ii}^{1/2} = \prod_{i=1}^n \|\mathbf{a}_i\|. \end{aligned}$$

Notice that if equality holds, then for each i, R_{ii} is the only possible nonzero entry of column i of R; i.e., R is diagonal, which implies $R^{\mathsf{H}}R$ is diagonal. Since $A^{\mathsf{H}}A = R^{\mathsf{H}}R$, the columns of A must be orthogonal. Conversely, if the columns of A are orthogonal, then $A^{\mathsf{H}}A$ and therefore $R^{\mathsf{H}}R$ are diagonal. Then since R is upper triangular, R must be diagonal. This forces equality in the above display.

(*vi*) If B is positive definite, let $A := B^{1/2}$ and apply the Hadamard inequality to write $\det B = |\det A|^2 \le \left(\prod_{i=1}^n \|\mathbf{a}_i\|\right)^2 = \prod_{i=1}^n \|\mathbf{a}_i\|^2$. Now $\|\mathbf{a}_i\|^2 = (A^{\mathsf{H}}A)_{ii} = B_{ii}$. Hence, $\det B \le \prod_{i=1}^n B_{ii}$.

Note 3.4. Block Matrix Inversion Formulas. If D is invertible, then

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} I & BD^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} \Phi & 0 \\ 0 & D \end{bmatrix} \begin{bmatrix} I & 0 \\ D^{-1}C & I \end{bmatrix},$$

where $\Phi := A - BD^{-1}C$ is the **Schur complement** of D. Assuming the block matrix on the left is invertible, we have

$$\begin{aligned} \begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} &= \begin{bmatrix} I & 0 \\ -D^{-1}C & I \end{bmatrix} \begin{bmatrix} \Phi^{-1} & 0 \\ 0 & D^{-1} \end{bmatrix} \begin{bmatrix} I & -BD^{-1} \\ 0 & I \end{bmatrix} \\ &= \left[\begin{array}{c|c} \Phi^{-1} & -\Phi^{-1}BD^{-1} \\ \hline -D^{-1}C\Phi^{-1} & D^{-1}C\Phi^{-1}BD^{-1} + D^{-1} \end{array}\right]. \end{aligned} \tag{3.21}$$

Similarly, if A is invertible,

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} I & 0 \\ CA^{-1} & I \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & \Psi \end{bmatrix} \begin{bmatrix} I & A^{-1}B \\ 0 & I \end{bmatrix},$$

where $\Psi := D - CA^{-1}B$ is the Schur complement of A. It follows that

$$\begin{aligned} \begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} &= \begin{bmatrix} I & -A^{-1}B \\ 0 & I \end{bmatrix} \begin{bmatrix} A^{-1} & 0 \\ 0 & \Psi^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -CA^{-1} & I \end{bmatrix} \\ &= \left[\begin{array}{c|c} A^{-1} + A^{-1}B\Psi^{-1}CA^{-1} & -A^{-1}B\Psi^{-1} \\ \hline -\Psi^{-1}CA^{-1} & \Psi^{-1} \end{array}\right]. \end{aligned} \tag{3.22}$$

Combining the top blocks of (3.21) with the bottom blocks of (3.22) yields

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \left[\begin{array}{c|c} [A - BD^{-1}C]^{-1} & -[A - BD^{-1}C]^{-1}BD^{-1} \\ \hline -[D - CA^{-1}B]^{-1}CA^{-1} & [D - CA^{-1}B]^{-1} \end{array}\right]. \tag{3.23}$$

Problems

1. Show that an orthonormal set of vectors must be linearly independent.

2. Let x and y be two unit vectors in a *real* inner-product space. Show that $x+y$ and $x-y$ are orthogonal.

3. Show that in any inner-product space, the **parallelogram law** holds:

$$\|x+y\|^2 + \|x-y\|^2 = 2(\|x\|^2 + \|y\|^2).$$

4. In a complex inner-product space, show that the **polarization identity** holds:

$$4\langle x,y\rangle = \|x+y\|^2 - \|x-y\|^2 + j\|x+jy\|^2 - j\|x-jy\|^2.$$

   What is the analogous formula in a real inner-product space?

5. Let X and Y be inner-product spaces with respective inner products $\langle\cdot,\cdot\rangle_X$ and $\langle\cdot,\cdot\rangle_Y$. For (x_1,y_1) and (x_2,y_2) in the product space $X\times Y$, put

$$\big\langle (x_1,y_1),(x_2,y_2)\big\rangle := \langle x_1,x_2\rangle_X + \langle y_1,y_2\rangle_Y.$$

   Determine whether or not this formula satisfies the properties of an inner product on $X\times Y$.

6. Use the Orthogonality Principle to prove the linearity of projections, Theorem 3.5.

7. Given a finite-energy waveform x, find the even waveform that best approximates x. Justify your answer.

8. Let f_0 be a given frequency. We call a finite-energy waveform lowpass if its Fourier transform is zero for $|f|>f_0$. Given a finite-energy waveform x, find the lowpass waveform that best approximates x. Justify your answer.

9. Let $w_1,\ldots,w_n$ be given vectors in an inner-product space, and let G denote the Gram matrix, whose entries are $G_{ij}=\langle w_j,w_i\rangle$. Assuming that the w_i are linearly independent, show that $Gc=0$ implies $c=0$. Conversely, if the only solution of $Gc=0$ is $c=0$, show that the w_i are linearly independent.

10. In an inner-product space X, let $w_1,\ldots,w_n$ be orthonormal. Fix $x\in X$, and put $\widehat{x} := \sum_{i=1}^n \langle x,w_i\rangle w_i$. Show that

$$\|\widehat{x}\|^2 = \sum_{i=1}^n \big|\langle x,w_i\rangle\big|^2.$$

11. On $[0,1]$, find the best first-degree polynomial approximation $\widehat{x}$ to $x(t) = t^3$. That is, find $\widehat{x}(t) = c_1 + c_2 t$ that minimizes

$$\int_0^1 |x(t) - \widehat{x}(t)|^2 \, dt.$$

How do your answers compare with the output of the following code

```
w1 = @(t)ones(size(t));
w2 = @(t)t;
x = @(t)t.^3;
tvec = linspace(0,1,50).';
xvec = x(tvec);
A = [ w1(tvec)  w2(tvec) ];
c = A\xvec
```

which implements the method of (3.16) and the paragraph following it?

12. On $[0,1]$ consider the waveform $x(t) = t^2$. Find the best first-degree polynomial approximation $\widehat{x}$ of x that also satisfies $\int_0^1 \widehat{x}(t) = 0$. That is, find $\widehat{x}(t) = c_1 + c_2 t$ that minimizes

$$\int_0^1 |x(t) - \widehat{x}(t)|^2 \, dt$$

and also satisfies $\int_0^1 \widehat{x}(t) = 0$.

13. Let G be any subset of an inner-product space X. (a) Show that $G^\perp$ is a subspace. (b) Show that $G \subset (G^\perp)^\perp$. (c) Show that $\operatorname{span} G \subset (G^\perp)^\perp$.

14. If $X = W \oplus W^\perp$, where X is an inner-product space and W is a subspace of X, show that $(W^\perp)^\perp = W$. *Hint:* Keep in mind that $(W^\perp)^\perp$ is a subset of the whole space X.

15. Let W be a subspace of an inner-product space X and such that for all $x \in X$, the projection of x onto W, $\widehat{x}$, exists. Hence, we may define the mapping $P : X \to X$ by $P(x) := \widehat{x}$. By Theorem 3.5, P is linear.

    (a) Show that for all x and y in X, $\|Px - Py\| \le \|x - y\|$.
    (b) Show that P is **idempotent**; i.e., $P^2 = P$, or, more explicitly, $P(Px) = Px$ for all $x \in X$.
    (c) Show that $\langle Px, y\rangle = \langle x, Py\rangle$. Such an operator is said to be **self adjoint**. *Hint:* Use the Orthogonality Principle to show that $\langle Px, y\rangle$ and $\langle x, Py\rangle$ are both equal to $\langle Px, Py\rangle$.
    (d) If $y = x + Px$, solve for x in terms of y and Py.

(e) Consider the operator $Hx := x - 2Px = (I - 2P)x$. Show that H is self adjoint. Show that $H^2 = I$.

Remarks. (*i*) The operator H can be considered a **reflection** about the subspace $W^\perp$ in the sense that for $x \in W^\perp$, $Hx = x$, while for $x \in W$, $Hx = x - 2x = -x$. For example, when $X = \mathbb{R}^3$ and W is one dimensional, $W^\perp$ is two dimensional like the surface of a **mirror**. A vector $x \in W$ is reflected by H to the new position $Hx = -x$, while a vector $x \in W^\perp$ lies in the mirror and is unaffected by H since $Hx = x$.

(*ii*) When $W = \text{span}\{v\}$, the vector $w := v/\|v\|$ is an orthonormal basis for W, and so $Px = \langle x, w\rangle w = \langle x, v\rangle v/\|v\|^2$. In this case, $Hx = x - 2\langle x, v\rangle v/\|v\|^2$ is called a **Householder transformation**. When $X = \mathbb{C}^d$, the quantity $I - 2vv^{\mathsf{H}}/\|v\|^2$ is called a **Householder matrix**. Although $Hx = (I - 2vv^{\mathsf{H}}/\|v\|^2)x$, multiplying x by this matrix is a very inefficient way to compute Hx.

16. Let u and v be unit vectors in an inner-product space. Let

$$W_1 := \text{span}\{u\} \qquad \text{and} \qquad W_2 := \text{span}\{v\}.$$

Let P_1 and P_2 be the corresponding projection operators.

(a) Show that $P_1 x = \langle x, u\rangle u$. (Of course it will follow that $P_2 x = \langle x, v\rangle v$.)

(b) Let $Tx := P_2(P_1 x)$. It is easy to see that $Tx = c\langle x, u\rangle v$, where $c := \langle u, v\rangle$. Show that $T^n x \to 0$ if $|c| < 1$.

(c) With c as in part (b), show that $|c| = 1$ implies $W_1 = W_2$.

17. Let $N \subset M$ be subspaces of an inner-product space X. Assume the corresponding projection operators P_N and P_M exist. Show that $P_N(x) = P_N(P_M(x))$ for all $x \in X$.

18. Let X denote the set of real-valued, continuous functions on $[0, 2]$ equipped with the usual inner product, $\langle x, y\rangle := \int_0^2 x(t)y(t)\,dt$ for $x, y \in X$. Put $W := \{x \in X : x(t) = 0 \text{ for } 1 \le t \le 2\}$.

(a) Prove that $W^\perp = \{y \in X : y(t) = 0, \text{ for } 0 \le t \le 1\}$.

(b) Does $X = W \oplus W^\perp$? Justify your answer.

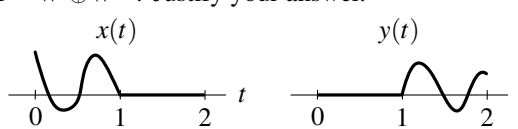


Figure 3.6. A typical waveform $x \in W$ (left), and a typical waveform $y \in W^\perp$ (right) in Problem 3.18.

19. Let X be an inner-product space, and let W be a subspace of X such that the projection of any vector x onto W exists. Denote this projection by $P_W x$. Consider the translated subspace

$$T := x_0 + W = \{x_0 + w : w \in W\}.$$

Given $x \in X$, determine whether or not

$$\widehat{x} := x_0 + P_W(x - x_0)$$

is the projection of x onto T.

20. Let U and V be subspaces of an arbitrary inner-product space X. Let W denote the orthogonal complement of $U \cap V$ regarded as a subspace of V; i.e.,

$$W := \{v \in V : v \perp U \cap V\}.$$

(a) Show that we can write $U \oplus W$. *Hint:* Problem 2.17.
(b) Show that we can write $(U \cap V) \oplus W$.
(c) Assume that $V = (U \cap V) \oplus W$. Show that $(U \oplus W)^{\perp} \subset V^{\perp}$.

21. Show that the Orthogonality Principle Theorem 3.3 can be derived from the Modified Orthogonality Principle Theorem 3.12. *Hint:* When z is any element of a subspace, $y = \pm z + \widehat{x}$ and $y = \pm jz + \widehat{x}$ also belong to the subspace and can be substituted into (3.18).

22. Let X denote the set of all real-valued continuous functions on $[0,2]$ equipped with the usual inner product, $\langle u, v \rangle := \int_0^2 u(t)v(t)\,dt$. Let

$$C := \{x \in X : x(t) \geq 0 \text{ for } 0 \leq t \leq 1\}.$$

Put $x(t) := t - 1$ for $0 \leq t \leq 2$. Does there exist an $\widehat{x} \in C$ such that

$$\|x - \widehat{x}\| \leq \|x - y\|, \quad \text{for all } y \in C\,?$$

Justify your answer.

23. Let $X = \ell^2$ denote the set of all real-valued, finite-energy sequences. The inner product is $\langle x, y \rangle := \sum_{k=1}^{\infty} x(k)y(k)$. Let

$$C := \{x \in X : x(k) \leq b \text{ for } k > N\},$$

where N is a given positive integer and $b > 0$ is a given bound. Given an arbitrary $x \in X$, does the projection of x onto C exist? If "yes," find it and justify your answer. If "no," explain why not.

24. Let X denote the set of real-valued waveforms on $\mathbb{R}$ having finite energy. Let $C_1 := \{x \in X : x(t) \geq 0 \text{ for all } t \geq 0\}$, and let $C_2 := \{x \in X : x(t) \leq 0 \text{ for all } t \leq 0\}$. Consider the waveform

$$x_0(t) := \begin{cases} e^t, & t \leq 0, \\ -2, & 0 < t < 3, \\ 1/t, & t \geq 3. \end{cases}$$

Does there exist an $\widehat{x}_0 \in C_1 \cup C_2$ such that $\|x_0 - \widehat{x}_0\| \leq \|x_0 - y\|$ for all $y \in C_1 \cup C_2$? If "no," explain why. If "yes," find $\widehat{x}_0$ and justify the steps of your analysis.

25. Given a real number x and a positive number t, show that the solution of

$$\operatorname*{argmin}_{y \in \mathbb{R}} \left(t|y| + \tfrac{1}{2}|x-y|^2 \right)$$

is given by the **soft threshold** or **shrinkage operator** operator,[c]

$$\eta_t(x) := \begin{cases} x - t, & x \geq t, \\ x + t, & x \leq -t, \\ 0, & |x| < t. \end{cases}$$

Hints: If $x \geq 0$, it is clear that minimizing value of y must be nonnegative. Similarly, if $x < 0$, the minimizing value of y must be nonpositive. Hence, you can treat each case separately. Use the fact that for $y \geq 0$, $|y| = y$, while for $y < 0$, $|y| = -y$.

[c] The shrinkage operator can also be written as $\eta_t(x) = (|x| - t)^+ \operatorname{sgn}(x)$, where $(\cdot)^+$ is the **positive part operator** defined by $x^+ := x$ for $x \geq 0$ and $x^+ := 0$ for $x < 0$. The **sign function** is $\operatorname{sgn}(x) := 1$ for $x > 0$, $\operatorname{sgn}(x) := -1$ for $x < 0$, and $\operatorname{sgn}(x) := 0$ for $x = 0$.

CHAPTER 4
Linear Operators

4.1. Definition and Examples

Let X and Y be vector spaces. A mapping $A: X \to Y$ is called a **linear operator** or a **linear transformation** if for all scalars c_1 and c_2 and all vectors $x_1, x_2 \in X$,

$$A(c_1x_1 + c_2x_2) = c_1Ax_1 + c_2Ax_2.$$

In the following examples, we give some common formulas that are used to define linear operators.

Example 4.1 (Matrix Operator). For $x = [x_1, \ldots, x_n]^{\mathsf{T}}$, define the length-$m$ column vector Ax by setting its ith entry to be

$$(Ax)_i := \sum_{j=1}^{n} a_{ij}x_j, \tag{4.1}$$

where a denotes the $m \times n$ matrix with given entries a_{ij}. In more compact terms, we write $Ax = ax$, where ax is understood to be the product of the $m \times n$ matrix a and the $n \times 1$ column vector x.

Example 4.2. Given an infinite sequence $x = (x_1, x_2, \ldots)$, define the infinite sequence Ax by

$$Ax := (x_1, x_2/2, x_3/3, \ldots).$$

In other words, $(Ax)_k = x_k/k$.

Example 4.3 (Integral Operator). Given a waveform x defined on the interval $[a,b]$, define the waveform Ax on the interval $[c,d]$ by

$$(Ax)(t) := \int_a^b k(t,\tau)x(\tau)\,d\tau, \quad t \in [c,d],$$

where $k(t,\tau)$ is a given function. Linear operators of this form are used in modeling time-varying wireless communication channels. In this context, k is called the time-varying impulse response. Operators of this form can also be generalized to model blur in image-processing systems. In this context, k corresponds to the point-spread function.

Example 4.4 (Modulation Operator). Let $w_1, \ldots, w_n$ be given **signaling waveforms**. For example, the w_k might be sinusoids of different frequencies. For $x = [x_1, \ldots, x_n]^{\mathsf{T}} \in \mathbb{C}^n$, define the waveform Ax to be the linear combination

$$Ax := \sum_{k=1}^{n} x_k w_k.$$

To be more explicit,

$$(Ax)(t) = \sum_{k=1}^{n} x_k w_k(t). \tag{4.2}$$

This **modulation operator** can be implemented as shown in Figure 4.1.

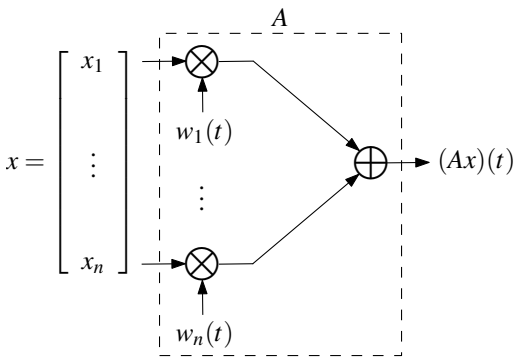


Figure 4.1. Block diagram for the modulation operator A in (4.2).

In (4.1), if we write $a_j(i)$ instead a_{ij}, then (4.1) is a discrete-time version of (4.2). In other words, the jth column of a matrix is a discrete-time, finite-duration signaling waveform, and matrix-vector multiplication can be thought of as a modulation operator.

Example 4.5 (Identity Operators). On any vector space X, the **identity operator** $I: X \to X$ is defined by $Ix := x$.

4.1.1. Missing or Incomplete Data

The situation in which it is most obvious that the problem of missing data can occur is that of an integral operator as in Example 4.3 or a modulation operator in

Example 4.4. Although these models call for continuous-time outputs, in practice, we usually only have finitely many output samples, say $y(t_i) = (Ax)(t_i)$ for $i = 1, \ldots, m$.

Even in the case of a matrix operator as in Example 4.1, we may be confronted with missing data. For example, in order to speed up the acquisition of **magnetic resonance imaging (MRI)** data, even though the model calls for measuring $y_i = (Ax)_i$ for $i = 1, \ldots, m$, we only measure y_i for a subset of indexes, say $i \in I$, because it would take too long to acquire y_i for all i. In this situation, it is typically the case that the number of measurements is much smaller than the dimension of the vector x. This leads to the situation that there are multiple solutions x of the equations $y_i = (Ax)_i$, $i \in I$. To see how this works, suppose that A is represented by a 5×5 matrix with entries a_{ij}. Then the full model has the form

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix}.$$

If we measure only y_1, y_3, and y_5, the new matrix-vector equation is

$$\begin{bmatrix} y_1 \\ y_3 \\ y_5 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix}.$$

This new equation has the form $z = Bx$, where $z := [y_1, y_3, y_5]^\mathsf{T}$ and B denotes operator corresponding to rows 1, 3, and 5 of the original matrix a. Because B has fewer rows than columns, if there are any solutions of $z = Bx$, then there are infinitely many solutions. Which one should we prefer? In this chapter, we learn how to find the solution of minimum energy, but we might also consider a solution of minimum one-norm (see Problem 5.48).

4.2. Terminology and Basic Results

The **kernel** or **null space** of a linear operator $A: X \to Y$ is the set

$$\ker A := \{x \in X : Ax = 0\}.$$

Notice that since $A0 = A(0+0) = A0 + A0$, we must have $A0 = 0$. Hence, the kernel always contains the zero vector of X. The **range** or **image** of A is the set

$$\operatorname{range} A := \{Ax : x \in X\} \subset Y.$$

It is easy to show that $\ker A$ is a subspace of X and that $\text{range} A$ is a subspace of Y (Problem 4.2). The dimension of the kernel is called the **nullity**, and the dimension of the range is called the **rank**.

Theorem 4.6 (Rank–Nullity). *Let X and Y be vector spaces, and let $A: X \to Y$ be a linear operator. If $\dim X < \infty$, then both $\ker A$ and $\text{range} A$ are finite dimensional, and*

$$\dim \ker A + \dim \text{range} A = \dim X.$$

Proof. Since $\ker A$ is a subspace of the finite-dimensional space X, $\dim \ker A \leq \dim X$ (recall Section 2.3). Put $n := \dim X$, and put $r := \dim \ker A$. Let $\{x_1, \ldots, x_r\}$ be a basis for $\ker A$. Extend this to a basis for X, say $\{x_1, \ldots, x_r, x_{r+1}, \ldots, x_n\}$. Then it is easy to show (Problem 4.3) that $\{Ax_{r+1}, \ldots, Ax_n\}$ is a basis for $\text{range} A$; i.e., $\dim \text{range} A = n - r = \dim X - \dim \ker A$. □

We can use the Rank–Nullity Theorem to give simple conditions to determine whether or not a linear operator is invertible. However, we must first discuss the general concept of an invertible function.

A function f mapping any set X into any set Y is said to be **invertible** if for every $y \in Y$, there is a unique $x \in X$ with $y = f(x)$. In this case, for each $y \in Y$, the corresponding unique value of x with $f(x) = y$ is denoted by $f^{-1}(y)$.

Notice that to show a function is invertible, we have to show two things. First, we have to show that for every $y \in Y$, there is an $x \in X$ with $y = f(x)$. A function with this property is said to be **onto**. The second thing we have to show is that for every $y \in Y$, the $x \in X$ with $y = f(x)$ is unique. In other words, we must show that

$$\text{for all } x_1, x_2 \in X, \quad f(x_1) = f(x_2) \text{ implies } x_1 = x_2.$$

A function with this property is said to be **one-to-one**.

You should verify that if a linear operator is invertible, its inverse is linear. You should also verify that a *linear* operator A is one-to-one if and only if it has the property

$$Ax = 0 \quad \text{implies} \quad x = 0.$$

In other words, a linear operator is one-to-one if and only its kernel contains only the zero vector. In this case, we say that the operator is **nonsingular**. Not surprisingly, an operator is said to be **singular** if its kernel contains a nonzero vector. When X is finite dimensional, the Rank–Nullity Theorem tells us that A is one-to-one if and only if it has maximum possible rank (which is $\dim X$). Such an operator is said to be **full rank**. To summarize,

$$A \text{ is one-to-one} \Leftrightarrow A \text{ is nonsingular} \Leftrightarrow \ker A = \{0\} \Leftrightarrow A \text{ is full rank},$$

where the last expression is reserved for the case $\dim X < \infty$.

Remark. With regard to matrices, the term "nonsingular" is usually restricted to *square* matrices so that, by the Rank–Nullity Theorem, "nonsingular" becomes a synonym for "invertible." However, we occasionally depart from this standard and say that a nonsquare matrix M is nonsingular, meaning only that $Mx = 0$ implies $x = 0$.

Example 4.7 (Some Operators That Are One-to-One). Given $w_1, \ldots, w_n$ belonging to a vector space Y, we can define a linear operator $A: \mathbb{C}^n \to Y$ by assigning to $x = [x_1, \ldots, x_n]^\mathsf{T} \in \mathbb{C}^n$, the linear combination $Ax := \sum_{k=1}^n x_k w_k$. We see that A is full rank if and only if $w_1, \ldots, w_n$ are linearly independent. The modulation operator of Example 4.4, in which $w_1, \ldots, w_n$ are waveforms in some waveform space Y, has the structure considered here. A more familiar example is provided by the case $Y = \mathbb{C}^m$. Multiplication of x by an $m \times n$ matrix whose m-dimensional columns are $w_1, \ldots, w_n$ also has the operator structure considered here.

Theorem 4.8 (Finite-Dimensional Invertibility). *Let X and Y be finite-dimensional vector spaces, and let $A: X \to Y$ be a linear operator. If $\dim X = \dim Y$, then A is nonsingular if and only if A is onto.*

Remark. The point of the theorem is that under the hypotheses, we only have to check one of the two properties (one-to-one or onto) to see if the linear operator is invertible.

Proof. Suppose that A is nonsingular. Then $\dim \ker A = 0$, and the Rank–Nullity Theorem implies $\dim \operatorname{range} A = \dim X = \dim Y$. Since $\operatorname{range} A$ is a subspace of Y and they have the same dimension, they are equal (recall Section 2.3). Conversely, if A is onto, then $\operatorname{range} A = Y$. This implies that $\dim \operatorname{range} A = \dim Y = \dim X$. Combining this with the Rank–Nullity Theorem implies $\dim \ker A = 0$; hence, $\ker A$ is the zero subspace, and A is nonsingular. □

Example 4.9. The condition in the foregoing theorem that the spaces be finite dimensional is essential. Let X denote the set of all infinite sequences of the form $x = (x_1, x_2, \ldots)$. Consider the **right-shift operator** $A: X \to X$ defined by

$$Ax := (0, x_1, x_2, \ldots).$$

Then A is clearly nonsingular. However, A is not onto because there is no $x \in X$ such that $Ax = (1, 0, 0, \ldots)$.

We also observe here that the **left-shift operator** $B: X \to X$ defined by

$$B(y_1, y_2, \ldots) := (y_2, y_3, \ldots)$$

has the property that $BA = I$ (the identity operator on X), but $AB \neq I$ (the identity operator on Y).

4.3. Adjoint Operators

Let $A: X \to Y$ be a linear operator, where both X and Y are inner-product spaces. If for each $y \in Y$, there is a unique vector $A^* y \in X$ such that

$$\langle Ax, y \rangle_Y = \langle x, A^* y \rangle_X, \quad \text{for all } x \in X, \tag{4.3}$$

then we call A^* the **adjoint** of A. Here we include subscripts on the inner products to emphasize that they are defined on different spaces. In the sequel, we usually drop the subscripts. Note also that while $A: X \to Y$, the adjoint $A^*: Y \to X$.

It is easy to see that any vector $A^* y$ that satisfies (4.3) is unique. Suppose $\langle Ax, y \rangle = \langle x, A^* y \rangle$ and $\langle Ax, y \rangle = \langle x, \tilde{A} y \rangle$ hold for all $x \in X$. By subtraction we see that

$$\langle x, A^* y - \tilde{A} y \rangle = 0$$

holds for all x, including $x = A^* y - \tilde{A} y$. But then $\|A^* y - \tilde{A} y\|^2 = 0$.

Because $A^* y$ is the unique solution of (4.3), it is often easy to find $A^* y$ by inspection. It is also easy to show that A^* is linear (Problem 4.6).

Example 4.10. For the matrix operator of Example 4.1, if we equip $\mathbb{C}^n$ and $\mathbb{C}^m$ with their usual Euclidean inner products, we see that

$$\langle Ax, y \rangle_{\mathbb{C}^m} = y^{\mathsf{H}}(ax) = (a^{\mathsf{H}} y)^{\mathsf{H}} x = \langle x, a^{\mathsf{H}} y \rangle_{\mathbb{C}^n}.$$

Thus, $A^* y = a^{\mathsf{H}} y$.

Example 4.11. Consider the integral operator of Example 4.3. We let X denote the set of all finite-energy waveforms on $[a, b]$ with the inner product $\langle x_1, x_2 \rangle = \int_a^b x_1(\tau)\overline{x_2(\tau)}\, d\tau$. Similarly, we let Y denote the set of all finite-energy waveforms on $[c, d]$ with the inner product $\langle y_1, y_2 \rangle = \int_c^d y_1(t)\overline{y_2(t)}\, dt$. Then for reasonable functions $k(t, \tau)$,

$$\langle Ax, y \rangle = \int_c^d (Ax)(t)\,\overline{y(t)}\, dt$$

$$
\begin{aligned}
&= \int_c^d \left[\int_a^b k(t,\tau)x(\tau)\,d\tau \right] \overline{y(t)}\,dt \\
&= \int_a^b x(\tau) \overline{\left[\int_c^d \overline{k(t,\tau)} y(t)\,dt \right]}\,d\tau \\
&= \langle x, A^* y \rangle,
\end{aligned}
$$

where

$$
(A^*y)(\tau) := \int_c^d \overline{k(t,\tau)} y(t)\,dt, \quad \tau \in [a,b].
$$

Example 4.12. Consider the modulation operator of Example 4.4. Let X denote $\mathbb{C}^m$ equipped with the usual Euclidean inner product. We assume the signaling waveforms $w_1,\ldots,w_n$ belong to a vector space Y of finite-energy waveforms on the time interval $[0,T]$. We equip Y with the inner product $\langle y_1, y_2 \rangle = \int_0^T y_1(t)\overline{y_2(t)}\,dt$. To determine the adjoint of A, write

$$
\begin{aligned}
\langle Ax, y \rangle &= \int_0^T (Ax)(t)\,\overline{y(t)}\,dt \\
&= \int_0^T \left[\sum_{k=1}^n x_k w_k(t) \right] \overline{y(t)}\,dt \\
&= \sum_{k=1}^n x_k \overline{\int_0^T y(t)\overline{w_k(t)}\,dt} \\
&= \sum_{k=1}^n x_k \overline{\langle y, w_k \rangle},
\end{aligned}
$$

which we recognize as the Euclidean inner product of the column vector x and the column vector whose kth entry is $\langle y, w_k \rangle$. We conclude that

$$
A^*y = \begin{bmatrix} \langle y, w_1 \rangle \\ \vdots \\ \langle y, w_n \rangle \end{bmatrix}.
$$

We can now make the following observations:

(*i*) A^*y is the column vector of inner products that is equivalent to the projection of y onto $W := \operatorname{span}\{w_1,\ldots,w_n\}$ (cf. Corollary 3.9).

(*ii*) A^*y consists of the inner products produced by a digital communications receiver when the transmitter employs signaling waveforms $w_1,\ldots,w_n$. Since A is the modulation operator, we call A^* the **demodulation operator** (cf. Example 3.10).

(*iii*) A^*y can be realized by a bank of multiplier-integrators as shown in Figure 4.2 or by a bank of matched filters as shown in Figure 4.3.

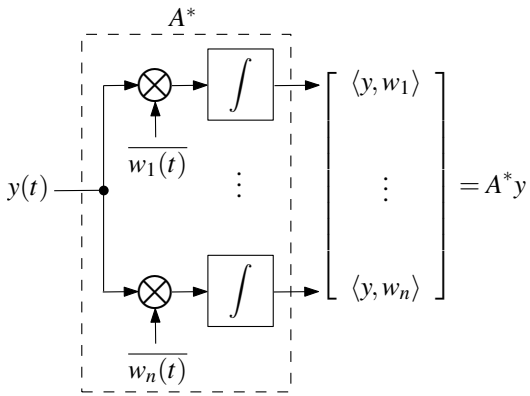


Figure 4.2. Implementation of the adjoint of the modulation operator of (4.2) and Figure 4.1.

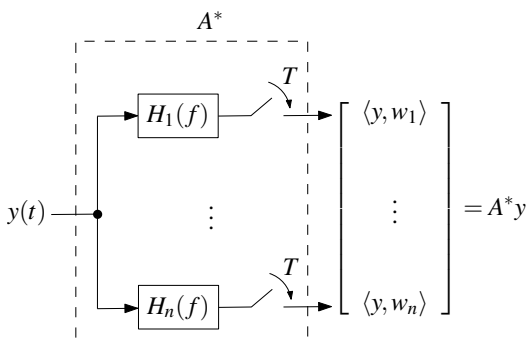


Figure 4.3. Matched filter equivalent of Figure 4.2. Here $H_k(f)$ is the Fourier transform of $\overline{w_k(T-t)}$.

Theorem 4.13. *Let X and Y be inner-product spaces. If $A: X \to Y$ is a linear operator whose adjoint $A^*: Y \to X$ exists, then*

(a) $\ker A^* = (\text{range}\, A)^\perp$.
(b) $(A^*)^* = A$.

(c) $\ker A = (\text{range} A^*)^\perp$.
(d) $\ker A^*A = \ker A$.
(e) $(\ker A)^\perp \supset \text{range} A^*$, *with equality if* $\text{rank} A^* < \infty$.
(f) $(\ker A^*)^\perp \supset \text{range} A$, *with equality if* $\text{rank} A < \infty$.
(g) *If* $\text{rank} A < \infty$ *and* A^* *is nonsingular, then* A *is onto.*
(h) *If* X *or* Y *is finite dimensional, then* $\text{rank} A$ *and* $\text{rank} A^*$ *are both finite, and* $\text{rank} A = \text{rank} A^*$.

Remark. The extension of parts (e) and (f) to Hilbert space can be found in Theorem 7.9. The extension of part (g) to Hilbert space can be found in Problem 7.15(b).

Proof. We leave the proof of parts (a)–(d) to Problem 4.12. To prove part (e), we begin by noting that from part (c), it follows that

$$(\ker A)^\perp = [(\text{range} A^*)^\perp]^\perp.$$

By Problem 3.13,

$$[(\text{range} A^*)^\perp]^\perp \supset \text{range} A^*.$$

However, if the range of A^* is finite dimensional, then by the remark following the Finite-Dimensional Projection Theorem, $X = \text{range} A^* \oplus (\text{range} A^*)^\perp$. This implies, by Problem 3.14, that

$$[(\text{range} A^*)^\perp]^\perp = \text{range} A^*.$$

To prove part (f), in part (e) replace A by A^* and apply part (b).

To prove part (g), apply part (f) and use $\ker A^* = \{0\}$ to write

$$\text{range} A = (\ker A^*)^\perp = \{0\}^\perp = Y.$$

To prove part (h), assume $\dim X < \infty$ (the proof for $\dim Y < \infty$ is similar and left to Problem 4.12). The finite dimensionality of X implies two facts. First, since $\ker A \subset X$, we must have $\dim \ker A < \infty$. Second, the Rank–Nullity Theorem applies and tells us that $\text{rank} A < \infty$ and that

$$\dim X = \dim \ker A + \dim \text{range} A.$$

Since $\text{range} A^* \subset X$, and X is finite dimensional, we see that $\text{rank} A^* < \infty$. Next, since $\dim \ker A < \infty$, the remark following the Finite-Dimensional Projection Theorem tells us that $X = \ker A \oplus (\ker A)^\perp$, which implies

$$\begin{aligned} \dim X &= \dim \ker A + \dim (\ker A)^\perp \\ &= \dim \ker A + \dim \text{range} A^*, \quad \text{by part (e).} \end{aligned}$$

Combining this with the previous display yields $\text{rank} A = \text{rank} A^*$. □

Example 4.14 (Digital Communication System). Consider a digital communication system in which a message $x \in \mathbb{C}^n$ is conveyed by transmitting the waveform Ax, where A is the modulation operator defined in Examples 4.4. The receiver applies the demodulation operator A^* to $y = Ax$ as pointed out in Example 4.12. When the signaling waveforms $w_1, \ldots, w_n$ are linearly independent, the modulation operator is full rank (i.e., $\ker A = \{0\}$), and, by Theorem 4.13, so is $A^*A: \mathbb{C}^n \to \mathbb{C}^n$. By the Finite-Dimensional Invertibility Theorem, A^*A is invertible. Hence, if $y = Ax$,

$$(A^*A)^{-1}(A^*y) = (A^*A)^{-1}(A^*A)x = x,$$

and we recover the message x. Such a system is shown in Figure 4.4. The receiver

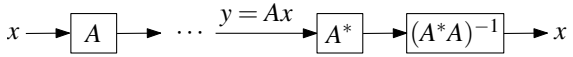


Figure 4.4. An ideal, noiseless communication system.

processing is greatly simplified if A^*A is diagonal. This is the case in **orthogonal frequency division multiplexing** (OFDM), in which

$$w_k(t) = e^{j2\pi(k/T)t}, \quad 0 \le t \le T.$$

Remark. The foregoing example can be slightly generalized. Before transmitting a vector $x \in \mathbb{C}^n$, first apply an invertible $n \times n$ matrix M; i.e., transmit $y = A(Mx)$. The receiver computes A^*y as before. Now observe that

$$M^{-1}(A^*A)^{-1}(A^*y) = M^{-1}(A^*A)^{-1}(A^*A)Mx = x.$$

We have now seen several examples of linear operators whose adjoints are easy to find, and we have seen several interesting properties of adjoints. Our next result gives a simple condition under which the adjoint is guaranteed to exist.

Theorem 4.15. *A linear operator mapping a finite-dimensional inner-product space into an arbitrary inner-product space has an adjoint.*

Proof. Suppose $A: X \to Y$, where X is a finite-dimensional inner-product space. Let $x_1, \ldots, x_n$ be an orthonormal basis for X. Then every $x \in X$ can be written in the form $x = \sum_{i=1}^n \langle x, x_i \rangle x_i$. So, for any $y \in Y$, we can write

$$\langle Ax, y \rangle = \left\langle A\left(\sum_{i=1}^n \langle x, x_i \rangle x_i\right), y \right\rangle = \sum_{i=1}^n \langle x, x_i \rangle \langle Ax_i, y \rangle = \left\langle x, \sum_{i=1}^n \langle y, Ax_i \rangle x_i \right\rangle.$$

Hence,

$$A^*y = \sum_{i=1}^n \langle y, Ax_i \rangle x_i. \qquad \square$$

4.3.1. An Operator without an Adjoint

Let X and Y denote the real-valued, continuous functions on $[0,2]$ and $[1,2]$, respectively, each equipped with the usual integral inner product over its corresponding interval. For $x \in X$, define $Ax \in Y$ by $(Ax)(t) := x(t)$ for $1 \le t \le 2$. In other words, A restricts the domain of x from $[0,2]$ to the subinterval $[1,2]$. This is illustrated for a particular waveform x in Figure 4.5.

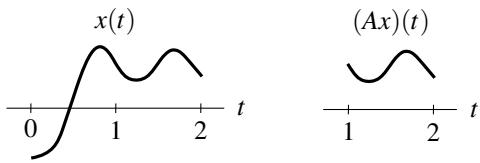


Figure 4.5. A waveform x on $[0,2]$ being restricted to the subinterval $[1,2]$ by the linear operator A.

To find the adjoint, we must solve $\langle Ax, y\rangle_Y = \langle x, A^*y\rangle_X$. For the spaces under consideration, this formula becomes

$$\int_1^2 (Ax)(t)y(t)\,dt = \int_0^2 x(t)(A^*y)(t)\,dt.$$

For the particular operator A here, we simplify the left-hand to obtain

$$\int_1^2 x(t)y(t)\,dt = \int_0^2 x(t)(A^*y)(t)\,dt. \tag{4.4}$$

Taking

$$(A^*y)(t) := \begin{cases} 0, & 0 \le t \le 1, \\ y(t), & 1 \le t \le 2, \end{cases}$$

solves (4.4), *but does not define a continuous function if* $y(1) \neq 0$; e.g., $y(t) = 1$ on $[1,2]$. However, this observation alone does not prove that A^*y does not exist for such y; it only proves that the above formula is not correct for such y. It is shown in the Notes at the end of the chapter[1] that the above formula must hold if A^*y exists. It follows that for any $y \in Y$ with $y(1) \neq 0$, A^*y does not exist.

4.3.2. Projection onto the Range of an Operator

Consider a linear operator $A: X \to Y$, where X and Y are inner-product spaces. Suppose that the range of A is a proper subspace of Y, and that a vector $y_0 \notin \text{range}\,A$, as shown in Figure 4.6. Then of course, there is no solution of the equation $Ax = y_0$. But how close can we get to y_0 with a vector of the form $Ax \in \text{range}\,A$? As illustrated in the figure, the projection of y_0 onto range$A$, denoted by $\widehat{y}_0$, would be the best we could do. The following lemma characterizes when this can be done.

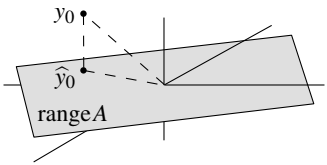


Figure 4.6. Projecting a point y_0 onto range$A = \{Ax : x \in X\}$. The projection is $\widehat{y}_0 = Ax_0$ for some $x_0 \in X$.

Lemma 4.16. *Consider a linear operator $A: X \to Y$, where X and Y are inner-product spaces. Assume that the adjoint $A^*: Y \to X$ exists. A given $y_0 \in Y$ can be projected onto the range of A if and only if there is a solution of the equation $A^*Ax = A^*y_0$. Furthermore, if any x_0 solves this equation, then Ax_0 is the projection of y_0 onto the range of A.*

Proof. The projection of y_0 onto range$A$, if it exists, will be a vector of the form $\widehat{y}_0 = Ax_0$ for some $x_0 \in X$. By the Orthogonality Principle, Ax_0 is the projection of y_0 if and only if

$$\langle y_0 - Ax_0, y\rangle = 0, \quad \text{for all } y \in \text{range}A,$$

if and only if

$$\langle y_0 - Ax_0, Ax\rangle = 0, \quad \text{for all } x \in X,$$

if and only if

$$\langle A^*(y_0 - Ax_0), x\rangle = 0, \quad \text{for all } x \in X,$$

if and only if

$$A^*(y_0 - Ax_0) = 0 \Leftrightarrow A^*Ax_0 = A^*y_0. \qquad \square$$

Remark. Whenever the range of A is finite dimensional, the projection of y_0 exists by the Finite-Dimensional Projection Theorem, and so a solution of $A^*Ax = A^*y_0$ is guaranteed to exist. For example, this is the case if either X or Y is finite dimensional. By Theorem 4.13(d), a solution of $A^*Ax = A^*y_0$ is unique if and only if A is nonsingular.

Example 4.17. Consider a linear operator $A: X \to Y$, where X and Y are inner-product spaces. If X is finite dimensional and A is nonsingular, show that the projection of y_0 onto the range of A is given by $A(A^*A)^{-1}A^*y_0$.

Solution. First note that Theorem 4.15 guarantees the existence of A^*. Next, since A is nonsingular, Theorem 4.13(d) tells us that A^*A is also nonsingular. Then since $\dim X < \infty$, the Finite-Dimensional Invertibility Theorem implies $(A^*A)^{-1}$ exists.

Hence, the expression $x_0 := (A^*A)^{-1}A^*y_0$ is well defined and satisfies $A^*Ax_0 = A^*y_0$. Therefore, by Lemma 4.16, $Ax_0 = A(A^*A)^{-1}A^*y_0$ is the required projection.

4.3.3. Minimum-Norm Solutions of Linear Equations

If the equation $Ax = y_0$ has a solution, say x_0, but A is singular, then there are nonzero vectors in $\ker A$, and for all nonzero $w \in \ker A$, $A(w + x_0) = Aw + Ax_0 = 0 + y_0 = y_0$. Hence, there are multiple solutions of $Ax = y_0$.

Lemma 4.18. *Consider a linear operator $A: X \to Y$, where X is an inner-product space. Assume that $X = \ker A \oplus (\ker A)^\perp$. If x_0 is any solution of $Ax = y_0$, then the projection of x_0 onto $(\ker A)^\perp$, denoted by $\widetilde{x}_0$, is the unique minimum-norm solution of $Ax = y_0$. See Figure* 4.7.

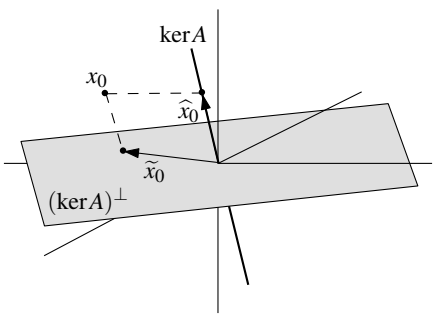


Figure 4.7. Decomposition of solution x_0 into its projection onto $\ker A$ and $(\ker A)^\perp$ to write $x_0 = \widehat{x}_0 + \widetilde{x}_0$.

Example 4.19 (An Equation without a Minimum-Norm Solution). The assumption in Lemma 4.18 that $X = \ker A \oplus (\ker A)^\perp$ is necessary. To see why, recall the restriction operator described in Section 4.3.1 that did not have a adjoint. Suppose $y_0 \in Y$ is taken as $y_0(t) = 1$ for $1 \le t \le 2$. Let's try to solve $Ax = y_0$. For example, the waveform x shown in Figure 4.8 solves this equation. Of course, the energy of this solution can be reduced by moving t_0 closer to 1. In general, the energy of any solution can be reduced by "making it drop to zero faster." Hence, there is no solution of minimum norm. Recalling Problem 3.18, we see that $\ker A = W$, $(\ker A)^\perp = W^\perp$, and we know that $X \neq W \oplus W^\perp$.

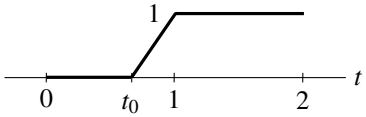


Figure 4.8. The solution $x(t)$ of $Ax = y_0$ in Example 4.19.

Proof of Lemma 4.18. Write $x_0 = \widehat{x}_0 + \widetilde{x}_0$, where $\widehat{x}_0 \in \ker A$, and $\widetilde{x}_0 \in (\ker A)^\perp$. Since $A\widehat{x}_0 = 0$, we have $y_0 = Ax_0 = A(\widehat{x}_0 + \widetilde{x}_0) = A\widehat{x}_0 + A\widetilde{x}_0 = A\widetilde{x}_0$. In other words, $\widetilde{x}_0$ also solves $Ax = y_0$. Now suppose x_1 is any other solution. Then by the same argument, $\widetilde{x}_1 \in (\ker A)^\perp$ also satisfies $A\widetilde{x}_1 = y_0$. Therefore, $A(\widetilde{x}_1 - \widetilde{x}_0) = A\widetilde{x}_1 - A\widetilde{x}_0 = y_0 - y_0 = 0$. In other words, $\widetilde{x}_1 - \widetilde{x}_0 \in \ker A \cap (\ker A)^\perp$. The only vector orthogonal to itself is the zero vector, and so $\widetilde{x}_1 = \widetilde{x}_0$. In other words, there is only one vector in $(\ker A)^\perp$ that solves $Ax = y_0$. Furthermore, since $\|x_0\|^2 = \|\widehat{x}_0 + \widetilde{x}_0\|^2 = \|\widehat{x}_0\|^2 + \|\widetilde{x}_0\|^2 \geq \|\widetilde{x}_0\|^2$, we see that $\widetilde{x}_0$ is the solution of minimum norm. □

Lemma 4.20. *Let $A: X \to Y$ be a linear operator between inner-product spaces X and Y. If $A^*: Y \to X$ exists and has finite rank, then $X = \ker A \oplus (\ker A)^\perp$.*

Proof. Because A^* has finite rank, the Finite-Dimensional Projection Theorem allows us to write $X = \text{range}\, A^* \oplus (\text{range}\, A^*)^\perp$. The finite-rank assumption also allows us to use Theorem 4.13(e) to write $\text{range}\, A^* = (\ker A)^\perp$. By Theorem 4.13(c), $(\text{range}\, A^*)^\perp = \ker A$. □

Example 4.21. Consider a linear operator $A: X \to Y$, where X and Y are inner-product spaces. Assume that Y is finite dimensional and that the adjoint A^* exists and is nonsingular. Show that the vector $A^*(AA^*)^{-1}y_0$ is well defined, belongs to $(\ker A)^\perp$, and solves $Ax = y_0$.

Solution. Since A^* is nonsingular, so is AA^* by Theorem 4.13(d). Since Y is finite dimensional, the Finite-Dimensional Invertibility Theorem tells us that the vector $A^*(AA^*)^{-1}y_0$ is well defined, and this vector obviously solves $Ax = y_0$. Furthermore, this vector lies in $\text{range}\, A^* = (\ker A)^\perp$ by Theorem 4.13(e).

Remark. Is the vector $A^*(AA^*)^{-1}y_0$ of the preceding example the unique minimum-norm solution of $Ax = y_0$? Yes, because Lemma 4.20 allows us to apply Lemma 4.18.

4.3.4. The Pseudoinverse

Consider a linear operator $A: X \to Y$, where X and Y are inner-product spaces. Assume that[a]

$$X = \ker A \oplus (\ker A)^{\perp} \quad \text{and} \quad Y = \text{range}\, A \oplus (\text{range}\, A)^{\perp}. \tag{4.5}$$

The **pseudoinverse** $A^{\dagger}: Y \to X$ is defined as follows. For $y_0 \in Y$, let $\widehat{y}_0$ denote the projection of y_0 onto range$A$. This projection always exists because of the assumed direct-sum decomposition of Y. Then $A^{\dagger} y_0$ is defined as the minimum-norm solution of the equation $Ax = \widehat{y}_0$. Since $\widehat{y}_0 \in \text{range}\, A$, this equation has solutions, and we can find the one of minimum norm because of the assumed direct-sum decomposition of X and Lemma 4.18.

Example 4.22. If X is finite dimensional[b] and A is nonsingular, show that $A^{\dagger} = (A^*A)^{-1}A^*$.

Solution. By Example 4.17, $\widehat{y}_0 = A(A^*A)^{-1}A^* y_0$. If we can show that the vector $(A^*A)^{-1}A^* y_0 \in (\ker A)^{\perp}$, then this vector is the required unique minimum-norm solution of $Ax = \widehat{y}_0$. Now, $A^* y \in \text{range}\, A^* \subset (\ker A)^{\perp}$ by Theorem 4.13(e). Also, since A is nonsingular, the decomposition $X = \ker A \oplus (\ker A)^{\perp}$ implies that $X = (\ker A)^{\perp}$. Hence, the mapping $(A^*A)^{-1}$ takes the vector $A^* y_0$ into $X = (\ker A)^{\perp}$. In other words, the vector $(A^*A)^{-1}A^* y_0 \in (\ker A)^{\perp}$ as required.

When X and Y are Euclidean spaces and $Ax = ax$ for some full-rank matrix a, it is more efficient in MATLAB to use the expression `a\y` than `(a'*a)^(-1)*a'*y`. If `a` is singular, you should use `pinv(a)*y` to get the minimum-norm solution of `ax=y`.

Example 4.23. Assume $X = \ker A \oplus (\ker A)^{\perp}$, that Y is finite dimensional, and that A^* exists and is nonsingular. Show that $A^{\dagger} = A^*(AA^*)^{-1}$.

Solution. Since $\dim Y < \infty$, the right-hand decomposition in (4.5) holds. Since A^* is nonsingular, combining Theorem 4.13(a) and the decomposition $Y = \text{range}\, A \oplus (\text{range}\, A)^{\perp}$ tells us that $Y = \text{range}\, A$. Hence, for every $y_0 \in Y$, there is a solution of $Ax = y_0$. By Example 4.21 and the remark following it, the minimum-norm solution is $A^*(AA^*)^{-1} y_0$.

---

[a] When the decomposition $Y = \text{range}\, A \oplus (\text{range}\, A)^{\perp}$ does not hold, $A^{\dagger} y$ is defined only for those y that can be projected onto range$A$. This issue is discussed in more detail in Section 7.5 on the singular-value decomposition.

[b] Since X is finite dimensional, so are $\ker A$ and range$A$. Hence, the decompositions in (4.5) hold on account of the Finite-Dimensional Projection Theorem and the remark following it.

Example 4.24. Let $P_{(\ker A)^\perp}$ denote the projection onto $(\ker A)^\perp$, and let $P_{\text{range}\,A}$ denote the projection onto range$A$. Show that

$$A^\dagger A = P_{(\ker A)^\perp} \quad \text{and} \quad AA^\dagger = P_{\text{range}\,A}.$$

Solution. Let $x_0 \in X$ be given, and put $y_0 := Ax_0$. By Lemma 4.18, $\widetilde{x}_0$, the projection of x_0 onto $(\ker A)^\perp$, is the minimum-norm solution of $Ax = y_0$. Furthermore, since $y_0 := Ax_0 \in \text{range}A$, $y_0 = \widehat{y}_0$, and thus, $\widetilde{x}_0$ is also the minimum-norm solution of $Ax = \widehat{y}_0$, which is the definition of $A^\dagger y_0$. Hence, $\widetilde{x}_0 = A^\dagger y_0 = A^\dagger Ax_0$.

Let $y_0 \in Y$ be given. By definition of the pseudoinverse, $A^\dagger y_0$ is a solution of $Ax = \widehat{y}_0$; i.e., $A(A^\dagger y_0) = \widehat{y}_0$. This formula says that $AA^\dagger y_0 = P_{\text{range}\,A} y_0$.

4.4. Self-Adjoint Linear Operators

A linear operator A mapping an inner-product space X to itself is said to be **self adjoint** if $\langle Ax, y\rangle = \langle x, Ay\rangle$ for all $x, y \in X$. In other words, A is self adjoint if $A = A^*$. We have already seen in Problem 3.15 that projection operators and Householder transformations are self adjoint. From Example 4.10 with $m = n$, we see that a complex matrix operator is self adjoint under the usual Euclidean inner product if the matrix is equal to its complex-conjugate transpose. Similarly, a real matrix operator is self adjoint if it is symmetric. In Example 4.11, if $a = c$ and $b = d$, then A is self adjoint if $\overline{k(\tau,t)} = k(t,\tau)$.

If an operator A is both self adjoint[c] and satisfies $\langle Ax, x\rangle \geq 0$ for all $x \in X$, we say that A is **positive semidefinite** and denote this by $A \geq 0$. A positive semidefinite operator that satisfies $\langle Ax, x\rangle > 0$ for all nonzero x is called **positive definite**, denoted by $A > 0$. A positive-definite operator must be nonsingular.

We can use a positive-definite operator to define a new inner product on X with the formula $\langle x, y\rangle_A := \langle Ax, y\rangle$. To verify that this formula satisfies the properties of an inner product, we proceed as follows. First, since $A = A^*$,

$$\overline{\langle x, y\rangle_A} = \overline{\langle Ax, y\rangle} = \overline{\langle x, Ay\rangle} = \langle Ay, x\rangle = \langle y, x\rangle_A.$$

Next, $\langle x, x\rangle_A = \langle Ax, x\rangle \geq 0$ and equal to zero if and only if $x = 0$. Finally, it is obvious that $\langle x, y\rangle_A = \langle Ax, y\rangle$ is linear in x.

Proposition 4.25. *If a self-adjoint operator is invertible, then its inverse is self adjoint.*

Proof. Write $\langle A^{-1}y, x\rangle = \langle A^{-1}y, A(A^{-1}x)\rangle = \langle A(A^{-1}y), A^{-1}x\rangle = \langle y, A^{-1}x\rangle$. □

[c] The requirement that A be self adjoint is superfluous in a complex inner product space. See Problem 4.20.

Proposition 4.26. *If $A: X \to Y$ is a nonsingular linear operator between inner-product spaces X and Y and has adjoint A^*, then A^*A is positive definite. If A^*A is invertible, then $(A^*A)^{-1}$ is self adjoint and positive definite.*

Proof. Concerning the positive-definiteness of A^*A, write $\langle A^*Ax, x\rangle = \langle Ax, Ax\rangle = \|Ax\|^2 \geq 0$, with equality if and only if $Ax = 0$; however, since A is nonsingular, $Ax = 0$ if and only if $x = 0$. Hence, A^*A is positive definite. Since A^*A is self adjoint, if it is invertible, then by the preceding proposition, its inverse is self adjoint. Finally,

$$\begin{aligned}\left\langle (A^*A)^{-1}x, x\right\rangle &= \left\langle (A^*A)^{-1}x, (A^*A)(A^*A)^{-1}x\right\rangle \\ &= \left\langle A(A^*A)^{-1}x, A(A^*A)^{-1}x\right\rangle \\ &= \|A(A^*A)^{-1}x\|^2 \geq 0,\end{aligned}$$

and since A is nonsingular, equal to zero if and only if $(A^*A)^{-1}x = 0$, which happens if and only if $x = 0$. □

4.5. Alternative Inner Products

Suppose that Q is a self-adjoint, positive-definite linear operator on an inner-product space X. For $x_1, x_2 \in X$, put $\langle x_1, x_2\rangle_Q := \langle Qx_1, x_2\rangle$. It is easy to check that $\langle\cdot,\cdot\rangle_Q$ satisfies the properties of an inner product. Hence, we also have a corresponding Q-norm, $\|x\|_Q := \langle x, x\rangle_Q^{1/2} = \langle Qx, x\rangle^{1/2}$.

Example 4.27. Recall that the equation of an **ellipse** is $x^2/a^2 + y^2/b^2 = 1$. This can be rewritten in matrix-vector form as

$$\begin{bmatrix} x & y\end{bmatrix}\begin{bmatrix} 1/a^2 & 0 \\ 0 & 1/b^2\end{bmatrix}\begin{bmatrix} x \\ y\end{bmatrix} = 1.$$

If we denote the above 2×2 positive-definite matrix by Q, and we let $\langle\cdot,\cdot\rangle$ denote the standard inner product on $\mathbb{R}^2$, then this equation can also be written as

$$\left\langle Q\begin{bmatrix} x \\ y\end{bmatrix}, \begin{bmatrix} x \\ y\end{bmatrix}\right\rangle = 1.$$

It follows that under this Q-norm on $\mathbb{R}^2$, balls are ellipse shaped.

Example 4.28. The definition of projection of a point onto a set depends on the norm being used to measure the distance from the point to the set. Consider the ellipse-shaped ball,

$$B_Q := \{x : \|x\|_Q \leq 1\}.$$

Now fix a point $y \notin B_Q$. If we measure the distance from y to B_Q using the Q-norm, then the projection of y onto B_Q is easily seen to be $y/\|y\|_Q$. However, if we measure the distance from y to B_Q using the original norm corresponding to $\langle \cdot, \cdot \rangle$, then the projection can be found using methods developed in the next chapter (see Examples 5.22 and 5.23).

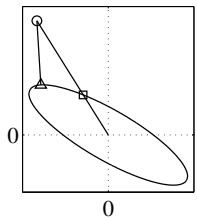


Figure 4.9. The point y (circle) has Q-norm projection given by the square and Euclidean-norm projection given by the triangle.

Let A be a linear operator mapping X into another inner-product space Y. Recalling that the adjoint depends on the inner product, let $\tilde{A}$ denote the adjoint of A when the Q-inner product is used on X. In other words, $\tilde{A}$ should satisfy

$$\langle Ax, y \rangle = \langle x, \tilde{A}y \rangle_Q,$$

where the inner product on the left is the inner product on Y. It is easy to show that $\tilde{A} = Q^{-1}A^*$ if A^* and Q^{-1} exist; e.g., if X is finite dimensional (apply Theorem 4.8 to Q and Theorem 4.15 to A).

To generalize the foregoing, let R be a self-adjoint, positive-definite linear operator on Y, and define a new inner product on Y by $\langle y_1, y_2 \rangle_R := \langle Ry_1, y_2 \rangle$. Now let $\tilde{A}$ solve

$$\langle Ax, y \rangle_R = \langle x, \tilde{A}y \rangle_Q.$$

It is easy to check that

$$\tilde{A} = Q^{-1}A^*R.$$

Let A be a nonsingular linear operator from a finite-dimensional inner-product space X into an inner-product space Y. By Example 4.17, the projection operator onto the range of A is given by $A(A^*A)^{-1}A^*$. Now suppose that we measure distance in Y with the R-norm. If we also use the Q-inner product on X, the projection operator is

$$A(\tilde{A}A)^{-1}\tilde{A} = A(A^*RA)^{-1}A^*R,$$

which does not depend on Q.

4.5.1. Inner Products of Matrices

Example 4.29 (Best Linear Unbiased Estimate (BLUE)). In many situations, we observe a random vector U, whose distribution is governed by an unknown parameter ξ. Since we do not know ξ, we estimate it by computing a function of the observed random vector U, and we denote the estimate by $\widehat{\Xi}$. If the estimate has the property that $\mathsf{E}[\widehat{\Xi}] = \xi$ for all possible values of ξ, we say that the estimate is **unbiased**. In this example, we consider an observation of the form $U = G\xi + W$, where W is a random noise vector with zero-mean and known, positive-definite, covariance matrix $Q := \mathsf{E}[WW^{\mathsf{H}}]$, and the matrix G is also known. The deterministic, unknown parameter ξ is to be estimated using a *linear* system of the form $\widehat{\Xi} = XU$. Since

$$\mathsf{E}[\widehat{\Xi}] = \mathsf{E}[XU] = \mathsf{E}[X(G\xi + W)] = XG\xi + X\mathsf{E}[W] = XG\xi$$

is equal to ξ for all ξ if and only if $XG = I$, we restrict attention to such matrices X. Next, among such matrices, we seek the one that minimizes the mean-squared error $\mathsf{E}[\|\xi - \widehat{\Xi}\|^2] = \mathsf{E}[\|\xi - XU\|^2]$. Since

$$\xi - XU = \xi - X(G\xi + W) = (I - XG)\xi - XW,$$

for X satisfying $XG = I$, the mean-squared error simplifies to[d]

$$\begin{aligned}\mathsf{E}[\|\xi - XU\|^2] = \mathsf{E}[\|XW\|^2] = \mathsf{E}[\mathrm{tr}\{(XW)^{\mathsf{H}}(XW)\}] &= \mathsf{E}[\mathrm{tr}\{(XW)(XW)^{\mathsf{H}}\}] \\ &= \mathrm{tr}\,\mathsf{E}[XWW^{\mathsf{H}}X^{\mathsf{H}}] \\ &= \mathrm{tr}(X\mathsf{E}[WW^{\mathsf{H}}]X^{\mathsf{H}}) \\ &= \mathrm{tr}(XQX^{\mathsf{H}}).\end{aligned}$$

It follows that we must solve the problem

$$\min_X \mathrm{tr}(XQX^{\mathsf{H}}) \quad \text{subject to} \quad XG = I. \tag{4.6}$$

We could try to solve this using Lagrange multipliers, or we could treat $\mathrm{tr}(XQX^{\mathsf{H}})$ as the "norm" of the matrix X and look for the minimum "norm" solution of the linear equation $XG = I$, which we can solve using the methods of this chapter. With this as motivation, we now turn to the details.

---

[d] Recall that the **trace** of a square matrix is the sum of its diagonal elements. If A is an $m \times n$ matrix and B is an $n \times m$ matrix, then AB is an $m \times m$ square matrix, BA is an $n \times n$ square matrix, and $\mathrm{tr}(AB) = \mathrm{tr}(BA)$. For any column vector x, the inner product $x^{\mathsf{H}}x$ is a scalar and therefore equal to its **trace**. Hence, $x^{\mathsf{H}}x = \mathrm{tr}(x^{\mathsf{H}}x) = \mathrm{tr}(xx^{\mathsf{H}})$.

If X and Y are matrices of the same size, then the standard inner product of X and Y is

$$\langle X, Y \rangle := \operatorname{tr}(XY^{\mathsf{H}}),$$

where tr denotes the **trace** (see previous footnote).

Now suppose that Q is a square matrix satisfying $Q^{\mathsf{H}} = Q$. Then Q is self adjoint when regarded as a linear operator on column vectors (recall Example 4.10). Similarly, Q is positive definite if $\langle Qx, x \rangle > 0$ for all nonzero x; in terms of matrices and column vectors, this says that $x^{\mathsf{H}}Qx > 0$ for all nonzero column vectors x. A matrix Q with this property is said to be a **positive-definite *matrix***. We now ask whether a positive-definite matrix Q has the property that $\operatorname{tr}(QXX^{\mathsf{H}}) > 0$ for all nonzero *matrices* X. In other words, if Q is positive definite when regarded as a linear operator on column vectors, does it follow that Q is positive definite when regarded as a linear operator on matrices? The answer is "yes," and can be seen as follows. Let X_k denote the kth column of a matrix X. Then for $m \times n$ matrices X, if Q is $m \times m$, we have $\operatorname{tr}(QXX^{\mathsf{H}}) = \operatorname{tr}(X^{\mathsf{H}}QX)$, which is equal to

$$\operatorname{tr}\left(\left[\begin{array}{c} X_1^{\mathsf{H}} \\ \hline \vdots \\ \hline X_n^{\mathsf{H}} \end{array}\right] Q[X_1|\cdots|X_n]\right) = \operatorname{tr}\left(\begin{bmatrix} X_1^{\mathsf{H}}QX_1 & \cdots & X_1^{\mathsf{H}}QX_n \\ \vdots & \ddots & \vdots \\ X_n^{\mathsf{H}}QX_1 & \cdots & X_n^{\mathsf{H}}QX_n \end{bmatrix}\right) = \sum_{j=1}^{n} X_j^{\mathsf{H}}QX_j.$$

It can similarly be shown that if Q is $n \times n$, then $\langle XQ, X \rangle > 0$ for all nonzero matrices X.

Consider a mapping $A: \mathbb{C}^{m\times n} \to \mathbb{C}^{p\times q}$ of the form $AX := FXG$ for suitably sized matrices F and G. Since

$$\langle AX, Y \rangle = \operatorname{tr}(FXGY^{\mathsf{H}}) = \operatorname{tr}(XGY^{\mathsf{H}}F) = \operatorname{tr}(X[F^{\mathsf{H}}YG^{\mathsf{H}}]^{\mathsf{H}}),$$

it is easy to show that that $A^*Y = F^{\mathsf{H}}YG^{\mathsf{H}}$ with respect to the standard matrix inner products. However, if $Q = Q^{\mathsf{H}}$ and $x^{\mathsf{H}}Qx > 0$ for all nonzero column vectors x, then with respect to the Q-inner product on $\mathbb{C}^{m\times n}$, $\langle X_1, X_2 \rangle_Q := \langle X_1Q, X_2 \rangle = \operatorname{tr}(X_1QX_2^{\mathsf{H}})$, it is easy to show that

$$A^*Z = F^{\mathsf{H}}ZG^{\mathsf{H}}Q^{-1}.$$

Now, given a matrix Y_0, consider the problem of finding the minimum Q-norm solution of $FXG = Y_0$.[e] With our definition of A, we know from Example 4.21 that the optimal solution is $X = A^*Z$, where Z solves $AA^*Z = Y_0$. So, we must solve

$$F(F^{\mathsf{H}}ZG^{\mathsf{H}}Q^{-1})G = Y_0.$$

[e] To solve (4.6), take F and Y_0 to be identity matrices.

Assuming F^{H} and G are nonsingular, we have (cf. Theorems 4.13(d) and 4.8)

$$Z = (FF^{\mathsf{H}})^{-1}Y_0(G^{\mathsf{H}}Q^{-1}G)^{-1},$$

and then

$$X = A^*Z = F^{\mathsf{H}}(FF^{\mathsf{H}})^{-1}Y_0(G^{\mathsf{H}}Q^{-1}G)^{-1}G^{\mathsf{H}}Q^{-1}.$$

We can now write the solution of (4.6) by taking F and Y_0 as identity matrices; i.e.,

$$X = (G^{\mathsf{H}}Q^{-1}G)^{-1}G^{\mathsf{H}}Q^{-1}.$$

Notes

Note 4.1. To show that $(A^*y)(t) = 0$ on $[0,1]$, we analyze (4.4) when $x(t) = 0$ on $[1,2]$. Then, to show $(A^*y)(t) = y(t)$ on $[1,2]$, we analyze (4.4) when $x(t) = 0$ on $[0,1]$.

Suppose $x(t) = 0$ on $[1,2]$, and observe that (4.4) simplifies to

$$0 = \int_0^1 x(t)(A^*y)(t)\,dt.$$

If we could take $x(t) := (A^*y)(t)$ on $[0,1]$ and $x(t) = 0$ on $[1,2]$, this would show that $0 = \int_0^1 |(A^*y)(t)|^2\,dt$, and it would follow that $(A^*y)(t) = 0$ on $[0,1]$. However, since this choice of x may not be continuous at $t = 1$, we have to argue more carefully. Fix any $0 < t_0 < 1$ and write the above equation as

$$0 = \int_0^{t_0} x(t)(A^*y)(t)\,dt + \int_{t_0}^1 x(t)(A^*y)(t)\,dt.$$

Now take $x(t) := (A^*y)(t)$ on $[0,t_0]$, and on $(t_0,1]$, let $x(t)$ be the straight line connecting $\big(t_0,(A^*y)(t_0)\big)$ to $(1,0)$ as illustrated in Figure 4.10; note that we maintain $x(t) = 0$ on $[1,2]$. Then the preceding equation becomes

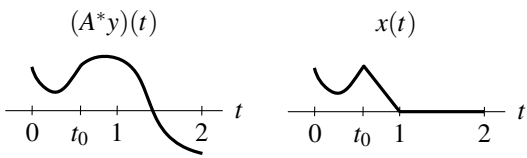


Figure 4.10. Given A^*y, construct x that is continuous on $[0,2]$, equal to A^*y on $[0,t_0]$, and zero on $[1,2]$.

$$0 = \int_0^{t_0} \big|(A^*y)(t)\big|^2\,dt + \int_{t_0}^1 x(t)(A^*y)(t)\,dt. \tag{4.7}$$

Observe that the second integral is bounded by

$$\left| \int_{t_0}^{1} x(t)(A^*y)(t)\,dt \right| \le (1-t_0) \max_{0\le t\le 1} \left|(A^*y)(t)\right|^2,$$

which tends to zero as $t_0 \to 1$. Hence, taking $t_0 \to 1$ in (4.7) shows that

$$0 = \int_0^1 \left|(A^*y)(t)\right|^2 dt,$$

and it follows that $(A^*y)(t) = 0$ on $[0,1]$.

Now suppose instead that $x(t) = 0$ on $[0,1]$. Then (4.4) becomes

$$\int_1^2 x(t)y(t)\,dt = \int_1^2 x(t)(A^*y)(t)\,dt,$$

which we rearrange as

$$\int_1^2 x(t)\big[y(t) - (A^*y)(t)\big]\,dt = 0. \tag{4.8}$$

We would like to take $x(t) = y(t) - (A^*y)(t)$ on $[1,2]$ and $x(t) = 0$ on $[0,1]$ because then (4.8) would imply $(A^*y)(t) = y(t)$ on $[1,2]$. However, this choice of x may not be continuous at $t = 1$. Instead, fix any $1 < t_1 < 2$, and consider the waveform $x(t) := y(t) - (A^*y)(t)$ on $[t_1,2]$, and on $[1,t_1)$, let $x(t)$ be the straight line connecting $(1,0)$ to $\big(t_1, y(t_1) - (A^*y)(t_1)\big)$. We also put $x(t) = 0$ on $[0,1]$. Then (4.8) becomes

$$\int_1^{t_1} x(t)\big[y(t) - (A^*y)(t)\big]\,dt + \int_{t_1}^2 \left|y(t) - (A^*y)(t)\right|^2 dt = 0.$$

Similar to the derivation in the previous paragraph, we can show that the first integral tends to zero as $t_1 \to 1$. It follows that $(A^*y)(t) = y(t)$ on $[1,2]$.

Problems

1. For $x = [x_1, \ldots, x_n]^{\mathsf{T}} \in \mathbb{R}^n$, define $Ax \in \mathbb{R}^{n-1}$ by

$$(Ax)_i := x_{i+1} - x_i, \quad i = 1, \ldots, n-1.$$

   (a) Find $\ker A$.
   (b) Determine whether or not A is onto.

2. Show that the kernel and the range of a linear operator are subspaces.

3. Show that $\{Ax_{r+1},\ldots,Ax_n\}$ used in the proof of the Rank–Nullity Theorem is a basis for the range of A.

4. Let $A\colon X \to Y$ be a linear operator. Given $y \in Y$, consider the problem of solving $Ax = y$. Show that the solution set $S := \{x \in X : Ax = y\}$ is affine.

5. Let X and Y be vector spaces, with $A\colon X \to Y$ and $B\colon Y \to X$ being linear operators.

   (a) If $BA = I$, show that A is nonsingular.

   (b) If $BA = I$ and A is onto, show that $AB = I$. (Observe that we are using I for two different objects. When we write $BA = I$, the symbol I denotes the identity operator on X, but when we write $AB = I$, the symbol I denotes the identity operator on Y.)

   (c) If $BA = I$ and X and Y are finite dimensional and have the same dimension, show that $AB = I$.

   Remark. The assumption in part (b) that A is onto is necessary. For example, if

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix},$$

   then BA is the 2×2 identity matrix, but

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

   is *not* the 3×3 identity matrix. Notice that A is not onto since there is no $x = [u,v]^{\mathsf{T}} \in \mathbb{R}^2$ such that $Ax = [0,0,1]^{\mathsf{T}} \in \mathbb{R}^3$.

6. Show that if $\langle Ax, y\rangle = \langle x, A^*y\rangle$ holds for all x and y, then A^* is linear.

7. Let X denote the real vector space of all continuous, real-valued waveforms on the interval $[0,1]$ with inner product $\langle x_1, x_2\rangle := \int_0^1 x_1(t)x_2(t)dt$, $x_1, x_2 \in X$. Define the linear operator $A : X \to \mathbb{R}$ by

$$Ax := \int_0^1 t^2 x(t)dt.$$

   Treat $\mathbb{R}$ as an inner product space with inner product $\langle y_1, y_2\rangle := y_1 y_2$ (i.e., ordinary multiplication of real numbers). Find a formula for $(A^*y)(t)$, and when $y = 3$, sketch a graph of $(A^*y)(t)$.

8. Let X denote the set of real-valued, finite-energy waveforms on $[0,1]$, with the usual inner product, $\langle u,v\rangle := \int_0^1 u(t)v(t)\,dt$ for $u,v \in X$. Define $A\colon X \to X$ by

$$(Ax)(t) := \int_0^t x(\theta)\,d\theta, \quad 0 \le t \le 1.$$

Find $(A^*y)(\theta)$ for $\theta \in [0,1]$.

9. Let $X = Y$ denote the set of all infinitely differentiable, real-valued waveforms x on $[0,T]$ such that x and all its derivatives vanish at $t = 0$ and at $t = T$. Let the inner product be given by $\langle x_1,x_2\rangle := \int_0^T x_1(t)x_2(t)\,dt$. Let $(Ax)(t) := \dot{x}(t)$, where $\dot{x}(t)$ is the usual derivative of $x(t)$ with respect to t. For $y \in Y$, find $(A^*y)(t)$.

10. Let $b_1,\ldots,b_m$ be vectors in a real inner product space Z, and define $B\colon Z \to \mathbb{R}^m$ by

$$Bz := \begin{bmatrix} \langle z,b_1\rangle \\ \vdots \\ \langle z,b_m\rangle \end{bmatrix}.$$

Show that if $\mathbb{R}^m$ is equipped with the standard Euclidean inner product, then

$$B^*\lambda = \sum_{i=1}^m \lambda_i b_i.$$

11. For the operator in Problem 4.1, find its adjoint if $\mathbb{R}^n$ and $\mathbb{R}^{n-1}$ are each equipped with the standard inner product.

12. Let X and Y be inner-product spaces, and let $A\colon X \to Y$ be a linear operator whose adjoint $A^*\colon Y \to X$ exists.

    (a) Show that $\ker A^* = (\text{range}\,A)^\perp$.
    (b) Show that $(A^*)^* = A$.
    (c) Show that $\ker A = (\text{range}\,A^*)^\perp$.
    (d) Show that $\ker A^*A = \ker A$.
    (e) Show that if $\text{rank}\,A^* < \infty$ and A is nonsingular, then A^* is onto. *Hint:* Use the proof of Theorem 4.13(g) as a guide.
    (f) Show that if $\dim Y < \infty$, then $\text{rank}\,A$ and $\text{rank}\,A^*$ are both finite and equal to each other. *Hint:* Use the proof of Theorem 4.13(h) as a guide.

13. Let $A\colon X \to Y$ be a linear transformation between inner-product spaces X and Y. Assume that the adjoint A^* exists. If A^* is onto, show that A is nonsingular.

14. Let $A: X \to Y$ be a linear transformation from a vector space X to a vector space Y. Let G be a linearly independent subset of X. If A is nonsingular, show that $H := A(G) := \{Ag : g \in G\}$ is linearly independent.

15. Let X and Y be inner-product spaces, and let $A: X \to Y$ be a linear operator. We say that A is **inner-product preserving** if $\langle Ax_1, Ax_2 \rangle = \langle x_1, x_2 \rangle$ for all $x_1, x_2 \in X$.

    (a) Show that A is inner-product preserving if and only if A is **norm preserving** in the sense that $\|Ax\| = \|x\|$ for all $x \in X$. *Hint:* The **polarization identity** of Problem 3.4 may be helpful.

    (b) If A has an adjoint, show that A is inner-product preserving if and only if $A^*A = I$.

    (c) Show that a norm-preserving operator is nonsingular.

    (d) If A is inner-product-preserving and onto, show that A is invertible and then that $\langle Ax, y \rangle = \langle x, A^{-1}y \rangle$. Hence, such an operator has an adjoint with $A^* = A^{-1}$. You may ***not*** use part (b) to solve part (d).

    (e) If A has an adjoint and $A^*A = I$, determine whether or not A is invertible.

16. Let $A: X \to Y$ be a linear operator between inner-product spaces, and assume that $A^*: Y \to X$ and $A^{-1}: Y \to X$ exist. Prove the following:

    (a) If $(A^*)^{-1}$ exists, then $(A^{-1})^* = (A^*)^{-1}$.

    (b) If $(A^{-1})^*$ exists, then $(A^*)^{-1} = (A^{-1})^*$.
    Hint for part (b): To show that $(A^{-1})^*A^* = I$, it suffices to show that $y - (A^{-1})^*A^*y = 0$. This holds if and only if $\langle y - (A^{-1})^*A^*y, z \rangle = 0$ for all z, which holds if and only if $\langle (A^{-1})^*A^*y, z \rangle = \langle y, z \rangle$. Similarly show $A^*(A^{-1})^* = I$.

17. A linear, time-invariant system with causal impulse response h is given. Find a causal input waveform x of minimum energy so that the system output at time $t = 1$ is equal to y; i.e., find a minimum-energy waveform x so that

$$\left[\int_0^t h(t-\tau)x(\tau)\,d\tau \right] \Bigg|_{t=1} = y.$$

18. Let $A: \mathbb{R}^n \to \mathbb{R}^m$ with $m \geq n$ be a linear transformation given by an $m \times n$ matrix with linearly independent columns. Let $W \subset \mathbb{R}^n$ be a nontrivial, proper subspace of $\mathbb{R}^n$ (i.e., $W \neq \{0\}$ and $W \neq \mathbb{R}^n$). For $x \in \mathbb{R}^n$, let $\widehat{x}$ denote the projection of x onto W. Let

$$V := \{Aw : w \in W\}.$$

Consider the statement:

Given $x \notin W$, the projection of Ax onto V is given by $A\widehat{x}$. (Use the standard inner products on $\mathbb{R}^n$ and $\mathbb{R}^m$.)

Is this statement:

(a) Always true?
(b) Sometimes true (give an example where it is true and another example where it is not true)?
(c) Never true?

19. Suppose A is self adjoint and satisfies $\langle Ax, x\rangle = 0$ for all x. Show that $Ax = 0$ for all x (in other words, A is the zero operator). *Hint:* Since $\langle A(u+v), u+v\rangle = 0$ for all u and v, what can you say about $\langle Au, v\rangle$?

20. In a complex inner-product space X, if an operator $A: X \to X$ satisfies $\langle Ax, x\rangle \in \mathbb{R}$ for all $x \in X$, show that A is self adjoint. *Hint:* The hypothesis implies that for all $\lambda \in \mathbb{C}$ and all $x, y \in X$, we have $\langle A(x+\lambda y), x+\lambda y\rangle \in \mathbb{R}$. Apply this observation to $\lambda = 1$ and $\lambda = j$.

21. In a complex inner-product space X, let $A: X \to X$ be a linear operator, and put $f(x) := \langle Ax, x\rangle$. Derive the **generalized polarization identity**

$$4\langle Ax, y\rangle = f(x+y) - f(x-y) + jf(x+jy) - jf(x-jy).$$

Use this result to show that if $\langle Bx, x\rangle = \langle Cx, x\rangle$ for all x, then $B = C$.

22. Derive the adjoint formula $\tilde{A} = Q^{-1}A^*R$ given in Section 4.5.

23. Let A be nonsingular, and let R be self adjoint and positive definite. Show that A^*RA is nonsingular.

24. ***Interference Rejection.*** Consider a measurement of the form $U = \Gamma\varphi + \Lambda\psi + W$, where W is a random noise vector with zero mean and known, positive-definite covariance matrix $Q := \mathsf{E}[WW^{\mathsf{H}}]$. The matrices Γ and Λ are known, full rank, and the combined matrix $G := [\Gamma\ \Lambda]$ is also full rank. The term $\Gamma\varphi$ represents a desired signal, while the term $\Lambda\psi$ represents **structured interference**. The deterministic parameters φ and ψ are unknown, and it is desired to estimate φ using a linear system of the form $\widehat{\Phi} = XU$, where the matrix X should minimize $\mathsf{E}[\|\varphi - \widehat{\Phi}\|^2]$ subject to the constraints $X\Gamma = I$ and $X\Lambda = 0$. For such X,

$$\mathsf{E}[\|\varphi - \widehat{\Phi}\|^2] = \mathsf{E}[\|\varphi - X(\Gamma\varphi + \Lambda\psi + W)\|^2] = \mathsf{E}[\|XW\|^2] = \mathrm{tr}(XQX^{\mathsf{H}}).$$

Hence, among all X that preserve φ and remove ψ, we seek the one that minimizes the residual error. Show that the optimal X is given by

$$\begin{bmatrix} I & 0 \end{bmatrix} (G^{\mathsf{H}}Q^{-1}G)^{-1}G^{\mathsf{H}}Q^{-1}.$$

Then put $B := Q^{-1/2}\Gamma$, $C := Q^{-1/2}\Lambda$, and use (3.14) to show that

$$(G^{\mathsf{H}}Q^{-1}G)^{-1}G^{\mathsf{H}}Q^{-1} = \begin{bmatrix} (B^{\mathsf{H}}P_C^{\perp}B)^{-1}B^{\mathsf{H}}P_C^{\perp} \\ (C^{\mathsf{H}}P_B^{\perp}C)^{-1}C^{\mathsf{H}}P_B^{\perp} \end{bmatrix} Q^{-1/2}.$$

25. ***Wronskians.*** Suppose functions $x_1,\dots,x_n$ are defined on a nonempty, open interval J, and assume that each function is $n-1$ times differentiable on J. For $t \in J$, define the $n \times n$ matrix $X(t)$ by $[X(t)]_{\ell,k} := x_k^{(\ell)}(t)$, where $k = 1,\dots,n$ and $\ell = 0,\dots,n-1$, and $x_k^{(0)}(t) := x_k(t)$. The **Wronskian** of $x_1,\dots,x_n$ is defined as $\det X(t)$. This is a scalar-valued function of $t \in J$.

(a) Show that if $x_1,\dots,x_n$ are linearly dependent, then the Wronskian is identically zero on J.

(b) The converse of (a) is false; i.e., it is *not* true that if the Wronskian is identically zero, then the functions must be linearly dependent. Show this as follows. Consider $x_1(t) = t^2$ and $x_2(t) = |t| \cdot t$ on some open interval containing $t = 0$. This example was given by Peano [44].

Remark. The contrapositive of (a) is more useful: If the Wronskian is not identically zero, then the functions are linearly independent.

26. ***Wronskians and Linearly Independent Solutions of Differential Equations.*** Consider the system of n coupled differential equations

$$\begin{aligned} y_1'(t) &= y_2(t) \\ y_2'(t) &= y_3(t) \\ &\vdots \\ y_{n-1}'(t) &= y_n(t) \\ y_n'(t) &= h(t, y_1(t), \dots, y_n(t)). \end{aligned}$$

Using the vector notation $y(t) := [y_1(t),\dots,y_n(t)]^{\mathsf{T}}$, we can write the system of differential equations as the single vector differential equation $y'(t) = \tilde{h}(t, y(t))$, where

$$\tilde{h}(t,z) := \begin{bmatrix} z_2 \\ \vdots \\ z_n \\ h(t, z_1, \dots, z_n) \end{bmatrix}, \quad z := [z_1,\dots,z_n]^{\mathsf{T}} \in \mathbb{R}^n.$$

Under suitable assumptions on h, there is a solution of the vector differential equation with initial condition $y(t_0) = \xi \in \mathbb{R}^n$. Given a vector solution $y(t)$, if

we let $x(t)$ denote the first component of $y(t)$; i.e., $x(t) := y_1(t)$, then

$$\begin{aligned} x'(t) &= y_1'(t) = y_2(t) \\ x''(t) &= y_2'(t) = y_3(t) \\ &\vdots \\ x^{(n-1)}(t) &= y_n(t) \\ x^{(n)}(t) &= y_n'(t) = h(t,x(t),x'(t),\ldots,x^{(n-1)}(t)). \end{aligned} \tag{4.9}$$

There are two important observations here. First, the definition $x(t) := y_1(t)$ together with the first $n-1$ lines of (4.9) show that $x^{(\ell)}(t) = y_{\ell+1}(t)$ for $\ell = 0,\ldots,n-1$. Second, the last equation in (4.9) shows that

$$x^{(n)}(t) = h(t,x(t),x'(t),\ldots,x^{(n-1)}(t)).$$

In other words, x solves the above nth-order scalar differential equation. Now the waveform x depends on the vector of initial conditions ξ used with $y'(t) = \tilde{h}(t,y(t))$ and $y(t_0) = \xi$. Let x_k denote the waveform x when the initial condition $\xi = e_k$, the kth standard unit vector in $\mathbb{R}^n$. Show that the solutions $x_1,\ldots,x_n$ are linearly independent.

Remark. Suppose that $\tilde{h}(t,z)$ is linear in z and we let ${}^k y$ denote the solution of $y'(t) = \tilde{h}(t,y(t))$ with initial condition $y(t_0) = e_k$. Then it is easy to check that $\sum_{k=1}^n \xi_k \cdot {}^k y(t)$ solves the differential equation with initial condition $\xi = [\xi_1,\ldots,\xi_n]^{\mathsf{T}}$. This can be seen by writing

$$\left(\sum_{k=1}^n \xi_k \cdot {}^k y(t)\right)' = \sum_{k=1}^n \xi_k \cdot {}^k y'(t) = \sum_{k=1}^n \xi_k \tilde{h}(t,{}^k y(t)) = \tilde{h}\left(t, \sum_{k=1}^n \xi_k \cdot {}^k y(t)\right)$$

and noting that

$$\left(\sum_{k=1}^n \xi_k \cdot {}^k y(t)\right)\Big|_{t=t_0} = \sum_{k=1}^n \xi_k \cdot {}^k y(t_0) = \sum_{k=1}^n \xi_k \cdot e_k = \xi.$$

CHAPTER 5
Optimization

Consider a real-valued function f defined on an arbitrary subset X_0 of an arbitrary set X. We say that $x_0 \in X_0$ minimizes f on X_0 if

$$f(x_0) \leq f(x) \text{ for all } x \in X_0.$$

For example, the problem of minimizing $f(x) = x \ln x$ for $x > 0$ falls into this framework if we put $X = \mathbb{R}$ and $X_0 = (0, \infty)$.

Sometimes, however, we want to restrict attention to a subset of X_0 characterized by a finite number of inequalities of the form $h_i(x) \leq 0$, $i = 1, \ldots, m$, where each h_i is a real-valued function defined on X_0. We say that $x_0 \in X_0$ minimizes f subject to the constraints $h_i(x) \leq 0$ for $x \in X_0$ if

$$f(x_0) \leq f(x) \text{ for all } x \in X_0 \text{ with } h_i(x) \leq 0,\ i = 1, \ldots, m.$$

Here f is called the **objective function** to distinguish it from the **constraint functions** h_i. For example, the problem of minimizing $f(x,y) = x \ln x + y \ln y$ for $x > 0$ and $y > 0$ satisfying $x^2 + y^2 \leq 1$ falls into this framework if we put $X = \mathbb{R}^2$ and let X_0 denote the strictly positive first quadrant; the objective function is $f(x,y) = x \ln x + y \ln y$ and the constraint function is $h(x,y) = x^2 + y^2 - 1$.

Notation. We put

$$H(x) := \begin{bmatrix} h_1(x) \\ \vdots \\ h_m(x) \end{bmatrix},$$

and write $H(x) \leq 0$ to mean that $h_i(x) \leq 0$ for $i = 1, \ldots, m$.

5.1. Introduction to Lagrange Multipliers

The generalization of the first quadrant in two-dimensional space to m-dimensional space is the **nonnegative orthant**,

$$\mathbb{R}^m_+ := \{\lambda \in \mathbb{R}^m : \lambda_i \geq 0, i = 1, \ldots, m\}.$$

The **Lagrangian** for the minimization problem subject to inequality constraints is the function $L : \mathbb{R}^m_+ \times X_0 \to \mathbb{R}$, defined by

$$L(\lambda, x) := f(x) + \lambda^{\mathsf{T}} H(x). \tag{5.1}$$

As we shall see, there is a close connection between unconstrained minimization of the Lagrangian over X_0 and minimization of f subject to the constraint $H(x) \leq 0$ for $x \in X_0$. In this section, we show that minimizing the Lagrangian is sufficient to minimize f subject to the constraint. As we know from calculus, the solution of minimization problems can often be made easier by using derivatives. For this reason, we will study the derivative of the Lagrangian with respect to x in Section 5.3.

Theorem 5.1 (Inequality Constraints). *If there exists a $\lambda_0 \in \mathbb{R}^m_+$ and an $x_0 \in X_0$ such that*

$$H(x_0) \leq 0 \qquad \text{and} \qquad \lambda_0^{\mathsf{T}} H(x_0) = 0, \tag{5.2}$$

and such that

$$L(\lambda_0, x_0) \leq L(\lambda_0, x), \quad \text{for all } x \in X_0, \tag{5.3}$$

then

$$f(x_0) \leq f(x), \quad \text{for all } x \in X_0 \text{ with } H(x) \leq 0. \tag{5.4}$$

Proof. Since $\lambda_0 \in \mathbb{R}^m_+$, for every $x \in X_0$ such that $H(x) \leq 0$, $\lambda_0^{\mathsf{T}} H(x) \leq 0$. Hence, we can write

$$\begin{aligned} f(x) &\geq f(x) + \lambda_0^{\mathsf{T}} H(x) \\ &= L(\lambda_0, x) \\ &\geq L(\lambda_0, x_0), && \text{by (5.3),} \\ &= f(x_0) + \lambda_0^{\mathsf{T}} H(x_0) \\ &= f(x_0), && \text{by (5.2).} \end{aligned}$$

□

Theorem 5.2 (Equality Constraints). *Let $G: X_0 \to Z$, where Z is a real or complex inner-product space. Put $L(\mu, x) := f(x) + \mathrm{Re}\langle \mu, G(x) \rangle$ for $\mu \in Z$ and $x \in X_0$. If there exists a $\mu_0 \in Z$ and an $x_0 \in X_0$ such that $G(x_0) = 0$ and*

$$L(\mu_0, x_0) \leq L(\mu_0, x), \quad \text{for all } x \in X_0, \tag{5.5}$$

then

$$f(x_0) \leq f(x), \quad \text{for all } x \in X_0 \text{ with } G(x) = 0.$$

Proof. Let $x \in X_0$ satisfy $G(x) = 0$. Then $\langle \mu_0, G(x) \rangle = 0$, and we can write

$$\begin{aligned} f(x) &= f(x) + \mathrm{Re}\langle \mu_0, G(x) \rangle \\ &= L(\mu_0, x) \\ &\geq L(\mu_0, x_0), && \text{by (5.5),} \\ &= f(x_0) + \mathrm{Re}\langle \mu_0, G(x_0) \rangle \\ &= f(x_0), && \text{since } G(x_0) = 0. \end{aligned}$$

□

Theorem 5.3 (Mixed Constraints). *Let $H: X_0 \to \mathbb{R}^m$, and let $G: X_0 \to Z$, where Z is a real or complex inner-product space. Put $L(\lambda, \mu, x) := f(x) + \lambda^{\mathsf{T}} H(x) + \mathrm{Re}\langle \mu, G(x) \rangle$. If there exists a $\lambda_0 \in \mathbb{R}^m_+$, a $\mu_0 \in Z$, and an $x_0 \in X_0$ such that* (5.2) *holds, $G(x_0) = 0$, and*

$$L(\lambda_0, \mu_0, x_0) \le L(\lambda_0, \mu_0, x), \quad \text{for all } x \in X_0,$$

then

$$f(x_0) \le f(x), \quad \text{for all } x \in X_0 \text{ with } H(x) \le 0 \text{ and } G(x) = 0.$$

Proof. Problem 5.2. □

5.2. Convex Functions

A real-valued function f is said to be **convex** if the line joining any two points $(x, f(x))$ and $(y, f(y))$ lies above the function values when the function argument lies on the line joining x and y. This is illustrated in Figure 5.1. A quick sketch of

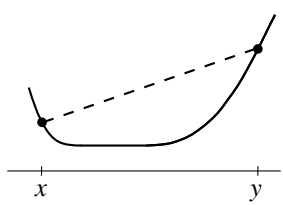


Figure 5.1. A convex function (solid curve) and the straight line passing through $(x, f(x))$ and $(y, f(y))$ (dashed line).

the graphs of functions like $f(x) = e^x$ and $f(x) = x^2$ provides additional examples. Functions of two or more variables can also be convex. For example, the bowl-shaped function $f(x, y) = x^2 + y^2$ shown in Figure 5.2 is convex.

A little reflection shows that we need to make our definition of convex function more precise. In particular, we must guarantee that f is defined for all points on the line joining x and y. A real-valued function f defined on a convex subset C of a real or complex vector space X is said to be **convex** if[a]

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y) \tag{5.6}$$

[a] If f satisfies the reverse inequality, then f is said to be **concave**. Equivalently, f is concave if $-f$ is convex. Some authors do not use the term "concave." Instead they write "convex $\cap$" for "concave," and they write "convex $\cup$" for "convex." In this context, the symbol $\cap$ is read "cap," and the symbol $\cup$ is read "cup."

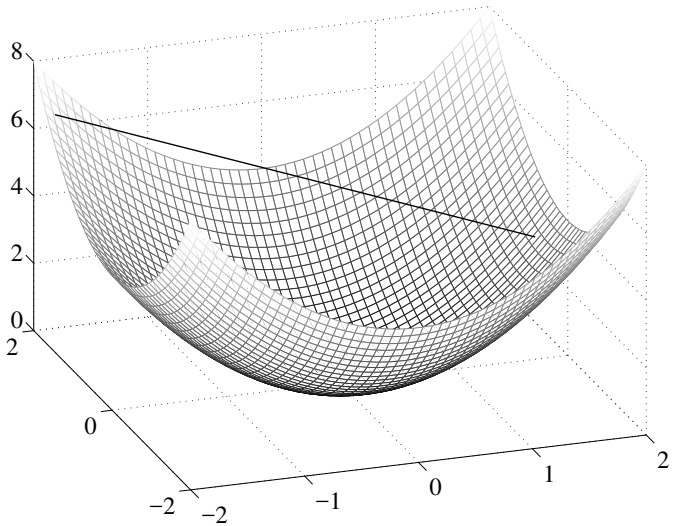


Figure 5.2. A convex function function of two variables and a line joining two points on its surface.

for all $x, y \in C$ and all $0 \leq \lambda \leq 1$. By induction, (5.6) extends to any finite convex combination of points in C. Specifically, a convex function f on C satisfies

$$f\left(\sum_{k=1}^{n} \lambda_k x_k\right) \leq \sum_{k=1}^{n} \lambda_k f(x_k) \tag{5.7}$$

when each $x_k \in C$ and the λ_k are nonnegative and sum to one (recall your solution of Problem 2.23).

If (5.6) is strict for $0 < \lambda < 1$ and $x \neq y$, we say that f is **strictly convex**. A strictly convex function can have at most one minimizer, but it may not have any minimizer at all. For example, $f(x) = x^2$ is uniquely minimized by $x = 0$, but $f(x) = e^x$ has no minimizer.[b]

The inequality (5.6) can also be written as

$$f(y + \lambda(x - y)) \leq f(y) + \lambda[f(x) - f(y)].$$

Interchanging the roles of x and y yields

$$f(x + \lambda(y - x)) \leq f(x) + \lambda[f(y) - f(x)].$$

[b] Although $e^x \geq 0$ for all $x \in \mathbb{R}$, there is no $x \in \mathbb{R}$ with $e^x = 0$.

If we rearrange this inequality, we find that

$$\frac{f(x+\lambda(y-x))-f(x)}{\lambda} \leq f(y)-f(x).$$

Let us put

$$(D^+f)(x,y-x) := \lim_{\lambda \downarrow 0} \frac{f(x+\lambda(y-x))-f(x)}{\lambda}, \tag{5.8}$$

assuming this limit exists. When this limit exists, it is called the **one-sided Gâteaux derivative** of f at x in the direction $y-x$. As we show near the end of the section, for a convex function f, this limit always exists either as a finite number or as $-\infty$. Thus, for a convex function,

$$(D^+f)(x,y-x) \leq f(y)-f(x),$$

which we can rewrite as

$$\boxed{f(y) \geq f(x)+(D^+f)(x,y-x).} \tag{5.9}$$

If there is a particular $x \in C$ for which the derivative in (5.9) is nonnegative for all $y \in C$, then $f(y) \geq f(x)$ for all $y \in C$; i.e., x is a **minimizer** of f on C. The converse is also true; if x minimizes f on C, then the limit in (5.8) is nonnegative. We summarize this finding precisely in the following theorem.

Theorem 5.4. *Let f be a convex function defined on a convex subset C of a real or complex vector space. A point $x \in C$ satisfies $(D^+f)(x,y-x) \geq 0$ for all $y \in C$ if and only if x minimizes f on C.*

If $X = \mathbb{R}$ and C is an interval, $y-x$ is a number. This means that we can write (5.8) as

$$\frac{f(x+\lambda(y-x))-f(x)}{\lambda(y-x)}(y-x) \to f'(x)(y-x),$$

assuming that f is differentiable in the usual sense for functions of one variable. In this case, we have $(D^+f)(x,y-x) = f'(x)(y-x)$.

Example 5.5. Consider the problem of minimizing the function $f(x) = e^{-x}$ for $x \in [0,1]$. Let us use Theorem 5.4 to show that the solution of this problem is achieved by $x = 1$. First, we must show f is convex. This can be done by the same reasoning that will be used later in Problem 5.15. It remains to show that $(D^+f)(1,y-1) \geq 0$ for $y \in [0,1]$. We simply observe that $(D^+f)(1,y-1) = f'(1)(y-1) = -e^{-1}(y-1)$ is greater than or equal to zero for $y \in [0,1]$. Hence, $x = 1$ minimizes f on $[0,1]$.

Just as we have related the Gâteaux derivative of a function of one variable to the ordinary derivative, we can relate the Gâteaux derivative of a function of n variables to ordinary partial derivatives.

Theorem 5.6. *Let f be a real-valued function defined on an open subset U of $\mathbb{R}^n$; e.g., $U = \mathbb{R}^n$. Denote the partial derivative of f with respect to its kth variable by f_k. If $f_1, \ldots, f_n$ exist and are continuous on U, then*

$$(D^+f)(x,\Delta x) = \sum_{k=1}^{n} f_k(x)\Delta x_k$$

for all $x \in U$ and all $\Delta x = [\Delta x_1, \ldots, \Delta x_n]^\mathsf{T} \in \mathbb{R}^n$.

Proof. See the remark following Note 5.1 at the end of the chapter. □

Example 5.7. Use Theorem 5.4 to find $(x,y) \in \mathbb{R}^2_+$ to minimize the function $f(x,y) := (x+1)^2 + (y-2)^2$.

Solution. By Problem 5.3, $\mathbb{R}^2_+$ is a convex set. We will show later in Example 5.21 that f is convex. The minimum of f over $\mathbb{R}^2$ occurs at $x = -1$ and $y = 2$. However, our job is to minimize $f(x,y)$ over $x, y \geq 0$. By Theorem 5.6,

$$(D^+f)(x_0,y_0,x-x_0,y-y_0) = 2(x_0+1)(x-x_0) + 2(y_0-2)(y-y_0).$$

This will be zero for all x and y if $x_0 = -1$ and $y_0 = 2$. However, $(-1,2)$ does not lie in $\mathbb{R}^2_+$, which is the set over which we minimizing f. Let us try $x_0 = 0$ and $y_0 = 2$. Then

$$(D^+f)(x_0,y_0,x-x_0,y-y_0) = 2x,$$

which is nonnegative for all x under consideration. Hence, $(0,2)$ minimizes f over $\mathbb{R}^2_+$.

In the preceding example, we minimized f over the set $\mathbb{R}^2_+$, which is not open. Fortunately f was defined on all of $\mathbb{R}^2$ and we could use Theorem 5.6 to compute the Gâteaux derivative. However, sometimes the function that we want to minimize is defined only on a non-open set. In this case, the partial derivatives may be $-\infty$ on the boundary. The following result modifies Theorem 5.6 to handle this case.

Theorem 5.8. *Let f be a convex function defined on $\mathbb{R}^n_+$. Assume that for each $x = [x_1, \ldots, x_n]^\mathsf{T} \in \mathbb{R}^n_+$ with $x_k > 0$, the partial derivative of f with respect to its kth variable, denoted by $f_k(x)$, exists as a finite number. If $x_k = 0$, we take $f_k(x)$ to be*

$$\lim_{t \downarrow 0} \frac{f(x_1, \ldots, x_{k-1}, t, x_{k+1}, \ldots, x_n) - f(x)}{t},$$

which is the Gâteaux derivative of f in the direction of the kth standard unit vector in $\mathbb{R}^n$. As we have seen, this limit may be $-\infty$. We assume that for all $x \in \mathbb{R}^n_+$,

$$\lim_{w \in \mathbb{R}^n_+, w \to x} f_k(w) = f_k(x),$$

where the right-hand side may be $-\infty$ if $x_k = 0$. If $x \in \mathbb{R}^n_+$ is such that all the partial derivatives $f_k(x)$ are finite, then

$$(D^+ f)(x, y-x) = \sum_{k=1}^{n} f_k(x)(y_k - x_k)$$

for all $y = [y_1, \ldots, y_n]^\mathsf{T} \in \mathbb{R}^n_+$.

Proof. See the Notes at the end of the chapter.[1] □

Existence of the Limit in (5.8)

Theorem 5.9. *A convex function f on a convex set C has a one-sided Gâteaux derivative at every $x \in C$ in the direction $y - x$ for every $y \in C$ if we allow $-\infty$ as a possible value of the derivative.*

Proof. If we can show that the quotient in (5.8) is a nonincreasing function of λ, then as $\lambda \downarrow 0$, either the quotients tend to $-\infty$ or they are bounded below. In the latter case, the quotients tend to a finite limit (see Problem 5.30 for precise details). To show that the quotients are nonincreasing, we must show that for $0 < \lambda_1 < \lambda_2 \leq 1$,

$$\frac{f(x + \lambda_1(y-x)) - f(x)}{\lambda_1} \leq \frac{f(x + \lambda_2(y-x)) - f(x)}{\lambda_2}. \tag{5.10}$$

Put $y_1 := x + \lambda_1(y-x)$ and $y_2 := x + \lambda_2(y-x)$. Rearrange the second definition as $y - x = (y_2 - x)/\lambda_2$ and substitute this into the first definition so that

$$y_1 = x + \frac{\lambda_1}{\lambda_2}(y_2 - x).$$

Since $0 < \lambda_1/\lambda_2 < 1$, and since f is convex,

$$f(y_1) \leq f(x) + \frac{\lambda_1}{\lambda_2}[f(y_2) - f(x)],$$

which we can rearrange as (5.10) as required. □

Theorem 5.10. *Let f be a convex function on an interval I of the real line. If a and b are interior points of I with $a < b$, then f is* **Lipschitz continuous** *on $[a,b]$. Hence, if x is not an endpoint of I, then f is continuous at x. Also, if x is not an endpoint of I, then f has finite left and right derivatives at x, denoted by $f'_-(x)$ and $f'_+(x)$, respectively, satisfying $f'_-(x) \le f'_+(x)$. Furthermore,*

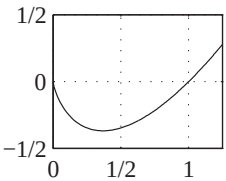
$$f(y) \ge f(x) + f'_\pm(x)(y-x), \quad y \in I.$$

Finally, f'_- and f'_+ are nondecreasing.

Proof. See the Notes at the end of the chapter.[2] □

Example 5.11. Let $X = \mathbb{R}$ and consider the convex set $C = [0,\infty)$. Define the real-valued function f on C by

$$f(x) := \begin{cases} x\ln x, & x > 0, \\ 0, & x = 0, \end{cases}$$



which is shown in the figure at the right. You will show that f is convex in Problem 5.16. Here we simply show that when $x = 0$, $(D^+f)(x, y-x) = -\infty$. For $x = 0$ and $y > 0$,

$$f\big(x+\lambda(y-x)\big) - f(x) = f(\lambda y) - f(0) = f(\lambda y) = \lambda y \ln(\lambda y),$$

and so

$$\frac{f\big(x+\lambda(y-x)\big) - f(x)}{\lambda} = \frac{\lambda y \ln(\lambda y)}{\lambda} = y\ln(\lambda y) \to -\infty$$

as $\lambda \downarrow 0$.

Example 5.12. The assumption in Theorem 5.10 that x is not an endpoint of the interval is critical for continuity. Consider the function on $[0,\infty)$ defined by $f(0) := 1$ and $f(x) := 0$ for $x > 0$. Then f is convex, but not continuous at the endpoint $x = 0$.

Theorem 5.13 (Jensen's Inequality). *Let f be a convex function defined on an interval, and let X be a random variable taking values in that interval. If $\mathsf{E}[|X|] < \infty$, then*

$$\mathsf{E}[f(X)] \ge f(\mathsf{E}[X]),$$

where the left-hand side may be $+\infty$, but cannot be $-\infty$.

Proof. A general proof is given in the Notes at the end of the chapter.[3] Here we point out that if X is a discrete random variable taking finitely many distinct values $x_1, \ldots, x_n$ with probabilities $\lambda_k := \mathsf{P}(X = x_k)$, then

$$\mathsf{E}[f(X)] = \sum_{k=1}^{n} f(x_k)\mathsf{P}(X = x_k) = \sum_{k=1}^{n} \lambda_k f(x_k),$$

while $\mathsf{E}[X] = \sum_k x_k \mathsf{P}(X = x_k) = \sum_k \lambda_k x_k$. Thus, (5.7) is exactly Jensen's inequality. □

5.2.1. The Gradient Descent Algorithm

Consider the convex function f shown in Figure 5.3. We would like to develop

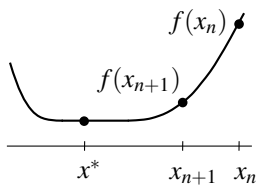


Figure 5.3. A convex function f that is minimized by x^* and a sequence x_n for which $f(x_n) \to f(x^*)$. We do not require that $x_n \to x^*$.

an algorithm to generate points x_n for which $f(x_n)$ decreases to $f(x^*)$, the minimum value of f. Since f may be flat, there may be many minimizers x^*, and so we cannot expect that x_n will converge to a particular one.

To begin, suppose that the function f is differentiable on a real inner-product space X. Given a point x_n and corresponding value $f(x_n)$, the value of f at a nearby point x is approximated by (Problem 5.21)

$$f(x) \approx f(x_n) + \langle x - x_n, \nabla_{x_n} f \rangle.$$

This suggests that we put

$$x_{n+1} = \operatorname*{argmin}_x \big[f(x_n) + \langle x - x_n, \nabla_{x_n} f \rangle + \alpha \|x - x_n\|^2 \big],$$

where $\alpha > 0$ is a **regularization parameter**. The inclusion of the term $\alpha\|x - x_n\|^2$ forces the minimizer to be close to x_n, which is where the above approximation is

valid. To find x_{n+1}, note that the quantity in brackets is quadratic in x. It follows that[c]

$$x_{n+1} = x_n - \tfrac{1}{2\alpha}\nabla_{x_n} f.$$

Setting $\eta := 1/(2\alpha)$ yields the **gradient descent algorithm**,

$$x_{n+1} = x_n - \eta \nabla_{x_n} f,$$

where $\eta > 0$ is a called the **step size**.

Example 5.14 (Gradient Descent and the Projection Problem). Projecting a vector $y_0 \in Y$ onto the range of an operator $A: X \to Y$ corresponds to minimizing the function $f(x) := \|y_0 - Ax\|^2$. We saw in Section 4.3.2 that under suitable assumptions, the minimizer is given by $x_0 = (A^*A)^{-1}A^*y_0$. However, the numerical implementation of this formula can involve considerable challenges. For this reason, a common alternative is to approximate x_0 using the gradient descent algorithm. To do this, we need a formula for $\nabla_x f$. Expanding $f(x) = \|y_0\|^2 - 2\langle x, A^*y_0\rangle + \langle Ax, Ax\rangle$, it is not hard to show that $\nabla_x f = 2A^*(Ax - y_0)$ (cf. Problem 5.21). The gradient descent algorithm becomes

$$x_{n+1} = x_n + 2\eta A^*(y_0 - Ax_n).$$

It is sometimes convenient to write this as a pair equations so that we can keep track of the error $e_n := y_0 - Ax_n$; i.e.,

$$\begin{aligned} e_n &= y_0 - Ax_n \\ x_{n+1} &= x_n + 2\eta A^* e_n. \end{aligned}$$

Theorem 5.15 (Gradient Descent). *Let f be an L-smooth (Problem 5.24) convex function defined on a real inner product space X. If $x^* = \operatorname{argmin}_x f(x)$, then the gradient descent algorithm starting at x_0 and using step size $\eta = 1/L$ satisfies*[d]

$$f(x_n) - f(x^*) \le \frac{2L}{n}\|x_0 - x^*\|^2.$$

[c] There are two ways to see this. One way is to complete the square by writing

$$\begin{aligned} \langle x - x_n, \nabla_{x_n} f\rangle + \alpha\|x - x_n\|^2 &= \alpha\left[\|x - x_n\|^2 + 2\langle x - x_n, \tfrac{1}{2\alpha}\nabla_{x_n} f\rangle + \|\tfrac{1}{2\alpha}\nabla_{x_n} f\|^2\right] - \alpha\|\tfrac{1}{2\alpha}\nabla_{x_n} f\|^2 \\ &= \alpha\|(x - x_n) + \tfrac{1}{2\alpha}\nabla_{x_n} f\|^2 - \alpha\|\tfrac{1}{2\alpha}\nabla_{x_n} f\|^2. \end{aligned}$$

The other way is to take the gradient of the above left-hand side with respect to x. The desired gradient is $\nabla_{x_n} f + 2\alpha(x - x_n)$ (see the Remarks in Problem 5.21). Setting this equal to zero and solving for x yields the asserted value of x_{n+1}.

[d] Any $0 < \eta < 2/L$ will make the error decay like $1/n$. However, taking η closer to $2/L$ makes the constant multiplying $(1/n)$ larger.

Proof. Consider two steps x_k and $x_{k+1} = x_k - \eta \nabla_{x_k} f$. We first show that $\|x_k - x^*\|$ is strictly decreasing in k. Write

$$\begin{aligned}
\|x_{k+1} - x^*\|^2 &= \|(x_k - \eta \nabla_{x_k} f) - x^*\|^2 \\
&= \|(x_k - x^*) - \eta \nabla_{x_k} f)\|^2 \\
&= \|x_k - x^*\|^2 - 2\eta \langle x_k - x^*, \nabla_{x_k} f \rangle + \eta^2 \|\nabla_{x_k} f\|^2 \\
&= \|x_k - x^*\|^2 - 2\eta \langle x_k - x^*, \nabla_{x_k} f - \nabla_{x^*} f \rangle + \eta^2 \|\nabla_{x_k} f\|^2,
\end{aligned}$$

since $\nabla_{x^*} f = 0$ on account of the assumption that x^* minimizes f on X (see the Remark in Problem 5.22). Next, use the lower bound

$$\langle x_k - x^*, \nabla_{x_k} f - \nabla_{x^*} f \rangle \geq \tfrac{1}{L} \|\nabla_{x_k} f - \nabla_{x^*} f\|^2, \quad \text{by Problem 5.26,}$$

to write

$$\begin{aligned}
\|x_{k+1} - x^*\|^2 &\leq \|x_k - x^*\|^2 - 2\eta \cdot \tfrac{1}{L} \|\nabla_{x_k} f - \nabla_{x^*} f\|^2 + \eta^2 \|\nabla_{x_k} f\|^2 \\
&= \|x_k - x^*\|^2 - (2/L^2) \|\nabla_{x_k} f\|^2 + \|\nabla_{x_k} f\|^2 / L^2, \qquad \text{since } \eta = 1/L, \\
&\leq \|x_k - x^*\|^2 - \|\nabla_{x_k} f\|^2 / L^2.
\end{aligned}$$

Hence $\|x_k - x^*\|$ is strictly decreasing.

Our second task is to upper bound $f(x_{k+1}) - f(x^*)$ in terms of $f(x_k) - f(x^*)$. Since $x_{k+1} - x_k = -\eta \nabla_{x_k} f$, and since f is L-smooth, we have from (5.36) in Problem 5.24 that

$$\begin{aligned}
f(x_{k+1}) &\leq f(x_k) - \eta \|\nabla_{x_k} f\|^2 + \tfrac{L}{2} \eta^2 \|\nabla_{x_k} f\|^2 \\
&= f(x_k) - \tfrac{1}{L} \|\nabla_{x_k} f\|^2 + \tfrac{1}{2L} \|\nabla_{x_k} f\|^2, \quad \text{since } \eta = 1/L, \\
&= f(x_k) - \tfrac{1}{2L} \|\nabla_{x_k} f\|^2.
\end{aligned}$$

Now subtract $f(x^*)$ from both sides to get

$$f(x_{k+1}) - f(x^*) \leq f(x_k) - f(x^*) - \tfrac{1}{2L} \|\nabla_{x_k} f\|^2.$$

To make the right-hand side larger, we need a lower bound on the gradient term. We can obtain this by using the convexity of f to write

$$\begin{aligned}
f(x^*) &\geq f(x_k) + \langle x_k - x^*, \nabla_{x_k} f \rangle \\
&\geq f(x_k) - \left| \langle x_k - x^*, \nabla_{x_k} f \rangle \right| \\
&\geq f(x_k) - \|x_k - x^*\| \, \|\nabla_{x_k} f\| \\
&\geq f(x_k) - \|x_0 - x^*\| \, \|\nabla_{x_k} f\|, \quad \text{since } \|x_k - x^*\| \text{ is decreasing.}
\end{aligned}$$

It follows that

$$\|\nabla_{x_k} f\| \geq \frac{f(x_k) - f(x^*)}{\|x_0 - x^*\|}.$$

Noting that the right-hand side is nonnegative since x^* minimizes f, it follows that

$$f(x_{k+1}) - f(x^*) \leq f(x_k) - f(x^*) - \frac{1}{2L} \cdot \frac{[f(x_k) - f(x^*)]^2}{\|x_0 - x^*\|^2}.$$

To make further progress, put $\Delta_k := f(x_k) - f(x^*)$ and $\lambda := 1/(2L\|x_0 - x^*\|^2)$ so that the above inequality takes the form $\Delta_{k+1} \leq \Delta_k - \lambda \Delta_k^2$. This inequality tells us two things. First, $\Delta_{k+1} < \Delta_k$. Second, it can be rewritten as

$$\lambda \frac{\Delta_k}{\Delta_{k+1}} \leq \frac{1}{\Delta_{k+1}} - \frac{1}{\Delta_k}.$$

Combining this with $\Delta_{k+1} < \Delta_k$ yields $\lambda < 1/\Delta_{k+1} - 1/\Delta_k$. Hence,

$$\frac{1}{\Delta_n} \geq \frac{1}{\Delta_{n-1}} + \lambda \geq \frac{1}{\Delta_{n-2}} + 2\lambda \geq \cdots \geq \frac{1}{\Delta_0} + n\lambda \geq n\lambda.$$

It follows that $\Delta_n \leq 1/(n\lambda) = 2L\|x_0 - x^*\|^2/n$. □

5.3. Lagrange Multipliers and Derivatives

We return to the optimization problems considered in Section 5.1.

Theorem 5.16. *Let X_0 be a convex subset of a real or complex vector space X. Assume that each of the real-valued functions $f, h_1, \ldots, h_m$ is convex on X_0. Suppose there is a point $x_0 \in X_0$ and a $\lambda_0 \in \mathbb{R}^m_+$ such that*

$$H(x_0) \leq 0 \qquad \text{and} \qquad \lambda_0^{\mathsf{T}} H(x_0) = 0. \tag{5.11}$$

Suppose also that the Lagrangian in (5.1) *satisfies*[e] $(D_x^+ L)(\lambda_0, x_0, y - x_0) \geq 0$ *for all $y \in X_0$. Then $f(x_0) \leq f(y)$ for all $y \in X_0$ with $H(y) \leq 0$.*

Proof. This is a simple application of Theorem 5.1 and the convexity property (5.9) applied to the Lagrangian. First note that (5.11) is the same as (5.2). It then suffices to establish that (5.3) holds. Since f and the h_i are convex, the Lagrangian $L(\lambda, x)$ is convex in x for $\lambda \in \mathbb{R}^m_+$ (Problem 5.32). Hence,

$$L(\lambda_0, y) \geq L(\lambda_0, x_0) + (D_x^+ L)(\lambda_0, x_0, y - x_0)$$

[e] The notatation $(D_x^+ L)$ indicates that the operator D^+ acts on the second argument of the Lagrangian.

holds for all $y \in X_0$. By hypothesis, the Gâteaux derivative is nonnegative, and so (5.3) follows as required. □

Remark. The foregoing proof uses the convexity of f and the h_i only to show that for *all* $\lambda \in \mathbb{R}^m_+$, the Lagrangian $L(\lambda, \cdot)$ is convex on X_0. However, all we really need is that for the *particular* λ_0 of interest, $L(\lambda_0, \cdot)$ is convex on X_0.

Theorem 5.17. *Let X and Z be vector spaces, both real or both complex, with Z being an inner-product space. Let X_0 be a convex subset of X. Assume that each of the real-valued functions f, $h_1, \ldots, h_m$ is convex on X_0. In addition, let $G(x) = Ax + b$, where $b \in Z$ and $A: X \to Z$ is linear. Suppose there exist $x_0 \in X_0$, $\lambda_0 \in \mathbb{R}^m_+$, and $\mu_0 \in Z$ such that*

$$H(x_0) \le 0, \quad \lambda_0^{\mathsf{T}} H(x_0) = 0, \quad \text{and} \quad G(x_0) = 0.$$

Suppose also that the Lagrangian

$$L(\lambda, \mu, x) := f(x) + \lambda^{\mathsf{T}} H(x) + \mathrm{Re}\langle \mu, G(x) \rangle$$

satisfies $(D_x^+ L)(\lambda_0, \mu_0, x_0, y - x_0) \ge 0$ for all $y \in X_0$. Then $f(x_0) \le f(y)$ for all $y \in X_0$ with $H(y) \le 0$ and $G(y) = 0$.

Proof. Problem 5.32. □

Examples

Example 5.18. We again consider the problem of minimizing $f(x) = e^{-x}$ for $x \in [0,1]$. To apply Theorem 5.16, we must select an appropriate convex set X_0 and appropriate constraint functions $h_i(x)$. There is more than one way to do this. For the first way, we choose $X_0 = [0, \infty)$ and $h_1(x) := x - 1 \le 0$. We already noted that f is convex. It is easy to check that h_1 is convex. The corresponding Lagrangian is $L(\lambda, x) = f(x) + \lambda(x-1)$, from which it follows that

$$(D_x^+ L)(\lambda, x, y - x) = [f'(x) + \lambda](y - x) = [-e^{-x} + \lambda](y - x). \tag{5.12}$$

We need this to be nonnegative for all $y \in [0, \infty)$. By (5.11), we also need $\lambda(x-1) = 0$. If $x = 0$, then $\lambda = 0$ and (5.12) becomes $-y$, which is *negative* for $y > 0$. Hence, $x = 0$ does not satisfy the conditions of the theorem. Suppose $x > 0$. Then $y - x$ can be positive or negative as y varies over $[0, \infty)$. To make the Gâteaux derivative in (5.12) nonnegative for all such y, we must make $e^{-x} = \lambda$; in particular, this means

$\lambda > 0$. To satisfy $\lambda(x-1) = 0$ then requires $x = 1$. We have now met the conditions of the theorem and so we may conclude that $x = 1$ minimizes e^{-x} on $[0,\infty)$ subject to the constraint $x \le 1$.

We now turn to a second way to apply Theorem 5.16. This time we take $X_0 = \mathbb{R}$, $h_1(x) := x - 1 \le 0$ and $h_2(x) = -x \le 0$. We know that f and h_1 are convex. It is easy to check that h_2 is also convex. Because there are two constraints, we need two Lagrange multipliers. The Lagrangian is $L(\lambda_1,\lambda_2,x) = f(x) + \lambda_1(x-1) + \lambda_2(-x)$, and

$$(D_x^+ L)(\lambda_1,\lambda_2,x,y-x) = [-e^{-x} + \lambda_1 - \lambda_2](y-x).$$

Since $y \in \mathbb{R}$, $y - x$ can be positive or negative. To make the Gâteaux derivative nonnegative for all $y \in \mathbb{R}$ requires

$$e^{-x} = \lambda_1 - \lambda_2. \tag{5.13}$$

We also need nonnegative λ_1 and λ_2 satisfying $\lambda_1(x-1) + \lambda_2(-x) = 0$, which we rewrite as

$$\lambda_1(x-1) = \lambda_2 x. \tag{5.14}$$

The constraint inequalities $h_1(x) = x - 1 \le 0$ and $h_2(x) = -x \le 0$ imply $0 \le x \le 1$. If $x = 0$, then (5.14) implies $\lambda_1 = 0$ and (5.13) becomes $1 = -\lambda_2$, which contradicts $\lambda_2 \ge 0$. If $0 < x < 1$, then (5.14) implies that either the multipliers have opposite signs, which we cannot allow, or they are both zero, which contradicts (5.13). Hence, the only possible choice for x is $x = 1$. In this case, (5.14) implies $\lambda_2 = 0$, and then (5.13) gives $\lambda_1 = e^{-1} > 0$. Thus, $x = 1$ minimizes f subject to the constraints.

To do more interesting examples, we need to compute Gâteaux derivatives for more complicated functions. If f is a convex function defined on the whole space X, then we will need to compute $(D^+ f)(x, y-x)$ for all $y \in X$. Hence, it suffices to compute $(D^+ f)(x,\Delta x)$ for arbitrary $\Delta x \in X$.

Example 5.19. Consider the function $f(x) = \operatorname{Re}\langle Ax, z\rangle$, where $A: X \to Z$ is a linear operator, Z is an inner-product space, and $z \in Z$ is given. Show that f is convex, and show that

$$(D^+ f)(x,\Delta x) = \operatorname{Re}\langle A\Delta x, z\rangle.$$

If X is also an inner-product space and A^* exists, then $(D^+ f)(x,\Delta x) = \operatorname{Re}\langle \Delta x, A^* z\rangle$.

Solution. To establish convexity, fix $x, y \in X$ and $0 \le \lambda \le 1$. Then

$$\begin{aligned} f\big(\lambda x + (1-\lambda)y\big) &= \operatorname{Re}\langle A[\lambda x + (1-\lambda)y], z\rangle \\ &= \operatorname{Re}\{\lambda\langle Ax, z\rangle + (1-\lambda)\langle Ay, z\rangle\} \\ &= \lambda \operatorname{Re}\langle Ax, z\rangle + (1-\lambda)\operatorname{Re}\langle Ay, z\rangle, \quad \text{since } \lambda \text{ is real,} \\ &= \lambda f(x) + (1-\lambda) f(y). \end{aligned}$$

To compute the Gâteaux derivative, observe that

$$f(x+\lambda\Delta x)-f(x)=\operatorname{Re}\langle A(x+\lambda\Delta x),z\rangle-\operatorname{Re}\langle Ax,z\rangle=\lambda\operatorname{Re}\langle A\Delta x,z\rangle.$$

Hence,

$$\lim_{\lambda\downarrow 0}\frac{f(x+\lambda\Delta x)-f(x)}{\lambda}=\lim_{\lambda\downarrow 0}\operatorname{Re}\langle A\Delta x,z\rangle=\operatorname{Re}\langle A\Delta x,z\rangle.$$

Example 5.20. Assume X is an inner-product space and that $B:X\to X$ is a self-adjoint, positive-semidefinite, linear operator. Show that $f(x):=\langle Bx,x\rangle$ is convex on X and that

$$(D^+f)(x,\Delta x)=2\operatorname{Re}\langle\Delta x,Bx\rangle.$$

We note that it will frequently be the case that $B=A^*A$ where $A:X\to Y$ and Y is another inner-product space such that $A^*:Y\to X$ exists. In this case, $f(x)=\langle Ax,Ax\rangle$, and $(D^+f)(x,\Delta x)=2\operatorname{Re}\langle\Delta x,A^*Ax\rangle$.

Solution. First note that since B is self adjoint, f is real valued, which is a prerequisite for f to be convex. To establish convexity, fix $x,y\in X$ and $0\le\lambda\le 1$. Note that $\lambda^2\le\lambda$. Then

$$\begin{aligned}f\big(x+\lambda(y-x)\big)&=\big\langle B[x+\lambda(y-x)],x+\lambda(y-x)\big\rangle\\&=\langle Bx,x\rangle+\lambda\langle B(y-x),x\rangle+\lambda\langle Bx,y-x\rangle+\lambda^2\langle B(y-x),y-x\rangle\\&\le\langle Bx,x\rangle+\lambda\langle B(y-x),x\rangle+\lambda\langle Bx,y-x\rangle+\lambda\langle B(y-x),y-x\rangle,\end{aligned}$$

where the last step uses the facts that $\lambda^2\le\lambda$ and $\langle B(y-x),y-x\rangle\ge 0$ (since B is positive semidefinite).[f] Combine the second and fourth inner products to get $\langle B(y-x),y\rangle$. Adding this to the third inner product yields $\lambda[\langle By,y\rangle-\langle Bx,x\rangle]$. It now follows that

$$f\big(x+\lambda(y-x)\big)\le f(x)+\lambda[f(y)-f(x)].$$

Next, we compute the Gâteaux derivative. Observe that

$$\begin{aligned}f(x+\lambda\Delta x)-f(x)&=\langle B[x+\lambda\Delta x],x+\lambda\Delta x\rangle-\langle Bx,x\rangle\\&=\lambda\langle B\Delta x,x\rangle+\lambda\langle Bx,\Delta x\rangle+\lambda^2\langle B\Delta x,\Delta x\rangle\\&=2\lambda\operatorname{Re}\langle\Delta x,Bx\rangle+\lambda^2\langle B\Delta x,\Delta x\rangle,\quad\text{since } B\text{ is self adjoint.}\end{aligned}$$

It now follows that

$$\frac{f(x+\lambda\Delta x)-f(x)}{\lambda}-2\operatorname{Re}\langle\Delta x,Bx\rangle=\lambda\langle B\Delta x,\Delta x\rangle,$$

which goes to zero as $\lambda\downarrow 0$.

[f] If B is positive definite, then f is strictly convex. This can be seen by considering $x\ne y$, $0<\lambda<1$, and observing that $\lambda^2\langle B(y-x),y-x\rangle<\lambda\langle B(y-x),y-x\rangle$.

Example 5.21. Show that $f(u,v) := (u+1)^2 + (v-2)^2$ is a convex function on $\mathbb{R}^2$.

Solution. The first step is to observe that

$$f(u,v) = \left\| \begin{bmatrix} u \\ v \end{bmatrix} - \begin{bmatrix} -1 \\ 2 \end{bmatrix} \right\|_{\mathbb{R}^2}^2.$$

If we put $x = [u,v]^{\mathsf{T}}$ and $y = [-1,2]^{\mathsf{T}}$, then f has the form

$$\|x-y\|^2 = \langle x-y, x-y\rangle = \langle x,x\rangle - 2\langle x,y\rangle + \langle y,y\rangle.$$

By Example 5.20, $\langle x,x\rangle$ is convex. By Example 5.19, $\langle x,y\rangle$ is a convex function of x. The last term is a constant function of x and is convex. By Problem 5.18, the sum of these terms is convex.

5.3.1. Norm Constrained Least Squares

Example 5.22 (Quadratically Constrained Least Squares). Consider the problem

$$\min_{x\in X} \|y - Ax\| \quad \text{subject to} \quad \|Qx\|^2 \le b,$$

where X and Y are inner-product spaces, $A: X \to Y$ and $Q: X \to X$ are linear operators, and $y \in Y$ and $b \ge 0$ are given. Assume that the adjoints A^* and Q^* exist. This is a common problem in communications and signal processing. The problem arises whenever the system output y is given, and the goal is to design a system input x to achieve $Ax \approx y$ as closely as possible, given a constraint on the energy available for x. The first step is to realize that the problem is unchanged if we replace $\|y-Ax\|$ by $\|y-Ax\|^2$. Next, to apply Theorem 5.16, we first put

$$f(x) = \|y-Ax\|^2 = \langle y-Ax, y-Ax\rangle = \|y\|^2 - 2\,\mathrm{Re}\langle Ax, y\rangle + \langle Ax, Ax\rangle.$$

By the same argument as in Example 5.20, f is convex. Next, we put $H(x) := \|Qx\|^2 - b = \langle Qx, Qx\rangle - b$. By Example 5.20, H is also convex. With $L(\lambda, x) = f(x) + \lambda H(x)$, we have from Examples 5.19 and 5.20 that

$$\begin{aligned}(D_x^+ L)(\lambda, x, \Delta x) &= -2\,\mathrm{Re}\langle \Delta x, A^*y\rangle + 2\,\mathrm{Re}\langle \Delta x, A^*Ax\rangle + \lambda 2\,\mathrm{Re}\langle \Delta x, Q^*Qx\rangle \\ &= 2\,\mathrm{Re}\langle \Delta x, -A^*y + A^*Ax + \lambda Q^*Qx\rangle.\end{aligned}$$

In order for this to be zero for all $\Delta x \in X$, it is necessary and sufficient that the right-hand argument of the inner product be zero; i.e.,

$$(\lambda Q^*Q + A^*A)x = A^*y. \tag{5.15}$$

We denote a solution of this equation by x_λ. When $\lambda = 0$, we must solve $A^*Ax = A^*y$. This is the equation considered in Lemma 4.16 If a solution x_0 exists, we must check to see if $\|Qx_0\|^2 \le b$. If this is so, the unconstrained solution solves the constrained problem and we are finished. If $\|Qx_0\|^2 > b$, we must try to solve (5.15) for positive values of λ. To satisfy the condition $\lambda(\|Qx\|^2 - b) = 0$ with positive λ, we must adjust λ so that $\|Qx_\lambda\|^2 = b$. In other words, we must solve (5.15) for increasing values of λ until $\|Qx_\lambda\|^2$ drops to b.

Example 5.23 (Projection onto an Ellipse). Recall the problem of projecting onto an ellipse as discussed in Example 4.28.

If we measure the distance from y to the ellipse B_Q in the Q-norm, then we must solve

$$\min_{x\in X} \|y-x\|_Q \quad \text{subject to} \quad \|x\|_Q^2 \le 1.$$

In this case,

$$L(\lambda, x) = \langle Q(y-x), y-x\rangle + \lambda(\langle Qx, x\rangle - 1),$$

and (5.15) becomes $(\lambda+1)Qx = Qy$. Rewriting this as $Q(y-(1+\lambda)x) = 0$ and using the fact that Q is nonsingular, we conclude that $x = y/(1+\lambda)$. If $\|y\|_Q > 1$, then we must have $1+\lambda = \|y\|_Q$.

Now suppose we measure the distance from y to B_Q in the norm induced by the original inner product. Then we must solve

$$\min_{x\in X} \|y-x\| \quad \text{subject to} \quad \|x\|_Q^2 \le 1.$$

In this case,

$$L(\lambda, x) = \langle y-x, y-x\rangle + \lambda(\langle Qx, x\rangle - 1),$$

and (5.15) becomes $(\lambda Q + I)x = y$.

Example 5.24 (1-norm Constrained Least Squares). Consider the problem

$$\min_{x\in \mathbb{R}^n} \|y-Ax\| \quad \text{subject to} \quad \|x\|_1 \le b,$$

where $\|x\|_1 := |x_1| + \cdots + |x_n|$ is the 1-norm on $\mathbb{R}^n$. We recast the problem as follows. First observe that

$$\sum_{k=1}^n |x_k| \le b \;\Leftrightarrow\; \exists \theta_k \ge 0 \text{ with } |x_k| \le \theta_k \text{ and } \sum_{k=1}^n \theta_k \le b.$$

The implication $\Rightarrow$ follows by taking $\theta_k := |x_k|$. The reverse implication is obvious. Since the condition $|x_k| \le \theta_k$ is equivalent to the pair of conditions $x_k \le \theta_k$ and $-x_k \le \theta_k$, the original problem is equivalent to

$$\min_{x\in\mathbb{R}^n,\theta\in\mathbb{R}^n} \|y-Ax\|^2 \quad \text{subject to} \quad \sum_{k=1}^{n} \theta_k \le b \ \text{ and } \begin{cases} x_k - \theta_k \le 0, \\ -x_k - \theta_k \le 0, \end{cases} \ k = 1,\dots,n.$$

We can rewrite this as follows. Put $z = [x^\mathsf{T}, \theta^\mathsf{T}]^\mathsf{T}$, $s := [0_{1\times n}, 1_{1\times n}]^\mathsf{T} \in \mathbb{R}^{2n}$,

$$\tilde{A} := \begin{bmatrix} A & 0_{m\times n} \end{bmatrix} \in \mathbb{R}^{m\times 2n}, \quad \text{and} \quad H := \begin{bmatrix} I_n & -I_n \\ -I_n & -I_n \end{bmatrix} \in \mathbb{R}^{2n\times 2n}.$$

With this notation, the above formulation is equivalent to

$$\min_{z\in\mathbb{R}^{2n}} \|y-\tilde{A}z\|^2 \quad \text{subject to} \quad \langle z, s\rangle \le b \text{ and } Hz \le 0.$$

This is a **quadratic programming problem**, which can be solved using the MATLAB function `quadprog`.

5.3.2. Water-Filling

Example 5.25 (Water-Filling: Existence of a Solution). Consider the problem

$$\max_{x\in\mathbb{R}^n_+} \sum_{k=1}^{n} F_k(x_k) \quad \text{subject to} \quad \sum_{k=1}^{n} x_k \le P,$$

where $P > 0$ is a given constant and the F_k are real-valued, concave functions on $[0,\infty)$ that satisfy some additional conditions mentioned below. A classical **water-filling** problem arises in determining the capacity of parallel Gaussian channels in information theory [10]. In that problem,

$$F_k(t) = \ln\Big(1 + \frac{t}{N_k}\Big), \quad t \ge 0, \tag{5.16}$$

where N_k is the noise power in the kth channel. The key features of the F_k that we need to solve the general problem are that their derivatives, denoted by f_k, be positive, strictly decreasing, continuous, and satisfy $f_k(t) \to 0$ as $t \to \infty$. Under these assumptions,

$$M_k := f_k(0) > 0$$

is the maximum value of f_k on $[0,\infty)$. Since f_k is stricly decreasing, it is a one-to-one mapping of $[0,\infty)$ to $(0,M_k]$. In fact, since derivatives have the **intermediate-value property** [35, p. 108, Th. 5.12], f_k is onto.[g] Therefore, $f_k^{-1}\colon (0,M_k] \to [0,\infty)$

[g] Alternatively, since we have assumed that the f_k are continuous, we could have appealed to the intermediate-value property of continuous functions [35, p. 93, Th. 4.23].

exists and is strictly decreasing (Problem 6.61(a)). For future reference, note that $f_k^{-1}(M_k) = 0$. When F_k is given by (5.16), $f_k(t) = 1/(t+N_k)$, $M_k = f_k(0) = 1/N_k$, and $f_k^{-1}(s) = (1/s) - N_k$.

We now turn to the optimization. First, maximizing $\sum_k F_k(x_k)$ is the same as minimizing $-\sum_k F_k(x_k)$. Since the F_k are concave, the $-F_k$ are convex. Hence, $-\sum_{k=1}^n F_k(x_k)$ is convex by Problem 5.18. Since $H(x) := \sum_k x_k - P$ is also convex, we can apply Theorem 5.16. The Lagrangian is

$$L(\lambda, x) := -\sum_{k=1}^n F_k(x_k) + \lambda \left(\sum_{k=1}^n x_k - P \right).$$

We must find $\lambda \geq 0$ and $x \in \mathbb{R}^n_+$ satisfying the three conditions

$$\sum_{k=1}^n x_k \leq P, \qquad \lambda \left(\sum_{k=1}^n x_k - P \right) = 0,$$

and

$$(D_x^+ L)(\lambda, x, y-x) = \sum_{k=1}^n \big[\lambda - f_k(x_k)\big](y_k - x_k) \geq 0, \quad \text{for all } y \in \mathbb{R}^n_+.$$

We first argue that $\lambda = 0$ is not going to work because the f_k are all positive. To see this, suppose $\lambda = 0$ and each $y_k > x_k$. Then

$$\sum_{k=1}^n \big[\lambda - f_k(x_k)\big](y_k - x_k) = -\sum_{k=1}^n f_k(x_k)(y_k - x_k) < 0.$$

So fix a $\lambda > 0$, and consider a k for which $M_k < \lambda$. For such k, if we take $x_k = 0$, then

$$\big[\lambda - f_k(x_k)\big](y_k - x_k) = \big[\lambda - f_k(x_k)\big] y_k \geq [\lambda - M_k] y_k \geq 0, \quad \text{for } y_k \geq 0.$$

For k with $M_k \geq \lambda$, λ is in the domain of f_k^{-1} and we can take $x_k = f_k^{-1}(\lambda)$; in other words, with this value of $x_k \geq 0$, $f_k(x_k) = \lambda$ and we have

$$\big[\lambda - f_k(x_k)\big](y_k - x_k) = 0(y_k - x_k) = 0, \quad \text{for } y_k \geq 0.$$

Choosing the x_k in this way guarantees that $(D_x^+ L)(\lambda, x, y-x) \geq 0$ for all $y \in \mathbb{R}^n_+$. The choice between $x_k = 0$ and $x_k = f_k^{-1}(\lambda)$ can be expressed compactly by writing $x_k = \xi_k(\lambda)$, where

$$\xi_k(\lambda) := \begin{cases} 0, & \lambda > M_k, \\ f_k^{-1}(\lambda), & 0 < \lambda \leq M_k. \end{cases}$$

If we put

$$g(\lambda) := \sum_{k=1}^n \xi_k(\lambda),$$

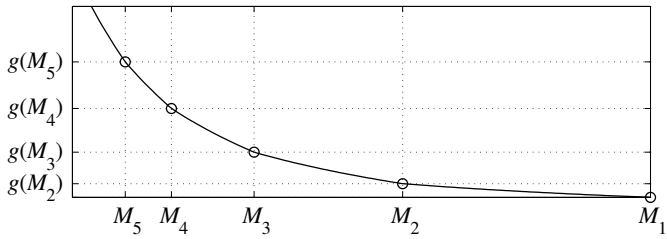


Figure 5.4. A typical function g when $M_n < \cdots < M_1$.

then $\sum_{k=1}^n x_k = \sum_{k=1}^n \xi_k(\lambda) = g(\lambda)$, and the constraint is satisfied if we can solve $g(\lambda) = P$. A typical function g is shown in Figure 5.4. From the graph, g is continuous, strictly decreasing on $(0, M_1]$, blows up near the origin, and is zero at M_1. By the intermediate-value theorem for continuous functions [35, p. 93, Th. 4.23], for any level $P > 0$, there is a $0 < \lambda < M_1$ such that $g(\lambda) = P$; the solution is unique because g is strictly decreasing. The properties of g that we need to justify these conclusions are that g be continuous, strictly decreasing on $(0, M_1]$, satisfy $g(\lambda) \to \infty$ as $\lambda \to 0$ and $g(M_1) = 0$. Fortunately, these properties hold in general as a consequence of the assumptions made above about the f_k.[4]

Remark. The value of the function $f_k^{-1}(\lambda)$ is just the solution of the equation $h(x) = 0$ when $h(x) = f_k(x) - \lambda$. The solution can be obtained in MATLAB with the commands

```
h = @(x)f(k,x)-lambda
x = fzero(h,Mk/2)
```

where `f(k,x)` is a function you write to compute $f_k(x)$, `fzero` is a MATLAB function, and $M_k/2$ is a starting point for `fzero`. Since it is easy to compute f_k^{-1} numerically, it is easy to compute ξ_k and g as well. To solve $g(\lambda) = P$, use the commands

```
gP = @(lambda)g(lambda)-P
lambda = fzero(gP,lstart)
```

where `lstart` might be $\max(M_1, \ldots, M_n)/2$.

Example 5.26 (Water-Filling: Why the Name?). Suppose $M_n < \cdots < M_1$. All we know in general is that the desired value of λ lies between zero and M_1. However, we can actually say a bit more without too much work. When $M_n < \cdots < M_1$, we have $g(M_n) > g(M_{n-1}) > \cdots > g(M_2) > g(M_1)$. Suppose we compute these n numbers and find that

$$g(M_{i+1}) > P \geq g(M_i).$$

Then the solution of $g(\lambda) = P$ must satisfy

$$M_{i+1} < \lambda \leq M_i.$$

The definitions of the ξ_k and g imply (see also Figures 5.5 and 5.4)

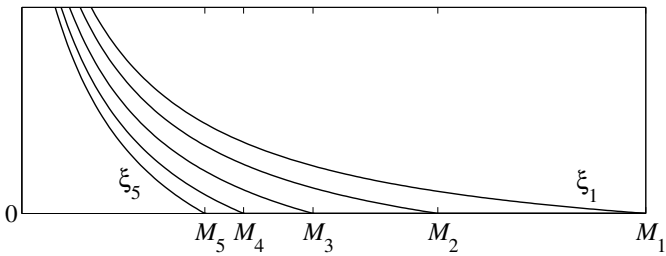


Figure 5.5. Typical functions ξ_k when $M_n < \cdots < M_1$.

$$g(\lambda) = \sum_{k=1}^{i} f_k^{-1}(\lambda), \quad M_{i+1} < \lambda \leq M_i, \tag{5.17}$$

where M_{n+1} taken as zero. Suppose we think of the terms $f_k^{-1}(\lambda)$ in (5.17) as representing the different depths of water in a swimming pool as shown at the left in Figure 5.6. In other words, we fill the pool with water by decreasing λ until the sum of the different depths is equal to P. For this reason, we call (5.17) the **water-filling equation**, and the process of finding λ is called **water-filling**.

When F_k is given by (5.16), $f_k^{-1}(\lambda) - f_{k+1}^{-1}(\lambda) = N_{k+1} - N_k$. If we put a platform of height N_1 underneath the entire swimming pool we get the right-hand picture in Figure 5.6.

Example 5.27 (Water-Filling: A Special Case). Suppose that $f_k^{-1}(s) = \beta(s) - c_k$, where the c_k are given constants, and β is a function that does not depend on k. Then using (5.17), we can rearrange $g(\lambda) = P$ as

$$\beta(\lambda) = \frac{1}{i}\left[P + \sum_{k=1}^{i} c_k\right].$$

In the classical water-filling problem, $\beta(\lambda) = 1/\lambda$, $c_k = N_k$, and we get

$$\lambda = \frac{i}{P + \sum_{k=1}^{i} N_k}.$$

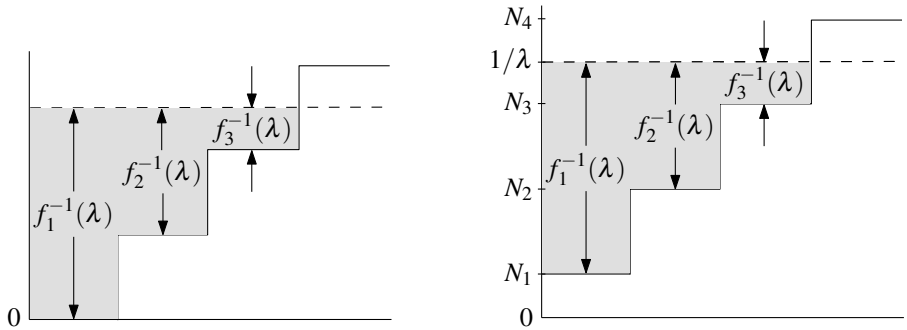


Figure 5.6. Left: Illustration of water-filling equation (5.17) when $i = 3$; the pool is filled by adjusting λ until the sum of the water depths is equal to P. Right: Illustration of water-filling adapted to (5.16) by setting the pool on a platform of height N_1.

Then $x_\ell = 0$ for $\ell > i$, and

$$x_\ell = f_\ell^{-1}(\lambda) = (1/\lambda) - N_\ell = \frac{1}{i}\left[P + \sum_{k=1}^{i} N_k\right] - N_\ell, \quad \ell = 1, \ldots, i.$$

This is would be a closed-form solution except for the determination of i. Fortunately, it is easy to find i by comparing P with the n numbers $g(M_i) = iN_i - \sum_{k=1}^{i} N_k$.

5.3.3. Portfolio Optimization

You are presented with the opportunity to invest money in a variety of assets; e.g., stocks, bonds, etc. For $k = 1, \ldots, n$, the kth asset has **rate of return** ρ_k. This means that if you invest d dollars in asset k, you will end up with an additional $d \cdot \rho_k$ dollars more than you started with. Equivalently, starting with d dollars, you end up with a total of $d(1 + \rho_k)$ dollars. The quantity $1 + \rho_k$ is called the **total return**.

Suppose you start out with initial wealth d, and you allocate a fraction x_k to asset k, where $x_1 + \cdots + x_n = 1$. In other words, you will invest all d dollars, with $d \cdot x_k$ dollars invested in asset k. In the end, you will have

$$\sum_{k=1}^{n} (d \cdot x_k)\rho_k = d \sum_{k=1}^{n} x_k \rho_k$$

more dollars than you started with. To maximize this expression, set $x_k = 1$ for some $k \in \operatorname{argmax}_k \rho_k$ and set all the other $x_k = 0$.

The challenge of the foregoing is that the rates of return ρ_k are *random*. We denote their means by $m_k := \mathsf{E}[\rho_k]$ so that the expected rate of return *per dollar invested* can

be written as

$$\mathsf{E}\left[\sum_{k=1}^{n} x_k \rho_k\right] = \sum_{k=1}^{n} x_k m_k = \langle x, m\rangle,$$

where $m := [m_1, \ldots, m_n]^\mathsf{T} \in \mathbb{R}^n$ is the vector of expected rates of return, and $x := [x_1, \ldots, x_n]^\mathsf{T} \in \mathbb{R}^n$ is called the **portfolio**. To maximize the expected rate of return, set $x_k = 1$ for some $k \in \operatorname{argmax}_k m_k$ and set all the other $x_k = 0$.

However, you probably would not feel comfortable investing all of your money in one stock, or even in the stock market as a whole. If prices fall, you lose money. So you would probably put some of your money in a savings account. Even though interest rates vary over time, you will not lose your original investment. The question becomes one of balancing return against how much risk you are willing to assume. The following formulation of this problem (**portfolio selection**) and its solution are due to Harry Markowitz [29], for which he was awarded the 1990 Nobel Prize in Economic Sciences (with Merton Miller and William Sharpe) [40].

Portfolio risk is defined as the standard deviation of the portfolio's rate of return; i.e., the square root of

$$\mathsf{var}\left(\sum_{k=1}^{n} x_k \rho_k\right) = \langle Cx, x\rangle,$$

where C is the **covariance matrix** of the rates of return; i.e., $C_{ij} := \mathsf{E}[(\rho_i - m_i)(\rho_j - m_j)]$. Given a desired expected portfolio rate of return r, among the vectors x satisfying

$$\langle x, m\rangle = r \quad \text{and} \quad \sum_{k=1}^{n} x_k = 1,$$

we seek the one that minimizes the portfolio risk, or equivalently the variance $\langle Cx, x\rangle$. To express the problem in compact notation, we combine the two equations in the preceding display into the matrix-vector equation $Bx = z$, where[h]

$$B := \begin{bmatrix} m_1 & m_2 & \cdots & m_n \\ 1 & 1 & \cdots & 1 \end{bmatrix} \quad \text{and} \quad z := \begin{bmatrix} r \\ 1 \end{bmatrix}. \tag{5.18}$$

Then our optimization problem becomes[i]

$$\min_{x \in \mathbb{R}^n} \tfrac{1}{2}\langle Cx, x\rangle \quad \text{subject to} \quad Bx = z. \tag{5.19}$$

This **quadratic programming problem** can be solved in closed form using Lagrange multipliers. As shown in Problem 5.43, the optimal solution x is obtained by solving

[h] Note that B^T is full rank as long as the m_k are not all the same; i.e., the top row of B is not proportional to the all ones vector. In this case, Theorem 4.13(g) implies B is onto.

[i] The factor of $1/2$ puts the problem in standard form and does not change the solution.

the linear equation equation

$$\begin{bmatrix} C & B^{\mathsf{T}} \\ B & 0 \end{bmatrix} \begin{bmatrix} x \\ \mu \end{bmatrix} = \begin{bmatrix} 0 \\ z \end{bmatrix}. \tag{5.20}$$

Assuming that the matrix on the left is invertible (which can be the case even if C is singular; see Problem 5.43),

$$x = \begin{bmatrix} I_{n\times n} & 0_{n\times 2} \end{bmatrix} \begin{bmatrix} C & B^{\mathsf{T}} \\ B & 0 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ z \end{bmatrix}. \tag{5.21}$$

Since z is given by (5.18), we see that the optimal x is a linear function of the desired rate of return r. Hence, the optimal variance $\langle Cx,x\rangle$ is a quadratic function of r, which we denote by $\sigma^2(r)$. In other words, the graph of $\sigma^2(r)$ is a parabola that opens upward (convex function of r), as illustrated in the graph at the left in Figure 5.7. Now suppose we are given a desired risk level σ_0. Then the equation $\sigma^2(r) = \sigma_0^2$ will usually have two solutions, say $r_1 < r_2$, corresponding to two different portfolios. These two portfolios induce the same risk, but the second one yields a higher return. This second portfolio is said to be **efficient**.

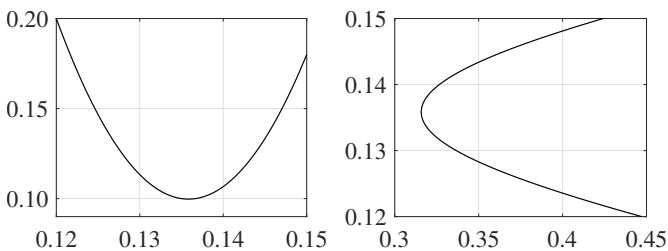


Figure 5.7. Left: Variance $\sigma^2(r)$ as a function of return r. For each portfolio x, the point $(\langle x,m\rangle, \langle Cx,x\rangle)$ must lie above or on the curve. Right: r plotted as a multivalued function of the risk σ. For each portfolio x, the point $(\langle Cx,x\rangle^{1/2}, \langle x,m\rangle)$ lies to the right of or on the curve.

Remarks. (i) Economists and financial engineers typically plot return as a multivalued function of the risk σ, and so their curves open to the right as shown at the right in Figure 5.7.

(ii) There is no guarantee that the solution x in (5.21) will have all nonnegative components. Having x_k negative corresponds to **shorting** a stock. To prevent this, we can restrict $x \in \mathbb{R}^n_+$. This yields the **quadratic programming problem**

$$\min_{x\in\mathbb{R}^n_+} \tfrac{1}{2}\langle Cx,x\rangle \quad \text{subject to} \quad Bx = [r,1]^{\mathsf{T}},$$

which can be solved using the MATLAB function `quadprog`.

The One-Fund Theorem

Suppose that $\rho_n \equiv m_n$; i.e., ρ_n is a constant random variable that has zero variance. For this reason, we call the nth asset the **risk-free asset**, and we call m_n the **risk-free rate of return**. The other assets are said to be **risky**. Because ρ_n has zero variance, the bottom row and right-most column of the covariance matrix C are zero. In other words, C has the form

$$C = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}.$$

We assume that D is positive definite and that at least one of $m_1, \ldots, m_{n-1}$ is different from m_n.

The **one-fund theorem** says that under a mild assumption made below, the optimal portfolio for any desired rate of return r can be expressed as the affine combination of a two fixed portfolios. The first one contains only the $n-1$ risky assets and achieves rate of return $r_0 \neq m_n$ with minimum risk, while the second one contains only the riskless asset and achieves rate of return m_n with zero risk.

Because of the partitioned form of C, it is convenient to put $x = [\xi^{\mathsf{T}} \; v]^{\mathsf{T}} \in \mathbb{R}^n$, where $\xi \in \mathbb{R}^{n-1}$, and $\mu := [\mu_1 \; \mu_2]^{\mathsf{T}} \in \mathbb{R}^2$. We also write

$$B = \begin{bmatrix} \mathbf{m}^{\mathsf{T}} & m_n \\ \mathbb{1}^{\mathsf{T}} & 1 \end{bmatrix}, \quad \text{where } \mathbf{m} := \begin{bmatrix} m_1 & \cdots & m_{n-1} \end{bmatrix}^{\mathsf{T}} \text{ and } \mathbb{1} := \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}^{\mathsf{T}} \in \mathbb{R}^{n-1}.$$

With this notation, we follow [8, Sec. 4] and observe that (5.20) becomes

$$\begin{aligned} D\xi + \mathbf{m}\mu_1 + \mathbb{1}\mu_2 &= 0 \\ m_n\mu_1 + \mu_2 &= 0 \\ \mathbf{m}^{\mathsf{T}}\xi + m_n v &= r \\ \mathbb{1}^{\mathsf{T}}\xi + v &= 1, \end{aligned} \tag{5.22}$$

where the first two equations correspond to the top row of (5.20), $Cx + B^{\mathsf{T}}\mu = 0$, and the last two equations correspond to the second row, $Bx = z$. At this point we have four linear equations in the four unknowns ξ, v, μ_1, and μ_2. The second equation can be used to solve for μ_2 and substituted into the first equation. The fourth equation can be solved for v and substituted into the third equation. This leaves us with two equations in the two unknowns ξ and μ_1. Doing the first substitution results in

$$\xi = -\mu_1 D^{-1}\{\mathbf{m} - m_n\mathbb{1}\}.$$

Doing the second substitution and rearranging yields

$$(\mathbf{m} - m_n\mathbb{1})^{\mathsf{T}}\xi = r - m_n.$$

Into this last formula substitute the one for ξ to get

$$-\mu_1 \big\langle D^{-1}\{\mathbf{m}-m_n\mathbb{1}\}, \mathbf{m}-m_n\mathbb{1}\big\rangle = r-m_n.$$

Solving for μ_1 and substituting into the above formula for ξ yields

$$\xi = \frac{r-m_n}{\big\langle D^{-1}\{\mathbf{m}-m_n\mathbb{1}\}, \mathbf{m}-m_n\mathbb{1}\big\rangle} D^{-1}\{\mathbf{m}-m_n\mathbb{1}\}. \tag{5.23}$$

Write $\xi(r)$ to emphasize the dependence of ξ on the desired rate of return r. The key observation is that given any $r_0 \neq m_n$,

$$\begin{aligned}\xi(r) &= \frac{r-m_n}{r_0-m_n} \cdot \frac{r_0-m_n}{\big\langle D^{-1}\{\mathbf{m}-m_n\mathbb{1}\}, \mathbf{m}-m_n\mathbb{1}\big\rangle} D^{-1}\{\mathbf{m}-m_n\mathbb{1}\} \\ &= \frac{r-m_n}{r_0-m_n}\xi(r_0).\end{aligned}$$

The next step is to use this formula for ξ in (5.22) to write

$$v = 1 - \mathbb{1}^{\mathsf{T}}\xi = 1 - \frac{r-m_n}{r_0-m_n}\mathbb{1}^{\mathsf{T}}\xi(r_0).$$

When $r = r_0$, $v = 1-\mathbb{1}^{\mathsf{T}}\xi(r_0)$. We would like to choose r_0 with $\mathbb{1}^{\mathsf{T}}\xi(r_0) = 1$ so that when $r = r_0$, the optimal portfolio has $v = 0$. Using (5.23), we find that

$$r_0 = m_n + \frac{\big\langle D^{-1}\{\mathbf{m}-m_n\mathbb{1}\}, \mathbf{m}-m_n\mathbb{1}\big\rangle}{\mathbb{1}^{\mathsf{T}} D^{-1}\{\mathbf{m}-m_n\mathbb{1}\}},$$

assuming the denominator is not zero.[j] Under this assumption, we can write

$$x = \begin{bmatrix}\xi \\ v\end{bmatrix} = \frac{r-m_n}{r_0-m_n}\begin{bmatrix}\xi(r_0) \\ 0\end{bmatrix} + \left(1 - \frac{r-m_n}{r_0-m_n}\right)\begin{bmatrix}0 \\ 1\end{bmatrix}.$$

Thus, the optimal portfolio for any desired rate of return r is an affine combination of a two fixed portfolios. The first one contains only the $n-1$ risky assets and achieves rate of return $r_0 \neq m_n$ with minimum risk; the second one contains only the riskless asset and achieves rate of return m_n with zero risk.

Notes

Note 5.1. We establish the formula for the Gâteaux derivative in Theorem 5.8. To begin, fix any $x = [x_1, \ldots, x_n]^{\mathsf{T}}$ and $y = [y_1, \ldots, y_n]^{\mathsf{T}}$ in $\mathbb{R}^n_+$, and write[k]

$$f\big(x+\lambda(y-x)\big) - f(x) = \sum_{k=1}^{n} f(z_k) - f(z_{k-1}), \tag{5.24}$$

[j] If the denominator is zero, then by (5.23), $\mathbb{1}^{\mathsf{T}}\xi = 0$ for all r.

[k] Here z_k is a vector, not a component of a vector.

where $z_0 := x$ and for $k = 1, \dots, n$,

$$z_k := \big[x_1 + \lambda(y_1 - x_1), \dots, x_k + \lambda(y_k - x_k), x_{k+1}, \dots, x_n\big]^{\mathsf{T}}. \tag{5.25}$$

Since x_i and y_i are nonnegative, it is clear that each component of z_k is nonnegative, and so $z_k \in \mathbb{R}^n_+$. In particular, $z_n = x + \lambda(y - x)$. Since z_k and z_{k-1} differ only in their kth component,

$$f(z_k) - f(z_{k-1}) = g\big(x_k + \lambda(y_k - x_k)\big) - g(x_k),$$

where g is the function of one variable

$$g(t) := f\big(x_1 + \lambda(y_1 - x_1), \dots, x_{k-1} + \lambda(y_{k-1} - x_{k-1}), t, x_{k+1}, \dots, x_n\big)$$

and t varies in the closed interval with end points x_k and $x_k + \lambda(y_k - x_k)$. Since g is differentiable on this interval, it is continuous on this interval [35, p. 104, Theorem 5.2]. Hence, we may apply the mean-value theorem [35, p. 108, Theorem 5.10] to write

$$g(x_k + \lambda(y_k - x_k)) - g(x_k) = \lambda g'(\tau)(y_k - x_k), \tag{5.26}$$

where τ lies strictly between x_k and $x_k + \lambda(y_k - x_k)$. Hence, $|\tau - x_k| < \lambda|y_k - x_k|$. In terms of the kth partial derivative of f, (5.26) says

$$f(z_k) - f(z_{k-1}) = \lambda f_k(w_k)(y_k - x_k),$$

where

$$w_k := \big[x_1 + \lambda(y_1 - x_1), \dots, x_{k-1} + \lambda(y_{k-1} - x_{k-1}), \tau, x_{k+1}, \dots, x_n\big]^{\mathsf{T}}.$$

Note that

$$\begin{aligned}\|w_k - x\|^2_{\mathbb{R}^n} = \lambda^2 \sum_{i=1}^{k-1} |y_i - x_i|^2 + |\tau - x_k|^2 &< \lambda^2 \sum_{i=1}^{k} |y_i - x_i|^2 \\ &\le \lambda^2 \sum_{i=1}^{n} |y_i - x_i|^2 = \lambda^2 \|y - x\|^2_{\mathbb{R}^n}.\end{aligned}$$

Hence, for small $\lambda > 0$, w_k is close to x. We can now write

$$\begin{aligned}f\big(x + \lambda(y - x)\big) - f(x) &= \sum_{k=1}^{n} f(z_k) - f(z_{k-1}) \\ &= \sum_{k=1}^{n} \lambda f_k(w_k)(y_k - x_k).\end{aligned}$$

It follows that

$$\frac{f\big(x+\lambda(y-x)\big)-f(x)}{\lambda}-\langle y-x,\nabla_x f\rangle_{\mathbb{R}^n}=\sum_{k=1}^{n}f_k(w_k)(y_k-x_k)-\sum_{k=1}^{n}f_k(x)(y_k-x_k)$$
$$=\sum_{k=1}^{n}[f_k(w_k)-f_k(x)](y_k-x_k).$$

Given $\varepsilon>0$, for small enough $\lambda>0$, $|f_k(w_k)-f_k(x)|<\varepsilon/n$ by continuity of the partial derivatives. Hence,

$$\left|\sum_{k=1}^{n}[f_k(w_k)-f_k(x)](y_k-x_k)\right|\le\frac{\varepsilon}{n}\sum_{k=1}^{n}\|y-x\|_{\mathbb{R}^n}=\varepsilon\|y-x\|_{\mathbb{R}^n}.$$

Remark. The foregoing proof can be adapted to establish the analogous result in Theorem 5.6 by making the following changes. First, replace $y-x$ with Δx. Second, change the sentence below (5.25) to say that for small λ, each $z_k\in U$. In fact, we can even allow $\lambda<0$ if we change "small $\lambda>0$" to "small $|\lambda|>0$." In this way, we can show that under the assumptions of Theorem 5.6, f has a ***two*-sided Gâteaux derivative**.

Note 5.2. We prove Theorem 5.10. We begin with some basic inequalities. For any $u<x<v$ in the interval I, we can write x in two different ways as

$$x=u+\lambda(v-u),\quad\text{with }\lambda=(x-u)/(v-u)\in(0,1),$$

and

$$x=v+\lambda(u-v),\quad\text{with }\lambda=(x-v)/(u-v)\in(0,1).$$

Since f is convex, we have

$$f(x)\le f(u)+\frac{x-u}{v-u}\big[f(v)-f(u)\big]\quad\text{and}\quad f(x)\le f(v)+\frac{x-v}{u-v}\big[f(u)-f(v)\big].$$

Rewrite the above display noting that $x-u$ is positive and $x-v$ is negative. Thus,

$$\frac{f(x)-f(u)}{x-u}\le\frac{f(v)-f(u)}{v-u}\quad\text{and}\quad\frac{f(v)-f(u)}{v-u}\le\frac{f(v)-f(x)}{v-x}.\tag{5.27}$$

It follows that

$$\frac{f(x)-f(u)}{x-u}\le\frac{f(v)-f(x)}{v-x}.\tag{5.28}$$

The inequalities in (5.27) and (5.28) are illustrated in Figure 5.8.

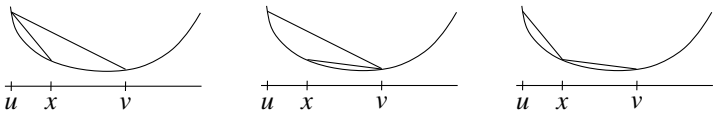


Figure 5.8. The left-hand graph illustrates the left-hand inequality in (5.27), which says that the slope of the chord above $[u,x]$ is less than or equal to the slope of the chord above $[u,v]$. The middle graph illustrates the right-hand inequality in (5.27), which says that the slope of the chord above $[u,v]$ is less than or equal to the slope of the chord above $[x,v]$. The right-hand graph illustrates (5.28), which says that the slope of the chord above $[u,x]$ is less than or equal to the slope of the chord above$[x,v]$.

If $u < x < v < w$, then making the substitution $(u,x,v) \to (x,v,w)$ in (5.28) yields

$$\frac{f(v)-f(x)}{v-x} \le \frac{f(w)-f(v)}{w-v}.$$

Combining this with (5.28) yields

$$\frac{f(x)-f(u)}{x-u} \le \frac{f(w)-f(v)}{w-v}. \tag{5.29}$$

Now suppose that $a < b$ are interior points of I. Then there are other interior points $p,q \in I$ with $p < a < b < q$. Consider points

$$p < a < s < t < b < q.$$

From (5.29) with $(u,x,v,w) = (p,a,s,t)$,

$$\frac{f(a)-f(p)}{a-p} \le \frac{f(t)-f(s)}{t-s}. \tag{5.30}$$

From the left-hand inequality in (5.27) with $(u,x,v) = (p,a,t)$,

$$\frac{f(a)-f(p)}{a-p} \le \frac{f(t)-f(a)}{t-a},$$

which shows that (5.30) holds even for $s = a$. Similarly, from (5.29) with $(u,x,v,w) = (s,t,b,q)$,

$$\frac{f(t)-f(s)}{t-s} \le \frac{f(q)-f(b)}{q-b}, \tag{5.31}$$

and from the right-hand inequality in (5.27) with $(u,x,v) = (s,b,q)$,

$$\frac{f(b)-f(s)}{b-s} \le \frac{f(q)-f(b)}{q-b}.$$

Thus, (5.31) holds even for $t = b$. Letting K denote the larger of the absolute values of the left-hand side of (5.30) and of the right-hand side of (5.31), we have

$$|f(t) - f(s)| \le K|t - s|, \quad s,t \in [a,b].$$

Thus, f is **Lipschitz continuous** on $[a,b]$.

We now turn to the derivatives. The left-hand inequality in (5.27) shows that for fixed u, $\frac{f(t)-f(u)}{t-u}$ is increasing in t (nondecreasing, to be precise). The right-hand inequality in (5.27) shows that for fixed v, $\frac{f(v)-f(t)}{v-t}$ is increasing in t. If we fix v on the right-hand side of (5.28), then the left-hand side is bounded above and is increasing in u as u increases to x. Similarly, if we fix u on the left-hand side of (5.28), then the right-hand side is bounded below and is decreasing in v as v decreases to x. Hence, the limits

$$f'_-(x) := \lim_{u \uparrow x} \frac{f(x) - f(u)}{x - u} \le \lim_{v \downarrow x} \frac{f(v) - f(x)}{v - x} =: f'_+(x) \tag{5.32}$$

exist and are finite. From (5.28),

$$\frac{f(x) - f(u)}{x - u} \le f'_+(x), \quad u < x.$$

On the other hand, since $f'_+(x)$ is the limit of decreasing quotients, the limit is less than or equal to any particular quotient; i.e.,

$$f'_+(x) \le \frac{f(v) - f(x)}{v - x}, \quad v > x.$$

Given any $y \in I$, by taking $u = y$ if $y < x$ and taking $v = y$ if $y > x$, we see that

$$f(y) \ge f(x) + f'_+(x)(y - x).$$

This also holds trivially for $y = x$. A similar argument can be made with $f'_+(x)$ replaced by $f'_-(x)$. □

To conclude, we show that f'_+ and f'_- are nondecreasing. In (5.29) let $x \to u$ and then $w \to v$ to get $f'_+(u) \le f'_+(v)$. Alternatively, letting $u \to x$ and then $v \to w$, yields $f'_-(x) \le f'_-(w)$.

Note 5.3. We prove Jensen's inequality, Theorem 5.13. Let C denote the interval of interest. Put $m := \mathsf{E}[X]$. Then m belongs to the interval C. If C is an interval that includes one of its endpoints, say c_0, then either $m = c_0$ or $m \ne c_0$. If $m = c_0$, then $X = c_0$ almost surely. In this case, $\mathsf{E}[f(X)] = \mathsf{E}[f(m)] = f(m) = f(\mathsf{E}[X])$. If m is not an endpoint of the interval, then convexity of f implies

$$f(x) \ge f(m) + f^*(m)(x - m), \quad \text{for all } x \in C, \tag{5.33}$$

where f^* denotes either the left-hand or right-hand derivative of f (cf. Theorem 5.10). Assuming $\mathsf{E}[f(X)]$ is finite, we have from (5.33) that

$$\mathsf{E}[f(X)] \geq f(m) + f^*(m) \underbrace{(\mathsf{E}[X] - m)}_{=0} = f(m) = f(\mathsf{E}[X]).$$

Below we show that $\mathsf{E}[f(X)^-] < \infty$.[l] Hence, $\mathsf{E}[f(X)]$ exists. It then follows that if $\mathsf{E}[f(X)]$ is not finite, it is $+\infty$; in this case, Jensen's inequality is obviously true. Let $B := \{x \in C : f(x) < 0\}$. Then[m]

$$\begin{aligned}
\mathsf{E}[f(X)^-] &= \mathsf{E}[\mathbf{1}_B(X) f(X)^-] \\
&= \mathsf{E}[-f(X)\mathbf{1}_B(X)] \\
&\leq \mathsf{E}[\{-f(m) - f^*(m)(X-m)\}\mathbf{1}_B(X)], \quad \text{by (5.33)}, \\
&= \{f^*(m)m - f(m)\}\mathsf{P}(X \in B) - f^*(m)\mathsf{E}[X\mathbf{1}_B(X)].
\end{aligned}$$

The last expectation on the right is finite since we have assumed $\mathsf{E}[|X|] < \infty$. Hence, $\mathsf{E}[f(X)^-] < \infty$.

Note 5.4. Properties of g in Water-Filling Example 5.25. Without loss of generality, suppose that $M_n < \cdots < M_1$. We must show that g is continuous on $(0,\infty)$, strictly decreasing on $(0,M_1]$, and satisfies $g(\lambda) \to \infty$ as $\lambda \to 0$ and $g(\lambda) = 0$ for $\lambda \geq M_1$. As noted in Example 5.25, the properties of f_k imply f_k^{-1} is continuous and strictly decreasing on $(0,M_k]$ with $f_k^{-1}(M_k) = 0$. Hence, ξ_k is continuous on $(0,\infty)$, strictly decreasing on $(0,M_k]$, and satisfies $\xi_k(\lambda) = 0$ for $\lambda \geq M_k$. It follows that g is continuous on $(0,\infty)$, strictly decreasing on $(0,M_1]$, and satisfies $g(\lambda) = 0$ for $\lambda \geq M_1$. Since $f_k(t) \to 0$ as $t \to \infty$, we have that as $\lambda \to 0$, $f_k^{-1}(\lambda) \to \infty$. Hence, the same is true for ξ_k and for g.

Problems

1. Show that (5.2) holds if and only if

$$L(\lambda, x_0) \leq L(\lambda_0, x_0), \quad \text{for all } \lambda \in \mathbb{R}_+^m.$$

Remark. Using this result, (5.2) and (5.3) can be combined into the pair of inequalities,

$$L(\lambda, x_0) \leq L(\lambda_0, x_0) \leq L(\lambda_0, x), \quad \text{for all } \lambda \in \mathbb{R}_+^m, x \in X_0. \tag{5.34}$$

In other words, (λ_0, x_0) is a **saddle point** of L. Theorem 5.1 is often stated with (5.2) and (5.3) replaced by (5.34), e.g., [41, p. 59].

[l] For any random variable Y, we define $Y^- := -Y$ if $Y < 0$ and $Y^- := 0$ otherwise.

[m] Here $\mathbf{1}_B$ denotes the **indicator function** defined by $\mathbf{1}_B(x) := 1$ if $x \in B$ and $\mathbf{1}_B(x) := 0$ if $x \notin B$.

2. Prove Theorem 5.3.

3. Show that $\mathbb{R}^d_+$ is a convex set.

4. Let $A: X \to Y$, where A is a linear transformation from the vector space X to the vector space Y. Suppose that C is a nonempty convex subset of X. Put

$$D := \{Ax : x \in C\}.$$

Determine whether or not D is convex.

5. Let f be a real-valued, convex function defined on the real line $\mathbb{R}$. Assume that $x_0 \in \mathbb{R}$ minimizes f on $\mathbb{R}$; i.e., $f(x_0) \le f(x)$ for all $x \in \mathbb{R}$. Determine whether or not f is nondecreasing on $[x_0, \infty)$; i.e., for $x_0 \le x_1 \le x_2$, determine whether or not $f(x_0) \le f(x_1) \le f(x_2)$.

6. Let f be a concave function on a convex set C. Suppose that C can be expressed as the convex hull of a finite set of points, say $C = \operatorname{co}\{v_1, \ldots, v_p\}$ (such a set is called a (convex) **polytope**). Let $f(v_k) = \min\{f(v_1), \ldots, f(v_p)\}$. Show that v_k minimizes f on C. Hence, minimization of a concave function on a polytope is easy if p is not too large.

7. Show that a strictly convex function can have at most one **minimizer**.

8. Let C be a convex subset of a vector space X, and let f be a real-valued, convex function defined on C. Suppose that there is an $x_0 \in C$ such that $f(x_0) \le f(x)$ for all $x \in C$. Determine whether or not $E := \{x \in C : f(x) = f(x_0)\}$ is a convex set.

9. ***Strong convexity.*** A convex function f on a convex set C in a real or complex inner-product space X is said to be **strongly convex** with parameter $\mu > 0$ if the function $g(x) := f(x) - \frac{\mu}{2}\|x\|^2$ is convex.

   (a) Show that if f_0 is convex on C and $x_0 \in X$, then

$$f_\mu(x) := f_0(x) + \tfrac{\mu}{2}\|x - x_0\|^2, \quad \mu > 0,$$

   is strongly convex.

   (b) Show that if f is strongly convex, then f satisfies the strengthening of (5.9),

$$f(y) \ge f(x) + (D^+ f)(x, y - x) + \tfrac{\mu}{2}\|y - x\|^2, \quad x, y \in C.$$

   (c) Show that a function f on C is strongly convex if and only if it satisfies the strengthening of (5.6),

$$f\big(\lambda x + (1-\lambda)y\big) + \lambda(1-\lambda) \cdot \tfrac{\mu}{2}\|x - y\|^2 \le \lambda f(x) + (1-\lambda) f(y),$$

for all $0 \le \lambda \le 1$ and all $x, y \in C$, *Hint:* First derive the identity that for all real a,

$$a\|x\|^2 + (1-a)\|y\|^2 - \|ax + (1-a)y\|^2 = a(1-a)\|x-y\|^2.$$

(d) Show that a strongly convex function is strictly convex.

Remarks. A convex function f_0 as in part (a) may not be strictly convex; e.g., $f_0(x) = x$ for $x \in \mathbb{R}$. However, the quadratic term added to form f_μ makes f_μ strongly convex and therefore strictly convex by part (d). A convex function f_0 may not be bounded below; e.g., $f_0(x) = -\ln x$ for $x \in (0, \infty)$. However, most convex functions cannot decrease faster than linearly.[n] If f_0 is such a function, then adding the quadratic term in the definition of f_μ means that $f_\mu(x) \to \infty$ as $\|x\| \to \infty$ (such a function is said to be **coercive**; see Problem 5.23). This implies that any minimum of f_μ must occur inside some closed ball of finite radius. With some additional assumptions, one can guarantee the existence of a minimizer of f_μ, which, as just noted, must be unique. For example, if f_0 is continuous, then so is f_μ; if X is finite dimensional, then by Theorems 6.41 and 6.44, f_μ has a unique minimizer. For a more general result, see [4, Corollary 11.16]. When f_0 is such that f_μ always has a unique minimizer, one can define the **proximal mapping**,

$$(\operatorname{prox} f_0)(x) := \operatorname*{argmin}_{y \in C} \left[f_0(y) + \tfrac{1}{2}\|y - x\|^2 \right], \quad x \in X.$$

For $\mu > 0$,

$$(\operatorname{prox}(\tfrac{1}{\mu} f_0))(x) = \operatorname*{argmin}_{y \in C} \left[\tfrac{1}{\mu} f_0(y) + \tfrac{1}{2}\|y - x\|^2 \right] = \operatorname*{argmin}_{y \in C} \left[f_0(y) + \tfrac{\mu}{2}\|y - x\|^2 \right].$$

10. Consider the function $f(x) := |x|$ for $x \in \mathbb{R}$. Show that

$$(D^+ f)(x, y-x) = \begin{cases} y - x, & x > 0, \\ x - y, & x < 0, \\ y - x, & x = 0 \text{ and } y > 0, \\ x - y, & x = 0 \text{ and } y < 0. \end{cases}$$

Remarks. (i) Since $(D^+ f)(0, \Delta x) = |\Delta x|$, we see that the one-sided Gâteaux derivative is not necessarily a **linear** function of Δx. (ii) Although the one-sided Gâteaux derivative of $f(x) = |x|$ exists at $x = 0$, the function $f(x) = |x|$ does not have a derivative at $x = 0$ in the usual calculus sense because the left and right derivatives there are not equal.

[n] This is illustrated in Problem 5.22. See [4, Theorem 9.19] for the general case.

11. Show that the one-sided Gâteaux derivative satisfies

$$(D^+f)\big(x,a(y-x)\big) = a(D^+f)(x,y-x), \quad a>0.$$

In other words, the one-sided Gâteaux derivative is **homogeneous** in its direction argument for positive constants.

12. Let X denote the real vector space of continuous waveforms on $[0,1]$. Let $t_0 \in [0,1]$ be given, and for $x \in X$, put $f(x) := [x(t_0)]^2$. Find the Gâteaux derivative of f at x in the direction Δx.

13. For convex functions on an interval of the real line, (5.9) becomes

$$f(y) \geq f(x) + f'(x)(y-x), \tag{5.35}$$

for x,y in the interval where f is defined, and assuming f is differentiable in the usual sense at x. Prove that if f is a real-valued function that is differentiable on an interval and if (5.35) holds for all x,y in the interval, then f is convex. *Hint:* Given any two points u,v in the interval, show that

$$f\big(u+\lambda(v-u)\big) \leq f(u) + \lambda[f(v)-f(u)], \quad 0<\lambda<1.$$

(The cases $\lambda = 0$ or 1 are trivial.) In (5.35) put $x = u+\lambda(v-u)$ and $y = u$. Repeat with the same x but now $y = v$. Use the two resulting inequalities to obtain a lower bound on $(1-\lambda)f(u)+\lambda f(v)$.

14. Let f be a real-valued function on an interval of the real line, and assume the usual derivative $f'(x)$ exists. Show that if f' is nondecreasing (resp. strictly increasing), then f is convex (resp. strictly convex). *Hint:* Use the ordinary mean-value theorem to show that (5.35) holds for all x,y. Treat the cases $x<y$ and $x>y$ separately. Apply the result of the preceding problem.

Remark. Conversely, if f is convex and differentiable, it is necessary that f' be nondecreasing, since in this case $f' = f'_- = f'_+$ and Theorem 5.10 tells us that f'_+ is nondecreasing.

15. Use the result of the previous problem to show that the following functions are strictly convex. (a) $f(x) = e^x$. (b) $f(x) = -\ln x$ for $x>0$. (c) $f(x) = x\ln x$ for $x>0$. (d) $f(x) = x^p$ for $x>0$ (assume $p>1$).

16. Put $f(x) := x\ln x$ for $x>0$ and $f(0)=0$. Show that f is strictly convex on $[0,\infty)$.

17. For $x \in \mathbb{R}^n_+$, put

$$f(x) := \sum_{k=1}^{n} x_k \ln x_k,$$

where $0\ln 0 := 0$. Use the result of the previous problem to show that f is convex.

18. Let C be a convex set on which convex functions $f_1, \ldots, f_n$ are defined. Let $c_1, \ldots, c_n$ be nonnegative constants, and put $g(x) := \sum_{k=1}^{n} c_k f_k(x)$. Show that g is convex. In other words, a nonnegative linear combination of convex functions is convex.

19. Is every strictly convex function strongly convex? *Hint:* Consider the function $f(x) := x^4$. Problem 5.14, including the remark following it may be helpful.

20. Let g be a real-valued function on a convex set C. Show that g is convex on C if and only if for each pair $x, y \in C$, the function $h(t) := g(tx + (1-t)y)$ for $t \in [0,1]$ is convex on $[0,1]$.

21. ***Gradient Vectors.*** For a real-valued function f defined on an open set U of a real inner-product space X, the **gradient** of f at $x \in U$ is the vector $\nabla_x f \in X$ having the property that given any $\varepsilon > 0$, for all sufficiently small $\|\Delta x\|$,[o]

$$\left| f(x + \Delta x) - f(x) - \langle \Delta x, \nabla_x f \rangle \right| \leq \varepsilon \|\Delta x\|.$$

Suppose S and T are bounded linear operators mapping a real inner-product space X to a real inner-product space Y.[p] If S and T have adjoints S^* and T^*, and if $f(x) := \langle Sx, Tx \rangle$, show that $\nabla_x f = (S^*T + T^*S)x$.

Remarks. In the special case that $T = S$ so that $f(x) = \|Sx\|^2$, the formula simplifies to $\nabla_x f = 2S^*Sx$, and if $Y = X$ and $T = S = I$ so that $f(x) = \|x\|^2$, then $\nabla_x f = 2x$. If f is an arbitrary real-valued function on $\mathbb{R}^n$ and f has continuous partial derivatives, a trivial adaptation of the method of Note 5.1 can be used to show that $\nabla_x f = [\partial f/\partial x_1, \ldots, \partial f/\partial x_n]^{\mathsf{T}}$.

22. We can generalize Problem 5.13 as follows. Let f be a real-valued function defined on an open subset U of a real inner-product space X. Let C be a convex subset of U, and assume $\nabla_x f$ exists for $x \in C$. Show that f is convex on C if and only if

$$f(y) \geq f(x) + \langle y - x, \nabla_x f \rangle, \quad x, y \in C.$$

Hints: To show that the inequality implies convexity, follow the hints for Problem 5.13. To show the necessity of the inequality, observe that by (5.9), it suffices to show that

$$(D^+ f)(x, y - x) = \langle y - x, \nabla_x f \rangle.$$

[o] If $x \in U$, then $f(x)$ is defined since we assumed f was defined on U. However, to make sure that $f(x + \Delta x)$ is also defined, at least for small Δx, we have required U to be an open set. A set U in an inner-product space is **open** if for every $x \in U$, all points close to x also lie in U.

[p] A linear operator S is bounded if for some finite, nonnegative constant C, $\|Sx\| \leq C\|x\|$ holds for all vectors x. This concept is explored in more detail in Section 7.2.

Remark. Combining the above formula with (5.8) shows that if x minimizes f on C, then $\langle y-x, \nabla_x f\rangle \geq 0$ for all $y \in C$. In the special case that $C = X$, we can apply this inequality to $y = x \pm \nabla_x f$ to show that $\pm\|\nabla_x f\|^2 \geq 0$, which implies $\nabla_x f = 0$.

23. Let f be as in the preceding problem, and further assume that f is strongly convex as in Problem 5.9. Show that $f(y) \to \infty$ as $\|y\| \to \infty$; i.e., f is **coercive**. *Hints:* If $g(x) := f(x) - \frac{\mu}{2}\|x\|^2$, then $\nabla_x g = \nabla_x f - \mu x$. The Cauchy–Schwarz inequality may also be helpful.

24. Let f be as in Problem5.22, and suppose that ∇f is **Lipschitz continuous**, say[q]

$$\|\nabla_y f - \nabla_x f\| \leq L\|y-x\|, \quad x, y \in C,$$

for some finite, nonnegative constant L. The following steps show that

$$f(y) \leq f(x) + \langle y-x, \nabla_x f\rangle + \tfrac{L}{2}\|y-x\|^2, \quad x, y \in C. \tag{5.36}$$

(a) Show that if $g(x) := \frac{L}{2}\|x\|^2 - f(x)$ is convex, then (5.36) holds. *Hint:* Apply Problem 5.22 to g and note that $\nabla_x g = Lx - \nabla_x f$.

(b) Show that g is convex. *Hints:* By Problem 5.20, it suffices to show that $h(t) := g(tx + (1-t)y)$ is convex for $t \in [0,1]$. By Problem 5.14, h will be convex if h' is nondecreasing. Use the fact that $h'(t) = \langle x-y, \nabla_{tx+(1-t)y} g\rangle$. Note also that the Cauchy–Schwarz inequality and the Lipschitz condition imply

$$\langle y-x, \nabla_u f - \nabla_v f\rangle \leq L\|y-x\|\,\|u-v\|.$$

Remarks. (*i*) Applying the convexity property of Problem 5.22 to (5.36) shows that

$$0 \leq f(y) - f(x) - \langle y-x, \nabla_x f\rangle \leq \tfrac{L}{2}\|y-x\|^2. \tag{5.37}$$

(*ii*) Using (5.36) with the additional assumption that f is *strongly* convex, allows us to apply Problems 5.9(b) to write

$$f(x) + \langle y-x, \nabla_x f\rangle + \tfrac{\mu}{2}\|y-x\|^2 \leq f(y) \leq f(x) + \langle y-x, \nabla_x f\rangle + \tfrac{L}{2}\|y-x\|^2.$$

Such a function of y is "sandwiched" between two quadratic functions of y that have the same constant and linear terms.

25. Let f be as in the preceding problem. Show that

$$f(x) - f(y) \leq \langle x-y, \nabla_x f\rangle - \tfrac{1}{2L}\|\nabla_x f - \nabla_y f\|^2, \quad x, y \in C.$$

[q] Such a function is said to be L-smooth.

Hints: The convexity of f implies that for any x and z, $f(z) \geq f(x) + \langle z-x, \nabla_x f\rangle$, or $f(x) - f(z) \leq \langle x-z, \nabla_x f\rangle$. By Problem 5.24, for any z and y, $f(z) - f(y) \leq \langle z-y, \nabla_y f\rangle + \frac{L}{2}\|z-y\|^2$. It follows that

$$\begin{aligned} f(x) - f(y) &= f(x) - f(z) + f(z) - f(y) \\ &\leq \langle x-z, \nabla_x f\rangle + \langle z-y, \nabla_y f\rangle + \tfrac{L}{2}\|z-y\|^2. \end{aligned}$$

Now substitute $z = y - \frac{1}{L}(\nabla_y f - \nabla_x f)$ and simplify.

26. Use the result of the preceding problem to show that

$$\tfrac{1}{L}\|\nabla_x f - \nabla_y f\|^2 \leq \langle x-y, \nabla_x f - \nabla_y f\rangle.$$

Hint: Write out the result of the preceding problem. Then write it again with x and y interchanged.

27. ***Monotone Gradient.*** Let f be a real-valued function defined on an open subset U of a real inner-product space X. Let C be a convex subset of U, and assume $\nabla_x f$ exists for $x \in C$. We say that ∇f is **monotone** on C if[r]

$$\langle y-x, \nabla_y f - \nabla_x f\rangle \geq 0, \quad x, y \in C.$$

Show that f is a convex function on C if and only if ∇f is monotone on C. *Hints:* By Problem 5.20, f is convex on C if and only if $h(t) := f(x+t(y-x))$ is convex on $[0,1]$, and by Problem 5.14, h is convex if and only if h' is nondecreasing. Note also that $h'(t) = \langle y-x, \nabla_{x+t(y-x)} f\rangle$.

28. ***Strongly Monotone Gradient.*** Let f be as in the preceding problem. We say that ∇f is ***strongly*** **monotone** with parameter $\mu > 0$ if

$$\langle y-x, \nabla_y f - \nabla_x f\rangle \geq \mu\|y-x\|^2, \quad x, y \in C.$$

Show that f has a strongly monotone gradient with parameter μ if and only if f is strongly convex with parameter μ as defined in Problem 5.9.

29. Let X_0 be a convex subset of a real or complex vector space X. Fix any $x_0 \in X_0$, and put $C := \{x \in X : x_0 + x \in X_0\}$.

    (a) Show that C is convex.

    (b) Show that if $x \in C$ and $0 \leq \eta \leq 1$, then $x_0 + \eta x \in X_0$.

[r] The notion of monotone gradient does not require that C be convex.

(c) For $x \in C$, put $\varphi(x) := (D^+ f)(x_0, x)$. Show that φ is a convex function on C. *Hint:* Fix $x_1, x_2 \in C$, $0 \le \lambda \le 1$, and $0 < \eta \le 1$. Observe that

$$\lambda(x_0 + \eta x_1) + (1-\lambda)(x_0 + \eta x_2) \in X_0.$$

This point can be rewritten as $x_0 + \eta[\lambda x_1 + (1-\lambda)x_2]$. Since f is convex,

$$f\big(x_0 + \eta[\lambda x_1 + (1-\lambda)x_2]\big) \le \lambda f(x_0 + \eta x_1) + (1-\lambda) f(x_0 + \eta x_2).$$

Remark. For some $x_0 \in X_0$ and $x \in C$, we may have $\varphi(x) = -\infty$.

30. Let $\psi(t)$ be nondecreasing and bounded below for $t > \tau$. Let L denote the greatest lower bound of $\{\psi(t) : t > \tau\}$. Prove that

$$\lim_{t \downarrow \tau} \psi(t) = L.$$

31. Let X denote the set of real-valued, continuous waveforms on $[0,1]$. If

$$f(x) := \int_0^1 x(t)^3 \, dt,$$

find the Gâteaux derivative $(D^+ f)(x, \Delta x)$ for $\Delta x \in X$.

32. Prove Theorem 5.17. You should first verify that the Lagrangian is a convex function of x.

33. Find real numbers v and w that minimize $f(v,w) := v^2 + w^2 + 4v - 2w + 5$ subject to the constraint $v^2 + w^2 \le 1$. *Hint:* Observe that $f(v,w) = (v+2)^2 + (w-1)^2$. Check your work with the MATLAB code

```
f = @(x) (x(1)+2)^2+(x(2)-1)^2;
x0 = [ 0 0 ]';
[xmin,fmin] = fmincon(f,x0,[],[],[],[],[],[],'nonlcon')
```

where `nonlcon.m` is a MATLAB M-file containing

```
function [c,ceq] = nonlcon(x)
c = x(1)^2 + x(2)^2 - 1;
ceq = [];
```

34. Consider the function $f(x,y) := x^2 - \ln x + \frac{1}{2}y^2 - \ln y$ for $x, y > 0$. Solve

$$\min_{x,y>0} f(x,y) \quad \text{subject to} \quad x^2 + y^2 \le c^2$$

under the assumption $0 < c^2 < 3/2$. Repeat for $c^2 \ge 3/2$. *Hint:* You can gain some insight to the problem by separately minimizing $x^2 - \ln x$ for $x > 0$ and $y^2/2 - \ln y$ for $y > 0$.

35. Let g be a bounded function with $|g(t)| > 0$ for $t \in [0,1]$, and let y be a given finite-energy waveform on $[0,1]$. Find the finite-energy waveform x that solves

$$\min_x \int_0^1 \big|y(t) - g(t)x(t)\big|^2 dt \quad \text{subject to} \quad \int_0^1 \big|x(t)\big|^2 dt \le b.$$

Express your solution x in terms of y, g, and a Lagrange multiplier. Also, in the special case $g(t) \equiv 1$, find the Lagrange multiplier and the corresponding solution x in terms of y and b only.

36. Let a and y be given vectors in an inner-product space V. Do not assume V finite dimensional. Solve

$$\min_{x \in \mathbb{C}} \|y - ax\|^2 \quad \text{subject to} \quad |x|^2 \le E,$$

where E is a given energy constraint.

37. Solve

$$\min_x \|x\|^2 \quad \text{subject to} \quad Ax = y,$$

where $A: X \to Y$ is a linear operator between inner-product spaces X and Y such that the adjoint A^* exists and $y \in Y$ is given. Discuss any additional assumptions you would like to make in order to guarantee that a solution exists.

38. Solve

$$\min_{x \in \mathbb{R}^n_+} \sum_{k=1}^n x_k \ln x_k \quad \text{subject to} \quad \sum_{k=1}^n x_k = 1.$$

Remark. A sequence of nonnegative numbers that sums to one is sometimes called a **probability mass function**. In this problem you are finding the probability mass function that minimizes the negative of the **entropy** of the probability mass function.

39. Solve

$$\max_{x \ge 0, y \ge 0} \{2x - x^2 + \tfrac{3}{2}y\} \quad \text{subject to} \quad x + y \le 1.$$

Check your work with the MATLAB code

```
f = @(x)-(2*x(1)-x(1).^2+(3/2)*x(2));
A = [ 1 1 ];
b = 1;
lb = [ 0 0 ]';
x0 = [ 0 0 ]';
[xmin,fmin] = fmincon(f,x0,A,b,[],[],lb)
```

40. Solve

$$\max_{x\geq 0, y\geq 0} \{2x - x^2 + 3y\} \quad \text{subject to} \quad x + y \leq 1.$$

Check your work by modifying the MATLAB code of the previous problem.

41. Consider the problem

$$\min_{x\in X} f(x) \quad \text{subject to} \quad \int_0^1 |x(t)|^2\, dt \leq 1,$$

where X and f are defined in Problem 5.31. Set up the Lagrangian, and find all choices of $\lambda \geq 0$ and $x \in X$ such that $(D_x^+ L)(\lambda, x, \Delta x) = 0$ for all Δx, the constraint is satisfied, and

$$\lambda \left(\int_0^1 |x(t)|^2\, dt - 1 \right) = 0.$$

42. Let X and Y be real inner-product spaces. Let $A: X \to Y$ and $B: X \to \mathbb{R}^m$ be linear operators. Assume that B^* exists and is nonsingular. Assume that A^* exists and that A^*A is invertible. For given $y \in Y$ and $z \in \mathbb{R}^m$, find $x \in X$ to minimize $\|y - Ax\|$ subject to $Bx = z$. *Hint:* It may be helpful to apply Proposition 4.26.

43. Let X and Z be complex inner-product spaces. Let $C: X \to X$ and $B: X \to Z$ be linear operators. Assume B^* exists and that C is positive semidefinite. Given $d \in X$ and $z \in Z$, consider the **quadratic programming problem**

$$\min_{x\in X} \tfrac{1}{2}\langle Cx, x\rangle + \operatorname{Re}\langle x, d\rangle \quad \text{subject to} \quad Bx = z.$$

Remark. This problem contains the previous one as a special case if we take $C = 2A^*A$ and $d = -2A^*y$.

(a) Show that if $x \in X$ and $\mu \in Z$ solve

$$\begin{bmatrix} C & B^* \\ B & 0 \end{bmatrix} \begin{bmatrix} x \\ \mu \end{bmatrix} = \begin{bmatrix} -d \\ z \end{bmatrix},$$

then x solves the above quadratic programming problem.

(b) Suppose that $x^{(k)}$ and $\mu^{(k)}$ solve

$$\begin{bmatrix} C & B^* \\ B & 0 \end{bmatrix} \begin{bmatrix} x^{(k)} \\ \mu^{(k)} \end{bmatrix} = \begin{bmatrix} -d \\ z^{(k)} \end{bmatrix}, \quad k = 1, \ldots, K.$$

Show that if $\sum_{k=1}^K \eta_k = 1$ and if $\widetilde{z} := \sum_{k=1}^K \eta_k z^{(k)}$, then $\widetilde{x} := \sum_{k=1}^K \eta_k x^{(k)}$ solves

$$\min_{x\in X} \tfrac{1}{2}\langle Cx,x\rangle + \mathrm{Re}\langle x,d\rangle \quad \text{subject to} \quad Bx = \widetilde{z}.$$

Remark. The importance of part (b) is that once we have solved the minimization problem K times for $z^{(1)},\ldots,z^{(K)}$, we can obtain the solution for any z in the affine hull of $z^{(1)},\ldots,z^{(K)}$ by taking the appropriate affine combination of the solutions $x^{(1)},\ldots,x^{(K)}$. For example, recall the discussion of **portfolio optimization** in Section 5.3.3 where $Z = \mathbb{R}^2$. Suppose we have found optimal portfolios $x^{(1)}$ and $x^{(2)}$ for

$$z^{(1)} = \begin{bmatrix} r_1 \\ 1 \end{bmatrix} \quad \text{and} \quad z^{(2)} = \begin{bmatrix} r_2 \\ 1 \end{bmatrix}.$$

Then for any real η, if we put $r_\eta := \eta r_1 + (1-\eta) r_2$, the optimal portfolio is $x_\eta := \eta x^{(1)} + (1-\eta)x^{(2)}$. This is the **Two-Fund Theorem**.

(c) Assume that B^* is nonsingular. Assume that whenever both $Bx = 0$ and $\langle Cx,x\rangle = 0$, we must have $x = 0$. Show that the matrix in part (a) is nonsingular.

(d) Let B be as in (5.18), and suppose C is an $n \times n$ matrix of the form

$$C = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix},$$

where D is an $(n-1)\times(n-1)$ positive-definite matrix. Show that whenever $Bx = 0$ and $\langle Cx,x\rangle = 0$, we must have $x = 0$.

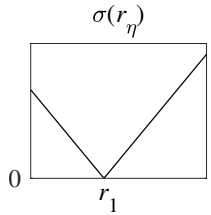
Remark. An important instance of the **Two-Fund Theorem** arises when the covariance matrix C is singular. Using the notation of the previous remark, suppose $x^{(1)} \in \ker C$. Then $Cx^{(1)} = 0$, and the risk $\sigma(r_1) = 0$. More generally, $Cx_\eta = (1-\eta)Cx^{(2)}$, and

$$\begin{aligned}\sigma^2(r_\eta) = \big\langle Cx_\eta, x_\eta \big\rangle = (1-\eta)\big\langle Cx^{(2)}, x_\eta \big\rangle &= (1-\eta)\big\langle x^{(2)}, Cx_\eta \big\rangle \\ &= (1-\eta)^2 \big\langle x^{(2)}, Cx^{(2)} \big\rangle.\end{aligned}$$

From the definition of r_η, we can write $\eta = (r_2 - r_\eta)/(r_2 - r_1)$ or $1-\eta = (r_\eta - r_1)/(r_2 - r_1)$. Hence,

$$\sigma^2(r_\eta) = \left[\frac{r_\eta - r_1}{r_2 - r_1}\right]^2 \big\langle x^{(2)}, Cx^{(2)} \big\rangle.$$

Assuming $r_1 < r_2$, the above formula implies that for $r_\eta \geq r_1$, the risk $\sigma(r_\eta)$ is a straight line with nonnegative slope, while for $r_\eta < r_1$, the risk is a straight line with nonpositive slope.



44. If x is given by (5.21) with z as in (5.18), show that $\langle Cx,x\rangle$ is a convex function of r, if C is positive semidefinite. *Hint:* Example 5.20.

45. **MATLAB.** Consider the convex set

$$C := \left\{ \begin{bmatrix} u \\ v \end{bmatrix} : u \in \mathbb{R} \text{ and } v \geq e^u \right\}.$$

Find the projection of the origin $x := [0,0]^\mathsf{T}$ onto C. *Hint:* Set this up as a Lagrange multiplier problem. You will have to solve a nonlinear scalar equation numerically.

46. In Example 5.22 additionally assume that Q^{-1} and $(Q^{-1})^*$ exist. Consider the problem

$$\min_{z\in X} \|y - AQ^{-1}z\|^2 \quad \text{subject to} \quad \|z\|^2 \leq b.$$

Suppose that you can find z and λ such that $\|z\|^2 \leq b$, $\lambda(\|z\|^2 - b) = 0$, and

$$[\lambda I + (AQ^{-1})^*(AQ^{-1})]z = (AQ^{-1})^* y.$$

Show that $x := Q^{-1}z$ solves

$$\min_{x\in X} \quad \text{subject to} \quad \|Qx\|^2 \leq b.$$

47. Consider the problem

$$\min_{x\in X} \|Qx\|^2 \quad \text{subject to} \quad \|y - Ax\|^2 \leq \varepsilon,$$

where X and Y are complex inner-product spaces, $A: X \to Y$ and $Q: X \to X$ are linear operators, and $y \in Y$ and $\varepsilon \geq 0$ are given. Assume that the adjoints A^* and Q^* exist.

(a) If the Gâteaux derivative of the Lagrangian must be zero in all directions Δx, determine the linear equation that x must satisfy. How does it compare to (5.15)?

(b) If Q is nonsingular and $\varepsilon < \|y\|^2$, show that the Lagrange multiplier cannot be zero.

48. Use the notation from Example 5.24 to formulate the problem of finding the minimum 1-norm solution of $Ax = b$ as a **quadratic programming problem**.

49. Let $\|\cdot\|$ denote the standard Euclidean norm on $\mathbb{R}^n$. For $y = [y_1, \ldots, y_n]^\mathsf{T} \in \mathbb{R}^n$, the 1-norm is defined by $\|y\|_1 := |y_1| + \cdots + |y_n|$. Given a positive number t and a vector $x = [x_1, \ldots, x_n]^\mathsf{T} \in \mathbb{R}^n$, show that the solution of

$$\operatorname*{argmin}_{y \in \mathbb{R}^n} \left(t\|y\|_1 + \tfrac{1}{2}\|x - y\|^2 \right)$$

is given by $[\eta_t(x_1), \ldots, \eta_t(x_n)]^\mathsf{T}$, where η_t is the **shrinkage operator** defined in Problem 3.25.

Remark. Using the notation in the remark of Problem 5.9, we can write

$$[\eta_t(x_1), \ldots, \eta_t(x_n)]^\mathsf{T} = (\operatorname{prox}(tf))(x),$$

where $f(x) = \|x\|_1$.

CHAPTER 6
Sequences, Limits, Completeness, and Compactness

To motivate the material in this chapter, we consider some questions raised by our earlier work.

Recall from our proof of the Projection Theorem for Finite-Dimensional Subspaces that if $w_1, \ldots, w_n$ are orthonormal vectors in an inner-product space X, then the projection of any $x \in X$ onto $W_n := \text{span}\{w_1, \ldots, w_n\}$ is given by

$$\sum_{k=1}^{n} \langle x, w_k \rangle w_k.$$

If we now have an infinite sequence of orthonormal vectors, $w_1, w_2, \ldots$, it is natural to expect that the projection of x onto $W := \text{span}\{w_1, w_2, \ldots\}$ is given by

$$\sum_{k=1}^{\infty} \langle x, w_k \rangle w_k := \lim_{n \to \infty} \sum_{k=1}^{n} \langle x, w_k \rangle w_k.$$

In other words, we define the infinite sum to be the limit of the finite partial sums. How do we know the limit exists? Even if the limit exists, how do we know it lies in W?

Consider the problem of minimizing or maximizing a real-valued function f defined on a subset X_0 of an arbitrary set X. How do we know that a maximum or minimum exists? For example, the function $f(x) := x$ for $x \in [0,1)$ has a unique minimizer at $x = 0$, but there is no maximum value of f on $[0,1)$. Or suppose we have an algorithm that generates a sequence $x_n \in X_0$ with the property that $f(x_n) \leq f(x_{n+1})$. Even if x_n converges to some limit x_0, it may happen that $x_0 \notin X_0$.

The purpose of this chapter is to provide tools for understanding and addressing the foregoing questions. Key concepts will be those of Cauchy sequence, closure of a set, sequentially compact set, and continuous function.

6.1. The Real Numbers

The Least Upper Bound and the Greatest Lower Bound

Sometimes a set of numbers has a largest or maximum element, and sometimes it does not. For example, the largest element of $[0,1]$ is 1, but $[0,1)$ does not have a

largest element. For this reason, we introduce the concept of least upper bound as the next best thing to the largest element of a set in case the set does not have a largest element.

We say that a real number r is an **upper bound** of a set B of real numbers if for all $b \in B$, $b \leq r$. We say that a real number $\underline{r}$ is the ***least*** **upper bound** of B if $\underline{r}$ is an upper bound of B, and if every number smaller than $\underline{r}$ is not an upper bound of B.

Since 1 is an upper bound of $[0,1)$, and since no number smaller than 1 is an upper bound of $[0,1)$, we see that 1 is the least upper bound of $[0,1)$. Similarly, 1 is the least upper bound of $[0,1]$.

Least Upper Bound Axiom **[35].** Every nonempty subset of real numbers that is bounded above has a least upper bound.

If B is a set of real numbers, we define the **supremum** of B as follows. If B is nonempty and bounded above, we define $\sup B$ to be the least upper bound of B. If B is nonempty and not bounded above, we write $\sup B = \infty$. If B is empty, we write $\sup B = -\infty$. (Conventions concerning $\pm\infty$ are discussed in the notes at the end of the chapter.)

A nonempty set B that is bounded below has a **greatest lower bound**. We define the **infimum** of B as follows. If B is nonempty and bounded below, we define $\inf B$ to be the greatest lower bound of B. If B is nonempty and not bounded below, we write $\inf B = -\infty$, and if B is empty, we write $\inf B = \infty$.

Limits

A sequence of real numbers x_n is said to **converge** to a real number x if given any $\varepsilon > 0$, we have for sufficiently large n that $|x_n - x| < \varepsilon$. In this case, we write $\lim_{n\to\infty} x_n = x$ or we write $x_n \to x$.

Example 6.1. Show that $1/2^n \to 0$.

Solution. Since $2^n \geq n$, we have $1/2^n < 1/n$. Hence, given $\varepsilon > 0$, we have for $n > 1/\varepsilon$ that $|1/2^n| = 1/2^n < \varepsilon$.

Example 6.2. If $x_n \to x$ and $y_n \to y$, prove that $x_n + y_n \to x + y$.

Solution. We need to show that for large n,

$$|(x_n + y_n) - (x + y)| < \varepsilon.$$

Before we can start the proof itself, we need to do some analysis to see how to approach it. We begin by using the triangle inequality to write

$$|(x_n+y_n)-(x+y)| = |(x_n-x)+(y_n-y)| \le |x_n-x|+|y_n-y|. \tag{6.1}$$

If we can show that $|x_n-x|<\varepsilon/2$ and $|y_n-y|<\varepsilon/2$, then we will have the inequality we need.

Proof. Let $\varepsilon>0$ be given. Since $x_n\to x$, we know that there is an N_1 such that for all $n\ge N_1$, $|x_n-x|<\varepsilon/2$. Similarly, since $y_n\to y$, there is an N_2 such that for all $n\ge N_2$, $|y_n-y|<\varepsilon/2$. Hence, for $n\ge N=\max(N_1,N_2)$, we can upper bound (6.1) by ε as required. □

Lemma 6.3. *Let B be a nonempty set of real numbers with least upper bound $\underline{r}$. Then there is a sequence $b_n\in B$ with b_n converging to $\underline{r}$.*

Proof. To say that $\underline{r}$ is the least upper bound of B implies that for every $n=1,2,3,\ldots$, the number $\underline{r}-1/n$ is not an upper bound of B. This means that there is some $b_n\in B$ with $\underline{r}-1/n<b_n\le\underline{r}$. Hence, $|b_n-\underline{r}|<1/n$. It is now clear that given any $\varepsilon>0$, we have for all $n>1/\varepsilon$ that $|b_n-\underline{r}|<\varepsilon$. □

Sequential Compactness

If $x_1,x_2,x_3,\ldots$ is a sequence of real numbers, and if $n_1,n_2,\ldots$ is a sequence of integers with the property that $n_k\to\infty$ in the sense that given any N, there is a K such that for all $k\ge K$, $n_k\ge N$, then we say that x_{n_k} is a **subsequence** of x_n.

Example 6.4. Consider a sequence $x_1,x_2,x_3,\ldots$ that starts out as π, e, 5, -2, 12, -8, ... We can summarize this in the table

m	1	2	3	4	5	6	⋯
x_m	π	e	5	-2	12	-8	⋯

Similarly, suppose $n_1,n_2,n_3,\ldots$ is a sequence of positive integers that starts out $3,1,6,2,2,4,\ldots$. This corresponds to the table

k	1	2	3	4	5	6	⋯
n_k	3	1	6	2	2	4	⋯

We now determine the subsequence $y_k:=x_{n_k}$ as follows. Since $n_1=3$, $y_1=x_{n_1}=x_3=5$. Since $n_2=1$, $y_2=x_{n_2}=x_1=\pi$. Since $n_3=6$, $y_3=x_{n_3}=x_6=-8$. Since

$n_4 = 2$, $y_4 = x_{n_4} = x_2 = e$. Since $n_5 = 2$, $y_5 = x_{n_5} = x_2 = e$. Since $n_6 = 4$, $y_6 = x_{n_6} = x_4 = -2$. Hence, the first part of the y_k sequence corresponds to the table

k	1	2	3	4	5	6	$\cdots$
y_k	5	π	-8	e	e	-2	$\cdots$

This example does not illustrate the requirement that $n_k \to \infty$.

Sequence Notation. When we write, "Consider a sequence x_n," it is important to understand that the subscript n is a dummy variable. We could just as well have written, "Consider a sequence x_m." In other words, in these two statements, n and m are "dummy variables." Strictly speaking, x is the name of a list of real numbers, and an expression such as x_5 refers to the 5th entry in the list named x.

When we write, "Consider a sequence x_n having subsequence x_{n_k}," the n in x_n and the k in x_{n_k} are dummy variables, but the n in both n_k and x_{n_k} is the name of a list of positive integers. So we could have written, "Consider a sequence x_m having subsequence x_{n_ℓ}."

A subset of real numbers is said to be **sequentially compact** if every sequence in the set has a converging subsequence whose limit lies in the subset. The following result implies that a closed interval $[a,b]$ is sequentially compact.

Theorem 6.5 (Bolzano–Weierstrass). *Every bounded sequence of real numbers has a subsequence that converges to a finite real number.*

Proof. Let x_n be a bounded sequence of real numbers. Then there are bounds $-\infty < a < b < \infty$ such that $a \le x_n \le b$ for all n. Put

$$y_m := \sup_{n \ge m} x_n.$$

This is shorthand for $y_m := \sup A_m$, where $A_m := \{x_n : n \ge m\}$. Since $x_n \le b$ for all n, b is an upper bound on A_m. Since y_m is the *least* upper bound of A_m, $y_m \le b$. Also note that $y_m \ge x_m \ge a$. Hence, $a \le y_m \le b$ for all $m = 1, 2, \ldots$.

Now put

$$z := \inf_{m \ge 1} y_m.$$

Since a is a lower bound on y_m, a must be less than or equal to the *greatest* lower bound of the y_m; i.e., $a \le z$. Combining this with $z \le y_1 \le b$ shows that $z \in [a,b]$.

We now show that there is a subsequence $x_{n_k} \to z$. For $k = 1, 2, \ldots$, we proceed as follows. Since z is a greatest lower bound, $z + 1/k$ is not a lower bound of the y_m. Hence, for some y_{m_k}, we have

$$z \le y_{m_k} < z + 1/k. \tag{6.2}$$

Note that since the y_m are nonincreasing, i.e., $y_{m+1} \le y_m$, we may assume $m_k \ge k$. Next, since

$$y_{m_k} = \sup_{n \ge m_k} x_n,$$

$y_{m_k} - 1/k$ is not an upper bound on $\{x_n : n \ge m_k\}$. Hence, for some $n_k \ge m_k$,

$$y_{m_k} - 1/k < x_{n_k} \le y_{m_k}. \tag{6.3}$$

It follows that

$$-1/k < x_{n_k} - y_{m_k}.$$

Now, from the left-hand inequality in (6.2), $y_{m_k} - z \ge 0$, and so

$$\begin{aligned} -1/k &< (x_{n_k} - y_{m_k}) + (y_{m_k} - z) \\ &= x_{n_k} - z \\ &\le y_{m_k} - z, \quad \text{by (6.3)}, \\ &< 1/k, \qquad \text{by (6.2)}. \end{aligned}$$

Hence, $-1/k < x_{n_k} - z < 1/k$, or $|x_{n_k} - z| < 1/k$. □

The number z in the preceding proof is called the **limit superior** of the sequence x_n; i.e.,

$$\limsup_n x_n := \inf_m \Big(\sup_{n \ge m} x_n \Big).$$

Similarly, the **limit inferior** is defined by

$$\liminf_n x_n := \sup_m \Big(\inf_{n \ge m} x_n \Big).$$

The notation $\overline{\lim}_n x_n$ and $\underline{\lim}_n x_n$ is also used for $\limsup_n x_n$ and $\liminf_n x_n$, respectively.

Remark. Observe that for each $m = 1, 2, \ldots$,

$$\inf_{n \ge m} x_n \le \sup_{n \ge m} x_n.$$

Since the left-hand side is a nondecreasing sequence in m, and the right-hand side is a nonincreasing sequence in m, it is easy to prove that both sides have limits in m.

The limit in m on the left is $\underline{\lim}_n x_n$, and the limit in m on the right is $\overline{\lim}_n x_n$. It then follows that $\underline{\lim}_n x_n \le \overline{\lim}_n x_n$. Now suppose that one has a sequence x_n for which one can show that $\underline{\lim}_n x_n \ge \overline{\lim}_n x_n$. Then $\underline{\lim}_n x_n = \overline{\lim}_n x_n$, and it can be proved that x_n converges to this common value. Conversely, if x_n converges, it can be proved that $\underline{\lim}_n x_n = \overline{\lim}_n x_n$ is the value of the limit.

6.2. Normed Vector Spaces and Metric Spaces

Normed Vector Spaces

To generalize properties of absolute value from numbers to vectors, we introduce the concept of a norm.

Given a real or complex vector space X, we say that $\|\cdot\|$ is a **norm** on X if the following three properties hold.

(i) For all $x \in X$, $0 \le \|x\| < \infty$, with $\|x\| = 0$ if and only if x is the zero vector.
(ii) For all $x \in X$ and all scalars a, $\|ax\| = |a|\,\|x\|$.
(iii) For all $x, y \in X$, $\|x+y\| \le \|x\| + \|y\|$. This is known as the **triangle inequality**.

We sometimes call $\|x\|$ the **length** of x. If x has length one, it is called a **unit vector**. A vector space on which a norm is defined is called a **normed vector space**.

The first thing to note is that if X is an inner-product space and if $\|x\| := \langle x, x\rangle^{1/2}$ as in Chapter 3, then $\|\cdot\|$ satisfies the foregoing three properties; the first two properties are obvious, and the third was established by Corollary 3.2.

Proposition 6.6. *A norm satisfies the inequality*

$$\big|\|x\| - \|y\|\big| \le \|x - y\|.$$

Proof. By the triangle inequality, we can write

$$\|x\| = \|(x-y)+y\| \le \|x-y\| + \|y\|,$$

from which it follows that $\|x\| - \|y\| \le \|x-y\|$. If we had started with $\|y\|$ instead of $\|x\|$, we would have ended up with $\|y\| - \|x\| \le \|y-x\|$. Since the larger of $\|x\| - \|y\|$ and $\|y\| - \|x\|$ is the absolute value of their difference, the proposition follows. □

If $\|\cdot\|$ is a norm on a vector space X, then we can define related norms on X^d several ways. Here are three ways. For $x = [x_1, \ldots, x_d]^\mathsf{T}$, put[a]

$$\|x\|_1 := \sum_{i=1}^{d} \|x_i\|, \quad \|x\|_2 := \sqrt{\sum_{i=1}^{d} \|x_i\|^2}, \quad \text{and} \quad \|x\|_\infty := \max_{1\le i\le d} \|x_i\|.$$

In the formula

$$\|x\|_\infty \le \|x\|_2 \le \|x\|_1 \le d\,\|x\|_\infty, \tag{6.4}$$

the first and third inequalities are obvious. For the middle inequality, write

$$\begin{aligned}\|x\|_1^2 &= (\|x_1\| + \cdots + \|x_d\|)^2 \\ &= \sum_{i=1}^{d} \|x_i\|^2 + \text{nonnegative cross terms} \\ &\ge \sum_{i=1}^{d} \|x_i\|^2 = \|x\|_2^2.\end{aligned}$$

Now take square roots. It is left to the problems to show that $\|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|_\infty$ satisfy the properties of a norm.

Example 6.7. Suppose $X = \mathbb{R}$ equipped with the absolute-value norm. Then we immediately get three norms on $\mathbb{R}^d$. When $d = 2$, it is interesting to sketch $\{x : \|x\| = 1\}$ under each norm. The "2-norm" is the usual Euclidean norm, which yields a circle of radius one. The reader should verify that the "1-norm" yields a diamond, and the "infinity norm" yields a square. When $d = 3$, the 2-norm yields a sphere, the infinity norm yields a box, and the 1-norm yields the surface of an **octahedron** as shown in Figure 6.1.

We can identify $\mathbb{C}$ under the usual absolute-value for complex numbers with $\mathbb{R}^2$ under the 2-norm. Hence, we immediately get the 1-norm, the 2-norm, and the infinity norm on $\mathbb{C}^d$.

Example 6.8. Not all norms come from an inner product. Recall that if a norm comes from an inner product, then the norm must satisfy the **parallelogram law** (Problem 3.3),

$$\|x+y\|^2 + \|x-y\|^2 = 2(\|x\|^2 + \|y\|^2).$$

However, the reader can check that with $x = [1,1]^\mathsf{T}$ and $y = [1,-1]^\mathsf{T}$, the parallelogram law does not hold for the infinity norm.

---

[a] We caution the reader that the norm symbol with subscripts is defined in different ways in other situations in the text; e.g., in the next subsection.

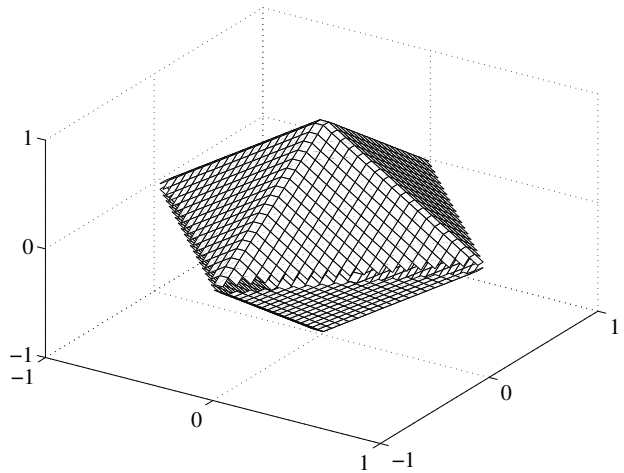


Figure 6.1. The unit "sphere" in $\mathbb{R}^3$ under the 1-norm.

6.2.1. The L^p Spaces

In Example 2.11 we introduced the L^p spaces ($1 \le p < \infty$) as the set of waveforms x on some fixed interval such that $\int |x(t)|^p \, dt < \infty$. If $x \in L^p$ and we put[b]

$$\|x\|_p := \left(\int |x(t)|^p \, dt \right)^{1/p},$$

then $\|x\|_p$ is finite on L^p. It can be shown that $\|\cdot\|_p$ satisfies the properties of a norm on L^p. In the L^p spaces, the triangle inequality is called the **Minkowski inequality** (see Problem 6.13).

An important tool in the study of L^p spaces is the **Hölder inequality** (see Problem 6.12). It says that if $x \in L^p$ and $y \in L^q$ and $\frac{1}{p} + \frac{1}{q} = 1$ with $1 < p < \infty$, then

$$\int |x(t)y(t)| \, dt \le \|x\|_p \, \|y\|_q.$$

In particular, this implies that the product $xy \in L^1$.

Example 6.9. Show that $L^p[a,b] \subset L^1[a,b]$ if $p > 1$.

[b] Do not confuse the meaning of $\|\cdot\|_1$ and $\|\cdot\|_2$ in the context of waveforms with that in the context of vectors in $\mathbb{R}^d$ or $\mathbb{C}^d$.

Solution. Suppose $x \in L^p[a,b]$. We must show that $\int_a^b |x(t)|\,dt < \infty$. Observe that $y(t) \equiv 1$ satisfies $\int_a^b |y(t)|^q\,dt = \int_a^b 1\,dt = b-a$. Hence, $\|y\|_q = (b-a)^{1/q}$. By Hölder's inequality,

$$\int_a^b |x(t)|\,dt = \int_a^b |x(t)\cdot 1|\,dt \le \|x\|_p (b-a)^{1/q} < \infty.$$

In particular, if $x \in L^2[a,b]$, then $x \in L^1[a,b]$.

Example 6.10. Show that $L^p(\mathbb{R}) \not\subset L^1(\mathbb{R})$ if $p > 1$.

Solution. Consider the waveform $x(t) := 1/t$ for $t \ge 1$ and $x(t) := 0$ for $t < 1$. Then

$$\int_{-\infty}^{\infty} |x(t)|^p\,dt = \int_1^{\infty} t^{-p}\,dt = \frac{1}{p-1} < \infty.$$

However,

$$\int_{-\infty}^{\infty} |x(t)|\,dt = \int_1^{\infty} t^{-1}\,dt = \ln t\Big|_1^{\infty} = \infty.$$

Example 6.11. Show that $L^1(\mathbb{R}) \not\subset L^p(\mathbb{R})$ if $p > 1$.

Solution. Consider the waveform $x(t) = t^{-1/p}$ for $0 < t \le 1$ and $x(t) = 0$ otherwise. Then

$$\int_{-\infty}^{\infty} |x(t)|\,dt = \int_0^1 t^{-1/p}\,dt = \frac{t^{1-1/p}}{1-1/p}\Big|_0^1 = \frac{p}{p-1}.$$

However,

$$\int_{-\infty}^{\infty} |x(t)|^p\,dt = \int_0^1 t^{-1}\,dt = \ln t\Big|_0^1 = 0 - (-\infty) = \infty.$$

Example 6.12 (L^p Spaces of Random Variables). The collection of random variables with finite absolute pth moment is also sometimes denoted by L^p. In other words, a random variable X is in L^p if $\mathsf{E}[|X|^p] < \infty$. Recall that the **moment generating function** of a random variable X is the function $M(s) := \mathsf{E}[e^{sX}]$. For real s, we can use the Hölder inequality to show that $\psi(s) := \ln M(s)$ is a convex function of s.[c] Fix any real s and t. Then for $0 < \lambda < 1$,

$$\psi\big(\lambda s + (1-\lambda)t\big) = \ln \mathsf{E}[e^{(\lambda s+(1-\lambda)t)X}] = \ln \mathsf{E}\big[\underbrace{(e^{sX})^{\lambda}}_{=:U}\,\underbrace{(e^{tX})^{1-\lambda}}_{=:V}\big].$$

[c] The function $\psi(s) := \ln M(s)$ is known as the **cumulant generating function** of X.

Now by the Hölder inequality for nonnegative U and V,

$$\mathsf{E}[UV] \le \big(\mathsf{E}[U^p]\big)^{1/p}\big(\mathsf{E}[V^p]\big)^{1/q}.$$

Taking $p := 1/\lambda$ and $q := 1/(1-\lambda)$, observe that

$$U^p = \big[(e^{sX})^\lambda\big]^p = e^{sX} \quad \text{and} \quad V^q = \big[(e^{tX})^{1-\lambda}\big]^q = e^{tX}.$$

Hence,

$$\mathsf{E}[UV] \le \big(\mathsf{E}[e^{sX}]\big)^{1/p}\big(\mathsf{E}[e^{tX}]\big)^{1/q},$$

and we have

$$\begin{aligned}\ln \mathsf{E}[UV] &\le \frac{1}{p}\ln \mathsf{E}[e^{sX}] + \frac{1}{q}\ln \mathsf{E}[e^{tX}] \\ &= \lambda\,\psi(s) + (1-\lambda)\,\psi(t).\end{aligned}$$

To conclude, simply note that $\ln \mathsf{E}[UV] = \psi(\lambda s + (1-\lambda t))$.

The foregoing example is important in studying the **Chernoff bound**. For $s > 0$,

$$\begin{aligned}\mathsf{P}(X \ge x) = \mathsf{P}(sX \ge sx) = \mathsf{P}(e^{sX} \ge e^{sx}) &\le \frac{\mathsf{E}[e^{sX}]}{e^{sx}}, \qquad \text{by Markov's inequality,} \\ &= e^{-sx}M(s) \\ &= \exp[-sx + \ln M(s)].\end{aligned}$$

To minimize this last exponential, it suffices to minimize $-sx + \ln M(s)$ as a function of s. This is a convex function of s and therefore has a unique minimum value. The optimal value of s can be found by taking the derivative, which is

$$-x + \frac{M'(s)}{M(s)},$$

and finding its zeros; e.g., by using the MATLAB function `fzero`.

6.2.2. Metric Spaces

In a normed vector space, the distance between two vectors x and y is taken as $\|x-y\|$. We have already seen three different norms, and therefore different notions of distance, on $\mathbb{R}^d$. To generalize ideas of distance from vector spaces to more abstract sets, we introduce the concept of a metric.

Let X be a nonempty set, and let $\rho : X \times X \to [0,\infty)$ have the following properties for any points x, y, and z in X:

(*i*) $\rho(x,y) \geq 0$, and $\rho(x,y) = 0$ if and only if $x = y$.
(*ii*) $\rho(x,y) = \rho(y,x)$.
(*iii*) $\rho(x,z) \leq \rho(x,y) + \rho(y,z)$. This is known as the **triangle inequality**.

The function ρ is called a metric, and the pair (X,ρ) is called a **metric space**. When ρ is understood, we just say that X is a metric space.

The first thing to note is that if X is a normed vector space, then $\rho(x,y) := \|x-y\|$ satisfies the three properties of a metric. The first two properties are obvious. To verify the third one, we simply use the triangle inequality for norms to write

$$\rho(x,z) = \|x-z\| = \|(x-y)+(y-z)\| \leq \|x-y\| + \|y-z\| = \rho(x,y) + \rho(y,z).$$

Since absolute value on $\mathbb{R}$ is obviously a norm, $|x-y|$ defines a metric on $\mathbb{R}$.

Example 6.13 (Discrete Metric). Given any nonempty set X, take $\rho(x,y) := 1$ for $x \neq y$ and $\rho(x,y) := 0$ for $x = y$. Then it is easy to see that ρ satisfies the three properties of a metric. This metric is called the **discrete metric**.

Example 6.14. If (X,ρ) is a metric space and d is a positive integer, we can make X^d into a metric space in several ways. For $x := [x_1,\ldots,x_d]^\mathsf{T}$ and $y := [y_1,\ldots,y_d]^\mathsf{T}$, put

$$\rho_1(x,y) := \sum_{i=1}^{d} \rho(x_i,y_i), \quad \rho_2(x,y) := \sqrt{\sum_{i=1}^{d} \rho(x_i,y_i)^2},$$

and

$$\rho_\infty(x,y) := \max_{1\leq i\leq d} \rho(x_i,y_i).$$

It is left to the problems to show that ρ_1, ρ_2, and ρ_∞ satisfy the properties of a metric. We also note that

$$\rho_\infty(x,y) \leq \rho_2(x,y) \leq \rho_1(x,y) \leq d\rho_\infty(x,y). \tag{6.5}$$

Example 6.15. The foregoing example can be generalized. Suppose $(X_1,\rho^{(1)}),\ldots$ are metric spaces. Then on $X_1 \times \cdots \times X_d$ we can define the metrics

$$\rho_\infty(x,y) := \max_{1\leq i\leq d} \rho^{(i)}(x_i,y_i), \quad \rho_1(x,y) := \sum_{i=1}^{d} \rho^{(i)}(x_i,y_i),$$

and

$$\rho_2(x,y) := \sqrt{\sum_{i=1}^{d} \rho^{(i)}(x_i,y_i)^2}.$$

It is easy to check that (6.5) still holds.

6.3. Open Sets and Closed Sets

The set $B(x,r) := \{y \in X : \rho(x,y) < r\}$ is called the **ball of radius r centered at x**. Note that $y \in B(x,r)$ if and only if $\rho(x,y) < r$.

Example 6.16 (Dependence of Ball on Definition of X). We emphasize that the structure of $B(x,\varepsilon)$ depends on how the whole space X is defined. Using the absolute value metric on $X = \mathbb{R}$, we see that

$$B(x,\varepsilon) = \{y \in \mathbb{R} : |x-y| < \varepsilon\} = (x-\varepsilon, x+\varepsilon).$$

In particular, $B(0,\varepsilon) = (-\varepsilon,\varepsilon)$. However, if $X = [0,\infty)$, then

$$B(0,\varepsilon) = \{y \in X : |y| < \varepsilon\} = \{y \geq 0 : |y| < \varepsilon\} = [0,\varepsilon).$$

A set $U \subset X$ is said to be **open** if for every $x \in U$, there is an $\varepsilon > 0$ with $B(x,\varepsilon) \subset U$. To say that U is **not open** means that there exists an $x \in U$ such that for all $\varepsilon > 0$, $B(x,\varepsilon) \not\subset U$.

Proposition 6.17. *The whole space X and the empty set $\varnothing$ are both open.*

Proof. To see that X is open, observe that for any $\varepsilon > 0$, $B(x,\varepsilon)$ is by definition a subset of X.

To prove that the empty set is open, we give a proof by contradiction. That is, we assume $\varnothing$ is not open and then derive a contradiction.

Suppose that $\varnothing$ is not open. Then there exists an $x \in \varnothing$ such that for every $\varepsilon > 0$, $B(x,\varepsilon) \not\subset \varnothing$. The statement $x \in \varnothing$ contradicts the fact that for the empty set, $x \notin \varnothing$ for all $x \in X$. □

The next proposition shows that $B(x,r)$ is an open set. Hence, $B(x,r)$ is often called the ***open* ball of radius r centered at x**.

Proposition 6.18. *The set $B(x,r)$ is open.*

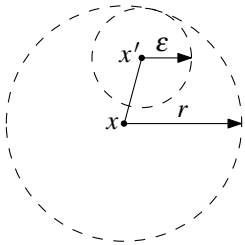


Figure 6.2. Diagram for showing $B(x,r)$ is open.

Proof. If $r \leq 0$, then $B(x,r) = \varnothing$, which is open by Proposition 6.17. So assume $r > 0$. Fix any $x' \in B(x,r)$. We must show that there is an $\varepsilon > 0$ with $B(x',\varepsilon) \subset B(x,r)$. Consider the diagram in Figure 6.2, which suggests that the distance from x to x' plus ε should be at most r. We claim that with $\varepsilon := r - \rho(x,x')$,[d] $B(x',\varepsilon) \subset B(x,r)$. To prove the claim, we fix an arbitrary $y \in B(x',\varepsilon)$, and we show that $y \in B(x,r)$. Fix any $y \in B(x',\varepsilon)$. Then $\rho(x',y) < \varepsilon$, and we can write

$$\begin{aligned}\rho(x,y) &\leq \rho(x,x') + \rho(x',y)\\ &< \rho(x,x') + \varepsilon\\ &= \rho(x,x') + r - \rho(x,x')\\ &= r.\end{aligned}$$

Hence, $y \in B(x,r)$. □

Definition 6.19 (Closed Set). A set $F \subset X$ is **closed** if its **complement**, $F^{\mathrm{c}} := \{x \in X : x \notin F\}$, is open.

It follows that $\varnothing$, X, and $B(x,r)^{\mathrm{c}} = \{y \in X : \rho(x,y) \geq r\}$ are all closed sets. In Problem 6.18 you will show that $\{y \in X : \rho(x,y) \leq r\}$ is a closed set for all $r \geq 0$. It follows that the singleton set $\{x_0\} = \{y \in X : \rho(x_0,y) \leq 0\}$ is closed.

Example 6.20 (Dependence of Complement on Definition of X). We emphasize that just as $B(x,\varepsilon)$ depends on how the whole space X is defined, the complement of a set also depends on how X is defined. If $X = \mathbb{R}$ and $E = (0,1]$, then $E^{\mathrm{c}} = (-\infty,0] \cup (1,\infty)$. However, if $X = (0,\infty)$, then $E^{\mathrm{c}} = (1,\infty)$. In the second case, E^{c} is open, but in the first case it is neither open nor closed (using the absolute value metric in both cases).

[d] Note that $x' \in B(x,r)$ implies that $r - \rho(x,x') > 0$.

Proposition 6.21. *Every union of open sets is an open set.*

Proof. Given a collection of open sets U_α, we must show that $\bigcup_\alpha U_\alpha$ is also open. To do this, we must show that for every $x \in \bigcup_\alpha U_\alpha$, there is a positive ε such that $B(x,\varepsilon) \subset \bigcup_\alpha U_\alpha$. So, fix any $x \in \bigcup_\alpha U_\alpha$. Then x must belong to at least one of the U_α, say $x \in U_{\alpha'}$. Since $U_{\alpha'}$ is open, there is some positive ε with $B(x,\varepsilon) \subset U_{\alpha'}$. Hence,

$$\begin{aligned} B(x,\varepsilon) &\subset U_{\alpha'} \\ &\subset U_{\alpha'} \cup \left(\bigcup_{\alpha \neq \alpha'} U_\alpha \right) \\ &= \bigcup_\alpha U_\alpha. \end{aligned}$$

□

Proposition 6.22. *If U_1 and U_2 are open sets, then so is $U_1 \cap U_2$.*

Proof. Fix any $x \in U_1 \cap U_2$. Since U_1 and U_2 are both open, there exist $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ such that $B(x,\varepsilon_1) \subset U_1$ and $B(x,\varepsilon_2) \subset U_2$. Take $\varepsilon := \min(\varepsilon_1, \varepsilon_2)$. We claim that $B(x,\varepsilon) \subset U_1 \cap U_2$. Since $\varepsilon \leq \varepsilon_1$,

$$B(x,\varepsilon) \subset B(x,\varepsilon_1) \subset U_1,$$

and since $\varepsilon \leq \varepsilon_2$,

$$B(x,\varepsilon) \subset B(x,\varepsilon_2) \subset U_2.$$

Since $B(x,\varepsilon)$ is contained in both U_1 and U_2, we have $B(x,\varepsilon) \subset U_1 \cap U_2$. □

Remark. Let $\mathscr{T}$ be a collection of subsets of an arbitrary set X. We say that $\mathscr{T}$ is a **topology** if the following three properties hold:

(i) X and $\varnothing$ belong to $\mathscr{T}$.
(ii) Every union of sets in $\mathscr{T}$ is a set in $\mathscr{T}$.
(iii) Every intersection of two sets in $\mathscr{T}$ is a set in $\mathscr{T}$.

On account of the foregoing propositions, it is clear that the collection of open sets defined in terms of a metric is a topology. For example, in Problem 6.16 you will show that under the discrete metric, every set is open; i.e., the topology consists of all subsets of X. This is the largest possible topology, and is known as the **discrete topology**. The smallest possible topology, called the **trival topology**, is $\mathscr{T} = \{\varnothing, X\}$, which contains only two sets. If X contains more than one point, then this topology cannot come from a metric. To see why, recall that in a metric space, singleton sets are closed, and so their complements are open. Given $x_0 \in X$, the set $\{x_0\}^c$ would have to be open, but is not one of the sets in the trival topology $\mathscr{T} = \{\varnothing, X\}$.

6.4. Closure, Boundary, and Interior

The **closure**[e] of a set E is

$$\overline{E} := \bigcap_{\substack{C: E \subset C \text{ and} \\ C \text{ is closed}}} C.$$

There are several observations to make about closures.

(i) The closure $\overline{E}$ is a closed set. By taking the complement of the definition, we see that $\overline{E}^{\mathrm{c}}$ is the union of open sets and therefore open.

(ii) We always have $E \subset \overline{E}$. Since $E \subset C$ for every set C in the above intersection, we have $E \subset \cap C =: \overline{E}$.

(iii) If F is any closed set containing E, then $\overline{E} \subset F$. Since $\overline{E}$ is a closed set, we summarize this observation in words by saying that "the closure of E the smallest closed set containing E." The result follows by applying the relation $A \cap F \subset F$ to

$$\overline{E} := \bigcap_{\substack{C: E \subset C \text{ and} \\ C \text{ is closed}}} = \left(\bigcap_{\substack{C \neq F: E \subset C \text{ and} \\ C \text{ is closed}}} \right) \cap F \subset F.$$

(iv) A set E is closed $\Leftrightarrow E = \overline{E}$. By our first observation, we see that $\Leftarrow$ holds. Conversely, suppose E is closed. Then we may take $F = E$ in the preceding observation to obtain $\overline{E} \subset E$. Combining this with the second observation yields $E = \overline{E}$.

Although the definition of closure is a bit abstract, it allows us to define the boundary and interior of a set in a way that makes them easy to work with.

The **boundary** of E is $\partial E := \overline{E} \cap \overline{E^{\mathrm{c}}}$. Notice that since the boundary is the intersection of two closed sets, the boundary is a closed set. Also, the boundary of a set is the same as the boundary of its complement; i.e., $\partial E = \partial(E^{\mathrm{c}})$.

The **interior** of E is $E^{\mathrm{o}} := (\overline{E^{\mathrm{c}}})^{\mathrm{c}}$. A point x is said to be an **interior point** of E if $x \in E^{\mathrm{o}}$. There are several observations to make about the interior.

(i) The interior E^{o} is an open set. This is immediate since it is defined as the complement of a closed set.

(ii) We always have $E^{\mathrm{o}} \subset E$. Since $(E^{\mathrm{o}})^{\mathrm{c}} = \overline{E^c} \supset E^c$, it follows that $E^{\mathrm{o}} \subset E$.

(iii) If U is an open set contained in E, then $U \subset E^{\mathrm{o}}$. Since $E \supset E^{\mathrm{o}}$ is open, we summarize this observation in words by saying that "the interior of E is the largest open set contained in E." The proof of the observation is left to Problem 6.21.

[e] The kind of closure here is sometimes called ***topological*** **closure** to distinguish it from the notions closure under linear combinations or scalar multiplication used earlier in defining subspaces, affine sets, and convex sets.

(*iv*) A set E is open $\Leftrightarrow E = E^{\mathrm{o}}$. By our first observation, we see that $\Leftarrow$ holds. Conversely, suppose E is open. Then E^{c} is closed, which implies $E^{\mathrm{c}} = \overline{E^{\mathrm{c}}}$, which implies $E = \left(\overline{E^{\mathrm{c}}}\right)^{\mathrm{c}} =: E^{\mathrm{o}}$.

From our observations about the closure and the interior of a set, it follows that a set is "sandwiched" between its interior and its closure; i.e.,

$$E^{\mathrm{o}} \subset E \subset \overline{E}.$$

We next show that the closure of a set is the disjoint union of its interior and its boundary. Since $E^{\mathrm{o}} \subset \overline{E}$,

$$\begin{aligned}\overline{E} &= E^{\mathrm{o}} \cup \overline{E} \backslash E^{\mathrm{o}}, \quad \text{which is a disjoint union,} \\ &= E^{\mathrm{o}} \cup \left(\overline{E} \cap \overline{E^{\mathrm{c}}}\right) \\ &= E^{\mathrm{o}} \cup \partial E.\end{aligned}$$

Since this holds for every set E, we can replace E by E^{c} and find that

$$\overline{E^{\mathrm{c}}} = (E^{\mathrm{c}})^{\mathrm{o}} \cup \partial E^{\mathrm{c}} = \overline{E}^{\mathrm{c}} \cup \partial E,$$

where we have used the definition of interior and the fact mentioned above that the boundary of a set is the same as the boundary of its complement.

6.5. Convergence

We say that a sequence $x_n \in X$ **converges** to $x \in X$ in the metric ρ if given any $\varepsilon > 0$, we have for all sufficiently large n that $\rho(x_n, x) < \varepsilon$. When the metric ρ is understood, we denote the convergence of x_n to x by $x_n \to x$ or by $\lim_{n\to\infty} = x$.

When you see expressions like $x_n \to x$ or $\lim_{n\to\infty} = x$, you always need to make sure you know the metric under which the convergence is being considered.

Example 6.23. Let X denote the real numbers. Let ρ_a denote the absolute-value metric, and let ρ_t denote the discrete metric. Given any $\varepsilon > 0$, $\rho_a(1/n, 0) = |1/n - 0| = 1/n < \varepsilon$ for $n > 1/\varepsilon$. Hence $x_n = 1/n$ converges to zero under the absolute-value metric. However, under the discrete metric, $\rho_t(1/n, 0) = 1 < \varepsilon$ *fails* for $\varepsilon < 1$. Hence, $x_n = 1/n$ does *not* converge to zero under the discrete metric.

Another important aspect of the definition of convergence is the requirement that the proposed limit lie in the set X under consideration.

Example 6.24. Let $X = (0, 1]$ and put $x_n := 1/n$. Although $|x_n - 0| = 1/n$ tends to zero, the element $0 \notin X$. Hence, the sequence $x_n = 1/n \in X = (0, 1]$ does not have a limit in X under the absolute-value metric.

Example 6.25 (Continuity of the Inner Product). Let x_n be a sequence in an inner-product space X. If $x_n \to x \in X$, show that for any fixed $y \in X$, $\langle x_n, y\rangle \to \langle x, y\rangle$.

Solution. Did you notice that this question involves two metric spaces? First, there is the inner-product space X with its norm induced by the inner product, and the norm inducing a metric on X. Second, there is the space of scalars, either $\mathbb{R}$ or $\mathbb{C}$, equipped with the absolute-value norm/metric.

We must show that the sequence of scalars $\langle x_n, y\rangle$ converges to the scalar $\langle x, y\rangle$. Given $\varepsilon > 0$, we must show that for sufficiently large n,

$$\big|\langle x_n, y\rangle - \langle x, y\rangle\big| < \varepsilon. \tag{6.6}$$

Observe that by the Cauchy–Schwarz inequality,

$$\big|\langle x_n, y\rangle - \langle x, y\rangle\big| = \big|\langle x_n - x, y\rangle\big| \le \|x_n - x\|\,\|y\|.$$

Since the vectors x_n converge to the vector x, we know that $\|x_n - x\| < \varepsilon/\|y\|$ for large n. Hence, (6.6) holds for large n.

The importance of the result of Example 6.25 can be seen more clearly if we use the notation $x = \lim_{n\to\infty} x_n$. Then we can write

$$\lim_{n\to\infty} \langle x_n, y\rangle = \Big\langle \lim_{n\to\infty} x_n, y \Big\rangle. \tag{6.7}$$

In the definition of convergence, note that an equivalent way to write $\rho(x_n, x) < \varepsilon$ is to write $x_n \in B(x, \varepsilon)$. We use this to prove the following important result.

Theorem 6.26 (Characterization of Closed Sets). *A set E in a metric space is closed $\Leftrightarrow$ every sequence of points in E that converges has its limit in E.*

Proof. ($\Rightarrow$): Let E be closed. We must show that if $x_n \to x$ with $x_n \in E$, then $x \in E$. For a proof by contradiction, suppose otherwise that there is some sequence $x_n \in E$ that converges to a limit $x \notin E$. Then $x \in E^c$ where E^c is open. Hence, there is an $\varepsilon > 0$ with $B(x, \varepsilon) \subset E^c$. However, since $x_n \to x$, for large n, $x_n \in B(x, \varepsilon) \subset E^c$. For these n, $x_n \in E^c$ and $x_n \in E$, which is a contradiction.

($\Leftarrow$): Suppose that every converging sequence from E has its limit in E. We must prove that E is closed. For a proof by contradiction, suppose otherwise that E is not closed. Then E^c is not open. Hence, there is an $x \in E^c$ such that there is no open ball about x contained in E^c. This implies that for each open ball of the form $B(x, 1/n)$, $B(x, 1/n) \not\subset E^c$; i.e., there is an $x_n \in B(x, 1/n) \cap E$. In other words, $x_n \in E$

and $\rho(x_n,x) < 1/n$; i.e., x_n is a sequence in E that converges to a point $x \notin E$. This contradicts the original assumption that every converging sequence from E has its limit it E. □

Example 6.27. In an inner-product space, show that the orthogonal complement of any set is closed.

Solution. Let W be any subset. We must show that $W^\perp$ is closed. By Theorem 6.26, we must show that an arbitrary sequence in $W^\perp$ that converges must have its limit in $W^\perp$. So let $x_n \in W^\perp$, and assume $x_n \to x \in X$. We must show that $x \in W^\perp$; i.e., we must show that $x \perp w$ for all $w \in W$. Fix any $w \in W$ and write

$$\begin{aligned}\langle x, w\rangle &= \left\langle \lim_{n\to\infty} x_n, w \right\rangle \\ &= \lim_{n\to\infty} \langle x_n, w\rangle, \quad \text{by (6.7)}, \\ &= \lim_{n\to\infty} 0, \quad \text{since } x_n \in W^\perp, \\ &= 0.\end{aligned}$$

This says that $x \perp w$ or that $x \in W^\perp$ as required.

Theorem 6.28 (Approximation). *Let E be a subset of a metric space. Given $x \in \overline{E}$, either $x \in E$, or if $x \notin E$, we can approximate x by some $y \in E$. More precisely, given $\varepsilon > 0$, there is a $y \in E$ with $\rho(x,y) < \varepsilon$. Hence, by taking $\varepsilon = 1/n$, there is an $x_n \in E$ with $\rho(x_n,x) < 1/n$. In other words, there is a sequence from E that converges to x.*

Proof. Let $x \in \overline{E}$ with $x \notin E$. We need to show that for every $\varepsilon > 0$, there is a $y \in E$ with $y \in B(x,\varepsilon)$. Suppose otherwise that this is not the case. Then for some $\varepsilon > 0$, $B(x,\varepsilon) \cap E = \varnothing$. Equivalently, $E \subset B(x,\varepsilon)^c$, which is a closed set. Hence, $\overline{E} \subset B(x,\varepsilon)^c$. However, we now have $x \in \overline{E} \subset B(x,\varepsilon)^c$, which is a contradiction. □

Example 6.29. Let X denote the space of finite-energy waveforms on $[-\pi,\pi]$, equipped with the usual integral inner product. Let E denote the subset of all finite Fourier series of the form

$$\sum_{k=-n}^{n} c_k e^{jkt}.$$

The fact that every finite-energy waveform on $[-\pi,\pi]$ is the limit in norm of its finite-sum Fourier series can be expressed by the formula $\overline{E} = X$.

6.5.1. The Sampling Theorem

The **sampling theorem** says that if x is a finite-energy waveform bandlimited to W, then

$$x(t) = \sum_{k=-\infty}^{\infty} x(k/f_s)\,\mathrm{sinc}(f_s[t-k/f_s]),$$

where $f_s \geq 2W$. To derive the sampling theorem, we start with the fact that the assumptions on x imply

$$x(t) = \int_{-\infty}^{\infty} X(f)e^{j2\pi ft}\,df = \int_{-W}^{W} X(f)e^{j2\pi ft}\,df,$$

where X is the Fourier transform of x. Since X is zero outside of $[-W,W]$, we can also view X as zero outside of $[-f_s/2, f_s/2]$ for any $f_s \geq 2W$. Since x has finite energy, so does X, and we can view X as belonging to $L^2[-f_s/2, f_s/2]$. From the theory of Fourier series, we can write

$$X(f) = \sum_{k=-\infty}^{\infty} c_k e^{-j2\pi kf/f_s}, \quad |f| \leq f_s/2,$$

where the Fourier-series coefficients c_k are easily found to be $c_k = x(k/f_s)/f_s$. The above equation is understood as saying $\|X_n - X\|_2 \to 0$, where X_n is the nth partial sum,

$$X_n(f) := \sum_{k=-n}^{n} c_k e^{-j2\pi kf/f_s}, \quad |f| \leq f_s/2.$$

Now put $Y_t(f) := e^{-j2\pi ft}$, and observe that

$$\begin{aligned}
x(t) &= \int_{-\infty}^{\infty} X(f)e^{j2\pi ft}\,df \\
&= \int_{-f_s/2}^{f_s/2} X(f)e^{j2\pi ft}\,df \\
&= \int_{-f_s/2}^{f_s/2} X(f)\overline{Y_t(f)}\,df \\
&= \langle X, Y_t\rangle, && \text{since } X, Y_t \in L^2[-f_s/2, f_s/2], \\
&= \lim_{n\to\infty} \langle X_n, Y_t\rangle, && \text{by continuity of the inner product,} \\
&= \lim_{n\to\infty} \sum_{k=-n}^{n} c_k \int_{-f_s/2}^{f_s/2} e^{-j2\pi kf/f_s} e^{j2\pi ft}\,df \\
&= \sum_{k=-\infty}^{\infty} x(k/f_s)/f_s \cdot f_s\,\mathrm{sinc}(f_s[t-k/f_s]).
\end{aligned}$$

6.5.2. Bounded Sets and Bounded Sequences

A set E in a metric space (X,ρ) is **bounded** if E is contained in some open ball. More precisely, E is bounded if for some $x \in X$ and some $0 \le r < \infty$, $E \subset B(x,r)$. Note that once this is true for some point x and some radius r, we can write for any y and $r' := r + \rho(x,y)$, $E \subset B(y,r')$. Hence, the center of the ball does not matter for assessing boundedness. When X is a normed vector space, we usually restrict attention to balls centered at the origin. Hence, a set E in a normed vector space is bounded if for some r, $\|x\| < r$ for all $x \in E$.

A sequence x_n is **bounded** if it lies entirely in some open ball.

Proposition 6.30. *In a metric space, a convergent sequence is bounded.*

Proof. Suppose $x_n \to x$. For $\varepsilon = 1$, there is some N such that for all $n \ge N$, $\rho(x_n,x) < 1$. Now put

$$r := \max\{1, \rho(x_1,x), \ldots, \rho(x_{N-1},x)\}.$$

Then $\rho(x_n,x) \le r$ for $n < N$ and for $n \ge N$. If we insist on strict inequality, we can replace r by $r+1$. □

Example 6.31. Prove that if two sequences of real numbers converge, then their product converges to the product of the limits.

Solution. Suppose $x_n \to x$ and $y_n \to y$. We must show that $x_n y_n \to xy$. We give two proofs.

The first proof requires the following preliminary analysis. Write

$$\begin{aligned} |x_n y_n - xy| &= |x_n y_n - xy_n + xy_n - xy| \\ &\le |x_n y_n - xy_n| + |xy_n - xy| \\ &= |x_n - x|\,|y_n| + |x|\,|y_n - y| \\ &\le r|x_n - x| + |x|\,|y_n - y|, \end{aligned}$$

where the last step uses the fact that since y_n converges, y_n is bounded by some r.

Proof. Since y_n converges, we may assume $|y_n| \le r$ for some finite r and all n. Let $\varepsilon > 0$ be given. Since $x_n \to x$, for large n we have $|x_n - x| < \varepsilon/(2r)$. Since $y_n \to y$, for large n we have $|y_n - y| < \varepsilon/(2|x|)$. For large n,

$$\begin{aligned} |x_n y_n - xy| &= |x_n y_n - xy_n + xy_n - xy| \\ &\le |x_n - x|\,|y_n| + |x|\,|y_n - y| \\ &\le r|x_n - x| + |x|\,|y_n - y| \\ &< \varepsilon. \end{aligned}$$

□

The second proof requires a different preliminary analysis. It easy to see that $|(x_n - x)(y_n - y)| = |x_n - x|\,|y_n - y| \to 0$. Next, observe that

$$(x_n - x)(y_n - y) = x_n y_n - x y_n - x_n y + xy.$$

Rearrange this as

$$x_n y_n = (x_n - x)(y_n - y) + x y_n + x_n y - xy.$$

It is easy to show that $xy_n \to xy$ and $x_n y \to xy$. Hence,

$$\lim_{n\to\infty} x_n y_n = 0 + xy + xy - xy = xy.$$

We leave it to the reader to write a complete proof.

6.6. Cauchy Sequences and Completeness

If $x_n \to x$, then given any $\varepsilon > 0$, there is an N such that for all $n \geq N$, we have $\rho(x_n, x) < \varepsilon/2$. Hence, if n and m are *both* greater than or equal to N, then

$$\rho(x_n, x_m) \leq \rho(x_n, x) + \rho(x, x_m) < \varepsilon/2 + \varepsilon/2 = \varepsilon.$$

A sequence with the property that given any $\varepsilon > 0$, there exists an N such that for all $n, m \geq N$, we have $\rho(x_n, x_m) < \varepsilon$ is said to be **Cauchy**. From the foregoing observation, every convergent sequence is Cauchy.

If a metric space has the property that every Cauchy sequence converges to a point in the space, then the metric space is said to be **complete**. The following lemma can be combined with the Bolzano–Weierstrass Theorem to show that the real numbers are complete.

Lemma 6.32. *If a Cauchy sequence in a metric space has a subsequence that converges to a point x, then the sequence itself converges to x.*

Proof. Problem 6.32. □

Theorem 6.33. *The real numbers are complete under the absolute-value metric.*

Proof. Problem 6.33. □

Theorem 6.34. *The spaces $\mathbb{R}^d$ and $\mathbb{C}^d$ are complete under any of the three norms $\|\cdot\|_1$, $\|\cdot\|_2$, or $\|\cdot\|_\infty$ introduced in Example 6.7.*

Proof. First consider $\mathbb{R}^d$. Let $x_n = [x_{n,1},\ldots,x_{n,d}]^\mathsf{T}$ be a Cauchy sequence in $\mathbb{R}^d$. In other words, given $\varepsilon > 0$, there is an N such that for $n,m \geq N$, we have $\|x_n - x_m\| < \varepsilon$, where $\|\cdot\|$ denotes any one of the three norms that were introduced in Example 6.7. From (6.4), we see that the infinity norm is the smallest of the three, and so for any $1 \leq i \leq d$,

$$\varepsilon > \|x_n - x_m\| \geq \|x_n - x_m\|_\infty \geq |x_{n,i} - x_{m,i}|.$$

It now follows that for each i, the ith sequence $\{x_{n,i}\}_{n=1}^\infty$ is Cauchy in $\mathbb{R}$ and therefore converges to a limit $x_{*,i} \in \mathbb{R}$. Put $x_* := [x_{*,1},\ldots,x_{*,d}]^\mathsf{T}$. Since the 1-norm is the largest in (6.4),

$$\|x_n - x_*\| \leq \|x_n - x_*\|_1 = \sum_{i=1}^{d} |x_{n,i} - x_{*,i}|.$$

Since each absolute value on the right goes to zero, $\|x_n - x_*\| \to 0$.

Now consider $\mathbb{C}^d$. By identifying $\mathbb{C}$ under the usual absolute value for complex numbers with $\mathbb{R}^2$ under the 2-norm, the preceding paragraph shows that $\mathbb{C}$ is complete. Armed with this result, the argument of the preceding paragraph applies essentially verbatim to show that $\mathbb{C}^d$ is complete. □

If a normed vector space is complete, it is called a **Banach space**. If an inner-product space is complete, it is called a **Hilbert space**. The L^p spaces discussed in Section 6.2.1 are complete spaces. This fact, sometimes known as the **Riesz–Fischer Theorem**, is proved, for example in [33].

Theorem 6.35. *If $w_1, w_2, \ldots$ are orthonormal vectors in a Hilbert space X, then for all $x \in X$,*

$$\widehat{x} := \sum_{k=1}^{\infty} \langle x, w_k\rangle w_k \text{ exists, and } \|\widehat{x}\|^2 = \sum_{k=1}^{\infty} \big|\langle x, w_k\rangle\big|^2 \leq \|x\|^2 < \infty.$$

Furthermore, $\widehat{x}$ is the projection of x onto $\overline{\operatorname{span}\{w_1, w_2, \ldots\}}$. Also, given scalars $c_1, c_2, \ldots$, the infinite sum

$$\sum_{k=1}^{\infty} c_k w_k \text{ exists } \Leftrightarrow \sum_{k=1}^{\infty} |c_k|^2 < \infty.$$

Proof. Let $w_1, w_2, \ldots$ be an infinite sequence of orthonormal vectors in a Hilbert space X. Put $W_n := \text{span}\{w_1, \ldots, w_n\}$. We know from the proof of the Finite-Dimensional Projection Theorem that the projection of $x \in X$ onto W_n is given by

$$\widehat{x}_n := \sum_{k=1}^{n} \langle x, w_k \rangle w_k.$$

We also point out that since

$$\|\widehat{x}_n\|^2 = \sum_{k=1}^{n} \big|\langle x, w_k \rangle\big|^2,$$

and since $\|\widehat{x}_n\| \le \|x\|$ (recall (3.8)), $\|\widehat{x}_n\|^2$ is nondecreasing and bounded above. By Problem 6.1, $\|\widehat{x}_n\|^2$ converges, which implies $\|\widehat{x}_n\|^2$ is a Cauchy sequence in $\mathbb{R}$.

Let $W := \text{span}\{w_1, w_2, \ldots\}$. In other words, W is the collection of all linear combinations involving finitely many of the w_k. We now show that $\widehat{x}_n$ converges to some $\widehat{x} \in \overline{W}$ and that $\widehat{x}$ is the projection of x onto $\overline{W}$. Since we are working in a Hilbert space, it suffices to show that $\widehat{x}_n$ is Cauchy; it will then follow that $\widehat{x}_n$ converges. Since $\widehat{x}_n \in W_n \subset \overline{W}$, which is closed, the limit, denoted by $\widehat{x}$, must lie in $\overline{W}$. The Orthogonality Principle can be used to show that $\widehat{x}$ is the projection.

For $m > n$,

$$\|\widehat{x}_m - \widehat{x}_n\|^2 = \left\| \sum_{k=n+1}^{m} \langle x, w_k \rangle w_k \right\|^2 = \sum_{k=n+1}^{m} \big|\langle x, w_k \rangle\big|^2 = \|\widehat{x}_m\|^2 - \|\widehat{x}_n\|^2.$$

Hence, $\widehat{x}_n$ is a Cauchy sequence of vectors if and only if $\|\widehat{x}_n\|^2$ is a Cauchy sequence of real numbers, which, as noted above, it is. Therefore $\widehat{x}_n$ is a Cauchy sequence of vectors, and since X is a Hilbert space, $\widehat{x}_n$ converges as required. We denote the limit by $\widehat{x}$. It remains to show that $\widehat{x}$ satisfies the Orthogonality Principle (Problem 6.38).

Since $\big|\|\widehat{x}_n\| - \|\widehat{x}\|\big| \le \|\widehat{x}_n - \widehat{x}\| \to 0$, and since $\|\widehat{x}_n\| \le \|x\|$, we have $\|\widehat{x}\| \le \|x\|$. Furthermore,

$$\infty > \|x\|^2 \ge \|\widehat{x}\|^2 = \lim_{n\to\infty} \|\widehat{x}_n\|^2 = \lim_{n\to\infty} \sum_{k=1}^{n} \big|\langle x, w_k \rangle\big|^2 = \sum_{k=1}^{\infty} \big|\langle x, w_k \rangle\big|^2.$$

The results about the c_k follow by repeating the appropriate parts of the foregoing analysis with the necessary changes. □

A collection of orthonormal vectors, $w_1, w_2, \ldots$ in a Hilbert space X is said to be a **complete orthonormal set** if the closure of their span is equal to the whole space; i.e., if $\overline{\text{span}\{w_1, w_2, \ldots\}} = X$. In this case, every $x \in X$ is equal to its projection onto $\overline{\text{span}\{w_1, w_2, \ldots\}}$, and therefore

$$x = \sum_{k=1}^{\infty} \langle x, w_k \rangle w_k \quad \text{and} \quad \|x\|^2 = \sum_{k=1}^{\infty} \left| \langle x, w_k \rangle \right|^2.$$

The formula on the right is usually called **Parseval's equation**. The best-known example of a complete orthonormal set is the collection of complex exponentials $\{e^{jkt}, k = 0, \pm 1, \pm 2, \ldots\}$ in $L^2[-\pi, \pi]$.

Remark. It is possible to construct an inner-product space containing an *uncountable* subset of orthonormal vectors. Consider the set X of waveforms $x(t)$ such that

$$\lim_{T \to \infty} \frac{1}{T} \int_{-T}^{T} |x(t)|^2 \, dt$$

exists and is finite. For $x, y \in X$, define the inner product[f]

$$\langle x, y \rangle := \lim_{T \to \infty} \frac{1}{T} \int_{-T}^{T} x(t) \overline{y(t)} \, dt.$$

For real ω, consider the waveforms $\varphi_\omega(t) := e^{j\omega t}/\sqrt{2}$. Then $\{\varphi_\omega : \omega \in \mathbb{R}\}$ is an uncountable orthonormal subset of X.

6.6.1. The Projection Theorem for Hilbert Space

Theorem 6.36 (Projection Theorem for Hilbert Space). *Let C be a nonempty, closed, convex subset of a Hilbert space X. Then for every $x \in X$, the unique projection of x onto C exists.*

Proof. If we can establish the existence of at least one projection, uniqueness follows from Theorem 3.12. Given $x \in X$, we must show that there exists an $\widehat{x} \in C$ with

$$\|x - \widehat{x}\| \leq \|x - y\|, \quad \text{for all } y \in C. \tag{6.8}$$

[f] Strictly speaking, we have to agree not to disinguish between two functions x and y for which

$$\lim_{T \to \infty} \frac{1}{T} \int_{-T}^{T} |x(t) - y(t)|^2 \, dt = 0.$$

Let $h := \inf_{y\in C}\|x-y\|$. From the definition of infimum, there is a sequence $y_n \in C$ with $\|x-y_n\| \to h$. We will show that y_n is a Cauchy sequence. Since X is a Hilbert space, y_n converges to some limit in X. Since C is closed, the limit, which we call $\widehat{x}$, must be in C by Theorem 6.26. We conclude the proof by showing that (6.8) holds.

To show y_n is Cauchy, we proceed as follows. By the parallelogram law (Problem 3.3),

$$\begin{aligned} 2(\|x-y_n\|^2 + \|x-y_m\|^2) &= \|2x-(y_n+y_m)\|^2 + \|y_m-y_n\|^2 \\ &= 4\left\|x-\frac{y_n+y_m}{2}\right\|^2 + \|y_m-y_n\|^2. \end{aligned}$$

Note that the vector $(y_n+y_m)/2 \in C$ since C is convex. It now follows that

$$2(\|x-y_n\|^2 + \|x-y_m\|^2) \geq 4h^2 + \|y_m-y_n\|^2.$$

Since $\|x-y_n\| \to h$, given $\varepsilon > 0$, there exists an N such that for all $n \geq N$, $\|x-y_n\| < h+\varepsilon$. Thus, for $m,n \geq N$,

$$\|y_m-y_n\|^2 < 2((h+\varepsilon)^2 + (h+\varepsilon)^2) - 4h^2 = 4\varepsilon(2h+\varepsilon).$$

This shows that y_n is Cauchy. To establish (6.8), first write

$$\|x-\widehat{x}\| \leq \|x-y_n\| + \|y_n-\widehat{x}\|.$$

Since $\|x-y_n\| \to h$ and $\|y_n-\widehat{x}\| \to 0$, taking limits on both sides of the inequality yields $\|x-\widehat{x}\| \leq h$. Since by definition of h, $h \leq \|x-y\|$ for all $y \in C$, (6.8) holds. □

Remark. If W is a closed subspace of a Hilbert space X, then by the Orthogonality Principle, $X = W \oplus W^{\perp}$. A similar remark was made following the Finite-Dimensional Projection Theorem.

6.6.2. Fixed Points and Contraction Mappings

Consider a function $f: X \to X$, where X is equipped with a metric ρ. We say that $x \in X$ is a **fixed point** of f if $f(x) = x$. The mapping f is a **contraction** if there is a constant $0 \leq c < 1$ such that for all $x,y \in X$,

$$\rho\big(f(x),f(y)\big) \leq c\,\rho(x,y). \tag{6.9}$$

Theorem 6.37 (Contraction Mapping Theorem). *A contraction f on a complete metric space X has a unique fixed point x. Furthermore, the fixed point is the limit of the sequence*

$$x_{n+1} := f(x_n), \quad n = 1,2,3,\ldots,$$

where x_1 is any point in X. Also, if c is the contraction constant,

$$\rho(x,x_n) \le \frac{\rho(x_2,x_1)}{1-c}\, c^{n-1}. \tag{6.10}$$

Example 6.38. Suppose we want to solve an equation of the form $g(x) = y$, where $g: X \to X$ and X is a vector space. Notice that $g(x) = y$ if and only if for any $\lambda \neq 0$,

$$\lambda[g(x) - y] + x = x.$$

Setting $f(x) := \lambda[g(x) - y] + x$, we have that $g(x) = y$ is equivalent to $f(x) = x$. In other words, we have converted the problem of solving $g(x) = y$ into a fixed-point problem for f. If we can choose λ so that f is a contraction, and if X is complete, then the contraction mapping theorem gives us an algorithm for finding the solution x. The theorem even gives us a bound to use as a stopping criterion for the algorithm.

When $X = \mathbb{R}$, and f is differentiable, we can use the **mean-value theorem** to find a constant c in (6.9) that is less than one. Recall that for any x and y,

$$f(x) - f(y) = f'(t)(x - y),$$

where t lies between x and y. Then

$$|f(x) - f(y)| = |f'(t)| \cdot |x - y|.$$

Since $f'(t)$ depends on λ, we can adjust λ so that for some $c < 1$, $|f'(t)| \le c$ for all possible values of t.

Example 6.39 (Differential Equations). Consider the differential equation

$$x'(t) = h(t, x(t)), \quad x(t_0) = \xi_0.$$

Assuming there exists a waveform x that satisfies the above conditions, we can integrate both sides from t_0 to t and get

$$\int_{t_0}^{t} x'(\tau)\, d\tau = \int_{t_0}^{t} h(\tau, x(\tau))\, d\tau.$$

Since the left-hand side is $x(t) - x(t_0) = x(t) - \xi_0$, we can write

$$x(t) = \xi_0 + \int_{t_0}^{t} h(\tau, x(\tau))\, d\tau. \tag{6.11}$$

Notice that if we differentiate the above expression with respect to t, we recover $x'(t) = h(t,x(t))$.[g] Hence, x solves the differential equation with initial condition $x(t_0) = \xi_0$ if and only if x solves (6.11). To prove the existence and uniqueness of a solution of (6.11), consider the function f defined on waveforms x by

$$(f(x))(t) := \xi_0 + \int_{t_0}^{t} h(\tau, x(\tau))\, d\tau.$$

In other words, given a waveform x, $f(x)$ is the waveform whose value at time t is given by the above formula. Assuming $h(t,\xi)$ is continuous, whenever x is a continuous waveform, $f(x)$ will also be a continuous waveform.[h] Since continuous waveforms on a closed interval form a Banach space, if we can show f is a contraction, then f will have a unique fixed point. But $f(x) = x$ implies $(f(x))(t) = x(t)$, which is exactly (6.11).

Proof of the Contraction Mapping Theorem. First note that fixed points of contraction mappings are always unique. If x and y are both fixed points, then

$$\rho(x,y) = \rho\big(f(x), f(y)\big) \le c\,\rho(x,y).$$

If $x \neq y$, then $\rho(x,y) > 0$ and we can divide through by $\rho(x,y)$ and find $1 \le c$, which is a contradiction. We conclude that $x = y$.

We now turn to the algorithm $x_{n+1} = f(x_n)$. We show that x_n is a Cauchy sequence. Since X is complete, there must be a limit x with $\rho(x_n,x) \to 0$. To begin, write

$$\rho(x_3,x_2) = \rho\big(f(x_2), f(x_1)\big) \le c\,\rho(x_2,x_1).$$

Next,

$$\rho(x_4,x_3) = \rho\big(f(x_3), f(x_2)\big) \le c^2\,\rho(x_2,x_1).$$

In general,

$$\rho(x_{n+1},x_n) \le c^{n-1}\,\rho(x_2,x_1).$$

For $m > n$, use the triangle inequality to write

$$\rho(x_m,x_n) \le \sum_{k=n}^{m-1} \rho(x_{k+1},x_k) \le \sum_{k=n}^{m-1} c^{k-1}\,\rho(x_2,x_1).$$

Since

$$\sum_{k=n}^{m-1} c^{k-1} = \sum_{k=1}^{m-1} c^{k-1} - \sum_{k=1}^{n-1} c^{k-1},$$

[g] Assuming h is continuous and that x is continuous, the fundamental theorem of calculus shows that differentiating (6.11) results in $x'(t) = h(t,x(t))$.

[h] In fact, $f(x)$ will be continuously differentiable since then $(f(x))'(t) = h(t,x(t))$ is continuous.

we can use the finite **geometric series** formula to write

$$\sum_{k=n}^{m-1} c^{k-1} = \frac{1-c^{m-1}}{1-c} - \frac{1-c^{n-1}}{1-c} = \frac{c^{n-1}-c^{m-1}}{1-c} \leq \frac{c^{n-1}}{1-c}.$$

Thus,

$$\rho(x_m, x_n) \leq \rho(x_2, x_1)\frac{c^{n-1}}{1-c}, \tag{6.12}$$

and we see that x_n is a Cauchy sequence. Since X is complete, there is an $x \in X$ with $x_n \to x$.

We next show that this limit x is a fixed point of f. By the triangle inequality,

$$\begin{aligned}\rho\big(x, f(x)\big) &\leq \rho(x, x_{n+1}) + \rho\big(x_{n+1}, f(x)\big) \\ &= \rho(x, x_{n+1}) + \rho\big(f(x_n), f(x)\big) \\ &\leq \rho(x, x_{n+1}) + c\,\rho(x_n, x) \to 0.\end{aligned}$$

Since $\rho\big(x, f(x)\big) = 0$, $f(x) = x$.

To conclude the proof, we establish the bound (6.10). Let $\varepsilon > 0$ be arbitrary. For $m > n$, use (6.12) to write

$$\begin{aligned}\rho(x, x_n) &\leq \rho(x, x_m) + \rho(x_m, x_n) \\ &\leq \rho(x, x_m) + \frac{c^{n-1}}{1-c}\rho(x_2, x_1).\end{aligned}$$

If we now also assume m is large enough that $\rho(x, x_m) < \varepsilon$, we have

$$\rho(x, x_n) \leq \varepsilon + \frac{c^{n-1}}{1-c}\rho(x_2, x_1).$$

Since ε is arbitrary, (6.10) follows. □

6.7. Sequential Compactness

The notion of sequential compactness of subsets of real numbers carries over verbally unchanged to metric spaces. The only difference is that convergence is in terms of the metric. Here is the precise definition. If ρ is a metric on X, then a subset D is said to be **sequentially compact** if every sequence $x_n \in D$ has a subsequence x_{n_k} that converges to a point $x \in D$; i.e., $\lim_{k\to\infty} \rho(x_{n_k}, x) = 0$.

Proposition 6.40. *In a metric space, sequentially compact sets are closed and bounded.*

Proof. Let ρ be a metric on a space X, and let $D \subset X$ be sequentially compact. To show D is closed, we use Theorem 6.26 on the characterization of closed sets. Let $x_n \in D$ converge to a point $x \in X$. We must show that $x \in D$. Since D is sequentially compact, there is a subsequence x_{n_k} converging to a point $y \in D$. However, since $x_n \to x$, any subsequence x_{n_k} must also converge to x (Problem 6.34). Since limits are unique, $x = y \in D$ (Problem 6.35).

To prove that D is bounded, we assume otherwise and obtain a contradiction.[i] Suppose D is not bounded. Fix any point $x \in D$. Then for every $n = 1, 2, \ldots$, the set D is not contained in $B(x,n)$. In other words, there is some point $x_n \in D$ that is outside $B(x,n)$; i.e., $\rho(x_n, x) \geq n$. However, since D is sequentially compact, there is a converging subsequence x_{n_k} and a point $y \in D$ with $\rho(x_{n_k}, y) \to 0$ as $k \to \infty$. Furthermore, by the definition of subsequence, $n_k \to \infty$ as $k \to \infty$. Hence,

$$n_k \leq \rho(x_{n_k}, x) \leq \rho(x_{n_k}, y) + \rho(y, x).$$

The right-hand side is bounded, but the left-hand side tends to infinity. □

In $\mathbb{R}^d$ and $\mathbb{C}^d$ we have the following converse result.

Theorem 6.41. *Closed and bounded subsets of $\mathbb{R}^d$ or $\mathbb{C}^d$ are sequentially compact under any of the three norms $\|\cdot\|_1$, $\|\cdot\|_2$, or $\|\cdot\|_\infty$ introduced in Example 6.7.*

Proof. First consider $\mathbb{R}^d$. To keep the notation from getting out of hand, we treat the case $d = 2$. Let $[x_n, y_n]^\mathsf{T}$ be a sequence in a closed and bounded set D. Since $|x_n| \leq \|[x_n, y_n]^\mathsf{T}\|_\infty$ is bounded, the Bolzano–Weierstrass Theorem tells us that there is a subsequence x_{n_k} converging to a real number x. Similarly, since $|y_{n_k}| \leq \|[x_{n_k}, y_{n_k}]^\mathsf{T}\|_\infty$ is bounded, there is a further subsequence $y_{n_{k_l}}$ converging to a real number y. At this point, it is important to observe that since $x_{n_k} \to x$, we also have $x_{n_{k_l}} \to x$. No matter which of the three norms we use, writing

$$\|[x_{n_{k_l}}, y_{n_{k_l}}]^\mathsf{T} - [x, y]^\mathsf{T}\| \leq |x_{n_{k_l}} - x| + |y_{n_{k_l}} - y|$$

shows that as $l \to \infty$, the sub-subsequence converges to $[x, y]^\mathsf{T}$, which must lie in D on account of Theorem 6.26.

[i] The reader may find it helpful to review the discussion of bounded sets in Section 6.5.2.

Now consider $\mathbb{C}^d$. By identifying $\mathbb{C}$ under the usual absolute value for complex numbers with $\mathbb{R}^2$ under the 2-norm, the preceding paragraph shows that closed and bounded subsets of $\mathbb{C}$ are sequentially compact. This implies that a bounded sequence of complex numbers, which lies inside some closed and bounded disk, has a converging subsequence; i.e., we have the "Bolzano–Weierstrass Theorem for complex numbers." Armed with this result, the argument of the preceding paragraph applies essentially verbatim to show that $\mathbb{C}^d$ is complete. □

The foregoing proof makes essential use of the fact that $\mathbb{R}^d$ is finite dimensional. To make the first term on the right in the above inequality less than $\varepsilon/2$, we need $l \geq L_1$. To make the second term less than $\varepsilon/2$, we need $l \geq L_2$. Then when $l \geq \max(L_1, L_2)$, both terms will be less than $\varepsilon/2$, and their sum will be less than ε. This extends to making d terms all less than ε/d. However, if there were infinitely many terms, with the kth one less than $\varepsilon/2^k$, we would need $l \geq \sup(L_1, L_2, \ldots)$, which may be infinite.

Example 6.42. Consider the space of bounded elements in $\mathbb{R}^\infty$. A typical $\mathbf{x} \in \mathbb{R}^\infty$ has the form $\mathbf{x} = (x_1, x_2, \ldots)$ with $\|\mathbf{x}\| := \sup_n |x_n| < \infty$. We show that the closed and bounded ball $\{\mathbf{x} : \|\mathbf{x}\| \leq 1\}$ is not sequentially compact. Put $\mathbf{x}_1 := (1,0,0,0,\ldots)$, $\mathbf{x}_2 := (0,1,0,0,\ldots)$, $\mathbf{x}_3 := (0,0,1,0,\ldots)$, and so on. For $m \neq n$, $\mathbf{x}_n - \mathbf{x}_m$ has 1 in position n, -1 in position m, and 0s elsewhere. Hence, $\|\mathbf{x}_n - \mathbf{x}_m\| = 1$. It follows that no subsequence can be Cauchy,[j] and hence, no subsequence can converge.

6.8. Continuous Functions

Consider a function $f: X \to Y$, where X is equipped with metric ρ and Y is equipped with metric m. We say that f is **continuous at a point** $x_0 \in X$ if

$$\begin{aligned} &\forall \varepsilon > 0, \exists \delta > 0, \forall x \in X, \\ &\qquad \rho(x, x_0) < \delta \Rightarrow m(f(x), f(x_0)) < \varepsilon. \end{aligned} \tag{6.13}$$

Letting B_ρ and B_m denote open balls in X and Y, respectively, we can rewrite the above implication as

$$x \in B_\rho(x_0, \delta) \Rightarrow f(x) \in B_m(f(x_0), \varepsilon).$$

In other words, $f(x)$ is close to $f(x_0)$ whenever x is close enough to x_0. "Close enough" is determined by δ, which depends on the point x_0.

[j] A little more detail is required. Given $\varepsilon = 1/2$, suppose there is a K such that for $k, l \geq K$, $\|\mathbf{x}_{n_k} - \mathbf{x}_{n_l}\| < 1/2$. In particular, $\|\mathbf{x}_{n_K} - \mathbf{x}_{n_l}\| < 1/2$. Since $n_l \to \infty$ as $l \to \infty$, we know that for large enough $l \geq K$, $n_l > n_K$. For such l, we have $1 = \|\mathbf{x}_{n_K} - \mathbf{x}_{n_l}\| < 1/2$.

If a function is continuous at every point in a subset of the whole space, then we say that the function is **continuous on the subset**.

As you will show in Problem 6.44, the product of two continuous functions is continuous. Since the function $f(x) = x$ for $x \in \mathbb{R}$ is continuous, it follows that x^2, x^3, etc. are continuous. More generally, since the sum of two continuous functions is continuous, polynomials are continuous functions.

Theorem 6.43 (Convergence Preservation). *A function between metric spaces is continuous if and only if the function is convergence preserving.*

Proof. Let $f: X \to Y$, where X is equipped with metric ρ and Y is equipped with metric m. Before proving anything, we must be clear about what we mean when we say f is convergence preserving at a point x_0. We mean that for *all* sequences x_n that converge to x_0 in the metric ρ, the image sequence $f(x_n)$ converges to $f(x_0)$ in the metric m.

There are two things to prove. First, if f is continuous at a point $x_0 \in X$, we must show that if x_n is any sequence converging to x_0, then $f(x_n) \to f(x_0)$. We leave this to the reader in Problem 6.45.

The second thing we must prove is that if f is convergence preserving at x_0, then f is continuous at x_0. Suppose otherwise that f is not continuous at x_0. Then there is some $\varepsilon_0 > 0$ such that for every $\delta = 1/n$, there is an x_n with $\rho(x_n, x_0) < \delta = 1/n$, but $m(f(x_n), f(x_0)) \geq \varepsilon_0$. Since $\rho(x_n, x_0) < 1/n$, x_n converges to x_0. However, for this sequence, $m(f(x_n), f(x_0)) \geq \varepsilon_0$ shows that $f(x_n)$ cannot converge to $f(x_0)$. This contradicts the assumption that f preserves convergence of *all* sequences that converge to x_0. □

Theorem 6.44. *A real-valued continuous function on a sequentially compact subset of a metric space achieves its maximum and minimum values on that subset.*

Proof. Let D denote the sequentially compact subset, and let $f: D \to \mathbb{R}$ be continuous. It is enough to prove that the maximum is achieved. We defer to later the fact that f is bounded above. With this assumption, we have $\overline{f} := \sup_{x \in D} f(x) < \infty$. By definition of supremum, there is a sequence $x_n \in D$ with $f(x_n) \to \overline{f}$. Since D is sequentially compact, there is a subsequence x_{n_k} with $x_{n_k} \to x_0 \in D$. Since f is continuous on D, f is convergence preserving at x_0. Therefore, $f(x_{n_k}) \to f(x_0)$ as $k \to \infty$. However, since $f(x_n) \to \overline{f}$, we have $f(x_{n_k}) \to \overline{f}$ as well. Since limits are unique (Problem 6.35), we have $f(x_0) = \overline{f}$.

It remains to show that f is bounded above on D. Suppose otherwise that for $n = 1, 2, \ldots$, there is an $x_n \in D$ with $f(x_n) > n$. Since D is sequentially compact,

there is a converging subsequence $x_{n_k} \to x_0 \in D$ as $k \to \infty$. Since f is continuous, $f(x_{n_k}) \to f(x_0)$. Since convergent sequences are bounded (Proposition 6.30), this contradicts $f(x_{n_k}) > n_k \to \infty$. □

A nice application of Theorem 6.44 is the following lemma, which we will then use to prove that finite-dimensional subspaces of a normed vector space are complete and therefore closed.

Lemma 6.45. *Let $w_1, \ldots, w_d$ be linearly independent vectors in a normed vector space, and for scalars $c_1, \ldots, c_d$, put*

$$w := \sum_{i=1}^{d} c_i w_i.$$

Let $\underline{c} := [c_1, \ldots, c_d]^\mathsf{T}$. Then there exist positive finite constants K_1 and K_2 such that

$$K_1 \|\underline{c}\|_\infty \le \|w\| \le K_2 \|\underline{c}\|_\infty,$$

where $\|\cdot\|_\infty$ denotes the infinity norm on $\mathbb{R}^d$ or $\mathbb{C}^d$ as appropriate.

Proof. First write

$$\|w\| = \|c_1 w_1 + \cdots + c_d w_d\| \le \sum_{i=1}^{d} |c_i| \, \|w_i\| \le \|\underline{c}\|_\infty \underbrace{\left(\sum_{i=1}^{d} \|w_i\| \right)}_{=: K_2}.$$

To obtain K_1 is more difficult. Define the real-valued function

$$f(\underline{c}) := \left\| \sum_{i=1}^{d} c_i w_i \right\|.$$

It is easy to show that f is a continuous. Consider minimizing f over $\{\underline{c} : \|\underline{c}\|_\infty = 1\}$. This closed and bounded set is sequentially compact by Theorem 6.41. Hence, the minimum of f on this set is achieved by some $\underline{c}^*$ with $\|\underline{c}^*\|_\infty = 1$. Put

$$K_1 := \min_{\underline{c} : \|\underline{c}\|_\infty = 1} f(\underline{c}) = f(\underline{c}^*).$$

Note that K_1 cannot be zero. If it were, then we would have $0 = f(\underline{c}^*) = \|c_1^* w_1 + \cdots + c_d^* w_d\| = 0$; then linear independence would force $\underline{c}^* = 0$, which would contradict $\|\underline{c}^*\| = 1$. To complete the proof, consider any $w = c_1 w_1 + \cdots + c_d w_d$. Then $w = 0$ if and only if $\underline{c} = 0$. For $w \neq 0$,

$$\|w\| = \|\underline{c}\|_\infty \left\| \sum_{i=1}^{d} \frac{c_i}{\|\underline{c}\|_\infty} w_i \right\| = \|\underline{c}\|_\infty f\left(\frac{\underline{c}}{\|\underline{c}\|_\infty} \right) \ge \|\underline{c}\|_\infty f(\underline{c}^*) = K_1 \|\underline{c}\|_\infty.$$ □

Theorem 6.46. *In a normed vector space, every finite-dimensional subspace is complete and therefore closed.*

Proof. Let W be a finite-dimensional subspace of a normed vector space X. Then W has a basis, say $w_1, \ldots, w_d$. If $w^{(n)}$ is any sequence in W, then for each n,

$$w^{(n)} = c_1^{(n)} w_1 + \cdots + c_d^{(n)} w_d$$

for a unique d-tuple of scalars, $\underline{c}^{(n)} := [c_1^{(n)}, \ldots, c_d^{(n)}]^{\mathsf{T}}$. By Lemma 6.45, for any n and m,

$$K_1 \|\underline{c}^{(n)} - \underline{c}^{(m)}\|_\infty \le \|w^{(n)} - w^{(m)}\| \le K_2 \|\underline{c}^{(n)} - \underline{c}^{(m)}\|_\infty.$$

Suppose $w^{(n)}$ is Cauchy. Then by the left-hand inequality, $\underline{c}^{(n)}$ is Cauchy in $\mathbb{R}^d$ or $\mathbb{C}^d$ as appropriate. Hence, there is a $\underline{c}$ with $\|\underline{c}^{(n)} - \underline{c}\|_\infty \to 0$. Put $w := c_1 w_1 + \cdots + c_d w_d$, and use Lemma 6.45 to write

$$K_1 \|\underline{c}^{(n)} - \underline{c}\|_\infty \le \|w^{(n)} - w\| \le K_2 \|\underline{c}^{(n)} - \underline{c}\|_\infty.$$

This shows that $w^{(n)} \to w \in W$. Hence W is complete.

It is similarly easy to show that W is closed. If $w^{(n)}$ converges to some $x \in X$, then $w^{(n)}$ is Cauchy, and by the foregoing argument $w^{(n)} \to w \in W$. By uniqueness of limits $x = w \in W$. Hence, W is closed. □

6.8.1. Uniform Continuity

Sometimes we have a stronger kind of continuity called **uniform continuity**. We say that f is **uniformly continuous** if $f(x)$ and $f(x_0)$ are close whenever x and x_0 are close enough, and "close enough" does not depend on either of the two points. More precisely, we say f is **uniformly continuous** if

$$\begin{aligned} &\forall \varepsilon > 0, \exists \delta > 0, \forall x, x_0 \in X, \\ &\qquad \rho(x, x_0) < \delta \Rightarrow m\big(f(x), f(x_0)\big) < \varepsilon. \end{aligned}$$

Notice that both x and x_0 appear *after* the existence of δ is mentioned. This is in contrast with (6.13), where x_0 is fixed *before* the existence of δ is mentioned.

Example 6.47. The function f defined in the proof of Lemma 6.45 is actually uniformly continuous, as you will show in Problem 6.54.

Notes

Note 6.1. We denote by $\overline{\mathbb{R}}$ the set consisting of $\mathbb{R}$ together with the symbols ∞ and $-\infty$. We call $\overline{\mathbb{R}}$ the **extended real numbers**. The symbols $\pm\infty$ interact with real numbers as follows. If x is a real number, then $-\infty < x < \infty$ and $x \pm \infty = \pm\infty$. If x is a positive real number, then $x \cdot (\pm\infty) = \pm\infty$. If x is a negative real number, then $x \cdot (\pm\infty) = \mp\infty$. We also use the convention $0 \cdot (\pm\infty) = 0$.

Problems

1. A sequence of real numbers x_n is said to be **monotonic increasing** if $x_n \leq x_{n+1}$ for all n. The sequence is said to be **monotonic decreasing** if $x_n \geq x_{n+1}$ for all n. If x_n is monotonic increasing and bounded above, prove that the sequence has a limit. *Hint:* Put $x := \sup_n x_n$, and prove that $x_n \to x$.

2. ***Geometric Series.*** Put $S_N := 1 + z + z^2 + \cdots + z^{N-1}$.

   (a) If $z \neq 1$, show that $S_N = (1-z^N)/(1-z)$. *Hint:* Write S_{N+1} in two different ways. First, $S_{N+1} = S_N + z^N$. Second, $S_{N+1} = 1 + zS_N$.

   (b) For $|z| < 1$, show that $\lim_{N\to\infty} S_N = 1/(1-z)$.

3. Consider the sequence $x_n := \sum_{k=0}^{n} 1/k!$.

   (a) Show that
   $$x_n \leq 1 + \sum_{m=0}^{n-1} \left(\frac{1}{2}\right)^m.$$

   (b) Show that $x_n \leq 3$.

   (c) Explain why
   $$e := \lim_{n\to\infty} \sum_{k=0}^{n} \frac{1}{k!}$$
   exists.

4. Let C be a convex subset of a vector space X, and let $\{f_\alpha\}$ be a family of convex functions defined on C. Show that
   $$f(x) := \sup_\alpha f_\alpha(x), \quad x \in C,$$
   is convex.

5. Given any function $L(\lambda, x)$, show that[k]

$$\sup_{\lambda} \inf_{x} L(\lambda, x) \leq \inf_{x} \sup_{\lambda} L(\lambda, x).$$

6. Suppose that in the previous problem, there exist λ_0 and x_0 such that (λ_0, x_0) is a **saddle point**; i.e.,

$$L(\lambda, x_0) \leq L(\lambda_0, x_0) \leq L(\lambda_0, x), \quad \text{for all } \lambda, x.$$

In this case, establish the reverse inequality,

$$\sup_{\lambda} \inf_{x} L(\lambda, x) \geq \inf_{x} \sup_{\lambda} L(\lambda, x).$$

Remark. If a saddle point (λ_0, x_0) exists, then

$$\sup_{\lambda} \inf_{x} L(\lambda, x) = \inf_{x} \sup_{\lambda} L(\lambda, x) = L(\lambda_0, x_0).$$

For other conditions that allow the interchange of supremum and infimum, see [22].

7. Suppose that

$$\sup_{\lambda} \inf_{x} L(\lambda, x) = \inf_{x} \sup_{\lambda} L(\lambda, x).$$

If λ_0 achieves the supremum on the left and x_0 achieves the infimum on the right, show that (λ_0, x_0) is a saddle point.

8. Suppose A and B are subsets of real numbers. Be sure to consider special cases of empty sets and sets that are not bounded below.

   (a) If $A \subset B$, show that $\inf B \leq \inf A$.
   (b) Show that $\inf A \cup B = \min\{\inf A, \inf B\}$. Do not assume $A \subset B$.

9. Let X be a normed vector space. For fixed $x_0 \in X$, define the function $\varphi(x) := \|x - x_0\|$. Determine whether or not φ is convex.

10. Show that $\|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|_\infty$ defined in Section 6.2 satisfy the properties of a norm.

[k] The expression is more precisely understood as

$$\sup_{\lambda} \Big[\inf_{x} L(\lambda, x) \Big] \leq \inf_{x} \Big[\sup_{\lambda} L(\lambda, x) \Big].$$

11. From (6.4), we have $(1/d)\|x\|_1 \le \|x\|_2 \le d\|x\|_\infty$. Derive the improved bounds $(1/\sqrt{d}\,)\|x\|_1 \le \|x\|_2 \le \sqrt{d}\,\|x\|_\infty$.

12. For $1 \le p < \infty$, if $x \in L^p$, put

$$\|x\|_p := \left(\int |x(t)|^p \, dt \right)^{1/p}.$$

Suppose $x \in L^p$ and $y \in L^q$ for $1 < p,q < \infty$ satisfying $\frac{1}{p} + \frac{1}{q} = 1$. Derive **Hölder's inequality**,

$$\int |x(t)y(t)| \, dt \le \|x\|_p \, \|y\|_q.$$

Hints: Rewrite the inequality as

$$\int \frac{|x(t)y(t)|}{\|x\|_p \, \|y\|_q} \, dt \;\le\; 1.$$

Since $\exp(\cdot)$ is a convex function on $\mathbb{R}$,

$$\exp\left[\tfrac{1}{p}\alpha + \tfrac{1}{q}\beta\right] \le \tfrac{1}{p}e^{\alpha} + \tfrac{1}{q}e^{\beta}.$$

Set $\alpha(t) := \ln(|x(t)|/\|x\|_p)^p$ and $\beta(t) := \ln(|y(t)|/\|y\|_q)^q$ and integrate.

13. Using the notation of the preceding problem, derive **Minkowski's inequality**,

$$\|x+y\|_p \le \|x\|_p + \|y\|_p,$$

where $1 \le p < \infty$. *Hint:* Observe that

$$\begin{aligned} \|x+y\|_p^p &= \int |x(t)+y(t)| \, |x(t)+y(t)|^{p-1} \, dt \\ &\le \int |x(t)| \, |x(t)+y(t)|^{p-1} \, dt + \int |y(t)| \, |x(t)+y(t)|^{p-1} \, dt, \end{aligned}$$

and apply Hölder's inequality.

14. **MATLAB.** ***Chernoff Bound.*** Let X be a **Laplace random variable** with density $f(x) = e^{-|x|}/2$. Use MATLAB to plot $\mathsf{P}(X > x)$ and the Chernoff bound on a semilog scale for $x \in [-5, 20]$. *Note:* When the Chernoff bound is greater than one, you should replace the bound by one since probabilities are always less than or equal to one.

15. Let ρ be a metric, and put $d(x,y) := \rho(x,y)/[1+\rho(x,y)]$. Show that d is also a metric. *Hint:* Show that $f(t) := t/(1+t)$ is an increasing function of t on $[0,\infty)$.

16. Show that with the discrete metric, every subset of X is an open subset.

17. Show that ρ_∞, ρ_1, and ρ_2 defined in Example 6.14 satisfy the properties of a metric.

18. Show that for $x \in X$ and $r \geq 0$, $F := \{y \in X : \rho(x,y) \leq r\}$ is a closed set.

19. Let ρ be a metric on a space X. On X^d consider the three metrics ρ_1, ρ_2, and ρ_∞ defined in Example 6.14. For $x \in X^d$, use (6.5) to show that

$$B_{\rho_\infty}(x, \varepsilon/d) \subset B_{\rho_1}(x,\varepsilon) \subset B_{\rho_2}(x,\varepsilon) \subset B_{\rho_\infty}(x,\varepsilon).$$

Hence, a set in X^d is open under any one of the three metrics if and only if the set is open under the other two metrics.

20. Using the setup of Example 6.15, show that if U_i is an open subset of X_i, then $U_1 \times \cdots \times U_d$ is an open subset of $X_1 \times \cdots \times X_d$ using the metric ρ_2.

21. Let U be a subset of E, and suppose that U is open. Show that $U \subset E^{\mathrm{o}}$.

22. If E is a subset of a metric space, show that $x \in E^{\mathrm{o}}$ if and only if there exists an $\varepsilon > 0$ with $B(x,\varepsilon) \subset E$.

23. Under the discrete metric, show that the boundary of every set is empty.

24. Give an example of a metric space in which for some $0 < r < \infty$, the closure of the open ball, $\overline{B(x,r)}$, is *not* equal to the closed ball $\{y \in X : \rho(x,y) \leq r\}$.

25. Let X denote the rational numbers, and for $x, y \in X$, let $\rho(x,y) := |x-y|$ be the usual absolute-value metric. Show that if r is a positive, *ir*rational number, then the open ball $B(0,r)$ is a closed set and the closed ball $\{x \in X : |x| \leq r\}$ is an open set.

26. Suppose that the real numbers, $\mathbb{R}$, is equipped with a metric such that every subset is both open and closed. If $x_n \to x$, determine whether or not $x_n = x$ for all sufficiently large n. **Caution:** Do *not* assume that the metric is the discrete metric!

27. Let U be a subset of a real inner-product space X. The **polar** of U is

$$U^- := \{x \in X : \langle x, u \rangle \leq 0 \text{ for all } u \in U\}.$$

Determine whether or not U^- is a closed set.

28. In a normed vector space, determine whether or not the closure of a convex set is convex.

29. In a normed vector space, determine whether or not the closure of the open ball is equal to the closed ball. In symbols, does $\overline{B} = C$, where $B := \{x \in X : \|x\| < 1\}$ and $C := \{x \in X : \|x\| \leq 1\}$?

30. Let X be a metric space with metric ρ. Suppose that A and B are two *nonempty*, *disjoint*, *closed* subsets of X. Put

$$d := \inf_{x \in A, y \in B} \rho(x,y).$$

Determine whether or not $d > 0$.

31. Show that a Cauchy is sequence is bounded.

32. Prove Lemma 6.32.

33. Use Lemma 6.32 and the Bolzano–Weierstrass Theorem 6.5 to prove that the real numbers are complete (Theorem 6.33).

34. In a metric space, show that if $\lim_{n\to\infty} x_n = x$ and x_{n_k} is any subsequence, then $\lim_{k\to\infty} x_{n_k} = x$

35. Show that limits are unique in a metric space. *Hint:* Let $x_n \to x$ and $x_n \to y$. Show that $\rho(x,y) = 0$.

36. Let $X := (0,1]$ be equipped with the discrete metric. Determine whether or not X is complete.

37. ***Continuity of the Inner Product.*** In an inner-product space, show that if $x_n \to x$ and $y_n \to y$, then $\langle x_n, y_n \rangle \to \langle x, y \rangle$.

38. Prove that $\widehat{x}$ in Theorem 6.35 satisfies the Orthogonality Principle for $\overline{W}$. *Hints:* First use (6.7) to show that $\langle x - \widehat{x}, w_m \rangle = 0$ for each m. Second, use this to show that $\langle x - \widehat{x}, w \rangle = 0$ for $w \in W := \text{span}\{w_1, w_2, \ldots\}$. Third, for $y \in \overline{W}$, use the Approximation Theorem 6.28 to prove that $\langle x - \widehat{x}, y \rangle = 0$.

39. **MATLAB.** Consider the task of developing an algorithm to compute $\sqrt{y}$ for $0 \leq y \leq 9$. Observe that $x = \sqrt{y}$ if and only if $x \in [0,3]$ solves $x^2 = y$. The desired contraction is $f(x) = \lambda[x^2 - y] + x$.

    (a) Determine all values of λ that make f a contraction.

    (b) What value of λ do you expect to work best; i.e., make the algorithm converge faster? Why?

(c) Write a MATLAB script to implement your algorithm. Use your algorithm to compute the sequence x_n to approximate $\sqrt{2}$. Your sequence will depend on λ, if you want your algorithm to converge faster, use a good value of λ. Your sequence will also depend on x_1. Use $x_1 = 1.5$. Print out the first few values of x_n to see that it is converging to the correct value.

40. ***Generalized Contraction Mapping Theorem.*** Suppose that $\rho(f(x), f(y)) \leq c\,\rho(x,y)$ for some $1 \leq c < \infty$ so that f may not be a contraction. Let f^m denote f iterated m times; e.g., $f^3(x) = f(f(f(x)))$. Show that if f^m is a contraction for some positive integer m, then f itself has a unique fixed point. *Hints:* Define $x_{n+1} = f(x_n)$ as in the proof of the contraction mapping theorem. Observe that if $x_n \to x$, then the proof that x is a fixed point of f goes through as before. Furthermore, any fixed point of f must be a fixed point of f^m and so must be unique since f^m is a contraction. The challenge is to prove x_n converges. To this end, note that every positive integer n can be written uniquely as $n = km + r$, where k is a nonnegative integer, and $r \in \{0, \ldots, m-1\}$. Hence,

$$x_{n+1} = f^n(x_1) = f^{km+r}(x_1) = ((f^m)^k)(f^r(x_1)).$$

41. Let E and F be sequentially compact subsets of a metric space X. Put $G := E \cap F$. Determine whether or not G is sequentially compact.

42. Let f be a real-valued function defined on a metric space. If f is continuous at a point, show that f is bounded in a neighborhood of that point. More specifically, if f is continuous at x_0, show that there is a positive constant $B < \infty$ and a $\delta > 0$ such that for all x with $\rho(x, x_0) < \delta$, we have $|f(x)| \leq B$.

43. If f is a strictly increasing function mapping a nonempty interval I *onto* an interval J, show that f is continuous.

44. Let f and g be real-valued mappings defined on a metric space. If f and g are both continuous at a point x_0, show that their product $f(x)g(x)$ is continuous at x_0.

45. Prove the first part of Theorem 6.43.

46. Let $x_n := \prod_{k=1}^{n} 3^{1/k^2}$. Find $\lim_{n\to\infty} x_n$ by using the fact that $\sum_{k=1}^{\infty} 1/k^2 = \pi^2/6$. *Hint:* Let $y_n := \ln x_n$, and use the fact that $\exp(\cdot)$ is a continuous function.

47. Let X and Y be metric spaces, and let $D \subset X$ be sequentially compact. Let $f: D \to Y$ be continuous, and put $f(D) := \{f(x) : x \in D\}$. Show that $f(D)$ is sequentially compact.

48. Let X be a space that is complete under the metric ρ, and let $E \subset X$ be closed. Let Y be another space with metric m, and let $f:E \to Y$ be continuous. Show that if

$$m\big(f(x),f(z)\big) \geq \rho(x,z), \quad \text{for all } x,z \in E,$$

then $f(E) := \{f(e) : e \in E\}$ is closed.

49. Let $f:\mathbb{R} \to \mathbb{R}$ be continuous. Put $B := \{(x,y) \in \mathbb{R}^2 : y \geq f(x)\}$. Determine whether or not B a closed subset of $\mathbb{R}^2$. Justify your answer.

50. Consider the situation in Problem 5.8. Make the additional assumptions that X is a normed vector space and that C is closed. Determine whether or not E is closed.

51. Let $f:X \to Y$ be onto. Assume that X is equipped with metric ρ and that Y is equipped with metric m. Show that if

$$m\big(f(x_1),f(x_2)\big) \geq \rho(x_1,x_2), \quad \text{for all } x_1,x_2 \in X,$$

then f is invertible and that f^{-1} is continuous.

52. ***Epigraph.*** If C is a subset of X and $f:C \to [-\infty,\infty]$ is an extended-real-valued function, the **epigraph** of f is

$$\operatorname{epi} f := \{(x,t) : x \in C, t \in \mathbb{R} \text{ and } f(x) \leq t\}.$$

Notice that if $(x,t) \in \operatorname{epi} f$, then since $t \in \mathbb{R}$, we have $f(x) \leq t < \infty$; however, it is possible that $f(x) = -\infty$. If we know that $f:C \to (-\infty,\infty]$, then $(x,t) \in \operatorname{epi} f$ implies $-\infty < f(x) < \infty$. If f takes only the value ∞, then $\operatorname{epi} f = \varnothing$. The epigraph is regarded as a subset of the **Cartesian product**,

$$X \times \mathbb{R} := \{(x,t) : x \in X \text{ and } t \in \mathbb{R}\}.$$

If X is a metric space with metric ρ, then we equip $X \times \mathbb{R}$ with the metric

$$m\big((x,t),(y,s)\big) := \sqrt{\rho(x,y)^2 + |t-s|^2}.$$

(a) Show that if C is closed and f is finite and continuous on C, then $\operatorname{epi} f$ is a closed subset of $X \times \mathbb{R}$.

(b) Show that if X is complete, then $X \times \mathbb{R}$ is complete.

(c) Suppose that C is a convex subset of a vector space X. If $f:C \to (-\infty,\infty]$, show that f is a convex function (satisfies (5.6) for all $x,y \in C$) if and only if $\operatorname{epi} f$ is a convex subset of $X \times \mathbb{R}$.

53. Let a space X be equipped with a metric ρ. For any subset $A \subset X$, and any $x \in X$, define $d(x,A) := \inf_{y\in A} \rho(x,y)$.

    (a) Show that $d(x,A) = 0 \Leftrightarrow x \in \overline{A}$.

    (b) For any set A, define $A_n := \{x \in X : d(x,A) < 1/n\}$. Show that $\bigcap_{n=1}^{\infty} A_n = \overline{A}$.

    (c) Show that for any set A, $d(x,A)$ is a continuous function of x.

    (d) The **indicator function** of a set A is $\mathbf{1}_A(x) := 1$ for $x \in A$ and $\mathbf{1}_A(x) := 0$ otherwise. For a closed set F, show that

    $$\lim_{n\to\infty} \mathbf{1}_{F_n}(x) = \mathbf{1}_F(x).$$

    (e) In general, indicator functions are not continuous. However, for closed subsets of a metric space, we can construct a continuous approximation as follows. The construction begins with the continuous function

    $$\varphi(t) := \begin{cases} 1, & t < 0, \\ 1-t, & 0 \le t \le 1, \\ 0, & t > 1. \end{cases}$$

    Next, given a closed set F, consider the continuous functions defined by $g_n(x) := \varphi(n\,d(x,F))$. Show that

    $$\mathbf{1}_F(x) \le g_n(x) \le \mathbf{1}_{F_n}(x).$$

    Also show that $\lim_{n\to\infty} g_n(x) = \mathbf{1}_F(x)$.

54. Prove that the function f introduced in the proof of Lemma 6.45 is uniformly continuous.

55. Consider the function $f\colon (0,\infty) \to \mathbb{R}$ defined by $f(x) := \ln x$. Determine whether or not f is uniformly continuous.

56. Consider the sets $X := (0,\infty)$ and $Y := \mathbb{R}$.

    (a) If $f\colon X \to Y$ is one-to-one, show that the formula

    $$\rho_f(x_1,x_2) := |f(x_1) - f(x_2)|, \quad x_1,x_2 \in X,$$

    defines a metric on X.

    (b) If $f(x) := \ln x$, determine whether or not the sequence $x_n := 1/n$ is Cauchy under the metric ρ_f defined in part (a).

    (c) If $f(x) := \ln x$, determine whether or not X is complete under the metric ρ_f defined in part (a).

(d) If $f(x) := e^{-x}$, determine whether or not X is complete under the metric ρ_f defined in part (a).

57. Consider the set $X := (-\pi, \pi]$ equipped with the metric

$$\rho(x,y) := \left|e^{jx} - e^{jy}\right|, \quad x,y \in (-\pi,\pi].$$

Determine whether or not X is complete.

58. Let D be a nonempty, closed and bounded subset of a normed vector space X. Show that if D is contained in a finite-dimensional subspace of X, then D is sequentially compact.

59. Let D be a nonempty, closed subset of a normed vector space X. Assume that D is contained in a finite-dimensional subspace of X. Show that if $x \in X$, then there is an $\widehat{x} \in D$ with $\|x - \widehat{x}\| \le \|x - y\|$ for all $y \in D$. *Hint:* As in the proof of the Projection Theorem for Hilbert Space, put $h := \inf_{y\in D} \|x - y\|$, and let $y_n \in D$ satisfy $\|x - y_n\| \to h$. Show that y_n is bounded and apply the result of the previous problem appropriately.

60. Consider the function

$$f(x) := [1 + (x/3)^2](2 + \cos x), \quad x \in \mathbb{R}.$$

Put $\mu := \inf_{x\in\mathbb{R}} f(x)$. Determine whether or not there exists a real number x_0 such that $f(x_0) = \mu$.

61. Let f be positive, strictly decreasing, and continuous on $[0,\infty)$, with $f(x) \to 0$ as $x \to \infty$. Put $M := f(0)$ so that f maps $[0,\infty)$ onto $(0,M]$. Then f^{-1} exists and maps $(0,M]$ to $[0,\infty)$. (a) Show that f^{-1} is strictly decreasing. (b) Fix any $y \in (0,M]$ and show that f^{-1} is continuous at y. *Hints:* Theorem 6.43, Proposition 6.30, and sequential compactness (The Bolzano–Weierstrass Theorem). You may also use the following result.

> **Lemma.** *To show $x_n \to x$, it suffices to show that for every subsequence x_{n_k}, there is a sub-subsequence $x_{n_{k_\ell}}$ that converges to x.*

62. When (X,ρ) is a metric space and $f: X \to [-\infty,\infty]$ is an extended-real-valued function, the notions of **lower semicontinuity** and **upper semicontinuity** are often useful. We say that f is **lower semicontinuous** at x_0 if whenever t is a real number with $f(x_0) > t$, then for all x close to x_0, we have $f(x) > t$ as well. We say that f is **upper semicontinuous** at x_0 if the inequalities $f(x_0) > t$ and $f(x) > t$ are changed to $f(x_0) < t$ and $f(x) < t$.

(a) If $f(x_0) = -\infty$, is f lower semicontinuous at x_0?

(b) On $X := [0,\infty)$, consider the function $f(x) := -\ln x$ for $x > 0$ and $f(0) := \infty$. Determine whether or not f is lower semicontinuous at $x = 0$.

(c) Show that f is lower semicontinuous at all $x \in X$ if and only if for each real t, the **level set** $\{x \in X : f(x) \le t\}$ is closed.

(d) Show that f is lower semicontinuous at $x_0 \Leftrightarrow$ if whenever $x_n \to x_0$ and $f(x_n) \to L$, we must have $f(x_0) \le L$. Note that L is allowed to be an extended real number. *Hint:* To prove $\Leftarrow$, you may use the fact that if $f(x_n)$ is a sequence, then there is a subsequence $f(x_{n_k}) \to L := \underline{\lim}_{n\to\infty} f(x_n)$, even if $\underline{\lim}_{n\to\infty} f(x_n) = -\infty$.

(e) Let D be a sequentially compact subset of X on which f is lower semicontinuous. Show that f achieves its minimum value on D.

(f) The **epigraph** of f was defined in Problem 6.52. Show that f is lower semicontinuous on X if and only if epi f is closed in $X \times \mathbb{R}$.

CHAPTER 7
Diagonalization and Singular-Value Decomposition of Linear Operators

7.1. Bounded Linear Functionals

The simplest linear operators are scalar valued and are called **linear functionals**. A linear functional defined on a normed vector space X is said to be **bounded** if there is a finite constant B such that

$$|f(x)| \le B\|x\|, \quad \text{for all } x \in X.$$

Observe that for any $x, x_0 \in X$,

$$|f(x) - f(x_0)| = |f(x - x_0)| \le B\|x - x_0\|.$$

We thus see that a bounded linear functional is uniformly continuous.

Remark. It is important to realize that the term "bounded linear functional" does *not* mean that $|f(x)|$ is bounded by a finite constant. Instead, what is bounded by a finite constant is the ratio $|f(x)|/\|x\|$ for nonzero x.

Proposition 7.1. *If a linear functional on a normed vector space is continuous at a point, then the linear functional is bounded (and thus uniformly continuous).*

Proof. Suppose a linear functional f is continuous at a point x_0. Fix any $\varepsilon > 0$. Let $\|x - x_0\| < \delta$ imply $|f(x) - f(x_0)| < \varepsilon$. Fix any $0 < \eta < \delta$. We claim that for all $z \in X$, $|f(z)| \le (\varepsilon/\eta)\|z\|$. For $z = 0$, the inequality obviously holds. For $z \neq 0$, put $w := \eta z/\|z\|$ so that

$$\|(w + x_0) - x_0\| = \|w\| = \eta < \delta \quad \text{and then} \quad |f(w + x_0) - f(x_0)| = |f(w)| < \varepsilon.$$

But, since

$$\varepsilon > |f(w)| = \left| \frac{\eta}{\|z\|} f(z) \right|,$$

we have

$$|f(z)| < \frac{\varepsilon}{\eta}\|z\|. \qquad \square$$

Example 7.2. Let $C[0,1]$ denote the set of all real-valued, continuous waveforms on $[0,1]$. If we equip $C[0,1]$ with the **uniform norm**

$$\|x\| := \max_{0\le t\le 1} |x(t)|,$$

show that the **point-evaluation linear functional** defined by $f(x) := x(0)$ is continuous.

Solution. First note that by Theorem 6.44, the maximum in the definition of $\|x\|$ is achieved by some $t \in [0,1]$. Hence, $\|x\|$ is well defined. Second, by Problem 7.1, the formula for $\|\cdot\|$ satisfies the properties of a norm on $C[0,1]$. It is easy to show that f is uniformly continuous. For $x, y \in C[0,1]$, write

$$|f(x) - f(y)| = |x(0) - y(0)| \le \max_{0\le t\le 1} |x(t) - y(t)| = \|x - y\|.$$

We emphasize that whether or not a mapping $f: X \to Y$ between metric spaces is continuous depends on the metric used on the space X and the metric used on the space Y.

Example 7.3. In the preceding example, if we replace the uniform norm by

$$\|x\| := \int_0^1 |x(t)|\,dt,$$

then the point-evaluation linear functional is *not* continuous. See Problem 7.2.

7.1.1. Linear Functionals Represented by Inner Products

Given any vector y in an inner-product space X, if we put $f(x) := \langle x, y\rangle$, then it is easy to check that f is a bounded linear functional. The linearity follows from the properties of the inner product. To establish boundedness, use the Cauchy–Schwarz inequality to write $|f(x)| = |\langle x, y\rangle| \le \|x\|\,\|y\|$. Furthermore, the representing vector y is unique. To see this, suppose $f(x) = \langle x, z\rangle$ as well. Then $0 = f(x) - f(x) = \langle x, y\rangle - \langle x, z\rangle = \langle x, y - z\rangle$. Since this holds for all $x \in X$, taking $x = y - z$ shows that $\|y - z\|^2 = 0$, which implies $y = z$.

Example 7.4. It is easy to show that if X is a finite-dimensional inner-product space, then *every* linear functional is given by an inner product (and is therefore bounded and continuous). To see this, let $w_1, \ldots, w_n$ be an orthonormal basis for X. Since $x = \sum_{k=1}^n \langle x, w_k\rangle w_k$,

$$f(x) = f\left(\sum_{k=1}^n \langle x, w_k\rangle w_k\right) = \sum_{k=1}^n \langle x, w_k\rangle f(w_k) = \left\langle x, \sum_{k=1}^n \overline{f(w_k)}\, w_k \right\rangle.$$

Hence, the representing vector is $\sum_{k=1}^{n} \overline{f(w_k)} w_k$.

In the infinite-dimensional setting we have seen that not all linear functionals are continuous. However, in a Hilbert space, every bounded linear functional can be represented by an inner product. This result is known as the Riesz Representation Theorem for Hilbert Space.

Theorem 7.5 (Riesz Representation for Hilbert Space). *Every bounded linear functional on a Hilbert space can be represented uniquely by an inner product.*

Proof. Let X be a Hilbert space, and let f be a bounded linear functional on X. We show there is a unique vector $y \in X$ such that $f(x) = \langle x, y \rangle$ for all $x \in X$. Since f is bounded, it is continuous, and it is easy to show that $\ker f$ is a closed subspace (Problem 7.4). By the Projection Theorem for Hilbert Space,

$$X = (\ker f) \oplus (\ker f)^{\perp}.$$

If $f(x) = 0$ for all $x \in X$, take $y = 0$. Otherwise, on account of the above decomposition of X, there is a $z \in (\ker f)^{\perp}$ with $f(z) \neq 0$. We claim that

$$y = \frac{\overline{f(z)}}{\|z\|^2} z$$

is the desired vector. To prove this, given any $x \in X$, put

$$w := x - \frac{f(x)}{f(z)} z.$$

Then $f(w) = 0$, and so $w \in \ker f$. Since $z \in (\ker f)^{\perp}$, $\langle w, z \rangle = 0$. Replacing w by its definition yields

$$\left\langle x - \frac{f(x)}{f(z)} z, z \right\rangle = 0, \quad \text{which implies} \quad \langle x, z \rangle = \frac{f(x)}{f(z)} \|z\|^2.$$

Solving for $f(x)$ yields

$$f(x) = \left\langle x, \frac{\overline{f(z)}}{\|z\|^2} z \right\rangle$$

as claimed. Uniqueness of such a representation was established earlier in the section. □

If f is a bounded linear functional defined on a normed vector space, we define the **norm** of f by[a]

$$\|f\| := \sup_{x \neq 0} \frac{|f(x)|}{\|x\|}. \tag{7.1}$$

The assumption that f is bounded implies that the above supremum is finite. Also, the definition of $\|f\|$ implies that for nonzero x,

$$\|f\| \geq \frac{|f(x)|}{\|x\|},$$

which implies $|f(x)| \leq \|f\| \, \|x\|$, which also holds when $x = 0$. (Why?)

One way to find the norm in (7.1) is the following two-step procedure. First find a finite constant B such that $|f(x)| \leq B\|x\|$ for all x; this shows that the supremum is finite and upper bounded by B; thus, $\|f\| \leq B$. Second, find a nonzero vector x_0 such that $|f(x_0)| \geq B\|x_0\|$. This shows that the supremum is lower bounded by B. We conclude that the supremum is equal to B; i.e., $\|f\| = B$. You may use this technique in Problem 7.6 to show that in an inner-product space, if $f(x) := \langle x, y \rangle$ for some fixed y, then $\|f\| = \|y\|$; i.e., for such linear functionals, the norm of the functional f is equal to the vector-space norm of the vector y used to define f.

It is easy to check that the definition of $\|f\|$ is equivalent to the formula $\|f\| = \sup_{\|x\|=1} |f(x)|$.

7.2. Bounded Linear Operators

A linear operator $A: X \to Y$ between normed vector spaces is **bounded** if there is a finite constant B such that

$$\|Ax\|_Y \leq B\|x\|_X, \quad \text{for all } x \in X.$$

We included the subscripts on the norms to emphasize the different spaces involved. In the sequel, we usually drop the subscripts.

It is easy to show that a linear operator is continuous at a point if and only if the operator is bounded, in which case it is uniformly continuous.

[a] The reader should check that the collection of bounded linear functionals on X is a vector space (called the **dual** of X and denoted by X^*) and that (7.1) satisfies the properties of a norm on this vector space.

If A is a bounded linear operator, we define the **norm** of A by[b]

$$\|A\| := \sup_{x\neq 0} \frac{\|Ax\|_Y}{\|x\|_X}. \tag{7.2}$$

Hence, $\|Ax\|_Y \leq \|A\|\,\|x\|_X$. It is easy to check that the definition of $\|A\|$ is equivalent to the formula $\|A\| = \sup_{\|x\|=1} \|Ax\|$.

The suggestions following (7.1) for determining the norm of a bounded linear functional apply more generally to the determination of the norm of a bounded linear operator.

Example 7.6. If $\int_c^d \left(\int_a^b |k(t,\tau)|^2\,d\tau\right) dt < \infty$, then we can show that

$$(Ax)(t) := \int_a^b k(t,\tau)x(\tau)\,d\tau, \quad c \leq t \leq d,$$

defines a bounded operator from $L^2[a,b]$ into $L^2[c,d]$. First write

$$|(Ax)(t)| \leq \int_a^b |k(t,\tau)|\,|x(\tau)|\,d\tau.$$

Now apply **Hölder's inequality** (Problem 6.12) to get

$$|(Ax)(t)| \leq \left(\int_a^b |k(t,\tau)|^2\,d\tau\right)^{1/2} \left(\int_a^b |x(\tau)|^2\,d\tau\right)^{1/2}.$$

It follows that

$$\begin{aligned}\int_c^d |(Ax)(t)|^2\,dt &\leq \int_c^d \left(\int_a^b |k(t,\tau)|^2\,d\tau\right)\left(\int_a^b |x(\tau)|^2\,d\tau\right) dt \\ &= \int_c^d \left(\int_a^b |k(t,\tau)|^2\,d\tau\right) \|x\|^2\,dt \\ &= \int_c^d \left(\int_a^b |k(t,\tau)|^2\,d\tau\right) dt\,\|x\|^2.\end{aligned}$$

Hence, $x \in L^2[a,b] \Rightarrow Ax \in L^2[c,d]$. Since the right-hand side is finite, A is a bounded operator. Furthermore, $\|A\| \leq \left(\int_c^d \left(\int_a^b |k(t,\tau)|^2\,d\tau\right) dt\right)^{1/2}$.

[b] The reader should verify that the set of bounded linear operators from X to Y is a vector space and that (7.2) satisfies the properties of a norm on this vector space.

We have seen many examples of linear operators whose adjoint is easy to find. However, if we want to talk about operators in general without a specific one in mind, it would be nice to know that the adjoint exists.

Theorem 7.7. *A bounded linear operator from a Hilbert space into an inner-product space always has an adjoint.*

Proof. Let $A: X \to Y$ be a bounded linear operator from a Hilbert space X to an inner-product space Y. Given $y \in Y$, we must find a vector A^*y that satisfies

$$\langle Ax, y\rangle_Y = \langle x, A^*y\rangle_X, \quad \text{for all } x \in X.$$

Put $f(x) := \langle Ax, y\rangle_Y$, and observe that since

$$|f(x)| = |\langle Ax, y\rangle_Y| \le \|Ax\|_Y \|y\|_Y \le \|A\|\|x\|_X \|y\|_Y = \big(\|A\| \|y\|_Y\big)\|x\|_X,$$

f is a bounded linear functional on X. By the Riesz Representation Theorem for Hilbert Space, there is a unique representing vector in X. This vector depends on both the operator A and the point $y \in Y$ that we started with. Therefore, we denote the representing vector by A^*y. Thus, $f(x) = \langle x, A^*y\rangle_X$ for all $x \in X$. □

Proposition 7.8. *Let A be a linear operator between inner product spaces, and assume the adjoint A^* exists. Then either A and A^* are both bounded or they are both unbounded. If they are bounded, they have the same norm; i.e., $\|A^*\| = \|A\|$. Furthermore, in the bounded case, $\|A^*A\| = \|A\|^2$.*

Proof. Suppose A is bounded. Then

$$\|A^*y\|^2 = \langle A^*y, A^*y\rangle = |\langle AA^*y, y\rangle| \le \|AA^*y\| \|y\| \le \|A\| \|A^*y\| \|y\|.$$

Dividing both sides by $\|A^*y\|$ shows that A^* is bounded and satisfies $\|A^*\| \le \|A\|$. Similarly, if A^* is bounded, then

$$\|Ax\|^2 = \langle Ax, Ax\rangle = |\langle A^*Ax, x\rangle| \le \|A^*Ax\| \|x\| \le \|A^*\| \|Ax\| \|x\|, \tag{7.3}$$

and so $\|A\| \le \|A^*\|$.

Having shown that $\|A^*\| = \|A\|$, it follows that $\|A^*Ax\| \le \|A^*\| \|Ax\| \le \|A\|^2 \|x\|$. Hence, $\|A^*A\| \le \|A\|^2$. Conversely,

$$\|Ax\|^2 = \langle Ax, Ax\rangle = |\langle A^*Ax, x\rangle| \le \|A^*Ax\| \|x\| \le \|A^*A\| \|x\|^2 \tag{7.4}$$

shows that $\|A\|^2 \le \|A^*A\|$. □

Remark. The reader should be sure to understand the differences in the right-most inequalities in (7.3) and (7.4) and the assumptions under which each one holds.

We next have the following extension of Theorem 4.13(e)(f).

Theorem 7.9. *Let X and Y be inner-product spaces. If $A:X \to Y$ is a linear operator whose adjoint $A^*:Y \to X$ exists, then*

(a) $(\ker A)^\perp = \overline{\text{range}A^*}$, if X is a Hilbert space.
(b) $(\ker A^*)^\perp = \overline{\text{range}A}$, if Y is a Hilbert space.

Proof. To prove part (a), use Theorem 4.13(c) to write $\ker A = (\text{range}A^*)^\perp$. Hence,

$$(\ker A)^\perp = [(\text{range}A^*)^\perp]^\perp = \overline{\text{range}A^*},$$

where the last step follows by applying Problem 7.14 to the Hilbert space X and its subspace $\text{range}A^*$.

To prove part (b), use Theorem 4.13(a) to write $\ker A^* = (\text{range}A)^\perp$. Hence,

$$(\ker A^*)^\perp = [(\text{range}A)^\perp]^\perp = \overline{\text{range}A},$$

where the last step follows by applying Problem 7.14 to the Hilbert space Y and its subspace $\text{range}A$. □

7.2.1. Convolution Operators

We say that

$$(Ax)(t) = (h*x)(t) := \int h(t-\tau)x(\tau)\,d\tau$$

is a **convolution operator** or a **linear, time-invariant system**. But we have to be more precise. First, we must specify the range of τ over which $x(\tau)$ is defined. Of course, if $x(\tau)$ is defined only over a finite interval, we can define $x(\tau)$ to be zero outside that interval. The second thing we must specify is the range of t that we are interested in. In addition to specifying the ranges of τ and t, we must specify properties of x such as $x \in L^1$ or $x \in L^2$. And we must specify properties of $h*x$ such as $h*x \in L^2$. It is these different considerations that motivate the following results, which at first glance all seem to say nearly the same things.

Let us begin by reviewing a few facts about Fourier transforms.

(*i*) Suppose $x \in L^1(\mathbb{R})$. Then $X(f) := \int_{-\infty}^{\infty} x(t)e^{-j2\pi ft}\,dt$ is well defined. Furthermore, if $X \in L^1(\mathbb{R})$, then $x(t) = \int_{-\infty}^{\infty} X(f)e^{j2\pi ft}\,df$. It is possible that $X \notin L^1(\mathbb{R})$; a simple example is provided by $x(t) = 1$ on $[-T,T]$ and $x(t) = 0$ elsewhere. Then $x \in L^1(\mathbb{R})$, but X is a sinc function, which is not in $L^1(\mathbb{R})$ by Example 7.13 below.

(*ii*) If $x \in L^2(\mathbb{R})$, then $X_n(f) := \int_{-n}^{n} x(t)e^{-j2\pi ft}\,dt$ belongs to $L^2(\mathbb{R})$ and is Cauchy. Since $L^2(\mathbb{R})$ is complete, there is a limit $X \in L^2(\mathbb{R})$. This limit X is the Fourier transform of x. Furthermore, using this X, if we put $x_n(t) := \int_{-n}^{n} X(f)e^{j2\pi ft}\,df$, then $x_n \in L^2(\mathbb{R})$, and $\|x_n - x\| \to 0$. If we had started with $x \in L^2(\mathbb{R}) \cap L^1(\mathbb{R})$, then the Fourier transform X obtained by the foregoing procedure is equal to the well-defined integral $\int_{-\infty}^{\infty} x(t)e^{-j2\pi ft}\,dt$.

(*iii*) If $x, y \in L^2(\mathbb{R})$ have Fourier transforms $X, Y \in L^2(\mathbb{R})$, then **Parseval's equation** tells us that[c]

$$\int_{-\infty}^{\infty} x(t)\overline{y(t)}\,dt = \int_{-\infty}^{\infty} X(f)\overline{Y(f)}\,df.$$

Writing these integrals as inner products, we have

$$\langle x, y\rangle = \langle X, Y\rangle.$$

Hence, the Fourier transform is an **isometry** from $L^2(\mathbb{R})$ to $L^2(\mathbb{R})$; furthermore, since the Fourier transform is invertible, it is onto and therefore **unitary** (see Problem 7.16).

Theorem 7.10. *Suppose that a linear, time-invariant system has an impulse response h with finite energy and whose transform H is bounded. Then the system is a bounded linear operator from $L^2(\mathbb{R})$ into $L^2(\mathbb{R})$.*

Proof. Let h and x belong to $L^2(\mathbb{R})$ and have corresponding transforms H and X. Then

$$\begin{aligned}(h*x)(t) &:= \int_{-\infty}^{\infty} h(t-\tau)x(\tau)\,d\tau = \int_{-\infty}^{\infty} x(\tau)\overline{\overline{h(t-\tau)}}\,d\tau \\ &= \langle x, \overline{h(t-\cdot)}\rangle = \left\langle X, \overline{H(\cdot)e^{j2\pi\cdot t}}\right\rangle = \int_{-\infty}^{\infty} H(f)X(f)e^{j2\pi ft}\,df.\end{aligned}$$

Since H and X are in $L^2(\mathbb{R})$, their product is in $L^1(\mathbb{R})$ by Hölder's inequality. Since we have also assumed H is bounded, $HX \in L^2(\mathbb{R})$. Thus, $HX \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$. Since the above equation shows that $h*x$ is the inverse Fourier transform of $HX \in L^2(\mathbb{R})$, we have $h*x \in L^2(\mathbb{R})$. To show that the operator is bounded, suppose $|H(f)| \le B$. Then

$$\|h*x\|^2 = \langle h*x, h*x\rangle = \langle HX, HX\rangle \le B^2\|X\|^2 = B^2\|x\|^2,$$

where the second and last equalities are Parseval's. □

[c] Since both integrands are in $L^1(\mathbb{R})$ by Hölder's inequality, both integrals are well defined.

Example 7.11. Consider an **ideal lowpass filter**. Its transfer function has finite energy and is bounded. Thus, Theorem 7.10 applies; i.e., an ideal lowpass filter maps $L^2(\mathbb{R})$ into $L^2(\mathbb{R})$.

Example 7.12 (Stable Systems). Recall that a linear, time-invariant system is **stable** if its impulse response h satisfies $\int_{-\infty}^{\infty} |h(t)|\,dt < \infty$. In other words, the system is stable if its impulse response $h \in L^1(\mathbb{R})$. It is easy to see that the transfer function of a stable system is bounded. Just write

$$|H(f)| = \left| \int_{-\infty}^{\infty} h(t) e^{-j2\pi ft}\,dt \right| \le \int_{-\infty}^{\infty} \left| h(t) e^{-j2\pi ft} \right| dt = \int_{-\infty}^{\infty} |h(t)|\,dt < \infty.$$

We also note that a stable system maps $L^p(\mathbb{R})$ into $L^p(\mathbb{R})$ and $\|h * x\|_p \le \|h\|_1 \|x\|_p$ by **Young's inequality** [14].

Example 7.13. An ideal lowpass filter is not a stable system. If h were in $L^1(\mathbb{R})$, then as shown in the Notes,[1] H would be continuous. But the lowpass-filter transfer function is not continuous at its cutoff frequency. Alternatively, since h is a sinc function, write

$$\begin{aligned} \int_0^{\infty} |\operatorname{sinc}(t)|\,dt = \sum_{k=0}^{\infty} \int_k^{k+1} \left| \frac{\sin(\pi t)}{\pi t} \right| dt &\ge \sum_{k=0}^{\infty} \frac{1}{\pi(k+1)} \int_k^{k+1} |\sin(\pi t)|\,dt \\ &= \frac{2}{\pi^2} \sum_{k=0}^{\infty} \frac{1}{k+1} = \infty, \end{aligned}$$

where the last step follows by noting that $\sum_{k=1}^{2^n} (1/k) \ge 1 + n/2$.

Corollary 7.14. *As in Theorem 7.10, consider a linear, time-invariant system with finite-energy impulse response h whose transform H is bounded. Then the system can be viewed as a mapping from $L^2[a,b]$ into $L^2(\mathbb{R})$. The system can also be viewed as mapping $L^2[a,b]$ into $L^2[c,d]$.*

Proof. To prove the first statement, treat $L^2[a,b]$ as a subspace of $L^2(\mathbb{R})$ by defining $x(t) := 0$ for $t \notin [a,b]$. For the second statement, put $y(t) := (h * x)(t)$ for $c \le t \le d$. Then observe that

$$\int_c^d |y(t)|^2\,dt = \int_c^d |(h*x)(t)|^2\,dt \le \int_{-\infty}^{\infty} |(h*x)(t)|^2\,dt < \infty. \qquad \square$$

The conditions of Corollary 7.14 guarantee that $h * x \in L^2(\mathbb{R})$, even if we only need $h * x \in L^2[c,d]$. Consider the transfer function $H(f) := 1/\sqrt{1-f^2}$ for $|f| < 1$ and $H(f) := 0$ for $|f| \geq 0$. Then $H \in L^1(\mathbb{R})$, but $H \notin L^2(\mathbb{R})$. If we put $h(t) := \int_{-\infty}^{\infty} H(f) e^{j2\pi ft}\, df$, then $h \notin L^1(\mathbb{R})$.[d] Even though the preceding results do not apply, the following theorem shows that convolution with h is a bounded linear operator from $L^2[a,b]$ into $L^2[c,d]$.

Theorem 7.15. *A linear, time-invariant system whose transfer function H satisfies $\int_{-\infty}^{\infty} |H(f)|\, df < \infty$ is a bounded linear operator from $L^2[a,b]$ into $L^2[c,d]$.*

Proof. The argument used in Example 7.12 shows that $h(t) := \int_{-\infty}^{\infty} H(f) e^{j2\pi ft}\, df$ is bounded. Hence, $k(t,\tau) := h(t-\tau)$ for $t \in [c,d]$ and $\tau \in [a,b]$ satisfies the condition of Example 7.6. □

7.2.2. Some Nonsingular Convolution Operators

Theorem 7.16. *As in Theorem 7.10, consider a linear, time-invariant system with finite-energy impulse response h whose transform H is bounded. If in addition $H(f) \neq 0$ for all f, then the system is a nonsingular mapping from $L^2(\mathbb{R})$ into $L^2(\mathbb{R})$.*

Proof. If $h * x$ is the zero waveform, then its transform HX is the zero frequency function. Since $H(f) \neq 0$ for all f, X must be the zero frequency function, which implies x is the zero waveform. □

The preceding theorem does not apply to an ideal lowpass filter. However, the next result implies that an ideal lowpass filter applied to time-limited waveforms is a nonsingular mapping from $L^2[a,b]$ into $L^2(\mathbb{R})$.

Theorem 7.17. *As in Theorem 7.10, consider a linear, time-invariant system with finite-energy impulse response h whose transform H is bounded. In addition, assume H is nonnegative and that there is some open interval (f_1, f_2) on which H is strictly positive and continuous. Then the system is a nonsingular mapping from $L^2[a,b]$ into $L^2(\mathbb{R})$.*

Proof. Assume $Ax = 0$. Using this and the nonnegativity of H,

$$0 = \langle Ax, x\rangle = \langle h*x, x\rangle = \langle HX, X\rangle = \int_{-\infty}^{\infty} H(f)|X(f)|^2\, df \geq \int_{f_1}^{f_2} H(f)|X(f)|^2\, df \geq 0.$$

[d] Suppose otherwise that $h \in L^1(\mathbb{R})$. Then by the argument in Example 7.12, $H(f)$ would be bounded, which contradicts the fact that $H(f) \to \infty$ as $|f| \to 1$.

Hence, $\int_{f_1}^{f_2} H(f)|X(f)|^2\,df = 0$. By hypothesis, H is continuous and positive on (f_1, f_2). Furthermore, by Problem 7.19, X is continuous too. The only way that a nonnegative continuous function can integrate to zero is for the integrand to be identically zero.[2] Since H is strictly positive, we must have $X(f) = 0$ on (f_1, f_2). By Problem 7.20, this implies $X(f) = 0$ *for all* f, which implies $x(t)$ is the zero waveform. □

Corollary 7.18. *Under the assumptions of Theorem* 7.17*, the system is a nonsingular mapping from* $L^2[a,b]$ *into* $L^2[a,b]$.

Proof. Proceed as in the previous proof, except that $\langle Ax, x\rangle$ is viewed as an inner product on $L^2[a,b]$, while $\langle h*x, x\rangle$ is viewed as an inner product on $L^2(\mathbb{R})$. Since the two inner products are the same, the rest of the proof goes through as before. □

Here is the analogous result under the conditions of Theorem 7.15.

Theorem 7.19. *As in Theorem 7.15, consider a linear, time-invariant system whose transfer function* H *satisfies* $\int_{-\infty}^{\infty} |H(f)|\,df < \infty$*. hold. In addition, assume* H *is nonnegative and that there is some open interval* (f_1, f_2) *on which* H *is strictly positive and continuous.*[e] *Then the system is a nonsingular mapping from* $L^2[a,b]$ *into* $L^2[a,b]$.

Proof. The proof is nearly the same as that for Theorem 7.17 and Corollary 7.18. The difference is that since we do not know if $h*x \in L^2(\mathbb{R})$, we cannot use Parseval's equation to show that $\langle Ax, x\rangle = \int_{-\infty}^{\infty} H(f)|X(f)|^2\,df$. Instead, for $x \in L^2[a,b]$, write

$$\begin{aligned}
\langle Ax, x\rangle &= \int_a^b (Ax)(t)\,\overline{x(t)}\,dt \\
&= \int_a^b \left[\int_a^b h(t-\tau)x(\tau)\,d\tau\right]\overline{x(t)}\,dt \\
&= \int_a^b \left[\int_a^b \left[\int_{-\infty}^{\infty} H(f)e^{j2\pi f(t-\tau)}\,df\right]x(\tau)d\tau\right]\overline{x(t)}\,dt \\
&= \int_{-\infty}^{\infty} H(f)|X(f)|^2\,df,
\end{aligned}$$

where changing the order of integration in the last equality is justified by Tonelli's Theorem and Fubini's Theorem [6], [14], [33], [34] on account of the fact that $H \in L^1(\mathbb{R})$ and $x \in L^2[a,b] \subset L^1[a,b]$. □

[e] The assumption that there is an open interval on which H is strictly positive and continuous can be replaced by the assumption that $\int_{-\infty}^{\infty} H(f)\,df > 0$ [47, Appendix A].

7.3. Eigenvalues

Let X be a vector space, and let $A: X \to X$ be a linear operator. If for some *nonzero* vector x and some scalar λ, $Ax = \lambda x$, we say that λ is an **eigenvalue** of A, and x is an **eigenvector** of A. We say (λ, x) is an **eigenpair** of A. The **eigenspace** corresponding to an eigenvalue λ is the set

$$\{x \in X : Ax = \lambda x\},$$

which is simply $\ker(\lambda I - A)$.

Example 7.20 (Eigenvalues and Operator Norms). If a bounded operator A has an eigenvalue λ, then $|\lambda| \leq \|A\|$. If $Ax = \lambda x$, we can write

$$|\lambda| \, \|x\| = \|\lambda x\| = \|Ax\| \leq \|A\| \, \|x\|.$$

Since eigenvectors are nonzero, we can divide by $\|x\|$ to obtain $|\lambda| \leq \|A\|$. In particular, suppose $X = \mathbb{C}^n$ and A is given by an $n \times n$ matrix with entries a_{ij}. Using the 1-norm in $\mathbb{C}^n$ turns $|\lambda| \leq \|A\|$ into

$$|\lambda| \leq \max_{1 \leq j \leq n} \sum_{i=1}^{n} |a_{ij}|, \quad \text{by Problem 7.10(a),}$$

while using the infinity norm on $\mathbb{C}^n$ yields

$$|\lambda| \leq \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}|, \quad \text{by Problem 7.10(b).}$$

Example 7.21. Let X be an arbitrary vector space with two subspaces U and V such that $X = U \oplus V$. In other words, every $x \in X$ has the form $x = u + v$ for unique $u \in U$ and unique $v \in V$. Let $Px := u$.

There are three important properties of P.

- Given $u \in U$, observe that the formula $u = u + 0$ is the unique way of writing u as the sum of a vector in U and a vector in V. Hence, $Pu = u$.
- Given $v \in V$, writing $v = 0 + v$ shows that $Pv = 0$.
- Writing $x = u + v$, we have $Px = u$, and then $P(Px) = P(u) = u = Px$. In other words, $P^2 = P$; i.e., P is **idempotent**.

The first property of P shows that every nonzero $u \in U$ is an eigenvector of P with eigenvalue $\lambda = 1$. The second property shows that every nonzero $v \in V$ is an eigenvector with eigenvalue $\lambda = 0$. Are there any other eigenvalues of P? The third property implies that the answer is "No." To see this, suppose $Px = \lambda x$. Now write

$$\lambda x = Px = P(Px) = P(\lambda x) = \lambda Px = \lambda(\lambda x) = \lambda^2 x.$$

It follows that $\lambda(1-\lambda)x = 0$. Since eigenvectors are nonzero, we must have either $\lambda = 0$ or $\lambda = 1$. There are no other possible eigenvalues for P.

We now turn to the question of whether there are any other eigenvectors besides those in U and those in V. Since $Px \in U$ for all x, if x is an eigenvector with eigenvalue $\lambda = 1$, then $Px = x$ implies $x \in U$. On the other hand, if x is an eigenvector with eigenvalue $\lambda = 0$, then $Px = 0$. But by definition of P, we must have started with $x = 0 + v$, which implies $x \in V$.

Example 7.22. Let $X = \mathbb{C}^n$, and let

$$a := \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}.$$

If we define $Ax := ax$ for $x \in \mathbb{C}^n$, then each λ_i is an eigenvalue of A, and $e_i := [0,\dots,1,\dots,0]^\mathsf{T}$, where the 1 is in the ith position, is a corresponding eigenvector.

Example 7.23. Let X be a Hilbert space, and let $\{\varphi_k\}_{k=1}^\infty$ be an orthonormal set of vectors in X. Let $\{\lambda_k\}_{k=1}^\infty$ be any *bounded* sequence of *nonzero* scalars. Consider the operator[f]

$$Ax := \sum_{k=1}^{\infty} \lambda_k \langle x, \varphi_k \rangle \varphi_k. \tag{7.5}$$

Notice that $A\varphi_i = \lambda_i \varphi_i$; i.e., the (λ_i, φ_i) are eigenpairs of A. **In fact, the λ_k are the *only* possible nonzero eigenvalues of A.** Suppose $Ax = \lambda x$ for some nonzero λ and nonzero x. Then

$$x = \frac{1}{\lambda} \sum_{k=1}^{\infty} \lambda_k \langle x, \varphi_k \rangle \varphi_k,$$

which implies $x \in \overline{\operatorname{span}\{\varphi_k\}}$. Hence, x is equal to its projection onto $\overline{\operatorname{span}\{\varphi_k\}}$, and so

$$x = \sum_{k=1}^{\infty} \langle x, \varphi_k \rangle \varphi_k$$

by Theorem 6.35. Since the coefficients of any expansion in the φ_k must be unique (why?), it follows that for all k with $\langle x, \varphi_k \rangle \neq 0$, we have $\lambda = \lambda_k$. There must be at least one such k because $x \neq 0$. If there is more than one such k, then λ is a repeated

[f] Since the λ_k are bounded, we have by Theorem 6.35 that the sum in (7.5) converges; i.e., A is well defined.

eigenvalue. Thus, if λ is a nonzero eigenvalue of an operator A of the form (7.5), the **eigenspace** corresponding to λ is

$$\overline{\operatorname{span}\{\varphi_k : \lambda_k = \lambda\}}.$$

The action of the matrix operator of Example 7.22 can be put in the form in (7.5) by observing that

$$ax = \begin{bmatrix} \lambda_1 \\ \vdots \\ 0 \end{bmatrix} x_1 + \cdots + \begin{bmatrix} 0 \\ \vdots \\ \lambda_n \end{bmatrix} x_n = \sum_{k=1}^{n} \lambda_k x_k e_k = \sum_{k=1}^{n} \lambda_k \langle x, e_k \rangle e_k.$$

For this reason, we say that an operator of the form (7.5) is **diagonalizable**. In other words, an operator A is diagonalizable if there is an orthonormal sequence of eigenpairs such that (7.5) holds.

Example 7.24. Not every operator can be put into the form in (7.5). Consider solving the eigenvalue problem

$$\begin{bmatrix} a & b \\ 0 & a \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \lambda \begin{bmatrix} x \\ y \end{bmatrix},$$

where $a \neq 0$, $b \neq 0$ and $[x,y]^\mathsf{T} \neq [0,0]^\mathsf{T}$. To begin the analysis, we write the matrix-vector equation as the two scalar equations

$$\begin{aligned} ax + by &= \lambda x \\ ay &= \lambda y. \end{aligned}$$

We claim that the only possible solution requires $y = 0$. Suppose otherwise that $y \neq 0$. Then the second equation implies $\lambda = a$. Using this in the first equation yields $ax + by = ax$. This implies $by = 0$. Since $b \neq 0$, we must have $y = 0$, contradicting our assumption that $y \neq 0$. Now, if $y = 0$ in the two equations, we find in the first one that $ax = \lambda x$. Since $y = 0$ forces $x \neq 0$ to get an eigenvector, we conclude that $\lambda = a$. Also, every eigenvector is of the form $[x,0]^\mathsf{T} = x\varphi$, where $\varphi := [1,0]^\mathsf{T}$ is a unit vector. Now, an operator of the form $Ax = \lambda \langle x, \varphi \rangle \varphi$ has a one-dimensional range. However, the matrix operator of this example is clearly invertible and therefore has a two-dimensional range.

Proposition 7.25. *A self-adjoint linear operator A on an inner-product space has the following properties.*

(a) *The inner product $\langle Ax,x\rangle$ is real.*
(b) *If A has any eigenvalues, they must be real.*
(c) *If A is positive semidefinite and has any eigenvalues, they must be nonnegative.*
(d) *If A has two distinct eigenvalues, then their corresponding eigenvectors must be orthogonal.*

Proof. Problem 7.30. □

We now return to operators of the form

$$Ax = \sum_{k=1}^{\infty} \lambda_k \langle x, \varphi_k\rangle \varphi_k$$

for bounded, nonzero λ_k and orthonormal φ_k as discussed in Example 7.23. Observe that

$$\langle Ax,x\rangle = \left\langle \sum_{k=1}^{\infty} \lambda_k \langle x,\varphi_k\rangle \varphi_k, x \right\rangle = \sum_{k=1}^{\infty} \lambda_k \langle x,\varphi_k\rangle\langle \varphi_k, x\rangle = \sum_{k=1}^{\infty} \lambda_k \big|\langle x,\varphi_k\rangle\big|^2.$$

Taking absolute values shows that

$$\big|\langle Ax,x\rangle\big| \le \Big(\sup_{k\ge 1} |\lambda_k|\Big)\|x\|^2.$$

However, if A is self adjoint, then $\langle Ax,x\rangle$ is real as are its eigenvalues; hence, the λ_k are also real. In this case, we can write

$$\langle Ax,x\rangle = \sum_{k=1}^{\infty} \lambda_k \big|\langle x,\varphi_k\rangle\big|^2 \le \Big(\sup_{k\ge 1} \lambda_k\Big) \sum_{k=1}^{\infty} \big|\langle x,\varphi_k\rangle\big|^2 \le \Big(\sup_{k\ge 1} \lambda_k\Big)\|x\|^2.$$

Keep in mind that the λ_k could be positive or negative. For example, if $\lambda_k = -1/k$, then the supremum is zero.

7.4. Diagonalization (The Spectral Theorem)

The Spectral Theorem says that linear operators on a Hilbert space that are compact (to be defined later) and self adjoint ($A^* = A$) can always be diagonalized. As shown in Problem 7.44(b), finite-rank operators are compact. Hence, all $n \times n$ real

or complex matrix operators are compact. It follows that any real symmetric matrix or any complex Hermitian matrix is diagonalizable. In a later section, we derive the singular-value decomposition (SVD) (to be defined later) by diagonalizing A^*A.

There are two key ideas needed for the proof of the Spectral Theorem. The first uses the following Invariant Subspace Lemma.

Lemma 7.26 (Invariant Subspaces). *Let X be an inner-product space, and suppose $A: X \to X$ is a linear operator having adjoint A^*. If W is a subspace of X for which $A: W \to W$, then $A^*: W^\perp \to W^\perp$.*

Proof. Fix any $v \in W^\perp$. Set $y = A^*v$. Then $y \in W^\perp$ if and only if $\langle w, y\rangle = 0$ for all $w \in W$. But $\langle w, y\rangle = \langle w, A^*v\rangle = \langle Aw, v\rangle = 0$ since $Aw \in W$ and $v \in W^\perp$. □

When A is self adjoint, the lemma says that if $A: W \to W$, then $A: W^\perp \to W^\perp$. Now suppose $A\varphi_1 = \lambda_1\varphi_1$ for some nonzero vector φ_1. Take $W := \text{span}\{\varphi_1\}$. Then a typical $x \in W$ has the form $x = c\varphi_1$, and $Ax = A(c\varphi_1) = cA\varphi_1 = c\lambda_1\varphi_1 \in \text{span}\{\varphi_1\} = W$. By the lemma, $A: \text{span}\{\varphi_1\}^\perp \to \text{span}\{\varphi_1\}^\perp$. Now suppose we can find a nonzero $\varphi_2 \in \text{span}\{\varphi_1\}^\perp$ with $A\varphi_2 = \lambda_2\varphi_2$. Then take $W := \text{span}\{\varphi_1, \varphi_2\}$. Again $A: W \to W$, and $A: W^\perp \to W^\perp$. Suppose we can find a nonzero $\varphi_3 \in W^\perp = \text{span}\{\varphi_1, \varphi_2\}^\perp$ with $A\varphi_3 = \lambda_3\varphi_3 \ldots$. Continuing in this way we obtain an orthogonal sequence of eigenvectors and corresponding eigenvalues.

The problem with the foregoing analysis is that we have no way of proving that A has *any* eigenvectors at all! Hence, the second key to proving the Spectral Theorem consists of the next two lemmas that establish the *existence* of eigenpairs. In particular, we show that either $\|A\|$ or $-\|A\|$ is an eigenvalue if A is self adjoint and compact on a Hilbert space.

Recall that for a bounded linear operator, $\|A\| = \sup_{\|x\|=1} \|Ax\|$. Thus, there exists a sequence $\{x_n\}$ with $\|x_n\| = 1$ and $\|Ax_n\| \to \|A\|$. In other words, the x_n lie on the unit sphere in X, and the real numbers $\|Ax_n\|$ converge to the real number $\|A\|$. Now, in general, we cannot conclude that the vectors Ax_n converge. However, if the vector space X is finite dimensional, the unit sphere is sequentially compact, and therefore there is a subsequence $\{x_{n_k}\}$ and a point x_0 on the unit sphere such that $x_{n_k} \to x_0$. Since A is bounded, it is continuous and therefore convergence preserving. Hence, $Ax_{n_k} \to Ax_0$, from which it follows via the triangle inequality that $\|Ax_{n_k}\| \to \|Ax_0\|$. Since $\|Ax_{n_k}\| \to \|A\|$, we see that $\|Ax_0\| = \|A\|$. To prove the existence of an eigenvalue in the Spectral Theorem, we only need that for some subsequence $\{x_{n_k}\}$, Ax_{n_k} converge to some $y_0 \in X$; we do not need that x_{n_k} itself converge. This leads to the following definition of a compact linear operator.

Let X and Y be normed vector spaces. Let $A: X \to Y$ be a linear operator. Then A is a **compact operator** if:

> Whenever $\{x_n\}$ is a sequence in X with $\|x_n\| = 1$, then there is a subsequence $\{x_{n_k}\}$ and there is a vector $y_0 \in Y$ with $Ax_{n_k} \to y_0$.

There are several equivalent ways to say this. For example, a linear operator is compact if the image of every unit-norm sequence has a converging subsequence. A linear operator is compact if the image of every sequence on the unit sphere has a converging subsequence. For another equivalent statement, see Problem 7.38.

Proposition 7.27.

(a) A compact linear operator is bounded.

(b) A bounded linear operator of finite rank (finite-dimensional range) is compact. Hence, every matrix operator is compact.

Proof. Problem 7.44. □

It can be proved that the integral operator of Example 7.6 is compact [17].

Lemma 7.28 (Characterization of $\|A\|$). *If A is bounded and self-adjoint, then*

$$\|A\| = \sup_{\|x\|=1} \big|\langle Ax, x\rangle\big|.$$

Proof. Let $m := \sup_{\|x\|=1} \big|\langle Ax, x\rangle\big|$. Using the Cauchy–Schwarz inequality with $\|x\| = 1$ yields $\big|\langle Ax, x\rangle\big| \le \|A\|$. Hence, $m \le \|A\|$. We must show that $m \ge \|A\|$. Observe that $A^* = A$ implies that for all $x, y \in X$,

$$\langle A(x \pm y), x \pm y\rangle = \langle Ax, x\rangle \pm 2\,\mathrm{Re}\langle Ax, y\rangle + \langle Ay, y\rangle.$$

Hence,

$$4\,\mathrm{Re}\langle Ax, y\rangle = \langle A(x+y), x+y\rangle - \langle A(x-y), x-y)\rangle.$$

Using the definition of m and the parallelogram law,

$$4\big|\mathrm{Re}\langle Ax, y\rangle\big| \le m\big(\|x+y\|^2 + \|x-y\|^2\big) = 2m\big(\|x\|^2 + \|y\|^2\big).$$

Taking $y = \frac{\|x\|}{\|Ax\|} Ax$, $\langle Ax, y\rangle = \|Ax\|\,\|x\|$, and $\|y\| = \|x\|$; hence,

$$4\|Ax\|\,\|x\| \le 2m\big(\|x\|^2 + \|x\|^2\big) = 4m\|x\|^2.$$

Therefore, $\|Ax\|/\|x\| \le m$ for $x \ne 0$, and $\|A\| \le m$. □

Corollary 7.29. *Let A be bounded and self-adjoint. If $\langle Ax,x\rangle = 0$ for all $x \in X$, then $A = 0$.*

Lemma 7.30 (Existence of Eigenpairs). *Let A be a compact, self-adjoint operator with $\|A\| > 0$. Then either $\|A\|$ or $-\|A\|$ is an eigenvalue of A.*

Remark. By combining this result with Example 7.20, we see that if A is compact, self-adjoint, and positive semidefinite, then $\|A\|$ is the largest eigenvalue of A, and so

$$0 \le \langle Ax,x\rangle \le \|A\|\,\|x\|^2 = \lambda_{\max}(A)\|x\|^2.$$

Proof of Lemma 7.30. On account of Lemma 7.28, we may assume the existence of a sequence $\{x_n\}$ with $\|x_n\| = 1$ and $\big|\langle Ax_n,x_n\rangle\big| \to \|A\|$. Now, $A^* = A \Rightarrow \langle Ax,x\rangle \in \mathbb{R}$. Thus $\big\{\langle Ax_n,x_n\rangle\big\}$ is a bounded sequence of real numbers, and by the Bolzano–Weierstrass Theorem contains a converging subsequence. To simplify the notation, we assume, without loss of generality, that the sequence itself converges to some real number λ. Of course, $|\lambda| = \|A\|$; i.e., $\lambda = \pm\|A\|$. Next, write

$$\begin{aligned} 0 \le \|Ax_n - \lambda x_n\|^2 &= \|Ax_n\|^2 - 2\lambda\langle Ax_n,x_n\rangle + \lambda^2 \\ &\le \lambda^2 - 2\lambda\langle Ax_n,x_n\rangle + \lambda^2. \end{aligned}$$

It follows that $\|Ax_n - \lambda x_n\|^2 \to 0$. Since A is compact, there is a subsequence such that Ax_{n_k} converges to some $y \in X$. So,

$$\|y - \lambda x_{n_k}\| \le \|y - Ax_{n_k}\| + \|Ax_{n_k} - \lambda x_{n_k}\| \to 0.$$

So, $\lim_{k\to\infty} x_{n_k} = y/\lambda$. Then

$$y := \lim_{k\to\infty} Ax_{n_k} = A\Big(\lim_{k\to\infty} x_{n_k}\Big) = A(y/\lambda),$$

and $Ay = \lambda y$. It remains to prove that $y \neq 0$. This follows by noting that since $y/\lambda = \lim_{k\to\infty} x_{n_k}$, $\|y\| = \lim_{k\to\infty}\|\lambda x_{n_k}\| = |\lambda| = \|A\| > 0$. □

Theorem 7.31 (Spectral Theorem). *Let X be a Hilbert space, and let $A\colon X \to X$ be a linear, compact, self-adjoint operator with $\|A\| > 0$. Then there is a family of eigenpairs $\{(\lambda_k,\varphi_k)\}$, possibly finite, such that $|\lambda_1| = \|A\|$, $\lambda_k \neq 0$, the $\{\varphi_k\}$ are orthonormal, and if the family $\{(\lambda_k,\varphi_k)\}$ is infinite, $|\lambda_k| \searrow 0$. Furthermore,*[g]

$$Ax = \sum_{k=1}^{\infty} \lambda_k\langle x,\varphi_k\rangle\varphi_k, \quad \text{for all } x \in X, \tag{7.6}$$

$$\overline{\operatorname{span}\{\varphi_k\}} = \overline{\operatorname{range} A} = (\ker A)^{\perp}, \tag{7.7}$$

$$X = \ker A \oplus \overline{\operatorname{span}\{\varphi_k\}}. \tag{7.8}$$

Remark. When considering an operator to which the Spectral Theorem does not apply, the operator may have *no* eigenvalues at all. Or, if the operator has eigenvalues, it may happen that $\sup_k |\lambda_k| < \|A\|$. See Problem 7.29.

Remark. When A is nonsingular, (7.8) implies that $X = \overline{\operatorname{span}\{\varphi_k\}}$; i.e., the eigenvectors φ_k form a complete orthonormal set. Hence, every $x \in X$ is equal to its projection onto $\overline{\operatorname{span}\{\varphi_k\}}$ and therefore satisfies

$$x = \sum_{k=1}^{\infty} \langle x, \varphi_k \rangle \varphi_k \quad \text{and} \quad \|x\|^2 = \sum_{k=1}^{\infty} \big|\langle x, \varphi_k \rangle\big|^2.$$

This would be the case for any positive-definite operator like the ones in Corollary 7.18 and Theorem 7.19.

Proof. We first show that (7.6) $\Rightarrow$ (7.7) $\Rightarrow$ (7.8). We then establish the representation (7.6).

(7.6) $\Rightarrow$ (7.7): Clearly (7.6) implies $\operatorname{range} A \subset \overline{\operatorname{span}\{\varphi_k\}}$, which implies

$$\overline{\operatorname{range} A} \subset \overline{\operatorname{span}\{\varphi_k\}}.$$

Also, since $A(\varphi_k/\lambda_k) = \varphi_k$, $\operatorname{span}\{\varphi_k\} \subset \operatorname{range} A$, which implies

$$\overline{\operatorname{span}\{\varphi_k\}} \subset \overline{\operatorname{range} A}.$$

Thus, the first equality in (7.7) holds. To prove the second equality, observe that

$$\begin{aligned} (\ker A)^{\perp} &= \overline{\operatorname{range} A^*}, && \text{by Theorem 7.9(a),} \\ &= \overline{\operatorname{range} A}, && \text{since } A^* = A. \end{aligned}$$

(7.7) $\Rightarrow$ (7.8): Since A is compact, it is bounded and therefore continuous. Hence, $\ker A$ is a *closed* subspace of the Hilbert space X, and the Projection Theorem applies. Combining this with the Orthogonality Principle and (7.7) permits us to write

$$X = \ker A \oplus (\ker A)^{\perp} = \ker A \oplus \overline{\operatorname{span}\{\varphi_k\}}.$$

We now prove the existence of the eigenpairs and the representation (7.6). Let $X_1 := X$. By Lemma 7.30, A has an eigenvalue λ_1 with $|\lambda_1| = \|A\|$. Let φ_1 be a corresponding eigenvector with $\|\varphi_1\| = 1$. Let $X_2 := (\operatorname{span}\{\varphi_1\})^{\perp}$. Since $A\colon \operatorname{span}\{\varphi_1\} \to \operatorname{span}\{\varphi_1\}$, Lemma 7.26 implies $A\colon X_2 \to X_2$. By Lemma 7.30 applied to the restriction of A to X_2, there exists an eigenpair (λ_2, φ_2) with $\|\varphi_2\| = 1$ and

$$|\lambda_2| = \sup_{\substack{x \in X_2 \\ x \neq 0}} \frac{\|Ax\|}{\|x\|} \leq \sup_{\substack{x \in X_1 \\ x \neq 0}} \frac{\|Ax\|}{\|x\|} = |\lambda_1| = \|A\|,$$

[g] Once it is shown that the λ_k are bounded, the sum in (7.6) converges by Theorem 6.35.
If $\{(\lambda_k, \varphi_k)\}$ is a finite family, the sum in (7.6) is finite, and no closures appear in (7.7) and (7.8).

where the inequality follows because $X_2 \subset X_1$. Now suppose that the eigenpairs $(\lambda_1, \varphi_1), \ldots, (\lambda_{n-1}, \varphi_{n-1})$ have been found with $|\lambda_1| \geq \cdots \geq |\lambda_{n-1}| > 0$. Set

$$X_n := (\text{span}\{\varphi_1, \ldots, \varphi_{n-1}\})^{\perp}.$$

Note that $X_n \subset X_{n-1}$ and that $A: X_n \to X_n$. If the restriction of A to X_n is the zero operator, i.e., if

$$\sup_{\substack{x \in X_n \\ x \neq 0}} \frac{\|Ax\|}{\|x\|} = 0,$$

then we stop. Otherwise, there exists an eigenpair (λ_n, φ_n) such that $\varphi_n \in X_n = (\text{span}\{\varphi_1, \ldots, \varphi_{n-1}\})^{\perp}$, and with $|\lambda_n|$ equal to the norm of the restriction of A to X_n, and $0 < |\lambda_n| \leq |\lambda_{n-1}|$.

If this procedure never terminates, we claim that $|\lambda_n| \searrow 0$. Suppose otherwise. Then there is some $\varepsilon_0 > 0$ such that for every k, there is some $n_k \geq k$ with $|\lambda_{n_k}| > \varepsilon_0$. Since A is a compact operator, and since $\|\varphi_{n_k}\| = 1$, there is a further subsequence for which $\{A\varphi_{n_{k_i}}\}$ is convergent, and therefore Cauchy. Since $A\varphi_{n_{k_i}} = \lambda_{n_{k_i}} \varphi_{n_{k_i}}$, $\{\lambda_{n_{k_i}} \varphi_{n_{k_i}}\}$ is Cauchy. However, this contradicts the fact that for all i and j with $n_{k_i} \neq n_{k_j}$,

$$\|\lambda_{n_{k_i}} \varphi_{n_{k_i}} - \lambda_{n_{k_j}} \varphi_{n_{k_j}}\|^2 = \lambda_{n_{k_i}}^2 + \lambda_{n_{k_j}}^2 \geq 2\varepsilon_0^2 > 0.$$

Thus, $|\lambda_n| \searrow 0$.

We now derive the representation (7.6). Fix any $x \in X$. Set

$$\widehat{x}_n := \sum_{k=1}^{n-1} \langle x, \varphi_k \rangle \varphi_k.$$

It is easy to see that $\widetilde{x}_n := x - \widehat{x}_n$ is orthogonal to $\varphi_1, \ldots, \varphi_{n-1}$; i.e., $\widetilde{x}_n \in X_n = (\text{span}\{\varphi_1, \ldots, \varphi_{n-1}\})^{\perp}$. By the Orthogonality Principle, it follows that $\widehat{x}_n$ is the projection of x onto $\text{span}\{\varphi_1, \ldots, \varphi_{n-1}\}$. Also, since $x = \widehat{x}_n + \widetilde{x}_n$, we can write

$$\|x\|^2 = \|\widehat{x}_n\|^2 + \|\widetilde{x}_n\|^2 \geq \|\widetilde{x}_n\|^2.$$

Since $\widetilde{x}_n \in X_n$,

$$\|A\widetilde{x}_n\| \leq |\lambda_n| \, \|\widetilde{x}_n\| \leq |\lambda_n| \, \|x\| \to 0.$$

Now,

$$A\widetilde{x}_n = A(x - \widehat{x}_n) = Ax - A\widehat{x}_n = Ax - A\left(\sum_{k=1}^{n-1} \langle x, \varphi_k \rangle \varphi_k\right) = Ax - \sum_{k=1}^{n-1} \lambda_k \langle x, \varphi_k \rangle \varphi_k.$$

Since $\|A\widetilde{x}_n\| \to 0$,

$$Ax = \sum_{k=1}^{\infty} \lambda_k \langle x, \varphi_k \rangle \varphi_k.$$

If the procedure for obtaining nonzero λ_k terminates with the last one being λ_{n-1}, say, then $A\widetilde{x_n} = 0$, and

$$Ax = \sum_{k=1}^{n-1} \lambda_k \langle x, \varphi_k \rangle \varphi_k. \qquad \square$$

Example 7.32. Let $A: \mathbb{C}^n \to \mathbb{C}^n$ (or $A: \mathbb{R}^n \to \mathbb{R}^n$), where $Ax = ax$ for some $n \times n$ matrix a. Assume $a^{\mathsf{H}} = a$ so that A is self-adjoint. Then by the Spectral Theorem,

$$\begin{aligned} ax = Ax = \sum_{k=1}^{r} \lambda_k \langle x, \varphi_k \rangle \varphi_k &= [\ \varphi_1\ |\cdots|\ \varphi_r\ ] \begin{bmatrix} \lambda_1 \langle x, \varphi_1 \rangle \\ \vdots \\ \lambda_r \langle x, \varphi_r \rangle \end{bmatrix} \\ &= [\ \varphi_1\ |\cdots|\ \varphi_r\ ] \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_r \end{bmatrix} \begin{bmatrix} \langle x, \varphi_1 \rangle \\ \vdots \\ \langle x, \varphi_r \rangle \end{bmatrix} \\ &= [\ \varphi_1\ |\cdots|\ \varphi_r\ ] \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_r \end{bmatrix} \left[\begin{array}{c} \varphi_1^{\mathsf{H}} \\ \hline \vdots \\ \hline \varphi_r^{\mathsf{H}} \end{array}\right] x. \end{aligned}$$

If $r < n$, let $\{\varphi_{r+1}, \ldots, \varphi_n\}$ be an orthonormal basis for $\ker A$. Then $\{\varphi_1, \ldots, \varphi_n\}$ is an orthonormal basis for the whole space. By defining $\lambda_{r+1} = \cdots = \lambda_n = 0$, we can repeat the foregoing calculation with r replaced by n to arrive at

$$ax = \sum_{k=1}^{n} \lambda_k \langle x, \varphi_k \rangle \varphi_k = [\ \varphi_1\ |\cdots|\ \varphi_n\ ] \begin{bmatrix} \lambda_1 & & & & & \\ & \ddots & & & & \\ & & \lambda_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix} \left[\begin{array}{c} \varphi_1^{\mathsf{H}} \\ \hline \vdots \\ \hline \varphi_n^{\mathsf{H}} \end{array}\right] x.$$

If we put $P := [\varphi_1|\cdots|\varphi_n]$, and if we let Λ denote the above diagonal matrix, then the preceding equation can be written as $ax = P\Lambda P^{\mathsf{H}} x$. Since x is arbitrary, we conclude that $a = P\Lambda P^{\mathsf{H}}$. Since $P^{\mathsf{H}} P = I$, it follows that $P^{\mathsf{H}} a P = \Lambda$. Note that although $[\varphi_1|\cdots|\varphi_r]$ is nonsingular, it maps r-dimensional space into n-dimensional space and is therefore not onto. However, P is nonsingular and $n \times n$; hence, by Problem 4.5(c), $PP^{\mathsf{H}} = I$.

To obtain the eigenvalues (but not the eigenvectors) of a in MATLAB, use the statement `lambda=eig(a)`, where `lambda` is an n-dimensional *column* vector containing *both* $\lambda_1,\dots,\lambda_r$ *and* the zero eigenvalues in some order, not necessarily ordered by absolute value. To obtain both the eigenvalues and the eigenvectors of a, use `[P,Lambda] = eig(a)`, where `Lambda` is the $n\times n$ diagonal matrix with `lambda` along the main diagonal. To extract the diagonal elements of `Lambda` as a *column* vector, use the command `lambda = diag(Lambda)`. Note that the columns of `P` are ordered to correspond with the ordering of the eigenvalues in `lambda`.

Once we have found `P` and `lambda`, we can use the fact that $a = P\Lambda P^{\mathsf{H}}$ to compute $y = ax$ in MATLAB with the statement `y = P*(lambda.*(P'*x))`. Recall that in MATLAB, `P'` returns P^{H}, while `P.'` returns P^{T}.

Example 7.33 (Square Root of an Operator). Let A satisfy the hypotheses of the Spectral Theorem. In addition, assume that A is positive semidefinite. Then by Proposition 7.25, all eigenvalues of A are nonnegative, and thus $\lambda_k \searrow 0$. The square root of A is the operator defined by

$$\sqrt{A}\,x := \sum_{k=1}^{\infty} \sqrt{\lambda_k}\langle x,\varphi_k\rangle\varphi_k.$$

It is easy to verify that $\sqrt{A}$ is self adjoint, and

$$\begin{aligned}
\sqrt{A}\big(\sqrt{A}\,x\big) &= \sum_{k=1}^{\infty} \sqrt{\lambda_k}\Big\langle \sqrt{A}x,\varphi_k\Big\rangle\varphi_k \\
&= \sum_{k=1}^{\infty} \sqrt{\lambda_k}\Big\langle \sum_{\ell=1}^{\infty} \sqrt{\lambda_\ell}\langle x,\varphi_\ell\rangle\varphi_\ell,\varphi_k\Big\rangle\varphi_k \\
&= \sum_{k=1}^{\infty} \sqrt{\lambda_k}\Big(\sum_{\ell=1}^{\infty} \sqrt{\lambda_\ell}\langle x,\varphi_\ell\rangle\langle\varphi_\ell,\varphi_k\rangle\Big)\varphi_k \\
&= \sum_{k=1}^{\infty} \lambda_k\langle x,\varphi_k\rangle\varphi_k \\
&= Ax.
\end{aligned}$$

When A is the matrix operator of Example 7.32 and Λ and P are as defined there, we define $\sqrt{\Lambda}$ to be the $n\times n$ matrix $\operatorname{diag}(\sqrt{\lambda_1},\dots,\sqrt{\lambda_r},0,\dots,0)$. Then

$$a = P\Lambda P^{\mathsf{H}} = P\sqrt{\Lambda}\sqrt{\Lambda}P^{\mathsf{H}} = P\sqrt{\Lambda}P^{\mathsf{H}}P\sqrt{\Lambda}P^{\mathsf{H}} = (P\sqrt{\Lambda}P^{\mathsf{H}})(P\sqrt{\Lambda}P^{\mathsf{H}}).$$

We therefore define $\sqrt{a} := P\sqrt{\Lambda}P^{\mathsf{H}}$. Note that $(\sqrt{a}\,)^{\mathsf{H}} = \sqrt{a}$.

Example 7.34 (Fredholm Equations of the Second Kind). If X is a Hilbert space and $A: X \to X$ is a compact linear operator, then an equation of the form $(I+A)x = y$ is called a **Fredholm equation of the *second* kind**. Note that the operator $I+A$ is one-to-one if and only if -1 is not an eigenvalue of A; in this case, any solution of the second-kind equation must be unique. When A is self-adjoint, we can use the Spectral Theorem to find the solution. Before doing so, we make a few preliminary observations. By the Spectral Theorem,

$$Ax = \sum_{k=1}^{\infty} \lambda_k \langle x, \varphi_k \rangle \varphi_k,$$

where $|\lambda_k| \searrow 0$. Assuming $\lambda_k \neq -1$ for all k implies the ratio $\lambda_k/(1+\lambda_k)$ is bounded, which implies

$$\sum_{k=1}^{\infty} \left| \frac{\lambda_k \langle y, \varphi_k \rangle}{1+\lambda_k} \right|^2 < \infty, \tag{7.9}$$

since $\sum_{k=1}^{\infty} |\langle y, \varphi_k \rangle|^2 < \infty$ by Theorem 6.35.

Suppose there is an $x \in X$ with $x + Ax = y$. Then $y - x = Ax$, where Ax has the expansion above, which implies $y - x = Ax \in \overline{\operatorname{span}\{\varphi_k\}}$. Hence,

$$y - x = \sum_{k=1}^{\infty} \langle y - x, \varphi_k \rangle \varphi_k.$$

We thus have two expansions of Ax in terms of the φ_k. Therefore, the corresponding coefficients must be equal; i.e., $\langle y - x, \varphi_k \rangle = \lambda_k \langle x, \varphi_k \rangle$, or

$$\langle x, \varphi_k \rangle = \frac{\langle y, \varphi_k \rangle}{1+\lambda_k}.$$

Now write

$$\begin{aligned} y - x = Ax &= \sum_{k=1}^{\infty} \lambda_k \langle x, \varphi_k \rangle \varphi_k \\ &= \sum_{k=1}^{\infty} \lambda_k \frac{\langle y, \varphi_k \rangle}{1+\lambda_k} \varphi_k, \end{aligned}$$

or

$$x = y - \sum_{k=1}^{\infty} \frac{\lambda_k \langle y, \varphi_k \rangle}{1+\lambda_k} \varphi_k. \tag{7.10}$$

Conversely, given any $y \in X$, the sum in (7.10) converges on account of (7.9), which means that the right-hand side of (7.10) is well defined. If we now *define* x by (7.10),

it is easy to check that $\langle x, \varphi_\ell \rangle = \langle y, \varphi_\ell \rangle / (1 + \lambda_\ell)$. Making this substitution in (7.10) yields

$$\begin{aligned} x &= y - \sum_{k=1}^{\infty} \lambda_k \langle x, \varphi_k \rangle \varphi_k \\ &= y - Ax, \qquad \text{by the Spectral Theorem,} \end{aligned}$$

thus showing that (7.10) solves $x + Ax = y$.

If $\dim X = n$ and there are $r < n$ nonzero eigenvalues of A, let $\varphi_{r+1}, \ldots, \varphi_n$ be an orthonormal basis for $\ker A$ so that $\varphi_1, \ldots, \varphi_n$ is an orthonormal basis for X. Then $y = \sum_{k=1}^{n} \langle y, \varphi_k \rangle \varphi_k$. If we also put $\lambda_{r+1} = \cdots = \lambda_n = 0$, then (7.10) can be written as

$$x = \sum_{k=1}^{n} \frac{\langle y, \varphi_k \rangle}{1 + \lambda_k} \varphi_k = \begin{bmatrix} \varphi_1 \mid \cdots \mid \varphi_n \end{bmatrix} \begin{bmatrix} \frac{1}{1+\lambda_1} & & \\ & \ddots & \\ & & \frac{1}{1+\lambda_n} \end{bmatrix} \left[\begin{array}{c} \varphi_1^{\mathsf{H}} \\ \hline \vdots \\ \hline \varphi_n^{\mathsf{H}} \end{array}\right] y.$$

To compute this expression efficiently in MATLAB, put $P := [\varphi_1 | \cdots | \varphi_n]$ and $\lambda := [\lambda_1, \ldots, \lambda_n]^{\mathsf{T}}$. Then `x = P*(P'*y./(1+lambda))`.

7.4.1. Simultaneous Diagonalization and Normal Operators[§]

For a compact, self-adjoint operator A on a Hilbert space X, the Spectral Theorem tells us that the nonzero eigenvalues λ_k tend to zero. Hence, a nonzero eigenvalue can be repeated at most a finite number of times. Equivalently, if λ is a nonzero eigenvalue of such an operator, the corresponding eigenspace $\{x : Ax = \lambda x\}$ is finite dimensional. For example, if the first eigenvalue is repeated n times; i.e., $\lambda_1 = \cdots = \lambda_n$,[h] the first eigenspace is

$$\Phi_1 := \operatorname{span}\{\varphi_1, \ldots, \varphi_n\}.$$

Since every $x \in \Phi_1$ satisfies $Ax = \lambda_1 x$, we denote by $\tilde{\lambda}_1 = \lambda_1$ the first distinct eigenvalue. If λ_{n+1} is repeated m times, then the second eigenspace is

$$\Phi_2 := \operatorname{span}\{\varphi_{n+1}, \ldots, \varphi_{n+m}\},$$

[§]This material is not needed in the sequel. It is not necessary to read this subsection in order to do the problems involving commuting or normal operators.

[h]We may have to reorder the eigenpairs before starting this procedure. For example, if the derivation of the Spectral Theorem produces eigenvalues $-5, 5, 5, -5, 4, -4, 4, \ldots$, we reorder as $5, 5, -5, -5, 4, 4, -4, \ldots$, etc.

and the second distinct eigenvalue is $\tilde{\lambda}_2 = \lambda_{n+1}$. If λ_{n+m+1} is repeated p times, then

$$\Phi_3 := \operatorname{span}\{\varphi_{n+m+1}, \ldots, \varphi_{n+m+p}\}$$

and $\tilde{\lambda}_3 = \lambda_{n+m+1}$ is the third distinct eigenvalue. Continuing in this way shows that (7.8) can be expanded as

$$X = \ker A \oplus \Phi_1 \oplus \Phi_2 \oplus \cdots,$$

where each Φ_k is finite dimensional, and for $x \in \Phi_k$, $Ax = \tilde{\lambda}_k x$.

Commuting Operators

Let $B: X \to X$ be another self-adjoint operator, and assume that A and B **commute**; i.e., $AB = BA$, or in more detail, for all x, $A(Bx) = B(Ax)$. By Problem 7.50, each Φ_k is invariant under B, as is $\ker A$. By applying the Spectral Theorem to the restriction of B to each Φ_k, we get an orthonormal basis for Φ_k, say $\{\widetilde{\varphi}_{ki}\}$, and eigenvalues $\{\mu_{ki}\}$ such that $B\widetilde{\varphi}_k = \mu_{ki}\widetilde{\varphi}_{ki}$.[i] Hence, for $x \in \Phi_k$,

$$x = \sum_i \langle x, \widetilde{\varphi}_{ki}\rangle \widetilde{\varphi}_{ki},$$

and it follows that

$$Bx = \sum_i \langle x, \widetilde{\varphi}_{ki}\rangle B\widetilde{\varphi}_{ki} = \sum_i \mu_{ki}\langle x, \widetilde{\varphi}_{ki}\rangle \widetilde{\varphi}_{ki}.$$

Of course, since $Ax = \tilde{\lambda}_k x$ for $x \in \Phi_k$, we also have

$$Ax = \sum_i \langle x, \widetilde{\varphi}_{ki}\rangle A\widetilde{\varphi}_{ki} = \sum_i \tilde{\lambda}_k\langle x, \widetilde{\varphi}_{ki}\rangle \widetilde{\varphi}_{ki}.$$

To this point, we have not assumed B is compact. However, if it is, we can also apply the Spectral Theorem to the restriction of B to $\ker A$, which may be infinite dimensional. Then there exist orthonormal vectors $\widetilde{\varphi}_{0i}$ and *nonzero* eigenvalues μ_{0i} such that for $x \in \ker A$,

$$x = x_{00} + \sum_i \langle x, \widetilde{\varphi}_{0i}\rangle \widetilde{\varphi}_{0i},$$

where $x_{00} \in \ker B \cap \ker A$ and is orthogonal to the $\widetilde{\varphi}_{0i}$, and

$$Bx = \sum_i \mu_{0i}\langle x, \widetilde{\varphi}_{0i}\rangle \widetilde{\varphi}_{0i}.$$

[i] We allow some of the μ_{ki} to be zero in case any elements of Φ_k belong to $\ker B$.

For $x \in X$, we can write

$$x = x_{00} + \sum_{k=0}^{\infty} \sum_i \langle x, \widetilde{\varphi}_{ki} \rangle \widetilde{\varphi}_{ki},$$

where x_{00} is orthogonal to all of the $\widetilde{\varphi}_{ki}$. Then applying B and A separately to this equation, we have

$$Bx = \sum_{k=0}^{\infty} \sum_i \mu_{ki} \langle x, \widetilde{\varphi}_{ki} \rangle \widetilde{\varphi}_{ki} \quad \text{and} \quad Ax = \sum_{k=1}^{\infty} \tilde{\lambda}_k \sum_i \langle x, \widetilde{\varphi}_{ki} \rangle \widetilde{\varphi}_{ki}.$$

The outer sum for Ax can be extended to $k = 0$ if we put $\tilde{\lambda}_0 := 0$. These formulas show that A and B are **simultaneously diagonalizable** with a common set of orthonormal eigenvectors. It should also be noted that if $\mu_{ki} = 0$, then $\tilde{\lambda}_k \neq 0$; i.e., each common eigenvector $\widetilde{\varphi}_{ki}$ corresponds to a nonzero eigenvalue for at least one of the two operators A or B.

Normal Operators

Suppose an operator $A: X \to X$ has adjoint A^*. Then A is said to be **normal** if it commutes with its adjoint; i.e., if $A^*A = AA^*$. In particular, if A is **unitary**, i.e., if $A^*A = I$ and $AA^* = I$, then A is normal. For an arbitrary operator $A: X \to X$ having adjoint A^*, the operators

$$B := \frac{A + A^*}{2} \quad \text{and} \quad C := \frac{A - A^*}{2j}$$

are self adjoint, and by Problem 7.51 they compute if and only if A is normal. Hence, if A is a compact normal operator on a Hilbert space, then B and C can be simultaneously diagonalized as

$$Bx = \sum_k \beta_k \langle x, \varphi_k \rangle \varphi_k \quad \text{and} \quad Cx = \sum_k \gamma_k \langle x, \varphi_k \rangle \varphi_k,$$

where the φ_k are orthonormal, and for each k, β_k and γ_k are real and not both zero. Since $B + jC = A$ and $B - jC = A^*$, if we put $\alpha_k := \beta_k + j\gamma_k$, then

$$Ax = \sum_k \alpha_k \langle x, \varphi_k \rangle \varphi_k \quad \text{and} \quad A^*x = \sum_k \overline{\alpha}_k \langle x, \varphi_k \rangle \varphi_k, \tag{7.11}$$

where $\alpha_k \neq 0$ for all k. Furthermore, every $x \in X$ has the unique representation

$$x = x_{00} + \sum_k \langle x, \varphi_k \rangle \varphi_k,$$

where $x_{00} \in \ker A = \ker A^*$ is orthogonal to all of the φ_k. The fact that $x_{00} \in \ker A$ follows from the two observations

$$\ker B \cap \ker C = \ker A \cap \ker A^* = \ker A = \ker A^*,$$

where the first equality follows from the definitions of B and C (Problem 7.52), and the second one follows because the normality of A implies $\ker A = \ker A^*$ (Problem 7.53).

7.5. The Singular-Value Decomposition (SVD)

Theorem 7.35 (Singular-Value Decomposition (SVD)). *Let X and Y be Hilbert spaces. Let $A: X \to Y$ be a compact linear operator with $\|A\| > 0$. Then there exist $\lambda_k \searrow 0$ and corresponding orthonormal sequences $\{\varphi_k\}$ in X and $\{\psi_k\}$ in Y such that $\psi_k = (1/\lambda_k)A\varphi_k$, and*

$$(A^*A)\varphi_k = \lambda_k^2 \varphi_k, \tag{7.12}$$

$$Ax = \sum_k \lambda_k \langle x, \varphi_k \rangle \psi_k, \quad \text{for all } x \in X, \tag{7.13}$$

$$\overline{\operatorname{span}\{\psi_k\}} = \overline{\operatorname{range} A} = (\ker A^*)^\perp, \tag{7.14}$$

$$X = \ker A \oplus \overline{\operatorname{span}\{\varphi_k\}}, \tag{7.15}$$

$$(AA^*)\psi_k = \lambda_k^2 \psi_k, \tag{7.16}$$

$$A^*y = \sum_k \lambda_k \langle y, \psi_k \rangle \varphi_k, \quad \text{for all } y \in Y, \tag{7.17}$$

$$\overline{\operatorname{span}\{\varphi_k\}} = \overline{\operatorname{range} A^*} = (\ker A)^\perp, \tag{7.18}$$

$$Y = \ker A^* \oplus \overline{\operatorname{span}\{\psi_k\}}. \tag{7.19}$$

The positive numbers $\{\lambda_k\}$ are called the **singular values** *of the operator, and $\lambda_1 = \|A\|$.*

Proof. We begin with a few observations. Since A is compact, it is bounded; then since X is a Hilbert space, A^* exists. Also we have shown in Proposition 7.8 that $\|A\| = \|A^*\|$, and so A^* is bounded. Since A is compact and A^* is bounded, A^*A is compact by Problem 7.44(c). Hence, we can apply the Spectral Theorem to

obtain a family of eigenpairs of A^*A. Since $\langle A^*Ax, x\rangle = \langle Ax, Ax\rangle \geq 0$, A^*A is positive semidefinite; hence, its nonzero eigenvalues must be positive. Let λ_k denote the positive square root of the kth nonzero eigenvalue of A^*A so that we can write the eigenpairs of A^*A as $\{(\lambda_k^2, \varphi_k)\}$ with $\lambda_k^2 \searrow 0$. Then (7.12) follows as does $\lambda_k \searrow 0$.

Now put $\psi_k := (1/\lambda_k)A\varphi_k$. It is easy to see that (7.12) implies (7.16), and that the $\{\psi_k\}$ are orthonormal because the $\{\varphi_k\}$ are orthonormal.

To prove (7.15), note that by the Spectral Theorem,

$$X = \ker A^*A \oplus \overline{\operatorname{span}\{\varphi_k\}}.$$

Since $\ker A^*A = \ker A$, (7.15) follows.

We now derive the representation (7.13). By (7.15) and Theorem 6.35, every $x \in X$ can be written in the form

$$x = x_0 + \sum_k \langle x, \varphi_k\rangle \varphi_k,$$

where $x_0 \in \ker A$. Applying A to this equation and recalling that $\psi_k := (1/\lambda_k)A\varphi_k$ yields (7.13).

Equation (7.14) is now derived. From (7.13), $\operatorname{range} A \subset \overline{\operatorname{span}\{\psi_k\}}$, which implies

$$\overline{\operatorname{range} A} \subset \overline{\operatorname{span}\{\psi_k\}}.$$

To obtain the reverse inclusion as well as the right-hand equality in (7.14), we proceed as follows. On account of (7.16), $\operatorname{span}\{\psi_k\} \subset \operatorname{range} AA^*$, and so

$$\begin{aligned}
\overline{\operatorname{span}\{\psi_k\}} &\subset \overline{\operatorname{range} AA^*} \\
&= (\ker AA^*)^\perp, \quad \text{by Theorem 7.9(a)}, \\
&= (\ker A^*)^\perp, \quad \text{by Theorem 4.13(d)}, \\
&= \overline{\operatorname{range} A}, \quad \text{by Theorem 7.9(b)},
\end{aligned}$$

and (7.14) follows.

We now prove (7.19). Since A^* is bounded, $\ker A^*$ is closed. Since Y is a Hilbert space, we can apply the Projection Theorem. Combining this with the Orthogonality Principle and (7.14) yields

$$Y = \ker A^* \oplus (\ker A^*)^\perp = \ker A^* \oplus \overline{\operatorname{span}\{\psi_k\}}.$$

Using the above decomposition of Y, for any $y \in Y$, we can write

$$y = y_0 + \sum_k \langle y, \psi_k\rangle \psi_k,$$

where $y_0 \in \ker A^*$. Now apply A^* to this equation, and note that

$$A^*\psi_k = A^*\Big(\frac{1}{\lambda_k} A\varphi_k\Big) = \frac{\lambda_k^2 \varphi_k}{\lambda_k} = \lambda_k \varphi_k.$$

Hence, (7.17) follows.

Equation (7.18) can be obtained in a manner analogous to that used to obtain (7.14).

Finally, we show that $\lambda_1 = \|A\| = \|A^*\|$. First, from (7.12), $\lambda_k^2 = \|(A^*A)\varphi_k\| \leq \|A\|^2$; hence, $\lambda_k \leq \|A\|$. Since $\lambda_k \leq \lambda_1$, we can use (7.13) to write

$$\|Ax\|^2 = \sum_k \lambda_k^2 \big|\langle x, \varphi_k\rangle\big|^2 \leq \sum_k \lambda_1^2 \big|\langle x, \varphi_k\rangle\big|^2 = \lambda_1^2 \sum_k \big|\langle x, \varphi_k\rangle\big|^2 \leq \lambda_1^2 \|x\|^2,$$

where the last step follows by Theorem 6.35. Thus, $\|A\| \leq \lambda_1$. □

Remark. If X is finite dimensional, then A has a **smallest singular value**, say λ_n. If A is nonsingular, then $\dim X = n$, and a trivial modification of the above display shows that

$$\|Ax\| \geq \lambda_n \|x\|.$$

Example 7.36. Let $A\colon \mathbb{C}^n \to \mathbb{C}^m$ (or $A\colon \mathbb{R}^n \to \mathbb{R}^m$), where $Ax = ax$ for some $m \times n$ matrix a. By the SVD, we can write

$$ax = \sum_{k=1}^{r} \lambda_k \langle x, \varphi_k\rangle \psi_k = [\ \psi_1\ |\cdots|\ \psi_r\ ] \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_r \end{bmatrix} \left[\begin{array}{c} \varphi_1^{\mathsf{H}} \\ \hline \vdots \\ \hline \varphi_r^{\mathsf{H}} \end{array}\right] x,$$

where $r \leq \min\{m,n\}$. If $r < n$, let $\varphi_{r+1}, \ldots, \varphi_n$ be an orthonormal basis for $\ker A$. If $r < m$, let $\psi_{r+1}, \ldots, \psi_m$ be an orthonormal basis for $\ker A^*$. Put $P := [\varphi_1|\cdots|\varphi_n]$ and $Q := [\psi_1|\cdots|\psi_m]$. We can then write $ax = QSP^{\mathsf{H}}x$, where S is the $m \times n$ matrix

$$S := \begin{bmatrix} \operatorname{diag}(\lambda_1, \ldots, \lambda_r) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

Since x is arbitrary, we have $a = QSP^{\mathsf{H}}$. Since $P^{\mathsf{H}}P = I$ and $Q^{\mathsf{H}}Q = I$, we see that $Q^{\mathsf{H}}aP = S$. Since Q is nonsingular and $m \times m$, we have $QQ^{\mathsf{H}} = I$. Since P is nonsingular and $n \times n$, we have $PP^{\mathsf{H}} = I$.

To obtain the SVD in MATLAB, use the command `[Q,S,P] = svd(a)`. The diagonal elements of `S` are nonincreasing. To obtain only the diagonal of `S` as a *column* vector, use instead the command `s = svd(a)`.

Example 7.37 (Fredholm Equations of the First Kind). Let $A\colon X \to Y$ be a compact linear operator between Hilbert spaces X and Y. Then an equation of the form

$Ax = y$ is called a **Fredholm equation of the *first* kind**. There is a solution if and only if $y \in \text{range}A$. If $y \notin \text{range}A$, let $\widehat{y}$ denote the projection of y onto

$$\overline{\text{range}A} = \overline{\text{span}\{\psi_k\}}.$$

Then

$$\widehat{y} = \sum_k \langle y, \psi_k \rangle \psi_k \tag{7.20}$$

by Theorem 6.35. By the SVD we can write

$$Ax = \sum_k \lambda_k \langle x, \varphi_k \rangle \psi_k.$$

Comparing these two equations shows that if there is a solution of $Ax = \widehat{y}$, then the corresponding coefficients must be equal, which implies

$$\langle x, \varphi_k \rangle = \frac{\langle y, \psi_k \rangle}{\lambda_k}.$$

Recall also (Section 4.3.3) that since $X = \ker A \oplus (\ker A)^\perp$, if x is any solution of $Ax = \widehat{y}$, then the minimum-norm solution, $\widetilde{x}$, is the projection of x onto $(\ker A)^\perp$. Since the SVD tells us that $(\ker A)^\perp = \overline{\text{span}\{\varphi_k\}}$,

$$\widetilde{x} = \sum_k \langle x, \varphi_k \rangle \varphi_k = \sum_k \frac{\langle y, \psi_k \rangle}{\lambda_k} \varphi_k. \tag{7.21}$$

We now see that if there is any solution of $Ax = \widehat{y}$, then the minimum-norm solution is given by the above formula. Furthermore,

$$\infty > \|\widetilde{x}\|^2 = \sum_k \left| \langle x, \varphi_k \rangle \right|^2 = \sum_k \frac{\left| \langle y, \psi_k \rangle \right|^2}{\lambda_k^2}.$$

Hence, the condition

$$\sum_k \frac{\left| \langle y, \psi_k \rangle \right|^2}{\lambda_k^2} < \infty \tag{7.22}$$

necessary in order for there to be a solution of $Ax = \widehat{y}$. Conversely, if (7.22) holds, then the sum on the right in (7.21) converges by Theorem 6.35. Denoting this sum by $\widetilde{x}$, we find that

$$A\widetilde{x} = A\left(\sum_k \frac{\langle y, \psi_k \rangle}{\lambda_k} \varphi_k \right) = \sum_k \frac{\langle y, \psi_k \rangle}{\lambda_k} A\varphi_k = \sum_k \langle y, \psi_k \rangle \frac{A\varphi_k}{\lambda_k} = \sum_k \langle y, \psi_k \rangle \psi_k = \widehat{y},$$

where the last equality follows from (7.20) and the second one because A is convergence preserving. The fact that (7.22) is necessary and sufficient for the existence of a solution to $Ax = \widehat{y}$ is known as **Picard's criterion**.

Now suppose X and Y are finite-dimensional Euclidean spaces, $Ax = ax$ for some $m \times n$ matrix a, and a has r singular values. In this case, (7.21) can be expressed as

$$\widetilde{x} = \begin{bmatrix} \varphi_1 \,|\cdots|\, \varphi_r \end{bmatrix} \begin{bmatrix} \frac{1}{\lambda_1} & & \\ & \ddots & \\ & & \frac{1}{\lambda_r} \end{bmatrix} \begin{bmatrix} \psi_1^{\mathsf{H}} \\ \hline \vdots \\ \hline \psi_r^{\mathsf{H}} \end{bmatrix} y. \tag{7.23}$$

To do this in MATLAB, use the commands

```
[Q,S,P] = svd(a);
s = diag(S);
s = s(s>0);          % Keep only positive entries in s
r = length(s);
xtilde = P(:,1:r)*(Q(:,1:r)'*y./s);
```

Compare with the last line of Example 7.34.

Remark. Comparing the preceding discussion with that in Section 4.3.4, we see that the **pseudoinverse** $A^\dagger$ is defined for all y satisfying Picard's criterion; i.e.,

$$A^\dagger y = \sum_k \frac{\langle y, \psi_k \rangle}{\lambda_k} \varphi_k, \quad \text{for } y \text{ satisfying (7.22).}$$

When $Ax = ax$ and a is an $m \times n$ matrix, $a^\dagger$ is the product of the three matrices multiplying y in (7.23). To compute $a^\dagger$ in MATLAB, you can use the **pseudoinverse** command `pinv(a)`.

7.5.1. Ill-Posed and Well-Posed Problems

A problem is said to be **well posed** if it has a unique solution and if the solution varies continuously as a function of the problem parameters. If a problem is not well posed, it is said to be **ill posed**. In this section, we consider the problem of solving linear equations such as $Ax = y$, where y is the parameter to be varied. If A is invertible, then we want to know if $x = A^{-1}y$ is a continuous function of y; of course, for linear operators, this is equivalent to asking of A^{-1} is a bounded operator. When A is singular, we can ask if the pseudoinverse is bounded.

First-Kind Equations

We show that when a compact operator A has an infinite number of positive singular values, solving $Ax = y$ is an **ill-posed problem**. Suppose that $y \in \text{range}\, A$. Then there is a solution of $Ax = y$, $y = \widehat{y}$ is given by (7.20), and (7.22) holds. Fix any $\delta > 0$, and let

$$y_i := y + \delta \psi_i.$$

Then $\|y_i - y\| = \delta$, and

$$\sum_k \frac{\left|\langle y_i, \psi_k\rangle\right|^2}{\lambda_k^2} = \frac{\left|\langle y, \psi_i\rangle + \delta\right|^2}{\lambda_i^2} + \sum_{k\neq i} \frac{\left|\langle y, \psi_k\rangle\right|^2}{\lambda_k^2},$$

which is clearly finite on account of (7.22). So the minimum-norm solution of $Ax = y$ is given by the right-hand side of (7.21). Similarly, the minimum-norm solution of $Ax = y_i$ is given by

$$\widetilde{x_i} := \sum_k \frac{\langle y_i, \psi_k\rangle}{\lambda_k} \varphi_k = \widetilde{x} + \frac{\delta}{\lambda_i} \varphi_i.$$

Hence,

$$\|\widetilde{x_i} - \widetilde{x}\| = \frac{\delta}{\lambda_i}.$$

We can now show that the mapping that takes $y \in \text{range}\, A$ into the minimum-norm solution of $Ax = y$ is not continuous. Fix any $y \in \text{range}\, A$. Given any $\varepsilon > 0$, one would like to find a $\delta > 0$ such that for all $y' \in \text{range}\, A$,

$$\|y' - y\| \leq \delta \quad \Rightarrow \quad \|\widetilde{x'} - \widetilde{x}\| < \varepsilon.$$

However, suppose such a δ exists. Choose i such that $\delta/\lambda_i \geq \varepsilon$. Then $\|y_i - y\| = \delta$, but $\|\widetilde{x_i} - \widetilde{x}\| = \delta/\lambda_i \geq \varepsilon$.

Second-Kind Equations

Second-kind Fredholm equations are well posed. This is easy to see when A is self-adjoint. Let x be given by (7.10), and let x' denote the right-hand side of (7.10) with y replaced by y' (recall that in (7.10), $\{(\lambda_k, \varphi_k)\}$ are the eigenpairs of A). Then

$$\|x - x'\| \leq \|y - y'\| + \left\| \sum_{k=1}^{\infty} \frac{\lambda_k \langle y' - y, \varphi_k\rangle}{1 + \lambda_k} \varphi_k \right\|.$$

Next, write

$$\left\| \sum_{k=1}^{\infty} \frac{\lambda_k \langle y' - y, \varphi_k\rangle}{1 + \lambda_k} \varphi_k \right\|^2 = \sum_{k=1}^{\infty} \left| \frac{\lambda_k}{1 + \lambda_k} \langle y' - y, \varphi_k\rangle \right|^2.$$

Assuming that $\lambda_k \neq -1$ for all k, and using the fact that $|\lambda_k| \to 0$, it is easy to see that

$$B := \sup_k \left| \frac{\lambda_k}{1+\lambda_k} \right| < \infty.$$

Hence, we can write

$$\begin{aligned} \left\| \sum_{k=1}^{\infty} \frac{\lambda_k \langle y' - y, \varphi_k \rangle}{1+\lambda_k} \varphi_k \right\|^2 &= \sum_{k=1}^{\infty} \left| \frac{\lambda_k}{1+\lambda_k} \langle y' - y, \varphi_k \rangle \right|^2 \\ &\leq B^2 \sum_{k=1}^{\infty} |\langle y' - y, \varphi_k \rangle|^2 \\ &\leq B^2 \|y' - y\|^2. \end{aligned}$$

Hence, $\|x - x'\| \leq (1+B)\|y - y'\|$.

Operators of Norm Less Than One[§]

We consider equations of the form $(I - A)x = y$. This looks like a second-kind Fredholm equation in which A is replaced by $-A$. However, in this section, we do not assume that A is compact or that A is self-adjoint. Instead, we assume only $\|A\| < 1$.[j] Of course, if A has any eigenvalues, their magnitudes must also be less than one. This is in contrast to the second-kind Fredholm equations in which we required only that all eigenvalues be different from -1; i.e., some eigenvalues could have magnitudes greater than one.

Let $A: X \to X$, where X is a Banach space and $\|A\| < 1$. We begin the analysis by showing that the series

$$\sum_{k=0}^{\infty} A^k y$$

converges. Put $B_n y := \sum_{k=0}^{n} A^k y$ and note that for $m > n$,

$$\|B_m y - B_n y\| = \left\| \sum_{k=n+1}^{m} A^k y \right\| \leq \|y\| \sum_{k=n+1}^{m} \|A\|^k.$$

Hence, $B_n y$ is a Cauchy sequence.[k] Since X is complete, $B_n y$ converges to some limit, which we denote by By. Note also that (cf. Proposition 6.6)

$$\|By\| = \lim_{n\to\infty} \|B_n y\| \leq \lim_{n\to\infty} \|y\| \sum_{k=0}^{n} \|A\|^k = \frac{\|y\|}{1 - \|A\|}.$$

[§] This material is not needed in the sequel.

[j] See Problem 7.34 for an easy way to construct such an operator using projections.

[k] By the **geometric series** (Problem 6.2), $z_n := \sum_{k=0}^{n} \|A\|^k$ converges and is therefore a Cauchy sequence of real numbers.

Hence, B is a bounded operator with $\|B\| \leq 1/(1-\|A\|)$. Next, observe that

$$(I-A)By = (I-A)\lim_{n\to\infty} B_n y = \lim_{n\to\infty}(I-A)\sum_{k=0}^{n} A^k y,$$

and

$$(I-A)\sum_{k=0}^{n} A^k y = \sum_{k=0}^{n} A^k y - \sum_{k=1}^{n+1} A^k y = y - A^{n+1}y.$$

Since $\|A\| < 1$, $A^{n+1}y \to 0$, and it follows that $(I-A)By = y$. In other words, taking $x = By$ solves $(I-A)x = y$. Furthermore, since B is continuous, the problem is well posed. To conclude, we note that a similar analysis shows that $B(I-A)x = x$. Hence, B is the inverse of $I-A$.

7.5.2. Best-Fit Subspace

Suppose we are given vectors $a_1, \ldots, a_n$ in an inner-product space Y, and we would like to find an r-dimensional subspace such that the sum of squared distances from the points to the subspace is minimized.

To set up the problem in more detail, suppose that $w_1, \ldots, w_r$ form an orthonormal basis for the subspace. Then the distance from a_j to the subspace is simply $\|a_j - \widehat{a}_j\|$, where $\widehat{a}_j$ is the orthogonal projection of a_j onto the subspace. Our goal is to choose orthonormal vectors $w_1, \ldots, w_r$ to minimize

$$\sum_{j=1}^{n} \|a_j - \widehat{a}_j\|^2 = \sum_{j=1}^{n} \|a_j\|^2 - \|\widehat{a}_j\|^2, \quad \text{by the error formula (3.7).}$$

Since only the $\widehat{a}_j$ depend on the w_i, it suffices to maximize

$$\sum_{j=1}^{n} \|\widehat{a}_j\|^2 = \sum_{j=1}^{n}\sum_{i=1}^{r} \big|\langle a_j, w_i\rangle\big|^2, \quad \text{by (3.11).} \tag{7.24}$$

Now consider the operator $A\colon \mathbb{C}^n \to Y$ by

$$Ax := \sum_{j=1}^{n} x_j a_j, \quad x \in \mathbb{C}^n.$$

An easy calculation shows that the adjoint of A is given by

$$A^* y = \begin{bmatrix} \langle y, a_1\rangle \\ \vdots \\ \langle y, a_n\rangle \end{bmatrix}, \quad y \in Y.$$

Now change the order of summation in (7.24) and use the fact that $\big|\langle a_j, w_i\rangle\big| = \big|\langle w_i, a_j\rangle\big|$. It follows that

$$\sum_{j=1}^{n} \|\widehat{a}_j\|^2 = \sum_{i=1}^{r} \|A^* w_i\|^2.$$

By the SVD applied to A, we can write $A^* w_i = \sum_k \lambda_k \langle w_i, \psi_k\rangle \varphi_k$, and so in particular,

$$\|A^* w_1\|^2 = \sum_k \lambda_k^2 \big|\langle w_1, \psi_k\rangle\big|^2 \le \lambda_1^2 \sum_k \big|\langle w_1, \psi_k\rangle\big|^2 \le \lambda_1^2 \|w_1\|^2 = \lambda_1^2,$$

with equality if $w_1 = \psi_1$. Since w_2 must be orthogonal to w_1,

$$\|A^* w_2\|^2 = \sum_{k\ge 2} \lambda_k^2 \big|\langle w_1, \psi_k\rangle\big|^2 \le \lambda_2^2 \sum_{k\ge 2} \big|\langle w_1, \psi_k\rangle\big|^2 \le \lambda_2^2 \|w_1\|^2 = \lambda_2^2,$$

with equality if $w_2 = \psi_2$. Continuing in this way, we conclude that the optimal subspace is $\operatorname{span}\{\psi_1, \ldots, \psi_r\}$.

Remark. To compute the SVD of A, we diagonalize the $n \times n$ Gram matrix with entries $G_{ij} := \langle a_j, a_i\rangle$ to obtain the eigenvalues λ_k^2 and eigenvectors $\varphi_k \in \mathbb{C}^n$ of G. Then $\psi_k = (1/\lambda_k) A\varphi_k \in Y$.

7.6. Regularization

The method of **regularization** was developed in order to address the ill-posedness of first-kind Fredholm equations. The method replaces the first-kind equation with a family of related second-kind equations, which are well posed.

For fixed $\alpha \ge 0$, consider the minimization problem

$$\inf_x \|y - Ax\|^2 + \alpha \|x\|^2.$$

If $\alpha = 0$, the infimum is $\|y - \widehat{y}\|$, where $\widehat{y}$ is the projection of y onto $\overline{\operatorname{range} A}$. When $\alpha > 0$, the solution must have the property that $\|x\|^2$ cannot be too large. By adjusting the value of α, one can trade off fidelity to the data (making $Ax \approx y$) and making the energy of the solution, $\|x\|^2$, small.

From our earlier work on convex functions, we see that for fixed $\alpha > 0$, the quantity we want to minimize is convex in x. Hence, it suffices to have the Gâteaux derivative equal to zero in all directions. From the calculations in Example 5.22, the solution of the minimization problem (for fixed α) is equivalent to the solution of

$$(\alpha I + A^* A)x = A^* y. \tag{7.25}$$

Note that A^*A is self adjoint and positive semidefinite. Assuming A is compact and $\alpha > 0$, this is a second-kind Fredholm equation whose solution, denoted by x_α, is (cf. (7.10))

$$x_\alpha = \frac{A^*y}{\alpha} - \sum_{k=1}^{\infty} \frac{\frac{\lambda_k^2}{\alpha}\left\langle \frac{A^*y}{\alpha}, \varphi_k \right\rangle}{1+\frac{\lambda_k^2}{\alpha}} \varphi_k,$$

where $\{(\lambda_k^2, \varphi_k)\}$ are the eigenpairs for A^*A. By the SVD (recall (7.17) and $\psi_k := (1/\lambda_k)A\varphi_k$),

$$\begin{aligned} x_\alpha &= \frac{1}{\alpha}\left(\sum_{k=1}^{\infty} \lambda_k \langle y, \psi_k \rangle \varphi_k - \sum_{k=1}^{\infty} \frac{\lambda_k^3}{\alpha+\lambda_k^2} \langle y, \psi_k \rangle \varphi_k \right) \\ &= \sum_{k=1}^{\infty} \frac{\lambda_k}{\alpha+\lambda_k^2} \langle y, \psi_k \rangle \varphi_k. \end{aligned} \tag{7.26}$$

It is shown in the next paragraph that if Picard's criterion (7.22) holds, then

$$\lim_{\alpha \downarrow 0} x_\alpha = \sum_{k=1}^{\infty} \frac{\langle y, \psi_k \rangle}{\lambda_k} \varphi_k = \widetilde{x},$$

which is exactly the minimum-norm solution of $Ax = \widehat{y}$ in (7.21). Hence,

$$A^\dagger y = \lim_{\alpha \downarrow 0} (\alpha I + A^*A)^{-1} A^* y.$$

Write

$$\begin{aligned} \|x_\alpha - \widetilde{x}\|^2 &= \left\| \sum_{k=1}^{\infty} \frac{-\alpha}{\lambda_k(\alpha+\lambda_k^2)} \langle y, \psi_k \rangle \varphi_k \right\|^2 \\ &= \sum_{k=1}^{\infty} \left(\frac{\alpha}{\alpha+\lambda_k^2} \right)^2 \frac{\left| \langle y, \psi_k \rangle \right|^2}{\lambda_k^2}. \end{aligned}$$

Taking limits as $\alpha \downarrow 0$ on both sides, and taking the limit on the right inside the sum, we have $\|x_\alpha - \widetilde{x}\| \to 0$. Taking the limit inside the sum is justified because the series converges uniformly in α; to see this, observe that the terms of the sum are uniformly dominated by $\left| \langle y, \psi_k \rangle \right|^2 / \lambda_k^2$, which is summable on account of Picard's criterion.

For future reference, note that (7.26) implies

$$\|x_\alpha\|^2 = \sum_{k=1}^{\infty} \frac{\lambda_k^2 \left| \langle y, \psi_k \rangle \right|^2}{(\alpha+\lambda_k^2)^2}. \tag{7.27}$$

This is a real-valued, nonnegative, nonincreasing function of the nonnegative, real variable α. In fact, $\|x_\alpha\|^2$ is a continuous function of $\alpha \geq 0$. Hence, if a certain value of $\|x_\alpha\|^2$ is desired, the appropriate value of α can be determined by a root-finding algorithm. Furthermore, since root-finding algorithms repeatedly compute $\|x_\alpha\|^2$ for different values of α, it is computationally efficient to use (7.27) (in a finite-dimensional setting with a finite number of terms) since the $|\langle y, \psi_k \rangle|^2$ can be precomputed.

In the finite-dimensional case when $Ax = ax$ for some $m \times n$ matrix a, (7.26) can be expressed as

$$x_\alpha = \begin{bmatrix} \varphi_1 & \cdots & \varphi_r \end{bmatrix} \begin{bmatrix} \frac{\lambda_1}{\alpha+\lambda_1^2} & & \\ & \ddots & \\ & & \frac{\lambda_r}{\alpha+\lambda_r^2} \end{bmatrix} \begin{bmatrix} \psi_1^{\mathsf{H}} \\ \hline \vdots \\ \hline \psi_r^{\mathsf{H}} \end{bmatrix} y.$$

Notice that when $\alpha = 0$, this formula reduces to (7.23). For $\alpha \geq 0$, this formula can be evaluated in MATLAB with the commands

```
[Q,S,P] = svd(A);
s = diag(S);
s = s(s>0);          % Keep only positive entries in s
r = length(s);
Qry = Q(:,1:r)'*y;
xalpha = P(:,1:r)*(s./(alpha+(s.^2)).*Qry);
```

Note also that (7.27) is given by

```
Qry2 = Qry.*conj(Qry);
((s./(alpha+(s.^2))).^2)'*Qry2
```

where we have written this expression in two separate lines to emphasize that `Qry2` should be precomputed if the second line will be computed many times with different values of `alpha`.

Example 7.38. To illustrate the benefits of regularization, we return to Example 3.11 and make a few minor changes. Instead of using 31 exact samples of $x(t) = \cos(2\pi t/5)$, we use 31 noisy samples. Also, instead of using only 21 shifts τ_j, we use 201 equally spaced shifts. To be precise, we replace the second, third, and last lines of the script in Example 3.11 with

```
xvec = (cos(2*pi*tvec/5) + randn(size(tvec))*.2).';
tau = [-10:.1:10];
plot(t,w)
```

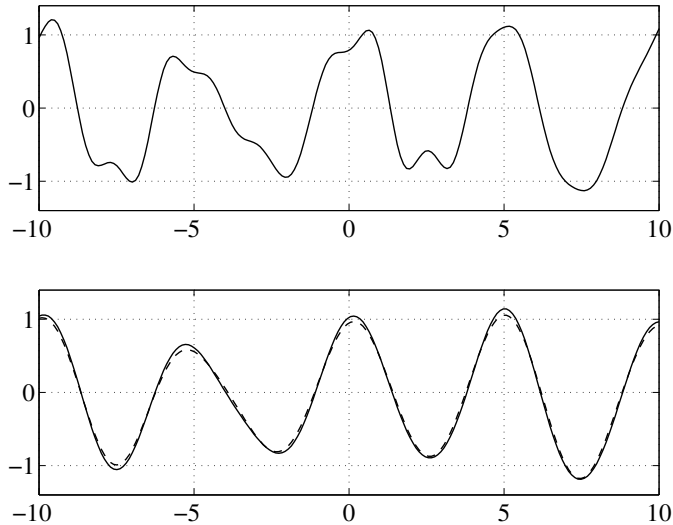


Figure 7.1. Least-squares approximation of noisy sinusoid (top), 2-norm regularized approximation (bottom, solid line), and 1-norm regularized approximation (bottom, dashed line).

Leaving the other lines of that script unchanged, we obtained the approximation of x shown at the top in Figure 7.1.

To obtain the regularized approximation shown at the bottom in Figure 7.1 (solid line), instead of using `c = A\xvec`, we chose `c` to be the minimizer of

$$\|\texttt{xvec} - \texttt{A} * \texttt{c}\|^2 + \alpha\|\texttt{c}\|^2. \tag{7.28}$$

with $\alpha = 0.2$.

7.6.1. 1-Norm Regularization

For $x \in \mathbb{R}^n$, consider the problem[l]

$$\inf_{x\in\mathbb{R}^n} \|y - Ax\|^2 + \alpha\|x\|_1.$$

[l] If A is a scalar multiple of the identity, the optimal value of x can be trivially expressed in terms of the shrinkage operator (see Problems 3.25 and 5.49).

Rather than reformulate this problem along the lines of Example 5.24, we follow [13, Section II].[m] Define the **positive** and **negative part operators** on a number t by

$$t^+ := \begin{cases} t, & t \geq 0, \\ 0, & t < 0, \end{cases} \quad \text{and} \quad t^- := \begin{cases} -t, & t < 0, \\ 0, & t \geq 0, \end{cases}$$

so that $t = t^+ - t^-$ and $|t| = t^+ + t^-$. For a vector x, define x^+ and x^- componentwise so that $x = x^+ - x^-$ and

$$\|x\|_1 = \sum_{k=1}^{n} x_k^+ + x_k^- = \langle \mathbb{1}^n, x^+ + x^- \rangle,$$

where $\mathbb{1}^n$ is the n-dimensional vector of all ones. Then our original problem can be expressed as

$$\min_{x \in \mathbb{R}^n} \|y - A(x^+ - x^-)\|^2 + \alpha \langle \mathbb{1}^n, x^+ + x^- \rangle.$$

Notice that the objective function is of the form

$$\|y - A(u - v)\|^2 + \alpha \langle \mathbb{1}^n, u + v \rangle, \quad u, v \in \mathbb{R}^n_+, \tag{7.29}$$

with the additional property that for each k, either u_k or v_k is zero. Now observe that

$$\alpha \langle \mathbb{1}^n, u + v \rangle = \alpha \sum_{k=1}^{n} u_k + v_k.$$

If for some k, u_k and v_k are both positive, we could reduce both of them *by the same amount*, and not change the value of $u - v$. Hence, the minimum of (7.29) automatically satisfies u_k or v_k is zero for each k. We conclude that our original problem is equivalent to

$$\min_{u, v \in \mathbb{R}^n_+} \|y - A(u - v)\|^2 + \alpha \langle \mathbb{1}^n, u + v \rangle.$$

To write this as a problem using the $2n$-dimensional vector $z := [u^\mathsf{T}, v^\mathsf{T}]^\mathsf{T}$, put $C := [A, -A]$ so that $Cz = A(u - v)$. Then with $B := 2C^\mathsf{T}C$ and $b := \alpha \mathbb{1}^{2n} - 2C^\mathsf{T}y$, the above problem is equivalent to (dropping the $\langle y, y \rangle$ term)

$$\min_{z \in \mathbb{R}^{2n}_+} \tfrac{1}{2} \langle Bz, z \rangle + \langle b, z \rangle.$$

We now have a **quadratic programming problem** in standard form that can be solved with the MATLAB commands

[m] The FISTA algorithm [5] is more efficient and simple to program, but its analysis is beyond our scope.

```
C = [ A -A ];
b = repmat(alpha,2*n,1)-2*(C.'*y);
B = 2*(C.')*C;
z = quadprog(B,b,[],[],[],[],zeros(2*n,1));
x = z(1:n)-z(n+1:end);
```

Example 7.39 (Continuation of Example 7.38). Instead of choosing `c` to minimize (7.28), we now choose `c` to minimize

$$\|\mathtt{xvec} - \mathtt{A} * \mathtt{c}\|^2 + \alpha\|\mathtt{c}\|_1$$

with $\alpha = 0.2$. The resulting approximation is the dashed line in the bottom graph of Figure 7.1. Since the approximations using 1-norm and 2-norm regularization are so similar, why should we prefer one over the other? The answer lies in looking at the resulting vector of coefficients. Figure 7.2 shows the coefficient vectors `c` recovered using 2-norm regularization (top) and 1-norm regularization (bottom). In the bottom graph, notice that most of the coefficients have absolute values that are several orders of magnitude smaller than the largest ones at the top of the graph. Let K denote the indexes k of the 14 largest `abs(c(k))`. Then a graph of

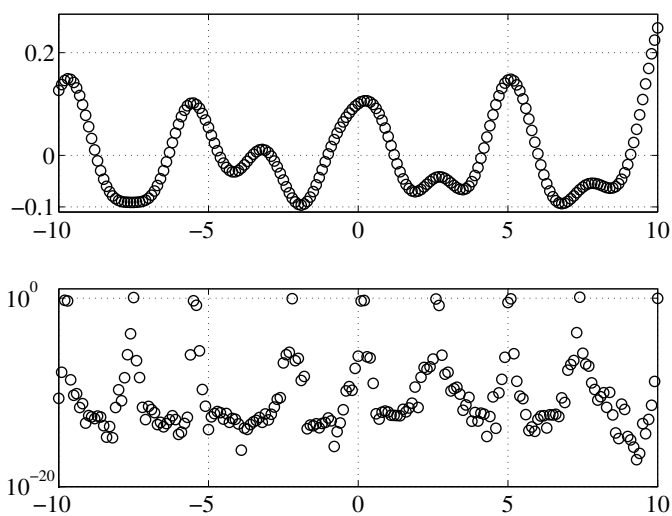


Figure 7.2. Coefficients `c(k)` from 2-norm regularization (top) and 1-norm (bottom). The horizontal coordinates are the shifts `tau(k)` (see Example 3.11).

$$\sum_{k\in K} c_k v(t-\tau_k)$$

is visually indistinguishable from the dashed line in Figure 7.1. Hence, 1-norm regularization can result in **sparse approximation**.

7.7. Numerical Methods

7.7.1. Gaussian Quadrature

If w is a nonnegative function defined on a given time interval and $0 < \int w(t)\,dt < \infty$, then we call w a **weight function**. Typical weight functions include

$$\begin{aligned} w(t) &= 1 && \text{on } [-1,1] \text{ or on } [0,1], \\ w(t) &= 1/\sqrt{1-t^2} && \text{on } [-1,1], \\ w(t) &= e^{-t} && \text{on } [0,\infty), \\ w(t) &= e^{-t^2} && \text{on } (-\infty,\infty). \end{aligned}$$

To approximate an integral of the form $\int x(t)w(t)\,dt$, we often use **numerical integration** or **quadrature** formulas having the structure

$$\int x(t)w(t)\,dt \approx \sum_{i=1}^{n} w_i x(t_i),$$

where the coefficients w_i are called **weights**, and the evaluation points t_i are called **nodes**. The **trapezoidal rule** and **Simpson's rule** fall in to this category on any interval $[a,b]$ with $w(t) = 1$. More generally, if the nodes are fixed, then it is easy to choose the weights so that the integral on the left and the sum on the right are exactly equal whenever x is a polynomial of degree less than n. However, if the nodes are chosen carefully, then we can achieve equality for all polynomials of degree less than $2n$; this is called ***Gaussian*** **quadrature**.

The theory of **orthogonal polynomials** says that if we apply the Gram–Schmidt procedure to the power functions $1, t, t^2, \ldots$ using the inner product

$$\langle x, y \rangle := \int x(t)\overline{y(t)}w(t)\,dt,$$

then the resulting orthonormal polynomials, which we call $\psi_0, \psi_1, \ldots$, are such that ψ_i has degree i and has i *real* roots. It is not hard to show that because we are dealing with polynomials, the Gram–Schmidt procedure simplifies to a three-term recursion of orthogonal polynomials, which we denote by φ_i. Of course, $\psi_i = \varphi_i / \|\varphi_i\|$.

The theory of Gaussian quadrature says that the required nodes are the n real roots of ψ_n. However, rather than using the recursion to generate the required polynomial and then finding its roots, the following theorem allows us to obtain both the nodes and the weights directly from the diagonalization of a simple matrix.

Theorem 7.40. *The nodes t_i and weights w_i of a Gaussian quadrature formula, based on orthogonal polynomials relative to a weight function w and generated by a three-term recursion with coefficients a_n and b_n, can be obtained from the eigenvalue decomposition of the symmetric, tridiagonal* ***Jacobi matrix***

$$J_n := \begin{bmatrix} a_0 & \sqrt{b_1} & & & \\ \sqrt{b_1} & a_1 & \sqrt{b_2} & & \\ & \sqrt{b_2} & \ddots & \ddots & \\ & & \ddots & a_{n-2} & \sqrt{b_{n-1}} \\ & & & \sqrt{b_{n-1}} & a_{n-1} \end{bmatrix}.$$

If $P^{\mathsf{H}} J_n P = \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, where $P^{\mathsf{H}} P = I$, then $t_i = \lambda_i$ and $w_i = p_i^2 \int w(t)\,dt$,[n] p_i is the first component of the ith column of P.

Proof. See the Notes[3] at the end of the chapter. □

Example 7.41. The a_n and b_n can be found in tables, e.g., [16, p. 29].[o] For the weight function $w(t) = 1$ on $[0,1]$, the three-term recursion with $a_k = 1/2$, $b_0 = 1$, and $b_k = 1/(4(4-k^{-2}))$ for $k \geq 2$ generates the ***shifted*** **Legendre polynomials** and leads to **Legendre–Gauss quadrature** on $[0,1]$. This suggests that we write the following MATLAB function to generate the nodes t_i and the weights w_i for the approximation

$$\int_0^1 x(t)\,dt \approx \sum_{i=1}^{n} w_i x(t_i).$$

```
function [t,w] = legendrequad01(n)
%
% Generate nodes and weights for
% Legendre-Gauss quadrature on [0,1].
% Note that t is a column vector
% and w is a row vector.
%
a = repmat(1/2,1,n);                  % diagonal of J
u = sqrt(1./(4*(4-1./[1:n-1].^2))); % upper diag of J
[P,Lambda] = eig(diag(u,1)+diag(a)+diag(u,-1));
[t,i] = sort(diag(Lambda));      % Sort roots in incr order
Ptop = P(1,:);                   % Extract top row of P and
Ptop = Ptop(i);                  % reorder to go with roots
w = Ptop.^2;
```

[n] Following Gautschi [16, p. 11], we put $b_0 := \int w(t)\,dt$.
[o] Gautschi [16] writes α_k and β_k for our a_k and b_k.

Suppose you have a MATLAB function `x`, and for a suitable value of `n`, you use the statement `[t,w]=legendrequad01(n)`. Since `t` is a column vector, `xvec = x(t)` is also a column vector. Then `w*xvec` multiplies the row vector $[w_1,\dots,w_n]$ and the column vector $[x(t_1),\dots,x(t_n)]^{\mathsf{T}}$ to produce the scalar value $\sum_{i=1}^{n} w_i x(t_i) \approx \int_0^1 x(t)\,dt$. Of course, many times we are interested in computing a whole family of integrals such as

$$y(s) := \int_0^1 x(t,s)\,dt.$$

To compute $y(s_j)$ for $j = 1,\dots,k$, we use the approximation

$$y(s_j) \approx \sum_{i=1}^{n} w_i x(t_i, s_j),$$

where n is chosen large enough for the approximation to work well for all values of s_j. To calculate the above sum efficiently in MATLAB, it is convenient to observe that it corresponds to multiplying the row vector $[w_1,\dots,w_n]$ by the $n \times k$ matrix whose $i\,j$ entry is $x(t_i,s_j)$. Denoting this matrix by `X`, we can compute the row vector `y` $= [y(s_1),\dots,y(s_k)]$ with the one-line statement `y=w*X`.

7.7.2. Eigenvalues and Eigenvectors of Integral Operators

Consider an integral operator of the form

$$(Ax)(t) = \int a(t,\tau)x(\tau)\,d\tau.$$

Given x, it is natural to approximate $(Ax)(t)$ with a numerical integration formula. For example, we could write

$$(Ax)(t) \approx \sum_{j=1}^{n} w_j a(t,t_j)x(t_j),$$

where the w_j and the t_j depend on the numerical integration technique, e.g., the **trapezoidal rule**, **Simpson's rule**, or **Gaussian quadrature**.

Now suppose that in the above integral t and τ belong to a common interval, and we want to solve the eigenvalue problem $A\varphi = \lambda\varphi$. We consider instead the problem [2, eq. (3.4)]

$$\sum_{j=1}^{n} w_j a(t,t_j)\widetilde{\varphi}(t_j) = \lambda\,\widetilde{\varphi}(t). \tag{7.30}$$

Upon taking $t = t_i$, we obtain

$$\sum_{j=1}^{n} \underbrace{w_j a(t_i,t_j)}_{=:\, M_{ij}} \widetilde{\varphi}(t_j) = \lambda \widetilde{\varphi}(t_i).$$

This suggests that we solve the finite-dimensional matrix-vector eigenvalue problem $M\mathbf{z} = \lambda\mathbf{z}$ and use the elements of $\mathbf{z}$ as values of the eigenfunction $\widetilde{\varphi}(t_i)$. There are many ways to define $\widetilde{\varphi}(t)$ for $t \neq t_i$; e.g., linear interpolation. However, (7.30) itself suggests the so-called **Nyström extension** [2, p. 170]

$$\widetilde{\varphi}(t) := \frac{1}{\lambda} \sum_{j=1}^{n} w_j a(t,t_j) z_j, \tag{7.31}$$

where z_j is the jth component of $\mathbf{z}$.

Proposition 7.42. *If $(\lambda,\mathbf{z})$ is an eigenpair of M, and if $\widetilde{\varphi}$ is defined by* (7.31)*, then $\widetilde{\varphi}$ satisfies* (7.30).

Proof. Assume $M\mathbf{z} = \lambda\mathbf{z}$ for nonzero $\mathbf{z}$. Then (7.31) implies $\widetilde{\varphi}(t_i) = (M\mathbf{z})_i/\lambda = (\lambda\mathbf{z})_i/\lambda = z_i$. Since i is arbitrary, we may replace z_j in (7.31) by $\widetilde{\varphi}(t_j)$; multiplying the result by λ shows that (7.30) holds. □

To make $\|\widetilde{\varphi}\| \approx 1$, observe that

$$\int |\widetilde{\varphi}(\tau)|^2 d\tau \approx \sum_{j=1}^{n} w_j |\widetilde{\varphi}(t_j)|^2 = \sum_{j=1}^{n} w_j |z_j|^2.$$

Dividing $\mathbf{z}$ by the square root of the right-hand side makes $\widetilde{\varphi}$ have energy one.

When $a(t,\tau)$ is real and symmetric; i.e., $a(t,\tau) = a(\tau,t)$, then the eigenvalue problem $M\mathbf{z} = \lambda\mathbf{z}$ has real eigenvalues and real eigenvectors.[p] In this case, $\widetilde{\varphi}$ is also real. After forcing $\widetilde{\varphi}$ to have unit energy, we can replace $\widetilde{\varphi}$ with $-\widetilde{\varphi}$ if we want to. For example, if are working on the interval $[0,1]$ we may want to require $\widetilde{\varphi}(0) > 0$. However, it may happen that $\widetilde{\varphi}(0) = 0$. Instead we may choose the sign of $\widetilde{\varphi}$ to be such that when plotting $\widetilde{\varphi}(t_i)$, we have $\widetilde{\varphi}(t_2) - \widetilde{\varphi}(t_1) \geq 0$; i.e., we make $\widetilde{\varphi}$ increasing at the origin.

To write a program in MATLAB that approximates the eigenvalues and eigenfunctions of an integral operator requires a little more attention to various details. If

[p] This can be seen as follows [2, p. 173]. The matrix $A_{ij} = a(t_i,t_j)$ is symmetric, but M is not. However, if we put $D = \mathrm{diag}(w_1,\ldots,w_n)$, then $M = AD$, and $M\mathbf{z} = \lambda\mathbf{z}$ if and only if $AD^{1/2}(D^{1/2}\mathbf{z}) = \lambda\mathbf{z}$. If we multiply this equation by $D^{1/2}$ and put $\mathbf{v} = D^{1/2}\mathbf{z}$, then we obtain $D^{1/2}AD^{1/2}\mathbf{v} = \lambda\mathbf{v}$.

$M\mathbf{z}_k = \lambda_k \mathbf{z}_k$, then we can obtain the λ_k and $Z = [\mathbf{z}_1|\cdots|\mathbf{z}_n]$ using the MATLAB command `eig(M)`. To use (7.31) to evaluate the kth eigenfunction at m points $t = \tau_i$ (not related to the quadrature nodes t_j) requires

$$\widetilde{\varphi}_k(\tau_i) = \frac{1}{\lambda_k} \sum_{j=1}^{n} w_j a(\tau_i, t_j) Z_{jk}.$$

If we let $\widetilde{\Phi}$ denote the matrix whose ik entry is $\widetilde{\varphi}_k(\tau_i)$, then

$$\widetilde{\Phi} = \begin{bmatrix} a(\tau_1,t_1) & \cdots & a(\tau_m,t_n) \\ \vdots & \ddots & \vdots \\ a(\tau_m,t_1) & \cdots & a(\tau_m,t_n) \end{bmatrix} \begin{bmatrix} w_1 & & \\ & \ddots & \\ & & w_n \end{bmatrix} Z \begin{bmatrix} \frac{1}{\lambda_1} & & \\ & \ddots & \\ & & \frac{1}{\lambda_n} \end{bmatrix}.$$

The formula for $\widetilde{\varphi}_k(\tau_i)$ shows that the columns of $\widetilde{\Phi}$ correspond to the columns of Z. In other words, by deleting columns of Z and the corresponding rows of $\operatorname{diag}(1/\lambda_1, \ldots, 1/\lambda_n)$, we can compute only some of the eigenfunctions. The point here is that once the eigenvalues and the Z matrix are computed, we can evaluate any of the eigenfunctions at any time points with a simple matrix equation.

Example 7.43. The function $\min(t,\tau)$ is the correlation function of the **Wiener process**. To employ the **Karhunen–Loève expansion** of the Wiener process on $[0,1]$ requires the solution of the eigenvalue problem

$$\int_0^1 \min(t,\tau)\varphi(\tau)\,d\tau = \lambda\varphi(t), \quad 0 \le t \le 1.$$

We can generate the nodes and weights for Legendre quadrature on $[0,1]$ using our MATLAB function given in Example 7.41. Using $n = 16$ and applying the Nyström method (see scripts in the Notes[4]), we obtained the results shown in Figures 7.3 and 7.4. The exact eigenvalues and eigenfunctions can be found in closed form (Problem 7.33).

Example 7.44 (Prolate Spheroidal Wave Functions). We begin with the eigenvalue problem on $L^2[-1,1]$,

$$\int_{-1}^{1} WT\,\operatorname{sinc}(WT(t-\tau))\varphi(\tau)\,d\tau = \lambda\varphi(t), \quad -1 \le t \le 1. \tag{7.32}$$

The left-hand side defines a compact, self-adjoint linear operator. From the discussion in Section 7.2.2, the operator is positive definite, which implies all of its eigenvalues are positive. Once an eigenpair (λ_k, φ_k) has been found, we put

$$\psi_k(t) := \frac{1}{\lambda_k} \int_{-1}^{1} WT\,\operatorname{sinc}(WT(t-\tau))\varphi_k(\tau)\,d\tau, \quad -\infty < t < \infty. \tag{7.33}$$

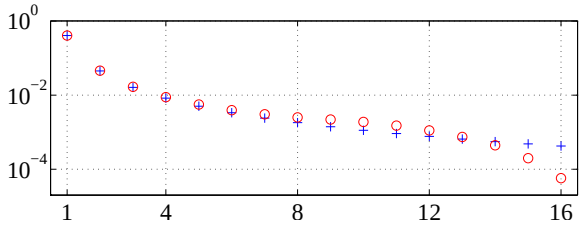


Figure 7.3. Exact eigenvalues (+) and approximate eigenvalues (o) for Example 7.43.

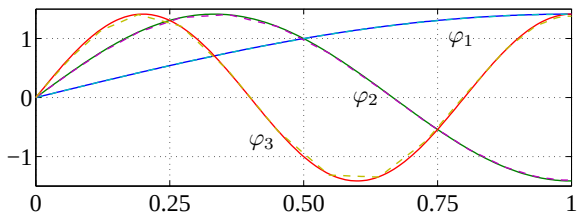


Figure 7.4. Exact eigenfunctions (solid lines), approximate eigenfunctions (dashed lines) for Example 7.43.

Of course, when $|t| \le 1$, the right-hand side is just $\varphi_k(t)$. In other words, we have extended the definition of φ_k to the whole real line. Hence, we can also write

$$\psi_k(t) = \frac{1}{\lambda_k} \int_{-1}^{1} WT \operatorname{sinc}(WT(t-\tau))\psi_k(\tau)\,d\tau, \quad -\infty < t < \infty.$$

The functions ψ_k are called angular **prolate spheroidal wave functions**. They depend on the bandwidth parameter W and the time-duration parameter T only through their product WT. Traditionally, however, these functions are parameterized by $c := \pi WT$, and the eigenvalues and eigenfunctions are indexed starting from 0 rather than 1. In addition, the φ_k are normalized so that $\int_{-1}^{1} |\varphi_k(t)|^2\,dt = \lambda_k$.

When $c = 1$, we used 8-point Legendre–Gauss quadrature on $[-1, 1]$ (see Problem 7.69) and found

$$\begin{aligned} \lambda_0 &= 5.7258178 \times 10^{-1} & \lambda_4 &= 3.717929 \times 10^{-8} \\ \lambda_1 &= 6.2791274 \times 10^{-2} & \lambda_5 &= 9.4914 \times 10^{-11} \\ \lambda_2 &= 1.2374793 \times 10^{-3} & \lambda_6 &= 1.67 \times 10^{-13} \\ \lambda_3 &= 9.200977 \times 10^{-6} & \lambda_7 &= 2. \times 10^{-16} \end{aligned}$$

which agrees with [39] when rounded to the number of digits shown. To obtain similar accuracy for $c = 8$ requires 12-point quadrature. Examples of the ψ_k are shown in Figure 7.5.

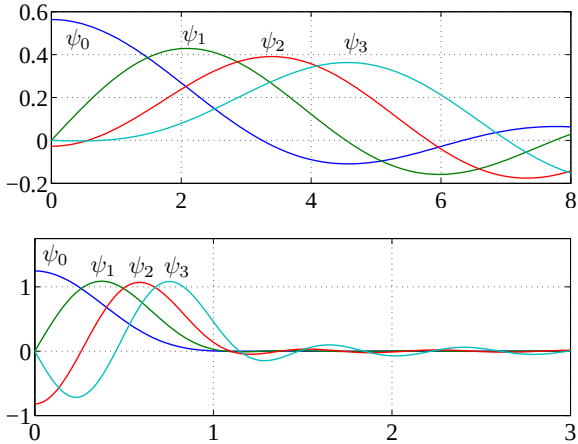


Figure 7.5. Prolate spheroidal wave functions for $c = 1$ (top) and $c = 8$ (bottom).

It is also interesting to plot the eigenvalues λ_k corresponding to different values of $2WT = 2c/\pi$ as shown in Figure 7.6. We see that for all k sufficiently less than $2WT$, the λ_k are nearly one, and for all k sufficiently larger than $2WT$, the λ_k are negligible. This property has been proved mathematically [24].

7.7.3. Solving Second-Kind Integral Equations

A trivial modification to the foregoing allows us to approximately solve $(I + A)x = y$ using the **Nyström** or **quadrature** method. The approximate equation is

$$x(t) + \sum_{j=1}^{n} w_j a(t,t_j) x(t_j) = y(t). \tag{7.34}$$

Taking $t = t_i$ yields

$$x(t_i) + \sum_{j=1}^{n} w_j a(t_i,t_j) x(t_j) = y(t_i).$$

This suggests that we solve the finite-dimensional matrix-vector problem $(I_n + M)\mathbf{z} = \mathbf{y}$, where I_n is the $n \times n$ identity matrix and $\mathbf{y} := [y(t_1), \ldots, y(t_n)]'$. Once $\mathbf{z}$ is obtained, define the continuous-time Nyström extension [2, p. 357]

$$\tilde{x}(t) := y(t) - \sum_{j=1}^{n} w_j a(t,t_j) z_j.$$

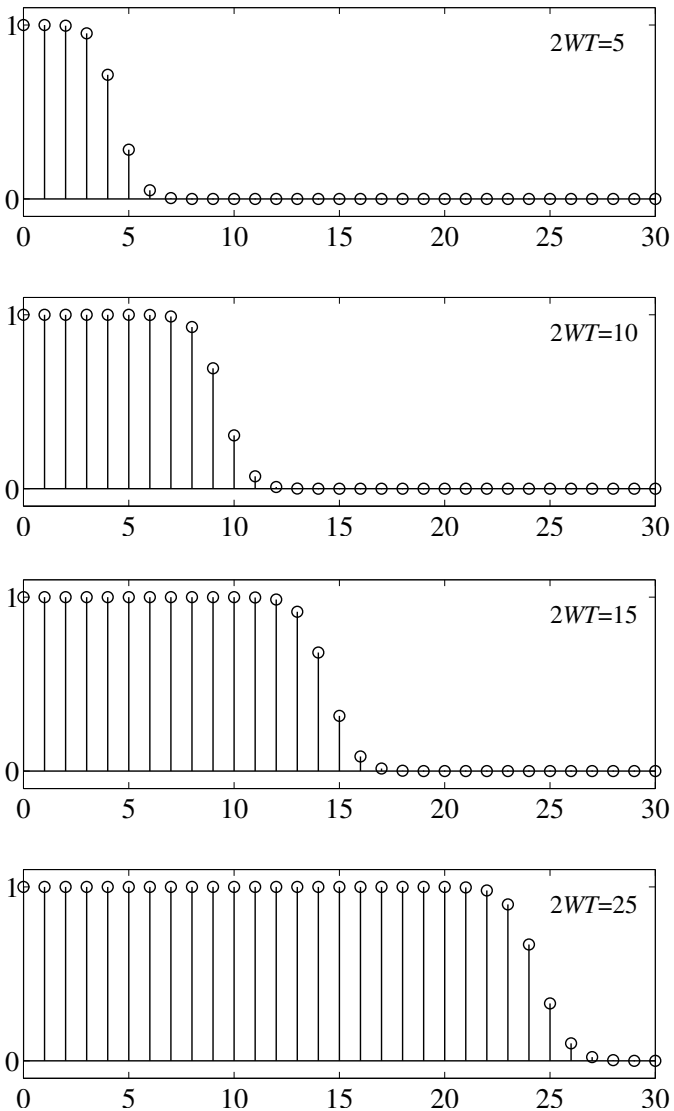


Figure 7.6. Eigenvalues λ_k corresponding to prolate spheroidal wave functions ψ_k for different values of $2WT = 2c/\pi$.

When $t = t_i$, this says that the column vector of samples of $\tilde{x}$ is equal to $\mathbf{y} - M\mathbf{z} = \mathbf{z}$; i.e., $\tilde{x}(t_i) = z_i$. Hence, we may replace z_j by $\tilde{x}(t_j)$ in the definition of $\tilde{x}$. This shows that continuous-time Nyström extension $\tilde{x}$ solves (7.34). It can be shown that if one considers a sequence of improving numerical integration schemes, then the solutions of (7.34) converge to the solution of $(I+A)x = y$ [23, Ch. 12].

Remark. The Nyström or quadrature methods used in this section and the preceding one are the easiest to implement. There are several other approaches discussed in the references [2], [11], [23], [46].

Notes

Note 7.1. The following result shows that Fourier transforms of L^1 time functions are uniformly continuous functions of frequency.

Theorem 7.45. *If $\int_{-\infty}^{\infty} |h(t)|\,dt < \infty$, then $H(f) := \int_{-\infty}^{\infty} h(t)e^{-j2\pi ft}\,dt$ is a uniformly continuous function of f.*

Proof. Write

$$\begin{aligned} |H(f+\nu) - H(f)| &= \left| \int_{-\infty}^{\infty} h(t)\left[e^{-j2\pi(f+\nu)t} - e^{-j2\pi ft}\right] dt \right| \\ &= \left| \int_{-\infty}^{\infty} h(t)e^{-j2\pi ft}\left[e^{-j2\pi\nu t} - 1\right] dt \right| \\ &\le \int_{-\infty}^{\infty} |h(t)|\left|e^{-j2\pi\nu t} - 1\right| dt. \end{aligned}$$

Since this last integrand is bounded by $2|h(t)|$, the integral tends to zero as $\nu \to 0$ by the dominated convergence theorem [6], [14], [33], [34]. □

Note 7.2. The following result is often useful in showing an integral is positive.

Theorem 7.46. *Suppose $\int_a^b x(t)\,dt = 0$. If x is continuous and nonnegative on $[a,b]$, then $x(t) = 0$ for all $t \in [a,b]$.*

Proof. Suppose otherwise that there is some $t_0 \in [a,b]$ with $x(t_0) > 0$. Since x is continuous at t_0, we know that for every $\varepsilon > 0$, there is a $\delta > 0$ such that for all $t \in [a,b]$ with $|t - t_0| < \delta$, $|x(t) - x(t_0)| < \varepsilon$. Rewrite this as

$$-\varepsilon < x(t) - x(t_0) < \varepsilon,$$

and rearrange the left-hand inequality as $x(t) > x(t_0) - \varepsilon$. Since $x(t_0) > 0$, we can take $\varepsilon = x(t_0)/2$ to get $x(t) > x(t_0)/2$ for all t within δ of t_0. We can now write

$$\int_a^b x(t)\,dt \geq \int_{t_0-\delta}^{t_0+\delta} x(t)\,dt \geq \int_{t_0-\delta}^{t_0+\delta} \frac{x(t_0)}{2}\,dt = \delta x(t_0) > 0.$$

If t_0 is an endpoint, the reader can modify the above inequality accordingly. □

Note 7.3. Proof of Theorem 7.40. (Sketch). The first step is to rewrite the **three-term recursion** $\varphi_0(t) = 1$, $\varphi_1(t) = (t - a_0)\varphi_0(t) = t - a_0$, and

$$\varphi_k(t) = (t - a_{k-1})\varphi_{k-1}(t) - b_{k-1}\varphi_{k-2}(t), \quad k \geq 2,$$

in terms of the ortho*normal* polynomials $\psi_k := \varphi_k/\|\varphi_k\|$. This leads to

$$\|\varphi_k\|\psi_k(t) = (t - a_{k-1})\|\varphi_{k-1}\|\psi_{k-1}(t) - b_{k-1}\|\varphi_{k-2}\|\psi_{k-2}(t).$$

Divide this equation by $\|\varphi_{k-1}\|$ and use the fact that $\sqrt{b_k} = \|\varphi_k\|/\|\varphi_{k-1}\|$ to obtain

$$\sqrt{b_k}\psi_k(t) = (t - a_{k-1})\psi_{k-1}(t) - \sqrt{b_{k-1}}\psi_{k-2}(t).$$

Rearrange this as

$$t\psi_{k-1}(t) = \sqrt{b_k}\psi_k(t) + a_{k-1}\psi_{k-1}(t) + \sqrt{b_{k-1}}\psi_{k-2}(t).$$

If we write out this formula for $k = 1, \ldots, n$, we get a system of n linear equations. To express this in matrix-vector notation, put

$$\Psi(t) := [\psi_0(t), \ldots, \psi_{n-1}(t)]^{\mathsf{T}}.$$

Then the system of linear equations can be written as

$$t\Psi(t) = J_n\Psi(t) + \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \sqrt{b_n}\psi_n(t) \end{bmatrix}.$$

Now, if $t = t_i$ is the ith root of φ_n, which is also the ith root of $\psi_n = \varphi_n/\|\varphi_n\|$, then the matrix-vector equation reduces to

$$t_i\Psi(t_i) = J_n\Psi(t_i),$$

which says that t_i is an eigenvalue of J_n with eigenvector $\Psi(t_i)$. Since by definition eigenvectors cannot be the zero vector, we should check this condition. It can be

shown that (we omit the proof) $(\sqrt{w_i}\Psi(t_i))^{\mathsf{T}}(\sqrt{w_i}\Psi(t_i)) = 1$. Hence, $\sqrt{w_i}\Psi(t_i)$ is a unit-norm eigenvector of J_n. Since we are working in a real vector space, $\sqrt{w_i}\Psi(t_i)$ must be equal to plus or minus the ith column vector of P. Since $\sqrt{w_i}\Psi(t_i) = \pm$ the ith column of P, their first components must obey this relation too. Since the first component of $\Psi(t_i)$ is $\psi_0(t_i) = 1/\|\varphi_0\|$, the theorem is proved. □

Remark. An easy corollary of Theorem 7.40 is that

$$\varphi_n(t) = \det(tI - J_n). \tag{7.35}$$

The right-hand side is a polynomial of degree n with leading coefficient one and whose roots are the eigenvalues of J_n. Hence, the right-hand side is exactly $(t - t_1)\cdots(t - t_n) = \varphi_n(t)$. An alternative way to prove (7.35) is the expand the determinant along the last column of $tI - J_n$ to show that $\det(tI - J_n)$ satisfies the same three-term recurrence as φ_n; hence, $\det(tI - J_n) = \varphi_n(t)$.

According to Gautschi [15], the fact that the roots of φ_n are the eigenvalues of J_n was known prior to the 1960s. The relationship of the weights to the *orthonormal* eigenvectors of J_n is found in Wilf [45, Ch. 2, Ex. 9]. Gautschi also says that this fact was known to Goertzel around 1954 and appeared in Gordon [19] in 1968. It was Golub and Welsch [18] who provided an efficient algorithm for solving the eigenvalue problem for J_n to obtain the eigenvalues t_i and the weights w_i.

Note 7.4. Here is the MATLAB script for Example 7.43 along with the associated functions `eigNystrom` and `eigfcnNystrom`.

```
%% Nystrom method for finding eigenvalues and eigenvectors
T = 1;  % Functions live on [0,T]
n = 16; % number of quadrature points
fprintf('%g-pt Legendre quadrature requested on [0,1]\n',n)
a = @(t,s)min(t,s); % kernel of operator
%% Get eigenvalues and auxiliary data
[lambda,Z,s,w] = eigNystrom(a,T,n);
%% Print eigenvalues
fprintf('Eigenvalues:\nindex  eigenvalue\n')
k = 1:length(lambda); % indexes of eigenvalues to print
emat = [ k ; lambda(k).' ];
fprintf('%2i %13.7e\n',emat)
%% Plot eigenvalues
subplot(2,1,1)
semilogy(k,lambda(k),'o'); grid on
set(gca,'GridLineStyle',':','GridAlpha',.6)
%% Now compute and plot a few eigenfunctions
t = linspace(0,T,200); % range to plot eigenfunctions
```

```
k = 1:3; % indexes of eigenfunctions to be plotted
fmat = eigfcnNystrom(k,t,a,T,lambda,Z,s,w);
subplot(2,1,2)
plot(t,fmat); grid on
set(gca,'GridLineStyle',':','GridAlpha',.6)


function [lambda,Z,s,w] = eigNystrom(kernel,T,n)
%
% kernel = either a string with the name of the kernel fcn
%          or the handle of an anonymous kernel function.
%
% T      = functions live on [0,T].
%
% n      = number of requested quadrature points.
%
% Consider the eigenvalue problem on [0,T]
%
% int_0^T a(t,s) x(s) ds = lambda x(t),   0 < t < T,
%
% which is equivalent to (let  s'=s/T; ds'=ds/T)
%
% int_0^1 T a(Tt',Ts') y(s') ds' = lambda y(t'), 0 < t' < 1,
%
% where y(s') := x(Ts') or x(s)=y(s/T) and t' := t/T.
%
% Using Legendre-Gauss quadrature on [0,1],
% we have the approximate problem
%
% sum_j w_j Ta(Ts_i,Ts_j) y(s_j)=lambda y(s_i), i,j=1,...,n.
%
% Let M denote the matrix with entries T a(Ts_i,Ts_j) w_j,
% and solve M z = lambda z.  Put eigenvectors (columns)
% into Z and corresponding eigenvalues into lambda, sorted
% largest to smallest.

[s,w] = legendrequad01(n);
simat = repmat(T*s,1,n);
sjmat = simat.';
wmat = repmat(T*w,n,1);
M = feval(kernel,simat,sjmat).*wmat;
[Z,Lambda] = eig(M);
lambda = real(diag(Lambda)); % Assume self-adjoint operator
[lambda,k] = sort(lambda,'descend'); % Sort eigenvalues
Z = Z(:,k); % from largest to smallest; correspondingly
```

```
            % rearrange eigenvectors.

% Normalize so that eigenfunctions will have unit energy
E2vec = w*(Z.*conj(Z));
nmat = repmat(1./sqrt(E2vec*T),n,1);
Z = Z.*nmat;

% If necessary, multiply eigenfunctions by -1 to make them
% increasing at the origin.
factor = sign(Z(2,k)-Z(1,k));
facmat = repmat(factor,n,1);
Z = facmat.*Z;


function fmat = eigfcnNystrom(k,t,kernel,T,lambda,Z,s,w)
%
% k = vector of indexes of eigenfunctions that you
%     want to evaluate.
%
% t = array of times at which you want to evaluate selected
%     eigenfunctions.
%
% kernel = either a string with the name of the kernel fcn
%          or the handle of an anonymous kernel function.
%
% T      = functions live on [0,T].
%
% [lambda,Z,s,w] = eigNystrom(kernel,T,n).
%
% Then the kth column of fmat contains the values of
% the kth eigenfunction evaluated at the times in t:
%
% phi_k(t_i)/lambda_k = sum_j T w_j a(t_i,T s_j) Z_{j,k}

sizt = size(t);          % Convert t to column vector
ltt = prod(sizt);
tt = reshape(t,ltt,1);
ZZ = Z(:,k);             % extract needed columns from Z
slen = length(s);        % = num rows of ZZ
% extract corresponding eigenvalues and make matrix
lamlami = repmat(1./(lambda(k).'),ltt,1); % for later
timat = repmat(tt,1,slen);
sjmat = repmat(T*s.',ltt,1);
wmat = repmat(T*w,ltt,1);
Mmat = feval(kernel,timat,sjmat).*wmat;
```

```
fmat = (Mmat*ZZ).*lamlami;
```

Problems

1. For $x \in C[0,1]$, the formula for the uniform norm is $\|x\| := \max_{0\le t\le 1} |x(t)|$. Show that this formula satisfies the properties of a norm given in Section 6.2.

2. Recall the point-evaluation linear functional f defined for $x \in C[0,1]$ by $f(x) := x(0)$. Show that f is not continuous when $C[0,1]$ is equipped with the integral norm
$$\|x\| := \int_0^1 |x(t)|\,dt.$$
Hint: Consider the sequence $f(x_n)$, where
$$x_n(t) := \begin{cases} 1-nt, & 0 \le t \le 1/n, \\ 0, & 1/n < t \le 1. \end{cases}$$
You may find it helpful to sketch $x_n(t)$. Show that if $x(t) := 0$ for all $t \in [0,1]$, then $\|x_n - x\| \to 0$.

3. Let $C[0,1]$ be equipped with the same integral norm as in the previous problem. For $n = 1, 2, \ldots$, define the linear functional
$$f_n(x) := n \int_0^{1/n} x(\tau)\,d\tau.$$
   (a) For fixed n, is f_n a bounded linear functional?
   (b) Define a new linear functional
   $$f(x) := \lim_{n\to\infty} f_n(x).$$
   Determine whether or not f is a bounded linear functional.

4. Show that the kernel of a continuous linear functional defined on a normed vector space is closed.

5. Let f and g be linear functionals on an arbitrary vector space X over $\mathbb{C}$. Assume that $\ker g \subset \ker f$. Prove that there exists a scalar λ such that $f(x) = \lambda g(x)$ for all $x \in X$. Do *not* assume X is an inner product space. *Hint:* If $\ker g \ne X$, there exists a $z \in X$ with $g(z) \ne 0$. Now fix any $x \in X$ and consider the vector
$$w := x - \frac{g(x)}{g(z)} z.$$

6. If $f(x) = \langle x, y\rangle$, show that $\|f\| = \|y\|$.

7. Show that every linear functional on a finite-dimensional normed vector space is bounded. *Hint:* Lemma 6.45 may be helpful.

8. Let X denote the real inner product space $L^2[0,\infty)$. Define the linear operator $A: X \to X$ by
$$(Ax)(t) := \int_0^\infty e^{-(t+\tau)} x(\tau)\, d\tau, \quad t \geq 0.$$
Find $\|A\|$.

9. Let X denote the set of bounded sequences of real numbers. For vectors $x = (x_1, x_2, \ldots) \in X$, define the norm
$$\|x\| := \sup_j |x_j| < \infty.$$
Define the operator A on X as follows. For $i, j = 1, 2, \ldots$, let a_{ij} be nonnegative, and assume that for each i, $\sum_{j=1}^\infty a_{ij} = 1$. For $x \in X$, define Ax by the formula
$$(Ax)_i := \sum_{j=1}^\infty a_{ij} x_j, \quad i = 1, 2, \ldots.$$
Show that $A: X \to X$, and find $\|A\| := \sup_{x \neq 0} \|Ax\|/\|x\|$.

10. ***Matrix Norms, Part 1.*** This problem illustrates how the operator norm of $A: X \to Y$ depends on the norms chosen for X and Y. *Hint:* The suggestions in the gray box following (7.1) for determining the norm of a bounded linear functional apply more generally to finding the norms of bounded linear operators.

    (a) Using the 1-norm on m- and n-dimensional space, show that the operator norm of an $m \times n$ matrix a is
    $$\|a\| = \max_{1 \leq j \leq n} \sum_{i=1}^m |a_{ij}|.$$
    Hint: If the above maximum is achieved by $j = k$, consider the value of $\|ae_k\|_1$, where e_k is the kth standard unit vector in n-dimensional space. To compute this norm in MATLAB, use the command `norm(a,1)`.

    (b) Using the infinity norm on m- and n-dimensional space, show that the operator norm of an $m \times n$ matrix a is
    $$\|a\| = \max_{1 \leq i \leq m} \sum_{j=1}^n |a_{ij}|.$$

Hint: If the above maximum is achieved by $i = k$, consider the value of $\|ax\|_\infty$, where $x_j = \overline{a}_{kj}/|a_{kj}|$, provided the denominator is nonzero; if it is, then x_j can be arbitrary, but take $x_j = 1$ to maintain $\|x\|_\infty = 1$. To compute this norm in MATLAB, use the command `norm(a,inf)`.

(c) Using the infinity norm on m-dimensional space and the 1-norm on n-dimensional space, show that

$$\|a\| = \max_{1\le i\le m, 1\le j\le n} |a_{ij}|.$$

The command `max(max(abs(a)))` computes this norm in MATLAB.

11. ***Matrix Norms, Part 2.*** Let a be an $m \times n$ matrix. If x is an n-dimensional vector, show that

$$\|ax\|^2 \le \left(\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2\right)\|x\|^2,$$

where $\|\cdot\|$ denotes the usual Euclidean 2-norm on m-dimensional and n-dimensional space as appropriate.

Remark. By identifying $m \times n$ matices with column vectors of length mn, it is clear that

$$\|a\|_F := \left(\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2\right)^{1/2}$$

defines on norm on the space of $m \times n$ matrices. This norm is called the **Frobenius norm**. Alternatively, since $\|a\|_F = \sqrt{\operatorname{tr}(aa^{\mathsf{H}})}$, the Frobenius norm is induced by the usual inner product on $m \times n$ matrices, $\langle a, b\rangle := \operatorname{tr}(ab^{\mathsf{H}})$. To compute the Frobenius norm in MATLAB, use the command `norm(a,'fro')`. Note also that since $\|ax\| \le \|a\|_F \|x\|$, the operator norm

$$\|a\| := \sup_{x\ne 0} \frac{\|ax\|}{\|x\|} \le \|a\|_F.$$

As we know from the SVD, the left-hand side is the largest singular value of a. To compute this norm in MATLAB, use the command `norm(a)` or `norm(a,2)`.

12. Let ℓ^2 denote the Hilbert space of square summable sequences of complex numbers. For $x = (x_1, x_2, \ldots)$ and $y = (y_1, y_2, \ldots) \in \ell^2$,

$$\langle x, y\rangle := \sum_{k=1}^{\infty} x_k \overline{y_k}, \quad \text{and} \quad \|x\|^2 = \langle x, x\rangle = \sum_{k=1}^{\infty} |x_k|^2.$$

Define a mapping $A\colon \ell^2 \to \ell^2$ by $Ax := (x_1, x_2/2, x_3/3, \ldots)$. In other words, $(Ax)_k = x_k/k$.

(a) Prove that A is a bounded operator, and find $\|A\|$.

(b) For $y \in \text{range}A$, define $By := (y_1, 2y_2, 3y_3, \ldots)$; i.e., $(By)_k = ky_k$. Clearly, $B(Ax) = x$ for all $x \in \ell^2$. Hence, B: range$A \to \ell^2$. Show that B is an unbounded operator.

(c) Show that B does not have an adjoint. *Hint:* Suppose otherwise that for all $x \in \ell^2$, there exists a unique vector $B^*x \in \text{range}A$ with

$$\langle By, x\rangle = \langle y, B^*x\rangle, \quad \text{for all } y \in \text{range}A.$$

If $y^{(n)}$ converges, then so does $\langle y^{(n)}, B^*x\rangle$. However, you can construct a sequence $y^{(n)} \in \text{range}A$ that converges to zero and for which $\langle By^{(n)}, x\rangle$ diverges for a particular choice of $x \in \ell^2$.

13. Let f be a bounded linear functional on a Hilbert space.

(a) Find the adjoint of f.

(b) Assuming $f \neq 0$, show that $\dim(\ker f)^\perp = 1$.

14. Show that if W is a subspace of a Hilbert space X, then $(W^\perp)^\perp = \overline{W}$. *Hint:* By the Projection Theorem for Hilbert Space, $X = \overline{W} \oplus (\overline{W})^\perp$, and so by Problem 3.14 we have $[(\overline{W})^\perp]^\perp = \overline{W}$. Now show that $(\overline{W})^\perp = W^\perp$.

15. Let A: $X \to Y$, and assume A^* exists.

(a) If A^* is onto, show that A is nonsingular.

(b) If Y is a Hilbert space, A^* is nonsingular, and range$A$ is closed, show that A is onto.

16. A mapping between metric spaces that preserves distance is called an **isometry**. Hence, a mapping between normed vector spaces that preserves norms is an isometry. Let X and Y be inner-product spaces, and assume A: $X \to Y$ has an adjoint. Then by Problem 4.15, A is an isometry if and only if $A^*A = I$. The operator A is called **unitary** if $A^*A = I$ and $AA^* = I$.

(a) Show that A is unitary if and only if A is an isometry and A is onto.

(b) Show that A is unitary if and only if $A^* = A^{-1}$.

(c) Let X denote the space of all finite-energy time functions, and let Y denote the space of all finite-energy frequency functions. Using the formulas

$$(Ax)(f) = \int_{-\infty}^{\infty} x(t)e^{-j2\pi ft}\,dt \quad \text{and} \quad (A^{-1}X)(t) = \int_{-\infty}^{\infty} X(f)e^{j2\pi ft}\,df$$

show that the Fourier transform A is a unitary operator; i.e., show that $A^* = A^{-1}$.

(d) Referring to Problem 3.15, show that a **Householder transformation** is unitary.

(e) Let A be the matrix in the Remark of Problem 4.5. Show that A is an isometry, but A is not unitary.

17. Let X and Y be normed vector spaces, and let $A:X \to Y$ be a bounded linear operator. Assume that A is invertible and that A^{-1} is bounded. Show that if W is a closed subset of X, then $\{Aw : w \in W\}$ is a closed subset of Y.

18. Let X and Y be normed vector spaces, and let $A:X \to Y$ be a bounded linear operator. Assume that A is invertible. Do not assume A^{-1} is bounded. Show that A^{-1} is a **closed operator** in the sense that whenever y_n is a convergent sequence in Y with limit y, and $A^{-1}y_n$ is a convergent sequence in X with limit x, then $A^{-1}y = x$.

Remark. If X and Y are both Banach spaces, then A^{-1} is continuous (and therefore bounded) by the **closed graph theorem** [17, p. 221].

19. Suppose $x \in L^2[a,b]$. By Hölder's inequality, $x \in L^1[a,b]$, and by Theorem 7.45 in the Notes, $X(f)$ is continuous. However, because $x \in L^2[a,b]$, here is an alternative proof of the continuity of $X(f)$. The proof of Theorem 7.45 used the inequality $|e^{j\theta} - 1| \le 2$. Instead, use the inequality

$$|e^{j\theta} - 1| = \left| j\theta \int_0^1 e^{j\theta\tau}\, d\tau \right| \le |\theta|$$

and apply Hölder's inequality.

20. Suppose $X(f) = \int_a^b x(t)e^{-j2\pi ft}\, dt$ for some $x \in L^1[a,b]$. Show that if $X(f) = 0$ on an interval, say $(f_0 - \delta, f_0 + \delta)$, then $X(f) = 0$ for all f. *Hint:* In the definition of $X(f)$, make the substitution

$$e^{-j2\pi ft} = e^{-j2\pi f_0 t} e^{-j2\pi(f-f_0)t}$$

and then expand the second factor on the right using the power series $e^z = \sum_{k=0}^{\infty} z^k/k!$. After a little rearrangement, identify an integral equal to $X^{(k)}(f_0)$. Then use the fact that the assumptions on X imply all its derivatives at f_0 are zero; i.e., $X^{(k)}(f_0) = 0$ for $k = 0, 1, \ldots$.

Remark. An easy consequence of this problem is that if a time-limited waveform is bandlimited, then its transform is identically zero, and therefore the waveform itself is zero. The dual result is that if a bandlimited waveform is time limited, then it is the zero waveform. Putting this all together shows that the only finite-energy waveform that is both time limited and bandlimited is the zero waveform.

21. ***When Projections Commute.*** Let M and N be closed subspaces of a Hilbert space X. Then $M \cap N$ is also a closed subspace of X. Show that $P_{M\cap N} = P_M P_N$ if and only if $P_M P_N = P_N P_M$.

22. Let X be a real or complex inner-product space, and let $A: X \to X$ be a linear operator whose adjoint A^* exists. If $A^*A = I$ and λ is an eigenvalue of A, determine whether or not $|\lambda| \leq 1$. Do not assume that A is a bounded operator.

23. If A is given by (7.5), find A^*. Then show that $A = A^*$ if and only if all the λ_k are real.

24. Consider the space $X := \mathbb{R}^9$ equipped with the standard Euclidean inner product. Let $\varphi_1, \ldots, \varphi_7$ be orthonormal vectors in X. Consider the mapping $A: X \to X$ defined by

$$Ax := \sum_{k=1}^{7} k\langle x, \varphi_k\rangle \varphi_k.$$

Find the distinct eigenvalues of A.

25. Let X denote the set of infinitely differentiable functions on $\mathbb{R}$, and for $x \in X$, let $D: X \to X$ be the derivative operator defined by $(Dx)(t) := \dot{x}(t)$, the ordinary derivative of x. Show that every real number λ is an eigenvalue of D, and find a corresponding eigenfunction.

26. Let X denote the set of all finite-energy waveforms on $[-\pi, \pi]$, and consider the operator $A: X \to X$ defined by

$$(Ax)(t) = \int_{-\pi}^{\pi} h(t-\tau)x(\tau)\,d\tau,$$

where h is a 2π-periodic function whose energy over one period is finite. Find the eigenvalues and eigenvectors of this operator, and find the representation analogous to (7.5).

27. Let X denote the set of all bounded functions on $(-\infty, \infty)$, and consider the operator $A: X \to X$ defined by

$$(Ax)(t) = \int_{-\infty}^{\infty} h(t-\tau)x(\tau)\,d\tau,$$

where $h \in L^1(\mathbb{R})$. Find an $h \in L^1(\mathbb{R})$ such that A has a continuum of eigenvalues, and for these eigenvalues, find corresponding eigenfunctions.

28. Show that the convolution operators of Theorems 7.10 and 7.19 are self adjoint if H is real. Show they are positive semidefinite if H is nonnegative.

29. Consider the convolution operator of Theorem 7.10.

    (a) Show that if H is real valued, then the operator is self adjoint.
    (b) Give an example for which no eigenvalues exist.
    (c) Give an example for which the only nonzero eigenvalue is one *and* such that the norm of the operator is greater than one.
    (d) Give an example for which there is a positive eigenvalue with infinitely many *orthogonal* eigenfunctions.

30. Prove Proposition 7.25.

31. Let X be a real or complex inner-product space, and let $A: X \to X$ be a self-adjoint linear operator. Suppose A has eigenvalues $\lambda_1, \lambda_2, \ldots$ with the property that $\lambda_n \to \infty$. Determine whether or not A is a bounded operator.

32. On finite-energy functions on $[0,T]$, consider the operator A defined by

$$(Ax)(t) := \int_0^T \min(t,\tau)x(\tau)\,d\tau, \quad 0 \le t \le T,$$

where $\min(t,\tau)$ denotes the smaller of t and τ; i.e.,

$$\min(t,\tau) := \begin{cases} \tau, & \tau \le t, \\ t, & \tau > t. \end{cases}$$

Show that A is positive definite. *Hints:* Evaluate $\langle Ax, x\rangle$ using integration by parts; it is convenient to put $u(t) := (Ax)(t)$ and $v(t) := -\int_t^T x(\theta)\,d\theta$. Clearly, $v'(t) = x(t)$, while expanding

$$u(t) = \int_0^t \tau x(\tau)\,d\tau + t\int_t^T x(\tau)\,d\tau$$

implies

$$u'(t) = \int_t^T x(\tau)\,d\tau = -v(t).$$

33. Find the eigenvalues and eigenfunctions of the operator A of the preceding problem. *Hints:*

    (*i*) By Problem 7.32, A is positive semidefinite, and so any of its eigenvalues must be nonnegative.
    (*ii*) Show that $(A\varphi)(t) = \lambda\varphi(t)$ expands to

$$\int_0^t \tau\varphi(\tau)\,d\tau + t\int_t^T \varphi(\tau)\,d\tau = \lambda\varphi(t). \tag{7.36}$$

(*iii*) Differentiate (7.36) with respect to t.

(*iv*) Differentiate your result in (*iii*) and obtain a differential equation for φ. Use (7.36) to solve for $\varphi(0)$. Use your result in (*iii*) to solve for $\dot{\varphi}(T)$. Now solve the differential equation for φ. What values of λ are permissible? Find all solutions λ_n and φ_n.

34. Let U and W be subspaces of an inner-product space X, and suppose that the projection operators onto U and W, denoted by P_U and P_W, exist. Define a new operator $A := P_U P_W$. Assuming that $U \cap W$ contains only the zero vector and that W is finite-dimensional, show that $\|A\| < 1$. *Hints:* The problem can be divided into three parts. First show that $\|A\| \leq 1$. Second, show that if $1 = \|A\| = \|Ax_0\|$ for some $x_0 \in W$ with $\|x_0\| = 1$, then a contradiction results. Third, show that $\|A\| = \|Ax_0\|$ for some $x_0 \in W$ with $\|x_0\| = 1$.

35. ***Levy–Desplanques Theorem.*** A square matrix A with entries a_{ij} is **diagonally dominant** if for each row i, $|a_{ii}| \geq \sum_{j \neq i} |a_{ij}|$. If the inequality is strict, then A is ***strictly*** **diagonally dominant**. Show that a strictly diagonally dominant matrix is nonsingular. *Hints:* Suppose $Ax = 0$ for some vector x. Let i be such that $|x_i| \geq |x_j|$ for all j. The ith row of $Ax = 0$ says that $0 = \sum_j a_{ij} x_j$. The inequality $|u + v| \geq |u| - |v|$ may be helpful.

36. ***Gershgorin Circle Theorem.*** Show that if a square matrix B has an eigenvalue λ, then for some i,

$$|\lambda - b_{ii}| \leq \sum_{j \neq i} |b_{ij}|.$$

In other words, λ lies in the **Gershgorin disc** centered at b_{ii} and having radius $\sum_{j \neq i} |b_{ij}|$. The usefulness of the theorem is that it gives bounds on the locations of eigenvalues in the complex plane. *Hint:* If $Bx = \lambda x$ for some *nonzero* x, then $A := (\lambda I - B)$ is singular and therefore cannot be strictly diagonally dominant.

37. Let $A: X \to Y$ be a compact operator, and suppose that z_n is a bounded sequence of nonzero vectors. Prove that there exists a sub-subsequence $z_{n_{k_i}}$ and a point $w \in Y$ such that $A z_{n_{k_i}} \to w$ as $i \to \infty$.

38. Show that A is a compact operator if and only if the closure of the image of the unit sphere, $\overline{\{Ax : \|x\| = 1\}}$, is a sequentially compact set in Y.

39. Let X be an infinite-dimensional inner product space. Show that the identity operator is not compact.

40. Show that if W is a closed, infinite-dimensional subspace of a Hilbert space, and P is the projection operator onto W, then P is not compact.

41. Let A be a compact, self-adjoint, linear operator on a Hilbert space X. Assume A is not the zero operator. If A has finite rank, use (7.6) to show that

$$\sup_{x \neq 0} \frac{\langle Ax, x \rangle}{\langle x, x \rangle}$$

is equal to the largest eigenvalue of A (which may be positive, zero, or negative). Give an example of how this result can fail if A has infinitely many nonzero eigenvalues.

Remark. The expression $\langle Ax, x \rangle / \langle x, x \rangle$ is called the **Rayleigh quotient**.

42. Let A be an $n \times n$ Hermitian matrix, and let α denote its kth principal submatrix; i.e., the upper-left $k \times k$ submatrix of A. Show that all the eigenvalues of α are less than or equal to the largest eigenvalue of A. *Hint:* Let $\lambda_{\max}$ denote the largest eigenvalue of A, and let μ be an eigenvalue of α with k-dimensional eigenvector φ. Then A has the form

$$A = \begin{bmatrix} \alpha & \beta \\ \beta^{\mathsf{H}} & \delta \end{bmatrix}.$$

Show that $\lambda_{\max} \geq \mu$.

Remark. This is a special case of the **eigenvalue interlacing theorem**.

43. Let $\mathscr{A}$ denote the set of all compact, self-adjoint linear operators of finite rank on a Hilbert space X. For $A \in \mathscr{A}$, let $\lambda_{\max}(A)$ denote the largest eigenvalue of A. Show that $\lambda_{\max}$ is a convex function on $\mathscr{A}$. *Hints:* The Problems 7.41 and 6.4 may be helpful.

44. In each of the following parts, assume the operators are defined on appropriate normed vector spaces.

    (a) Show that a compact linear operator is bounded.
    (b) Show that a bounded linear operator of finite rank (finite-dimensional range) is compact. *Hint:* Problem 6.58.
    (c) If A is compact and B is bounded, show that BA is compact.
    (d) Show that if A and B are compact, then $A + B$ is also compact.
    (e) If A is compact and B is bounded, show that AB is compact. *Hint:* Problem 7.37.

45. If $A : X \to X$, then $A^n x := A(A^{n-1}x)$, where $A^1 x := Ax$. In other words, $A^2 x = A(Ax)$, $A^3 x = A(A(Ax))$, and so on. Assume X is a Hilbert space and that $A \neq 0$ is a compact, self-adjoint linear operator. If the largest-magnitude eigenvalue of A has magnitude less than one, determine whether or not $A^n x \to 0$ holds for all $x \in X$.

46. Let X and Y be vector spaces, and let $A: X \to Y$ and $B: Y \to X$ be linear operators. Suppose that λ is a nonzero eigenvalue of BA. Determine whether or not λ is an eigenvalue of AB.

47. Let A be a compact, self-adjoint operator with eigenpairs (λ_n, φ_n). Determine how (7.10) changes if we try to solve $(cI + A)x = y$ for some nonzero constant c. Specify all nonzero values of c for which this can be done.

48. Suppose A is self adjoint and positive semidefinite. Show that if A is nonsingular, then A is positive definite provided that A has a square root; i.e., assume $A = B^2$ where $B: X \to X$ is self adjoint.

    Remark. By the Spectral Theorem, if A is compact, self-adjoint, and positive semidefinite, then A has a square root by Example 7.33. Hence, $n \times n$ matrices that are Hermitian and positive semidefinite have square roots. The assumption of compactness is not necessary for the existence of a square root [1, Section 23.1].

49. In Example 7.33 we defined the square root of a compact, self-adjoint, positive-semidefinite, linear operator A. In this problem, we find the square root of $R := I + A$. Let A have eigenpairs (λ_k, φ_k) from the Spectral Theorem. Assume $\lambda_k \geq -1$ for all k, and put

$$Mx := \sum_{k=1}^{\infty} \sqrt{1 + \lambda_k} \langle x, \varphi_k \rangle \varphi_k.$$

    Let P denote the projection onto $\ker A$. Show that

$$R^{1/2} := P + M$$

    is self adjoint and satisfies $R^{1/2}(R^{1/2}x) = Rx$.

50. Let $A : X \to X$ have eigenvalue λ. Let $B: X \to X$ as well, and assume that B **commutes** with A; i.e., assume $AB = BA$, or in more detail, for all x, $A(Bx) = B(Ax)$. Show that B maps the eigenspace of λ to itself; in other words, show that if $Ax = \lambda x$, then $A(Bx) = \lambda(Bx)$. Thus, if x is an eigenvector of A with eigenvalue λ, and if Bx is nonzero, then Bx is another eigenvector of A with eigenvalue λ.

51. Let A have adjoint A^*. We say A is a **normal operator** if A commutes with its adjoint; i.e., A is normal if $A^*A = AA^*$. Put

$$B := \frac{A + A^*}{2} \quad \text{and} \quad C := \frac{A - A^*}{2j}.$$

    Show that A is normal if and only if B and C commute.

52. Let A have adjoint A^*. If B and C are defined as in the previous problem, show that Bx and Cx are both zero if and only if Ax and A^*x are zero.

53. Let A have adjoint A^* and assume A is normal. Show that $\ker A = \ker A^*$. Show that α is an eigenvalue of A if and only if $\overline{\alpha}$ is an eigenvalue of A^*. *Hint:* For the second part, expand $\|A^*\varphi - \overline{\alpha}\varphi\|^2$.

54. Let A be a normal operator with distinct eigenvalues α and α' and corresponding eigenvectors φ and φ'. Show that φ and φ' are orthogonal.

55. ***Relation between Eigenvalues and Singular Values.*** Let $A: X \to X$ be a compact normal operator on a Hilbert space X. Using (7.11), show that

$$A^*Ax = \sum_k |\alpha_k|^2 \langle x, \varphi_k \rangle \varphi_k.$$

We recognize this as the result of applying the Spectral Theorem to A^*A, which is how we derived the SVD of A. Hence, the $|\alpha_k|$ are the singular values of A. In particular, for a compact normal operator A, its norm is $\|A\| = \max |\alpha_k|$.

56. If A is not normal, it can happen that $\|A\|$ is strictly larger than the maximum of the absolute values of the eigenvalues of A. Find the singular values of the matrix of Example 7.24 when $a = b = 1$. As noted in the example, the only eigenvalue of A is a. Cf. Problem 7.29(c).

57. If A has adjoint A^*, show that A is normal if and only if $\|Ax\| = \|A^*x\|$ for all x. *Hint:* Problem 4.21 may be helpful.

58. Let X and Y be inner-product spaces. Fix nonzero vectors $c \in X$ and $w \in Y$. Define the linear operator $A: X \to Y$ by $Ax := \langle x, c \rangle w$.

    (a) Compute A^*y.
    (b) Find the eigenvalues and eigenvectors of A^*A.
    (c) Use the results of part (b) to write the SVD in the form of (7.13) for this particular operator A.
    (d) Simplify your results for (a)-(c) when $X = \mathbb{C}$ and $c = 1$.
    (e) Write out your answers to part (d) when $Y = L^2[0, \infty)$ and $w(t) = e^{-t}$.

59. Let $A: X \to Y$ be a linear operator between inner-product spaces X and Y. Assume $\dim X < \infty$ and $\|A\| > 0$. Put $\lambda_1 := \sup_{\|x\|=1} \|Ax\|$. Of course, $\lambda_1 = \|A\|$.

    (a) Give a direct proof that there exists a unit vector $\varphi_1 \in X$ with $\|A\varphi_1\| = \lambda_1$.
    (b) Put $\psi_1 := (1/\lambda_1)A\varphi_1$. Then ψ_1 is a unit vector. Show that $A^*\psi_1 = \lambda_1 \varphi_1$. *Hint:* Show that $\|A^*\psi_1 - \lambda_1 \varphi_1\|^2 \leq 0$.

(c) Consider the subspaces

$$X_2 := (\operatorname{span}\{\varphi_1\})^\perp \quad \text{and} \quad Y_2 := (\operatorname{span}\{\psi_1\})^\perp.$$

Show that $A: X_2 \to Y_2$. More specifically, given $x \in X_2$, prove that $Ax \in Y_2$.

60. Consider the matrix

$$A := \begin{bmatrix} a & d \\ b & e \\ c & f \end{bmatrix}.$$

If the columns of A are nonzero and orthogonal, find the singular values of A. Your answers should be explicit functions of the entries of A.

61. Consider the operator A on $m \times n$ matrices X defined by $AX := UXV$. Here U and V are given matrices which are assumed to satisfy $U^{\mathsf{H}}U = 9I_m$ and $VV^{\mathsf{H}} = 4I_n$, where I_m and I_n denote identity matrices of appropriate size. Using the standard inner product on matrices, $\langle R, S\rangle = \operatorname{tr}(RS^{\mathsf{H}})$, find all of the (positive) singular values of A. If any are repeated, note their multiplicity.

62. Show that $\|x_\alpha\|^2$ in (7.27) is a continuous function of $\alpha \geq 0$. *Hint:* The fact that $\|x_\alpha - \widetilde{x}\|^2 \to 0$ may be helpful in proving $\|x_\alpha\|^2$ is continuous at $\alpha = 0$.

63. **MATLAB.** Consider the equation $Ax = y$, where

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 5 \end{bmatrix} \quad \text{and} \quad y = \begin{bmatrix} 2 \\ 1 \end{bmatrix}.$$

(a) Since A is invertible, use the command `A\y` to find the solution. What is your solution, and what is the norm of your solution?

(b) With $\alpha = 0.005$, find the regularized solution and its norm.

(c) Now suppose

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

With y as before, compare the solutions using (*i*) `A\y`, (*ii*) the regularized solutions with $\alpha = 0.1$, $\alpha = 0.01$, and $\alpha = 0.001$, (*iii*) the solution using `pinv(A)*y`.

64. **MATLAB.** In the text we discussed how to approximate $\int_0^1 x(t)\,dt$. Use the following script to compute $y(s) := \int_0^1 \cos(\pi t s)\,dt$ numerically with 9-point Legendre–Gauss quadrature and plot $y(s)$ for 200 equally-spaced values of $s \in [0, 10]$. For comparison, the exact value of the integral is $\sin(\pi s)/(\pi s) = \operatorname{sinc}(s)$. The M-file `legendrequad01` was given in Example 7.41.

```
[t,w] = legendrequad01(9);      % 9-pt quadrature
s = linspace(0,10,200);         % 200 values of s
x = @(t,s)cos(pi*t*s);          % Define fcn
yapprox = w*x(t,s);             % Approx. integral
yexact = sinc(s);               % Exact integral
plot(s,yexact,s,yapprox,'--')
```

65. **MATLAB.** Modify the script of the preceding problem to handle the function $y(s) := \int_0^1 \cos(ts+s+t)\,dt$ using 4-point Legendre–Gauss quadrature. *Hint:* In the preceding problem, `t` is a column vector and `s` is a row vector. Hence, `t*s` is a matrix with $i\,j$ entry $t_i s_j$. In this problem we need a matrix `tsmat` whose $i\,j$ entry is $t_i + s_j$. This can be created with the statements

```
tmat = repmat(t,1,length(s));
smat = repmat(s,length(t),1);
tsmat = tmat + smat;
```

However, in this case it is more convenient to use `bsxfun` and replace lines 3–5 from the preceding problem with

```
x = @(t,s)cos(t.*s+t+s);
yapprox = w*bsxfun(x,t,s);
yexact = ???; % Exact formula for the integral.
```

The command `bsxfun` virtually creates the matrices `tmat` and `smat` and passes them to `x`. That is why in the new definition of `x` we used `t.*s` instead of `t*s` as in the preceding problem. **Turn in your derivation of the exact formula for the integral along with your code and plot.**

66. **MATLAB.** The weight function $w(t) = e^{-t}$ for $t \geq 0$ leads to the **Laguerre polynomials** and **Laguerre–Gauss quadrature**, that is, approximation of integrals of the form $\int_0^\infty x(t)e^{-t}\,dt$. Here is a MATLAB function to generate the nodes and weights to approximate such an integral.

```
function [t,w] = laguerrequad(n)
%
% Generate nodes and weights for
% Laguerre-Gauss quadrature.
% Note that t is a column vector
% and w is a row vector.
%
a = 2*[0:n-1]+1; % diagonal of J
u = [1:n-1];     % upper diagonal of J
[P,Lambda] = eig(diag(u,1)+diag(a)+diag(u,-1));
[t,i] = sort(diag(Lambda));
Ptop = P(1,:);
```

```
Ptop = Ptop(i);
w = Ptop.^2;
```

Use Laguerre–Gauss quadrature to approximate and plot the function $y(s) := \int_0^\infty \cos(ts)e^{-t}\,dt$ for 200 values of $s \in [0,6]$. What value of n did you need to give a satisfactory plot?

67. The Fourier transform of $x(t) := e^{-\lambda|t|}$ is $X(f) = 2\lambda/[\lambda^2 + (2\pi f)^2]$. Using the fact that $x(t)$ is even, compute $X(f)$ using Laguerre–Gauss quadrature on $[0,\infty)$. Plot the exact formula for $X(f)$ and your approximation on $[-\lambda,\lambda]$ when $\lambda = 5$. How many nodes n do you need to get a good approximation of the exact answer?

68. Reproduce Figure 7.4 from Example 7.43. Start with the code in Note 7.4 at the end of the chapter. You will need results from Problem 7.33.

69. Modify the function `legendrequad01` in Example 7.41 to produce the nodes and weights for Legendre–Gauss quadrature on $[-1,1]$. You may use the fact that Legendre polynomials on $[-1,1]$ are generated by the three-term recursion with $a_k = 0$, $b_0 = 2$, and $b_k = 1/(4-k^{-2})$. Call your new function `legendrequad`.

70. Reproduce the eigenvalue list and Figure 7.5 from Example 7.44. Start with the code in Note 7.4 at the end of the chapter. In the function `eigfcnNystrom` change the call to `legendrequad01` to a call to the function `legendrequad` that you wrote for the preceding problem. No changes are needed to the other function `eigfcnNystrom`.

71. ***Orthonormality of the Prolate Spheroidal Wave Functions.*** Show that the prolate spheroidal wave functions $\psi_k \in L^2(\mathbb{R})$ defined in Example 7.44 are satisfy $\langle \psi_k, \psi_j \rangle = \delta_{kj}$. *Hint:* In (7.33), regard $\varphi_k(\tau)$ as being zero for $|\tau| > 1$. With $h(t) := WT\,\mathrm{sinc}(WTt)$, (7.33) says that $\psi_k = (h * \varphi_k)/\lambda_k$. Since h is the impulse response of the ideal lowpass filter with bandwidth $WT/2$, the ψ_j are bandlimited, which implies $h * \psi_j = \psi_j$. Use the fact that $\langle \varphi_k, \varphi_j \rangle = \lambda_k \delta_{kj}$.

CHAPTER 8
Applications

8.1. Quadratically Constrained Least Squares with the SVD

Recall that the **regularization** problem in Section 7.6 required the solution of

$$(\alpha I + A^*A)x = A^*y, \tag{8.1}$$

which is a special case of (5.15) when Q is the identity.[a] Hence, there is a close connection between regularization and quadratically constrained least squares. In regularization, we fix a value of $\alpha > 0$ and solve (8.1) for the corresponding value of x. In quadratically constrained least squares, we fix an energy constraint $b > 0$ and choose the Lagrange multiplier $\alpha \geq 0$ so that the corresponding solution x of (8.1) satisfies

$$\|x\| \leq b \quad \text{and} \quad \alpha(\|x\|^2 - b) = 0.$$

When A is given by a matrix, we can adapt the MATLAB code in Section 7.6 to choose α to satisfy the above Lagrange multiplier conditions and return the corresponding value of x.

The code below includes the function $f(\alpha) := \|x_\alpha\|^2 - b$, where x_α is the solution of (8.1). The definition of f makes use of the fact that $\|x_\alpha\|^2$ is given by (7.27).

```
function [xalpha,alpha] = qcls(A,y,b)
% Solve Quadratically Constrained Least Squares using SVD
% Solve min_x ||y-Ax||^2  subject to  ||x||^2 <= b.
alpha = 0;
[Q,S,P] = svd(A);
s = diag(S);
s = s(s>0);         % Keep only positive entries in s
r = numel(s);
if r>0
   Qry = Q(:,1:r)'*y;
   Qry2 = Qry.*conj(Qry);
   f = @(alpha)((s./(alpha+(s.^2))).^2)'*Qry2 - b;
   fprintf('pseudoinverse yields ||x_0||^2 = %g,\n',f(0)+b)
   if f(0)>0 % Is ||x_0||^2 > b?
      fprintf('which is > energy constraint %g.\n',b)
      alpha = fzero(f,0);
```

[a] If Q is not the identity but is invertible, see Problem 5.46.

```
        fprintf('Using alpha = %g yields\n',alpha)
        xalpha = P(:,1:r)*(s./(alpha+(s.^2)).*Qry)
    else % Compute pseudoinverse of y.
        fprintf('which is <= energy constraint %g\n',b)
        fprintf('Using alpha = 0 yields\n')
        xalpha = P(:,1:r)*(Qry./s)
    end
else
    fprintf('qcls: A matrix is all zeros\n')
    xalpha = zeros(size(P(:,1)));
end
```

Example 8.1. Consider the singular matrix A and vector y given by

$$A := \begin{bmatrix} 6 & 6 & 3 \\ -2 & 6 & 1 \end{bmatrix} \quad \text{and} \quad y := \begin{bmatrix} -3 \\ 8 \end{bmatrix}.$$

Setting `b = 3` and running the command `qcls(A,y,b)` yields

```
pseudoinverse yields ||x_0||^2 = 2.64236,
which is <= energy constraint 3
Using alpha = 0 yields

xalpha =

   -1.3472
    0.9028
   -0.1111
```

Repeating with `b = 0.1` yields

```
pseudoinverse yields ||x_0||^2 = 2.64236,
which is > energy constraint 0.1.
Using alpha = 114.51 yields

xalpha =

   -0.2489
    0.1946
   -0.0136
```

8.2. Finite-Duration Pulses of Maximum In-Band Energy

In many communication systems, we need waveforms that are both time limited and bandlimited. Unfortunately, Problem 7.20 implies that the zero function is the only finite-energy waveform that is both strictly time limited and strictly bandlimited. Therefore, we try to find strictly time-limited waveforms that are *approximately* bandlimited. For convenience, we take the time interval to be $[-T/2, T/2]$, and we take the frequency band to be $[-W, W]$.

In order to set up the problem mathematically, we need a few preliminary observations. For any finite-energy waveform $x \in L^2(\mathbb{R})$, let X denote its Fourier transform. By **Parseval's equation**, the total energy in the waveform satisfies

$$\|x\|^2 := \int_{-\infty}^{\infty} |x(t)|^2 \, dt = \int_{-\infty}^{\infty} |X(f)|^2 \, df =: \|X\|^2.$$

The **in-band energy** is

$$\int_{-W}^{W} |X(f)|^2 \, df.$$

If we let H denote the transfer function of the ideal lowpass filter

$$H(f) := \begin{cases} 1, & |f| \le W, \\ 0, & |f| > W, \end{cases}$$

then the in-band energy is

$$\int_{-W}^{W} |X(f)|^2 \, df = \int_{-\infty}^{\infty} H(f)|X(f)|^2 \, df = \int_{-\infty}^{\infty} H(f)X(f)\overline{X(f)} \, df = \langle HX, X \rangle.$$

Since the inverse Fourier transform of H is $h(t) = 2W \operatorname{sinc}(2Wt)$, we have by Parseval's equation that the in-band energy is $\langle HX, X \rangle = \langle h * x, x \rangle$, where the $*$ denotes convolution.

If we define the operator $A: L^2[-T/2, T/2] \to L^2[-T/2, T/2]$ by

$$(Ax)(t) = \int_{-T/2}^{T/2} h(t-\tau)x(\tau) \, d\tau, \quad |t| \le T/2, \tag{8.2}$$

then for any $x \in L^2(\mathbb{R})$ that satisfies $x(t) = 0$ for $|t| > T/2$, we can write its in-band energy as

$$\langle h * x, x \rangle = \int_{-\infty}^{\infty} (h*x)(t)\overline{x(t)} \, dt = \int_{-T/2}^{T/2} (h*x)(t)\overline{x(t)} \, dt = \langle Ax, x \rangle,$$

where the inner product on the left is on $L^2(\mathbb{R})$, and the inner product on the right is on $L^2[-T/2,T/2]$. Since A is a compact, self-adjoint, positive-definite linear operator, it has a complete orthonormal set of eigenfunctions $\{\beta_k\}_{k=0}^{\infty}$ and corresponding positive eigenvalues $\lambda_k \searrow 0$, and we have the representation[b]

$$Ax = \sum_k \lambda_k \langle x, \beta_k \rangle \beta_k.$$

Hence, the in-band energy is

$$\langle Ax, x \rangle = \sum_k \lambda_k \langle x, \beta_k \rangle \langle \beta_k, x \rangle = \sum_k \lambda_k |\langle x, \beta_k \rangle|^2.$$

Since the λ_k are nonincreasing,

$$\langle Ax, x \rangle \le \lambda_0 \sum_k |\langle x, \beta_k \rangle|^2 = \lambda_0 \|x\|^2,$$

where the last step follows because the β_k form a complete orthonormal set for $L^2[-T/2,T/2]$. Furthermore, the above inequality holds with equality if x is proportional to β_0. Thus, waveforms proportional to the eigenfunction β_0 have the greatest possible fraction of their energy confined to the passband $[-W,W]$. That fraction is λ_0.[c]

To determine the eigenfunctions and eigenvalues of the operator A in (8.2), consider the eigenvalue problem

$$\int_{-T/2}^{T/2} 2W \operatorname{sinc}(2W[t-\tau]) x(\tau)\, d\tau = \lambda x(t), \quad |t| \le T/2. \tag{8.3}$$

Make the change of variable $\theta = 2\tau/T$, $d\theta = 2\,d\tau/T$ to get

$$\int_{-1}^{1} WT \operatorname{sinc}(2W[t - T\theta/2]) x(T\theta/2)\, d\theta = \lambda x(t), \quad |t| \le T/2.$$

Now make the identification $t = Ts/2$ to get

$$\int_{-1}^{1} WT \operatorname{sinc}(WT[s-\theta]) x(T\theta/2)\, d\theta = \lambda x(Ts/2), \quad |s| \le 1.$$

[b] Recall Theorem 7.19, Problem 7.28, and the Spectral Theorem.

[c] We note that $\lambda_0 < 1$. This can be seen as follows. The ideal lowpass filter is a projection; hence, for $x \in L^2(\mathbb{R})$, $\|h*x\| \le \|x\|$ by (3.8), with equality if and only if x is bandlimited. Also, for x with $x(t) = 0$ for $|t| > T/2$, $\|Ax\|_{L^2[-T/2,T/2]} \le \|h*x\|_{L^2(\mathbb{R})} \le \|x\|$ implies all eigenvalues of A satisfy $|\lambda_k| \le 1$. Furthermore, since a nonzero, time-limited x cannot be bandlimited (by the remark in Problem 7.20), we must have $\|Ax\|_{L^2[-T/2,T/2]} \le \|h*x\|_{L^2(\mathbb{R})} < \|x\|$. Hence, the eigenvalues of A are strictly less than one.

If we put $\varphi(\theta) := x(T\theta/2)$, then our equation becomes

$$\int_{-1}^{1} WT\,\mathrm{sinc}(WT[s-\theta])\varphi(\theta)\,d\theta = \lambda\varphi(s), \quad |s| \le 1.$$

From Example 7.44, we know that the solutions φ_k of this equation are the restrictions to $[-1,1]$ of the prolate spheroidal wave functions ψ_k. Hence,[d]

$$\beta_k(t) = \begin{cases} \psi_k(2t/T)/\sqrt{\lambda_k}, & |t| \le T/2, \\ 0, & |t| > T/2. \end{cases} \tag{8.4}$$

8.2.1. A 2*WT* Theorem

We saw above that the in-band energy of a time-limited signal x is

$$\langle h * x, x\rangle = \langle Ax, x\rangle = \sum_k \lambda_k |\langle x, \beta_k\rangle|^2.$$

Now recall from Figure 7.6 in Example 7.44 that for k sufficiently less than $2WT$, the λ_k are approximately 1, and for k sufficiently larger than $2WT$, the λ_k are very close to zero. Hence, with n sufficiently less than $2WT$, and $x \in \mathrm{span}\{\beta_0, \ldots, \beta_{n-1}\}$, the in-band energy is

$$\sum_k \lambda_k |\langle x, \beta_k\rangle|^2 = \sum_{k=0}^{n-1} \lambda_k |\langle x, \beta_k\rangle|^2 \approx \sum_{k=0}^{n-1} |\langle x, \beta_k\rangle|^2 = \|x\|^2.$$

In other words, for these signals, essentially all of the energy lies in-band. More precisely,

$$\sum_{k=0}^{n-1} \lambda_k |\langle x, \beta_k\rangle|^2 \ge \lambda_{n-1} \sum_{k=0}^{n-1} |\langle x, \beta_k\rangle|^2 = \lambda_{n-1}\|x\|^2,$$

and so the ratio of the in-band energy to the total energy satisfies

$$\frac{\displaystyle\sum_{k=0}^{n-1} \lambda_k |\langle x, \beta_k\rangle|^2}{\|x\|^2} \ge \lambda_{n-1}.$$

For example, if $2WT = 10$ and we take $n = 9$, then $\lambda_{n-1} = \lambda_8 = 0.93$. In other words, signals in the 9-dimensional subspace $\mathrm{span}\{\beta_0, \ldots, \beta_8\}$ have 93% of their energy in the passband.

[d] Recall from Example 7.44 that $\|\varphi_k\|^2 = \lambda_k$, and we require $\|\beta_k\| = 1$.

8.3. Reproducing Kernel Hilbert Spaces

A **reproducing kernel Hilbert space (RKHS)** X is a Hilbert space of real or complex valued functions x such that point evaluation is a bounded linear functional. By the Riesz Representation Theorem for Hilbert Space, the linear functional that evaluates functions at argument t can be represented by a vector $K_t \in X$, and for all $x \in X$, we can write $x(t) = \langle x, K_t \rangle$. Since $K_t \in X$, we can evaluate $K_t(\cdot)$ itself at any point τ via $K_t(\tau) = \langle K_t, K_\tau \rangle$. This formula explains the name reproducing kernel Hilbert space, because K reproduces itself via the foregoing inner product. We also have by the Cauchy–Schwarz inequality that

$$|x(t)| = |\langle x, K_t \rangle| \leq \|x\| \langle K_t, K_t \rangle^{1/2} = \sqrt{K_t(t)}\, \|x\|.$$

Example 8.2 (The RKHS of Bandlimited Functions). Let X denote the subset of $L^2(\mathbb{R})$ consisting of waveforms that are bandlimited to $[-W, W]$. For $x \in X$, the Hölder inequality yields

$$\begin{aligned} |x(t)| = \left| \int_{-W}^{W} X(f) e^{j2\pi ft}\, df \right| &\leq \left(\int_{-W}^{W} |X(f)|^2\, df \right)^{1/2} \left(\int_{-W}^{W} |e^{j2\pi ft}|^2\, df \right)^{1/2} \\ &= \|X\|_2 \sqrt{2W} = \|x\|_2 \sqrt{2W}, \end{aligned}$$

where the last step follows by Parseval's equation. This shows that point evaluation is a bounded linear functional. To find the kernel function, we use the fact that if $x \in X$ is passed through an ideal lowpass filter of bandwidth W, the signal x is unchanged. So if $h(t) := 2W \operatorname{sinc}(2Wt)$,

$$x(t) = (h * x)(t) = \int_{-\infty}^{\infty} h(t-\tau) x(\tau)\, d\tau = \int_{-\infty}^{\infty} x(\tau) \overline{K_t(\tau)}\, d\tau = \langle x, K_t \rangle,$$

where $K_t(\tau) := \overline{h(t-\tau)} = h(t-\tau)$ since h is real.

Given an RKHS X with kernel function $K_t(\tau)$ and a finite set of arguments $t_1, \ldots, t_n$, consider the finite-dimensional subspace $X_0 := \operatorname{span}\{K_{t_1}, \ldots, K_{t_n}\}$. By the Projection Theorem, every $x \in X$ can be written as $x = \widehat{x} + \widetilde{x}$, where $\widehat{x}$ is the projection of x onto X_0, and $\widetilde{x} \in X_0^\perp$ must be orthogonal to $K_{t_1}, \ldots, K_{t_n}$. Hence,

$$x(t_i) = \langle x, K_{t_i} \rangle = \langle \widehat{x} + \widetilde{x}, K_{t_i} \rangle = \langle \widehat{x}, K_{t_i} \rangle + \langle \widetilde{x}, K_{t_i} \rangle = \langle \widehat{x}, K_{t_i} \rangle = \widehat{x}(t_i).$$

Furthermore,

$$\|x\|^2 = \|\widehat{x} + \widetilde{x}\|^2 = \|\widehat{x}\|^2 + \|\widetilde{x}\|^2.$$

These two formulas have a powerful implication for regularization problems of the form

$$\min_{x\in X} \sum_{i=1}^{n} J_i(x(t_i)) + \alpha\|x\|^2. \tag{8.5}$$

This can be seen by writing

$$\sum_{i=1}^{n} J_i(x(t_i)) + \alpha\|x\|^2 = \sum_{i=1}^{n} J_i(\widehat{x}(t_i)) + \alpha\left[\|\widehat{x}\|^2 + \|\widetilde{x}\|^2\right] \geq \sum_{i=1}^{n} J_i(\widehat{x}(t_i)) + \alpha\|\widehat{x}\|^2.$$

It follows that in solving (8.5), we can restrict the search to the n-dimensional subspace X_0; i.e., we may restrict attention to vectors $x_0 \in X_0$, which have the form

$$x_0(t) = \sum_{\ell=1}^{n} c_\ell K_{t_\ell}(t),$$

where the coefficients c_ℓ have to be found. Since

$$\|x_0\|^2 = \left\langle \sum_{\ell=1}^{n} c_\ell K_{t_\ell}, \sum_{i=1}^{n} c_i K_{t_i} \right\rangle = \sum_{\ell=1}^{n}\sum_{i=1}^{n} c_\ell \overline{c}_i K_{t_\ell}(t_i),$$

it is only necessary to solve

$$\min_{c_1,\ldots,c_n} \sum_{i=1}^{n} J_i\left(\sum_{\ell=1}^{n} c_\ell K_{t_\ell}(t_i)\right) + \alpha \sum_{\ell=1}^{n}\sum_{i=1}^{n} c_\ell \overline{c}_i K_{t_\ell}(t_i).$$

Example 8.3. Suppose we want to find a waveform $x \in X$ that satisfies $x(t_i) = y_i$ for $i = 1,\ldots,n$. If $J_i(\xi) := |y_i - \xi|^2$, then (8.5) becomes

$$\min_{x\in X} \sum_{i=1}^{n} |y_i - x(t_i)|^2 + \alpha\|x\|^2.$$

Now put $M_{i\ell} := K_{t_\ell}(t_i)$, $\mathbf{y} := [y_1,\ldots,y_n]^{\mathsf{T}}$, and $\mathbf{c} := [c_1,\ldots,c_n]^{\mathsf{T}}$. Then for $x_0 \in X_0$,

$$\begin{bmatrix} x_0(t_1) \\ \vdots \\ x_0(t_n) \end{bmatrix} = M\mathbf{c},$$

and

$$\sum_{i=1}^{n} |y_i - x_0(t_i)|^2 + \alpha\|x_0\|^2 = \|\mathbf{y} - M\mathbf{c}\|^2_{\mathbb{C}^n} + \alpha\mathbf{c}^{\mathsf{H}} M\mathbf{c}.$$

Minimization of the right-hand side as a function of $\mathbf{c}$ using the Gâteaux derivative leads to formulas very similar to those used in solving the quadratically constrained

least squares problem in Example 5.22. Since $M_{i\ell} = \langle K_{t_\ell}, K_{t_i} \rangle$, $M = M^{\mathsf{H}}$ is a Gram matrix and therefore positive semidefinite. Assuming M is positive definite, the solution is

$$\mathbf{c} = (\alpha M + M^{\mathsf{H}} M)^{-1} M^{\mathsf{H}} \mathbf{y} = (\alpha M + MM)^{-1} M\mathbf{y} = [M(\alpha I + M)]^{-1} M\mathbf{y} = (\alpha I + M)^{-1} \mathbf{y}.$$

8.4. Matched Filters for Known Signals

In radar systems, a known pulse v is transmitted, reflected by an object, and returned to a receiver in the presence of additive noise z. In an attempt to enhance the signal and attenuate the noise, the receiver applies the measurement waveform $v + z$ to a filter; i.e., a linear, time-invariant system with impulse response h. The filter response to v is

$$v_o(t) := \int h(t-\tau) v(\tau)\, d\tau,$$

and the response to z is

$$z_o(t) := \int h(t-\tau) z(\tau)\, d\tau.$$

We assume z is a zero-mean, random noise waveform with known correlation function

$$r(t_1, t_2) := \mathsf{E}[z(t_1)\overline{z(t_2)}\,].$$

The correlation function of z_o is easily shown to be

$$\mathsf{E}[z_o(t_1)\overline{z_o(t_2)}\,] = \int h(t_1 - t) \left[\int r(t, \tau) \overline{h(t_2 - \tau)}\, d\tau \right] dt.$$

Our goal is to design the system defined by h so as to maximize the output signal-to-noise ratio at time t_0,[e]

$$\mathrm{SNR} := \frac{|v_o(t_0)|^2}{\mathsf{E}[|z_o(t_0)|^2]}.$$

To analyze this expression, it is convenient to introduce the following notation. Put

$$\xi(\tau) := \overline{h(t_0 - \tau)}$$

so that $v_o(t_0) = \langle v, \xi \rangle$. Since $\overline{r(t_2, t_1)} = r(t_1, t_2)$,

$$(Rx)(t) := \int r(t, \tau) x(\tau)\, d\tau,$$

[e] More precisely, we seek to maximize the ratio of the instantaneous output signal power at time t_0 to the expected instantaneous noise power at time t_0.

is a self-adjoint linear operator. Then

$$\mathsf{E}[|z(t_0)|^2] = \langle R\xi, \xi\rangle.$$

Under suitable assumptions (cf. Problem 7.49), $R^{1/2}$, R^{-1}, and $R^{-1/2}$ exist. Hence,

$$\begin{aligned}
\mathrm{SNR} = \frac{|\langle v,\xi\rangle|^2}{\langle R\xi,\xi\rangle} = \frac{|\langle R^{1/2}R^{-1/2}v,\xi\rangle|^2}{\langle R^{1/2}R^{1/2}\xi,\xi\rangle} &= \frac{|\langle R^{-1/2}v, R^{1/2}\xi\rangle|^2}{\langle R^{1/2}\xi, R^{1/2}\xi\rangle} \\
&= \frac{|\langle R^{-1/2}v, R^{1/2}\xi\rangle|^2}{\|R^{1/2}\xi\|^2} \\
&\le \frac{\|R^{-1/2}v\|^2\,\|R^{1/2}\xi\|^2}{\|R^{1/2}\xi\|^2} \\
&= \|R^{-1/2}v\|^2,
\end{aligned}$$

where the inequality holds by the Cauchy–Schwarz inequality. This upper bound does not depend on ξ, which we are trying to optimize. Furthermore, equality holds if $R^{-1/2}v = \lambda R^{1/2}\xi$ for any complex scalar λ. Taking $\lambda = 1$ for convenience, we should choose $\xi = R^{-1}v$. More explicitly, for fixed t_0, we should take $h(t_0 - \cdot)$ as the complex conjugate of the solution of

$$\int r(t,\tau)\xi(\tau)\,d\tau = v(t)$$

for the given waveform v. Since ξ, and therefore $h(t_0 - \cdot)$ depend on v, we say that $h(t_0 - \cdot)$ is "matched" to the signal v.

Example 8.4. If the noise is wide-sense stationary; i.e., $r(t,\tau) = c(t-\tau)$ for some function c. If all signals live on $(-\infty,\infty)$, then the above integral equation says $c * \xi = v$. In terms of Fourier transforms, $C(f)\Xi(f) = V(f)$ or $\Xi(f) = V(f)/C(f)$. If z is white noise, then $C(f) \equiv \sigma^2$ is a positive constant, and so $\xi(\tau) = v(\tau)/\sigma^2$, and then

$$h(t_0 - \tau) = \overline{v(\tau)}/\sigma^2.$$

Since τ can be any real number, put $\tau = t_0 - t$ to get $h(t) = \overline{v(t_0 - t)}/\sigma^2$.

8.5. Matched Filters for Random Signals

In a multipath radio channel, the receiver input is again of the form $v + z$, but this time both v and z are random. Let z be as in the preceding section. We assume that v is zero mean and has known correlation function

$$k(t_1,t_2) := \mathsf{E}[v(t_1)\overline{v(t_2)}\,].$$

As before, we propose to pass the input though a linear, time-invariant system. However, to design the linear system, we use the signal-to-noise ratio

$$\text{SNR} := \frac{\mathsf{E}[|v_o(t_0)|^2]}{\mathsf{E}[|z_o(t_0)|^2]}.$$

Notice that this definition differs from the previous one by that addition of the expectation in the numerator.

Setting

$$(Kx)(t) := \int k(t,\tau)x(\tau)\,d\tau,$$

it follows from the analysis in the preceding section that, with ξ as defined there,[f]

$$\begin{aligned}
\text{SNR} = \frac{\langle K\xi,\xi\rangle}{\langle R\xi,\xi\rangle} = \frac{\langle K\xi, R^{-1/2}R^{1/2}\xi\rangle}{\|R^{1/2}\xi\|^2} &= \frac{\langle R^{-1/2}K\xi, R^{1/2}\xi\rangle}{\|R^{1/2}\xi\|^2} \\
&= \frac{\big\langle (R^{-1/2}KR^{-1/2})R^{1/2}\xi, R^{1/2}\xi\big\rangle}{\|R^{1/2}\xi\|^2} \\
&= \left\langle (R^{-1/2}KR^{-1/2})\frac{R^{1/2}\xi}{\|R^{1/2}\xi\|}, \frac{R^{1/2}\xi}{\|R^{1/2}\xi\|}\right\rangle \\
&\le \|R^{-1/2}KR^{-1/2}\|, \quad \text{by Lemma 7.28.}
\end{aligned}$$

This upper bound does not depend on ξ, which we are trying to optimize. Assuming $R^{-1/2}KR^{-1/2}$ is compact, the Spectral Theorem implies that the upper bound is achieved if $R^{1/2}\xi/\|R^{1/2}\xi\|$ is a unit-norm eigenvector corresponding to the largest eigenvalue of the $R^{-1/2}KR^{-1/2}$.

If K is compact and $R^{-1/2}$ is bounded, then Problem 7.37 can be used to show that $R^{-1/2}KR^{-1/2}$ is compact. If (λ,φ) is the desired eigenpair, then it remains to solve $R^{1/2}\xi = \mu\varphi$ for any scalar μ. If z is white noise so that $r(t,\tau) = \sigma^2\delta(t-\tau)$, then $\xi = \mu\varphi/\sigma$. In other words, we need the first eigenpair of the operator K.

8.6. Conjugate Gradient Direction Algorithms

We describe an algorithm for solving $Qx = y$ when Q is a symmetric, positive-definite matrix. This kind of equation results from finite-dimensional instances of several problems we considered in previous chapters:

- Example 4.17 when there is no solution of $y = Ax$ and A is full rank.
- Example 4.21 when there are multiple solutions of $y = Ax$ and A^* is full rank.

[f] The SNR analysis in this paragraph was suggested by B. N. Bhaskar.

- The quadratically constrained least squares problem of Example 5.22 when Q in (5.15) is full rank and $\lambda > 0$.
- Regularization when $\alpha > 0$ in (7.25).

Preliminary Observations

Let X be an n-dimensional inner-product space with n orthogonal vectors denoted by $\xi_0, \ldots, \xi_{n-1}$. Then every $x \in X$ has the unique representation

$$x = \sum_{k=0}^{n-1} \frac{\langle x, \xi_k \rangle}{\langle \xi_k, \xi_k \rangle} \xi_k.$$

Similarly, for any two vectors x and x_0, we can write

$$x - x_0 = \sum_{k=0}^{n-1} \frac{\langle x - x_0, \xi_k \rangle}{\langle \xi_k, \xi_k \rangle} \xi_k \qquad \text{or} \qquad x = x_0 + \sum_{k=0}^{n-1} \frac{\langle x - x_0, \xi_k \rangle}{\langle \xi_k, \xi_k \rangle} \xi_k.$$

This suggests that we define the sequence

$$x_m := x_0 + \sum_{k=0}^{m-1} \frac{\langle x - x_0, \xi_k \rangle}{\langle \xi_k, \xi_k \rangle} \xi_k, \quad m = 1, \ldots, n.$$

Observe that $x_n = x$ and

$$x_{m+1} = x_m + \frac{\langle x - x_0, \xi_m \rangle}{\langle \xi_m, \xi_m \rangle} \xi_m. \tag{8.6}$$

Notice that $x_{m+1} - x_m$ is just a constant times ξ_m. Equivalently, $x_{k+1} - x_k$ is a constant times ξ_k. Keeping this in mind, we have

$$\langle x_m - x_0, \xi_m \rangle = \left\langle \sum_{k=0}^{m-1} x_{k+1} - x_k, \xi_m \right\rangle = \sum_{k=0}^{m-1} \langle x_{k+1} - x_k, \xi_m \rangle = 0$$

by orthogonality. Hence, $\langle x - x_0, \xi_m \rangle = \langle x - x_m + x_m - x_0, \xi_m \rangle = \langle x - x_m, \xi_m \rangle$. Making this substitution in (8.6), we have

$$x_{m+1} = x_m + \frac{\langle x - x_m, \xi_m \rangle}{\langle \xi_m, \xi_m \rangle} \xi_m, \quad m = 0, 1, \ldots. \tag{8.7}$$

Hence, given any $x_0 \in X$, if we apply the recursion (8.7), we get $x_n = x$.

Solving the Equation

Now suppose that we repeat the foregoing analysis with the alternative inner product $\langle \cdot, \cdot \rangle_Q$, where Q is self adjoint and positive definite under the original inner product. Furthermore, assume that x is the unique solution of $Qx = y$ for some given $y \in X$. Then (8.7) becomes

$$\begin{aligned} x_{m+1} &= x_m + \frac{\langle Q(x - x_m), \xi_m \rangle}{\langle Q\xi_m, \xi_m \rangle} \xi_m \\ &= x_m + \frac{\langle y - Qx_m, \xi_m \rangle}{\langle Q\xi_m, \xi_m \rangle} \xi_m, \end{aligned} \tag{8.8}$$

and $x_n = x$ solves $Qx = y$.

To explain the "gradient" terminology, observe that $Qx = y$ if and only if

$$0 = \|Qx - y\|^2 = \|Q^{1/2}(Q^{1/2}x - Q^{-1/2}y)\|^2,$$

which holds if and only if $\|Q^{1/2}x - Q^{-1/2}y\|^2 = 0$. Hence, $Qx = y$ if and only if x minimizes

$$\|Q^{1/2}x - Q^{-1/2}y\|^2 = \langle Qx, x \rangle - 2\langle y, x \rangle + \langle Q^{-1}y, y \rangle,$$

which is equivalent to minimizing

$$f(x) := \tfrac{1}{2}\langle Qx, x \rangle - \langle y, x \rangle,$$

which is a **quadratic programming problem**. Suppose we are at the point x_m, and we want to study f along the direction ξ_m. Consider the **line search problem** of choosing t to minimize $g(t) := f(x_m + t\xi_m)$. With this value of t, we put $x_{m+1} := x_m + t\xi_m$. The optimal value of t makes $g'(t) = 0$. Since $g'(t) = \langle \xi_m, \nabla f(x_m + t\xi_m) \rangle$ and $\nabla f(x) = Qx - y$, the optimal value of t is precisely the quotient in (8.8). Hence, the recursion (8.8) minimizes the quadratic f in at most n steps.

Finding the Conjugate Directions

Remember that the vectors ξ_k must be orthogonal under the Q-inner product; i.e., $\langle Q\xi_k, \xi_m \rangle = 0$ for $k \neq m$. Such a set of vectors is said to be **conjugate** with respect to Q. Of course, one approach to determining a set of ξ_k would be to apply the Gram–Schmidt procedure to the standard basis. However, this is *not* recommended. Much better and simpler methods are given in [28, Section 9.3] and [30, Section 5.1]. For example, below is a MATLAB implementation based on [30]. ***In practice, the conjugate gradient algorithm should only be used on problems that are too large to solve using more traditional methods [30, p. 112].***

```
function x = conjgrad(Q,y,x0)
% Conjugate Gradient Algorithm based on Nocedal & Wright.
% Solve Q x = y, where Q is symmetric and positive definite.
x = x0;
r = Q*x-y;
norm2rold = r'*r;
xi = -r;
Qxi = Q*xi;
for k=1:length(y)
   alpha = norm2rold/(xi'*Qxi);
   x = x + alpha*xi;
   r = r + alpha*Qxi;
   norm2r = r'*r;
   xi = -r + (norm2r/norm2rold)*xi;
   % Prepare for next time thru loop
   Qxi = Q*xi;
   norm2rold = norm2r;
end
```

You can test this function with the following script.

```
n = 5;
A = randi([-9,9],n,n);
Q = A'*A
lambda = eig(Q)' % verify Q>0
x = randi([-9,9],n,1);
y = Q*x;
x0 = randi([-9,9],n,1);
xhat = conjgrad(Q,y,x0);
fprintf('      x       xhat\n')
fprintf('   -------   -------\n')
disp([x xhat])
err = x-xhat;
fprintf('||x-xhat||^2 = %g\n',err'*err)
```

8.7. Hermite Functions

We show that the **Hermite functions**, defined below, are eigenfunctions of the Fourier transform. Consider the so-called "generating function,"

$$g(t,z) := e^{2tz-z^2} = e^{-(t-z)^2} e^{t^2} = e^{t^2} w(t-z),$$

where $w(\theta) := e^{-\theta^2}$. Let us regard g as a function of z and expand g in a Taylor series; i.e.,[g]

$$g(t,z) = \sum_{n=0}^{\infty} \frac{z^n}{n!} \left(\frac{\partial^n g}{\partial z^n}\bigg|_{z=0} \right).$$

Note that $(\partial/\partial z)w(t-z) = -w'(t-z)$, and in general, the nth partial derivative with respect to z is $(-1)^n w^{(n)}(t-z)$. It now follows that

$$\frac{\partial^n g}{\partial z^n}\bigg|_{z=0} = (-1)^n e^{t^2} w^{(n)}(t).$$

Using the definition of w yields the more common presentation of this expression as

$$\frac{\partial^n g}{\partial z^n}\bigg|_{z=0} = (-1)^n e^{t^2} \frac{d^n}{dt^n} e^{-t^2}.$$

It is easy to see that this formula is a polynomial in t. Just observe that if $p(t)$ is a polynomial in t, then $(d/dt)[p(t)e^{-t^2}]$ has the form $q(t)e^{-t^2}$ where $q(t)$ is another polynomial in t. For this reason,

$$H_n(t) := (-1)^n e^{t^2} \frac{d^n}{dt^n} e^{-t^2}$$

is called the nth **Hermite polynomial**. With this notation,

$$g(t,z) = \sum_{n=0}^{\infty} H_n(t) \frac{z^n}{n!}.$$

We now define the **Hermite functions**[h]

$$\mathscr{H}_n(t) := e^{-t^2/2} H_n(t).$$

For fixed z, consider the Fourier transform of the time function $e^{-t^2/2} g(t,z)$. Write[i]

$$\begin{aligned} \int_{-\infty}^{\infty} e^{-t^2/2} g(t,z) e^{-j\omega t}\, dt &= \int_{-\infty}^{\infty} e^{-t^2/2} \left[\sum_{n=0}^{\infty} H_n(t) \frac{z^n}{n!} \right] e^{-j\omega t}\, dt \\ &= \sum_{n=0}^{\infty} \frac{z^n}{n!} \int_{-\infty}^{\infty} e^{-t^2/2} H_n(t) e^{-j\omega t}\, dt \end{aligned}$$

[g] If z is regarded as a complex variable, then it is easy to see that $g(t,z)$ is an **entire function** of z and therefore the Taylor expansion exists and is valid for all complex z [9].

[h] Many authors use the phrase "Hermite functions" to refer to the Hermite polynomials, so it is important to pay careful attention to terminology.

[i] The interchange of integral and infinite sum can be justified [26, p. 64].

$$= \sum_{n=0}^{\infty} \frac{z^n}{n!} \int_{-\infty}^{\infty} \mathscr{H}_n(t) e^{-j\omega t}\, dt$$
$$= \sum_{n=0}^{\infty} \frac{z^n}{n!} \widehat{\mathscr{H}_n}(\omega),$$

where $\widehat{\mathscr{H}_n}$ is the Fourier transform of $\mathscr{H}_n$. On the other hand, by direct calculation,[j]

$$\begin{aligned}
\int_{-\infty}^{\infty} e^{-t^2/2} g(t,z) e^{-j\omega t}\, dt &= \int_{-\infty}^{\infty} e^{-t^2/2} e^{2tz - z^2} e^{-j\omega t}\, dt \\
&= e^{z^2} \int_{-\infty}^{\infty} e^{-(t-2z)^2/2} e^{-j\omega t}\, dt \\
&= e^{z^2} \cdot e^{-j\omega(2z)} \cdot \sqrt{2\pi}\, e^{-\omega^2/2} \\
&= e^{2\omega(-jz) - (-jz)^2} \cdot \sqrt{2\pi}\, e^{-\omega^2/2} \\
&= g(\omega, -jz) \cdot \sqrt{2\pi}\, e^{-\omega^2/2} \\
&= \sum_{n=0}^{\infty} \frac{(-jz)^n}{n!} H_n(\omega) \sqrt{2\pi}\, e^{-\omega^2/2} \\
&= \sum_{n=0}^{\infty} \frac{z^n}{n!} (-j)^n \sqrt{2\pi}\, \mathscr{H}_n(\omega).
\end{aligned}$$

Comparing the two expansions, we see that

$$\widehat{\mathscr{H}_n}(\omega) = (-j)^n \sqrt{2\pi} \mathscr{H}_n(\omega).$$

If we let $\mathscr{F}: L^2(\mathbb{R}) \to L^2(\mathbb{R})$ denote the Fourier transform, then the above equation says

$$\mathscr{F}\, \mathscr{H}_n = (-j)^n \sqrt{2\pi}\, \mathscr{H}_n.$$

Thus, the Hermite functions are eigenfunctions of the Fourier transform operator regarded as a mapping from $L^2(\mathbb{R})$ to $L^2(\mathbb{R})$. The eigenvalue of the nth eigenfunction is $(-j)^n \sqrt{2\pi}$. The eigenvalues change slightly if a different definition of the transform is used; e.g., if j is replaced by $-j$ or ω is replaced by $2\pi f$. In any case, the eigenvalues have constant absolute value and do not decay to zero. Hence, the Fourier transform is not a compact operator.

It remains to show that the Hermite functions are orthogonal. Since

$$\int_{-\infty}^{\infty} \mathscr{H}_n(t) \mathscr{H}_m(t)\, dt = \int_{-\infty}^{\infty} H_n(t) H_m(t) e^{-t^2}\, dt,$$

[j] Recall that the Fourier transform of $e^{-t^2/2}$ is $\sqrt{2\pi} e^{-\omega^2/2}$, and the transform of $x(t-\theta)$ is $X(\omega) e^{-j\omega\theta}$.

the result will follow if we can show that the Hermite polynomials are orthogonal with respect to the **weight function** $w(t) = e^{-t^2}$. However, this fact is well known [26, p. 65].

Problems

1. ***Discrete Prolate Spheroidal Sequences (DPSS).*** Discuss how you would formulate and try to solve the discrete-time version of the problem in Section 8.2. In other words, re-write Section 8.2 for the discrete-time case.

2. ***Discrete Prolate Spheroidal Wave Functions (DPSWF).*** Let $x(n)$ be a finite-energy, discrete-time sequence whose DTFT is bandlimited to $W < 1/2$. The amount of energy in the time window $0, \ldots, N-1$ is

$$\sum_{n=0}^{N-1} |x(n)|^2 = \sum_{n=-\infty}^{\infty} g(n)x(n)\overline{x(n)} = \langle gx, x \rangle,$$

where

$$g(n) := \begin{cases} 1, & n = 0, \ldots, N-1, \\ 0, & \text{otherwise.} \end{cases}$$

Show that the optimal x is related to the eigenvalue problem

$$\int_{-W}^{W} H(f-\nu)Y(\nu)\,d\nu = \lambda Y(f), \quad |f| \le W,$$

where $H(f) := \sin(\pi N f)/\sin(\pi f)$. The extension of the solutions $Y(f)$ to $f \in (-\infty, \infty)$ are called the **discrete prolate spheroidal wave functions**.

3. Let X be an RKHS of time functions with kernel function $K_t(\tau)$. Show that if $\|x_n - x\| \to 0$, then for each time t, $\lim_{n\to\infty} x_n(t) = x(t)$.

4. Show that the kernel of an RKHS is unique. More explicitly, show that if for every time t and every vector x we have $\langle x, K_t \rangle = \langle x, H_t \rangle$, then for all t and τ, $K_t(\tau) = H_t(\tau)$.

5. In an RKHS with kernel $K_t(\tau)$, given any sequence of times $t_1, \ldots, t_n$, the matrix M defined by $M_{ij} := K_{t_j}(t_i)$ is positive semidefinite.

6. Establish the **three-term recursion** for the Hermite polynomials,

$$H_{m+1}(t) = 2tH_m(t) - 2mH_{m-1}(t), \quad m = 0, 1, 2, \ldots,$$

where it is understood that $H_{-1}(t) \equiv 0$. *Hint:* Observe that

$$\frac{\partial}{\partial z} g(t,z) = (2t - 2z) g(t,z).$$

On the right-hand side substitute the power series for g and do a little algebra. On the left-hand side differentiate the power series for g term by term. Changing the index of summation and equating coefficients of like powers of z yields the recursion.

APPENDIX A
How Proofs Work

We introduce basic concepts of logic and how they are applied to proofs. In the first section, we present the **sentential calculus**, which involves implication ($\Rightarrow$, **conditional**), **negation** ($\neg$), **conjunction** ($\wedge$, logical and), and **disjunction** ($\vee$, inclusive or). In the second section, we present the **quantifier calculus**, which involves the notions of "**for all**" ($\forall$) and "**there exists**" ($\exists$). In the concluding section, we illustrate the apparatus we have developed with applications in mathematics.

A.1. Sentential Calculus

A.1.1. Basic Notation

To see logical structure more clearly, it is convenient to use symbolic notation for sentences. For example, let

$$P := \text{Bob studies} \quad \text{and} \quad Q := \text{Bob does well on the exam.}$$

Then we read

$$P \Rightarrow Q$$

as, "If Bob studies, then Bob does well on the exam."

To negate the sentence P, we write $\neg P$. For example, if we let

$$P := \text{Betty brings a second pen} \quad \text{and} \quad Q := \text{Betty runs out of ink,}$$

then we read

$$P \Rightarrow \neg Q$$

as, "If Betty brings a second pen, then it is not the case that Betty runs out of ink." More informally, we would say, "If Betty brings a second pen, then Betty does not run out of ink."

A.1.2. Basic Inference Rules and Methods of Proof

Suppose we believe the statement, "If Bob studies, then Bob does well on the exam." Suppose also that we believe, "Bob studies." Then we must also believe,

"Bob does well on the exam." This general rule of inference is known as **modus ponens** (MP). We express it symbolically as[a]

$$\frac{\begin{array}{l} P \Rightarrow Q \\ P \end{array}}{Q}$$

Another rule of inference is **double negation**

$$\frac{\neg(\neg P)}{P}$$

An important technique for establishing many results is **proof by contradiction**. In this technique, if you want to prove P, you begin by assuming $\neg P$. Then you carry out a sequence of steps in which at some point you obtain a statement Q and at another point you obtain $\neg Q$. Then the proof of P is complete.

Example A.1. We can use the foregoing ideas to establish the *derived* inference rule **modus tollens**, which says

$$\frac{\begin{array}{r} P \Rightarrow Q \\ \neg Q \end{array}}{\neg P}$$

Solution.

1.	$P \Rightarrow Q$	premise
2.	$\neg Q$	premise
3.	Show $\neg P$	assertion
4.	$\neg(\neg P)$	assumption for proof by contradiction
5.	P	apply DN to line 4
6.	Q	apply MP to lines 1 and 5
7.	$\neg Q$	repeat line 2
8.	End Show 3	proof by contradiction

The foregoing solution is an example of a formal line-by-line proof or derivation. Notice that each line is numbered, and each line is explained or justified by citing a rule of inference, with reference to any relevant line numbers.

Important Rules. Once a "Show" has a matching "End Show," no lines in between can be referenced elsewhere. In the preceding example, if the derivation continued

[a] The statements above the horizontal line are called **premises**. The statement below the line is called the **conclusion**.

beyond line 8, none of the lines 4–7 could be used or referenced after line 8. You may use or reference a "Show" line only after its matching "End Show" line has been obtained. Notice also we have introduced a new inference rule called **repetition** (R). This allowed us to have both Q and $\neg Q$ between the "Show" and the "End Show," which we need for a proof by contradiction.

To prove a conditional of the form $P \Rightarrow Q$, the first step (if necessary) is to assume P and then proceed to derive Q. We call this method **proof of a conditional**.

Example A.2. Prove the alternative double negation rule $P \Rightarrow \neg(\neg P)$.

Solution.

1.	Show $P \Rightarrow \neg(\neg P)$	assertion
2.	P	assumption for conditional derivation
3.	Show $\neg(\neg P)$	assertion
4.	$\neg[\neg(\neg P)]$	assumption for proof by contradiction
5.	$\neg P$	DN 4
6.	P	R 2
7.	End Show 3	proof by contradiction
8.	End Show 1	proof of conditional

Example A.3. Prove that $Q \Rightarrow (P \Rightarrow Q)$.

Solution.

1.	Show $Q \Rightarrow (P \Rightarrow Q)$	assertion
2.	Q	assumption for conditional derivation
3.	Show $P \Rightarrow Q$	assertion
4.	Q	R 2
5.	End Show 3	proof of conditional
6.	End Show 1	proof of conditional

Notice that in the inner proof, we did not need to assume P in order to prove $P \Rightarrow Q$.

A.1.3. More Notation, Inference Rules, and Methods

If two statements P and Q are both true, we write $P \wedge Q$. The two **simplification** (S) rules are

$$\frac{P \wedge Q}{P} \quad \text{and} \quad \frac{P \wedge Q}{Q}$$

In other words, if P and Q are both true, then each of them is true by itself.

To prove $P \wedge Q$, we use the **adjunction** (Adj) inference rule

$$\frac{\begin{array}{c} P \\ Q \end{array}}{P \wedge Q}$$

As our first application of $\wedge$, we write $P \Leftrightarrow Q$ to mean $(P \Rightarrow Q) \wedge (P \Leftarrow Q)$.[b] We call $\Leftrightarrow$ a **biconditional**. By the simplification inference rule, we obtain the **biconditional-conditional** (BC) inference rules

$$\frac{P \Leftrightarrow Q}{P \Rightarrow Q} \quad \text{and} \quad \frac{P \Leftrightarrow Q}{P \Leftarrow Q}$$

By the adjunction rule, we obtain the **conditional-biconditional** (CB) inference rule

$$\frac{\begin{array}{c} P \Rightarrow Q \\ P \Leftarrow Q \end{array}}{P \Leftrightarrow Q}$$

Example A.4. Show that $(P \Leftrightarrow Q) \Leftrightarrow (\neg P \Leftrightarrow \neg Q)$.

Solution. By CB, it suffices to prove

$$(P \Leftrightarrow Q) \Rightarrow (\neg P \Leftrightarrow \neg Q) \quad \text{and} \quad (P \Leftrightarrow Q) \Leftarrow (\neg P \Leftrightarrow \neg Q).$$

We prove the first and leave the second to the reader. We also omit the more obvious justifications.

1.	Show $(P \Leftrightarrow Q) \Rightarrow (\neg P \Leftrightarrow \neg Q)$	
2.	$\quad P \Leftrightarrow Q$	assumption
3.	$\quad P \Rightarrow Q$	BC 2
4.	$\quad P \Leftarrow Q$	BC 2
5.	$\quad$ Show $\neg P \Leftarrow \neg Q$	
6.	$\quad\quad \neg Q$	assumption
7.	$\quad\quad \neg P$	MT 3
8.	$\quad$ End Show 5	
9.	$\quad$ Show $\neg P \Rightarrow \neg Q$	
10.	$\quad\quad \neg P$	assumption
11.	$\quad\quad \neg Q$	MT 4
12.	$\quad$ End Show 9	
13.	$\quad \neg P \Leftrightarrow \neg Q$	CB 5, 9
14.	End Show 1	

[b] We often write $P \Leftarrow Q$ instead of $Q \Rightarrow P$.

Example A.5. Prove that $(P \Rightarrow Q) \Leftrightarrow \neg(P \wedge \neg Q)$.

Solution.

1.	Show $(P \Rightarrow Q) \Leftrightarrow \neg(P \wedge \neg Q)$	
2.	Show $\Rightarrow$	*note the abreviated form*
3.	$P \Rightarrow Q$	assumption
4.	Show $\neg(P \wedge \neg Q)$	
5.	$P \wedge \neg Q$	assumption followed by DN
6.	P	S 5
7.	$\neg Q$	S 5
8.	Q	MP 3, 6
9.	End Show 4	
10.	End Show 2	
11.	Show $\Leftarrow$	*note the abreviated form*
12.	$\neg(P \wedge \neg Q)$	assumption
13.	Show $P \Rightarrow Q$	
14.	P	assumption
15.	Show Q	
16.	$\neg Q$	assumption
17.	P	R 14
18.	$P \wedge \neg Q$	Adj. 16, 17
19.	$\neg(P \wedge \neg Q)$	R 12
20.	End Show 15	
21.	End Show 13	
22.	End Show 11	
23.	End Show 1	CB 2, 11

If either P or Q is true or if both are true, we write $P \vee Q$. To prove $P \vee Q$, we use the **addition** (Add) inference rules

$$\frac{P}{P \vee Q} \quad \text{and} \quad \frac{Q}{P \vee Q}$$

In other words, if P is true, then either P or Q is true. Similarly, if Q is true, then either P or Q is true. Finally, to make inferences from disjunctions, we use the inference rules **modus tollendo ponens** (MTP)

$$\frac{P \vee Q \quad \neg P}{Q} \quad \text{and} \quad \frac{P \vee Q \quad \neg Q}{P}$$

In other words, if either P or Q is true and P is not true, then Q must be true. Similarly, if Q is not true, then P must be true.

Example A.6. Prove that $(P \vee Q) \Leftrightarrow (\neg P \Rightarrow Q)$.

Solution.

1.	Show $(P \vee Q) \Leftrightarrow (\neg P \Rightarrow Q)$	
2.	Show $\Rightarrow$	
3.	$P \vee Q$	assumption
4.	Show $\neg P \Rightarrow Q$	
5.	$\neg P$	assumption
6.	Show Q	
7.	$\neg Q$	assumption
8.	P	MTP 3, 7
9.	$\neg P$	R 5
10.	End Show 6	
11.	End Show 4	
12.	End Show 2	
13.	Show $\Leftarrow$	
14.	$\neg P \Rightarrow Q$	assumption
15.	Show $P \vee Q$	
16.	$\neg (P \vee Q)$	assumption
17.	Show P	
18.	$\neg P$	assumption
19.	Q	MP 14, 18
20.	$P \vee Q$	Add. 19
21.	$\neg (P \vee Q)$	R 16
22.	End Show 17	
23.	$P \vee Q$	Add. 17
24.	End Show 15	contradiction 16, 23
25.	End Show 13	
26.	End Show 1	CB 2, 13

As an exercise, the reader should try to prove De Morgan's laws

$$\neg (P \wedge Q) \Leftrightarrow (\neg P \vee \neg Q)$$

and

$$\neg (P \vee Q) \Leftrightarrow (\neg P \wedge \neg Q).$$

A.2. Quantifier Calculus

A.2.1. Variables

We denote logical variables by x, y, and z. By Mx we mean a statement about x. For example, we may denote by Mx the statement "x is a Martian." Or we may let Tx denote the statement "x has three legs."

A.2.2. Quantifiers

The **universal quantifier** is denoted by $\forall$. It can be read as "for all," "for every," or "for any." The **existential quantifier** is denoted by $\exists$. It can be read as "there exists (at least one, and maybe more)," or "there is (at least one, and maybe more)."

For example, to denote the statement that Martians exist, we write $\exists xMx$, which means there exists at least one x such that x is a Martian. To denote the statement, "All Martians have three legs," we write $\forall x(Mx \Rightarrow Tx)$. We read these symbols as, "For all x, if x is a Martian, then x has three legs."

A.2.3. Inference Rules

Suppose Fx is a statement about x. The inference rule of **universal instantiation** (UI) is

$$\frac{\forall xFx}{Fy}$$

where y is any variable we find useful, even x.

To prove $\forall xFx$, we start with an arbitrary variable, say x, or any other variable that will not be confused with earlier uses, and then show that Fx is true. At this point we end the proof of $\forall xFx$ with the citation **universal derivation**. Hence, the proof has the structure

Show $\forall xFx$
  Show Fx
    $\vdots$
  End Show
End Show universal derivation

To prove $\exists xFx$, we must show that some variable has the property F. We may then conclude that $\exists xFx$, with the inference rule cited being **existential generalization** (EG).

In the following example, we show that if there are no Martians, then all Martians have three legs.

Example A.7. Show that $\neg \exists x Mx \Rightarrow \forall x(Mx \Rightarrow Tx)$.

Solution.

1.	Show $\neg \exists x Mx \Rightarrow \forall x(Mx \Rightarrow Tx)$	
2.	$\neg \exists x Mx$	assumption for conditional derivation
3.	Show $\forall x(Mx \Rightarrow Tx)$	
4.	Show $Mx \Rightarrow Tx$	
5.	Mx	assumption for conditional derivation
6.	$\exists x Mx$	EG 5
7.	$\neg \exists x Mx$	R 2
8.	End Show 4	contradiction 6, 7
9.	End Show 3	universal dervation
10.	End Show 1	

The inference rule of **existential instantiation** (EI) is

$$\frac{\exists x Fx}{Fy}$$

where y is a variable that has not been used earlier in the proof. As you become more adept at proofs, you may re-use variables as long as you do not confuse the new use with an old one.

In the following example, we show that if all Martians have three legs, and if there is at least one Martian, then there is a Martian with three legs.

Example A.8. Show that $[\forall x(Mx \Rightarrow Tx) \wedge \exists x Mx] \Rightarrow \exists x(Mx \wedge Tx)$.

Solution.

1.	Show $[\forall x(Mx \Rightarrow Tx) \wedge \exists x Mx] \Rightarrow \exists x(Mx \wedge Tx)$	
2.	$\forall x(Mx \Rightarrow Tx) \wedge \exists x Mx$	assumption
3.	$\exists x Mx$	S 2
4.	Mz	EI 3
5.	$\forall x(Mx \Rightarrow Tx)$	S 2
6.	$Mz \Rightarrow Tz$	UI 5
7.	Tz	MP 4, 6
8.	$Mz \wedge Tz$	Adj. 4, 7
9.	$\exists x(Mx \wedge Tx)$	EG 8
10.	End Show 1	

Notice that in line 4 we introduced a new variable as required for EI.

The *derived* inference rules known as **quantifier negation** (QN) are

$$\neg\forall xFx \Leftrightarrow \exists x(\neg Fx) \quad \text{and} \quad \neg\exists xFx \Leftrightarrow \forall x(\neg Fx).$$

We prove the first and leave the proof of the second to the reader as an exercise.

1.	Show $\neg\forall xFx \Leftrightarrow \exists x(\neg Fx)$	
2.	Show $\Rightarrow$	
3.	$\neg\forall xFx$	assumption
4.	Show $\exists x(\neg Fx)$	
5.	$\neg\exists x(\neg Fx)$	assumption
6.	Show $\forall xFx$	
7.	Show Fx	
8.	$\neg Fx$	assumption
9.	$\exists x(\neg Fx)$	EG 8
10.	$\neg\exists x(\neg Fx)$	R 5
11.	End Show 7	contradiction 9, 10
12.	End Show 6	
13.	$\neg\forall xFx$	R 3
14.	End Show 4	contradiction 6, 13
15.	End Show 2	
16.	Show $\Leftarrow$	
17.	$\exists x(\neg Fx)$	assumption
18.	Show $\neg\forall xFx$	
19.	$\forall xFx$	assumption followed by DN
20.	$\neg Fz$	EI 17 **— note new variable!**
21.	Fz	UI 19
22.	End Show 18	contradiction 20, 21
23.	End Show 16	
24.	End Show 1	CB 2, 16

A.3. Applications to Mathematics

We now use symbolic logic to do some basic mathematics. Recall that a sequence of real numbers x_n converges to a limit x if for every $\varepsilon > 0$, there is an integer N such that for all $n \geq N$, $|x_n - x| < \varepsilon$. We can write this symbolically as

$$\forall\varepsilon > 0, \exists N\forall n \geq N, |x_n - x| < \varepsilon. \tag{A.1}$$

However, this is not as explicit as we need to use our methods. We write instead

$$\forall\varepsilon[\varepsilon > 0 \Rightarrow \exists N\forall n(n \geq N \Rightarrow |x_n - x| < \varepsilon)]. \tag{A.2}$$

Example A.9. How can we express the statement that x_n does *not* converge to x?

Solution. We apply QN to (A.2) and use Example A.5 and DN to get

$$\exists\varepsilon[\varepsilon > 0 \wedge \neg\,\exists N\forall n(n \geq N \Rightarrow |x_n - x| < \varepsilon)].$$

With two more applications of QN and then Example A.5 and DN, we get

$$\exists\varepsilon[\varepsilon > 0 \wedge \forall N\exists n(n \geq N \wedge |x_n - x| \geq \varepsilon)].$$

From this last expression in the example, by EI, we know that there is an ε_0 such that $\varepsilon_0 > 0$ and

$$\forall N\exists n(n \geq N \wedge |x_n - x| \geq \varepsilon_0).$$

Now apply UI with $N = 1, 2, \ldots$. When $N = k$, we have

$$\exists n(n \geq k \wedge |x_n - x| \geq \varepsilon_0).$$

By EI with n replaced by n_k, we can write

$$n_k \geq k \wedge |x_{n_k} - x| \geq \varepsilon_0.$$

This shows that if x_n does not converge to x, then there is some $\varepsilon_0 > 0$ and there is a subsequence[c] x_{n_k} with $|x_{n_k} - x| \geq \varepsilon_0$ holding for all $k = 1, 2, \ldots$.

Continue on the next page.

[c] The definition of a subsequence requires that $n_k \to \infty$; this is guaranteed by the condition $n_k \geq k$.

Example A.10. We now show that if x_n converges to x and if y_n converges to y, then $x_n + y_n$ converges to $x + y$.

1.	$\forall\varepsilon[\varepsilon > 0 \Rightarrow \exists N\forall n(n \geq N \Rightarrow \lvert x_n - x\rvert < \varepsilon)]$	premise
2.	$\forall\varepsilon[\varepsilon > 0 \Rightarrow \exists N\forall n(n \geq N \Rightarrow \lvert y_n - y\rvert < \varepsilon)]$	premise
3.	Show $\forall\varepsilon[\varepsilon > 0 \Rightarrow \exists N\forall n(n \geq N \Rightarrow \lvert(x_n + y_n) - (x + y)\rvert < \varepsilon)]$	assertion
4.	Show $\varepsilon > 0 \Rightarrow \exists N\forall n(n \geq N \Rightarrow \lvert(x_n + y_n) - (x + y)\rvert < \varepsilon)$	
5.	$\varepsilon > 0$	
6.	$\varepsilon/2 > 0 \Rightarrow \exists N\forall n(n \geq N \Rightarrow \lvert x_n - x\rvert < \varepsilon/2)$	UI 1
7.	$\varepsilon/2 > 0$	by 5
8.	$\exists N\forall n(n \geq N \Rightarrow \lvert x_n - x\rvert < \varepsilon/2)$	MP 6, 7
9.	$\forall n(n \geq N_1 \Rightarrow \lvert x_n - x\rvert < \varepsilon/2)$	EG 8
10.	$\varepsilon/2 > 0 \Rightarrow \exists N\forall n(n \geq N \Rightarrow \lvert y_n - y\rvert < \varepsilon/2)$	UI 2
11.	$\exists N\forall n(n \geq N \Rightarrow \lvert y_n - y\rvert < \varepsilon/2)$	MP 7, 10
12.	$\forall n(n \geq N_2 \Rightarrow \lvert y_n - y\rvert < \varepsilon/2)$	EG 12
13.	Show $\forall n(n \geq \max(N_1, N_2) \Rightarrow \lvert(x_n + y_n) - (x + y)\rvert < \varepsilon)$	
14.	Show $n \geq \max(N_1, N_2) \Rightarrow \lvert(x_n + y_n) - (x + y)\rvert < \varepsilon$	
15.	$n \geq \max(N_1, N_2)$	
16.	$n \geq N_1$	by 15
17.	$n \geq N_1 \Rightarrow \lvert x_n - x\rvert < \varepsilon/2$	UI 9
18.	$n \geq N_2$	by 15
19.	$n \geq N_2 \Rightarrow \lvert y_n - y\rvert < \varepsilon/2$	UI 12
20.	$\lvert x_n - x\rvert < \varepsilon/2$	MP 16, 17
21.	$\lvert y_n - y\rvert < \varepsilon/2$	MP 18, 19
22.	$\lvert(x_n + y_n) - (x + y)\rvert = \lvert(x_n - x) + (y_n - y)\rvert$	math
23.	$\lvert(x_n + y_n) - (x + y)\rvert \leq \lvert x_n - x\rvert + \lvert y_n - y\rvert$	tri. ineq.
24.	$\lvert(x_n + y_n) - (x + y)\rvert < \varepsilon/2 + \varepsilon/2 = \varepsilon$	20, 21, 23
25.	End Show 14	
26.	End Show 13	
27.	$\exists N\forall n(n \geq N \Rightarrow \lvert(x_n + y_n) - (x + y)\rvert < \varepsilon)$	EG 13
28.	End Show 4	
29.	End Show 3	

Mathematicians do not write proofs as in the preceding example. Here is how that proof would normally look (keep in mind that $x_n \to x$ is defined by (A.1)).

Theorem. If $x_n \to x$ and $y_n \to y$, then $(x_n + y_n) \to (x + y)$.

Proof. Let $\varepsilon > 0$ be given. Since $x_n \to x$, there is an N_1 such that for all $n \geq N_1$,

$$\lvert x_n - x\rvert < \varepsilon/2. \tag{A.3}$$

Similarly, since $y_n \to y$, there is an N_2 such that for all $n \geq N_2$,

$$|y_n - y| < \varepsilon/2. \tag{A.4}$$

Then for $n \geq \max(N_1, N_2)$, (A.3) and (A.4) both hold, and we have by the triangle inequality that

$$|(x_n + y_n) - (x + y)| = |(x_n - x) + (y_n - y)| \leq |x_n - x| + |y_n - y| < \varepsilon/2 + \varepsilon/2 = \varepsilon. \tag{A.5}$$

Starting with an arbitrary $\varepsilon > 0$, we showed that there exists an integer N, namely $\max(N_1, N_2)$, such that for all $n \geq N$, (A.5) holds. □

When confronted with an assertion such as that in the above theorem, the proof does not magically construct itself. We start with some analysis. In this case, we start with the definition of the limit of a sequence and apply it to $x_n + y_n$ and $x + y$. We then see that we need to end up with

$$|(x_n + y_n) - (x + y)| < \varepsilon.$$

If we know the triangle inequality, we proceed as in (A.5). This shows that we can get a bound of ε if we can get a bound of $\varepsilon/2$ as in (A.3) and (A.4). We know we can do this because we are assuming that $x_n \to x$ and $y_n \to y$. We are now in a position to write down a careful proof.

APPENDIX B
Mathematical Induction

Consider a statement about positive integers, for example

$$\mathrm{P}(n): \quad \sum_{k=1}^{n} k = \frac{n(n+1)}{2}, \tag{B.1}$$

where $n \geq 1$. Using (B.1), we see that $\mathrm{P}(1)$ is given by

$$\sum_{k=1}^{1} k = \frac{1(1+1)}{2},$$

which is obviously true. When $\mathrm{P}(n)$ is given by (B.1), $\mathrm{P}(n+1)$ is given by

$$\begin{aligned} \sum_{k=1}^{n+1} k &= \frac{[n+1]([n+1]+1)}{2} \\ &= \frac{(n+1)(n+2)}{2}. \end{aligned} \tag{B.2}$$

For a general statement $\mathrm{P}(n)$, to prove that it is true for all positive integers using **mathematical induction** on n is to carry out the following two-step procedure:

1. Show that $\mathrm{P}(1)$ is true.
2. Fix an arbitrary $n \geq 1$ and show that if $\mathrm{P}(n)$ is true, then $\mathrm{P}(n+1)$ is true; i.e., show that for $n \geq 1$,
$$\mathrm{P}(n) \Rightarrow \mathrm{P}(n+1).$$

We note that sometimes it is more convenient to prove that for all $n \geq 2$, we have $\mathrm{P}(n-1) \Rightarrow \mathrm{P}(n)$.

When $\mathrm{P}(n)$ is given by (B.1), we have already noted that $\mathrm{P}(1)$ is true. We now show that $\mathrm{P}(n) \Rightarrow \mathrm{P}(n+1)$. Suppose $\mathrm{P}(n)$ is true. We must show that $\mathrm{P}(n+1)$ is true; i.e., we most show that (B.2) holds. So we write

$$\begin{aligned} \sum_{k=1}^{n+1} k &= \sum_{k=1}^{n} k + (n+1) \\ &= \frac{n(n+1)}{2} + (n+1), \quad \text{by the induction hypothesis } \mathrm{P}(n), \\ &= \frac{n(n+1)}{2} + \frac{2(n+1)}{2} \\ &= \frac{(n+1)(n+2)}{2}. \end{aligned}$$

Example B.1. Show that if $h \geq -1$, then

$$\mathrm{P}(n): \quad (1+h)^n \geq 1+nh \tag{B.3}$$

holds for $n \geq 1$.

Solution. First, when $n = 1$, (B.3) becomes $(1+h) \geq 1+h$, which is obviously true. Suppose $(1+h)^n \geq 1+nh$. We must show that (B.3) holds for n replaced by $n+1$; i.e., we must show that

$$(1+h)^{n+1} \geq 1+(n+1)h.$$

To derive this, write

$$\begin{aligned}(1+h)^{n+1} &= (1+h)^n(1+h) \\ &\geq (1+nh)(1+h), \quad \text{by the induction hypothesis (B.3),} \\ &= 1+nh+h+nh^2 \\ &\geq 1+(n+1)h, \qquad \text{since } nh^2 \geq 0.\end{aligned} \tag{B.4}$$

Note that the inequality in (B.4) requires $(1+h) \geq 0$; i.e., $h \geq -1$.

Sometimes it is convenient to start the induction at $n = 0$.

Example B.2 (Division Algorithm). Show that if d is a positive integer, then

$$\mathrm{P}(n): \quad n = dq+r, \quad 0 \leq r < d, q \geq 0,$$

holds for $n \geq 0$.

Solution. If $n = 0$, we can take $q = r = 0$. Now suppose the result is true for some $n \geq 0$. To show it is true for $n+1$, use the induction hypothesis to write

$$n = dq+r, \quad 0 \leq r < d.$$

Adding one to both sides of the above equation shows that

$$n+1 = dq+(r+1).$$

If $r+1 < d$, we have the desired representation of $n+1$. Otherwise $r+1 = d$, and the above display becomes $n+1 = d(q+1)$.

Problems

1. Show that for $n \geq 1$,
$$\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}.$$

2. Show that for $n \geq 1$,
$$\sum_{k=1}^{n} (2k-1)^2 = \frac{n(2n-1)(2n+1)}{3}.$$

3. Recall that $\binom{n}{k} := n!/(k!\,(n-k)!)$. Derive the **Leibniz rule**,
$$(xy)^{(n)} = \sum_{k=0}^{n} \binom{n}{k} x^{(k)} y^{(n-k)}, \quad n \geq 0,$$
where $x^{(n)}$ is the nth derivative of x. In particular, $x^{(0)} = x$, $x^{(1)} = x'$, $x^{(2)} = x''$, etc. *Hint:* The easily verified identity
$$\binom{n}{k} + \binom{n}{k-1} = \binom{n+1}{k}, \quad n \geq k \geq 1,$$
may be useful.

APPENDIX C
Compact Sets

Let X be a metric space, and let E be a subset of X. If $\{U_\alpha\}$ is a collection of open subsets of X such that

$$E \subset \bigcup_\alpha U_\alpha, \tag{C.1}$$

then $\{U_\alpha\}$ is called an **open cover** of E.

The set E is said to be **compact** if every open cover of E can be reduced to a finite subcover. In other words, whenever (C.1) holds for a collection of open sets $\{U_\alpha\}$, there exists a finite subcollection, say $\{U_{\alpha_1}, \ldots, U_{\alpha_n}\}$ for some finite n, such that

$$E \subset \bigcup_{i=1}^{n} U_{\alpha_i}.$$

In the problems at then end of this appendix, you are asked to show that compact sets have the following properties.

- A closed subset of a compact set is compact.
- A compact subset is bounded.
- A compact set must be closed.

Theorem C.1 (Finite Intersection Property). *In a metric space X, for each α, let K_α be a compact set. Suppose that for every finite collection of α's, say $\alpha_1, \ldots, \alpha_n$,*

$$\bigcap_{i=1}^{n} K_{\alpha_i} \neq \varnothing.$$

Then

$$\bigcap_\alpha K_\alpha \neq \varnothing.$$

Proof. To obtain a contradiction, suppose

$$\bigcap_\alpha K_\alpha = \varnothing.$$

Setting $G_\alpha := K_\alpha^{\mathrm{c}}$, this is equivalent to[a]

$$\bigcup_\alpha G_\alpha = X \supset K_{\alpha_0},$$

[a] Since compact sets are closed, G_α is open.

for any fixed α_0. Since K_{α_0} is compact, for some finite set of α's, say $\alpha_1, \ldots, \alpha_n$,

$$K_{\alpha_0} \subset \bigcup_{i=1}^{n} G_{\alpha_i}.$$

Hence,

$$\varnothing = K_{\alpha_0} \cap \left(\bigcup_{i=1}^{n} G_{\alpha_i} \right)^{\mathrm{c}} = K_{\alpha_0} \cap K_{\alpha_1} \cap \cdots \cap K_{\alpha_n}. \qquad \square$$

Given a set E and a point $x \in X$ (x need not be in E), we say that x is an **accumulation point of** E (or **cluster point of** E or **limit point of** E) if for every open set containing x, say O_x, there is a point $y \neq x$ with $y \in O_x \cap E$.

Lemma C.2. *Let K be a compact subset of a metric space, and suppose that E is an infinite subset of K. Then at least one point in K is a limit point of E.*

Proof. Suppose that no point of K is an accumulation point of E. Then for each $x \in K$, there is an open set U_x containing no points of E (except for x itself when $x \in E$). The sets U_x form an open cover of K, and hence can be reduced to a finite subcover of $K \supset E$. But no such subcover could contain more than finitely many points from E, and hence, $K \not\supset E$. $\square$

The next result shows that if K is a compact subset of a metric space, then K is sequentially compact.

Corollary C.3. *If x_n is a sequence contained in a compact subset K of a metric space X, then there is a converging subsequence x_{n_k} with limit $x \in K$.*

Proof. First suppose that the set $E = \{x_n\}$ is infinite. Then by the preceding lemma, there is a point $x \in K$ that is an accumulation point of E. Consider the open sets $B(x, 1/k)$. By the definition of accumulation point, there is a point $x_{n_k} \in B(x, 1/k) \cap E$ with $x_{n_k} \neq x$. Since, $\rho(x_{n_k}, x) < 1/k$, $x_{n_k} \to x$. Also, note that $n_k \to \infty$.

Now suppose that E is finite. Then for at least one point of E, say x, we must have $x_n = x$ for infinitely many n. Let $n_1 < n_2 < \cdots$ denote these distinct values of n. We then have $x_{n_k} \to x$ as $k \to \infty$. $\square$

We also have the converse result that in a metric space, if K is sequentially compact, then K is compact. This is an easy consequence of the following two technical lemmas.

For any set K and any $\varepsilon > 0$, K can be covered by open balls by writing

$$K \subset \bigcup_{x \in K} B(x, \varepsilon). \tag{C.2}$$

Lemma C.4. *If K is a sequentially compact subset of a metric space (X,ρ), then for every $\varepsilon > 0$, the cover of open balls in* (C.2) *can be reduced to a finite subcover; i.e., there exist $x_1,\dots,x_n$ in K such that*

$$K \subset \bigcup_{i=1}^{n} B(x_i,\varepsilon).$$

Proof. Suppose that for some $\varepsilon_0 > 0$, there is no such finite subcover. Fix any $x_1 \in K$. Then we cannot have $K \subset B(x_1,\varepsilon_0)$. Hence, there is an $x_2 \in K \cap B(x_1,\varepsilon_0)^{\mathrm{c}}$. In particular, $\rho(x_2,x_1) \geq \varepsilon_0$. Now suppose $x_1,\dots,x_n$ have been chosen so that for $i \neq j$ with $i,j \in \{1,\dots,n\}$, $\rho(x_i,x_j) \geq \varepsilon_0$. Since there can be no finite subcover of open balls of radius ε_0, we cannot have

$$K \subset \bigcup_{i=1}^{n} B(x_i,\varepsilon_0).$$

Hence, there must be an

$$x_{n+1} \in K \cap \left(\bigcup_{i=1}^{n} B(x_i,\varepsilon_0) \right)^{\mathrm{c}}.$$

In particular, for $i = 1,\dots,n$, we must have $\rho(x_{n+1},x_i) \geq \varepsilon_0$. Clearly, this sequence is not Cauchy, nor can any subsequence be Cauchy. However, by the sequential compactness of K, there must be a converging (and hence, Cauchy) subsequence, which is a contradiction. □

The next lemma allows us to go from the special case of open balls to arbitrary open covers.

Lemma C.5. *If K is a sequentially compact subset of a metric space (X,ρ), then for every open cover $\mathscr{U} = \{U\}$ of K, there is an $\varepsilon_0 > 0$ such that for all $x \in K$, there is an open set $U_x \in \mathscr{U}$ with*

$$B(x,\varepsilon_0) \subset U_x.$$

Remark. A number ε_0 having the property asserted in the lemma is called a **Lebesgue number** for the cover.

Proof. To obtain a contradiction, suppose that for every $\varepsilon_0 = 1/n$, there is an $x_n \in K$ such that

$$B(x_n,1/n) \not\subset U, \quad \text{for all } U \in \mathscr{U}. \tag{C.3}$$

By sequential compactness, there is a converging subsequence $x_{n_k} \to x \in K$ as $k \to \infty$. Since $x \in K$, and $\mathscr{U}$ is an open cover of K, $x \in U_0$ for some $U_0 \in \mathscr{U}$. Since U_0 is

open and contains x, for large m, we must have $B(x,2/m) \subset U_0$. By convergence of the subsequence, we must have $x_{n_k} \in B(x,1/m)$ for large k. We now show that $B(x_{n_k},1/n_k) \subset B(x,2/m)$ for large k. Fix any $y \in B(x_{n_k},1/n_k)$. Then $\rho(y,x_{n_k}) < 1/n_k$. Hence,

$$\begin{aligned}\rho(y,x) &\leq \rho(y,x_{n_k}) + \rho(x_{n_k},x)\\ &< 1/n_k + 1/m,\end{aligned}$$

which is less than $2/m$ for large k since $n_k \to \infty$ by the definition of a subsequence. We thus have, for large k,

$$B(x_{n_k},1/n_k) \subset B(x,2/m) \subset U_0,$$

which contradicts (C.3). □

Theorem C.6. *A subset of a metric space is sequentially compact ⇔ it is compact.*

Proof. ($\Rightarrow$): Suppose K is sequentially compact and has an open cover $\mathscr{U}$. For this cover, let $\varepsilon_0 > 0$ be given by Lemma C.5. Then for every $x \in K$, there is a $U_x \in \mathscr{U}$ with

$$B(x,\varepsilon_0) \subset U_x. \tag{C.4}$$

Next, in Lemma C.4 take $\varepsilon = \varepsilon_0$ to obtain

$$\begin{aligned}K &\subset \bigcup_{i=1}^{n} B(x_i,\varepsilon_0).\\ &\subset \bigcup_{i=1}^{n} U_{x_i}, \quad \text{taking } x = x_i \text{ in (C.4).}\end{aligned}$$

($\Leftarrow$): This is simply Corollary C.3. □

Theorem C.7 (Heine–Borel). *A subset E of a finite-dimensional Euclidean space is compact ⇔ the subset is both closed and bounded.*

Proof. ($\Rightarrow$): By Problem C.2, E is bounded. By Problem C.3, E must be closed.

($\Leftarrow$): By Theorem 6.41, if a subset of finite-dimensional Euclidean space is closed and bounded, it is sequentially compact, and therefore compact by Theorem C.6. □

Problems

1. Let E be a compact subset of a metric space, and let F be a closed subset of E. Show that F is compact.

2. Let K be a compact subset of a metric space. Show that K is bounded in the sense that there is a finite, nonnegative number b and some $x_0 \in X$ such that $\rho(x,x_0) \leq b$ for all $x \in K$.

3. Show that if E is a compact set in a metric space, then E must be closed. *Hint:* Show that E^c is open; i.e., fix any $x \in E^c$ and construct an open set U containing x. Then prove that $U \subset E^c$ by showing that $U \cap E = \varnothing$.

Bibliography

[1] G. Bachman and L. Narici, *Functional Analysis*. Mineola, NY: Dover, 2000.

[2] C. T. H. Baker, *The Numerical Treatment of Integral Equations*. Oxford, UK: Oxford Univ. Press, 1977.

[3] A. V. Balakrishnan, *Applied Functional Analysis*, 2nd ed. New York: Springer, 1981.

[4] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. New York: Springer, 2011.

[5] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[6] P. Billingsley, *Probability and Measure*, 3rd ed. New York: Wiley, 1995.

[7] S. Bubeck, "Convex optimization: Algorithms and complexity," *Foundations and Trends in Machine Learning*, vol. 8, no. 3–4, pp. 231–357, 2015.

[8] J. V. Burke, "Markowitz mean-variance portfolio theory," [Online]. Available: `https://sites.math.washington.edu/~burke/crs/408/fin-proj/`, accessed Apr. 23, 2017.

[9] R. V. Churchill, J. W. Brown, and R. F. Verhey, *Complex Variables and Applications*, 3rd ed. New York: McGraw-Hill, 1976.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley, 2006.

[11] L. M. Delves and J. L. Mohamed, *Computational Methods for Integral Equations*. Cambridge, UK: Cambridge Univ. Press, 1985.

[12] L. Elden, *Matrix Methods in Data Mining and Pattern Recognition*. Philadelphia: SIAM, 2007.

[13] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 586–597, Dec. 2007.

[14] G. B. Folland, *Real Analysis: Modern Techniques and Their Applications*, 2nd ed. New York: Wiley, 1999.

[15] W. Gautschi, "The interplay between classical analysis and (numerical) linear algebra — A tribute to Gene H. Golub," *Electronic Transactions on Numerical Analysis*, vol. 13, pp. 119–147, 2002. `http://www.emis.ams.org/journals/ETNA/`

[16] W. Gautschi, *Orthogonal Polynomials: Computation and Approximation*. Oxford, UK: Oxford Univ. Press, 2004.

[17] I. Gohberg and S. Goldberg, *Basic Operator Theory*. Boston: Birkhäuser, 1980.

[18] G. H. Golub and J. H. Welsch, "Calculation of Gauss quadrature rules," *Math. Comp.*, vol. 23, no. 106, pp. 221–230, addendum: supplement pp. A1–A10, 1969.

[19] R. G. Gordon, "Error bounds in equilibrium statistical mechanics," *J. Math. Phys.*, vol. 9, no. 5, pp. 655–663, 1968.

[20] B. C. Hall, *Quantum Theory for Mathematicians*. New York: Springer, 2013.

[21] D. Kalish, R. Montague, and G. Mar, *Logic: Techniques of Formal Reasoning*, 2nd ed. New York: Harcourt, 1980.

[22] H. Komiya, "Elementary proof for Sion's minimax theorem," *Kodai Math. J.*, vol. 11, no. 1, pp. 5–7, 1988.

[23] R. Kress, *Numerical Analysis*, 2nd ed. New York: Springer, 1998.

[24] H. J. Landau and H. Widom, "Eigenvalue distribution of time and frequency limiting," *J. Math. Anal. Appl.*, vol. 77, pp. 469–481, 1980.

[25] S. R. Lay, *Convex Sets and Their Applications*. Mineola, NY: Dover, 2007.

[26] N. N. Lebedev, *Special Functions and Their Applications*. New York: Dover, 1972.

[27] D. G. Luenberger, *Optimization by Vector Space Methods*. New York: Wiley, 1969.

[28] D. G. Luenberger and Yinyu Ye, *Linear and Nonlinear Programming*, 3rd ed. New York: Springer 2010.
[29] H. M. Markowitz, *Portfolio Selection: Efficient Diversification of Investments*. New Haven: Cowles Foundation for Research in Economics at Yale Univ., 1959.
[30] J. Nocedal and S. J. Wright *Numerical Optimization*, 2nd ed. New York: Springer, 2006.
[31] O. L. Mangasarian, *Nonlinear Programming*. New York: McGraw-Hill, 1969.
[32] R. T. Rockafellar, *Convex Analysis*. Princeton, NJ: Princeton Univ. Press, 1970.
[33] H. L. Royden, *Real Analysis*, 2nd ed. New York: MacMillan, 1968.
[34] W. Rudin, *Real and Complex Analysis*, 2nd ed. New York: McGraw-Hill, 1974.
[35] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed. New York: McGraw-Hill, 1976.
[36] D. Slepian, "On bandwidth," *Proc. IEEE*, vol. 64, no. 3, pp. 292–300, Mar. 1976.
[37] D. Slepian, "Prolate spheroidal wave functions, Fourier analysis, and uncertainty — V: The discrete case," *Bell Syst. Tech. J.*, vol. 57, no. 5, pp. 1371–1430, May–June, 1978.
[38] D. Slepian, "Some comments on Fourier analysis, uncertainty and modeling," *SIAM Rev.*, vol. 25, no. 3, pp. 379–393, July 1983.
[39] D. Slepian and E. Sonnenblick, "Eigenvalues associated with prolate spheroidal wave functions of zero order," *Bell Syst. Tech. J.*, vol. 44, pp. 1745–1759, Oct. 1965.
[40] "The Prize in Economics 1990 - Press Release," [Online]. Available: `http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/1990/press.html`, accessed Mar. 20, 2017.
[41] S. Vajda, *Theory of Linear and Nonlinear Programming*. London: Longman, 1974.
[42] J. van Tiel, *Convex Analysis: An Introductory Text*. Chichester: Wiley, 1984.
[43] P. Whittle, *Optimization under Constraints: Theory and Applications of Nonlinear Programming*. London, U.K.: Wiley, 1971.
[44] Wikipedia, "Wronskian — Wikipedia, The Free Encyclopedia," [Online]. Available: `http://en.wikipedia.org/w/index.php?title=Wronskian&oldid=579284887`, accessed Mar. 20, 2014.
[45] H. S. Wilf, *Mathematics for the Physical Sciences*. New York: Wiley, 1962. `http://www.math.upenn.edu/%7Ewilf/website/Mathematics_for_the_Physical_Sciences.html`
[46] G. Milton Wing, *A Primer on Integral Equations of the First Kind*. Philadelphia: SIAM, 1991.
[47] D. C. Youla, "The use of the method of maximum likelihood in estimating continuous-modulated intelligence which has been corrupted by noise," *Trans. IRE Prof. Group on Inform. Theory*, PGIT-3, pp. 90–106, Mar. 1954.

Index