Contents

CHAPTER 1

Numbers

Engineers, scientists, and applied mathematicians think about numbers all the time, but usually in a utilitarian way. We manipulate them, calculate with them, make plans based on them, drive cars and fly in airplanes whose design depended on them, and so on. We don't spend a lot of time thinking about what numbers "actually are" or "where they come from." Numbers are just kind of "out there." We all have our own ways of visualizing numbers, and we understand on some level how different sorts of numbers are related. We think of the integers, for example, as forming a subset of the rational numbers, which in turn form a subset of the real numbers. And the real numbers constitute a subset of the complex numbers. Once in a while, it pays to spend some time pondering numbers a bit more deeply than we usually do. That's what I'll attempt in what follows.

Sets, mappings, cardinality, and the natural numbers

First let's talk about sets. You have to be careful when you define what "set" means if you want to have a "set theory" that works in the sense that it doesn't lead to logical contradictions. For example, if you allow any arbitrary collection of objects to be a set, then Russell's Paradox comes into play.

Here's how Russell's Paradox goes. Specifying a set entails, in some sense, listing the elements of the set. So you can think of a set simply as a list. Here's an example of a list:

(1) Collegetown Bagels
(2) Radiohead
(3) One teaspoon of salt
(4) Nitroglycerin

Here's another one:

(1) Sibley Dome
(2) Stella's
(3) This list
(4) The Foo Fighters
(5) The album *Daydream Nation* by Sonic Youth

The first list, although it enumerates some rather unrelated things, is not as unusual as the second list. The second list contains itself as a list item. Let's call a list *anomalous* if it lists itself as an item. Next, define L_R as the list of all non-anomalous lists. Question: is L_R anomalous? I.e., does L_R list itself as an item? If L_R lists itself as an item, then L_R is by definition anomalous — but that's a contradiction, since L_R lists only non-anomalous lists by construction. On the other hand, if L_R does not list itself as an item, then L_R is by definition non-anomalous — but that, too, is a contradiction, since L_R lists *all* non-anomalous lists by construction, hence would have to list itself if it were non-anomalous.

Behold Russell's Paradox. What went wrong? Essentially, we attempted to define "set" too broadly. It won't do merely to say that a set is any collection of objects. Only certain collections of objects can qualify as sets if we are to have a set theory immune to Russell-type paradoxes. Different stipulations of exactly what collections to deem sets underlie different approaches to set theory. Many such different approaches are mathematically viable and several are particularly popular with working mathematicians. One approach that has become somewhat of an industry standard is the so-called *ZFC axiomatization* of set theory. ZFC stands for "Zermelo-Fraenkel axioms together with the Axiom of Choice." Search online for ZFC and see what you hit.

Let's assume henceforth that we're employing a working version of set theory in which all the elementary set operations make sense. We have the empty set ϕ. If A and B are two sets, then

$$A \cup B = \{c : c \in A \text{ or } c \in B\}$$

and

$$A \cap B = \{c : c \in A \text{ and } c \in B\} .$$

Read these last two expressions, respectively, as "the union of A and B is the set of all c such that c is an element of A or c is an element of B" and "the intersection of A and B is the set of all c such that c is an element of A and c is an element of B." If A and B are sets, the *Cartesian product* $A \times B$ of A and B is the set of all ordered pairs (a, b) where $a \in A$ and $b \in B$. In math_speak:

$$A \times B = \{(a, b) : a \in A \text{ and } b \in B\} .$$

What about numbers? Define the set $\mathbb{N}$ of *natural numbers* as

$$\mathbb{N} = \{0, 1, 2, 3, 4, \ldots\} .$$

I include zero in $\mathbb{N}$ even though most books don't. Any decent version of set theory includes $\mathbb{N}$ among its sets. But wait, what "is" a natural number? What is 17, for example? It's really a concept — "the idea of 17-ness" or something like that. If you want to define what a number is more concretely, you have many options. Here, for example, is a way to construct the natural numbers, almost literally, out of nothing. I first read about it in Rudy Rucker's remarkable book *Infinity and the Mind*, which I recommend strongly to anyone interested in further reading on sets, numbers, infinity, and related ideas. The empty set ϕ makes sense to everyone, so let's start there. Define the natural number 0 as ϕ. Then define 1 as $\{\phi\}$ — i.e., the set whose only element is the empty set. Define 2 as $\{\{\phi\}\}$, 3 as $\{\{\{\phi\}\}\}$, etc.

Rucker motivates this construction of the natural numbers with the idea that you won't go far wrong if you adopt the position that "everything in the universe is a set." That idea has philosophical merit, and Rucker speaks to it eloquently. We'll see how useful it is when discussing and defining the rational, real, and complex numbers. You may think it's silly, and that's okay. You'll be fine for now as long as you have a conceptual architecture that accommodates natural numbers comfortably.

Now let's talk about mappings between sets. The notation for a mapping f from a set A to a set B is

$$f : A \longrightarrow B \; .$$

This means that to each $a \in A$ there corresponds a unique element of B called $f(a)$. People also write

$$a \mapsto f(a)$$

when A and B are clear from the context. A mapping $f : A \longrightarrow B$ is

- *injective* when no two distinct elements of A map to the same element of B under f. In technical terms: f is injective when for all $a_1, a_2 \in A$, if $a_1 \neq a_2$, then $f(a_1) \neq f(a_2)$.
- *surjective* when for every $b \in B$ there's some $a \in A$ such that $f(a) = b$; in other words, f maps A *onto* B; i.e. the mapping f "hits" every element of B.
- *bijective* when it's both injective and surjective. A bijective mapping $f : A \longrightarrow B$ establishes a one-to-one correspondence between the elements of A and the elements of B.

A set A is *finite* when either $A = \phi$ or for some $N \in \mathbb{N}$ there is a bijective mapping

$$f : \{0, 1, 2, \ldots, N-1\} \longrightarrow A \; ,$$

and in this case we say that A *has cardinality* N. By convention, the empty set ϕ has cardinality 0. A set A is *infinite* when A is not finite. What about the word "cardinality?" The cardinality of a set is, roughly speaking, the size of the set. If A is a finite set, the cardinality of A is simply the number of elements in A. Cardinality gets interesting and touchy when you deal with infinite sets, as we'll discover shortly.

If A is any set, the *power set of* A, written $\mathcal{P}(A)$ or 2^A, is the set of all subsets of A. In math_speak,

$$\mathcal{P}(A) = \{S : S \subset A\} \; .$$

Read that last line as, "The power set of A is the set of all S such that S is a subset of A." For example, if $A = \{a_1, a_2\}$ is any set with two elements, the power set of A is

$$\{\phi, \{a_1\}, \{a_2\}, \{a_1, a_2\}\} \; .$$

For any set A, the empty set ϕ and A itself are subsets of A, hence elements of $\mathcal{P}(A)$.

The cardinalities of A and $\mathcal{P}(A)$ when A is finite bear a simple relationship, namely: if A has N elements, then $\mathcal{P}(A)$ has 2^N elements. To see why this is true, suppose

$$A = \{a_1, a_2, \ldots, a_N\} \; .$$

Let B_N be the set of binary strings of length N. A typical element of B takes the form

$$b_1 b_2 b_3 \cdots b_N \; ,$$

where each b_i is either 0 or 1. Define a mapping $f : \mathcal{P}(A) \longrightarrow B_N$ as follows: for each $S \in \mathcal{P}(A)$, $f(S) \in B_N$ is the binary string b prescribed by

$$b_i = \begin{cases} 1 & \text{if } a_i \in S \\ 0 & \text{if } a_i \notin S \end{cases}$$

for $1 \le i \le N$. The mapping f is bijective, and, since B_N has 2^N elements, so does $\mathcal{P}(A)$. Note that f maps the empty set to the string of all 0's and f maps A to the string of all 1's.

We won't struggle to define exactly what we mean by "the cardinality of A" when A is infinite. We'll be more interested in statements about how cardinalities of different infinite sets compare to each other. Here are the key definitions. Given any two sets A and B, we say that

- A and B have the same cardinality — which we abbreviate by writing $\mathrm{card}(A) = \mathrm{card}(B)$ — when there exists a bijective mapping $f : A \longrightarrow B$. In other words, A and B have the same cardinality when we can establish a one-to-one correspondence between the elements of A and the elements of B.
- A has cardinality less than or equal to that of B — abbreviated $\mathrm{card}(A) \le \mathrm{card}(B)$ — when there exists an injective mapping $f : A \longrightarrow B$.

It's comforting to note that when $A \subset B$, $\mathrm{card}(A) \le \mathrm{card}(B)$. To see why, observe that the mapping $f : A \to B$ defined by $f(a) = a$ for every $a \in A$ is trivially injective. Observe also that if $\mathrm{card}(A) = \mathrm{card}(B)$, then $\mathrm{card}(A) \le \mathrm{card}(B)$. Furthermore, it's easy to show that cardinality comparisons obey the transitive laws

$$\mathrm{card}(A) \le \mathrm{card}(B) \text{ and } \mathrm{card}(B) \le \mathrm{card}(C) \Longrightarrow \mathrm{card}(A) \le \mathrm{card}(C)$$

and

$$\mathrm{card}(A) = \mathrm{card}(B) \text{ and } \mathrm{card}(B) = \mathrm{card}(C) \Longrightarrow \mathrm{card}(A) = \mathrm{card}(C)\ .$$

I'd like to stress here that the "cardinality" of an infinite set is not a number, *per se.* The statement "$\mathrm{card}(A) \le \mathrm{card}(B)$" is not a relationship between numbers when A and B are infinite. Rather, it's a statement about the existence of a mapping establishing a one-to-one correspondence between A and a subset of B. It's true (but not obvious) that, using our definitions, if $\mathrm{card}(A) \le \mathrm{card}(B)$ and $\mathrm{card}(B) \le \mathrm{card}(A)$, then $\mathrm{card}(A) = \mathrm{card}(B)$. The Schröder-Bernstein Theorem establishes this fact. It's also true that a set A is infinite if and only if $\mathrm{card}(\mathbb{N}) \le \mathrm{card}(A)$ — i.e., if and only if there exists an injective mapping $f : \mathbb{N} \longrightarrow A$.

Comparing cardinalities of infinite sets generates a whole hierarchy of "levels of infinity" that we won't have time to explore. As it happens, the set $\mathbb{N}$ of natural numbers is in some sense the smallest of all infinite sets. We say a set A is *countably infinite* when $\mathrm{card}(A) = \mathrm{card}(\mathbb{N})$ — i.e., there exists a bijective mapping $f : \mathbb{N} \longrightarrow A$. A set A is *uncountably infinite* when $\mathrm{card}(\mathbb{N}) \le \mathrm{card}(A)$ but $\mathrm{card}(\mathbb{N}) \ne \mathrm{card}(A)$. Thus when A is uncountably infinite, we can find an injective mapping from $\mathbb{N}$ into A but not a bijective mapping from $\mathbb{N}$ onto A. It turns out that the power set $\mathcal{P}(A)$ of any set A has cardinality strictly greater than the cardinality of A, from which it follows that uncountably infinite sets of arbitrarily large cardinality exist — for example, $\mathcal{P}(\mathbb{N})$, $\mathcal{P}(\mathcal{P}(\mathbb{N}))$, etc.

We've seen already that the cardinality of the power set $\mathcal{P}(A)$ of a nonempty finite set A is strictly greater than the cardinality of A. Why is the same thing true when A is infinite? Here's an argument that has a decidedly Russellian flavor. Let A be a (possibly infinite) set. Consider the mapping from A into $\mathcal{P}(A)$ defined by

$$a \mapsto \{a\}\ .$$

This is clearly an injective mapping from A into $\mathcal{P}(A)$, and we conclude that $\text{card}(A) \leq \text{card}(\mathcal{P}(A))$. So the power set of A has cardinality at least as big as the cardinality of A.

If we had $\text{card}(A) = \text{card}(\mathcal{P}(A))$, we'd be able to find a bijective mapping $f : A \longrightarrow \mathcal{P}(A)$. Suppose for the moment we have such a mapping. Define a subset $X \subset A$ as follows:

$$X = \{a \in A : a \notin f(a)\} .$$

Read this as, "X is the set of all a in A such that a is not in $f(a)$." The definition makes sense because $f(a)$ is a subset of A for every $a \in A$, so we can ask whether a is an element of $f(a)$. Since the putative mapping f is bijective, we can find some $a_o \in A$ so that $f(a_o) = X$. Question: is a_o in X? If $a_o \in X$, then, by definition of X, $a_o \notin f(a_o) = X$, so $a_o \notin X$ — a contradiction. Similarly, if $a_o \notin X$, then, by definition of X, $a_o \in f(a_o) = X$, so $a_o \in X$ — another contradiction. So we have a paradox-like situation that contradicts the existence of a bijective mapping from A onto $\mathcal{P}(A)$, which means that $\text{card}(A) \neq \text{card}(\mathcal{P}(A))$ even when A is infinite. Since $\text{card}(A) \leq \text{card}(\mathcal{P}(A))$, it follows that $\mathcal{P}(A)$ has strictly greater cardinality than A.

Enough for now about sets and cardinality. Let's assume we have a grip on $\mathbb{N}$ and talk from now on about natural numbers the way we've always talked about them. A few observations:

- The elements of $\mathbb{N}$ have a natural ordering.
- We have two natural commutative and associative algebraic operations on the elements of $\mathbb{N}$ — multiplication and addition. 0 serves as an identity element for addition, 1 serves as an identity element for multiplication, and multiplication distributes over addition.
- We have a natural notion of distance between elements of $\mathbb{N}$. If $n > m$, the distance between m and n is $n - m$.

Observe also that no element of $\mathbb{N}$ except 0 has an additive inverse in $\mathbb{N}$, and no element except 1 has a multiplicative inverse in $\mathbb{N}$. Accordingly, $\mathbb{N}$ contains a lot of numbers that we can manipulate in standard ways, but it seems to be missing some things that would be useful were they present.

The integers and rational numbers

If we throw in additive inverses for all the elements of $\mathbb{N}$ — namely, $\{-1, -2, -3, \ldots\}$ — we get the set of *integers*, for which $\mathbb{Z}$ is the standard notation. The set $\mathbb{Z}$ is countably infinite. To see this, check for yourself that the mapping $f : \mathbb{N} \longrightarrow \mathbb{Z}$ defined by

$$f(n) = \begin{cases} n/2 & \text{if } n \text{ is even} \\ -(n+1)/2 & \text{if } n \text{ is odd} \end{cases}$$

is bijective. Curiously, although $\mathbb{Z}$ contains an obvious "natural copy" of $\mathbb{N}$ and would seem *a priori* to be "twice the size of $\mathbb{N}$," it is precisely "the same size" as $\mathbb{N}$ from the standpoint of cardinality. That kind of thing happens a lot when you're dealing with infinite sets.

The bijection from $\mathbb{N}$ onto $\mathbb{Z}$ in the preceding paragraph suggests how one might "define the integers" in a set-theoretic fashion. To wit: 0 is ϕ; 1 is $\{\{\phi\}\}$; -1 is $\{\phi\}$; 2 is $\{\{\{\{\phi\}\}\}\}$; -2 is $\{\{\{\phi\}\}\}$; etc. Of course, this way of "defining" the integers would require that we modify our earlier set-theoretic "definition" of the natural numbers. I'd rather not belabor this sort of thing, but I'd like you to ponder it while keeping in mind the "Everything in the universe is a set" doctrine I alluded to earlier.

Some observations about the integers:

- The elements of $\mathbb{Z}$ have a natural ordering.
- We have two natural commutative and associative algebraic operations on the elements of $\mathbb{Z}$ — multiplication and addition. 0 serves as an identity element for addition, 1 serves as an identity element for multiplication, and multiplication distributes over addition.
- We have a natural notion of distance between elements of $\mathbb{Z}$. The distance between any m and n in $\mathbb{Z}$ is $|n - m|$.
- Every element of $\mathbb{Z}$ has an additive inverse.

Observe that no element of $\mathbb{Z}$ save 1 and -1 has a multiplicative inverse. If we throw a multiplicative inverse for every element of $\mathbb{Z}$, we get some of the rational numbers. We get all the rational numbers if we throw in a multiplicative inverse for every integer and also make sure to throw in enough other things so the set we construct is closed under addition and multiplication.

The set $\mathbb{Q}$ of *rational numbers* contains all ratios m/n of integers m and n where $n \neq 0$. But wait, there's some ambiguity here. What if $m_1/n_1 = m_2/n_2$? Furthermore, what do we mean by a "ratio of integers," anyway? This last question prompts us yet again to look admiringly, or perhaps with annoyance, upon the "Everything in the universe is a set" doctrine.

Let's define $\mathbb{Z}_0$ as the set of all nonzero integers. Consider the set

$$\widetilde{\mathbb{Q}} = \mathbb{Z} \times \mathbb{Z}_0 = \{(m, n) : m, n \in \mathbb{Z} \text{ and } n \neq 0\} .$$

For any integers m_o and n_o with $n_o \neq 0$, define the ratio of m_o to n_o, i.e. the rational number m_o/n_o, as the following subset of $\widetilde{\mathbb{Q}}$:

$$\{(m, n) : mn_o = nm_o\} .$$

In this way, we think of a rational number as a *set* of pairs of integers. Two pairs of integers lie in the same such set precisely when their "ratios" (in the middle-school sense) are the same. Mathematicians refer to these sets as *equivalence classes* of pairs of integers, where two pairs are equivalent when the products of each pair's first element with the other pair's second element are the same. The bottom line is that, if you want, you can think of each rational number as a set — a particular set of integer pairs. In this fashion, the rational numbers have an embodiment as a subset of $\mathcal{P}(\widetilde{\mathbb{Q}})$, the power set of $\widetilde{\mathbb{Q}}$. If you think that's heavy, just wait until we talk about the reals.

It's probably a good idea at this point to lighten up on the ontological abstraction and agree to talk from now on about rational numbers the way we've always talked about them. Some observations about the rational numbers:

- The elements of $\mathbb{Q}$ have a natural ordering.
- We have two natural commutative and associative algebraic operations on the elements of $\mathbb{Q}$ — multiplication and addition. 0 serves as an identity element for addition, 1 serves as an identity element for multiplication, and multiplication distributes over addition.
- We have a natural notion of distance between elements of $\mathbb{Q}$. If $q_1 = m_1/n_1$ and $q_2 = m_2/n_2$ are rational numbers, the distance between q_1

and q_2 is the rational number

$$|q_1 - q_2| = \left| \frac{m_1}{n_1} - \frac{m_2}{n_2} \right| = \left| \frac{m_1 n_2 - m_2 n_1}{n_1 n_2} \right| \ .$$

- Every element of $\mathbb{Q}$ has an additive inverse, and every nonzero element of $\mathbb{Q}$ has a multiplicative inverse.

Although we were able to "define" $\mathbb{Q}$ abstractly as a set of subsets of pairs of integers, it makes life easier if we think of $\mathbb{Q}$ as containing $\mathbb{N}$ and $\mathbb{Z}$ as subsets in the usual way. From this we conclude that $\text{card}(\mathbb{N}) \leq \text{card}(\mathbb{Q})$ — i.e., the set of rational numbers is at least countably infinite. One might wonder, since infinitely many rational numbers lie between any two integers, whether $\mathbb{Q}$ is uncountable.

In fact, $\mathbb{Q}$ is countable. I'll demonstrate this fact by first constructing an injective mapping from $\mathbb{Q}$ into $\mathbb{Z} \times \mathbb{Z}$ and then showing that $\text{card}(\mathbb{Z} \times \mathbb{Z}) = \text{card}(\mathbb{N})$. Begin by expressing every rational number $q = m/n$ in lowest terms with the negative sign in the numerator when $q < 0$. Set $f(m/n) = (m, n)$ for all such $m/n \in \mathbb{Q}$. The mapping f is clearly injective, proving that $\text{card}(\mathbb{Q}) \leq \text{card}(\mathbb{Z} \times \mathbb{Z})$. Next observe that $\text{card}(\mathbb{Z} \times \mathbb{Z}) = \text{card}(\mathbb{N} \times \mathbb{N})$ because the mapping g that "doubles up" the bijection we discovered earlier between $\mathbb{N}$ and $\mathbb{Z}$ is bijective. Specifically,

$$g(m,n) = \begin{cases} (m/2, n/2) & \text{if } m \text{ and } n \text{ are both even} \\ (-(m+1)/2, n/2) & \text{if } m \text{ is odd and n is even} \\ (m/2, -(n+1)/2) & \text{if } m \text{ is even and n is odd} \\ (-(m+1)/2, -(n+1)/2) & \text{if } m \text{ and } n \text{ are both odd.} \end{cases}$$

So far we have

$$\text{card}(\mathbb{Q}) \leq \text{card}(\mathbb{Z} \times \mathbb{Z}) = \text{card}(\mathbb{N} \times \mathbb{N}) \ .$$

Finally, construct a bijective mapping $h : \mathbb{N} \to \mathbb{N} \times \mathbb{N}$ by "threading" the natural numbers through the rectangular grid of points (m, n) of natural-number pairs as in Figure 1. Thus $\text{card}(\mathbb{N} \times \mathbb{N}) = \text{card}(\mathbb{N})$, and

$$\text{card}(\mathbb{Q}) \leq \text{card}(\mathbb{Z} \times \mathbb{Z}) = \text{card}(\mathbb{N} \times \mathbb{N}) = \text{card}(\mathbb{N}) \ ,$$

completing the argument that $\mathbb{Q}$ is countably infinite.

Convergent sequences, Cauchy sequences, and the real numbers

To understand why we "need" the real numbers, we have to figure out what's "missing" from the rational numbers. Building up from $\mathbb{N}$ to $\mathbb{Z}$ we supplied additive inverses. To get from $\mathbb{Z}$ to $\mathbb{Q}$ we supplied multiplicative inverses, among other things. What about getting the reals from the rationals?

If A is any set, a *sequence* of elements of A is an ordered list of elements of A indexed by $\mathbb{N}$. We use the notation $\{a_n\}$ to denote such a sequence. So

$$\{a_n\} = a_0, a_1, a_2, a_3, \dots \ .$$

The a_n need not all be different; in fact, if A is finite, they can't be. It's even possible for all the a_n to be the same. We'll be addressing real and complex sequences formally in Chapter 3, but for now let's focus on sequences $\{q_n\}$ of rational numbers. Here are a few examples. In each case, I give a specification of the nth number in a sequence $\{q_n\}$.

- $q_n = \frac{1}{n^{17}+3}$

- $q_n = \frac{n^2}{2n^2+1}$
- $q_n = (-3)^n$
- $q_n = (-1)^n$

Consider what happens in these sequences as n gets larger and larger. In the first sequence, q_n gets smaller and smaller (i.e. closer to 0) as n increases. In the second sequence, q_n gets closer and closer to $1/2$ as n increases. That's intuitively clear, but you can also see it from

$$\left| q_n - \frac{1}{2} \right| = \frac{1}{4n^2+2} ,$$

which obviously gets ever closer to zero as n increases. The terms in the third sequence alternate in sign and grow larger and larger in absolute value as n increases. The fourth sequence just alternates between $+1$ and -1.

The first two sequences above seem to be "going somewhere" as n increases, whereas the last two don't. We say that a sequence $\{q_n\}$ of rational numbers *converges* to $\bar{q} \in \mathbb{Q}$ when the distance between q_n and $\bar{q}$ approaches zero as $n \to \infty$, in which case we write

$$\lim_{n\to\infty} q_n = \bar{q} .$$

Here's a precise mathematical definition of convergence: for every integer $K > 0$ there exists an integer $N > 0$ such that $|q_n - \bar{q}| < 1/K$ for every $n > N$. (The $1/K$ plays the role of "ϵ" in the usual definition of convergence. I'm being picky and perhaps pedantic by using $1/K$ instead of ϵ because, as yet, we haven't even constructed the real numbers.)

A sequence $\{q_n\}$ of rational numbers is called a *Cauchy sequence* when the terms in the sequence get closer and closer together as n increases. Specifically, $\{q_n\}$ is a Cauchy sequence when for every integer $K > 0$ there exists an integer $N > 0$ such that $|q_m - q_n| < 1/K$ whenever m and n are both bigger than N. Note that every convergent sequence is necessarily a Cauchy sequence. To see why, suppose $\{q_n\}$ converges to $\bar{q}$. This implies that for any $K > 0$ we can find $N > 0$ so that $|q_n - \bar{q}| < 1/2K$ when $n > N$. For this choice of N, if m and n are both bigger than N, we have

$$|q_m - q_n| \leq |q_m - \bar{q}| + |\bar{q} - q_n| < \frac{1}{K} .$$

Since K is arbitrary, we've shown that $\{q_n\}$ is a Cauchy sequence.

Unfortunately, not every Cauchy sequence of rational numbers converges to a rational limit. It feels as if something were missing from $\mathbb{Q}$. When the terms in $\{q_n\}$ get closer and closer together, squishing in on each other as n gets larger, one would hope that the sequence would be homing in on something. Here's a Cauchy sequence of rational numbers that doesn't converge to a rational limit (um ... it's supposed to be pieces of the standard decimal expression for π):

$$3,\ 3.1,\ 3.14,\ 3.141,\ 3.1415,\ 3.14159,\ 3.141592,\ 3.1415926,\ \ldots$$

Another Cauchy sequence $\{q_n\}$ of rational numbers that has no rational limit has nth term

$$q_n = \left(1 + \frac{1}{n}\right)^n .$$

This last sequence converges to e. You'll just have to believe me when I tell you that π and e aren't rational.

Accordingly, we need to augment the rational numbers if we want a set of numbers in which every Cauchy sequence has a limit. The set we come up with is the *real numbers*, for which we use the notation $\mathbb{R}$. Roughly speaking, $\mathbb{R}$ is just "the set of all limits of Cauchy sequences of rational numbers." However, just as when we defined the rationals as "the set of all ratios of integers," we really need to be careful because many different Cauchy sequences can share the same limit.

Here's one possible abstract construction of the real numbers from sequences of rationals. First define $\widetilde{\mathbb{R}}$ as the set of all Cauchy sequences of rational numbers. So, each element of $\widetilde{\mathbb{R}}$ is an infinite sequence of rationals. As if that weren't complicated enough, divide $\widetilde{\mathbb{R}}$ up into subsets as follows: sequences $\{q_n\}$ and $\{r_n\}$ are in the same subset of $\widetilde{\mathbb{R}}$ whenever $\lim_{n\to\infty} |q_n - r_n| = 0$. (Intuitively, sequences $\{q_n\}$ and $\{r_n\}$ lie in the same subset precisely when they close in on each other as $n \to \infty$ — meaning they appear to be "headed for the same limit." Technically, these subsets are equivalence classes of Cauchy sequences, where we regard two sequences as equivalent if they close in on each other.) Each subset we construct in this way "is," in some sense, a real number. That number is the limit toward which all the sequences $\{q_n\}$ in the subset appear to be headed. Again, we're back to "Everything in the universe is a set." A rational number is a set of integer pairs. A real number is a set of Cauchy sequences of rational numbers. And so on. Sigh.

All right, it's time once again to let some air out of the balloon and float back down toward earth. I mean, we all have a gut feeling for the real numbers, right? To manipulate them, we really don't need to know "what they are." In truth, the "definition" I've given for $\mathbb{R}$ is only one of many possible, and it doesn't reflect how the real numbers emerged historically. What lies ahead will go more smoothly if we think of the set of rational numbers (and the set of natural numbers, and the set of integers) as subsets of the set of real numbers in the usual way. So let's do that from now on.

Some observations about the real numbers:

- The elements of $\mathbb{R}$ have a natural ordering.
- We have two natural commutative operations on the elements of $\mathbb{R}$ — multiplication and addition. 0 serves as an identity element for addition, 1 serves as an identity element for multiplication, and multiplication distributes over addition.
- We have a natural notion of distance between elements of $\mathbb{R}$. The distance between any a and b in $\mathbb{R}$ is $|a - b|$.
- Every element of $\mathbb{R}$ has an additive inverse, and every nonzero element of $\mathbb{R}$ has a multiplicative inverse.
- Every Cauchy sequence of real numbers has a real-number limit.

The first three observations hold for $\mathbb{Q}$. The last one is new. In technical terms, it states that the real numbers constitute a *complete metric space.*

A *metric space* is just a set A with a distance function that satisfies these conditions:

- For every $a \in A$ and $b \in A$, the distance between a and b is a nonnegative real number, and the distance between a and b is zero if and only if $a = b$.
- For every $a \in A$, $b \in A$, and $c \in A$, the distance between a and c is less than or equal to the sum of the distance between a and b and the distance between b and c. People call this the *triangle inequality.*

In a metric space A with appropriate distance function, we can talk about Cauchy sequences $\{a_n\}$ of elements of A the same way we used standard notions of distance to talk about Cauchy sequences of real or rational numbers. Cauchy sequences in A are sequences $\{a_n\}$ whose terms get closer and closer together as n increases, where "closer" is with respect to the distance function on A. A metric space A is *complete* when every Cauchy sequence $\{a_n\}$ from A has a limit $\bar{a} \in A$. The real numbers turn out to be complete in this way, essentially by construction.

It's useful to know that every Cauchy sequence of real numbers has a limit, but it's not the kind of thing you need to remember how to prove. In fact, you can't really prove it from first principles because any way of building the real numbers from scratch, including the way I outlined in the foregoing, rigs things so that the set you get is automatically complete. It's fair to say that every Cauchy sequence of real numbers has a limit *by definition.* I'll have more to say about this in Chapter 3.

What about the cardinality of $\mathbb{R}$? As it happens, $\mathbb{R}$ is uncountably infinite. I'll present Cantor's classic proof of that fact in due course.

The complex numbers

So far, we've talked about $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$. What about the complex numbers? Where did they come from, and why would anyone want to go there? Recall how $\mathbb{Z}$ augments $\mathbb{N}$ with additive inverses, $\mathbb{Q}$ augments $\mathbb{Z}$ with multiplicative inverses, and $\mathbb{R}$ augments $\mathbb{Q}$ with limits for all the Cauchy sequences. The real numbers are indeed complete (in a technical sense, as a metric space). What's missing?

A *polynomial over* $\mathbb{R}$ *of degree* n is an expression of the form

$$a_0x^n + a_1x^{n-1} + a_2x^{n-2} + \cdots + a_{n-1}x + a_n \ ,$$

where a_i, $0 \leq i \leq n$, are real numbers with $a_0 \neq 0$ and x is a "variable." We can think of a polynomial simply as a formal expression involving some numbers and the letter x, or we can think of it as defining a mapping $f : \mathbb{R} \longrightarrow \mathbb{R}$ by means of

$$f(x) = a_0x^n + a_1x^{n-1} + a_2x^{n-2} + \cdots + a_{n-1}x + a_n \ .$$

A *root* of such a polynomial is a value of x that maps to zero under f.

The polynomial $x^2 - 3x + 2$ has roots 1 and 2; we can factor the polynomial as $(x-1)(x-2)$. The polynomial $x^2 + 10x + 25$ has only -5 as a root, but -5 has *multiplicity* 2 as a root because we can factor the polynomial as $(x+5)^2$. The polynomial $x^2 - 11$ has roots $\sqrt{11}$ and $-\sqrt{11}$; we can factor it as $(x-\sqrt{11})(x+\sqrt{11})$. The polynomial $x^2 + 13$ has no real roots. To see this, note that $x^2 + 13 \geq 13$ for every $x \in \mathbb{R}$, so we never have $x^2 + 13 = 0$.

If you take the real numbers and throw in enough extra things so that every polynomial has at least one root, you get the *complex numbers*, the standard notation for which is $\mathbb{C}$. Defining j as $\sqrt{-1}$ — i.e., j and $-j$ are the roots of $x^2 + 1$ — allows one to express any complex number z in the form

$$z = a + jb$$

where a and b are real numbers known respectively as the real and imaginary parts of z. You all know how to manipulate complex numbers once you've represented them this way. Set-theoretically, you can think of the complex numbers as "the same" as $\mathbb{R} \times \mathbb{R}$ with a special way of multiplying two pairs of real numbers to get another such pair, namely

$$(a_1, b_1) \times (a_2, b_2) = (a_1 a_2 - b_1 b_2, a_1 b_2 + b_1 a_2) \ .$$

I'll demonstrate shortly that the cardinality of $\mathbb{R} \times \mathbb{R}$, and hence the cardinality of $\mathbb{C}$, is the same as the cardinality of $\mathbb{R}$. In particular, both $\mathbb{R}$ and $\mathbb{C}$ are uncountably infinite.

The Fundamental Theorem of Algebra states that every polynomial over $\mathbb{C}$ of degree n, i.e. any expression of the form

$$a_0 z^n + a_1 z^{n-1} + a_2 z^{n-2} + \cdots + a_{n-1} z + a_n \ ,$$

where $a_0 \neq 0$ and $a_i \in \mathbb{C}$, $0 \leq i \leq n$, has exactly n roots counting multiplicities. More precisely, we can factor any such polynomial as

$$a_0 (z - z_1)(z - z_2) \cdots (z - z_n)$$

where the roots z_i are not necessarily all different. The number of times any given root appears in this factorization is called its *multiplicity* as a root of the polynomial. As it happens, if a polynomial over $\mathbb{C}$ has real coefficients, its roots come in complex-conjugate pairs in the sense that if $z_o = a_o + jb_o$ is a root, then so is $\bar{z}_o = a_o - jb_o$, the complex conjugate of z_o.

So now we have $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$. I've tried to indicate how mathematicians over the years have sought to define and construct these sets of numbers. I do, however, want to reassure you that as long as you know how to manipulate the numbers and understand their important properties as I've outlined them here, you'll be okay just thinking of them in the usual way. What I mean is that, as far as we're concerned, it's fine to think of these sets as standing in the relation

$$\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C} \ .$$

When you contemplate this chain of inclusions, try to keep in mind what you gain each time you move one step to the right in the chain.

Decimal expansions of real numbers

My main goal in the remainder of this chapter is to demonstrate that $\mathbb{R}$, the set of real numbers, is uncountably infinite. I'll employ a classic proof technique known as *Cantor's diagonal argument.* Invoking Cantor's argument requires some other results about the real numbers that are interesting and useful in their own right. In particular, decimal expansions of real numbers play a critical role.

Before discussing decimal expansions, I'd like to remind you about an extraordinarily handy piece of machinery that you've learned about before. It's the *geometric series.* We'll talk more formally about series in Chapter 3, but for now consider this expression involving a real or complex number γ:

$$\sum_{n=0}^{\infty} \gamma^n \ .$$

The official meaning of the series expression is

$$\lim_{N \to \infty} \sum_{n=0}^{N-1} \gamma^n \ .$$

Of course, the limit's existence is not guaranteed. Observe, however, that for any $N > 0$ we have

$$(1-\gamma)\sum_{n=0}^{N-1}\gamma^n = 1-\gamma^N \ ,$$

so that when $\gamma \neq 1$

$$\sum_{n=0}^{N-1}\gamma^n = \frac{1-\gamma^N}{1-\gamma} \ .$$

If $|\gamma| < 1$, the right-hand side converges as $N \to \infty$ to $1/(1-\gamma)$, and hence so does the left-hand side. The bottom line is that

$$\sum_{n=0}^{\infty}\gamma^n = \frac{1}{1-\gamma} \quad \text{if} \quad |\gamma| < 1 \ .$$

Now let's talk about decimal expansions of real numbers. We'll focus on the open unit interval

$$(0,1) = \{x \in \mathbb{R} : 0 < x < 1\} \ .$$

It turns out that any $x \in (0,1)$ has at least one decimal expansion

$$x = .a_1a_2a_3a_4a_5\dots \ ,$$

where each a_n is a natural number between 0 and 9, inclusive. The expansion means

$$x = \sum_{n=1}^{\infty} a_n 10^{-n} = \lim_{N\to\infty}\sum_{n=1}^{N} a_n 10^{-n} \ .$$

A decimal expansion of x is one way of representing x as the limit of a sequence of rational numbers. If you define

$$q_N = \sum_{n=1}^{N} a_n 10^{-n} \ ,$$

then each q_N is rational and the sequence $\{q_N\}$ is a Cauchy sequence (check this for yourself) that converges to x as $n \to \infty$.

Some decimal expansions terminate; for those expansions, there's a smallest $M > 0$ such that $a_n = 0$ for $n > M$. It happens that if x has a terminating decimal expansion, then x has at least one other decimal expansion. To wit, suppose $x = .a_1a_2\dots a_M$ is a terminating expansion for x with $a_M \neq 0$ and $a_n = 0$ for all $n > M$. Then a non-terminating expansion for x is

$$x = .a_1a_2a_3\dots a_{M-1}(a_M-1)999999\dots \ ,$$

where the 9's go on forever. The Mth decimal place has $a_M - 1$ in it; in other words, you decrement the last nonzero decimal place in x's terminating expansion by 1, and you replace all the trailing zeroes in x's terminating expansion by 9's. As an example,

$$.183746285647 = .1837462856469999999999999999999999\dots$$

The geometric series makes this work. Re-write the second decimal expansion above as

$$a_1 10^{-1} + a_2 10^{-2} + \cdots + a_{M-1}10^{-(M-1)} + a_M 10^{-M} - 10^{-M} + \sum_{n=M+1}^{\infty} 9\times 10^{-n} \ .$$

Change the index in the last sum to $m = n - (M+1)$ and you get

$$\left(9 \times 10^{-(M+1)}\right) \sum_{m=0}^{\infty} 10^{-m} = \left(9 \times 10^{-(M+1)}\right) \frac{1}{1 - 1/10} = 10^{-M} \; .$$

So the sum of all the 9-terms in the expansion adds up to 10^{-M}, which cancels the -10^{-M} in the expansion, which makes the expansion equal to

$$a_1 10^{-1} + a_2 10^{-2} + \cdots + a_{M-1} 10^{-(M-1)} + a_M 10^{-M} \; ,$$

which is just our first (terminating) expansion for x. Thus you can expand any such x either as a terminating decimal or as one that ends in an infinite string of 9's.

In fact, a strong converse is also true, namely that if a number x between 0 and 1 has more than one decimal expansions, then

- x has exactly two expansions, and
- one expansion terminates, one expansion ends in all 9's, and the two expansions are related as above.

Here's a proof. Suppose x has two expansions

$$x = .a_1 a_2 a_3 \ldots$$

and

$$x = .b_1 b_2 b_3 \ldots$$

There will be some smallest value of $M > 0$ so that $a_M \neq b_M$; i.e., the two expansions will agree for $n < M$ (note that $M = 1$ is a possibility). Since both expansions evaluate to x, and since the expansions agree for $n < M$, we must have

$$\sum_{n=M}^{\infty} a_n 10^{-n} = \sum_{n=M}^{\infty} b_n 10^{-n} \; .$$

Suppose without loss of generality that $a_M > b_M$. Re-arrange the last equation to read

$$(a_M - b_M) 10^{-M} = \sum_{n=M+1}^{\infty} (b_n - a_n) 10^{-n} \; .$$

Since $a_M - b_M \geq 1$ and $b_n - a_n \leq 9$ for all k, we have

$$10^{-M} \leq (a_M - b_M) 10^{-M} = \sum_{n=M+1}^{\infty} (b_n - a_n) 10^{-n} \leq 9 \sum_{n=M+1}^{\infty} 10^{-n} = 10^{-M} \; ,$$

where the last equality comes from the geometric series as before. Accordingly, we have equality along the whole chain of inequalities. But equality holds in the left inequality only when $b_M = a_M - 1$, and equality holds in the right inequality only when $b_n - a_n = 9$ for all $n > M$, which can happen only when $a_n = 0$ and $b_n = 9$ for all $n > M$. Conclusion: the a-expansion terminates and the b-expansion ends in all 9's, and the two expansions are related as stipulated. The two bulleted assertions follow immediately since it's clear that no x has two different terminating expansions.

One remaining loose end: I haven't supplied a formal proof of the existence of at least one decimal expansion for each $x \in (0,1)$. Here's one argument. To find

a_1, all you have to do is shift the decimal place to the right by one digit and take the lower integer part of the result. Formally,

$$a_1 = \lfloor 10x \rfloor \ .$$

To find a_2, do this:

$$a_2 = \lfloor 100(x - a_1 10^{-1}) \rfloor \ .$$

More generally,

$$a_n = \lfloor 10^n (x - \sum_{m=1}^{n-1} a_m 10^{-m}) \rfloor \ .$$

This procedure gives you the digits from one decimal expansion for x. If x has two decimal expansions, the procedure yields the terminating one.

Cantor's diagonal argument

The ultimate goal is to show that $\mathbb{R}$ is uncountably infinite. Observe first that the mapping

$$x \mapsto \tan \pi(x - 1/2) \ \text{ for } \ x \in (0, 1)$$

establishes a bijection between the interval $(0, 1)$ and $\mathbb{R}$, so $(0, 1)$ and $\mathbb{R}$ have the same cardinality and it suffices to show that $(0, 1)$ is uncountably infinite.

Now let's see why $(0, 1)$ is uncountably infinite. I'll demonstrate that $(0, 1)$ contains an uncountably infinite subset C, which will imply in turn that $(0, 1)$ is itself uncountable because its cardinality is at least as large as C's. Let C be the set of values of $x \in (0, 1)$ whose decimal expansions contain no 9's in them. Each $x \in C$ has exactly one decimal expansion. If C were countable, we could define a bijective mapping $f : \mathbb{N} \to C$. From this mapping we could make an exhaustive list of C's elements, like so:

$$\begin{array}{rcl} f(0) & = & .138254736527\ldots \\ f(1) & = & .385643772314\ldots \\ f(2) & = & .632564026533\ldots \\ f(3) & = & .382655341628\ldots \\ & \cdots & \end{array}$$

Now for Cantor's sleight of hand. It turns out that *no such listing can exhaust* all the decimal expansions of x-values lying in C. How so? Construct a decimal $\widehat{x}$ as follows: for each n, generate the number that appears in the nth decimal place of $\widehat{x}$ using this recipe: take the number in the nth decimal place of $f(n)$ on the list above, add 1 to it, and mod out by 9 (example, $8 + 1 \bmod 9$ comes out to be 0). So, using the list above, $\widehat{x}$ starts out like this:

$$\widehat{x} = .2037\ldots$$

The decimal $\widehat{x}$, while a perfectly valid 9-free decimal, is not on the list. If it were on the list, it would appear as $f(n_o)$ for some $n_o \in \mathbb{N}$. By construction, the number in the n_oth decimal place of $\widehat{x}$ differs from the number in the n_oth decimal place of $f(n_o)$, so $\widehat{x}$ can't be $f(n_o)$. The moral of the story is that the putative bijection f that enabled us to construct the list cannot exist, and C is uncountably infinite. Hence, so is $(0, 1)$ and so is $\mathbb{R}$.

C is uncountable, but how big is it?

What kind of subset of the unit interval is C? By that question, I mean, is there a nice way to describe it geometrically? How "long" is it, or at least what's the total length of all the "little pieces" that make it up? The answer is astounding, at least to me. Certainly, C contains none of the numbers between .9 and 1 because all those numbers' decimal expansions lead off with a 9. So, C is contained in $(0, .9)$. Since numbers in C can't have a 9 in the second decimal place, C can't include the intervals between .09 and .1, between .19 and .2, between .29 and .3, ... , between .89 and .9. There are nine such intervals and each has length 10^{-2}. You can keep this up. You discover that eliminating numbers with a 9 in the third decimal place requires throwing away 81 intervals of length 10^{-3}, etc., etc.

All told, to pare down the unit interval to C, you throw out a whole bunch of little intervals the sum of whose lengths is

$$\frac{1}{10} + \frac{9}{100} + \frac{81}{1000} + \frac{9^3}{10^4} + \cdots$$

What do all those terms add up to? If you factor out 1/10, you find that it's a geometric series, and that the sum comes out to be

$$\frac{1}{10}\sum_{n=0}^{\infty}(9/10)^n = \left(\frac{1}{10}\right)\left(\frac{1}{1-9/10}\right) = 1\ .$$

In other words, to get down to C, you eliminate from $(0, 1)$ a total length equal to the length of the entire interval, which is 1. People say for this reason that C is "a set of measure zero." On the other hand, C is a huge set, right? It's uncountably infinite after all.

It helps to think about it probabilistically. Suppose we build a real number by choosing its decimal digits one by one at random. If we draw each of 0 through 9 with probability 1/10, what's the probability that we'll *never* draw a 9? It's basically $(9/10)^\infty$, i.e. zero. So the numbers with no 9's in their expansions, while they constitute the uncountable set C, are rare indeed. Ah, the joy of infinity.

The sets $\mathbb{R}$ and $\mathbb{C}$ have the same cardinality

We noted earlier that $\mathbb{C}$ has the same cardinality of $\mathbb{R} \times \mathbb{R}$, and I promised you a proof that $\text{card}(\mathbb{R} \times \mathbb{R})$ is the same as $\text{card}(\mathbb{R})$. Here's that proof.

First of all, since $(0, 1)$ has the same cardinality as $\mathbb{R}$, $(0, 1) \times (0, 1)$ has the same cardinality as $\mathbb{R} \times \mathbb{R}$. For each $(x, y) \in (0, 1) \times (0, 1)$, pick a decimal expansion for x and one for y. If either x or y has two expansions, pick the terminating one(s). Now interleave the digits in the two expansions to get $z \in (0, 1)$. For example,

$$(x, y) = (.382137\ldots, .113535\ldots) \mapsto z = .318123153375\ldots\ .$$

If you think about it, you'll see that this defines a bijective mapping from $(0, 1) \times (0, 1)$ onto $(0, 1)$. Thus, those two sets have the same cardinality, and therefore so do $\mathbb{R} \times \mathbb{R}$ and $\mathbb{R}$. It follows that $\mathbb{C}$ has the same cardinality as $\mathbb{R}$. It's worth noting that by interleaving decimal expansions one can show that

$$\text{card}(\mathbb{R}^n) = \text{card}(\mathbb{R})$$

for any $n > 0$.

The power of the continuum

The cardinality of $\mathbb{R}$ is known colorfully as the *power of the continuum.* Where that cardinality lies in the so-called hierarchy of infinite cardinals is a mystery. A famous open conjecture regarding card($\mathbb{R}$) is the *Continuum Hypothesis,* a discussion of which would take us too far afield. For now, we'll have to satisfy ourselves with a neat characterization of card($\mathbb{R}$), namely

$$\text{card}(\mathbb{R}) = \text{card}(\mathcal{P}(\mathbb{N})) \ ;$$

i.e., the cardinality of the reals is the same as the cardinality of the power set of the natural numbers. I'll try to justify that statement here.

Recall that the power set of an N-element set A has cardinality 2^N. The key was a bijective mapping from $\mathcal{P}(A)$ onto B_N, the set of all binary strings of length N. We can take a similar approach to investigating the cardinality of $\mathcal{P}(\mathbb{N})$.

Let B_∞ be the set of all (one-sided) infinite binary strings

$$b_1 b_2 b_3 b_4 b_5 \dots \ ,$$

where each b_n is 0 or 1. Define a mapping $\beta : \mathcal{P}(\mathbb{N}) \to B_\infty$ by setting, for all $S \subset \mathbb{N}$,

$$[\beta(S)]_n = \begin{cases} 1 & \text{if } n \in S \\ 0 & \text{if } n \notin S \ . \end{cases}$$

Thus, for any subset S of $\mathbb{N}$, $\beta(S)$ is a binary string that has 1's precisely in those positions corresponding to natural numbers that are in S. Note that the empty set ϕ maps under β to the string of all 0's and the set $\mathbb{N}$ maps to the string of all 1's. It is clear that β is a bijective mapping. Accordingly, $\mathcal{P}(\mathbb{N})$ has the same cardinality as B_∞.

Meanwhile, every $x \in (0, 1)$ has at least one *binary expansion*, which is an expression of the form

$$x = .b_1 b_2 b_3 b_4 \dots$$

where each b_n is 0 or 1. The meaning of the expansion is

$$x = \sum_{n=1}^{\infty} b_n 2^{-n} \ .$$

The theory of binary expansions (existence and not-quite uniqueness) runs parallel to that of decimal expansions. To summarize:

- Every $x \in (0, 1)$ has either one binary expansion or two.
- $x \in (0, 1)$ has two binary expansions if and only if x has a terminating binary expansion. In this case, the non-terminating expansion for x ends in an infinite string of 1's. Furthermore, if the last 1 in the terminating expansion for x occurs at position M, you can find the non-terminating expansion for x by changing that 1 to a zero and appending the infinite string of 1's. For example,

$$.1011101 = .10111001111111111111 \dots \ .$$

- You can construct a binary expansion for any $x \in (0,1)$ inductively (i.e. starting from $n = 1$ and working rightward) via

$$b_n = \left\lfloor 2^n (x - \sum_{m=1}^{n-1} b_m 2^{-m}) \right\rfloor \ , \ n \geq 1.$$

The upshot is that the interval $(0,1)$ is in one-to-one correspondence with the set of all strings in B_∞ that don't end in an infinite sequence of 1's. Technically, this means that there exists an injective mapping from $(0,1)$ into B_∞. Since B_∞ is in one-to-one correspondence with $\mathcal{P}(\mathbb{N})$, we conclude that there exists an injective mapping

$$f : (0,1) \longrightarrow \mathcal{P}(\mathbb{N}) \ .$$

As a consequence, $\text{card}((0,1)) \leq \text{card}(\mathcal{P}(\mathbb{N}))$.

It turns out that these cardinalities are actually equal. One clever argument (which I'm stealing from George Simmons's book *Introduction to Topology and Modern Analysis*) goes as follows. Map each binary string in B_∞ to an infinite sequence of 3's and 5's by replacing every 0 with a 3 and every 1 with a 5. (Nothing is special about 3 and 5, but 9 is forbidden.) Then think of that string of 3's and 5's as the decimal expansion of some $x \in (0,1)$. So, for example,

$$10011011000\ldots \mapsto .53355355333\ldots \in (0,1) \ .$$

This correspondence defines a mapping

$$g : \mathcal{P}(\mathbb{N}) \longrightarrow (0,1)$$

because $\mathcal{P}(\mathbb{N})$ and B_∞ are in one-to-one correspondence. The mapping g is injective because any x expandable decimally in a string of 3's and 5's has only one such expansion.

Accordingly, we have injective mappings going both ways between $\mathcal{P}(\mathbb{N})$ and the interval $(0,1)$. It follows that $\text{card}((0,1)) \leq \text{card}(\mathcal{P}(\mathbb{N}))$ and $\text{card}(\mathcal{P}(\mathbb{N})) \leq \text{card}((0,1))$, so the two sets have the same cardinality. Since $\text{card}(\mathbb{R}) = \text{card}(0,1)$, we conclude that

$$\text{card}(\mathbb{R}) = \text{card}(\mathcal{P}(\mathbb{N})) \ .$$

CHAPTER 2

Working with Integers: Prime Numbers and Modular Arithmetic

Understanding the integers is important for both theoretical and practical reasons. Integer-based analogues underpin many seemingly arcane constructions in abstract algebra (the study of groups, rings, fields, and all that). Modular arithmetic and integer factorization feature prominently in combinatorics and cryptography. Investigating the integers can also be fun irrespective of potential application or lack thereof. This chapter covers mainly elementary material, but I hope it works as a starting point. As you'll see, getting to the heart of nontrivial applications requires surprisingly little number theory.

Prime numbers: the basics

Given two natural numbers a and b, we say that *b is a divisor of a* when $a = mb$ for some natural number m. The standard notation for "b is a divisor of a" is $b|a$. Often we just say "b divides a" for short. Observe that every $b \in \mathbb{N}$ divides 0 and that 1 divides every $a \in \mathbb{N}$. Note also that if $b|a$ and $c|b$, then $c|a$. A natural number p is *prime* when the only natural-number divisors of p are 1 and p itself. By convention, 1 is not a prime number even though it satisfies the technical definition. The first few prime numbers are 2, 3, 5, 7, 11, and 13. Note that 2 is the only even prime number. Can you see why?

In what follows, I'll be using *induction* a fair amount to prove things. It takes some practice to get the hang of using inductive arguments, but the effort pays off. Induction is, among other things, a versatile tool for proving facts about the natural numbers. Here's an example of a typical inductive argument. I'll prove that every $a \in \mathbb{N}$, $a > 1$, has at least one prime divisor. You start with

- the base case $a = 2$: clearly, 2 has a prime divisor (2 itself).

Then you move on to

- the induction step: suppose we have shown that every $a \leq n$ has at least one prime divisor. Consider $a = n + 1$. If $n + 1$ is prime, we're done, since $n + 1$ is then a prime divisor of itself. If $n + 1$ is not prime, then we can write

  $$n + 1 = bc$$

  for some natural numbers b and c with $1 < b, c \leq n$. But since we've shown already that every such b and c has at least one prime divisor, and since b and c here are both divisors of $n + 1$, we conclude that $n + 1$ must have at least one prime divisor.

I hope you see how the induction works. We know that the theorem is true for $n = 2$ by the base case. What about $n = 3$? It's true for $n = 2$, and the induction step shows that if it's true for $n = 2$, then it's true for $n = 3$; hence it's also true for $n = 3$. What about $n = 4$? We know now that it's true for $n = 2$ and $n = 3$, and the induction step enables us to conclude that it's also true for $n = 4$. And so the dominoes fall.

Euclid used the result we just proved to demonstrate that there are infinitely many prime numbers. His argument proceeds as follows. Suppose that we have a list of K primes. Index them as p_1, p_2, $\dots$, p_K. Consider the number

$$R = 1 + p_1 p_2 p_3 \cdots p_K \ .$$

Our "theorem" above guarantees that R has at least one prime divisor p, and p could not possibly be among the p_j on our list. If it were on the list, then p would divide both R and $p_1 p_2 \cdots p_K$, implying that

$$p | (R - p_1 p_2 p_3 \cdots p_K) \text{ i.e. } p | 1 \ ,$$

which is impossible. So p is not on our list. In particular, our list doesn't contain every prime. More trenchantly, nothing about our list is special, so no finite list of primes can be exhaustive. In other words, infinitely many primes exist.

Two positive natural numbers a and b are said to be *relatively prime* or *coprime* when they have no common divisors except 1. One of the workhorses of number theory is the following result.

2.1 Theorem: If a and b are positive natural numbers and are coprime, then there exist integers m and n (note: negative integers allowed) such that $ma+nb = 1$.

Proof: Define a set I of integers as follows:

$$I = \{k \in \mathbb{Z} : k = ma + nb \text{ for some } m, n \in \mathbb{Z}\} \ .$$

I obviously contains some positive elements. Let $d \in \mathbb{N}$ be the smallest positive element of I; suppose $d = m_o a + n_o b$. I'll show that $d = 1$.

First of all, since both a and b are in I, $d \leq a$ and $d \leq b$. If d is not a divisor of a, then since $d \leq a$ we get a positive remainder $r > 0$ when we divide d into a. In other words,

$$a = jd + r$$

for some positive $r \in \mathbb{N}$ with $r < d$. But this means that

$$a = j(m_o a + n_o b) + r$$

which in turn implies that

$$r = (1 - j m_o) a - j n_o b = m_1 a + n_1 b \ ,$$

so $r \in I$, as well. This is a contradiction since $r < d$ and we defined d as the smallest positive element of I. It follows that d must be a divisor of a after all. A similar argument shows that $d | b$, as well. Since a and b are coprime, this means $d = 1$. The bottom line is that $m_o a + n_o b = 1$. □

We can generalize Theorem 2.1 in several useful ways. First, a definition. The *greatest common divisor* of two natural numbers a and b is the largest natural number that's a divisor of both a and b. The standard notation for the greatest common divisor of a and b is $\gcd(a, b)$. Note that $\gcd(p, a) = 1$ when p is prime and p is not a divisor of a. Furthermore, $\gcd(a, b) = 1$ precisely when a and b are coprime.

Observe that for any positive natural numbers a and b, $\overline{a} = a/\gcd(a, b)$ and $\overline{b} = b/\gcd(a, b)$ are natural numbers and are coprime. Dividing by $\gcd(a, b)$ cancels out any common divisors a and b might have. More rigorously, suppose m divides both $\overline{a}$ and $\overline{b}$. Then $m \gcd(a, b)$ divides both a and b, so $m = 1$ because $\gcd(a, b)$ is the largest common divisor of a and b by definition. If $k > 2$ and a_1, a_2, $\ldots$, a_k are natural numbers, define $\gcd(a_1, a_2, \ldots, a_k)$ as the largest natural number that divides all of the a_j. You can show easily that $\overline{a}_1$, $\overline{a}_2$, $\ldots$, $\overline{a}_k$ have no common divisors other than 1, where

$$\overline{a}_j = \frac{a_j}{\gcd(a_1, a_2, \ldots, a_k)} \text{ for } 1 \leq j \leq k \; .$$

Now for three extensions of Theorem 2.1.

- If a and b are positive natural numbers, there exist integers m and n such that
$$ma + nb = \gcd(a, b) \; .$$
To see this, form $\overline{a}$ and $\overline{b}$ as above and apply Theorem 2.1 to them.
- If a_1, a_2, $\ldots$, a_k are positive natural numbers that have no divisors other than 1 common to all of them, then there exist integers n_1, n_2, $\ldots$, n_k such that
$$n_1 a_1 + n_2 a_2 + \cdots + n_k a_k = 1 \; .$$
To demonstrate this, just mimic the proof of Theorem 2.1.
- If a_1, a_2, $\ldots$, a_k are positive natural numbers, then there exist integers n_1, n_2, $\ldots$, n_k such that
$$n_1 a_1 + n_2 a_2 + \cdots + n_k a_k = \gcd(a_1, a_2, \ldots, a_k) \; .$$
I'll let you prove this one for yourself.

Another useful property of primes is the following.

2.2 Theorem: If p is prime and $p|ab$, where a and b are positive natural numbers, then $p|a$ or $p|b$ or both.

Proof: Suppose p is not a divisor of a. Then p and a are coprime, since p and 1 are the only divisors of p. By Theorem 2.1, we can find integers m and n so that $ma + np = 1$. Multiply this last equation by b and you get

$$m(ab) + n(p)b = b \; .$$

Since p is a divisor of both expressions in parentheses on the left-hand side, it follows that p must be a divisor of b. I've shown that if p is not a divisor of a, then p must be a divisor of b. A symmetric argument shows that if p is not a divisor of b, then p must be a divisor of a. You can conclude that p must be a divisor of either a or b or both. □

The intuition behind Theorem 2.2 is that if p is prime and p is a divisor of ab, you can't "split up" p into two factors one of which is a divisor of a and one of which is a divisor of b. A simple inductive argument leads to the following generalization of Theorem 2.2.

- If p is prime and p is a divisor of $a_1 a_2 a_3 \cdots a_k$, where a_j, $1 \leq j \leq k$, are all positive natural numbers, then p is a divisor of at least one of the a_j, $1 \leq j \leq k$.

Prime factorization

It's time now for what is arguably the most important result concerning primes.

2.3 Theorem: If a is a natural number bigger than 1, then a has a unique factorization into the product of powers of prime numbers. Specifically, you can find $L > 0$ along with distinct primes p_1, p_2, $\ldots$, p_L and positive powers m_1, m_2, $\ldots$, m_L so that

$$a = p_1^{m_1} p_2^{m_2} \cdots p_L^{m_L}$$

and, furthermore, the numbers L, p_j, $1 \leq j \leq L$, and m_j, $1 \leq j \leq L$ are determined uniquely by a.

I'll give a slightly hand-wavey argument for Theorem 2.3, but the gaps in the argument are easy to fill in. The existence part of Theorem 2.3 is "an easy induction." Here's how it goes. First the base case: clearly $a = 2$ has a factorization into the product of positive prime powers (namely, $L = 1$, $p_1 = 2$, $m_1 = 1$). Now the induction step: suppose we have shown that every $a \leq n$ has a factorization of the form that Theorem 2.3 calls for. Consider $a = n+1$. If $a = n+1$ is prime, then we're done. If not,, $a = n + 1$ factors into a product bc, where $1 < b, c \leq n$. By the induction assumption, b and c have factorizations into products of positive powers of distinct primes. Folding these factorizations together yields a factorization of $a = n+1$ into positive powers of distinct primes. One concludes, by induction, that every $a > 1$ has a factorization into a product of positive powers of distinct primes.

What about uniqueness? Suppose some $a > 1$ has factorizations

$$a = p_1^{m_1} p_2^{m_2} \cdots p_L^{m_L} = q_1^{j_1} q_2^{j_2} \cdots q_K^{j_K} \ ,$$

where all the p's are distinct primes and all the q's are distinct primes and all the m's and j's are positive. I'm not assuming anything about relationships between K, L, the p's, and the q's. Consider now p_1. Since p_1 is a divisor of a, p_1 is a divisor of the product of q-powers. By Theorem 2.2 (or at least its extension), p_1 is a divisor either of $q_1^{j_1}$ or of the product

$$q_2^{j_2} \cdots q_K^{j_K}$$

of the remaining q-powers. If p_1 is a divisor of $q_1^{j_1}$, it's easy to show by Theorem 2.2 (or its extension) that $p_1 = q_1$. If p_1 is not a divisor of $q_1^{j_1}$, then you can move one step down the line of q's and conclude that p_1 is a divisor either of $q_2^{j_2}$ or of the product

$$q_3^{j_3} \cdots q_K^{j_K} \ .$$

And so on.

I hope you can see that the end result is that p_1 must be one of the q's — and the same exact argument works for any of the other p_i, $1 \le i \le L$. You can run the argument the other way to show that any of the q's has to be one of the p's. The bottom line is that a prime appears among the p's if and only if it appears among the q's, so $L = K$, for one thing, and we can, if necessary, re-number the q's so that they match up with their companion p's, which leads to revised expressions for a:

$$a = p_1^{m_1} p_2^{m_2} \cdots p_L^{m_L} = p_1^{j_1} p_2^{j_2} \cdots p_L^{j_L} \ .$$

It remains to show that the powers appearing in the factorizations are equal — i.e., $m_1 = j_1$, $m_2 = j_2$, etc. Start with the last identity and for all k divide both by the smaller power of p_k appearing in one of the factorizations. If for some p_k we have $m_k \ne j_k$, p_k will cancel from one side and remain on the other side. For example, if we had

$$p_1^2 p_2^3 p_3 p_4^7 = p_1^5 p_2 p_3^2 p_4^7 \ ,$$

the cancellation maneuver would lead to

$$p_2^2 = p_1^3 p_3 \ .$$

But such an identity is impossible by the same reasoning (applying Theorem 2.2 and its extension) that led to the existence proof above — the same primes would have to appear on both sides of the identity in order for it to be valid. Ergo, we have a contradiction, and we conclude that all the powers have to match up, so the prime-power factorization of a is unique.

Modular integer arithmetic and Euler's Theorem

Given a positive integer a and an integer k, there exists a unique r with $0 \le r < a$ for which $k = ma + r$ for some integer m. If $k \ge 0$, r is just the remainder you get when you divide a into k. If $k < 0$, you can find r by adding a to k repeatedly until you hit a natural number between 0 and $a - 1$. We call this number r the *mod-a value of* k, or "k mod a" for short. The notation I'll be using for it is $\langle\!\langle k \rangle\!\rangle_a$.

My $\langle\!\langle k \rangle\!\rangle_a$ notation and other features of the approach I'll be taking in what follows are somewhat nonstandard, but I'd like to keep things as concrete as possible and avoid unnecessary algebraic detours. To test your understanding of the k mod a concept, make sure you see why $\langle\!\langle k \rangle\!\rangle_a = k$ when $0 \le k < a$; why $\langle\!\langle 17 \rangle\!\rangle_{11} = 6$; why $\langle\!\langle ma \rangle\!\rangle_a = 0$ for every $m \in \mathbb{Z}$; and why $\langle\!\langle -1 \rangle\!\rangle_a = a - 1$. A standard twelve-hour clock face embodies a quotidian implementation of mod-12 addition. If the clock face reads 11:00 now, then 31 hours from now it will read 6:00 because $\langle\!\langle 11 + 31 \rangle\!\rangle_{12} = \langle\!\langle 42 \rangle\!\rangle_{12} = 6$.

Taking $\langle\!\langle \quad \rangle\!\rangle_a$ of any expression involving integers is known as *modding out by a* or *reducing mod a*. Conveniently, reducing mod a treats addition and multiplication respectfully. By this I mean that for any integers k and l

- $\langle\!\langle k + l \rangle\!\rangle_a = \langle\!\langle\, \langle\!\langle k \rangle\!\rangle_a + \langle\!\langle l \rangle\!\rangle_a \,\rangle\!\rangle_a$ and
- $\langle\!\langle kl \rangle\!\rangle_a = \langle\!\langle\, \langle\!\langle k \rangle\!\rangle_a \ \langle\!\langle l \rangle\!\rangle_a \,\rangle\!\rangle_a$.

To see how these identities arise, suppose $k = m_1 a + r_1$ and $l = m_2 a + r_2$ with $0 \le r_1, r_2 < a$, so $\langle\!\langle k \rangle\!\rangle_a = r_1$ and $\langle\!\langle l \rangle\!\rangle_a = r_2$. Then

$$k + l = (m_1 + m_2)a + r_1 + r_2$$

and

$$kl = (m_1 m_2 + r_1 m_2 + m_1 r_2)a + r_1 r_2 \ .$$

Reducing mod a yields

$$\langle\!\langle k + l \rangle\!\rangle_a \ = \ \langle\!\langle r_1 + r_2 \rangle\!\rangle_a \ = \ \langle\!\langle \, \langle\!\langle k \rangle\!\rangle_a \ + \ \langle\!\langle l \rangle\!\rangle_a \, \rangle\!\rangle_a$$

and

$$\langle\!\langle kl \rangle\!\rangle_a \ = \ \langle\!\langle r_1 r_2 \rangle\!\rangle_a \ = \ \langle\!\langle \, \langle\!\langle k \rangle\!\rangle_a \ \langle\!\langle l \rangle\!\rangle_a \, \rangle\!\rangle_a \ \ .$$

More generally, when performing any integer computation that involves only multiplication and addition and ends with reduction mod a, you are free to reduce any intermediate terms mod a if you find it convenient — you'll still end up with the same result. For example, if you want to compute $\left\langle\!\left\langle k^3(l+m) \right\rangle\!\right\rangle_a$, you can first find $\langle\!\langle k \rangle\!\rangle_a$, then find $\left\langle\!\left\langle k^2 \right\rangle\!\right\rangle_a \ = \ \left\langle\!\left\langle \, \langle\!\langle k \rangle\!\rangle_a{}^2 \right\rangle\!\right\rangle_a$, then compute the final result via

$$\left\langle\!\left\langle k^3(l+m) \right\rangle\!\right\rangle_a \ = \ \left\langle\!\left\langle k \, \left\langle\!\left\langle k^2 \right\rangle\!\right\rangle_a \ \langle\!\langle l + m \rangle\!\rangle_a \right\rangle\!\right\rangle_a \ \ .$$

Given an integer $a > 1$, let $\mathbb{Z}_a = \{0, 1, 2, \ldots, a-1\}$. Thus $\mathbb{Z}_a$ is the set of all possible mod-a values of integers. For any k and l in $\mathbb{Z}_a$, $\langle\!\langle k + l \rangle\!\rangle_a$ and $\langle\!\langle kl \rangle\!\rangle_a$ are also in $\mathbb{Z}_a$. Let's define two operations on $\mathbb{Z}_a$ by

$$k \,\overline{+}\, l = \ \langle\!\langle k + l \rangle\!\rangle_a$$

and

$$k \,\overline{\times}\, l = \ \langle\!\langle kl \rangle\!\rangle_a$$

for every k and l in $\mathbb{Z}_a$. Both operations are clearly commutative and associative, and the "multiplication" operation distributes over the "addition" operation in the sense that

$$k \,\overline{\times}\, (l \,\overline{+}\, m) = (k \,\overline{\times}\, l) \ \overline{+} \ (k \,\overline{\times}\, m) \ \text{ for all } \ l \text{ and } m \in \mathbb{Z}_a \ .$$

Furthermore, 0 is an identity element for $\overline{+}$ and 1 is an identity element for $\overline{\times}$. The number 0 is clearly its own "additive inverse," and every positive $k \in \mathbb{Z}_a$ has "additive inverse" $a - k$ because $k + (a-k) = a$ so $k \,\overline{+}\, (a-k) = 0$.

The set $\mathbb{Z}_a$ endowed with the operations $\overline{+}$ and $\overline{\times}$ has the algebraic structure of a *commutative ring.* Whereas every $k \in \mathbb{Z}_a$ has an "additive inverse," it is not true in general that every $k \neq 0$ has a "multiplicative inverse." Suppose, for example, that k and a have a common divisor $d \in \mathbb{Z}_a$ with $d > 1$. Suppose $a = dq$ and $k = ld$. Then

$$k \,\overline{\times}\, q = \ \langle\!\langle ldq \rangle\!\rangle_a \ = \ \langle\!\langle la \rangle\!\rangle_a \ = 0 \ .$$

In this case, k cannot have a "multiplicative inverse" m, because then we would have

$$0 = m \,\overline{\times}\, (k \,\overline{\times}\, q) = (m \,\overline{\times}\, k) \,\overline{\times}\, q = q \ ,$$

and we know $q \neq 0$. If, on the other hand, k and a are coprime, then by Theorem 2.1 we can find integers m and n so that $mk + na = 1$. Reducing mod a yields $\langle\!\langle mk \rangle\!\rangle_a = 1$, from which it follows that $\langle\!\langle m \rangle\!\rangle_a \ \overline{\times}\, k = 1$, so k has "multiplicative inverse" $\langle\!\langle m \rangle\!\rangle_a$.

Let's dispense with the quotation marks and agree that when we talk about addition or multiplication in the context of $\mathbb{Z}_a$ we're referring to the operation $\overline{+}$ or $\overline{\times}$. We've demonstrated that $k \in \mathbb{Z}_a$ has a multiplicative inverse if and only if k and a are coprime. Denote the set of all such $k \in \mathbb{Z}_a$ by $\mathbb{Z}_a^*$. The number of elements of $\mathbb{Z}_a^*$ is known as $\phi(a)$ and the mapping $a \mapsto \phi(a)$ is called *Euler's phi function* or the *totient function.* In any event, $\mathbb{Z}_a^*$ is a subset of $\mathbb{Z}_a$ and happens to be closed under the operation of $\overline{\times}$. To see why, note that if k_1 and k_2 are in $\mathbb{Z}_a^*$ with respective multiplicative inverses k_1^{-1} and k_2^{-1}, then

$$\left(k_1^{-1} \,\overline{\times}\, k_2^{-1}\right) \,\overline{\times}\, \left(k_1 \,\overline{\times}\, k_2\right) = \left(k_1^{-1} \,\overline{\times}\, k_1\right) \,\overline{\times}\, \left(k_2^{-1} \,\overline{\times}\, k_2\right) = 1 \; ,$$

so $k_1 \,\overline{\times}\, k_2$ has multiplicative inverse of $k_1^{-1} \,\overline{\times}\, k_2^{-1}$, and $k_1 \,\overline{\times}\, k_2$ therefore lies in $\mathbb{Z}_a^*$. Furthermore, $1 \in \mathbb{Z}_a^*$. Mathematically speaking, $\mathbb{Z}_a^*$ has the structure of a *commutative group* with group operation $\overline{\times}$ and identity element 1.

As we'll see presently, the following theorem lies at the heart of certain modern cryptographic schemes. It's actually a special case of a central result from group theory, but I'll prove it using an elementary argument.

2.4 Euler's Theorem: Let $a > 1$ be a positive integer and define $\mathbb{Z}_a^*$ and $\phi(a)$ as in the foregoing. Then every $k \in \mathbb{Z}_a^*$ satisfies

$$\left\langle\!\!\left\langle k^{\phi(a)} \right\rangle\!\!\right\rangle_a = 1 \; .$$

Proof: The theorem is trivially true when $k = 1$ and also when $a = 2$, in which case $\mathbb{Z}_a^* = \{1\}$, so let's assume that $a > 2$ and $k > 1$. Because $\mathbb{Z}_a^*$ is closed under the $\overline{\times}$ operation, we know that $\langle\!\langle k^j \rangle\!\rangle_a \in \mathbb{Z}_a^*$ for every $j > 0$. Thus

$$H_0 = \left\{1, k, \langle\!\langle k^2 \rangle\!\rangle_a \, , \langle\!\langle k^3 \rangle\!\rangle_a \, , \ldots\right\}$$

is a subset of $\mathbb{Z}_a^*$ and is therefore finite. It follows that there exist positive integers i and j with $i > j$ such that $\langle\!\langle k^i \rangle\!\rangle_a = \langle\!\langle k^j \rangle\!\rangle_a$. Multiplying that relation by $\left\langle\!\!\left\langle \left(k^{-1}\right)^j \right\rangle\!\!\right\rangle_a$, where k^{-1} is k's multiplicative inverse, reveals that $\langle\!\langle k^{i-j} \rangle\!\rangle_a = 1$. In particular, there exists at least one positive integer s such that $\langle\!\langle k^s \rangle\!\rangle_a = 1$. The smallest such positive integer, which I'll denote by r, is called the *order of k,* mod a. I claim that

$$H_0 = \left\{1, k, \langle\!\langle k^2 \rangle\!\rangle_a \, , \ldots , \langle\!\langle k^{r-1} \rangle\!\rangle_a \right\}$$

and that H_0 contains exactly r numbers. To see why, note that when $0 < j < r$, $\langle\!\langle k^j \rangle\!\rangle_a \neq 1$ because by construction r is the smallest positive integer j for which $\langle\!\langle k^j \rangle\!\rangle_a = 1$. If $0 < i < j < r$, we can't have $\langle\!\langle k^i \rangle\!\rangle_a = \langle\!\langle k^j \rangle\!\rangle_a$ because multiplying through by $\langle\!\langle k^{r-j} \rangle\!\rangle_a$ would yield

$$\left\langle\!\!\left\langle k^{r-(j-i)} \right\rangle\!\!\right\rangle_a = \langle\!\langle k^r \rangle\!\rangle_a = 1 \; ,$$

implying that a lower power than r of k — namely $r - (j - i)$ — would equal 1 mod a, again contradicting the definition of r. Accordingly, H_0 contains exactly r numbers, each of which is of the form $\langle\!\langle k^j \rangle\!\rangle_a$ for some $0 \leq j < r$.

I'll demonstrate in what follows that r is a divisor of $\phi(a)$, from which we can conclude that $\langle\!\langle k^{\phi(a)} \rangle\!\rangle_a = 1$. To see this, note that if $\phi(a) = rm$ for some positive integer m, then

$$\left\langle\!\!\left\langle k^{\phi(a)} \right\rangle\!\!\right\rangle_a = \langle\!\langle (k^r)^m \rangle\!\rangle_a = \langle\!\langle (\langle\!\langle k^r \rangle\!\rangle_a)^m \rangle\!\rangle_a = 1 \; .$$

The simplest case arises when $H_0 = \mathbb{Z}_a^*$, in which case r, the number of elements of H_0, is the same as $\phi(a)$, the number of elements of $\mathbb{Z}_a^*$, so that

$$\left\langle\!\!\left\langle k^{\phi(a)} \right\rangle\!\!\right\rangle_a = \langle\!\langle k^r \rangle\!\rangle_a = 1 \ .$$

If H_0 is a proper subset of $\mathbb{Z}_a^*$, we can find $n_1 \in \mathbb{Z}_a^*$ that doesn't lie in H_0. Define H_1 via

$$H_1 = \left\{ n_1, \langle\!\langle n_1 k \rangle\!\rangle_a \, , \left\langle\!\!\left\langle n_1 k^2 \right\rangle\!\!\right\rangle_a , \ldots , \left\langle\!\!\left\langle n_1 k^{r-1} \right\rangle\!\!\right\rangle_a \right\} \ .$$

Then

- H_1 contains r distinct numbers. That's because if we had $\left\langle\!\!\left\langle n_1 k^i \right\rangle\!\!\right\rangle_a = \left\langle\!\!\left\langle n_1 k^j \right\rangle\!\!\right\rangle_a$ for some $0 \le i < j < r$, then we could multiply through by the multiplicative inverse of n_1 to obtain $\left\langle\!\!\left\langle k^i \right\rangle\!\!\right\rangle_a = \left\langle\!\!\left\langle k^j \right\rangle\!\!\right\rangle_a$, which we know isn't true. Furthermore,
- H_1 is disjoint from H_0. If that weren't the case, we'd have $\left\langle\!\!\left\langle n_1 k^i \right\rangle\!\!\right\rangle_a = \left\langle\!\!\left\langle k^j \right\rangle\!\!\right\rangle_a$ for some i and j with $0 \le i, j < r$. Multiplying through by $\left\langle\!\!\left\langle k^{r-i} \right\rangle\!\!\right\rangle_a$ would yield $n_1 = \left\langle\!\!\left\langle k^{j+r-i} \right\rangle\!\!\right\rangle_a$, implying that $n_1 \in H_0$, which we know isn't true.

Thus the set $H_0 \cup H_1$ contains exactly $2r$ numbers. If that set encompasses all of $\mathbb{Z}_a^*$, then $\phi(a) = 2r$, so

$$\left\langle\!\!\left\langle k^{\phi(a)} \right\rangle\!\!\right\rangle_a = \left\langle\!\!\left\langle k^{2r} \right\rangle\!\!\right\rangle_a = 1 \ .$$

If $H_0 \cup H_1$ is a proper subset of $\mathbb{Z}_a^*$, we can find $n_2 \in \mathbb{Z}_a^*$ that lies neither in H_0 nor in H_1 and form

$$H_2 = \left\{ n_2, \langle\!\langle n_2 k \rangle\!\rangle_a \, , \left\langle\!\!\left\langle n_2 k^2 \right\rangle\!\!\right\rangle_a , \ldots , \left\langle\!\!\left\langle n_2 k^{r-1} \right\rangle\!\!\right\rangle_a \right\} \ .$$

Reasoning as in the preceding paragraph we can show easily that

- H_2 contains r distinct numbers, and
- H_0, H_1, and H_2 are mutually disjoint.

Thus the set $H_0 \cup H_1 \cup H_2$ contains $3r$ elements. If that set encompasses all of $\mathbb{Z}_a^*$, then $\phi(a) = 3r$, so

$$\left\langle\!\!\left\langle k^{\phi(a)} \right\rangle\!\!\right\rangle_a = \left\langle\!\!\left\langle k^{3r} \right\rangle\!\!\right\rangle_a = 1 \ .$$

You can see where this is going. If necessary, we plow on by choosing n_3, n_4, ... , n_{m-1} and forming mutually disjoint H_3, H_4, ... , H_{m-1}, each of which contains r numbers, until we've exhausted $\mathbb{Z}_a^*$ in the sense that

$$\mathbb{Z}_a^* = H_0 \cup H_1 \cup H_2 \cup \cdots H_{m-1} \ .$$

At that point we conclude that $\phi(a) = mr$, so

$$\left\langle\!\!\left\langle k^{\phi(a)} \right\rangle\!\!\right\rangle_a = \langle\!\langle k^{mr} \rangle\!\rangle_a = 1 \ ,$$

and we're done. □

Note that when p is prime, every positive integer $k < p$ is coprime with p, so $\phi(p) = p - 1$ and

$$\mathbb{Z}_p^* = \{1, 2, 3, \ldots, p-1\} \ ,$$

which is simply the set of all nonzero elements of $\mathbb{Z}_p$. Taking $a = p$ in Euler's Theorem yields the following result, which is easy to prove using basic facts about prime numbers.

2.5 Fermat's Little Theorem: If $p > 0$ is prime, then $\left\langle\!\left\langle k^{p-1} \right\rangle\!\right\rangle_p = 1$ for all $0 < k < p$.

Rudimentary cryptography and the notion of a key

Most people first encounter cryptography in the puzzle section of the newspaper. Presented with a string of letters and spaces such as

$$\text{U JROQ QWWP},$$

the reader is challenged to decipher it into a grammatical piece of English text under the assumption that each of the letters in the string stands unequivocally for some other letter. In essence, underlying the so-called cryptogram is a fixed permutation of the 26-letter English alphabet, and the reader's job is to figure out that permutation, or at least enough of it to decipher the given string. The solution need not be unique. You can check that the string above deciphers to I LOVE EGGS under one permutation and I HATE EGGS under another. People don't solve these puzzles by brute-forcing their way through the 26! possible permutations of the English alphabet. Statistics of English-letter frequencies and rules of English narrow the set of feasible solutions. A one-letter word in English must be either I or A (or maybe O in poetic contexts), and E, T, and O occur most frequently, so an educated guesser might begin by assuming that U stands for I and Q stands for E and seeing where that leads. Simple *substitution cyphers* of the this type have been around for millennia, and people — not to mention computers — have gotten good at solving them.

Julius Caesar allegedly used a substitution cypher to encode messages (presumably in Latin) for the purpose of secure communication. His cypher cycled the alphabet mod 3 in the sense that D stood for A, E for B, F for C, etc. An adversary intercepting one of Caesar's encrypted messages faced the same problem that the newspaper-puzzle solver faces. You can imagine what quick work a clever modern human or reasonable computer would make of a message encoded using Caesar's cypher. Indeed, computers have changed the game because they can perform brute-force computations that lie beyond human capabilities. Say, for example, that the messages you want to send are strings of digits. There are $10! = 3,628,800$ permutations of the ten digits, and it's easy to visualize a computer churning through all of them to decode a message sent encrypted using a simple substitution cypher. Things are significantly worse if you want to send bit strings, in which case only one nontrivial substitution cypher exists.

Elaborating on the simple substitution cypher leads to an important cryptographic advance, the *polyalphabetic substitution cypher.* Such an encryption scheme encodes different symbols in the message using different substitution cyphers and decides which substitution cypher to use on each symbol by referring to a *key.* Here's a simple example. Define E_n, $1 \leq n \leq 9$, as the substitution cypher for the English alphabet that cycles the English alphabet mod n, Caesar-style. Let the key be $k = 14853$. Suppose we want to encode the message MODULAR ARITHMETIC RULES. We write out the key repeatedly atop our message, i.e.

1 4 8 5 3 1 4 8 5 3 1 4 8 5 3 1 4 8 5 3 1 4

MODULAR ARITHMETIC RULES

and then use substitution cypher E_n to encrypt each letter above which n sits, so for example we encrypt RULES as ZZOFW. Using long keys and/or many different substitution cyphers in a scheme of this kind washes out the statistical and English-rules evidence that people use to solve simple substitution cyphers, making decrypting significantly more difficult.

Here's one more example of a polyalphabetic substitution cypher. Suppose that the messages we want to send are bit strings and the key k is also a bit string. For definiteness suppose that $k = 110011010$. If our message is $m = 11100010110011$, we encrypt it by extending the key by repetition to a bit string as long as the message and then performing bitwise binary addition, or an XOR operation, between that string and m. The encrypted message in this case is 001111101010. Imagine in general that the key k is as long as the messages we want to send, and that we generate k randomly by flipping a coin for each bit of k. Then no matter how organized a bit-string message is, the encrypted message will look like a random string of bits. Note that this XOR-encryption technique is indeed a polyalphabetic substitution cipher. A 0 in the key directs us to apply the trivial substitution cypher to the message bit whose coding the 0 regulates, whereas a 1 in the key directs us to apply the bit-flip substitution cypher to the relevant message bit.

Let's step back and consider more generally how encryption techniques such as those I've just described might help people communicate securely. We begin with a set of agents who want to communicate with each other so that an eavesdropper who intercepts transmissions between the agents won't be able to determine the semantic content of the agents' communications. The agents settle on an encryption scheme, for example a polyalphabetic substitution cypher with attendant key, and keep the scheme, particularly the value of the key, secret among themselves. Whatever scheme they employ should have at least three properties:

- Agents should be able to encrypt messages easily using their knowledge of the scheme and the value of the key.
- Agents should be able to decrypt encoded messages easily using their knowledge of the scheme and the value of the key.
- An eavesdropper should not be able to decrypt encoded messages without knowing the value of the key, even if the eavesdropper knows some other general features of the encryption scheme (e.g. that the agents are using substitution cyphers, say)

Polyalphabetic substitution cyphers, especially when they employ many substitution cyphers and use long keys, meet all three criteria. The renowned Enigma machine used by the Germans during World War II implemented polyalphabetic substitution cyphers, and cracking the Germans' code depended not only on the work of some brilliant mathematicians, including Alan Turing, but on some lucky breaks.

The Hellman-Diffie-Merkle (HDM) key-establishment protocol

Ultra-complex polyalphabetic substitution cyphers and other related encryption schemes employing keys, while powerful tools for encryption, come encumbered with a difficult task, namely, delivering the key to all the agents. You can imagine getting everyone together in a room and agreeing once and for all on the value of the key, but that won't work in the real world. Agents leave or get fired from the group. Sometimes they're careless. Whatever the cause, keys have a way of getting out, and the agents will need to re-set the key at least occasionally and more likely on a regular basis. Agents in far-flung locations can't be expected to assemble periodically to do an in-person re-set, so they need to figure out a way to circulate new key values securely among themselves. But that's yet another secure-communication problem layered on top of the one they started with, and you can visualize how these meta-problems proliferate — shall we use a special key to encrypt the particular messages in which we circulate the new key? — ad infinitum.

Martin Hellman and Whitfield Diffie proposed in 1976 a solution based on modular arithmetic to the key-establishment problem. I'll describe their scheme and then attempt to explain in general terms why it's effective. All agents begin by agreeing on a large prime number p and a base $b \in \mathbb{Z}_p^*$. Each agent then picks privately, and keeps as a secret, a number $e \in \mathbb{Z}_p^*$. What happens when one agent wants to communicate with another? Most people call the two communicating agents Alice and Bob and call the adversarial eavesdropper Eve (see what they did there?), but I'll assume Frodo wants to communicate with Sam and doesn't want Gollum to hear. Frodo sends Sam $\langle\!\langle b^{e_F} \rangle\!\rangle_p$, where e_F is Frodo's private e-value. When Sam receives that number, he knows Frodo wants to communicate, so Sam sends Frodo $\langle\!\langle b^{e_S} \rangle\!\rangle_p$, where e_S is Sam's e-value. Then Frodo takes what Sam has sent him and computes

$$\left\langle\!\!\left\langle\, \langle\!\langle b^{e_S} \rangle\!\rangle_p{}^{e_F} \right\rangle\!\!\right\rangle_p = \langle\!\langle b^{e_F e_s} \rangle\!\rangle_p = k$$

while Sam proceeds similarly with what Frodo sent him and computes

$$\left\langle\!\!\left\langle\, \langle\!\langle b^{e_F} \rangle\!\rangle_p{}^{e_S} \right\rangle\!\!\right\rangle_p = \langle\!\langle b^{e_F e_s} \rangle\!\rangle_p = k \ ,$$

which is the same k that Frodo computed. Then Frodo and Sam communicate using their favorite encryption scheme, perhaps a polyalphabetic substitution cypher, employing k as the key.

Eavesdropper Gollum can't decipher Frodo's and Sam's subsequent communications unless he knows what encryption scheme they're using, as indeed he might, along with the value of k. Figuring out k turns out to be hard for him, even if he knows the p and b that Frodo and Sam agreed upon to start with. To compute k, Gollum must solve for at least one of e_F and e_S having seen only $\langle\!\langle b^{e_F} \rangle\!\rangle_p$ and $\langle\!\langle b^{e_S} \rangle\!\rangle_p$. Even if he knows p and b, he needs to compute the mod-p logarithm to the base b of a number, and that problem turns out to have worst-case complexity that grows linearly in p and thus exponentially in the number of bits or digits required to specify p, which means that if Frodo and Sam pick p large enough, Gollum's k-computation problem is computationally intractable.

Hellman and Diffie's result revolutionized cryptography. Hellman proposed that Ralph Merkle be credited with important early work leading to the Hellman-Diffie scheme, which is why Merkle's name appears alongside theirs. It's worth noting two disadvantages to their scheme. First, it requires an initial handshake between any two agents who want to communicate — they need to do one back-and-forth to establish the key value, but after that each of them can fire off one-way messages at will. Second, it's vulnerable to a so-called Man-in-the-Middle Attack. Suppose Saruman knows p and b and intercepts Frodo's original key-establishment transmission to Sam. Then, posing as Sam, Saruman sends Frodo $\langle\!\langle b^{e_\sigma}\rangle\!\rangle_p$, where e_σ is Saruman's chosen e-value. Frodo thinks $\langle\!\langle b^{e_\sigma}\rangle\!\rangle_p$ is $\langle\!\langle b^{e_S}\rangle\!\rangle_p$, and therefore computes $k = \langle\!\langle b^{e_\sigma e_F}\rangle\!\rangle_p$, which Saruman also computes, whereupon Frodo and Saruman begin communicating using key k, provided of course that Saruman also knows the encryption scheme Frodo and Sam plan to use and thus knows how to employ k. Saruman can also impersonate Frodo and send Sam $\langle\!\langle b^{e_\sigma}\rangle\!\rangle_p$, whereupon Sam and Saruman establish key $\langle\!\langle b^{e_S e_\sigma}\rangle\!\rangle_p$ and start communicating. Frodo and Sam then think they're talking with each other while Saruman mediates their interaction to his diabolical delight. To get around the Man-in-the-Middle vulnerability, people have developed sophisticated protocols for agent authentication that I won't discuss here.

Private keys, public keys, and RSA encryption

So far I've used the word "key" to refer to a piece of information shared by two communicating agents and unavailable to an eavesdropper. I'd like to loosen that characterization a bit and think of a key as being a collection of pieces of information, some private to individual agents, some shared between communicating agents, and some publicly available. When Sam and Frodo establish a key using HDM, we can think of e_F and e_S as parts of the key that are private to Frodo and Sam, respectively, and $k = \langle\!\langle b^{e_F e_S}\rangle\!\rangle_p$ as a part of the key that Frodo and Sam both know.

A secure-communication scheme involving only private keys is *Shamir's Three-Pass Protocol,* due to Adi Shamir. To understand the basic idea, imagine a lockbox with two locks. Frodo has the key to one lock and Sam has the key to the other. When Frodo wants to send a message to Sam, Frodo puts the message in the unlocked box, locks the lock to which he has the key, and sends the locked box to Sam. Sam receives the box, locks the other lock, and sends the now double-locked box back to Frodo. Frodo unlocks his lock and sends the now single-locked box back to Sam. Finally, Sam unlocks the single locked lock, to which he has the key, and retrieves the message. Let's see how Shamir's protocol implements this ingenious sequence of actions using modular arithmetic.

Assume that the messages agents want to send are positive integers and that all agents have agreed on a large prime p. Each agent chooses e and d in $\mathbb{Z}^*_{p-1}$ such that $\langle\!\langle ed\rangle\!\rangle_{p-1} = 1$. Note that such an e and d are easy to choose. For e, an agent could pick any prime that doesn't divide $p-1$, and then set

$$d = \left\langle\!\!\left\langle e^{\phi(p-1)-1}\right\rangle\!\!\right\rangle_{p-1} ,$$

which makes $\langle\!\langle ed\rangle\!\rangle_{p-1} = 1$ by Euler's Theorem 2.4. Suppose Frodo wants to send message m to Sam. We can assume $m < p$ by breaking m into pieces if necessary. Shamir's Three-Pass Protocol works as follows.

- Frodo sends Sam $\langle\!\langle m^{e_F}\rangle\!\rangle_p$.

- Sam receives the transmission and sends back to Frodo
$$\left\langle\!\!\left\langle \langle\!\langle m^{e_F}\rangle\!\rangle_p{}^{e_S} \right\rangle\!\!\right\rangle_p = \langle\!\langle m^{e_F e_S}\rangle\!\rangle_p \ .$$
- Frodo receives Sam's transmission and sends back to Sam
$$\begin{aligned}\left\langle\!\!\left\langle \langle\!\langle m^{e_F e_S}\rangle\!\rangle_p{}^{d_F} \right\rangle\!\!\right\rangle_p &= \langle\!\langle m^{e_F d_F e_S}\rangle\!\rangle_p \\ &= \left\langle\!\!\left\langle m^{(l(p-1)+1)e_S} \right\rangle\!\!\right\rangle_p \text{ for some } l \in \mathbb{N} \\ &= \left\langle\!\!\left\langle \left\langle\!\!\left\langle m^{l(p-1)} \right\rangle\!\!\right\rangle_p \langle\!\langle m^{e_S}\rangle\!\rangle_p \right\rangle\!\!\right\rangle_p \\ &= \langle\!\langle m^{e_S}\rangle\!\rangle_p \ ,\end{aligned}$$
where the second line holds because $\langle\!\langle e_F d_F\rangle\!\rangle_{p-1} = 1$ and the fourth line holds because
$$\left\langle\!\!\left\langle m^{l(p-1)} \right\rangle\!\!\right\rangle_p = \left\langle\!\!\left\langle \langle\!\langle m^{p-1}\rangle\!\rangle_p{}^{l} \right\rangle\!\!\right\rangle_p = 1 \ ,$$
since by Fermat's Little Theorem 2.5 we have $\left\langle\!\!\left\langle m^{p-1} \right\rangle\!\!\right\rangle_p = 1$.
- Finally, Sam computes m by taking what he receives from Frodo and computing
$$\begin{aligned}\left\langle\!\!\left\langle \langle\!\langle m^{e_S}\rangle\!\rangle_p{}^{d_S} \right\rangle\!\!\right\rangle_p &= \langle\!\langle m^{e_S d_S}\rangle\!\rangle_p \\ &= \left\langle\!\!\left\langle m^{j(p-1)+1} \right\rangle\!\!\right\rangle_p \text{ for some } j \in \mathbb{N} \\ &= \langle\!\langle m\rangle\!\rangle_p = m \ ,\end{aligned}$$
where Fermat's Little Theorem proves the third line.

Gollum the eavesdropper sees the message m raised to various powers mod p, but he has no way of determining m without knowing at least one of Frodo's and Sam's d-values. If he knows p and can figure out one of their e-values from what he sees, he could easily calculate the corresponding d. As in the case of HDM, figuring out an e-value requires computing logarithms mod p. Gollum sees $\overline{m} = \langle\!\langle m^{e_F}\rangle\!\rangle_p$ and
$$\langle\!\langle m^{e_F e_S}\rangle\!\rangle_p = \langle\!\langle \overline{m}^{e_S}\rangle\!\rangle_p \ ,$$
so obtaining e_S entails taking the mod p logarithm to the base $\overline{m}$ of something, which is computationally intractable for large p. Like HDM, Shamir's Three-Pass Protocol is vulnerable to a Man-in-the-Middle Attack. If Saruman knows p and chooses e_σ and d_σ appropriately, he can intercept Frodo's initial transmission, impersonate Sam via a return message, and finally receive $\langle\!\langle m^{e_\sigma}\rangle\!\rangle_p$, allowing him to compute m. The Three-Pass Protocol also requires three transmissions per message sent and, unlike HDM, doesn't allow for one-way transmissions after an initial handshake.

We owe the last encryption scheme I'll discuss to Ronald Rivest, Adi Shamir, and Leonard Adelman. RSA encryption, as it's known, features both public and private keys. To set things up, each agent

- picks two large primes p and q with $p \neq q$ and keeps these private;
- picks e and d in $\mathbb{Z}^*_{(p-1)(q-1)}$ so that $\langle\!\langle ed\rangle\!\rangle_{(p-1)(q-1)} = 1$;
- publishes e along with $N = pq$, while keeping d private.

An agent's published (e, N)-pair is the agent's *public RSA key,* and all agents' public RSA keys appear in some sort of directory.

If Gandalf wants to send a message $m \in \mathbb{N}$ to Frodo, Gandalf looks up Frodo's public RSA key (e_F, N_F) and sends Frodo $\langle\!\langle m^{e_F} \rangle\!\rangle_{N_F}$. Here I'm assuming that $m < N_F$, which Gandalf can guarantee by breaking m into pieces if necessary. Upon receiving Gandalf's transmission, Frodo computes

$$\begin{aligned} \left\langle\!\!\left\langle \left(\langle\!\langle m^{e_F} \rangle\!\rangle_{N_F} \right)^{d_F} \right\rangle\!\!\right\rangle_{N_F} &= \langle\!\langle m^{e_F d_F} \rangle\!\rangle_{N_F} \\ &= \left\langle\!\!\left\langle m^{l(p-1)(q-1)+1} \right\rangle\!\!\right\rangle_{pq} \quad \text{for some } \ l \in \mathbb{N} \ , \end{aligned}$$

where to minimize subscripts I've denoted Frodo's chosen primes by p and q so that $N_F = pq$. Observe now that $\phi(pq) = (p-1)(q-1)$ because $\mathbb{Z}_{pq}$ contains pq elements, and the only ones not coprime with pq are multiples of p (exactly $q-1$ of those) and multiples of q (exactly $p-1$ of those), and 0 is the only number in $\mathbb{Z}_{pq}$ that's a multiple of both p and q, whereby $\phi(pq)$, the number of elements in $\mathbb{Z}^*_{pq}$, is

$$\phi(p, q) = pq - (p + q - 1) = (p-1)(q-1) \ .$$

Thus by Euler's Theorem 2.4 we have

$$\left\langle\!\!\left\langle m^{l(p-1)(q-1)} \right\rangle\!\!\right\rangle_{pq} = \left\langle\!\!\left\langle \left(\left\langle\!\!\left\langle m^{\phi(pq)} \right\rangle\!\!\right\rangle_{pq} \right)^{l} \right\rangle\!\!\right\rangle_{pq} = 1 \ .$$

It follows that

$$\left\langle\!\!\left\langle m^{l(p-1)(q-1)+1} \right\rangle\!\!\right\rangle_{pq} = \langle\!\langle m \rangle\!\rangle_{pq} = m \ .$$

I've glossed over one detail: for Euler's Theorem to apply, we need $m \in \mathbb{Z}^*_{pq}$. For large p and q, the fraction of m-values in $\mathbb{Z}_{pq}$ that don't lie in $\mathbb{Z}^*_{pq}$ is small and therefore easily avoided.

How does RSA foil eavesdroppers? Saruman sees Gandalf's encrypted message to Frodo. Saruman knows Frodo's publicly available (e, N)-pair and knows that Gandalf's transmission takes the form $\langle\!\langle m^e \rangle\!\rangle_N$. To decrypt the transmission, Saruman needs Frodo's d-value, which only Frodo knows. Saruman could figure out d easily if he knew $(p-1)(q-1)$, which would require that he knew p and q. Knowing $N = pq$, as it happens, doesn't help Saruman compute p and q individually. Prime factorization, even of a number known to be the product of two primes, has complexity that grows linearly in the size of the number and hence exponentially in then number of bits required to specify the number. For large N, the calculation is intractable. Furthermore, RSA possesses a significant advantage over the other two secure-communication schemes we've investigated. RSA requires no initial handshake or real-time back-and-forth between agents. Any agent can send a message anytime to any other agent whose public RSA key appears in the directory.

Finally, to wrap things up, you may have been wondering about the computations the agents must perform to encrypt and decrypt transmissions in the schemes I've described. For the schemes to be useful these computations need to be quick and easy relative to the tasks confronting eavesdroppers. All encryption and decryption computations take the form $\langle\!\langle n^r \rangle\!\rangle_l$ for possibly rather large positive integers n, r, and l. As it happens, fast algorithms for calculating such modular powers exist. One such algorithm, the method of repeated squares, modularizes the well known computer-science technique of exponentiation by squaring.

CHAPTER 3

Working with Real and Complex Numbers

My main purpose in this chapter is to summarize fairly briskly the central results on sequences and series of real and complex numbers. We've touched on some of the concepts already (e.g. convergent sequences, Cauchy sequences, etc.) in Chapter 1, but I think it's useful to collect everything in one place in a kind of bulleted-list format. The lack of a lot of intervening text makes at times for some high-density mathematics, but I hope the layout will facilitate easy reference.

To start with, I'll assume you have some basic familiarity with the real numbers $\mathbb{R}$ and the complex numbers $\mathbb{C}$. I'll assume you understand the algebra of complex numbers at a standard pre-calculus level (real and imaginary parts, addition and multiplication, magnitude and argument, etc.). I'll cleave to electrical-engineering convention and use notation j for $\sqrt{-1}$. For a real number a, $|a|$ denotes the absolute value of a; for a complex number c, $|c|$ denotes the magnitude of c. So if $c = a + jb$, with a and b in $\mathbb{R}$, then

$$|c| = \sqrt{a^2 + b^2}\ .$$

The distance between two real numbers a and b is the absolute value of $a - b$ and the distance between two complex numbers c_1 and c_2 is the magnitude of $c_1 - c_2$. To avoid having to type "real or complex numbers" a zillion times, I'll use the notation $\mathbb{F}$ to denote the phrase "$\mathbb{R}$ or $\mathbb{C}$." The "F" is supposed to mean "field."

Sequences and their convergence

A *sequence in* $\mathbb{F}$ is an ordered list of elements of $\mathbb{F}$ indexed by $\mathbb{N}$. We use notation such as $\{a_n\}$ or $\{c_n\}$ to denote such a sequence. So, for example,

$$\{a_n\} = a_0, a_1, a_2, a_3, \dots\ .$$

We say that a sequence $\{c_n\}$ in $\mathbb{F}$ *converges* to $\bar{c} \in \mathbb{F}$ when the distance between c_n and $\bar{c}$ approaches zero as $n \to \infty$. In this case, we write

$$\lim_{n\to\infty} c_n = \bar{c}\ .$$

A precise mathematical definition of convergence: $\{c_n\}$ converges to $\bar{c}$ when for every $\epsilon > 0$ there exists an integer $N > 0$ such that $|c_n - \bar{c}| < \epsilon$ for every $n > N$. It turns out that a sequence of complex numbers converges if and only if its real-part and imaginary-part sequences both converge, in which case the limit of the real parts is the real part of the limit, and the limit of the imaginary parts is the imaginary part of the limit.

3.1 Fact: A sequence $\{c_n = a_n + jb_n\}$ in $\mathbb{C}$, where a_n and b_n are real for all n, converges to $\bar{c} = \bar{a} + j\bar{b} \in \mathbb{C}$, where $\bar{a}$ and $\bar{b}$ are real, if and only if the real sequences $\{a_n\}$ and $\{b_n\}$ converge respectively to $\bar{a}$ and $\bar{b}$ in $\mathbb{R}$.

Proof: First of all, for every $n \in \mathbb{N}$,

$$|c_n - \bar{c}| = \sqrt{|a_n - \bar{a}|^2 + |b_n - \bar{b}|^2}\ ,$$

If $\{c_n\}$ converges to $\bar{c}$, then for every $\epsilon > 0$ we can find $N > 0$ so that $|c_n - \bar{c}| < \epsilon$ when $n > N$. Hence for $n > N$, we have

$$\sqrt{|a_n - \bar{a}|^2 + |b_n - \bar{b}|^2} < \epsilon\ ,$$

which implies that both $|a_n - \bar{a}| < \epsilon$ and $|b_n - \bar{b}| < \epsilon$ for every $n > N$. Accordingly, $\{a_n\}$ converges to $\bar{a}$ and $\{b_n\}$ converges to $\bar{b}$.

Conversely, if $\{a_n\}$ converges to $\bar{a}$ and $\{b_n\}$ converges to $\bar{b}$, then for every $\epsilon > 0$ we can find $N > 0$ so that both $|a_n - \bar{a}| < \epsilon/\sqrt{2}$ and $|b_n - \bar{b}| < \epsilon/\sqrt{2}$ when $n > N$. Hence for $n > N$, we have

$$|c_n - \bar{c}| < \sqrt{\epsilon^2/2 + \epsilon^2/2} = \epsilon\ .$$

Accordingly, $\{c_n\}$ converges to $\bar{c}$. □

A sequence $\{c_n\}$ in $\mathbb{F}$ is called a *Cauchy sequence* when the terms in the sequence get closer and closer together as n increases. Specifically, $\{c_n\}$ is a Cauchy sequence when for every $\epsilon > 0$ there exists an integer $N > 0$ such that $|c_m - c_n| < \epsilon$ whenever m and n are both bigger than N.

3.2 Fact: A sequence $\{c_n\}$ in $\mathbb{F}$ is a Cauchy sequence if and only if it is a convergent sequence.

"Proof:" I've put quotes around the word Proof because you really can't prove from first principles that every Cauchy sequence converges. As I noted in Chapter 1, the real and complex numbers are just set up so that every Cauchy sequence converges. The converse — that every convergent sequence is Cauchy — is easy to demonstrate.

Suppose, then, that $\{c_n\}$ converges to $\bar{c}$. Then for every $\epsilon > 0$ we can find some $N > 0$ such that $|c_n - \bar{c}| < \epsilon/2$ when $n > N$. So if m and n are both bigger than N, we have

$$|c_m - c_n| = |c_m - \bar{c} + \bar{c} - c_n| \leq |c_m - \bar{c}| + |\bar{c} - c_n| < \epsilon/2 + \epsilon/2 = \epsilon\ .$$

We have shown: if $\{c_n\}$ converges, then for every $\epsilon > 0$ there exists $N > 0$ such that $|c_m - c_n| < \epsilon$ whenever m and n are both bigger than N, which means that $\{c_n\}$ is a Cauchy sequence. □

Series and their convergence

If $\{c_n\}$ is a sequence in $\mathbb{F}$, the *infinite series with nth term c_n* is the formal expression

$$\sum_{n=0}^{\infty} c_n \ .$$

Associated with that expression is another sequence in $\mathbb{F}$, the sequence $\{s_n\}$ of *partial sums* defined by

$$s_n = \sum_{m=0}^{n} c_m$$

for each $n \in \mathbb{N}$. We say that *the infinite series* $\sum_{n=0}^{\infty} c_n$ *converges* when the sequence $\{s_n\}$ of partial sums is a convergent sequence in $\mathbb{F}$, in which case we say that the *sum* of the series is $\lim_{n\to\infty} s_n$.

We'll also encounter two-sided infinite series of the form

$$\sum_{n=-\infty}^{\infty} c_n \ ,$$

where $\{c_n : n \in \mathbb{Z}\}$ is a sequence in $\mathbb{F}$ whose index n extends infinitely in both directions. A two-sided infinite series converges when both of the series

$$\sum_{n=0}^{\infty} c_n$$

and

$$\sum_{n=-\infty}^{-1} c_n = \sum_{m=1}^{\infty} c_{-m}$$

converge. In this case, the sum of the doubly infinite series is the sum of the sums of the two one-sided series. This convergence condition is not the same as requiring that the sequence $\{s_n\}$ of partial sums defined by

$$s_n = \sum_{m=-n}^{n}$$

converges. Consider, for example, the two-sided series wherein $c_n = n$ for all $n \in \mathbb{Z}$. For that series, $s_n = 0$ for every n, but the series does not converge by our definition.

If $\{c_n\}$ is a (one-sided) sequence in $\mathbb{F}$, $\{c_n\}$ is said to be *summable* when the infinite series $\sum_{n=0}^{\infty} c_n$ converges, and $\{c_n\}$ is said to be *absolutely summable* when the infinite series $\sum_{n=0}^{\infty} |c_n|$ converges. Similarly, if $\{c_n\}$ is a two-sided sequence in $\mathbb{F}$, $\{c_n\}$ is summable when $\sum_{n=-\infty}^{\infty} c_n$ converges and absolutely summable when $\sum_{n=-\infty}^{\infty} |c_n|$ converges. Many sequences are summable but not absolutely summable. An example is the sequence $\{c_n = (-1)^n/(n+1)\}$. It turns out that the infinite series

$$\sum_{n=0}^{\infty} c_n = 1 - 1/2 + 1/3 - 1/4 + \cdots$$

converges to $\ln(2)$, but the infinite series

$$\sum_{n=0}^{\infty} |c_n| = 1 + 1/2 + 1/3 + 1/4 + \cdots$$

does not converge. If you Google on “alternating harmonic series” you can read more about this example. In contrast, every absolutely summable sequence is summable, as the following quite useful result asserts.

3.3 Fact: If a one- or two-sided sequence $\{c_n\}$ in $\mathbb{F}$ is absolutely summable, then $\{c_n\}$ is also summable.

Proof: First suppose $\{c_n\}$ is one-sided and let $s_n = \sum_{m=0}^{n} c_n$ and $\widetilde{s}_n = \sum_{m=0}^{n} |c_n|$. Assume that $\{c_n\}$ is absolutely summable. Then the sequence $\{\widetilde{s}_n\}$ converges. By Fact 3.2, $\{\widetilde{s}_n\}$ is a Cauchy sequence. Now consider the sequence $\{s_n\}$. Let m and n be natural numbers and assume without loss of generality that $n \geq m$. Then

$$|s_m - s_n| = \left| \sum_{l=m+1}^{n} c_l \right| \leq \sum_{l=m+1}^{n} |c_l| = |\widetilde{s}_m - \widetilde{s}_n| \; .$$

Since $\{\widetilde{s}_n\}$ is a Cauchy sequence, given $\epsilon > 0$ we can find $N \in \mathbb{N}$ such that when m and n are bigger than N, we have $|\widetilde{s}_m - \widetilde{s}_n| < \epsilon$. For that same N, we therefore have $|s_m - s_n| < \epsilon$ when m and n are bigger than N. It follows that $\{s_n\}$ itself is a Cauchy sequence, which converges by Fact 3.2, and we conclude that $\{c_n\}$ is summable.

If $\{c_n\}$ is an absolutely summable two-sided sequence, apply the foregoing argument to each of the two one-sided sequences $\{c_n : n \geq 0\}$ and $\{c_{-n} : n > 0\}$ to prove their summability, from which it follows that $\{c_n\}$ is summable. □

Upper and lower bounds

Mathematical models for real-world phenomena are rarely if ever exact. Engineers and applied scientists working with such models need to be able to estimate quantitatively how inexact the models are. Developing these estimates requires bounding things. Even theoretical results about mathematical models make assertions about upper and lower bounds on quantities of interest. I’m talking about statements like “If you employ such-and-such a communication scheme, your probability of a one-bit error is bounded from above by .001.” Or, “If your input amplitude is bounded from above by R, your output will be bounded from above by ΓR.” So it’s important to develop a facility for working with bounds.

An *upper bound* for a set $A \subset \mathbb{R}$ is a real number $\bar{v}$ such that $a \leq \bar{v}$ for every $a \in A$. Similarly, a *lower bound* for $A \subset \mathbb{R}$ is a real number $\underline{v} \in \mathbb{R}$ such that $a \geq \underline{v}$ for every $a \in A$. A set of real numbers is *bounded from above* when it has an upper bound and *bounded from below* when it has a lower bound. A set is simply *bounded* when it has both an upper bound and a lower bound.

If $B \subset A$, then every upper or lower bound for A is also an upper or lower bound for B, so if A is bounded from above or below, then B is too. It’s also easy to see that $A \subset \mathbb{R}$ is bounded if and only if A is a subset of some bounded interval.

The idea is that A is bounded if and only if $A \subset [\underline{v}, \bar{v}]$ where $\underline{v}$ and $\bar{v}$ are lower and upper bounds for A. Bounded intervals themselves are archetypal bounded sets.

Every finite set $A \subset \mathbb{R}$ has upper and lower bounds $\bar{v}$ and $\underline{v}$ that are elements of A. These are the *maximum* and *minimum* elements in A, and we denote them by $\max(A)$ and $\min(A)$. Infinite bounded sets don't necessarily have maxima and minima. Canonical examples are open intervals such as $(3, 7)$. On the other hand, $\max([0, 1]) = 1$ and $\min([0, 1]) = 0$.

A property of real numbers analogous to the "every Cauchy sequence converges" property is the so-called *least upper bound property.* It's analogous to the Cauchy thing in the sense that you can't really prove it since the real numbers are "rigged" so that it holds. In fact, as we'll see in Theorem 3.9 below, these two special built-in features of real numbers are equivalent in the sense that one holds if and only if the other does. The basic idea is that every bounded set of real numbers, even if it lacks a maximum and/or minimum, has upper and lower bounds that are "tight" in some sense.

3.4 Fact: If $A \subset \mathbb{R}$ is bounded from above, A has a least upper bound. In other words, if A is bounded from above there exists an upper bound $\bar{\bar{v}}$ for A such that $\bar{\bar{v}} \leq \bar{v}$ for every upper bound $\bar{v}$ for A. We write $\sup(A)$ for the least upper bound $\bar{\bar{v}}$. "Sup" stands for "supremum." If $A \subset \mathbb{R}$ is bounded from below, A has a greatest lower bound. In other words, if A is bounded from below there exists a lower bound $\underline{\underline{v}}$ for A such that $\underline{\underline{v}} \geq \underline{v}$ for every lower bound $\underline{v}$ for A. We write $\inf(A)$ for the greatest lower bound $\underline{\underline{v}}$. "inf" stands for "infimum." □

Least upper bounds are sort of like maxima and greatest lower bounds are sort of like minima. If a set A does indeed have a maximum, then that maximum is equal to $\sup(A)$. Similarly, if A has a minimum, then that minimum is equal to $\inf(A)$. It's instructive to prove these assertions, and I suggest you try your hand at it and see how it applies to the closed interval $[0, 1]$. As for the open interval $A = (3, 7)$, it's clear that every $\bar{v} \geq 7$ is an upper bound for A and every $\underline{v} \leq 3$ is a lower bound for A. Furthermore, no number less than 7 is an upper bound and no number greater than 3 is a lower bound. It follows that $\sup(A) = 7$ and $\inf(A) = 3$.

Working with bounds takes some practice and can be tricky. I'd like now to explore in some detail an example I've found helpful to study carefully and understand thoroughly. A well traveled aphorism in optimization theory asserts that "the max of the min is no greater than the min of the max." What does this mean? First let's suppose we have an $(m \times n)$ real matrix P. The entries in P constitute a finite set, so $\max(\{P_{ij}\})$ and $\min(\{P_{ij}\})$ exist — they're the largest and smallest entries in P. Consider now the following two procedures. The first procedure first identifies the smallest entry in each row of P (say, by circling it), and then maximizes over all the circled entries, one of which sits in each of P's rows. The second procedure first circles the largest entry in each column of P, and then minimizes over all the circled entries, one of which sits in each of P's columns.

Formally, the first procedure computes

$$\max\left(\{\min\left(\{P_{ij} : 1 \leq j \leq n\}\right) : 1 \leq i \leq m\}\right) \tag{1}$$

and the second procedure computes

$$\min\left(\{\max\left(\{P_{ij} : 1 \le i \le m\}\right) : 1 \le j \le n\}\right) \ . \tag{2}$$

Common abbreviations for these are $\max_i \min_j P_{ij}$ and $\min_j \max_i P_{ij}$ respectively. In (1), the inner minimization computes, for each i, the smallest element in row i of P. The outer maximization then computes the maximum of all these row minima. The inner maximization in (2) computes, for each j, the largest element in column j of P, and the outer minimization then computes the minimum of all these column maxima. For every i and j we have

$$\min(\{P_{ij} : 1 \le j \le n\}) \le P_{ij} \le \max(\{P_{ij} : 1 \le i \le m\}) \ .$$

This is shorthand for mn inequalities. The term on the far left depends only on i and the term on the far right depends only on j and we can vary i and j independently. It follows that every "column min" is less than or equal to every "row max," so the maximum of the column mins, namely (1), cannot exceed the minimum of the column maxes (2).

The following parable might sharpen your intuition. Imagine that we've entrusted two competitors, a maximizer and a minimizer, with selecting an element of P. The maximizer wants the selected element to be big, and her job is to pick the row containing the element. The minimizer wants the selected element to be small, and his job is to pick the column containing the element. They don't pick simultaneously; instead, one goes first and the other follows. You might expect that each competitor is happier when he or she goes first than when he or she goes second. In (1), the minimizer goes first and gets an outcome more pleasing to him (i.e. smaller) than when the maximizer goes first in (2), leading to an outcome more pleasing to her (i.e. larger) than in (1).

Going first affords a competitor the opportunity to narrow down the choices available to the competitor who goes second. Try it yourself on a few matrices and you'll see what's going on. Then find a friend and play the game where the two of you make your choices simultaneously. Do it a few times and see whether the same outcome arises every time.

The "max min $\le$ min max" result extends to sup and inf. Suppose A and B are sets of real numbers and $f : A \times B \to \mathbb{R}$ is a function. Suppose that the range of f lies within some bounded set $C \subset \mathbb{R}$. For each $a \in A$, set

$$C_a = \{c \in C : c = f(a,b) \text{ for some } b \in B\} \ .$$

Since C is bounded, so is C_a, and in particular C_a has a greatest lower bound. Define a function $g : A \to \mathbb{R}$ by

$$g(a) = \inf(C_a) \ .$$

Note that $g(a) \le f(a,b)$ for every $a \in A$ and $b \in B$. Now given $b \in B$, set

$$C_b = \{c \in C : c = f(a,b) \text{ for some } a \in A\} \ .$$

C_b is also a bounded set and therefore has a least upper bound. Define $h : C_b \to \mathbb{R}$ by

$$h(b) = \sup(C_b) \ .$$

Then $f(a,b) \le h(b)$ for every $a \in A$ and $b \in B$.

We've arrived at the following chain of inequalities.

$$g(a) \le f(a,b) \le h(b) \ \text{ for all } a \in A \ , \ b \in B \ .$$

Now set

$$D_1 = \{d \in \mathbb{R} : d = g(a) \text{ for some } a \in A\}$$

and

$$D_2 = \{d \in \mathbb{R} : d = h(b) \text{ for some } b \in B\}\ .$$

Our chain of inequalities shows that every number in D_2 is an upper bound for D_1 It follows that $\sup(D_1)$ is a lower bound for D_2; otherwise $\sup(D_1)$ would exceed some number in D_2, which is impossible since $\sup(D_1)$ cannot exceed any upper bound for D_1. Since $\sup(D_1)$ is a lower bound for D_2, it cannot exceed the greatest lower bound for D_2, so

$$\sup(D_1) \le \inf(D_2)\ .$$

People summarize this discussion with the inequality

$$\sup_{a\in A}\left(\inf_{b\in B} f(a,b)\right) \le \inf_{b\in B}\left(\sup_{a\in A} f(a,b)\right)\ .$$

In words, if you first "max out" $f(a,b)$ over a for each fixed b, and then take the "min" over b of all those "max" values, you get something at least as big as you would if you reversed the process by "minimizing" $f(a,b)$ over b for each fixed a and then "maxing" out over a all the "min" values you obtained. It's worth considering a simple example. Suppose $A = B = (0,1)$ and

$$f(a,b) = \begin{cases} 1 & \text{if } a \ge b \\ 0 & \text{if } a < b\ . \end{cases}$$

With notation as above, $C_a = C_b = \{0,1\}$ for every a and $b \in (0,1)$. So $g(a) = 0$ for every a and $h(b) = 1$ for every b. Hence $D_1 = \{0\}$ and $D_2 = \{1\}$, and $\sup(D_1) = 0$ while $\inf(D_2) = 1$.

Monotonic sequences

From Fact 3.4 we get two extremely useful results.

3.5 Fact: Every sequence $\{a_n\}$ of real numbers that is *bounded from above* and *monotonically increasing* — i.e., $a_n \le a_{n+1}$ for all $n \in \mathbb{N}$ — has a limit $\bar{a}$, and $\bar{a} = \sup(\{a_n\})$.

Proof: Let $\bar{\bar{v}} = \sup(\{a_n\})$. Note that for every $\epsilon > 0$ we can find some $N_o \in \mathbb{N}$ such that

$$|a_{N_o} - \bar{\bar{v}}| = \bar{\bar{v}} - a_{N_o} < \epsilon\ .$$

Otherwise, all the $\{a_n\}$ would be at least ϵ below $\bar{\bar{v}}$, meaning that $\bar{\bar{v}} - \epsilon$ would be an upper bound for $\{a_n\}$, which is impossible since $\bar{\bar{v}}$ is the least upper bound. Now, since $a_n \ge a_{N_o}$ for all $n > N_o$, we conclude that

$$|a_n - \bar{\bar{v}}| = \bar{\bar{v}} - a_n < \epsilon$$

for every $n > N_o$. To recap, we've shown that for every $\epsilon > 0$ there exists $N_o > 0$ such that $|a_n - \bar{v}| < \epsilon$ for all $n > N_o$. This means that $\lim_{n\to\infty} a_n = \bar{\bar{v}}$. So the sequence does have a limit $\bar{a}$, and $\bar{a} = \bar{\bar{v}}$, the least upper bound of the numbers in the sequence. □

Similarly:

3.6 Fact: Every sequence $\{a_n\}$ of real numbers that is *bounded from below* and *monotonically decreasing* — i.e., $a_n \geq a_{n+1}$ for all $n \in \mathbb{N}$ — has a limit $\bar{a}$, and $\bar{a} = \inf(\{a_n\})$. □

Fact 3.5 and Fact 3.3 taken together underpin a powerful criterion for summability of one-and two-sided sequences. To demonstrate that a sequence $\{c_n\}$ is summable, then by Fact 3.3 it suffices to show that $\{c_n\}$ is absolutely summable. But how do we do that? Absent a "candidate limit" for $\sum_n |c_n|$, how can we show the series converges? We'll make frequent use of the following result.

3.7 Fact: Let $\{c_n\}$ be a (one-sided) sequence from $\mathbb{F}$, where $\mathbb{F}$ is $\mathbb{R}$ or $\mathbb{C}$. $\{c_n\}$ is absolutely summable if and only if the sequence $\{\widetilde{s}_n\}$ defined by

$$\widetilde{s}_n = \sum_{m=0}^{n} |c_m|$$

is bounded from above, i.e., if and only if there exists some $R > 0$ such that $\widetilde{s}_n \leq R$ for every $n \geq 0$. If $\{c_n\}$ is a two-sided sequence, $\{c_n\}$ is absolutely summable if and only if the sequence $\{\widetilde{s}_n\}$ defined by

$$\widetilde{s}_n = \sum_{m=-n}^{n} |c_m|$$

is bounded from above.

Proof: Note that in either case $\{\widetilde{s}_n\}$ is a monotonically increasing sequence of real numbers. By Fact 3.5, if it is bounded from above, it converges. Hence if $\{\widetilde{s}_n\}$ is bounded from above, it converges, so $\{c_n\}$ is absolutely summable. Conversely, if $\{c_n\}$ is absolutely summable, then $\{\widetilde{s}_n\}$ converges, so it's bounded from above because it increases monotonically to its limit, which is therefore an upper bound for $\{\widetilde{s}_n\}$. □

I'll leave as exercises proofs of the following two assertions about least upper bounds and greatest lower bounds, which follow swiftly from the definitions.

- If $A \subset \mathbb{R}$ is bounded from above, we can find a sequence $\{a_n\}$ of numbers in A so that
$$\lim_{n\to\infty} a_n = \sup(A) \ .$$
Furthermore, we can choose $\{a_n\}$ to be monotonically increasing.
- If $A \subset \mathbb{R}$ is bounded from below, we can find a sequence $\{a_n\}$ of numbers in A so that
$$\lim_{n\to\infty} a_n = \inf(A) \ .$$

Furthermore, we can choose $\{a_n\}$ to be monotonically decreasing.

Two considerably more sophisticated verities from real-number lore are

- For any $a \in \mathbb{R}$, there exists a sequence $\{q_n\}$ of rational numbers such that
$$\lim_{n\to\infty} q_n = a \ .$$
Furthermore, we can choose the sequence $\{q_n\}$ to be monotonically increasing or decreasing.
- Any $a \in \mathbb{R}$ has the following characterizations:
$$a = \sup(\{q \in \mathbb{Q} : q \le a\})$$
and
$$a = \inf(\{q \in \mathbb{Q} : q \ge a\}) \ .$$

These last two items address approximating real numbers with rational numbers, which is important in applications involving the sort of finite-precision arithmetic one encounters when working with computers interfacing with the "real" world. In a very ... real ... sense, both results hold by construction — i.e., the real numbers are rigged so that both are true automatically. Nonetheless, they're good ones to remember and know how to use.

lim sup and lim inf

Facts 3.4, 3.5, and 3.6 have important consequences for bounded sequences. Suppose $\{a_n\}$ is bounded sequence of real numbers satisfying $|a_n| \le R$ for some $R > 0$ and every $n \in \mathbb{N}$. Consider the associated sequences $\{\bar{a}_n\}$ and $\{\underline{a}_n\}$ with nth terms
$$\bar{a}_n = \sup(\{a_m : m \ge n\}) \ \text{ and } \ \underline{a}_n = \inf(\{a_m : m \ge n\}) \ .$$
Fact 3.4 guarantees the existence of $\bar{a}_n$ and $\underline{a}_n$ for every $n \in \mathbb{N}$. Observe that $\bar{a}_n \ge -R$ and $\underline{a}_n \le R$ for every $n \in \mathbb{N}$. Furthermore, $\{\bar{a}_n\}$ is a monotonically decreasing sequence and $\{\underline{a}_n\}$ is a monotonically increasing sequence. To see this, note that for each n
$$\{a_m : m \ge n+1\} \subset \{a_m : m \ge n\} \ ,$$
so the sup of the set on the right is an upper bound for the smaller set on the left and is therefore at least as large as the left-hand set's sup. Accordingly, $\bar{a}_{n+1} \le \bar{a}_n$ and $\underline{a}_{n+1} \ge \underline{a}_n$ for every $n \in \mathbb{N}$.

Facts 3.5 and 3.6 imply that $\{\bar{a}_n\}$ and $\{\underline{a}_n\}$ both converge. We denote these sequences' limits by
$$\limsup_{n\to\infty} a_n = \lim_{n\to\infty} \bar{a}_n$$
and
$$\liminf_{n\to\infty} a_n = \lim_{n\to\infty} \underline{a}_n \ .$$
People generally call these things the "lim sup" and "lim inf" of the sequence $\{a_n\}$. These quantities exist for any bounded sequence, even a non-convergent one. While every convergent sequence is bounded, not every bounded sequence converges. How can you tell whether a given bounded sequence does converge?

3.8 Fact: A bounded sequence $\{a_n\}$ of real numbers converges to limit $\bar{a}$ if and only if

$$\limsup_{n\to\infty} a_n = \liminf_{n\to\infty} a_n = \bar{a} \ .$$

Proof: First suppose that $\{a_n\}$ converges to limit $\bar{a}$. For any $\epsilon > 0$ you can find $N \in \mathbb{N}$ such that when $n > N$ we have $|a_n - \bar{a}| < \epsilon/2$. This means that all the numbers a_n for $n > N$ lie inside the open interval of width ϵ centered on $\bar{a}$. Accordingly, when $n > N$, $|\sup(\{a_m : m \geq n\}) - \bar{a}| < \epsilon$ and $|\inf(\{a_m : m \geq n\}) - \bar{a}| < \epsilon$. Since ϵ was arbitrary, we conclude that

$$\lim_{n\to\infty}(\sup\{a_m : m \geq n\}) = \limsup_{n\to\infty} a_n = \bar{a}$$

and

$$\lim_{n\to\infty}(\inf\{a_m : m \geq n\}) = \liminf_{n\to\infty} a_n = \bar{a} \ ,$$

Conversely, if $\limsup_{n\to\infty} a_n = \liminf_{n\to\infty} a_n = \bar{a}$, we know that for any $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that all the numbers a_n for $n > N$ lie in the open interval of width ϵ centered on $\bar{a}$. This is the same as saying that $|a_n - \bar{a}| < \epsilon$ for $n > N$. It follows that $\lim_{n\to\infty} a_n = \bar{a}$ since ϵ was arbitrary. □

When applying Fact 3.8 to demonstrate convergence of a sequence $\{a_n\}$, people often prove only that $\limsup a_n \leq \liminf a_n$. The strategy works because $\liminf a_n \leq \limsup a_n$ already, under any circumstances. To see why, just note that $\underline{a}_n \leq \bar{a}_n$ for every n, so the limiting value of $\underline{a}_n$, which is $\liminf a_n$, cannot exceed the limiting value of $\bar{a}_n$, which is $\limsup a_n$.

Equivalence of the two mysterious properties of $\mathbb{R}$

I have emphasized that the real numbers are "rigged" so that every Cauchy sequence converges (Fact 3.2) and so that every set of real numbers bounded from above (or below) has a least upper (or greatest lower) bound (Fact 3.4). It turns out that these facts are logically equivalent in the sense that each one implies the other. Here is a formal statement of the result.

3.9 Theorem: The following two properties of the real numbers are equivalent in the sense that one implies the other:

- Every Cauchy sequence of real numbers converges (Fact 3.2)
- Every set of real numbers bounded from above has a least upper bound and every set of real numbers bounded from below has a greatest lower bound (Fact 3.4).

Proof: First let's assume Fact 3.4 and prove that every Cauchy sequence converges. Keep in mind that Fact 3.4 is enough to prove the existence of lim sup and lim inf of a bounded sequence. If $\{a_n\}$ is a Cauchy sequence, given $\epsilon > 0$ we can find $N \in \mathbb{N}$ such that $|a_n - a_m| < \epsilon/3$ when m and n are both bigger than N. This implies not only that $\{a_n\}$ is a bounded sequence but that some fixed interval of length $\epsilon/3$ contains every a_n for $n > N$. Furthermore, for every m and n bigger than N, we have $|v_n - w_m| \leq \epsilon/3$, where

$$v_n = \sup\left(\{a_k : k \geq n\}\right)$$

and

$$w_m = \inf\left(\{a_k : k \geq m\}\right) \ .$$

Since

$$\lim_{n\to\infty} v_n = \bar{\bar{a}} = \limsup_{n\to\infty} a_n$$

and

$$\lim_{m\to\infty} w_m = \underline{\underline{a}} = \liminf_{n\to\infty} a_n \ ,$$

we can find m and n large enough so, in addition, $|v_n - \bar{\bar{a}}| < \epsilon/3$ and $|w_m - \underline{\underline{a}}| < \epsilon/3$. It follows that

$$|\bar{\bar{a}} - \underline{\underline{a}}| \leq |\bar{\bar{a}} - v_n| + |v_n - w_m| + |w_m - \underline{\underline{a}}| < \epsilon \ .$$

We have shown that $|\bar{\bar{a}} - \underline{\underline{a}}| < \epsilon$ for every $\epsilon > 0$, and the only way that could happen is for $\bar{\bar{a}} = \underline{\underline{a}}$. Fact 3.8 implies in turn that $\{a_n\}$ converges.

Now for the converse. Assume that every Cauchy sequence converges. Let A be a set of real numbers bounded from above; I'll show that A has a least upper hound. (The argument for the existence of $\inf(A)$ when A is bounded from below proceeds similarly.) First let $\bar{v}$ be an upper bound for A. If $\bar{v}$ is the least upper bound for A, then we're done. If not, there exists some largest $n_0 \in \mathbb{N}$ for which $\bar{v}_0 = \bar{v} - n_0$ is an upper bound for A but $\bar{v} - (n_0 + 1)$ is not. If $\bar{v}_0$ is the least upper bound for A, then we're done. If not, there exists some smallest $n_1 \in \mathbb{N}$ such that

$$\bar{v}_1 = \bar{v}_0 - 2^{-n_1}$$

is an upper bound for A. Note that $n_1 \geq 1$. If $\bar{v}_1$ is the least upper bound for A, then we're done. If not, then there exists some smallest $n_2 \in \mathbb{N}$ such that

$$\bar{v}_2 = \bar{v}_1 - 2^{-n_2}$$

is an upper bound for A. Note that $\bar{v}_0 - 2^{-n_2}$ is also an upper bound for A, so $n_2 \geq n_1$ by definition of n_1. In fact, $n_2 > n_1$, since if $n_2 = n_1$, we would have $\bar{v}_2 = \bar{v}_0 - 2^{-(n_1-1)}$, again contradicting the definition of n_1. Observe that $n_2 \geq 2$ because $n_1 \geq 1$. If $\bar{v}_2$ is the least upper bound for A, then we're done; otherwise there exists some smallest $n_3 \in \mathbb{N}$ such that

$$\bar{v}_2 = \bar{v}_2 - 2^{-n_3}$$

is an upper bound for A. It is easy to show that $n_3 > n_2$ and hence that $n_3 \geq 3$ in the same way we showed that $n_2 > n_1$.

Continuing in this fashion, we get a sequence of upper bounds

$$\bar{v}_m = \bar{v}_0 - \sum_{k=1}^{m} 2^{-n_k}$$

of upper bounds for A where $n_k \geq k$ for all k. If any of the $\bar{v}_m$ is the least upper bound for A, then we're done. Otherwise, the sequence $\{\bar{v}_m\}$ is an infinite sequence of upper bounds for A. It is a Cauchy sequence because, assuming $l > m > N$,

$$|\bar{v}_m - \bar{v}_l| = \sum_{k=m+1}^{l} 2^{-n_k} < \sum_{k=N}^{\infty} 2^{-k} = 2^{-N+1} ,$$

which we can make as small as we want by choosing N large enough. Note that I used the fact that $n_k \geq k$ for all k. So $\{\bar{v}_m\}$ converges to a limit $\bar{\bar{v}}$, which I claim is the least upper bound for A. $\bar{\bar{v}}$ is an upper bound for A because every $\bar{v}_m$ is. If $\bar{\bar{v}}$ weren't the least upper bound for A, we could find $m \in \mathbb{N}$ such that $\bar{\bar{v}} - 2^{-m}$ is an upper bound for A. This would imply that $\bar{v}_{m+1} - 2^{-m}$ is also an upper bound for A, which is impossible by construction of n_{m+1}, taking into account that $n_{m+1} \geq m + 1$. □

CHAPTER 4

Linear Algebra I

You've been exposed to linear algebra on some level, but I'm uncomfortable with the approach that many "math for engineers" linear algebra courses take to the subject. In my experience they tend to zigzag awkwardly between talk about abstract vector spaces, linear mappings, and bases on the one hand and matrices, linear equations, and $\mathbb{F}^n$ on the other. The abstract theory is beautiful and, if packaged well, easy to understand. Some recent books, for example Sheldon Axler's excellent if audaciously titled *Linear Algebra Done Right,* present treatments in the same spirit as the one I offer here, although far more comprehensive. Axler goes through some minor contortions to avoid any non-essential invocation of matrices and thereby achieves a pristinely "coordinate-free" exposition. In what follows, I'll be somewhat less uncompromising, but not much less.

Vector spaces

As usual, $\mathbb{F}$ is $\mathbb{R}$ or $\mathbb{C}$. A *vector space over* $\mathbb{F}$ is a set V on which are defined two operations. The *addition* operation takes any v and w in V and produces another element $v + w$ of V. The *scalar multiplication* operation takes any $c_o \in \mathbb{F}$ and $v \in V$ and produces another element $c_o v$ of V. We call the elements of V *vectors.* Addition is commutative and associative, and V contains an identity element 0 — the *zero vector* — for addition. Scalar multiplication distributes over addition and also has these three properties: $0v = 0$ for every $v \in$V; $1v = v$ for every $v \in V$; and $c_1(c_2v) = (c_1c_2)v$ for every c_1, $c_2 \in \mathbb{F}$ and every $v \in V$. I'll always write $-v$ for $(-1)v$. I'll also use the symbol 0 to denote both $0 \in \mathbb{F}$ and $0 \in V$ and hope that the context makes things clear. For example, in the equation $0v = 0$ the left-hand zero has to be in $\mathbb{F}$ and the right-hand zero has to be in V.

Here are a few examples of vector spaces. Let 0 be the number 0 in $\mathbb{F}$ and set $V = \{0\}$. Then V is a vector space over $\mathbb{F}$ with addition and scalar multiplication defined as ordinary addition and multiplication in $\mathbb{F}$. A less trivial example is $\mathbb{F}^n$, the set of all column n-vectors with entries from $\mathbb{F}$. Addition and scalar multiplication on $\mathbb{F}^n$ are defined entry-wise. Observe that $\mathbb{C}^n$ is also a vector space over $\mathbb{R}$ with addition and scalar multiplication defined in the usual way. Finally, let $\mathbb{F}^{\mathbb{Z}}$ be the set of all two-sided infinite sequences with elements from $\mathbb{F}$. A typical $v \in \mathbb{F}^{\mathbb{Z}}$ looks like

$$\ldots\ ,\ \alpha_{-2}\ ,\ \alpha_{-1}\ ,\ \alpha_0\ ,\ \alpha_1\ ,\ \alpha_2\ ,\ \alpha_3\ ,\ \ldots\ ,$$

where $\alpha_m \in \mathbb{F}$ for all $m \in \mathbb{Z}$. If $c_o \in \mathbb{F}$, then $c_o v$ is

$$\ldots\ ,\ c_o\alpha_{-2}\ ,\ c_o\alpha_{-1}\ ,\ c_o\alpha_0\ ,\ c_o\alpha_1\ ,\ c_o\alpha_2\ ,\ c_o\alpha_3\ ,\ \ldots\ .$$

If $w \in \mathbb{F}^{\mathbb{Z}}$ looks like v above except with β's instead of α's, then $v + w$ is

$$\ldots\ ,\ \alpha_{-2} + \beta_{-2}\ ,\ \alpha_{-1} + \beta_{-1}\ ,\ \alpha_0 + \beta_0\ ,\ \alpha_1 + \beta_1\ ,\ \alpha_2 + \beta_2\ ,\ \alpha_3 + \beta_3\ ,\ \ldots\ .$$

The vector space $\mathbb{F}^{\mathbb{Z}}$ will feature prominently in our discussion of discrete-time signals and systems.

If V is a vector space over $\mathbb{F}$, a *linear combination* of vectors in V is an expression of the form

$$c_1 v_1 + c_2 v_2 + \cdots + c_m v_m\ ,$$

where $c_j \in \mathbb{F}$ and $v_j \in V$ for $1 \leq j \leq m$. Any linear combination, of course, specifies a vector in V. A *subspace* of V is a subset W of V with the following property: every linear combination of vectors in W is also a vector in W. In other words, W is a subspace of V if and only if W is closed under the taking of linear combinations. Saying that W is a subspace of V is the same as saying that W is itself a vector space with the same vector operations that make V a vector space. Note that every subspace W contains the zero vector because W is closed under scalar multiplication, in particular multiplication by the scalar 0.

The subsets $\{0\}$ and V of any vector space V are subspaces of V. A more interesting example is the subspace W of $\mathbb{R}^n$ consisting of all $v \in \mathbb{R}^n$ that satisfy $c^T v = 0$ for some $c \in \mathbb{R}^n$. Please check for yourself that this W, known as a *hyperplane* in $\mathbb{R}^n$, is indeed closed under the taking of linear combinations (it's easy). If $n = 3$ and you think of vectors in $\mathbb{R}^3$ as little arrows emanating from the origin, the hyperplane W is the set of arrows perpendicular to the arrow c. The word hyperplane describes a lot of other subsets of $\mathbb{R}^n$ that aren't subspaces. For example, given $c \in \mathbb{R}^n$, the set H of all $v \in \mathbb{R}^n$ satisfying $c^T v = 17$ is also a hyperplane in $\mathbb{R}^n$ but is not a subspace. For one thing, H is missing the zero vector, and you can verify readily that H is not closed under addition.

Every vector space V other than $\{0\}$ is uncountably infinite. To see why, let v be any nonzero vector in V and note that all the vectors in the uncountably infinite set $\{c_o v : c_o \in \mathbb{F}\}$ are different. On the other hand, intuition tells us that $\mathbb{F}^3$ is somehow bigger or richer than $\mathbb{F}^2$ and that $\mathbb{F}^{\mathbb{Z}}$ is richer than $\mathbb{F}^n$ for any n. Making "size comparisons" between vector spaces requires a measuring device more refined than cardinality, and developing such a device is our next mission.

Spanning sets, finite dimensionality, and linear independence

If S is a subset of a vector space V, the *span of* S, which I'll denote by span(S), is the set of all linear combinations of vectors in S. span(S) is a subspace of V because any linear combination of linear combinations of vectors in V is another linear combination of vectors in V and therefore in span(S). If span(S) $= V$, we say that *S is a spanning set for V* or that *S spans V*. Observe that V itself is a spanning set for V.

A spanning set S for a vector space V, at least potentially, supplies a stripped-down characterization of V, namely as the set of all linear combinations of vectors in S. The smaller the spanning set, the sleeker the characterization — in particular, V itself as a spanning set for V is not helpful in this regard. But consider the extreme case where every vector in V is a scalar multiple of some single vector $v_o \in V$. Then we can describe V succinctly as "the set of all scalar multiples of v_o." In this case, $\{v_o\}$ is a spanning set for V.

Every vector space other than $\{0\}$ has many infinite spanning sets, but only some vector spaces have finite spanning sets. A vector space V with at least one finite spanning set is called *finite-dimensional.* Succinctly characterizing a finite-dimensional V entails finding spanning sets for V that are as small as possible. I'll now describe two procedures aimed at constructing small spanning sets for a finite-dimensional vector space V. I'll assume throughout that $V \neq \{0\}$, since that vector space's only spanning set is $\{0\}$.

Procedure 1: If V is finite-dimensional and $S = \{v_1, \ldots, v_m\}$ is a spanning set for V, you might be able to construct a smaller spanning set by removing unnecessary vectors from S. Suppose, for example, that one of the vectors in S — say v_m — can be written as a linear combination of the others. If you eliminate v_m from S, the resulting smaller set still spans V because any linear combination of all the vectors can be re-written as a linear combination of $v_1, \ldots, v_{m-1}$ by substituting for every appearance of v_m its expression as a linear combination of the other $m-1$ vectors. If none of the vectors in S is a linear combination of the others, the set $\widehat{S}$ you obtain by removing any vector from S will no longer span V — in particular, the vector you've removed won't be in span$(\widehat{S})$. □

Saying that one vector in S can be written as a linear combination of the others is the same as saying that the vectors in S are *linearly dependent* in the sense that there exist $c_1, \ldots, c_m$ in $\mathbb{F}$ at least one of which is nonzero that satisfy

$$c_1v_1 + c_2v_2 + \cdots + c_mv_m = 0 .$$

Expressing one vector in S as a linear combination of the others leads directly to such a relation. Conversely, if such a relation holds, you can divide out a nonzero c_k and solve for the corresponding v_k as a linear combination of the other vectors. We call a set of vectors *linearly independent* when it is not linearly dependent. I'll use interchangeably the terminologies "$v_1, \ldots, v_m$ are linearly (in)dependent" and "$S = \{v_1, \ldots, v_m\}$ is a linearly (in)dependent set."

From Procedure 1 we can draw three significant conclusions. First, if S is a spanning set for V and the vectors in S are linearly dependent, then you can generate a smaller spanning set for V by removing a vector from S. Second, if the vectors in S are linearly independent, you can't remove a vector from S and have a spanning set left over. Furthermore, since a linearly dependent spanning set S can be reduced by one vector, you can start with such an S, remove an unnecessary vector, check the resulting smaller spanning set for linear dependence, remove another vector if necessary, and so on. The process can't go on forever since S is finite. Eventually you'll end up with a linearly independent spanning set for V. Behold our third and most important conclusion: every finite-dimensional vector space has a linearly independent spanning set.

Procedure 2: You could also consider trying to build a linearly independent spanning set for V from the ground up. Start with any nonzero $v_1 \in V$ and form

$\{v_1\}$. If that set spans V, you're done. If not, find $v_2 \in V$ not in span($\{v_1\}$). Then $\{v_1, v_2\}$ will be a linearly independent set. If this set spans V, you're done. If not, find v_3 outside span($\{v_1, v_2\}$), etc. All the sets you construct using this procedure are linearly independent as the following inductive argument shows. Certainly $\{v_1\}$ is a linearly independent set. Suppose now that $S_k = \{v_1, \ldots, v_k\}$ is linearly independent and that v_{k+1} is not in span(S_k). If $S_{k+1} = \{v_1, \ldots, v_{k+1}\}$ were linearly dependent, we could write

$$c_1v_1 + c_2v_2 + \cdots c_kv_k + c_{k+1}v_{k+1} = 0$$

where at least one coefficients is nonzero. We can't have $c_{k+1} = 0$ because that would contradict linear independence of S_k. Accordingly, we can divide out by c_{k+1} and solve for v_{k+1} as a linear combination of the other vectors, which contradicts $v_{k+1} \notin \text{span}(S_k)$. It follows that $S_{k+1} = \{v_1, \ldots, v_{k+1}\}$ is a linearly independent set. With any luck, S_k will be a spanning set for V for k sufficiently large. □

All optimism aside, what we've established so far does not guarantee that Procedure 2 will lead eventually to a linearly independent spanning set for V. For example, might the procedure fail to terminate? Might some infelicitous choice of v_1 or unfortunate selection of v_{k+1} somewhere along the way lead to ever larger linearly independent subsets of V that don't span V? More benignly, might some starting vector v_1 or selection routine for v_{k+1} give rise to linearly independent spanning sets larger than those arising from other starting vectors and selection routines? A similar question dogs Procedure 1, which we know terminates in a linearly independent spanning set for V. Might some fortuitous initial spanning set S or some clever ordering of vector removals lead to smaller linearly independent spanning sets than other choices of S and sequencings of vector removals? The answer to all these questions is No. Proving that statement requires some additional work.

Bases and dimension

The next result marks my only appeal to the "mundane" language of linear equations. I could dodge it for the sake of purity, but it's so fundamental and so handy to know that I feel compelled to include it. It crystallizes the storied mantra that an underdetermined set of homogeneous linear equations has a nontrivial solution.

4.1 Lemma: Let $\mathbb{F}$ be $\mathbb{R}$ or $\mathbb{C}$. If $n > m$, the system of equations

$$\begin{aligned}
c_{11}x_1 + c_{12}x_2 + c_{13}x_3 + \cdots + c_{1n}x_n &= 0 \\
c_{21}x_1 + c_{22}x_2 + c_{23}x_3 + \cdots + c_{2n}x_n &= 0 \\
c_{31}x_1 + c_{32}x_2 + c_{33}x_3 + \cdots + c_{3n}x_n &= 0 \\
\cdots &= \cdots \\
\cdots &= \cdots \\
c_{m1}x_1 + c_{m2}x_2 + c_{m3}x_3 + \cdots + c_{mn}x_n &= 0\,,
\end{aligned}$$

where $c_{ij} \in \mathbb{F}$ for all $1 \le i \le m$ and $1 \le j \le n$, has a solution $x_1, x_2, \ldots, x_n$ where at least one x_j, $1 \le j \le n$, is nonzero.

Proof: Fix $k > 0$ and suppose $n = m + k$. The proof proceeds by induction on m. If $m = 1$, you have one equation in $k + 1$ unknowns. If $c_{11} = 0$, set $x_1 = 1$ and $x_j = 0$ for all $j > 1$ and you have a nonzero solution to the equation. If $c_{11} \neq 0$, set $x_2 = 1$, $x_1 = -c_{12}/c_{11}$, and $x_j = 0$ for all other j and you have a nonzero solution. That takes care of the case $m = 1$.

Now suppose we've proven the result for m equations in $m + k$ unknowns. Consider the case of $m + 1$ equations in $n = m + 1 + k$ unknowns. If all the x_1-coefficients — i.e. all the c_{i1} for $1 \le i \le m+1$ — are zero, setting $x_1 = 1$ and $x_j = 0$ for $j > 1$ yields a nonzero solution to the system of equations. If not all the c_{i1} are zero, re-order the equations if necessary so $c_{11} \neq 0$ and replace the $m + 1$ equations with the equivalent set of equations you obtain by leaving the first equation alone and replacing each other equation i with

$$(\text{equation } i) \; - \; (c_{i1}/c_{11}) \times \; (\text{equation } 1) \; .$$

The new set of equations is equivalent to the old one because the equation-replacement procedure is reversible, so any solution to the new set of equations is also a solution to the original set.

None of the equations 2 through $m + 1$ in the new set has an x_1-term. Accordingly, these equations constitute a set of m equations in the $n - 1 = m + k$ unknowns $x_2, x_3, \ldots, x_n$. By the induction assumption, these equations have a solution d_j, $2 \le j \le m + k + 1$, where not all of the d_j are zero. Finish the job by solving for x_1 from the first equation by dividing out c_{11}, i.e.

$$x_1 = -\left(1/c_{11}\right)\left(c_{12}d_2 + c_{13}d_3 + \cdots + c_{1n}d_n\right) \; .$$

So the conclusion of lemma is true for $m = 1$ and is true for $m + 1$ when it is true for m, and by induction it is therefore true for all m. □

Lemma 4.1 makes it easy to prove the following central result.

4.2 Lemma: Let V be a finite-dimensional vector space over $\mathbb{F}$. If $S = \{v_1, v_2, \ldots, v_m\}$ is a spanning set for V and $\{w_1, w_2, \ldots, w_n\}$ is a linearly independent set, then $n \le m$.

Proof: Since S spans V, we can find c_{ij}, $1 \le i \le m$ and $1 \le j \le n$, such that

$$w_j = \sum_{i=1}^{m} c_{ij} v_i \; \text{ for } \; 1 \le j \le n \; .$$

Suppose now that $n > m$, contrary to what we want to prove. By Lemma 4.1 we can find d_j, $1 \le j \le n$, not all zero that

$$\sum_{j=1}^{n} c_{ij} d_j = 0 \; \text{ for } \; 1 \le i \le m \; .$$

It follows that

$$\sum_{j=1}^{n} d_j w_j = \sum_{j=1}^{n} d_j \left(\sum_{i=1}^{m} c_{ij} v_i \right) = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} c_{ij} d_j \right) v_i = 0 \ ,$$

which contradicts linear independence of the w_j. We must conclude that the starting assumption $n > m$ was incorrect, so $n \leq m$. □

Lemma 4.2 is a key that unlocks many doors. We saw earlier that any finite-dimensional vector space $V \neq \{0\}$ has a linearly independent spanning set, but we wondered whether linearly independent spanning sets of different sizes might exist. If S_1 and S_2 are two linearly independent spanning sets, then Lemma 4.2 tells us that

- because S_1 is a spanning set and S_2 is a linearly independent set, S_1 contains at least as many vectors as S_2, and
- because S_2 is a spanning set and S_1 is a linearly independent set, S_2 contains at least as many vectors as S_1.

Conclusion: any two linearly independent spanning sets for V contain the same number of vectors. That number is called the *dimension* of V, and I'll often denote it by $\dim(V)$.

Carrying on, we can deduce that linearly independent subsets of a finite-dimensional vector space V can get only so large. If V has dimension n, then V has a spanning set containing n vectors, so by Lemma 4.2 no linearly independent subset of V can contain more than n vectors. As a consequence, a vector space V that has arbitrarily large linearly independent sets must be infinite-dimensional.

Furthermore, spanning sets of a finite-dimensional V can get only so small. If V has dimension n, then V has a linearly independent subset (a spanning set, in fact) containing n vectors, so by Lemma 4.2 no spanning set for V can contain fewer than n vectors. Thus the dimension of V is both the largest possible size of a linearly independent set and the smallest possible size of a spanning set.

A *basis* for an n-dimensional vector space V is an ordered n-tuple $(v_1, v_2, \ldots, v_n)$ of vectors in V with the property that $\{v_1, v_2, \ldots, v_n\}$ is a linearly independent spanning set for V. The fact that bases are ordered distinguishes them from linearly independent spanning sets. For example, if (v_1, v_2) is a basis for a 2-dimensional V, then (v_2, v_1) is technically a different basis. In any event, "basis" is almost synonymous with "linearly independent spanning set."

Observe that if $(v_1, v_2, \ldots, v_n)$ is a basis for V, then every $v \in V$ can be written as a linear combination of the v_j because $\{v_1, \ldots, v_n\}$ spans V. What's more, v's expansion in terms of the basis vectors is unique. If v had two such expansions, then subtracting one from the other would lead to a linear-dependence relation between the vectors in the basis, which is impossible since the vectors are linearly independent.

The following result, which we have proven over the course of the foregoing discussion, anchors the theory of finite-dimensional vector spaces.

4.3 Theorem: If $V \neq \{0\}$ is a finite-dimensional vector space over $\mathbb{F}$, then V has a basis. Any two bases for V contain the same number of vectors. That number is called the dimension of V. If V has dimension n, then no linearly independent subset of V contains more than n vectors and no spanning set for V contains fewer than n vectors. If $(v_1, \dots, v_n)$ is a basis for V, then every $v \in V$ can be written in exactly one way as a linear combination of $v_1, \dots, v_n$. □

Procedure 2, which I described earlier, was targeted at building a linearly independent spanning set for a finite-dimensional V by starting with $\{v_1\}$ and adding vectors while maintaining linear independence of the vectors at each step. It was not obvious *a priori* that Procedure 2 would necessarily terminate. Now, in the light of Theorem 4.3, we know that the procedure does indeed terminate in a linearly independent spanning set for V. If V has dimension n, then once we've arrived via Procedure 2 at a linearly independent set $S_n = \{v_1, v_2, \dots, v_n\}$, we know that S_n must span V or else we could find a linearly independent subset of V with $n+1$ vectors in it, contradicting $\dim(V) = n$. This "basis construction" procedure is important and I'll make frequent use of it.

4.4 Theorem: If $\{v_1, v_2, \dots, v_k\}$ is a linearly independent subset of an n-dimensional vector space V and $k < n$, we can find $n-k$ vectors $v_{k+1}, \dots, v_n$ so that $(v_1, v_2, \dots, v_n)$ is a basis for V.

Proof: We've really pretty much proven this already, but here's what's going on in a nutshell. Since $k < n$, the set can't span V by Theorem 4.3, and we can therefore add vectors to the set one at a time using Procedure 2, all the while maintaining linear independence. Eventually we will reach a linearly independent spanning set for V, which must contain n vectors by Theorem 4.3. Ordering the vectors in this last set yields a basis for V. □

As you probably know, the dimension of $\mathbb{F}^n$ is n. To prove this formally, let $e^i \in \mathbb{F}^n$ be the vector with ith element 1 and all other elements zero. I'll show that $(e^1, \dots, e^n)$ is a basis for $\mathbb{F}^n$. First, for $v \in \mathbb{F}^n$, let $[v]_i$ be the ith entry in v. Note that

$$v = \sum_{i=1}^{n} [v]_i e^i .$$

Since v is an arbitrary vector in $\mathbb{F}^n$, $\{e^1, \dots, e^n\}$ spans $\mathbb{F}^n$. Furthermore, for any $c_1, \dots, c_n$ in $\mathbb{F}$, the linear combination

$$c_1 e^1 + \cdots + c_n e^n$$

has ith entry c_i, and hence is zero if only if $c_i = 0$ for all i. It follows that $\{e^1, \dots, e^n\}$ is a linearly independent set as well as a spanning set for $\mathbb{F}^n$, and $(e^1, \dots, e^n)$ is therefore a basis for $\mathbb{F}^n$.

What about $\mathbb{F}^{\mathbb{Z}}$? For each $i \in \mathbb{Z}$, let e^i be the sequence with a 1 in the ith position and a zero in every other position. Given $n > 0$ along with $c_1, \dots c_n$ in $\mathbb{F}$, the sequence

$$v = c_1 e^1 + c_2 e^2 + \cdots + c_n e^n$$

has zeroes in every position i save $1 \leq i \leq n$, and c_i appears in position i for i in that range. So $v = 0$ if and only if $c_i = 0$ for $1 \leq i \leq n$, proving that $\{e^1, \ldots, e^n\}$ is a linearly independent set. It follows that for every $n > 0$ $\mathbb{F}^{\mathbb{Z}}$ has a linearly independent subset containing n vectors and, by Theorem 4.3, $\mathbb{F}^{\mathbb{Z}}$ is not finite-dimensional. $\mathbb{F}^{\mathbb{Z}}$ is arguably the archetypal infinite-dimensional vector space.

If W is a subspace of an n-dimensional vector space V, then W is also finite-dimensional. If W weren't finite-dimensional, then we could build arbitrarily large linearly independent subsets of W, one vector at a time. These would also be linearly independent subsets of V, and some would contain more than $\dim(V)$ vectors, and we know such sets can't exist because of Theorem 4.3. Since W is finite-dimensional, if $W \neq \{0\}$ it has a basis $(w_1, \ldots, w_k)$, and $k \leq \dim(V)$ by Theorem 4.3. It follows $\dim(W) \leq \dim(V)$.

Finally, Theorem 4.3 implies that the only n-dimensional subspace of an n-dimensional vector space V is V itself. Suppose $(w_1, \ldots, w_n)$ is a basis for such a subspace W. If $W \neq V$, we can find $w_{n+1} \in V$ not expressible as a linear combination of the other w_j, making $\{w_1, \ldots, w_n, w_{n+1}\}$ an impossibly large linearly independent subset of V. Accordingly, $W = V$.

Vector sums and disjoint subspaces

The *vector sum* of a collection $W_1, \ldots, W_k$ of subspaces of a vector space V is the subspace of V defined by

$$W_1 + \cdots + W_k = \operatorname{span}(W_1 \cup \cdots \cup W_k) .$$

This is one of several equivalent ways to define the vector sum of the W_j. Another is

$$W_1 + \cdots + W_k = \{w_1 + \cdots + w_k : w_j \in W_j \text{ for } 1 \leq j \leq k\} .$$

The definitions are equivalent because any vector in the span of the union of the W_j, being a linear combination of vectors from the W_j, can be parsed as the sum of k vectors, one from each subspace W_j. Conversely, any such k-fold sum must certainly lie in the span of the union of the W_j.

For each j,

$$W_j \subset W_1 + \cdots + W_k ,$$

so every W_j is a subspace of the vector sum of the W_j. The vector sum is in fact the smallest subspace of V containing all the W_j. These containment relations imply that the dimension of the vector sum of the W_j is at least as large as the dimensions of all the W_j when all the subspaces are finite-dimensional. Typically, the dimension of the vector sum exceeds all the dimensions of the W_j. More on that in a moment.

The vector sum provides a means for assembling larger subspaces of V from smaller ones. Simply taking the union of subspaces won't do. For example, if v_1 and v_2 are nonzero linearly independent vectors in V and we set $W_1 = \operatorname{span}(\{v_1\})$ and $W_2 = \operatorname{span}(\{v_2\})$, then $W_1 \cup W_2$ is not a subspace of V because it does not contain $v_1 + v_2$. On the other hand, the intersection of any collection of subspaces of V is a subspace of V. This is because any linear combination of vectors in $W_1 \cap \cdots \cap W_k$ is a linear combination of vectors in W_j for every j and hence a member of W_j for every j and hence a member of the intersection of the W_j.

Since every subspace of V contains the zero vector, the intersection of any collection of subspaces of V is nonempty. If $W_1 \cap W_2 = \{0\}$, meaning that the intersection of W_1 and W_2 is as empty as possible, we call W_1 and W_2 *disjoint* subspaces of V. More generally, if $k > 2$ and $W_1, \ldots , W_k$ are subspaces of V, we say that the W_j are *mutually disjoint* when each W_j is disjoint from the vector sum of the other $k-1$ subspaces. Mutual disjointness is a condition far more restrictive than pairwise disjointness, which requires only that W_i and W_j be disjoint for any $i \neq j$. For example, if V is 2-dimensional with basis (v_1, v_2), the three subspaces $\operatorname{span}(\{v_1\})$, $\operatorname{span}(\{v_2\})$, and $\operatorname{span}(\{v_1 + v_2\})$ are pairwise disjoint but not mutually disjoint because the vector sum of any two of them is the entire vector space V. Here is a convenient criterion for mutual disjointness.

4.5 Lemma: Subspaces $W_1, \ldots , W_k$ of a vector space V over $\mathbb{F}$ are mutually disjoint if and only if the relation

$$w_1 + \cdots + w_k = 0$$

with $w_j \in W_j$ for all j holds only when $w_j = 0$ for all j.

Proof: Note first that if the subspaces aren't mutually disjoint, then there exists some i and some vector $w \neq 0$ lying in both W_i and the vector sum of the other subspaces. For convenience, suppose $i = 1$. Setting $w_1 = w$ and $x = -w$ yields a relation $w_1 + x = 0$ featuring nonzero vectors w_1 in W_1 and x in the vector sum $W_2 + \cdots + W_k$. We can write x as $w_2 + \cdots + w_k$ with $w_j \in W_j$ for $2 \leq j \leq k$, yielding a relation

$$w_1 + w_2 + \cdots + w_k = 0$$

between vectors some of which are nonzero. Conversely, if such a relation holds, assuming without loss of generality that $w_1 \neq 0$, we have

$$w_1 = -w_2 - \cdots - w_k \ ,$$

and the single nonzero vector represented differently by the two sides of this last equation must lie in both W_1 and in $W_2 + \cdots + W_k$, so the subspaces aren't mutually disjoint. □

Suppose that W_j, $1 \leq j \leq k$, are finite-dimensional subspaces of a vector space V. Let W_j have dimension d_j and suppose you've chosen a basis for each W_j. If you form a set S of vectors by merging the vectors in all these bases, the set S will contain $\sum_{j=1}^{k} d_j$ vectors that span $W_1 + \cdots + W_k$. Accordingly, by Theorem 4.3,

$$\dim(W_1 + \cdots + W_k) \leq d_1 + \cdots + d_k = \dim(W_1) + \cdots + \dim(W_k) \ .$$

As it happens, this inequality holds with equality if and only if the W_j are mutually disjoint.

4.6 Theorem: If $W_1, \ldots , W_k$ be finite-dimensional subspaces of a vector space V over $\mathbb{F}$, then

$$\dim(W_1 + \cdots + W_k) \leq \dim(W_1) + \cdots + \dim(W_k)$$

with equality if and only if the W_j are mutually disjoint.

Proof: I've demonstrated already that the inequality always holds. Choose for each W_j a basis $(w_1^j, \dots, w_{d_j}^j)$, where $d_j = \dim(W_j)$. Merging all these bases together yields a set S containing $d_1 + \cdots + d_k$ vectors that span $W_1 + \cdots + W_k$. If the W_j are mutually disjoint, then S is a linearly independent set. To see why, suppose some linear combination of the vectors in S is zero. We can re-write the linear combination in the form

$$w_1 + w_2 + \cdots + w_k$$

where w_j is for each j a linear combination of the chosen basis vectors for W_j. By mutual disjointness, $w_j = 0$ for all j. Since the basis vectors for W_j are linearly independent, $w_j = 0$ implies that all the coefficients in the linear combination yielding w_j are zero. Since this last assertion holds for all j, all the coefficients in the original linear combination of vectors in S are zero. It follows that S is a linearly independent spanning set for $W_1 + \cdots + W_k$, which consequently has dimension $d_1 + \cdots + d_k$.

Conversely, if the W_j are not mutually disjoint, then by Lemma 4.5 we can find a relation the form

$$w_1 + w_2 + \cdots + w_k = 0$$

where $w_j \in W_j$ for all j with at least one w_j nonzero. We can write each w_j as a linear combination of the chosen basis vectors for W_j, so the relation yields a nontrivial linear combination of the vectors in S totaling zero, implying that S is a linearly dependent set. Since S is a linearly dependent spanning set for the vector sum $W_1 + \cdots + W_k$, the dimension of the vector sum must be lower than the number of vectors in S, which is $d_1 + \cdots d_k$. □

A sharper result holds when $k = 2$. If W_1 and W_2 are disjoint, then

$$\dim(W_1 + W_2) = \dim(W_1) + \dim(W_2)$$

by Theorem 4.6. If W_1 and W_2 are not disjoint, then

$$\dim(W_1 + W_2) = \dim(W_1) + \dim(W_2) - \dim(W_1 \cap W_2)\ .$$

To prove this identity, let $(v_1, \dots, v_d)$ is a basis for $W_1 \cap W_2$. Use Theorem 4.4 to generate a basis $(v_1, \dots, v_d, w_1^1, \dots, w_{m-d}^1)$ for W_1 and a basis $(v_1, \dots, v_d, w_1^2, \dots, w_{n-d}^2)$ for W_2. Then

$$S = \{w_1^1, \dots, w_{n-d}^1, v_1, \dots, v_d, w_1^2, \dots, w_{n-d}^2\}$$

spans $W_2 + W_2$ because it contains spanning sets for both W_j. S is also linearly independent. A relation

$$c_1^1 w_1^1 + \cdots + c_{m-d}^1 w_{m-d}^1 + b_1 v_1 + \cdots + b_d v_d + c_1^2 w_1^2 + \cdots + c_{n-d}^2 w_{m-d}^1 = 0$$

yields

$$c_1^1 w_1^1 + \cdots + c_{m-d}^1 w_{m-d}^1 + b_1 v_1 + \cdots + b_d v_d = -c_1^2 w_1^2 - \cdots - c_{n-d}^2 w_{m-d}^1\ .$$

The left-hand side lies in W_1 and the right-hand side lies in W_2, so the right-hand side must lie in $W_1 \cap W_2$ and hence be representable as a linear combination of v_1,

$\ldots$, v_d which is impossible because the v_j and the w_i^2 are linearly independent. It follows that S is a linearly independent spanning set for $W_1 + W_2$, so

$$(w_1^1, \ldots , w_{m-d}^1, v_1, \ldots , v_d, w_1^2, \ldots , w_{n-d}^2)$$

is a basis for $W_1 + W_2$, and $W_1 + W_2$ therefore has dimension

$$(m - d) + d + (n - d) = \dim(W_1) + \dim(W_2) - \dim(W_1 \cap W_2)$$

by Theorem 4.3.

Linear mappings, range, and nullspace

If V and W are vector spaces over $\mathbb{F}$, a mapping $T : V \to W$ is *linear* when

$$T(v_1 + v_2) = T(v_1) + T(v_2)$$

for every v_1 and v_2 in V, and

$$T(c_o v) = c_o T(v)$$

for every $v \in V$ and $c_o \in \mathbb{F}$. From this definition it follows that a linear mapping T satisfies $T(0) = 0$ and respects arbitrary linear combinations in the sense that

$$T(c_1 v_1 + \cdots + c_k v_k) = c_1 T(v_1) + \cdots + c_k T(v_k)$$

for every $k > 0$, every c_1, $\ldots$, c_k in $\mathbb{F}$, and every v_1, $\ldots$, v_k in V. The simplest linear mapping from V to W is the zero mapping, which sends every $v \in V$ to $0 \in W$. If $W = V$, the identity mapping, which sends every $v \in V$ to v itself, is another particularly simple example of a linear mapping from V to W.

Like all mappings, linear mappings can be composed to yield other mappings. What's special about linear mappings is that their compositions are also linear. If $T : V \to W$ and $S : W \to X$ are linear mappings between vector spaces over $\mathbb{F}$, the composition ST is the mapping from V to X defined by

$$ST(v) = S(T(v)) \text{ for all } v \in V \ .$$

It is easy to show that ST is indeed linear. If T is a linear mapping from V to V, we can compose T with itself repeatedly and obtain the linear mapping $T^k : V \to V$ defined each every $k > 0$ as the k-fold composition of T with itself. By convention, T^0 is the identity mapping.

Associated with any linear mapping $T : V \to W$ are two important subspaces.

- The *range* of T which I'll denote by range(T), is the set of all $w \in W$ such that $T(v) = w$ for some $v \in V$. Observe that if w_1, $\ldots$, w_k are in range(T) and $v_j \in V$ satisfies $T(v_j) = w_j$ for $1 \le j \le k$, then

$$\begin{aligned} T(c_1 v_1 + \cdots c_k v_k) &= c_1 T(v_1) + \cdots + c_k T(v_k) \\ &= c_1 w_1 + \cdots c_k w_k \ , \end{aligned}$$

  so $c_1 w_1 + \cdots + c_k w_k$ is also in range(T) for every c_1, $\ldots$, c_k in $\mathbb{F}$. So range(T) is indeed a subspace of W because it is closed under the taking of linear combinations.
- The *nullspace* of T, which I'll denote by nullspace(T), is the set of all $v \in V$ such that $T(v) = 0$. Note that if v_1, $\ldots$, v_k are in nullspace(T), then

$$T(c_1 v_1 + \cdots + c_k v_k) = c_1 T(v_1) + \cdots + c_k T(v_k) = 0 \ ,$$

so $c_1v_1 + \cdots + c_kv_k$ is also in nullspace(T) for every $c_1, \ldots, c_k$ in $\mathbb{F}$. The nullspace of T is therefore a subspace of V because it is closed under the taking of linear combinations.

Evidently, T is surjective if and only if range(T) = W. Injectivity of T has a nice characterization in terms of nullspace(T).

4.7 Fact: Let V and W be vector spaces over $\mathbb{F}$ and let $T : V \to W$ be a linear mapping. T is injective if and only if nullspace(T) = $\{0\}$.

Proof: If T is injective, then any nonzero $v \in V$ has to map under T to something different from what 0 maps to, which is 0. Accordingly, if $v \neq 0$ then $T(v) \neq 0$, so nullspace(T) = $\{0\}$. Conversely, if nullspace(T) $\neq \{0\}$, we can find some $v_o \neq 0$ in nullspace(T). By definition of nullspace, $T(v_o) = 0$, so two different vectors map to the zero vector in W, implying that T is not injective. □

If V and W are vector spaces and $T : V \to W$ is a bijective mapping, an inverse mapping $S : W \to V$ exists regardless of whether T is linear. The mapping S has the following description: for every $w \in W$, $S(w)$ is the unique $v \in V$ satisfying $T(v) = w$. Note that S satisfies $S(T(v)) = v$ for every $v \in V$ and $T(S(w)) = w$ for every $w \in W$. If T is a linear mapping, the inverse mapping S is also linear.

4.8 Theorem: Let V and W be vector spaces over $\mathbb{F}$. A linear mapping $T : V \to W$ is bijective if and only if T is linearly invertible in the sense that there exists a linear mapping $S : W \to V$ such that $S(T(v)) = v$ for every $v \in V$ and $T(S(w)) = w$ for all $w \in W$.

Proof: If T is invertible with linear inverse S, then $T(S(w)) = w$ for all $w \in W$ implies that T is surjective, since every $w \in W$ is in range(T). Furthermore, if $T(v) = 0$, then $S(T(v)) = v$ implies that $v = 0$, so nullspace(T) = $\{0\}$ and T is injective by Fact 4.7. We conclude that if T is invertible, then T is bijective.

Conversely, suppose T is bijective and let S be the resulting inverse mapping from S to V. We need to show that S is linear. If $w_1, \ldots, w_k$ are vectors in W and $v_j \in V$ is the unique vector in V that maps under T to w_j for $1 \leq j \leq k$, we know that $S(w_j) = v_j$ for $1 \leq j \leq k$. Since T is linear, $c_1v_1 + \cdots + c_kv_k$ maps to $c_1w_1 + \cdots + c_kw_k$ for every $c_1, \ldots, c_k$ in $\mathbb{F}$. Accordingly,

$$\begin{aligned} S(c_1w_1 + \cdots + c_kw_k) &= c_1v_1 + \cdots + c_kv_k \\ &= c_1S(w_1) + \cdots + c_kS(w_k) \ , \end{aligned}$$

and S is therefore linear. Observe that if $V = W$, so T and S map V to itself, the composed mappings ST and TS are both the identity mapping on V. □

An injective linear mapping $T : V \to W$ maps linearly independent sets in V to linearly independent sets in W. To see why, let $v_1, \ldots, v_k$ be linearly independent vectors in V. If

$$0 = c_1T(v_1) + \cdots + c_kT(v_k) = T(c_1v_1 + \cdots + c_kv_k) \ ,$$

then $c_1v_1+\cdots+c_kv_k$ must lie in the nullspace of T, and hence must be zero by Fact 4.7. Since the v_j are linearly independent, all the c_j must be zero, and we conclude that $T(v_1)$, ... , $T(v_k)$ are linearly independent. When V is infinite-dimensional and W is finite-dimensional, V contains arbitrarily large linearly independent sets while W's linearly independent sets are limited in size by Theorem 4.3. Thus no injective linear mapping $T : V \to W$ exists. We can say a lot more when V is finite-dimensional.

4.9 Theorem: Let V and W be vector spaces over $\mathbb{F}$ with V finite-dimensional. If $T : V \to W$ is a linear mapping, then

$$\dim(\text{nullspace}(T)) + \dim(\text{range}(T)) = \dim(V) \ ,$$

where by convention $\dim(\{0\}) = 0$.

Proof: Suppose V has dimension n and nullspace(T) has dimension d. Let $(v_1, \ldots, v_d)$ be a basis for nullspace(T). By Theorem 4.4, we can find vectors v_{d+1}, ... , v_n so that $(v_1, \ldots, v_d, v_{d+1}, \ldots, v_n)$ is a basis for V. I claim that $(T(v_{d+1}), \ldots, T(v_n))$ is a basis for range(T). First of all, the vectors $T(v_j)$, $d+1 \leq j \leq n$, span range(T). That's because since $(v_1, \ldots, v_n)$ is a basis for V, every $w \in$ range(T) can be written as

$$\begin{aligned} T(c_1v_1 + \cdots + c_nv_n) &= c_1T(v_1) + \cdots c_dT(v_d) + c_{d+1}T(v_{d+1}) + \cdots + c_nT(v_n) \\ &= c_{d+1}T(v_{d+1}) + \cdots + c_nT(v_n) \end{aligned}$$

for some c_1, ... c_n in $\mathbb{F}$, where the last equality holds because $v_j \in$ nullspace(T) for $1 \leq j \leq d$. Second, the vectors $T(v_j)$, $d+1 \leq j \leq n$, are linearly independent. That's because if

$$0 = c_{d+1}T(v_{d+1}) + \cdots + c_nT(v_n) = T(c_{d+1}v_{d+1} + \cdots + c_nv_n) \ ,$$

then $c_{d+1}v_{d+1} + \cdots + c_nv_n$ lies in the nullspace of T, and hence can be written as a linear combination of v_1, ... , v_d, which leads to a relation of the form

$$c_1v_1 + \cdots + c_dv_d + c_{d+1}v_{d+1} + \cdots + c_nv_n = 0 \ ,$$

and all the c_j are therefore zero by linear independence of the v_j.

The bottom line is that $\{v_{d+1}, \ldots, v_n\}$ is a linearly independent spanning set for range(T). It follows that range(T) has dimension $n-d$, so $\dim(V) = d + (n-d) = \dim(\text{nullspace}(T)) + \dim(\text{range}(T))$. □

We noted earlier that there exists no injective linear mapping from an infinite-dimensional vector space to a finite-dimensional vector space. Theorem 4.9 makes possible some additional related assertions. First of all, since the dimension of range(T) is bounded from above by the dimension of V, no surjective linear mapping $T : V \to W$ exists when V is finite-dimensional and W is infinite-dimensional. The relationship between the dimensions of V and W also restricts what kinds of linear mappings $T : V \to W$ can exist when both V and W are finite-dimensional. A linear mapping $T : V \to W$ cannot be injective if $\dim(V) > \dim(W)$ because in that case

$$\dim(\text{nullspace}(T)) = \dim(V) - \dim(\text{range}(T)) \geq \dim(V) - \dim(W) > 0 ,$$

so nullspace$(T) \neq \{0\}$ and T is not injective by Fact 4.7. Similarly, T cannot be surjective if $\dim(V) < \dim(W)$. Theorem 4.9 implies in this case that

$$\dim(\text{range}(T)) = \dim(V) - \dim(\text{nullspace}(T)) < \dim(W) ,$$

meaning that range$(T) \neq W$, precluding surjectivity of T. Since a bijective mapping is both injective and surjective, it follows that a linear mapping $T : V \to W$ can't be bijective unless $\dim(V) = \dim(W)$.

Perhaps the most striking corollary of Theorem 4.9 is the following fundamental result about linear mappings between finite-dimensional vector spaces.

4.10 Theorem: If V and W are vector spaces over $\mathbb{F}$ with the same finite dimension, then the following conditions on a linear mapping $T : V \to W$ are equivalent in the sense that any one of them implies the other three:

- T is bijective
- T is injective
- T is surjective
- T is linearly invertible in the sense that there exists a linear mapping $S : W \to V$ such that $S(T(v)) = v$ for every $v \in V$ and $T(S(w)) = w$ for all $w \in W$.

Proof: For convenience, let $n = \dim(V) = \dim(W)$. Theorem 4.8 establishes the equivalence of bijectivity and linear invertibility af T. Bijectivity implies both injectivity and surjectivity by definition. If T is injective, then nullspace$(T) = \{0\}$, so dim(range$(T)) = n$ by Theorem 4.9. Since W also has dimension n, it follows that range(T) is an n-dimensional subspace of the n-dimensional vector space W, hence range$(T) = W$ and T is surjective. Accordingly, injectivity implies surjectivity, and bijectivity follows because it's just the conjunction of injectivity and surjectivity. Finally, assume T is surjective. Then range$(T) = W$, so by Theorem 4.9 we have

$$\dim(\text{nullspace}(T)) + n = n .$$

It follows that nullspace$(T) = \{0\}$ and T is injective by Fact 4.7. Thus surjectivity implies injectivity, and bijectivity holds again because it's just the conjunction of injectivity and surjectivity. □

If you're reading carefully, you've noticed that I've demonstrated only what kinds of linear mappings *can't* exist between finite-dimensional vector spaces whose dimensions bear various relationships. For example, I've shown only that a bijective linear mapping $T : V \to W$ can't exist unless $\dim(V) = \dim(W)$. In fact, plenty such mappings exist. If $(v_1, \ldots, v_n)$ is a any basis for V and $(w_1, \ldots, w_n)$ is any basis for W, then there exists a unique bijective linear mapping $T : V \to W$ satisfying $T(v_j) = w_j$ for $1 \leq j \leq n$. This is because every $v \in V$ can be written in exactly one way as a linear combination of the v_j by Theorem 4.3, so we can define T unambiguously by

$$T(c_1v_1 + \cdots + c_nv_n) = c_1w_1 + \cdots + c_nw_n$$

for every $c_1, \dots, c_n$ in $\mathbb{F}$.

Similarly, if $(w_1, \dots, w_m)$ is a basis for W and $n < m$, the exact same prescription provides an injective mapping $T : V \to W$. If on the other hand $n > m$, you get a surjective mapping $T : V \to W$ by setting

$$T(c_1v_1 + \cdots + c_nv_n) = c_1w_1 + \cdots + c_mw_m$$

for every $c_1, \dots, c_n$ in $\mathbb{F}$. Note that this last mapping sends each v_j for $m + 1 \le j \le n$ to zero. In fact,

$$\text{nullspace}(T) = \text{span}(\{v_{m+1}, \dots, v_n\}) .$$

The foregoing examples employ a useful technique for constructing a linear mapping from V to W when V is finite-dimensional, namely defining the linear mapping on basis for V and extending it to all of V by linearity. Specifying what a linear mapping does to the vectors in a basis for V specifies the mapping completely because every vector in V has a unique representation as a linear combination of the basis vectors, and the linear mapping must respect that linear combination. More precisely, if $(v_1, \dots, v_n)$ is a basis for V and $w_1, \dots, w_n$ are any vectors in W, there exists a unique linear mapping $T : V \to W$ such that $T(v_j) = w_j$ for $1 \le j \le n$. Where does T send an arbitrary $v \in V$? First find the unique c_j, $1 \le j \le n$, for which

$$v = c_1v_1 + \cdots + c_nv_n ,$$

and then you know by linearity that

$$T(v) = c_1w_1 + \cdots + c_nw_n .$$

If V and W are vector spaces over $\mathbb{F}$, let $\text{Hom}(V, W)$ be the set of all linear mappings from V to W. “Hom” is short for “homomorphism,” which means roughly “something that preserves form.” $\text{Hom}(V, W)$ is itself a vector space over $\mathbb{F}$. The zero vector in $\text{Hom}(V, W)$ is the zero mapping from V to W. The vector operations arise as follows. If T_1 and T_2 are in $\text{Hom}(V, W)$, define the linear mapping $T_1 + T_2$ by

$$(T_1 + T_2)(v) = T_1(v) + T_2(v) \text{ for all } v \in V .$$

If $T \in \text{Hom}(V, W)$ and $c_o \in \mathbb{F}$, define the linear mapping c_oT by

$$(c_oT)(v) = c_o(T(v)) \text{ for all } v \in V .$$

If V and W are finite-dimensional, then so in $\text{Hom}(V, W)$.

4.11 Theorem: If V and W are vector spaces over $\mathbb{F}$ with $\dim(V) = n$ and $\dim(W) = m$, then $\text{Hom}(V, W)$ has dimension mn.

Proof: Let $(v_1, \dots, v_n)$ be a basis for V and $(w_1, \dots, w_m)$ a basis for W. Define linear mappings $E_{ij} \in \text{Hom}(V, W)$ for $1 \le i \le m$ and $1 \le j \le n$ as follows.

$$E_{ij}(v_k) = \begin{cases} w_i & \text{if } k = j \\ 0 & \text{if } k \ne j . \end{cases}$$

In other words, E_{ij} sends v_j to w_i and sends all the other basis vectors v_k to zero. Defining the E_{ij} on a basis for V, as we’ve observed, specifies them completely. I claim that

$$\{E_{ij} : 1 \le i \le m, 1 \le j \le n\}$$

is a linearly independent spanning set for $\operatorname{Hom}(V, W)$, which therefore has dimension mn.

To prove that the E_{ij} span $\operatorname{Hom}(V, W)$, we must show how to express any $T \in \operatorname{Hom}(V, W)$ as a linear combination of the E_{ij}. Given such a T, we can write each $T(v_j)$ uniquely as a linear combination of the w_i. In other words, we can find t_{ij} in $\mathbb{F}$ such that

$$T(v_j) = t_{1j}w_1 + \cdots + t_{mj}w_m$$

for each j, $1 \leq j \leq n$. I'll leave it for you to show that

$$T = \sum_{i=1}^{m} \sum_{j=1}^{n} t_{ij} E_{ij}$$

and conclude that the E_{ij} span $\operatorname{Hom}(V, W)$.

As for linear independence of the E_{ij}, suppose

$$0 = \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} E_{ij} \ .$$

Apply the linear mapping on the right-hand side of this relation to v_k and you discover that

$$\begin{aligned} 0 &= \left(\sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} E_{ij} \right) (v_k) \\ &= \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} E_{ij}(v_k) \\ &= \sum_{i=1}^{m} c_{ik} w_i \ , \end{aligned}$$

which proves that $c_{ik} = 0$ for all i because the w_i are linearly independent. Since that maneuver works for every k, it follows that $c_{ik} = 0$ for all i and k, and the E_{ij} are therefore linearly independent. □

Norms and their associated convergence notions

A *norm* on a vector space V over $\mathbb{F}$ is a mapping

$$v \mapsto \|v\|$$

from V into $\mathbb{R}$ with three properties. First of all, $\|v\| \geq 0$ for all $v \in V$ and $\|v\| = 0$ if and only if $v = 0$. Second, $\|c_o v\| = |c_o| \|v\|$ for all $v \in V$ and $c_o \in \mathbb{F}$. Finally, we have the so-called triangle inequality

$$\|v + w\| \leq \|v\| + \|w\| \text{ for all } v, w \in V \ .$$

I'll refer to $\|v\|$ as the norm of v.

Given a norm on V, you can think of $\|v\|$ as representing the length of v as measured by that norm. The norm in turn spawns a distance function on V: the distance between v and w is $\|v - w\|$. Along with the distance function comes a notion of convergence: a sequence $\{v_n\}$ in V *converges to* $v \in V$ *with respect to the given norm* when

$$\lim_{n\to\infty} \|v_n - v\| = 0 \ .$$

In analogy with real and complex numbers, a sequence $\{v_n\}$ in V is a *Cauchy sequence with respect to the given norm* when for every $\epsilon > 0$ there exists an $N > 0$ such that $\|v_m - v_n\| < \epsilon$ when m and n are bigger than N.

Every sequence in V that converges with respect to a given norm on V is a Cauchy sequence with respect to the given norm. The proof of that statement is essentially identical to the proof of the corresponding statement in Fact 3.2. In contrast to real- and complex-number sequences, a sequence in V that's Cauchy with respect to a given norm need not be convergent with respect to that norm. Further complicating the convergence analysis of vector sequences is the fact that you can define many workable norms on a single vector space, none of which has the cachet associated with absolute value and magnitude for $\mathbb{R}$ and $\mathbb{C}$. A sequence in V might well be convergent or Cauchy with respect one norm and not with respect to another.

Suppose that $\| \ \|_a$ and $\| \ \|_b$ are two norms on a vector space V and that there exists $R > 0$ such that

$$\|v\|_b \leq R\|v\|_a \ \text{ for all } v \in V \ .$$

I claim that if a sequence $\{v_n\}$ converges to v with respect to $\| \ \|_a$ then it also converges to v with respect to $\| \ \|_b$. To see this, suppose $\epsilon > 0$ is given. If $\{v_n\}$ converges with respect to $\| \ \|_a$, we can find $N > 0$ such that

$$\|v_n - v\|_a < \epsilon/R$$

when $n > N$. It follows that

$$\|v_n - v\|_b \leq R\|v_n - v\|_a < \epsilon$$

when $n > N$, and $\{v_n\}$ therefore converges with respect to $\| \ \|_b$ since ϵ was arbitrary. Similarly, if a sequence $\{v_n\}$ is Cauchy with respect to $\| \ \|_a$ then it is also Cauchy with respect to $\| \ \|_b$. Again, given $\epsilon > 0$ we can find $N > 0$ such that

$$\|v_n - v_m\|_a < \epsilon/R$$

when m and n are bigger than N. It follows that

$$\|v_n - v_m\|_b \leq R\|v_n - v_m\|_a < \epsilon$$

when m and n are bigger than N, and $\{v_n\}$ is therefore Cauchy with respect to $\| \ \|_b$ since ϵ was arbitrary.

Two norms $\| \ \|_a$ and $\| \ \|_b$ on V are *equivalent* when there exist constants Q and R such that

$$\|v\|_b \leq R\|v\|_a \ \text{ for all } v \in V$$

and

$$\|v\|_a \leq Q\|v\|_b \ \text{ for all } v \in V$$

The discussion in the preceding paragraph reveals that two equivalent norms give rise to the same convergent sequences in the sense that a sequence $\{v_n\}$ converges to v with respect to one of the norms if and only if it converges to v with respect to the other. In Chapter 5 we'll meet examples of non-equivalent norms on infinite-dimensional vector spaces of signals. Non-equivalent norms and their attendant difficulties turn out not to be issues in finite-dimensional vector spaces. I won't prove that result, but it's a good one to know.

4.12 Theorem: If V be a finite-dimensional vector space over $\mathbb{F}$, then any two norms on V are equivalent. Consequently, a sequence $\{v_n\}$ converges to v with respect to some norm on V if and only if it converges to v with respect to every norm on V. □

To give Theorem 4.12 something to work with, I'll present three popular norms available an any n-dimensional vector space V once you've chosen a basis $(v_1, \ldots, v_n)$ for V. We know from Theorem 4.3 that every $v \in V$ has a unique representation as a linear combination

$$v = [v]_1 v_1 + \cdots + [v]_n v_n \ ,$$

where $[v]_j \in \mathbb{F}$ for $1 \leq j \leq n$. Define the *max norm* or *infinity norm* of v with respect to the given basis by

$$\|v\|_\infty = \max\left(\{|[v]_j| : 1 \leq j \leq n\}\right) \ .$$

Define the *sum norm* or *1-norm* of v with respect to the given basis by

$$\|v\|_1 = \sum_{j=1}^{n} |[v]_j| \ .$$

Define the *Euclidean norm* or *2-norm* of v with respect to the given basis by

$$\|v\|_2 = \left(\sum_{j=1}^{n} |[v]_j|^2\right)^{1/2} \ .$$

It's fairly straightforward to show that these are all norms on V. The only tricky part is proving the triangle inequality for the 2-norm, a task that I'll set aside until Chapter 9. We know from Theorem 4.12 that all three of these norms are equivalent, but let's prove it by hand. It's pretty obvious that

$$\|v\|_\infty \leq \|v\|_1 \text{ for all } v \in V$$

and that

$$\|v\|_1 \leq n\|v\|_\infty \text{ for all } v \in V \ ,$$

so the sum norm and max norms are equivalent. Meanwhile,

$$\|v\|_\infty^2 \leq \sum_{j=1}^{n} |[v]_j|^2 = \|v\|_2^2 \text{ for all } v \in V \ ,$$

so $\|v\|_\infty \leq \|v\|_2$ for all v, and

$$\|v\|_2^2 \leq n\|v\|_\infty^2 \text{ for all } v \in V \ ,$$

so $\|v\|_2 \leq \sqrt{n}\|v\|_\infty$ for all v, proving that the 2-norm and max norm are equivalent. Furthermore,

$$\begin{aligned}\|v\|_2^2 &= \sum_{j=1}^{n} |[v_j]|^2 \\ &\leq \left(\sum_{j=1}^{n} |[v]_j|\right)^2 = \|v\|_1^2 \ ,\end{aligned}$$

so $\|v\|_2 \leq \|v\|_1$ for all $v \in V$. Finally,

$$\|v\|_1 \leq n\|v\|_\infty \leq n\|v\|_2 \ \text{ for all } v \in V \ ,$$

so the sum norm and 2-norm are equivalent.

CHAPTER 5

Discrete-time Signals and Convolution

It's time to put some of the material from Chapter 3 to work. I think you'll begin to appreciate the central role that the basic facts about sequences, series, and convergence play in the study of mathematical models whose critical importance to modern applications is beyond dispute. Indeed, discrete-time signals constitute the currency mediating all transactions in the digital world we inhabit.

Discrete-time signals and their elementary properties

We view the integers $\mathbb{Z}$ as a mathematical model for "discrete time." Integer n corresponds to "time n." Integer 0 corresponds to "time 0." If $m > n$, then "time m is later than time n." It's not generally helpful to think of these "integer times" as being embedded somehow in a familiar "continuous time axis" or as having standard time units such as seconds or milliseconds or whatever. Integer times are just indices with a natural ordering.

Having settled on $\mathbb{Z}$ to model discrete time, let's define an $\mathbb{F}$-*valued discrete-time signal* as a function with domain $\mathbb{Z}$ that takes values in $\mathbb{F}$ — i.e., a discrete-time signal is some $x : \mathbb{Z} \to \mathbb{F}$. As usual, $\mathbb{F}$ is either $\mathbb{R}$ or $\mathbb{C}$. If you want, you can view a typical $\mathbb{F}$-valued discrete-time signal x as a two-sided infinite sequence of numbers from $\mathbb{F}$, i.e.

$$\ldots, x(-3), x(-2), x(-1), x(0), x(1), x(2), \ldots$$

where $x(n) \in \mathbb{F}$ for all $n \in \mathbb{Z}$. Think of $x(n)$ as the value of the signal x at time n. I'll denote the set of all $\mathbb{F}$-valued discrete-time signals by $\mathbb{F}^{\mathbb{Z}}$. When discussing a discrete-time signal $x : \mathbb{Z} \to \mathbb{F}$, I'll be consistent in using x to denote the whole signal and $x(n)$ to denote the value of x at the specific time n.

I won't list a whole bunch of examples of signals, but three special signals deserve mention. The *zero signal* is the signal x with specification $x(n) = 0$ for every $n \in \mathbb{Z}$. The *discrete-time unit impulse* is the signal δ with specification

$$\delta(n) = \begin{cases} 1 & \text{if } n = 0 \\ 0 & \text{if } n \neq 0 \,. \end{cases}$$

The *discrete-time unit step* is the signal u with specification

$$u(n) = \begin{cases} 1 & \text{if } n \geq 0 \\ 0 & \text{if } n < 0 \,. \end{cases}$$

I'll always use notation 0 for the zero signal, δ for the unit impulse, and u for the unit step.

A signal x is *right-sided* when there exists $N_1 \in \mathbb{Z}$ such that $x(n) = 0$ when $n < N_1$. A signal x is *left-sided* when there exists $N_2 \in \mathbb{Z}$ such that $x(n) = 0$ when

$n > N_2$. A signal x has *finite duration* when there exist integers N_1 and N_2 such that $x(n) = 0$ when $n < N_1$ and $x(n) = 0$ when $n > N_2$. Clearly, a signal has finite duration if and only if it is both right- and left-sided. The unit step and unit impulse are archetypal right-sided and finite-duration signals, respectively.

The set $\mathbb{F}^{\mathbb{Z}}$ of all discrete-time signals has a natural vector-space structure. The zero signal is the zero vector, and linear combinations of signals are defined pointwise in the sense that if x_1 and x_2 are two signals in $\mathbb{F}^{\mathbb{Z}}$ and c_1 and c_2 are scalars in $\mathbb{F}$, then the signal $y = c_1x_1 + c_2x_2$ has specification

$$y(n) = c_1x_1(n) + c_2x_2(n) \text{ for all } n \in \mathbb{Z} . \tag{3}$$

Observe that the set of all right-sided signals is a subspace of $\mathbb{F}^{\mathbb{Z}}$ under these vector operations. In other words, the set of right-sided signals is closed under the taking of linear combinations. The same is true of the set of left-sided signals and the set of finite-duration signals. I suggest you prove these facts for yourself.

Also defined on the vector space $\mathbb{F}^{\mathbb{Z}}$ is the operation of time shifting. Given $x \in \mathbb{F}^{\mathbb{Z}}$ and $k_o \in \mathbb{Z}$, define $\text{Shift}_{k_o}(x)$ as the signal with specification

$$\text{Shift}_{k_o}(x)(n) = x(n - k_o) \text{ for all } n \in \mathbb{Z} .$$

I'll let you check that Shift_{k_o} is a linear operation on $\mathbb{F}^{\mathbb{Z}}$ for every $k_o \in \mathbb{Z}$ in the sense that

$$\text{Shift}_{k_o}(c_1x_1 + c_2x_2) = c_1\text{Shift}_{k_o}(x_1) + c_2\text{Shift}_{k_o}(x_2)$$

for every x_1 and x_2 in $\mathbb{F}^{\mathbb{Z}}$ and every c_1 and c_2 in $\mathbb{F}$. When $k_o > 0$, you can view $\text{Shift}_{k_o}(x)$ as "x delayed by time k_o." In this case, if you could graph $\text{Shift}_{k_o}(x)(n)$ as a function of n it would look just like the graph of $x(n)$ as a function of n shifted to the right by k_o. Observe that the set of right-sided signals, the set of left-sided signals, and the set of finite-duration signals are all closed under shifting in the sense that $\text{Shift}_{k_o}(x)$ has finite duration or is right- or left-sided when x has the same property.

Bounded and absolutely summable signals: the spaces l^∞ and l^1

We say that $x \in \mathbb{F}^{\mathbb{Z}}$ is a *bounded signal* when there exists $R > 0$ such that $|x(n)| \le R$ for every $n \in \mathbb{Z}$. The *infinity norm* of a bounded signal x is defined by

$$\|x\|_\infty = \sup\{|x(n)| : n \in \mathbb{Z}\} .$$

We use the notation l^∞ for the set of all bounded signals in $\mathbb{F}^{\mathbb{Z}}$. As it happens, l^∞ is a subspace of $\mathbb{F}^{\mathbb{Z}}$ in the sense that it is closed under the taking of linear combinations as in (3). To see this, note that if $y = c_1x_1 + c_2x_2$, where x_1 and x_2 are in l^∞ and c_1 and c_2 are in $\mathbb{F}$, then

$$|y(n)| = |c_1x_1(n) + c_2x_2(n)| \le |c_1||x_1(n)| + |c_2||x_2(n)| \text{ for all } n \in \mathbb{Z} ,$$

so

$$|y(n)| \le |c_1|\|x_1\|_\infty + |c_2|\|x_2\|_\infty \text{ for all } n \in \mathbb{Z}$$

by definition of the infinity norm. It follows that y is bounded and that

$$\|y\|_\infty \le |c_1|\|x_1\|_\infty + |c_2|\|x_2\|_\infty .$$

One further comment: the infinity norm is actually a norm on the vector space l^∞ in the technical sense, and this is easy to demonstrate (the triangle inequality, for example, follows from a manipulation like the one above with $c_1 = c_2 = 1$).

We say that $x \in \mathbb{F}^\mathbb{Z}$ is an *absolutely summable signal* when $\sum_{n=-\infty}^{\infty} |x(n)|$ converges. The *1-norm* of an absolutely summable signal x is defined by

$$\|x\|_1 = \sum_{n=-\infty}^{\infty} |x(n)| \ .$$

We use the notation l^1 for the set of all absolutely summable signals in $\mathbb{F}^\mathbb{Z}$.

Like l^∞, the set l^1 is a subspace of $\mathbb{F}^\mathbb{Z}$ in the sense that it is closed under the taking of linear combinations via (3). Moreover, the 1-norm is indeed a norm on l^1. These assertions are direct consequences of Fact 3.7. To see at least partly how that goes, suppose x_1 and x_2 are in l^1 and c_1 and c_2 are in $\mathbb{F}$. Set $y = c_1x_1 + c_2x_2$. Then for every $N \in \mathbb{Z}$

$$\sum_{n=-N}^{N} |y(n)| = \sum_{n=-N}^{N} (|c_1x_1(n) + c_2x_2(n)|) \leq |c_1| \sum_{n=-N}^{N} |x_1(n)| + |c_2| \sum_{n=-N}^{N} |x_2(n)| \ .$$

The two sums on the right-hand side are bounded from above respectively by $\|x_1\|_1$ and $\|x_2\|_1$, so

$$\sum_{n=-N}^{N} |y(n)| \leq |c_1|\|x_1\|_1 + |c_2|\|x_2\|_1 = R \ \text{ for all } \ N \in \mathbb{Z} \ .$$

Fact 3.7 implies that the sequence $\{y(n)\}$ is absolutely summable, so $y \in l^1$.

Observe that an absolutely summable signal must be bounded — in other words, $l^1 \subset l^\infty$. Observe also that every finite-duration signal is both bounded and absolutely summable. Often we refer to a signal $x \in l^\infty$ as "an l^∞-signal" and $x \in l^1$ as "an l^1-signal." Since l^1 is a subspace of l^∞, both the 1-norm and the infinity norm are serviceable norms on l^1. In Chapter 4 during the discussion leading up to Theorem 4.12, I promised you an example of non-equivalent norms on an infinite-dimensional vector space, and this is it. Here is a sequence $\{x_k\}$ of l^1 signals that converges to 0 in the infinity norm but not in the 1-norm. For each $k > 0$, let x_k be the l^1-signal with specification

$$x_k(n) = \begin{cases} 1/k & 0 \leq n < k \\ 0 & \text{otherwise.} \end{cases}$$

Since $\|x_k\|_\infty = 1/k$ for all $k > 0$, the sequence $\{x_k\}$ converges to 0 in the infinity norm. On the other hand, $\|x_k\|_1 = 1$ for all $k > 0$, so $\{x_k\}$ does not converge to 0 in the 1-norm.

It turns out that l^1 and l^∞ are particular instantiations of what are known as the *l^p-spaces* of discrete-time signals. Given a real number $p \geq 1$, let l^p be the set of all $x \in \mathbb{F}^\mathbb{Z}$ for which

$$\sum_{n=-\infty}^{\infty} |x(n)|^p$$

converges. It turns out that l^p is a subspace of $\mathbb{F}^\mathbb{Z}$ and that the prescription

$$\|x\|_p = \left(\sum_{n=-\infty}^{\infty} |x(n)|^p \right)^{1/p} \qquad \text{for all } \ x \in l^p$$

defines a norm on l^p.

Of special importance among the l^p-spaces is l^2, which will play a major role in our discussion of Fourier series as orthogonal expansions. l^2 is the set of *square summable* signals. You might wonder how the l^p-spaces stand in relation to one another. I noted above that $l^1 \subset l^\infty$. It turns out that if $p \le q$, then $l^p \subset l^q$, and that $l^p \subset l^\infty$ for every real number $p \ge 1$. In particular,

$$l^1 \subset l^2 \subset l^\infty \ .$$

The right-hand inclusion is easy to verify. The left-hand inclusion follows from the following bit of trickery. First note that if $x \in l^1$, then $|x(n)| > 1$ can hold for at most finitely many n; otherwise, x would not be an absolutely summable signal. Let R_1 be the sum of all $|x(n)|^2$ with $|x(n)| > 1$. For all other n, $|x(n)|^2 \le |x(n)|$. You can conclude that for every $N \in \mathbb{Z}$,

$$\sum_{n=-N}^{N} |x(n)|^2 \le R_1 + \sum_{n=-N}^{N} |x(n)| \le R_1 + \|x\|_1 \ .$$

While this upper bound is quite crude, it's enough to imply via Fact 3.7 that $x \in l^2$.

And finally, a word on notation and pronunciation. The "l" in l^p is the first letter of the last name of Henri Lebesgue, a great French mathematician of the early twentieth century. l^p is pronounced like "ell pea" or "little ell pea." The reason for the latter is that the continuous-time versions of these spaces, the so-called L^p-spaces, came first, and L^p is always pronounced "ell pea."

Convolution

Given two signals x_1 and x_2 in $\mathbb{F}^{\mathbb{Z}}$, the *convolution of* x_1 *and* x_2, if it exists, is the signal $x \in \mathbb{F}^{\mathbb{Z}}$ with specification

$$x(n) = \sum_{k=-\infty}^{\infty} x_1(k)x_2(n-k) \ , \quad n \in \mathbb{Z} \ . \tag{4}$$

Some notations I'll be using for the convolution of x_1 and x_2 are $x_1 * x_2$ and $\mathrm{Conv}(x_1, x_2)$. Alternative terminologies for the convolution of x_1 and x_2 are "the convolution of x_1 *with* x_2" and "x_1 convolved with x_2." In accordance with our notation for discrete-time signals, $x_1 * x_2$ denotes a "whole signal" and $x_1 * x_2(n)$ denotes the value of that signal at time n, so we can rewrite (4) as

$$x_1 * x_2(n) = \sum_{k=-\infty}^{\infty} x_1(k)x_2(n-k) \ , \quad n \in \mathbb{Z} \ .$$

The convolution of x_1 and x_2 exists if and only if the sum in (4) converges for every $n \in \mathbb{Z}$. The sum is a series that's potentially "infinite in both directions" such as the ones we met during our discussion of l^1, and we handle it accordingly.

Let's begin with an elementary observation about convolution. If $x_1 * x_2$ exists, then the sum in (4) converges for every $n \in \mathbb{Z}$. Change the index of summation as follows:

$$x_1 * x_2(n) = \sum_{k=-\infty}^{\infty} x_1(k)x_2(n-k) = \sum_{m=-\infty}^{\infty} x_1(n-m)x_2(m) = \sum_{k=-\infty}^{\infty} x_1(n-k)x_2(k) \ .$$

Setting $m = n - k$ yields the middle equality. To get the last, re-name the "dummy index of summation" m as k. The bottom line is that on the right-hand side of equation (4), it doesn't matter where we put the k and where we put the $(n-k)$ — the result is the same. One could dignify this observation by saying something along the lines of, "convolution, defined by (4), is a commutative operation in the sense that if $x_1 * x_2$ exists, then $x_1 * x_2 = x_2 * x_1$." That's fine, but it's a bit unnecessary in my view.

A slightly less elementary observation about convolution is that it is an *associative* operation in the sense that if $x_1*(x_2*x_3)$ exists, then so does $(x_1*x_2)*x_3$, and vice versa, and both convolutions are the same. Proving this fact is an exercise in summation manipulation. I'll be cavalier about interchanging orders of summation here, but the interchanges are legal since all the sums converge. Assuming that $x_1 * (x_2 * x_3)$ exists, we have

$$\begin{aligned}
x_1 * (x_2 * x_3)(n) &= \sum_{k=-\infty}^{\infty} x_1(k)(x_2 * x_3(n-k)) \\
&= \sum_{k=-\infty}^{\infty} x_1(k)\left(\sum_{m=-\infty}^{\infty} x_2(m)x_3((n-k)-m)\right) \\
&= \sum_{k=-\infty}^{\infty} x_1(k)\left(\sum_{m=-\infty}^{\infty} x_2((n-k)-m)x_3(m)\right) \\
&= \sum_{m=-\infty}^{\infty}\left(\sum_{k=-\infty}^{\infty} x_1(k)x_2((n-m)-k)\right)x_3(m) \\
&= \sum_{m=-\infty}^{\infty} (x_1 * x_2(n-m))x_3(m) \\
&= (x_1 * x_2) * x_3(n)\ .
\end{aligned}$$

The equalities hold for every $n \in \mathbb{Z}$, so

$$x_1 * (x_2 * x_3) = (x_1 * x_2) * x_3\ .$$

(Note: to get the third equality in the chain above, I used the "commutativity" of convolution that I alluded to earlier, which allowed me to switch the roles of $(n-k)-m$ and m.)

Convolution is also *bilinear* in the sense that

$$x_1 * (c_2x_2 + c_3x_3) = c_2x_1 * x_2 + c_3x_1 * x_3$$

and

$$(c_1x_1 + c_2x_2) * x_3 = c_1x_1 * x_3 + c_2x_2 * x_3$$

for every c_1, c_2, and c_3 in $\mathbb{F}$. I'm assuming here that all the indicated convolutions exist, and existence is our next order of business.

Criteria for existence of convolutions

Given x_1 and x_2 in $\mathbb{F}^{\mathbb{Z}}$, how can we tell whether $x_1 * x_2$ exists? Note that equation (4) is really shorthand for an infinite list of equations — one for each $n \in \mathbb{Z}$ — and each of those equations involves an infinite sum whose existence is an issue. Aside from checking for convergence of each of these infinitely many infinite sums, how might we proceed? In what follows, I'll state and prove several useful criteria for the existence of $x_1 * x_2$.

Here's an example of signals x_1 and x_2 whose convolution fails to exist. Let x_1 be the constant signal whose value is 13 for every $n \in \mathbb{Z}$, i.e.

$$x_1(n) = 13 \quad \text{for all } n \in \mathbb{Z} \ .$$

Let $x_2 = u$, the discrete-time unit step, which has specification

$$u(n) = \begin{cases} 1 & \text{if } \ n \geq 0 \\ 0 & \text{if } \ n < 0 \ . \end{cases}$$

Attempting to compute $x_1 * x_2(n)$ using (4) leads to

$$x_1 * x_2(n) = \sum_{k=-\infty}^{\infty} x_1(k)x_2(n-k) = \sum_{k=-\infty}^{\infty} 13u(n-k) = \sum_{k=-\infty}^{n} 13 = \infty \ .$$

The third equality holds because $u(n-k) = 0$ when $k > n$ and $u(n-k) = 1$ when $k \leq n$. The last equality holds because the sum of an infinite number of 13's does not converge.

Now for something more positive. Certain restrictions on x_1 and x_2 guarantee that $x_1 * x_2$ exists. Below I demonstrate the validity of four useful criteria each of which provides a sufficient condition for $x_1 * x_2$ to exist.

5.1 Criterion: If either x_1 or x_2 has finite duration, then $x_1 * x_2$ exists. If both x_1 and x_2 have finite duration, then $x_1 * x_2$ also has finite duration.

Proof: In this case, the sum in (4) has finitely many nonzero terms for every $n \in \mathbb{Z}$, which means convergence is not an issue. To see this, suppose that x_1 has finite duration and that $x_1(n) = 0$ when $n < N_1$ and when $n > N_2$. Then

$$\sum_{k=-\infty}^{\infty} x_1(k)x_2(n-k) = \sum_{k=N_1}^{N_2} x_1(k)x_2(n-k)$$

for every $n \in \mathbb{Z}$. A similar argument applies when x_2 has finite duration. The bottom line is that the sums in (4) converge for every $n \in \mathbb{Z}$, so $x_1 * x_2$ exists.

Suppose that both x_1 and x_2 have finite duration; specifically, assume $x_1(n) = 0$ when $n < N_1$ and when $n > N_2$ and that $x_2(n) = 0$ when $n < M_1$ and when $n > M_2$. We still have

$$x_1 * x_2(n) = \sum_{k=-\infty}^{\infty} x_1(k)x_2(n-k) = \sum_{k=N_1}^{N_2} x_1(k)x_2(n-k) \ .$$

If $n < N_1 + M_1$, then $n - k < M_1$ for every k in the range of summation, which means that $x_2(n-k) = 0$ for all such k and the sum is therefore zero. In other words, $x_1 * x_2(n) = 0$ when $n < N_1 + M_1$. Similarly, if $n > N_2 + M_2$, then $n - k > M_2$ for all k in the range of summation, meaning that $x_2(n-k) = 0$ for all such k and the sum is zero once again, implying that $x_1 * x_2(n) = 0$ when $n > N_2 + M_2$. We conclude that $x_1 * x_2$ has finite duration. □

It follows from Criterion 5.1 that for any $x \in \mathbb{F}^{\mathbb{Z}}$, the convolution $\delta * x$ exists. Indeed, as I noted earlier, δ is arguably the archetypal finite-duration signal. Of great importance is the fact that $\delta * x = x$ for every $x \in \mathbb{F}^{\mathbb{Z}}$, so that δ serves as an identity element for the operation of convolution. To see this, note that for every $n \in \mathbb{Z}$

$$\delta * x(n) = \sum_{k=-\infty}^{\infty} \delta(k)x(n-k) = x(n-0) = x(n)$$

since $\delta(k) = 0$ when $k \neq 0$ and $\delta(0) = 1$.

5.2 Criterion: If x_1 and x_2 are both right-sided or both left-sided, then $x_1 * x_2$ exists. Furthermore, in this case $x_1 * x_2$ has the same "sidedness" as x_1 and x_2.

Proof: I'll present the argument only in the case when both signals are right-sided; the left-sided version is similar. Suppose, then, that $x_1(n) = 0$ when $n < N_1$ and $x_2(n) = 0$ when $n < M_1$. Then

$$\begin{aligned} \sum_{k=-\infty}^{\infty} x_1(k)x_2(n-k) &= \sum_{k=N_1}^{\infty} x_1(k)x_2(n-k) \\ &= \begin{cases} \sum_{k=N_1}^{n-M_1} x_1(k)x_2(n-k) & \text{if } n \geq N_1 + M_1 \\ 0 & \text{if } n < N_1 + M_1 \;. \end{cases} \end{aligned}$$

The first equality holds because $x_1(n) = 0$ when $n < N_1$. The second is a bit more involved. First, note that if $n < N_1 + M_1$, then $n - k < M_1$ for every k in the range of summation, so $x_2(n-k) = 0$ for all such k, and every term in the sum is zero. If $n \geq N_1 + M_1$, then the sum features $x_2(n - N_1)$, $x_2(n - N_1 - 1)$, etc. Some of these might be nonzero, but $x_2(n-k) = 0$ for all the k-values satisfying $n-k < M_1$, which is the same as $k > n - M_1$. So the terms in the sum corresponding with k-values in the range $n - M_1 < k < \infty$ are all zero. This argument proves that $x_1 * x_2$ exists (since all the sums in (4) have finitely many nonzero terms) and that $x_1 * x_2$ is right-sided (since $x_1 * x_2(n) = 0$ when $n < N_1 + M_1$). □

5.3 Criterion: Given two signals x_1 and x_2, if one signal is bounded and the other is absolutely summable, then $x_1 * x_2$ exists and is a bounded signal. Furthermore, the infinity norm of $x_1 * x_2$ satisfies

$$\|x_1 * x_2\|_\infty \leq \|\text{the } l^1 \text{ signal}\|_1 \, \|\text{the } l^\infty \text{ signal}\|_\infty \;.$$

Proof: I'll prove this in the case that x_1 is an l^1 signal and x_2 is an l^∞ signal. Given $n \in \mathbb{Z}$, we can conclude from Fact 3.3 and Fact 3.7 that the sum in (4) converges if we can find $R > 0$ such that

$$\sum_{k=-K}^{K} |x_1(k)| \, |x_2(n-k)| \leq R$$

for every $K \in \mathbb{Z}$. Since $|x_2(n-k)| \leq \|x_2\|_\infty$ for every n and k,

$$\sum_{k=-K}^{K} |x_1(k)| \, |x_2(n-k)| \leq \left(\sum_{n=-K}^{K} |x_1(k)| \right) \|x_2\|_\infty \;.$$

The sum in parentheses is, in turn, bounded from above by $\|x_1\|_1$, from which it follows that

$$\sum_{k=-K}^{K} |x_1(k)|\,|x_2(n-k)| \leq \|x_1\|_1\,\|x_2\|_\infty\ ,$$

implying not only that the sum in (4) converges for every $n \in \mathbb{Z}$ (so $x_1 * x_2$ exists), but also that

$$|x_1 * x_2(n)| \leq \|x_1\|_1\,\|x_2\|_\infty$$

for every $n \in \mathbb{Z}$. It follows that $x_1 * x_2$ is a bounded signal (i.e. an l^∞-signal), and that $\|x_1 * x_2\|_\infty \leq \|x_1\|_1\,\|x_2\|_\infty$. □

5.4 Criterion: If x_1 and x_2 are both square-summable signals, then $x_1 * x_2$ exists and is a bounded signal. Furthermore, the infinity norm of $x_1 * x_2$ satisfies

$$\|x_1 * x_2\|_\infty \leq \frac{\|x_1\|_2^2 + \|x_2\|_2^2}{2}\ .$$

Proof: I'll proceed as in the proof of Criterion 5.3. Again, given $n \in \mathbb{Z}$, we can conclude from Fact 3.3 and Fact 3.7 that the sum in (4) converges if we can find $R > 0$ such that

$$\sum_{k=-K}^{K} |x_1(k)|\,|x_2(n-k)| \leq R$$

for every $K \in \mathbb{Z}$. From $(|x_1(k)| - |x_2(n-k)|)^2 \geq 0$ it follows directly that

$$|x_1(k)||x_2(n-k)| \leq \frac{|x_1(k)|^2 + |x_2(n-k)^2|}{2} \quad \text{for all } \ k \in \mathbb{Z}\ .$$

Thus

$$\begin{aligned}
\sum_{k=-K}^{K} |x_1(k)|\,|x_2(n-k)| &\leq \frac{1}{2}\sum_{k=-K}^{K} |x_1(k)|^2 + \frac{1}{2}\sum_{k=-K}^{K} |x_2(n-k)|^2 \\
&\leq \frac{\|x_1\|_2^2 + \|x_2\|_2^2}{2} \quad \text{for all } \ K \in \mathbb{Z}\ .
\end{aligned}$$

It follows that the sum in (4) converges for every $n \in \mathbb{Z}$, so $x_1 * x_2$ exists, and also that for every $n \in \mathbb{Z}$

$$\begin{aligned}
|x_1 * x_2(n)| &= \lim_{K\to\infty} \left| \sum_{k=-K}^{K} x_1(k)x_2(n-k) \right| \\
&\leq \lim_{K\to\infty} \sum_{k=-K}^{K} |x_1(k)||x_2(n-k)| \\
&\leq \frac{\|x_1\|_2^2 + \|x_2\|_2^2}{2}\ ,
\end{aligned}$$

whereby $\|x_1 * x_2\|_\infty \leq \left(\|x_1\|_2^2 + \|x_2\|_2^2\right)/2$. □

In Chapter 9 we'll see how the Schwarz Inequality enables us to tighten the upper bound on $\|x_1 * x_2\|_\infty$ in Criterion 5.4.

5.5 Criterion: If x_1 and x_2 are both absolutely summable signals, then $x_1 * x_2$ exists and is an absolutely summable signal. Furthermore, the 1-norm of $x_1 * x_2$ satisfies

$$\|x_1 * x_2\|_1 \leq \|x_1\|_1 \, \|x_2\|_1 \; .$$

Proof: The existence of $x_1 * x_2$ in this case follows directly from Criterion 5.3 because every absolutely summable signal, as we have noted, is also bounded. Proving that $x_1 * x_2$ is an l^1-signal takes a little more work.

Given $N \in \mathbb{Z}$,

$$\begin{aligned} \sum_{n=-N}^{N} |x_1 * x_2(n)| &= \sum_{n=-N}^{N} \left| \sum_{k=-\infty}^{\infty} x_1(k) x_2(n-k) \right| \\ &\leq \sum_{n=-N}^{N} \sum_{k=-\infty}^{\infty} |x_1(k)| \, |x_2(n-k)| \\ &= \sum_{k=-\infty}^{\infty} \left(|x_1(k)| \sum_{n=-N}^{N} |x_2(n-k)| \right) \; . \end{aligned}$$

Interchanging the order of summation is legal because the sum over n is finite and the sum over k converges for every n. The inner sum on the last line is bounded from above by $\|x_2\|_1$, which implies that

$$\sum_{n=-N}^{N} |x_1 * x_2(n)| \leq \left(\sum_{k=-\infty}^{\infty} |x_1(k)| \right) \|x_2\|_1 = \|x_1\|_1 \, \|x_2\|_1 \; .$$

This inequality holds for every $N \in \mathbb{Z}$, and Fact 3.7 implies not only that $\sum_{n=-\infty}^{\infty} |x_1 * x_2(n)|$ converges (meaning that $x_1 * x_2$ is an absolutely summable signal) but also that

$$\|x_1 * x_2\|_1 \leq \|x_1\|_1 \, \|x_2\|_1 \; ,$$

which completes the proof. □

Criterion 5.5 demonstrates that l^1 is closed under the taking of convolutions. Accordingly, l^1 is a vector space equipped with a commutative and associative "multiplication" operation (namely convolution) that behaves appropriately with respect the vector-space operations on l^1. Such a vector space is called a *commutative algebra*, and people often characterize l^1 as a "convolution algebra."

Examples

It's worth computing a few example convolutions by hand. All the calculations proceed similarly, as you'll see.

5.6 Example: $x_1 = x_2 = u$. Since x_1 and x_2 are both right-sided, Criterion 5.2 applies, so $x_1 * x_2$ exists. For any $n \in \mathbb{Z}$,

$$x_1 * x_2(n) = \sum_{k=-\infty}^{\infty} u(k)u(n-k) = \sum_{k=0}^{\infty} u(n-k) \ ;$$

the last equality holds because $u(k) = 1$ for $k \geq 0$ and $u(k) = 0$ for $k < 0$. If $n < 0$, $u(n-k) = 0$ for every k in the range of summation $0 \leq k \leq \infty$, so the whole sum is zero when $n < 0$. When $n \geq 0$, $u(n-k) = 1$ when $0 \leq k \leq n$ and $u(n-k) = 0$ when $n < k < \infty$. It follows that

$$x_1 * x_2(n) = \begin{cases} 0 & \text{if } \ n < 0 \\ \sum_{k=0}^{n} 1 = n+1 & \text{if } \ n \geq 0 \ . \end{cases}$$

Another way of writing this last equation is

$$x_1 * x_2(n) = (n+1)u(n)$$

for every $n \in \mathbb{Z}$. If you think about what $x_1 * x_2$ looks like in this case, you can see why people say that "the convolution of two unit steps is a ramp."

5.7 Example: $x_1 = u$ and x_2 is the signal with specification

$$x_2(n) = \begin{cases} 3^{-n} & n \geq 0 \\ 0 & n < 0 \ . \end{cases}$$

Note that $x_2(n) = 3^{-n}u(n)$ for every $n \in \mathbb{Z}$. Again, Criterion 5.2 applies since both x_1 and x_2 are right-sided. In fact, Criterion 5.3 also applies since x_1 is bounded and x_2 is absolutely summable, which is easy to check (hint: geometric series). To compute $x_1 * x_2$, follow a procedure similar to the one we followed in the previous example. For any $n \in \mathbb{Z}$,

$$x_1 * x_2(n) = \sum_{k=-\infty}^{\infty} u(k)3^{-(n-k)}u(n-k) = \sum_{k=0}^{\infty} 3^{-(n-k)}u(n-k) \ ,$$

where the last equality holds because $u(k) = 1$ for $k \geq 0$ and $u(k) = 0$ for $k < 0$. If $n < 0$, $u(n-k) = 0$ for every k in the range of summation $0 \leq k \leq \infty$, so the whole sum is zero when $n < 0$. When $n \geq 0$, $u(n-k) = 1$ when $0 \leq k \leq n$ and $u(n-k) = 0$ when $n < k < \infty$. It follows that

$$x_1 * x_2(n) = \begin{cases} 0 & \text{if } \ n < 0 \\ \sum_{k=0}^{n} 3^{-(n-k)} & \text{if } \ n \geq 0 \ . \end{cases}$$

Using simple geometric-series reasoning, it follows that

$$x_1 * x_2(n) = \begin{cases} 0 & \text{if } \ n < 0 \\ \frac{3}{2} - \frac{1}{2}3^{-n} & \text{if } \ n \geq 0 \ . \end{cases}$$

Another way of writing this last equation is

$$x_1 * x_2(n) = \left(\frac{3}{2} - \frac{1}{2}3^{-n}\right) u(n)$$

for every $n \in \mathbb{Z}$.

5.8 Example: $x_1 = u$ and x_2 is the signal with specification

$$x_2(n) = 3^{-|n|} = \begin{cases} 3^{-n} & \text{if } n \geq 0 \\ 3^n & \text{if } n < 0 \, . \end{cases}$$

Note that we can also specify x_2 for every $n \in \mathbb{Z}$ via

$$x_2(n) = 3^{-n}u(n) + 3^n u(-n-1) \, .$$

Note also that x_1 is an l^∞ signal and x_2 is an l^1-signal, so Criterion 5.3 guarantees that $x_1 * x_2$ exists. It helps to set $x_3(n) = 3^{-n}u(n)$ and $x_4(n) = 3^n u(-n-1)$ because we computed $x_1 * x_3$ in the previous example. Now let's find $x_1 * x_4$.

By definition, for any $n \in \mathbb{Z}$,

$$x_1 * x_4(n) = \sum_{k=-\infty}^{\infty} u(k)3^{n-k}u(-(n-k)-1) = \sum_{k=0}^{\infty} 3^{n-k}u(-n+k-1)$$

where the last equality holds because $u(k) = 1$ for $k \geq 0$ and $u(k) = 0$ for $k < 0$. If $n < 0$, then $u(-n+k-1) = 1$ for all k in the range of summation $0 \leq k < \infty$. If $n \geq 0$, then $u(-n+k-1) = 0$ for $0 \leq k < n+1$ and $u(-n+k-1) = 1$ for all k in the range $n+1 \leq k < \infty$. Accordingly,

$$x_1 * x_4(n) = \begin{cases} \sum_{k=0}^{\infty} 3^{n-k} & \text{if } n < 0 \\ \sum_{k=n+1}^{\infty} 3^{n-k} & \text{if } n \geq 0 \, . \end{cases}$$

Geometric-series manipulation reveals that

$$x_1 * x_4(n) = \begin{cases} \frac{3}{2}3^n & \text{if } n < 0 \\ \frac{1}{2} & \text{if } n \geq 0 \, . \end{cases}$$

Plugging the result of the previous example into the equation $x_1{*}x_2 = x_1{*}x_3 + x_1{*}x_4$ yields the following specification for $x_1 * x_2$:

$$x_1 * x_2(n) = \begin{cases} \frac{3}{2}3^n & \text{if } n < 0 \\ 2 - \frac{1}{2}3^{-n} & \text{if } n \geq 0 \, . \end{cases}$$

Alternatively, for every $n \in \mathbb{Z}$,

$$x_1 * x_2(n) = \left(2 - \frac{1}{2}3^{-n}\right)u(n) + \frac{3}{2}3^n u(-n-1) \, .$$

Just for completeness, I'd like to show you another way to do this example. Interchanging k and $n-k$ in equation (4) yields

$$x_1 * x_2(n) = \sum_{k=-\infty}^{\infty} u(n-k)3^{-|k|} = \sum_{k=-\infty}^{n} 3^{-|k|}$$

for every $n \in \mathbb{Z}$. The last equality holds because the $u(n-k)$ merely chops of the top of the sum at $k = n$. As a result,

$$x_1 * x_2(n) = \begin{cases} \sum_{k=-\infty}^{n} 3^k & \text{if } n < 0 \\ \sum_{k=-\infty}^{-1} 3^k + \sum_{k=0}^{n} 3^{-k} & \text{if } n \geq 0 \, . \end{cases}$$

Fortunately, as you can check, this turns out to be the same answer in a slightly different form.

5.9 Example: $x_1 = u$ and x_2 is the signal with specification $x_2(n) = 3^n$ for all $n \in \mathbb{Z}$. I'm including this example partly because it satisfies none of the criteria we've discussed for convolution existence. Those criteria, in other words, aren't exhaustive. Here, x_1 is bounded and right-sided, but x_2 is neither right-sided nor absolutely summable. Nonetheless, $x_1 * x_2$ exists. For every $n \in \mathbb{Z}$,

$$x_1 * x_2(n) = \sum_{k=-\infty}^{\infty} u(k) 3^{n-k} = 3^n \sum_{k=0}^{\infty} 3^{-k} = \frac{3}{2} 3^n \; .$$

It's of interest to note that $x_2(n)$ and $x_1 * x_2(n)$ both take the form (constant) $\times\, 3^n$.

CHAPTER 6

Discrete-time LTI Systems

Linear time-invariant systems serve as effective models for a variety of real-world processes that arise in applications. The models are useful not only in electrical and computer engineering — particularly in the areas of signal processing, communication, and control — but in other fields of science and engineering including mechanical engineering, operations research, economics, and even biology. The ideas are more transparent in the context of discrete-time models, which is why we start there.

Definition and examples

As in Chapter 5, the integers $\mathbb{Z}$ model discrete time. A discrete-time signal over $\mathbb{F}$ is a function x with domain $\mathbb{Z}$ that takes values in $\mathbb{F}$. As usual, $\mathbb{F}$ is either $\mathbb{R}$ or $\mathbb{C}$. Alternatively and equivalently, a discrete-time signal is a doubly infinite sequence $\{x(n)\}$ where $x(n) \in \mathbb{F}$ for all $n \in \mathbb{Z}$. Think of $x(n)$ as the value of the signal x at time n. As in Chapter 5, I denote the set of all discrete-time signals by $\mathbb{F}^{\mathbb{Z}}$.

The real-world processes we're interested in modeling take discrete-time input signals and generate discrete-time output signals. An appealing way to represent such a process is as a mapping

$$S : X \longrightarrow \mathbb{F}^{\mathbb{Z}}$$

where X is a subset of $\mathbb{F}^{\mathbb{Z}}$ that represents the set of possible input signals for the system. The idea of the mapping S is that when $x \in X$ is the input signal to the system, $S(x) \in \mathbb{F}^{\mathbb{Z}}$ is the output signal that arises. One usually assumes that the input space X is "rich enough" to include a lot of signals of interest. We'll always require that X contain at least all the finite-duration signals and also that X be closed under shifting in the sense that when $x \in X$ and $k_o \in \mathbb{Z}$, the signal $\mathrm{Shift}_{k_o}(x)$ is also in X.

As we saw in Chapters 4 and 5, $\mathbb{F}^{\mathbb{Z}}$ has a natural vector-space structure with componentwise addition and scalar multiplication. If a system's input set X is closed under the taking of linear combinations — i.e. is a subspace of the vector space $\mathbb{F}^{\mathbb{Z}}$ — and $S : X \to \mathbb{F}^{\mathbb{Z}}$ is a linear mapping, we call the system linear. Furthermore, if the system has the property that shifting its input signal by time-shift k_o always gives rise to the same time-shift k_o in the system's output, we call the system time-invariant. Here is the formal definitions.

6.1 Definition: A *discrete-time input-output linear time-invariant system over* $\mathbb{F}$ consists of the following:

- A subset X of $\mathbb{F}^{\mathbb{Z}}$ representing the system's set of possible input signals. X is a subspace of $\mathbb{F}^{\mathbb{Z}}$ that contains all the finite-duration signals and is shift-invariant in the sense that if $x \in X$ then $\mathrm{Shift}_{k_o}(x) \in X$ for all $k_o \in \mathbb{Z}$.
- A mapping $S : X \to \mathbb{F}^{\mathbb{Z}}$ that is linear, i.e.

$$S(c_1x_1 + c_2x_2) = c_1S(x_1) + c_2S(x_2) \text{ for all } x_1, x_2 \in X \text{ and } c_1, c_2 \in \mathbb{F}$$

and shift-invariant, i.e.

$$S(\mathrm{Shift}_{k_o}(x)) = \mathrm{Shift}_{k_o}(S(x)) \text{ for all } x \in X \text{ and } k_o \in \mathbb{Z} \ .$$

"LTI" always means "linear time-invariant." I'll always use "LTI system" to mean "input-output LTI system." Now for some examples of discrete-time LTI systems. The *zero system* has input space $X = \mathbb{F}^{\mathbb{Z}}$ and system mapping $S : X \to \mathbb{F}^{\mathbb{Z}}$ defined by $S(x) = 0$ for all $x \in X$, where 0 here denotes the zero signal. The *identity system* also has input space $X = \mathbb{F}^{\mathbb{Z}}$, but its system mapping S has specification $S(x) = x$ for all $x \in X$. For any $k_1 \in \mathbb{Z}$, the *pure k_1-shift system* has input space $X = \mathbb{F}^{\mathbb{Z}}$ and system mapping S defined by $S(x) = \mathrm{Shift}_{k_1}(x)$ for every $x \in X$. I'll leave it to you to show that all three of these systems are LTI.

A slightly more interesting example is a system I call the *causal sliding-window M-fold averager.* For this system, the input space X is again $\mathbb{F}^{\mathbb{Z}}$. The mapping $S : X \to \mathbb{F}^{\mathbb{Z}}$ takes each input signal $x \in X$ to the output signal $S(x) \in \mathbb{F}^{\mathbb{Z}}$ whose value at time n is the average of the previous M values of x including $x(n)$. In equation form,

$$S(x)(n) = \frac{1}{M}\sum_{k=0}^{M-1} x(n-k)$$

for every $x \in X$ and $n \in \mathbb{Z}$. The sliding-window averager is ubiquitous in signal-processing applications. It has a way of smoothing out local rapid variations in inputs. You can show fairly easily that the system is LTI. Observe that we could also describe the sliding-window averager's system mapping S via

$$S(x) = \frac{1}{M}\sum_{k=0}^{M-1} \mathrm{Shift}_k(x) \text{ for all } x \in X \ .$$

This "whole-signal" description of S, at least for me, makes it slightly less obvious exactly what the system does to an input signal x.

Another LTI system one encounters frequently in applications is the *discrete-time integrator.* This system takes an input signal x and outputs the signal $S(x)$ with specification

$$S(x)(n) = \sum_{m=-\infty}^{n} x(m) = \sum_{k=0}^{\infty} x(n-k) \text{ for every } n \in \mathbb{Z} \ .$$

The input space X contains precisely all those signals x for which the sums in the last equation converge for every $n \in \mathbb{Z}$. The sum defining $S(x)(n)$ is a bit like a discrete-time version of the "integral of x from time $-\infty$ up to time n," which helps explain the name of the system. Like the sliding-window averager, the discrete-time integrator's system mapping S admits a "whole-signal" description, namely

$$S(x) = \sum_{k=0}^{\infty} \mathrm{Shift}_k(x) \ \text{ for all } \ x \in X \ ,$$

which in my view obscures the system's function.

Removing from Definition 6.1 the linearity and shift-invariance conditions on the mapping S leaves us with the definition of an input-output system that's not necessarily linear or time-invariant. What do such systems look like? Sometimes the lack of linearity and/or time-invariance is obvious. For example, if the system takes any input $x \in \mathbb{F}^{\mathbb{Z}}$ to output $S(x)$ with specification

$$S(x)(n) = \frac{1}{M} \sum_{k=0}^{M-1} x^3(n-k) \ \text{ for all } \ n \in \mathbb{Z} \ ,$$

then linearity clearly fails to hold. On the other hand, if $S(x)$ has specification

$$S(x)(n) = 3 + \frac{1}{M} \sum_{k=0}^{M-1} x(n-k) \ \text{ for all } \ n \in \mathbb{Z} \ ,$$

you might not recognize S as nonlinear right away. But keep in mind that $S(x) = 0$ must hold when S is linear, and $S(0)$ for this example is the constant signal with value 3 for every $n \in \mathbb{Z}$. Both of these systems are time-invariant in the sense that

$$S\left(\mathrm{Shift}_{k_o}(x)\right) = \mathrm{Shift}_{k_o}(S(x))$$

for all k_o and x, as is the system that takes any input signal x and puts out the signal $S(x)$ with specification

$$S(x) = \frac{\max\left(\{x(n-k) : 0 \le k < 5\}\right) + \min\left(\{x(n-k) : 0 \le k < 5\}\right)}{2}$$

for all $n \in \mathbb{Z}$. This system, another kind of sliding-window averager, takes the average of the maximum and minimum input values in a window of length 5 instead of the average of all input values lying in the window. You can see it's not linear by setting $x_1 = \delta + \mathrm{Shift}_1(\delta)$ and $x_2 = \mathrm{Shift}_3(\delta) + \mathrm{Shift}_4(\delta)$ and noting that

$$S(x_1 + x_2)(5) = S(x_1)(5) = S(x_2)(5) = 1/2 \ .$$

What about linear systems for which time-invariance fails? Consider the following modification of the sliding-window 2-fold averager: for every $x \in \mathbb{F}^{\mathbb{Z}}$, $S(x)$ has specification

$$S(x)(n) = \frac{n}{2}\left(x(n) + x(n-1)\right) \ \text{ for all } \ n \in \mathbb{Z} \ .$$

You can check that $S(\delta)(1) = 1/2$ and $S(\delta)(n) = 0$ for all other $n \in \mathbb{Z}$. Meanwhile, $S(\mathrm{Shift}_1)(\delta)$ has specification

$$S(\mathrm{Shift}_1(\delta))(n) = \begin{cases} 1/2 & \text{when } n = 1 \\ 1 & \text{when } n = 2 \\ 0 & \text{otherwise,} \end{cases}$$

so $S(\mathrm{Shift}_1(\delta)) \neq \mathrm{Shift}_1(S(\delta))$ and the system isn't time-invariant, although it's linear. Another linear but not time-invariant system, known as a *decimator,* comes up frequently in digital signal processing. It takes an input signal $x \in \mathbb{F}^{\mathbb{Z}}$ and generates output signal $S(x)$ with specification

$$S(x)(n) = x(7n) \text{ for all } n \in \mathbb{Z} ,$$

where I've chosen 7 just for definiteness — any positive integer would do. The system takes all the input values at times that are multiples of 7 and "compresses" them into the signal $S(x)$. Meanwhile, the system ignores all the other input values. Note that $S(\delta) = \delta$ but $S(\mathrm{Shift}_1(\delta)) = 0$, so time-invariance fails. An easy way to see that the system annihilates $\mathrm{Shift}_1(\delta)$ is to note that $\mathrm{Shift}_1(\delta)(n) = 0$ when n is any multiple of 7. Observe also that

$$S(\mathrm{Shift}_7(x)) = \mathrm{Shift}_1(S(x)) \text{ for all } x \in \mathbb{F}^{\mathbb{Z}} ,$$

so the system maps a shift by 7 in the input to a shift by 1 in the output.

Impulse response and FIR systems

By virtue of Definition 6.1, the input space X of any LTI system contains every finite-duration signal and therefore contains the impulse δ. Thus it makes sense to talk about $h = S(\delta)$, the output that arises when δ is the input. We call h the system's *impulse response* for obvious reasons.

To appreciate the critical importance of a system's impulse response, observe first that you can write any finite-duration signal $x \in \mathbb{F}^{\mathbb{Z}}$ as a finite linear combination of shifted impulses, namely

$$x = \sum_{k=-\infty}^{\infty} x(k)\mathrm{Shift}_k(\delta) ,$$

where the sum has finitely many nonzero terms because x has finite duration. Suppose we use x as input signal for some LTI system with input space X and system mapping $S : X \to \mathbb{F}^{\mathbb{Z}}$. Then

$$\begin{aligned} S(x) &= S\left(\sum_{k=-\infty}^{\infty} x(k)\mathrm{Shift}_k(\delta)\right) \\ &= \sum_{k=-\infty}^{\infty} x(k)S(\mathrm{Shift}_k(\delta)) \\ &= \sum_{k=-\infty}^{\infty} x(k)\mathrm{Shift}_k(S(\delta)) \\ &= \sum_{k=-\infty}^{\infty} x(k)\mathrm{Shift}_k(h) , \end{aligned}$$

where $h = S(\delta)$ is the impulse response of the system. The equality on the second line holds because the sum inside the parentheses on the first line has finitely many terms, so we can, by linearity of the system, interchange S and the linear combination with impunity. Note that the $x(k)$-terms play the role of coefficients in a linear combination of shifted impulse signals — i.e. the shifted impulses are whole signals and the $x(k)$-terms are just numbers. The equality on the third line follows from time-invariance of the system.

Evaluating at time $n \in \mathbb{Z}$ the terms at the beginning and end of the previous equation yields

$$\begin{aligned} S(x)(n) &= \sum_{k=-\infty}^{\infty} x(k)\mathrm{Shift}_k(h)(n) \\ &= \sum_{k=-\infty}^{\infty} x(k)h(n-k) \end{aligned}$$

for every $n \in \mathbb{Z}$. But this is the same as

$$S(x) = h * x \ .$$

In other words, the response of the system to any finite-duration input signal x is the convolution of the system's impulse response h with x. This fact is sufficiently important to dignify as a theorem.

6.2 Theorem: Given a LTI system with input space X and system mapping $S : X \to \mathbb{F}^{\mathbb{Z}}$, let $h = S(\delta)$ be the system's impulse response. Then $S(x) = h * x$ for every finite-duration signal $x \in X$. □

Theorem 6.2 comes close to asserting that every LTI system is "convolutional" in the sense that $S(x) = h * x$ for every input signal $x \in X$. Strictly speaking, however, Theorem 6.2 applies only to finite-duration input signals. As it happens, all the LTI systems we'll encounter in applications will satisfy a stronger "for all $x \in X$ (finite-duration or not)" version of Theorem 6.2. What does it take for a system to be sufficiently well behaved for the stronger result to apply? To get some insight, it's helpful to examine an example system that's not well behaved in this sense.

Let $X \subset \mathbb{F}^{\mathbb{Z}}$ be the set of all signals x for which $\lim_{m\to\infty} x(m)$ exists. X is closed under linear combinations and is therefore a subspace of $\mathbb{F}^{\mathbb{Z}}$. X is also closed under time-shifting, and X contains all the finite-duration signals since each such signal x satisfies $\lim_{m\to\infty} x(m) = 0$. For every $x \in X$, define $S : X \to \mathbb{F}^{\mathbb{Z}}$ to be the constant signal whose value at every $n \in \mathbb{Z}$ is given by

$$S(x)(n) = \lim_{m\to\infty} x(m) \ .$$

S is clearly a linear mapping, and $S(\mathrm{Shift}_{k_o}(x)) = S(x) = \mathrm{Shift}_{k_o}(S(x))$ for every $x \in X$, so we have a LTI system here. Its impulse response h is the constant signal whose value at every time $n \in \mathbb{Z}$ is

$$h(n) = S(\delta)(n) = \lim_{m\to\infty} \delta(m) = 0 \ .$$

In other words, h is the zero signal. Certainly, $S(x) = h * x = 0$ for every finite-duration signal x, but $S(x) \neq h * x$ for any $x \in X$ satisfying $\lim_{m\to\infty} x(m) \neq 0$.

Roughly speaking, what we need to assume about a LTI system for the stronger version of Theorem 6.2 to hold is that the response of the system to an infinite-duration input signal x is, in some sense, the "limit" of the system's response to finite-duration "approximations" of x. I don't want to sweat about the meanings of "approximation" and "limit," but the sort of condition I'm thinking of is along the following lines. For $x \in X$ and positive integers M and N, let $\mathrm{Trunc}_{M,N}(x)$ be the finite-duration signal with specification

$$\mathrm{Trunc}_{M,N}(x)(n) = \begin{cases} x(n) & \text{if } -M \le n \le N \\ 0 & \text{otherwise.} \end{cases}$$

$\mathrm{Trunc}_{M,N}(x)$ is a finite-duration truncation of the signal x. If it is true for every $x \in X$ that

$$\lim_{M,N\to\infty} S\left(\mathrm{Trunc}_{M,N}(x)\right) = S(x) \ ,$$

with an appropriate definition of "limit," then the system will be such that $S(x) = h * x$ for every $x \in X$. Observe that the system in the example above does not satisfy this limiting condition.

Henceforth, we'll assume that all the LTI systems we deal with satisfy the stronger version of Theorem 6.2. In other words, we'll assume always that any system under consideration is such that $S(x) = h * x$ for every input signal $x \in X$, where h is the system's impulse response. We'll assume in addition that the input space X is "as large as possible" in the sense that it contains every signal whose convolution with h exists. Let's denote by $\mathcal{D}_h$ the set of all $x \in \mathbb{F}^{\mathbb{Z}}$ for which the convolution $h * x$ exists. Since we'll stipulate that a LTI system with impulse response h has input space $X = \mathcal{D}_h$, we'd better check to make sure $\mathcal{D}_h$ has the constraints that Definition 6.1 imposes on input spaces.

It's easy to show that $\mathcal{D}_h$ is closed under linear combinations (i.e. is a subspace of $\mathbb{F}^{\mathbb{Z}}$). Furthermore, $\mathcal{D}_h$ contains all the finite-duration signals since by convolution-existence Criterion 5.1 $h * x$ exists for every finite-duration signal x. $\mathcal{D}_h$ is also closed under time shifting. To see this, note that for any $k_o \in \mathbb{Z}$ we have, for every $n \in \mathbb{Z}$,

$$\begin{aligned} h * (\mathrm{Shift}_{k_o}(x))(n) &= \sum_{k=-\infty}^{\infty} h(k)\mathrm{Shift}_{k_o}(x)(n-k) \\ &= \sum_{k=-\infty}^{\infty} h(k)x(n-k-k_o) \\ &= \sum_{k=-\infty}^{\infty} h(k)x((n-k_o)-k) \\ &= h * x(n-k_o) \\ &= \mathrm{Shift}_{k_o}(h * x)(n) \ . \end{aligned}$$

Since $x \in \mathcal{D}_h$, $h * x$ exists, so $\mathrm{Shift}_{k_o}(h * x)(n)$ is well defined for every $n \in \mathbb{Z}$. The chain of equalities above demonstrates that $h * (\mathrm{Shift}_{k_o}(x))(n)$ is also well defined for every $n \in \mathbb{Z}$. We conclude that $\mathrm{Shift}_{k_o}(x) \in \mathcal{D}_h$ if $x \in \mathcal{D}_h$. Since $\mathcal{D}_h$ is closed under linear combination and time-shifting and includes all the finite-duration signals, it is a suitable input space for a LTI system. Let's formalize the foregoing discussion as follows.

6.3 Standing Assumption: Every LTI system we deal with has system mapping S specified by $S(x) = h * x$ for every input signal x, where $h = S(\delta)$ is the system's impulse response. Furthermore, the system's input space X is $\mathcal{D}_h$, the set of all $x \in \mathbb{F}^{\mathbb{Z}}$ for which $h * x$ exists.

The impulse response h, as a consequence, tells the entire story about any system we care to study. The system's input space is $X = \mathcal{D}_h$, and finding $S(x)$ for any input x simply requires convolving h with x. Not only is h the "response of the system to an impulse," it's "what you convolve with inputs to get outputs." Observe that Standing Assumption 6.3 has a flip side of sorts. If $h \in \mathbb{F}^{\mathbb{Z}}$ is any signal, you can define a LTI system satisfying Standing Assumption 6.3 as follows: let the system's input space X be $\mathcal{D}_h$ and define the system mapping $S : \mathcal{D}_h \to \mathbb{F}^{\mathbb{Z}}$ via $S(x) = h * x$ for every $x \in X$. The system so constructed has impulse response h since $S(\delta) = h * \delta = h$, the last equality holding because δ serves as an identity element for convolution.

Let's figure out the impulse responses of our example systems. In each case, we use the definition of the system mapping S to compute $h = S(\delta)$. The zero system's impulse response is clearly $h = 0$, the zero signal. For the identity system, since $S(x) = x$ for every input signal x, it follows that $S(\delta) = \delta$. Thus the impulse response of the identity system is $h = \delta$. Since for the pure k_1-shift system we have $S(x) = \text{Shift}_{k_1}(x)$ for every x, $h = S(\delta) = \text{Shift}_{k_1}(\delta)$. In other words, h is the signal with specification

$$h(n) = \begin{cases} 1 & \text{if } \ n = k_1 \\ 0 & \text{otherwise.} \end{cases}$$

The causal sliding-window M-fold averager is a little more interesting. Its impulse response $h = S(\delta)$ is the signal whose value at time n is

$$h(n) = \frac{1}{M} \sum_{k=0}^{M-1} \delta(n-k)$$

for every $n \in \mathbb{Z}$. Thinking about δ, you see that $h(n)$ is equal to $1/M$ precisely when the value of n is such that $k = n$ lies in the range of the sum on the right-hand side; otherwise, $h(n) = 0$. Accordingly,

$$h(n) = \begin{cases} \frac{1}{M} & \text{when } \ 0 \le n \le M-1 \\ 0 & \text{otherwise.} \end{cases}$$

For the discrete-time integrator, $h = S(\delta)$ is the signal whose value at time n is

$$h(n) = \sum_{k=-\infty}^{n} \delta(k)$$

for every $n \in \mathbb{Z}$. Because $\delta(k) = 0$ for every $k < 0$, $h(n) = 0$ for every $n < 0$. Because $\delta(0) = 1$ and $\delta(k) = 0$ for $k > 0$, $h(n) = 1$ for every $n \ge 0$, since for each such n the sum defining $h(n)$ contains exactly one 1 and all the rest zeroes. Accordingly,

$$h(n) = \begin{cases} 1 & \text{for } n \ge 0 \\ 0 & \text{for } n < 0 \, . \end{cases}$$

Alternatively, h is the discrete-time unit step u.

Note that the impulse response of each of the first four example systems is actually a finite-duration signal. A discrete-time I/O system with a finite-duration impulse response is an *FIR system.* The abbreviation stands for "finite impulse response" and the terminology is standard, although I'm not fond of it. I'd prefer

"FDIR" (for finite-duration impulse response), but whatever. Now for a simple observation: if h has finite duration, then $h * x$ exists for every $x \in \mathbb{F}^{\mathbb{Z}}$ by Criterion 5.1. Accordingly, $\mathcal{D}_h = \mathbb{F}^{\mathbb{Z}}$, so any FIR system admits any signal $x \in \mathbb{F}^{\mathbb{Z}}$ as an input. For FIR systems, then, the set X of all admissible inputs is $\mathbb{F}^{\mathbb{Z}}$ itself.

Causality and BIBO stability

Since the impulse response h determines the entire input-output behavior of a LTI system satisfying Standing Assumption 6.3, one might expect that important "system properties" have embodiments as "properties of the signal h," and that is indeed the case. Two instances of this correspondence arise in relation to the system properties of causality and bounded-input bounded-output stability.

A discrete-time LTI system is causal if, roughly speaking, the current value of the output signal depends only on the current and past values of the input signal and not on future values of the input signal. Technically, we need to do a little better than that.

6.4 Definition: A LTI system $S : X \to \mathbb{F}^{\mathbb{Z}}$ is *causal* when for every $n \in \mathbb{Z}$ the following holds: if x_1 and x_2 are two input signals in X such that $x_1(k) = x_2(k)$ for every $k \leq n$, then $S(x_1)(k) = S(x_2)(k)$ for every $k \leq n$.

In other words, when a system is causal and two input signals "agree" up to and including time n, the outputs to which they give rise will also "agree" up to and including time n — and that assertion holds for every $n \in \mathbb{Z}$. It's easy to prove a condition on a system's impulse response h that holds if and only if the system is causal.

6.5 Theorem: A LTI system is causal if and only if its impulse response h satisfies $h(n) = 0$ for $n < 0$.

Proof: First, suppose a given system is causal. Since the system is linear, $S(0) = 0$, where 0 stands for the zero signal. Now, $\delta(m) = 0$ for every $m < 0$, so δ agrees with the all-zero signal up to and including time $n = -1$. Since the system is causal, its impulse response $h = S(\delta)$ must agree with $S(0)$ up to and including time $n = -1$, so

$$h(n) = 0 \text{ for } n < 0 .$$

Conversely, suppose $h(m) = 0$ for every $m < 0$. Given any $n \in \mathbb{Z}$ and any $x \in X$,

$$S(x)(n) = h * x(n) = \sum_{k=-\infty}^{\infty} h(n-k)x(k) = \sum_{k=-\infty}^{n} h(n-k)x(k) ,$$

where the last equality holds because $h(n-k) = 0$ when $k > n$. Consequently, the output at time n in response to input x depends only on the values of $x(k)$ for $k \leq n$ and not on the values of $x(k)$ for $k > n$. This condition holds for every

$x \in X$ and every $n \in \mathbb{Z}$, so if x_1 and x_2 are two input signals that agree up to and including time n, then $S(x_1)$ and $S(x_2)$ must also agree up to and including time n. It follows that the system is causal. □

Theorem 6.5 makes total sense when you think about it. Forcing a system with an impulse δ amounts to giving the system a little "kick" at time $n = 0$ and doing nothing to the system before or after that time. If the system's impulse response h has a nonzero value for some negative time — e.g., if $h(-17) = 3$ — then somehow the system must anticipate the upcoming kick. A system that responds in anticipation of future inputs is not causal.

The zero system, the identity system, and the causal sliding-window M-fold averager are all causal LTI systems. It's intuitively clear that all those systems satisfy the informal definition of causality. Each system's "current output value" depends explicitly on "current and/or past input values" and not on "future input values." (That dependence is trivial in the case of the zero system.)

Is the k_1-shift system causal? Well, it depends. If $k_1 \geq 0$, then for every $n \in \mathbb{Z}$

$$S(x)(n) = \mathrm{Shift}_{k_1}(x)(n) = x(n - k_1)$$

depends only on values of $x(k)$ for $k \leq n$, so the system is causal. The shift system acts as a pure delay in this case. If $k_1 < 0$, then $x(n - k_1)$ depends on $x(m)$ for $m > n$, so the system is not causal. In this case the system acts as a pure predictor. Note also that since the impulse response of the shift system is $\mathrm{Shift}_{k_1}(\delta)$, Theorem 6.5 tells us immediately that the shift system is causal if and only if $k_1 \geq 0$. I'll add that the badly behaved LTI system I introduced in the run-up to Standing Assumption 6.3 is not causal. It would have been causal had it used $\lim_{n\to-\infty}$ instead of $\lim_{n\to\infty}$.

How might non-causal systems be relevant to applications? Suffice it to say that if the index n on our discrete-time signals denotes some kind of truly temporal variable, where $x(n)$ actually "comes after" $x(k)$ when $n > k$, then a non-causal system is not physically realizable, at least given our present understanding of how the universe operates. Nonetheless, many applications involve discrete-time LTI system models wherein the ostensible "time index" n is not temporal at all. Image processing and data post-processing give rise to such applications. In those contexts, non-causal LTI systems play important roles.

What about stability? Roughly speaking, a system is stable if nothing crazy happens when you drive the system with well behaved inputs. As a technical definition, of course, that won't do. People have settled on a notion of stability that's appropriate to discrete-time LTI systems. It goes essentially like this: a system is stable if every bounded signal $x \in \mathbb{F}^{\mathbb{Z}}$ is an admissible input for the system and if, in addition, the output $S(x)$ arising from such a bounded input signal x is also a bounded signal.

6.6 Definition: A LTI system with input space X and system mapping $S : X \to \mathbb{F}^{\mathbb{Z}}$ is *bounded-input bounded-output stable* or *BIBO stable* when X contains all the bounded signals in $\mathbb{F}^{\mathbb{Z}}$ (i.e. $l^\infty \subset X$) and, for every bounded $x \in X$, $S(x)$ is also a bounded signal.

Like causality, BIBO stability of a system has a neat characterization in terms of the system's impulse response. I'll prove a weak version of this characterization first, and then state a stronger version without proof.

6.7 Theorem: A LTI system with impulse response h is BIBO stable if and only if h is absolutely summable — i.e., if and only if $h \in l^1$.

Proof: Assume first that h is absolutely summable. By Criterion 5.3, $h * x$ exists for every bounded signal $x \in \mathbb{F}^{\mathbb{Z}}$ and, furthermore, is a bounded signal satisfying

$$\|h * x\|_\infty \leq \|h\|_1 \, \|x\|_\infty \; .$$

In particular, every bounded x is in $\mathcal{D}_h$ and, since $\mathcal{D}_h = X$ by Standing Assumption 6.3, every bounded signal is an admissible input to the system. $h * x$ is just $S(x)$, so it follows that every bounded x leads to an output $S(x)$ that is also bounded. In fact,

$$\|S(x)\|_\infty \leq \|h\|_1 \, \|x\|_\infty \; .$$

Conversely, suppose h is not absolutely summable. I'll construct a bounded input signal x for which $h * x$ fails to exist, which contradicts BIBO stability of the system. Define

$$x(m) = \begin{cases} \overline{h(-m)}/|h(-m)| & \text{when } h(m) \neq 0 \\ 0 & \text{when } h(m) = 0 \; , \end{cases}$$

where the overbar denotes complex conjugate. Note that when $h(-m)$ is real-valued, $\overline{h(-m)}/|h(-m)| = \operatorname{sgn}(h(-m))$, where sgn means "sign". The signal x is bounded; in fact, $\|x\|_\infty = 1$ provided $h \neq 0$. I claim that $x \notin \mathcal{D}_h$, which means that x is not admissible as an input to the system. To see why, attempt to compute $S(x)(0)$. You get

$$\begin{aligned} S(x)(0) &= h * x(0) \\ &= \sum_{k=-\infty}^{\infty} h(k)x(0-k) \\ &= \sum_{\{k:h(k)\neq 0\}} h(k)\overline{h(k)}/|h(k)| \\ &= \sum_{k=-\infty}^{\infty} |h(k)| \; , \end{aligned}$$

and the last series fails to converge because h isn't absolutely summable. Since x is a bounded input and is inadmissible as an input to the system, the system is not BIBO stable. □

It turns out that the following significantly stronger version of Theorem 6.7 holds. I won't prove it, but I'll attempt to explain what makes it difficult to prove.

6.8 Theorem: A LTI system with input space X and system mapping S is BIBO stable if and only if every bounded right-sided signal is in X and $S(x)$ is a bounded signal for every bounded right-sided signal x. □

What makes Theorem 6.8 stronger than Theorem 6.7? It states that to check for BIBO stability (or, equivalently, for absolute summability of h), we need only make sure that $h * x$ is bounded for every bounded *right-sided* signal x. If so, we can conclude that $h * x$ is bounded for *all* bounded signals x, right-sided or not. This is an important gloss on the BIBO stability concept that plays a role in more advanced applications.

The hard part of proving Theorem 6.8 is showing that if $h * x$ exists and is bounded for every bounded right-sided signal x, then $h \in l^1$. We can construct a bounded right-sided signal resembling x in the proof of Theorem 6.7 that enables us to conclude that $\sum_{n=-\infty}^{0} |h(n)|$ must converge for $h * x$ to exist. It's trickier to prove that $\sum_{k=0}^{\infty} |h(k)|$ converges when $h * x$ exists for all bounded right-sided signals x. Suppose, for example, that the system under consideration is causal, in which case h is right-sided by Theorem 6.5. Criterion 5.2 tells us that $h * x$ exists for every right-sided signal x, so there's no way to build a bounded right-sided signal x for which $h * x$ fails to exist when $\sum_{k=0}^{\infty} |h(k)|$ diverges. Finishing the proof requires a sophisticated result from functional analysis called the Uniform Boundedness Theorem.

Let's wrap things up by checking for BIBO stability of the various example systems. The first three systems are easy for the following reason: from Theorem 6.7 it follows that every FIR system is BIBO stable, since every such system has a trivially absolutely summable impulse response. Accordingly, the zero and identity systems, the shift system(s), and the causal sliding-window M-fold averager are all BIBO stable. You can understand this at a more elementary level just by contemplating the definition of BIBO stability and asking whether a bounded input signal leads to a bounded output signal for each of these systems. In each case, the answer is fairly obviously Yes.

The discrete-time integrator, on the other hand, is not BIBO stable. Consider driving the system with bounded input signal $x = u$. You discover that

$$S(u)(n) = \sum_{k=-\infty}^{n} u(k) = \begin{cases} n+1 & \text{if } n \geq 0 \\ 0 & \text{if } n < 0 \, . \end{cases}$$

Alternatively, $S(u)(n) = (n+1)u(n)$ for every $n \in \mathbb{Z}$, so $S(u)$ is not bounded even though u is.

CHAPTER 7

Continuous-time Signals and Convolution

Once upon a time, signals and systems classes dealt exclusively with continuous-time signals. With the growing ascendancy of computers during the twentieth century, discrete-time signals and relationships between discrete- and continuous-time signals began achieving prominence. That development proved beneficial to pedagogy because the mathematically cleaner discrete-time arena provides a friendlier setting in which to engage key concepts. I have tried with Chapters 5 and 6 to set the stage for the material in this chapter and the next. All the central ideas from the earlier chapters have continuous-time analogues, and the way the theory unfolds might provoke a bit of *déjà vu.* Nonetheless, complications abound in continuous time. Ironing out every wrinkle is impossible at this level, but we can come close.

Decent signals

We view the real numbers $\mathbb{R}$ as a mathematical model for "continuous time." Real number t corresponds to "time t." Real number 0 corresponds to "time 0." If $s > t$, then "time s is later than time t." Having used $\mathbb{R}$ to model continuous time, let's define an $\mathbb{F}$*-valued continuous-time signal* as a function with domain $\mathbb{R}$ that takes values in $\mathbb{F}$ — i.e., a continuous-time signal is a mapping $x : \mathbb{R} \to \mathbb{F}$. As usual, $\mathbb{F}$ is either $\mathbb{R}$ or $\mathbb{C}$. We denote the set of all continuous-time signals by $\mathbb{F}^{\mathbb{R}}$.

Working with continuous-time signals is touchier than working with discrete-time signals. All sorts of worrisome analytical issues arise involving, for example, continuity and differentiability. Some continuous-time signals are quite nasty. Fortunately, those signals play a limited role in applications, and we will end up essentially wishing them out of the picture by restricting our attention largely to what I'll be calling decent signals.

7.1 Definition: A *decent signal* is a signal $x \in \mathbb{F}^{\mathbb{R}}$ that has the following properties:

(1) x is continuous except possibly for jump discontinuities.
(2) x has at most finitely many jumps in any bounded interval $[t_1, t_2] \subset \mathbb{R}$.
(3) x is bounded on any bounded interval $[t_1, t_2] \subset \mathbb{R}$. I.e., for any such interval, there exists $R > 0$ such that $|x(t)| \leq R$ for every $t \in [t_1, t_2]$.

Requirement 1 eliminates a number of signals we'd rather not deal with. Consider, for example, the signal $x_1 \in \mathbb{R}^{\mathbb{R}}$ with specification

$$x_1(t) = \begin{cases} 0 & \text{if } t \text{ when rational} \\ 1 & \text{if } t \text{ when irrational.} \end{cases}$$

The signal x_1 has no points of continuity. In any bounded interval, x_1 has uncountably many "jumps," if you even want to call them jumps. Requirement 2 rules out the slightly less pathological signal x_2 with specification

$$x_2(t) = \begin{cases} 0 & \text{when } t < 0 \\ 0 & \text{when } 1/(n+1) < t \le 1/n \text{ and } n \text{ is even} \\ 1 & \text{when } 1/(n+1) < t \le 1/n \text{ and } n \text{ is odd} \\ 1 & \text{when } t \ge 1 \text{ .} \end{cases}$$

x_2 has countably infinitely many jumps in the interval $[0, 1]$, but is flat between the jumps. Requirement 3 eliminates signals that blow up somewhere other than as $t \to \pm\infty$. Consider, for example, the signal x_3 with specification

$$x_3(t) = \begin{cases} \frac{1}{t} & \text{when } t \neq 0 \\ 0 & \text{when } t = 0 \text{ .} \end{cases}$$

$|x_3(t)|$ blows up as $t \to 0$ from either side, so x is unbounded on any interval containing time $t = 0$.

Every continuous signal is a decent signal. Many discontinuous signals are decent as well, including some that we'll encounter frequently. The *continuous-time unit step* is the signal u that has specification

$$u(t) = \begin{cases} 1 & \text{when } t \ge 0 \\ 0 & \text{when } t < 0 \text{ .} \end{cases}$$

I hope you don't object to my using the same notation u for both the discrete- and continuous-time unit steps. u is not continuous, but its only discontinuity is a jump at $t = 0$. One comment on the unit step: I've defined it so $u(0) = 1$. As it happens, in all the manipulations we do that involve decent signals, it will *not matter at all* how we define the signals' values at jump-discontinuity points. I could just as well have set $u(0) = 0$ or even $u(0) = 1/2$. As you'll see, it will always be safe to regard two decent signals that agree everywhere except possibly at jumps as constituting, in every practical sense, "the same signal."

Another discontinuous decent signal that we'll be using often enough so it deserves a special name is the signal p_a with specification

$$p_a(t) = \begin{cases} 1 & \text{when } -a/2 \le t < a/2 \\ 0 & \text{otherwise .} \end{cases}$$

If you graph $p_a(t)$ against t, you'll see why we call p_a a rectangular pulse with height 1 and width a centered at $t = 0$. Once again, it doesn't matter how we define $p_a(t)$ at $t = \pm a/2$. The way I've specified p_a and u makes it true that

$$p_a(t) = u(t + \frac{a}{2}) - u(t - \frac{a}{2}) \text{ for all } t \in \mathbb{R} \text{ .}$$

Like $\mathbb{F}^{\mathbb{Z}}$, $\mathbb{F}^{\mathbb{R}}$ has a natural vector space structure. The zero vector in $\mathbb{F}^{\mathbb{R}}$ is the zero signal, and for signals x_1 and x_2 and constants c_1 and c_2 in $\mathbb{F}$, the linear combination $y = c_1x_1 + c_2x_2$ is the signal with specification

$$y(t) = c_1x_1(t) + c_2x_2(t) \text{ for all } t \in \mathbb{R} \text{ .}$$

Fortunately, the set of decent signals is closed under taking such linear combinations, and therefore forms a subspace of $\mathbb{F}^{\mathbb{R}}$.

Associated with any $t_o \in \mathbb{R}$ is the time-shift mapping $\mathrm{Shift}_{t_o} : \mathbb{F}^{\mathbb{R}} \to \mathbb{F}^{\mathbb{R}}$ defined by

$$\mathrm{Shift}_{t_o}(x)(t) = x(t - t_o)$$

for every signal $x \in \mathbb{F}^{\mathbb{R}}$ and every $t \in \mathbb{R}$. Note that these shift mappings are linear mappings on $\mathbb{F}^{\mathbb{R}}$. Also, conveniently, the set of decent signals is closed under shifting.

Now for the natural continuous-time versions of finite duration and right- and left-sidedness. A signal x is *right-sided* when there exists T_1 such that $x(t) = 0$ for $t < T_1$. A signal x is *left-sided* when there exists T_2 such that $x(t) = 0$ for $t > T_2$. A signal $x \in \mathbb{F}^{\mathbb{R}}$ has *finite duration* when there exist T_1 and T_2 such that $x(t) = 0$ for $t < T_1$ and for $t > T_2$. Clearly, a signal has finite duration if and only if it is both right- and left-sided. The set of right-sided signals, the set of left-sided signals, and the set of finite-duration signals are all subspaces of $\mathbb{F}^{\mathbb{R}}$ and are all closed under shifting.

Bounded and absolutely integrable signals: the spaces L^∞ and L^1

Integrals, as you might expect, play a central role in continuous-time signal analysis. By focusing largely on decent signals, we avoid most of the nagging problems associated with reconciling various notions of integration. In particular, for decent signals, Riemann and Lebesgue integration are essentially identical. Here is one nice property of decent signals that we'll use frequently.

7.2 Fact If $x \in \mathbb{F}^{\mathbb{R}}$ is a decent signal, then the integral of x over any bounded interval is well defined and finite. In other words, for any real numbers T_1 and T_2,

$$\int_{T_1}^{T_2} x(t)dt$$

is well defined and finite.

Several important facts about infinite sums have continuous-time versions that pertain to integrals over unbounded intervals. A signal $x \in \mathbb{F}^{\mathbb{R}}$ is *integrable* when both of the limits

$$\lim_{T\to\infty} \int_0^T x(t)dt \quad \text{and} \quad \lim_{S\to\infty} \int_{-S}^0 x(t)dt$$

exist, in which case we define

$$\int_{-\infty}^{\infty} x(t)dt = \lim_{S\to\infty} \int_{-S}^0 x(t)dt + \lim_{T\to\infty} \int_0^T x(t)dt \; .$$

A signal $x \in \mathbb{F}^{\mathbb{R}}$ is *absolutely integrable* when $\int_{-\infty}^{\infty} |x(t)|dt$ exists. The *1-norm* of an absolutely integrable signal x is defined by

$$\|x\|_1 = \int_{-\infty}^{\infty} |x(t)|dt \; .$$

I'll use the notation L^1 for the set of all absolutely integrable signals in $\mathbb{F}^{\mathbb{R}}$. Fact 3.3 has the following continuous-time analogue.

7.3 Fact If $x \in \mathbb{F}^{\mathbb{R}}$ is absolutely integrable, then x is also integrable. As a result, x is integrable if there exists $R > 0$ such that for every $T > 0$ we have

$$\int_{-T}^{T} |x(t)| dt \leq R \, .$$

A signal $x \in \mathbb{F}^{\mathbb{R}}$ is *bounded* when there exists $R > 0$ such that $|x(t)| \leq R$ for every $t \in \mathbb{R}$. The *infinity norm* of a bounded signal x is defined by

$$\|x\|_\infty = \sup\{|x(t)| : t \in \mathbb{R}\} \, .$$

I'll use the notation L^∞ for the set of all bounded signals in $\mathbb{F}^{\mathbb{R}}$.

Before we get too far along, I'd like to point out that my terminology and notation are not quite standard. First of all, nobody talks about "decent signals." I made up that nomenclature so we'd have a single word to encapsulate the conditions in Definition 7.1. Second, the official technical definitions of L^1 and L^∞ depend on sophisticated concepts from Lebesgue integration and measure theory. (The "L" happens to stand for "Lebesgue.") It's impossible to elucidate all these details given the tools currently at our disposal. Please bear with me for now, and then take a real-analysis course if you'd like to find out what you're missing.

The sets L^1 and L^∞ as I have defined them are both closed under linear combination and are therefore subspaces of $\mathbb{F}^{\mathbb{R}}$. Furthermore, both L^1 and L^∞ are closed under time-shifting. Recall that in discrete time we had for every p with $1 < p < \infty$ the set l^p of discrete-time signals for which

$$\sum_{n=-\infty}^{\infty} |x(n)|^p$$

converges. The continuous-time analogues are the sets L^p of signals for which

$$\int_{-\infty}^{\infty} |x(t)|^p dt$$

exists. Of particular importance is the set L^2 of square-integrable signals. Like L^1 and L^∞, L^2 is a subspace of $\mathbb{F}^{\mathbb{R}}$ and is closed under time-shifting. The *2-norm* of a signal $x \in L^2$ is defined by

$$\|x\|_2 = \left(\int_{-\infty}^{\infty} |x(t)|^2 dt \right)^{1/2} \, .$$

Recall that in discrete time we have $l^1 \subset l^\infty$ — i.e., every absolutely summable signal is bounded. More generally, it turns out that

$$l^1 \subset l^p \subset l^\infty$$

for every $p > 1$. In continuous time, no such nice relationships hold between the L's. Consider the signal x with specification

$$x(t) = \begin{cases} 0 & \text{if } t \leq 0 \\ 1/\sqrt{t} & \text{if } 0 < t \leq 1 \\ 1/t^2 & \text{if } t > 1 \ . \end{cases}$$

Note that x is absolutely integrable because

$$\int_{-\infty}^{\infty} |x(t)|dt = \int_0^1 1/\sqrt{t}dt + \int_1^{\infty} 1/t^2 dt = 3 \ .$$

Clearly, x is not bounded because $x(t)$ blows up as t approaches 0 from above. Furthermore, x is not square integrable because

$$\int_0^1 (1/\sqrt{t})^2 dt = \int_0^1 (1/t)dt = \ln(t)]_0^1 = \infty \ .$$

Observe that the example signal x is not a decent signal because it is unbounded on any interval that contains $t = 0$. Is it true, you might ask, that every decent L^1-signal is also in L^∞? The answer is No, and here's an example. Let x be the signal with the following specification: $x(t) = 0$ for all t except for t-values that lie in narrow intervals around nonzero integer values of t. Specifically, $x(t) = 0$ except that

$$x(t) = 3^{|n|} \ \text{ when } \ n - \left(3^{-2|n|}/2\right) \leq t \leq n + \left(3^{-2|n|}/2\right)$$

for every nonzero $n \in \mathbb{Z}$. If you graph $x(t)$ against t, it looks like a bunch of rectangular pulses centered on nonzero integer t-values whose widths narrow and heights increase as $|t|$ increases (see Figure 1). This x is a decent signal. Moreover, $x \in L^1$ because

$$\int_{-\infty}^{\infty} |x(t)|dt = \sum_{n=-\infty}^{-1} (3^{-n})(3^{2n}) + \sum_{n=1}^{\infty} (3^n)(3^{-2n}) = 2\sum_{n=1}^{\infty} 3^{-n} = 1 \ .$$

The point is that the area under the nth pulse is $3^{-|n|}$ for every $n \in \mathbb{Z}$. However, it's clear that x is not bounded, so $x \notin L^\infty$. Nor is x an L^2 signal because

$$\int_{-\infty}^{\infty} |x(t)|^2 dt = \sum_{n=-\infty}^{-1} (3^{-n})^2(3^{2n}) + \sum_{n=1}^{\infty} (3^n)^2(3^{-2n}) = 2\sum_{n=1}^{\infty} 1 = \infty \ .$$

Convolution

Given two signals x_1 and x_2 in $\mathbb{F}^{\mathbb{R}}$, the *convolution of* x_1 *and* x_2, if it exists, is the signal $x_1 * x_2 \in \mathbb{F}^{\mathbb{R}}$ with specification

$$(5) \qquad x_1 * x_2(t) = \int_{-\infty}^{\infty} x_1(\tau)x_2(t-\tau)d\tau \ \text{ for all } \ t \in \mathbb{R} \ .$$

Alternative terminologies for the convolution of x_1 and x_2 are "the convolution of x_1 *with* x_2" and "x_1 convolved with x_2." For our purposes, the convolution of x_1 and x_2 exists precisely when the integral in (5) exists for every $t \in \mathbb{R}$, which is the same as saying that for all $t \in \mathbb{R}$ the function

$$\tau \mapsto x_1(\tau)x_2(t-\tau)$$

is an integrable function of τ.

Let's begin with an elementary observation about convolution. If $x_1 * x_2$ exists, we can change the variable of integration in (5) as follows:

$$x_1 * x_2(t) = \int_{-\infty}^{\infty} x_1(\tau)x_2(t-\tau)d\tau = \int_{-\infty}^{\infty} x_1(t-\zeta)x_2(\zeta)d\zeta = \int_{-\infty}^{\infty} x_1(t-\tau)x_2(\tau)d\tau \ .$$

Setting $\zeta = t - \tau$ yields the first equality. To get the second, re-name the "dummy variable of integration" ζ as τ. The bottom line is that on the right-hand side of equation (5), it doesn't matter where we put the τ and where we put the $(t-\tau)$ — the result is the same. As in discrete time, one could add a bit of window dressing by proclaiming that convolution, defined by (5), is a commutative operation in the sense that if $x_1 * x_2$ exists, then $x_1 * x_2 = x_2 * x_1$.

A slightly less elementary observation about convolution is that it is an *associative* operation in the sense that if $x_1 * (x_2 * x_3)$ exists, then so does $(x_1 * x_2) * x_3$, and vice versa, and both convolutions are the same. Proving this fact is an exercise in manipulating integrals. I'll be a bit casual about interchanging orders of integration here. The interchanges are legal since all the integrals exist. Assuming that $x_1 * (x_2 * x_3)$ exists, we have

$$\begin{aligned}
x_1 * (x_2 * x_3)(t) &= \int_{-\infty}^{\infty} x_1(\tau)(x_2 * x_3(t-\tau))d\tau \\
&= \int_{-\infty}^{\infty} x_1(\tau)\left(\int_{-\infty}^{\infty} x_2(\zeta)x_3((t-\tau)-\zeta)d\zeta\right)d\tau \\
&= \int_{-\infty}^{\infty} x_1(\tau)\left(\int_{-\infty}^{\infty} x_2((t-\tau)-\zeta)x_3(\zeta)d\zeta\right)d\tau \\
&= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} x_1(\tau)x_2((t-\zeta)-\tau)d\tau\right)x_3(\zeta)d\zeta \\
&= \int_{-\infty}^{\infty}(x_1 * x_2(t-\zeta))x_3(\zeta)d\zeta \\
&= (x_1 * x_2) * x_3(t) \ .
\end{aligned}$$

The equalities hold for every $t \in \mathbb{R}$, so

$$x_1 * (x_2 * x_3) = (x_1 * x_2) * x_3 \ .$$

Note: to get the third equality in the chain above, I used the "commutativity" of convolution that I alluded to earlier, which allowed me to switch the roles of $(t-\tau)-\zeta$ and ζ. Convolution is also *bilinear* in the sense that

$$x_1 * (c_2x_2 + c_3x_3) = c_2x_1 * x_2 + c_3x_1 * x_3$$

and

$$(c_1x_1 + c_2x_2) * x_3 = c_1x_1 * x_3 + c_2x_2 * x_3$$

for every c_1, c_2, and c_3 in $\mathbb{F}$ provided all the indicated convolutions exist.

Criteria for existence of convolutions

Given x_1 and x_2 in $\mathbb{F}^{\mathbb{R}}$, how can we tell whether $x_1 * x_2$ exists, aside from checking for convergence of an infinite number of integrals in equation (5)? In what follows, I'll state and prove several useful criteria for the existence of $x_1 * x_2$. Before doing that, I'll show you an example of signals x_1 and x_2 whose convolution fails to exist. Let x_1 be the constant signal whose value is 17 for every $t \in \mathbb{R}$, i.e.

$$x_1(t) = 17 \quad \text{for all } t \in \mathbb{R} \ .$$

Let $x_2 = u$, the continuous-time unit step, which has specification

$$u(t) = \begin{cases} 1 & \text{if } \ t \geq 0 \\ 0 & \text{if } \ t < 0 \ . \end{cases}$$

Attempting to compute $x_1 * x_2(t)$ using (5) leads to

$$x_1 * x_2(t) = \int_{-\infty}^{\infty} x_1(\tau) x_2(t-\tau) d\tau = \int_{-\infty}^{\infty} 17 u(t-\tau) d\tau = \int_{-\infty}^{t} 17 d\tau = \infty \ .$$

The third equality holds because $u(t-\tau) = 0$ when $\tau > t$ and $u(t-\tau) = 1$ when $\tau \leq t$. So much for the possibility of $x_1 * x_2$'s nonexistence. As in discrete time, certain conditions on x_1 and x_2 guarantee that $x_1 * x_2$ exists. Assuming x_1 and x_2 are decent makes things work.

7.4 Criterion If x_1 and x_2 are decent signals and either has finite duration, then $x_1 * x_2$ exists. If both x_1 and x_2 have finite duration, then $x_1 * x_2$ also has finite duration.

Proof: In this case, the integral in (5) is over a bounded τ-interval for every $t \in \mathbb{R}$, which means convergence is not an issue in view of Fact 7.2. To see this, suppose that x_1 has finite duration and that $x_1(t) = 0$ when $t < T_1$ and when $t > T_2$. Then

$$\int_{-\infty}^{\infty} x_1(\tau) x_2(t-\tau) d\tau = \int_{T_1}^{T_2} x_1(\tau) x_2(t-\tau) d\tau$$

for every $t \in \mathbb{R}$. A similar argument applies when x_2 has finite duration. The bottom line is that the integral in (5) converges for every $t \in \mathbb{R}$, so $x_1 * x_2$ exists.

Suppose that both x_1 and x_2 have finite duration; specifically, assume $x_1(t) = 0$ when $t < T_1$ and when $t > T_2$ and that $x_2(t) = 0$ when $t < S_1$ and when $t > S_2$. We still have

$$x_1 * x_2(t) = \int_{-\infty}^{\infty} x_1(\tau) x_2(t-\tau) d\tau = \int_{T_1}^{T_2} x_1(\tau) x_2(t-\tau) d\tau \ .$$

If $t < T_1 + S_1$, then $t - \tau < S_1$ for every τ in the range of integration, which means that $x_2(t-\tau) = 0$ for all such τ and the integral is therefore zero. In other words, $x_1 * x_2(t) = 0$ when $t < T_1 + S_1$. Similarly, if $t > T_2 + S_2$, then $t - \tau > S_2$ for all τ in the range of integration, meaning that $x_2(t-\tau) = 0$ for all such τ and the integral is zero once again, implying that $x_1 * x_2(t) = 0$ when $t > T_2 + S_2$. It follows that $x_1 * x_2$ has finite duration. □

7.5 Criterion: If x_1 and x_2 are decent signals that are both right-sided or both left-sided, then $x_1 * x_2$ exists. Furthermore, in this case $x_1 * x_2$ has the same "sidedness" as x_1 and x_2.

Proof: I'll present the argument only in the case when both signals are right-sided; the left-sided version is similar. Suppose, then, that $x_1(t) = 0$ when $t < T_1$ and $x_2(t) = 0$ when $t < S_1$. Then

$$\begin{aligned}\int_{-\infty}^{\infty} x_1(\tau)x_2(t-\tau)d\tau &= \int_{T_1}^{\infty} x_1(\tau)x_2(t-\tau)d\tau \\ &= \begin{cases} \int_{T_1}^{t-S_1} x_1(\tau)x_2(t-\tau)d\tau & \text{if } t \geq T_1 + S_1 \\ 0 & \text{if } t < T_1 + S_1 \ . \end{cases}\end{aligned}$$

The first equality holds because $x_1(t) = 0$ when $t < T_1$. The second is a bit more involved. First, note that if $t < T_1 + S_1$, then $t - \tau < S_1$ for every τ in the range of integration, so $x_2(t-\tau) = 0$ for all such τ and the integral is zero. Suppose that $t \geq T_1 + S_1$. Remember that $x_2(t-\tau) = 0$ for all the τ-values satisfying $t - \tau < S_1$, which is the same as $\tau > t - S_1$. So the part of the integrand corresponding to τ-values in the range $t - S_1 < \tau < \infty$ is identically zero. This argument proves not only that $x_1 * x_2$ exists (since all the integrals in (5) are over bounded τ-intervals), but also that $x_1 * x_2$ is right-sided (since $x_1 * x_2(t) = 0$ when $t < T_1 + S_1$). □

7.6 Criterion Given two decent signals x_1 and x_2, if one signal is bounded (i.e. is an L^∞-signal) and the other is absolutely integrable (i.e. is an L^1-signal), then $x_1 * x_2$ exists and is a bounded signal. Furthermore, the infinity norm of $x_1 * x_2$ satisfies

$$\|x_1 * x_2\|_\infty \leq \|\text{the } L^1 \text{ signal}\|_1 \, \|\text{the } L^\infty \text{ signal}\|_\infty \ .$$

Proof: I'll prove this in the case that x_1 is an L^1-signal and x_2 is an L^∞ signal. Given $t \in \mathbb{R}$, we can conclude from Fact 7.3 that the integral in (5) exists if we can find $R > 0$ such that

$$\int_{-T}^{T} |x_1(\tau)||x_2(t-\tau)|d\tau \leq R$$

for every $T > 0$. Since $|x_2(t-\tau)| \leq \|x_2\|_\infty$ for every t and τ,

$$\int_{-T}^{T} |x_1(\tau)||x_2(t-\tau)|d\tau \leq \left(\int_{-T}^{T} |x_1(\tau)|d\tau\right) \|x_2\|_\infty \ .$$

The integral in parentheses is, in turn, bounded from above by $\|x_1\|_1$, from which it follows that for every $T > 0$

$$\int_{-T}^{T} |x_1(\tau)||x_2(t-\tau)|d\tau \leq \|x_1\|_1 \, \|x_2\|_\infty \ ,$$

implying not only that the integral in (5) converges for every $t \in \mathbb{R}$ (so $x_1 * x_2$ exists), but also that

$$|x_1 * x_2(t)| \leq \|x_1\|_1 \, \|x_2\|_\infty$$

for every $t \in \mathbb{R}$. It follows that $x_1 * x_2$ is a bounded signal (i.e. an L^∞-signal), and that $\|x_1 * x_2\|_\infty \leq \|x_1\|_1 \, \|x_2\|_\infty$. □

7.7 Criterion: If x_1 and x_2 are both decent square-integrable signals, then $x_1 * x_2$ exists and is a bounded signal. Furthermore, the infinity norm of $x_1 * x_2$ satisfies

$$\|x_1 * x_2\|_\infty \leq \frac{\|x_1\|_2^2 + \|x_2\|_2^2}{2} \ .$$

Proof: I'll proceed as in the proof of Criterion 7.6. Again, given $t \in \mathbb{R}$, we can conclude from Fact 7.3 that the integral in (5) converges if we can find $R > 0$ such that

$$\int_{-T}^{T} |x_1(\tau)|\,|x_2(t-\tau)| d\tau \leq R$$

for every $T > 0$. From $(|x_1(\tau)| - |x_2(t-\tau)|)^2 \geq 0$ it follows directly that for each t-value

$$|x_1(\tau)||x_2(t-\tau)| \leq \frac{|x_1(\tau)|^2 + |x_2(t-\tau)^2|}{2} \quad \text{for all} \ \ \tau \in \mathbb{R} \ .$$

Thus for each t-value we have

$$\begin{aligned} \int_{-T}^{T} |x_1(\tau)|\,|x_2(t-\tau)| d\tau \ &\leq \ \frac{1}{2}\int_{-T}^{T} |x_1(\tau)|^2 d\tau + \frac{1}{2}\int_{-T}^{T} |x_2(t-\tau)|^2 d\tau \\ &\leq \ \frac{\|x_1\|_2^2 + \|x_2\|_2^2}{2} \quad \text{for all} \ \ T > 0 \ . \end{aligned}$$

It follows that the integral in (5) converges for every $t \in \mathbb{R}$, so $x_1 * x_2$ exists, and also that for every $t \in \mathbb{R}$

$$\begin{aligned} |x_1 * x_2(t)| \ &= \ \lim_{T\to\infty} \left| \int_{-T}^{T} x_1(\tau) x_2(t-\tau) d\tau \right| \\ &\leq \ \lim_{T\to\infty} \int_{-T}^{T} |x_1(\tau)||x_2(t-\tau)| d\tau \\ &\leq \ \frac{\|x_1\|_2^2 + \|x_2\|_2^2}{2} \ , \end{aligned}$$

whereby $\|x_1 * x_2\|_\infty \leq \left(\|x_1\|_2^2 + \|x_2\|_2^2\right)/2$. □

7.8 Criterion If x_1 and x_2 are both decent absolutely integrable signals, then $x_1 * x_2$ exists and is an absolutely integrable signal. Furthermore, the 1-norm of $x_1 * x_2$ satisfies

$$\|x_1 * x_2\|_1 \leq \|x_1\|_1 \, \|x_2\|_1 \ .$$

Proof: The existence of $x_1 * x_2$ in this case does not follow from Criterion 7.6 because, as we have noted, an absolutely integrable continuous-time signal need not be bounded. In other words, we need to do more work to prove Criterion 7.8 than we needed to do when proving the corresponding discrete-time result. I won't be able to give you all the details of the proof, because it relies on an advanced result called Fubini's Theorem. The gist of the argument, however, is not hard to explain.

Let S and T be given positive real numbers. Given $t \in \mathbb{R}$,

$$\begin{aligned}
\int_{-T}^{T}\left(\int_{-S}^{S}|x_1(\tau)||x_2(t-\tau)|d\tau\right)dt &= \int_{-S}^{S}\left(\int_{-T}^{T}|x_1(\tau)||x_2(t-\tau)|dt\right)d\tau \\
&= \int_{-S}^{S}|x_1(\tau)|\left(\int_{-T}^{T}|x_2(t-\tau)|dt\right)d\tau \\
&\leq \left(\int_{-S}^{S}|x_1(\tau)|d\tau\right)\|x_2\|_1 \\
&\leq \|x_1\|_1\,\|x_2\|_1\ .
\end{aligned}$$

Interchanging the order of integration on the first line is legal because the signals are decent (cf. Fact 7.2). The inner integral on the right-hand side of the second line is bounded from above by $\|x_2\|_1$, which leads to the inequality on the third line.

Fact 7.3 allows us to take the limit as $S \to \infty$ on the right-hand side of the first line, which yields

$$\begin{aligned}
\int_{-\infty}^{\infty}\left(\int_{-T}^{T}|x_1(\tau)||x_2(t-\tau)|dt\right)d\tau &= \int_{-T}^{T}\left(\int_{-\infty}^{\infty}|x_1(\tau)||x_2(t-\tau)|d\tau\right)dt \\
&\leq \|x_1\|_1\,\|x_2\|_1\ .
\end{aligned}$$

The aforementioned Fubini's Theorem allows us to interchange the order of integration and also implies that the inner integral on the right-hand side of the first line is finite for "almost every" $t \in \mathbb{R}$ ("almost every" has a technical definition that I won't go into right now). Since by equation (5)

$$|x_1 * x_2(t)| \leq \int_{-\infty}^{\infty}|x_1(\tau)||x_2(t-\tau)|d\tau\ ,$$

$x_1 * x_2(t)$ exists for (almost) all $t \in \mathbb{R}$.

Furthermore, by Fact 7.3, we can take the limit as $T \to \infty$ in the inequality

$$\int_{-T}^{T}\left(\int_{-\infty}^{\infty}|x_1(\tau)||x_2(t-\tau)|d\tau\right)dt \leq \|x_1\|_1\,\|x_2\|_1$$

to get

$$\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}|x_1(\tau)||x_2(t-\tau)|d\tau\right)dt \leq \|x_1\|_1\,\|x_2\|_1\ .$$

The inner integral is an upper bound for $|x_1 * x_2(t)|$, so the whole expression on the left is an upper bound on $\int_{-\infty}^{\infty}|x_1 * x_2(t)|dt$. It follows that $x_1 * x_2$ is indeed absolutely integrable, and $\|x_1 * x_2\|_1 \leq \|x_1\|_1\,\|x_2\|_1$. □

I glossed over a major detail in the argument for Criterion 7.8, which is the bit about "$x_1 * x_2(t)$ exists for *almost all* $t \in \mathbb{R}$." That's another one of those Lebesgue measure theory things. For our purposes, think of Criterion 7.8 as saying that $x_1 * x_2$ does indeed exist if both x_1 and x_2 are decent L^1 signals — and $x_1 * x_2$ will be an L^1 signal — but $x_1 * x_2$ need not be a decent signal. Example 7.13 below illustrates such an eventuality, which one rarely encounters in applications. I would prefer, for now, that you assume Criterion 7.8 applies literally as written in every situation we encounter. Reassuringly, the convolutions $x_1 * x_2$ in Criteria 7.4, 7.5, 7.6, and 7.7 all turn out to be decent signals, a fact I won't prove.

Examples of computing convolutions

Nothing like a few examples to leaven the mood.

7.9 Example: $x_1 = x_2 = u$. Since x_1 and x_2 are both right-sided, Criterion 7.5 applies, so $x_1 * x_2$ exists. For any $t \in \mathbb{R}$,

$$x_1 * x_2(t) = \int_{-\infty}^{\infty} u(\tau)u(t-\tau)d\tau = \int_0^{\infty} u(t-\tau)d\tau \ ,$$

where the last equality holds because $u(\tau) = 1$ for $\tau \geq 0$ and $u(\tau) = 0$ for $\tau < 0$. If $t < 0$, $u(t-\tau) = 0$ for every τ in the range of integration $0 \leq \tau \leq \infty$, so the whole integral is zero when $t < 0$. When $t \geq 0$, $u(t-\tau) = 1$ when $0 \leq \tau \leq t$ and $u(t-\tau) = 0$ when $t < \tau < \infty$. It follows that

$$x_1 * x_2(t) = \begin{cases} 0 & \text{if } t < 0 \\ \int_0^t 1 d\tau = t & \text{if } t \geq 0 \ . \end{cases}$$

Another way of writing the last equation is

$$x_1 * x_2(t) = tu(t) \ \text{ for all } \ t \in \mathbb{R} \ .$$

If you think about what $x_1 * x_2$ looks like in this case, you can see why people say that "the convolution of two unit steps is a ramp."

7.10 Example: $x_1 = u$ and x_2 is the signal with specification

$$x_2(t) = \begin{cases} e^{-3t} & t \geq 0 \\ 0 & t < 0 \ . \end{cases}$$

Note that $x_2(t) = e^{-3t}u(t)$ for every $t \in \mathbb{R}$. Again, Criterion 7.5 applies since both x_1 and x_2 are right-sided. In fact, Criterion 7.6 also applies since x_1 is bounded and x_2 is absolutely integrable, which is easy to check. To compute $x_1 * x_2$, follow a procedure similar to the one we followed in Example 7.9. For any $t \in \mathbb{R}$,

$$x_1 * x_2(t) = \int_{-\infty}^{\infty} u(\tau)e^{-3(t-\tau)}u(t-\tau)d\tau = \int_0^{\infty} e^{-3(t-\tau)}u(t-\tau)d\tau \ ,$$

where the last equality holds because $u(\tau) = 1$ for $\tau \geq 0$ and $u(\tau) = 0$ for $\tau < 0$. If $t < 0$, $u(t-\tau) = 0$ for every τ in the range of integration $0 \leq \tau \leq \infty$, so the whole integral is zero when $t < 0$. When $t \geq 0$, $u(t-\tau) = 1$ when $0 \leq \tau \leq t$ and $u(t-\tau) = 0$ when $t < \tau < \infty$. It follows that

$$x_1 * x_2(t) = \begin{cases} 0 & \text{if } t < 0 \\ \int_0^t e^{-3(t-\tau)} d\tau & \text{if } t \geq 0 \ . \end{cases}$$

Evaluating the integral yields

$$x_1 * x_2(t) = \begin{cases} 0 & \text{if } t < 0 \\ \frac{1}{3} - \frac{1}{3}e^{-3t} & \text{if } t \geq 0 \ . \end{cases}$$

Another way of writing the last equation is

$$x_1 * x_2(t) = \left(\frac{1}{3} - \frac{1}{3}e^{-3t}\right) u(t) \text{ for all } t \in \mathbb{R} .$$

7.11 Example: $x_1 = u$ and x_2 is the signal with specification

$$x_2(t) = e^{-3|t|} = \begin{cases} e^{-3t} & \text{if } t \geq 0 \\ e^{3t} & \text{if } t < 0 . \end{cases}$$

Observe that we can also specify x_2 for every $t \in \mathbb{R}$ via

$$x_2(t) = e^{-3t}u(t) + e^{3t}u(-t) .$$

Note that x_1 is a bounded signal and x_2 is an L^1-signal, so Criterion 7.6 guarantees that $x_1 * x_2$ exists. It helps to set $x_3(t) = e^{-3t}u(t)$ and $x_4(t) = e^{3t}u(-t)$ because we computed $x_1 * x_3$ in the Example 7.10. Now let's find $x_1 * x_4$.

By definition, for any $t \in \mathbb{R}$,

$$x_1 * x_4(t) = \int_{-\infty}^{\infty} u(\tau)e^{3(t-\tau)}u(-(t-\tau))d\tau = \int_{0}^{\infty} e^{3(t-\tau)}u(-t+\tau)d\tau$$

since $u(\tau) = 1$ for $\tau \geq 0$ and $u(\tau) = 0$ for $\tau < 0$. If $t < 0$, then $u(-t+\tau) = 1$ for all τ in the range of summation $0 \leq \tau < \infty$. If $t \geq 0$, then $u(-t+\tau) = 0$ for $0 \leq \tau < t$ and $u(-t+\tau) = 1$ for all τ in the range $t \leq \tau < \infty$. Accordingly,

$$x_1 * x_4(t) = \begin{cases} \int_0^\infty e^{3(t-\tau)}d\tau & \text{if } t < 0 \\ \int_t^\infty e^{3(t-\tau)}d\tau & \text{if } t \geq 0 . \end{cases}$$

It follows that

$$x_1 * x_4(t) = \begin{cases} \frac{1}{3}e^{3t} & \text{if } t < 0 \\ \frac{1}{3} & \text{if } t \geq 0 . \end{cases}$$

Plugging the result of Example 7.10 into the equation $x_1 * x_2 = x_1 * x_3 + x_1 * x_4$ yields the following specification for $x_1 * x_2$:

$$x_1 * x_2(t) = \begin{cases} \frac{1}{3}e^{3t} & \text{if } t < 0 \\ \frac{2}{3} - \frac{1}{3}e^{-3t} & \text{if } t \geq 0 . \end{cases}$$

Alternatively, for every $t \in \mathbb{R}$,

$$x_1 * x_2(t) = \left(\frac{2}{3} - \frac{1}{3}e^{-3t}\right) u(t) + \frac{1}{3}e^{3t}u(-t) .$$

Just for completeness, I'd like to show you another way to do this example. Interchanging τ and $t - \tau$ in equation (5) yields

$$x_1 * x_2(t) = \int_{-\infty}^{\infty} u(t-\tau)e^{-3|\tau|}d\tau = \int_{-\infty}^{t} e^{-3|\tau|}d\tau$$

for every $t \in \mathbb{R}$ because the $u(t-\tau)$ merely chops of the top of the sum at $\tau = t$. As a result,

$$x_1 * x_2(t) = \begin{cases} \int_{-\infty}^t e^{3\tau}d\tau & \text{if } t < 0 \\ \int_{-\infty}^0 e^{3\tau}d\tau + \int_0^t e^{-3\tau}d\tau & \text{if } t \geq 0 . \end{cases}$$

Fortunately, as you can check, this turns out to be the same answer in a slightly different form.

7.12 Example: $x_1 = u$ and x_2 is the signal with specification $x_2(t) = e^{3t}$ for all $t \in \mathbb{R}$. I'm including this example partly because it satisfies none of the criteria I've presented for convolution existence. Those criteria, in other words, aren't exhaustive. Here, x_1 is bounded and right-sided, but x_2 is neither right-sided nor bounded nor absolutely integrable. Nonetheless, $x_1 * x_2$ exists. For every $t \in \mathbb{R}$,

$$x_1 * x_2(t) = \int_{-\infty}^{\infty} u(\tau)e^{3(t-\tau)}d\tau = e^{3t}\int_0^{\infty} e^{-3\tau} = \frac{1}{3}e^{3t} .$$

It's of interest to note that $x_2(t)$ and $x_1 * x_2(t)$ both take the form (constant) $\times\, e^{3t}$.

7.13 Example: Let x_1 and x_2 both be the signal x in Figure 1. As I noted earlier, x is absolutely integrable, so Criterion 7.8 applies to this example — that is, $x_1 * x_2$, which is the same as $x * x$, must exist and be absolutely integrable. Look what happens, though, when we try to compute $x_1 * x_2(0)$.

$$\begin{aligned}
x_1 * x_2(0) &= \int_{-\infty}^{\infty} x(\tau)x(0-\tau)d\tau \\
&= \int_{-\infty}^{\infty} |x(t)|^2 dt \\
&= \sum_{n=-\infty}^{-1} (3^{-n})^2(3^{2n}) + \sum_{n=1}^{\infty}(3^n)^2(3^{-2n}) \\
&= 2\sum_{n=1}^{\infty} 1 = \infty .
\end{aligned}$$

The equality on the second line holds because $x(\tau) = x(-\tau)$ for all $\tau \in \mathbb{R}$.

Accordingly, $x_1 * x_2(0)$ does not exist. You can prove, in fact, that $x_1 * x_2(n)$ does not exist for any $n \in \mathbb{Z}$ — that is, $x_1 * x_2(t)$ is not defined for integer values of t. It also turns out that $x_1 * x_2(t) = 0$ for a lot of t-values. For example, consider $t = 1/2$.

$$x_1 * x_2(1/2) = \int_{-\infty}^{\infty} x(\tau)x(1/2-\tau)d\tau = 0$$

because none of the small intervals over which the pulses in $x(\tau)$ are nonzero overlaps an interval over which a pulse in $x(1/2-\tau)$ is nonzero, so the integrand is identically zero. See Figure 2. You can show that for any non-integer value of t, the product $x(\tau)x(t-\tau)$ is nonzero on only a finite number of bounded intervals, which implies that the integral defining $x * x(t)$ converges for every $t \notin \mathbb{Z}$. Figure 3 displays a graph of $x * x(t)$ vs. t.

Let's take another look at Criterion 7.8 in the light of this example. The signal x is a decent absolutely integrable signal, and the integral defining $x*x(t)$ converges for all but a countable set of t-values. For all practical purposes, $x * x$ exists, and technically, in a Lebesgue-integration sense, $x * x(t)$ exists "for almost all t." That would be a more accurate way to have formulated Criterion 7.8, but since signals such as x don't feature prominently in applications, I'd rather not fret about these fine points.

To make $x * x$ into a true signal, we need to define $x * x(t)$ for integer values of t. We can do that in many ways, and again, in a Lebesgue-integration sense, it doesn't matter how we do it. The easiest solution is arguably to set $x * x(n) = 0$ for all $n \in \mathbb{Z}$. You can check that $x * x$ is indeed an absolutely integrable signal as we expect given Criterion 7.8, although $x * x$ is certainly not a decent signal.

CHAPTER 8

Continuous-time LTI Systems

The path we're about to take is somewhat bumpier than the one we followed in Chapter 6, but the trailside scenery is similar. Along the way we'll confront the ubiquitous but unnerving continuous-time unit impulse, a.k.a. the Dirac δ-function. The impulse is only one of the obstacles that make life in continuous time challenging. Please rest assured that the ragged edges we'll encounter are minor annoyances that play essentially no role when it comes to applying the theory in the real world. The intuition underlying the results is the same as in discrete time even though the results themselves aren't quite as crisp.

Definition and examples

We're interested in modeling real-world processes that take continuous-time input signals and generate continuous-time output signals. An appealing way to represent such a process is as a mapping

$$S : X \longrightarrow \mathbb{F}^{\mathbb{R}}$$

where X is a subset of $\mathbb{F}^{\mathbb{R}}$ that represents the set of possible input signals for the system. The idea of the mapping S is that when $x \in X$ is the input signal to the system, $S(x) \in \mathbb{F}^{\mathbb{R}}$ is the output signal that arises. One usually assumes that the input space X is "rich enough" to include a lot of signals of interest. We'll always require that X contain at least all the finite-duration decent signals and also that X be closed under time-shifting in the sense that $\text{Shift}_{t_o}(x) \in X$ whenever $x \in X$ and $t_o \in \mathbb{R}$.

As we saw in Chapter 7, $\mathbb{F}^{\mathbb{R}}$ has a natural vector-space structure. If a system's input set X is closed under the taking of linear combinations — i.e. is a subspace of the vector space $\mathbb{F}^{\mathbb{R}}$ — and $S : X \to \mathbb{F}^{\mathbb{R}}$ is a linear mapping, we call the system linear. Furthermore, if the system has the property that shifting its input signal by time-shift t_o always gives rise to the same time-shift t_o in the system's output, we call the system time-invariant. Here is the formal definition. As always, $\mathbb{F}$ means $\mathbb{R}$ or $\mathbb{C}$ and $\mathbb{F}^{\mathbb{R}}$ is the set of all continuous-time $\mathbb{F}$-valued signals.

8.1 Definition A *continuous-time input-output linear time-invariant system over* $\mathbb{F}$ consists of the following:

- A subset X of $\mathbb{F}^{\mathbb{R}}$ representing the system's set of possible input signals. X is a subspace of $\mathbb{F}^{\mathbb{R}}$ that contains all the finite-duration decent signals and is shift-invariant in the sense that if $x \in X$ then $\text{Shift}_{t_o}(x) \in X$ for all $t_o \in \mathbb{R}$.

- A mapping $S : X \to \mathbb{F}^{\mathbb{R}}$ that is linear, i.e.

$$S(c_1x_1 + c_2x_2) = c_1S(x_1) + c_2S(x_2) \text{ for all } x_1, x_2 \in X \text{ and } c_1, c_2 \in \mathbb{F}$$

and shift-invariant, i.e.

$$S(\text{Shift}_{t_o}(x)) = \text{Shift}_{t_o}(S(x)) \text{ for all } x \in X \text{ and } t_o \in \mathbb{R} \, .$$

As in discrete time, I'll use "LTI" to mean "linear time-invariant" and "LTI system" to mean "input-output LTI system." The simple example systems we studied in discrete time have natural continuous-time analogues. The *zero system* has input space $X = \mathbb{F}^{\mathbb{R}}$ and system mapping $S : X \to \mathbb{F}^{\mathbb{R}}$ defined by $S(x) = 0$ for all $x \in X$, where 0 here is the zero signal. The *identity system* also has $X = \mathbb{F}^{\mathbb{R}}$ but its system mapping S has specification $S(x) = x$ for all $x \in X$. The *pure t_1-shift system* features a given fixed time-shift $t_1 \in \mathbb{R}$. Its input function space is $X = \mathbb{F}^{\mathbb{R}}$ and has system mapping S given by $S(x) = \text{Shift}_{t_1}(x)$ for every $x \in X$. I'll leave it to you to show that all three of these systems are LTI.

A slightly more interesting example is a continuous-time version of the discrete-time causal sliding-window M-fold averager. For this system, the input space X is the set of all decent signals in $\mathbb{F}^{\mathbb{R}}$. The mapping $S : X \to \mathbb{F}^{\mathbb{R}}$ takes each input signal $x \in X$ to the output signal $S(x) \in \mathbb{F}^{\mathbb{R}}$ whose value at time t is the average of the input x over the interval of length T preceding time t, where $T > 0$ is given. In equation form,

$$S(x)(t) = \frac{1}{T}\int_{t-T}^{t} x(\tau)d\tau$$

for every $x \in X$ and $t \in \mathbb{R}$. One can show fairly easily that the system is LTI. We stipulate that the input space X contain only decent signals to guarantee that the integrals in the definition of S exist.

Another LTI system one encounters frequently in applications is the integrator. That system takes an input signal x and outputs the signal $S(x)$ with specification

$$S(x)(t) = \int_{-\infty}^{t} x(\tau)d\tau \text{ for every } t \in \mathbb{R} \, .$$

The input space X contains precisely all those decent signals x for which the integrals in the last equation exist for every $t \in \mathbb{R}$.

As in discrete time, it pays to contemplate a few examples of systems that aren't linear or time-invariant. Linearity clearly fails for the system whose system mapping takes any decent signal to output signal $S(x)$ with specification

$$S(x)(t) = \frac{1}{T}\int_{t-T}^{t} \cos(x(\tau))d\tau \text{ for all } t \in \mathbb{R} \, .$$

Note that this system is time-invariant, as is the system with output signal $S(x)$ specified by

$$S(x)(t) = 7 - \frac{1}{T}\int_{t-T}^{t} x(\tau)d\tau \text{ for all } t \in \mathbb{R}$$

for any decent input signal x. This second system is nonlinear because, for one thing, $S(0) \neq 0$.

You can frequently determine that a system isn't time-invariant by checking how it processes elementary input signals such as the unit step u and the unit pulse p_a. For example, the system whose input space is the set of all decent signals and whose output signal in response to input x is the signal $S(x)$ with specification

$$S(x)(t) = \frac{t}{T}\int_{t-T}^{t} x(\tau)d\tau \ \text{ for all } \ t \in \mathbb{R}$$

is linear but not time-invariant. You can check that $S(u)$ has specification

$$S(u)(t) = \begin{cases} 0 & \text{when } t < 0 \\ t^2/T & \text{when } 0 \leq t < T \\ t & \text{when } t \geq T \end{cases}$$

while $S(\text{Shift}_1(u))$ has specification

$$S(\text{Shift}_1(u))(t) = \begin{cases} 0 & \text{when } t < 1 \\ t(t-1)/T & \text{when } 1 \leq t < T+1 \\ t & \text{when } t \geq T+1 \ , \end{cases}$$

so $S(\text{Shift}_1(u)) \neq \text{Shift}_1(S(u))$. Another linear system for which time-invariance fails takes a decent input signal x to output signal $S(x)$ with specification

$$S(x)(t) = x(3t-1) \ \text{ for all } \ t \in \mathbb{R} \ .$$

$S(u)$ has specification

$$S(u)(t) = \left\{ \begin{array}{ll} 0 & \text{when } t < 1/3 \\ 1 & \text{when } t \geq 1/3 \end{array} \right\} = u(t-1/3) \ \text{ for all } \ t \in \mathbb{R}$$

while $S(\text{Shift}_1(u))$ has specification

$$\begin{aligned} S(\text{Shift}_1(u))(t) &= \text{Shift}_1(u)(3t-1) \\ &= u(3t-2) \\ &= \begin{cases} 0 & \text{when } t < 2/3 \\ 1 & \text{when } t \geq 2/3 \end{cases} \\ &= u(t-2/3) \ \text{ for all } \ t \in \mathbb{R} \ , \end{aligned}$$

whereby

$$S(\text{Shift}_1(u)) = \text{Shift}_{1/3}(S(u)) \neq \text{Shift}_1(S(u)) \ ,$$

and time-invariance fails.

Strictly convolutional systems

To construct a more general example of a LTI system, begin by letting $h \in \mathbb{F}^{\mathbb{R}}$ be any decent signal. Let $\mathcal{D}_h$ denote the set of all decent signals $x \in \mathbb{F}^{\mathbb{R}}$ for which the convolution $h*x$ exists. It's easy to show that $\mathcal{D}_h$ is closed under linear combinations (i.e. is a subspace of $\mathbb{F}^{\mathbb{R}}$). $\mathcal{D}_h$ also contains all the finite-duration decent signals by Criterion 7.4. In fact, $\mathcal{D}_h$ is also closed under time shifting. To see this, note that for any $t_o \in \mathbb{R}$ we have, for every $t \in \mathbb{R}$,

$$\begin{aligned} h * (\text{Shift}_{t_o}(x))(t) &= \int_{-\infty}^{\infty} h(\tau)\text{Shift}_{t_o}(x)(t-\tau)d\tau \\ &= \int_{-\infty}^{\infty} h(\tau)x(t-\tau-t_o)d\tau \\ &= \int_{-\infty}^{\infty} h(\tau)x((t-t_o)-\tau)d\tau \\ &= h * x(t-t_o) \ . \end{aligned}$$

Since $x \in \mathcal{D}_h$, $h * x$ exists, so $h * x(t-t_o)$ is well defined for every $t \in \mathbb{R}$. By the chain of equalities above, $h * (\text{Shift}_{t_o}(x))(t)$ is also well defined for every $t \in \mathbb{R}$, from which it follows that $h * \text{Shift}_{t_o}(x)$ exists. We conclude that $\text{Shift}_{t_o}(x) \in \mathcal{D}_h$ if $x \in \mathcal{D}_h$.

Since $\mathcal{D}_h$ is closed under linear combination and time-shifting, it is a suitable input space for a LTI system. Let's define $S : \mathcal{D}_h \to \mathbb{F}^{\mathbb{R}}$ by

$$S(x) = h * x \ \text{ for all } \ x \in \mathcal{D}_h \ .$$

It's clear from linearity of convolution that S is a linear mapping from $\mathcal{D}_h$ into $\mathbb{F}^{\mathbb{R}}$. As for time-invariance, we saw above that for any $x \in \mathcal{D}_h$ and any $t_o \in \mathbb{R}$,

$$h * \text{Shift}_{t_o}(x)(t) = h * x(t-t_o) \ \text{ for all } \ t \in \mathbb{R} \ ,$$

which is the same as saying that

$$S(\text{Shift}_{t_o}(x)) = h * \text{Shift}_{t_o}(x) = \text{Shift}_{t_o}(h * x) = \text{Shift}_{t_o}(S(x))$$

for every $x \in \mathcal{D}_h$, so S is a shift-invariant mapping. The bottom line is that for any decent signal $h \in \mathbb{F}^{\mathbb{R}}$, the system with input space $X = \mathcal{D}_h$ and system mapping defined by $S(x) = h * x$ is LTI. One might call such a system a "convolutional LTI system" for obvious reasons. It turns out that convolutional systems are almost — but not quite — as universal in continuous time as they are in discrete time. It is here that the continuous-time theory starts displaying significant complications relative to the discrete-time theory.

8.2 Definition: A continuous-time LTI system with input space X and system mapping $S : X \to \mathbb{F}^{\mathbb{R}}$ is *strictly convolutional* when there exists a decent signal $h \in \mathbb{F}^{\mathbb{R}}$ for which $X = \mathcal{D}_h$ and $S(x) = h * x$ for every $x \in X$.

The averager system and the integrator system are both strictly convolutional systems in the sense of Definition 8.2. For the averager system, you can show that

$$S(x)(t) = \frac{1}{T}\int_{t-T}^{t} x(\tau)d\tau = \int_{-\infty}^{\infty} h(t-\tau)x(\tau)d\tau = h * x(t)$$

for every $x \in X$ and $t \in \mathbb{R}$, where

$$h(t) = \begin{cases} 1/T & \text{if } \ 0 \le t < T \\ 0 & \text{otherwise.} \end{cases}$$

Since h has finite duration, $\mathcal{D}_h$ is the set of all decent signals, which is the input space for the averager system by our original definition. Meanwhile, the integrator's output in response to an admissible input signal x has specification

$$S(x)(t) = \int_{-\infty}^{t} x(\tau)d\tau = \int_{-\infty}^{\infty} u(t-\tau)x(\tau)d\tau = h * x(t)$$

for every $t \in \mathbb{R}$, where $h = u$. The input space X for the integrator system is the set of all decent $x \in \mathbb{F}^{\mathbb{R}}$ for which $u * x$ exists, which is $\mathcal{D}_h$ since $h = u$.

Unfortunately, many reasonable continuous-time LTI systems aren't strictly convolutional. Examples include the identity system and the pure t_1-shift system. The identity system is actually a special case of a t_1-shift system with $t_1 = 0$. Let's see why time-shift systems, including the identity system, fail to be strictly convolutional. Given $t_1 \in \mathbb{R}$, suppose we had a decent signal $h \in \mathbb{F}^{\mathbb{R}}$ for which

$$S(x) = \text{Shift}_{t_1}(x) = h * x$$

for every decent signal $x \in \mathbb{F}^{\mathbb{R}}$. Consider driving the system with the decent finite-duration input signal $x = p_a$, where $a > 0$ is given. The given input signal, graphed against t, looks like a rectangular pulse of width a and height 1 centered at $t = 0$. Observe that, for every $t \in \mathbb{R}$,

$$S(p_a)(t) = h * p_a(t) = \int_{-\infty}^{\infty} h(t-\tau)p_a(\tau)d\tau = \int_{t-a/2}^{t+a/2} h(\zeta)d\zeta \ .$$

Since h is a decent signal, we can find $R > 0$ such that $|h(\zeta)| \leq R$ for all ζ satisfying $|\zeta - t_1| \leq 1$. It follows that when $a < 1$ we have

$$|S(p_a)(t_1)| \leq aR \ .$$

On the other hand, by definition of the shift system, we also require

$$S(p_a)(t) = \text{Shift}_{t_1}(p_a)(t) = p_a(t - t_1)$$

for every $t \in \mathbb{R}$. In particular,

$$S(p_a)(t_1) = p_a(0) = 1 \ .$$

If $a < 1/R$ this is impossible. It follows that no decent h exists for which $\text{Shift}_{t_1}(x) = h * x$ even for all decent signals x, much less for all $x \in \mathbb{F}^{\mathbb{R}}$.

Impulse response and the continuous-time impulse

The impulse response of discrete-time LTI system is the signal you convolve with any input to the system to get the corresponding output. In continuous time, only a strictly convolutional system has such an object associated with it, i.e. a signal h that you convolve with inputs to generate outputs. In analogy with discrete time, we define that h as the strictly convolutional system's impulse response. So we've defined "impulse response" at least for a large class of continuous-time systems without having defined "continuous-time impulse." The impulse response of a discrete-time system is literally the system's output in response to an impulse at the input. Why designate similarly the signal h appearing in Definition 8.2?

To discover the answer, suppose we have a strictly convolutional system with system mapping $S : \mathcal{D}_h \to \mathbb{F}^{\mathbb{R}}$ specified by

$$S(x) = h * x \ \text{ for all } \ x \in \mathcal{D}_h \ .$$

Let h_a be the system's response to input $x = (1/a)p_a$, where $a > 0$ is small. For every $t \in \mathbb{R}$ we have

$$h_a(t) = h * ((1/a)p_a)(t) = \frac{1}{a}\int_{-\infty}^{\infty} h(\tau)p_a(t-\tau)d\tau = \frac{1}{a}\int_{t-a/2}^{t+a/2} h(\tau)d\tau \ .$$

In other words, $h_a(t)$ is, for each $t \in \mathbb{R}$, the average value of $h(\tau)$ over the τ-interval $t - a/2 \leq \tau \leq t + a/2$. As $a \to 0$, this average value converges to $h(t)$ whenever t is a continuity point of h and it converges to the average of h across the jump if t is a jump point of h. Most importantly, it always converges because h is decent. Let's call the limiting signal h_0, i.e.

$$h_0(t) = \lim_{a\to 0} S((1/a)p_a)(t) = \lim_{a\to 0} h_a(t) = \lim_{a\to 0}\frac{1}{a}\int_{t-a/2}^{t+a/2} h(\tau)d\tau \ \text{ for all } \ t \in \mathbb{R} \ .$$

Since h_0 agrees with h except at t-values where h jumps, h_0 is essentially the same signal as h. In particular, h_0 is decent, $\mathcal{D}_h = \mathcal{D}_{h_0}$, and $h_0 * x = h * x$ for every decent signal $x \in \mathcal{D}_h$.

Now let's brazenly interchange the limit and the S in the definition of h_0, i.e.

$$h_0 = \lim_{a\to 0} S((1/a)p_a) = S\left(\lim_{a\to 0}\frac{1}{a}p_a\right) \ .$$

By taking this reckless step we've created something of a monster, namely the expression in large parentheses on the right-hand side, which is known as the *continuous-time unit impulse* δ or the *Dirac* δ*-function.* Our maneuver enables us to regard h_0 as the response of the system to a unit-impulse input — i.e., $h_0 = S(\delta)$ — which would explain why we call h_0 or the essentially equivalent signal h the impulse response of the strictly convolutional system we started with.

But let's be careful. Disturbingly, δ as we've described it has specification

$$\delta(t) = \lim_{a\to 0}\frac{1}{a}p_a(t) = \begin{cases} \infty & \text{if } \ t = 0 \\ 0 & \text{if } \ t \neq 0 \ . \end{cases}$$

As a consequence, δ isn't really a signal at all, so it's questionable whether we have the right to use it as an input to a LTI system. The good news about δ is that we find it almost exclusively under integral signs, and we can attach a rigorous meaning to any expression wherein δ makes such an appearance. Specifically,

$$(6) \qquad \int \delta(\tau)[\cdots]d\tau \ \text{ means } \ \lim_{a\to 0}\frac{1}{a}\int p_a(\tau)[\cdots]d\tau \ .$$

The quantity in brackets and limits of integration could be anything.

Of particular importance is the following chain of equalities stemming from (6), in which I assume that x is a decent signal.

$$\begin{aligned}\int_{-\infty}^{\infty}\delta(\tau)x(t-\tau)d\tau &= \lim_{a\to 0}\frac{1}{a}\int_{-\infty}^{\infty} p_a(\tau)x(t-\tau)d\tau \\ &= \lim_{a\to 0}\frac{1}{a}\int_{-a/2}^{a/2} x(t-\tau)d\tau \\ &= \lim_{a\to 0}\frac{1}{a}\int_{t-a/2}^{t+a/2} x(\zeta)d\zeta = x(t) \ ,\end{aligned}$$

with the last equality holding if t is a point of continuity for x. If t is a jump point of x, the value of the limit in the last line is the average value of x across the jump. At least formally, this means that $\delta * x = x$ for every continuous signal x, and $\delta * x = x$ *almost* holds (except possibly at x's jump points) for every decent signal x. In that sense, δ acts essentially as an identity element for continuous-time convolution, created by fiat via (6). Consequently, one could regard the identity system, which is not strictly convolutional, as "quasi-convolutional" in the sense that

$$S(x) = x = \delta * x$$

for every decent signal x.

Note also that for given $t_1 \in \mathbb{R}$ and any decent signal x we have

$$\begin{aligned}\int_{-\infty}^{\infty} \mathrm{Shift}_{t_1}(\delta)(\tau)x(t-\tau)d\tau &= \int_{-\infty}^{\infty} \delta(\tau - t_1)x(t-\tau)d\tau \\ &= \int_{-\infty}^{\infty} \delta(\zeta)x(t-t_1-\zeta)d\zeta \\ &= \lim_{a\to 0}\frac{1}{a}\int_{-\infty}^{\infty} p_a(\zeta)x(t-t_1-\zeta)d\zeta \\ &= \lim_{a\to 0}\frac{1}{a}\int_{-a/2}^{a/2} x(t-t_1-\zeta)d\zeta \\ &= \lim_{a\to 0}\frac{1}{a}\int_{t-t_1-a/2}^{t-t_1+a/2} x(\mu)d\mu = x(t-t_1)\end{aligned}$$

whenever $t - t_1$ is a point of continuity for x. (The third line in this last sequence of equations follows from (6).) At least formally, this means that

$$\mathrm{Shift}_{t_1}(\delta) * x = \mathrm{Shift}_{t_1}(x)$$

for decent signals x. So the t_1-shift system is also "quasi-convolutional" provided we admit impulses as things to convolve with. A note on terminology: people often refer to $\mathrm{Shift}_{t_1}(\delta)$ as an impulse *occurring at time* t_1.

I'd like to hold off for now on discussing impulses any further. Treating impulses rigorously requires the same techniques from measure theory that underpin Lebesgue integration. Suffice it to say that impulses supply us with a *de facto* identity element for convolution and method for shifting signals via convolution. These appurtenances make it possible to deal with the identity and t_1-shift systems almost as if they were strictly convolutional systems.

In that spirit, I'd like now to introduce a standing assumption that will remain in force whenever we talk about continuous-time LTI systems from now on. Making the assumption doesn't limit us appreciably, and most treatments cleave to it without even bothering to state it explicitly. What it posits, essentially, is that every system we'll deal with be decomposable into a strictly convolutional part and a shift part.

8.3 Standing Assumption: Every LTI system we encounter has a system mapping S of the following form.

$$S(x) = h_0 * x + \sum_{k=0}^{M} d_k \mathrm{Shift}_{t_k}(x) \ ,$$

where $h_0 \in \mathbb{F}^{\mathbb{R}}$ is a decent signal; M is a nonnegative integer; $t_0 = 0$; $t_k \in \mathbb{R}$ are distinct and nonzero when $k > 0$; and $d_k \in \mathbb{F}$ for each k, with $d_k \neq 0$ when $k > 0$. Furthermore, the system's input space X is the set of all decent signals $x \in \mathbb{F}^{\mathbb{R}}$ for which $h_0 * x$ exists — i.e. $X = \mathcal{D}_{h_0}$. We define the system's impulse response as

$$h = h_0 + \sum_{k=0}^{M} d_k \mathrm{Shift}_{t_k}(\delta) \ ,$$

which means that $S(x) = h * x$ for every $x \in X$.

More often than not, we'll be working with strictly convolutional systems. For any such system, $h = h_0$ — that is, the system's impulse response contains no pure-shift component. Sometimes we'll encounter systems that have $d_0 \neq 0$ but no d_k-terms $k > 0$. Because $t_0 = 0$, such systems' input-output behavior features a "pure identity" component, and the system mapping decomposes schematically as

$$S = (\text{strictly convolutional part}) \ + \ d_0 \times (\text{identity part}) \ .$$

The impulse response h of a system satisfying Standing Assumption 8.3 tells the entire story about the system's input-output behavior. To find $S(x)$ for any input x, simply convolve h with x. As in discrete time, h is not only "what you convolve with inputs to get outputs" but is also "the response of the system to an impulse." To see why, note that

$$S(\delta) = h_0 * \delta + \sum_{k=0}^{N} d_k \mathrm{Shift}_{t_k}(\delta) = h_0 + \sum_{k=0}^{N} d_k \mathrm{Shift}_{t_k}(\delta) = h \ .$$

To get the second equality, I used that fact that $h_0 * \delta = h_0$ because h_0 is a decent signal and δ acts as an identity element for convolution with such signals.

We figured out earlier the impulse responses of the strictly convolutional averager and integrator systems. Now let's consider the other prototype systems and also redo those earlier examples using $h = S(\delta)$. Because $S(x) = 0$ for every input x to the zero system, $S(\delta) = 0$ and its impulse response is $h = 0$. For the identity system, since $S(x) = x$ for every input signal x, it follows that $S(\delta) = \delta$. Thus the impulse response of the identity system is $h = \delta$. Strictly speaking, we should stipulate that the input space X, defined previously as $\mathbb{F}^{\mathbb{R}}$, contain only decent signals to make $\delta * x = x$ for every $x \in X$. The pure t_1-shift system has $S(x) = \mathrm{Shift}_{t_1}(x)$ for every x, so $h = S(\delta) = \mathrm{Shift}_{t_1}(\delta)$. Again, strictly speaking, we should stipulate that the input space X, defined previously as $\mathbb{F}^{\mathbb{R}}$, contain only decent signals.

As for the averager, $h = S(\delta)$, so h is the signal whose value at time t is

$$h(t) = \frac{1}{T}\int_{t-T}^{t} \delta(\tau) d\tau$$

for every $t \in \mathbb{R}$. Invoking the rigorous meaning of the right-hand side yields

$$h(t) = \frac{1}{T}\lim_{a\to 0}\int_{t-T}^{t} (1/a)p_a(\tau) d\tau = \begin{cases} 1/T & \text{if } 0 < t < T \\ 0 & \text{if } t < 0 \text{ or } t > T \\ 1/2T & \text{if } t = 0 \text{ or } t = T \ . \end{cases}$$

Ignoring the slight disagreements at the jump points for h, this is the same as the answer we found before. For the integrator, $h = S(\delta)$ is the signal whose value at time t is

$$h(t) = \int_{-\infty}^{t} \delta(\tau) d\tau$$

for every $t \in \mathbb{R}$, meaning that

$$h(t) = \lim_{a \to 0} \int_{-\infty}^{t} (1/a) p_a(\tau) d\tau = \begin{cases} 1 & \text{if } t > 0 \\ 0 & \text{if } t < 0 \\ 1/2 & \text{if } t = 0 \ . \end{cases}$$

Ignoring the $t = 0$ issue, we see that $h = u$ as we calculated earlier.

Causality and BIBO stability

Since the impulse response h determines the entire input-output behavior of a LTI system satisfying Standing Assumption 8.3, you might expect that important "system properties" have embodiments as "properties of h," and that is indeed the case. Two instances of this correspondence arise in relation to the system properties of causality and bounded-input bounded-output stability.

In analogy with discrete time, a continuous-time LTI system is causal if, roughly speaking, the current value of the output signal depends only on the current and past values of the input signal and not on future values of the input signal. Here is the formal definition.

8.4 Definition: A LTI system $S : X \to \mathbb{F}^{\mathbb{R}}$ is *causal* when for every $t \in \mathbb{R}$ the following holds: if x_1 and x_2 are two input signals in X such that $x_1(\tau) = x_2(\tau)$ for every $\tau \leq t$, then $S(x_1)(\tau) = S(x_2)(\tau)$ for every $\tau \leq t$.

In other words, when a system is causal and two inputs "agree" up to and including time t, the outputs to which they give rise will also "agree" up to and including time t, and that condition holds for every $t \in \mathbb{R}$. As in discrete time, we can prove a condition on a system's impulse response h that holds if and only if the system is causal.

8.5 Theorem: A LTI system satisfying Standing Assumption 8.3 is causal if and only if its impulse response h satisfies

(1) $h_0(t) = 0$ for $t < 0$, and
(2) $t_k > 0$ for $1 \leq k \leq M$.

Proof: First, suppose a system is causal. Since the system is linear, $S(0) = 0$, where 0 stand for the "all-zero signal." For any $a > 0$, $p_a(\tau) = 0$ for every $\tau < -a/2$, so p_a agrees with the all-zero signal up to and including any time $t < -a/2$. Since the system is causal, for any $a > 0$ the signal $S((1/a)p_a)$ must agree with $S(0) = 0$ up to and including any time $t < -a/2$. We need to show that all the t_k for $1 \leq k \leq M$ are positive — note that when $M = 0$ we have nothing to prove here. Suppose $t_1 < 0$. Pick $a > 0$ small enough so that $|t_k - t_1| > a/2$ for all $k > 1$, which we can do because the t_k are distinct. For such a choice of a, $\text{Shift}_{t_1}((1/a)p_a)(t_1) = 1/a$ and $\text{Shift}_{t_k}((1/a)p_a)(t_1) = 0$ for all $k \neq 1$, so

$$S((1/a)p_a)(t_1) = h_0 * ((1/a)p_a)(t_1) + \frac{d_1}{a} .$$

When a is very small, the right-hand side is not equal to zero because the second term swamps the first. This contradicts causality when we choose a small enough so that $t_1 < -a/2$. We conclude that $t_1 \geq 0$. A similar argument shows that all the t_k for $1 \leq k \leq M$ must be positive.

Similarly, if $h_0(\tau) \neq 0$ for some $\tau < 0$, because h_0 is decent we can assume that $h(\tau) \neq 0$ over some interval $[-T_2, -T_1]$ of τ-values where T_1 and T_2 are positive and $T_1 < T_2$. Consider using the following input to the system:

$$x(\tau) = \begin{cases} \overline{h_0(-\tau)} & \text{if } T_1 \leq \tau \leq T_2 \\ 0 & \text{otherwise,} \end{cases}$$

where $\overline{h_0(\tau)}$ is the complex conjugate of $h_0(\tau)$. Note that $x(\tau) = 0$ for all $\tau < T_1$, so by causality we must have $S(x)(\tau) = 0$ for $\tau < T_1$; in particular, we need $S(x)(0) = 0$. But when we calculate $S(x)(0)$ we find that

$$S(x)(0) = h_0 * (x)(0) = \int_{-\infty}^{\infty} h_0(0-\tau)x(\tau)d\tau = \int_{T_1}^{T_2} |h_0(-\tau)|^2 d\tau > 0 ,$$

which is a contradiction. (Note: the shift terms in h, if any, don't contribute to $S(x)(0)$ because the input x does not "turn on" until time T_1 and we know already that all the nonzero t_k are positive.) The bottom line is that if the system is causal, then all the t_k from the shift terms (if any) in h must be positive, and in addition

$$h_0(t) = 0 \text{ when } t < 0 .$$

Conversely, suppose $h_0(t) = 0$ for every $t < 0$ and all the t_k for $1 \leq k \leq M$ are positive — recall also that $t_0 = 0$. Given any $t \in \mathbb{R}$ and any $x \in X$,

$$S(x)(t) = h_0 * (x)(t) + \sum_{k=0}^{M} d_k \text{Shift}_{t_k}(x)(t) = \int_{-\infty}^{t} h_0(t-\tau)x(\tau)d\tau + \sum_{k=0}^{M} d_k x(t-t_k)$$

where the last equality holds because $h_0(t-\tau) = 0$ when $\tau > t$. Consequently, the output at time t in response to input x depends only on the values of $x(\tau)$ for $\tau \leq t$ and not on the values of $x(\tau)$ for $\tau > t$. This condition holds for every $x \in X$ and every $t \in \mathbb{R}$, so if x_1 and x_2 are two input signals that agree up to and including time t, then $S(x_1)$ and $S(x_2)$ must also agree up to and including time t. It follows that the system is causal. □

The zero system, the identity system, the sliding-window averager, and the integrator are all causal LTI systems. All those systems clearly satisfy the informal definition of causality. Each system's "current output value" depends explicitly on "current and/or past input values" and not on "future input values." Is the shift system with $S = \text{Shift}_{t_1}$ causal? It depends. Theorem 8.5 supplies the answer: if $t_1 \geq 0$, then Yes; if $t_1 < 0$, then No.

What about stability? Roughly speaking, a system is stable if nothing crazy happens when you drive the system with well behaved inputs. As in discrete time, people have settled on a notion of stability that goes essentially like this: a system is stable if every bounded decent signal $x \in \mathbb{F}^{\mathbb{R}}$ is an admissible input for the system and if, in addition, the output $S(x)$ arising from such a signal x is also a bounded signal.

8.6 Definition: A LTI system with input space X and system mapping $S : X \to \mathbb{F}^{\mathbb{R}}$ is *bounded-input bounded-output stable* or *BIBO stable* when X contains all the bounded decent signals in $\mathbb{F}^{\mathbb{R}}$ and, for every bounded decent $x \in X$, $S(x)$ is also a bounded signal.

Like causality, BIBO stability of a system has a neat characterization in terms of the system's impulse response. I'll prove the standard version of this characterization first, and then state a stronger version without proof.

8.7 Theorem: A LTI system satisfying Standing Assumption 8.3 with impulse response h is BIBO stable if and only if h_0 is absolutely integrable — i.e., if and only if $h_0 \in L^1$.

Proof: Assume first that h_0 is absolutely integrable. By convolution-existence Criterion 7.6, $h_0 * x$ exists for every bounded decent signal $x \in \mathbb{F}^{\mathbb{R}}$ and furthermore is a bounded signal satisfying

$$\|h_0 * x\|_\infty \le \|h_0\|_1 \, \|x\|_\infty \ .$$

In particular, every bounded decent x lies in $\mathcal{D}_{h_0}$, and, since $\mathcal{D}_{h_0} = X$ by Standing Assumption 8.3, every bounded decent signal is an admissible input to the system. Now,

$$S(x) = h_0 * x + \sum_{k=0}^{M} d_k \mathrm{Shift}_{t_k}(x) \ ,$$

and the kth shift term is bounded from above by $|d_k| \, \|x\|_\infty$ if x is bounded. It follows that every bounded decent input x leads to a bounded output $S(x)$. In fact,

$$\|S(x)\|_\infty \le \left(\|h_0\|_1 + \sum_{k=0}^{N} |d_k| \right) \|x\|_\infty \ .$$

The system is therefore BIBO stable.

Conversely, suppose the system is BIBO stable but h_0 is not absolutely integrable. I'll construct a bounded decent input signal x for which $h_0 * x$ doesn't exist, which means $x \notin \mathcal{D}_{h_0}$. By Standing Assumption 8.3, the input space X for the system is $\mathcal{D}_{h_0}$, so the existence of such an x contradicts BIBO stability of the system. Define

$$x(\tau) = \begin{cases} \overline{h_0(-\tau)}/|h_0(-\tau)| & \text{when } \tau \le 0 \text{ and } h_0(-\tau) \ne 0 \\ 0 & \text{otherwise} \ , \end{cases}$$

where $\overline{h_0(-\tau)}$ is the complex conjugate of $h_0(-\tau)$. The signal x is bounded; in fact, $\|x\|_\infty = 1$. Furthermore, x is decent because h_0 is. I claim that $x \notin \mathcal{D}_{h_0}$. To see why, attempt to compute $S(x)(0)$. You get

$$\begin{aligned} S(x)(0) &= h_0 * x(0) \\ &= \int_{-\infty}^{\infty} h_0(\tau)x(0-\tau)d\tau \\ &= \int h_0(\tau)\overline{h_0(\tau)}/|h_0(\tau)|d\tau \\ &= \int_{-\infty}^{\infty} |h_0(\tau)|d\tau \ , \end{aligned}$$

where the integral on the third line is over the range where $h_0(\tau) \neq 0$. The last integral doesn't exist because h_0 isn't absolutely integrable, so $S(x)(0)$ isn't well defined and thus $x \notin \mathcal{D}_{h_0}$. We conclude that h_0 must indeed be absolutely integrable for the system to be BIBO stable. □

Once again by analogy to discrete time, a significantly stronger refinement of Theorem 8.7 holds.

8.8 Theorem: A LTI system with input space X and system mapping S is BIBO stable if and only if every bounded decent right-sided signal is in X and $S(x)$ is a bounded signal for every bounded decent right-sided signal x. □

What makes Theorem 8.8 stronger than Theorem 8.7? As in discrete time, to check for BIBO stability (or, equivalently, for absolute integrability of h), we need only make sure that $h * x$ is bounded for every bounded *right-sided* decent signal x. If so, we can conclude that $h * x$ is bounded for *all* bounded decent signals x, right-sided or not.

The hard part of proving Theorem 8.8 is showing that if $h * x$ exists and is bounded for every bounded right-sided decent signal x, then $h \in L^1$. We can construct a bounded right-sided signal resembling x in the proof of Theorem 8.7 that enables us to conclude that $\int_{-\infty}^{0} |h(t)|dt$ must exist for $h*x$ to exist. It's trickier to prove that $\int_{0}^{\infty} |h(t)|dt$ exists when $h * x$ exists for all bounded right-sided decent signals x. Suppose, for example, that the system under consideration is causal, in which case h is right-sided by Theorem 8.5. Criterion 5.5 tells us that $h * x$ exists for every decent right-sided signal x, so there's no way to build a bounded decent right-sided signal x for which $h * x$ fails to exist when $\int_{0}^{\infty} |h(t)|dt$ blows up. As in discrete time, finishing the proof requires the Uniform Boundedness Theorem from functional analysis.

Let's wrap things up by checking for BIBO stability of the various example systems. It's easy to see that the zero and identity systems, the shift system(s), and the averager are all BIBO stable. You can understand this on an elementary level just by contemplating the definition of BIBO stability and asking whether a bounded input signal leads to a bounded output signal for each of these systems. In each case, the answer is unequivocally Yes.

The integrator, on the other hand, is not BIBO stable. Consider driving the system with bounded input signal $x = u$. You find that

$$S(u)(t) = \int_{-\infty}^{t} u(\tau) d\tau = \begin{cases} t & \text{if } t \geq 0 \\ 0 & \text{if } t < 0 \ . \end{cases}$$

Alternatively, $S(u)(t) = tu(t)$ for every $t \in \mathbb{R}$. While $S(u)$ is well defined (i.e. $u \in X$), $S(u)$ not bounded even though u is.

CHAPTER 9

Fourier Series as Orthogonal Expansions

The study of periodic signals provided early inspiration for fundamental developments in continuous-time signal and system analysis during the twentieth century. The key concept that emerged from that study is the frequency content of a signal. Understanding signals in terms of their frequency content allows novel approaches to problems involving not only the signals themselves but also how LTI systems act upon them. While frequency content is most easily understood in the context of periodic signals, we'll see in subsequent chapters that it also makes sense for a variety of non-periodic signals.

Periodic signals

You know intuitively that a periodic signal cycles repetitively through a set of values. More formally, we say that a signal $x \in \mathbb{F}^{\mathbb{R}}$ is *periodic* when there exists $T > 0$ such that $\mathrm{Shift}_T(x) = x$. In this case, we say that T *is a period* of x. As usual, $\mathbb{F}$ means $\mathbb{R}$ or $\mathbb{C}$ and $\mathbb{F}^{\mathbb{R}}$ is the set of all continuous-time $\mathbb{F}$-valued signals.

Every constant signal is trivially periodic, and every $T > 0$ is a period of such a signal. Sines and cosines are in a sense the prototypical periodic signals. Given $\Omega_o > 0$, the signals $t \mapsto \cos \Omega_o t$ and $t \mapsto \sin \Omega_o t$ are periodic, and every T of the form $m2\pi/\Omega_o$, with m a positive integer, is a period of both of these. Another paradigmatic periodic signal is $t \mapsto e^{j\Omega_o t}$, which has the same periods as the cosine and sine above. Other periodic signals we encounter frequently are periodic square waves, triangle waves, and sawtooth waves.

If T is a period of x, then any integer multiple of T is also a period of x. It follows that every periodic signal has arbitrarily large periods. Note, however, that all the examples in the preceding paragraph except the constant signal have some "smallest period." A non-constant periodic signal lacking a "smallest period" is the signal x with specification

$$x(t) = \begin{cases} 0 & \text{when } t \text{ is rational} \\ 1 & \text{when } t \text{ is irrational.} \end{cases}$$

Every rational number $T > 0$ is a period of x because the sum of two rational numbers is rational and the sum of a rational number and an irrational number is irrational. This x, which we met in Chapter 7, is not a decent signal. As it happens, every decent non-constant periodic signal has a "smallest period."

9.1 Fact: If $x \in \mathbb{F}^{\mathbb{R}}$ is a decent non-constant periodic signal, then there exists a smallest $T_o > 0$ that is a period of x.

Proof: I'll show that if x has arbitrarily small periods, then x must be constant. Given an arbitrary $T > 0$, we can find a sequence $\{T_1, T_2, \ldots\}$ of periods of x such that T_n increases to T in the limit as $n \to \infty$. To construct this sequence, form T_{n+1} from T_n by adding on a sufficiently small period of x so that $T_{n+1} < T$. For any $t \in \mathbb{R}$ and $n > 0$, we have $x(t - T_n) = x(t)$ since T_n is a period of x. If $t - T$ is a point of continuity of x, it follows that $x(t - T) = x(t)$ because $\lim_{n\to\infty} x(t-T_n) = x(t-T)$. What if $t-T$ is a jump point of x? As it happens, x can't have any jumps. If t_o were a jump point of x, then for some $\epsilon > 0$ we would have $x(t_o - \delta) \neq x(t_o + \delta)$ for every $\delta < \epsilon$. But x has a period $\tau < \epsilon$, by assumption, which implies that $x(t_o - \tau/2) = x(t_o + \tau/2)$, a contradiction. Accordingly, x must be continuous since it is a decent signal with no jumps, and $x(t - T) = x(t)$ for every $t \in \mathbb{R}$, which implies that x is constant since T was arbitrary. In summary, if x is decent and not constant, then it can't have arbitrarily small periods. □

Fact 9.1 enables us to define the *fundamental period* of any decent periodic signal x as the smallest period of x. If the fundamental period of x is T_o, the *fundamental frequency* of x is $2\pi/T_o$, which I'll usually denote by Ω_o. As a reality check, confirm for yourself that the fundamental period of $t \mapsto \cos \Omega_o t$, $t \mapsto \sin \Omega_o t$, and $t \mapsto e^{j\Omega_o t}$ is indeed $2\pi/\Omega_o$, so all three of these signals have fundamental frequency Ω_o. I noted earlier that every positive integer multiple of $2\pi/\Omega_o$ is a period of all these signals. In fact, those are the signals' only periods.

9.2 Fact: If $x \in \mathbb{F}^{\mathbb{R}}$ is a decent periodic signal with fundamental period T_o, then every period of x is a positive integer multiple of T_o.

Proof: Suppose T is a period of x. By definition of fundamental period, $T \geq T_o$, so we can write $T = mT_o + R$ for some positive integer m and some R that satisfies $0 \leq R < T_o$. Since T and mT_o are both periods of x, it follows that

$$x = \mathrm{Shift}_T(x) = \mathrm{Shift}_{mT_o+R}(x) = \mathrm{Shift}_R(\mathrm{Shift}_{mT_o}(x)) = \mathrm{Shift}_R(x) ,$$

so if $R > 0$ then R is also a period of x. But this is a contradiction since $R < T_o$ and T_o is x's fundamental period. Hence $R = 0$ and $T = mT_o$. □

What about linear combinations of periodic signals? Clearly, if two periodic signals x_1 and x_2 both have T as a period, then any signal of the form $c_1x_1 + c_2x_2$ also has T as a period. On the other hand, if T is a period of x_1 but not of x_2, we have no reason to expect that T will be a period of $c_1x_1 + c_2x_2$. More fundamentally, is $c_1x_1 + c_2x_2$ even necessarily periodic? The answer is somewhat delicate but not terribly surprising.

9.3 Fact: If x_1 has fundamental period T_1 and x_2 has fundamental period T_2 and both signals are decent, then following statements hold.

(1) If T_1/T_2 is a rational number, then every linear combination of the form $c_1x_1 + c_2x_2$ with c_1 and c_2 in $\mathbb{F}$ is periodic.

(2) If, for *some* choice of nonzero constants c_1 and c_2 in $\mathbb{F}$, the signal $x = c_1x_1 + c_2x_2$ is periodic, then T_1/T_2 is a rational number, and the stronger conclusion in (1) therefore holds.

Proof: Consider item (1) first. if $T_1/T_2 = m/n$ for some positive integers m and n, let $T = nT_1 = mT_2$. Then T is a period of both x_1 and x_2, and it follows that any linear combination $c_1x_1 + c_2x_2$ has T as a period and is therefore periodic.

As for (2), say we have nonzero c_1 and c_2 such that $x = c_1x_1 + c_2x_2$ is periodic. Let T be a period of x. Subtracting $\mathrm{Shift}_T(x)$ from x yields

$$-c_1(x_1 - \mathrm{Shift}_T(x_1)) = c_2(x_2 - \mathrm{Shift}_T(x_2)) \ .$$

We have two cases to consider.

Case 1: The signals on either side are constant. It turns out in this case that both signals are zero, which means that T is a period of both x_1 and x_2, implying that $T = nT_1 = mT_2$ for some positive integers m and n by Fact 9.2, so that $T_1/T_2 = m/n$ is a rational number. Why are both signals zero? Consider the left-hand side. Suppose $x_1 - \mathrm{Shift}_T(x_1) = c_0 \neq 0$. This would imply that $x_1(T) = x_1(0) + c_0$, $x_1(2T) = x_1(0) + 2c_0$, $x_1(3T) = x_1(0) + 3c_0$, etc. Accordingly, x_1 would have to be unbounded. But since x_1 is a decent signal, x_1 is bounded on the interval $[0, T_1]$ and therefore on all of $\mathbb{R}$ since x_1 is T_1-periodic. So we have a contradiction.

Case 2: The signals on either side are not constant. In this case, they have a fundamental period T_o. Since T_1 is a period of the left-hand side and T_2 is a period of the right-hand side, it follows from Fact 9.2 that $T_1 = mT_o$ and $T_2 = nT_o$ for some positive integers m and n, so, once again, $T_1/T_2 = m/n$ is a rational number. □

Suppose decent signals x_1 and x_2 have respective fundamental periods T_1 and T_2, where T_1/T_2 is a rational number. It turns out that for most choices of nonzero constants c_1 and c_2, the fundamental period of $x = c_1x_1 + c_2x_2$ is the lowest common integer multiple of T_1 and T_2. To figure out that number, first write $T_1/T_2 = m/n$ in lowest terms. Then $T_o = nT_1 = mT_2$ is the lowest common integer multiple of T_1 and T_2. One needs to choose c_1 and c_2 carefully for x to have a fundamental period lower than T_o. For example, if $x_1(t) = \cos(t/2) + \cos(t/3)$ and $x_2(t) = \cos(t/3)$ for all $t \in \mathbb{R}$, then $T_1 = 12\pi$ and $T_2 = 6\pi$, but the fundamental period of $x = x_1 - x_2$ is 4π. On the other hand, for arbitrarily small ϵ, the fundamental period of $x = (1+\epsilon)x_1 - x_2$ is 12π.

Given $T_o > 0$, let X_{T_o} be the set of all decent periodic signals that have T_o as a period. Observe that X_{T_o} contains all the constant signals. Some of the non-constant signals in X_{T_o} will have fundamental period T_o but others will have smaller fundamental periods. By Fact 9.2, the only possible fundamental periods for signals in X_{T_o} are numbers of the form T_o/m, where m is a positive integer. Accordingly, the ratio of the fundamental periods of any two signals in X_{T_o} will be rational. It's clear in any event that X_{T_o} is closed under the taking of linear combinations, so X_{T_o} is a vector space of signals.

Fourier series

Given $T_o > 0$, let $\Omega_o = 2\pi/T_o$. For any complex numbers c_0 and c_1, the signal

$$t \mapsto c_0 + c_1 e^{j\Omega_o t}$$

is periodic and has T_o as a period. If $c_1 \neq 0$, the signal actually has fundamental period T_o. Similarly, for any $c_{-1} \in \mathbb{C}$, the signal

$$t \mapsto c_{-1} e^{-j\Omega_o t} + c_0 + c_1 e^{j\Omega_o t}$$

is periodic and has T_o as a period. The same is true for the signal

$$t \mapsto c_{-2} e^{-j2\Omega_o t} + c_{-1} e^{-j\Omega_o t} + c_0 + c_1 e^{j\Omega_o t} + c_2 e^{j2\Omega_o t} .$$

In fact, for any choices of the complex constants c_k, the signal

$$t \mapsto \sum_{k=-n}^{n} c_k e^{jk\Omega_o t}$$

is periodic and has T_o as a period. Continuing in this fashion, we can assemble a variety of periodic signals, all of which have T_o as a period, by forming linear combinations of terms of the form $e^{jk\Omega_o t}$. The theory of Fourier series tells us that you can build essentially any well behaved periodic signal in this way. Here is the principal result.

9.4 Theorem: Let $x \in \mathbb{C}^{\mathbb{R}}$ be a decent periodic signal that has T_o as a period and is piecewise differentiable on all the intervals between its jumps. Let $\Omega_o = 2\pi/T_o$. Then there exist complex constants c_k, $k \in \mathbb{Z}$, such that the sequence

$$S_N(t) = \sum_{k=-N}^{N} c_k e^{jk\Omega_o t}$$

converges as $N \to \infty$ for every $t \in \mathbb{R}$ as follows:

- if t is a continuity point of x, then $\lim_{N\to\infty} S_N(t) = x(t)$, and
- if t is a jump point of x, then $S_N(t)$ converges as $N \to \infty$ to the mean value of x across the jump, namely $\left(x(t_+) + x(t_-)\right)/2$.

Theorem 9.4, whose proof is beyond our scope, asserts that any decent periodic signal has an "expansion" as an "infinite linear combination" of pure complex exponential sinusoids. The series in Theorem 9.4 is called a *Fourier series* for x in honor of nineteenth-century French mathematician Jean-Baptiste Joseph Fourier. The theorem statement summarizes the detailed meaning of the word "expansion." Often, I'll be a bit casual and write

$$x(t) = \sum_{k=-\infty}^{\infty} c_k e^{jk\Omega_o t} \quad \text{for all } t \in \mathbb{R} , \tag{7}$$

even when x is not continuous. The equation is fully accurate only when x is continuous and piecewise differentiable. Although I can't provide the proof of Theorem

9.4, I can show you how to compute the coefficients c_k given that Theorem 9.4 holds. The computation rests on the following observation.

9.5 Fact: With notation as in the foregoing, if $m \in \mathbb{Z}$, then

$$\int_0^{T_o} e^{jm\Omega_o t} dt = \begin{cases} T_o & \text{if } m = 0 \\ 0 & \text{if } m \neq 0 \, . \end{cases}$$

Proof: Just do the integral. If $m = 0$, the integrand is 1 for all $t \in [0, T_o]$, so the integral evaluates to T_o. If $m \neq 0$, the integral evaluates to

$$\frac{1}{jm\Omega_o}(e^{jm\Omega_o T_o} - 1) = 0 \, ,$$

where the last equality holds because $\Omega_o T_o = 2\pi$ and $e^{j2\pi} = 1$. □

Given a decent periodic signal x that has T_o as a period and satisfies the conditions of Theorem 9.4, let's first compute c_0 using Theorem 9.4. Integrate both sides of (7) from 0 to T_0. The manner in which the series converges ensures that we can integrate the right-hand side term-by-term. By Fact 9.5 every term integrates to zero except the c_0 term. We get

$$\int_0^{T_o} x(t) dt = c_0 T_o \, ,$$

or

$$c_0 = \frac{1}{T_o} \int_0^{T_o} x(t) dt \, .$$

Consequently, c_0 is simply the average value of x over one period, which is also the average value of the entire signal x.

Now suppose we want to compute c_k for some $k \neq 0$. Telescoping in on the series from (7) near the kth term yields

$$x(t) = \cdots + c_{k-1} e^{j(k-1)\Omega_o t} + c_k e^{jk\Omega_o t} + c_{k+1} e^{j(k+1)\Omega_o t} + c_{k+2} e^{j(k+2)\Omega_o t} + \cdots$$

for $t \in \mathbb{R}$. Multiply both sides by $e^{-jk\Omega_o t}$ to obtain

$$x(t)e^{-jk\Omega_o t} = \cdots + c_{k-1} e^{-j\Omega_o t} + c_k + c_{k+1} e^{j\Omega_o t} + c_{k+2} e^{j2\Omega_o t} + \cdots$$

and integrate both sides from 0 to T_o. Fact 9.5 guarantees that all terms on the right-hand side except the solitary c_k will integrate to zero, from which it follows that

$$\int_0^{T_o} x(t) e^{-jk\Omega_o t} dt = c_k T_o \, ,$$

or

$$c_k = \frac{1}{T_o} \int_0^{T_o} x(t) e^{-jk\Omega_o t} dt \, .$$

These formulas enable you to calculate the Fourier coefficients c_k given that you know x in one form or another. Sometimes you have an explicit formula for x. Other times you have a graph of x from which you have to figure out an explicit formula. A work-saver in some circumstances is the following observation: the integrals in the c_k-formulas are from 0 to T_o, but you can actually perform the integrals over any convenient interval of length T_o. This is because the integrands are T_o-periodic, so shifting the interval of integration doesn't change the integrals' values. I'll make the observation explicit by writing

$$c_0 = \frac{1}{T_o}\int_{T_o} x(t)dt$$

and

$$c_k = \frac{1}{T_o}\int_{T_o} x(t)e^{-jk\Omega_o t}dt \ ,$$

where the notation $\int_{T_o}$ means "integral over any interval of length T_o."

Suppose x is a decent periodic signal with T_o as a period and we expand x in a Fourier series as above, i.e.

$$x(t) = \sum_{k=-\infty}^{\infty} c_k e^{jk\Omega_o t} \ \text{for all} \ \ t \in \mathbb{R} \ ,$$

where $\Omega_o = 2\pi/T_o$. Note that T_o need not be the fundamental period of x. It turns out that you can calculate x's fundamental period as follows. Let $\bar{k}$ be the greatest common divisor of all k-values for which $c_k \neq 0$. Then x has fundamental frequency $\bar{k}\Omega_o$ and fundamental period $2\pi/\bar{k}\Omega_o$. Thus, for example, if in the Fourier series above we have $c_k = 0$ for all odd values of k but c_4 and c_6 are nonzero, the fundamental frequency of x is $2\Omega_o$ and the fundamental period of x is $2\pi/2\Omega_o = T_o/2$. Observe that the Fourier series might contain no term at x's fundamental frequency — in this last example we could have had $c_2 = c_{-2} = 0$, and x's fundamental frequency would still have been $2\Omega_o$ provided both c_4 and c_6 were nonzero.

You might have seen Fourier series for real-valued signals in sine-cosine form. It's easy to derive the sine-cosine form from the complex exponential form in Theorem 9.5. First note that if x is real-valued, then c_0 is real and, for every $k \neq 0$, $c_{-k} = \bar{c}_k$, where $\bar{c}_k$ denotes the complex conjugate of c_k. The reason is that

$$\bar{c}_k = \frac{1}{T_o}\int_{T_o} \overline{x(t)e^{-jk\Omega_o t}}dt = \frac{1}{T_o}\int_{T_o} x(t)e^{jk\Omega_o t}dt = \frac{1}{T_o}\int_{T_o} x(t)e^{-j(-k)\Omega_o t}dt = c_{-k} \ .$$

The second equality holds because x is real, so $\overline{x(t)} = x(t)$. Now group the terms in the Fourier series together in pairs to obtain

$$x(t) = c_0 + \sum_{k=1}^{\infty}(c_k e^{jk\Omega_o t} + c_{-k}e^{-jk\Omega_o t}) \ .$$

Because $c_{-k} = \bar{c}_k$, The kth term in the expansion simplifies to

$$\begin{aligned} 2\mathrm{Re}\{c_k e^{jk\Omega_o t}\} &= 2\mathrm{Re}\{c_k\}\cos k\Omega_o t - 2\mathrm{Im}\{c_k\}\sin k\Omega_o t \\ &= a_k \cos k\Omega_o t + b_k \sin k\Omega_o t \ , \end{aligned}$$

where

$$a_k = 2\,\mathrm{Re}\left\{\frac{1}{T_o}\int_{T_o} x(t)e^{-jk\Omega_o t}dt\right\} = \frac{2}{T_o}\int_{T_o} x(t)\cos k\Omega_o t dt$$

and

$$b_k = -2\,\mathrm{Im}\left\{\frac{1}{T_o}\int_{T_o} x(t)e^{-jk\Omega_o t}dt\right\} = \frac{2}{T_o}\int_{T_o} x(t)\sin k\Omega_o t dt \ .$$

With a_k and b_k given by these formulas, the sine-cosine form of the Fourier series for x is

$$x(t) = c_0 + \sum_{k=1}^{\infty} a_k \cos k\Omega_o t + \sum_{k=1}^{\infty} b_k \sin k\Omega_o t \ , \quad t \in \mathbb{R} \ .$$

Sound in general and music in particular are good sources of intuition about periodic signals and Fourier series. Think of a periodic signal as representing a musical tone. The higher the fundamental frequency of the periodic signal, the higher the pitch of the corresponding musical tone. In music, a given pitch is *one octave higher* than another pitch if its fundamental frequency is twice that of the other pitch. On a piano, the note "A above middle-C" has fundamental frequency $f_o = 440$Hz, which corresponds to $\Omega_o = 880\pi$. People generally call it "A-440." The next higher A on the piano keyboard has fundamental frequency 1760π, and the next higher A after that one has fundamental frequency 3520π. The piano keyboard, in this way, plots pitches on a log-frequency scale.

If someone plays an A-440 on a piano, it sounds different from an A-440 played on a French horn or a Fender Stratocaster or a Moog synthesizer. All these notes have the same fundamental frequency $\Omega_o = 880\pi$. What distinguishes them is their so-called *higher harmonics.* Suppose x is some particular A-440. Let $\Omega_o = 880\pi$ and expand x in a Fourier series

$$x(t) = \sum_{k=-\infty}^{\infty} c_k e^{jk\Omega_o t} \ , \quad t \in \mathbb{R}.$$

The sum of the $k = \pm 1$ terms, i.e.

$$t \mapsto c_{-1} e^{-j\Omega_o t} + c_1 e^{j\Omega_o t}$$

is called the *fundamental component* of x. The signal

$$t \mapsto c_{-k} e^{-jk\Omega_o t} + c_k e^{jk\Omega_o t}$$

is called the *kth harmonic component* of x. The relative magnitudes of the various c_k determine how x's "signal energy" distributes itself over the various frequencies $k\Omega_o$.

The A-440 notes on different instruments have different Fourier-series coefficients, and that's at least part of why they sound different. A brassier instrument might have relatively higher-magnitude c_k for large k than would a mellower-sounding instrument. Of course, energy distribution over harmonic components does not tell the whole story about tone-color contrasts between different instruments. The *timbre* of an A-440 played on a given instrument depends not just on Fourier series but on subtle variations in fundamental frequency and volume and a host of other things. One final comment: observe that $k\Omega_o$, the frequency of the kth harmonic, is not generally the frequency of an A because it is not an integer number of octaves above A-440 unless $k = 2^m$. Thus higher harmonics are not simply higher-octave versions of the "fundamental note."

Inner-product spaces and Hilbert spaces

Perhaps the cleanest and most intuitively appealing mathematical approach to Fourier series is through orthogonal expansions in inner product spaces. I won't be able to give you all the details here, but the essential ideas are relevant not only to Fourier series but to other orthogonal expansions that arise in applications.

9.6 Definition: A *complex inner product space* is a vector space V over $\mathbb{C}$ together with a mapping $(v, w) \to \langle v, w\rangle$ from $V \times V$ into $\mathbb{C}$ that has the following properties:

- $\langle v, v\rangle \geq 0$ for every $v \in V$, and $\langle v, v\rangle = 0$ if and only if $v = 0$.
- $\langle w, v\rangle = \overline{\langle v, w\rangle}$ for every v and w in V.
- For every v_1, v_2, v_3 in V and every c_1, c_2, c_3 in $\mathbb{C}$,

$$\begin{aligned} \langle c_1v_1 + c_2v_2, v_3\rangle &= c_1\langle v_1, v_3\rangle + c_2\langle v_2, v_3\rangle \\ \langle v_1, c_2v_2 + c_3v_3\rangle &= \bar{c}_2\langle v_1, v_2\rangle + \bar{c}_3\langle v_1, v_3\rangle\ . \end{aligned}$$

The complex number $\langle v, w\rangle$ is called the *inner product of* v *with* w.

An extraordinarily useful property of any inner product space is the Schwarz Inequality, also known as the Cauchy-Schwarz Inequality.

9.7 Schwarz Inequality: With notation as in the foregoing, if V is an inner product space, then

$$|\langle v, w\rangle| \leq \sqrt{\langle v, v\rangle}\,\sqrt{\langle w, w\rangle}$$

for every v and w in V.

Proof: The inequality clearly holds when $w = 0$, so assume $w \neq 0$. Consider

$$\begin{aligned} 0 &\leq \left\langle v - \frac{\langle v, w\rangle}{\langle w\,,\,w\rangle} w\,,\, v - \frac{\langle v, w\rangle}{\langle w, w\rangle} w \right\rangle \\ &= \left\langle v - \frac{\langle v, w\rangle}{\langle w, w\rangle} w, v \right\rangle - \frac{\langle w, v\rangle}{\langle w, w\rangle} \left\langle v - \frac{\langle v, w\rangle}{\langle w, w\rangle} w, w \right\rangle\ . \end{aligned}$$

The second term on the second line is zero and the first term evaluates to $\langle v, v\rangle - |\langle v, w\rangle|^2/\langle w, w\rangle$, from which it follows that

$$|\langle v, w\rangle|^2 \leq \langle v, v\rangle\,\langle w, w\rangle\ ,$$

which is equivalent to the Schwarz Inequality. □

We learned in Chapter 4 about norms on vector spaces and their associated notions of convergence. Conveniently, the inner product on an inner product space V gives rise to a special associated norm on V defined by

$$\|v\| = \sqrt{\langle v, v\rangle} \ \text{ for every } \ v \in V\ .$$

Proving that $v \mapsto \|v\|$ defines a norm on V depends on the Schwarz Inequality. The triangle inequality

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

arises from

$$\begin{aligned}
\|v+w\|^2 &= \langle v+w\,,\,v+w\rangle \\
&= \|v\|^2 + \langle v,w\rangle + \langle w,v\rangle + \|w\|^2 \\
&= \|v\|^2 + 2\mathrm{Re}\{\langle v,w\rangle\} + \|w\|^2 \\
&\leq \|v\|^2 + 2|\langle v,w\rangle| + \|w\|^2 \\
&\leq \|v\|^2 + 2\|v\|\,\|w\| + \|w\|^2 \\
&= (\|v\| + \|w\|)^2 \ ,
\end{aligned}$$

where the inequality on the second-to-last line holds because of the Schwarz Inequality.

The norm in turn spawns a distance function on V: the distance between v and w is $\|v-w\|$. Along with the distance function comes a notion of convergence: a sequence $\{v_n\}$ in V converges to $v \in V$ when

$$\lim_{n\to\infty} \|v_n - v\| = 0 \ .$$

In analogy to real and complex numbers, a sequence $\{v_n\}$ in V is a *Cauchy sequence* when for every $\epsilon > 0$ there exists an $N > 0$ such that $\|v_m - v_n\| < \epsilon$ when m and n are bigger than N. Every convergent sequence in V is a Cauchy sequence. If, in addition, every Cauchy sequence in V converges, V is called a *Hilbert space.* Hilbert spaces are named after the great nineteenth- and twentieth-century German mathematician David Hilbert.

A familiar inner product space is $\mathbb{C}^n$, the set of all column n-vectors with complex entries. The inner product between two such vectors v and w is $\langle v,w\rangle = w^H v$, where w^H is the *Hermitian conjugate* of w defined by $\overline{w}^T$, the complex conjugate of the transpose of w. Note that $\langle v,w\rangle = \sum_{i=1}^n [v]_i \overline{[w]_i}$, where $[v]_i$ and $[w]_i$ are respectively the ith entries in v and w. In particular, the norm associated with the inner product is the standard Euclidean norm on $\mathbb{C}^n$ given by

$$\|v\| = \langle v,v\rangle^{1/2} = \left(\sum_{i=1}^n |[v]_i|^2\right)^{1/2} \ ,$$

where $[v]_i$ is the ith entry in v. It's easy to show that every Cauchy sequence in $\mathbb{C}^n$ has a limit, so $\mathbb{C}^n$ is a Hilbert space.

Another inner product space is the set l^2 of all square-summable complex-valued discrete-time signals that we first met in Chapter 5. Elements of l^2 are signals $x \in \mathbb{C}^{\mathbb{Z}}$ for which $\sum_{n=-\infty}^{\infty} |x(n)|^2$ converges. We saw in Chapter 5 that l^2 is indeed a vector space. For l^2 signals x and y, define the inner product of x and y via

$$\langle x,y\rangle = \sum_{n=-\infty}^{\infty} x(n)\overline{y(n)} \ .$$

The norm associated with this inner product is the standard l^2 norm

$$\langle x,x\rangle^{1/2} = \left(\sum_{n=-\infty}^{\infty} |x(n)|^2\right)^{1/2} = \|x\|_2 \ .$$

Similarly, the set L^2 of complex-valued square-integrable continuous-time signals, which we learned in Chapter 7 is a vector space, is an inner product space with inner product defined by

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

for all x and y in L^2 and associated norm

$$\langle x, x\rangle^{1/2} = \left(\int_{-\infty}^{\infty} |x(t)|^2 dt\right)^{1/2} = \|x\|_2 \ .$$

For the record, l^2 and L^2 are both Hilbert spaces, a fact that's not simple to prove, especially for L^2.

The Schwarz Inequality enables us to improve on the upper bounds we established in Criterion 5.4 on the infinity norm of the convolution of two l^2-signals and in Criterion 7.7 on the infinity norm of the convolution of two L^2-signals. Suppose x_1 and x_2 are l^2 signals. Given $n \in \mathbb{Z}$, let $y \in \mathbb{C}^{\mathbb{Z}}$ have specification

$$y(k) = \overline{x_2(n-k)} \ \text{ for all } \ k \in \mathbb{Z} \ .$$

Observe that $\|y\|_2 = \|x_2\|_2$ and that

$$\begin{aligned} x_1 * x_2(n) &= \sum_{k=-\infty}^{\infty} x_1(k)x_2(n-k) \\ &= \sum_{k=-\infty}^{\infty} x_1(k)\overline{y(k)} \\ &= \langle x_1, y\rangle \ . \end{aligned}$$

From the Schwarz Inequality it follows that

$$|x_1 * x_2(n)| \le \|x_1\|_2 \, \|y\|_2 = \|x_1\|_2 \, \|x_2\|_2 \ \text{ for all } \ n \in \mathbb{Z} \ ,$$

whereby

$$\|x_1 * x_2\|_\infty \le \|x_1\|_2 \, \|x_2\|_2 \ ,$$

a tighter bound than the one in Criterion 5.4. A parallel continuous-time argument yields

$$\|x_1 * x_2\|_\infty \le \|x_1\|_2 \, \|x_2\|_2 \ ,$$

an improvement on Criterion 7.7, when x_1 and x_2 are L^2-signals.

Orthogonal expansions

Inner-product spaces come equipped with a geometric structure that ordinary vector spaces lack. Inner products generalize the dot product between vectors in $\mathbb{R}^3$ that you learned about in pre-calculus. Remember how two little arrows emanating from the origin are perpendicular if and only if their dot product is zero? Here's a more sophisticated rendition of that property.

9.8 Definition: Two vectors v and w in an inner product space V are *orthogonal* when $\langle v, w\rangle = 0$. A finite or countably infinite set $\{w_k\}$ of vectors is an *orthonormal set* when

$$\langle w_k, w_l\rangle = \begin{cases} 1 & \text{if } \ k = l \\ 0 & \text{if } \ k \neq l \ . \end{cases}$$

An orthonormal set $\{w_k\}$ is *complete* when the only $v \in V$ satisfying $\langle v, w_k \rangle = 0$ for all k is $v = 0$.

The vectors in an orthonormal set are "mutually perpendicular" and have "unit length" in the sense that $\|w_k\| = \sqrt{\langle w_k, w_k \rangle} = 1$ for every k. The customary standard basis $\left(e^1, e^2, \dots e^n\right)$ for $\mathbb{C}^n$ with the usual inner product is a complete orthonormal set in $\mathbb{C}^n$. Recall that $\left[e^k\right]_i = 1$ when $i = k$ and 0 when $i \neq k$. In l^2 with the usual inner product, that the set $\{w_k : k \in \mathbb{Z}\}$ of square-integrable signals defined by

$$w_k = \mathrm{Shift}_k(\delta) \ \text{ for all } \ k \in \mathbb{Z}$$

is certainly an orthonormal set. Note in addition that for any $x \in l^2$ we have

$$\langle x, w_k \rangle = \sum_{n=-\infty}^{\infty} x(n)\overline{\mathrm{Shift}_k(\delta)(n)} = \sum_{n=-\infty}^{\infty} x(n)\delta(n-k) = x(k)$$

for every $k \in \mathbb{Z}$. Thus the only x satisfying $\langle x, w_k \rangle = 0$ for every $k \in \mathbb{Z}$ is $x = 0$ since $x = 0$ is the only x satisfying $x(k) = 0$ for all $k \in \mathbb{Z}$. It follows that $\{w_k\}$ is a complete orthonormal set in l^2.

A finite orthonormal set $\{w_1, \dots, w_n\}$ is also a linearly independent set — just take the inner product of each w_k with any expression of the form

$$c_1 w_1 + \cdots + c_n w_n = 0 \ ,$$

and you'll discover that $c_k = 0$. It follows that $\{w_k : 1 \leq k \leq n\}$ spans an n-dimensional subspace W of V, so $(w_1, w_2, \dots, w_n)$ is an orthonormal basis of W. As it happens, every finite-dimensional subspace of an inner product space possesses an orthonormal basis.

9.9 Fact: If W is an n-dimensional subspace of an inner product space V, then W has an orthonormal basis $(w_1, w_2, \dots, w_n)$.

Proof: Theorem 4.3 guarantees the existence of a basis $(v_1, v_2, \dots, v_n)$ for W. We'll construct an orthonormal basis inductively using what's known as the *Gram-Schmidt procedure.* Define $u_1 = v_1$ and $w_1 = u_1/\|u_1\|$, Note that $\|w_1\| = 1$ and $\mathrm{span}(\{w_1\}) = \mathrm{span}(\{v_1\})$. Suppose we've defined orthonormal vectors w_1, w_2, ... , w_k such that

$$\mathrm{span}\left(\{w_1, w_2, \dots, w_k\}\right) = \mathrm{span}\left(\{v_1, v_2, \dots, v_k\}\right) \ ,$$

where $k < n$. Set

$$u_{k+1} = v_{k+1} - \sum_{l=1}^{k} \langle v_{k+1}, w_l \rangle w_l \ .$$

$u_{k+1} \neq 0$ because the sum on the right-hand side lies in the span of v_l for $l \leq k$ and the v_l for $l \leq k+1$ are linearly independent. Now set $w_{k+1} = u_{k+1}/\|u_{k+1}\|$. It's easy to check that $\{w_1, w_2, \dots, w_{k+1}\}$ is an orthonormal set whose span is the same as $\mathrm{span}\left(\{v_1, v_2, \dots, v_{k+1}\}\right)$. Keep this up until $k+1 = n$ and you end up with an orthonormal spanning set $\{w_1, w_2, \dots, w_n\}$ for W. The set, being orthonormal, is linearly independent, so $(w_1, w_2, \dots, w_n)$ is an orthonormal basis of W. □

Given $v \in V$ and a finite-dimensional subspace W of V, it is often of interest to find the vector in W "closest to v" in the sense that it minimizes $\|v - w\|^2$ over $w \in W$. The vector

$$\widehat{v} = \sum_{k=1}^{n} \langle v, w_k \rangle w_k$$

solves that problem when $\{w_1, \ldots, w_n\}$ is an orthonormal basis for W. To see this, note first that $v - \widehat{v}$ is orthogonal to each w_l, since

$$\langle v - \widehat{v}, w_l \rangle = \left\langle v - \sum_{k=1}^{n} \langle v, w_k \rangle w_k \, , \, w_l \right\rangle = \langle v, w_l \rangle - \langle v, w_l \rangle \|w_l\|^2 = 0 \, .$$

The second equality holds because $\langle w_k, w_l \rangle = 0$ when $k \neq l$ and the last equality holds because $\|w_l\| = 1$. Next, observe that we can write an arbitrary $w \in W$ as

$$w = \widehat{v} + \sum_{k=1}^{n} \Delta_k w_k$$

for some complex numbers Δ_k, so

$$\begin{aligned} \|v - w\|^2 &= \langle v - w, v - w \rangle \\ &= \left\langle (v - \widehat{v}) - \sum_{k=1}^{n} \Delta_k w_k \, , \, (v - \widehat{v}) - \sum_{k=1}^{n} \Delta_k w_k \right\rangle \\ &= \|v - \widehat{v}\|^2 + \sum_{k=1}^{n} |\Delta_k|^2 \, , \end{aligned}$$

where the last equality holds because $v - \widehat{v}$ is orthogonal to each w_k and the w_k are orthonormal. It's obvious from the last line that choosing $\Delta_k = 0$ for all k minimizes $\|v - w\|^2$. The vector $\widehat{v}$ is called the *orthogonal projection* of v on the subspace spanned by $\{w_1, \ldots, w_n\}$.

A complete orthonormal set $\mathcal{W} = \{w_k\}$ in a Hilbert space V serves as an orthonormal basis for V in the sense that every $v \in V$ possesses an expansion of the form

$$v = \sum_{k} c_k w_k$$

for some coefficients $c_k \in \mathbb{C}$. If $\mathcal{W}$ is finite, say $\mathcal{W} = \{w_1, w_2, \ldots, w_n\}$, then the sum in the last equation is finite. If $\mathcal{W}$ is countably infinite, say $\mathcal{W} = \{w_k : k \in \mathbb{Z}\}$, then the rigorous meaning of the expression

$$v = \sum_{k=-\infty}^{\infty} c_k w_k$$

is

$$\lim_{n \to \infty} \left\| v - \sum_{k=-n}^{n} c_k w_k \right\| = 0 \, .$$

As you might expect given the orthogonal-projection discussion above, the coefficients c_k are given by the formula

$$c_k = \langle v, w_k \rangle \text{ for all } k \, .$$

To get an intuitive grasp of what's happening when $\mathcal{W}$ is infinite and V is therefore infinite-dimensional, set

$$W_n = \text{span}\left(\{w_k : -n \leq k \leq n\}\right) \quad \text{for all } \; n \in \mathbb{N} \; .$$

W_n is a $(2n+1)$-dimensional subspace of V for each n, and these subspaces are nested in the sense that

$$W_0 \subset W_1 \subset \cdots W_{n-1} \subset W_n \subset W_{n+1} \cdots$$

Furthermore, for each n

$$S_n(v) = \sum_{k=-n}^{n} \langle v, w_k \rangle w_k$$

is the orthogonal projection of v onto the subspace W_n. As n grows, W_n gets larger and $S_n(v)$ gets closer to v. Completeness of the orthonormal set $\mathcal{W}$ ensures that W_n expands as $n \to \infty$ to encompass all of V and that $S_n(v)$ converges to v, i.e.

$$\lim_{n\to\infty} \left\| v - \sum_{k=-n}^{n} \langle v, w_k \rangle w_k \right\| = 0 \; .$$

Proving this central result requires two supporting facts. The first, regarding the convergence of inner products and infinite sequence, is an important consequence of the Schwarz Inequality. The second establishes a one-to-one correspondence between l^2-sequences and "infinite linear combinations" of a countably infinite set of orthonormal vectors.

9.10 Fact: If the sequence $\{v_n\}$ in an inner product space V converges to $v \in V$, then $\langle v_n, w \rangle$ converges to $\langle v, w \rangle$ for every $w \in V$. More generally, if the sequences $\{v_n\}$ and $\{w_n\}$ converge, respectively, to v and w, then $\langle v_n, w_n \rangle$ converges to $\langle v, w \rangle$.

Proof: The first assertion follows from the second if we take $w_n = w$ for all n, so let's just prove the second assertion. Observe that

$$\langle v_n, w_n \rangle - \langle v, w \rangle = \langle v_n - v, w_n - w \rangle + \langle v_n - v, w \rangle + \langle v, w_n - w \rangle \; .$$

Take absolute values of both sides and apply the Schwarz inequality to the right-hand side to obtain

$$|\langle v_n, w_n \rangle - \langle v, w \rangle| \leq \|v_n - v\| \, \|w_n - w\| + \|v_n - v\| \, \|w\| + \|v\| \, \|w_n - w\| \; .$$

All the terms on the right-hand side go to zero as $n \to \infty$ because $\|v_n - v\|$ and $\|w_n - w\|$ go to zero as $n \to \infty$. It follows that $\lim_{n\to\infty} |\langle v_n, w_n \rangle - \langle v, w \rangle| = 0$. □

9.11 Fact: Let $\{w_k : k \in \mathbb{Z}\}$ be an orthonormal set in a Hilbert space V. Let $\{c_k : k \in \mathbb{Z}\}$ be a sequence of complex numbers. For each $n \in \mathbb{Z}$, let $S_n = \sum_{k=-n}^{n} c_k w_k$. Then the sequence $\{S_n\}$ converges, i.e. there exists $v \in V$ such that

$$\lim_{n\to\infty} \|S_n - v\| = \lim_{n\to\infty} \left\| \sum_{k=-n}^{n} c_k w_k - v \right\| = 0 \; ,$$

if and only if $\{c_k\}$ is an l^2-sequence.

Proof: First observe that if $m > n$, we have

$$\begin{aligned}\|S_m - S_n\|^2 &= \left\langle \sum_{k=-m}^{-n-1} c_k w_k + \sum_{k=n+1}^{m} c_k w_k \, , \sum_{k=-m}^{-n-1} c_k w_k + \sum_{k=n+1}^{m} c_k w_k \right\rangle \\ &= \sum_{k=-m}^{-n-1} |c_k|^2 + \sum_{k=n+1}^{m} |c_k|^2 \, ,\end{aligned}$$

where the last equality follows from the orthonormality of $\{w_k\}$. Saying $\{c_k\}$ is an l^2-sequence is the same as saying that $\sum_{k=-\infty}^{\infty} |c_k|^2$ converges, which happens if and only if for every $\epsilon > 0$ we can find $N \in \mathbb{Z}$ such that when $m > n > N$ we have

$$\sum_{k=-m}^{-n-1} |c_k|^2 + \sum_{k=n+1}^{m} |c_k|^2 < \epsilon \, .$$

In turn, this is the same as saying that $\{S_n\}$ is a Cauchy sequence in V. It follows that $\{c_k\}$ is an l^2-sequence if and only if $\{S_n\}$ is a Cauchy sequence, which is true if and only if $\{S_n\}$ converges to some $v \in V$ since V is a Hilbert space. □

9.12 Theorem: Let V be a Hilbert space and let $\mathcal{W} = \{w_k\}$ be a finite or countably infinite complete orthonormal set in V. If $\mathcal{W} = \{w_1, \ldots, w_n\}$ is finite, then $v = \sum_{k=1}^{n} \langle v, w_k \rangle w_k$ for every $v \in V$. If $\mathcal{W} = \{w_k : k \in \mathbb{Z}\}$ is countably infinite, then for any $v \in V$

$$\lim_{n \to \infty} \|v - S_n(v)\| = 0 \, ,$$

where

$$S_n(v) = \sum_{k=-n}^{n} \langle v, w_k \rangle w_k \, .$$

Proof: If $\mathcal{W}$ is finite, then $v - \sum_{k=1}^{n} \langle v, w_k \rangle w_k$ is orthogonal to each w_l and is therefore zero by completeness of $\mathcal{W}$. That disposes of the finite-$\mathcal{W}$ case. Suppose, then, that $\mathcal{W}$ is infinite. For each $n \in \mathbb{Z}$, $S_n(v)$ is the orthogonal projection on $\operatorname{span}(\{w_k : -n \le k \le n\})$, so

$$\langle v - S_n(v), w_l \rangle = 0 \text{ when } -n \le l \le n \, .$$

Since $S_n(v)$ is a linear combination of $\{w_k : -n \le k \le n\}$, it follows that $v - S_n(v)$ is orthogonal to $S_n(v)$, so

$$\|v\|^2 = \langle S_n(v) + (v - S_n(v)) \, , \, S_n(v) + (v - S_n(v)) \rangle = \|S_n(v)\|^2 + \|v - S_n(v)\|^2 \, .$$

As a result,

$$\|S_n(v)\|^2 \le \|v\|^2$$

for every $n \in \mathbb{Z}$. Observe next that

$$\|S_n(v)\|^2 = \langle S_n(v), S_n(v) \rangle = \sum_{k=-n}^{n} |\langle v, w_k \rangle|^2$$

by orthonormality of $\{w_k\}$, so

$$\sum_{k=-n}^{n} |\langle v, w_k\rangle|^2 \leq \|v\|^2 \quad \text{for all } \; n \in \mathbb{Z} \; ,$$

implying that

$$\sum_{k=-\infty}^{\infty} |\langle v, w_k\rangle|^2$$

converges.

Applying Fact 9.11 with the substitutions $c_k = \langle v, w_k\rangle$ and $S_n = S_n(v)$ allows us to conclude that the sequence $\{S_n(v)\}$ converges to some limit. That limit turns out to be v itself. To see why, make the following observations.

- The sequence $\{v - S_n(v)\}$ must also converge, since $\{S_n(v)\}$ does.
- Let $v_0 = \lim_{n\to\infty}(v - S_n(v))$. Because for every $k \in \mathbb{Z}$ we have $\langle v - S_n(v), w_k\rangle = 0$ for all $n \geq k$, we know that $\lim_{n\to\infty}\langle v - S_n(v), w_k\rangle = 0$ for every $k \in \mathbb{Z}$. It follows from Fact 9.10 that $\langle v_0, w_k\rangle = 0$ for every $k \in \mathbb{Z}$.
- Hence $v_0 = 0$ because $\mathcal{W}$ is a complete orthonormal set.

The bottom line is that

$$v = \lim_{n\to\infty} S_n(v) \; ,$$

i.e.

$$\lim_{n\to\infty} \|v - S_n(v)\| = 0 \; ,$$

which completes the proof. □

Fourier series are orthogonal expansions

Back now to Fourier series. We've noted already that, given $T_o > 0$, the set X_{T_o} of complex-valued decent signals having T_o as a period is closed under linear combination and is therefore a vector space of signals. In fact, X_{T_o} is an inner product space with inner product

$$\langle x_1, x_2\rangle = \frac{1}{T_o}\int_0^{T_o} x_1(t)\overline{x_2(t)}dt \;\; \text{for all } \; x_1, x_2 \in X_{T_o}$$

and associated norm

$$\|x\| = \left(\frac{1}{T_o}\int_0^{T_o} |x(t)|^2 dt\right)^{1/2} \quad \text{for all } \; x \in X_{T_o} \; .$$

For each $k \in \mathbb{Z}$, let w_k be the signal $t \mapsto e^{jk\Omega_o t}$, where $\Omega_o = 2\pi/T_o$. Then the countably infinite set $\mathcal{W} = \{w_k : k \in \mathbb{Z}\}$ is an orthonormal set in X_{T_o} by Fact 9.5. Furthermore, $\mathcal{W}$ is actually a complete orthonormal set in X_{T_o}. This last statement is tough to prove, so you'll have to take my word on it.

Inconveniently, X_{T_o} isn't a Hilbert space because there exist Cauchy sequences in X_{T_o} that don't have limits in X_{T_o}. We circumvent this difficulty by embedding X_{T_o} in a Hilbert space endowed with the "same" inner product as X_{T_o}. Let $L^2_{T_o}$ be the set of all periodic signals x — decent or not — that have T_o as a period and for which $\int_0^{T_o} |x(t)|^2 dt$ is finite. $L^2_{T_o}$ includes X_{T_o} and the inner product on X_{T_o} extends nicely to $L^2_{T_o}$. Furthermore, $L^2_{T_o}$ is a Hilbert space and $\mathcal{W}$ is a complete orthonormal set in $L^2_{T_o}$. I'll caution you that I'm glossing over some important Lebesgue-measure issues that pertain to this setup, but you won't go wrong by ignoring them for now.

Theorem 9.12 applies to the Hilbert space $L^2_{T_o}$ and the complete orthonormal set $\mathcal{W}$. For any $x \in L^2_{T_o}$, we have

$$\langle x, w_k \rangle = \frac{1}{T_o} \int_0^{T_o} x(t) \overline{e^{jk\Omega_o t}} dt = \frac{1}{T_o} \int_0^{T_o} x(t) e^{-jk\Omega_o t} dt \ ,$$

and, in the notation of Theorem 9.12, $S_n(x)$ is the signal

$$t \mapsto \sum_{k=-n}^{n} \langle x, w_k \rangle e^{jk\Omega_o t} \ .$$

Theorem 9.12 implies that

$$\lim_{n\to\infty} \|x - S_n(x)\|^2 = \lim_{n\to\infty} \frac{1}{T_o} \int_0^{T_o} \left| x(t) - \sum_{k=-n}^{n} \langle x, w_k \rangle e^{jk\Omega_o t} \right|^2 dt = 0 \ .$$

The foregoing argument applies to every $x \in L^2_{T_o}$. In particular, it applies to every $x \in X_{T_o}$. The formula for $\langle x, w_k \rangle$ is the same as the formula for the Fourier-series coefficient c_k. What's new about the orthogonal-expansion take on Fourier series for decent signals is Theorem 9.12's assertion that the Fourier series converges in the "L^2 sense," also called the mean-square sense. That assertion applies to the Fourier series for any signal in X_{T_o}, not just those satisfying the regularity conditions of Theorem 9.4.

Applying the argument in the proof of Theorem 9.12 to Fourier series yields as a by-product a classic result known as Parseval's Theorem. Referring to that proof and substituting x for v, where x is a decent T_o-periodic signal, yields

$$\|S_n(x)\|^2 = \sum_{k=-n}^{n} |\langle x, w_k \rangle|^2 \ \text{ for all } n \in \mathbb{Z} \ ,$$

where w_k is the signal $t \mapsto e^{jk\Omega_o t}$. Since $\lim_{n\to\infty} S_n(x) = x$, it follows from Fact 9.10 that

$$\lim_{n\to\infty} \|S_n(x)\|^2 = \lim_{n\to\infty} \langle S_n(x), S_n(x) \rangle = \|x\|^2 \ .$$

Accordingly,

$$\|x\|^2 = \frac{1}{T_o} \int_0^{T_o} |x(t)|^2 dt = \lim_{n\to\infty} \sum_{k=-n}^{n} |\langle x, w_k \rangle|^2 = \sum_{k=-\infty}^{\infty} |\langle x, w_k \rangle|^2 \ .$$

Thus we have equality between the norm of x in X_{T_o} and the l^2-norm of the sequence of Fourier coefficients for x. From the identity

$$\langle v, w \rangle = \frac{\|v+w\|^2 - \|v\|^2 - \|w\|^2}{2} + j\frac{\|v+jw\|^2 - \|v\|^2 - \|w\|^2}{2} \ ,$$

which holds for all v and w in any complex inner product space V, we obtain the following.

9.13 Parseval's Theorem: Let x and y be decent periodic signals that have T_o as a period. Let $\Omega_o = 2\pi/T_o$ and let

$$c_k = \frac{1}{T_o}\int_0^{T_o} x(t)e^{-jk\Omega_o t}dt \text{ for all } k \in \mathbb{Z}$$

and

$$d_k = \frac{1}{T_o}\int_0^{T_o} y(t)e^{-jk\Omega_o t}dt \text{ for all } k \in \mathbb{Z}\ .$$

Then $\{c_k\}$ and $\{d_k\}$ are l^2 sequences. Furthermore,

$$\frac{1}{T_o}\int_0^{T_o} x(t)\overline{y(t)}dt = \sum_{k=-\infty}^{\infty} c_k\overline{d_k}$$

and, in particular,

$$\frac{1}{T_o}\int_0^{T_o} |x(t)|^2 dt = \sum_{k=-\infty}^{\infty} |c_k|^2\ .$$

Finally, it would be mildly irresponsible of me not to show you at least one example of a signal $x \in L^2_{T_o}$ that's not a decent signal and therefore not in X_{T_o}. Accordingly, let x be the signal with fundamental period T_o whose specification on the interval $(0, T_o]$ is $x(t) = t^{-1/3}$. The full specification for x is

$$x(t) = (t - nT_o)^{-1/3} \text{ for } nT_o < t \leq (n+1)T_o \text{ and } n \in \mathbb{Z}\ .$$

The signal x isn't decent because it's not bounded on bounded intervals. Nonetheless, as you can check for yourself, $x \in L^2_{T_o}$.

Why study Fourier series?

Compelling answers abound, but I'll focus on two of them. First, from the standpoint of signal analysis, one might argue that pure sinusoids are paradigmatic periodic signals. Fourier series provide the means to "decompose" an arbitrary periodic signal into an "infinite linear combination" of pure sinusoids. A central goal of signal analysis is to explain complicated things in terms of simpler things, and Fourier series play an important role in doing just that.

Fourier series also enable us to get a grip on the concept of *frequency content* of periodic signals. If x is periodic and has fundamental period T_o, the Fourier series

$$x(t) = \sum_{k=-\infty}^{\infty} c_k e^{jk\Omega_o t} \quad \text{for all } t \in \mathbb{R}\ ,$$

where $\Omega_o = 2\pi/T_o$, tells us that x has frequency content concentrated on the discrete set of frequencies $\{k\Omega_o : k \in \mathbb{Z}\}$. The magnitudes of the various c_k indicate how the signal energy in x parcels out over these various frequencies. People refer to the Fourier series for x as a *frequency-domain description* of x for that reason. Describing a periodic signal x in terms of its frequency content characterizes x completely. A frequency-domain description of x doesn't just "tell you something about x." Instead, it tells you everything about x, but in a manner different from specifying x, graphically or analytically, as a periodic time function. It's simply another way of looking at x, as if from a new angle.

Of comparable importance are systems-analysis reasons for studying Fourier series. Suppose S is the system mapping of a LTI system with input space X, and suppose that $x \in X$ is a decent periodic signal that has T_o as a period. Then

$$\mathrm{Shift}_{T_o}(S(x)) = S(\mathrm{Shift}_{T_o}(x)) = S(x) \ .$$

The first equality holds because of time-invariance and the second because T_o is a period of x. So if x is periodic and T_o is a period of x, then T_o is also a period of $S(x)$. In particular, $S(x)$ is periodic and in essentially every case will be decent.

Suppose, then, that $x \in X$ is decent and has T_o as a period. We can expand x in a Fourier series

$$x(t) = \sum_{k=-\infty}^{\infty} c_k e^{jk\Omega_o t} \ \text{ for all } \ t \in \mathbb{R} \ ,$$

where $\Omega_o = 2\pi/T_o$. Since $S(x)$ also has T_o as a period we can, assuming $S(x)$ is decent, expand it in its own Fourier series

$$S(x)(t) = \sum_{k=-\infty}^{\infty} c'_k e^{jk\Omega_o t} \ \text{ for all } \ t \in \mathbb{R} \ .$$

How are the c'_k related to the c_k?

The crucial observation is that any signal of the form $t \mapsto e^{j\Omega_1 t}$, where $\Omega_1 \in \mathbb{R}$, is an "eigen-input" for any LTI system that admits it as an input. Suppose a LTI system has impulse response h. Define x via $x(t) = e^{j\Omega_1 t}$ for every $t \in \mathbb{R}$. If x is an admissible input for the system — i.e., if $x \in \mathcal{D}_h$ — then $S(x) = h * x$, so

$$\begin{aligned} S(x)(t) &= \int_{-\infty}^{\infty} h(\tau)x(t-\tau)d\tau \\ &= \int_{-\infty}^{\infty} h(\tau)e^{j\Omega_1(t-\tau)}d\tau \\ &= \left(\int_{-\infty}^{\infty} h(\tau)e^{-j\Omega_1 \tau}d\tau\right) e^{j\Omega_1 t} \ \text{ for all } \ t \in \mathbb{R} \ . \end{aligned}$$

In other words, $S(x) = (\text{constant}) \times x$, which is why I call x an "eigen-input" for the system. The quantity in the parentheses in the last equation depends on Ω_1, the fundamental frequency of x. If we think of varying Ω_1 over all $\Omega \in \mathbb{R}$, we get a function $\widehat{H}$ of Ω.

9.14 Definition: Let h be the impulse response of a LTI system and suppose that for every $\Omega \in \mathbb{R}$ the signal $t \mapsto e^{j\Omega t}$ lies in $\mathcal{D}_h$ and is therefore an admissible system input. In this case, we say that *the system has a frequency response* and define the *frequency response* of the system as the function $\Omega \mapsto \widehat{H}(\Omega)$ specified by

$$\widehat{H}(\Omega) = \int_{-\infty}^{\infty} h(\tau)e^{-j\Omega\tau}d\tau \ \text{ for all } \ \Omega \in \mathbb{R} \ .$$

If a LTI system has frequency response $\widehat{H}$, then the amount by which the system "scales" an eigen-input of the form $t \mapsto x(t) = e^{j\Omega_1 t}$ is the value of $\widehat{H}$ at frequency Ω_1. In other words,

$$S(x) = \widehat{H}(\Omega_1)x$$

for such an input x. The system tends to amplify pure sinusoids of frequency Ω_1 when $\widehat{H}(\Omega_1)$ is large and attenuate pure sinusoids of frequency Ω_1 when $\widehat{H}(\Omega_1)$ is small. In this way, a LTI system with a frequency response acts as a *frequency-selective filter.* I'd like to emphasize that not every LTI system has a frequency response. The continuous-time integrator doesn't have one because it doesn't admit pure sinusoidal inputs. On the other hand, every BIBO-stable system, in particular every system with a decent finite-duration impulse response, has a frequency response.

Let's return now to the Fourier series for $S(x)$, where x is a signal that has T_o as a period and Fourier series

$$x(t) = \sum_{k=-\infty}^{\infty} c_k e^{jk\Omega_o t} \ \text{ for all } \ t \in \mathbb{R} \ ,$$

where $\Omega_o = 2\pi/T_o$. The kth term in the series is the signal $t \mapsto c_k e^{jk\Omega_o t}$, which is an eigen-input of the kind we've been discussing. If we use $t \mapsto e^{jk\Omega_o t}$ as input to the system, the output that arises is the signal $t \mapsto \widehat{H}(k\Omega_o)e^{jk\Omega_o t}$. Provided the series converges appropriately, we can conclude that

$$S(x)(t) = \sum_{k=-\infty}^{\infty} \left(c_k \widehat{H}(k\Omega_o)\right) e^{jk\Omega_o t} \ \text{ for all } \ t \in \mathbb{R} \ .$$

The Fourier series serves in this way as an "eigenfunction expansion" in that it decomposes a periodic input x into an "infinite linear combination" of eigen-inputs to the system. The system processes these eigen-inputs in a simple way, and the expansion of x gives rise to a similar expansion for $S(x)$. If c_k is the kth Fourier coefficient for x, then the kth Fourier coefficient c'_k of $S(x)$ is simply $c'_k = c_k \widehat{H}(k\Omega_o)$ for every $k \in \mathbb{Z}$. In essence, the system re-shapes the frequency content of a T_o-periodic input according to the relative magnitudes of $\widehat{H}(k\Omega_o)$ for various values of k. Thus the system's frequency response gives us a handle on how the output's energy distribution over harmonic components differs from or resembles the input's energy distribution over harmonics. Fourier series make these ideas transparent, and that's part of why we study them.

CHAPTER 10

Continuous-time Fourier Transforms

As we saw in Chapter 9, the notion of frequency content makes perfect sense for periodic signals. If x is a periodic signal that has T_o as a period, x might have frequency content at any number of frequencies. All those frequencies are of the form $k\Omega_o$, where $\Omega_o = 2\pi/T_o$ and $k \in \mathbb{Z}$. The coefficients in the Fourier series for x show how the frequency content in x is distributed over these frequencies. Why might one expect non-periodic signals to have frequency content? I'll attempt first to demonstrate by means of an example that such an expectation is not unreasonable. After that, I'll try to segué as smoothly as possible from Fourier series to the Fourier transform, which is the tool people use to analyze non-periodic signals from the standpoint of frequency content. A comprehensive and truly rigorous treatment of Fourier transforms would require an entire book of its own, and excellent such books exist. We'll need to cut some corners and set some issues aside, and I'll try to be conscientious about mentioning when we do that.

Motivation, definition, and "derivation"

An example I enjoy pondering is a finite-duration A-440. Recall from Chapter 9 that a true A-440 is a periodic signal with fundamental period $T_o = 1/440$ and fundamental frequency $\Omega_o = 880\pi$. If I play an A-440 on an instrument and you listen to it, what you hear is not a true A-440 but a finite-duration signal that "sounds like an A-440" while I'm playing it. That signal, one would like to think, has significant "frequency content" around frequency 880π.

The Fourier transform enables us to make mathematical sense of frequency content for non-periodic signals like the finite-duration A-440. I'll try first to motivate the definition with a sketchy sort of pseudo-derivation. Suppose first that x is a decent finite-duration signal and $T_o > 0$ is such that $x(t) = 0$ when $|t| \geq T_o/2$. x is about as non-periodic as you can imagine, but we can extend x to form a decent periodic signal x_r that has T_o as a period by adding infinitely many shifted replicas of x. Technically,

$$x_r(t) = \sum_{n=-\infty}^{\infty} x(t - nT_o) \text{ for all } t \in \mathbb{R} .$$

For any specific t, at most one term in the infinite series is nonzero due to the finite-duration condition on x, so the sum converges trivially. Note that

$$x_r(t) = x(t) \text{ for } -T_o/2 \leq t \leq T_o/2 .$$

Now let $\Omega_o = 2\pi/T_o$ and expand x_r in a Fourier series

$$x_r(t) = \sum_{k=-\infty}^{\infty} c_k e^{jk\Omega_o t} \text{ for all } t \in \mathbb{R} .$$

For every $k \in \mathbb{Z}$,

$$c_k = \frac{1}{T_o}\int_{-T_o/2}^{T_o/2} x_r(t)e^{-jk\Omega_o t}dt = \frac{1}{T_o}\int_{-T_o/2}^{T_o/2} x(t)e^{-jk\Omega_o t}dt ,$$

where the last equality holds because $x_r = x$ on the interval $[-T_o/2, T_o/2]$. Accordingly, ignoring the issue of jumps in $x_r(t)$, we have

$$x_r(t) = \sum_{k=-\infty}^{\infty} \left(\frac{1}{T_o}\int_{-T_o/2}^{T_o/2} x(t)e^{-jk\Omega_o t}dt\right) e^{jk\Omega_o t} \text{ for all } t \in \mathbb{R}$$

and therefore

$$x(t) = \sum_{k=-\infty}^{\infty} \left(\frac{1}{T_o}\int_{-T_o/2}^{T_o/2} x(t)e^{-jk\Omega_o t}dt\right) e^{jk\Omega_o t} \text{ for } -T_o/2 \le t \le T_o/2 .$$

The final step is to let $T_o \to \infty$ in the last equation. The right-hand side becomes a Riemann approximation of an integral. To see how it works, first replace $1/T_o$ with $\Omega_o/2\pi$ to obtain

$$(8)\quad x(t) = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty} \left(\int_{-T_o/2}^{T_o/2} x(t)e^{-jk\Omega_o t}dt\right) e^{jk\Omega_o t} \times \Omega_o \text{ for } -T_o/2 \le t \le T_o/2 .$$

As $T_o \to \infty$, the sum becomes an integral, Ω_o becomes $d\Omega$, $k\Omega_o$ becomes Ω, and $T_o/2 \le t \le T_o/2$ becomes $t \in \mathbb{R}$. The end result is

$$(9)\qquad x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x(t)e^{-j\Omega t}dt\right) e^{j\Omega t}d\Omega \text{ for all } t \in \mathbb{R} .$$

The bogus part of the argument leading from (8) to (9) is the step where the condition $-T_o/2 \le t \le T_o/2$ becomes "for all $t \in \mathbb{R}$." The right-hand side of (8), although it coincides with $x(t)$ over ever-widening intervals as T_o increases, does not converge nicely to $x(t)$ for all $t \in \mathbb{R}$ as $T_o \to \infty$. Nonetheless, (9) provides a strong motivation for the following formal definition.

10.1 Definition: A complex-valued signal $t \mapsto x(t)$ and a complex-valued function $\Omega \mapsto \widehat{X}(\Omega)$ are a *Fourier transform pair* when one or both of the following equations holds:

$$(\mathcal{F})\qquad \widehat{X}(\Omega) = \int_{-\infty}^{\infty} x(t)e^{-j\Omega t}dt \text{ for all } \Omega \in \mathbb{R}$$

$$(\mathcal{F}^{-1})\qquad x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \widehat{X}(\Omega)e^{j\Omega t}d\Omega \text{ for all } t \in \mathbb{R}$$

In this case, we also say that $\widehat{X}$ is the *Fourier transform* of x, and write

$$x \xleftrightarrow{\mathcal{F}} \widehat{X} .$$

It's impossible at this level to deal rigorously with all the subtleties of the Fourier transform. Definition 10.1 is somewhat unconventional in that it talks about Fourier transform pairs instead of defining "Fourier-transformable signals" first, then proving that equation $\mathcal{F}^{-1}$ holds in certain special cases, etc. Definition 10.1 works better for the purposes of signal and system applications, at least in my view. I caution you that in what follows I'll often assume for ease of exposition in a specific context that one or the other of equations $\mathcal{F}$ or $\mathcal{F}^{-1}$ holds for a particular Fourier transform pair.

For convenience in applications, it's useful to extend Definition 10.1 in a couple of directions. First, consider the unit impulse δ. Plugging $x = \delta$ into $\mathcal{F}$ yields

$$\widehat{X}(\Omega) = \int_{-\infty}^{\infty} \delta(t) e^{-j\Omega t} dt = \lim_{a \to 0} \frac{1}{a} \int_{-a/2}^{a/2} e^{-j\Omega t} dt = \lim_{a \to 0} \frac{1}{(a/2)\Omega} \sin(a/2)\Omega = 1$$

for all $\Omega \in \mathbb{R}$. In deriving this formula, I used equation (6) from Chapter 8 along with $\sin\theta/\theta \to 1$ as $\theta \to 0$. Accordingly, by convention if you will, the unit impulse δ has as Fourier transform the constant function 1. Schematically,

$$\delta \overset{\mathcal{F}}{\longleftrightarrow} 1 \ .$$

It's painfully obvious that the integral in $\mathcal{F}^{-1}$ fails to converge in any ordinary sense for this example. Similarly, applying equation $\mathcal{F}^{-1}$ to the "function" $\Omega \mapsto 2\pi\delta(\Omega)$ gives

$$1 = \frac{1}{2\pi} \int_{-\infty}^{\infty} 2\pi\delta(\Omega) e^{j\Omega t} d\Omega$$

for every $t \in \mathbb{R}$. Thus arises another generalized Fourier transform pair

$$1 \overset{\mathcal{F}}{\longleftrightarrow} 2\pi\delta \ .$$

You can check that equation $\mathcal{F}$ fails for this example. We'll see shortly that these generalizations along with various operational rules for Fourier transforms will enable us to amass a variety of Fourier transform pairs involving impulses either in time or frequency for which only one of $\mathcal{F}$ or $\mathcal{F}^{-1}$ holds.

It proves beneficial to generalize Definition 10.1 even further by allowing Fourier transform pairs for which neither equation $\mathcal{F}$ nor $\mathcal{F}^{-1}$ holds. The examples I'm thinking of are along the lines of the signal $x = \delta + 1$. We have Fourier transforms for 1 and δ, but for each of them one of the two defining equations in Definition 10.1 fails. We would like it to be true that

$$1 + \delta \overset{\mathcal{F}}{\longleftrightarrow} 2\pi\delta + 1 \ ,$$

so we simply "make it so." This amounts to extending Definition 10.1 by linearity to a wider class of signals than the definition covers officially. It's clear, for example, that if $\mathcal{F}$ holds for two signals x_1 and x_2 and their respective Fourier transforms $\widehat{X}_1$ and $\widehat{X}_2$, then $\mathcal{F}$ also holds for $c_1x_1 + c_2x_2$ for any complex numbers c_1 and c_2. But what if only $\mathcal{F}$ holds for x_1 and only $\mathcal{F}^{-1}$ holds for x_2? The following, which states the first of many properties of Fourier transforms, essentially wishes that problem away.

10.2 Linearity: The Fourier transform is linear in the sense that if $x_1 \overset{\mathcal{F}}{\longleftrightarrow} \widehat{X}_1$ and $x_2 \overset{\mathcal{F}}{\longleftrightarrow} \widehat{X}_2$, then

$$c_1 x_1 + c_2 x_2 \overset{\mathcal{F}}{\longleftrightarrow} c_1 \widehat{X}_1 + c_2 \widehat{X}_2$$

for every c_1, $c_2 \in \mathbb{C}$. This property is clear when either $\mathcal{F}$ or $\mathcal{F}^{-1}$ applies to both x_1 and x_2, but we extend Definition 10.1 so that it holds in general. More precisely, we consider x and $\widehat{X}$ to be a Fourier-transform pair if we can decompose x into a sum of signals $\{x_k\}$ and $\widehat{X}$ into a sum of functions $\widehat{X}_k$ so that, for each k, $x_k \overset{\mathcal{F}}{\longleftrightarrow} \widehat{X}_k$ in the strict sense of Definition 10.1.

It's worth noting but not easy to prove that decent signals and their Fourier transforms are essentially in one-to-one correspondence in the sense that if $x \overset{\mathcal{F}}{\longleftrightarrow} \widehat{X}$, then x determines $\widehat{X}$ completely even if $\mathcal{F}$ fails and $\widehat{X}$ determines x completely even if $\mathcal{F}^{-1}$ fails. In this sense, a signal and its Fourier transform are simply alternative descriptions of the "same abstract object." It's like looking at one "thing" from two different angles. The operative lingo, as in Chapter 9, is that "$t \mapsto x(t)$ is a time-domain description of a signal, whereas $\Omega \mapsto \widehat{X}(\Omega)$ is a frequency-domain description of that same signal." Alternative terminology that I'll often use calls $\widehat{X}$ the *spectrum* of x or the *spectral description* of x.

We'll see presently, in the context of the finite-duration A-440 and other examples, how $\widehat{X}$ provides a useful representation of the frequency content of x. For now, compare what equation $\mathcal{F}^{-1}$ does for x to what a Fourier-series expansion does for a T_o-periodic signal. The Fourier series expresses the periodic signal as a superposition of pure sinusoids at a *discrete* set of frequencies $\{k\Omega_o : k \in \mathbb{Z}\}$. The weights in the superposition are the Fourier-series coefficients. Meanwhile, equation $\mathcal{F}^{-1}$ expresses x as a *continuum superposition* of pure sinusoids. The weights in this continuum superposition are the values of $\widehat{X}(\Omega)$ as Ω ranges over frequency space. People say for these reasons that a periodic signal has a *discrete spectrum* or *pure-point spectrum* whereas a non-periodic x with a continuous Fourier transform $\widehat{X}$ has a *continuous spectrum.*

It is difficult to characterize precisely the set of signals that possess Fourier transforms in the sense of Definition 10.1. The following sufficient conditions will have to do for now. Although they don't tell the whole story by any means, they're useful things to know. Some are easier to prove than others, but I won't prove any of them.

10.3 Existence criteria for Fourier transforms:

- If x is a decent finite-duration signal, then $\widehat{X}$ exists and equation $\mathcal{F}$ holds. Furthermore, $\widehat{X}$ is a bounded function of Ω. Similarly, if $\Omega \mapsto \widehat{X}(\Omega)$ is a decent function of Ω that satisfies $\widehat{X}(\Omega) = 0$ when $|\Omega| \geq \Omega_m$ for some $\Omega_m > 0$, then there exists a bounded signal x for which $x \overset{\mathcal{F}}{\longleftrightarrow} \widehat{X}$, and equation $\mathcal{F}^{-1}$ holds.

- Suppose x is an infinitely differentiable signal that decreases rapidly as $|t| \to \infty$ in the sense that
$$\lim_{|t|\to\infty} t^m D^n x(t) = 0$$
for every m and n in $\mathbb{N}$, where $D^n x$ is the nth derivative of x. Then $\widehat{X}$ exists and both equations $\mathcal{F}$ and $\mathcal{F}^{-1}$ hold. In addition, $\Omega \mapsto \widehat{X}(\Omega)$ is an infinitely differentiable function of Ω that decreases rapidly as $|\Omega| \to \infty$ in the sense that
$$\lim_{|\Omega|\to\infty} \Omega^m D^n \widehat{X}(\Omega) = 0$$
for every m and n in $\mathbb{N}$, where $D^n \widehat{X}$ is the nth derivative of $\widehat{X}$ with respect to Ω.
- If x is an absolutely integrable signal, then $\widehat{X}$ exists and equation $\mathcal{F}$ holds. Furthermore, $\widehat{X}$ is a bounded function of Ω; in fact,
$$|\hat{X}(\Omega)| \le \|x\|_1 = \int_{-\infty}^{\infty} |x(t)|dt \;\text{ for all }\; \Omega \in \mathbb{R}\,.$$
- If x is a square-integrable signal, then $\widehat{X}$ exists and is a square-integrable function of Ω. Both equations $\mathcal{F}$ and $\mathcal{F}^{-1}$ hold in a "mean-square sense," which means
$$\lim_{T\to\infty} \int_{-\infty}^{\infty} \left| \widehat{X}(\Omega) - \int_{-T}^{T} x(t)e^{-j\Omega t}dt \right|^2 d\Omega = 0$$
and
$$\lim_{\overline{\Omega}\to\infty} \int_{-\infty}^{\infty} \left| x(t) - \frac{1}{2\pi}\int_{-\overline{\Omega}}^{\overline{\Omega}} \widehat{X}(\Omega)e^{j\Omega t}d\Omega \right|^2 dt = 0\,.$$

When x is square-integrable, we have a result analogous to Parseval's Theorem for Fourier series.

10.4 Plancherel's Theorem: If x and y are square-integrable signals with respective Fourier transforms $\widehat{X}$ and $\widehat{Y}$, then
$$\int_{-\infty}^{\infty} x(t)\overline{y(t)}dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} \widehat{X}(\Omega)\overline{\widehat{Y}(\Omega)}d\Omega$$
and, in particular,
$$\int_{-\infty}^{\infty} |x(t)|^2 dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} |\widehat{X}(\Omega)|^2 d\Omega\,.$$

Operational rules and prototype examples

To develop facility with manipulating expressions involving signals and their Fourier transforms, it's important to learn various operational rules that simplify the manipulations. In "proving" each of these rules, I'll be casual about whether equation $\mathcal{F}$ or $\mathcal{F}^{-1}$ holds. In particular, I'll assume in each case the validity of whichever equation makes the proof easier. Rest assured that you can start each proof with the other equation if you wish.

10.5 Time-shift rule: If $x \xleftrightarrow{\mathcal{F}} \widehat{X}$ then, for any $t_o \in \mathbb{R}$, $y = \mathrm{Shift}_{t_o}(x)$ has Fourier transform $\widehat{Y}$ specified by

$$\widehat{Y}(\Omega) = e^{-j\Omega t_o}\widehat{X}(\Omega) \text{ for all } \Omega \in \mathbb{R} .$$

To see this, assume $\mathcal{F}^{-1}$ holds. Then

$$y(t) = x(t-t_o) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \widehat{X}(\Omega)e^{j\Omega(t-t_o)}d\Omega = \frac{1}{2\pi}\int_{-\infty}^{\infty} \left(e^{-j\Omega t_o}\widehat{X}(\Omega)\right)e^{j\Omega t}d\Omega ,$$

and this is just equation $\mathcal{F}^{-1}$ for y, revealing that $\widehat{Y}(\Omega) = e^{-j\Omega t_o}\widehat{X}(\Omega)$ for all $\Omega \in \mathbb{R}$. □

10.6 Frequency-shift rule: If $x \xleftrightarrow{\mathcal{F}} \widehat{X}$ then, for any $\Omega_o \in \mathbb{R}$, the signal y with specification $y(t) = e^{j\Omega_o t}x(t)$ for all $t \in \mathbb{R}$ has Fourier transform $\widehat{Y}$ specified by

$$\widehat{Y}(\Omega) = \widehat{X}(\Omega - \Omega_o) \text{ for all } \Omega \in \mathbb{R} .$$

To demonstrate this, start with equation $\mathcal{F}$ for x and plug in $\Omega - \Omega_o$ for Ω. You get

$$\widehat{X}(\Omega-\Omega_o) = \int_{-\infty}^{\infty} x(t)e^{-j(\Omega-\Omega_o)t}dt = \int_{-\infty}^{\infty} \left(e^{j\Omega_o t}x(t)\right)e^{-j\Omega t}dt = \int_{-\infty}^{\infty} y(t)e^{-j\Omega t}dt ,$$

and this is equation $\mathcal{F}$ for y, revealing that $\widehat{Y}(\Omega) = \widehat{X}(\Omega - \Omega_o)$ for all $\Omega \in \mathbb{R}$. □

10.7 Time-derivative rule: If $x \xleftrightarrow{\mathcal{F}} \widehat{X}$ and x is differentiable, then the signal $y = Dx$, where D denotes time derivative, has Fourier transform $\widehat{Y}$ with specification

$$\widehat{Y}(\Omega) = j\Omega\widehat{X}(\Omega) \text{ for all } \Omega \in \mathbb{R} .$$

Start the argument with equation $\mathcal{F}^{-1}$ for x and you get

$$y(t) = Dx(t) = D\left(\frac{1}{2\pi}\int_{-\infty}^{\infty} \widehat{X}(\Omega)e^{j\Omega t}d\Omega\right) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \left(j\Omega\widehat{X}(\Omega)\right)e^{j\Omega t}d\Omega ,$$

and this is equation $\mathcal{F}^{-1}$ for y, revealing that $\widehat{Y}(\Omega) = j\Omega\widehat{X}(\Omega)$ for all $\Omega \in \mathbb{R}$. □

10.8 Frequency-derivative rule: If $x \xleftrightarrow{\mathcal{F}} \widehat{X}$ and $\widehat{X}$ is a differentiable function of Ω, then $\widehat{Y} = D\widehat{X}$, where D denotes derivative with respect to Ω, is the Fourier transform of the signal y with specification

$$y(t) = -jtx(t) \text{ for all } t \in \mathbb{R} .$$

Start the argument with equation $\mathcal{F}$ for x and you get

$$\widehat{Y}(\Omega) = D\widehat{X}(\Omega) = D\left(\int_{-\infty}^{\infty} x(t)e^{-j\Omega t}dt\right) = \int_{-\infty}^{\infty} (-jtx(t))\, e^{-j\Omega t}d\Omega \ ,$$

and this is just equation $\mathcal{F}$ for $y \overset{\mathcal{F}}{\longleftrightarrow} \widehat{Y}$, where $\widehat{Y} = D\widehat{X}$ and $y(t) = -jtx(t)$ for every $t \in \mathbb{R}$. □

10.9 Scaling rule: If $x \overset{\mathcal{F}}{\longleftrightarrow} \widehat{X}$ and $a > 0$, then the signal y with specification $y(t) = x(at)$ for all $t \in \mathbb{R}$ has Fourier transform $\widehat{Y}$ with specification

$$\widehat{Y}(\Omega) = \frac{1}{a}\widehat{X}\left(\frac{\Omega}{a}\right) \quad \text{for all } \ \Omega \in \mathbb{R} \ .$$

Start with equation $\mathcal{F}$ for y. You find that

$$\begin{aligned} \widehat{Y}(\Omega) &= \int_{-\infty}^{\infty} x(at)e^{-j\Omega t}dt \\ &= \int_{-\infty}^{\infty} x(\tau)e^{-j\Omega(\tau/a)}(1/a)d\tau \\ &= \frac{1}{a}\left(\int_{-\infty}^{\infty} x(\tau)e^{-j(\Omega/a)\tau}d\tau\right) \end{aligned}$$

for all Ω, so $\widehat{Y}(\Omega) = \frac{1}{a}\widehat{X}\left(\frac{\Omega}{a}\right)$ for all $\Omega \in \mathbb{R}$. □

10.10 Convolution rule: If $x_1 \overset{\mathcal{F}}{\longleftrightarrow} \widehat{X}_1$ and $x_2 \overset{\mathcal{F}}{\longleftrightarrow} \widehat{X}_2$ and $x = x_1 * x_2$ exists, then $x \overset{\mathcal{F}}{\longleftrightarrow} \widehat{X}_1\widehat{X}_2$. In other words, the Fourier transform takes convolution in the time domain to simple function multiplication in the frequency domain.

In "proving" this one, I'll assume first that both x_1 and x_2 are absolutely integrable, which implies that we can interchange orders of integration with impunity in the equations below. Start with equation $\mathcal{F}$ for x.

$$\begin{aligned} \widehat{X}(\Omega) &= \int_{-\infty}^{\infty} x(t)e^{-j\Omega t}dt \\ &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} x_1(\tau)x_2(t-\tau)d\tau\right)e^{-j\Omega t}dt \\ &= \int_{-\infty}^{\infty} x_1(\tau)\left(\int_{-\infty}^{\infty} x_2(t-\tau)e^{-j\Omega t}dt\right)d\tau \\ &= \int_{-\infty}^{\infty} x_1(\tau)\left(e^{-j\Omega\tau}\widehat{X}_2(\Omega)\right)d\tau \\ &= \left(\int_{-\infty}^{\infty} x_1(\tau)e^{-j\Omega\tau}d\tau\right)\widehat{X}_2(\Omega) \\ &= \widehat{X}_1(\Omega)\widehat{X}_2(\Omega) \ \text{ for all } \ \Omega \in \mathbb{R} \ . \end{aligned}$$

The crucial step on the fourth line follows from the time-shift rule applied to the inner integral.

When both x_1 and x_2 are square-integrable, an entirely different argument applies. First fix $t \in \mathbb{R}$ and let y be the signal with specification $y(\tau) = \overline{x_2(t-\tau)}$

for all $\tau \in \mathbb{R}$. Observe that

$$\begin{aligned}\langle x_1, y\rangle &= \int_{-\infty}^{\infty} x_1(\tau)\overline{y(\tau)}d\tau \\ &= \int_{-\infty}^{\infty} x_1(\tau)x_2(t-\tau)d\tau \\ &= x_1 * x_2(t)\ .\end{aligned}$$

Furthermore, y has Fourier transform $\widehat{Y}$ with specification

$$\begin{aligned}\widehat{Y}(\Omega) &= \int_{-\infty}^{\infty} y(\tau)e^{-j\Omega\tau}d\tau \\ &= \int_{-\infty}^{\infty} \overline{x_2(t-\tau)}e^{-j\Omega\tau}d\tau \\ &= \int_{-\infty}^{\infty} \overline{x_2(\zeta)}e^{-j\Omega(t-\zeta)}d\zeta \\ &= \overline{\widehat{X}_2(\Omega)}e^{-j\Omega t}\end{aligned}$$

for all $\Omega \in \mathbb{R}$. By Plancherel's Theorem 10.4,

$$\langle x, y\rangle = \frac{1}{2\pi}\int_{-\infty}^{\infty} \widehat{X}(\Omega)\overline{\widehat{Y}(\Omega)}d\Omega\ ,$$

so

$$x_1 * x_2(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \widehat{X}_1(\Omega)\widehat{X}_2(\Omega)e^{j\Omega t}d\Omega \ \text{ for all } \ t \in \mathbb{R}\ ,$$

which is just equation $\mathcal{F}^{-1}$ for $x_1 * x_2$. □

10.11 Modulation rule: If $x_1 \overset{\mathcal{F}}{\longleftrightarrow} \widehat{X}_1$ and $x_2 \overset{\mathcal{F}}{\longleftrightarrow} \widehat{X}_2$ and $\widehat{X} = \widehat{X}_1 * \widehat{X}_2$ exists, then the signal x such that $x \overset{\mathcal{F}}{\longleftrightarrow} \widehat{X}$ is

$$x = 2\pi x_1 x_2\ .$$

In other words, the Fourier transform (almost) takes multiplication in the time domain to convolution in the frequency domain.

In proving this one, I'll assume first that both $\widehat{X}_1$ and $\widehat{X}_2$ are absolutely integrable functions of Ω, which implies that we can interchange orders of integration with impunity in the equations below. Start with equation $\mathcal{F}^{-1}$ for x.

$$\begin{aligned}x(t) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \widehat{X}(\Omega)e^{j\Omega t}d\Omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} \widehat{X}_1(\zeta)\widehat{X}_2(\Omega-\zeta)d\zeta\right)e^{j\Omega t}d\Omega \\ &= \int_{-\infty}^{\infty} \widehat{X}_1(\zeta)\left(\frac{1}{2\pi}\int_{-\infty}^{\infty} \widehat{X}_2(\Omega-\zeta)e^{j\Omega t}d\Omega\right)d\zeta \\ &= \int_{-\infty}^{\infty} \widehat{X}_1(\zeta)\left(e^{j\zeta t}x_2(t)\right)d\zeta \\ &= \left(\int_{-\infty}^{\infty} \widehat{X}_1(\zeta)e^{j\zeta t}d\zeta\right)x_2(t) \\ &= 2\pi x_1(t)x_2(t) \ \text{ for all } \ t \in \mathbb{R}\ .\end{aligned}$$

The step on the fourth line follows from the Frequency-shift rule applied to the inner integral. When x_1 and x_2 are square-integrable, an argument based on Plancherel's Theorem similar to the one in the proof of 10.10 applies. □

10.12 Symmetry properties: Suppose $x \overset{\mathcal{F}}{\longleftrightarrow} \widehat{X}$.

- If x is real-valued, then $\widehat{X}(-\Omega) = \overline{\widehat{X}(\Omega)}$ for all $\Omega \in \mathbb{R}$.
- If x is an even function of t, i.e., if $x(-t) = x(t)$ for all $t \in \mathbb{R}$, then $\widehat{X}$ is an even function of Ω, i.e. $\widehat{X}(-\Omega) = \widehat{X}(\Omega)$ for all $\Omega \in \mathbb{R}$.
- If x is an odd function of t, i.e., if $x(-t) = -x(t)$ for all $t \in \mathbb{R}$, then $\widehat{X}$ is an odd function of Ω, i.e. $\widehat{X}(-\Omega) = -\widehat{X}(\Omega)$ for all $\Omega \in \mathbb{R}$.
- If x is real-valued and even, then $\widehat{X}$ is also real-valued and even.
- If x is real-valued and odd, then $\widehat{X}$ is pure imaginary-valued and odd.

I'll prove the first of these explicitly starting from $\mathcal{F}$.

$$\widehat{X}(-\Omega) = \int_{-\infty}^{\infty} x(t)e^{-j(-\Omega)t}dt = \overline{\int_{-\infty}^{\infty} \overline{x}(t)e^{-j\Omega t}dt} = \overline{\int_{-\infty}^{\infty} x(t)e^{-j\Omega t}dt} = \overline{\widehat{X}(\Omega)} \ .$$

The last two follow from the first three. The second and third are simple exercises in changing variables in integrals. □

The operational rules enable us to construct a variety of handy prototype examples of Fourier transform pairs. First recall that $1 \overset{\mathcal{F}}{\longleftrightarrow} 2\pi\delta$. Applying the Frequency-shift rule 10.6, we find that for any $\Omega_o \in \mathbb{R}$

$$e^{j\Omega_o t} \quad \overset{\mathcal{F}}{\longleftrightarrow} \quad 2\pi\delta(\Omega - \Omega_o) \ .$$

Here and in what follows I'm taking some notational liberties. Formally, what I mean is that the signal x with specification $x(t) = e^{j\Omega_o t}$ for all $t \in \mathbb{R}$ has Fourier transform $\widehat{X}$ with specification $\widehat{X}(\Omega) = 2\pi\delta(\Omega)$. Invoking Euler's Formulas leads to

$$\cos \Omega_o t \quad \overset{\mathcal{F}}{\longleftrightarrow} \quad \pi\delta(\Omega - \Omega_o) + \pi\delta(\Omega + \Omega_o)$$

and

$$\sin \Omega_o t \quad \overset{\mathcal{F}}{\longleftrightarrow} \quad \frac{\pi}{j}\delta(\Omega - \Omega_o) - \frac{\pi}{j}\delta(\Omega + \Omega_o) \ .$$

If x is a periodic signal with a Fourier series

$$x(t) = \sum_{k=-\infty}^{\infty} c_k e^{jk\Omega_o t} \ \text{ for all } \ t \in \mathbb{R} \ ,$$

then the foregoing examples enable us to think of x as having a Fourier transform $\widehat{X}$ with specification

$$\widehat{X}(\Omega) = \sum_{k=-\infty}^{\infty} 2\pi c_k \delta(\Omega - k\Omega_o) \ .$$

This $\widehat{X}$ is an *impulse train* in the frequency domain. It illuminates graphically the intuitive assertion that the frequency content of x is concentrated entirely on the discrete set of frequencies $\{k\Omega_o : k \in \mathbb{Z}\}$.

The Fourier transform pair

$$e^{-\alpha t}u(t) \quad \overset{\mathcal{F}}{\longleftrightarrow} \quad \frac{1}{\alpha + j\Omega}\ ,$$

which holds for $\alpha > 0$, spawns a whole sequence of prototype examples. First, let's make sure we believe that this one is indeed a Fourier transform pair. Let $x(t) = e^{-\alpha t}u(t)$ for $t \in \mathbb{R}$. Apply equation $\mathcal{F}$ to find $\widehat{X}$.

$$\widehat{X}(\Omega) = \int_{-\infty}^{\infty} x(t)e^{-j\Omega t}dt = \int_{0}^{\infty} e^{-(\alpha+j\Omega)t}dt = \frac{-1}{\alpha + j\Omega}e^{-(\alpha+j\Omega)t}\bigg|_{0}^{\infty} = \frac{1}{\alpha + j\Omega}\ .$$

The upper limit in the second-to-last expression evaluates to zero because $\alpha > 0$ and $|e^{-j\Omega t}| = 1$ for all t. Apply the frequency-derivative rule to this example and you get

$$-jtx(t) \quad \overset{\mathcal{F}}{\longleftrightarrow} \quad \frac{-j}{(\alpha + j\Omega)^2}\ ,$$

which because of linearity is the same as

$$tx(t) \quad \overset{\mathcal{F}}{\longleftrightarrow} \quad \frac{1}{(\alpha + j\Omega)^2}\ .$$

Apply the frequency-derivative rule again.

$$-jt^2x(t) \quad \overset{\mathcal{F}}{\longleftrightarrow} \quad \frac{-2j}{(\alpha + j\Omega)^3}\ ,$$

which is the same as

$$\frac{t^2}{2}x(t) \quad \overset{\mathcal{F}}{\longleftrightarrow} \quad \frac{1}{(\alpha + j\Omega)^3}\ .$$

Repeated applications of the frequency-derivative rule lead to the following list of prototype examples: for every integer $m > 0$,

$$\frac{t^m}{m!}e^{-\alpha t}u(t) \quad \overset{\mathcal{F}}{\longleftrightarrow} \quad \frac{1}{(\alpha + j\Omega)^{m+1}}\ .$$

Rectangular pulses star in the following two examples, both of which play central roles in applications of Fourier transforms to signal and system analysis.

10.13 Example: $x = p_a$ for some $a > 0$. Since x is decent and has finite duration, it has a Fourier transform $\widehat{X}$ that we can figure out from equation $\mathcal{F}$.

$$\begin{aligned}
\widehat{X}(\Omega) &= \int_{-\infty}^{\infty} x(t)e^{-j\Omega t}dt \\
&= \int_{-a/2}^{a/2} e^{-j\Omega t}dt \\
&= \frac{e^{j(a\Omega/2)} - e^{-j(a\Omega/2)}}{j\Omega} \\
&= \frac{2\sin(a\Omega/2)}{\Omega} \quad \text{for all } \ \Omega \in \mathbb{R}\ .
\end{aligned}$$

The $\widehat{X}$ of this example takes the form of what people call a *sinc function.* Despite the Ω in the denominator, $\widehat{X}(0)$ is finite. Using $\sin\theta/\theta \to 1$ as $\theta \to 0$, you find that

$$\widehat{X}(0) = \lim_{\Omega\to 0} \frac{a\sin(a\Omega/2)}{(a\Omega/2)} = a \ .$$

$\widehat{X}$ has zero-crossings where $a\Omega/2$ is an integer multiple of π, i.e. at all Ω-values of the form $2m\pi/a$ for $m \in \mathbb{Z}$. The peaks in $\widehat{X}$ decay toward zero as $|\Omega| \to \infty$ like $1/|\Omega|$. See Figure 1.

10.14 Example: $\widehat{X}$ is the function $\Omega \mapsto p_{\Omega_o}(\Omega)$ for some $\Omega_o > 0$. In other words, $\widehat{X}$ is a rectangular pulse in the frequency domain. We can use equation $\mathcal{F}^{-1}$ to figure out the signal x that has $\widehat{X}$ as Fourier transform.

$$\begin{aligned} x(t) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \widehat{X}(\Omega)e^{j\Omega t}d\Omega \\ &= \frac{1}{2\pi}\int_{-\Omega_o/2}^{\Omega_o/2} e^{j\Omega t}d\Omega \\ &= \frac{e^{j(\Omega_o t/2)} - e^{-j(\Omega_o t/2)}}{j2\pi t} \\ &= \frac{\sin(\Omega_o t/2)}{\pi t} \quad \text{for all } \ t \in \mathbb{R} \ . \end{aligned}$$

The x of this example is a sinc function in the time domain. Again using $\sin\theta/\theta \to 1$ as $\theta \to 0$, you find that

$$x(0) = \lim_{t\to 0} \frac{\Omega_o}{2\pi}\frac{\sin(\Omega_o t/2)}{(\Omega_o t/2)} = \frac{\Omega_o}{2\pi} \ .$$

x has zero-crossings where $\Omega_o t/2t$ is an integer multiple of π, i.e. at all t-values of the form $2m\pi/\Omega_o$ for $m \in \mathbb{Z}$. The peaks in x decay toward zero as $|t| \to \infty$ like $1/|t|$. See Figure 2.

Example 10.13 helps us make sense of the finite-duration A-440. A reasonable mathematical model for a particularly simple finite-duration A-440 is the signal y with specification

$$y(t) = p_a(t)\cos(880\pi t) = \begin{cases} \cos(880\pi t) & -a/2 \le t < a/2 \\ 0 & \text{otherwise.} \end{cases}$$

A more general model would replace the cosine with an arbitrary periodic signal x with fundamental frequency 880π. Euler's formula implies that

$$y(t) = (p_a(t)/2)e^{j880\pi t} + (p_a(t)/2)e^{-j880\pi t} \ .$$

By the Frequency-shift rule 10.6 along with Example 10.13,

$$\widehat{Y}(\Omega) = \frac{\sin\frac{a}{2}(\Omega - 880\pi)}{\Omega - 880\pi} + \frac{\sin\frac{a}{2}(\Omega + 880\pi)}{\Omega + 880\pi} \ ,$$

so $\widehat{Y}$ is the sum of two sinc functions, one centered on $\Omega = 880\pi$ and the other on $\Omega = -880\pi$. Looking at Figure 3 might convince you that it makes sense to say that y has significant frequency content in the neighborhood of frequency $\Omega = 880\pi$, which is 440 Hz. That property of y's frequency content supports nicely our original intuition about the finite-duration A-440.

Think now about what happens as we extend the duration of the finite-duration A-440, which is the same as letting a get larger. The peaks of the two sinc functions that constitute $\widehat{Y}$ are both $a/2$. The zero-crossings in each sinc function occur at frequencies of the form $\pm 880\pi + 2m\pi/a$ for $m \in \mathbb{Z}$. Adding the two sinc functions causes the peaks and zero crossings to change slightly, but it's fairly clear what happens when a gets large. The central peaks at $\Omega = \pm 880\pi$ get taller and the zero crossings cluster around $\Omega = \pm 880\pi$, The picture of $\widehat{Y}$ starts looking more and more like a pair of impulses located at $\Omega = \pm 880\pi$, which is not surprising because as a increases y looks more and more like a true 440 Hz cosine whose Fourier transform has specification $\pi\delta(\Omega - 880\pi) + \pi\delta(\Omega + 880\pi)$.

Heisenberg's Inequality and bandlimited signals

The examples $\delta \overset{\mathcal{F}}{\longleftrightarrow} 1$ and $1 \overset{\mathcal{F}}{\longleftrightarrow} 2\pi\delta$ are extreme instances of a general rule, namely: if a signal is sharply focused in the time domain, its Fourier transform tends to be spread out, whereas a signal with Fourier transform sharply focused in Ω-space tends to be spread out in time. The Scaling rule 10.9 gives this qualitative property some quantitative teeth. From that rule it follows that for any $a > 0$ we have

$$t \mapsto \sqrt{a}x(at) \quad \overset{\mathcal{F}}{\longleftrightarrow} \quad \Omega \mapsto \frac{1}{\sqrt{a}}\widehat{X}(\frac{\Omega}{a}) \ .$$

Consider starting with a nominal x. For large a, $t \mapsto \sqrt{a}x(at)$ is a taller and more sharply focused version of x. Correspondingly, $\Omega \mapsto (1/\sqrt{a})\widehat{X}(\Omega/a)$ is a squashed-down spread-out version of $\widehat{X}$. For small a, $t \mapsto \sqrt{a}x(at)$ is a more spread-out version of x, and $\Omega \mapsto (1/\sqrt{a})\widehat{X}(\Omega/a)$ is a taller more sharply focused version of $\widehat{X}$. So increasing a causes the time-domain picture to sharpen its focus and the frequency-domain picture to flatten and spread out, and vice versa for decreasing a.

As it happens, the scaling rule is related to the inequality that constitutes Heisenberg's Uncertainty Principle in quantum mechanics. Roughly speaking, the position and momentum wave functions for a particle are a Fourier-transform pair. Accurate knowledge of position corresponds to a sharply focused position wave function and hence a spread-out momentum wave function, which corresponds in turn to inaccurate knowledge of momentum — and vice versa. An example of a pertinent quantitative result is the following.

10.15 Heisenberg's Inequality: Suppose x is a square-integrable signal and $x \overset{\mathcal{F}}{\longleftrightarrow} \widehat{X}$. Then

$$\left(\int_{-\infty}^{\infty} t^2 |x(t)|^2 dt\right)^{1/2} \left(\int_{-\infty}^{\infty} \Omega^2 |\widehat{X}(\Omega)|^2 d\Omega\right)^{1/2} \geq \sqrt{\frac{\pi}{2}} \|x\|_2^2 \ ,$$

and equality holds only if x is a Gaussian signal, i.e. $x(t) = C_o e^{-\alpha t^2}$ for all $t \in \mathbb{R}$ for some $C_o \in \mathbb{C}$ and $\alpha > 0$.

I'll omit the proof of Heisenberg's Inequality, which rests on the Schwarz Inequality 9.7 and is easiest when you assume that x is infinitely differentiable and rapidly decreasing in the sense that $t^m D^n x(t) \to 0$ as $|t| \to \infty$ for all nonnegative integers m and n, where $D^n x$ denotes the nth derivative of x. To understand the intuition behind Heisenberg's Inequality, think of $f = |x|^2/\|x\|_2^2$ as a probability distribution on t-space and $g = |\widehat{X}|^2/\|\widehat{X}\|_2^2$ as a probability distribution on Ω-space. Suppose for convenience that the means of the probability distributions f and g lie at $t = 0$ and $\Omega = 0$ respectively. Then the standard deviations of the f and g distributions are

$$\sigma_f = \frac{1}{\|x\|_2}\left(\int_{-\infty}^{\infty} t^2|x(t)|^2 dt\right)^{1/2}$$

and

$$\sigma_g = \frac{1}{\|\widehat{X}\|_2}\left(\int_{-\infty}^{\infty} \Omega^2|\widehat{X}(\Omega)|^2 d\Omega\right)^{1/2} .$$

Heisenberg's Inequality states that

$$\sigma_f\sigma_g \geq \sqrt{\frac{\pi}{2}}\frac{\|x\|_2^2}{\|x\|_2\|\widehat{X}\|_2} = \sqrt{\frac{\pi}{2}}\frac{1}{\sqrt{2\pi}} = \frac{1}{2} .$$

I invoked Plancherel's Theorem 10.4 here, which implies that $\|x\|_2/\|\widehat{X}\|_2 = 1/\sqrt{2\pi}$. The standard deviations measure the spreads of the distributions about their means, and Heisenberg's Inequality puts a lower bound on the product of those standard deviations. The smaller one is, the larger the other has to be.

One could argue that any finite-duration signal is, on some level, "sharply focused in the time domain." The frequency-domain analogue of "finite duration" will feature prominently in our discussion of sampling and reconstruction later on.

10.16 Definition: Suppose $x \overset{\mathcal{F}}{\longleftrightarrow} \widehat{X}$. We say that x is a *bandlimited* signal, or that x *has finite bandwidth,* when there exists some $\Omega_m > 0$ such that $\widehat{X}(\Omega) = 0$ for $|\Omega| \geq \Omega_m$. In this case, we also say that x is *bandlimited to within* Ω_m. The *bandwidth* of a bandlimited signal x is

$$\Omega_m^* = \inf\Big(\{\Omega_m > 0 : \widehat{X}(\Omega) = 0 \text{ when } |\Omega| > \Omega_m\}\Big) .$$

Many, if not most, signals of practical interest are bandlimited. Any signal audible to humans has frequency content confined entirely between 0 and 20,000 Hz, so in the terminology of Definition 10.16 it is bandlimited to within $\Omega_m = 40,000\pi$. Other instances of band_speak include *narrowband* and *broadband*, both of which have obvious meanings as modifiers of bandlimited signals. If you think of a bandlimited signal as being "sharply focused in the frequency domain," the following result won't surprise you.

10.17 Theorem: Suppose $x \overset{\mathcal{F}}{\longleftrightarrow} \widehat{X}$ and $x \neq 0$. If x has finite duration, then x is not bandlimited. If x is bandlimited, then x does not have finite duration.

Proof: Suppose x has finite duration and let $T > 0$ be such that $x(t) = 0$ for $|t| \geq T$. Suppose x is also bandlimited to within $\Omega_m > 0$. I'll show that $x = 0$. Assuming $\widehat{X}$ is a reasonable function of Ω, equation $\mathcal{F}^{-1}$ holds and we can write

$$x(t) = \frac{1}{2\pi} \int_{-\Omega_m}^{\Omega_m} \widehat{X}(\Omega) e^{j\Omega t} d\Omega \ \text{ for all } \ t \in \mathbb{R} \ .$$

Since $x(t)$ is identically zero for all $|t| \geq T$, $D^n x(t)$ is identically zero for all $|t| > T$, where $D^n x(t)$ is the nth derivative of x evaluated at time t. (Note: I'm not assuming that x is a differentiable signal.) Pick some $t_1 > T$ and evaluate the nth time-derivative of the equation above at time t_1 and you get

$$\int_{-\Omega_m}^{\Omega_m} \widehat{X}(\Omega)(j\Omega)^n e^{j\Omega t_1} d\Omega = 0 \ \text{ for all } \ n \in \mathbb{N} \ .$$

Finally, observe that

$$\begin{aligned} x(t) &= \int_{-\Omega_m}^{\Omega_m} \widehat{X}(\Omega) e^{j\Omega(t-t_1)} e^{j\Omega t_1} d\Omega \\ &= \sum_{n=0}^{\infty} \frac{(t-t_1)^n}{n!} \int_{-\Omega_m}^{\Omega_m} \widehat{X}(\Omega)(j\Omega)^n e^{j\Omega t_1} d\Omega = 0 \ \text{ for all } \ t \in \mathbb{R} \ . \end{aligned}$$

I obtained the second line by expanding one of the exponentials under the integral sign. Conclude that if x has finite duration and is also bandlimited, then $x = 0$. □

Frequency response, filters, and amplitude modulation

In case you haven't noticed already, we actually met our first example of a Fourier transform in Chapter 9 while discussing systems-analysis reasons for studying Fourier series. If a LTI system has a frequency response in the sense of Definition 9.14, then the system's impulse response and frequency response are a Fourier-transform pair, and equation $\mathcal{F}$ gives the frequency response in terms of the impulse response. So if a system has a frequency response $\widehat{H}$, you can figure out $\widehat{H}$ in at least two ways:

- First find the output y of the system when the input is $t \mapsto e^{j\Omega t}$. You'll find that $y(t) = \widehat{H}(\Omega)e^{j\Omega t}$ for all $t \in \mathbb{R}$. Do this for every $\Omega \in \mathbb{R}$ and you get $\widehat{H}$.
- First find the impulse response h of the system, then find $\widehat{H}$ by taking the Fourier transform of h using equation $\mathcal{F}$.

We encountered the frequency response when analyzing how LTI systems respond to pure sinusoids and more general periodic signals. The convolution rule for Fourier transforms enables us to analyze and interpret a system's responses to more general inputs in terms of the system's frequency response. Suppose a LTI system has system mapping S and impulse response h and possesses a frequency response

$\widehat{H}$. Let x be an input to the system and suppose $x \overset{\mathcal{F}}{\longleftrightarrow} \widehat{X}$. Since $h \overset{\mathcal{F}}{\longleftrightarrow} \widehat{H}$ and $S(x) = h * x$, the Convolution rule 10.10 implies that

$$S(x) \overset{\mathcal{F}}{\longleftrightarrow} \widehat{H}\widehat{X} .$$

In this way, the frequency response re-shapes the frequency content of an input x, as embodied in $\widehat{X}$, into the frequency content of the corresponding output, as embodied in $\widehat{S(x)} = \widehat{H}\widehat{X}$. Thus every system with a frequency response acts as a frequency-selective filter of sorts. If $\widehat{H}(\Omega)$ is large for Ω near Ω_1, then the system will boost the part of the input spectrum near frequency Ω_1. If $\widehat{H}(\Omega)$ is small for Ω near Ω_1, then the system will attenuate the part of the input spectrum near frequency Ω_1.

In signals and systems applications, one often encounters discussions featuring ideal filters of various kinds. While none of these filters is physically realizable, a lot of work in signal processing focuses on approximating their behavior with physically realizable LTI systems. The three ideal filters are

- The ideal low-pass filter: this is the LTI system with frequency response
$$\widehat{H}(\Omega) = \begin{cases} 1 & |\Omega| \le \Omega_2 \\ 0 & \text{otherwise,} \end{cases}$$
where $\Omega_2 > 0$ is some given frequency.
- The ideal high-pass filter: this is the LTI system with frequency response
$$\widehat{H}(\Omega) = \begin{cases} 1 & |\Omega| \ge \Omega_1 \\ 0 & \text{otherwise,} \end{cases}$$
where $\Omega_1 > 0$ is some given frequency.
- The ideal bandpass filter: this is the LTI system with frequency response
$$\widehat{H}(\Omega) = \begin{cases} 1 & \Omega_1 \le |\Omega| \le \Omega_2 \\ 0 & \text{otherwise,} \end{cases}$$
where $\Omega_1 > 0$ and $\Omega_2 > 0$ are given frequencies with $\Omega_1 < \Omega_2$.

The frequencies Ω_1 and Ω_2 in these examples are called *cutoff frequencies* for the filters. Each filter has a *passband* and a *stopband*. The passband has the following characterization: Ω_o is in the passband when the input signal $t \mapsto e^{j\Omega_o t}$ passes through the filter unchanged. The stopband has the following characterization: Ω_o is in the stopband when the filter annihilates the input signal $t \mapsto e^{j\Omega_o t}$. Because the output of a filter with frequency response $\widehat{H}$ in response to input $t \mapsto e^{j\Omega_o t}$ is $t \mapsto \widehat{H}(\Omega_o)e^{j\Omega_o t}$, and because each of the ideal filters has $\widehat{H}(\Omega) = 1$ over some range of frequencies, that range of frequencies therefore constitutes the filter's passband. Similarly, each filter has $\widehat{H}(\Omega) = 0$ over some range of frequencies, and that range of frequencies constitutes the filter's stopband.

If x is a more general input signal and $x \overset{\mathcal{F}}{\longleftrightarrow} \widehat{X}$, we know from the Convolution rule 10.10 that the output y of the system with frequency response $\widehat{H}$ due to input x has Fourier transform $\widehat{Y}$ with specification

$$\widehat{Y}(\Omega) = \widehat{H}(\Omega)\widehat{X}(\Omega) \text{ for all } \Omega \in \mathbb{R} .$$

Thus the ideal filters serve to "chop" the spectrum of x in such a way that $\widehat{Y}$ looks exactly like $\widehat{X}$ over the range of frequencies lying in the filters' passbands and $\widehat{Y}(\Omega) = 0$ for all Ω lying in the filters' stopbands. An ideal filter passes unchanged the "frequency content of x" that lies in the filter's passband and annihilates the "frequency content of x" that lies in its stopband.

It would be nice if we could build LTI systems that behaved just like these ideal filters, but, as I mentioned earlier, we can't. What I mean is that no causal LTI system has frequency response that matches that of any of the three ideal filters. An elementary way to see this is to refer to the Symmetry properties 10.12 of Fourier transforms. The frequency responses of the ideal filters are all real-valued and even functions of Ω. Accordingly, the impulse responses of the LTI systems corresponding with the ideal filters must be real-valued and even functions of t. Such signals cannot be impulse responses of causal LTI systems by Theorem 8.5.

More generally, no causal LTI system with a decent real-valued impulse response h has a real-valued frequency response $\widehat{H}$. Referring again to the Symmetry properties 10.12, note that if h is real-valued and $h \overset{\mathcal{F}}{\longleftrightarrow} \widehat{H}$, then $\widehat{H}(-\Omega) = \overline{\widehat{H}(\Omega)}$ for all Ω. If we write $\widehat{H}$ in polar form, i.e.

$$\widehat{H}(\Omega) = |\widehat{H}(\Omega)|e^{j\phi(\Omega)} ,$$

we can conclude that $|\widehat{H}(-\Omega)| = |\widehat{H}(\Omega)|$ and $\phi(-\Omega) = -\phi(\Omega)$ for all Ω. If $\widehat{H}$ is real-valued, then $\phi(\Omega) = 0$ or $\pm\pi$ for all Ω, which makes $\Omega \mapsto e^{j\phi(\Omega)}$ a real-valued and even function of Ω since ϕ is an odd function of Ω. In turn, that makes $\widehat{H}$ itself a real-valued and even function of Ω, so h is a real-valued and even signal, implying by virtue of Theorem 8.5 that system with impulse response h is not causal.

The crucial observation is that the frequency response of any physically realizable LTI system will have some nontrivial phase $\Omega \mapsto \phi(\Omega)$. In essence, what's ideal about the ideal filters is that they have zero phase. A number of typical real-world filter-design strategies consist of two steps:

- Figure out a desired frequency-response magnitude $\Omega \mapsto |\widehat{H}_{\text{des}}(\Omega)|$
- Design a causal LTI system whose frequency-response magnitude matches the desired magnitude and whose phase is "not too damaging."

Phase wouldn't be much of a problem if we were interested only in pure sinusoidal inputs to the system. Suppose we have $|\widehat{H}_{\text{des}}|$. Let y_{des} be the output that arises from applying input x specified by $x(t) = e^{j\Omega_o t}$ to the (unrealizable) system with frequency response $|\widehat{H}_{\text{des}}|$, so

$$y_{\text{des}}(t) = |\widehat{H}_{\text{des}}(\Omega_o)|e^{j\Omega_o t} \text{ for all } t \in \mathbb{R} .$$

The output y that arises when we apply the same input to any system with frequency response

$$\widehat{H}(\Omega) = |\widehat{H}_{\text{des}}(\Omega)|e^{j\phi(\Omega)} \text{ for all } \Omega \in \mathbb{R}$$

has specification

$$\begin{aligned} y(t) &= \widehat{H}(\Omega_o)e^{j\Omega_o t} \\ &= |\widehat{H}_{\text{des}}(\Omega_o)|e^{j(\Omega_o t+\phi(\Omega_o))} \\ &= |\widehat{H}_{\text{des}}(\Omega_o)|e^{j\Omega_o(t+\phi(\Omega_o)/\Omega_o)} \\ &= \text{Shift}_T(y_{\text{des}}(t)) , \end{aligned}$$

where $T = -\phi(\Omega_o)/\Omega_o$. Thus the response to x of any system whose frequency response has the desired magnitude along with some phase is simply a time shift of the response to x of the desired phase-free system.

As soon as we force a system with an input whose frequency content is distributed over a range of frequencies — a discrete range for an arbitrary periodic signal and a continuous range for a more general signal — phase in the frequency response becomes an issue. Roughly speaking, different frequency components of the input get phase-shifted by different amounts. What might it mean for phase in a frequency response to be "not too damaging?"

Suppose we can solve the filter-design problem above by building a system whose frequency response has *linear phase* in the sense that, for some $T > 0$,

$$\widehat{H}(\Omega) = |\widehat{H}_{\text{des}}(\Omega)|e^{-j\Omega T} \text{ for all } \Omega \in \mathbb{R} .$$

Then for an arbitrary input x, the output y of the system in response to x will have Fourier transform

$$\widehat{Y}(\Omega) = e^{-j\Omega T}|\widehat{H}_{\text{des}}(\Omega)|\widehat{X}(\Omega) .$$

By the Time-shift rule 10.5, $y = \text{Shift}_T(y_{\text{des}})$, where y_{des} is the response to x of the ideal phase-free system. In other words, the response to an arbitrary input of a system whose frequency response has linear phase will be a delayed version of the response to that input of the unrealizable system with phase-free frequency response. One could therefore argue that linear phase has relatively benign effects.

Applications demand not only that we design filters that perform according to some specifications but also that we devise schemes to mitigate the undesirable impacts of filters we're stuck with. Consider the problem of transmitting a relatively low-frequency signal, say an audio signal, from a source to a distant destination or destinations. Acoustical approaches to the problem are obvious non-starters. Converting the audio signal to an electromagnetic signal to be transmitted wirelessly through the atmosphere is a step in the right direction, but the atmosphere acts as a high-pass filter of electromagnetic radiation, and the frequency content of our electromagnetic signal lies well outside of the atmospheric filter's passband. How do we circumvent the atmospheric obstacle?

Suppose x is bandlimited to within a relatively small $\Omega_m > 0$ — for example, $\Omega_m \approx 40,000\pi$ for an audio signal. Let $\Omega_c > 0$ be large enough so that the intervals of frequencies $[\Omega_c - \Omega_m, \Omega_c + \Omega_m]$ and $[-\Omega_c - \Omega_m, -\Omega_c + \Omega_m]$ lie within the passband of the atmosphere, which for simplicity I'll view as an ideal high-pass filter. Form the signal z with specification

$$\begin{aligned} z(t) &= x(t)\cos\Omega_c t \\ &= \frac{1}{2}x(t)e^{j\Omega_c t} + \frac{1}{2}x(t)e^{-j\Omega_c t} \text{ for all } t \in \mathbb{R} . \end{aligned}$$

By the Frequency-shift rule 10.6,

$$\widehat{Z}(\Omega) = \frac{1}{2}\widehat{X}(\Omega - \Omega_c) + \frac{1}{2}\widehat{X}(\Omega + \Omega_c) ,$$

so the nonzero part of z's spectrum lies in the filter's passband, and if we transmit z it passes through the filter unchanged. Essentially what we've done is piggy-back the low-frequency signal x onto a high-frequency signal z. In this particular example, z is a high-frequency cosine with amplitude modulated by the low–frequency signal x, which is why people allude to this scheme as *amplitude modulation* and call Ω_c the *carrier frequency* associated with the scheme. The recipient of the transmitted signal z can recover x, at least approximately, by first forming the signal y

with specification

$$\begin{aligned} y(t) &= z(t)\cos\Omega_c t \\ &= \frac{1}{2}z(t)e^{j\Omega_c t} + \frac{1}{2}z(t)e^{-j\Omega_c t} \text{ for all } t \in \mathbb{R} . \end{aligned}$$

The Frequency-shift rule yields

$$\widehat{Y}(\Omega) = \frac{1}{2}\widehat{Z}(\Omega - \Omega_c) + \frac{1}{2}\widehat{Z}(\Omega + \Omega_c) ,$$

and you can verify easily that $\widehat{Y}(\Omega) = \widehat{X}(\Omega)/2$ for $-\Omega_m \le \Omega \le \Omega_m$. Passing y through an ideal low-pass filter with frequency response

$$\widehat{H}(\Omega) = \begin{cases} 2 & \text{when } |\Omega| \le \Omega_m \\ 0 & \text{otherwise} \end{cases}$$

yields x exactly, but of course you can't build an ideal low-pass filter, so you need to approximate it and therefore can produce at best an approximation of x.

The scheme I've just described, known as *amplitude modulation with synchronous demodulation*, doesn't work in practice because it requires every recipient of the signal z to have access to a cosine signal synchronized perfectly with the cosine signal at the source — and it gets worse when you try to account for source-to-destination time delays, which differ for the different destinations that arise in a broadcast setting. Amplitude modulation with *asynchronous demodulation* gets around this difficulty by forming its z-signal differently and prescribing a more realistic procedure for recovering x approximately from z. Assume again that x is bandlimited to within $\Omega_m > 0$ and that we've selected Ω_c as above. Choose any $m \in (0, 1)$ for which $1 + mx(t) > 0$ for all $t \in \mathbb{R}$, where for simplicity I'm considering only real-valued x. Such an m, called the *modulation index*, exists when x is bounded, which is a reasonable assumption. The transmitted signal z has specification

$$z(t) = (1 + mx(t))\cos\Omega_c t \text{ for all } t \in \mathbb{R} .$$

Thus z is the sum of a cosine whose frequency lies within the channel pass-band and the signal $t \mapsto mx(t)\cos\Omega_c t$, which we know already passes through the channel unadulterated. Accordingly, z arrives unchanged at the destination(s). Recovering x from z — i.e. demodulation — entails following the positive peaks in z to get an approximation of $1 + mx$ and from that an approximation of x. No synchronous cosine generator is required, unlike the previous scheme where $z(t) = x(t)\cos\Omega_c t$ for all t. Why, you might ask, couldn't we have demodulated asynchronously there, as well? The answer is simple: following the positive peaks in $z(t) = x(t)\cos\Omega_c t$ yields an approximation of $|x|$, which isn't the same as x if x ever takes on negative values.

Finally, consider the multiple-access problem that arises when several agents want to transmit their own low-frequency signals simultaneously through the same high-pass channel using amplitude modulation. Say agent i, for $1 \le i \le N$, wants to send signal x_i, and assume that all the signals x_i are bandlimited to within Ω_m. Every agent wants to transmit its signal through the channel by modulating the amplitude of a high-frequency cosine, so agent i transmits a signal z_i with spectrum confined to $\Omega_i - \Omega_m \le |\Omega| \le \Omega_i + \Omega_m$, where Ω_i is agent i's carrier frequency. Recipients receive the signal $z = \sum_{i=1}^{N} z_i$, and a recipient looking to recover x_i by demodulation must be able to extract z_i from z.

To make that extraction possible, we assign carrier frequencies to the transmitting agents so as to allow them to share the channel bandwidth in such a way that their transmissions don't interfere with each other. Specifically, we require $|\Omega_i - \Omega_j| \geq 2\Omega_m$ for all i and j. This condition guarantees that the spectra of the various z_i are nonzero over disjoint intervals in Ω-space. A recipient of the signal z can extract z_i by sending z through a band-pass filter that annihilates all the z_j for $j \neq i$ and can subsequently recover x_i approximately from z_i as before. Figure 4 might help you understand this scheme, known as *frequency-division multiplexing.*

AM radio, roughly speaking, solves the problem of broadcasting multiple audio signals simultaneously over long distances by implementing frequency-division multiplexing of amplitude-modulated signals. In the US, the FCC assigns carrier frequencies to transmitting agents in local AM markets wherein agents' transmissions might interfere with each other without regulation. The numbers on an AM radio station-selection dial represent carrier frequencies in kilohertz. Centering that dial on Ω_i — more accurately on $\Omega_i/2\pi$ — slides the passband of a variable band-pass filter so as to capture the signal z_i associated with the agent using the carrier frequency Ω_i. I won't elaborate on various bandwidth-conserving and fidelity-enhancing embellishments of the basic schemes I've described. Our goal is to understand the essential mathematics and also arguably to appreciate how mathematics motivates clever solutions to real-world engineering problems.

CHAPTER 11

The Discrete-Time Fourier Transform and Sampling

We've developed some facility with frequency-domain concepts in continuous time. Expanding a periodic signal in a Fourier series — a discrete superposition of pure sinusoids — reveals how the signal's frequency content is distributed over a discrete set of frequencies. We can't expand a non-periodic signal in a Fourier series, but we can often express it as a "continuum superposition" of pure sinusoids using the continuous-time Fourier transform by means of equation $\mathcal{F}^{-1}$. Like Fourier coefficients for periodic signals, the Fourier transform of a continuous-time signal indicates how the signal's frequency content, also known as its spectral content, is distributed over frequency space. How do we make sense of the frequency-domain idea for discrete-time signals?

Definition of the DTFT

The fundamental question is, what might it mean for a discrete-time signal to have frequency content at some frequency ω_o? The most elementary periodic signal in continuous time is a pure sinusoid like $t \mapsto e^{j\Omega_o t}$, which has all its frequency content at Ω_o. So how about declaring that the analogous discrete-time signal $n \mapsto e^{jn\omega_o}$ has frequency content at ω_o? Let's make that declaration while keeping in mind that its meaning isn't obvious. For starters, the sequence $\{e^{jn\omega_o}\}$ is rarely a periodic sequence of numbers, so the words "frequency" and "period" no longer dance in step with each other. One source of difficulty is that, unlike the continuous time variable t, the discrete time index n is "unit-free." We think of t as being measured in seconds, so frequencies Ω are measured in radians per second. In discrete time, we have no corresponding unit-like notion associated with the frequency variable ω. Observe also that $e^{jn\omega_o} = e^{jn(\omega_o+2\pi k)}$ for all integers n and k. So if $n \mapsto e^{jn\omega_o}$ has frequency content at ω_o, then it ought to have the same frequency content at $\omega_o + 2\pi k$ for every k. It's reasonable to expect the same to be true of an arbitrary discrete-time signal x — that is, if x has frequency content at ω_o, then x ought to have the same frequency content at $\omega_o + 2\pi k$.

Let's now see what happens if we attempt to define the discrete-time Fourier transform by imitating the continuous-time theory. Let $x \in \mathbb{C}^{\mathbb{Z}}$ be a complex-valued discrete-time signal. The equation

$$(\mathcal{DTFT}) \qquad \widehat{X}(\omega) = \sum_{n=-\infty}^{\infty} x(n)e^{-jn\omega} \quad \text{for all } \omega \in \mathbb{R}$$

is the analogue of equation $\mathcal{F}$ in continuous time. The sum in $\mathcal{DTFT}$ need not converge — it depends on properties of the signal x. Note that if it does converge for all $\omega \in \mathbb{R}$, the function $\omega \mapsto \widehat{X}(\omega)$ is periodic in ω, and has 2π as a period. Equation $\mathcal{DTFT}$ is tantamount to a Fourier-series expansion of the function $\widehat{X}$. If you compare $\mathcal{DTFT}$ with the Fourier series

$$\widehat{X}(\omega) = \sum_{k=-\infty}^{\infty} c_k e^{jk\omega} \quad \text{for all } \ \omega \in \mathbb{R} \ ,$$

for $\widehat{X}$, you see that $x(k) = c_{-k}$ for all $k \in \mathbb{Z}$. Thus

$$x(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \widehat{X}(\omega) e^{-j(-n)\omega} d\omega \quad \text{for all } \ n \in \mathbb{Z} \ ,$$

which is the same as

$$(\mathcal{DTFT}^{-1}) \qquad x(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \widehat{X}(\omega) e^{jn\omega} d\omega \quad \text{for all } \ n \in \mathbb{Z} \ .$$

Here is the formal definition of the discrete-time Fourier transform.

11.1 Definition: Let $x \in \mathbb{C}^{\mathbb{Z}}$ be a discrete-time signal and let $\omega \mapsto \widehat{X}(\omega)$ be a complex-valued function of the real variable ω that has 2π as a period. We say that x and $\widehat{X}$ are a *discrete-time Fourier transform pair* when one or both of the equations $\mathcal{DTFT}$ or $\mathcal{DTFT}^{-1}$ holds. In this case, we also say that $\widehat{X}$ is the *discrete-time Fourier transform*, or the *DTFT*, of x, and write

$$x \xleftrightarrow{\mathcal{DTFT}} \widehat{X} \ .$$

It's worth pondering what the 2π-periodicity of $\widehat{X}$ means intuitively. Based on our declaration that the pure discrete-time sinusoid $n \mapsto e^{jn\omega_o}$ has frequency content at ω_o and therefore at every frequency $\omega_o + 2\pi k$, we developed the expectation that the frequency content of an arbitrary discrete-time signal x ought to be the same in the neighborhood of each frequency $\omega_o + 2\pi k$ as it is in the neighborhood of frequency ω_o. Thus $\widehat{X}$, which is supposed to represent the frequency content of the signal x, is unsurprisingly 2π-periodic. The periodicity of $\widehat{X}$ allows us to determine $\widehat{X}$ completely by specifying it on the ω-interval $-\pi \leq \omega \leq \pi$. Getting the entire $\widehat{X}$ from such a partial specification entails a simple 2π-periodic extension. For example, if $\widehat{X}$ resembles the graph in Figure 1 on the interval $|\omega| \leq \pi$, then $\widehat{X}$ resembles the graph in Figure 2 as a function of all ω.

I won't spend a lot of time talking about what signals have DTFTs and what signals don't. Suffice it to say that not every signal has a DTFT and not every 2π-periodic function of ω is the DTFT of some signal. For example, the signal x with specification $x(n) = 3^{|n|}$ for all $n \in \mathbb{Z}$ doesn't have a DTFT. It's fairly easy to see for this example that the sum in $\mathcal{DTFT}$ won't converge for most $\omega \in \mathbb{R}$. It's true but not so easy to see that you can't find a function $\widehat{X}$ so that $\mathcal{DTFT}^{-1}$ holds for the given x. In view of our observation that the values of a signal x are Fourier-series coefficients for $\widehat{X}$, the question of whether a signal x has a DTFT is equivalent to the question of whether x's values constitute the set of Fourier-series coefficients for some 2π-periodic function of ω. That question is deep, but some partial answers are worth mentioning.

If x is an absolutely summable signal, then the sum in $\mathcal{DTFT}$ converges for every $\omega \in \mathbb{R}$ by Fact 3.3 and therefore defines a function $\widehat{X}$. Actually, $\widehat{X}$ turns out to be a continuous function of ω in this case. If x is a square-summable signal, then the series in $\mathcal{DTFT}$ converges in the mean-square sense to a function $\widehat{X}$ for which

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} |\widehat{X}(\omega)|^2 d\omega$$

exists. Mean-square convergence amounts to

$$\lim_{N\to\infty} \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|\widehat{X}(\omega) - \sum_{n=-N}^{N} x(n)e^{-jn\omega}\right|^2 d\omega = 0 \ .$$

These l^2 results hinge on x's status as a list of Fourier-series coefficients for $\widehat{X}$ and follow from reasoning similar to that in Chapter 9 leading up to Parseval's Theorem 9.13. In fact, when x is an l^2-signal, we have

$$\sum_{n=-\infty}^{\infty} |x(n)|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |\widehat{X}(\omega)|^2 d\omega$$

and, more generally,

$$\sum_{n=-\infty}^{\infty} x(n)\overline{y(n)} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \widehat{X}(\omega)\overline{\widehat{Y}(\omega)} d\omega \tag{10}$$

for every x and y in l^2.

Finally, if $\widehat{X}$ is any decent function of ω with 2π as a period, there exists a discrete-time signal x whose DTFT is $\widehat{X}$. I'm using the word "decent" here in the technical sense of Definition 7.1. If $\widehat{X}$ is decent, then the integrals in equation $\mathcal{DTFT}^{-1}$ pose no problems and define a signal x.

The discrete-time impulse δ is a genuine signal, and computing its DTFT using equation $\mathcal{DTFT}$ yields

$$\delta \xleftrightarrow{\mathcal{DTFT}} 1 \ .$$

In other words, the DTFT of a discrete-time impulse is the constant function of ω with constant value 1. Allowing for impulses in ω-space gives us DTFTs for pure discrete-time sinusoids. Since any DTFT $\widehat{X}$ must be 2π-periodic in ω, the simplest DTFT featuring an impulse is the 2π-periodic impulse train with specification

$$\widehat{X}(\omega) = \sum_{k=-\infty}^{\infty} 2\pi\delta(\omega - \omega_o + 2\pi k) \ ,$$

where ω_o is some real number. Note that unless ω_o is an odd multiple of π, in which case $e^{jn\omega_o} = (-1)^n$ for all n, precisely one of the impulses in the train — say the one located at $\omega_o + 2\pi k^*$ — falls in the central interval $|\omega| < \pi$. Investigating equation $\mathcal{DTFT}^{-1}$ for this example yields

$$\begin{aligned}
\frac{1}{2\pi}\int_{-\pi}^{\pi} 2\pi \sum_{k=-\infty}^{\infty} \delta(\omega - \omega_o + 2\pi k) e^{jn\omega} d\omega &= \int_{-\pi}^{\pi} \sum_{k=-\infty}^{\infty} \delta(\omega - \omega_o + 2\pi k) e^{jn\omega} d\omega \\
&= \int_{-\pi}^{\pi} \delta(\omega - \omega_o + 2\pi k^*) e^{jn\omega} d\omega \\
&= e^{jn(\omega_o + 2\pi k^*)} = e^{jn\omega_o} \ \text{ for all } \ n \in \mathbb{Z} \ .
\end{aligned}$$

The equality on the second line holds because the interval of integration includes none of the impulses in the sum on the first line except the one corresponding to $k = k^*$. We conclude that the impulse train $\widehat{X}$ is the DTFT of the pure discrete-time sinusoid x with specification $x(n) = e^{jn\omega_o}$ for all $n \in \mathbb{Z}$, which supports our intuition that the pure discrete-time sinusoid x has all its frequency content at frequencies in the discrete set $\{\omega_o + 2\pi k : k \in \mathbb{Z}\}$.

Like the continuous-time Fourier transform, the DTFT has numerous associated operational rules. Two of these are particularly important.

11.2 Time-shift Rule: If $x \xleftrightarrow{\mathcal{DTFT}} \widehat{X}$, then for any $n_o \in \mathbb{Z}$ the DTFT $\widehat{Y}$ of the signal $y = \text{Shift}_{n_o}(x)$ has specification

$$\widehat{Y}(\omega) = e^{-jn_o\omega}\widehat{X}(\omega) \ \text{ for all } \ \omega \in \mathbb{R} \ .$$

To see this, assume equation $\mathcal{DTFT}^{-1}$ holds. Then

$$y(n) = x(n - n_o) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \widehat{X}(\omega) e^{j(n-n_o)\omega} d\omega = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left(e^{-jn_o\omega}\widehat{X}(\omega)\right) e^{jn\omega} d\omega$$

for all $n \in \mathbb{Z}$, and this is just equation $\mathcal{DTFT}^{-1}$for y, revealing that $\widehat{Y}(\omega) = e^{-jn_o\omega}\widehat{X}(\omega)$ for all $\omega \in \mathbb{R}$. □

11.3 Convolution Rule: If $x_1 \xleftrightarrow{\mathcal{DTFT}} \widehat{X}_1$ and $x_2 \xleftrightarrow{\mathcal{DTFT}} \widehat{X}_2$ and the convolution of x_1 and x_2 exists, then

$$x = x_1 * x_2 \xleftrightarrow{\mathcal{DTFT}} \widehat{X} = \widehat{X}_1 \widehat{X}_2 \ .$$

In other words, the DTFT takes convolution in the time domain to simple function multiplication in the frequency domain.

In proving this one, I'll assume first that both x_1 and x_2 are absolutely summable, which implies that we can interchange orders of summation with impunity in the equations below. Start with equation $\mathcal{DTFT}$ for x.

$$\begin{aligned}
\widehat{X}(\omega) &= \sum_{n=-\infty}^{\infty} x(n)e^{-jn\omega} \\
&= \sum_{n=-\infty}^{\infty} \left(\sum_{k=-\infty}^{\infty} x_1(k)x_2(n-k) \right) e^{-jn\omega} \\
&= \sum_{k=-\infty}^{\infty} x_1(k) \left(\sum_{n=-\infty}^{\infty} x_2(n-k)e^{-jn\omega} \right) \\
&= \sum_{k=-\infty}^{\infty} x_1(k) \left(e^{-jk\omega}\widehat{X}_2(\omega) \right) \\
&= \left(\sum_{k=-\infty}^{\infty} x_1(k)e^{-jk\omega} \right) \widehat{X}_2(\omega) \\
&= \widehat{X}_1(\omega)\widehat{X}_2(\omega) \ \text{ for all } \ \omega \in \mathbb{R} \ .
\end{aligned}$$

The crucial step on the fourth line follows from the Time-Shift Rule applied to the inner sum.

When both x_1 and x_2 are square-summable, an entirely different argument applies. First fix $n \in \mathbb{Z}$ and let y be the signal with specification $y(k) = \overline{x_2(n-k)}$ for all $k \in \mathbb{Z}$. Observe that $y \in l^2$, and

$$\begin{aligned}
\langle x_1, y \rangle &= \sum_{k=-\infty}^{\infty} x_1(k)\overline{y(k)} \\
&= \sum_{k=-\infty}^{\infty} x_1(k)x_2(n-k) \\
&= x_1 * x_2(n) \ .
\end{aligned}$$

Furthermore, y has DTFT $\widehat{Y}$ with specification

$$\begin{aligned}
\widehat{Y}(\omega) &= \sum_{k=-\infty}^{\infty} y(k)e^{-jk\omega} \\
&= \sum_{k=-\infty}^{\infty} \overline{x_2(n-k)}e^{-jk\omega} \\
&= \sum_{m=-\infty}^{\infty} \overline{x_2(m)}e^{-j\omega(n-m)} \\
&= \overline{\widehat{X}_2(\omega)}e^{-jn\omega}
\end{aligned}$$

for all $\omega \in \mathbb{R}$. Applying the Parseval-like (10) yields

$$\langle x_1, y \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} \widehat{X}_1(\omega)\overline{\widehat{Y}(\omega)}d\omega \ ,$$

so

$$x_1 * x_2(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \widehat{X}_1(\omega)\widehat{X}_2(\omega)e^{jn\omega}d\omega \ \text{ for all } \ n \in \mathbb{Z} \ ,$$

which is just equation $\mathcal{DTFT}^{-1}$ for $x_1 * x_2$, revealing that $x_1 * x_2$ has DTFT $\widehat{X}_1\widehat{X}_2$. □

The DTFT plays a role in discrete-time LTI systems analysis analogous to the role that the continuous-time Fourier transform plays in continuous-time LTI systems analysis.

11.4 Definition: We say that a discrete-time LTI system with impulse response $h \in \mathbb{C}^{\mathbb{Z}}$ *has a frequency response* when the input signal $n \mapsto e^{jn\omega_o}$ is in the input space $X = \mathcal{D}_h$ for every $\omega_o \in \mathbb{R}$. In this case, we define the *frequency response* of the system as the DTFT $\widehat{H}$ of the impulse response h.

While not every system has a frequency response, every FIR system has one, as does every BIBO stable system, since by Theorem 6.7 the impulse response of a BIBO stable system is absolutely summable. If a system has a frequency response, consider what happens when we use input signal x with specification $x(n) = e^{jn\omega_o}$ as input to the system. We discover that

$$\begin{aligned} S(x)(n) &= h * x(n) \\ &= \sum_{k=-\infty}^{\infty} h(k)x(n-k) \\ &= \sum_{k=-\infty}^{\infty} h(k)e^{j(n-k)\omega_o} \\ &= \left(\sum_{k=-\infty}^{\infty} h(k)e^{-jk\omega_o}\right)e^{jn\omega_o} \\ &= \widehat{H}(\omega_o)e^{jn\omega_o} \quad \text{for all} \quad n \in \mathbb{Z}\ . \end{aligned}$$

In other words, $S(x) = \widehat{H}(\omega_o)x$, so x is an "eigen-input" to the system. On the other hand, if $x \in \mathcal{D}_h$ and x has DTFT $\widehat{X}$, then the Convolution Rule 11.3 implies that $S(x) = h * x$ has DTFT $\widehat{H}\widehat{X}$. In these senses, as in continuous time, you can regard a discrete-time LTI system with a frequency response as a frequency-selective filter.

Sampling and Interpolation

One way of generating discrete-time signals is by sampling continuous-time signals. Given a continuous-time signal x_c and a sampling interval $T > 0$, consider what happens when we form the discrete-time signal x specified by

$$x(n) = x_c(nT) \quad \text{for all} \quad n \in \mathbb{Z}\ .$$

Certainly, x_c tells us everything about x. What about the converse? What does x tell us about x_c? You can think of x as a sparse, skeletal version of x_c, comprising only a discrete set of numbers — a sequence of dots, if you will. *A priori*, you might not expect it to contain a lot of information about the continuum of numbers that constitutes x_c. Our intuition tells us that for x to determine a lot about x_c, the sampling interval will need to be small enough so that x captures the "interesting features" in x_c. For example, suppose $x_c(t) = \sin t$ and $T = \pi$. Then $x(n) = \sin(n\pi) = 0$ for every n, and so x — since it's just a long string of zeroes — tells us essentially nothing interesting about x_c. On the other hand, if T is much smaller, the oscillations in $x(n)$ as a function of n will resemble those in the sinusoid x_c. The Shannon-Nyquist Sampling Theorem enables us to make quantitative sense of all this.

Think now of any discrete-time signal x as comprising a sequence of dots that mark its values. By connecting those dots to form a continuous-time signal, we are performing *interpolation.* Specifically, given a discrete-time signal x and some $T > 0$, I'll call any continuous-time signal y_c that satisfies $y_c(nT) = x(n)$ for all $n \in \mathbb{Z}$ a *T-interpolation* of x. The signal y_c with specification

$$y_c(t) = x(n) + \frac{t - nT}{T}(x(n+1) - x(n))$$

for all $t \in [nT, (n+1)T]$ and $n \in \mathbb{Z}$ is the linear T-interpolation of x. Another T-interpolation of x is the signal y_c with specification

$$y_c(t) = x(n) \; + \; (x(n+1) - x(n)) \sin\left(\frac{\pi}{2T}(t - nT)\right)$$

for all $t \in [nT, (n+1)T]$ and $n \in \mathbb{Z}$. If x arose originally from T-sampling a continuous-time signal x_c — i.e. if $x(n) = x_c(nT)$ for all $n \in \mathbb{Z}$ — we have no reason to expect that either of these T-interpolations y_c will bear any resemblance to x_c other than agreeing with x_c at the sampling instants $\{nT : n \in \mathbb{Z}\}$.

You perform interpolation whenever you watch a movie. A movie is a discrete sequence of still images, and the people who filmed the movie generated that sequence by sampling a continuous-time visual signal. Standard Hollywood movies are filmed at 24 frames per second, which means that movie-makers sample visual signals at angular frequency $\Omega_s = 48\pi$ with sampling interval $T_s = 1/24$ sec. When you watch a movie, you generally manage to interpolate in such a way as to arrive at a fairly good idea of what the original continuous-time visual signal looked like, but consider the following notable exception. Everyone has seen a Western featuring covered wagons lumbering frontierward, their giant spoked wheels appearing to turn in bizarrely unphysical ways. The wagon moves from left to right across the screen, but the wheel appears sometimes to be rotating clockwise, which is what it was actually doing while being filmed, and sometimes to be rotating counterclockwise or even standing still. As you "see" the wheel moving "incorrectly" on the screen, you interpolate the movie x and manufacture a continuous-time signal y_c that's not the same as the original continuous-time signal x_c that the movie-makers sampled.

Let's put some numbers to the movie example. Suppose someone films a vehicle moving left to right across the screen. The filming occurs at the standard rate of 24 frames per second. One of the wheels on the vehicle has a radial marking whose position indicates the wheel's orientation. Physics tells us that the wheel was rotating clockwise while being filmed. Now watch the movie. Referring to Figure 3, Case 1 shows the sequence of frames you see when the vehicle was moving so that the wheel made a complete rotation every 1/24 second. Case 2 shows what you see when the wheel was making one rotation every 1/18 second, which is the same as 3/4 of a rotation every 1/24 second. Case 3 corresponds to a rotation every

1/12 second, or half a rotation every 1/24 second, and Case 4 to a rotation every 1/6 second, or one quarter of a rotation every 1/24 second. I would argue that in Case 1 the wheel looks to you as if it were standing still; in Case 2 the wheel looks as if it were making 1/4 rotation counterclockwise every 1/24 second; in Case 3 your brain can't decide which way the wheel is rotating, and in fact snaps back and forth between the two possibilities; and in Case 4 the wheel looks as if it were doing what it was actually doing in real life. We'll return to the movie example after developing the mathematics that explains it.

Deconstruction, reconstruction, and the Sampling Theorem

To understand the relationship between a continuous-time signal x_c and the discrete-time signal x obtained by T-sampling x_c, we need to focus on what's happening in the frequency domain. Our first order of business will be to derive a formula expressing the DTFT of x in terms of the continuous-time Fourier transform of x_c. With apologies to Derrida, I'm calling that formula the deconstruction equation.

11.5 The Deconstruction Equation: Given a continuous-time signal x_c and a sampling interval $T > 0$, let x be the discrete-time signal with specification $x(n) = x_c(nT)$, for all $n \in \mathbb{Z}$. If x has a DTFT $\widehat{X}$, then $\widehat{X}$ has specification

$$(\mathcal{D}) \qquad \widehat{X}(\omega) = \sum_{k=-\infty}^{\infty} \frac{1}{T}\widehat{X}_c\left(\frac{\omega}{T} + k\frac{2\pi}{T}\right) \text{ for all } \omega \in \mathbb{R} ,$$

where $\widehat{X}_c$ is the continuous-time Fourier transform of x_c.

Proof: Start with $\mathcal{F}^{-1}$ for x_c, which yields

$$x_c(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \widehat{X}_c(\Omega)e^{j\Omega t}d\Omega \text{ for all } t \in \mathbb{R} .$$

Plug in $t = nT$ and $x(n) = x_c(nT)$ and you get

$$x(n) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \widehat{X}_c(\Omega)e^{j\Omega nT}d\Omega \text{ for all } n \in \mathbb{Z} .$$

Now split up the interval of integration into chunks of Ω-length $2\pi/T$ centered on the values $\Omega = k2\pi/T$:

$$x(n) = \sum_{k=-\infty}^{\infty} \frac{1}{2\pi}\int_{k\frac{2\pi}{T}-\frac{\pi}{T}}^{k\frac{2\pi}{T}+\frac{\pi}{T}} \widehat{X}_c(\Omega)e^{j\Omega nT}d\Omega \text{ for all } n \in \mathbb{Z} .$$

Change variables in the kth term by setting $\mu = \Omega - k\frac{2\pi}{T}$:

$$x(n) = \sum_{k=-\infty}^{\infty} \frac{1}{2\pi}\int_{-\frac{\pi}{T}}^{\frac{\pi}{T}} \widehat{X}_c\left(\mu + k\frac{2\pi}{T}\right)e^{j(\mu nT+2\pi kn)}d\mu \text{ for all } n \in \mathbb{Z} .$$

Now use $e^{j2\pi kn} = 1$ along with the change of variable $\omega = \mu T$; also, assume that you can move the sum inside the integral, and do it:

$$x(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\sum_{k=-\infty}^{\infty}\frac{1}{T}\widehat{X}_c\left(\frac{\omega}{T}+k\frac{2\pi}{T}\right)\right)e^{jn\omega}d\omega \;\text{ for all }\; n \in \mathbb{Z}\ .$$

This last equation is just $\mathcal{DTFT}^{-1}$ for x and reveals that the sum in large parentheses is precisely $\widehat{X}(\omega)$, whereby

$$\widehat{X}(\omega) = \sum_{k=-\infty}^{\infty}\frac{1}{T}\widehat{X}_c\left(\frac{\omega}{T}+k\frac{2\pi}{T}\right) \;\text{ for all }\; \omega \in \mathbb{R}\ ,$$

which is the deconstruction equation $\mathcal{D}$. □

The foregoing proof dodges some significant analytical obstacles. For one thing, it assumes that equation $\mathcal{F}^{-1}$ holds for x_c. That's not a problem if x_c is bandlimited and $\widehat{X}_c$ is reasonable. The proof also plays fast and loose with infinite sums and interchanging summation and integration. Again, that's not a problem when x_c is bandlimited, because in that case the ostensibly infinite sums have only finitely many nonzero terms for any $\omega \in \mathbb{R}$.

Perhaps more important, the proof doesn't address the question of under what circumstances the signal x has a DTFT. For some signals x_c, it's possible to choose a sampling interval T that leads to an x with no DTFT. For example, let x_c be the signal with the following specification: $x_c(t) = 0$ for all t except for t-values that lie in narrow intervals around nonzero integer values of t, so $x_c(t) = 0$ except that

$$x_c(t) = 3^{|n|} \;\text{ when }\; n - \left(3^{-2|n|}/2\right) \le t \le n + \left(3^{-2|n|}/2\right)$$

for some $n \in \mathbb{Z}$. We met this signal back in Chapter 7. It's an L^1-signal so it has a Fourier transform and equation $\mathcal{F}$ holds. But if $T = 1$, the discrete-time signal x with specification

$$x(n) = x_c(nT) = 3^{|n|} \;\text{ for all }\; n \in \mathbb{Z}$$

has no DTFT. Similar problems occur when T is any integer. A non-integer T, however, leads to a finite-duration x that does have a DTFT. Accordingly, the difficulty arises not simply because x_c is a strange signal but also because of the interplay between the strangeness of x_c and the choice of sampling interval.

Setting aside these misgivings, let's consider what the deconstruction equation says. The first thing to notice is that $\mathcal{D}$ holds for any Fourier-transformable x_c with a reasonable spectrum $\widehat{X}_c$ and any sampling interval T that gives rise to a discrete-time signal x that has a DTFT. No additional assumptions about x_c and T are necessary; in particular, x_c need not be bandlimited. Next, observe that $\mathcal{D}$ exhibits $\widehat{X}$ as the infinite sum of scaled, shifted replicas of $\widehat{X}_c$. The scaling is both in amplitude — by the leading $1/T$ factor — and in frequency variable — by the transformation $\omega \leftrightarrow \Omega T$.

You can build $\widehat{X}$ from $\widehat{X}_c$ as follows. First construct a replica of $\widehat{X}_c$ centered at each Ω-value of the form $k\frac{2\pi}{T}$ and add all the replicas together to obtain

$$F(\Omega) = \sum_{k=-\infty}^{\infty}\widehat{X}_c\left(\Omega + k\frac{2\pi}{T}\right)\ .$$

Then multiply by $1/T$ and re-scale the frequency axis by means of the substitution $\omega = \Omega T$, which yields

$$\widehat{X}(\omega) = \frac{1}{T} F\left(\frac{\omega}{T}\right) \ .$$

In general, the $\widehat{X}_c$-replicas you add to get F will overlap as in Figure 4(a) in the sense that each replica's "interval of nonzeroness" will intersect the interval of nonzeroness of each of its neighboring replicas and perhaps even replicas centered farther away. This phenomenon is called *aliasing,* and I'll have more to say about it later. It's clear that aliasing is unavoidable unless x_c is bandlimited to begin with. If x_c fails to be bandlimited, then $\widehat{X}_c(\Omega)$ will be nonzero "all the way out" in Ω-space, and the replicas we form will always overlap. So to avoid aliasing, we require at least that x_c be bandlimited.

Actually, we need even more. For adjacent replicas to avoid hitting each other, we require that T be small enough so that $\widehat{X}_c(\Omega) = 0$ when $|\Omega| \geq \pi/T$ as in Figure 4(b). If this condition holds, then the replicas' intervals of nonzeroness are all disjoint, and the infinite sum in equation $\mathcal{D}$ is trivial in the sense that for any specific ω-value at most one term in the sum is nonzero. In particular, $\mathcal{D}$ implies in this case that

$$\widehat{X}(\omega) = \frac{1}{T}\widehat{X}_c\left(\frac{\omega}{T}\right) \ \text{ when } -\pi \leq \omega \leq \pi \ .$$

That is, only the $k = 0$-term in the series contributes to $\widehat{X}(\omega)$ on the central interval $|\omega| \leq \pi$. Equivalently,

$$\widehat{X}_c(\Omega) = T\widehat{X}(\Omega T) \ \text{ when } -\frac{\pi}{T} \leq \Omega \leq \frac{\pi}{T} \ .$$

Since we know already that $\widehat{X}_c(\Omega) = 0$ when $|\Omega| \geq \pi/T$, we arrive at a complete specification of $\widehat{X}_c$ in terms of $\widehat{X}$, namely

$$\widehat{X}_c(\Omega) = \begin{cases} T\widehat{X}(\Omega T) & \text{when } -\frac{\pi}{T} \leq \Omega \leq \frac{\pi}{T} \\ 0 & \text{when } |\Omega| \geq \frac{\pi}{T} \ . \end{cases}$$

The last equation has profound implications. It says that $\widehat{X}$ determines $\widehat{X}_c$ completely when the sampling interval T is small enough relative to the bandwidth of x_c. In the time domain, this means that x determines completely the continuous-time signal x_c of which x is a sampled version, provided the sampling is fast enough. In this case we have the following equation, a "reconstruction equation," which follows directly from $\mathcal{F}^{-1}$:

$$x_c(t) = \frac{1}{2\pi}\int_{-\frac{\pi}{T}}^{\frac{\pi}{T}} T\widehat{X}(\Omega T)e^{j\Omega t}d\Omega \ \text{ for all } \ t \in \mathbb{R} \ . \tag{11}$$

We can summarize the discussion so far as follows.

11.6 Shannon-Nyquist Sampling Theorem, Version 1: Let $x_c \in \mathbb{C}^{\mathbb{R}}$ have Fourier transform $\widehat{X}_c$, and let $T > 0$ be given. Let x be the discrete-time signal with specification $x(n) = x_c(nT)$ for all $n \in \mathbb{Z}$. If x_c is bandlimited and T is small enough so that $\widehat{X}_c(\Omega) = 0$ for $|\Omega| \geq \frac{\pi}{T}$, then x determines x_c completely. □

The idea is that x determines $\widehat{X}$; in turn, under the indicated assumptions, $\widehat{X}$ determines x_c by means of the explicit formula (11).

We can reformulate the conditions of the Sampling Theorem in a couple of ways. Suppose x_c is bandlimited. Following Definition 10.16, the bandwidth of x_c is

$$\Omega_m^* = \inf\{\Omega_m > 0 : \widehat{X}_c(\Omega) = 0 \text{ when } |\Omega| \geq \Omega_m\} \ .$$

If $\pi/T > \Omega_m^*$, then $\widehat{X}(\Omega) = 0$ when $|\Omega| \geq \pi/T$. It follows that we can recover x_c from x if $\pi/T > \Omega_m^*$. The frequency $\Omega_s = 2\pi/T$ is the sampling frequency, and the sampling is fast enough when

$$\Omega_s > 2\Omega_m^* \ .$$

In other words, sampling "faster than twice the bandwidth of x_c" always suffices to generate a sampled record that determines x_c uniquely. The frequency $\Omega_{\text{Nyq}} = 2\Omega_m^*$ is called the *Nyquist frequency* or *Nyquist rate* for x_c. The Sampling Theorem asserts that sampling a bandlimited signal faster than its Nyquist rate generates a sequence of samples from which you can reconstruct the signal.

We can reconstitute (11) in the time domain and see exactly how the values of $x(n)$ for $n \in \mathbb{Z}$ determine $x_c(t)$ for all $t \in \mathbb{R}$. Rewrite (11) using equation $\mathcal{DTFT}$ for $\widehat{X}(\Omega T)$ and you get

$$\begin{aligned} x_c(t) &= \frac{1}{2\pi}\int_{-\frac{\pi}{T}}^{\frac{\pi}{T}} T\left(\sum_{n=-\infty}^{\infty} x(n)e^{-j\Omega nT}\right)e^{j\Omega t}d\Omega \\ &= \sum_{n=-\infty}^{\infty} x(n)\left(\frac{1}{2\pi}\int_{-\frac{\pi}{T}}^{\frac{\pi}{T}} Te^{j\Omega(t-nT)}d\Omega\right) \quad \text{for all } \ t \in \mathbb{R} \ . \end{aligned}$$

Evaluating the integrals yields

$$x_c(t) = \sum_{n=-\infty}^{\infty} x(n)\frac{\sin(\frac{\pi}{T}(t-nT))}{\frac{\pi}{T}(t-nT)} \quad \text{for all } \ t \in \mathbb{R} \ . \tag{12}$$

Equation (12) exhibits x_c explicitly in terms of x without frequency-domain intervention, as it were. The right-hand side of (12) is a a *sinc-function interpolation* between the values of x. As a reality check, let's evaluate it at time $t = mT$ and make sure we get $x(m)$. For each $n \neq m$, the nth term in the expansion evaluates to zero because

$$\frac{\sin(\frac{\pi}{T}(mT-nT))}{\frac{\pi}{T}(mT-nT)} = \frac{\sin((m-n)\pi)}{(m-n)\pi} = 0 \ .$$

The $n = m$ term evaluates to $x(m)$ because

$$\lim_{t\to mT} \frac{\sin(\frac{\pi}{T}(t-mT))}{\frac{\pi}{T}(t-mT)} = 1 \ .$$

Sinc-function interpolation: the movies revisited

What happens when the conditions of the Sampling Theorem 11.6 aren't satisfied? In particular, what happens if aliasing occurs because T is too large or because x_c isn't even bandlimited? Equation $\mathcal{D}$ still holds, but equations (11) and (12) are no longer valid. We can imagine taking the sample sequence x and its accompanying DTFT $\widehat{X}$ and plugging them blindly into (11) and (12). The signal that emerges in this case will not be x_c but will instead be some other related signal x_R. The signal x_R has two noteworthy properties. First of all, $x_R(nT) = x(n) = x_c(nT)$ for all n, so x_R does indeed interpolate between the samples. This is because, as we've noted already, the expansion on the right-hand side of

$$(\mathcal{R}2) \qquad x_R(t) = \sum_{n=-\infty}^{\infty} x(n)\frac{\sin(\frac{\pi}{T}(t-nT))}{\frac{\pi}{T}(t-nT)} \ \text{ for all } \ t \in \mathbb{R}$$

evaluates to $x(n)$ when $t = nT$, so $x_R(t)$ agrees with $x_c(t)$ at every sampling instant $t = nT$. Second, x_R is bandlimited; in fact, $\widehat{X}_R(\Omega) = 0$ when $|\Omega| > \frac{\pi}{T}$. This is because

$$(\mathcal{R}1) \qquad x_R(t) = \frac{1}{2\pi}\int_{-\frac{\pi}{T}}^{\frac{\pi}{T}} T\widehat{X}(\Omega T)e^{j\Omega t}d\Omega \ \text{ for all } \ t \in \mathbb{R} ,$$

which implies by equation $\mathcal{F}^{-1}$ that

$$\widehat{X}_R(\Omega) = \begin{cases} T\widehat{X}(\Omega T) & \text{when } -\frac{\pi}{T} \leq \Omega \leq \frac{\pi}{T} \\ 0 & \text{when } |\Omega| > \frac{\pi}{T} \ . \end{cases}$$

We can think of equations $\mathcal{R}1$ and $\mathcal{R}2$ as representing mathematically the operation of a special kind of interpolator. Any interpolator implements a particular recipe for "connecting the dots" in x, and our special interpolator is a device that takes a discrete-time signal x and outputs a continuous-time signal x_R by superposing sinc functions as in $\mathcal{R}2$ or with frequency-domain mediation as in $\mathcal{R}1$. The discrete-time signal x serving as "input" to $\mathcal{R}1$ and $\mathcal{R}2$ need not come from sampling some pre-specified continuous-time signal. Most important, the special interpolation x_R is the only T-interpolation of x that's bandlimited to within π/T. You can see this by noting that if x_c is bandlimited to within π/T, and $x_c(nT) = x(n)$ for all n, then (11) and (12) imply that x_c is the output of the sinc-function T-interpolation driven by x — in other words, $x_c = x_R$. This last observation illuminates the Sampling Theorem from a new angle.

11.7 Shannon-Nyquist Sampling Theorem, Version 2: Suppose x is a discrete-time signal that has a DTFT $\widehat{X}$ and let $T > 0$ be given. There exists exactly one continuous-time signal x_R with the following two properties:

- $x_R(nT) = x(n)$ for all $n \in \mathbb{Z}$
- $\widehat{X}_R(\Omega) = 0$ when $|\Omega| > \pi/T$.

Again, the discrete-time signal x in Theorem 11.7 need not come from sampling some pre-specified continuous-time signal. The theorem requires only that x be a discrete-time signal possessing a DTFT. It asserts that among the myriad T-interpolations of such an x, exactly one is bandlimited to within π/T. That special T-interpolation is the sinc-function interpolation x_R arising from $\mathcal{R}1$ and $\mathcal{R}2$.

Why do people say that "aliasing occurs" when you sample a continuous-time signal x_c slower than its Nyquist rate? If the sampling interval is T and you pass the samples through the sinc-function T-interpolator, you get x_R instead of x_c. In this way, the samples masquerade as a sequence of samples of x_R even though you got them originally by sampling x_c. One might say that the signal x_c "assumes the alias x_R" by virtue of your having sampled it too slowly.

Now let's go back to the movies. You can regard a real-life continuous-time wheel spinning clockwise at frequency Ω_o as defining a continuous-time signal x_c with specification

$$x_c(t) = e^{j\Omega_o t} \text{ for all } t \in \mathbb{R} .$$

Its Fourier transform $\widehat{X}$ has specification

$$\widehat{X}_c(\Omega) = 2\pi\delta(\Omega - \Omega_o) .$$

The wheel turns, the film crew swings into gear, and the camera creates a discrete-time signal x by sampling x_c at frequency $\Omega_s = 48\pi$, which corresponds to 24 frames per second and an inter-sample interval $T = 1/24$ seconds. By equation $\mathcal{D}$, the DTFT $\widehat{X}$ of x has specification

$$\begin{aligned} \widehat{X}(\omega) &= 24 \sum_{k=-\infty}^{\infty} \widehat{X}_c\left(24\omega + k48\pi\right) \\ &= \sum_{k=-\infty}^{\infty} 48\pi\delta\left(24\omega - \Omega_o + k48\pi\right) . \end{aligned}$$

What do you "see" when you watch the movie? I would argue that your visual apparatus, with a bit of help from your experience of the world, acts like a sinc-function interpolator in the sense that you "see" the continuous-time signal x_R given by $\mathcal{R}1$, i.e.

$$\begin{aligned} x_R(t) &= \frac{1}{2\pi}\int_{-\frac{\pi}{T}}^{\frac{\pi}{T}} T\widehat{X}(\Omega T)e^{j\Omega t}d\Omega \\ &= \sum_{k=-\infty}^{\infty} \int_{-24\pi}^{24\pi} \delta\left(\Omega - \Omega_o + k48\pi\right) e^{j\Omega t} d\Omega . \end{aligned}$$

Let's consider the values of Ω_o corresponding to three of the different wheel speeds we considered earlier. In Case 1, the wheel was rotating clockwise 24 times per second when filmed, so $\Omega_o = 48\pi$. The impulses under the integral sign are situated at Ω-values 0, $\pm 48\pi$, $\pm 96\pi$, etc. Only one of these impulses — the one at $\Omega = 0$ — lies in the interval of integration, which means that

$$x_R(t) = \int_{-24\pi}^{24\pi} \delta(\Omega)e^{j\Omega t}dt = 1 \text{ for all } t \in \mathbb{R} .$$

In other words, the wheel looks as if it were standing still.

In Case 2, where the wheel turned 18 times per second, $\Omega_o = 36\pi$, so

$$x_R(t) = \int_{-24\pi}^{24\pi} \delta\left(\Omega - 36\pi + k48\pi\right) e^{j\Omega t} d\Omega .$$

The impulses under the integral sign lie at Ω-values -60π, -12π, 36π, 84π, etc. Again, only one of these impulses — the one at $\Omega = -12\pi$ — lies in the interval of integration, and

$$x_R(t) = \int_{-24\pi}^{24\pi} \delta(\Omega + 12\pi)e^{j\Omega t}dt = e^{-j12\pi t} \text{ for all } t \in \mathbb{R} .$$

In this case x_R corresponds to counterclockwise rotation at frequency $2\pi \times 6$, i.e. six revolutions per second. The wheel looks as if it were rotating "backward."

Skipping over the annoying borderline Case 3, which leads to indecision on your part when you're watching the movie, consider Case 4, where the continuous-time wheel spins six times clockwise every second, corresponding to $\Omega_o = 12\pi$ and

$$x_R(t) = \int_{-24\pi}^{24\pi} \delta\left(\Omega - 12\pi + k48\pi\right) e^{j\Omega t} d\Omega \ .$$

This time the impulses occur at Ω-values -36π, 12π, 60π, 108π, etc., and the only one under the integral sign sits at 12π, so

$$x_R(t) = \int_{-24\pi}^{24\pi} \delta(\Omega - 12\pi) e^{j\Omega t} dt = e^{j12\pi t} = x_c(t) \ \text{ for all } \ t \in \mathbb{R} \ .$$

Only in this case does your mind reconstruct x_c when you watch the movie and allow your visual apparatus implement its interpolation ritual.

Mathematically, these results harmonize with the Sampling Theorem. In each case, the sampling frequency is 48π and the bandwidth of the continuous-time signal x_c is Ω_o, so its Nyquist rate is $2\Omega_o$. The Nyquist rate for x_c in Case 1 is 96π, so the sampling frequency does not exceed the Nyquist rate. The same is true in Case 2, where $\Omega_o = 36\pi$ and the Nyquist rate for x_c is 72π. In Case 4, $\Omega_o = 12\pi$, so the Nyquist rate for x_c is 24π, and the sampling frequency exceeds the Nyquist rate, so we expect the interpolation procedure embodied in $\mathcal{R}1$ to yield $x_R = x_c$, and indeed it does.

If the math doesn't surprise us, what about the implications for cognitive science? Why do we settle on x_R as an explanation for what we observe as we watch the movie? I would submit that even if we saw the frame sequence from Case 2 in slow motion we would think of x_R — that is, counterclockwise rotation of one quarter turn between frames — before considering other viable options such as clockwise rotation at three quarters of a turn between frames. It would seem as if we were seeking the most parsimonious explanation for the data. Perhaps natural selection tends to favor sentient beings that find, and settle quickly on, simple explanations for observed phenomena. If you wake up in the middle of the night and see a lion prowling near your wilderness campsite, you'll probably think "Lion!" before considering the possibility that someone has dressed up in a lion suit to scare you. Deciding quickly on the simple explanation might make the difference between survival and the alternative. A more subtle question in the present context is why x_R, which you could construe as the "lowest-bandwidth signal" connecting the dots in the sample sequence, might explain a sample sequence "most parsimoniously" in settings more general than movies of rotating wheels.

Now for one last bit of terminology that streamlines discussions of sampling and interpolation involving pure sinusoids such as rotating wheels. When a continuous-time signal x_c is a pure sinusoid with specification $x_c(t) = e^{j\Omega_o t}$ and you sample it every T seconds, you get a discrete-time signal x that could have arisen from T-sampling any number of other continuous-time signals, among which are pure sinusoids of frequencies different from Ω_o. For any $k \in \mathbb{Z}$, the signal $x_{c,k}$ with specification

$$x_{c,k}(t) = e^{j(\Omega_o + k2\pi/T)t} \ \text{ for all } \ t \in \mathbb{R} \ ,$$

which has fundamental frequency $\Omega_k = \Omega_o + k2\pi/T$, satisfies

$$x_{c,k}(nT) = x_c(nT) = x(n) \ \text{ for all } \ n \in \mathbb{Z} \ .$$

Let's call the signals $x_{c,k}$ the *continuous-time aliases of x_c with respect to the sampling interval T.* When Ω_o isn't an odd multiple of π, exactly one of the Ω_k, say Ω_{k^*}, satisfies $-\pi < \Omega_{k^*} < \pi$. We call the signal x_{c,k^*} the *principal continuous-time alias of x_c with respect to the sampling interval T.* If you sample a pure sinusoid x_c every T seconds and obtain the discrete-time signal x, the continuous-time signal x_R that emerges from a sinc-function T-interpolator driven by x is the principal continuous-time alias of x_c with respect to the sampling interval T. As one noteworthy consequence, if a discrete-time signal x could have come from T-sampling a pure sinusoid, then the sinc-function T-interpolation of x is a sinusoid.

Pulse sampling and time-division multiplexing

Given $T > 0$ and $a > 0$ with $a < T/2$, let Π be the pulse train with specification

$$\Pi(t) = \sum_{n=-\infty}^{\infty} p_a(t - nT) \ \text{ for all } \ t \in \mathbb{R} \ .$$

If $x_c \in \mathbb{C}^{\mathbb{R}}$ is an arbitrary signal, we can form another signal $z = x_c\Pi$, which has specification

$$\begin{aligned} z(t) &= \sum_{n=-\infty}^{\infty} x_c(t)p_a(t - nT) \\ &= \begin{cases} x_c(t) & \text{when } \ nT - a/2 \le t < nT + a/2 \ \text{ and } \ n \in \mathbb{Z} \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The signal z is called a *pulse-sampled version* of x_c. It comprises a lot of little pieces of x_c, each centered on an integer multiple of T, with zero in between. Ostensibly z contains more information about x_c than a sampled record $\{x_c(nT) : n \in \mathbb{Z}\}$ contains. In fact, that sampled record is embedded in z, so by the Sampling Theorem 11.6 we could conceivably reconstruct x_c from z by extracting the sampled record provided x_c is bandlimited within π/T.

As it happens, we can actually reconstruct x_c from z in a more straightforward fashion when x_c is suitably bandlimited. The Fourier series for Π is

$$\Pi(t) = \sum_{k=-\infty}^{\infty} c_k e^{jk\frac{2\pi}{T}t} \ \text{ for all } \ t \in \mathbb{R} \ ,$$

and you can check that $c_0 = \frac{a}{T}$. Since $z = x_c\Pi$, we have

$$z(t) = \sum_{k=-\infty}^{\infty} c_k x_c(t) e^{jk\frac{2\pi}{T}t} \ \text{ for all } \ t \in \mathbb{R} \ ,$$

and the Frequency-shift rule 10.6 for continuous-time Fourier transforms implies that

$$\widehat{Z}(\Omega) = \sum_{k=-\infty}^{\infty} c_k \widehat{X}_c\left(\Omega - k\frac{2\pi}{T}\right) \ .$$

If x_c is bandlimited and T small enough — i.e. if $\widehat{X}_c(\Omega) = 0$ when $|\Omega| \geq \frac{\pi}{T}$ — then the scaled shifted replicas of $\widehat{X}_c$ that form $\widehat{Z}$ don't collide. In particular,

$$\widehat{Z}(\Omega) = \frac{a}{T}\widehat{X}_c(\Omega) \text{ when } -\frac{\pi}{T} \leq \Omega \leq \frac{\pi}{T} \; .$$

If we send z through the ideal low-pass filter whose frequency response $\widehat{H}$ has specification

$$\widehat{H}(\Omega) = \begin{cases} T/a & \text{when } |\Omega| \leq \pi/T \\ 0 & \text{otherwise,} \end{cases}$$

then, since $c_0 = a/T$, what comes out of the filter is exactly x_c. Note that the pulse sampling and subsequent recovery by low-pass filtering take place entirely in the world of continuous-time signals and systems even though the enabling spectral ideas are identical to those underlying the Sampling Theorem.

The signal z contains a lot of dead space between successive pulse samples of x_c when a/T is small. We can fill that space with pulse-sampled versions of other signals and thereby solve the following engineering problem: given N signals x_{c1}, $\ldots$, x_{cN}, with each x_{cm} bandlimited to within π/T, devise a way to encode the information from all the signals into a single signal z from which we can recover all N signals, at least in principle. Here's how it works. Choose $a > 0$ so a is much smaller than T/N. Let z be the signal with specification

$$z(t) = \sum_{m=1}^{N} x_{cm}(t)\Pi\left(t - (m-1)\frac{T}{N}\right) \text{ for all } t \in \mathbb{R} \; .$$

The signal z comprises interleaved pulse-sampled versions of the x_{cm}. The pulse-sampled version of x_{cm} embedded in z is shifted by $(m-1)T/n$ for each m. Imagine forming z by using a switch to cycle repeatedly through the signals x_{c1}, $\ldots$, x_{cN}, the switch dwelling for time a on x_{cm} every time it lands on x_{cm}. You can unpack the signal z and recover a pulse-sampled version of each x_{cm}, which you can then pass through an ideal low-pass filter to obtain x_{cm} itself.

The technique I've just described is called *time-division multiplexing.* It solves a multiple-access problem just as frequency-division multiplexing solves another multiple-access problem. Think of N agents, agent m producing signal x_{cm}, all of whom want to transmit their signals simultaneously over a channel that acts like the identity or pure t_1-shift system. The agents share the channel by dividing up time just as the agents participating in frequency-division multiplexing divide up channel bandwidth. It would seem that by choosing a small enough you could pack as many bandlimited signals as you wanted into a multiplexed signal such as z. Practical limitations include the fact that the channels over which you will want to transmit z won't be ideal t_1-shift systems, so z won't pass through unscathed. Smaller a means smaller inter-sample time lags in z and lower energy in each individual signal pulse sample. These disadvantageous features make it more likely that narrow pulse samples of the x_{cm} will get swamped by noise or mashed together and polluted beyond recognition by the distortion the channel introduces. And even if you have access to a clean version of z, you need ideal low-pass filters to reconstruct the x_{cm} from their pulse-sampled versions.

Practical difficulties aside, time-division multiplexing and its variants feature prominently in a variety of modern telecommunications systems. The primitive scheme I've presented here, while only the tip of the iceberg, illustrates in frequency-domain terms how sampling and reconstruction procedures actually work. The DTFT and the Sampling Theorem provide the mathematical underpinning, and it's fair to say that the telecommunications revolution of the last half-century would never have gained traction without them.

CHAPTER 12

The DFT and the FFT

Real-world discrete-time signal processing deals exclusively with finite-duration signals. Infinite-duration signals are useful mathematically, but nobody has ever watched such a signal play out in its entirety. When addressing signals and systems problems involving infinite-duration signals, one's goal is often to generate useful approximate results by manipulating finite-duration signals in a computationally efficient way. The DFT and FFT are tools that facilitate such manipulations. In essence, the DFT reduces a variety of signal-processing calculations to finite-dimensional linear algebra, and the FFT serves as an efficient procedure for computing the DFT. For ease of exposition, I'll continue to assume that all the signals we encounter are complex-valued. This doesn't cost us any generality, since a real-valued signal is just a special kind of complex-valued signal.

N-point signals, cyclic shifting, and circular convolution

Given a positive natural number N, an *N-point signal* is a discrete-time signal $x \in \mathbb{C}^{\mathbb{Z}}$ satisfying $x(n) = 0$ for $n < 0$ and for $n \geq N$. Thus an N-point signal is simply a finite-duration signal x whose "duration interval" is contained in the range $0 \leq n < N$. Observe that $x(n)$ could still be zero for some n-values in that range. In particular, when $M > N$, we can regard any N-point signal x as an M-point signal that just happens to satisfy $x(n) = 0$ for $N \leq n < M$. I'll denote by $\mathcal{S}_N$ the set of all N-point signals. Observe that any $x \in \mathcal{S}_N$ is specified completely by N numbers $x(0)$, $x(1)$, $\ldots$, $x(N-1)$, so the set of all N-point signals stands in one-to-one correspondence with $\mathbb{C}^N$, the set of all N-vectors with entries in $\mathbb{C}$. If x is an N-point signal, I'll denote by $\underline{x}$ the column N-vector

$$\begin{bmatrix} x(0) & x(1) & x(2) & . & . & . & x(N-1) \end{bmatrix}^T .$$

The set $\mathcal{S}_N$ is closed under the taking of linear combinations in $\mathbb{C}^{\mathbb{Z}}$, so it forms a vector space under the usual vector operations on $\mathbb{C}^{\mathbb{Z}}$. It's easy to check that those vector operations map nicely to the usual vector operations on $\mathbb{C}^N$ under the signal-vector correspondence in the sense that if $\underline{x}$ and $\underline{y}$ are the vectors corresponding to N-point signals x and y, then for any c_1 and c_2 in $\mathbb{C}$ the vector $c_1\underline{x}+c_2\underline{y}$ corresponds to the N-point signal $c_1x + c_2y$. Furthermore, since every N-point signal is square-summable, $\mathcal{S}_N$ inherits from l^2 the inner product we studied in Chapters 5 and 9. For x and y in $\mathcal{S}_N$,

$$\langle x, y\rangle = \sum_{n=-\infty}^{\infty} x(n)\overline{y(n)} = \sum_{n=0}^{N-1} x(n)\overline{y(n)} \ ,$$

where the last equality holds because x and y are N-point signals. Happily, the inner product on $\mathcal{S}_N$ maps under the signal-vector correspondence to the standard inner product on $\mathbb{C}^N$ in the sense that

$$\langle x, y\rangle = \underline{y}^H \underline{x} \ \text{ for all } \ x \text{ and } y \text{ in } \mathcal{S}_N \ .$$

In what follows I'll make considerable use of the notation $\langle\!\langle l \rangle\!\rangle_N$ for $l \in \mathbb{Z}$ and nonzero $N \in \mathbb{N}$, which we first met in Chapter 2. Read that notation as "l mod N." It denotes the unique natural number m such that $0 \leq m < N$ and $m = pN + l$ for some $p \in \mathbb{Z}$. If $l \geq 0$, $\langle\!\langle l \rangle\!\rangle_N$ is the remainder you obtain after dividing l by N. For example, $\langle\!\langle 7 \rangle\!\rangle_5 = 2$ whereas $\langle\!\langle 15 \rangle\!\rangle_5 = 0$. Note that $\langle\!\langle l \rangle\!\rangle_N = l$ if and only if $0 \leq l < N$. When $l < 0$, add N to l repeatedly and $\langle\!\langle l \rangle\!\rangle_N$ will be the first number m you obtain that satisfies $0 \leq m < N$. For example, $\langle\!\langle -3 \rangle\!\rangle_7 = 4$ and $\langle\!\langle -13 \rangle\!\rangle_5 = 2$. I'll also make a habit of specifying an N-point signal x only on the time interval $0 \leq n < N$, with the understanding that $x(n) = 0$ for all other $n \in \mathbb{Z}$.

Given a positive integer N, an N-point signal x, and any n_o with $0 \leq n_o < N$, the *cyclic shift of x by n_o* is the N-point signal $\text{CShift}_{n_o}(x)$ with specification

$$\begin{aligned} \text{CShift}_{n_o}(x)(n) &= x(\,\langle\!\langle n - n_o \rangle\!\rangle_N\,) \\ &= \begin{cases} x(n - n_o) & \text{when } n_o \leq n < N \\ x(N + n - n_o) & \text{when } 0 \leq n < n_o \ . \end{cases} \end{aligned}$$

It's clear what cyclic shifting does to an N-point signal. If you think of the signal as a left-to-right ordered list of its values at times $0 \leq n < N$, cyclic shifting by n_o slides all the entries in the list to the right by n_o, then takes the ones that "fall off the edge of the page" and splices them back in their original order at the left end of the list.

Cyclic shifting is obviously a linear operation on the set $\mathcal{S}_N$ of N-point signals. In terms of the vector representation of N-point signals, we can regard cyclic shifting of an N-point signal as multiplying the signal vector on the left with a special matrix. Define the $(N \times N)$ matrix $\underline{C}_N$ by

$$\underline{C}_N = \begin{bmatrix} 0 & 0 & 0 & . & . & . & 0 & 1 \\ 1 & 0 & 0 & 0 & . & . & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & . & . & 0 \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ 0 & . & . & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & . & . & 1 & 0 & 0 \\ 0 & 0 & . & . & . & 0 & 1 & 0 \end{bmatrix} .$$

Observe that $[\underline{C}_N]_{pq} = 1$ if and only if $\langle\!\langle p - q \rangle\!\rangle_N = 1$. I leave it for you to verify that

$$\underline{\text{CShift}_1(x)} = \underline{C}_N \underline{x} \ \text{ for all } \ x \in \mathcal{S}_N \ .$$

Since cyclic shifting by n_o is the same as cyclic shifting by 1 performed n_o times in succession, it follows that when $0 \leq n_o < N$ we have

$$\underline{\text{CShift}_{n_o}(x)} = \underline{C}_N^{n_o} \underline{x} \ \text{ for all } \ x \in \mathcal{S}_N \ ,$$

where $\underline{C}_N^{n_o}$ is the matrix product of $\underline{C}_N$ with itself n_o times, and by convention $\underline{C}_N^0 = I_N$, the $(N \times N)$ identity matrix.

Given a positive integer N and N-point signals h and x, the *N-point circular convolution of h and x* is the N-point signal $\mathrm{CConv}_N(h,x)$ with specification

$$\mathrm{CConv}_N(h,x)(n) = \sum_{m=0}^{N-1} h(m)x(\,\langle\!\langle n-m\rangle\!\rangle_N\,) \ \text{ for } \ 0 \le n < N \ .$$

Like ordinary convolution, circular convolution is a commutative operation in the sense that $\mathrm{CConv}_N(h,x) = \mathrm{CConv}_N(x,h)$. To see this, observe that

$$\begin{aligned}\mathrm{CConv}_N(h,x)(n) &= \sum_{m=0}^{N-1} h(m)x(\,\langle\!\langle n-m\rangle\!\rangle_N\,) \\ &= \sum_{m=0}^{n} h(m)x(n-m) + \sum_{m=n+1}^{N-1} h(m)x(N+n-m) \\ &= \sum_{l=0}^{n} h(n-l)x(l) + \sum_{l=n+1}^{N-1} h(N+n-l)x(l) \\ &= \sum_{l=0}^{N-1} x(l)h(\,\langle\!\langle n-l\rangle\!\rangle_N\,) \\ &= \mathrm{CConv}(x,h)(n) \text{ for } 0 \le n < N \ .\end{aligned}$$

On the third line, I changed indices of summation to $l = n - m$ in the first sum and $l = N + n - m$ in the second sum.

As with cyclic shifting, we can represent circular convolution using matrices and vectors associated with signals. Note first that from the identity

$$\mathrm{CConv}_N(h,x)(n) = \sum_{m=0}^{N-1} x(m)h(\,\langle\!\langle n-m\rangle\!\rangle_N\,) \ \text{ for } \ 0 \le n < N$$

along with the fact that for each m

$$h(\,\langle\!\langle n-m\rangle\!\rangle_N\,) = \mathrm{CShift}_m(h)(n) \ \text{ for } \ 0 \le n < N$$

it follows that

$$\mathrm{CConv}_N(h,x) = \sum_{m=0}^{N-1} x(m)\mathrm{CShift}_m(h) \ ,$$

where the terms on either side of the last equation are whole N-point signals. Think of the $x(m)$-values on the right-hand side as coefficients in a linear combination of cyclically shifted h's. Now define $\underline{H}$ as the $(N \times N)$ matrix whose qth column is the signal vector for $\mathrm{CShift}_{q-1}(h)$. If $N = 4$, for example,

$$\underline{H} = \begin{bmatrix} h(0) & h(3) & h(2) & h(1) \\ h(1) & h(0) & h(3) & h(2) \\ h(2) & h(1) & h(0) & h(3) \\ h(3) & h(2) & h(1) & h(0) \end{bmatrix} .$$

Then

$$\underline{\mathrm{CConv}_N(h,x)} = \underline{H}\underline{x} \ \text{ for all } \ h,x \text{ in } \mathcal{S}_N \ . \tag{13}$$

A matrix such as $\underline{H}$, whose qth column is $\underline{C}_N^{q-1}$ times its first column for each q, is called a circulant matrix. Observe that $\underline{C}_N$ itself is a circulant matrix.

The N-point DFT of an N-point signal

An N-point signal has finite duration, so it has a DTFT. Let x be an N-point signal and let $\widehat{X}$ be its DTFT, i.e.

$$\widehat{X}(\omega) = \sum_{n=-\infty}^{\infty} x(n)e^{-jn\omega} = \sum_{n=0}^{N-1} x(n)e^{-jn\omega} \ \text{ for all } \ \omega \in \mathbb{R} \ .$$

Consider "sampling" $\widehat{X}$ at the N equally spaced ω-values $k2\pi/N$, where k ranges from 0 to $N-1$. You get

$$\widehat{X}\left(k\frac{2\pi}{N}\right) = \sum_{n=0}^{N-1} x(n)e^{-jnk\frac{2\pi}{N}} \ \text{ for } \ 0 \le k < N \ .$$

For notational convenience, set

$$\psi_N = e^{j\frac{2\pi}{N}} \ \text{ and } \ \widehat{X}_k = \widehat{X}\left(k\frac{2\pi}{N}\right) \ \text{ for } \ 0 \le k < N \ .$$

The formula for $\widehat{X}_k$ in terms of x becomes

$$(\mathcal{DFT}) \qquad \widehat{X}_k = \sum_{n=0}^{N-1} x(n)\psi_N^{-nk} \ \text{ for } \ 0 \le k < N \ .$$

12.1 Definition: The N*-point DFT* of the N-point signal x is the ordered N-tuple of complex numbers $\widehat{X}_0$, $\widehat{X}_1$, ... , $\widehat{X}_{N-1}$ given by equation $\mathcal{DFT}$.

The initials "DFT" stand for "discrete Fourier transform," but everyone just says "DFT." If you form a column N-vector $\underline{\widehat{X}}$ with the $\widehat{X}_k$ as entries, i.e.

$$\underline{\widehat{X}} = \left[\begin{array}{cccccc} \widehat{X}_0 & \widehat{X}_1 & \widehat{X}_2 & . & . & . & \widehat{X}_{N-1} \end{array}\right]^T ,$$

then you'll find that

$$\underline{\widehat{X}} = \Psi_N \underline{x} \ ,$$

where Ψ_N is the $(N \times N)$ matrix whose (p,q)-entry is

$$[\Psi_N]_{pq} = \psi_N^{-(p-1)(q-1)} \ .$$

Here's what Ψ_N looks like.

$$\Psi_N = \left[\begin{array}{cccccc} 1 & 1 & 1 & . & . & 1 & 1 \\ 1 & \psi_N^{-1} & \psi_N^{-2} & \psi_N^{-3} & . & . & \psi_N^{-(N-1)} \\ 1 & \psi_N^{-2} & \psi_N^{-4} & \psi_N^{-6} & . & . & \psi_N^{-2(N-1)} \\ 1 & \psi_N^{-3} & \psi_N^{-6} & . & . & . & \psi_N^{-3(N-1)} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 1 & \psi_N^{-(N-1)} & \psi_N^{-2(N-1)} & . & . & . & \psi_N^{-(N-1)(N-1)} \end{array}\right] .$$

As it happens, Ψ_N is invertible. You can see this in at least two ways. First of all, Ψ_N is a so-called Vandermonde matrix. It takes the form

$$V_N(\alpha_0, \alpha_1, \dots, \alpha_{N-1}) = \begin{bmatrix} 1 & \alpha_0 & \alpha_0^2 & . & . & \alpha_0^{N-1} \\ 1 & \alpha_1 & \alpha_1^2 & . & . & \alpha_1^{N-1} \\ 1 & \alpha_2 & \alpha_2^2 & . & . & \alpha_2^{N-1} \\ . & . & . & . & . & \\ . & . & . & . & . & \\ 1 & \alpha_{N-1} & \alpha_{N-1}^2 & . & . & \alpha_{N-1}^{N-1} \end{bmatrix}$$

with $\alpha_k = \psi_N^{-k}$ for $0 \le k < N$. A nice fact about the Vandermonde matrix is that it's invertible if and only if all the α_k are different, which is the case for Ψ_N. The Vandermonde argument proves Ψ_N's invertibility, but it doesn't provide a formula for the inverse of Ψ_N. It turns out that

$$\Psi_N^{-1} = (1/N)\overline{\Psi_N} \ ,$$

where overbar denotes complex conjugate. The argument rests on the identity

$$\psi_N^{lN} = 1 \ \text{ for all } \ l \in \mathbb{Z} \ , \tag{14}$$

which holds because $\psi_N^{lN} = e^{j2\pi l} = 1$ for every integer l. Observe that

$$\begin{aligned} \left[(1/N)\overline{\Psi_N}\Psi_N\right]_{pq} &= \frac{1}{N}\sum_{r=1}^{N} \left[\overline{\Psi_N}\right]_{pr} \left[\Psi_N\right]_{rq} \\ &= \frac{1}{N}\sum_{r=1}^{N} \psi_N^{(p-1)(r-1)}\psi_N^{-(r-1)(q-1)} \\ &= \frac{1}{N}\sum_{r=1}^{N} \psi_N^{(p-q)(r-1)} \end{aligned}$$

for $1 \le p, q \le N$. If $p = q$, each term in the sum is 1, so the value of the right-hand side is 1. If $p \ne q$, the sum is a partial geometric series and evaluates to

$$\sum_{m=0}^{N-1} \left(\psi_N^{p-q}\right)^m = \frac{1 - \psi_N^{(p-q)N}}{1 - \psi_N^{p-q}} = 0$$

by identity (14). Accordingly,

$$\left[\left(\frac{1}{N}\overline{\Psi_N}\right)\Psi_N\right]_{pq} = \begin{cases} 1 & \text{when } p = q \\ 0 & \text{when } p \ne q \ , \end{cases}$$

which is the same as saying

$$\left(\frac{1}{N}\overline{\Psi_N}\right)\Psi_N = I_N$$

or, equivalently,

$$\Psi_N^{-1} = \frac{1}{N}\overline{\Psi_N} \ .$$

Since Ψ_N is invertible, you can recover an N-point signal from its N-point DFT. In terms of the vectors $\underline{x}$ and $\underline{\widehat{X}}$,

$$\underline{x} = \Psi_N^{-1}\underline{\widehat{X}} = \frac{1}{N}\overline{\Psi_N}\underline{\widehat{X}} \ .$$

In components, this last equation reads

$$(\mathcal{DFT}^{-1}) \qquad x(n) = \frac{1}{N}\sum_{k=0}^{N-1} \widehat{X}_k \psi_N^{nk} \text{ for } 0 \le n < N \,.$$

Thus N-point signals and their N-point DFTs stand in one-to-one correspondence. The signal determines the DFT through equation $\mathcal{DFT}$ and the DFT determines the signal through equation $\mathcal{DFT}^{-1}$. While this observation might not seem profound, it has a noteworthy consequence that reads like a "reverse sampling theorem."

12.2 Theorem: The DTFT $\widehat{X}$ of any $x \in \mathcal{S}_N$ is determined for all $\omega \in \mathbb{R}$ by its values at the N equally spaced ω-values 0, $2\pi/N$, $4\pi/N$, $\ldots$, $(N-1)2\pi/N$. Alternatively, given any N complex numbers c_0, c_1, c_2, $\ldots$, c_{N-1}, there exists a unique $x \in \mathcal{S}_N$ whose DTFT $\widehat{X}$ satisfies $\widehat{X}(k2\pi/N) = c_k$ for $0 \le k < N$.

Proof: The first statement follows from the fact that the N values $\widehat{X}(k2\pi/N)$, $0 \le k < N$, determine the N-point signal x via equation $\mathcal{DFT}^{-1}$, and x in turn determines $\widehat{X}$. As for the second statement, let x be the N-point signal with N-point DFT $\widehat{X}_k = c_k$, $0 \le k < N$. Then x's DTFT $\widehat{X}$ satisfies $\widehat{X}(k2\pi/N) = c_k$ for $0 \le k < N$, and x is the only N-point signal with that property by equation $\mathcal{DFT}^{-1}$. □

Many authors approach the DFT from the standpoint of orthogonal expansions. Although I've adopted a different line of attack, it's useful to understand the orthogonal-expansion approach. We've noted already that $\mathcal{S}_N$ is an inner product space with the usual l^2 inner product. Since $\mathcal{S}_N$ has dimension N, it possesses orthonormal bases by Fact 9.9. The N-tuple of signals

$$(\delta, \mathrm{Shift}_1(\delta), \mathrm{Shift}_2(\delta), \ldots, \mathrm{Shift}_{N-1}(\delta))$$

is one orthonormal basis for $\mathcal{S}_N$, and it maps to the standard basis of $\mathbb{C}^N$ under the signal-vector correspondence. Another orthonormal basis for $\mathcal{S}_N$ arises as follows. For each k, $0 \le k < N$, let w_k be the N-point signal with specification

$$w_k(n) = \frac{1}{\sqrt{N}}\psi_N^{nk} \text{ for } 0 \le n < N \,.$$

The N-tuple $(w_0, w_1, w_2, \ldots, w_{N-1})$ is an orthonormal basis for $\mathcal{S}_N$. To verify orthonormality, note that given p and q we have

$$\begin{aligned} \langle w_p, w_q \rangle &= \sum_{n=0}^{N-1} w_p(n)\overline{w_q(n)} \\ &= \frac{1}{N}\sum_{n=0}^{N-1} \psi_N^{n(p-q)} \,. \end{aligned}$$

If $p = q$, every term in the sum on the second line is 1, so $\langle w_p, w_q \rangle = 1$. If $p \neq q$, the sum takes the form of a partial geometric series and evaluates to

$$\sum_{n=0}^{N-1} \left(\psi_N^{p-q}\right)^n = \frac{1 - \psi_N^{N(p-q)}}{1 - \psi_N^{p-q}} = 0$$

by identity (14). So $\langle w_p, w_q \rangle = 0$ if $p \neq q$. Since the w_k are orthonormal, they're linearly independent and hence form a basis for $\mathcal{S}_N$. Consequently, every $x \in \mathcal{S}_N$ has an orthogonal expansion

$$x = \sum_{k=0}^{N-1} \langle x, w_k \rangle \, w_k \ .$$

Because

$$\begin{aligned} \langle x, w_k \rangle &= \sum_{n=0}^{N-1} x(n)\overline{w_k(n)} \\ &= \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x(n)\psi_n^{-nk} \\ &= \frac{1}{\sqrt{N}} \widehat{X}_k \ \text{ for } \ 0 \leq k < N \ , \end{aligned}$$

the orthogonal expansion for any N-point signal x is

$$x = \sum_{k=0}^{N-1} \left(\frac{1}{\sqrt{N}} \widehat{X}_k \right) w_k \ , \tag{15}$$

which is equivalent to

$$x(n) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \widehat{X}_k w_k(n) = \frac{1}{N} \sum_{k=0}^{N-1} \widehat{X}_k \psi_N^{nk} \ \text{ for } \ 0 \leq n < N \ ,$$

and this is simply equation $\mathcal{DFT}^{-1}$. In short, taking the N-point DFT of an N-point signal x amounts to finding the coefficients in the orthogonal expansion of x in terms of the orthonormal basis $(w_0, \ldots, w_{N-1})$ for $\mathcal{S}_N$, and equation $\mathcal{DFT}^{-1}$ exhibits the orthogonal expansion itself.

Let's re-cast these orthogonal-expansion ideas in terms of the linear algebra of signal vectors. Recall that the inner product on $\mathcal{S}_N$ maps to

$$\langle \underline{x}, \underline{y} \rangle = \underline{y}^H \underline{x} \ ,$$

the usual inner product on $\mathbb{C}^N$. The signal vectors $\underline{w}_k$ corresponding to the signals w_k above constitute an orthonormal basis for $\mathbb{C}^N$ with respect to this inner product. Thus we can expand any $\underline{x} \in \mathbb{C}^N$ as

$$\begin{aligned} \underline{x} &= \sum_{k=0}^{N-1} \langle \underline{x}, \underline{w}_k \rangle \, \underline{w}_k \\ &= \sum_{k=0}^{N-1} \left(\underline{w}_k{}^H x \right) w_k \\ &= \sum_{k=0}^{N-1} \left(\frac{1}{\sqrt{N}} \widehat{X}_k \right) \underline{w}_k \ , \end{aligned}$$

which is the vector version of (15).

Operational rules and a symmetry property

Associated with the DFT are operational rules analogous to those for the DTFT. Most important to us are two that I'll state and prove first using the language of N-point signals and then reformulate in terms of signal vectors. Before moving on to those rules, let's record a simple observation about N-point DFTs of real-valued N-point signals.

12.3 Fact: If x is a real-valued N-point signal with N-point DFT $\widehat{X}_k$, $0 \leq k < N$, then $\widehat{X}_0$ is real, and

$$\widehat{X}_{N-k} = \overline{\widehat{X}_k} \text{ when} 1 \leq k < N \ .$$

The proof is simple. $\widehat{X}_0 = \sum_{n=0}^{N-1} x(n)$ is obviously real when x is. For $1 \leq k < N$ we have

$$\begin{aligned} \widehat{X}_{N-k} &= \sum_{n=0}^{N-1} x(n)\psi_N^{-n(N-k)} \\ &= \sum_{n=0}^{N-1} x(n)\psi_N^{nk} \\ &= \overline{\widehat{X}_k} \ , \end{aligned}$$

where the second line holds because $\psi_N^{-nN} = 1$ and the third line because x is real-valued. □

Now for the aforementioned operational rules.

12.4 Cyclic Shift Rule: If x is an N-point signal whose N-point DFT is $\{\widehat{X}_k : 0 \leq k < N\}$, then for any n_o satisfying $0 \leq n_o < N$, $y = \text{CShift}_{n_o}(x)$ has N-point DFT specified by

$$\widehat{Y}_k = \psi_N^{-n_o k}\widehat{X}_k \ \text{ for } \ 0 \leq k < N \ .$$

To see this, write out the formula for $\widehat{Y}_k$.

$$\begin{aligned}
\widehat{Y}_k &= \sum_{n=0}^{N-1} y(n)\psi_N^{-nk} \\
&= \sum_{n=0}^{N-1} x(\,\langle\!\langle n-n_o\rangle\!\rangle_N\,)\psi_N^{-nk} \\
&= \sum_{n=n_o}^{N-1} x(n-n_o)\psi_N^{-nk} + \sum_{n=0}^{n_o-1} x(N+n-n_o)\psi_N^{-nk} \\
&= \sum_{m=0}^{N-n_o-1} x(m)\psi_N^{-(m+n_o)k} + \sum_{m=N-n_o}^{N-1} x(m)\psi_N^{-(m-N+n_o)k} \\
&= \psi_N^{-n_o k} \sum_{m=0}^{N-1} x(m)\psi_N^{-mk} \\
&= \psi_N^{-n_o k}\widehat{X}_k\ .
\end{aligned}$$

On the fourth line, I changed summation index to $m = n - n_o$ in the first sum and $m = N + n - n_o$ in the second sum. To get the fifth line from the fourth line, I used

$$\psi_N^{-(m-N+n_o)k} = \psi_N^{-(m+n_o)k}\psi_N^{Nk} = \psi_N^{-(m+n_o)k}\ ,$$

where $\psi_N^{Nk} = 1$ follows from identity (14). □

12.5 Circular Convolution Rule: If h and x are N-point signals with respective N-point DFTs $\{\widehat{H}_k : 0 \le k < N\}$ and $\{\widehat{X}_k : 0 \le k < N\}$, then $y = \mathrm{CConv}_N(h,x)$ has DFT specified by

$$\widehat{Y}_k = \widehat{H}_k\,\widehat{X}_k \ \text{ for } \ 0 \le k < N\ .$$

Again it pays to plug and chug starting with the definition of $\widehat{Y}_k$.

$$\begin{aligned}
\widehat{Y}_k &= \sum_{n=0}^{N-1} y(n)\psi_N^{-nk} \\
&= \sum_{n=0}^{N-1}\left(\sum_{m=0}^{N-1} h(m)x(\,\langle\!\langle n-m\rangle\!\rangle_N\,)\right)\psi_N^{-nk} \\
&= \sum_{m=0}^{N-1} h(m)\left(\sum_{n=0}^{N-1} x(\,\langle\!\langle n-m\rangle\!\rangle_N\,)\psi_N^{-nk}\right) \\
&= \left(\sum_{m=0}^{N-1} h(m)\psi_N^{-mk}\right)\widehat{H}_k \\
&= \widehat{H}_k\,\widehat{X}_k\ .
\end{aligned}$$

To get the third line from the second line, I interchanged order of summation. To get the fourth line from the third line, I applied the Cyclic Shift Rule to the inner sum on the third line. □

Reformulating the Cyclic Shift Rule and Circular Convolution Rule in terms of signal vectors unearths a significant and widely applicable linear-algebraic fact. Let's start with the signal-vector rendering of $y = \mathrm{CShift}_{n_o}(x)$, which reads

$$\underline{y} = \underline{C}_N^{n_o}\, \underline{x} \ .$$

Multiplying a vector by Ψ_N produces the vector of values of the N-point DFT of the vector's corresponding signal. Thus from

$$\Psi_N\, \underline{y} = \Psi_N\, \underline{C}_N^{n_o}\, \underline{x} = \left(\Psi_N\, \underline{C}^{n_o}\, \Psi_N^{-1}\right)\, \Psi_N\, \underline{x} \ \text{ for all } \ \underline{x} \in \mathbb{C}^N$$

it follows that

$$\underline{\widehat{Y}} = \left(\Psi_N\, \underline{C}_N\, \Psi_N^{-1}\right)\, \underline{\widehat{X}} \ \text{ for all } \ \underline{\widehat{X}} \in \mathbb{C}^N \ .$$

By the Cyclic Shift Rule, we have

$$\widehat{Y}_k = \psi_N^{-n_o k} \widehat{X}_k \ \text{ for } \ 0 \le k < N \ ,$$

which means that the matrix in parentheses must be diagonal. Specifically,

$$\left[\Psi_N\, \underline{C}_N\, \Psi_N^{-1}\right]_{pq} = \begin{cases} \psi_N^{n_o(p-1)} & \text{if } p = q \\ 0 & \text{if } p \ne q \ . \end{cases}$$

In the terminology of Chapter 14, taking the N-point DFT diagonalizes the linear mapping on $\mathbb{C}^N$ associated with cyclic shifting. That linear mapping arises from multiplication of vectors by a circulant matrix, in this case a power of $\underline{C}_N$. As it happens, taking the DFT diagonalizes any linear operation on $\mathbb{C}^N$ induced in similar fashion by a circulant matrix.

The Circular Convolution Rule also corresponds to such an operation on signal vectors. Start with (13) and multiply both sides by Ψ_N. That yields

$$\Psi_N\, \underline{y} = \Psi_N\, \underline{H}\, \underline{x} = \left(\Psi_N\, \underline{H}\, \Psi_N^{-1}\right)\, \Psi_N\, \underline{x} \ \text{ for all } \ \underline{x} \in \mathbb{C}^N \ ,$$

meaning

$$\underline{\widehat{Y}} = \left(\Psi_N\, \underline{H}\, \Psi_N^{-1}\right)\, \underline{\widehat{X}} \ \text{ for all } \ \underline{\widehat{X}} \in \mathbb{C}^N \ .$$

The Circular Convolution Rule stipulates that $\widehat{Y}_k = \widehat{H}_k \widehat{X}_k$ for all k, so the matrix in parentheses must be the diagonal matrix whose pth diagonal entry is $\widehat{H}_{p-1}$ for $1 \le p \le N$. Note that since h is an arbitrary N-point signal, $\underline{H}$ is an arbitrary circulant matrix. It follows that the DFT matrix Ψ_N diagonalizes any circulant matrix. Proving the following formal elaboration of this last observation requires material from Chapter 14.

12.6 Fact: Let $\underline{H}$ he an $(N \times N)$ circulant matrix whose first column is

$$\underline{h} = \left[\ h(0) \quad h(1) \quad h(2) \quad . \quad . \quad . \quad h(N-1) \ \right]^T$$

and whose qth column is $\underline{C}_N^{q-1} \underline{h}$ for $2 \le q \le N$. The eigenvalues of $\underline{H}$ are

$$\widehat{H}_k = \sum_{n=0}^{N-1} h(n) \psi_N^{-nk} \ \text{ for } \ 0 \le k < N \ ,$$

and for each k an eigenvector of $\underline{H}$ corresponding to eigenvalue $\widehat{H}_k$ is

$$\underline{w}_k = \left[\ 1 \quad \psi_N^k \quad \psi_N^{2k} \quad . \quad . \quad . \quad \psi_N^{(N-1)k} \ \right]^T \ .$$

Three applications of the DFT

People have developed extremely fast algorithms for computing and inverting DFTs. If you can reduce a signal-processing calculation to one involving only DFT computations, chances are you'll be able to do your calculation more efficiently than by "brute force." In what follows, I'll describe three such applications of DFTs.

First consider the problem of finding the ordinary convolution $y = h * x$ of two finite-duration signals h and x. Suppose for simplicity that $h(n) = x(n) = 0$ for $n < 0$. If that's not the case, you can always find n_1 and n_2 so that $\mathrm{Shift}_{n_1}(x)(n) = \mathrm{Shift}_{n_2}(x)(n) = 0$ for $n < 0$, then compute $\mathrm{Shift}_{n_1}(h) * \mathrm{Shift}_{n_2}(x)$, and finally note that

$$y = h * x = \mathrm{Shift}_{-n_1-n_2}\left(\mathrm{Shift}_{n_1}(h) * \mathrm{Shift}_{n_2}(x)\right) \ .$$

Since h and x have finite duration, there exist positive integers P and L so that $h(n) = 0$ for $n \geq P$ and $x(n) = 0$ for $n \geq L$. Thus h is a P-point signal and x is an L-point signal. The discussion of Criterion 5.1 taught us that $h * x(n) = 0$ for $n < 0$ and for $n \geq P + L - 1$, so $h = h * x$ is an N-point signal with $N = P + L - 1$.

We can also think of h and x themselves as N-point signals that just happen to be zero for $n \geq P$ and $n \geq L$, respectively. Adopting that view of h and x, we can find their N-point circular convolution, and we can do that using N-point DFTs as follows.

- Find the N-point DFTs of h and x viewed as N-point signals, where $N = P + L - 1$, with P and L defined as above. Call these DFTs $\{\widehat{H}_k\}$ and $\{\widehat{X}_k\}$, respectively.
- Let $\widehat{Z}_k = \widehat{H}_k\,\widehat{X}_k$ for $0 \leq k < N$. Find the N-point signal z whose N-point DFT is $\{\widehat{Z}_k : 0 \leq k < N\}$. By the Circular Convolution Rule 12.5, z is the circular convolution of h and x viewed as N-point signals.

As it happens, the ordinary convolution $y = h * x$ is the same as the circular convolution $z = \mathrm{CConv}_N(h, x)$, the latter computed as above by viewing h and x as N-point signals. Let's see why. The ordinary convolution x of h and x has specification

$$\begin{aligned} y(n) &= \sum_{m=-\infty}^{\infty} h(m)x(n-m) \\ &= \sum_{m=0}^{n} h(m)x(n-m) \\ &= \begin{cases} 0 & \text{when } n < 0 \\ \sum_{m=0}^{n} h(m)x(n-m) & \text{when } 0 \leq n < N \\ 0 & \text{when } n \geq N \ . \end{cases} \end{aligned}$$

The range of summation in the second line takes into account that $h(m) = 0$ when $m < 0$ and $x(n-m) = 0$ when $m > n$. The last line holds because $h(m)x(n-m) = 0$ for every m in the range of summation when $n < 0$ or when $n \geq N = P + L - 1$.

Meanwhile, the N-point signal $z = \mathrm{CConv}_N(h,x)$ has specification $z(n) = 0$ for $n < 0$, $z(n) = 0$ for $n \geq N$, and

$$\begin{aligned} z(n) &= \sum_{m=0}^{N} h(m)x(\,\langle\!\langle n-m \rangle\!\rangle_N\,) \\ &= \sum_{m=0}^{n} h(m)x(n-m) + \sum_{m=n+1}^{N} h(m)x(N+n-m) \;\text{ for }\; 0 \leq n < N\,. \end{aligned}$$

Consider the second sum on the second line. Because $h(m) = 0$ for $m \geq P$, the only possibly nonzero terms in the sum are those for which $m < P$. For those terms, $N + n - m > L - 1 + n$. Since $n \geq 0$, $N + n - m > L - 1$ for these terms, so $x_2(N + n - m) = 0$. It follows that the second sum is zero and therefore that $z(n) = y(n)$ for all n.

So the DFT helps us compute ordinary convolutions of finite-duration signals. A technique called *block convolution* extends this DFT application somewhat. Suppose we want to compute $h{*}x$ when h has finite duration but x does not. I'm using h and x for the two signals to suggest a typical context for block convolution, namely computing the output of an FIR system in response to a possibly infinite-duration input. Suppose for simplicity that $h(n) = 0$ when $n < 0$ and $h(n) = 0$ when $n \geq P$, so h is a P-point signal.

Block convolution begins by dividing x into blocks of length L. Typically, L is much larger than P, but not always — L is a "user choice." For each $r \in \mathbb{Z}$, define the L-point signal x_r as follows:

$$x_r(n) = \begin{cases} x(rL+n) & \text{for } 0 \leq n < L \\ 0 & \text{otherwise.} \end{cases}$$

The possibly nonzero values in x_r are the values of the signal x that occur in the "rth time-block of length L," which begins at time rL and extends through time $(r+1)L - 1$. Because

$$x = \sum_{r=-\infty}^{\infty} \mathrm{Shift}_{rL}(x_r)\ ,$$

we know that

$$\begin{aligned} h * x &= \sum_{r=-\infty}^{\infty} h * \mathrm{Shift}_{rL}(x_r) \\ &= \sum_{r=-\infty}^{\infty} \mathrm{Shift}_{rL}(h * x_r) \\ &= \sum_{r=-\infty}^{\infty} \mathrm{Shift}_{rL}(y_r)\ , \end{aligned}$$

where $y_r = h * x_r$. Since h is a P-point signal and each x_r is an L-point signal, we can use DFTs to compute the y_r's, which are all $(L + P - 1)$-point signals.

Finishing the computation of $h * x$ entails shifting each y_r by rL and then adding all those shifted signals together. That may seem like a tall order, but it's not difficult if P isn't too large. The key observation is that if L is significantly larger than P, then the "duration interval" of each shifted y_r overlaps the duration intervals only of its "nearest neighbors," the shifted y_{r-1} and y_{r+1}. These overlaps are all of length $P-1$, so adding all the shifted y_r's together requires that you perform batches of $P-1$ additions widely separated in time. A glance at Figure 1 might prove helpful. For example, y_0 starts at 0 and ends at $L+P-2$ and y_1 starts at L and ends at $2L+P-1$. So the overlap between the duration intervals of y_0 and y_1 extends only from L to $L+P-2$, an interval of length $P-1$.

Block convolution is particularly useful when the signal x parses into blocks that look substantially alike. For example, suppose x is periodic and L is a period of x. Then every x_r is the same L-point signal, and you need compute $y_r = h * x_r$ only once — i.e. $y_r = y_0$ for all r. The real work comes with the splicing, but note that all the splicing computations are the same because each entails splicing the left end of a a y_0 with the right end of another y_0.

A final application of DFTs plays an important role in FIR filter design, so I'll frame the discussion in those terms. Suppose you want to design an FIR filter whose frequency response is a reasonable approximation of some desired frequency response $\widehat{H}_{\text{des}}$. The desired frequency response $\widehat{H}_{\text{des}}$ might not be achievable by an FIR filter, but in any event $\widehat{H}_{\text{des}}$ is the DTFT of some possibly infinite-duration impulse response h_{des}. A popular technique for coming up with an FIR filter that, with any luck, behaves much like the filter with frequency response $\widehat{H}_{\text{des}}$ is *frequency-sampling design.* Choose an integer $N > 0$ and record the values of $\widehat{H}_{\text{des}}$ at the N equally spaced frequencies $k2\pi/N$ for $0 \leq k < N$. Then find the N-point signal h whose N-point DFT is

$$\widehat{H}_k = \widehat{H}_{\text{des}}\left(k\frac{2\pi}{N}\right) \quad \text{for} \quad 0 \leq k < N\ .$$

Note that larger N means more "frequency samples" of $\widehat{H}_{\text{des}}$ distributed more densely in the ω-interval $[0, 2\pi]$. Denote the frequency response of the FIR filter with impulse response h in the usual way as $\omega \mapsto \widehat{H}(\omega)$. By definition of the DFT, $\widehat{H}_k = \widehat{H}(k2\pi/N)$ for all k, so we will always have

$$\widehat{H}\left(k\frac{2\pi}{N}\right) = \widehat{H}_{\text{des}}\left(k\frac{2\pi}{N}\right) \quad \text{for} \quad 0 \leq k < N\ .$$

Thus $\widehat{H}$ matches $\widehat{H}_{\text{des}}$ at the frequency-sampling points, but $\widehat{H}(\omega)$ might deviate significantly from $\widehat{H}_{\text{des}}(\omega)$ at other ω-values.

Back in the time domain, one might hope that if N is large enough, then h will resemble h_{des} on the time interval $0 \leq n < N$. Suppose, for instance, that most of the "action" in h_{des} occurs on the time interval $0 \leq n < N$. An extreme example would be if h_{des} were actually an N-point signal, in which case we would have $h = h_{\text{des}}$ by Theorem 12.2. Generally, though, h_{des} has infinite duration. Let's find a formula for h in terms of h_{des} that illuminates what's happening in the time domain. By equation $\mathcal{DFT}^{-1}$,

$$\begin{aligned} h(n) &= \frac{1}{N}\sum_{k=0}^{N-1} \widehat{H}_k \psi_N^{nk} \\ &= \frac{1}{N}\sum_{k=0}^{N-1} \widehat{H}_{\text{des}}\left(k\frac{2\pi}{N}\right) e^{jnk\frac{2\pi}{N}} \quad \text{for} \quad 0 \leq n < N\ . \end{aligned}$$

Assuming equation $\mathcal{DTFT}$ holds for h_{des}, which it will, for example, if h_{des} is an l^1-signal, we have

$$\widehat{H}_{\text{des}}\left(k\frac{2\pi}{N}\right) = \sum_{m=-\infty}^{\infty} h_{\text{des}}(m)e^{-jmk\frac{2\pi}{N}} \text{ for } 0 \le k < N .$$

Accordingly,

$$\begin{aligned} h(n) &= \frac{1}{N}\sum_{k=0}^{N-1}\left(\sum_{m=-\infty}^{\infty} h_{\text{des}}(m)e^{-jmk\frac{2\pi}{N}}\right)e^{jnk\frac{2\pi}{N}} \\ &= \sum_{m=-\infty}^{\infty} h_{\text{des}}(m)\left(\frac{1}{N}\sum_{k=0}^{N-1} e^{j(n-m)k\frac{2\pi}{N}}\right) \end{aligned}$$

for $0 \le n < N$. Interchanging the order of summation is fine if h_{des} is an l^1-signal. The inner sum on the second line looks more familiar if we re-write it as

$$\sum_{k=0}^{N-1} \psi_N^{(n-m)k} .$$

If $n - m$ is an integer multiple of N, say $m = n - rN$ for some $r \in \mathbb{Z}$, all the terms in the sum are 1 and the expression in parentheses evaluates to 1. If $n - m$ is not an integer multiple of N, then $\psi_N^{n-m} \ne 1$ and the inner sum is a partial geometric series that evaluates to

$$\frac{1 - \psi_N^{(n-m)N}}{1 - \psi_N^{n-m}} = 0$$

by identity (14). Consequently, the only m-values that contribute to the outer sum are those of the form $m = n - rN$ for $r \in \mathbb{Z}$. For those m-values, the expression in parentheses evaluates to 1, and it follows that

$$h(n) = \sum_{r=-\infty}^{\infty} h_{\text{des}}(n - rN) = \sum_{r=-\infty}^{\infty} \text{Shift}_{rN}(h_{\text{des}})(n) \text{ for } 0 \le n < N . \tag{16}$$

What to make of equation (16)? It expresses h on the time interval $0 \le n < N$ as the infinite sum of shifted replicas of the desired impulse response h_{des}. Recall our hope that if h_{des} is most active on that time interval, then h will resemble h_{des} on that interval. Figure 2 illustrates schematically what's going on. Relative inactivity of h_{des} outside the interval $0 \le n < N$ means that the shifted replicas of h_{des} other than the one corresponding to $r = 0$ don't "pollute" the $r = 0$-replica too much, at least on the time interval $0 \le n < N$. In this case, $h(n)$ does indeed approximate $h_{\text{des}}(n)$ for $0 \le n < N$.

Nonetheless, if $h_{\text{des}}(n) \ne 0$ for any n outside the interval $0 \le n < N$, as generally happens in the filter-design setting, replicas $\text{Shift}_{rN}(h_{\text{des}})$ for $r \ne 0$ will contribute to the sum defining the N-point signal h on the interval $0 \le n < N$. When that happens, people often say that "frequency sampling has led to time-aliasing." The terminology makes sense in light of our earlier discussion of aliasing as it arises from sampling signals slower than their Nyquist rates, wherein we saw collisions, albeit in ω-space, between shifted replicas of various things. Not surprisingly, larger N leads to less time-aliasing and a better approximation of $h_{\text{des}}(n)$ for $0 \le n < N$. This amounts to a time-domain version of our earlier observation that larger N means more frequency samples of $\widehat{H}_{\text{des}}$ distributed more densely over $[0, 2\pi]$ and, one might hope, a tighter match between $\widehat{H}$ and $\widehat{H}_{\text{des}}$.

The FFT

What makes these DFT applications a source of real computational economy is the availability of fast algorithms for computing DFTs and inverse DFTs. These algorithms are known as FFTs, or Fast Fourier Transforms. James Cooley and John Tukey, the latter credited with coining the term “bit” for “binary digit,” published the first FFT algorithms in the 1960s. I’ll describe one particular FFT that illustrates most of the important ideas.

Suppose x is an N-point signal and $N = 2^L$ for some positive integer L. Assuming N is a power of 2 loses us no generality since we can regard any N-point signal as a 2^L-point signal provided $N \leq 2^L$. Define the two $N/2$-point signals x_e and x_o as follows:

$$x_e(m) = x(2m) \text{ for } 0 \leq m < \frac{N}{2}$$

and

$$x_o(m) = x(2m+1) \text{ for } 0 \leq m < \frac{N}{2}\ .$$

The $N/2$-point DFTs of x_e and x_o have specifications

$$\left(\widehat{X_e}\right)_k = \sum_{m=0}^{N/2-1} x_e(m)\psi_{\frac{N}{2}}^{-mk} \text{ for } 0 \leq k < \frac{N}{2}$$

and

$$\left(\widehat{X_o}\right)_k = \sum_{m=0}^{N/2-1} x_o(m)\psi_{\frac{N}{2}}^{-mk} \text{ for } 0 \leq k < \frac{N}{2}\ .$$

By equation $\mathcal{DFT}$,

$$\widehat{X}_k = \sum_{n=0}^{N-1} x(n)\psi_N^{-nk} \text{ for } 0 \leq k < N\ .$$

Separate the sum on the right-hand side into terms indexed by even values of n and terms indexed by odd values of n and you obtain

$$\begin{aligned}
\widehat{X}_k &= \sum_{n \text{ even}} x(n)\psi_N^{-nk} + \sum_{n \text{ odd}} x(n)\psi_N^{-nk} \\
&= \sum_{m=0}^{N/2-1} x(2m)\psi_N^{-2mk} + \sum_{m=0}^{N/2-1} x(2m+1)\psi_N^{-(2m+1)k} \\
&= \sum_{m=0}^{N/2-1} x_e(m)\psi_{\frac{N}{2}}^{-mk} + \psi_N^{-k} \sum_{m=0}^{N/2-1} x_o(m)\psi_{\frac{N}{2}}^{-mk} \text{ for } 0 \leq k < N\ .
\end{aligned}$$

I used $\psi_N^2 = \psi_{N/2}$ to get the last line. This equation almost says that

$$\widehat{X}_k = \left(\widehat{X_e}\right)_k + \psi_N^{-k}\left(\widehat{X_o}\right)_k\ ,$$

but not quite. The $N/2$-point DFTs of x_e and x_o are defined only for indices $0 \le k < N/2 - 1$, and k runs from 0 to $N-1$ in the equation above for $\widehat{X}_k$. The difficulty evaporates because $\left(\psi_{N/2}\right)^{N/2} = 1$, so

$$\sum_{m=0}^{N/2-1} x_e(m)\psi_{\frac{N}{2}}^{-mk} = \sum_{m=0}^{N/2-1} x_e(m)\psi_{\frac{N}{2}}^{-m(k-N/2)} = \left(\widehat{X}_e\right)_{k-N/2}$$

and

$$\sum_{m=0}^{N/2-1} x_o(m)\psi_{\frac{N}{2}}^{-mk} = \sum_{m=0}^{N/2-1} x_o(m)\psi_{\frac{N}{2}}^{-m(k-N/2)} = \left(\widehat{X}_o\right)_{k-N/2}$$

when $N/2 \le k < N$. Accordingly, you can compute the N-point DFT of the N-point signal x by first finding the $N/2$-point DFTs of x_e and x_o and then assembling them according to

$$\widehat{X}_k = \begin{cases} \left(\widehat{X}_e\right)_k + \psi_N^{-k}\left(\widehat{X}_o\right)_k & \text{when } 0 \le k < \frac{N}{2} \\ \left(\widehat{X}_e\right)_{k-N/2} + \psi_N^{-k}\left(\widehat{X}_o\right)_{k-N/2} & \text{when } \frac{N}{2} \le k < N\ . \end{cases}$$

The last equation has the more compact representation

$$(\mathcal{FFT}) \qquad \widehat{X}_k = \left(\widehat{X}_e\right)_{\langle\langle k\rangle\rangle_{N/2}} + \psi_N^{-k}\left(\widehat{X}_o\right)_{\langle\langle k\rangle\rangle_{N/2}} \quad \text{for } 0 \le k < N\ .$$

Figure 3 illustrates this maneuver schematically for the case $N = 8$. Why is it useful? Keep in mind that multiplications consume far more computational time than additions. Computing the N-point DFT directly using equation $\mathcal{DFT}$ requires N^2 multiplications nominally, although to be fair some of those are trivial multiplications by 1. Computing an $N/2$-point DFT directly requires $(N/2)^2$ multiplications nominally. Thus computing the N-point DFT of x by computing the $N/2$-point DFTs of x_e and x_o directly and then assembling them via equation $\mathcal{FFT}$ requires a nominal

$$2\left(\frac{N}{2}\right)^2 + N$$

multiplications. When $N = 8$, direct computation requires 64 multiplications whereas computation using equation $\mathcal{FFT}$ requires only 40 multiplications, and the savings are even more significant for larger values of N.

But let's not stop there. We can compute the $N/2$-point DFTs of x_e and x_o by dividing those signals into even- and odd-indexed $N/4$-point subsignals, computing those subsignals' $N/4$-point DFTs, and assembling them as in $\mathcal{FFT}$. Two computations each of which requires a nominal $(N/2)^2$ multiplications can each be done with a nominal

$$2\left(\frac{N}{4}\right)^2 + \frac{N}{2}$$

multiplications, which means that computing the original N-point DFT now requires only

$$2\left(2\left(\frac{N}{4}\right)^2 + \frac{N}{2}\right) + N = 4\left(\frac{N}{4}\right)^2 + 2N$$

multiplications. Since $N = 2^L$, we can carry out this divide-and-conquer procedure a total of L times, thereby reducing the original nominal N^2 multiplications to

$$2^L \left(\frac{N}{2^L}\right)^2 + LN = N + N \log_2 N$$

multiplications when all is said and done. The economies are stunning for large N. For example, if $N = 2^{10}$, the nominal 1,048,576-multiplication total shrinks by nearly two orders of magnitude to 11,264.

Figure 4 completes the picture for $N = 8$. You can see how the FFT comprises three stages. The first stage computes the 2-point DFTs of four 2-point signals. The second stage pairs up the outputs of the first stage and assembles each pair into a 4-point DFT. The third and final stage assembles the two 4-point DFTs from the second stage into the 8-point DFT of the original signal x. The FFT for general $N = 2^L$ operates similarly as an L-stage process. The first stage computes a bunch of 2-point DFTs, the second stage pairs these up into 4-point DFTs, and so on, until the Lth stage implements equation $\mathcal{FFT}$ and assembles two $N/2$-point DFTs into the N-point DFT of the original signal.

A closer look at Figure 4 reveals additional computational advantages that the FFT algorithm provides. You can think of the output of each stage as an ordered list of N numbers, where $N = 8$ in Figure 4. The algorithm generates the output of stage $m + 1$ by assembling the items in the output of stage m, and the assembly occurs *pairwise and in place* in the following sense. If the output of stage 1 is $y_1(0)$, $y_1(1)$, ... , $y_1(N - 1)$, then to get $y_2(0)$ and $y_2(2)$ you combine $y_1(0)$ and $y_1(2)$ according to

$$\begin{aligned} y_2(0) &= y_1(0) + y_1(2) \\ y_2(2) &= y_1(0) + \psi_4^{-2} y_1(2) \ . \end{aligned}$$

Similarly, $y_1(1)$ and $y_1(3)$ determine $y_2(1)$ and $y_2(3)$; $y_1(4)$ and $y_1(6)$ determine $y_2(4)$ and $y_2(6)$; and $y_1(5)$ and $y_1(7)$ determine $y_2(5)$ and $y_2(7)$. The same kind of pairwise in-place computation occurs at every stage of the algorithm for general N. The indices $\{1, 2, \ldots, N\}$ break up into pairs (p, q) for which the p- and q-indexed elements of the mth list determine the p- and q-the elements of the $(m + 1)$th list. You can check that the (p, q)-pairing of the elements in the output list of stage m always obeys

$$q = p + 2^m \ \text{ for } \ 1 \le m \le L - 1$$

when $N = 2^L$. The pairwise in-place nature of the computations makes the FFT a good candidate for parallel processing.

A further economy arises because of the nature of the individual pairwise in-place computations I've just described. Consider the 2-point DFTs that the first stage of the algorithm calculates. For example, in the notation of the preceding paragraph,

$$\begin{aligned} y_1(0) &= x(0) + x(2) \\ y_1(1) &= x(0) + \psi_2^{-1} x(2) = x(0) - x(2) \ , \end{aligned}$$

where the last equality holds because $\psi_2 = e^{j\pi} = -1$. So you can think of each two-point DFT as coming from the so-called *butterfly diagram* in Figure 5(a). The only multiplication that occurs is by -1, which in a computer usually entails the flip of a sign bit. A general in-place pairwise computation takes the form

$$\begin{aligned} y_{m+1}(p) &= y_m(p) + \left(\psi_{2^{m+1}}\right)^{-p} y_m(q) \\ y_{m+1}(q) &= y_m(p) + \left(\psi_{2^{m+1}}\right)^{-q} y_m(q) \;, \end{aligned}$$

where $q = p + 2^m$. Since $\left(\psi_{2^{m+1}}\right)^{2^m} = -1$, we can implement this computation with the diagram in Figure 5(b), which features only one true multiplication and contains a standard butterfly embedded in it.

Finally, you've probably noticed that the x-values in Figure 4 appear in a special order reading from top to bottom. Accordingly, implementing the FFT requires shuffling the x-values before processing, and you might expect the appropriate shuffle for large N to be complicated and computationally demanding. Miraculously, that's not the case. I've listed to the left of each x-value in Figure 2 a pair of binary numbers. The left number is the x-value's time index and the right number is its position on the list counting from top to bottom. As you can see, the shuffling process entails a simple bit reversal. If you think about the general L-stage even-odd divide-and-conquer procedure for $N = 2^L$, you'll be able to convince yourself that the bit-reversal shuffling procedure works in general and not just for $N = 8$. If the x-values arrive in an ordered array with entries time-indexed, the pre-processing shuffle simply bit-reverses the indices of the array entries.

As I mentioned before, the FFT algorithm I've presented is only one of many available. This particular algorithm is known as a *decimation in time FFT.* The word "decimation" refers to the separations between even- and odd-indexed items that occur at each stage. Qualitatively, the FFT reduces the computation of an N-point DFT to a pre-processing permutation followed by several stages of in-place pairwise parallel computations each of which looks like Figure 5(b). Quantitatively, it results in a nominal reduction of N^2 multiplications to $N + N\log_2 N$ multiplications. Ingenious, powerful, revolutionary — all those descriptors apply.

CHAPTER 13

The z-Transform

Plenty of discrete-time signals don't have discrete-time Fourier transforms, and plenty of discrete-time LTI systems don't have frequency responses. The z-transform allows us to bring transform analysis to bear on problems involving such signals and systems. The definition of the z-transform resembles the definition of the DTFT morphologically, and most z-transform applications and operational rules have DTFT analogues. While an exhaustive treatment of the z-transform would require more tools from complex analysis than we have at our disposal, we'll be able to penetrate the theory to a depth sufficient for common applications. I'll assume without loss of generality, as in recent chapters, that all the signals we encounter are complex-valued.

Definition of the z-transform

We'll need to understand infinite series of the form

$$\sum_{n=-\infty}^{\infty} x(n)z^{-n} \,, \tag{17}$$

where z is a complex variable and x is a discrete-time signal. Such a series is called a *power series* because each of its terms is a multiple of an integer power of z. We'll want to know whether the series converges for at least some $z \in \mathbb{C}$ and, if so, for what values of z the series converges. The special form of power series leads to a fairly tidy theory of convergence. Recall from Chapter 3 that the two-sided infinite series in (17) converges if and only if both of the one-sided series

$$\sum_{n=0}^{\infty} x(n)z^{-n} \tag{18}$$

and

$$\sum_{n=-\infty}^{-1} x(n)z^{-n} \tag{19}$$

converge. Understanding convergence of these one-sided series hinges on the geometric series, which we first met in Chapter 1. When $\gamma \in \mathbb{C}$, the geometric series

$$\sum_{n=0}^{\infty} \gamma^n$$

converges to $1/(1-\gamma)$ if $|\gamma| < 1$ and diverges if $|\gamma| > 1$.

13.1 Fact: If (18) converges when $z = z_o$, then it converges for all z satisfying $|z| > |z_o|$, In fact, for such z, the series converges absolutely in the sense that $\sum_{n=0}^{\infty} |x(n)||z|^{-n}$ converges. If (19) converges when $z = z_o$, then it converges for all z satisfying $|z| < |z_o|$, In fact, for such z, the series converges absolutely in the sense that $\sum_{n=-\infty}^{-1} |x(n)||z|^{-n}$ converges.

Proof: If $\sum_{n=0}^{\infty} x(n)z_o^{-n}$ converges, the sequence $(x(n)z_o^{-n})_{n\geq 0}$ is bounded from above in magnitude, say by $M > 0$. If $|z| > |z_o|$, then

$$\begin{aligned}
\sum_{n=0}^{N} |x(n)z^{-n}| &= \sum_{n=0}^{N} |x(n)z_o^{-n}|(|z_o|/|z|)^n \\
&\leq M\sum_{n=0}^{N}(|z_o|/|z|)^n \\
&\leq M\sum_{n=0}^{\infty}(|z_o|/|z|)^n \\
&= M/(1-|z_o|/|z|)\ ,
\end{aligned}$$

where the last equality holds by geometric-series reasoning since $|z_o|/|z| < 1$. It follows from Fact 3.7 that the infinite sequence $(x(n)z^{-n})_{n\geq 0}$ is absolutely summable, proving the assertion in the theorem statement about absolute convergence. The sequence is summable by Fact 3.3, so (18) converges. Similarly, if $\sum_{n=-\infty}^{-1} x(n)z_o^{-n}$ converges and $(x(n)z_o^{-n})_{n<0}$ is bounded from above in magnitude by M, then for $|z| < |z_o|$ we have

$$\begin{aligned}
\sum_{n=-N}^{-1} |x(n)z^{-n}| &= \sum_{n=-N}^{-1} |x(n)z_o^{-n}|(|z_o|/|z|)^n \\
&\leq M\sum_{m=1}^{N}(|z|/|z_o|)^m \\
&\leq M\sum_{m=0}^{\infty}(|z|/|z_o|)^m \\
&= M/(1-|z|/|z_o|)\ ,
\end{aligned}$$

so the infinite sequence $(x(n)z^{-n})_{n<0}$ is absolutely summable by Fact 3.7, proving the assertion in the theorem statement about absolute convergence. The sequence is summable by Fact 3.3, so (19) converges. □

Fact 13.1 implies that if the series in (17) converges for some $z_o \in \mathbb{C}$, then the "right side" of the series converges for every z satisfying $|z| > |z_o|$ and the "left side" converges for every z satisfying $|z| < |z_o|$. Assuming such a z_o exists, define R_a as follows:

$$R_a = \inf\left(\left\{|z| : \sum_{n=0}^{\infty} x(n)z^{-n} \text{ converges}\right\}\right)\ .$$

Observe that $R_a = 0$ is possible. By definition of R_a, the series (18) diverges for every $z \in \mathbb{C}$ satisfying $|z| < R_a$. On the other hand, if $|z_1| > R_a$ there exists some

z_o for which (18) converges at z_o and for which $|z_1| > |z_o| > R_a$, and it follows from Fact 13.1 that (18) converges at $z = z_1$. Define R_b similarly via

$$R_b = \sup\left(\left\{|z| : \sum_{n=-\infty}^{-1} x(n)z^{-n} \text{ converges}\right\}\right) .$$

By convention we take $R_b = \infty$ if the set whose sup we're taking is unbounded. By definition of R_b, the series (19) diverges if $|z| > R_b$. On the other hand, if $|z_1| < R_b$ there exists some z_o for which (19) converges at $z = z_o$ and for which $|z_1| < |z_o| < R_b$, and it follows from Fact 13.1 that (19) converges at $z = z_1$.

The two-sided infinite series (17) therefore converges at least for z-values in the region $R_a < |z| < R_b$. Such z-values exist if and only if $R_a < R_b$, in which case the indicated set of z-values constitutes an annular or doughnut-shaped region centered on $z = 0$ in the complex plane. The series may or may not converge for some z-values satisfying $|z| = R_a$ or $|z| = R_b$, but we pay little attention to those borderline values. In most cases, the series diverges for at least one z-value satisfying $|z| = R_a$ and at least one z-value satisfying $|z| = R_b$. Consider, however, the signal x with specification

$$x(n) = \begin{cases} \frac{1}{n^2} & \text{if } n > 0 \\ 0 & \text{if } n \leq 0 \; . \end{cases}$$

For this signal, $R_a = 1$ and $R_b = \infty$, and the series $\sum_{n=-\infty}^{\infty} x(n)z^{-n}$ converges for all z satisfying $|z| = 1$ because $\left(x(n)\right)_{n\in\mathbb{Z}}$ is absolutely summable. We're ready now for the formal definition of the z-transform.

13.2 Definition: Let x be a complex-valued discrete-time signal. We say that x is *z-transformable* when the following conditions hold:

- There exists some $z \in \mathbb{C}$ for which $\sum_{n=-\infty}^{\infty} x(n)z^{-n}$ converges.
- $R_a < R_b$, where

$$R_a = \inf\left(\left\{|z| : \sum_{n=0}^{\infty} x(n)z^{-n} \text{ converges}\right\}\right) .$$

and

$$R_b = \sup\left(\left\{|z| : \sum_{n=-\infty}^{-1} x(n)z^{-n} \text{ converges}\right\}\right) .$$

In that case, we define the *z-transform of x* in two-part fashion as follows:

$$\underbrace{X(z) = \sum_{n=-\infty}^{\infty} x(n)z^{-n}}_{\text{Formula}} \qquad \underbrace{R_a < |z| < R_b}_{\text{Region of convergence}} \quad .$$

We also write

$$x \overset{\mathbf{z}}{\longleftrightarrow} X \quad (\text{ROC})_X \; ,$$

where $(\text{ROC})_X$ abbreviates the specification $R_a < |z| < R_b$.

Fact 13.1 has three noteworthy consequences pertinent to Definition 13.2. First, x is z-transformable if and only if there exists a whole range of R-values for which $|x(n)|R^{-n}$ decays exponentially to zero as $n \to \pm\infty$. That requirement surfaces in the stipulation $R_a < R_b$ in Definition 13.2. The signal y with specification

$$y(n) = \begin{cases} 0 & \text{if } n = 0 \\ \frac{1}{n^2} & \text{if } n \neq 0 \end{cases}$$

fails that test. Even though y satisfies

$$\sum_{n=-\infty}^{\infty} y(n)\, 1^{-n} = \sum_{n=-\infty}^{\infty} y(n) = \frac{\pi^2}{3}$$

and thus complies with the first bullet in Definition 13.2, $\left(y(n)z^{-n}\right)_{n\geq 0}$ is unbounded when $|z| < 1$ and $\left(y(n)z^{-n}\right)_{n<0}$ is unbounded when $|z| > 1$, so $R_a = R_b = 1$ for y, and y has no z-transform. Second, the series (17), which appears as the formula part of the z-transform of a signal x, converges at least for all z in $(\mathrm{ROC})_X$ and diverges for all z lying "strictly outside" $(\mathrm{ROC})_X$ in the sense that $|z| < R_a$ or $|z| > R_b$. The series might, in addition, converge for some z-values satisfying $|z| = R_a$ or $|z| = R_b$. Third, and perhaps most important, the series (17) converges absolutely on $(\mathrm{ROC})_X$ in the sense that

$$\sum_{n=-\infty}^{\infty} |x(n)||z_o|^{-n}$$

converges for all $z_o \in (\mathrm{ROC})_X$.

If you like, you can think of the z-transform of x as a complex-valued function whose domain is the set $(\mathrm{ROC})_X$, which is a proper subset of the complex plane. As we'll see, it's convenient at times to think of the formula part of the z-transform as a function $z \mapsto F(z)$ with a domain larger than $(\mathrm{ROC})_X$ accompanied by the stipulation that the infinite series $\sum_{n=-\infty}^{\infty} x(n)z^{-n}$ converges to $F(z)$ only for $z \in (\mathrm{ROC})_X$. That distinction is subtle, but I hope the ensuing discussion illuminates it.

Prototype examples, operational rules, and z-transform inversion

We'll build up a list of what I call prototype examples of z-transforms. These examples show how easy it is to compute R_a and R_b in important special cases. As you might imagine, geometric-series reasoning makes life simple for us. Here are the first few prototype examples.

13.3 Example: $x = \delta$. In this case, the infinite series in Definition 13.2 converges trivially to 1 for all $z \in \mathbb{C}$, since $x(0) = 1$ and $x(n) = 0$ for all $n \neq 0$. Accordingly, $R_a = 0$ and $R_b = \infty$, and the z-transform of x is

$$X(z) = 1 \qquad 0 < |z| < \infty\ .$$

By convention, we don't include $z = 0$ in $(\mathrm{ROC})_X$ even though the series converges when $z = 0$.

13.4 Example: x is the signal with specification $x(n) = z_o^n u(n)$ for $n \in \mathbb{Z}$, where $z_o \in \mathbb{C}$ is given and nonzero. The series for $X(z)$ is

$$X(z) = \sum_{n=-\infty}^{\infty} x(n)z^{-n} = \sum_{n=0}^{\infty}(z_o/z)^n \ ,$$

which is a geometric series that converges if $|z| > |z_o|$ and diverges if $|z| < |z_o|$. Accordingly, the z-transform of x is

$$X(z) = \frac{1}{1-(z_o/z)} = \frac{z}{z-z_o} \qquad |z_o| < |z| < \infty \ .$$

13.5 Example: x is the signal with specification $x(n) = -z_o^n u(-n-1)$ for $n \in \mathbb{Z}$, where $z_o \in \mathbb{C}$ is given and nonzero. The series for $X(z)$ is

$$x(z) = \sum_{n=-\infty}^{\infty} x(n)z^{-n} = -\sum_{n=-\infty}^{-1}(z/z_o)^{-n} = -\sum_{m=1}^{\infty}(z/z_o)^m \ ,$$

which is a geometric series minus its $m = 0$-term, and it converges if $|z| < |z_o|$ and diverges if $|z| > |z_o|$. Accordingly, the z-transform of x is

$$X(z) = -\left(\frac{1}{1-(z/z_o)} - 1\right) = \frac{z}{z-z_o} \qquad 0 < |z| < |z_o| \ .$$

Examples 13.4 and 13.5 illustrate why it's important to lug the region of convergence along with the formula for the z-transform. The two examples feature the same simplified formula for $X(z)$ but different regions of convergence. Recall also how I mentioned earlier that one might embody the formula part of the z-transform of x in a function $z \mapsto F(z)$ whose domain of definition is larger than $(\mathrm{ROC})_X$. The operative F in these two examples has specification

$$F(z) = \frac{z}{z-z_o} \ .$$

$F(z)$ is well defined for every z except $z = z_o$, but the power series defining $X(z)$ in Examples 13.4 and 13.5 converge to $F(z)$ only for z in the respective regions of convergence.

Like the continuous- and discrete-time Fourier transforms, the z-transform obeys operational rules that streamline computations and contribute to the transform's utility. In proving a few of these rules, I'll try to give you a sense of how regions of convergence come into play.

13.6 Linearity: Suppose

$$x_1 \overset{\mathbf{z}}{\longleftrightarrow} X_1 \qquad (\mathrm{ROC})_{X_1}$$

and

$$x_2 \overset{\mathbf{z}}{\longleftrightarrow} X_2 \qquad (\mathrm{ROC})_{X_2} \ .$$

Suppose $(\mathrm{ROC})_{X_1} \cap (\mathrm{ROC})_{X_2} \neq \phi$. Then for any c_1 and c_2 in $\mathbb{C}$, the signal $x = c_1 x_1 + c_2 x_2$ has z-transform with specification

$$X(z) = c_1 X_1(z) + c_2 X_2(z) \qquad (\mathrm{ROC})_X \supset (\mathrm{ROC})_{X_1} \cap (\mathrm{ROC})_{X_2} \ .$$

Usually, $(\mathrm{ROC})_X = (\mathrm{ROC})_{X_1} \cap (\mathrm{ROC})_{X_2}$, but sometimes the containment is proper. If $(\mathrm{ROC})_{X_1} \cap (\mathrm{ROC})_{X_2} = \phi$, then $x = c_1 x_1 + c_2 x_2$ does not have a z-transform when c_1 and c_2 are both nonzero.

Proof: Let the respective regions of convergence be $R_a < |z| < R_b$ for X_1 and $Q_a < |z| < Q_b$ for X_2. If their intersection is empty, then either $R_a \geq Q_b$ or $Q_a \geq R_b$. Suppose $R_a \geq Q_b$, in which case we have

$$Q_a < Q_b \leq R_a < R_b \ .$$

Since $\sum_{n=0}^{\infty} x_1(n) z^{-n}$ diverges and $\sum_{n=0}^{\infty} x_2(n) z^{-n}$ converges when $|z| < R_a$ and $|z| > Q_a$, and since $Q_a < Q_b \leq R_a$, the series

$$\sum_{n=0}^{\infty} \left(c_1 x_1(n) + c_2 x_2(n)\right) z^{-n}$$

diverges when $Q_a < |z| < R_a$ when both c_1 and c_2 are nonzero. The series actually diverges for all z with $|z| < R_a$ because if it converged for some z_o with $|z_o| \leq Q_a$ it would converge whenever $|z| > |z_o|$ by Fact 13.1. Similarly, since $\sum_{n=-\infty}^{-1} x_1(n) z^{-n}$ converges and $\sum_{n=-\infty}^{-1} x_2(n) z^{-n}$ diverges when $|z| > Q_b$ and $|z| < R_b$, and since $Q_b \leq R_a < R_b$, the series

$$\sum_{n=-\infty}^{-1} \left(c_1 x_1(n) + c_2 x_2(n)\right) z^{-n}$$

diverges when $Q_b < |z| < R_b$ and both c_1 and c_2 are nonzero. It actually diverges when $|z| \geq R_b$ since if it converged for some z_o with $|z_o| \geq R_b$ it would also converge for all z with $|z| < |z_o|$ by Fact 13.1. Hence if $R_a > Q_b$, there exists no z for which the series (17) converges. If $R_a = Q_b$, it is possible that (17) converges, but only when $|z| = R_a$. In either case, $c_1 x_1 + c_2 x_2$ doesn't satisfy the conditions of Definition 13.2 and therefore doesn't have a z-transform. The foregoing argument began with the assumption that $R_a \geq Q_b$, and a parallel argument interchanging the roles of x_1 and x_2 works when $Q_a \geq R_b$.

If the regions of convergence intersect, then the intersection takes the form $P_a < |z| < P_b$, where $P_a = \max\left(\{R_a, Q_a\}\right)$ and $P_b = \min\left(\{R_b, Q_b\}\right)$. In particular, (17) converges when $P_a < |z| < P_b$. Accordingly, $c_1 x_1 + c_2 x_2$ is z-transformable and its z-transform has region of convergence that contains the region $P_a < R < P_b$. The region of convergence could be larger because it's possible that

$$\inf\left(\left\{|z| : \sum_{n=0}^{\infty} \left(c_1 x_1(n) + c_2 x_2(n)\right) z^{-n} \text{ converges}\right\}\right) < P_a$$

or

$$\sup\left(\left\{|z| : \sum_{n=-\infty}^{-1} \left(c_1 x_1(n) + c_2 x_2(n)\right) z^{-n} \text{ converges}\right\}\right) > P_b \ .$$

In any event, it's clear in this case that the formula for the z-transform of $c_1 x_1 + c_2 x_2$ is $c_1 X_1 + c_2 X_2$. □

You may wonder under what circumstances the z-transform of a linear combination of x_1 and x_2 has a region of convergence larger than the intersection of $(\mathrm{ROC})_{X_1}$ with $(\mathrm{ROC})_{X_2}$. Here's a simple example. Let $x_1 = u$ and $x_2 = \mathrm{Shift}_1(u)$. By Example 13.4 with $z_o = 1$, $(\mathrm{ROC})_{X_1}$ is $1 < |z| < \infty$, and it's easy to show that $(\mathrm{ROC})_{X_2}$ is the same. Meanwhile, $x_1 - x_2 = \delta$, whose z-transform has region of convergence $0 < |z| < \infty$.

13.7 Time-shift Rule: Suppose x has z-transform X with region of convergence $(\mathrm{ROC})_X$. Then for any $n_o \in \mathbb{Z}$ the signal $y = \mathrm{Shift}_{n_o}(x)$ has z-transform with specification

$$Y(z) = z^{-n_o} X(z) \qquad (\mathrm{ROC})_Y = (\mathrm{ROC})_X\ .$$

Proof: Let's plug y into the formula for the z-transform and see what happens.

$$\begin{aligned} Y(z) &= \sum_{m=-\infty}^{\infty} y(m) z^{-m} \\ &= \sum_{m=-\infty}^{\infty} x(m-n_o) z^{-m} \\ &= \sum_{n=-\infty}^{\infty} x(n) z^{-(n+n_o)} \\ &= z^{-n_o} X(z) \end{aligned}$$

for all $z \in (\mathrm{ROC})_X$. It's obvious from that chain of equalities that the series for $Y(z)$ converges for exactly the same z-values as the series for $X(z)$, so $(\mathrm{ROC})_Y$ is the same as $(\mathrm{ROC})_X$. □

13.8 Convolution Rule: Suppose

$$x_1 \overset{\mathbf{z}}{\longleftrightarrow} X_1 \qquad (\mathrm{ROC})_{X_1}$$

and

$$x_2 \overset{\mathbf{z}}{\longleftrightarrow} X_2 \qquad (\mathrm{ROC})_{X_2}\ .$$

If $(\mathrm{ROC})_{X_1} \cap (\mathrm{ROC})_{X_2} \neq \phi$, then the convolution $x = x_1 * x_2$ exists and has z-transform with specification

$$X(z) = X_1(z) X_2(z) \qquad (\mathrm{ROC})_X \supset (\mathrm{ROC})_{X_1} \cap (\mathrm{ROC})_{X_2}\ .$$

Usually, $(\mathrm{ROC})_X = ROC_{X_1} \cap (\mathrm{ROC})_{X_2}$, but sometimes the containment is proper.

Proof: Suppose z_o lies in $(\mathrm{ROC})_{X_1} \cap (\mathrm{ROC})_{X_2}$. Then the signals y_1 and y_2 with specifications $y_1(n) = x_1(n) z_o^{-n}$ and $y_2(n) = x_2(n) z_o^{-n}$ are both absolutely summable signals since, as we noted earlier by appealing to Fact 13.1, earlier, the series

$$\sum_{n=-\infty}^{\infty} |x_1(n)||z_o|^{-n}$$

converges when $z_o \in (\mathrm{ROC})_{X_1}$ and

$$\sum_{n=-\infty}^{\infty} |x_2(n)||z_o|^{-n}$$

converges when $z_o \in (\mathrm{ROC})_{X_2}$. By Criterion 5.5, $y_1 * y_2$ exists and is also absolutely summable. Meanwhile,

$$\begin{aligned} y_1 * y_2(n) &= \sum_{k=-\infty}^{\infty} y_1(k) y_2(n-k) \\ &= z_o^{-n} \sum_{k=-\infty}^{\infty} x_1(k) x_2(n-k) \\ &= z_o^{-n} x_1 * x_2(n) \ \text{ for all } \ n \in \mathbb{Z} \,, \end{aligned}$$

and it follows not only that $x = x_1 * x_2$ exists but also that $n \mapsto x_1 * x_2(n) z_o^{-n}$ is absolutely summable and hence summable by Fact 3.3. Thus

$$\sum_{n=-\infty}^{\infty} x_1 * x_2(n) z_o^{-n}$$

converges for all z_o in $(\mathrm{ROC})_{X_1} \cap (\mathrm{ROC})_{X_2}$. Accordingly, $x = x_1 * x_2$ has a z-transform whose region of convergence includes $(\mathrm{ROC})_{X_1} \cap (\mathrm{ROC})_{X_2}$. When z lies in that intersection,

$$\begin{aligned} X(z) &= \sum_{n=-\infty}^{\infty} x(n) z^{-n} \\ &= \sum_{n=-\infty}^{\infty} \left(\sum_{k=-\infty}^{\infty} x_1(k) x_2(n-k) \right) z^{-n} \\ &= \sum_{k=-\infty}^{\infty} x_1(k) \left(\sum_{n=-\infty}^{\infty} x_2(n-k) z^{-n} \right) \\ &= \left(\sum_{k=-\infty}^{\infty} x_1(k) z^{-k} \right) X_2(z) \ \text{ because } z \in (\mathrm{ROC})_{X_2} \\ &= X_1(z) X_2(z) \ \text{ because } z \in (\mathrm{ROC})_{X_1} \ . \end{aligned}$$

Interchanging the order of summation is legal because of the way all the series converge. The Time-shift Rule 13.7 applied to the inner summation on the third line yields the expression on the fourth line. It follows that $X(z) = X_1(z) X_2(z)$ and that $(\mathrm{ROC})_X$ includes $(\mathrm{ROC})_{X_1} \cap (\mathrm{ROC})_{X_2}$. □

Only rarely does $x_1 * x_2$ exist when x_1 and x_2 have z-transforms with non-intersecting regions of convergence. For example, if $x_1 = u$ and x_2 is the signal with specification

$$x_2(n) = \begin{cases} \frac{1}{n^2} & \text{when } n < 0 \\ 0 & \text{when } n \geq 0 \ , \end{cases}$$

then $(\mathrm{ROC})_{X_1}$ is $1 < |z| < \infty$ and $(\mathrm{ROC})_{X_2}$ is $0 < |z| < 1$. Nonetheless, $x_1 * x_2$ exists and has specification

$$x_1 * x_2(n) = \begin{cases} \frac{\pi^2}{6} & \text{if } n \geq 0 \\ \sum_{k=-\infty}^{n} \frac{1}{k^2} & \text{if } n < 0 \ . \end{cases}$$

Observe that $x_1 * x_2$ is not z-transformable because $\sum_{n=0}^{\infty} x_1 * x_2(n) z^{-n}$ converges if and only if $|z| > 1$ whereas $\sum_{n=-\infty}^{-1} x_1 * x_2(n) z^{-n}$ diverges when $|z| > 1$ because $|x_1 * x_2(n)| \geq 1/n^2$ for all $n < 0$.

13.9 z-differentiation Rule: If x has z-transform X with region of convergence $(\mathrm{ROC})_X$, then the signal y with specification

$$y(n) = (n-1)x(n-1) \ \text{ for all } \ n \in \mathbb{Z}$$

has z-transform with specification

$$Y(z) = -\frac{d}{dz} X(z) \qquad (\mathrm{ROC})_Y = (\mathrm{ROC})_X \ .$$

Proof: As it happens, $\left(y(n)z^{-n}\right)_{n \in \mathbb{Z}}$ decays geometrically as $n \to \infty$ for the same z-values as does $\left(x(n)z^{-n}\right)_{n \in \mathbb{Z}}$, so y is z-transformable and $(\mathrm{ROC})_Y$ is the same as $(\mathrm{ROC})_X$. To find the formula for $Y(z)$, start with the series formula for $X(z)$ and differentiate it term-by-term, which is legal because of the way the series converges on $(\mathrm{ROC})_X$.

$$\begin{aligned} -\frac{d}{dz} X(z) &= -\frac{d}{dz} \sum_{m=-\infty}^{\infty} x(m) z^{-m} \\ &= \sum_{m=-\infty}^{\infty} m x(m) z^{-m-1} \\ &= \sum_{n=-\infty}^{\infty} (n-1) x(n-1) z^{-n} \\ &= \sum_{n=-\infty}^{\infty} y(n) z^{-n} \ , \end{aligned}$$

and the series on the last line is the formula for $Y(z)$. □

The Time-shift Rule 13.7 and the z-differentiation Rule 13.9 provide means to expand the list of prototype examples. I'll take notational liberties in what follows by writing signal and z-transform specifications on either end of the z-transform arrow rather than whole signals and whole functions of z, for instance rendering Example 13.4 as

$$z_o^n u(n) \xleftrightarrow{\mathbf{z}} \frac{z}{z - z_o} \qquad |z_o| < |z| < \infty \ .$$

Let's start with that example and apply the Time-shift Rule to obtain

$$z_o^{(n-1)}u(n-1) \overset{\mathbf{z}}{\longleftrightarrow} \frac{1}{z-z_o} \qquad |z_o| < |z| < \infty \ .$$

By the z-differentiation Rule we have

$$(n-1)z_o^{n-2}u(n-2) \overset{\mathbf{z}}{\longleftrightarrow} \frac{1}{(z-z_o)^2} \qquad |z_o| < |z| < \infty \ ,$$

and another application of the time-shift rule leads to

$$nz_o^{n-1}u(n-1) \overset{\mathbf{z}}{\longleftrightarrow} \frac{z}{(z-z_o)^2} \qquad |z_o| < |z| < \infty \ .$$

Observe finally that $nz_o^{n-1}u(n-1) = nz_o^{n-1}u(n)$ for all $n \in \mathbb{Z}$. The final form of our first new prototype example is

$$(1a) \qquad nz_o^{n-1}u(n) \overset{\mathbf{z}}{\longleftrightarrow} \frac{z}{(z-z_o)^2} \qquad |z_o| < |z| < \infty \ .$$

Now go through the same drill starting from example $(1a)$. You get

$$(n-1)z_o^{(n-2)}u(n-1) \overset{\mathbf{z}}{\longleftrightarrow} \frac{1}{(z-z_o)^2} \qquad |z_o| < |z| < \infty$$

from the Time-shift Rule and then

$$(n-1)(n-2)z_o^{n-3}u(n-2) \overset{\mathbf{z}}{\longleftrightarrow} \frac{2}{(z-z_o)^3} \qquad |z_o| < |z| < \infty$$

from the z-differentiation Rule and finally

$$(2a) \qquad \frac{n(n-1)}{2}z_o^{n-2}u(n) \overset{\mathbf{z}}{\longleftrightarrow} \frac{z}{(z-z_o)^3} \qquad |z_o| < |z| < \infty \ ,$$

where I first used linearity to transplant the factor of 2 and then applied the fact that $n(n-1)u(n-1) = n(n-1)u(n)$ for all $n \in \mathbb{Z}$.

If you keep this up you end up with the following prototype example for every nonnegative integer k:

$$(ka) \qquad \binom{n}{k} z_o^{n-k}u(n) \overset{\mathbf{z}}{\longleftrightarrow} \frac{z}{(z-z_o)^{k+1}} \qquad |z_o| < |z| < \infty \ .$$

Similar plug-and-chug using the Time-shift Rule and the z-differentiation Rule starting from Example 13.5 leads to

$$(kb) \qquad -\binom{n}{k} z_o^{n-k}u(-n-1) \overset{\mathbf{z}}{\longleftrightarrow} \frac{z}{(z-z_o)^{k+1}} \qquad 0 < |z| < |z_o|$$

for every nonnegative integer k. Observe that for every k, the z-transforms in (ka) and (kb) feature the same formula but different regions of convergence. Note also that Examples 13.4 and 13.5 instantiate $(0a)$ and $(0b)$ respectively.

Setting the prototype examples aside for the moment, let's confirm that a signal's z-transform determines the signal unambiguously. A good way to do that is to demonstrate how to recover a signal x from its z-transform. Suppose x has z-transform with region of convergence given by $R_a < |z| < R_b$. For any $\widehat{R}$ lying strictly between R_a and R_b, the signal y with specification $y(n) = \widehat{R}^{-n}x(n)$ decays geometrically as $n \to \pm\infty$, so it's an absolutely summable signal and therefore has DTFT with specification

$$\widehat{Y}(\omega) = \sum_{n=-\infty}^{\infty} y(n)e^{-jn\omega} = \sum_{n=-\infty}^{\infty} x(n)\left(\widehat{R}e^{j\omega}\right)^{-n} \quad \text{for all} \ \ \omega \in \mathbb{R} \ .$$

$\widehat{Y}$ is a continuous function of ω because y is absolutely summable. Observe that the sum in the rightmost term is $X(\widehat{R}e^{j\omega})$, i.e. the z-transform of x evaluated along the circle of radius $\widehat{R}$ centered at $z = 0$. Applying equation $\mathcal{DTFT}^{-1}$ from Chapter 11 to y yields

$$y(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \widehat{Y}(\omega)e^{jn\omega}d\omega = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(\widehat{R}e^{j\omega})e^{jn\omega}d\omega \ \ \text{for all} \ \ n \in \mathbb{Z} \ ,$$

and, since $x(n) = y(n)\widehat{R}^n$ for all n,

$$(\mathcal{Z}^{-1}) \qquad x(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X\left(\widehat{R}e^{j\omega}\right)\widehat{R}^n e^{jn\omega}d\omega \ \ \text{for all} \ \ n \in \mathbb{Z} \ ,$$

Complex-analysis aficionados might prefer to express the right-hand side of equation $\mathcal{Z}^{-1}$ as a contour integral along the counterclockwise-directed circle of radius $\widehat{R}$ centered at $z = 0$, which I'll denote by $\widehat{C}$. Along $\widehat{C}$, $z = \widehat{R}e^{j\omega}$, so $dz = j\widehat{R}e^{j\omega}d\omega$, or, equivalently, $d\omega = dz/jz$. Accordingly,

$$x(n) = \frac{1}{2\pi j}\oint_{\widehat{C}} X(z)z^{n-1}dz \ \ \text{for all} \ \ n \in \mathbb{Z} \ .$$

Soon we'll learn an easier method for recovering x from X in the special case that X is given on $(\mathrm{ROC})_X$ by a proper rational function of z. For now, let's see how the inversion formula $\mathcal{Z}^{-1}$ applies to the prototype examples. The region of convergence for example (ka) is $|z_o| < |z| < \infty$, so $\widehat{R}$ in the inversion formula must satisfy $\widehat{R} > |z_o|$. Similarly, applying the inversion formula to example (kb) requires $\widehat{R} < |z_o|$. It follows that if $X(z)$ is the z-transform formula from example (ka) or example (kb), then for every $n \in \mathbb{Z}$ we have

$$(20) \qquad \frac{1}{2\pi}\int_{-\pi}^{\pi} X\left(\widehat{R}e^{j\omega}\right)\widehat{R}^n e^{jn\omega}d\omega = \begin{cases} \binom{n}{k}z_o^{n-k}u(n) & \text{if } \widehat{R} > |z_o| \\ -\binom{n}{k}z_o^{n-k}u(-n-1) & \text{if } \widehat{R} < |z_o| \ . \end{cases}$$

The z-transform and LTI systems

First let's establish the relationship between the z-transform and the discrete-time Fourier transform. Suppose x is z-transformable and $(\mathrm{ROC})_X$ contains the unit circle $|z| = 1$, which is the set of all z-values of the form $e^{j\omega}$ for $\omega \in \mathbb{R}$. Evaluating $X(z)$ at $z = e^{j\omega}$ yields

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x(n)e^{-jn\omega} = \widehat{X}(\omega) \ \ \text{for all} \ \ \omega \in \mathbb{R} \ .$$

Thus x has a DTFT if x is z-transformable and the region of convergence of x's z-transform includes the unit circle $|z| = 1$ in the complex plane. Note, however, that a z-transformable x can have a DTFT even when this condition fails. For example, the signal x with specification

$$x(n) = \begin{cases} \frac{1}{n^2} & \text{for } n > 0 \\ 0 & \text{for } n \leq 0 \end{cases}$$

is absolutely summable and therefore has a DTFT, but its z-transform has region of convergence $1 < |z| < \infty$, which doesn't include the unit circle. Furthermore, as we noted earlier, the signal x with specification

$$x(n) = \begin{cases} \frac{1}{n^2} & \text{when } n \neq 0 \\ 0 & \text{when } n = 0 \end{cases}$$

is absolutely summable and has a DTFT but doesn't have a z-transform. Nor does a signal whose DTFT contains impulses — for example, a constant signal or a discrete-time sinusoid of the form $n \mapsto e^{jn\omega_o}$ — possess a z-transform. At the same time, plenty of signals that have z-transforms don't have DTFTs. Any z-transformable x for which $(\text{ROC})_X$ is "bounded away" from the unit circle in the sense that $R_a > 1$ or $R_b < 1$ has no DTFT.

These complications aside, we learned in Chapter 11 that the frequency response of a discrete-time LTI system, if it exists, is the DTFT of the system's impulse response h. We discovered that a system's frequency response facilitates analyzing how the system responds to pure discrete-time sinusoidal inputs and, more generally, to inputs possessing DTFTs. The z-transform enables us to expand the scope of transform-domain analysis to encompass a wide class of systems, including many that lack frequency responses, and a wide class of input signals, including many that lack DTFTs.

13.10 Definition: We say that a LTI system with impulse response h *has a transfer function* when h is z-transformable, in which case the *transfer function* of the system is the z-transform of h.

Suppose h is the impulse response of a LTI system with system mapping S, and h is z-transformable, so the system has transfer function

$$H(z) = \sum_{n=-\infty}^{\infty} h(n) z^{-n} \qquad (\text{ROC})_H \,.$$

Consider driving the system with input x specified by $x(n) = z_o^n$ for all $n \in \mathbb{Z}$, where z_o is a given complex number. The input x is admissible if and only if x lies in $\mathcal{D}_h$, the set of all signals whose convolution with h exists. In that case,

$S(x) = h * x$, so

$$\begin{aligned} S(x)(n) &= \sum_{k=-\infty}^{\infty} h(k)x(n-k) \\ &= \sum_{k=-\infty}^{\infty} h(k)z_o^{n-k} \\ &= \left(\sum_{k=-\infty}^{\infty} h(k)z_o^{-k} \right) z_o^n \ \text{ for all } \ n \in \mathbb{Z} \,. \end{aligned}$$

It follows that $x \in \mathcal{D}_h$ if $z_o \in (\text{ROC})_H$, in which case the expression in parentheses is simply $H(z_o)$ and the overall equation reads

$$S(x)(n) = H(z_o)z_o^n = H(z_o)x(n) \ \text{ for all } \ n \in \mathbb{Z} \,.$$

Since $S(x) = \text{constant} \times x$, x is an “eigen-input” to the system much like pure discrete-time sinusoids are eigen-inputs to systems possessing frequency responses. It’s possible that $x \in \mathcal{D}_h$ even when $z_o \notin (\text{ROC})_H$, but only if z_o lies on one of the boundary circles of $(\text{ROC})_H$, because the series $\sum_{k=-\infty}^{\infty} h(k)z_o^{-k}$ diverges when z_o is “strictly outside” $(\text{ROC})_H$.

Definition 11.4 states that a LTI system with impulse response h has a frequency response when every pure discrete-time sinusoidal input $n \mapsto e^{jn\omega_o}$ lies in $\mathcal{D}_h$. The computations in the preceding paragraphs demonstrate that if every z-value of the form $e^{j\omega_o}$ lies in $(\text{ROC})_H$ — which is the same as saying that the unit circle in the complex plane lies in $(\text{ROC})_H$ — then the system has a frequency response $\widehat{H}$, and $\widehat{H}(\omega) = H(e^{j\omega})$ for all $\omega \in \mathbb{R}$. Note that it’s possible for a system to possess a frequency response $\widehat{H}$ and a transfer function H even when the unit circle doesn’t lie in $(\text{ROC})_H$. To recycle an example we’ve seen before, if h has specification

$$h(n) = \begin{cases} \frac{1}{n^2} & \text{if } n > 0 \\ 0 & \text{if } n \leq 0 \,, \end{cases}$$

then the system has a frequency response because h is absolutely summable, and the system has a transfer function H whose region of convergence is $1 < |z| < \infty$, which doesn’t include the unit circle. Worse yet, a system can have a frequency response and not even have a transfer function. An example is the system with the absolutely summable impulse response

$$h(n) = \begin{cases} \frac{1}{n^2} & \text{if } n \neq 0 \\ 0 & \text{if } n < 0 \,, \end{cases}$$

which we learned earlier doesn’t have a z-transform.

Next, suppose a LTI system with system mapping S and impulse response h has a transfer function, and suppose x is a z-transformable signal. The Convolution Rule 13.8 implies that if $(\text{ROC})_H \cap (\text{ROC})_X \neq \phi$, then $h * x$ exists, so x is an admissible input for the system and the corresponding output $y = S(x) = h * x$ has z-transform with specification

$$Y(z) = H(z)X(z) \qquad (\text{ROC})_Y \supset (\text{ROC})_H \cap (\text{ROC})_X \,.$$

As I noted before while discussing the Convolution Rule, a nonempty intersection between two z-transforms’ regions of convergence is a sufficient but not quite necessary condition for the existence of the corresponding signals’ convolution. In the present context, $x \in \mathcal{D}_h$ is possible even when $(\mathrm{ROC})_H$ and $(\mathrm{ROC})_X$ don't intersect, but only rarely.

If a LTI system with impulse response h is causal, then $h(n) = 0$ for all $n < 0$ by Theorem 6.5, so the series (19) converges trivially for all z. It follows that if h is z-transformable, i.e. if the system has a transfer function, then $(\mathrm{ROC})_H$ must take the form $R_a < |z| < \infty$. Note that many non-causal systems have transfer functions whose regions of convergence take that form. A simple example is the system with impulse response $h = \mathrm{Shift}_{-1}(\delta)$. In any event, suppose we have a causal system whose transfer function H has region of convergence $R_a < |z| < \infty$. If $R_a > 1$, Fact 13.1 implies that the signal h isn't absolutely summable, so by Theorem 6.7 the system isn't BIBO stable. On the other hand, suppose $R_a < 1$, which is the same as saying that the unit circle lies in $(\mathrm{ROC})_H$. Then Fact 13.1 implies that h is absolutely summable and the system is therefore BIBO stable by Theorem 6.7. It's possible for a causal system to be BIBO stable even when the region of convergence for its transfer function doesn't include the unit circle, but $R_a = 1$ is the only option in this case. Again, the signal with specification

$$h(n) = \begin{cases} \frac{1}{n^2} & \text{if } n > 0 \\ 0 & \text{if } n \leq 0\ , \end{cases}$$

is absolutely summable, so the system with impulse response h is BIBO stable, but $(\mathrm{ROC})_H$ is $1 < |z| < \infty$. We'll see shortly that this borderline scenario can't arise when the system's transfer function H is given on its region of convergence by a rational function of z.

Rational functions and rational z-transforms

A *rational function of z* is a function F with specification

$$F(z) = \frac{p(z)}{q(z)}\ ,$$

where p and q are polynomials in z with complex coefficients. F is a *proper rational function of z* when the degree of p is less than or equal to the degree of q. When the degree of p is strictly less than the degree of q, F is a *strictly proper rational function of z*. If p and q have no common factors, as we may always assume, then the roots of q are called the *poles* of F. Clearly, $F(z)$ is well-defined if and only if z is not a pole of F. Any linear combination

$$F = c_1F_1 + c_2F_2 + \cdots + c_nF_n$$

of rational functions F_i is a rational function, and if all the F_i are proper or strictly proper, then so is F. It's easy to see that every pole of F must be a pole of at least one of the F_i, since if $F(z_o)$ is undefined, then $F_i(z_o)$ must be undefined for at least one i. In most cases, every pole of every F_i is also a pole of F, but certain special linear combinations lead to pole cancellations. For example,

$$\frac{z}{z-1} - \frac{1}{z-1} = 1\ .$$

We say that a z-transformable signal x has a *rational z-transform* when the z-transform of x is given on its region of convergence by a rational function of z. Each of the prototype-example signals we considered earlier has a proper rational

z-transform. If x has a rational z-transform and $X(z)$ is a proper rational function of z, then $X(z)/z$ is a strictly proper rational function of z, so we can expand $X(z)/z$ in partial fractions, thereby expressing $X(z)/z$ as the sum of terms like

$$\frac{c_o}{(z-z_o)^{k+1}}$$

where $k \geq 0$ and z_o is either 0 or a pole of X. The reason $z_o = 0$ can arise even when 0 is not a pole of X is that we divided by z prior to doing the partial fraction expansion. Multiplying through by z yields an expansion of $X(z)$ itself as a sum of terms like c_o or

$$\frac{c_o z}{(z-z_o)^{k+1}} ,$$

where z_o is a pole of X. If $X(z)$ is not proper, $X(z)/z$ won't be strictly proper. In that case, you can use long division to write

$$\frac{X(z)}{z} = p_o(z) + \frac{p_1(z)}{q(z)},$$

where the second term is strictly proper. Expand the second term in partial fractions and multiply both sides by z. You end up expressing $X(z)$ itself as a superposition of terms like $c_o z/(z-z_o)^{k+1}$, as before, along with positive-power-of-z-terms of the form $c_o z^l$ that come from $zp_o(z)$.

The foregoing analysis tempts us to say that any signal x with a rational z-transform is some linear combination of prototype-example signals, whose z-transform formulas take the form $z/(z-z_o)^{k+1}$, and left-shifted impulses, since

$$\text{Shift}_{-l}(\delta) \quad \overset{\mathbf{z}}{\longleftrightarrow} \quad z^l \qquad 0 < |z| < \infty$$

for every $l > 0$. However, we have yet to take $(\text{ROC})_X$ into account. Since $X(z)$ must be finite for every $z \in (\text{ROC})_X$, no pole of X may lie in $(\text{ROC})_X$. Accordingly, the annular region $(\text{ROC})_X$ separates the poles of X into two sets in the following fashion. If $(\text{ROC})_X$ is $R_a < |z| < R_b$, the only possible locations for poles of X are $0 < |z| \leq R_a$ and $R_b \leq |z| < \infty$. The poles in the first set I'll call the *inward poles* and those in the second set the *outward poles* of X.

The inversion formula (20) determines x unambiguously. Plugging a term of the form $c_o z/(z-z_o)^{k+1}$ into the inversion formula leads to the prototype-example signal from (ka) if z_o is an inward pole of X and the prototype-example signal from (kb) if z_o is an outward pole of X. This is because if z_o is an inward pole, then $\widehat{R}$ in equation (20) must satisfy $\widehat{R} > |z_o|$ whereas if z_o is an outward pole, then $\widehat{R} < |z_o|$. Obviously we need not plug X into the inversion formula when we know what the answer will be. We can find the signal x whose z-transform is X with region of convergence $(\text{ROC})_X$ by simply going through the partial-fractions drill and expressing x via "table lookup" as a linear combination of prototype-example signals and left-shifted impulses.

I'll do an example in a moment, but first I'd like to demonstrate that when x has a rational z-transform and $(\text{ROC})_X$ is $R_a < |z| < R_b$ then, if $R_a > 0$, X must have at least one pole on the circle of radius R_a, and, if $R_b < \infty$, X must have at least one pole on the circle of radius R_b. In other words, $(\text{ROC})_X$ is "bounded by poles" of X. Start by writing x as a sum

$$x = x_1 + x_2 + \cdots + x_n$$

where none of the x_i is the zero signal and where each z-transform formula X_i has a single pole. Each of the x_i is a linear combination of prototype-example signals whose z-transforms share the same pole, and we may assume that the poles of X_i and X_j are different when $i \neq j$. This last assumption guarantees that the poles of the X_i are all poles of X. By the linearity property 13.6, the regions of convergence $(\text{ROC})_{X_i}$ of all the z-transforms X_i must have a non-empty intersection, and $(\text{ROC})_X$ will contain that intersection.

Suppose next that $(\text{ROC})_{X_i}$ is $R_{ai} < |z| < R_{bi}$ for $1 \leq i \leq n$. Observe that for each positive R_{ai} and each finite R_{bi} the circle of radius R_{ai} or R_{bi} passes through the pole of X_i, which is also a pole of X by construction. The intersection of all the regions of convergence will be $Q_a < |z| < Q_b$, where $Q_a = \max(\{R_{ai}\})$ and $Q_b = \min(\{R_{bi}\})$. Note that $Q_a = 0$ and $Q_b = \infty$ are possible. If $(\text{ROC})_X$ is $R_a < |z| < R_b$, we must have $R_a \leq Q_a$ and $R_b \geq Q_b$ because $|ROC_X$ must contain the region $Q_a < |z| < Q_b$. It turns out that $R_a = Q_a$ and $R_b = Q_b$. If $Q_a = 0$, then $R_a = Q_a$ trivially. If $Q_a > 0$, then X has a pole on the circle of radius Q_a, so we can't have $R_a < Q_a$, and $R_a = Q_a$ once again. If $Q_b = \infty$, then $R_b = Q_b$ trivially. If Q_b is finite, then X has a pole on the circle of radius Q_b, so we can't have $R_b > Q_b$, and $R_b = Q_b$ once again. It follows that $R_a = Q_a$ and $R_b = Q_b$, so X has at least one pole on the circle of radius R_a if $R_a > 0$ and at least one pole on the circle of radius R_b if R_b is finite.

The foregoing discussion has repercussions for signals with rational z-transforms. I'll leave it to you to prove the following assertions:

- If x is right-sided and has a rational z-transform, then $(\text{ROC})_X$ is the part of the complex plane outside all the poles of X; i.e., $(\text{ROC})_X$ is the set of all $z \in \mathbb{C}$ satisfying $|z| > |z_o|$, where z_o is the pole of X with largest magnitude.
- If x is left-sided and has a rational z-transform, $(\text{ROC})_X$ is the part of the complex plane inside all the poles of X; i.e., $(\text{ROC})_X$ is the set of all $z \in \mathbb{C}$ satisfying $|z| < |z_o|$, where z_o is the pole of X with smallest magnitude.
- If x has finite duration, then $(\text{ROC})_X$ is $0 < |z| < \infty$.

An important consequence is the following result concerning causal BIBO stable systems with rational transfer functions.

13.11 Theorem: Let h be the impulse response of a causal LTI system with system mapping S and suppose h has a proper rational z-transform. If

$$h \xleftrightarrow{\mathbf{z}} H \qquad R_a < |z| < \infty \ ,$$

then the system is BIBO stable if and only if no pole of H lies in $|z| \geq 1$. Alternatively, the system is BIBO stable if and only if $R_a < 1$.

Proof: Since the system is causal, h is right-sided so $(\text{ROC})_H$ is the part of the complex plane outside of all the poles of H. Now suppose z_o is a pole of H that lies strictly outside the unit circle $|z| = 1$. It follows that $1 < |z_o| \leq R_a$, so by definition of R_a, $\sum_{n=0}^{\infty} h(n)z^{-n}$ diverges when $z = 1$, meaning that h is not summable, much less absolutely summable, which in turn implies via Theorem 6.7 that the system is not BIBO stable. If z_o is a pole of H that lies on the unit circle, then $n \mapsto z_o^n$ is a bounded signal inadmissible as an input to the system since

$$h * x(n) = \sum_{k=-\infty}^{\infty} h(k) z_o^{n-k} = \left(\sum_{k=-\infty}^{\infty} h(k) z_o^{-k} \right) z_o^n \ ,$$

and the expression in parentheses blows up because z_o is a pole of H. Accordingly, for the system to be BIBO stable, every pole of H must have magnitude less than 1, and $R_a < 1$. Conversely, if every pole z_o of H satisfies $|z_o| < 1$, then $R_a < 1$, so the unit circle lies in $(\mathrm{ROC})_H$ which implies, as we noted earlier, that h is absolutely summable, so the system is BIBO stable by Theorem 6.7. □

Now for the promised example of rational z-transform inversion.

13.12 Example: Suppose the z-transform of x has specification

$$X(z) = \frac{(z+1) + (7z^2 - 3z)(z-1)(z-2)(z-3)}{(z-1)(z-2)(z-3)} \qquad 2 < |z| < 3 \ .$$

The poles of $X(z)$ are 1, 2, and 3. 1 and 2 are inward poles and 3 is the only outward pole. Following the recipe, you divide both sides by z and use long division to get

$$\frac{X(z)}{z} = 7z - 3 + \frac{z+1}{z(z-1)(z-2)(z-3)} \ .$$

Expand the second term in partial fractions and you obtain

$$\frac{X(z)}{z} = 7z - 3 + \frac{-(1/6)}{z} + \frac{1}{z-1} + \frac{-(3/2)}{z-2} + \frac{2/3}{z-3} \ ,$$

so

$$X(z) = 7z^2 - 3z - \frac{1}{6} + \frac{z}{z-1} + \frac{-(3/2)z}{z-2} + \frac{(2/3)z}{z-3} \ .$$

Finally, invoke the prototype examples to obtain

$$x(n) = 7\delta(n+2) - 3\delta(n+1) - \frac{1}{6}\delta(n) + u(n) - \frac{3}{2}2^n u(n) - \frac{2}{3}3^n u(-n-1) \ \text{ for all } \ n \in \mathbb{Z} \ .$$

Reminder: you can always check your answer to a computation of this kind by taking the z-transform of the x you obtain and making sure it agrees with what you started with.

Signal flow graphs

Suppose we have a causal discrete-time LTI system with a proper rational transfer function. The system's impulse response h is right-sided because the system is causal, so $(\mathrm{ROC})_H$, the region of convergence for the system's transfer function, is the part of the complex plane outside all the poles of H. In other words, knowing the formula for H determines the whole transfer function under the assumption that the system is causal. For that reason, people don't worry much about regions of convergence when discussing causal systems with rational transfer functions, saying simply that the system "has transfer function H." Suppose, then, that a causal system has transfer function H and

$$H(z) = \frac{p(z)}{q(z)} \; ,$$

where

$$p(z) = \sum_{k=0}^{m} p_k z^{m-k}$$

and

$$q(z) = z^m + \sum_{k=1}^{m} q_k z^{m-k} \; .$$

For the purposes of this discussion, we need not assume that the polynomials p and q have no factors in common. We can re-write $H(z)$ as a ratio of polynomials in z^{-1} by multiplying top and bottom by z^{-m}:

$$H(z) = \frac{\sum_{k=0}^{m} p_k z^{-k}}{1 + \sum_{k=1}^{m} q_k z^{-k}} \; .$$

If x is a z-transformable input to the system and y is the corresponding output, then $Y(z) = H(z)X(z)$ implies that

$$\left(1 + \sum_{k=1}^{m} q_k z^{-k} \right) Y(z) = \left(\sum_{k=0}^{m} p_k z^{-k} \right) X(z) \; .$$

By the time-shift rule, this last expression corresponds to the time-domain relationship

$$y(n) + \sum_{k=1}^{m} q_k y(n-k) = \sum_{k=0}^{m} p_k x(n-k) \;\text{ for all }\; n \in \mathbb{Z} \; . \tag{21}$$

Equation (21) is a linear difference equation relating the input and output signals of the system. Linear difference equations play a role in discrete time similar to the role that linear differential equations play in continuous time. A given causal LTI system with a rational transfer function has infinitely many such difference-equation "implementations." For example, one difference equation implementing the sliding-window M-fold averager is

$$y(n) = \frac{1}{M} \sum_{k=0}^{M-1} x(n-k) \;\text{ for all }\; n \in \mathbb{Z} \; .$$

In the notation of the preceding paragraph, $m = M - 1$, $q_k = 0$ for $1 \leq k \leq m$, and $p_k = 1/M$ for $0 \leq k \leq m$. Another difference equation implementing the averager is

$$y(n) - y(n-1) = \frac{1}{M} x(n) - \frac{1}{M} x(n-M) \;\text{ for all }\; n \in \mathbb{Z} \; . \tag{22}$$

For this one, $m = M$, $q_1 = 1$, $q_k = 0$ for $2 \leq k \leq m$, $p_0 = p_m = 1/M$, and $p_k = 0$ for $2 \leq k \leq m - 1$.

A signal flow graph is a connected labeled directed graph that helps us visualize how to "realize" a difference equation (21) in "pseudo-code," or software. At time n, the value of some signal at time n "sits" at each node in the graph. If a node has incoming branches, the signal value sitting at that node is the sum of the signals arriving over all the incoming branches. The label on a branch determines what happens to the signal flowing along it.

- A branch without a label simply passes the value of a signal from the node it exits to the node it enters. See Figure 1(a).
- A branch labeled by a number c_o multiplies by c_o the value of the signal at the node it exits and sends that value to the node it enters. See Figure 1(b).
- A branch labeled z^{-1} delays by one time unit the signal at the node it exits and passes that to the node it enters. See Figure 1(c).

To get a feel for how signal flow graphs work, it pays to start with FIR systems. Suppose a causal FIR system's impulse response h satisfies $h(n) = 0$ when $n > m$. The transfer function of the system is

$$H(z) = \sum_{k=0}^{m} h(k)z^{-k} = \frac{\sum_{k=0}^{m} h(k)z^{m-k}}{z^m} \qquad 0 < |z| < \infty \ .$$

One difference equation of the form (21) that implements the FIR system is

$$y(n) = \sum_{k=0}^{m} h(k)x(n-k) \ .$$

It's pretty clear that the signal flow graph in Figure 2, which takes the form of a so-called *tapped delay line,* realizes this difference equation. You can think of the graph as a snapshot taken at time n of on algorithm simulating the system. The signal values sitting at the outputs of the delay branches represent the contents of the algorithm's memory at time n. Computing $y(n)$ at time n requires not only $x(n)$ but also $x(n-k)$ for $1 \le k \le m$, so those past values must be available at time n. Most people use graphs of the type in Figure 2 to describe FIR systems. The sliding-window M-fold averager is, of course, such a system.

I'll describe in what follows three methods for constructing signal flow graphs realizing more general causal LTI systems implemented by difference equations of the form (21). I'll stick to the case $m = 2$ for simplicity, but it will be clear how to extend the construction to larger m. Equation (21) when $m = 2$ is

$$y(n) + q_1y(n-1) + q_2y(n-2) = p_0x(n) + p_1x(n-1) + p_2x(n-2) \ .$$

The transfer function of this system is

$$\begin{aligned} H(z) &= \frac{p_0z^2 + p_1z + p_2}{z^2 + q_1z + q_2} \\ &= \frac{p_0 + p_1z^{-1} + p_2z^{-2}}{1 + q_1z^{-1} + q_2z^{-2}} \ , \end{aligned}$$

with region of convergence given by $R_a < |z| < \infty$, where R_a is the largest of the magnitudes of the poles of $H(z)$. The formulas for the z-transforms of the system's input x and output y are related by $Y(z) = H(z)X(z)$.

To obtain the so-called *Direct Form I* signal flow graph for the system, first define

$$g(n) = p_0x(n) + p_1x(n-1) + p_2x(n-2) \ \text{ for all } \ n \in \mathbb{Z} \ .$$

Rewrite the difference equation as

$$y(n) = g(n) - q_1y(n-1) - q_2y(n-2) \ ,$$

and you'll see that the signal flow graph in Figure 3 realizes the system. The graph employs four delay branches. Again, you can think of the delay-branch outputs as the contents of memory at time n. To compute $y(n)$ using the algorithm represented in Figure 3, you need $y(n-1)$, $y(n-2)$, and $g(n)$; in order to compute $g(n)$, you need $x(n)$, $g(n-1)$, and $g(n-2)$.

The *Direct Form II* signal flow graph arises as follows. First set

$$W(z) = \frac{1}{1 + q_1 z^{-1} + q_2 z^{-2}} X(z) \ .$$

Then we have

$$\left(1 + q_1 z^{-1} + q_2 z^{-2}\right) W(z) = X(z) \ ,$$

which in the time domain reads

$$w(n) = x(n) - q_1 w(n-1) - q_2 w(n-2) \ .$$

The graph in Figure 4(a) generates $w(n)$; note that the outputs of the delay branches are just time-delayed $w(n)$-terms. Next, it follows from $Y(z) = H(z)X(z)$ that $Y(z) = \left(p_0 + p_1 z^{-1} + p_2 z^{-2}\right) W(z)$, which translates to the time-domain expression

$$y(n) = p_0 w(n) + p_1 w(n-1) + p_2 w(n-2) \ .$$

The resulting Direct Form II signal flow graph appears in Figure 4(b). The Direct Form II realization requires less memory than the Direct Form I. To compute $y(n)$ at time n, you need $w(n)$ along with two stored values $w(n-1)$ and $w(n-2)$. Together with $x(n)$, those stored values also suffice to compute $w(n)$ at time n.

To get the *Transposed Direct Form II* signal flow graph, first manipulate the difference equation relating x and y to obtain

$$y(n) = p_0 x(n) + \left(p_1 x(n-1) - q_1 y(n-1)\right) + \left(p_2 x(n-2) - q_2 y(n-2)\right) \ .$$

If you empty your mind and stare for a while at this equation alongside the graph in Figure 5, you'll see that the graph generates y from x just as the other ones do. I'd suggest that you "chase some signals around the diagram" to see what happens to them. Note that you can get the Transposed Direct Form II from the Direct Form II by reversing all the arrows and interchanging the roles of x and y. Like the Direct Form II, the Transposed Direct Form II employs only two delay branches and therefore requires storing only two signal values in memory at any given time.

I mentioned earlier that any LTI system with a rational transfer can be implemented in infinitely many ways using a difference equation of the form (21). This is easy to see when you consider that

$$H(z) = \frac{p(z)}{q(z)} = \frac{p(z)r(z)}{q(z)r(z)}$$

for every nonzero polynomial $r(z)$. Any two distinct choices of $r(z)$ will lead to different difference-equation implementations of the system with transfer function

H. We came up earlier with two difference equations implementing the sliding-window M-fold averager. For that system,

$$\begin{aligned} H(z) &= \frac{\frac{1}{M}\left(z^{M-1}+z^{M-2}+\cdots+1\right)}{z^{M-1}} \\ &= \frac{\frac{1}{M}\left(z^{M-1}+z^{M-2}+\cdots+1\right)(z-1)}{z^{M-1}(z-1)} \\ &= \frac{\frac{1}{M}\left(z^{M}-1\right)}{z^{M}-z^{M-1}}\ . \end{aligned}$$

The representations of $H(z)$ on the first and third lines correspond with the two difference equations we discussed earlier.

If difference-equation implementations and their attendant signal flow graphs describe distinct possible software realizations of a LTI system, how might one such realization offer computational advantages over another? Back when memory was costly, people focused attention largely on constructing signal-flow-graph realizations containing as few delay branches as possible. One can show that if the polynomials p and q have no common factors, then any signal flow graph realizing the corresponding difference equation (21) employs at least m delay branches. In that sense, m is the "order" of the system with transfer function $H = p/q$. Memory considerations aside, errors arising from finite-precision arithmetic can compromise the behavior of software realizations of LTI systems. The fewer multiplications and additions a realization requires, the less deleterious those errors will be.

Consider, for example, realizing the M-fold averager with the signal flow graph in Figure 2 with $h(k) = 1/M$ for all k. Ostensibly, computing $y(n)$ requires M multiplications and $M-1$ additions. Moving the $1/M$ coefficient to the input branch or the output branch reduces the number of multiplications but not the number of additions. On the other hand, the Direct Form II signal flow graph for the averager implemented by the difference equation (22) requires only two additions and at most three multiplications for any M no matter how large. For the price of some additional memory we have bought ourselves substantial immunity to finite-precision effects.

The unilateral z-transform

The unilateral z-transform of a signal $x \in \mathbb{C}^{\mathbb{Z}}$, if it exists, is the z-transform of the signal xu. Since xu is a right-sided signal, the region of convergence for the unilateral z-transform of x takes the form $R_a < |z| < \infty$. The formula for the unilateral z-transform of x is

$$X_I(z) = \sum_{n=0}^{\infty} x(n)z^{-n}\ .$$

The unilateral z-transform of x determines the signal xu unambiguously via equation $\mathcal{Z}^{-1}$. Alternatively, x's unilateral z-transform determines $x(n)$ for all $n \geq 0$ but says nothing about $x(n)$ for $n < 0$. A signal can have a unilateral z-transform but no z-transform. Typical examples are constant signals and every signal x with specification $x(n) = z_o^n$, which has unilateral z-transform

$$X_I(z) = \frac{z}{z - z_o} \qquad |z_o| < |z| < \infty \; .$$

The unilateral z-transform has its own time-shift rule, namely, if x has unilateral z-transform X_I with region of convergence $(\mathrm{ROC})_X$, then the signal $\mathrm{Shift}_1(x)$ has unilateral z-transform with specification

$$z^{-1}X_I(z) + x(-1)$$

and the same region of convergence $(\mathrm{ROC})_X$. I'll leave the simple proof rule as an exercise. Applying the rule iteratively shows that $\mathrm{Shift}_2(x)$ has unilateral z-transform with specification

$$z^{-2}X_I(z) + z^{-1}x(-1) + x(-2) \; ,$$

$\mathrm{Shift}_3(x)$ has unilateral z-transform with specification

$$z^{-3}X_I(z) + z^{-2}x(-1) + z^{-1}x(-2) + x(-3) \; ,$$

and so on.

The time-shift rule makes the unilateral z-transform useful as a tool for solving difference equations of the form (21) subject to initial conditions. Suppose in equation (21) we have $x(n) = 0$ for all $n < 0$ and we want to compute $y(n)$ for all $n \geq 0$ given $x(n)$ for all $n \geq 0$. That computation requires specifying the values of $y(-1)$, $y(-2), \ldots$, and $y(-m)$. Given those initial conditions, we can solve recursively for $y(0)$, $y(1)$, etc. Suppose for simplicity that $m = 2$, in which case we can re-write (21) as

$$y + q_1\mathrm{Shift}_1(y) + q_2\mathrm{Shift}_2(y) = p_0x + p_1\mathrm{Shift}_1(x) + p_2\mathrm{Shift}_2(x) \; .$$

Assuming x and y have unilateral z-transforms, we can apply the time-shift rule to conclude that

$$Y_I(z) = \frac{p(z)}{q(z)}\left(X_I(z) + c(z)\right) \; ,$$

where $p(z) = p_0z^2 + p_1z + p_2$, $q(z) = z^2 + q_1z + q_2$, and

$$c(z) = \left(p_1 + p_2z^{-1}\right)x(-1) + p_2x(-2) - \left(q_1 + q_2z^{-1}\right)y(-1) - q_2y(-2) \; .$$

It follows that if X_I is a proper rational function, then so is Y_I, and we can compute $y(n)$ for $n \geq 0$ using the recipe for inverting rational z-transforms. The foregoing discussion generalizes easily to $m > 2$.

CHAPTER 14

Linear Algebra II: Eigenvalues, Eigenvectors, and all that

A linear mapping T that maps a vector space V into itself demands closer inspection than a linear mapping between arbitrary vector spaces. Numerous questions about such a T make sense only because T maps V into V. For example, the zeroes on the left- and right-hand sides of the equation $T(0) = 0$ are the same; accordingly, 0 is a fixed point of the mapping $T : V \to V$. Does T have other fixed points? Do there exist nonzero vectors v such that, for example, $T(v) = -v$ or $T(v) = 3v$? If T mapped V into some other vector space W, these and other similar questions would never arise. Understanding the properties of $T : V \to V$ entails posing and answering such questions in a systematic way. The theory is cleanest when V is a finite-dimensional complex vector space, and I'll focus primarily on that case. At the end of the chapter I'll attempt to show how the abstract results underpin certain more or less familiar notions such as eigenvalues and eigenvectors of real and complex square matrices.

Invariant subspaces

Let V be a vector space over $\mathbb{F}$, where $\mathbb{F}$ is $\mathbb{R}$ or $\mathbb{C}$. I'll use the notation $\mathrm{End}(V)$ to denote the set of all linear mappings from V to V, which in the notation of Chapter 4 is $\mathrm{Hom}(V, V)$. Incidentally, "End" stands for "endomorphism." $\mathrm{End}(V)$ is a vector space over $\mathbb{F}$ and by Theorem 4.11 has dimension n^2 if V has dimension n. Defined on $\mathrm{End}(V)$ is the multiplication-type operation that arises from composing two linear mappings S and T in $\mathrm{End}(V)$ to obtain the mapping $ST \in \mathrm{End}(V)$ with specification

$$ST(v) = S(T(v)) \text{ for all } v \in V .$$

That operation, which is not commutative in general, makes $\mathrm{End}(V)$ a *noncommutative algebra* over $\mathbb{F}$. By composing a linear mapping T with itself, we obtain powers of T such as T^2, which has specification

$$T^2(v) = T(T(v)) \text{ for all } v \in V .$$

For every $k \geq 0$, T^{k+1} has specification

$$T^{k+1}(v) = T\left(T^k(v)\right) \text{ for all } v \in V ,$$

where T^0, by convention, is the identity mapping on V, which I'll denote by I in what follows.

Given $T \in \mathrm{End}(V)$, a subspace W of V is *invariant under* T when $T(w) \in W$ for every $w \in W$. Thus W is invariant under T when T maps W into itself. If you think of W as a vector space in its own right, the restriction of T to the subspace W defines a linear mapping from W to W, i.e. a member of $\mathrm{End}(W)$. In a sense, the restriction of T to W constitutes a "sub-mapping" of T. Note that the zero subspace and V itself are invariant under any $T \in \mathrm{End}(V)$, so every T has at least two trivial invariant subspaces. The range and nullspace of T are also invariant under T since $T(v) \in \mathrm{range}(V)$ for every $v \in V$ and *a fortiori* for every $v \in \mathrm{range}(V)$, while

$$T(v) = 0 \in \mathrm{nullspace}(V) \ \text{ for all } \ v \in \mathrm{nullspace}(V) \ .$$

Observe also that if W is invariant under T, then W is also invariant under T^k for every $k > 0$ because when $w \in W$, so is $T(w)$, hence so is $T(T(w)) = T^2(w)$, and therefore so is $T(T^2(w)) = T^3(w)$, and so on.

Suppose now that $W_1, \ldots , W_s$ are nonzero mutually disjoint subspaces of V and suppose also that

$$V = W_1 + \cdots + W_s \ .$$

By Lemma 4.5, every $v \in V$ has a unique expansion of the form

$$v = w_1 + \cdots + w_s \ ,$$

where $w_k \in W_k$ for all k, so to understand what T does to vectors in V, it suffices to understand what T does to vectors in each of the W_k. If W_k is invariant under T for every k, and $T_k \in \mathrm{End}(W_k)$ is the restriction of T to the subspace W_k, then

$$\begin{aligned} T(v) &= T(w_1) + \cdots + T(w_s) \\ &= T_1(w_1) + \cdots + T_s(w_s) \ . \end{aligned}$$

Accordingly, understanding the mapping $T \in \mathrm{End}(V)$ entails understanding the "sub-mappings" $T_k \in \mathrm{Hom}(W_k)$. One might hope that the T_k would be simpler in general that T because each W_k, being a proper subspace of V, is in some sense smaller than V, making the possibilities for mappings in $\mathrm{End}(W_k)$ more limited than the possibilities for mappings in $\mathrm{End}(V)$.

Our central project in this chapter will be to start with an arbitrary $T \in \mathrm{End}(V)$, where V is a finite-dimensional complex vector space, and find mutually disjoint subspaces $W_1, \ldots , W_s$ of V, all invariant under T, whose vector sum is V and on each of which T acts in a relatively simple fashion. By doing that, we'll arrive at a "decomposition" of T into simpler sub-mappings $T_k \in \mathrm{End}(W_k)$, where T_k is the restriction of T to W_k. Sometimes, the T_k take a particularly elementary form.

Eigenvalues, eigenvectors, and eigenspaces

Although we'll focus on finite-dimensional complex vector spaces, the following definition makes sense for arbitrary vector spaces.

14.1 Definition: Let V be a vector space over $\mathbb{F}$, where $\mathbb{F}$ is $\mathbb{R}$ or $\mathbb{C}$. If $T \in \mathrm{End}(V)$, a vector $v_o \in V$ is an *eigenvector* of T when $v_o \neq 0$ and there exists $\lambda_o \in \mathbb{F}$ for which

$$T(v_o) = \lambda_o v_o \ .$$

In this case, we say that λ_o is an *eigenvalue* of T and that v_o is an *eigenvector of T corresponding to eigenvalue λ_o*.

You can think of $T \in \mathrm{End}(V)$ as "moving vectors around" in V by means of the mapping $v \mapsto T(v)$. An eigenvector v_o of T is a vector that T "moves" in the simplest way possible, namely by multiplying v_o by a scalar $\lambda_o \in \mathbb{F}$.

Observe that any eigenvector v_o of T spans a one-dimensional subspace of V invariant under T, namely $\mathrm{span}(\{v_o\})$. This is because we can write any v in $\mathrm{span}(\{v_o\})$ as $v = c_o v_o$ for some $c_o \in \mathbb{F}$, and

$$T(c_o v_o) = c_o T(v_o) = c_o \lambda_o v_o \in \mathrm{span}(\{v_o\}) \ .$$

Because $T(c_o v_o) = \lambda_o(c_o v_o)$, every nonzero $v \in \mathrm{span}(\{v_o\})$ is an eigenvector of T corresponding to eigenvalue λ_o. In fact, any one-dimensional subspace W invariant under T is $\mathrm{span}(\{v_o\})$ for some eigenvector v_o of T. To see why, let v_o be any nonzero vector in W, which means $W = \mathrm{span}(\{v_o\})$. Since $T(v_o) \in W$, we can find $\lambda_o \in \mathbb{F}$ such that $T(v_o) = \lambda_o v_o$, which means v_o is an eigenvector of T corresponding to eigenvalue λ_o.

Note that v_o is an eigenvector of T corresponding to eigenvalue λ_o if and only if $v_o \neq 0$ and

$$\begin{aligned} 0 &= T(v) - \lambda_o v \\ &= (T - \lambda_o I)(v) \ , \end{aligned}$$

where I is the identity mapping on V. Thus saying that v_o is an eigenvector of T corresponding to eigenvalue λ_o is the same as saying that v_o is a nonzero vector in $\mathrm{nullspace}(T - \lambda_o I)$. In particular, for λ_o to be an eigenvalue of T, $T - \lambda_o I$ must have a nonzero nullspace, and the nonzero vectors in that nullspace are precisely the eigenvectors corresponding to eigenvalue λ_o. For that reason, we call $\mathrm{nullspace}(T - \lambda_o I)$ the *eigenspace* of T corresponding to eigenvalue λ_o. I'll use the notation $E(\lambda_o)$ for that, i.e.

$$E(\lambda_o) = \mathrm{nullspace}(T - \lambda_o I)$$

whenever λ_o is an eigenvalue of T. The eigenspace $E(\lambda_o)$ is invariant under T because if $v \in E(\lambda_o)$, then

$$(T - \lambda_o I)(T(v)) = \left(T^2 - \lambda_o T\right)(v) = T\left((T - \lambda_o I)(v)\right) = T(0) = 0 \ ,$$

so $T(v)$ is also in $E(\lambda_o)$.

The following central result holds for any vector space.

14.2 Theorem: If $\lambda_1, \lambda_2, \ldots, \lambda_k$ are distinct eigenvalues of $T \in \mathrm{End}(V)$, where V is a real or complex vector space, then the corresponding eigenspaces $E(\lambda_1), E(\lambda_2), \ldots, E(\lambda_k)$ are mutually disjoint subspaces of V.

Proof: Suppose $v_j \in E(\lambda_j)$ for $1 \leq j \leq k$ and that

$$v = v_1 + v_2 + \cdots + v_k = 0 \ .$$

Note that

$$\begin{aligned}(T-\lambda_1 I)(v) &= (T-\lambda_1 I)(v_1)+\sum_{j=2}^{k}(T-\lambda_1 I)(v_j)\\ &= \sum_{j=2}^{k}(\lambda_j-\lambda_1)v_j\end{aligned}$$

because $T(v_j)=\lambda_j v_j$ for all j. Next, apply to the vector on the last line the linear mappings $T-\lambda_2 I$, $T-\lambda_3 I$, ... , $T-\lambda_{k-1}I$ in succession. One by one you kill off the v_j-terms for $2\le j<k$, arriving at

$$(\lambda_k-\lambda_1)(\lambda_k-\lambda_2)\cdots(\lambda_k-\lambda_{k-1})v_k=0\ ,$$

implying that $v_k=0$ because $\lambda_j\ne\lambda_k$ for all $j<k$. In a similar fashion you can prove that $v_j=0$ for all $1\le j<k$. It follows from Lemma 4.5 that $E(\lambda_1)$, $E(\lambda_2)$, ... , $E(\lambda_k)$ are mutually disjoint subspaces of V. □

An important consequence of Theorem 14.2 follows from Theorem 4.6. Suppose that V has dimension n and that $\lambda_1,\dots,\lambda_k$ are distinct eigenvalues of $T\in\mathrm{End}(V)$. Since $E(\lambda_j)\ne\{0\}$ for all j, each $E(\lambda_j)$ is at least a one-dimensional subspace of V. Mutual disjointness and Theorem 4.6 imply that

$$\dim\left(E(\lambda_1)+\cdots+E(\lambda_k)\right)\ge k\ .$$

Since the subspace on the left-hand side is a subspace of V, its dimension is at most n, from which it follows that $k\le n$. Thus when V is finite-dimensional, no linear mapping in $\mathrm{End}(V)$ has more than $\dim(V)$ distinct eigenvalues.

When does a linear mapping have eigenvectors and eigenvalues at all? If $T\in\mathrm{End}(V)$ is the zero mapping, then $T(v)=0$ for every $v\in V$, so every nonzero vector $v\in V$ is an eigenvector of T corresponding to eigenvalue 0. If $I\in\mathrm{End}(V)$ is the identity mapping on V, then $I(v)=v$ for every $v\in V$, so every nonzero vector $v\in V$ is an eigenvalue of I corresponding to eigenvalue 1. In these special cases, T has only one eigenvalue, but every vector in V is an eigenvector corresponding to that eigenvalue.

When V is a vector space over $\mathbb{R}$, a linear mapping $T\in\mathrm{End}(V)$ need not have any eigenvalues and eigenvectors. Suppose, for example, that V is a two-dimensional vector space over $\mathbb{R}$, (v_1,v_2) is a basis for V, and $T\in\mathrm{End}(V)$ has specification

$$T(c_1v_1+c_2v_2)=c_2v_1-c_1v_2\ \text{ for all }\ c_1\ ,c_2\in\mathbb{R};.$$

If $v_o=c_1v_1+c_2v_2$ were an eigenvector of T corresponding to eigenvalue λ_o, we would have

$$T(v_o)=c_2v_1-c_1v_2=\lambda_oc_1v_1+\lambda_oc_2v_2$$

with at least one of the c_j nonzero. Linear independence of v_1 and v_2 yields $c_2=\lambda_oc_1$ and $-c_1=\lambda_oc_2$, implying $c_1=-\lambda_o^2c_1$ and $c_2=-\lambda_o^2c_2$. Since $\lambda_o^2=-1$ is impossible when $\lambda_o\in\mathbb{R}$, we must have $c_1=c_2=0$, which contradicts $v_o\ne0$. On the other hand, every $T\in\mathrm{End}(V)$ has at least one eigenvector when V is a finite-dimensional complex vector space.

14.3 Fact: If V is a finite-dimensional vector space over $\mathbb{C}$ and $T \in \mathrm{End}(V)$, then T has at least one eigenvector.

Proof: Suppose V has dimension n. Recall from Theorem 4.11 that for any real or complex vector spaces V and W the vector space $\mathrm{Hom}(V, W)$ has dimension mn when V has dimension n and W has dimension m. In particular, $\mathrm{End}(V) = \mathrm{Hom}(V, V)$ has dimension n^2 when V has dimension n. It follows from Theorem 4.3 that I, T, T^2, $\ldots$, T^{n^2} are linearly dependent in $\mathrm{End}(V)$, so we can find constants $c_j \in \mathbb{C}$ not all zero for which

$$c_0 I + c_1 T + c_2 T^2 + \cdots + c_{n^2} T^{n^2} = 0 \ .$$

Let k be the largest value of j for which $c_j \neq 0$ and consider the polynomial

$$\lambda^k + \frac{c_{k-1}}{c_k}\lambda^{k-1} + \frac{c_{k-2}}{c_k}\lambda^{k-2} + \cdots + \frac{c_0}{c_k} \ .$$

The Fundamental Theorem of Algebra implies that the polynomial factors as

$$(\lambda - \lambda_1)(\lambda - \lambda_2) \cdots (\lambda - \lambda_k)$$

where the complex numbers λ_j are not necessarily distinct. It follows from simple polynomial algebra that the left-hand side of the equation

$$T^k + \frac{c_{k-1}}{c_k}T^{k-1} + \frac{c_{k-2}}{c_k}T^{k-2} + \cdots + \frac{c_0}{c_k}I = 0$$

factors similarly, so that

$$(T - \lambda_1 I)(T - \lambda_2 I) \cdots (T - \lambda_k I) = 0 \ .$$

At least one of the linear mappings in parentheses must have a nonzero nullspace. If none of them did, then $v \neq 0$ would imply that $(T - \lambda_k I)(v) \neq 0$, implying in turn that $(T - \lambda_{k-1} I)(T - \lambda_k I)(v) \neq 0$, and so on, leading eventually to

$$(T - \lambda_1 I)(T - \lambda_2 I) \cdots (T - \lambda_k I)(v) \neq 0 \ ,$$

which is impossible. Accordingly, $T - \lambda_j I$ must have a nonzero nullspace for at least one j, implying that T has at least one eigenvalue and hence at least one eigenvector. □

The argument in the proof of Fact 14.3 fails for real vector spaces because there exist polynomials with real coefficients that possess no real roots. The canonical example is $\lambda^2 + 1 = 0$, which we encountered in the earlier example of a linear mapping on a real vector space that had no eigenvalues.

Geometric multiplicity and diagonalizability

Remember that our mission is to find for a given $T \in \mathrm{End}(V)$ a set of mutually disjoint subspaces of V whose vector sum is V, each of which is invariant under T and on each of which T "acts" in a relatively straightforward fashion. Theorem 14.2 nominates the eigenspaces of T as candidates for those subspaces. We'll find that the eigenspaces do the job under certain circumstances, but sometimes we'll need to look further.

If V is a finite-dimensional vector space over $\mathbb{R}$ or $\mathbb{C}$ and $T \in \mathrm{End}(V)$ has distinct eigenvalues $\lambda_1, \ldots, \lambda_s$, the *geometric multiplicity* of eigenvalue λ_j, which I'll denote by m_j, is the dimension of the eigenspace $E(\lambda_j)$. By Theorems 14.2 and 4.6, the dimension of the subspace

$$W = E(\lambda_1) + E(\lambda_2) + \cdots + E(\lambda_s)$$

is $m_1 + m_2 + \cdots + m_s$. If V has dimension n, then the only n-dimensional subspace of V is V itself, so we have $W = V$ if and only if the geometric multiplicities of T's eigenvalues sum to n. In that case, we say that T is *diagonalizable.* Stringing together bases for T's eigenspaces as in the proof of Theorem 4.6 generates a basis for V consisting solely of eigenvectors of T. Thus a diagonalizable linear mapping $T \in \mathrm{End}(V)$ has the property that its eigenvectors span V.

When T is diagonalizable, the eigenspaces of T provide the invariant-subspace decomposition of V we've been looking for. If T is diagonalizable and has distinct eigenvalues $\lambda_1, \ldots, \lambda_s$, then

$$V = E(\lambda_1) + E(\lambda_2) + \cdots + E(\lambda_s) ,$$

and each $E(\lambda_j)$ is invariant under T. Furthermore, since $T(v) = \lambda_j v$ for every $v \in E(\lambda_j)$, the restriction T_j of T to the subspace $E(\lambda_j)$ is particularly simple. We can carry the process one step further in this case by choosing a basis $(v_{j1}, \ldots, v_{jm_j})$ for each $E(\lambda_j)$ and noting that each of the n one-dimensional subspaces

$$V_{jk} = \mathrm{span}\left(\{v_{jk}\}\right) \quad , \; 1 \le j \le s \quad , 1 \le k \le m_j$$

is invariant under T. Furthermore, they have vector sum V because, for each j,

$$E(\lambda_j) = V_{j1} + V_{j2} + \cdots + V_{jm_j} .$$

Thus a diagonalizable T has the property that V is the vector sum of mutually disjoint one-dimensional subspaces, each invariant under T. Alternatively, T is diagonalizable when T has enough eigenvectors to span V, which is tantamount to saying that T's eigenspaces are "big enough" to sum to V in the sense of vector sum. One case deserves special mention.

14.4 Fact: If V is an n-dimensional real or complex vector space and $T \in \mathrm{End}(V)$ has n distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, then T is diagonalizable.

Proof: By Theorem 14.2, the eigenspaces $E(\lambda_1)$, $E(\lambda_2)$, $\ldots$, $E(\lambda_n)$ are mutually disjoint subspaces of V. Theorem 4.6 implies that

$$\dim(E(\lambda_1) + E(\lambda_2) + \cdots + E(\lambda_n)) = m_1 + m_2 + \cdots + m_n .$$

The vector sum on the left-hand side is a subspace of V and therefore has dimension at most n. Since $m_j \ge 1$ for each j, we must have $m_j = 1$ for all j, which means that the eigenspaces' dimensions sum to n, making T diagonalizable. □

I'd like to stress that Fact 14.4 provides a sufficient but by no means necessary condition for diagonalizability. As I mentioned earlier, if T is the zero or identity mapping on V, then T has only a single eigenvalue but is diagonalizable because its single eigenspace is V itself. Of course, not every T is diagonalizable. If V

is a vector space over $\mathbb{R}$, for example, any $T \in \mathrm{End}(V)$ lacking eigenvalues and eigenvectors is obviously not diagonalizable.

The following less trivial example foreshadows our general approach to non-diagonalizable linear mappings. Let V be a two-dimensional vector space over $\mathbb{F}$ with basis (v_1, v_2) and let $S \in \mathrm{End}(V)$ have specification

$$S(c_1v_1 + c_2v_2) = c_2v_1 \text{ for all } c_1,\ c_2 \in \mathbb{F}\ .$$

In particular, $S(v_1) = 0$ and $S(v_2) = v_1$. Since $S(v_1) = 0$, $\lambda_1 = 0$ is an eigenvalue of S. You can check that $\mathrm{nullspace}(S) = E(0) = \mathrm{span}(\{v_1\})$, so the geometric multiplicity of λ_1 is $m_1 = 1$. Moreover, S has no eigenvalues other than $\lambda_1 = 0$. Note first that $S^2(v) = 0$ for all $v \in V$. If $S(v_o) = \lambda_o v_o$ and $v_o = c_1v_1 + c_2v_2$, then

$$0 = S^2(v_o) = \lambda_o^2 c_1 v_1 + \lambda_o^2 c_2 v_2\ ,$$

implying that $\lambda_o = 0$ if at least one of c_1 and c_2 is nonzero, which must be the case if v_o is an eigenvector of S. Thus S has only the single eigenspace $E(0)$ which, having dimension 1, is a proper subspace of V, and S is therefore not diagonalizable.

Nilspaces and nilpotence

Let V be an n-dimensional vector space over $\mathbb{F}$ and suppose $S \in \mathrm{End}(V)$. We know that $\mathrm{nullspace}(S)$ and $\mathrm{range}(S)$ are both invariant under S. Furthermore, Theorem 4.9 tells us that

$$\dim(\mathrm{nullspace}(S)) + \dim(\mathrm{range}(S)) = n\ .$$

If $\mathrm{nullspace}(S)$ and $\mathrm{range}(S)$ were disjoint, Theorem 4.9 would give us the disjoint vector sum

$$V = \mathrm{nullspace}(S) + \mathrm{range}(S)\ .$$

But the nullspace and range of S are not in general disjoint. For example, the nullspace and range of the S from the preceding paragraph are both equal to $\mathrm{span}(\{v_1\})$. Let's define the *nilspace* of $S \in \mathrm{End}(V)$ as the set of all $v \in V$ satisfying $S^k(v) = 0$ for some $k > 0$. I warn you that "nilspace" is my own coinage. The nilspace of S contains the nullspace of S. The nilspace of S is also invariant under S because if $S^k(v) = 0$, then $S^{k-1}(S(v)) = 0$, so $S(v)$ is in the nilspace of S whenever v is.

Now let d be the dimension of $\mathrm{nilspace}(S)$. I claim that $S^d(v) = 0$ for every $v \in \mathrm{nilspace}(S)$. To see why, suppose $v \in \mathrm{nilspace}(S)$ but $S^d(v) \neq 0$. We know then that $S^l(v) \neq 0$ when $l \leq d$ but that $S^k(v) = 0$ for some smallest $k > d$ by definition of the nilspace, and thus $S^l(v) = 0$ for all $l > k$. Consider a relation of the form

$$c_0v + c_1S(v) + c_2S^2(v) + \cdots + c_dS^d(v) = 0\ ,$$

where the c_l are in $\mathbb{F}$. Operating on both sides with S^{k-1} yields

$$c_0S^{k-1}(v) = 0\ ,$$

implying that $c_0 = 0$ since $S^{k-1}(v) \neq 0$ by assumption. Operating on the resulting relation

$$c_1S(v) + c_2S^2(v) + \cdots + c_dS^d(v) = 0$$

with S^{k-2} yields

$$c_1S^{k-1}(v) = 0\ ,$$

so $c_1 = 0$, as well. You can keep this up and show that all the c_l must be zero. Thus the only linear combination of $\{S^l(v) : 0 \le l \le d\}$ equal to zero is the trivial linear combination, implying that $\{S^l(v) : 0 \le l \le d\}$ is a linearly independent subset of nilspace(S), which is impossible by Theorem 4.3 since nilspace(S) has dimension d. It follows that nilspace$(S) \subset$ nullspace $\left(S^d\right)$ when nilspace(S) has dimension d. Since nullspace $\left(S^d\right) \subset$ nilspace(S) by definition of the nilspace, we conclude that nilspace$(S) =$ nullspace $\left(S^d\right)$. Let's summarize the foregoing discussion, along with an important embellishment, as follows.

14.5 Theorem: Let V be an n-dimensional vector space over $\mathbb{F}$, where $\mathbb{F}$ is $\mathbb{R}$ or $\mathbb{C}$, and suppose $S \in \mathrm{End}(V)$. Define

$$\text{nilspace}(S) = \left\{v \in V : S^k(v) = 0 \text{ for some } k > 0\right\}$$

and suppose nilspace(S) has dimension d. Then nilspace(S) = nullspace $\left(S^d\right)$. Furthermore, nilspace(S) and range $\left(S^d\right)$ are disjoint subspaces of V, both invariant under S, and

$$V = \text{nilspace}(S) + \text{range}\left(S^d\right) \ .$$

Proof: We've shown already that nilspace(S) = nullspace $\left(S^d\right)$ and noted that nilspace(S) is invariant under S. The range of S^d is also invariant under S because if $w = S^d(v)$, then $S(w) = S^d(S(v)) \in$ range $\left(S^d\right)$. As for disjointness, if $w \in$ range $\left(S^d\right)$, and $w = S^d(v)$, and w also lies in nilspace(S), then $S^d(w) = 0$, so $S^{2d}(v) = 0$. It follows that $v \in$ nilspace(S) and therefore that $S^d(v) = 0$, implying that $w = 0$. Accordingly, nilspace(S) and range $\left(S^d\right)$ have only the zero vector in common and are disjoint subspaces of V. By Theorem 4.6,

$$\dim\left(\text{nilspace}(S) + \text{range}\left(S^d\right)\right) = \dim\left(\text{nilspace}(S)\right) + \dim\left(\text{range}\left(S^d\right)\right) \ .$$

Meanwhile, since nilspace(S) = nullspace $\left(S^d\right)$, Theorem 4.9 yields

$$\dim\left(\text{nilspace}(S)\right) + \dim\left(\text{range}\left(S^d\right)\right) = n \ ,$$

from which it follows that

$$\dim\left(\text{nilspace}(S) + \text{range}\left(S^d\right)\right) = n \ ,$$

implying that

$$V = \text{nilspace}(S) + \text{range}\left(S^d\right)$$

since the only n-dimensional subspace of V is V itself. □

Theorem 14.5 leaves open the possibility that nilspace$(S) = V$, which is the same as saying that $d = n$ and, because $S^d(v) = 0$ for all $v \in V$, $S^d = S^n = 0$. In this case, we say that S is a *nilpotent* linear mapping. An important property of nilpotent mappings is the following.

14.6 Fact: Let V be an n-dimensional vector space over $\mathbb{F}$, where $\mathbb{F}$ is $\mathbb{R}$ or $\mathbb{C}$, and suppose $S \in \mathrm{End}(V)$ is nilpotent, i.e. $S^n = 0$. Then $\lambda_o = 0$ is an eigenvalue of

S, and S has no other eigenvalues.

Proof: Suppose $S \in \operatorname{End}(V)$ is nilpotent. If $S = 0$, then, as we've noted already, $\lambda_o = 0$ is the only eigenvalue of S and every $v \in V$ is a corresponding eigenvector. If $S \neq 0$, there exists some $v \in V$ for which $S(v) \neq 0$. Since $S^n = 0$, there exists some $k < n$ for which $S^k(v) \neq 0$ but $S^{k+1}(v) = 0$, which makes $S^k(v)$ an eigenvector of S corresponding to eigenvalue $\lambda_o = 0$ because

$$S\left(S^k(v)\right) = S^{k+1}(v) = 0$$

and $S^k(v) \neq 0$. S has no eigenvalue other than $\lambda_o = 0$ because if $\lambda_1 \neq 0$ were such an eigenvalue with corresponding eigenvector v_1, we would have $S^n(v_1) = \lambda_1^n v_1 \neq 0$, contradicting $S^n = 0$. □

The converse of Fact 14.6 holds when V is a complex vector space but not when V is a real vector space. For example, suppose V is a three-dimensional vector space over $\mathbb{R}$ and (v_1, v_2, v_3) is a basis for V. Define $T \in \operatorname{End}(V)$ as follows:

$$T(c_1v_1 + c_2v_2 + c_3v_3) = -c_1v_2 + c_2v_1 \text{ for all } c_1\ , c_2 \in \mathbb{R}\ .$$

Because $T(v_3) = 0$, v_3 is an eigenvector of T corresponding to eigenvalue 0. If $v_o = c_1v_1 + c_2v_2 + c_3v_3$ is an arbitrary eigenvector of T corresponding to eigenvalue λ_o, we must have

$$-c_1v_2 + c_2v_1 = \lambda_o c_1 v_1 + \lambda_o c_2 v_2 + \lambda_o c_3 v_3\ .$$

Linear independence of v_1 and v_2 then implies that $c_2 = \lambda_o c_1$ and $c_1 = -\lambda_o c_2$, so $c_2 = -\lambda_o^2 c_2$ and $c_1 = -\lambda_o^2 c_1$. Since $\lambda_o^2 = -1$ is impossible when $\lambda_o \in \mathbb{R}$, $c_1 = c_2 = 0$, which means $v_o = c_3v_3$ and $\lambda_o = 0$. Accordingly, $\lambda_o = 0$ is T's only eigenvalue, but T is not nilpotent because, for example, $T^3(v_1) = v_2$, so $T^3 \neq 0$. Fact 14.3 ensures that such an eventuality never arises when V is a complex vector space.

14.7 Fact: Let V be an n-dimensional vector space over $\mathbb{C}$. If $S \in \operatorname{End}(V)$ has only $\lambda_o = 0$ as an eigenvalue, then S is nilpotent.

Proof: If S is not nilpotent, then $\operatorname{nilspace}(S) \neq V$ and by Theorem 14.5 we have the vector-sum decomposition

$$V = \operatorname{nilspace}(S) + \operatorname{range}\left(S^d\right)\ ,$$

where d is the dimension of $\operatorname{nilspace}(S)$. Note that $W = \operatorname{range}\left(S^d\right)$ is nonzero and is invariant under S, so S restricts to a linear mapping $S_1 \in \operatorname{End}(W)$. By Fact 14.3 there exists a nonzero $w_o \in W$ and some $\lambda_o \in \mathbb{C}$ such that

$$S_1(w_o) = \lambda_o w_o\ .$$

Since $S_1(w_o) = S(w_o)$, w_o must also be an eigenvector of S, and λ_o must also be an eigenvalue of S, so $\lambda_o = 0$ by assumption and therefore $w_o \in \operatorname{nullspace}(S)$. But W and $\operatorname{nilspace}(S)$ are disjoint by Theorem 14.5, and $\operatorname{nullspace}(S) \subset \operatorname{nilspace}(S)$, so $w_o = 0$, which contradicts our starting assumption that $\operatorname{range}\left(S^d\right)$ is nonzero. It follows that $S^d = 0$ and $\operatorname{nilspace}(S) = V$, so S is nilpotent. □

Observe that if V is any n-dimensional vector space and $S \in \mathrm{End}(V)$ is nilpotent, then S isn't diagonalizable unless $S = 0$. To see why, note that $S \neq 0$ implies that the eigenspace of S corresponding to eigenvalue 0 — namely, the nullspace of S — has dimension less than n. Since 0 is the only eigenvalue of S by Fact 14.6, the eigenvectors of S can't span V.

Generalized eigenvectors and algebraic multiplicity

Nonzero nilpotent mappings are the quintessential examples of non-diagonalizable mappings, and they furnish a clue as to how to approach general non-diagonalizable mappings T. By Theorem 14.5, the nilspace of a nilpotent linear mapping $S \in \mathrm{End}(V)$ is always all of V even when the nullspace of S — i.e. the eigenspace of S corresponding to eigenvalue 0 — is not. When $T \in \mathrm{End}(V)$ isn't diagonalizable, the eigenspaces of T don't "add up" to V in the sense of vector sum. Each eigenspace is the nullspace of $T - \lambda_j I$ for some eigenvalue λ_j of T. Perhaps we'll have better luck spanning V with vectors from the nilspaces of $T - \lambda_j I$ rather than the nullspaces.

14.8 Definition: Let V be a finite-dimensional vector space over $\mathbb{R}$ or $\mathbb{C}$ and let λ_o be an eigenvalue of T. The *generalized eigenspace* of T corresponding to eigenvalue λ_o, denoted by $G(\lambda_o)$, is the nilspace of $T - \lambda_o I$, i.e.

$$G(\lambda_o) = \{v \in V : (T - \lambda_o I)^k(v) = 0 \text{ for some } k > 0\} .$$

Every nonzero $v \in G(\lambda_o)$ is called a *generalized eigenvector* of T corresponding to eigenvalue λ_o, and the dimension of $G(\lambda_o)$ is called the *algebraic multiplicity* of λ_o.

Observe that if λ_o is an eigenvalue of T, then

$$E(\lambda_o) = \mathrm{nullspace}(T - \lambda_o I) \subset \mathrm{nilspace}(T - \lambda_o I) = G(\lambda_o) .$$

It follows that every eigenvector of T is also a generalized eigenvector of T and that

$$\dim(E(\lambda_o)) \leq \dim(G(\lambda_o)) ,$$

so the geometric multiplicity of λ_o cannot exceed the algebraic multiplicity of λ_o.

Suppose now that V is a complex vector space and $T \in \mathrm{End}(V)$ has only a single eigenvalue λ_1. I claim that $S = T - \lambda_1 I$ is nilpotent. If λ_o is an eigenvalue of S and v_o a corresponding eigenvector, then

$$T(v_o) = (S + \lambda_1 I)(v_o) = (\lambda_o + \lambda_1)v_o ,$$

implying that v_o is an eigenvalue of T corresponding to eigenvalue $\lambda_o + \lambda_1$. It follows that $\lambda_0 = 0$ because λ_1 is T's only eigenvalue. Thus S has sole eigenvalue 0, and S is nilpotent by Fact 14.7. Since S is nilpotent, the nilspace of S — which is the generalized eigenspace of T corresponding to eigenvalue λ_1 — is all of V. Thus the generalized eigenvectors of T span V even if T's eigenvalues don't. As it happens, this last assertion holds for arbitrary $T \in \mathrm{End}(V)$.

While the assertion holds in general only when V is a complex vector space, the proof I give below makes frequent use of the following basic facts about linear mappings in $\mathrm{End}(V)$, where V is a vector space over $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$. First, if v_o is an eigenvector of T corresponding to eigenvalue λ_o, then by simple polynomial algebra we have

$$(T - \lambda I)^k(v_o) = (\lambda_o - \lambda)^k v_o$$

for all $\lambda \in \mathbb{F}$ and every integer $k > 0$. For example,

$$\begin{aligned} (T - \lambda_o I)^2(v_o) &= T^2(v_o) - 2\lambda T(v_o) + \lambda^2 v_o \\ &= \lambda_o^2 v_o - 2\lambda_o \lambda v_o + \lambda^2 v_o \\ &= (\lambda_o - \lambda)^2 v_o \ . \end{aligned}$$

Second, also by simple polynomial algebra, the mappings T and $(T - \lambda I)^k$ commute, i.e. $T(T - \lambda_I)^k = (T - \lambda I)^k T$ for every $\lambda \in \mathbb{F}$ and integer $k > 0$.

14.9 Theorem: Let V be a finite-dimensional vector space over $\mathbb{C}$ and suppose $T \in \mathrm{End}(V)$ has distinct eigenvalues $\lambda_1, \ \dots \ , \lambda_s$. For each j, let $G(\lambda_j)$ be the generalized eigenspace corresponding to eigenvalue λ_j. Then the $G(\lambda_j)$ are mutually disjoint subspaces of V each of which is invariant under T. Furthermore,

$$V = G(\lambda_1) + G(\lambda_2) + G(\lambda_3) + \cdots + G(\lambda_s) \ .$$

Proof: To see why the $G(\lambda_j)$ are mutually disjoint, suppose $v_j \in G(\lambda_j)$ for $1 \leq j \leq s$ and that

$$v = v_1 + v_2 + \cdots + v_s = 0 \ .$$

Let $d_j = \dim(G(\lambda_j))$ denote the algebraic multiplicity of λ_j for each j. We know that $(T - \lambda_j I)^{d_j}(v_j) = 0$ for all j by Theorem 14.5. Note that

$$\begin{aligned} (T - \lambda_1 I)^{d_1}(v) &= (T - \lambda_1 I)^{d_1}(v_1) + \sum_{j=2}^{k} (T - \lambda_1 I)^{d_1}(v_j) \\ &= \sum_{j=2}^{k} (\lambda_j - \lambda_1)^{d_1} v_j \end{aligned}$$

because $T(v_j) = \lambda_j v_j$ for all j. Next, apply to the vector on the last line in succession the linear mappings $(T - \lambda_2 I)^{d_2}$, $(T - \lambda_3 I)^{d_3}$, $\dots$, $(T - \lambda_{s-1} I)^{d_{s-1}}$ and one by one you kill off the v_j-terms for $2 \leq j < s$, arriving at

$$(\lambda_s - \lambda_1)^{d_1}(\lambda_s - \lambda_2)^{d_2} \cdots (\lambda_s - \lambda_{s-1})^{d_{s-1}} v_s = 0 \ ,$$

implying that $v_s = 0$ because $\lambda_j \neq \lambda_s$ for all $j < s$. In a similar fashion you can prove that $v_j = 0$ for all $1 \leq j < s$. It follows from Lemma 4.5 that the $G(\lambda_j)$ are mutually disjoint subspaces of V.

Next, by Theorem 14.5, $v \in G(\lambda_j)$ if and only if $(T - \lambda_j I)^{d_j}(v) = 0$. Since

$$(T - \lambda_j I)^{d_j}(T(v)) = T\left((T - \lambda_j I)^{d_j}(v)\right) = 0$$

whenever $v \in G(\lambda_j)$, we have $T(v) \in G(\lambda_j)$ whenever $v \in G(\lambda_j)$, so $G(\lambda_j)$ is invariant under T for all j.

It remains to show that the vector sum of the $G(\lambda_j)$ is V. I'll prove the result by induction on s, the number of distinct eigenvalues of T. We've noted already that the result holds for $s = 1$, in which case T has a single eigenvalue λ_1 and $T - \lambda_1 I$ is nilpotent, implying that $G(\lambda_1) = V$. Suppose we have shown that any complex vector space W can be written as the vector sum of the generalized eigenspaces of any $T \in \mathrm{End}(W)$ possessing $s-1$ distinct eigenvalues. Referring now to T and V in the theorem statement, it follows from Theorem 14.5 that

$$V = G(\lambda_1) + W \ ,$$

where $W = \mathrm{range}\left((T-\lambda_1 I)^{d_1}\right)$. The two subspaces in the vector sum are disjoint, and W is invariant under T because for $w \in W$ with $w = (T-\lambda_1 I)^{d_1}(v)$ we have

$$T(w) = T\left((T-\lambda_1 I)^{d_1}(v)\right) = (T-\lambda_1 I)^{d_1}(T(v)) \in W \ .$$

Thus T restricts to a mapping $T_1 \in \mathrm{End}(W)$ with specification

$$T_1(w) = T(w) \ \text{ for all } \ w \in W \ .$$

I claim that the eigenvalues of T_1 are $\lambda_2, \ldots, \lambda_s$. To see why this is true, note first that any eigenvector of T_1 is also an eigenvector of T corresponding to the same eigenvalue, so the eigenvalues of T_1 must lie among the eigenvalues of T. Second, if v_o is an eigenvector of T corresponding to eigenvalue λ_1, then $v_o \in G(\lambda_1)$, so $v_o \notin W$ because $G(\lambda_1)$ and W are disjoint and $v_o \neq 0$. Thus the eigenvalues of T_1 must lie among $\lambda_2, \ldots, \lambda_s$. Finally, if $v_o \in V$ is an eigenvector of T corresponding to λ_j with $j > 1$, then

$$(T-\lambda_1 I)^{d_1}(v_o) = (\lambda_j - \lambda_1)^{d_1} v_o \ ,$$

so $v_o \in W$ because $W = \mathrm{range}\left((T-\lambda_1 I)^{d_1}\right)$ and $\lambda_j \neq \lambda_1$. Since $T_1(v) = T(v)$ for all $v \in W$, v_o is also an eigenvector of T_1 corresponding to eigenvalue λ_j, and, in particular, λ_j is an eigenvalue of T_1. Thus T_1 has exactly the $s-1$ eigenvalues $\lambda_2, \ldots, \lambda_s$.

By the induction assumption, W is the vector sum of the mutually disjoint generalized eigenspaces of T_1 corresponding to eigenvalues $\lambda_2, \ldots, \lambda_s$. But these are also generalized eigenspaces of T because when $w \in W$ and $l \geq 0$,

$$(T-\lambda_j I)^l(w) = 0 \Longleftrightarrow (T_1 - \lambda_k I_W)^l(w) = 0 \ \text{ for } \ 2 \leq j \leq s \ ,$$

where I_W is the identity mapping on W. Accordingly,

$$W = G(\lambda_2) + G(\lambda_3) + \cdots + G(\lambda_s) \ ,$$

and therefore

$$V = G(\lambda_1) + G(\lambda_2) + G(\lambda_3) + \cdots + G(\lambda_s)$$

since $V = G(\lambda_1) + W$. □

Theorem 14.9 states that when V is a finite-dimensional complex vector space and $T \in \mathrm{End}(V)$, the generalized eigenvectors of T span V even if T's eigenvectors don't. Note that when λ_j is an eigenvalue of T, $E(\lambda_j) \subset G(\lambda_j)$, so

$$m_j = \dim(E(\lambda_j)) \leq \dim(G(\lambda_j)) = d_j$$

and

$$E(\lambda_1) + \cdots + E(\lambda_s) \subset G(\lambda_1) + \cdots + G(\lambda_s) = V$$

when T has distinct eigenvalues $\lambda_1, \ldots, \lambda_s$. By definition, T is diagonalizable when the vector sum of the $E(\lambda_j)$ is V, which occurs only when

$$m_1 + \cdots + m_s = d_1 + \cdots + d_s = n$$

when V has dimension n. Since $m_j \leq d_j$ for all j, this happens only when $m_j = d_j$ for all j, which is the same as saying that $E(\lambda_j) = G(\lambda_j)$ for all j. In short, T is diagonalizable if and only if every generalized eigenvector of T is also an eigenvector of T. Alternatively, T is diagonalizable if and only if the geometric multiplicity of every eigenvalue λ_j of T equals λ_j's algebraic multiplicity.

Theorem 14.9 has abundant noteworthy consequences. Recall that in the proof of Fact 14.3 we observed that if V is an n-dimensional vector space over $\mathbb{F}$ and $T \in \mathrm{End}(V)$ there exist constants c_k such that

$$c_0 I + c_1 T + c_2 T^2 + \cdots + c_{n^2} T^{n^2} = 0 \ .$$

In other words, we can find a polynomial in T of degree n^2 that evaluates to 0. As it happens, there exists a polynomial of degree n with the same property.

14.10 Definition: Let V be an n-dimensional vector space over $\mathbb{C}$ and suppose $T \in \mathrm{End}(V)$ has distinct eigenvalues $\lambda_1, \ldots, \lambda_s$ with respective algebraic multiplicities $d_1, \ldots, d_s$. The *characteristic polynomial* of T is

$$p_T(\lambda) = (\lambda - \lambda_1)^{d_1} (\lambda - \lambda_2)^{d_2} \cdots (\lambda - \lambda_s)^{d_s} \ .$$

Note that $p_T(\lambda)$ has degree n because $d_1 + \cdots + d_s = n$ by Theorem 14.9.

14.11 Cayley-Hamilton Theorem: Let V be an n-dimensional vector space over $\mathbb{C}$ and suppose $T \in \mathrm{End}(V)$ has distinct eigenvalues $\lambda_1, \ldots, \lambda_s$ with respective algebraic multiplicities $d_1, \ldots, d_s$. The characteristic polynomial of T "evaluated at T" equals zero in the sense that

$$p_T(T) = (T - \lambda_1 I)^{d_1} (T - \lambda_2 I)^{d_2} \cdots (T - \lambda_s I)^{d_s} = 0 \ .$$

Proof: By Theorem 14.9,

$$V = G(\lambda_1) + G(\lambda_2) + \cdots + G(\lambda_s) \ ,$$

where $G(\lambda_j)$ is the generalized eigenspace of T corresponding to λ_j for each j. Since the $G(\lambda_j)$ are mutually disjoint, Lemma 4.5 permits us to write every $v \in V$ uniquely as

$$v = v_1 + v_2 + \cdots + v_s$$

with $v_j \in G(\lambda_j)$ for all j. Since $G(\lambda_j)$ is the nilspace of $T - \lambda_j I$ and has dimension d_j, Theorem 14.5 tells us that

$$(T - \lambda_j I)^{d_j}(v_j) = 0$$

for all j. The factors in the product defining $p_T(\lambda)$ commute, so operating on v_j with $p_T(T)$ yields 0 for all j. For example,

$$\begin{aligned} p_T(\lambda)(v_1) &= (T - \lambda_1 I)^{d_1} (T - \lambda_2 I)^{d_2} \cdots (T - \lambda_s I)^{d_s}(v_1) \\ &= (T - \lambda_2 I)^{d_2} \cdots (T - \lambda_s I)^{d_s} (T - \lambda_1 I)^{d_1}(v_1) \\ &= 0 \ . \end{aligned}$$

Thus $p_T(T)(v_j) = 0$ for all j, so $p_T(T)(v) = 0$. It follows $p_T(T) = 0$ because v was an arbitrary vector in V. □

You might wonder under what circumstances you can find a polynomial $q(\lambda)$ of degree less than n satisfying $q(T) = 0$. The following special example is worth mentioning. If T is diagonalizable and has distinct eigenvalues $\lambda_1, \ldots, \lambda_s$, then you can write any $v \in V$ as

$$v = v_1 + v_2 + \cdots + v_s$$

with $v_j \in E(\lambda_j)$ for all j. Furthermore,

$$(T - \lambda_j I)(v_j) = 0$$

for all j. Applying the argument in the proof of the Cayley-Hamilton Theorem 14.11, you discover that

$$(T - \lambda_1 I)(T - \lambda_2 I) \cdots (T - \lambda_s I)(v) = 0 \text{ for all } v \in V ,$$

so if $s < n$ you've discovered a polynomial in T of degree less than n that evaluates to zero.

Eigenvalues as growth rates: the diagonalizable case

If V is a complex vector space and v_o is an eigenvector of $T \in \mathrm{End}(V)$ corresponding to eigenvalue λ_o, then $T^k(v_o) = \lambda_o^k v$ for every integer $k > 0$, so if $\| \ \|$ is any norm on V,

$$\left\|T^k(v_o)\right\| = |\lambda_o|^k \left\|v_o\right\| \text{ for all } k > 0 .$$

In particular, $\left\|T^k(v_o)\right\| \to 0$ as $k \to \infty$ if $|\lambda_o| < 1$ and $\left\|T^k(v_o)\right\| \to \infty$ as $k \to \infty$ if $|\lambda_o| > 1$. Thus the magnitudes of T's eigenvalues specify the growth or decay rates of the quantity $\left\|T^k(v)\right\|$ at least when v is an eigenvector of T. If V is finite-dimensional and T is diagonalizable, then any $v \in V$ has an expansion

$$v = v_1 + v_2 + \cdots + v_n$$

where n is the dimension of V, each v_j is an eigenvector of T corresponding to eigenvalue λ_j, and the λ_j are not necessarily distinct. Thus

$$T^k(v) = \lambda_1^k v_1 + \lambda_2^k v_2 + \cdots + \lambda_n^k v_n \text{ for all } k \geq 0 .$$

It follows that if all of T's eigenvalues have magnitudes less than 1, then $T^k(v) \to 0$ as $k \to \infty$ for every $v \in V$. If we reason more quantitatively, our intuition might lead us to expect that the largest of T's eigenvalues' magnitudes dictates the worst-case growth or decay rate of the quantity $\left\|T^k(v)\right\|$ as $k \to \infty$ when T is diagonalizable. Is that expectation justified? If so, does it extend to non-diagonalizable T?

If V is a finite-dimensional complex vector space and $T \in \mathrm{End}(V)$, the *spectral radius* of T, which I'll denote by $\rho(T)$, is the maximum of the magnitudes of the eigenvalues of T, i.e.

$$\rho(T) = \max\left(\{|\lambda_o| : \lambda_o \text{ is an eigenvalue of } T\}\right) .$$

Suppose T is diagonalizable and $(v_1, v_2, \ldots, v_n)$ is a basis for V consisting of eigenvectors of T corresponding to not necessarily distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. Let $\| \ \|_1$, $\| \ \|_2$, and $\| \ \|_\infty$ denote, respectively, the 1-norm, the 2-norm, and the infinity-norm associated with the eigenvector basis, as defined at the end of Chapter 4. If $v \in V$ has representation

$$v = c_1 v_1 + c_2 v_2 + \cdots + c_n v_n \ ,$$

then

$$\begin{aligned}
\left\|T^k(v)\right\|_1 &= \left\|c_1\lambda_1^k v_1 + c_2\lambda_2^k v_2 + \cdots + c_n\lambda_n^k v_n\right\|_1 \\
&= \sum_{j=1}^{n} |\lambda_j|^k |c_j| \\
&\leq (\rho(T))^k \left(\sum_{j=1}^{n} |c_j|\right) \\
&= (\rho(T))^k \|v\|_1 \ .
\end{aligned}$$

Similarly,

$$\begin{aligned}
\left\|T^k(v)\right\|_2 &= \left\|c_1\lambda_1^k v_1 + c_2\lambda_2^k v_2 + \cdots + c_n\lambda_n^k v_n\right\|_2 \\
&= \left(\sum_{j=1}^{n} |\lambda_j|^{2k} |c_j|^2\right)^{1/2} \\
&\leq (\rho(T))^k \left(\sum_{j=1}^{n} |c_j|^2\right)^{1/2} \\
&= (\rho(T))^k \|v\|_2
\end{aligned}$$

and

$$\begin{aligned}
\left\|T^k(v)\right\|_\infty &= \left\|c_1\lambda_1^k v_1 + c_2\lambda_2^k v_2 + \cdots + c_n\lambda_n^k v_n\right\|_\infty \\
&= \max\left(\{|\lambda_j|^k |c_j| : 1 \leq j \leq n\}\right) \\
&\leq (\rho(T))^k \max\left(\{|c_j| : 1 \leq j \leq n\}\right) \\
&= (\rho(T))^k \|v\|_\infty \ .
\end{aligned}$$

In short, $\left\|T^k(v)\right\|$ grows at worst as fast, or decays at least as rapidly, as $(\rho(T))^k\|v\|$ when $\| \ \|$ is the 1-, 2-, or infinity-norm associated with the eigenvector basis. If $\| \ \|$ is any norm on V, then by Theorem 4.12 there exist A and B such that

$$\|v\| \leq A\|v\|_1 \ \text{ and } \ \|v\|_1 \leq B\|v\|$$

for every $v \in V$. It follows that for any norm $\| \ \|$ on V we have

$$\begin{aligned}
\left\|T^k(v)\right\| &\leq A\left\|T^k(v)\right\|_1 \\
&\leq A(\rho(T))^k\|v\|_1 \\
&\leq AB(\rho(T))^k\|v\|
\end{aligned}$$

for all $v \in V$. The foregoing discussion proves the following result about diagonalizable linear mappings.

14.12 Theorem: Let V be a finite-dimensional vector space over $\mathbb{C}$. If $T \in \mathrm{End}(V)$ is diagonalizable and $\rho(T)$ is the spectral radius of T, then for any norm $\| \ \|$ on V there exists $M > 0$ such that

$$\left\|T^k(v)\right\| \leq M(\rho(T))^k \|v\|$$

for every $k > 0$ and $v \in V$. □.

Theorem 14.12 confirms our suspicion that eigenvalues influence growth and decay rates of $\left\|T^k(v)\right\|$ as $k \to \infty$, at least when T is diagonalizable. The diagonalizability assumption is indispensable. For example, if $S \in \mathrm{End}(V)$ is nilpotent, then $\rho(S) = 0$ by Fact 14.6, and if S is nonzero there exists $v \in V$ for which $S(v) \neq 0$, which precludes $\|S(v\| \leq M\rho(S)\|v\|$ for all v. Fortunately, we can approximate Theorem 14.12 arbitrarily closely for non-diagonalizable T.

Eigenvalues as growth rates: the non-diagonalizable case

Here's what I mean when I say we can approximate Theorem 14.12 arbitrarily closely for non-diagonalizable linear mappings.

14.13 Theorem: Let V be a finite-dimensional vector space over $\mathbb{C}$. If $\rho(T)$ is the spectral radius of $T \in \mathrm{End}(V)$ and $\zeta > \rho(T)$, then for any norm $\| \ \|$ on V there exists $M > 0$ such that

$$\left\|T^k(v)\right\| \leq M\zeta^k \|v\|$$

for every $k > 0$ and $v \in V$. □.

For diagonalizable T, Theorem 14.13 follows directly from the stronger Theorem 14.12. The intricate construction I'll employ to prove Theorem 14.13 for general T is of independent interest. For one thing, it buttresses the iconic Jordan canonical form for complex square matrices.

Given a finite-dimensional real or complex vector space V, suppose $S \in \mathrm{End}(V)$ has a d-dimensional nilspace, where $d > 0$. A *Jordan basis* for nilspace(S) is a basis $(v_1, v_2, \ldots, v_d)$ with the following properties: $S(v_d) = 0$ and, for each j, either $S(v_j) = 0$ or $S(v_j) = v_{j+1}$. Any Jordan basis for nilspace(S) concatenates chains of vectors each of which takes the form

$$v_{j_o}, v_{j_o+1}, \ldots, v_{j_o+l} \ ,$$

where

$$S(v_j) = \begin{cases} v_{j+1} & \text{when } j_o \leq j < j_o + l \\ 0 & \text{when } j = j_o + l \ . \end{cases}$$

For any such chain I'll call v_{j_o} the *root* of the chain and v_{j_o+l} the *terminus* of the chain. Jordan bases take their name from French mathematician Camille Jordan, who lived during the late nineteenth and early twentieth centuries.

I'll demonstrate shortly how to build a Jordan basis for an arbitrary nilspace(S), but for now I'd like to point out that the termini of the chains in any Jordan basis constitute a linearly independent spanning set for nullspace(S), which means in particular that the number of chains in the Jordan basis is the same as the dimension of nullspace(S). To see why, note first that because nullspace(S) $\subset$ nilspace(S), any $v \in$ nullspace(S) can be expressed as a linear combination

$$v = \sum_{j=1}^{d} c_j v_j$$

of the Jordan basis vectors. Because S annihilates the terminus vectors, $S(v) = 0$ reveals a zero linear combination of the non-terminus basis vectors, all of whose coefficients must be zero by linear independence. Accordingly, the only possibly nonzero coefficients in the original linear combination yielding v are the terminus vectors' coefficients, which means those vectors span nullspace(S). Furthermore, they're linearly independent because they came from a basis in the first place.

I'll give a constructive argument below proving existence of Jordan bases for arbitrary nonzero nilspaces. For now, let's assume Jordan bases exist and see how they help us finish proving Theorem 14.13 for arbitrary non-diagonalizable T.

Suppose T has distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_s$. If $S_j = T - \lambda_j I$ for each j, then the generalized eigenspace $G(\lambda_j)$ is nilspace(S_j) for each j. By Theorem 14.9, the $G(\lambda_j)$ are mutually disjoint and

$$V = G(\lambda_1) + G(\lambda_2) + \cdots + G(\lambda_s) .$$

Form a basis for V by first finding a Jordan basis for each $G(\lambda_j)$ — which we can do because $G(\lambda_j) =$ nilspace(S_j) — and then stringing those Jordan bases together. Supposing V has dimension n, let $(v_1, v_2, v_3, \ldots, v_n)$ be the resulting basis. Each v_k is in nilspace(S_j) for some j. Accordingly, for each k, there's some j for which

$$(T - \lambda_j I)(v_k) = v_{k+1} \text{ or } 0$$

because v_k is a member of a Jordan basis for $G(\lambda_j)$. So for each k we can find a j for which

$$T(v_k) = \lambda_j v_k + v_{k+1} \text{ or } \lambda_j v_k .$$

Given $\epsilon > 0$, define a new basis for V as follows:

$$(w_1, w_2, w_3, \ldots, w_n) = (\epsilon^{n-1} v_1, \epsilon^{n-2} v_2, \epsilon^{n-3} v_3, \ldots, \epsilon v_{n-1}, v_n) ,$$

and let $\| \ \|_1$ be the 1-norm associated with the w-basis. Note that for each k we can find a j for which

$$\begin{aligned} T(w_k) &= \lambda_j \epsilon^{n-k} v_k + \epsilon^{n-k} v_{k+1} \text{ or } \lambda_j \epsilon^{n-k} v_k \\ &= \lambda_j w_k + \epsilon w_{k+1} \text{ or } \lambda_j w_k . \end{aligned}$$

Because $\|w_k\|_1 = 1$ for all k,

$$\|T(w_k)\|_1 \le \rho(T)\|w_k\|_1 + \epsilon\|w_{k+1}\|_1 = \rho(T) + \epsilon \text{ for all } k$$

where $\rho(T)$ is the spectral radius of T. If $v \in V$ has expansion $v = \sum_{k=1}^{n} w_k$, then

$$\begin{aligned} \|T(v)\|_1 &\le \sum_{k=1}^{n} |c_k| \, \|T(w_k)\|_1 \\ &\le (\rho(T) + \epsilon) \sum_{k=1}^{n} |c_k| \\ &= (\rho(T) + \epsilon)\|v\|_1 . \end{aligned}$$

Since $\epsilon > 0$ was arbitrary, we conclude that for any $\zeta > \rho(T)$ we can find a basis for V with associated 1-norm $\| \; \|_1$ so that

$$\left\|T^k(v)\right\|_1 \leq \zeta^k \|v\|_1 \text{ for all } v \in V ,$$

and the conclusion of Theorem 14.13 follows in turn from Theorem 4.12 on equivalence of norms. In particular, if $\rho(T) < 1$ we know that $T^k(v) \to 0$ as $k \to \infty$ for all $v \in V$.

Existence of Jordan bases

It's time to prove that a Jordan basis exists for for nilspace(S) whenever V is finite-dimensional and $S \in \operatorname{End}(V)$. My argument is constructive and produces an array of vectors. The vectors in the array constitute a linearly independent spanning set for nilspace(S), and the resulting Jordan basis consists of the list of vectors you obtain by concatenating the rows of the array. Each row in the array displays exactly one chain in the Jordan basis, so the number of rows in the array is the dimension of nullspace(S). A typical row in the array looks like

$$w_{i0} \quad w_{i1} \quad w_{i2} \quad . \quad . \quad . \quad w_{il} \; ,$$

where $w_{ik} = S^k(w_{i0})$ for $0 \leq k \leq l$. The root of the chain in this row is w_{i0} and the terminus is w_{il}, so $S(w_{il}) = 0$. Now let's get to work.

Begin by setting

$$N_k = \text{nullspace}(S) \cap \text{range}\left(S^k\right) \quad \text{for } 0 \leq k \leq d$$

and let $n_k = \dim(N_k)$ for each k. Note that $N_0 = \text{nullspace}(S)$. Furthermore, because nullspace$(S) \subset$ nilspace(S) and because nilspace(S) and range(S^d) are disjoint by Theorem 14.5, $N_d = \{0\}$. Let d^* be the smallest k-value for which $N_k = \{0\}$. Because range$\left(S^k\right) \supset$ range$\left(S^{k+1}\right)$ for all $k \geq 0$, we have the chain of inclusions

$$\text{nullspace}(S) = N_0 \supset N_1 \supset N_2 \supset \cdots \supset N_{d^*} = \{0\} \; .$$

Two extreme cases are worth mentioning. In one case $d^* = d$, which is the largest possible value for d^*. By definition of d^* we have

$$N_{d-1} = \text{nullspace}(S) \cap \text{range}\left(S^{d-1}\right) \neq \{0\} \; ,$$

If w is a nonzero vector in N_{d-1}, then $S(w) = 0$ and $w = S^{d-1}(v)$ for some $v \in V$. Since $S^d(v) = 0$, v lies in nilspace(S), and an argument such as the one featured in the discussion leading up to Theorem 14.5 enables us to conclude that

$$\left\{v, S(v), S^2(v), \ldots, S^{d-1}(v)\right\}$$

is a linearly independent subset of the d-dimensional nilspace(S) and hence a linearly independent spanning set for nilspace(S). In this case, since $S^k(v) \neq 0$ when $k < d$, the nullspace of S is the one-dimensional subspace span$\left(\left\{S^{d-1}(v)\right\}\right)$. Furthermore,

$$N_k = \text{nullspace}(S) \text{ when } 0 \leq k < d \; ,$$

so all the N_k are all the same one-dimensional subspace of V when $k < d$. In particular, nilspace(S) has dimension d and nullspace(S) has dimension 1. The other extreme case has $d^* = 1$, which is the smallest possible value for d^*. In this case, by definition of d^*,

$$N_1 = \text{nullspace}(S) \cap \text{range}(S) = \{0\} ,$$

implying that every $v \in V$ satisfying $S^k(v) = 0$ for some $k > 0$ also satisfies $S(v) = 0$, which is the same as saying that nilspace(S) = nullspace(S).

Time now to build the promised array. Recall that each row in the array will exhibit a chain from a Jordan basis for nilspace(S). The procedure I'm about to describe fills the array from top to bottom and essentially from right to left — you'll see what I mean, although it gets a bit hairy at times. The procedure features a nonnegative integer variable k that indexes columns of the array and a set of vectors $\mathcal{N}$ that fills up as we progress. I'll denote the number of vectors in $\mathcal{N}$ by $|\mathcal{N}|$

14.14 Procedure: Initialize by setting $k = d^* - 1$ and $\mathcal{N} = \phi$, so $|\mathcal{N}| = 0$. Now go to Step 1.

Step 1: If the vectors in $\mathcal{N}$ span N_k, go to Step 2. If not, extend $\mathcal{N}$ to a linearly independent spanning set for N_k by adding vectors. List those vectors in column k of the array in rows $|\mathcal{N}| + 1$ through n_k. The vector you place in row i is of the form $S^k(w_{i0})$ for some $w_{i0} \in R_0$. Fill in row i for $|\mathcal{N}| + 1 \leq i \leq n_k$ as follows:

$$w_{i0} \quad S(w_{i0}) \quad S^2(w_{i0}) \quad . \quad . \quad . \quad S^k(w_{i0}) \,.$$

Now go to Step 2.

Step 2: If $k = 0$, stop. If $k > 0$, decrement k by 1 and return to Step 1. □

Now for a couple of comments. Note first that every vector in the array generated by Procedure 14.14 lies in nilspace(S). Say, for example, that you fill in row i as indicated in the description of Step 1. Since $S^k(w_{i0}) \in$ nullspace(S), $S^{k+1}(w_{i0}) = 0$, so every vector in row i is annihilated by some power of S between 1 and $k + 1$. Second, note that when the algorithm stops the set $\mathcal{N}$ is a linearly independent spanning set for nullspace(S). When you start instance k of Step 1, $\mathcal{N}$ contains a linearly independent spanning set for N_{k+1}. It might happen that $N_k = N_{k+1}$, in which case you add no vectors to $\mathcal{N}$ at that point. Performing instance $k = 0$ of Step 1 completes $\mathcal{N}$ to a linearly independent spanning set for N_0 = nullspace(S). The vectors in $\mathcal{N}$ are the rightmost vectors in the rows of the array.

Let's see how things go in the extreme cases I mentioned earlier. If $d^* = d$, instance $k = d^* - 1$ of Step 1 fills row 1 of the array with a chain of vectors

$$w_{10} \quad S(w_{10}) \quad S^2(w_{10}) \quad . \quad . \quad . \quad S^{d-1}(w_{10})$$

that constitute a linearly independent spanning set for nilspace(S). Subsequent instances of Step 1 add no additional vectors to the array because all the N_k are the same subspace span$\left(\left\{S^{d-1}(w_{10})\right\}\right)$. For the other extreme $d^* = 1$, the procedure terminates after one round because the first instance of Step 1 occurs when $k = 0$. The resulting array consists of a single column containing d vectors that constitute a linearly independent spanning set for nullspace(S), which is the same in this case as nilspace(S).

To understand in general why the vectors produced by Procedure 14.14 form a linearly independent spanning set for nilspace(S), let's first apply an inductive argument to confirm that the vectors are linearly independent. Consider the rightmost column of the array, indexed by $k = d^* - 1$. By construction in Step 1, the vectors in that column are linearly independent. Now suppose we've proven linear independence of the vectors in columns strictly to the right of column k (i.e. columns indexed by $k + 1$ or greater). Applying S to a linear dependence relation between vectors in columns to the right of and including column k results in a dependence relation between vectors in columns strictly to the right of column k because the array's structure features chains across the rows. The coefficients in this second dependence relation must be zero by induction, but some coefficients in the original dependence relation — the coefficients of the rightmost vectors in the array — disappear when we apply S because S annihilates the rightmost vectors in the array. Thus there remains an unaccounted-for zero linear combination of the rightmost vectors in the array, but that's no cause for alarm because those vectors, which are members of $\mathcal{N}$, form a linearly independent spanning set for nullspace(S), and their coefficients must therefore be zero as well. The bottom line: only the trivial linear combination of vectors in the array yields zero, and the vectors in the array are linearly independent.

To see why the vectors in the array span nilspace(S), first let's count them up. I claim that column k of the array contains exactly n_k vectors. You can see this easily by referring to Step 1. When we reach instance k of Step 1, which is our last opportunity to put vectors in column k, one of two things can happen. Either we put no new vectors in column k, which means $n_k = n_{k+1}$, or we fill out column k through row n_k. In either case, we end up with exactly n_k vectors in column k. Accordingly, the array contains

$$n_0 + n_1 + n_2 + \cdots + n_{d^*-1}$$

vectors when all is said and done. If we can show that this number matches the dimension of nilspace(S), we'll know that the vectors in the array span nilspace(S).

To that end, let $R_0 = \text{nilspace}(S)$ and let

$$R_k = \text{nilspace}(S) \cap \text{range}\left(S^k\right)$$

for each $k > 0$. As with the N_k, $R_k \supset R_{k+1}$ for all $k \geq 0$. Furthermore, $R_d = \{0\}$ by Theorem 14.5. In fact, $R_{d^*} = \{0\}$. To see why, suppose k is such that $R_{k-1} \neq \{0\}$ but $R_k = \{0\}$. Then there's some $v \in V$ such that $S^{k-1}(v) \neq 0$ and $S^k(v) = 0$, so not only is $S^{k-1}(v)$ in R_{k-1}, but $S^{k-1}(v) \in N_{k-1}$, so $k \leq d^*$ because $N_{d^*} = \{0\}$. Thus $R_{d^*} = \{0\}$ and we have the chain of inclusions

$$\text{nilspace}(S) = R_0 \supset R_1 \supset R_2 \supset \cdots \supset R_{d^*} = \{0\}\ .$$

14.15 Fact: With notation as in the foregoing,

$$n_k + \dim(R_{k+1}) = \dim(R_k) \text{ for } 0 \leq k < d^*\ ,$$

Furthermore, $\dim(R_{d^*-1}) = n_{d^*-1}$, so

$$\dim(\text{nilspace}(S)) = \dim(R_0) = n_0 + n_1 + n_2 + \cdots + n_{d^*-1}\ .$$

Proof: Each R_k is invariant under S, so we can consider the restriction $S_k \in \mathrm{End}(R_k)$ of the mapping S to the subspace R_k. A vector w lies in R_1 if and only if w lies in nilspace(S) and $w = S(v)$ for some $v \in V$. Stipulating $w \in$ nilspace(S) requires $v \in$ nilspace(S) as well, because $S^d(w) = 0$ implies that $S^{d+1}(v) = 0$. Accordingly, $v \in R_0$, which means that R_1 is precisely the set of vectors you get by applying S to vectors in R_0 — i.e. $R_1 = \mathrm{range}(S_{)})$. A similar argument demonstrates that

$$R_{k+1} = \{S(v) : v \in R_k\} = \mathrm{range}(S_k) \ \text{ for } \ 1 \le k < d^* \ .$$

Next, observe that

$$N_k = \mathrm{nullspace}(S_k) \ \text{ for } \ 1 \le k < d^* \ ,$$

so from Theorem 4.9 we obtain the recursion

$$n_k + \dim(R_{k+1}) = \dim(R_k) \ \text{ for } \ 0 \le k < d^* \ .$$

Since $R_{d^*} = \{0\}$, $\dim(R_{d^*-1}) = n_{d^*-1}$. The formula for dim(nilspace(S)) results from iterating the recursion from $k = d^* - 1$ to $k = 0$ starting with boundary condition $\dim(R_{d^*-1}) = n_{d^*-1}$. □

Thus the vectors in the array generated by Procedure 14.14 constitute a linearly independent spanning set for nilspace(S). Laying the rows end to end results in an ordered list of vectors that parses into chains — to wit, a Jordan basis for nilspace(S).

Eigenvalues and eigenvectors of matrices

For $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$, denote by $\mathbb{F}^{n\times n}$ the set of all $(n \times n)$ matrices with entries in $\mathbb{F}$. Any $A \in \mathbb{F}^{n\times n}$ defines a linear mapping $T_A \in \mathrm{End}(\mathbb{C}^n)$ via the prescription

$$T_A(v) = Av \ \text{ for all } \ v \in \mathbb{C}^n \ .$$

Here I'm regarding $\mathbb{C}^n$ as the n-dimensional complex vector space consisting of all n-dimensional column vectors with complex entries. Note that T_A as I've defined it lies in $\mathrm{End}(\mathbb{C}^n)$ even when A has real entries. If $0_{n\times n}$ is the $(n \times n)$ matrix of zeroes, then $T_{0_{n\times n}}$ is clearly the zero mapping on $\mathbb{C}^n$, and if $I_{n\times n}$ is the $(n \times n)$ identity matrix, then $T_{I_{n\times n}}$ is the identity mapping on $\mathbb{C}^n$. If A and B are in $\mathbb{F}^{n\times n}$, then

$$T_{AB}(v) = ABv = A(Bv) = T_A(T_B(v)) \ \text{ for all } \ v \in \mathbb{C}^n \ ,$$

from which follows the convenient fact that $T_{AB} = T_A T_B$.

A matrix $A \in \mathbb{F}^{n\times n}$ is *invertible* when there exists a matrix $A^{-1} \in \mathbb{F}^{n\times n}$ such that $A^{-1}A = AA^{-1} = I_{n\times n}$. If A is invertible, then

$$T_A T_{A^{-1}} = T_{AA^{-1}} = T_{I_{n\times n}}$$

and

$$T_{A^{-1}} T_A = T_{A^{-1}A} = T_{I_{n\times n}} \ .$$

Since $T_{I_{n\times n}}$ is the identity mapping on $\mathbb{C}^n$, it follows that if A is invertible, then T_A is linearly invertible and is therefore bijective by Theorem 4.8. By Fact 4.7, T_A is injective if and only if nullspace$(T_A) = 0$, which by definition of T_A is equivalent to the statement that $Av = 0$ if and only if $v = 0$. Similarly, T_A is surjective if and only if for every $w \in \mathbb{C}^n$ there exists $v \in \mathbb{C}^n$ such that $Av = w$. Theorem 4.10 establishes the equivalence of invertibility, bijectivity, surjectivity, and injectivity of T_A, thus almost proving the following parallel assertion about matrices.

14.16 Theorem: The following conditions on $A \in \mathbb{F}^{n\times n}$ are equivalent in the sense that any one of them implies the other two:

- A is invertible
- The only $v \in \mathbb{C}^n$ satisfying $Av = 0$ is $v = 0$
- For every $w \in \mathbb{C}^n$ there exists a $v \in \mathbb{C}^n$ such that $Av = w$

—

Proof: We've shown so far that the last two bullet points are equivalent and that *if* A is invertible, then T_A is invertible, so both of the last two bullet points hold. How do we know that A is invertible when the last two bullet points hold? Assuming they do, let e^j be the jth column of $I_{n\times n}$ for $1 \leq j \leq n$. Invoking the third bullet point, we can solve for $b^j \in \mathbb{C}^n$ such that $Ab^j = e^j$ for all j. The matrix B with jth column b^j then satisfies

$$AB = I_{n\times n} \ .$$

This last equation implies that the only $v \in \mathbb{C}^n$ satisfying $Bv = 0$ is $v = 0$, which is equivalent to saying that the columns of B are linearly independent, from which it follows that $(b^1, b^2, \ldots, b^n)$ is a basis for the n-dimensional vector space $\mathbb{C}^n$. Because $AB = I_{n\times n}$,

$$BAB = (BA)B = B \ ,$$

so $(BA)b^j = b^j$ for all j. By writing an arbitrary $v \in \mathbb{C}^n$ as a linear combination of the b^j, you'll see that $BAv = v$ for every $v \in \mathbb{C}^n$, from which it follows that $BA = I_{n\times n}$. Thus $AB = BA = I_{n\times n}$, so A is invertible and $A^{-1} = B$. □

If you've been reading carefully, you might have noticed a loose end hanging off of Theorem 14.16. If A has real entries and is invertible, does A^{-1} necessarily have real entries? The answer is yes, and perhaps your best bet for figuring out why is to proceed as follows. Note first that when A is real, T_A restricts to linear mapping $T_A^{\mathbb{R}}$ in End$(\mathbb{R}^n)$. Now carry through the argument leading up to and including the statement and proof of Theorem 14.16 substituting $T_A^{\mathbb{R}}$ for T_A and $\mathbb{R}$ for $\mathbb{C}$ and for $\mathbb{F}$ everywhere you can. Believe me, it works.

An eigenvalue of a matrix $A \in \mathbb{F}^{n\times n}$ is simply an eigenvalue of the corresponding linear mapping T_A. Similarly, eigenvectors and generalized eigenvectors of A. are just eigenvectors and generalized eigenvectors of T_A. Trudging onward, we identify the eigenspaces, generalized eigenspaces, and algebraic and geometric multiplicities of eigenvalues of A with the corresponding objects associated with T_A. We also say that the matrix A is diagonalizable if and only if T_A is diagonalizable.

If A has real entries, i.e. $A \in \mathbb{R}^{n\times n}$, its eigenvalues and eigenvectors possess some important symmetry properties.

14.17 Fact: If $A \in \mathbb{R}^{n\times n}$, its eigenvalues and eigenvectors come in complex-conjugate pairs in the following sense. If $\lambda_o \in \mathbb{C}$ is an eigenvalue of A, then so is $\overline{\lambda_o}$. If v_o is an eigenvector corresponding to eigenvalue λ_o, then $\overline{v_o}$ is an eigenvector corresponding to eigenvalue $\overline{\lambda_o}$, where $\overline{v_o}$ is the elementwise complex conjugate of v_o.

Proof: If λ_o is an eigenvalue of A and v_o a corresponding eigenvector, then $v_o \neq 0$ and $Av_o = \lambda_o v_o$. Take the elementwise complex conjugate of this last relation and you obtain $A\overline{v_o} = \overline{\lambda_o}\overline{v_o}$ since A has real entries, and the result follows because $\overline{v_o} \neq 0$. □

Given $A \in \mathbb{F}^{n\times n}$, it's possible that some or all of A's eigenvalues are real. If A is real and has a real eigenvalue, then we can always find bases for the corresponding eigenspace and generalized eigenspace consisting solely of real vectors.

14.18 Fact: If $\lambda_o \in \mathbb{R}$ is an eigenvalue of $A \in \mathbb{R}^{n\times n}$, then there exist bases for $E(\lambda_o)$ and $G(\lambda_o)$ consisting solely of real vectors.

Proof: As usual, let $m_o = \dim(E(\lambda_o))$ and $d_o = \dim(G(\lambda_o))$ be, respectively, the geometric and algebraic multiplicities of λ_o. If $(v_1, v_2, \ldots, v_{m_o})$ is a basis for $E(\lambda_o)$, we can write

$$v_k = \mathrm{Re}\{v_k\} + j\mathrm{Im}\{v_k\}$$

for all k, where $j = \sqrt{-1}$. Since each v_k is an eigenvector of A corresponding to λ_o, we have $Av_k = \lambda_o v_k$ for all k, so

$$A\,\mathrm{Re}\{v_k\} + jA\,\mathrm{Im}\{v_k\} = \lambda_o\mathrm{Re}\{v_k\} + j\lambda_o\mathrm{Im}\{v_k\}\ .$$

Equating real and imaginary parts and keeping in mind that both λ_o and A are real, we have

$$A\,\mathrm{Re}\{v_k\} = \lambda_o\mathrm{Re}\{v_k\} \ \text{ and } \ A\,\mathrm{Im}\{v_k\} = \lambda_o\mathrm{Im}\{v_k\}$$

for all k. It follows that every nonzero vector in the set

$$\{\mathrm{Re}\{v_k\} : 1 \le k \le m_o\} \cup \{\mathrm{Im}\{v_k\} : 1 \le k \le m_o\}$$

is an eigenvector of A corresponding to eigenvalue λ_o, hence lies in $E(\lambda_o)$. It's clear that the vectors in this set are real and span $E(\lambda_o)$. Procedure 1 from the discussion preceding Lemma 4.2 reduces it to a linearly independent spanning set that we can assemble into a basis for $E(\lambda_o)$ consisting solely of real vectors.

A similar argument works for the generalized eigenspace $G(\lambda_o)$. If $(v_1, v_2, \ldots, v_{d_o})$ is a basis for $G(\lambda_o)$, then $(A - \lambda_o I)^{d_o} v_k = 0$ for all k, so

$$(A - \lambda_o I)^{d_o}\mathrm{Re}\{v_k\} + j(A - \lambda_o I)^{d_o}\mathrm{Im}\{v_k\} = 0\ .$$

Equating real and imaginary parts and keeping in mind that both λ_o and A are real, we have

$$(A - \lambda_o I)^{d_o}\mathrm{Re}\{v_k\} = (A - \lambda_o I)^{d_o}\mathrm{Im}\{v_k\} = 0$$

for all k. It follows that every nonzero vector in the set

$$\{\mathrm{Re}\{v_k\} : 1 \le k \le d_o\} \cup \{\mathrm{Im}\{v_k\} : 1 \le k \le d_o\}$$

is a generalized eigenvector of A corresponding to eigenvalue λ_o, hence lies in $G(\lambda_o)$. The vectors in this set are real and span $G(\lambda_o)$, so we can reduce it to a linearly independent spanning set and build a basis for $G(\lambda_o)$ consisting solely of real vectors. □

The interplay between matrices and linear mappings goes both ways. If V is any n-dimensional vector space over $\mathbb{F}$, $\mathbf{v} = (v_1, v_2, \ldots, v_n)$ is a basis for V, and $T \in \operatorname{End}(V)$, then we can find numbers $t_{ij} \in \mathbb{F}$ such that

$$T(v_j) = \sum_{i=1}^{n} t_{ij} v_i \ \text{ for } \ 1 \le j \le n \ .$$

The *matrix of* T *with respect to the basis* $\mathbf{v}$ is the matrix $\operatorname{Mat}_{\mathbf{v}}(T) \in \mathbb{F}^{n\times n}$ with specification

$$[\operatorname{Mat}_{\mathbf{v}}(T)]_{ij} = t_{ij} \ \text{ for } \ 1 \le i, j \le n \ .$$

I'll leave it for you to verify that the matrices with respect to any basis for V of the zero and identity mappings on V are $0_{n\times n}$ and $I_{n\times n}$, respectively. If $A \in \mathbb{F}^{n\times n}$ is an arbitrary matrix and $\mathbf{e}$ is the standard basis for $\mathbb{C}^n$ that we encountered in Chapter 4, then

$$\operatorname{Mat}_{\mathbf{e}}(T_A) = A \ .$$

In other words, A is the matrix of T_A with respect to the standard basis for $\mathbb{C}^n$. You can see this easily by noting that

$$T_A(e^j) = Ae^j = \sum_{i=1}^{n} [A]_{ij} e^i \ \text{ for all } \ j$$

because Ae^j is simply the jth column of A.

If for a $v \in V$ with expansion $v = \sum_{j=1}^{n} c_j v_j$ we define $\operatorname{vec}_{\mathbf{v}}(v)$ as the column vector in $\mathbb{F}^n$ with jth element c_j, it's easy to show that

$$\operatorname{vec}_{\mathbf{v}}(T(v)) = \operatorname{Mat}_{\mathbf{v}}(T) \operatorname{vec}_{\mathbf{v}}(v) \ \text{ for all } \ v \in V \ .$$

It's also a simple matter to show that for any S and T in $\operatorname{End}(V)$ we have

$$\operatorname{Mat}_{\mathbf{v}}(ST) = \operatorname{Mat}_{\mathbf{v}}(S) \operatorname{Mat}_{\mathbf{v}}(T) \ .$$

A bijective $T \in \operatorname{End}(V)$ possesses a linear inverse mapping by Theorem 4.10. Let's call that inverse mapping T^{-1}. Since $T^{-1}T = TT^{-1} = I$, it must be the case that

$$\operatorname{Mat}_{\mathbf{v}}\left(T^{-1}\right) \operatorname{Mat}_{\mathbf{v}}(T) = \operatorname{Mat}_{\mathbf{v}}(T) \operatorname{Mat}_{\mathbf{v}}\left(T^{-1}\right) = I_{n\times n} \ ,$$

so $\operatorname{Mat}_{\mathbf{v}}(T)$ is invertible in the matrix sense when T is bijective, and

$$\left(\operatorname{Mat}_{\mathbf{v}}(T)\right)^{-1} = \operatorname{Mat}_{\mathbf{v}}\left(T^{-1}\right) \ .$$

Suppose $T \in \operatorname{End}(V)$ is diagonalizable. Let λ_1, λ_2, ... , λ_n be the not necessarily distinct eigenvalues of T and let v_1, v_2, ... , v_n be corresponding linearly independent eigenvectors that span V. Let $\mathbf{v} = (v_1, v_2, \ldots, v_n)$ be the resulting basis for V. Since $T(v_j) = \lambda_j v_j$ for all j,

$$[\operatorname{Mat}_{\mathbf{v}}(T)]_{ij} = \begin{cases} \lambda_j & \text{when } i = j \\ 0 & \text{when } i \ne j \ . \end{cases}$$

Thus when T is diagonalizable you can find a basis $\mathbf{v}$ for V so that the matrix of T with respect to $\mathbf{v}$ is a diagonal matrix, whence the term "diagonalizable."

To discover the implications of diagonalizabilty for matrices, recall that saying A is diagonalizable is the same as saying T_A is diagonalizable, which means that there exists a basis $\mathbf{v} = (v_1, v_2, \ldots, v_n)$ for $\mathbb{C}^n$ consisting solely of eigenvectors of T_A, where v_j corresponds to eigenvalue λ_j for each j. Form the matrix $X \in \mathbb{C}^{n \times n}$ whose jth column is v_j for each j. Linear independence of the v_j implies that the only $v \in \mathbb{C}^n$ satisfying $Xv = 0$ is $v = 0$. Accordingly, by Theorem 14.16, X is invertible. Furthermore, since $Av_j = T_A(v_j) = \lambda_j v_j$ for all j, we have

$$AX = X\Lambda \ ,$$

where $\Lambda \in \mathbb{C}^{n \times n}$ is the diagonal matrix with λ_j at diagonal position j. Thus arise the two equations

$$X^{-1}AX = \Lambda \quad \text{and} \quad A = X\Lambda X^{-1} \ ,$$

which you probably recognize from your earlier training in linear algebra.

CHAPTER 15

Singular-Value Decomposition

The real world offers countless opportunities for error. Noise corrupts measurements, computers perform arithmetic with finite precision, and mathematical models are approximations at best. At times we even make mistakes. Some errors are undoubtedly worse than others, and we're often called upon to assess the relative impacts of different errors and of different types of errors on the effectiveness of solutions to engineering problems. The singular-value decomposition breathes quantitative life into some such assessments in the context of problems involving matrices. Suppose, for example, that A is a complex $(m \times n)$ matrix and we're interested in the equation $Av = w$, where $v \in \mathbb{C}^n$ and $w \in \mathbb{C}^m$. Think of v as raw data and w as the output we get by "processing" v through A. What's the effect on w if we specify v incorrectly, and how does the answer depend on A and on the exact nature of the incorrect v-specification? Turning things around and making w the raw data, even if we lack an exact solution to $Av = w$ we might have an algorithm that produces an approximate solution, i.e. $\widehat{v}$ such that $A\widehat{v} \approx w$. What's the best approximation possible, and how does that depend on w and A? How does an error in specifying w affect $\widehat{v}$, and how does the answer depend on A and on the exact nature of the incorrect w-specification? These are the kinds of questions the singular-value decomposition answers.

Hermitian matrices and their eigenspaces

To pin things down, I'll frame the exposition in terms of complex matrices and comment on how the results specialize when matrices are real. By $\mathbb{C}^n$ I mean the set of all column n-vectors with entries in $\mathbb{C}$, and by $\mathbb{C}^{m\times n}$ I mean the set of all $(m\times n)$ matrices with entries in $\mathbb{C}$. The $(n\times n)$ identity matrix I'll denote by $I_{n\times n}$. In Chapter 9 we noted that $\mathbb{C}^n$ is an inner product space with inner product

$$\langle v, w\rangle = w^H v = \sum_{i=1}^{n} [v]_i\, \overline{[w]_i} \;\text{ for all }\; v, w \in \mathbb{C}^n\ ,$$

where w^H is the Hermitian conjugate — i.e. the conjugate transpose — of w. In what follows, when I use the word "orthonormal" I'll be referring to orthonormality with respect to that inner product. I'll use $\|\ \|$ to denote the norm that arises from that inner product, i.e.

$$\|v\| = \langle v, v\rangle^{1/2} = \left(\sum_{i=1}^{n} |[v]_i|^2\right)^{1/2} \quad \text{for all }\; v \in \mathbb{C}^n\ .$$

For general $A \in \mathbb{C}^{m\times n}$, the Hermitian conjugate of A is the matrix $A^H \in \mathbb{C}^{n\times m}$ defined by $A^H = \overline{A^T}$ or, in terms of elements, by

$$\left[A^H\right]_{ij} = \overline{[A]_{ji}} \ \text{ for } \ 1 \leq i \leq n \text{ and } 1 \leq j \leq m \ .$$

Note that if A is real, then $A^H = A^T$. A necessarily square matrix $Q \in \mathbb{C}^{n\times n}$ is a *Hermitian matrix* if and only if $Q^H = Q$. A real Hermitian matrix Q is symmetric, i.e. $Q = Q^T$.

Now for a quick reminder of an elementary fact about matrix multiplication. If A and B are matrices whose product AB makes sense, then $(AB)^T = B^T A^T$. That's because

$$\left[(AB)^T\right]_{ij} = \sum_{k=1}^{n} [A]_{jk}\,[B]_{ki} = \sum_{k=1}^{n} \left[B^T\right]_{ik} \left[A^T\right]_{kj} = \left[B^T A^T\right]_{ij}$$

for all i and j when A has n columns. Taking conjugates reveals that $(AB)^H = B^H A^H$. Similarly, the transpose of an arbitrary matrix product is the product of the factors' transposes with the order reversed, i.e.

$$\left(A_1 A_2 \cdots A_N\right)^T = A_N^T \cdots A_2^T A_1^T \ ,$$

and the same goes for Hermitian conjugates of matrix products.

15.1 Fact: If $Q \in \mathbb{C}^{n\times n}$ is Hermitian, then all the eigenvalues of Q are real.

Proof: Let $v_o \in \mathbb{F}^n$ be an eigenvector of Q corresponding to eigenvalue λ_o. Because Q is Hermitian,

$$\overline{v_o^H Q v_o} = \left(v_o^H Q v_o\right)^H = v_o^H Q^H v_o = v_o^H Q v_o \ .$$

The first equality holds because the transpose of a number — in this case $v_o^H Q v_o$ — is the number itself, so the Hermitian conjugate of a number is just the conjugate of the number. Meanwhile, because $Qv_o = \lambda_o v_o$,

$$v_o^H Q v_o = \lambda_o \|v_o\|^2 \quad \text{and} \quad \overline{v_o^H Q v_o} = \overline{\lambda_o \|v_o\|^2} = \overline{\lambda_o} \|v_o\|^2 \ ,$$

from which it follows that $\overline{\lambda_o} = \lambda_o$ because $v_o \neq 0$. Hence λ_o is real. □

15.2 Fact: If $Q \in \mathbb{C}^{n\times n}$ is Hermitian, then Q is diagonalizable.

Proof: Recall from Chapter 14 that a matrix Q is diagonalizable if and only if every generalized eigenvector of Q is also an eigenvector of Q. Suppose λ_o is an eigenvalue of a Hermitian Q and v_o is a corresponding generalized eigenvector, which means that

$$\left(Q - \lambda_o I_{n\times n}\right)^k v_o = 0$$

for some $k > 0$. If, for example, $(Q - \lambda_o I_{n\times n})^2 v_o = 0$, multiplying on the left by v_o^H yields

$$\begin{aligned} v_o^H (Q - \lambda_o I_{n\times n})^2 v_o &= v_o^H (Q - \lambda_o I_{n\times n})^H (Q - \lambda_o I_{n\times n}) v_o \\ &= ((Q - \lambda_o I_{n\times n}) v_o)^H (Q - \lambda_o I_{n\times n}) v_o \\ &= |(Q - \lambda_o I_{n\times n}) v_o|^2 = 0 \ , \end{aligned}$$

where the first equality holds because λ_o is real by Fact 15.1 and Q is Hermitian. Thus $(Q - \lambda_o I_{n\times n}) v_o = 0$, and v_o is an eigenvector of Q corresponding to eigenvalue λ_o. If

$$(Q - \lambda_o I_{n\times n})^3 v_o = 0 \ ,$$

then multiplying on the left by $v_o^H (Q - \lambda_o I_{n\times n})$ yields

$$v_o^H (Q - \lambda_o I_{n\times n})^4 v_o = 0 \ ,$$

implying since $Q = Q^H$ and $\lambda_o \in \mathbb{R}$ that

$$v_o^H \left((Q - \lambda_o I_{n\times n})^H\right)^2 (Q - \lambda_o I_{n\times n})^2 v_o = \left|(Q - \lambda_o I_{n\times n})^2 v_o\right|^2 = 0 \ ,$$

so $(Q - \lambda_o I_{n\times n})^2 v_o = 0$, and v_o is an eigenvector of Q corresponding to eigenvalue λ_o by our earlier work. I hope you can see how to carry this argument further to show that, for any $k > 0$,

$$(Q - \lambda_o I_{n\times n})^k v_o = 0 \implies Qv_o = \lambda_o v_o \ ,$$

so every generalized eigenvector of Q is also an eigenvector of Q, and Q is therefore diagonalizable. □

Thus if $Q \in \mathbb{C}^{n\times n}$ is Hermitian with distinct eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_s$ and corresponding eigenspaces $E(\lambda_1), E(\lambda_2), \dots, E(\lambda_s)$, we have the vector sum decomposition

$$\mathbb{C}^n = E(\lambda_1) + E(\lambda_2) + \cdots + E(\lambda_s) \ .$$

The $E(\lambda_j)$ are mutually disjoint by Theorem 14.2, but much more is true. Eigenvectors of Q corresponding to distinct eigenvalues are not just linearly independent but orthogonal.

15.3 Fact: Let $Q \in \mathbb{C}^{n\times n}$ be Hermitian and let $\lambda_1, \lambda_2, \dots, \lambda_s$ be the distinct eigenvalues of Q. If $v_j \in E(\lambda_j)$ and $v_k \in E(\lambda_k)$ with $j \neq k$, $\langle v_j, v_k\rangle = 0$.

Proof: Start with

$$Qv_j = \lambda_j v_j \ \text{ and } \ Qv_k = \lambda_k v_k \ ,$$

from which it follows that

$$(v_k)^H Qv_j = \lambda_j (v_k)^H v_j = \lambda_j \langle v_j, v_k\rangle$$

and

$$(v_j)^H Qv_k = \lambda_k (v_j)^H v_k = \lambda_k \langle v_k, v_j\rangle \ .$$

Because Q is Hermitian, taking the complex conjugate of the second expression keeping in mind that λ_k is real yields

$$\lambda_k \langle v_j, v_k \rangle = (v_k)^H Q v_j = \lambda_j \langle v_j, v_k \rangle \ ,$$

so $\langle v_k, v_j \rangle = 0$ because $\lambda_j \neq \lambda_k$. □

By Fact 9.9 we can find for each eigenspace an orthonormal basis. Stringing these bases together results in an orthonormal basis for $\mathbb{C}^n$ because any two basis vectors lying in different eigenspaces are orthogonal by Fact 15.3. Thus we've arrived at a proof of the following fundamental result.

15.4 Theorem: If $Q \in \mathbb{C}^{n\times n}$ is Hermitian, there exists an orthonormal basis for $\mathbb{C}^n$ consisting solely of eigenvectors of Q. □

When Q is real and Hermitian, which is the same as saying that Q is real and symmetric, we can refine Theorem 15.4 by applying Fact 14.18. First find a basis for each $E(\lambda_j)$ consisting solely of real vectors. Then follow the Gram-Schmidt procedure outlined in the proof of Fact 9.9 to transform that basis into an orthonormal basis for $E(\lambda_j)$. Since the Gram-Schmidt procedure operating on real vectors produces real vectors, the resulting orthonormal basis for $E(\lambda_j)$ consists solely of real vectors. Accordingly, associated with any real Hermitian Q is an orthonormal basis for $\mathbb{C}^n$ consisting solely of real eigenvectors of Q.

Suppose now that Q is Hermitian and $(v_1, v_2, \ldots, v_n)$ is an orthonormal basis for $\mathbb{C}^n$ consisting of eigenvectors of Q, where v_j corresponds to eigenvalue λ_j for $1 \leq j \leq n$. Here I'm not assuming that the λ_j are distinct. Form a matrix $U \in \mathbb{C}^{n\times n}$ whose jth column is v_j for each j. Note that

$$\left[U^H U\right]_{ij} = v_i^H v_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \ , \end{cases}$$

so $U^H U = I_{n\times n}$. Thus U is invertible and $U^{-1} = U^H$. A matrix with that property is called a *unitary matrix.* Since the jth column of QU is $\lambda_j v_j$ for all j,

$$QU = U\Lambda \ ,$$

where Λ is the diagonal matrix with eigenvalue λ_j at position (j, j) for all j. Accordingly, there exists a unitary matrix U such that

$$U^H Q U = \Lambda \ \text{ and } \ Q = U \Lambda U^H \ ,$$

which is why people often say that a Hermitian matrix is unitarily diagonalizable. If Q is real, we can take U to be real, so $U^H = U^T$ and therefore $U^{-1} = U^T$. A matrix whose inverse is its transpose is called an *orthogonal matrix,* which is why people say that a real symmetric matrix is orthogonally diagonalizable.

The singular-value decomposition

Every $A \in \mathbb{C}^{m\times n}$ defines a linear mapping $T_A \in \mathrm{Hom}(\mathbb{C}^n, \mathbb{C}^m)$ by means of the prescription

$$T_A(v) = Av \ \text{ for all } \ v \in \mathbb{C}^n \ .$$

The *nullspace of* A is the nullspace of the mapping T_A and the *range of* A is the range of T_A. In more pedestrian terms,

$$\text{nullspace}(A) = \{v \in \mathbb{C}^n : Av = 0\}$$

and

$$\text{range}(A) = \{w \in \mathbb{C}^m : w = Av \text{ for some } v \in \mathbb{C}^n\} \ .$$

The *rank* of A is the dimension of range(A). By Theorem 4.9,

$$\text{rank of } A = n - \dim(\text{nullspace}(A)) \ .$$

If $A \in \mathbb{C}^{m\times n}$, then the matrix $A^HA \in \mathbb{C}^{n\times n}$ is Hermitian. Furthermore, the rank of A^HA is the same as the rank of A. To see this, first observe that A and A^HA have the same nullspace because

$$A^HAv = 0 \Longrightarrow v^HA^HAv = 0 \Longleftrightarrow (Av)^H\, Av = 0 \Longleftrightarrow \|Av\|^2 = 0 \Longleftrightarrow Av = 0 \ ,$$

so the nullspace of A^HA is contained in the nullspace of A. The reverse inclusion is obvious. Thus

$$\text{rank}(A) = n - \dim(\text{nullspace}(A)) = n - \dim\left(\text{nullspace}\left(A^HA\right)\right) = \text{rank}\left(A^HA\right) \ .$$

Since A^HA is Hermitian, all its eigenvalues are real. In fact, they're all nonnegative. To see why, note first that if λ_o is an eigenvalue of A^HA and v_o a corresponding eigenvector, then

$$0 \le \|Av_o\|^2 = v_o^HA^HAv_o = \lambda_o\|v_o\|^2 \ ,$$

which implies that $\lambda_o \ge 0$ because $v_o \ne 0$. Suppose that A, and hence A^HA, have rank r. The nullspace of A^HA has dimension $n - r$, and a nonzero vector v is in the nullspace of A^HA if and only if it is an eigenvector of A^HA corresponding to eigenvalue 0. Accordingly, if we invoke Theorem 15.4 and pick an orthonormal basis $(v_1, v_2, \ldots, v_n)$ for $\mathbb{C}^n$ consisting solely of eigenvectors for A^HA, we can choose the v_j so that $(v_{r+1}, v_{r+2}, \ldots, v_n)$ is an orthonormal basis for the nullspace of A^HA. If we do that, then each v_j for $1 \le j \le r$ is an eigenvector of A^HA corresponding to a nonzero eigenvalue λ_j of A^HA. If $r = n$, which means that the nullspace of A is $\{0\}$, then all the v_j for $1 \le j \le n$ are eigenvectors corresponding to nonzero eigenvalues of A^HA.

Since the eigenvalues of A^HA are real and nonnegative, we can order the eigenvectors $\{v_j : 1 \le j \le r\}$ so that the necessarily strictly positive eigenvalues to which they correspond occur in decreasing order

$$\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_r > 0 \ .$$

15.5 Definition: Let $A \in \mathbb{C}^{m\times n}$ have rank r. The *singular values* of A are the positive square roots $\{\sigma_j : 1 \le j \le r\}$ of the nonzero eigenvalues $\{\lambda_j : 1 \le j \le r\}$ of A^HA. I.e. $\sigma_j = \sqrt{\lambda_j}$ for $1 \le j \le r$.

Next define vectors $w_j \in \mathbb{C}^m$ by

$$w_j = \frac{1}{\sigma_j} A v_j \ \text{ for } \ 1 \leq j \leq r \ .$$

The w_j are orthonormal because

$$\begin{aligned} \langle w_j, w_k \rangle &= \left\langle \sigma_j^{-1} A v_j, \sigma_k^{-1} A v_k \right\rangle \\ &= (\sigma_j \sigma_k)^{-1} (v_k)^H A^H A v_j \\ &= \lambda_j (\sigma_j \sigma_k)^{-1} \langle v_j, v_k \rangle \\ &= \begin{cases} 1 & \text{if } j = k \\ 0 & \text{if } j \neq k \ , \end{cases} \end{aligned}$$

where the third line holds because v_j is an eigenvector of $A^H A$ corresponding to eigenvalue λ_j and the last line follows from the definition of σ_j and σ_k along with the orthonormality of the v_j.

Because the w_j are orthonormal, they're linearly independent. The w_j also lie in the range of A since, for each j, w_j is a scalar multiple of Av_j. Since the range of A has dimension r, $(w_1, w_2, \ldots, w_r)$ is an orthonormal basis for the range of A. If $v \in \mathbb{C}^n$ is any vector, then $w = Av$ is in the range of A, so we can expand w in terms of the orthonormal basis as

$$w = \sum_{j=1}^{r} \langle w, w_j \rangle \ w_j \ .$$

Plugging in $w = Av$ and $w_j = (1/\sigma_j) A v_j$ yields

$$\begin{aligned} Av &= \sum_{j=1}^{r} \left\langle Av, \frac{1}{\sigma_j} A v_j \right\rangle w_j \\ &= \sum_{j=1}^{r} \frac{1}{\sigma_j} \left((v_j)^H A^H A v \right) w_j \\ &= \sum_{j=1}^{r} \left(\frac{\lambda_j}{\sigma_j} (v_j)^H v \right) w_j \\ &= \left(\sum_{j=1}^{r} \sigma_j w_j (v_j)^H \right) v \ \text{ for all } \ v \in \mathbb{C}^n \ . \end{aligned}$$

The third line follows from the second line because v_j is an eigenvector of $A^H A$ corresponding to eigenvalue λ_j. The term on the third line multiplying w_j is a scalar, so I moved the w_j to the left to obtain the fourth line.

The fourth line is of interest because the expression in parentheses is actually another way of writing the matrix A. The chain of equalities reads

$$Av = (\quad) \, v \ \text{ for all } \ v \in \mathbb{C}^n \ ,$$

which implies that the expression in parentheses is equal to A.

15.6 Definition: Let $A \in \mathbb{C}^{m \times n}$ have rank r and define the σ_j, v_j, and w_j for $1 \leq j \leq r$ as in the foregoing, with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. The *singular-value* *decomposition* or *SVD* of A is

$$A = \sum_{j=1}^{r} \sigma_j w_j \left(v_j\right)^H \ . \tag{23}$$

Observe that equation (23) expresses the matrix A as the sum of r $(m \times n)$ matrices each of which takes the somewhat unusual form

$$\sigma_j \begin{bmatrix} | \\ w_j \\ | \end{bmatrix} \begin{bmatrix} — & (v_j)^H & — \end{bmatrix} .$$

Each of these matrices has rank 1 because

$$\left(\sigma_j w_j \left(v_j\right)^H\right) v = \sigma_j \left\langle v, v_j \right\rangle \, w_j \ \text{ for all } \ v \in \mathbb{C}^n \ ,$$

so the range of the matrix $\sigma_j w_j \left(v_j\right)^H$ is the one-dimensional subspace span $(\{w_j\})$. We can re-write (23) in matrix form as follows:

$$A = W \Sigma V^H \ , \tag{24}$$

where

$$W = \begin{bmatrix} | & | & | & . & | \\ w_1 & w_2 & w_3 & . & w_r \\ | & | & | & . & | \end{bmatrix} \in \mathbb{C}^{m \times r} \ ,$$

$$V = \begin{bmatrix} | & | & | & . & | \\ v_1 & v_2 & v_3 & . & v_r \\ | & | & | & . & | \end{bmatrix} \in \mathbb{C}^{n \times r} \ ,$$

and

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & . & . & 0 \\ 0 & \sigma_2 & 0 & . & 0 \\ 0 & 0 & . & . & 0 \\ 0 & . & 0 & \sigma_{r-1} & 0 \\ 0 & . & . & 0 & \sigma_r \end{bmatrix} \in \mathbb{R}^{r \times r} \ .$$

Let's perform a quick reality check on equation (23) by multiplying the right-hand side by v_k and making sure we get Av_k.

$$\begin{aligned} \left(\sum_{j=1}^{r} \sigma_j w_j \left(v_j\right)^H\right) v_k &= \sum_{j=1}^{r} \sigma_j w_j \left\langle v_k, v_j \right\rangle \\ &= \begin{cases} \sigma_k w_k & \text{if } 1 \le k \le r \\ \quad 0 & \text{if } r+1 \le k \le n \end{cases} \\ &= Av_k \ \text{ for } \ 1 \le k \le n \ . \end{aligned}$$

The second line follows from the first line by orthonormality of the v_j. The last line holds by definition of w_k along with the fact that $(v_{r+1}, \ldots, v_n)$ is a basis for the nullspace of A by construction.

Equation (23) reveals an alternative interpretation of the way the linear mapping T_A operates. The subspace of $\mathbb{C}^n$ spanned by $\{v_j : 1 \le j \le r\}$ maps bijectively onto the range of A, which is the subspace of $\mathbb{C}^n$ spanned by $\{w_j : 1 \le j \le r\}$. That's because $Av_j = \sigma_j w_j$ for $1 \le j \le r$, so the mapping T_A takes an orthonormal basis for the first subspace onto a basis for the second. You can therefore visualize the overall mapping T_A as comprising two steps:

Step 1 Follow the procedure described after Fact 9.9 to project $v \in \mathbb{C}^n$ orthogonally onto the subspace spanned by $\{v_j : 1 \leq j \leq r\}$. You obtain the vector

$$\sum_{j=1}^{r} \langle v, v_j \rangle \, v_j \; .$$

Step 2 Map the vector from Step 1 into $\mathbb{C}^m$ using $Av_j = \sigma_j w_j$ to get

$$\sum_{j=1}^{r} \sigma_j \langle v, v_j \rangle w_j = \left(\sum_{j=1}^{r} \sigma_j w_j \, (v_j)^H \right) v = Av \; .$$

Taking the Hermitian conjugate of equations (23) and (24) yields

$$A^H = \sum_{j=1}^{r} \sigma_j v_j \, (w_j)^H \tag{25}$$

and the equivalent matrix equation

$$A^H = V \Sigma W^H \; . \tag{26}$$

If you think about it, you'll see that these last two equations constitute the singular-value decomposition of A^H. In particular, the singular values of A^H are the same as the singular values of A. Furthermore, $(v_1, v_2, \ldots, v_r)$ is an orthonormal basis for the range of A^H, and the mapping $w \mapsto A^H w$ maps the range of A, which is a subspace of $\mathbb{C}^m$ and has orthonormal basis $(w_1, w_2, \ldots, w_r)$, onto the range of A^H.

You might noticed that the entire foregoing exposition has some ragged edges. For example, I've referred to "**the** singular-value decomposition" throughout, as if the SVD were uniquely determined. In fact, it's not. The ambiguity stems from the fact that we have some wiggle room when specifying the v_j. Although all the λ_j and hence the singular values σ_j are uniquely determined, the eigenvectors of $A^H A$ corresponding to the eigenvalues λ_j are not. It gets even worse (or better, depending on your point of view) when one or more eigenvalues is repeated. I'd like to avoid getting hung up on questions of uniqueness here.

The SVD and numerical computations: the condition number

The singular-value decomposition determines various important quantitative properties of the linear mapping T_A specified by $T_A(v) = Av$. Suppose $A \in \mathbb{C}^{m \times n}$ has rank r. Recall that the vectors v_j appearing in (23) are the first r vectors in an orthonormal basis $(v_1, \ldots, v_n)$ for $\mathbb{C}^n$ and that $(v_{r+1}, \ldots, v_n)$ is a basis for the nullspace of A. Any $v \in \mathbb{C}^n$ has orthogonal expansion

$$v = \sum_{j=1}^{n} \langle v, v_j \rangle \, v_j \; .$$

By orthonormality of the v_j,

$$\|v\| = \left(\sum_{j=1}^{n} |\langle v, v_j \rangle|^2 \right)^{1/2} .$$

Now apply (23) to obtain

$$Av = \sum_{j=1}^{r} \sigma_j \langle v, v_j \rangle w_j .$$

Orthonormality of the w_j and the ordering $\sigma_1 \geq \cdots \geq \sigma_r$ imply that

$$\begin{aligned} \|Av\| &= \left(\sum_{j=1}^{r} \sigma_j^2 \, |\langle v, v_j \rangle|^2 \right)^{1/2} \\ &\leq \sigma_1 \left(\sum_{j=1}^{r} |\langle v, v_j \rangle|^2 \right)^{1/2} \\ &\leq \sigma_1 \|v\| , \end{aligned}$$

with equality holding if v is "aligned" with v_1 in the sense that $\langle v, v_j \rangle = 0$ for $j > 1$. Thus the largest singular value σ_1 is the largest factor by which the mapping T_A can "expand" a vector $v \in \mathbb{C}^n$. You can also check that if A has rank n, which by Theorem 4.9 is the same thing as saying that the nullspace of A is $\{0\}$, then the same line of reasoning yields the lower bound

$$\|Av\| \geq \sigma_n \|v\| \text{ for all } v \in \mathbb{C}^n ,$$

with equality holding if v is "aligned" with v_n. When A is a square invertible $(n \times n)$ matrix, which by Theorem 14.16 means that A has rank n and nullspace $\{0\}$, we therefore have

$$\sigma_n \|v\| \leq \|Av\| \leq \sigma_1 \|v\| \text{ for all } v \in \mathbb{C}^n . \tag{27}$$

The left inequality holds with equality if v is a multiple of v_n and the right inequality holds with equality if v is a multiple of v_1.

Furthermore, if $A \in \mathbb{C}^{n \times n}$ is invertible equation (23) reads

$$A = \sum_{j=1}^{n} \sigma_j w_j \left(v_j \right)^H$$

and equation (24) involves only square invertible matrices. Since the columns of the matrix V from (24) are orthonormal, $V^H V = I_{n \times n}$, so V is a unitary matrix. The same goes for W. The matrix Σ is a diagonal matrix with the positive number σ_j in the (j,j)-position for all j. In particular, Σ is invertible, and Σ^{-1} is the diagonal matrix with $1/\sigma_j$ in the (j,j)-position for all j. Inverting equation (24) yields $A^{-1} = V \Sigma^{-1} W^H$ because V and W are unitary matrices. Expand in terms of the columns of V and W and you get

$$A^{-1} = \sum_{j=1}^{n} \sigma_j^{-1} v_j \left(w_j \right)^H , \tag{28}$$

which is a sort of inverse version of equation (23). Given $w \in \mathbb{C}^n$,

$$w = \sum_{j=1}^{n} \langle w, w_j \rangle \, w_j \ ,$$

and

$$A^{-1}w = \sum_{j=1}^{n} \sigma_j^{-1} \langle w, w_j \rangle \, v_j \ .$$

Reasoning as we did when bounding $\|Av\|$, we find that

$$\sigma_1^{-1}\|w\| \le \left\|A^{-1}w\right\| \le \sigma_n^{-1}\|w\| \ \text{ for all } \ w \in \mathbb{C}^n \ . \tag{29}$$

The left inequality holds with equality if w is a multiple of w_1 and the right inequality holds with equality if w is a multiple of w_n.

In applications, one encounters frequently the problem of solving the equation $Av = w$ for $v \in \mathbb{C}^n$ given $w \in \mathbb{C}^n$ and $A \in \mathbb{C}^{n\times n}$. If A is invertible, then a unique solution exists, namely $v = A^{-1}w$, but computing v numerically can be problematic. Rather than computing A^{-1}, modern equation solvers employ variants of Gauss elimination and other algorithmic techniques to compute v, resulting in huge computational savings. Even so, finite-precision effects such as roundoff errors can have a significant impact on the accuracy of answers. The SVD illuminates how properties of A influence that impact.

Suppose, for example, that our goal is to solve the equation $A\widehat{v} = \widehat{w}$ for $\widehat{v}$ given $\widehat{w}$. Think of $\widehat{w}$ as the "nominal data" and $\widehat{v}$ as the "nominal answer." In real life, the "data" might arise from noisy measurements or might suffer numerical roundoff in preparation for computation. Accordingly, what we're really doing is solving $Av = w$ for v given w, where $w = \widehat{w} + \widetilde{w}$ is a perturbed version of the nominal data $\widehat{w}$. It's natural to ask what effect the perturbation $\widetilde{w}$ has on the computation. In other words, how far off is v from $\widehat{v}$?

An accepted measure of the sensitivity of the computation to perturbations in the data is the ratio

$$\frac{\text{percent change in } v}{\text{percent change in } w} \ ,$$

suitably interpreted as

$$\mathcal{S} = \frac{\|v - \widehat{v}\|/\|\widehat{v}\|}{\|w - \widehat{w}\|/\|\widehat{w}\|} \ .$$

We've defined $w - \widehat{w} = \widetilde{w}$ already, so now set $v - \widehat{v} = \widetilde{v}$. Then

$$v = A^{-1}w \ \text{ and } \ \widehat{v} = A^{-1}\widehat{w} \implies \widetilde{v} = A^{-1}\widetilde{w} \ .$$

The sensitivity measure becomes

$$\mathcal{S} = \frac{\|A^{-1}\widetilde{w}\|/\|A^{-1}\widehat{w}\|}{\|\widetilde{w}\|/\|\widehat{w}\|} \ .$$

Thus $\mathcal{S}$ depends only on the nominal data $\widehat{y}$, the perturbation in the data $\widetilde{y}$, and the matrix A.

By (29),

$$\left\|A^{-1}\widetilde{w}\right\| \le \sigma_n^{-1}\|\widetilde{w}\| \ \text{ and } \ \left\|A^{-1}\widehat{w}\right\| \ge \sigma_1^{-1}\|\widehat{w}\| \ ,$$

so the sensitivity measure $\mathcal{S}$ satisfies

$$\mathcal{S} \le \frac{\left(\sigma_n^{-1}\|\widetilde{w}\|\right) / \left(\sigma_1^{-1}\|\widehat{w}\|\right)}{\|\widetilde{w}\|/\|\widehat{w}\|} = \frac{\sigma_1}{\sigma_n} \ .$$

The ratio $\kappa = \sigma_1/\sigma_n$ is known as the *condition number* of A. Note that $1 \leq \kappa < \infty$. If κ is large, people call A an *ill-conditioned matrix.* The reason for the terminology is that the sensitivity measure $\mathcal{S}$ actually achieves its upper bound κ for certain choices of the nominal data $\widehat{w}$ and perturbation $\widetilde{w}$. Conditions for equality in (29) imply that $\mathcal{S} = \kappa$ if $\widehat{w}$ lies in the direction of w_1 and $\widetilde{w}$ lies in the direction of w_n. Thus you can always find choices of nominal data and perturbation that lead to maximal computational sensitivity κ. If κ is large, a small percentage change in the data w might lead to a massive percentage change in the solution v to the equation $Av = w$ — undoubtedly not a good thing.

The condition number of A also dictates the worst-case sensitivity of the computation of w to perturbations of v in the equation $w = Av$. In the notation of the foregoing discussion, the sensitivity of w to changes in v is

$$\frac{\|w - \widehat{w}\| / \|\widehat{w}\|}{\|v - \widehat{v}\| / \|\widehat{v}\|} = \frac{\|A\widetilde{v}\| / \|A\widehat{v}\|}{\|\widetilde{v}\| / \|\widehat{v}\|} \ ,$$

which is the inverse of the sensitivity of v to changes in w. Invoking (27) bounds this quantity from above by

$$\frac{(\sigma_1 \|\widetilde{v}\|) \,/\, (\sigma_n \|\widehat{v}\|)}{\|\widetilde{v}\| / \|\widehat{v}\|} = \frac{\sigma_1}{\sigma_n} = \kappa \ .$$

The upper bound is attained when $\widehat{v}$ lies in the direction of v_n and $\widetilde{v}$ lies in the direction of v_1.

The Moore-Penrose pseudo-inverse

I'd like to return now to the general case where $A \in \mathbb{C}^{m \times n}$ has rank r and discuss an important construction that hinges on the SVD. Recall equation (28) for the inverse of an invertible $A \in \mathbb{C}^{n \times n}$. If we manipulate analogously the SVD of an arbitrary rank-r matrix $A \in \mathbb{C}^{m \times n}$, we get

$$A^{\#} = \sum_{j=1}^{r} \sigma_j^{-1} v_j \left(w_j\right)^H \ .$$

$A^{\#}$ is named the *Moore-Penrose pseudo-inverse* of A after American mathematician E. H. Moore and British mathematical physicist Roger Penrose. Note that $A^{\#} \in \mathbb{C}^{n \times m}$ and that $A^{\#} = A^{-1}$ when A is square and invertible, so the pseudo-inverse is indeed the actual inverse when the inverse exists. What, you might ask, is inverse-like about $A^{\#}$ when A is not invertible and perhaps not even square?

We observed earlier that $(w_1, w_2, \ldots, w_r)$ is an orthonormal basis for the range of A and $(v_1, v_2, \ldots, v_r)$ is an orthonormal basis for the range of A^H, and that the mapping $v \mapsto Av$ maps the range of A^H bijectively onto the range of A. These observations underpin the assertion that the matrix $A^{\#}$ acts like an inverse in the following sense: it "inverts" the one-to-one mapping from the range of A^H onto the range of A that $v \mapsto Av$ induces. More precisely,

$$AA^{\#}w = w \qquad \text{for every } w \text{ in the range of } A$$

and

$$A^{\#}Av = v \qquad \text{for every } v \text{ in the range of } A^H \ .$$

To see how this happens, suppose w is in the range of A. We can write

$$w = \sum_{k=1}^{r} c_k w_k \ .$$

Thus

$$A^{\#} w = \left(\sum_{j=1}^{r} \sigma_j^{-1} v_j \left(w_j\right)^H \right) \left(\sum_{k=1}^{r} c_k w_k \right) = \sum_{k=1}^{r} \frac{c_k}{\sigma_k} v_k \ ,$$

where the last equality holds because the w_j are orthonormal. Hence

$$A A^{\#} w = \left(\sum_{j=1}^{r} \sigma_j w_j \left(v_j\right)^H \right) \left(\sum_{k=1}^{r} \frac{c_k}{\sigma_k} v_k \right) = \sum_{k=1}^{r} c_k w_k = w \ ,$$

where the penultimate equality holds because the v_k are orthonormal.

Similarly, if v is in the range of A^H, we can write

$$v = \sum_{k=1}^{r} c_k v_k \ ,$$

so

$$A v = \left(\sum_{j=1}^{r} \sigma_j w_j \left(v_j\right)^H \right) \left(\sum_{k=1}^{r} c_k v_k \right) = \sum_{k=1}^{r} c_k \sigma_k w_k \ ,$$

where the last equality holds because the v_j are orthonormal. Hence

$$A^{\#} A v = \left(\sum_{j=1}^{r} \sigma_j^{-1} v_j \left(w_j\right)^H \right) \left(\sum_{k=1}^{r} c_k \sigma_k w_k \right) = \sum_{k=1}^{r} c_k v_k = v \ ,$$

where the penultimate equality holds because the w_k are orthonormal.

I noted earlier that since the choices of v_j are not uniquely determined, the SVD of A is itself not uniquely determined. Still, no matter how you choose the v_j, which in turn determine the w_j, you get the same matrix A when you plug the v_j and w_j into the right-hand side of equation (23). Similarly $A^{\#}$, the Moore-Penrose pseudo-inverse of A, is uniquely determined even though the vectors participating in the formula for $A^{\#}$ are not. That is, the formula for $A^{\#}$ produces in the same matrix for every legal choice of the v_j.

To understand why, suppose that σ_o is one of the singular values of A and that $\lambda_o = \sigma_o^2$ has multiplicity d_o as an eigenvalue of $A^H A$, so $\dim\left(E(\lambda_o)\right) = d_o$. The formula for $A^{\#}$ contains d_o terms featuring σ_o, and their sum has the neat matrix representation

$$\sigma_o^{-1} V_o W_o^H \ ,$$

where the columns of $V_o \in \mathbb{C}^{n \times d_o}$ form an orthonormal basis for $E(\lambda_o)$ and $W_o = \sigma_o^{-1} A V_o$. If V_o' is another matrix whose columns form an orthonormal basis for $E(\lambda_o)$, then $V_o' = V_o U$ for some $U \in \mathbb{C}^{d_o \times d_o}$. Because the columns of V_o and V_o' are orthonormal,

$$I_{d_o \times d_o} = \left(V_o'\right)^H V_o' = U^H V_o^H V_o U = U^H U \ ,$$

so the matrix U is unitary. If we use V_o' and $W_o' = \sigma_o^{-1} A V_o'$ instead of V_o and W_o in the formula for $A^{\#}$, the sum of the σ_o-terms is

$$\sigma_o^{-1} V_o' \left(W_o'\right)^H = \sigma_o^{-1} V_o U U^H W_o^H = \sigma_o^{-1} V_o W_o^H \ .$$

Thus different choices of orthonormal bases for the eigenspaces of A^HA, although they change the v_j and corresponding w_j, do not alter the sum on the right-hand side of the formula for $A^\#$.

This last observation provides an appropriate lead-in to a final result concerning pseudo-inverses of full-rank matrices. By Theorem 4.9, the rank of $A \in \mathbb{C}^{m\times n}$ cannot exceed m or n. We say that A has full rank if the rank of A is the minimum of m and n. When A has full rank, we have a nice formula for $A^\#$.

15.7 Fact: If $m \le n$ and $A \in \mathbb{C}^{m\times n}$ has rank m, then AA^H is invertible, and $A^\# = A^H(AA^H)^{-1}$. If $m \ge n$ and $A \in \mathbb{C}^{m\times n}$ has rank n, then A^HA is invertible, and $A^\# = (A^HA)^{-1}A^H$.

Proof: Suppose A has rank m. In equation (24), the matrix W is square, and $W^HW = I_{m\times m}$ because W's columns form an orthonormal basis for $\mathbb{C}^m$. Furthermore, V is $(n \times m)$ and $V^HV = I_{m\times m}$ because the columns of V are orthonormal. It follows that

$$AA^H = W\Sigma V^H V\Sigma W^H = W\Sigma^2 W^H \ ,$$

so

$$\left(AA^H\right)^{-1} = W\Sigma^{-2}W^H = \sum_{k=1}^{m} \sigma_k^{-2} w_k \left(w_k\right)^H \ .$$

Meanwhile,

$$A^H = \sum_{j=1}^{m} \sigma_j v_j \left(w_j\right)^H \ ,$$

so

$$A^H \left(A^HA\right)^{-1} = \sum_{j=1}^{m} \sigma_j^{-1} v_j \left(w_j\right)^H = A^\#$$

because the w_j are orthonormal.

Assuming instead that A has rank n, the matrix V in (24) is square, and $V^HV = I_{n\times n}$ because V's columns form an orthonormal basis for $\mathbb{C}^n$. Furthermore, W is $(m \times n)$ and $W^HW = I_{m\times m}$ because the columns of W are orthonormal. It follows that

$$A^HA = V\Sigma W^H W\Sigma V^H = V\Sigma^2 V^H \ ,$$

so

$$\left(A^HA\right)^{-1} = V\Sigma_n^{-2}V^H = \sum_{k=1}^{n} \sigma_k^{-2} v_k \left(v_k\right)^H \ .$$

At the same time,

$$A^H = \sum_{j=1}^{n} \sigma_j v_j \left(w_j\right)^H \ ,$$

so

$$\left(AA^H\right)^{-1} A^H = \sum_{j=1}^{n} \sigma_j^{-1} v_j \left(w_j\right)^H = A^\# \ ,$$

this time because the v_j are orthonormal. Both of these formulas reduce to $A^\# = A^{-1}$ when A is square and invertible. □

The matrix $A^{\#}$ spawns a linear mapping $T_{A^{\#}}$ from $\mathbb{C}^m$ into $\mathbb{C}^n$. Note that $T_{A^{\#}}(w_j) = A^{\#}w_j = \sigma_j^{-1}v_j$ for all j in the notation we've been using. The mapping $T_{A^{\#}}$ parses into two steps as follows:

Step 1 Project $w \in \mathbb{C}^m$ orthogonally onto range(A). Since $(w_1, w_2, \ldots, w_r)$ is an orthonormal basis for range(A), this step results in the vector

$$\sum_{j=1}^{r} \langle w, w_j \rangle \; w_j \; .$$

Step 2 Map the vector from Step 1 into $\mathbb{C}^n$ using $A^{\#}w_j = \sigma_j^{-1}v_j$ to get

$$\sum_{j=1}^{r} \sigma_j^{-1} \langle w, w_j \rangle \, v_j = \left(\sum_{j=1}^{r} \sigma_j^{-1} v_j \, (w_j)^H \right) w = A^{\#}w \; .$$

The SVD and linear least squares optimization

The singular-value decomposition supplies handy solutions to a variety of linear least-squares optimization problems. Such problems pervade applications of linear algebra to science and engineering, and they come in many forms. Here I'll consider just three examples.

15.8 Problem: Let $A \in \mathbb{C}^{m \times n}$ have rank r. Find $v \in \mathbb{C}^n$ that minimizes $\|Av\|$ subject to the constraint $\|v\| \geq 1$. The vector v might represent, for example, our allocation of work among n agents, the constraint $\|v\| \geq 1$ might indicate that we need to get at least a certain amount of work done, and $\|Av\|$ might represent the time it takes for the agents to complete the work when we allocate it according to v.

Solution: If $r < n$, then A has a nonzero nullspace by Theorem 4.9, and any unit-norm vector v_o in the nullspace of A solves the problem trivially since $Av_o = 0$. Accordingly we'll assume that A has rank n and SVD

$$A = \sum_{j=1}^{n} \sigma_j w_j \, (v_j)^H \; .$$

Let's proceed as we did in the run-up to (27). Since $(v_1, v_2, \ldots, v_n)$ is an orthonormal basis for $\mathbb{C}^n$,

$$v = \sum_{j=1}^{n} \langle v, v_j \rangle \, v_j \quad \text{for all} \;\; v \in \mathbb{C}^n$$

and

$$Av = \sum_{j=1}^{n} \sigma_j \langle v, v_j \rangle \, w_j \quad \text{for all} \;\; v \in \mathbb{C}^n \; .$$

Orthonormality of the v_j implies that

$$\|v\| = \left(\sum_{j=1}^{n} |\langle v, v_j \rangle|^2 \right)^{1/2}$$

and, because the w_j are orthonormal,

$$\|Av\| = \left(\sum_{j=1}^{n} \sigma_j^2 |\langle v, v_j \rangle|^2 \right)^{1/2} \geq \sigma_n \left(\sum_{j=1}^{n} \| \langle v, v_j \rangle|^2 \right)^{1/2} = \sigma_n \|v\| \ .$$

Observe that equality holds if $v = \langle v, v_n \rangle \, v_n$. The constraint $\|v\| \geq 1$ means $\|Av\| \geq \sigma_n$ for every v satisfying the constraint, and you can see that choosing $v = v_n$ minimizes $\|Av\|$ subject to the constraint. Note that v_n is not the unique solution to the problem since $c_o v_n$ also works when $|c_o| = 1$. In fact, any unit-norm eigenvector of $A^H A$ corresponding to eigenvalue $\lambda_n = \sigma_n^2$ will also do the job. □

15.9 Problem: Let $A \in \mathbb{C}^{m \times n}$ have rank r. Find $v \in \mathbb{C}^n$ that maximizes $\|Av\|$ subject to the constraint $\|v\| \leq 1$. The vector v might represent, for example, our allocation of money among n investments, the constraint $\|v\| \leq 1$ might indicate that our funds are limited, and $\|Av\|$ might represent the total return on our investments if we allocate them according to v.

Solution: Suppose A has rank r and SVD

$$A = \sum_{j=1}^{r} \sigma_j w_j \left(v_j\right)^H \ .$$

Recall that $v_1, \ v_2, \ \ldots \ , \ v_r$ are orthonormal eigenvectors of $A^H A$ corresponding to nonzero eigenvalues and that they originate from an orthonormal basis $(v_1, v_2, \ldots, v_n)$ for $\mathbb{C}^n$, where $(v_{r+1}, \ldots, v_n)$ is a basis for the nullspace of A. Again, let's proceed as we did when deriving (27). We have

$$v = \sum_{j=1}^{n} \langle v, v_j \rangle \, v_j \ \text{ for all } \ v \in \mathbb{C}^n$$

and

$$Av = \sum_{j=1}^{r} \sigma_j \langle v, v_j \rangle \, w_j \ \text{ for all } \ v \in \mathbb{C}^n \ .$$

Orthonormality of the v_j implies that

$$\|v\| = \left(\sum_{j=1}^{n} |\langle v, v_j \rangle|^2 \right)^{1/2}$$

and, because the w_j are orthonormal,

$$\|Av\| = \left(\sum_{j=1}^{r} \sigma_j^2 |\langle v, v_j \rangle|^2 \right)^{1/2} \leq \sigma_1 \left(\sum_{j=1}^{r} \| \langle v, v_j \rangle|^2 \right)^{1/2} \leq \sigma_1 \|v\| \ .$$

Observe that equality holds if $v = \langle v, v_1 \rangle v_1$. The constraint $\|v\| \leq 1$ means $\|Av\| \leq \sigma_1$ for every v satisfying the constraint, and you can see that choosing $v = v_1$ maximizes $\|Av\|$ subject to the constraint. Note that v_1 is not the unique solution to the problem since $c_o v_1$ also works when $|c_o| = 1$. In fact, any unit-norm eigenvector of $A^H A$ corresponding to eigenvalue $\lambda_1 = \sigma_1^2$ does the job. □

15.10 Problem: Let $A \in \mathbb{C}^{m \times n}$ have rank r. Given $w \in \mathbb{C}^m$, find a vector $v \in \mathbb{C}^n$ of smallest norm that minimizes $\|Av - w\|$. Note that when $w \in \text{range}(A)$ the problem reduces to finding a minimim-norm v such that $Av = w$.

Solution: Split the problem in two. First find the vector $\widehat{w}$ in the range of A that's closest to w, i.e. find $\widehat{w}$ so that

$$\|\widehat{w} - w\| \leq \|w' - w\| \text{ for every } w' \in \text{range}(A) \ .$$

From the discussion preceding Fact 9.11 we can infer that $\widehat{w}$ is the orthogonal projection of w onto range of A. Since $(w_1, w_2, \ldots, w_r)$ is an orthonormal basis for the range of A,

$$\widehat{w} = \sum_{j=1}^{r} \langle w, w_j \rangle \, w_j \ .$$

Exactly those vectors $v \in \mathbb{C}^n$ that map to $\widehat{w}$ minimize $\|Av - w\|$ over $v \in \mathbb{C}^n$. Since $w_j = \sigma_j^{-1} A v_j$ for all j, one vector mapping to $\widehat{w}$ is

$$\begin{aligned} \widehat{v} &= \sum_{j=1}^{r} \sigma_j^{-1} \langle w, w_j \rangle \, v_j \\ &= \sum_{j=1}^{r} \sigma_j^{-1} \left((w_j)^H \, w \right) \, v_j \\ &= \left(\sum_{j=1}^{r} \sigma_j^{-1} v_j \, (w_j)^H \right) w \\ &= A^{\#} w \ . \end{aligned}$$

I claim that $\widehat{v}$ is the unique solution to the problem. Any other solution $v \in \mathbb{C}^n$ must also satisfy $Av = \widehat{w}$, so we can write

$$v = \widehat{v} + v_o$$

where $v_o \in \text{nullspace}(A)$. Since $\widehat{v}$ is a linear combination of $v_1, \, v_2, \, \ldots \, , v_r$, $\widehat{v}$ is orthogonal to every vector in the nullspace of A, so

$$\|v\|^2 = \|\widehat{v}\|^2 + \|v_o\|^2 \geq \|\widehat{v}\|^2 \ ,$$

with equality if and only if $v_o = 0$. Thus $\widehat{v} = A^{\#} w$ is the unique vector v of smallest norm that minimizes $\|Av - w\|$. □

The SVD and matrix norms

You've probably noticed that $\mathbb{C}^{m\times n}$ has a natural vector-space structure with vector operations performed elementwise in the sense that

$$[c_1A + c_2B]_{ij} = c_1\,[A]_{ij} + c_2\,[B]_{ij} \quad \text{for } 1 \le i \le m \text{ and } 1 \le j \le n$$

when A and B are in $\mathbb{C}^{m\times n}$ and c_1 and c_2 are in $\mathbb{C}$. If you set up a correspondence between $(m \times n)$ complex matrices and mn-dimensional complex vectors by listing matrices' elements in some fixed order as elements of vectors, you'll see that the elementwise vector-space structure on $\mathbb{C}^{m\times n}$ reflects the customary vector-space structure on $\mathbb{C}^{mn}$. The standard inner product on $\mathbb{C}^{mn}$ also ports over to $\mathbb{C}^{m\times n}$ via the prescription

$$\langle A, B\rangle = \sum_{i=1}^{n}\sum_{j=1}^{n}\overline{[B]_{ij}}[A]_{ij} \ \text{ for all } \ A,\ B \in \mathbb{C}^{m\times n}\ .$$

It proves convenient at times to express this inner product on $\mathbb{C}^{m\times n}$ in another way. The *trace* of a square matrix is the sum of its diagonal elements, so

$$\begin{aligned}\langle A, B\rangle &= \sum_{i=1}^{n}\left(\sum_{j=1}^{n}\overline{[B]_{ij}}[A]_{ij}\right)\\ &= \sum_{j=1}^{n}\left(\sum_{i=1}^{n}\left[B^H\right]_{ji}[A]_{ij}\right)\\ &= \sum_{j=1}^{n}\left[B^H A\right]_{jj}\\ &= \text{Trace}\left(B^H A\right) \ \text{ for all } \ A,\ B \in \mathbb{C}^{m\times n}\ .\end{aligned}$$

Trace has some nice properties. In particular, if A_1 and A_2 are matrices whose product makes sense and is square, then

$$\text{Trace}\,(A_1A_2) = \text{Trace}\,(A_2A_1)\ .$$

Note that A_1A_2 and A_2A_1 are both square but might have different sizes. The norm arising from our inner product on $\mathbb{C}^{m\times n}$, called the *Frobenius norm*, is given by

$$\|A\|_{\text{F}} = \left(\text{Trace}\left(A^H A\right)\right)^{1/2} = \left(\sum_{i=1}^{n}\sum_{j=1}^{n}\left|[A]_{ij}\right|^2\right)^{1/2} \quad \text{for all } \ A \in \mathbb{C}^{m\times n}\ .$$

In other words, $\|A\|_{\text{F}}$ is the square root of the sum of the squares of A's elements. It's worth mentioning that the Frobenius norm has the following invariance property: if $U_1 \in \mathbb{C}^{m\times m}$ and $U_2 \in \mathbb{C}^{n\times n}$ are unitary, then $\|U_1AU_2\|_{\text{F}} = \|A\|_{\text{F}}$. This is because

$$\begin{aligned}\|U_1AU_2\|_{\text{F}}^2 &= \text{Trace}\left(U_2^H A^H U_1^H U_1 A U_2\right)\\ &= \text{Trace}\left(U_2^H A^H A U_2\right)\\ &= \text{Trace}\left(U_2 U_2^H A^H A\right)\\ &= \text{Trace}\left(A^H A\right) = \|A\|_{\text{F}}^2\ ,\end{aligned}$$

where I used commutativity of the trace in the third line along with the fact that U_1 and U_2 are unitary.

Suppose now that $A \in \mathbb{C}^{m\times n}$ has rank r. Starting from the matrix form of the singular-value decomposition of A as in (24), you discover that

$$\begin{aligned} \|A\|_{\mathrm{F}}^2 &= \operatorname{Trace}\left(A^H A\right) \\ &= \operatorname{Trace}\left(V\Sigma W^H W \Sigma V^H\right) \\ &= \operatorname{Trace}\left(V\Sigma^2 V^H\right) \\ &= \operatorname{Trace}\left(V^H V \Sigma^2\right) \\ &= \operatorname{Trace}\left(\Sigma^2\right) = \sum_{j=1}^{r} \sigma_j^2 \end{aligned}$$

for every $A \in \mathbb{C}^n$, where again I used commutativity of the trace in the fourth line along with $V^H V = W^H W = I_{r\times r}$. In other words, the Frobenius norm of A is the square root of the sum of the squares of the singular values of A.

You're probably developing a sense that the singular-value decomposition of $A \in \mathbb{C}^{m\times n}$ appears to indicate what "directions" in $\mathbb{C}^n$ and $\mathbb{C}^m$ have more "influence" than others on the behavior of the linear mapping $v \mapsto Av$. In our analysis of numerical computational sensitivity, for example, we found that data-specification errors cause the most damage when aligned with the v_j or w_j corresponding to large singular values. When the singular values of A have a wide spread in magnitude, it's fair to say that most of the "action" in the linear mapping $v \mapsto Av$ occurs in the restriction of the mapping to the subspaces spanned by the v_j corresponding to the larger singular values. The SVD enables us to make quantitative sense of that intuition.

Suppose $A \in \mathbb{C}^{m\times n}$ has rank r and SVD given by (23). If $r' \le r$, set

$$\widehat{A}_{r'} = \sum_{j=1}^{r'} \sigma_j w_j \left(v_j\right)^H \quad .$$

The sum defining $\widehat{A}_{r'}$ incorporates the terms in A's SVD corresponding the largest r' singular values. $\widehat{A}_{r'}$ has rank r' because its range is span $(\{w_1, w_2, \ldots, w_{r'}\})$, and we might expect it to be a reasonable approximation of A if the singular values σ_j for $j > r'$ are small relative to σ_j for $j \le r'$.

15.11 Theorem: Suppose $A \in \mathbb{C}^{m\times n}$ has rank r. If $r' \le r$, the matrix $\widehat{A}_{r'}$ defined above is the matrix closest to A in the sense of Frobenius norm among all complex $(m \times n)$ matrices of rank at most r'. Furthermore, $\|A - \widehat{A}_{r'}\|_{\mathrm{F}} = \left(\sum_{j=r'+1}^{r} \sigma_j^2\right)^{1/2}$.

Proof: Recall that the v_j appearing in (23) come from an orthonormal basis $(v_1, v_2, \ldots, v_n)$ for $\mathbb{C}^n$, where $(v_{r+1}, \ldots, v_n)$ is an orthonormal basis for the nullspace of A. Define $V_{n\times n} \in \mathbb{C}^{n\times n}$ as the matrix whose jth column is v_j for $1 \le j \le n$. Next, if $r < m$ extend $(w_1, w_2, \ldots, w_r)$, with notation as in (23), to an orthonormal basis $(w_1, w_2, \ldots, w_m)$ for $\mathbb{C}^m$ by letting $(w_{r+1}, \ldots, w_{r+m})$ be a basis for the nullspace of W^H. Let $W_{m\times m} \in \mathbb{C}^{m\times m}$ be the matrix whose jth column is w_j for $1 \le j \le m$. Observe that if $r = n$, then $V = V_{n\times n}$ and if $r = m$ then $W = W_{m\times m}$, again with notation as in (24). Note that $V_{n\times n}$ and $W_{m\times m}$ are both unitary since each has orthonormal columns.

It follows easily that

$$A = W_{m\times m}\Sigma_{m\times n}\left(V_{n\times n}\right)^H \ ,$$

where $\Sigma_{m\times n} \in \mathbb{C}^{m\times n}$ has zero entries everywhere except at position (j, j) for $1 \leq j \leq r$, where $\left[\Sigma_{m\times n}\right]_{jj} = \sigma_j$. If $B \in \mathbb{C}^{m\times n}$, then

$$\begin{aligned} \|A - B\|_{\mathrm{F}} &= \left\| W_{m\times m}\Sigma_{m\times n}\left(V_{n\times n}\right)^H - B \right\|_{\mathrm{F}} \\ &= \left\| \Sigma_{m\times n} - \left(W_{m\times m}\right)^H B V_{n\times n} \right\|_{\mathrm{F}} \ , \end{aligned}$$

where I used the invariance of Frobenius norm under pre- and post-multiplication by unitary matrices. Since the Frobenius norm of a matrix is the sum of the magnitudes squared of the matrix's entries, it's clear that we want to choose B so that the entries of $\left(W_{m\times m}\right)^H B V_{n\times n}$ are as close as possible to those of $\Sigma_{m\times n}$. Simultaneously we need to make sure that B has rank at most r.

By setting

$$\left[\left(W_{m\times m}\right)^H B V_{n\times n}\right]_{ij} = 0 \text{ when } i \neq j$$

and

$$\left[\left(W_{m\times m}\right)^H B V_{n\times n}\right]_{ij} = \begin{cases} \sigma_j & \text{when } i = j \leq r' \\ 0 & \text{when } i = j > r' \end{cases}$$

we match all the zeroes in $\Sigma_{m\times n}$ and also match the r' largest nonzero entries. Matching more nonzero entries would result in a matrix B with rank larger than r', and matching any other subset of at most r' nonzero entries, while producing a B of rank at most r', would do a less effective job of minimizing the Frobenius norm of $A - B$. Observe that $\widehat{\Sigma}_{r'} = \left(W_{m\times m}\right)^H B V_{n\times n}$ is the same as $\Sigma_{m\times n}$ except that $\left[\widehat{\Sigma}_{r'}\right]_{jj} = 0$ when $j > r'$. Note also that

$$\|A - B\|_{\mathrm{F}} = \left\|\Sigma_{m\times n} - \widehat{\Sigma}_{r'}\right\|_{\mathrm{F}} = \left(\sum_{j=r'+1}^{r} \sigma_j^2\right)^{1/2} \ .$$

Since $W_{m\times m}$ and $V_{n\times n}$ are unitary,

$$B = W_{m\times m}\widehat{\Sigma}_{r'}\left(V_{n\times n}\right)^H \ ,$$

and therefore the B that minimizes $\|A - B\|_{\mathrm{F}}$ is

$$B = \sum_{j=1}^{r'} \sigma_j w_j \left(v_j\right)^H = \widehat{A}_{r'} \ ,$$

which is what we set out to prove. □

The Frobenius norm sizes up an $(m \times n)$ matrix A as an array of numbers, but other matrix norms — the so-called induced norms — assess more effectively the norm-like properties of the linear mapping $v \mapsto Av$. While discussing these, I'll revert to the customary notation for the 1-norm, 2-norm, and infinity-norm with respect to the standard basis for $\mathbb{C}^n$. To refresh your memory of Chapter 4,

$$\|v\|_1 = \sum_{i=1}^{n} |[v]_i| \quad \text{for all} \quad v \in \mathbb{C}^n \ ,$$

$$\|v\|_2 = \left(\sum_{i=1}^{n} |[v]_i|^2 \right)^{1/2} \quad \text{for all} \quad v \in \mathbb{C}^n \ ,$$

and

$$\|v\|_\infty = \max\left(\{|[v]_i| : 1 \leq i \leq n\}\right) \quad \text{for all} \quad v \in \mathbb{C}^n \ .$$

Analyzing Problem 15.9 led to the discovery that

$$\sigma_1 = \max\left(\{\|Av\|_2 : \|v\|_2 = 1\}\right)$$

when $A \in \mathbb{C}^{m\times n}$ and σ_1 is the largest singular value of A. Thus σ_1 is the largest factor by which A can multiply the 2-norm of a n-dimensional vector v of unit 2-norm when producing Av. We call σ_1 the *induced 2-norm of* A and denote it by $\|A\|_2$. For any nonzero $v \in \mathbb{C}^n$,

$$\left\| A\left(\frac{v}{\|v\|_2}\right) \right\|_2 \leq \sigma_1 \ ,$$

so

$$\|Av\|_2 \leq \sigma_1 \|v\|_2 = \|A\|_2 \|v\|_2 \quad \text{for all} \quad v \in \mathbb{C}^n \ .$$

Furthermore, if $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{n\times q}$, then

$$\|ABv\|_2 \leq \|A\|_2 \|Bv\|_2 \leq \|A\|_2 \|B\|_2 \|v\|_2 \quad \text{for all} \quad v \in \mathbb{C}^q \ .$$

Since this chain of inequalities holds for all v satisfying $\|v\| = 1$,

$$\|AB\|_2 \leq \|A\|_2 \|B\|_2 \ .$$

The induced 2-norm is indeed a norm on $\mathbb{C}^{m\times n}$. Let's just verify that it satisfies the triangle inequality. Suppose A and B are in $\mathbb{C}^{m\times n}$ and $v \in \mathbb{C}^n$. Then

$$\begin{aligned} \|(A+B)v\|_2 &\leq \|Av\|_2 + \|Bv\|_2 \\ &\leq \|A\|_2\|v\|_2 + \|B\|_2\|v\|_2 \ , \end{aligned}$$

and maximizing the left-hand side over v with $\|v\|_2 = 1$ results in

$$\|A+B\|_2 \leq \|A\|_2 + \|B\|_2 \ .$$

The 1-norm and infinity-norm also induce norms on $(m\times n)$ matrices. Following the definition of the induced 2-norm, set

$$\|A\|_1 = \max\left(\{\|Av\|_1 : \|v\|_1 = 1\}\right)$$

and

$$\|A\|_\infty = \max\left(\{\|Av\|_\infty : \|v\|_\infty = 1\}\right) \ .$$

Let's make sure these maxima exist. Note first that

$$\begin{aligned}
\|Av\|_1 &= \sum_{i=1}^{m} \left|[Av]_i\right| \\
&= \sum_{i=1}^{m} \left|\sum_{j=1}^{n} [A]_{ij}[v]_j\right| \\
&\leq \sum_{j=1}^{n} \left(\sum_{i=1}^{m} |[A]_{ij}|\right) |[v]_j| \\
&\leq \max\left(\left\{\sum_{i=1}^{m} |[A]_{ij}| : 1 \leq j \leq n\right\}\right) \|v\|_1 \ .
\end{aligned}$$

Thus when $\|v\|_1 = 1$,

$$\|Av\|_1 \leq \max\left(\left\{\sum_{i=1}^{m} |[A]_{ij}| : 1 \leq j \leq n\right\}\right) \ ,$$

so $\|Av\|_1$ is bounded from above by the maximum of the 1-norms of the columns of A viewed as m-vectors. That upper bound is attained when we choose $v = e^{j_o}$, the standard basis vector with a 1 in position j_o and zeroes elsewhere, where j_o indexes the column of A with maximum 1-norm. Accordingly,

$$\|A\|_1 = \max\left(\left\{\sum_{i=1}^{m} |[A]_{ij}| : 1 \leq j \leq n\right\}\right) \ .$$

Similarly, for every $v \in \mathbb{C}^n$ and every i we have

$$\begin{aligned}
|[Av]_i| &= \left|\sum_{j=1}^{n} [A]_{ij}[v]_j\right| \\
&\leq \sum_{j=1}^{n} |[A]_{ij}| \ |[v]_j| \\
&\leq \left(\sum_{j=1}^{n} |[A]_{ij}|\right) \|v\|_\infty \ ,
\end{aligned}$$

so

$$\|Av\|_\infty \leq \max\left(\left\{\sum_{j=1}^{n} |[A]_{ij}| : 1 \leq i \leq m\right\}\right) \|v\|_\infty \ .$$

Thus when $\|v\|_\infty = 1$, $\|Av\|_\infty$ is bounded from above by the maximum of the 1-norms of the rows of A viewed as n-vectors. We can attain that upper bound as follows. Let i_o index the row of A with maximum 1-norm. Define $v \in \mathbb{C}^n$ by

$$[v]_j = \begin{cases} \overline{[A]_{i_oj}} / |[A]_{i_oj}| & \text{if } [A]_{i_oj} \neq 0 \\ 0 & \text{if } [A]_{i_oj} = 0 \ . \end{cases}$$

Note that $\|v\|_\infty = 1$ and

$$\|Av\|_\infty = \sum_{j=1}^{n} |[A]_{i_oj}| \ ,$$

from which to follows that

$$\|A\|_\infty = \max\left(\left\{\sum_{j=1}^{n} |[A]_{ij}| : 1 \leq i \leq m\right\}\right) .$$

You can verify easily that

$$\|Av\|_1 \leq \|A\|_1 \|v\|_1 \text{ and } \|Av\|_\infty \leq \|A\|_\infty \|v\|_\infty$$

for every $v \in \mathbb{C}^n$ and $A \in \mathbb{C}^{m\times n}$ and that

$$\|AB\|_1 \leq \|A\|_1 \, \|B\|_1 \text{ and } \|AB\|_\infty \leq \|A\|_\infty \, \|B\|_\infty$$

whenever the product AB makes sense.

As it happens, any norms $\| \quad \|_{\mathbb{C}^n}$ and $\| \quad \|_{\mathbb{C}^m}$ on $\mathbb{C}^n$ and $\mathbb{C}^m$ induce a norm on $\mathbb{C}^{m\times n}$ via

$$\|A\| = \max\left(\{\|Av\|_{\mathbb{C}^m} : \|v\|_{\mathbb{C}^n} = 1\}\right) .$$

Proving this assertion requires an argument akin to the one I didn't supply for Theorem 4.12. On a more abstract level, any norms $\| \quad \|_V$ and $\| \quad \|_W$ on finite-dimensional vector spaces V and W induce a norm on $\mathrm{Hom}(V, W)$ via

$$\|T\| = \max\left(\{\|T(v)\|_W : \|v\|_V = 1\}\right) .$$

Squaring away the details of these and other related constructions would take us too far afield.

In Chapter 14 we encountered the spectral radius $\rho(T)$ of a linear mapping $T \in \mathrm{End}(V)$, where V is a finite-dimensional complex vector space, and explored its implications for the asymptotic behavior of $T^k(v)$ as $k \to \infty$. By analogy, if A is a square complex matrix, we define the spectral radius of A as the magnitude of A's largest eigenvalue and denote it by $\rho(A)$. The matrix version of Theorem 14.13 states that if $\| \quad \|$ is any norm on $\mathbb{C}^n$ and $\zeta > \rho(A)$ there exists $M > 0$ such that

$$\|A^k v\| \leq M\zeta^k \|v\| \text{ for all } v \in \mathbb{C}^n .$$

In particular, $A^k v \to 0$ as $k \to \infty$ for every v if $\rho(A) < 1$. Matrices can have small spectral radii but large induced norms. For example,

$$A = \begin{bmatrix} 0 & 10^{23} \\ 0 & 0 \end{bmatrix}$$

has zero spectral radius but $\|A\|_1 = \|A\|_\infty = \|A\|_2 = 10^{23}$. Thus induced norms, despite the inequality $\|Av\| \leq \|A\| \, \|v\|$, tend not to illuminate the asymptotics of $A^k v$. One noteworthy relationship holds in general between induced norms of a matrix and its spectral radius.

15.12 Fact: If $\| \quad \|$ is any norm on $\mathbb{C}^n$ and $A \in \mathbb{C}^{n\times n}$, then

$$\|A\| \geq \rho(A) ,$$

where $\|A\|$ is the norm of A induced by $\| \quad \|$ and $\rho(A)$ is the spectral radius of A.

Proof: Let λ_o be an eigenvector of A with $|\lambda_o| = \rho(A)$ and v_o a corresponding eigenvector with $\|v_o\| = 1$. Then $\|Av_o\| = |\lambda_o| \, \|v_o\| = \rho(A)$, and thus $\|A\| = \max(\{Av : \|v\| = 1\}) \geq \rho(A)$. □

Fact 15.12 enables us to establish crude upper bounds on the magnitudes of a matrix's eigenvalues just by inspecting the matrix. For example, if A has nonnegative entries and the entries in any row of A sum to 1, then no eigenvalue of A has magnitude greater than 1 because $\|A\|_\infty = 1$ must be at least as large as the spectral radius of A by Fact 15.12.