CHAPTER 1

Signals: Analog, Discrete, and Digital

Analog, discrete, and digital signals are the raw material of signal processing and analysis. Natural processes, whether dependent upon or independent of human control, generate analog signals; they occur in a continuous fashion over an interval of time or space. The mathematical model of an analog signal is a function defined over a part of the real number line. Analog signal conditioning uses conventional electronic circuitry to acquire, amplify, filter, and transmit these signals. At some point, digital processing may take place; today, this is almost always necessary. Perhaps the application requires superior noise immunity. Intricate processing steps are also easier to implement on digital computers. Furthermore, it is easier to improve and correct computerized algorithms than systems comprised of hard-wired analog components. Whatever the rationale for digital processing, the analog signal is captured, stored momentarily, and then converted to digital form. In contrast to an analog signal, a discrete signal has values only at isolated points. Its mathematical representation is a function on the integers; this is a fundamental difference. When the signal values are of finite precision, so that they can be stored in the registers of a computer, then the discrete signal is more precisely known as a digital signal. Digital signals thus come from sampling an analog signal, and—although there is such a thing as an analog computer—nowadays digital machines perform almost all analytical computations on discrete signal data.

This has not, of course, always been the case; only recently have discrete techniques come to dominate signal processing. The reasons for this are both theoretical and practical.

On the practical side, nineteenth century inventions for transmitting words, the telegraph and the telephone—written and spoken language, respectively—mark the beginnings of engineered signal generation and interpretation technologies. Mathematics that supports signal processing began long ago, of course. But only in the nineteenth century did signal theory begin to distinguish itself as a technical, engineering, and scientific pursuit separate from pure mathematics. Until then, scientists did not see mathematical entities—polynomials, sinusoids, and exponential functions, for example—as sequences of symbols or carriers of information. They were envisioned instead as ideal shapes, motions, patterns, or models of natural processes.

Signal Analysis: Time, Frequency, Scale, and Structure, by Ronald L. Allen and Duncan W. Mills
ISBN: 0-471-23441-9 Copyright © 2004 by Institute of Electrical and Electronics Engineers, Inc.

The development of electromagnetic theory and the growth of electrical and electronic communications technologies began to divide these sciences. The functions of mathematics came to be studied as bearing information, requiring modification to be useful, suitable for interpretation, and having a meaning. The life story of this new discipline—signal processing, communications, signal analysis, and information theory—would follow a curious and ironic path. Electromagnetic waves consist of coupled electric and magnetic fields that oscillate in a sinusoidal pattern and are perpendicular to one another and to their direction of propagation. Fourier discovered that very general classes of functions, even those containing discontinuities, could be represented by sums of sinusoidal functions, now called a Fourier series [1]. This surprising insight, together with the great advances in analog communication methods at the beginning of the twentieth century, captured the most attention from scientists and engineers.

Research efforts into discrete techniques were producing important results, even as the analog age of signal processing and communication technology charged ahead. Discrete Fourier series calculations were widely understood, but seldom carried out; they demanded quite a bit of labor with pencil and paper. The first theoretical links between analog and discrete signals were found in the 1920s by Nyquist,[1] in the course of research on optimal telegraphic transmission mechanisms [2]. Shannon[2] built upon Nyquist's discovery with his famous sampling theorem [3]. He also proved something to be feasible that no one else even thought possible: error-free digital communication over noisy channels. Soon thereafter, in the late 1940s, digital computers began to appear. These early monsters were capable of performing signal processing operations, but their speed remained too slow for some of the most important computations in signal processing—the discrete versions of the Fourier series. All this changed two decades later when Cooley and Tukey disclosed their fast Fourier transform (FFT) algorithm to an eager computing public [4–6]. Digital computations of Fourier's series were now practical on real-time signal data, and in the following years digital methods would proliferate. At the present time, digital systems have supplanted much analog circuitry, and they are the core of almost all signal processing and analysis systems. Analog techniques handle only the early signal input, output, and conditioning chores.

There are a variety of texts available covering signal processing. Modern introductory systems and signal processing texts cover both analog and discrete theory [7–11]. Many reflect the shift to discrete methods that began with the discovery of the FFT and was fueled by the ever-increasing power of computing machines. These often concentrate on discrete techniques and presuppose a background in analog

[1]As a teenager, Harry Nyquist (1887–1976) emigrated from Sweden to the United States. Among his many contributions to signal and communication theory, he studied the relationship between analog signals and discrete signals extracted from them. The term *Nyquist rate* refers to the sampling frequency necessary for reconstructing an analog signal from its discrete samples.

[2]Claude E. Shannon (1916–2001) founded the modern discipline of information theory. He detailed the affinity between Boolean logic and electrical circuits in his 1937 Masters thesis at the Massachusetts Institute of Technology. Later, at Bell Laboratories, he developed the theory of reliable communication, of which the sampling theorem remains a cornerstone.

signal processing [12–15]. Again, there is a distinction between discrete and digital signals. Discrete signals are theoretical entities, derived by taking instantaneous—and therefore exact—samples from analog signals. They might assume irrational values at some time instants, and the range of their values might be infinite. Hence, a digital computer, whose memory elements only hold limited precision values, can only process those discrete signals whose values are finite in number and finite in their precision—digital signals. Early texts on discrete signal processing sometimes blurred the distinction between the two types of signals, though some further editions have adopted the more precise terminology. Noteworthy, however, are the burgeoning applications of digital signal processing integrated circuits: digital telephony, modems, mobile radio, digital control systems, and digital video to name a few. The first high-definition television (HDTV) systems were analog; but later, superior HDTV technologies have relied upon digital techniques. This technology has created a true digital signal processing literature, comprised of the technical manuals for various DSP chips, their application notes, and general treatments on fast algorithms for real-time signal processing and analysis applications on digital signal processors [16–21]. Some of our later examples and applications offer some observations on architectures appropriate for signal processing, special instruction sets, and fast algorithms suitable for DSP implementation.

This chapter introduces signals and the mathematical tools needed to work with them. Everyone should review this chapter's first six sections. This first chapter combines discussions of analog signals, discrete signals, digital signals, and the methods to transition from one of these realms to another. All that it requires of the reader is a familiarity with calculus. There are a wide variety of examples. They illustrate basic signal concepts, filtering methods, and some easily understood, albeit limited, techniques for signal interpretation. The first section introduces the terminology of signal processing, the conventional architecture of signal processing systems, and the notions of analog, discrete, and digital signals. It describes signals in terms of mathematical models—functions of a single real or integral variable. A specification of a sequence of numerical values ordered by time or some other spatial dimension is a time domain description of a signal. There are other approaches to signal description: the frequency and scale domains, as well as some—relatively recent—methods for combining them with the time domain description. Sections 1.2 and 1.3 cover the two basic signal families: analog and discrete, respectively. Many of the signals used as examples come from conventional algebra and analysis.

The discussion gets progressively more formal. Section 1.4 covers sampling and interpolation. Sampling picks a discrete signal from an analog source, and interpolation works the other way, restoring the gaps between discrete samples to fashion an analog signal from a discrete signal. By way of these operations, signals pass from the analog world into the discrete world and vice versa. Section 1.5 covers periodicity, and foremost among these signals is the class of sinusoids. These signals are the fundamental tools for constructing a frequency domain description of a signal. There are many special classes of signals that we need to consider, and Section 1.6 quickly collects them and discusses their properties. We will of course expand upon and deepen our understanding of these special types of signals

throughout the book. Readers with signal processing backgrounds may quickly scan this material; however, those with little prior work in this area might well linger over these parts.

The last two sections cover some of the mathematics that arises in the detailed study of signals. The complex number system is essential for characterizing the timing relationships in signals and their frequency content. Section 1.7 explains why complex numbers are useful for signal processing and exposes some of their unique properties. Random signals are described in Section 1.8. Their application is to model the unpredictability in natural signals, both analog and discrete. Readers with a strong mathematics background may wish to skim the chapter for the special signal processing terminology and skip Sections 1.7 and 1.8. These sections can also be omitted from a first reading of the text.

A summary, a list of references, and a problem set complete the chapter. The summary provides supplemental historical notes. It also identifies some software resources and publicly available data sets. The references point out other introductory texts, reviews, and surveys from periodicals, as well as some of the recent research.

1.1 INTRODUCTION TO SIGNALS

There are several standpoints from which to study signal analysis problems: empirical, technical, and theoretical. This chapter uses all of them. We present lots of examples, and we will return to them often as we continue to develop methods for their processing and interpretation. After practical applications of signal processing and analysis, we introduce some basic terminology, goals, and strategies.

Our early methods will be largely experimental. It will be often be difficult to decide upon the best approach in an application; this is the limitation of an intuitive approach. But there will also be opportunities for making technical observations about the right mathematical tool or technique when engaged in a practical signal analysis problem. Mathematical tools for describing signals and their characteristics will continue to illuminate this technical side to our work. Finally, some abstract considerations will arise at the end of the chapter when we consider complex numbers and random signal theory. Right now, however, we seek only to spotlight some practical and technical issues related to signal processing and analysis applications. This will provide the motivation for building a significant theoretical apparatus in the sequel.

1.1.1 Basic Concepts

Signals are symbols or values that appear in some order, and they are familiar entities from science, technology, society, and life. Examples fit easily into these categories: radio-frequency emissions from a distant quasar; telegraph, telephone, and television transmissions; people speaking to one another, using hand gestures; raising a sequence of flags upon a ship's mast; the echolocation chirp of animals such as bats and dolphins; nerve impulses to muscles; and the sensation of light patterns

striking the eye. Some of these signal values are quantifiable; the phenomenon is a measurable quantity, and its evolution is ordered by time or distance. Thus, a residential telephone signal's value is known by measuring the voltage across the pair of wires that comprise the circuit. Sound waves are longitudinal and produce minute, but measurable, pressure variations on a listener's eardrum. On the other hand, some signals appear to have a representation that is at root not quantifiable, but rather symbolic. Thus, most people would grant that sign language gestures, maritime signal flags, and even ASCII text could be considered signals, albeit of a symbolic nature.

Let us for the moment concentrate on signals with quantifiable values. These are the traditional mathematical signal models, and a rich mathematical theory is available for studying them. We will consider signals that assume symbolic values, too, but, unlike signals with quantifiable values, these entities are better described by relational mathematical structures, such as graphs.

Now, if the signal is a continuously occurring phenomenon, then we can represent it as a function of a time variable t; thus, $x(t)$ is the value of signal x at time t. We understand the units of measurement of $x(t)$ implicitly. The signal might vary with some other spatial dimension other than time, but in any case, we can suppose that its domain is a subset of the real numbers. We then say that $x(t)$ is an *analog signal*. Analog signal values are read from conventional indicating devices or scientific instruments, such as oscilloscopes, dial gauges, strip charts, and so forth.

An example of an analog signal is the seismogram, which records the shaking motion of the ground during an earthquake. A precision instrument, called a *seismograph*, measures ground displacements on the order of a micron (10^{-6} m) and produces the seismogram on a paper strip chart attached to a rotating drum. Figure 1.1 shows the record of the Loma Prieta earthquake, centered in the Santa Cruz mountains of northern California, which struck the San Francisco Bay area on 18 October 1989.

Seismologists analyze such a signal in several ways. The total deflection of the pen across the chart is useful in determining the temblor's magnitude. Seismograms register three important types of waves: the *primary*, or *P waves*; the *secondary*, or *S waves*; and the *surface waves*. P waves arrive first, and they are compressive, so their direction of motion aligns with the wave front propagation [22]. The transverse S waves follow. They oscillate perpendicular to the direction of propagation. Finally, the large, sweeping surface waves appear on the trace.

This simple example illustrates processing and analysis concepts. Processing the seismogram signal is useful to remove noise. Noise can be minute ground motions from human activity (construction activity, heavy machinery, vehicles, and the like), or it may arise from natural processes, such as waves hitting the beach. Whatever the source, an important signal processing operation is to smooth out these minute ripples in the seismogram trace so as to better detect the occurrence of the initial indications of a seismic event, the P waves. They typically manifest themselves as seismometer needle motions above some threshold value. Then the analysis problem of finding when the S waves begin is posed. Figure 1.1 shows the result of a signal analysis; it slices the Loma Prieta seismogram into its three constituent wave

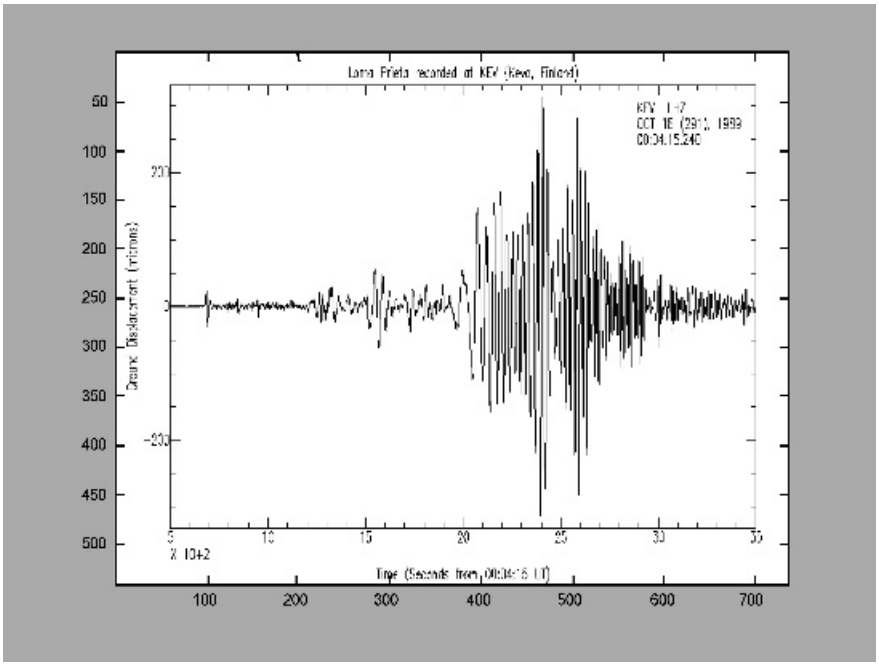


Fig. 1.1. Seismogram of the magnitude 7.1 Loma Prieta earthquake, recorded by a seismometer at Kevo, Finland. The first wiggle—some eight minutes after the actual event—marks the beginning of the low-magnitude P waves. The S waves arrive at approximately $t =$ 1200 s, and the large sweeping surface waves begin near $t = 2000$ s.

trains. This type of signal analysis can be performed by inspection on analog seismograms.

Now, the time interval between the arrival of the P and S waves is critical. These undulations are simultaneously created at the earthquake's epicenter; however, they travel at different, but known, average speeds through the earth. Thus, if an analysis of the seismogram can reveal the time that these distinct wave trains arrive, then the time difference can be used to measure the distance from the instrument to the earthquake's epicenter. Reports from three separate seismological stations are sufficient to locate the epicenter. Analyzing smaller earthquakes is also important. Their location and the frequency of their occurrence may foretell a larger temblor [23]. Further, soundings in the earth are indicative of the underlying geological strata; seismologists use such methods to locate oil deposits, for example [24]. Other similar applications include the detection of nuclear arms detonations and avalanches. For all of these reasons—scientific, economic, and public safety—seismic signal intepretation is one of the most important areas in signal analysis and one of the areas in which new methods of signal analysis have been pioneered. These further signal interpretation tasks are more troublesome for human interpreters. The signal behavior that distinguishes a small earthquake from a distant nuclear detonation is not apparent. This demands thorough computerized analysis.

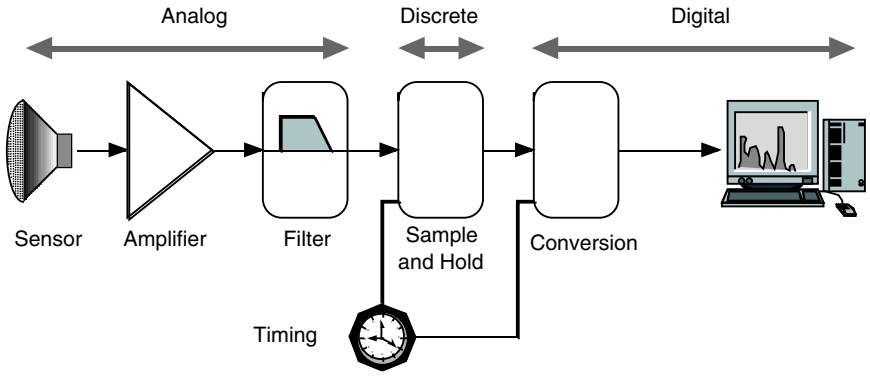


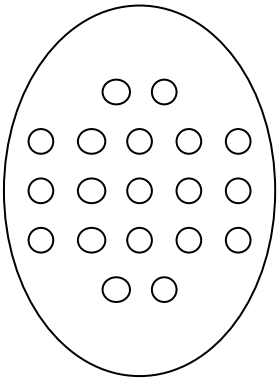
Fig. 1.2. Signal acquisition into a computer. Analog, discrete, and digital signals each occur—at least in principle—within such a system.

Suppose, therefore, that the signal is a discrete phenomenon, so that it occurs only at separate time instants or distance intervals and not continuously. Then we represent it as a function on a subset of the integers $x(n)$ and we identify $x(n)$ as a *discrete signal*. Furthermore, some discrete signals may have only a limited range of values. Their measurable values can be stored in the memory cells of a digital computer. The discrete signals that satisfy this further constraint are called *digital signals*.

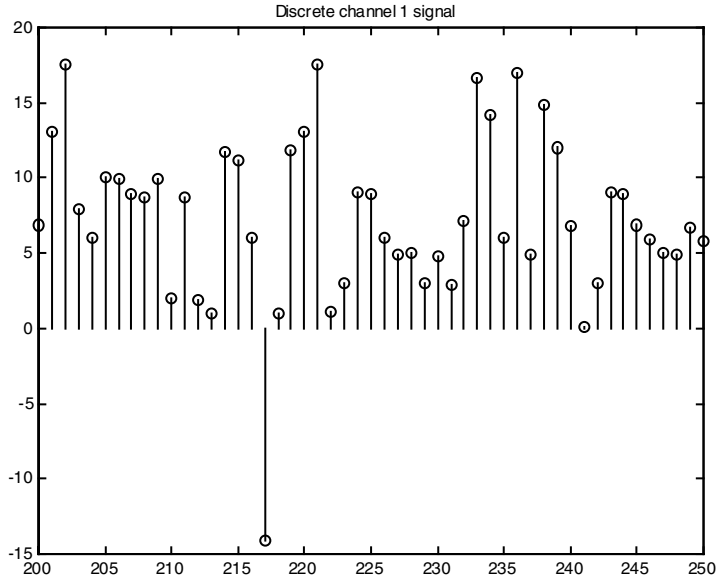
Each of these three types of signals occurs at some stage in a conventional computerized signal acquisition system (Figure 1.2). Analog signals arise from some quantifiable, real-world process. The signal arrives at an interface card attached to the computer's input–output bus.

There are generally some signal amplification and conditioning components, all analog, at the system's front end. At the sample and hold circuit, a momentary storage component—a capacitor, for example—holds the signal value for a small time interval. The sampling occurs at regular intervals, which are set by a timer. Thus, the sequence of quantities appearing in the sample and hold device represents the discrete form of the signal. While the measurable quantity remains in the sample and hold unit, a digitization device composes its binary representation. The extracted value is moved into a digital acquisition register of finite length, thereby completing the analog-to-digital conversion process. The computer's signal processing software or its input–output driver reads the digital signal value out of the acquisition register, across the input–output bus, and into main memory. The computer itself may be a conventional general-purpose machine, such as a personal computer, an engineering workstation, or a mainframe computer. Or the processor may be one of the many special purpose *digital signal processors* (DSPs) now available. These are now a popular design choice in signal processing and analysis systems, especially those with strict execution time constraints.

Some natural processes generate more than one measurable quantity as a function of time. Each such quantity can be regarded as a separate signal, in which case



(a)



(b)

Fig. 1.3. A multichannel signal: The electroencephalogram (EEG) taken from a healthy young person, with eyes open. The standard EEG sensor arrangement consists of 19 electrodes (a). Discrete data points of channel one (b). Panels (c) and (d) show the complete traces for the first two channels, $x_1(n)$ and $x_2(n)$. These traces span an eight second time interval: 1024 samples. Note the jaggedness superimposed on gentler wavy patterns. The EEG varies according to whether the patient's eyes are open and according to the health of the individual; markedly different EEG traces typify, for example, Alzheimer's disease.

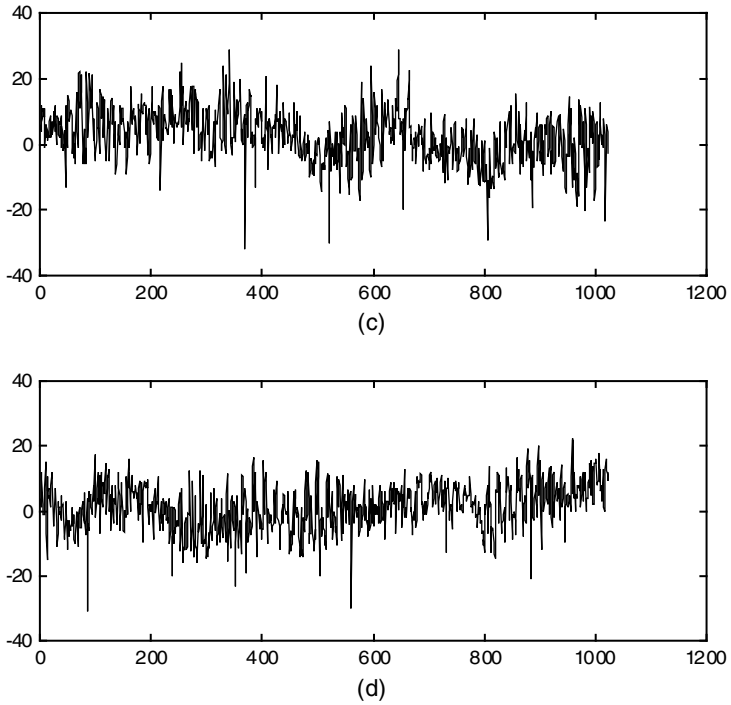


Fig. 1.3 (*Continued*)

they are all functions of the same independent variable with the same domain. Alternatively, it may be technically useful to maintain the multiple quantities together as a vector. This is called a *multichannel* signal. We use boldface letters to denote multichannel signals. Thus, if $\boldsymbol{x}$ is analog and has N channels, then $\boldsymbol{x}(t) = (x_1(t), x_2(t), \ldots, x_N(t))$, where the analog $x_i(t)$ are called the *component* or *channel* signals. Similarly, if $\boldsymbol{x}$ is discrete and has N channels, then $\boldsymbol{x}(n) = (x_1(n), x_2(n), \ldots, x_N(n))$.

One biomedical signal that is useful in diagnosing brain injuries, mental illness, and conditions such as Alzheimer's disease is the *electroencephalogram* (EEG) [25], a multichannel signal. It records electrical potential differences, or voltages, that arise from the interactions of massive numbers of neurons in different parts of the brain. For an EEG, 19 electrodes are attached from the front to the back of the scalp, in a two–five–five–five–two arrangement (Figure 1.3).

The EEG traces in Figure 1.3 are in fact digital signals, acquired one sample every 7.8 ms, or at a sampling frequency of 128 Hz. The signal appears to be continuous in nature, but this is due to the close spacing of the samples and linear interpolation by the plotting package.

Another variation on the nature of signals is that they may be functions of more than one independent variable. For example, we might measure air

temperature as a function of height: $T(h)$ is an analog signal. But if we consider that the variation may occur along a north-to-south line as well, then the temperature depends upon a distance measure x as well: $T(x, h)$. Finally, over an area with location coordinates (x, y), the air temperature is a continuous function of three variables $T(x, y, h)$. When a signal has more than one independent variable, then it is a *multidimensional* signal. We usually think of an "image" as recording light intesity measurements of a scene, but multidimensional signals—especially those with two or three independent variables—are usually called *images*. Images may be discrete too. Temperature readings taken at kilometer intervals on the ground and in the air produce a discrete signal $T(m, n, k)$. A discrete signal is a sequence of numerical values, whereas an image is an array of numerical values. Two-dimensional image elements, especially those that represent light intensity values, are called *pixels*, an acronym for *picture elements*. Occasionally, one encounters the term *voxel*, which is a three-dimensional signal value, or a *volume element*.

An area of multidimensional signal processing and analysis of considerable importance is the intepretation of images of landscapes acquired by satellites and high altitude aircraft. Figure 1.4. shows some examples. Typical tasks are to automatically distinguish land from sea; determine the amount and extent of sea ice; distinguish agricultural land, urban areas, and forests; and, within the agricultural regions, recognize various crop types. These are remote sensing applications.

Processing two-dimensional signals is more commonly called picture or image processing, and the task of interpreting an image is called image analysis or computer vision. Many researchers are involved in robotics, where their efforts couple computer vision ideas with manipulation of the environment by a vision-based machine. Consequently, there is a vast, overlapping literature on image processing [26–28], computer vision [29–31], and robotics [32].

Our subject, signal analysis, concentrates on the mathematical foundations, processing, and especially the intepretation of one-dimensional, single-valued signals. Generally, we may select a single channel of a multichannel signal for consideration; but we do not tackle problems specific to multichannel signal interpretation. Likewise, we do not delve deeply into image processing and analysis. Certain images do arise, so it turns out, in several important techniques for analyzing signals. Sometimes a daunting one-dimensional problem can be turned into a tractable two-dimensional task. Thus, we prefer to pursue the one-dimensional problem into the multidimensional realm only to the point of acknowledging that a straightforward image analysis will produce the intepretation we seek.

So far we have introduced the basic concepts of signal theory, and we have considered some examples: analog, discrete, multichannel, and multidimensional signals. In each case we describe the signals as sequences of numerical values, or as a function of an independent time or other spatial dimension variable. This constitutes a time-domain description of a signal. From this perspective, we can display a signal, process it to produce another signal, and describe its significant features.

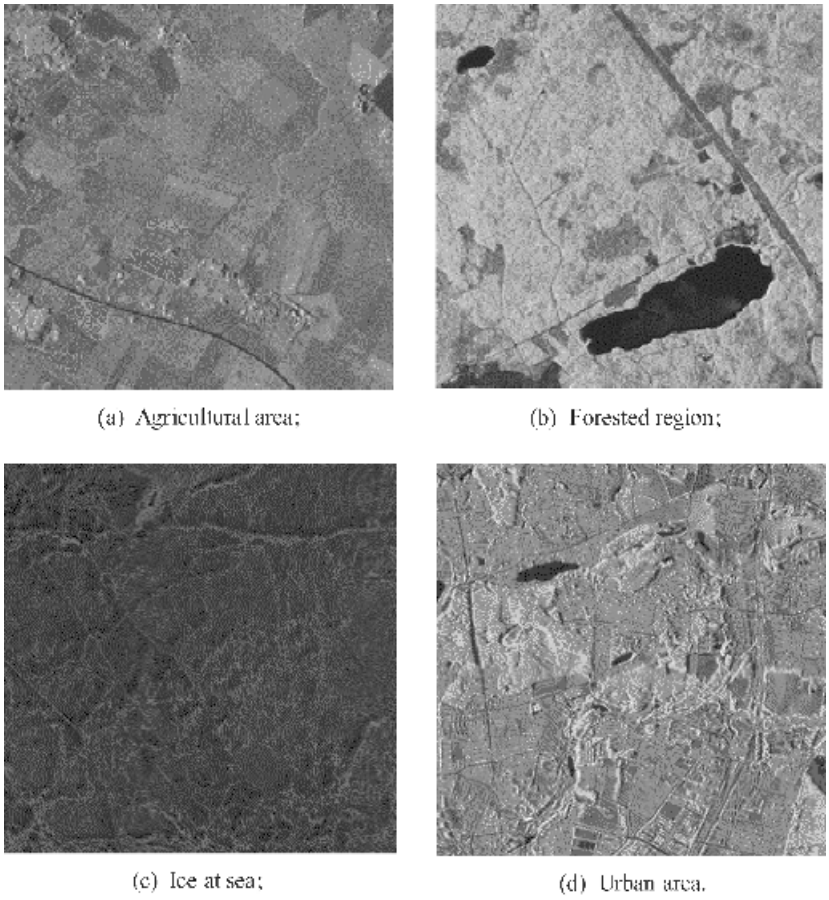


Fig. 1.4. Aerial scenes. Distinguishing terrain types is a typical problem of image analysis, the interpretation of two-dimensional signals. Some problems, however, admit a one-dimensional solution. A sample line through an image is in fact a signal, and it is therefore suitable for one-dimensional techniques. (a) Agricultural area. (b) Forested region. (c) Ice at sea. (d) Urban area.

1.1.2 Time-Domain Description of Signals

Since time flows continuously and irreversibly, it is natural to describe sequential signal values as given by a time ordering. This is often, but not always, the case; many signals depend upon a distance measure. It is also possible, and sometimes a very important analytical step, to consider signals as given by order of a salient event. Conceiving the signal this way makes the dependent variable—the signal value—a function of time, distance, or some other quantity indicated between successive events. Whether the independent variable is time, some other spatial dimension, or a counting of events, when we represent and discuss a signal in terms of its ordered values, we call this the *time-domain* description of a signal.

Note that a precise time-domain description may elude us, and it may not even be possible to specify a signal's values. A fundamentally unknowable or random process is the source of such signals. It is important to develop methods for handling the randomness inherent in signals. Techniques that presuppose a theory of signal randomness are the topic of the final section of the chapter.

Next we look further into two application areas we have already touched upon: biophysical and geophysical signals. Signals from representative applications in these two areas readily illustrate the time-domain description of signals.

1.1.2.1 Electrocardiogram Interpretation. Electrocardiology is one of the earliest techniques in biomedicine. It also remains one of the most important. The excitation and recovery of the heart muscles cause small electrical potentials, or voltages, on the order of a millivolt, within the body and measurable on the skin. Cardiologists observe the regularity and shape of this voltage signal to diagnose heart conditions resulting from disease, abnormality, or injury. Examples include cardiac dysrhythmia and fibrillation, narrowing of the coronary arteries, and enlargement of the heart [33]. Automatic interpretation of ECGs is useful for many aspects of clinical and emergency medicine: remote monitoring, as a diagnostic aid when skilled cardiac care personnel are unavailable, and as a surgical decision support tool.

A modern electrocardiogram (ECG or EKG) contains traces of the voltages from 12 leads, which in biomedical parlance refers to a configuration of electrodes attached to the body [34]. Refer to Figure 1.5. The voltage between the arms is Lead I, Lead II is the potential between the right arm and left leg, and Lead III reads between the left arm and leg. The WCT is a common point that is formed by connecting the three limb electrodes through weighting resistors. Lead aVL measures potential difference between the left arm and the WCT. Similarly, lead aVR is the voltage between the right arm and the WCT. Lead aVF is between the left leg and the WCT. Finally, six more electrodes are fixed upon the chest, around the heart. Leads V1 through V6 measure the voltages between these sensors and the WCT. This circuit

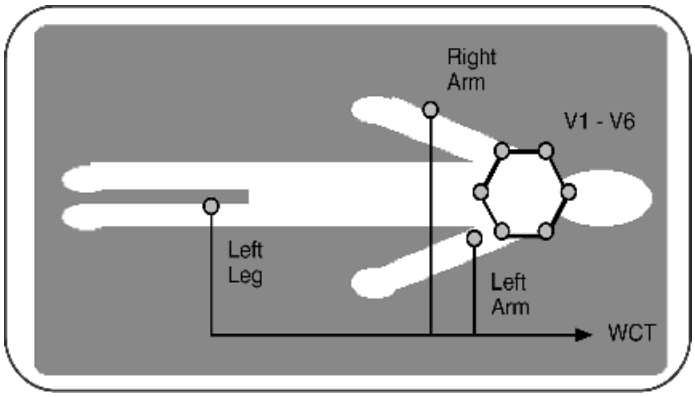


Fig. 1.5. The standard ECG configuration produces 12 signals from various electrodes attached to the subject's chest, arms, and leg.

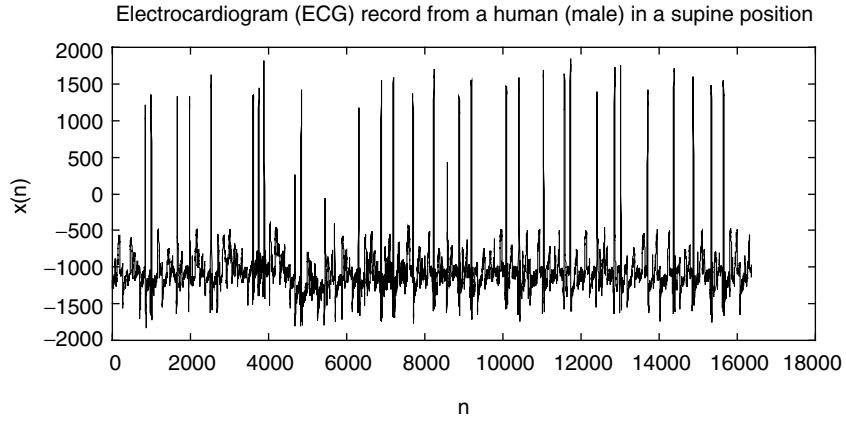


Fig. 1.6. One lead of an ECG: A human male in supine position. The sampling rate is 1 kHz, and the samples are digitized at 12 bits per sample. The irregularity of the heartbeat is evident.

arrangement is complicated; in fact, it is redundant. Redundancy provides for situations where a lead produces a poor signal and allows some cross-checking of the readings. Interpretation of 12-lead ECGs requires considerable training, experience, and expert judgment.

What does an ECG trace look like? Figure 1.6 shows an ECG trace from a single lead. Generally, an ECG has three discernible pulses: the P wave, the QRS complex, and the T wave. The P wave occurs upon excitation of the auricles of the heart, when they draw in blood from the body and lungs. The large-magnitude QRS complex occurs during the contraction of the vertricles as they contract to pump blood out of the heart. The Q and S waves are negative pulses, and the R wave is a positive pulse. The T wave arises during repolarization of the ventricles. The ECG signal is originally analog in nature; it is the continuous record of voltages produce across the various leads supported by the instrument. We could attach a millivoltmeter across an electrode pair and watch the needle jerk back and forth. Visualizing the signal's shape is easier with an oscilloscope, of course, because the instrument records the trace on its cathode ray tube. Both of these instruments display analog waveforms. If we could read the oscilloscope's output at regular time instants with perfect precision, then we would have—in principle, at least—a discrete representation of the ECG. But for computer display and automatic interpretation, the analog signal must be converted to digital form. In fact, Figure 1.6 is the result of such a digitization. The signal $v(n)$ appears continuous due to the large number of samples and the interpolating lines drawn by the graphics package that produced the illustration.

Interpreting ECGs is often difficult, especially in abnormal traces. A wide literature describing the 12-lead ECG exists. There are many guides to help technicians, nurses, and physicians use it to diagnose heart conditions. Signal processing and analysis of ECGs is a very active research area. Reports on new techniques, algorithms, and comparison studies continue to appear in the biomedical engineering and signal analysis literature [35].

One technical problem in ECG interpretation is to assess the regularity of the heart beat. As a time-domain signal description problem, this involves finding the separation between peaks of the QRS complex (Figure 1.6). Large time variations between peaks indicates dysrhythmia. If the time difference between two peaks, $v(n_1)$ and $v(n_0)$, is $\Delta_T = n_1 - n_0$, then the instantaneous heart rate becomes $60(\Delta_T)^{-1}$ beats/m. For the sample in Figure 1.6, this crude computation will, however, produce a wildly varying value of doubtful diagnostic use. The application calls for some kind of averaging and summary statistics, such as a report of the standard deviation of the running heart rate, to monitor the dysrhythmia.

There remains the technical problem of how to find the time location of QRS peaks. For an ideal QRS pulse, this is not too difficult, but the signal analysis algorithms must handle noise in the ECG trace. Now, because of the noise in the ECG signal, there are many local extrema. Evidently, the QRS complexes represent signal features that have inordinately high magnitudes; they are mountains above the forest of small-scale artifacts. So, to locate the peak of a QRS pulse, we might select a threshold M that is bigger than the small artifacts and smaller than the QRS peaks. We then deem any maximal, contiguous set of values $S = \{(n, v(n))\colon v(n) > M\}$ to be a QRS complex. Such regions will be disjoint. After finding the maximal value inside each such QRS complex, we can calculate Δ_T between each pair of maxima and give a running heart rate estimate. The task of dividing the signal up into disjoint regions, such as for the QRS pulses, is called signal *segmentation*. Chapter 4 explores this time domain procedure more thoroughly.

When there is poor heart rhythm, the QRS pulses may be jagged, misshapen, truncated, or irregulary spaced. A close inspection of the trace in Figure 1.7 seems to reveal this very phenomenon. In fact, one type of ventricular disorder that is

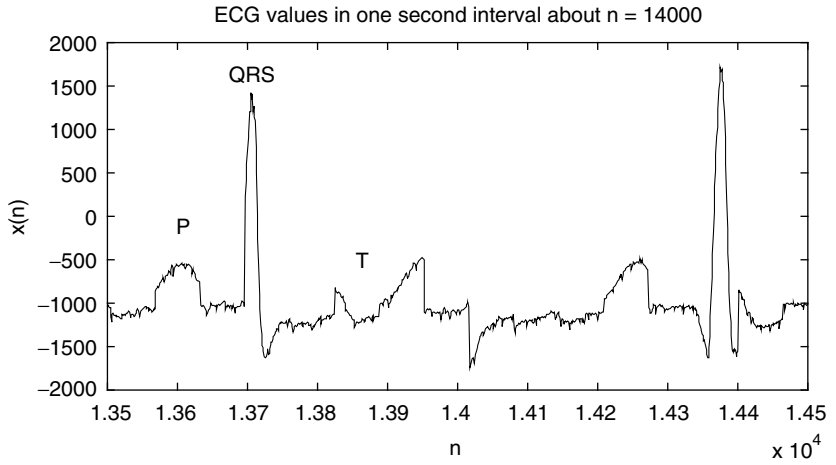


Fig. 1.7. Electrocardiogram of a human male, showing the fundametal waves. The 1-s time span around sample n = 14,000 is shown for the ECG of Figure 1.6. Note the locations of the P wave, the QRS complex, and—possibly—the T wave. Is there a broken P wave and a missing QRS pulse near the central time instant?

detectable in the ECG, provided that it employs a sufficiently high sampling rate, is *splintering* of the QRS complex. In this abnormal condition, the QRS consists of many closely spaced positive and negative transitions rather than a single, strong pulse. Note that in any ECG, there is a significant amount of signal noise. This too is clearly visible in the present example. Good peak detection and pulse location, especially for the smaller P and T waves, often require some data smoothing method. Averaging the signal values produces a smoother signal $w(n)$:

$$w(n) = \frac{1}{3}[v(n-1) + v(n) + v(n+1)]. \tag{1.1}$$

The particular formula (1.1) for processing the raw ECG signal to produce a less noisy $w(n)$ is called *moving average smoothing* or *moving average filtering*. This is a typical, almost ubiquitous signal processing operation. Equation (1.1) performs averaging within a symmetric window of width three about $v(n)$. Wider windows are possible and often useful. A window that is too wide can destroy signal features that bear on interpretation. Making a robust application requires judgment and experimentation.

Real-time smoothing operations require asymmetric windows. The underlying reason is that a symmetric smoothing window supposes knowledge of future signal values, such as $v(n + 1)$. To wit, as the computer monitoring system acquires each new ECG value $v(n)$, it can calculate the average of the last three values:

$$w(n) = \frac{1}{3}[v(n-2) + v(n-1) + v(n)]; \tag{1.2}$$

but at time instant n, it cannot possibly know the value of $v(n +1)$, which is necessary for calculating (1.1). If the smoothing operation occurs offline, after the entire set of signal values of interest has already been acquired and stored, then the whole range of signal values is accessible by the computer, and calculation (1.1) is, of course, feasible. When smoothing operations must procede in lockstep with acquisition operations, however, smoothing windows that look backward in time (1.2) must be applied.

Yet another method from removing noise from signals is to produce a signal whose values are the median of a window of raw input values. Thus, we might assign

$$w(n) = \text{Median}\{v(n-2), v(n-1), v(n), v(n+1), v(n+2)\} \tag{1.3}$$

so that $w(n)$ is the input value that lies closest to the middle of the range of five values around $v(n)$. A median filter tends to be superior to a moving average filter when the task is to remove isolated, large-magnitude spikes from a source signal. There are many variants. In general, smoothing is a common early processing step in signal analysis systems. In the present application, smoothing reduces the jagged noise in the ECG trace and improves the estimate of the QRS peak's location.

Contemplating the above algorithms for finding QRS peaks, smoothing the raw data, and estimating the instantaneous heart rate, we can note a variety of design choices. For example, how many values should we average to smooth the data? A span too small will fail to blur the jagged, noisy regions of the signal. A span too large may erode some of the QRS peaks. How should the threshold for segmenting QRS pulses be chosen? Again, an algorithm using values too small will falsely identify noisy bumps as QRS pulses. On the other hand, if the threshold values chosen are too large, then valid QRS complexes will be missed. Either circumstance will cause the application to fail. Can the thresholds be chosen automatically? The chemistry of the subject's skin could change while the leads are attached. This can cause the signal as a whole to trend up or down over time, with the result that the original threshold no longer works. Is there a way to adapt the threshold as the signal average changes so that QRS pulses remain detectable? These are but a few of the problems and tradeoffs involved in time domain signal processing and analysis.

Now we have illustrated some of the fundamental concepts of signal theory and, through the present example, have clarified the distinction between signal processing and analysis. Filtering for noise removal is a processing task. Signal averaging may serve our purposes, but it tends to smear isolated transients into what may be a quite different overall signal trend. Evidently, one aberrent upward spike can, after smoothing, assume the shape of a QRS pulse. An alternative that addresses this concern is median filtering. In either case—moving average or median filtering—the algorithm designer must still decide how wide to make the filters and discover the proper numerical values for thresholding the smoothed signal. Despite the analytical obstacles posed by signal noise and jagged shape, because of its prominence, the QRS complex is easier to characterize than the P and T waves.

There are alternative signal features that can serve as indicators of QRS complex location. We can locate the positive or negative transitions of QRS pulses, for example. Then the midpoint between the edges marks the center of each pulse, and the distance between these centers determines the instantaneous heart rate. This changes the technical problem from one of finding a local signal maximum to one of finding the positive- or negative-transition edges that bound the QRS complexes. Signal analysis, in fact, often revolves around edge detection. A useful indicator of edge presence is the discrete derivative, and a simple threshold operation identifies the significant changes.

1.1.2.2 Geothermal Measurements. Let us investigate an edge detection problem from geophysics. Ground temperature generally increases with depth. This variation is not as pronounced as the air temperature fluctuations or biophysical signals, to be sure, but local differences emerge due to the geological and volcanic history of the spot, thermal conductivity of the underlying rock strata, and even the amount of radioactivity. Mapping changes in ground termperature are important in the search for geothermal energy resources and are a supplementary indication of the underlying geological structures. If we plot temperature versus depth, we have a

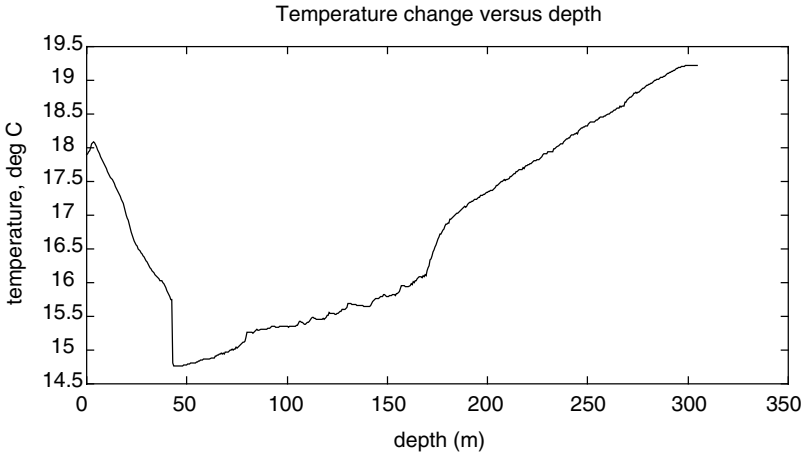


Fig. 1.8. A geothermal signal. The earth's temperature is sampled at various depths to produce a discrete signal with a spatially independent variable.

signal—the *geothermal gradient*—that is a function of distance, not time. It ramps up about 10°C per kilometer of depth and is a primary indicator for geothermal prospecting. In general, the geothermal gradient is higher for oceanic than for continental crust. Some 5% of the area of the United States has a gradient in the neighborhood of 40°C per kilometer of depth and has potential for use in geothermal power generation.

Mathematically, the geothermal gradient is the derivative of the signal with respect to its independent variable, which in this case measures depth into the earth. A very steep overall gradient may promise a geothermal energy source. A localized large magnitude gradient, or edge, in the temperature profile marks a geological artifact, such as a fracture zone. An example of the variation in ground temperature as one digs into the earth is shown in Figure 1.8.

The above data come from the second of four wells drilled on the Georgia–South Carolina border, in the eastern United States, in 1985 [36]. The temperature first declines with depth, which is typical, and then warmth from the earth's interior appears. Notice the large-magnitude positive gradients at approximately 80 and 175 m; these correspond to fracture zones. Large magnitude deviations often represent physically significant phenomena, and therein lies the importance of reliable methods for detecting, locating, and interpreting signal edges. Finding such large deviations in signal values is once again a time-domain signal analysis problem.

Suppose the analog ground temperature signal is $g(s)$, where s is depth into the earth. We seek large values of the derivative $g'(s) = dg/ds$. Approximating the derivative is possible once the data are digitized. We select a sampling interval $D > 0$ and set $x(n) = g(nD)$; then $x'(n) = x(n+1) - x(n-1)$ approximates the geothermal gradient at depth nD meters. It is further necessary to identify a threshold M for what constitutes a significant geothermal gradient. Threshold selection may rely upon expert scientific knowledge. A geophysicist might suggest significant gradients

for the region. If we collect some statistics on temperature gradients, then the outlying values may be candidates for threshold selection. Again, there are local variations in the temperature profile, and noise does intrude into the signal acquisition apparatus. Hence, preliminary signal smoothing may once again be useful. Toward this end, we may also employ discrete derivative formulas that use more signal values:

$$x'(n) = \frac{1}{12}[x(n-2) - 8x(n-1) + 8x(n+1) - x(n+2)]\,. \tag{1.4}$$

Standard numerical analysis texts provide many alternatives [37]. Among the problems at the chapter's end are several edge detection applications. They weigh some of the alternatives for filtering, threshold selection, and finding extrema.

For now, let us remark that the edges in the ECG signal (Figure 1.6) are far steeper than the edges in the geothermal trace (Figure 1.8). The upshot is that the signal analyst must tailor the discrete derivative methods to the data at hand. Developing methods for edge detection that are robust with respect to sharp local variation of the signal features proves to be a formidable task. Time-domain methods, such as we consider here, are usually appropriate for edge detection problems. There comes a point, nonetheless, when the variety of edge shapes, the background noise in the source signals, and the diverse gradients cause problems for simple time domain techniques. In recent years, researchers have turned to edge detection algorithms that incorporate a notion of the size or scale of the signal features. Chapter 4 has more to say about time domain signal analysis and edge detection, in particular. The later chapters round out the story.

1.1.3 Analysis in the Time-Frequency Plane

What about signals whose values are symbolic rather than numeric? In ordinary usage, we consider sequences of signs to be signals. Thus, we deem the display of flags on a ship's mast, a series of hand gestures between baseball players, DNA codes, and, in general, any sequence of codes to all be "signals." We have already taken note of such usages. And this is an important idea, but we shall not call such a symbolic sequence a signal, reserving for that term a narrow scientific definition as an ordered set of numbers. Instead, we shall define a sequence of abstract symbols to be a structural interpretation of a signal.

It is in fact the conversion of an ordered set of numerical values into a sequence of symbols that constitutes a signal interpretation or analysis. Thus, a microphone receives a logitudinal compressive sound wave and converts it into electrical impulses, thereby creating an analog signal. If the analog speech signal is digitized, processed, and analyzed by a speech recognition engine, then the output in the form of ASCII text characters is a symbolic sequence that interprets, analyzes, or assigns meaning to the signal. The final result may be just the words that were uttered. But, more likely, the speech interpretation algorithms will generate a variety of intermediate representations of the signal's structure. It is common to build a large hierarchy of interpretations: isolated utterances; candidate individual word sounds within the utterances; possible word recognition results; refinements from grammatical rules and application context; and, finally, a structural result.

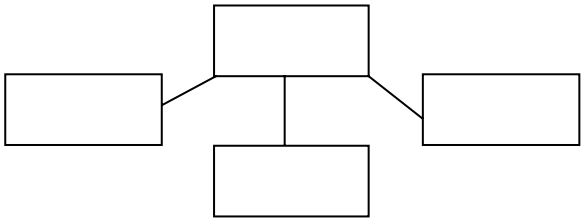


Fig. 1.9. Elementary graph structure for seismograms. One key analytical parameter is the time interval between the P waves and the S waves.

This framework applies to the applications covered in this section. A simple sequence of symbols representing the seismometer background, P waves, S waves, and surface waves may be the outcome of a structural analysis of a seismic signal (Figure 1.9).

The nodes of such a structure may have further information attached to them. For instance, the time-domain extent of the region, a confidence measure, or other analytical signal features can be inserted into the node data structure. Finding signal edges is often the prelude to a structural description of a signal. Figure 1.10

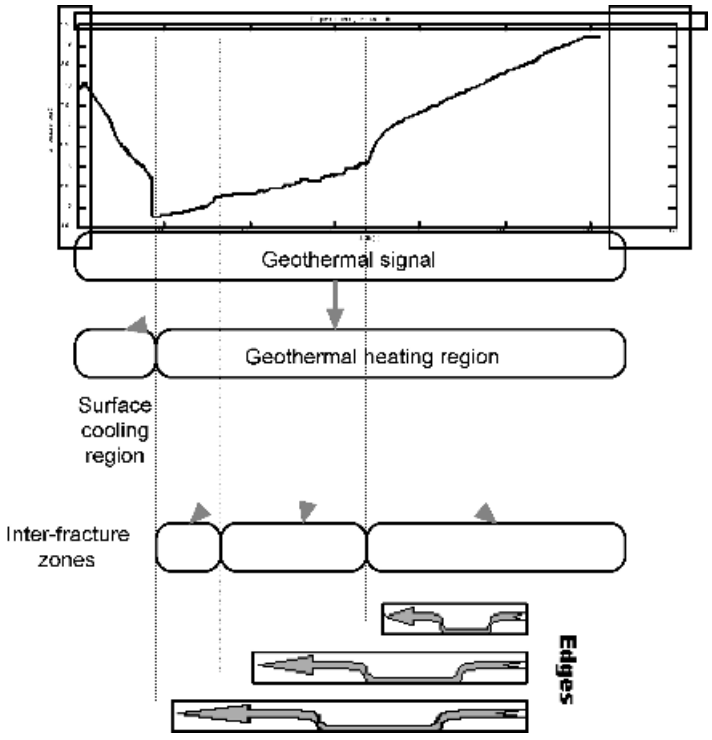


Fig. 1.10. Hypothetical geothermal signal structure. The root note of the interpretive structure represents the entire time-domain signal. Surface strata exhibit a cooling trend. Thereafter, geothermal heating effects are evident. Edges within the geothermal heating region indicate narrow fracture zones.

illustrates the decomposition of the geothermal profile from Figure 1.8 into a relational structure.

For many signal analysis problems, more or less flat relational structures that divide the signal domain into distinct regions are sufficient. Applications such as natural language understanding require more complicated, often hierarchical graph structures. Root nodes describe the coarse features and general subregions of the signal. Applying specialized algorithms to these distinct regions decomposes them further. Some regions may be deleted, further subdivided, or merged with their neighbors. Finally, the resulting graph structure can be compared with existing structural models or passed on to higher-level artificial intelligence applications.

1.1.4 Other Domains: Frequency and Scale

While we can achieve some success in processing and analyzing signals with elementary time-domain techniques, applied scientists regularly encounter applications demanding more sophisticated treatment. Thinking for a moment about the seismogram examples, we considered one aspect of their interpretation: finding the time difference between the arrival of the P and S waves. But how can one distinguish between the two wave sets? The distinction between them, which analysis algorithms must find, is in their oscillatory behavior and the magnitude of the oscillations. There is no monotone edge, such as characterized the geothermal signal. Rather, there is a change in the repetitiveness and the sweep of the seismograph needle's wiggling. When the oscillatory nature of a signal concerns us, then we are interested in its periodicity—or in other words, the reciprocal of period, the *frequency.*

Frequency-domain signal descriptions decompose the source signals into sinusoidal components. This strategy does improve upon pure time domain methods, given the appropriate application. A frequency-domain description uses some set of sinusoidal signals as a basis for describing a signal. The frequency of the sinusoid that most closely matches the signal is the principal frequency component of the signal. We can delete this principal frequency component from the source signal to get a difference signal. Then, we iterate. The first difference signal is further frequency analyzed to get a secondary periodic component and, of course, a second difference signal. The sinusoidal component identification and extraction continue until the difference signal consists of nothing but small magnitude, patternless, random perturbations—noise. This is a familiar procedure. It is just like the elementary linear algebra problem of finding the expansion coefficients of a given vector in terms of a basis set.

Thus, a frequency-domain approach is suitable for distinguishing the P waves from the S waves in seismogram interpretation. But, there is a caveat. We cannot apply the sinusoidal signal extraction to the whole signal, but rather only to small pieces of the signal. When the frequency components change radically on the separate, incoming small signal pieces, then the onset of the S waves must be at hand. The subtlety is to decide how to size the small signal pieces that will be subject to frequency analysis. If the seismographic station is far away, then the time interval

between the initial P waves and the later S waves is large, and fairly large subintervals should suffice. If the seismographic station is close to the earthquake epicenter, on the other hand, then the algorithm must use very small pieces, or it will miss the short P wave region of the motion entirely. But if the pieces are made too small, then they may contain too few discrete samples for us to perform a frequency analysis. There is no way to know whether a temblor that has not happened yet will be close or far away. And the dilemma is how to size the signal subintervals in order to analyze all earthquakes, near and far, and all possible frequency ranges for the S and P waves.

It turns out that although such a frequency-domain approach as we describe is adequate for seismic signals, the strategy has proven to be problematic for the interpretation of electrocardiograms. The waves in abnormal ECGs are sometimes too variable for successful frequency-domain description and analysis.

Enter the notion of a *scale-domain* signal description. A scale-domain description of a signal breaks it into similarly shaped signal fragments of varying sizes. Problems that involve the time-domain size of signal features tend to favor this type of representation. For example, a scale-based analysis can offer improvements in electrocardiogram analysis; in this field it is a popular redoubt for researchers that have experimented with time domain methods, then frequency-domain methods, and still find only partial success in interpreting ECGs.

We shall also illustrate the ideas of frequency- and scale-domain descriptions in this first chapter. A complete understanding of the methods of frequency- and scale-domain descriptions requires a considerable mathematical expertise. The next two sections provide some formal definitions and a variety of mathematical examples of signals. The kinds of functions that one normally studies in algebra, calculus, and mathematical analysis are quite different from the ones at the center of signal theory. Functions representing signals are often discontinuous; they tend to be irregularly shaped, blocky, spiky, and altogether more ragged than the smooth and elegant entities of pure mathematics.

1.2 ANALOG SIGNALS

At the scale of objects immediately present to human consciousness and at the macroscopic scale of conventional science and technology, measurable phenomena tend to be continuous in nature. Hence, the raw signals that issue from nature—temperatures, pressures, voltages, flows, velocities, and so on—are commonly measured through analog instruments. In order to study such real-world signals, engineers and scientists model them with mathematical functions of a real variable. This strategy brings the power and precision of mathematical analysis to bear on engineering questions and problems that concern the acquisition, transmission, interpretation, and utilization of natural streams of numbers (i.e., signals).

Now, at a very small scale, in contrast to our perceived macroscopic world, natural processes are more discrete and quantized. The energy of electromagnetic radiation exists in the form of individual quanta with energy $E = h/\lambda$, where h is

Planck's constant,[3] and λ is the wavelength of the radiation. Phenomena that we normally conceive of possessing wave properties exhibit certain particle-like behaviors. On the other hand, elementary bits of matter, electrons for instance, may also reveal certain wave-like aspects. The quantization of nature at the subatomic and atomic levels leads to discrete interactions at the molecular level. Lumping ever greater numbers of discretized interactions together, overall statistics take priority over particular interactions, and the continuous nature of the laws of nature at a large scale then become apparent.[4] Though nature is indeed discrete at the microlevel, the historical beginnings of common sense, engineering, and scientific endeavor involve reasoning with continuously measurable phenomena. Only recently, within the last century have the quantized nature of the interactions of matter and energy become known. And only quite recently, within our own lifetimes, have machines become available to us—digital computers—that require for their application the discretization of their continuous input data.

1.2.1 Definitions and Notation

Analog signal theory proceeds directly from the analysis of functions of a real variable. This material is familiar from introductory calculus courses. Historically, it also precedes the development of discrete signal theory. And this is a curious circumstance, because the formal development of analog signal theory is far more subtle—some would no doubt insist the right term is perilous—than discrete time signal processing and analysis.

Definition (Analog Signals). An *analog signal* is a function $x: \mathbb{R} \rightarrow \mathbb{R}$, where $\mathbb{R}$ is the set of real numbers, and $x(t)$ is the signal value at time t. A *complex-valued* analog signal is a function $x: \mathbb{R} \rightarrow \mathbb{C}$. Thus, $x(t) = x_r(t) + jx_i(t)$, where $x_r(t)$ is the real part of $x(t)$; $x_i(t)$ is the imaginary part of $x(t)$; both of these are real-valued signals; and $j^2 = -1$.

Thus, we simply identify analog signals with functions of a real variable. Ordinarily, analog signals, such the temperature of an oven varying over time, take on real values. In other cases, where signal timing relationships come into question, or the frequency content of signals is an issue, complex-valued signals are often used. We will work with both real- and complex-valued signals in this section. Section 1.7 considers the complex number system, complex-valued signals, and the mathematics of complex numbers in more detail. Complex-valued signals arise primarily in the study of signal frequency.

[3]To account for the observation that the maximum velocity of electrons dislodged from materials depended on the frequency of incident light, Max Planck (1858–1947) conjectured that radiant energy consists of discrete packets, called *photons* or *quanta*, thus discovering the quantum theory.

[4]This process draws the attention of philosophers (N. Hartmann, *New Ways of Ontology*, translator R. C. Kuhn, Chicago: Henry Regnery, 1953) and scientists alike (W. Zurek, "Decoherence and the transition from quantum to classical," *Physics Today*, vol. 44, no. 10, pp. 36–44, October 1991).

Of course, the independent variable of an analog signal does not have to be a time variable. The pneumatic valve of a bicycle tire follows a sinusoidal course in height above ground as the rider moves down the street. In this case the analog signal is a function of distance ridden rather than time passed. And the geothermal gradient noted in the previous section is an example of a signal that is a function of depth in the earth's crust.

It is possible to generalize the above definition to include multichannel signals that take values in $\mathbb{R}^n$, $n \geq 2$. This is a straightforward generalization for all of the theory that we develop. Another way to generalize to higher dimensionality is to consider signals with domains contained in $\mathbb{R}^n$, $n \geq 2$. This is the discipline of image processing, at least for $n = 2$, 3, and 4. As a generalization of signal processing, it is not so straightforward as multichannel theory; the extra dimension in the independent signal variable leads to complications in signal interpretation and imposes severe memory and execution time burdens for computer-based applications.

We should like to point out that modeling natural signals with mathematical functions is an inherently flawed step; many functions do not correspond to any real-world signal. Mathematical functions can have nonzero values for arbitrarily large values of their independent variable, whereas in reality, such signals are impossible; every signal must have a finite past and eventually decay to nothing. To suppose otherwise would imply that the natural phenomenon giving rise to the signal could supply energy indefinitely. We can further imagine that some natural signals containing random noise cannot be exactly characterized by a mathematical rule associating one independent variable with another dependent variable.

But, is it acceptable to model real-world signals with mathematical models that eventually diminish to zero? This seems unsatisfactory. A real-world signal may decay at such a slow rate that in choosing a function for its mathematical model we are not sure where to say the function's values are all zero. Thus, we should prefer a theory of signals that allows signals to continue forever, perhaps diminishing at an allowable rate. If our signal theory accomodates such models, then we have every assurance that it can account for the wildest natural signal that the real world can offer. We will indeed pursue this goal, beginning in this first chapter. With persistence, we shall see that natural signals do have mathematical models that reflect the essential nature of the real-world phenomenon and yet are not limited to be zero within finite intervals. We shall find as well that the notion of randomness within a real-world signal can be accommodated within a mathematical framework.

1.2.2 Examples

The basic functions of mathematical analysis, known from algebra and calculus, furnish many elementary signal models. Because of this, it is common to mix the terms "signal" and "function." We may specify an analog signal from a formula that relates independent variable values with dependent variable values. Sometimes the formula can be given in closed form as a single equation defining the signal values. We may also specify other signals by defining them piecewise on their domain. Some functions may best be described by a geometric definition. Still other

functions representing analog signals may be more convenient to sketch rather than specify mathematically.

1.2.2.1 Polynomial, Rational, and Algebraic Signals. Consider, for example, the *polynomial* signal,

$$x(t) = \sum_{k=0}^{N} a_k t^k . \tag{1.5}$$

$x(t)$ has derivatives of all orders and is continuous, along with all of its derivatives. It is quite unlike any of nature's signals, since its magnitude, $|x(t)|$, will approach infinity as $|t|$ becomes large. These signals are familiar from elementary algebra, where students find their roots and plot their graphs in the Cartesian plane. The domain of a polynomial $p(t)$ can be divided into disjoint regions of concavity: concave upward, where the second derivative is positive; concave downward, where the second derivative is negative; and regions of no concavity, where the second derivative is zero, and $p(t)$ is therefore a line. If the domain of a polynomial $p(t)$ contain- san interval $a < t < b$ where $\frac{d^2}{dt^2}p(t) = 0$ for all $t \in (a, b)$, then $p(t)$ is a line. However familiar and natural the polynomials may be, they are not the signal family with which we are most intimately concerned in signal processing. Their behavior for large $|t|$ is the problem. We prefer mathematical functions that more closely resemble the kind of signals that occur in nature: Signals $x(t)$ which, as $|t|$ gets large, the signal either approaches a constant, oscillates, or decays to zero. Indeed, we expend quite an effort in Chapter 2 to discover signal families—called *function* or *signal spaces*—which are faithful models of natural signals.

The concavity of a signal is a very important concept in certain signal analysis applications. Years ago, the psychologist F. Attneave [38] noted that a scattering of simple curves suffices to convey the idea of a complex shape—for instance, a cat. Later, computer vision researchers developed the idea of assemblages of simple, oriented edges into complete theories of low-level image understanding [39–41]. Perhaps the most influential among them was David Marr, who conjectured that understanding a scene depends upon the extraction of edge information [39] over a range of visual resolutions from coarse to fine multiple scales. Marr challenged computer vision researchers to find processing and analysis paradigms within biological vision and apply them to machine vision. Researchers investigated the applications of concavity and convexity information at many different scales. Thus, an intricate shape might resolve into an intricate pattern at a fine scale, but at a coarser scale might appear to be just a tree. How this can be done, and how signals can be smoothed into larger regions of convexity and concavity without increasing the number of differently curved regions, is the topic of scale-space analysis [42,43]. We have already touched upon some of these ideas in our discussion of edges of the QRS complex of an electrocardiogram trace and in our discussion of the geothermal gradient. There the scale of an edge corresponded to the number of points incorporated in the discrete derivative computation. This is precisely the notion we are

trying to illustrate, since the scale of an edge is a measure of its time-domain extent. Describing signal features by their scale is most satisfactorily accomplished using special classes of signals (Section 1.6). At the root of all of this deep theory, however, are the basic calculus notion of the sign of the second derivative and the intuitive and simple polynomial examples.

Besides motivating the notions of convexity and concavity as component building blocks for more complicated shapes, polynomials are also useful in signal theory as interpolating functions. The theory of splines generalizes linear interpolation. It is one approach to the modern theory of wavelet transforms. Interpolating the values of a discrete signal with continuous polynomial sections—connecting the dots, so to speak—is the opposite process to sampling a continuous-domain signal.

If $p(t)$ and $q(t)$ are polynomials, then $x(t) = p(t)/q(t)$ is a *rational* function. Signals modeled by rational functions need to have provisions made in their definitions for the times t_0 when $q(t_0) = 0$. If, when this is the case, $p(t_0) = 0$ also, then it is possible that the limit,

$$\lim_{t \to t_0} \frac{p(t)}{q(t)} = r_0 = x(t_0), \tag{1.6}$$

exists and can be taken to be $x(t_0)$. This limit does exist when the order of the zero of $p(t)$ at $t = t_0$ is at least the order of the zero of $q(t)$ at $t = t_0$.

Signals that involve a rational exponent of the time variable, such as $x(t) = t^{1/2}$, are called *algebraic* signals. There are often problems with the domains of such signals; to the point, $t^{1/2}$ does not take values on the negative real numbers. Consequently, we must usually partition the domain of such signals and define the signal piecewise. One tool for this is the upcoming unit step signal $u(t)$.

1.2.2.2 Sinusoids. A more real-to-life example is a *sinusoidal* signal, such as $\sin(t)$ or $\cos(t)$. Of course, the mathematician's sinusoidal signals are synthetic, ideal creations. They undulate forever, whereas natural periodic motion eventually deteriorates. Both $\sin(t)$ and $\cos(t)$ are differentiable: $\frac{d}{dt}\sin(t) = \cos(t)$ and $\frac{d}{dt}\cos(t) = -\sin(t)$. From this it follows that both have derivatives of all orders and have Taylor[5] series expansions about the origin:

$$\sin(t) = t - \frac{t^3}{3!} + \frac{t^5}{5!} - \frac{t^7}{7!} + \cdots \tag{1.7a}$$

$$\cos(t) = 1 - \frac{t^2}{2!} + \frac{t^4}{4!} - \frac{t^6}{6!} + \cdots . \tag{1.7b}$$

[5]The idea is due to Brook Taylor (1685–1731), an English mathematician, who—together with many others of his day—sought to provide rigorous underpinnings for Newton's calculus.

So, while $\sin(t)$ and $\cos(t)$ are most intuitively described by the coordinates of a point on the unit circle, there are also formulas (1.7a)–(1.7b) that define them. In fact, the Taylor series, where it is valid for a function $x(t)$ on some interval $a < t < b$ of the real line, shows that a function is the limit of a sequence of polynomials: $x(a)$, $x(a) + x^{(1)}(a)(t-a)$, $x(a) + x^{(1)}(t-a) + x^{(2)}(t-a)^2/2!, \ldots$, where we denote the nth-order derivative of $x(t)$ by $x^{(n)}(t)$.

The observation that $\sin(t)$ and $\cos(t)$ have a Taylor series representation (1.7a)–(1.7b) inspires what will become one of our driving principles. The polynomial signals may not be very lifelike, when we consider that naturally occurring signals will tend to wiggle and then diminish. But sequences of polynomials, taken to a limit, converge to the sinusoidal signals. The nature of the elements is completely changed by the limiting process. This underscores the importance of convergent sequences of signals, and throughout our exposition we will always be alert to examine the possibility of taking signal limits. Limit processes constitute a very powerful means for defininining fundamentally new types of signals.

From their geometric definition on the unit circle, the sine and cosine signals are periodic; $\sin(t + 2\pi) = \sin(t)$ and $\cos(t + 2\pi) = \cos(t)$ for all $t \in \mathbb{R}$. We can use the trigonometric formulas for $\sin(s + t)$ and $\cos(s + t)$, the limit $\frac{\sin(t)}{t} \to 1$ as $t \to 0$, and the limit $\frac{\cos(t) - 1}{t} \to 0$ as $t \to 0$ to discover the derivatives and hence the Taylor series. Alternatively, we can define $\sin(t)$ and $\cos(t)$ by (1.7a)–(1.7b), whence we derive the addition formulas; define π as the unique point $1 < \pi/2 < 2$, where $\cos(t) = 0$; and, finally, show the periodicity of sine and cosine [44].

1.2.2.3 **Exponentials.**

Exponential signals are of the form

$$x(t) = Ce^{at}, \tag{1.8}$$

where C and a are constants, and e is the real number b for which the exponential $x(t) = b^t$ has derivative $\frac{d}{dt}x(t) = 1$ for $t = 0$. For $C = a = 1$, we often write $x(t) = \exp(t)$. The derivative of $\exp(t)$ is itself. This leads to the Taylor series expansion about $t = 0$:

$$e = \exp(t) = 1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \cdots = \sum_{k=0}^{\infty} \frac{t^k}{k!}. \tag{1.9}$$

Notice once more that a polynomial limit process creates a signal of a completely different genus. Instead of a periodic signal, the limit in (1.9) grows rapidly as $t \to \infty$ and decays rapidly as $t \to -\infty$.

If $C > 0$ and $a > 0$ in (1.9), then the graph of the exponential signal is an ever-increasing curve for $t > 0$ and an ever-decaying curve for $t < 0$. Since it has non-zero derivatives of arbitrarily high orders, such an exponential grows faster than any polynomial for positive time values. For $a < 0$, the graph of the exponential reflects across the y-axis (Figure 1.11).

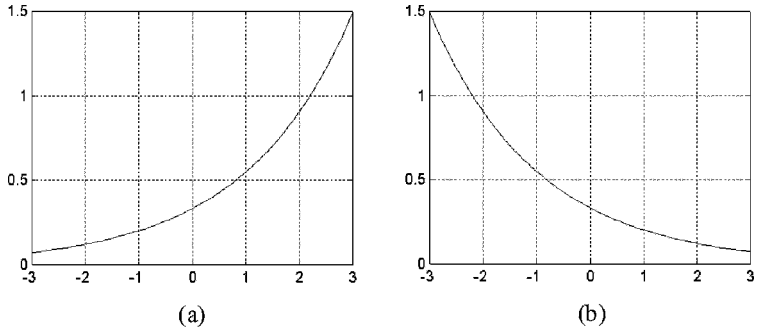


Fig. 1.11. Analog exponential signals. Panel (a) shows the exponential exp(t/2)/3, and (b) is its reflection across the y-axis, exp(–t/2)3.

A particularly valuable relation for signal theory is the *Laplace*[6] *identity*; we take the exponent in (1.9) to be purely imaginary:

$$e^{js} = \exp(js) = \cos(s) + j\sin(s), \tag{1.10}$$

where s is real. Why this is true can be seen from the unit circle in the complex plane (Figure 1.12) and by examining the expansion (1.10) of e^{js} in the series (1.9). First, substitute js for t in the expansion (1.9). Next, group the real and

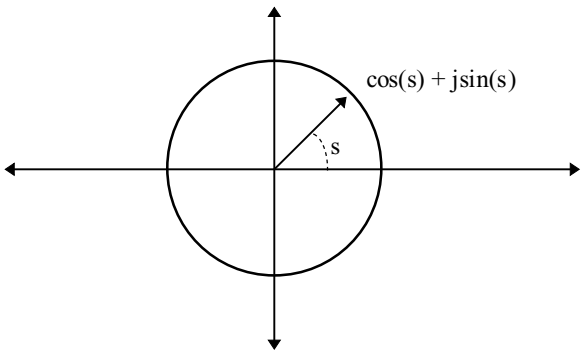


Fig. 1.12. Laplace's relation on the unit circle of the complex plane. By comparing Taylor series expansions, we find $e^{js} = \cos(s) + j\sin(s)$, and this corresponds to a point at arc distance s, counterclockwise on the unit circle from the positive x-axis.

[6]Pierre Simon Laplace (1749–1827), a French mathematician, physicist, and astronomer, theorized (along with German philosopher Immanuel Kant) that the solar system coalesced from a rotating gas cloud. The Laplace transform (Chapter 9) is named for him.

imaginary terms together. Observe that the sine and cosine Taylor series are in fact intermixed into the expansion of e^{js}, just as (1.10) expresses. The Laplace identity generalizes to any complex exponent $x + jy$: $e^{x+jy} = e^{x}[\cos(y) + j\sin(y)]$, where x and y are real. This is the most important formula in basic algebra.

The exponential signal is important in solving differential equations, such as arise from the study of heat transport and electromagnetism. For instance, the *heat diffusion equation* describes the propagation of heat $T(t, s)$ along a straight wire at time t and distance s from the end of the wire:

$$\frac{\partial T}{\partial t} = D\frac{\partial^2 T}{\partial s^2}, \tag{1.11}$$

where D is the *diffusion constant*. Solutions to (1.11) are $T(t, s) = e^{-\lambda t}e^{-jks}$, where λ and k are such that $D = \lambda/k^2$. The diffusion equation will make an unexpected appearance in Chapter 4 when we consider how to smooth a signal so that new regions of concavity do not appear as the smoothing progresses. Now, in electromagnetism, the electric and magnetic fields are vectors, $\boldsymbol{E}$ and $\boldsymbol{H}$, respectively, that depend upon one another. Maxwell's equations[7] for a vacuum describe this interaction in terms of space and time derivatives of the field vectors as follows:

$$\nabla \times \boldsymbol{E} = -\mu_0 \frac{\partial \boldsymbol{H}}{\partial t}, \tag{1.12a}$$

$$\nabla \times \boldsymbol{H} = \varepsilon_0 \frac{\partial \boldsymbol{E}}{\partial t}, \tag{1.12b}$$

$$\nabla \bullet \boldsymbol{E} = \nabla \bullet \boldsymbol{H} = 0. \tag{1.12c}$$

Equations (1.12a)–(1.12b) tell us that the curl of each field is proportional to the time derivative of the other field. The zero divergences in (1.12c) hold when there is no charge present. Constants μ_0 and ε_0 are the *magnetic permeability* and *electric permittivity* of space, respectively. By taking a second curl in (1.12a) and a second time derivative in (1.12b), separate equations in $\boldsymbol{E}$ and $\boldsymbol{H}$ result; for example, the electric field must satisfy

$$\nabla^2 \boldsymbol{E} = \mu_0 \varepsilon_0 \frac{\partial^2 \boldsymbol{E}}{\partial t^2}. \tag{1.13a}$$

For one spatial dimension, this becomes

$$\frac{\partial^2 \boldsymbol{E}}{\partial s^2} = \mu_0 \varepsilon_0 \frac{\partial^2 \boldsymbol{E}}{\partial t^2}. \tag{1.13b}$$

[7]Scottish physicist James Clerk Maxwell (1831–1879) is known best for the electromagnetic theory, but he also had a significant hand in the mechanical theory of heat.

Solutions to (1.13b) are sinusoids of the form $E(t, s) = A\cos(bs - \omega t)$, where $(b/\omega)^2 = \mu_0\varepsilon_0$, and b, ω, and A are constants.

Another signal of great importance in mathematics, statistics, engineering, and science is the *Gaussian.*[8]

Definition (Analog Gaussian). The *analog Gaussian* signal of mean μ and standard deviation σ is

$$g_{\mu,\sigma}(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}. \tag{1.14}$$

These terms are from statistics (Section 1.7). For now, however, let us note that the Gaussian $g_{\mu,\sigma}(t)$ can be integrated over the entire real line. Indeed, since (1.14) is always symmetric about the line $t = \mu$, we may take $\mu = 0$. The trick is to work out the square of the integral, relying on Fubini's theorem to turn the consequent iterated integral into a double integral:

$$\begin{aligned}\left(\int_{-\infty}^{\infty} g_{0,\sigma}(t)\,dt\right)^2 &= \left(\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{-\frac{t^2}{2\sigma^2}}dt\right)\left(\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{-\frac{s^2}{2\sigma^2}}ds\right)\\ &= \frac{1}{2\pi\sigma^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{s^2+t^2}{2\sigma^2}}dtds\end{aligned} \tag{1.15}$$

Changing to polar coordinates cracks the hard integral on the right-hand side of (1.15): $r^2 = t^2 + s^2$ and $dtds = rdrd\theta$. Hence,

$$\left(\int_{-\infty}^{\infty} g_{0,\sigma}(t)\,dt\right)^2 = \frac{1}{2\pi\sigma^2}\int_0^{2\pi}\int_0^{\infty} e^{-\frac{r^2}{2\sigma^2}} r\,drd\theta = \frac{1}{2\pi}\int_0^{2\pi}\left(-e^{-\frac{r^2}{2\sigma^2}}\Big|_0^{\infty}\right)d\theta \tag{1.16}$$

and we have

$$\left(\int_{-\infty}^{\infty} g_{0,\sigma}(t)\,dt\right)^2 = \frac{1}{2\pi}\int_0^{2\pi}(0-(-1))\,d\theta = \frac{1}{2\pi}(\theta)\Big|_0^{2\pi} = 1. \tag{1.17}$$

[8]Karl Friedrich Gauss (1777–1855) is a renowned German mathematician, physicist, and astronomer.

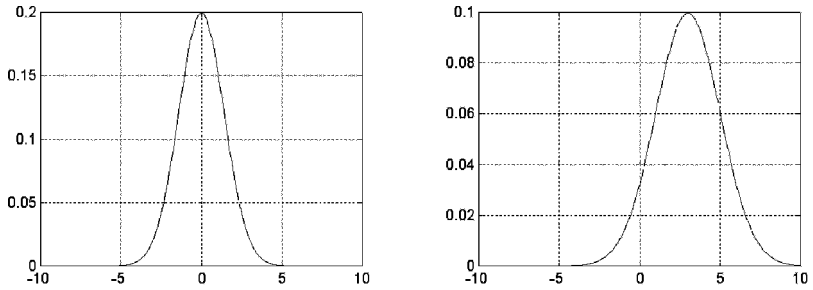


Fig. 1.13. Gaussian signals. The pulse on the left is $g_{0,2}(t)$, the Gaussian with mean μ equals 0, and the standard deviation σ equals 2. The pulse $g_{3,4}(t)$ is on the right. It has a wider spread than $g_{0,2}(t)$, and takes its smaller maximum value at $t = 3$.

Thus, there is unit area under the Gaussian curve. The Gaussian $g_{\mu,\,\sigma}(t)$ (1.14) is a bell-shaped curve (Figure 1.13), peaking at $t = \mu$, and symmetric about this line. The Gaussian decays forever as $|t| \to \infty$, but $g_{\mu,\,\sigma}(t) > 0$ for any $t \in \mathbb{R}$.

We define

$$g(t) = g_{0,\,1}(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}, \tag{1.18}$$

so that any Gaussian (1.14) is a scaled, shifted, and dilated version of $g(t)$:

$$g_{\mu,\,\sigma}(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma} g\left(\frac{t-\mu}{\sigma}\right). \tag{1.19}$$

The multiplying factor $(1/\sigma)$ governs the *scaling*, which may increase or decrease the height of the Gaussian. The same factor inside $g((t-\mu)/\sigma)$ *dilates* the Gaussian; it adjusts the spread of the bell curve according to the scale factor so as to preserve the unit area property. The peak of the bell shifts by the mean μ.

If we multiply a complex exponential $\exp(-j\omega t)$ by a Gaussian function, we get what is known as a *Gabor*[9] *elementary* function or signal [45].

[9]Dennis Gabor (1900–1979) analyzed these pulse-like signals in his 1946 study of optimal time and frequency signal representations. He is more famous outside the signal analysis discipline for having won the Nobel prize by inventing holography.

Definition (Gabor Elementary Functions). The *Gabor elementary function* $G_{\mu,\sigma,\omega}(t)$ is

$$G_{\mu,\,\sigma,\,\omega}(t) \;=\; g_{\mu,\,\sigma}(t)e^{j\omega t}. \tag{1.20}$$

Note that the real part of the Gabor elementary function $G_{\mu,\sigma,\omega}(t)$ in (1.20) is a cosine-like undulation in a Gaussian envelope. The imaginary part is a sine-like curve in a Gaussian envelope of the same shape (Figure 1.14). The time-frequency Gabor transform (Chapter 10) is based on Gabor elementary functions.

Interest in these signals surged in the mid-1980s when psychophysicists noticed that the modeled some aspects of the brain's visual processing. In particular, the receptive fields of adjacent neurons in the visual cortex seem to have profiles that resemble the real and imaginary parts of the Gabor elementary function. A controversy ensued, and researchers—electrical engineers, computer scientists, physiologists, and psychologists—armed with the techniques of mixed-domain signal decomposition continue to investigate and debate the mechanisms of animal visual perception [46, 47].

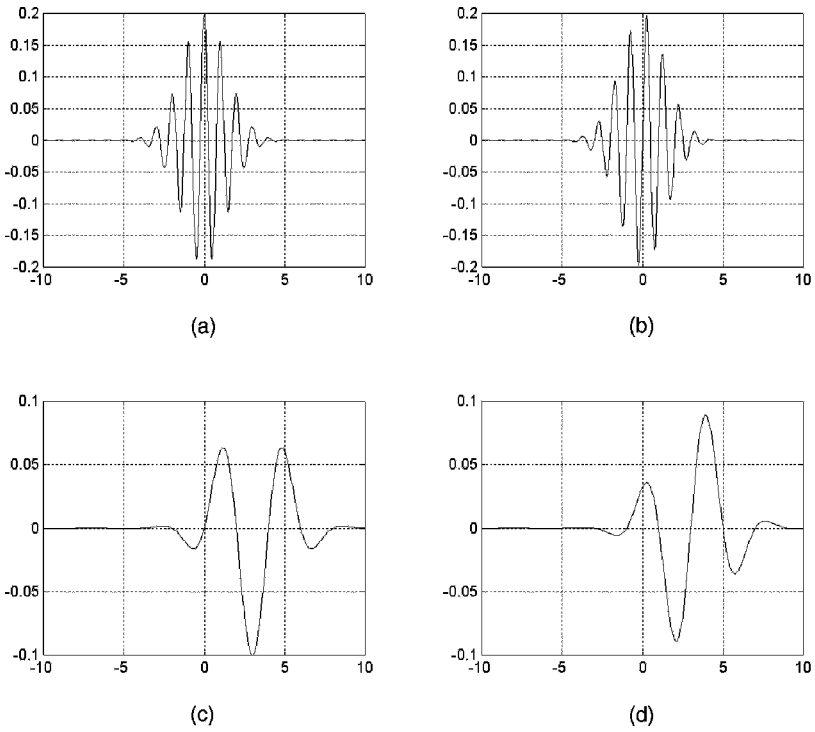


Fig. 1.14. Gabor elementary signals, real and imaginary parts. The pair on the top (a, b) are the real and imaginary parts of $g_{0,2}(t)\exp(j2\pi t)$. Below (c, d) is the Gabor pulse $G_{3,4,.5\pi}$. Note that if two Gabor elementary signals have the same sinusoidal frequency, but occupy Gaussian envelopes of different variances, then they have fundamentally different shapes.

1.2.3 Special Analog Signals

Several of the analog signal examples above are familiar from elementary algebra and calculus. Others, perhaps the Gabor elementary functions, are probably unfamiliar until one begins the formal study of signal processing and analysis. Some very simple analog signals play pivotal roles in the theoretical development.

1.2.3.1 Unit Step. We introduce the unit step and some closely related signals. The unit step signal (Figure 1.15) finds use in chopping up analog signals. It is also a building block for signals that consist of rectangular shapes and square pulses.

Definition (Unit Step). The *unit step* signal $u(t)$ is defined:

$$u(t) = \begin{cases} 1 & \text{if } t \geq 0, \\ 0 & \text{if } t < 0. \end{cases} \tag{1.21}$$

To chop up a signal using $u(t)$, we take the product $y(t) = x(t)u(t - c)$ for some $c \in \mathbb{R}$. The nonzero portion of $y(t)$ has some desired characteristic. Typically, this is how we zero-out the nonintegrable parts of signals such as $x(t) = t^{-2}$.

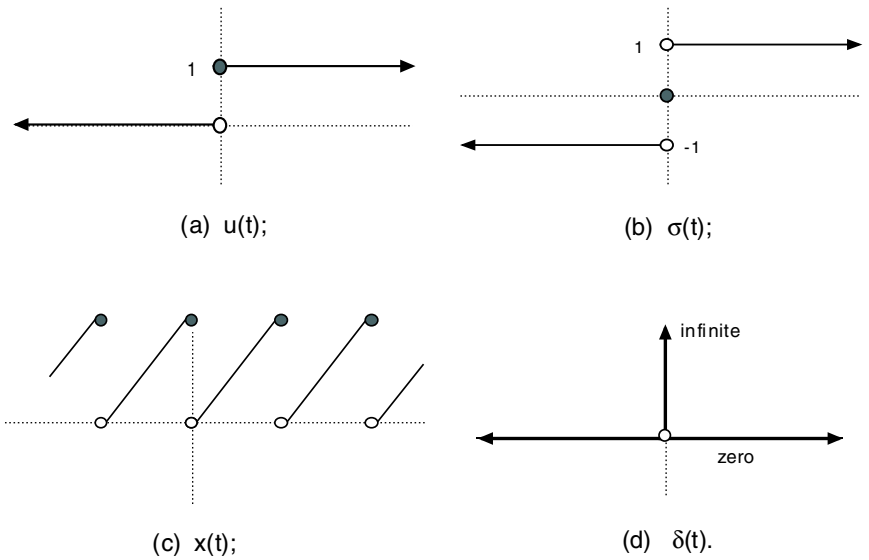


Fig. 1.15. Special Utility signals. (a) $u(t)$. (b) $\sigma(t)$. (c) $x(t)$. (d) $\delta(t)$. The unit step (a), signum (b), and sawtooth (c) are useful for constructing other signals and modeling their discontinuities. The Dirac delta (d) is "infinitely high" at $t = 0$ and zero otherwise; thus, it is not a bona fide analog signal. Chapters 3 and 5 provide the mathematical underpinnings of a valid, formal treatment of $\delta(t)$.

Definition (Signum). The *signum* signal $\sigma(t)$ is a cousin to the unit step:

$$\sigma(t) = \begin{cases} 1 & \text{if } t > 0, \\ 0 & \text{if } t = 0, \\ -1 & \text{if } t < 0. \end{cases} \tag{1.22}$$

Definition (Sawtooth). A sawtooth signal is a piecewise linear signal (Figure 1.15). For example, the infinite sawtooth $x(t)$ is

$$x(t) = \begin{cases} t & \text{if } t \geq 0, \\ 0 & \text{if } t < 0. \end{cases} \tag{1.23}$$

1.2.3.2 Dirac Delta. The *Dirac*[10] *delta* is really more of a fiction than a function. Nonetheless, it is a useful fiction. It can be made mathematically precise without losing its utility, and its informal development is familiar to many scientists and engineers.

For $n > 0$ let us define a sequence of analog signals $\delta_n(t)$:

$$\delta_n(t) = \begin{cases} \dfrac{n}{2} & \text{if } t \in \left[-\dfrac{1}{n}, \dfrac{1}{n}\right], \\ 0 & \text{if otherwise}. \end{cases} \tag{1.24}$$

The signals (1.24) are increasingly tall square spikes centered around the origin. Consider a general analog signal $x(t)$ and the integral over $\mathbb{R}$ of $x(t)\delta_n(t)$:

$$\int_{-\infty}^{\infty} x(t)\delta_n(t)\,dt = \int_{-1/n}^{1/n} x(t)\frac{n}{2}\,dt = \frac{1}{1/n - (-1/n)} \int_{-1/n}^{1/n} x(t)\,dt \tag{1.25}$$

The last term in (1.25) is the average value of $x(t)$ over $[-1/n, 1/n]$. As $n \to \infty$,

$$\lim_{n \to \infty} \int_{-\infty}^{\infty} x(t)\delta_n(t)\,dt = x(0). \tag{1.26}$$

The casual thought is to let $\delta(t)$ be the limit of the sequence $\{\delta_n(t): n > 0\}$ and conclude that the limit operation (1.26) can be moved inside the integral (Figure 1.16):

$$\lim_{n \to \infty} \int_{-\infty}^{\infty} x(t)\delta_n(t)\,dt = \int_{-\infty}^{\infty} x(t) \lim_{n \to \infty} \delta_n(t)\,dt = \int_{-\infty}^{\infty} x(t)\delta(t)\,dt = x(0). \tag{1.27}$$

[10]British physicist Paul Adrian Maurice Dirac (1902–1984) developed the theory of quantum electrodynamics. He received the Nobel prize in 1933.

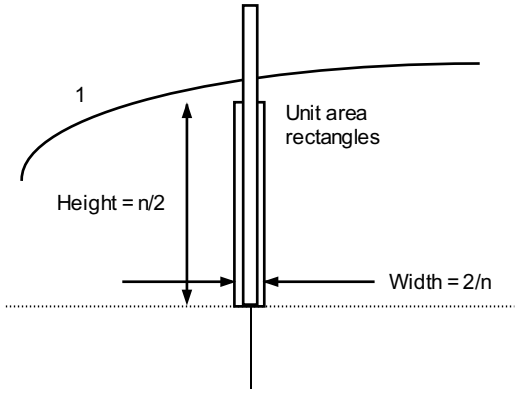


Fig. 1.16. Informal conception of the Dirac delta function. It is useful to think of $\delta(t)$ as the limit of a sequence of rectangles growing higher and narrower.

This idea is fundamentally mistaken, however. There is no pointwise limit of the sequence $\delta_n(t)$ at $t = 0$, and the limit of this signal sequence does not exist. The interchange of limit operations attempted in (1.27) is invalid. It is perhaps best to think of the final integral in (1.27) as an abbreviation for the valid limit operation in (1.26).

The Dirac delta can be shifted to any point t_0 in the domain of signal $x(t)$ and a similar argument applied. This gives the informal *sifting property* of the Dirac delta:

$$\int_{-\infty}^{\infty} x(t)\delta(t - t_0)\, dt \;=\; x(t_0)\,. \tag{1.28}$$

Again, mathematical prudence suggests that we think of the sifting property as special way of writing a limit of integrals. We can add another story to this mythology: The Dirac delta is the derivative of the unit step $u(t)$. Let $n > 0$ and consider the following sequence of continuous signals $u_n(t)$ approximating the unit step.

$$u_n(t) \;=\; \begin{cases} 1 & \text{if} \quad \dfrac{1}{n} < t, \\ \dfrac{(nt+1)}{2} & \text{if } -\dfrac{1}{n} \le t \le \dfrac{1}{n}, \\ 0 & \text{if} \quad t < -\dfrac{1}{n}. \end{cases} \tag{1.29}$$

Note that as $n \to \infty$, $u_n(t) \to u(t)$ for all $t \neq 0$. Also, for all $t \notin [-1/n, 1/n]$,

$$\frac{d}{dt}u_n(t) \;=\; \delta_n(t)\,. \tag{1.30}$$

We set aside our mathematical qualms and take limits as $n \to \infty$ of both sides of (1.30). The *derivative property* of the unit step results:

$$\lim_{n \to \infty} \frac{d}{dt} u_n(t) = \frac{d}{dt} \lim_{n \to \infty} u_n(t) = \frac{d}{dt} u(t) = \lim_{n \to \infty} \delta_n(t) = \delta(t). \tag{1.31}$$

The convergence of a sequence of functions must be *uniform* in order that interchange of limit operations, such as (1.27) and (1.30), be valid. Advanced calculus texts cover this theory [44]. The mathematical theory of *distributions* [48, 49] provides a rigorous foundation for the idea of a Dirac delta, as well as the sifting and derivative properties.

1.3 DISCRETE SIGNALS

Now that we have looked at some functions that serve as models for real-world analog signals, let us assume that we have a method for acquiring samples. Depending upon the nature of the analog signal, this may be easy or difficult. To get a discrete signal that represents the hourly air temperature, noting the reading on a thermometer is sufficient. Air temperature varies so slowly that hand recording of values works just fine. Rapidly changing analog signals, in contrast, require faster sampling methods. To acquire digital samples over one million times per second is not at all easy and demands sophisticated electronic design.

In signal processing, both analog and digital signals play critical roles. The signal acquisition process takes place at the system's front end. These are electronic components connected to some sort of transducer: a microphone, for instance. An analog value is stored momentarily while the digitization takes place. This sample-and-hold operation represents a discrete signal. An analog-to-digital converter turns the stored sample into a digital format for computer manipulation. We will, however, not ordinarily deal with digital signals, because the limitation on numerical precision that digital form implies makes the theoretical development too awkward. Thus, the discrete signal—actually an abstraction of the key properties of digital signals that are necessary for mathematical simplicity and flexibility—turns out to be the most convenient theoretical model for real-life digital signals.

1.3.1 Definitions and Notation

Unlike analog signals, which have a continuous domain, the set of real numbers $\mathbb{R}$, discrete signals take values on the set of integers $\mathbb{Z}$. Each integer n in the domain of x represents a time instant at which the signal has a value $x(n)$. Expressions such as $x(2/3)$ make no sense for discrete signals; the function is not even defined there.

Definition (Discrete and Digital Signals). A *discrete-time* (or simply *discrete*) signal is a real-valued function $x\colon \mathbb{Z} \to \mathbb{R}$. *x(n)* is the signal value at time instant n. A *digital* signal is an integer-valued function $x\colon \mathbb{Z} \to [-N, N]$, with domain $\mathbb{Z}$, $N \in \mathbb{Z}$, and $N > 0$. A complex-valued discrete-time signal is a function $x\colon \mathbb{Z} \to \mathbb{C}$, with domain $\mathbb{Z}$ and range included in the complex numbers $\mathbb{C}$.

Digital signals constitute a special class within the discrete signals. Because they can take on only a finite number of output values in the dependent variable, digital signals are rarely at the center of signal theory analyses. It is awkward to limit signal values to a finite set of integers, especially when arithmetic operations are performed on the signal values. Amplification is an example. What happens when the amplified value exceeds the maximum digital value? This is saturation, a very real problem for discrete signal processing systems. Some approach for avoiding saturation and some policy for handling it when it does occur must enter into the design considerations for engineered systems. To understand the theory of signals, however, it is far simpler to work with real-valued signals that may become arbitrarily small, arbitrarily large negative, and arbitrarily large positive. It is simply assumed that a real machine implementing signal operations would have a sufficiently high dynamic range within its arithmetic registers.

Notation. We use variable names such as "n", "m", and "k" for the independent variables of discrete signals. We prefer that analog signal independent variables have names such as "t" and "s". This is a tradition many readers will be comfortable with from Fortran computer programming. On those occasions when the discussion involves a sampling operation, and we want to use like names for the analog source and discrete result, we will subscript the continuous-domain signal: $x_a(t)$ is the analog source, and $x(n) = x_a(nT)$ is the discrete signal obtained from $x_a(t)$ by taking values every T time units.

With discrete-time signals we can tabulate or list signal values—for example, $x(n) = [3, 2, 1, \underline{5}, 1, 2, 3]$ (Figure 1.17). The square brackets signify that this is a discrete signal definition, rather than a set of integers. We must specify where the independent variable's zero time instant falls in the list. In this case, the value at $n = 0$ is

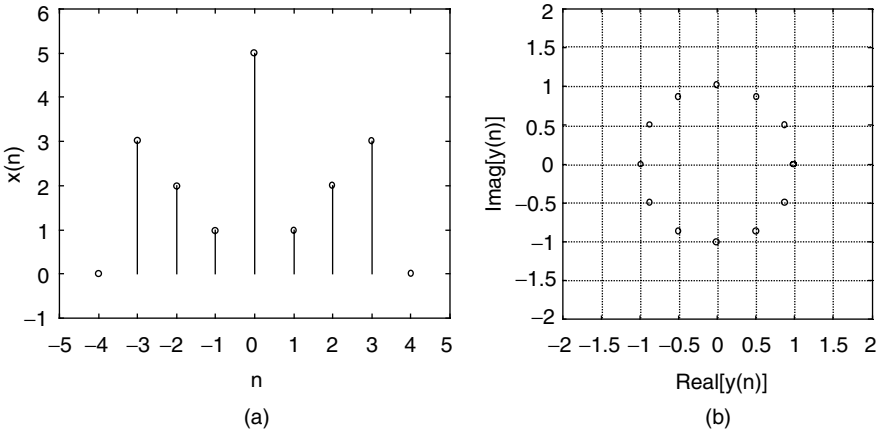


Fig. 1.17. Discrete signals. Panel (a) shows the signal $x(n) = [3, 2, 1, \underline{5}, 1, 2, 3]$. Signals may also be complex-valued, in which case their graphs are plotted in the complex plane. In (b), points of the signal $y(n) = \cos(n\pi/6) + j\sin(n\pi/6)$ are shown as pairs, (Real[$y(n)$], Imag[$y(n)$]).

underlined, and $x(0) = 5$, $x(-1) = 1$, $x(1) = 1$, and so on. For time instants not shown, signal values are zero. Thus, for discrete signals with an infinite number of nonzero values, we must provide a formula or rule that relates time instances to signal values, just as with analog signals.

1.3.2 Examples

We can derive straightforward discrete equivalents from the examples of analog signals above. A few curious properties relating to periodicity and derivatives arise.

1.3.2.1 Polynomials and Kindred Signals. There are discrete polynomial, rational, and algebraic signals. Discrete polynomials have the form

$$x(n) = \sum_{k=0}^{N} a_k n^k . \tag{1.32}$$

We cannot form the instantaneous derivative of $x(n)$ in (1.32) as with analog signals; instead discrete approximations must suffice. A variety of sometimes useful, but often problematic, notions of discrete derivatives do exist. For example, the left-hand discrete derivative of (1.32) is defined by $x_{\text{left}}(n) = x(n) - x(n-1)$. And a right-hand derivative exists too: $x_{\text{right}}(n) = x(n+1) - x(n)$. We can continue taking discrete derivatives of discrete derivatives. Note that there is no worry over the existence of a limit, since we are dividing the difference of successive signal values by the distance between them, and that can be no smaller than unity. Thus, discrete signals have (discrete) derivatives of all orders.

The domain of a polynomial $p(n)$ can be divided into disjoint regions of concavity: concave upward, where the second discrete derivative is positive; concave downward, where the second discrete derivative is negative; and regions of no concavity, where the second discrete derivative is zero, and $p(n)$ is therefore a set of dots on a line. Here is a first example, by the way, of how different analog signals can be from their discretely sampled versions. In the case of nonlinear analog polynomials, inflection points are always isolated. For discrete polynomials, though, there can be whole multiple point segments where the second derivative is zero.

If $p(n)$ and $q(n)$ are polynomials, then $x(n) = p(n)/q(n)$ is a *discrete rational* function. Signals modeled by discrete rational functions need to have provisions made in their definitions for the times n_0 when $q(n_0) = 0$. If, when this is the case, $p(n_0) = 0$ also, then it is necessary to separately specify the value of $x(n_0)$. There is no possibility of resorting to a limit procedure on $x(n)$ for a signal value, as with analog signals. Of course, if both $p(n)$ and $q(n)$ derive via sampling from analog ancestors, then such a limit, if it exists, could serve as the missing datum for $x(n_0)$.

Signals that involve a rational exponent of the time variable, such as $x(n) = n^{1/2}$, are called *discrete algebraic* signals. Again, there are problems with the domains of such signals; $n^{1/2}$ does not take values on the negative real numbers, for example. Consequently, we must usually partition the domain of such signals and define the signal piecewise.

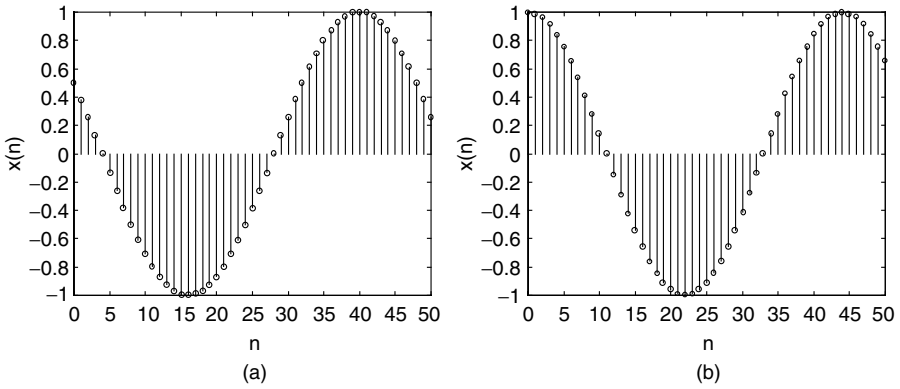


Fig. 1.18. Discrete sinusoids. Panel (a) shows the signal $x(n) = \cos(\omega n + \phi)$, with $\omega = \pi/24$ and $\phi = \pi/3$. Signal $x(n)$ has period $T = 48$. Panel (b,) on the other hand, shows the signal $y(n) = \cos(\omega n + \phi)$, with $\omega = 1/7$. It is not periodic.

1.3.2.2 Sinusoids. Discrete sinusoidal signals, such as $\sin(\omega n)$ or $\cos(\omega n)$, arise from sampling analog sinusoids (Figure 1.18). The function $\sin(\omega n + \phi)$ is the *discrete sine* function of *radial frequency* ω and phase ϕ. We will often work with $\cos(\omega n + \phi)$—and call it a sinusoid also—instead of the sine function. Note that—somewhat counter to intuition—discrete sinusoids may not be periodic! The periodicity depends upon the value of ω in $\cos(\omega n + \phi)$. We will study this nuance later, in Section 1.5.

1.3.2.3 Exponentials. Discrete exponential functions take the form

$$x(n) = Ce^{an} = C\exp(an), \tag{1.33}$$

where C and a are constants. Discrete exponentials (Figure 1.19) are used in frequency domain signal analysis (Chapters 7–9).

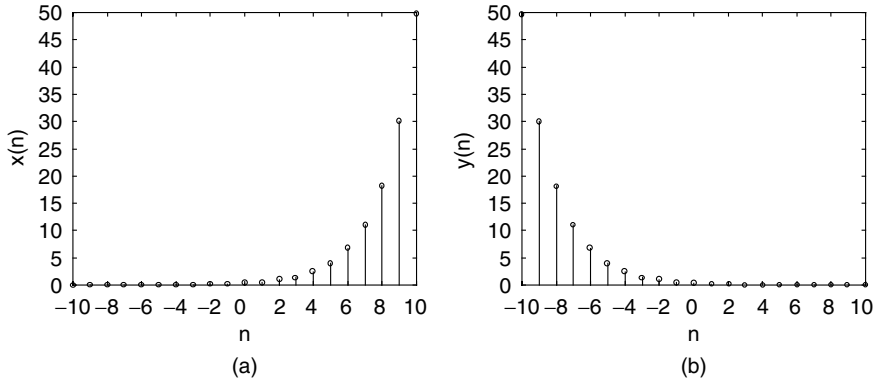


Fig. 1.19. Discrete exponential signals. Panel (a) shows the exponential $x(n) = \exp(n/2)/3$, and (b) is its reflection across the y-axis, $y(n) = \exp(-n/2)/3$.

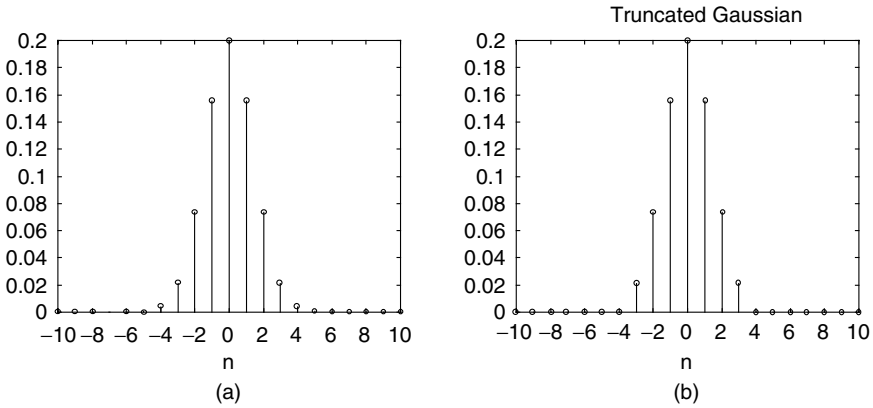


Fig. 1.20. Discrete Gaussian signals. The discrete pulse in (a) is $g_{0,2}(n)$, the Gaussian with mean μ equals 0 and standard deviation σ equals 2. Panel (b) illustrates a typical truncated Gaussian pulse.

The *discrete Gaussian* signal is

$$g_{\mu,\sigma}(n) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(n-\mu)^2}{2\sigma^2}}. \tag{1.34}$$

Truncating the discrete Gaussian so that it is zero outside of some interval $[-N, N]$ is a discrete signal processing commonplace (Figure 1.20). This assumes that $g_{\mu,\sigma}(n)$ is small for $|n| > N$. As a signal with a finite number of nonzero values, the truncated discrete Gaussian serves as a noise removal filter. We can use it instead of the moving average filter (Section 1.1), giving preference to local signal values, for example. Also, for noise removal it makes sense to normalize the nonzero values so that their sum is unity. This preserves the average value of the raw signal.

There are also discrete versions of the Gabor elementary functions:

$$G_{\mu,\sigma,\omega}(n) = g_{\mu,\sigma}(n) e^{j\omega n}. \tag{1.35}$$

1.3.3 Special Discrete Signals

Discrete delta and unit step present no theoretical difficulties.

Definition (Discrete Delta). The *discrete delta* or *impulse* signal $\delta(n)$ is

$$\delta(n) = \begin{cases} 1 & \text{if } n = 0, \\ 0 & \text{if } n \neq 0. \end{cases} \tag{1.36}$$

There is a *sifting property* for the discrete impulse:

$$\sum_{n=-\infty}^{\infty} x(n)\delta(n-k) = x(k)\,. \tag{1.37}$$

Discrete summation replaces the analog integral; this will become familiar.

Definition (Discrete Unit Step). The *unit step* signal $u(n)$ is

$$u(n) = \begin{cases} 1 & \text{if } n \geq 0, \\ 0 & \text{if } n < 0. \end{cases} \tag{1.38}$$

Note that if $m > k$, then $b(n) = u(n-k) - u(n-m)$ is a square pulse of unit height on $[k, m-1]$. Products $s(n)b(n)$ extract a chunk of the original discrete signal $s(n)$. And translated copies of $u(n)$ are handy for creating new signals on the positive or negative side of a signal: $y(n) = x(n)u(n-k)$.

1.4 SAMPLING AND INTERPOLATION

Sampling and interpolation take us back and forth between the analog and digital worlds. Sampling converts an analog signal into a digital signal. The procedure is straightforward: Take the values of the analog source at regular intervals. Interpolation converts a discrete signal into an analog signal. Its procedure is almost as easy: make some assumption about the signal between known values—linearity for instance—and fill them in accordingly. In the sampling process, much of the analog signal's information appears to be lost forever, because an infinite number of signal values are thrown away between successive sampling instants. On the other hand, interpolation appears to make some assumptions about what the discrete signal ought to look like between samples, when, in fact, the discrete signal says nothing about signal behavior between samples. Both operations would appear to be fundamentally flawed.

Nevertheless, we shall eventually find conditions upon analog signals that allow us to reconstruct them exactly from their samples. This was the discovery of Nyquist [2] and Shannon [3] (Chapter 7).

1.4.1 Introduction

This section explains the basic ideas of signal sampling: Sampling interval, sampling frequency, and quantization. The *sampling interval* is the time (or other spatial dimension measure) between samples. For a time signal, the *sampling frequency* is measured in hertz (Hz); it is the reciprocal of the sampling interval, measured in seconds (s). If the signal is a distance signal, on the other hand, with the sampling interval given in meters, then the sampling frequency is in units of (meters).$^{-1}$

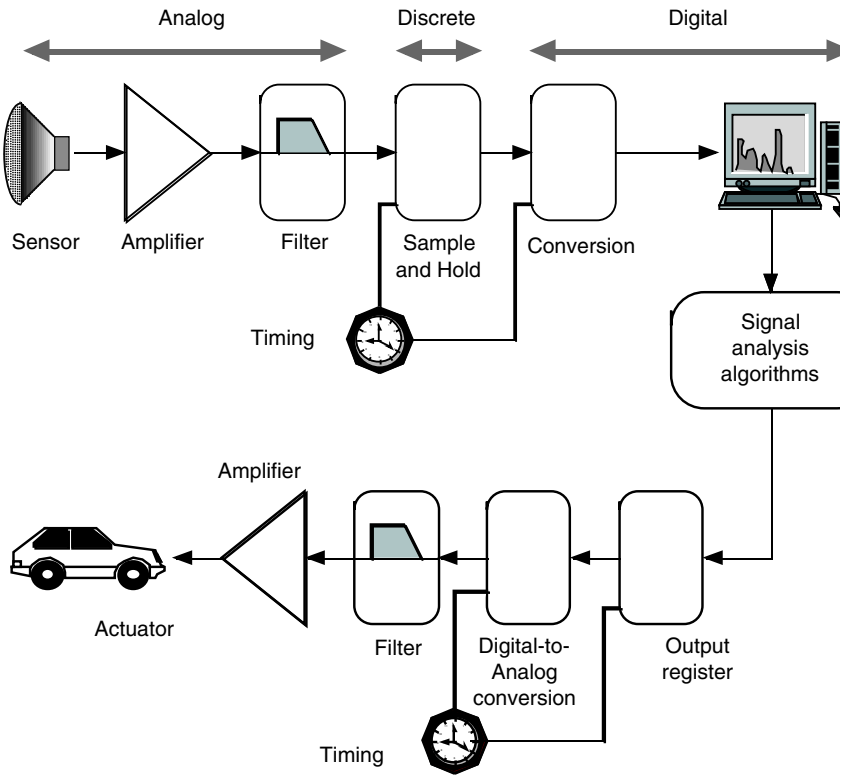


Fig. 1.21. Analog-to-digital conversion. Conversion of analog signals to digital signals requires several steps. Once digitized, algorithms running on the computer can analyze the signal. There may exist a closed loop with the analog world. For example, a digital output signal is converted back into analog form to control an actuator, such as anti-skid brakes on an automobile.

Discrete signals are more convenient for theoretical work, but for computer processing only a finite number of bits can represent the value in binary form. The signal must be digitized, or, in other words, the signal values must be *quantized*. By squeezing the signal value into an N-bit register, some fraction of the true signal value is lost, resulting in a *quantization error*. The number of possible digital signal values is called the *dynamic range* of the conversion.

If $x_a(t)$ is an analog signal, then $x(n) = x_a(n)$ defines a discrete signal. The time interval between $x(n)$ values is unity. We can also take more widely or narrowly spaced samples from $x_a(t)$: $x(n) = x_a(nT)$, where $T > 0$. In an actual system, electronic clock circuits set the sampling rate (Figure 1.21).

An N-bit register can hold non-negative digital values from 0 to $2^N - 1$. The smallest value is present when all bits are clear, and the largest value is when all bits are set. The two's complement representation of a digital value most common for storing signed digital signal values. Suppose there are N bits available in the input register, and the quantized signal's bit values are b_{N-1}, b_{N-2}, …, b_1, b_0. Then, the digital value is

$$D = -b_{N-1}2^{N-1} + b_{N-2}2^{N-2} + \cdots + b_2 2^2 + b_1 2^1 + b_0 2^0. \tag{1.39}$$

In this form, a register full of zeros represents a digital zero value; a single bit in the low-order position, $b_0 = 1$, represents unity; and a register having all bits set contains -1. The dynamic range of an N-bit register is 2^N.

There are several popular analog-to-digital converter (ADC) designs: Successive approximation, flash, dual-slope integration, and sigma–delta. The popular successive approximation converter operates like a balance beam scale. Starting at half of the digital maximum, it sets a bit, converts the tentative digital value to analog, and compares the analog equivalent to the analog input value. If the analog value is less than the converted digital guess, then the bit remains set; otherwise, the bit is cleared. The process continues with the next highest bit position in succession until all bits are tested against the input value. Thus, it adds and removes half-gram weights, quarter-gram weights, and so forth, until it balances the two pans on the beam. Successive approximation converters are accurate, slow, and common. A flash converter implements a whole bank of analog comparators. These devices are fast, nowadays operating at sampling rates of over 250 MHz. However, they have a restricted dynamic range. Dual-slope integration devices are slower, but offer better noise rejection. The sigma–delta converters represent a good design compromise. These units can digitize to over 20 bits and push sampling rates to almost 100 MHz.

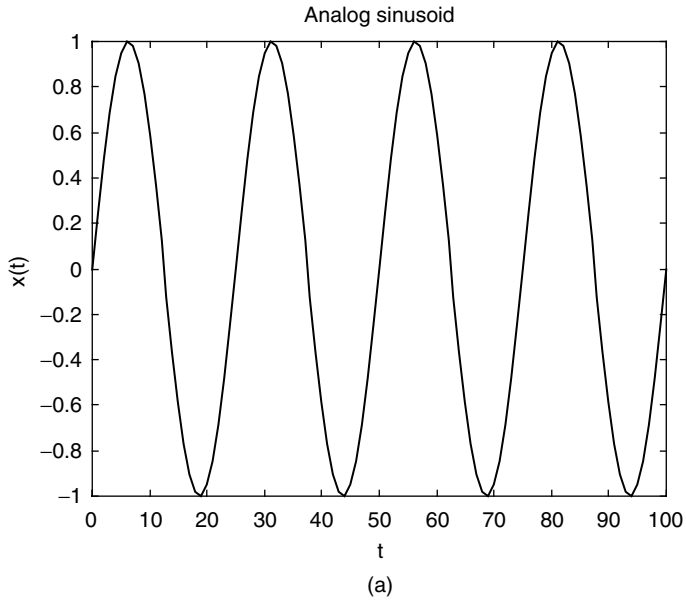
1.4.2 Sampling Sinusoidal Signals

Let us consider sampling a sinusoid, $x_a(t) = \cos(\omega t)$, as in Figure 1.22. We sample it at a variety of rates T: $x(n) = x_a(nT)$. For high sampling rates, the discrete result resembles the analog original. But as the sampling interval widens, the resemblance fades. Eventually, we cannot know whether the original analog signal, or, possibly, one of much lower frequency was the analog source for $x(n)$.

To answer the simple question—what conditions can we impose on an analog signal $x_a(t)$ in order to recover it from discrete samples $x(n)$?—requires that we develop both the analog and discrete Fourier transform theory (Chapters 5–7).

1.4.3 Interpolation

Why reconstruct an analog signal from discrete samples? Perhaps the discrete signal is the original form in which a measurement comes to us. This is the case with the geothermal signals we considered in Section 1.1. There, we were given temperature values taken at regular intervals of depth into the earth. It may be of interest—especially when the intervals between samples are very wide or irregular—to provide estimates of the missing, intermediate values. Also, some engineered systems use digital-to-analog converters to take a discrete signal back out into the analog world again. Digital communication and entertainment devices come to mind. So, there is an impetus to better understand and improve upon the analog conversion process.



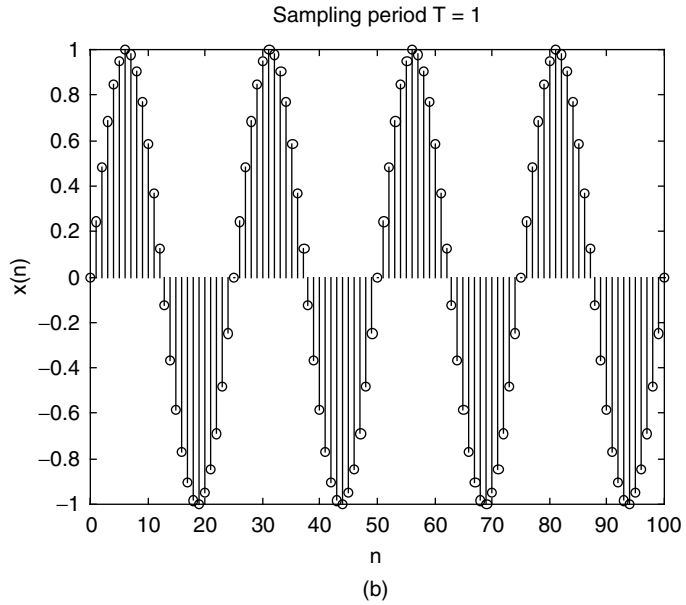
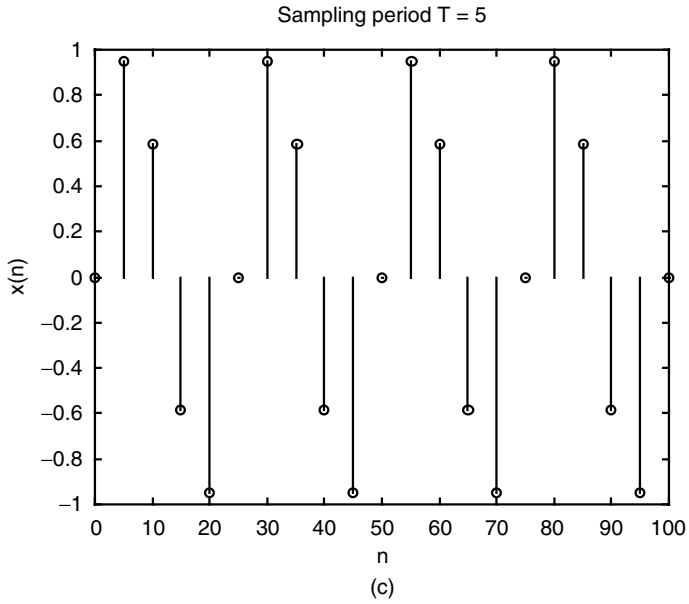


Fig. 1.22. Impossible to reconstruct. The original sinusoid (a) is first sampled at unit intervals (b). Sampling at a slower rate (c) suggests the same original $x(t)$ But when the rate falls, lower-frequency analog sinusoids could be the original signal (d).



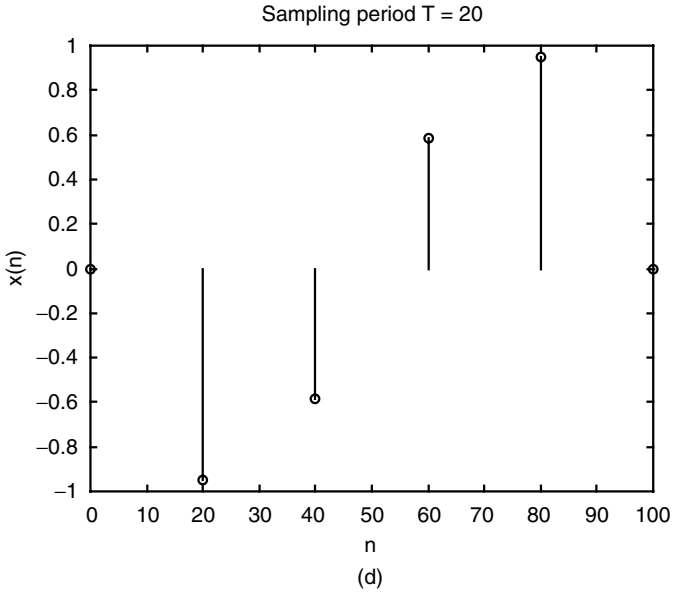


Fig. 1.22 (*Continued*)

1.4.3.1 Linear. Perhaps the simplest method of constructing an analog signal from discrete values is to use *linear interpolation* (Figure 1.23). Let $y_n = x(n)$, and define

$$x(t) = y_n + (y_{n+1} - y_n)(t - n) \tag{1.40}$$

for $t \in (n, n+1)$. The given samples are called *knots*, as if we were tying short sections of rope together. In this case, the analog signal passes through the knots. Observe that this scheme leaves corners—discontinuities in the first derivative—at the knots. The analog signal constructed from discrete samples via linear interpolation may therefore be unrealistic; nature's signals are usually smooth.

1.4.3.2 Polynomial. Smooth interpolations are possible with quadratic and higher-order polynomial interpolation.

Theorem (Lagrange[11] Interpolation). There is a unique polynomial $p(t)$ of degree $N > 0$ whose graph $(t, p(t))$ contains the distinct points $P_k = (n_k, x(n_k))$ for

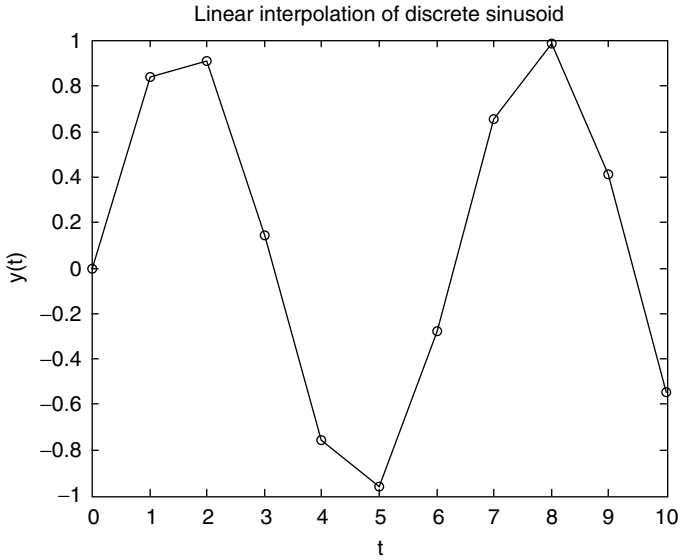


Fig. 1.23. Linear interpolation. Consider the signal $x(n) = \sin(n)$. Linear interpolation of this discrete sinusoid produces a jagged analog result.

[11]Joseph Louis Lagrange (1736–1813)—professor at Turin, Berlin, and Paris—was, with Euler, one of the great number theorists of the eighteenth century.

$0 \leq k \leq N$:

$$p(t) = \sum_{k=0}^{N} x(n_k) \prod_{\substack{m=0,\\ m \neq k}}^{N} \frac{(t - n_m)}{(n_k - n_m)} \tag{1.41}$$

Proof: Clearly $p(t)$ is of degree N and its graph passes through the points P_k. If $q(t)$ is another polynomial like this, then $d(t) = p(t) - q(t)$ is of degree N and has zeros at the $N + 1$ places: n_k, $0 \leq k \leq N$. Since a nonzero polynomial of degree N has at most N roots, the difference $d(t)$ must be identically zero. ■

Lagrange interpolation is not completely satisfactory. If N is large, then the method depends on a quotient of products of many terms; it is thereby subject to numerical round-off errors in digital computation. Furthermore, it is only valid for a restricted interval of points in a discrete signal. We may delete some points, since the interval between successive knots does not need to be unity in (1.41). However, we are still left with a polynomial approximation to a small part of the signal. If the signal is zero outside an interval, then however well the interpolant matches the nonzero signal region, $p(t)$ still grows large in magnitude with $|t|$.

An alternative is to compute quadratic polynomials on sequential triples of points. This provides some smoothness, but at every other knot, there is the possibility of a discontinuity in the derivative of the interpolants. Another problem is that the interpolation results vary depending on the point at which one starts choosing triples. To the point, how should we smooth the unit step signal $u(n)$ with interpolating quadratics? The interpolants are lines, except near the origin. If we begin with the triple (–2, –1, 0) and fit it with a quadratic polynomial, then our first interpolant will be concave up. If, on the other hand, we begin interpolating with the triad (–1, 0, 1), then our first quadratic approximation will be concave down. The exercises explore this and other questions entailed by quadratic interpolation. There is, however, a better way.

1.4.4 Cubic Splines

Perhaps the best method to make an analog signal out of discrete samples is to interpolate with spline[12] functions. The idea is to use a cubic polynomial between each successive pair of knots, $(n_k, x(n_k))$ and $(n_{k+1}, x(n_{k+1}))$ and yet match the first derivatives of cubics on either side of a knot.

To understand how this might work, let us first allow that the distance between known points need not be unity. Perhaps the system that collects the discrete samples cannot guarantee a regular sampling interval. The irregular sampling

[12]Architects and engineers once used a jointed, flexible ruler—known as a *spline*—to draw curves. These tools were made from sections of wood or plastic, attached together by brass rivets; modern splines, however, are almost always made of software.

notwithstanding, it is desirable to compose an analog signal that models the discrete data as a continuous, naturally occurring phenomenon. Later, it may even be useful to sample the analog model at regular intervals. So we assume that there are $N+1$ data points $(n_k, x(n_k))$, $0 \le k \le N$. Next, we set $\Delta_k = n_{k+1} - n_k$, $y_k = x(n_k)$, and we consider a polynomial $p_k(t)$ between the knots (n_k, y_k) and (n_{k+1}, y_{k+1}), for $0 \le k < N$. If the polynomial is quadratic or of higher degree and it contains the knots, then we need additional conditions to specify it. We prefer no sharp corners at the knots; thus, let us also stipulate that the derivatives of successive polyomials, $p_k'(t)$ and $p_{k+1}'(t)$, agree on their common knots: $p_k'(n_{k+1}) = p_{k+1}'(n_{k+1})$. Now there are four conditions on any interpolant: It must pass through two given knots and agree with its neighbors on endpoint derivatives. This suggests a cubic, since there are four unknowns. Readers might suspect that something more is necessary, because a condition on a polynomial's derivative is much less restrictive than requiring it to contain a given point.

Indeed, we need two further conditions on the second derivative in order to uniquely determine the interpolating polynomial. This reduces the search for a set of interpolating cubics to a set of linear equations. The two supplementary conditions are that we must have continuity of the second derivatives at knots, and we must specify second derivative values at the endpoints, (n_0, y_0) and (n_N, y_N). Then the equations are solvable [37, 50].

We write the interpolant on the interval $[n_k, n_{k+1}]$ in the form

$$p_k(t) = a_k(t-n_k)^3 + b_k(t-n_k)^2 + c_k(t-n_k) + y_k. \tag{1.42}$$

Then the derivative is

$$p_k'(t) = 3a_k(t-n_k)^2 + 2b_k(t-n_k) + c_k, \tag{1.43}$$

and the second derivative is

$$p_k''(t) = 6a_k(t-n_k) + 2b_k. \tag{1.44}$$

We define $D_k = p_k''(n_k)$ and $E_k = p_k''(n_{k+1})$. From (1.44),

$$D_k = 2b_k \tag{1.45}$$

and

$$E_k = 6a_k(n_{k+1} - n_k) + 2b_k = 6a_k\Delta_k + 2b_k, \tag{1.46}$$

with $\Delta_k = n_{k+1} - n_k$. Thus, we can express b_k and a_k in terms of Δ_k, which is known, and D_k and E_k, which are as yet unknown. Using (1.42), we can write c_k as follows:

$$c_k = \frac{y_{k+1} - y_k}{\Delta_k} - a_k\Delta_k^2 - b_k\Delta_k. \tag{1.47}$$

For $0 \le k < N$, (1.45)–(1.47) imply

$$c_k = \frac{y_{k+1} - y_k}{\Delta_k} - \frac{E_k - D_k}{6}\Delta_k - \frac{D_k}{2}\Delta_k = \frac{y_{k+1} - y_k}{\Delta_k} - \frac{\Delta_k}{6}(E_k + 2D_k). \quad (1.48)$$

To make a system of linear equations, we need to express this in terms of the second derivatives, D_k and E_k. The derivatives, $p_k'(t)$ and $p_{k+1}'(t)$, are equal for $t = n_{k+1}$; this is a required property of the interpolating cubics. Hence,

$$p_k'(n_{k+1}) = 3a_k(\Delta_k)^2 + 2b_k(\Delta_k) + c_k = p_{k+1}'(n_{k+1}) = c_{k+1}. \quad (1.49)$$

Inserting the expressions for c_{k+1}, c_k, b_k, and a_k in terms of second derivatives gives

$$6\frac{y_{k+2} - y_{k+1}}{\Delta_{k+1}} - 6\frac{y_{k+1} - y_k}{\Delta_k} = \Delta_k(2E_k + D_k) + \Delta_{k+1}(2D_{k+1} + E_{k+1}). \quad (1.50)$$

Now, $E_k = p_k''(n_{k+1}) = D_{k+1} = p_{k+1}''(n_{k+1})$, by invoking the continuity assumption on second derivatives. We also set $E_{N-1} = D_N$, producing a linear equation in D_k, D_{k+1}, and D_{k+2}:

$$6\frac{y_{k+2} - y_{k+1}}{\Delta_{k+1}} - 6\frac{y_{k+1} - y_k}{\Delta_k} = \Delta_k(2D_{k+1} + D_k) + \Delta_{k+1}(2D_{k+1} + D_{k+2}). \quad (1.51)$$

This system of equations has $N + 1$ variables, $D_0, D_1, \ldots, D_N$. Unfortunately, (1.51) has only $N - 1$ equations, for $k = 0, 1, \ldots, N - 2$. Let us lay them out as follows:

$$\begin{bmatrix} \Delta_0 & 2(\Delta_0 + \Delta_1) & \Delta_1 & 0 & 0 & \ldots & 0 \\ 0 & \Delta_1 & 2(\Delta_1 + \Delta_2) & \Delta_2 & 0 & \ldots & 0 \\ 0 & 0 & \Delta_2 & \ldots & \ldots & \ldots & 0 \\ 0 & 0 & 0 & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \Delta_{N-2} & 0 \\ 0 & 0 & 0 & 0 & \Delta_{N-2} & 2(\Delta_{N-2} + \Delta_{N-1}) & \Delta_{N-1} \end{bmatrix} \begin{bmatrix} D_0 \\ D_1 \\ D_2 \\ D_4 \\ \ldots \\ D_N \end{bmatrix}$$

$$= 6 \begin{bmatrix} \frac{y_2 - y_1}{\Delta_1} - \frac{y_1 - y_0}{\Delta_0} \\ \frac{y_3 - y_2}{\Delta_2} - \frac{y_2 - y_1}{\Delta_1} \\ \ldots \\ \ldots \\ \ldots \\ \frac{y_N - y_{N-1}}{\Delta_{N-1}} - \frac{y_{N-1} - y_{N-2}}{\Delta_{N-2}} \end{bmatrix}. \quad (1.52)$$

From linear algebra, the system (1.52) may have no solution or multiple solutions [51, 52]. It has a unique solution only if the number of variables equals the number of equations. Then there is a solution if and only if the rows of the coefficient matrix—and consequently its columns—are linearly independent. Thus, we must reduce the number of variables by a pair, and this is where the final condition on second derivatives applies.

We specify values for D_0 and D_N. The most common choice is to set $D_0 = D_N = 0$; this gives the so-called *natural spline* along the knots (n_0, y_0), (n_1, y_1), …, (n_N, y_N). The coefficient matrix of the linear system (1.52) loses its first and last columns, simplifying to the symmetric system (1.53). Other choices for D_0 and D_N exist and are often recommended [37, 50]. It remains to show that (1.53) always has a solution.

$$\begin{bmatrix} 2(\Delta_0+\Delta_1) & \Delta_1 & 0 & 0 & 0 & \dots & 0 \\ \Delta_1 & 2(\Delta_1+\Delta_2) & \Delta_2 & 0 & 0 & \dots & 0 \\ 0 & \Delta_2 & 2(\Delta_2+\Delta_3) & \Delta_3 & \dots & \dots & 0 \\ 0 & 0 & \Delta_3 & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \Delta_{N-2} \\ 0 & 0 & \dots & \dots & 0 & \Delta_{N-2} & 2(\Delta_{N-2}+\Delta_{N-1}) \end{bmatrix} \begin{bmatrix} D_1 \\ D_2 \\ D_3 \\ D_4 \\ \dots \\ D_{N-1} \end{bmatrix}$$

$$= 6 \begin{bmatrix} \frac{y_2 - y_1}{\Delta_1} - \frac{y_1 - y_0}{\Delta_0} \\ \frac{y_3 - y_2}{\Delta_2} - \frac{y_2 - y_1}{\Delta_1} \\ \dots \\ \dots \\ \dots \\ \frac{y_N - y_{N-1}}{\Delta_{N-1}} - \frac{y_{N-1} - y_{N-2}}{\Delta_{N-2}} \end{bmatrix}. \tag{1.53}$$

Theorem (Existence of Natural Splines). Suppose the points (n_0, y_0), (n_1, y_1), …, (n_N, y_N) are given and $n_0 < n_1 < \dots < n_N$. Let $\Delta_k = n_{k+1} - n_k$. Then the system $A\boldsymbol{v} = \boldsymbol{y}$ in (1.53) has a solution $\boldsymbol{v} = [D_1, D_2, \dots, D_{N-1}]^T$.

Proof: Gaussian elimination solves the system, using row operations to convert $\boldsymbol{A} = [A_{r,c}]$ into an upper-triangular matrix. The elements on the diagonal of the coefficient matrix are called the *pivots*. The first pivot is $P_1 = 2\Delta_0 + 2\Delta_1$, which is *positive*. We multiply the first row of $\boldsymbol{A}$ by the factor $f_1 = \frac{-\Delta_1}{2\Delta_0 + 2\Delta_1}$ and add it to

the second row, thereby annihilating $A_{2,1}$. A second pivot P_2 appears in place of $A_{2,2}$:

$$\begin{aligned} P_2 &= 2\Delta_1 + 2\Delta_2 + \Delta_1 f_1 = \Delta_1 \frac{-\Delta_1}{(2\Delta_0 + 2\Delta_1)} + 2\Delta_1 + 2\Delta_2 \\ &= \frac{3\Delta_1^2 + 4\Delta_0\Delta_1}{(2\Delta_0 + 2\Delta_1)} + 2\Delta_2. \end{aligned} \tag{1.54}$$

We update the vector $\boldsymbol{y}$ according to the row operation as well. Notice that $P_2 > 2\Delta_2 > 0$. The process produces another positive pivot. Indeed, the algorithm continues to produce positive pivots P_r. These are more than double the coefficient $A_{r+1,r}$, which the next row operation will annihilate. Thus, this process will eventually produce an upper-triangular matrix. We can find the solution to (1.53) by back substitution, beginning with D_{N-1} on the upper-triangular result. ■

Figure 1.24 shows how nicely cubic spline interpolation works on a discrete sinusoid. Besides their value for reconstructing analog signals from discrete samples, splines are important for building multiresolution signal decompositions that support modern wavelet theory [53] (Chapters 11 and 12).

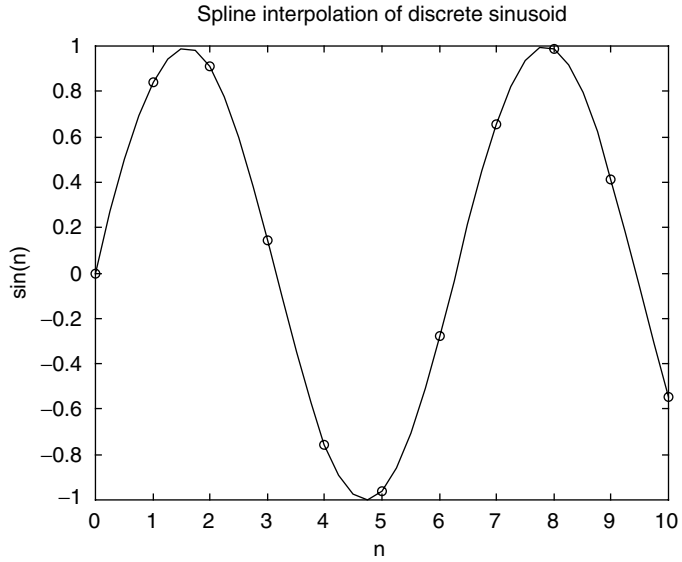


Fig. 1.24. Cubic spline interpolation. Again, discrete samples of the signal $x(n) = \sin(n)$ are used for the knots. Cubic spline interpolation offers a smooth model of the undulations and clearly captures the sinusoidal behavior of the original analog signal.

1.5 PERIODIC SIGNALS

Periodic signals, whether analog or discrete, repeat their values over intervals. The most familar ones are sinusoids. These signals arise from the mechanics of circular motion, in electric and magnetic interactions, and they are found in many natural phenomena. For instance, we considered the solution of Maxwell's equations, which describe the relation between the electric and magnetic fields. There we showed how to derive from the field equations a set of differential equations whose solution involves sinsusoidal functions. Radio waves propagate through empty space as electric and magnetic sinusoids at right angles to one another.

1.5.1 Fundamental Period and Frequency

The interval over which a signal repeats itself is its period, and the reciprocal of its period is its frequency. Of course, if a signal repeats itself over an interval, then it also repeats itself over any positive integral multiple of that interval; we must characterize a periodic signal by the smallest such interval of repetition.

Definition (Periodicity). An analog signal $x(t)$ is *periodic* if there is a $T > 0$ with $x(t + T) = x(t)$ for all t. A discrete signal $x(n)$ is periodic if there is an integer $N > 0$ with $x(n) = x(n + N)$ for all n. The smallest value for which a signal is periodic is called the *fundamental period*.

Definition (Analog Sinusoid). The signal

$$x(t) = A\cos(\Omega t + \phi) \tag{1.55}$$

is an *analog sinusoid*. A is the *amplitude* of $x(t)$, which gives its maximum value; Ω is its *frequency* in radians per second; and ϕ is its *phase* in radians. $\Omega = 2\pi F$, where F is the frequency in hertz.

Example. If $x(t) = A\cos(\Omega t + \phi)$, then $x(t) = x(t + \frac{2\pi}{\Omega})$, from the 2π-periodicity of the cosine function. So $T = \frac{2\pi}{\Omega} = \frac{1}{F}$ is the fundamental period of $x(t)$.

The sinusodal signals in nature are, to be sure, never the perfect sinusoid that our mathematical models suggest. Electromagnetic propagation through space comes close to the ideal, but always present are traces of matter, interference from other radiation sources, and the minute effects of gravity. Noise corrupts many of the phenomena that fall under our analytical eye, and often the phenomena are only vaguely sinusoidal. An example is the periodic trends of solar activity—in particular, the 11-year sunspot cycle, which we consider in more detail in the next section. But signal noise is only one aspect of nature's refusal to strictly obey our mathematical forumulas.

Natural sinusoidal signals decay. Thus, for it to be a faithful mathematical model of a naturally occurring signal, the amplitude of the sinusoid (1.55) should decrease. Its fidelity to nature's processes improves with a time-varying amplitude:

$$x(t) = A(t)\cos(\Omega t + \phi)\,. \tag{1.56}$$

The earliest radio telephony technique, *amplitude modulation* (AM), makes use of this idea. The AM radio wave has a constant *carrier frequency*, $F = \frac{\Omega}{2\pi}$ Hz, but its amplitude $A(t)$ is made to vary with the transmitted signal. Electronic circuits on the receiving end tune to the carrier frequency. The amplitude cannot jump up and down so quickly that it alters the carrier frequency, so AM is feasible only if F greatly exceeds the frequency of the superimposed amplitude modulation. This works for common AM content—voice and music—since their highest useful frequencies are about 8 and 25 kHz, respectively. In fact, limiting voice frequencies to only 4 kHz produces a very lifelike voice audio, suitable for telephony. Accordingly, the AM radio band, 550 kHz to 1600 kHz, is set well above these values. The signal looks like a sine wave whose envelope (the curve that follows local signal maxima) matches the transmitted speech or music.

Natural and engineered systems also vary the frequency value in (1.55). The basic *frequency-modulated* (FM) signal is the chirp, wherein the frequency increases linearly. Animals—birds, dolphins, and whales, for example—use frequency varying signals for communication. Other animals, such as bats, use chirps for echolocation. Some natural languages, such as Chinese, use changing tones as a critical indication of word meaning. In other languages, such as English and Russian, it plays only an ancillary role, helping to indicate whether a sentence is a question or a statement. Thus, we consider signals of the form

$$x(t) = A\cos(2\pi F(t) + \phi)\,, \tag{1.57}$$

where $F(t)$ need not be linear. An FM signal (1.57) is not a true sinusoid, but it provides the analyst with a different kind of signal model, suitable for situations where the frequency is not constant over the time region of interest. Applications that rely on FM signals include such systems as radars, sonars, seismic prospecting systems, and, of course, communication systems.

A *phase-modulated* signal is of the form

$$x(t) = A\cos(2\pi F + \phi(t))\,. \tag{1.58}$$

There is a close relation between phase and frequency modulation, namely, that the derivative of the phase function $\phi(t)$ in (1.58) is the *instantaneous frequency* of the signal $x(t)$ [54, 55]. The idea of instantaeous frequency is that there is a sinusoid that best resembles $x(t)$ at time t. It arose as recently as the late 1930s in the context of FM communication systems design, and its physical meaning has been the subject of some controversy [56]. If we fix F in (1.58) and allow $\phi(t)$ to vary, then the

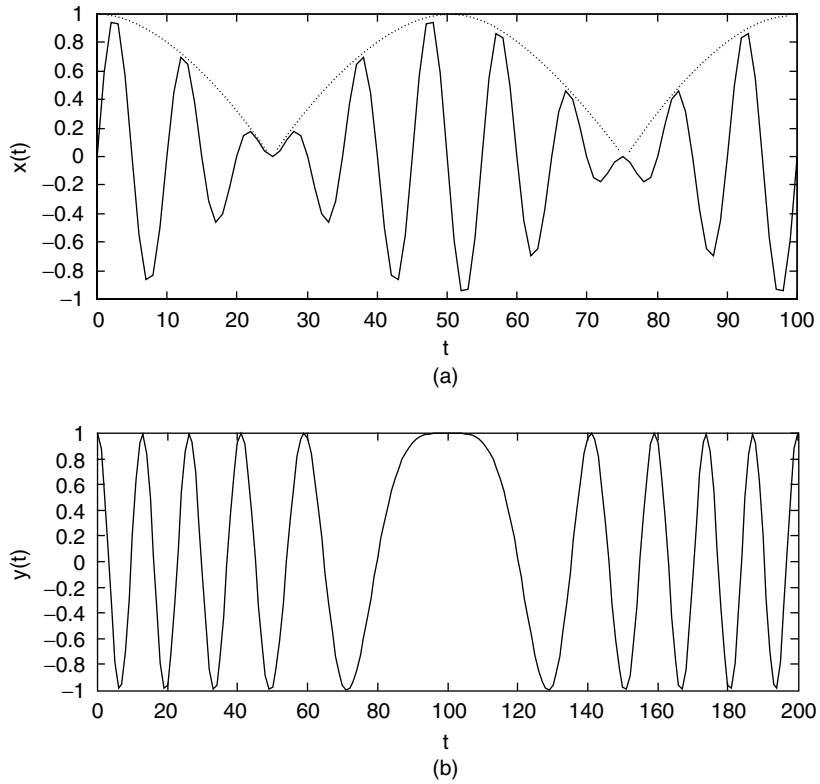


Fig. 1.25. AM and FM signals. Panel (a) shows a sinusoidal carrier modulated by a sinusoidal signal. The information-bearing part of the signal is given by the envelope of the signal, shown by dotted lines. In (b), a simple FM signal with the frequency varying sinusoidally is shown. Note that the oscillations bunch up and spread out as time passes, indicating rising and falling signal frequency, respectively.

frequency of the cosine wave changes. Over Δt seconds, the radial frequency of $x(t)$ changes by amount $[\Omega + \phi(t + \Delta t)] - [\Omega + \phi(t)] = \phi(t + \Delta t) - \phi(t)$, where $\Omega = 2\pi f$. The average change in Hertz frequency over this time interval is $\frac{\phi(t+\Delta t)-\phi(t)}{2\pi\Delta t}$. As $\Delta t \to 0$, this value becomes the derivative $\frac{d}{dt}\phi(t)$, the instantaneous frequency of the phase modulated signal $x(t)$.

If it seems odd that the derivative of phase is the signal frequency, then perhaps thinking about the familiar Doppler[13] effect can help reveal the connection. Suppose a train whistle makes a pure sinusoidal tone. If the train is standing still, then someone within earshot hears a sound of pure tone that varies neither in amplitude nor pitch. If the train moves while the whistle continues to blow, however, then the

[13]Austrian physicist Christian Doppler (1803–1853) discovered and described this phenomenon, first experimenting with trumpeters on a freight train.

tone changes. Coming toward us, the train whistle mechanically reproduces the same blast of air through an orifice, but the signal that we hear is different. The pitch increases as the train comes toward us. That means the signal frequency is increasing, but all that it takes to accomplish that is to move the train. In other words, a change in the phase of the whistle signal results in a different frequency in sound produced. A similar effect occurs in astronomical signals with the red shift of optical spectral lines from distant galaxies. In fact, the further they are away from us, the more their frequency is shifted. This means that the further they are from earth, the more rapidly is the phase changing. Objects further away move away faster. This led Hubble[14] to conclude that the universe is expanding, and the galaxies are spreading apart as do inked dots on an inflating balloon.

Some signals, natural and synthetic, are superpositions of sinusoids. In speech analysis, for example, it is often possible to model vowels as the sum of two sinusoidal components, called *formants*:

$$x(t) = x_1(t) + x_2(t) = A_1 \cos(\Omega_1 t + \phi_1) + A_2 \cos(\Omega_2 t + \phi_2) . \tag{1.59}$$

Generally, $x(t)$ in (1.59) is not sinusoidal, unless $\Omega_1 = \Omega_2 = \Omega$. A geometric argument demonstrates this. If the radial frequencies of the sinusoidal components are equal, then the vectors $\boldsymbol{v}_1 = (A_1 \cos(\Omega t + \phi_1), A_1 \sin(\Omega t + \phi_1))$ and $\boldsymbol{v}_2 = (A_2 \cos(\Omega t + \phi_2), A_2 \sin(\Omega t + \phi_2))$ rotate around the origin at equal speeds. This forms a parallelogram structure, rotating about the origin at the same speed as $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ (Figure 1.26), namely Ω radians per unit time. The x-coordinates of $\boldsymbol{v}_1$ and

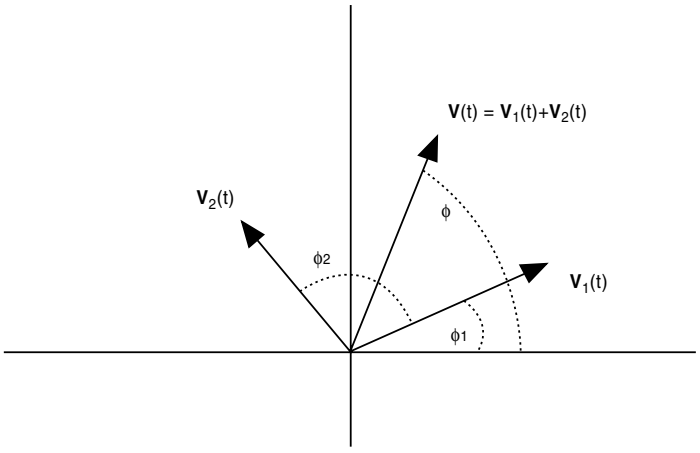


Fig. 1.26. Sinusoidal summation. Vector $\boldsymbol{v} = \boldsymbol{v}_1 + \boldsymbol{v}_2$ has length $\|\boldsymbol{v}\| = A$. Its x-coordinate, as a function of t, is a sinusoid of radial frequency Ω and phase $\phi = (\phi_2 - \phi_1)/2$.

[14]Working at the Mount Wilson Observatory near Los Angeles, Edwin Powell Hubble (1889–1953) discovered that galaxies are islands of stars in the vast sea of space, the red shift relates velocity to distance, and the universe itself expands in all directions.

$\boldsymbol{v}_2$ are the values of $x_1(t)$ and $x_2(t)$, respectively. The sum $x(t)$ is the x-coordinate of $\boldsymbol{v}_1 + \boldsymbol{v}_2$. Now, $\|\boldsymbol{v}_1 + \boldsymbol{v}_2\| = \|\boldsymbol{v}\| = A$, where

$$\begin{aligned} A^2 &= A_1^2 + A_2^2 + 2A_1A_2\cos(\Omega t + \phi_1)\cos(\Omega t + \phi_2) \\ &\quad + 2A_1A_2\sin(\Omega t + \phi_1)\cos(\Omega t + \phi_2) \\ &= A_1^2 + A_2^2 + 2A_1A_2\cos(\phi_1 - \phi_2). \end{aligned} \tag{1.60}$$

We also see from Figure 1.26 that the sum lies half way between $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. Thus, the phase of $\mathbf{v}$ is $\phi = \frac{\phi_2 - \phi_1}{2}$, and we have $x(t) = A\cos(\Omega t + \phi)$.

1.5.2 Discrete Signal Frequency

Due to the gap between successive signal values, discrete periodic signals have several properties that distinguish them from analog periodic waveforms:

(i) Discrete periodic signals have lower limits on their period; it makes no sense to have a discrete signal with period less than unity, because the discrete world does not even define signals at intervals smaller than unity.

(ii) A discrete signal with unit period is constant.

(iii) For sinusoids, the restriction to unit periods or more means that they have a maximum frequency: $|\Omega| = \pi$.

(iv) Not all sinusoids are periodic; periodicity only obtains when the frequency of the sampled signal is matched to the sampling interval.

This section covers these idiosyncrasies.

Proposition (Discrete Period). The smallest period for a discrete signal is $T = 1$. The largest frequency for a discrete sinusoid is $|\Omega| = \pi$, or equivalently, $|F| = 1$, where $\Omega = 2\pi F$ is the frequency in radians per sample.

Proof: Exercise. ■

Proposition (Periodicity of Discrete Sinusoids). Discrete sinusoid $x(n) = A\cos(\Omega n + \phi)$, $A \neq 0$, is periodic if and only if $\Omega = 2\pi p$, where $p \in \mathbb{Q}$, the rational numbers.

Proof: First, suppose that $\Omega = 2\pi p$, where $p \in \mathbb{Q}$. Let $p = m/k$ where $m, k \in \mathbb{N}$. If $m = 0$, then $x(n)$ is periodic; in fact, it is constant. Therefore, suppose $m \neq 0$ and choose $N = |k/m|$. Then,

$$\begin{aligned} x(n+N) &= A\cos\left(\Omega n + \Omega\left|\frac{k}{m}\right| + \phi\right) = A\cos\left(\Omega n + 2\pi\frac{m}{k}\left|\frac{k}{m}\right| + \phi\right) \\ &= A\cos(\Omega n + 2\pi(\pm 1) + \phi) = A\cos(\Omega n + \phi) = x(n) \end{aligned}$$

by the 2π-periodicity of the cosine function. Thus, $x(n)$ is periodic with period N. Conversely, suppose that for some $N > 0$, $x(n + N) = x(n)$ for all n. Then, $A\cos(\Omega n + \phi) = A\cos(\Omega n + \Omega N + \phi)$. Since $A \neq 0$, we must have

$\cos(\Omega n + \phi) = \cos((\Omega n + \phi) + \Omega N)$. And, since cosine can only assume the same values on intervals that are integral multiples of π, we must have $\Omega N = m\pi$ for some $m \in \mathbb{N}$. Then, $\Omega = m\pi/N$, so that Ω is a rational multiple of π. ■

Let us reinforce this idea. Suppose that $x(n) = x_a(Tn)$, where $x_a(t) = A\cos(\Omega t + \phi)$, with $A \neq 0$. Then $x_a(t)$ is an analog periodic signal. But $x(n)$ is not necessarily periodic. Indeed, $x(n) = A\cos(\Omega nT + \phi)$, so by the proposition, $x(n)$ is periodic only if ΩT is a rational multiple of π. Also, the discrete sinusoid $x(n) = \cos(2\pi fn)$ is periodic if and only if the frequency $f \in \mathbb{Q}$. The analog signal $s(t) = \cos(2\pi ft)$, $f > 0$, is always periodic with period $1/f$. But if $f = m/k$, with $m, k \in \mathbb{N}$, and m and k are relatively prime, then $x(n)$ has period k, not $1/f$. It takes time to get used to the odd habits of discrete sinusoids.

1.5.3 Frequency Domain

Having introduced analog and discrete sinusoids, fundamental period, and sinusoidal frequency, let us explain what it means to give a frequency-domain description of a signal. We already know that signals from nature and technology are not always pure sinusoids. Sometimes a process involves superpositions of sinusoids. The signal amplitude may vary too, and this behavior may be critical to system understanding. A variant of the pure sinusoid, the amplitude-modulated sine wave, models this situation. Another possibility is a variable frequency characteristic in the signal, and the frequency-modulated sine wave model accounts for it. There is also phase modulation, such as produced by a moving signal source. Finally, we must always be cognizant of, prepare for, and accommodate noise within the signal. How can we apply ordinary sinusoids to the study of these diverse signal processing and analysis applications?

1.5.3.1 Signal Decomposition. Many natural and synthetic signals contain regular oscillatory components. The purpose of a frequency-domain description of a signal is to identify these components. The most familar tool for aiding in identifying periodicities in signals are the sinusoidal signals, $\sin(t)$ and $\cos(t)$. Thus, a frequency-domain description presupposes that the signal to be analyzed consists of a sum of a few sinusoidal components. Perhaps the sum of sinusoids does not exactly capture the signal values, but what is left over may be deemed noise or background. We consider two examples: sunspot counts and speech. If we can identify some simple sinusoidal components, then a frequency-domain description offers a much simpler signal description. For instead of needing a great number of time domain values to define the signal, we need only a few triplets of real numbers—the amplitude, frequency, and phase of each substantive component—in order to capture the essence of the signal.

Consider the signal $x(n)$ of Figure 1.27, which consists of a series of irregular pulses. There appears to be no rule or regularity of the values that would allow us to describe it by more that a listing of its time-domain values. We note that certain regions of the signal appear to be purely sinusoidal, but the juxtaposition of unfamiliar, oddly shaped pulses upsets this local pattern.

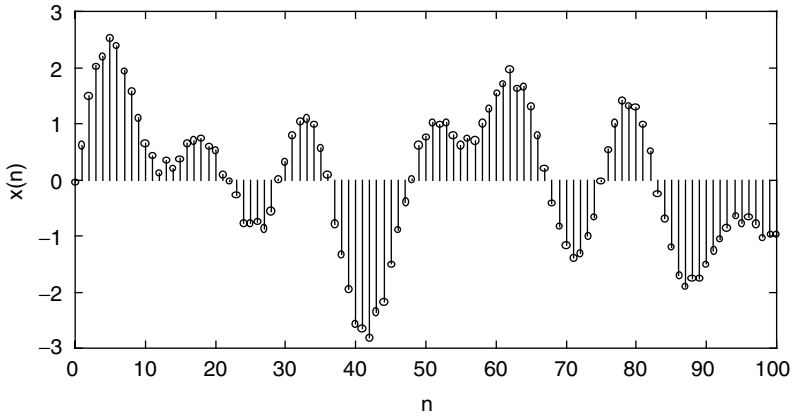


Fig. 1.27. Efficiency of freqency-domain signal description. An irregular signal, appearing to lack any regular description beyond a listing of its time-domain values, turns out to be the sum of three sinusoids and a small background noise signal.

The signal does, however, have a concise description as a sum of three sinusoids and a small noise component:

$$x(n) = \sin\left(\frac{2\pi n}{15}\right) + \sin\left(\frac{2\pi n}{25}\right) + \sin\left(\frac{2\pi n}{50}\right) + N(n)\,. \tag{1.61}$$

Thus, signal description, and hence the further analysis of signals, is often made more powerful by representing signals in terms of an established set of prototype signals. For a frequency domain signal description, sinusoidal models are appropriate. Some background noise may elude an immediate description in terms of sinusoidal components. Depending on the application, of course, this residual signal may be negligible because it has a much lower magnitude than the source signal of interest. If the sinusoidal trends are localized within a signal, the arithmetic of superposition may allow us to describe them even with sinusoids of longer duration, as Figure 1.27 illustrates.

1.5.3.2 Sunspots. One periodic phenomenon with which we are familiar is the sunspot cycle, an example of a natural periodic trend. Sunspots appear regularly on the sun's surface as clusters of dark spots and, at their highest intensities, disrupt high-frequency radio communication. Scientists have scrutinized sunspot activity, because, at its height during the 11-year cycle, it occasionally hampers very high frequency radio communication. Until recently, the mechanism responsible for this phenomenon was not known. With temperature of 3800 °K, they are considerably cooler than the rest of the sun's surface, which has an average temperature of some 5400 °K.

Ancient people sometimes observed dark spots on the solar disk when it was obscured by fog, mist, or smoke. Now we check for them with a simple telescope

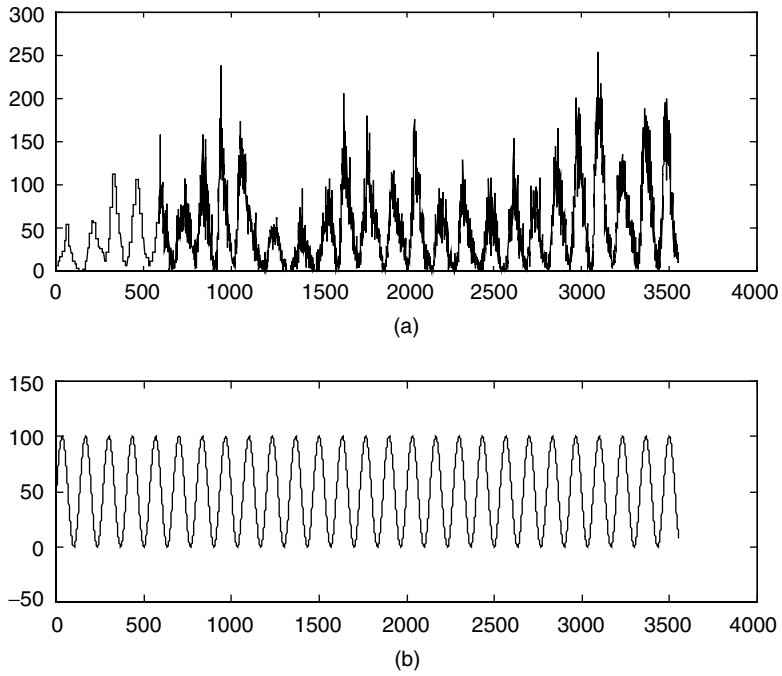


Fig. 1.28. Wolf sunspot numbers. Panel (a) plots the time-domain values of $w(n) = 10G(n) + S(n)$ for each month from 1700 through 1995. We compare the oscillation with sinusoids, for example, when period $T = 11.1$ years as in (b).

that projects the sun's image onto a white plate. Galileo[15] was the first to do so. His observations upset the prevailing dogma of seventeenth century Europe insisting that the sun was a perfect disk. Standardized sunspot reports began in the mid-1700s, and the earlier values given in our data plots (Figure 1.28) are assumptions based on informal observations.

Sunspot activity can be formulated as a discrete signal by counting the number of groups of sunspots. In 1848, the Swiss astronomer, Johann Rudolph Wolf, introduced a daily measurement of sunspot number. His method, which is still used today, counts the total number of spots visible on the face of the sun and the number of groups into which they cluster, because neither quantity alone satisfactorily measures sunspot activity. The *Wolf*[16] *sunspot number* is $w(n) = 10G(n) + S(n)$, where $G(n)$ is the average number of sunspot groups and $S(n)$ is the average number of spots. Individual observational results do vary greatly, however, since the measurement strongly depends on interpretation, experience, and the stability of the earth's atmosphere

[15]In addition to finding solar blemishes in 1610, Galileo Galilei (1564–1642) used his telescope to resolve the Milky Way into faint stars and, with his discovery of the phases of Venus, confirmed Copernicus's heliocentric theory.

[16]After Swiss astronomer Johann Rudolph Wolf (1816–1893).

above the observing site. The use of the earth as a platform from which to record these numbers contributes to their variability, too, because the sun rotates and the evolving spot groups are distributed unevenly across solar longitudes. To compensate for these limitations, each daily international number is computed as a weighted average of measurements made from a network of cooperating observatories.

One way to elucidate a frequency-domain description of the sunspot signal $w(n)$ is to compare it with a sinusoidal signal. For example, we can align sinusoids of varying frequency with $w(n)$, as shown in Figure 1.28b. Thus, the sinusoids are models of an ideal sunspot cycle. This ideal does not match reality perfectly, of course, but by pursuing the mathematical comparison between the trignonometric model and the raw data, we can get a primitive frequency-domain description of the sunspot cycle. What we want to derive in a frequency-domain description of a signal is some kind of quantification of how much a signal resembles model sinusoids of various frequencies. In other words, we seek the relative weighting of supposed frequency components within the signal. Thinking of the signal values as a very long vector, we can compute the inner product of $w(n)$ with the unit vectors whose values are given by sinusoids

$$s(n) = 50 + 50\sin\left(\frac{2\pi n}{12T}\right), \tag{1.62}$$

where T varies from 0.1 year to 16 years. Then we compute the difference $e_T(n) = w(n) - s(n)$, an error term which varies with periodicity of the sinusoid (1.62) determined by T. Now, we take evalutate the norm of the vector $e_T(n)$, which has length $12 \times 296 = 3552 : \|e_T\|$. If we plot the norm of the error vectors with respect to the supposed period T of the sinusoidal model, we see that there is a pronounced minimum near $T = 11$ (Figure 1.29).

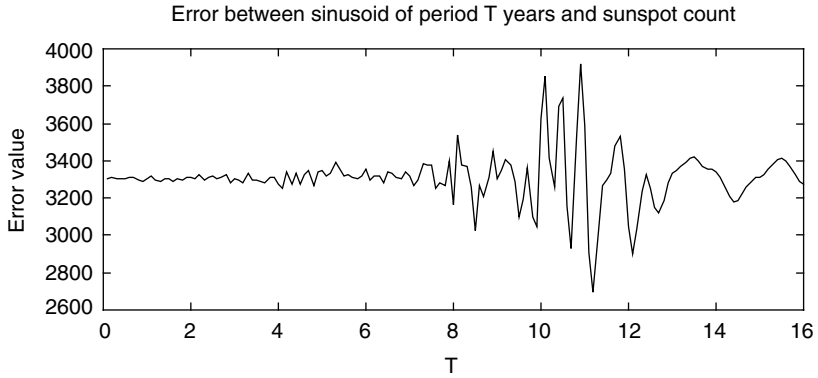


Fig. 1.29. Comparing sinusoids with sunspot numbers. The minimal error between such sinusoids and the sunspot oscillations occurs just above $T = 11$. Note that there are other local minima both below and above $T = 11$. In point of fact, sometimes the sunspot cycle peaks a few months early and sometimes a few months later.

Note that we have exploited a few specific facts about the sunspot numbers in this analysis. In particular, we have not worried about the relative position of the sinusoidal model. That is, we have not taken the relative phases of the sinusoid $s(n)$ and the discrete sunspot signal $w(n)$ into account. Could a slightly shifted sinusoid result in a smaller error term? This is indeed quite a likely possibility, and we avoid it for two reasons. First, it turns the minimization of the error term into a two-dimensional problem. This is a practical, application-oriented discussion. Although we want to explain the technical issues that arise in signal analysis problems, we do not want to stray into two-dimensional problems. Our focus is one-dimensional—signal analysis, not image analysis—problems. In fact, with more powerful tools, such as the discrete Fourier transform (Chapter 7), we we can achieve a frequency domain analysis in one dimension that handles the relative phase problem.

1.5.3.3 Speech. Scientists and engineers have long attempted to build commercial speech recognition products. Such products now exist, but their applicability remains limited. Computers are so fast, mathematics so rich, and the investigations so deep: How can there be a failure to achieve? The answers seem to lie in the fundamental differences between how signal analyzing computers and human minds—or animal minds in general, for that matter—process the acoustical signals they acquire. The biological systems process data in larger chunks with a greater application of top-down, goal-directed information than is presently possible with present signal analysis and artificial intelligence techniques.

An interesting contrast to speech recognition is speech generation. Speech synthesis is in some sense the opposite of recognition, since it begins with a structural description of a signal—an ASCII text string, for instance—and generates speech sounds therefrom. Speech synthesis technology has come very far, and now at the turn of the century it is found in all kinds of commercial systems: telephones, home appliances, toys, personal computer interfaces, and automobiles. This illustrates the fundamental asymmetry between signal synthesis and analysis. Speech recognition systems have become increasingly sophisticated, some capable of handling large vocabularies [57–59]. In recent years, some of the recognition systems have begun to rely on artificial neural networks, which mimic the processing capabilities of biological signal and image understanding systems [60].

To begin to understand this fundamental difference and some of the daunting problems faced by speech recognition researchers, let us consider an example of digitized voice (Figure 1.30). Linguists typically classify speech events according to whether the vocal cords vibrate during the pronunciation of a speech sound, called a *phone* [61]. Phones are speech fragments. They are realizations of the basic, abstract components of a natural language, called *phonemes*. Not all natural languages have the same phonemes. Speakers of one language sometimes have extreme difficulties hearing and saying phonemes of a foreign tongue. And within a given language, some phonemes are more prevalent than others. The most common strategy for speech recognition technology is to break apart a digital speech sample into separate phones and then identify phonemes among

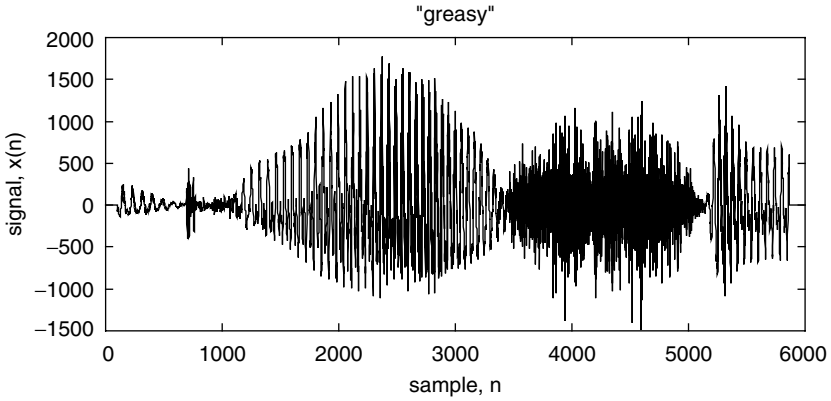


Fig. 1.30. A digitized voice signal, the word "greasy". The sampling rate is 16 kHz. It is difficult, even to the trained eye, to recognize the spoken content of a signal from its time-domain signal values.

them [59, 62]. The speech fragment of Figure 1.30 contains five phonemes. Sometimes the preliminary sequence of phonemes makes no sense; a sophisticated algorithm may merge or further segment some of the phones for a better result. A higher-level process uses the phoneme stream to extract whole words. Perhaps some posited word does not fit the application's context. Still higher-level algorithms—and at this point the application has removed itself from signal processing and analysis proper to the realm of artificial intelligence—may substitute one phoneme for another to improve the interpretation. This is called *contextual analysis*. Although we will consider speech analysis in more detail in later chapters, topics such as contextual analysis blend into artificial intelligence, and are outside the scope of our presentation. It is nevertheless interesting to note that computers are generally better than humans at recognizing individual phonemes, while humans are far superior when recognizing complete words [63], a hint of the power of contextual methods.

Linguists separate phonemes into two categories: *voiced* and *unvoiced*, according to whether the vocal cords vibrate or not, respectively. Vowels are voiced, and it turns out that a frequency-domain description helps to detect the presence of a vowel sound. Vowel sounds typically contain two sinusoidal components, and one important early step in speech processing is to determine the frequency components of the signal. We can see this in a digital speech sample of a vowel (Figure 1.31). There are clearly two trends of oscillatory behavior in the signal.

Thus, depending upon the relative strength of the two components and upon their actual frequency, a frequency domain description can identify this phone as the /u/ phoneme. There are complicating factors of course. The frequency components change with the gender of the speaker. And noise may corrupt the analysis. Nevertheless, a frequency-domain description is an important beginning point in phoneme recognition.

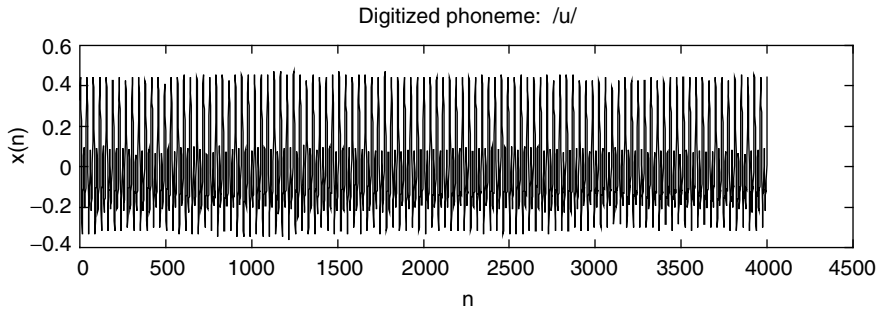


Fig. 1.31. A vowel phoneme, the sound "*u*", sampled at 8 kHz. Note the two sinusoids in the time-domain trace. The lower frequency component has a higher amplitude. Principal frequency components are at approximately 424 Hz and at 212 Hz.

1.5.4 Time and Frequency Combined

Until we develop the formal theory for the frequency analysis of signals—the Fourier transform, in particular—we continue to illustrate signals in terms of their time-domain values. We can nonetheless identify the oscillatory components by inspection. Parts of a given signal resemble a sinusoid of a given frequency. Sometimes the oscillations continue throughout the time span of the signal. For such signals, the Fourier transform (Chapters 5 and 6, and introduced from a practical standpoint in Chapter 4) is the appropriate tool.

Sometimes, however, the oscillatory components die out. This makes the game interesting, because our ordinary sinusoidal models continue oscillating forever. We can arbitrarily limit the time domain extent for our sinusoidal models, and this happens to be effective for applications such as speech analysis. Special mathematical tools exist that can decompose a signal into a form that exposes its frequency components within distinct time intervals. Among these tools is the Gabor transform, one of several time-frequency transforms that we explore in Chapter 10. Until the theoretical groundwork is laid for understanding these transforms, however, we must content ourselves with intuitive methods for describing the frequency content of signals.

Let us also remark that frequency-domain methods have been extensively explored in electrocardiography and electroencephalography—applications we considered at the beginning of the chapter. This seems a natural approach for the ECG, since the heartbeat is a regular pulse. Perhaps surprisingly, for heartbeat irregularities, frequency-domain techniques have been found to be problematic. In EEG work, the detection of certain transients and their development into regular waves is important for diagnosing epilepsy. And here again frequency-domain tools—even those that employ an explicit mixed time-frequency analysis strategy—do not address all of the difficulties [35, 64–66]. Like problems arise in seismic signal interpretation [67, 68]. In fact, problems in working with time-frequency analysis

methods for seismic soundings analysis in petroleum prospecting led to the discovery of the scale-based wavelet transform in the mid-1980s [69].

Researchers have thus begun to investigate methods that employ signal shape and scale as a tool, rather than frequency. Unlike the sinusoidal components that are the basis for a frequency-domain signal description, the components for a scale-based description are limited in their time-domain extent. The next section considers several special signal classes, among them several types which are time-limited. Of these, the finite-energy signals are particularly attractive, as later chapters demonstrate, for signal descriptions based on scale.

1.6 SPECIAL SIGNAL CLASSES

This section covers some special signal classes: finitely supported signals, even and odd signals, absolutely summable signals, finite energy signals, and finite average power signals. The finite-energy signals are by far the most important. This signal class has a particularly elegant structure. The finite energy signals are usually at the center of theoretical signal processing and analysis discussions.

It is from such signal families that the notion of a scale-domain description of a signal arises. A scale-domain description decomposes a signal into parts based on shape over a given length of time. All of the parts contain the same shape, even though the time-domain extent of the shape element varies. In order for a shape element to be so localized, the component signal must eventually die out; it becomes zero, or at least effectively zero. Thus, signals that oscillate forever, as do the sinusoids, do not directly serve a scale-domain analysis. Signals that diminish near infinity, such as Gaussians, Gabor elementary functions, and the like, are used for scale-domain description.

1.6.1 Basic Classes

Useful distinguishing properties of signals are their symmetry and their behavior near infinity.

1.6.1.1 Even and Odd Signals. One of the important characteristics of a signal is its symmetry. Symmetries allow us to simplify the description of a signal; we only need to know about the shape of the signal over some restricted domain. Uncovering symmetries can also be a first step to decomposing a signal into constituent parts. For brevity, this section primarily discusses discrete signals, but for analog signals, similar definitions and properties follow easily.

Definition (Symmetry, Even and Odd Signals). A discrete signal $x(n)$ is *symmetric* about the time instant $n = p$ if $x(p + n) = x(p - n)$ for all $n \in \mathbb{Z}$. And $x(n)$ is *antisymmetric* about the time instant p if $x(p + n) = -x(p - n)$ for all nonzero $n \in \mathbb{Z}$. A discrete signal $x(n)$ is *even* if it is symmetric about $n = 0$. Similarly, if $x(n)$ is antisymmetric about $n = 0$, then x is *odd*.

Corresponding definitions exist for symmetries of analog signals $x(t)$.

Definition (Even and Odd Part of Signals). Let $x(n)$ be a discrete signal. Then the *even part* of $x(n)$ is

$$x_e(n) = \frac{x(n) + x(-n)}{2}. \tag{1.63a}$$

The *odd part* of $x(n)$ is

$$x_o(n) = \frac{x(n) - x(-n)}{2}. \tag{1.63b}$$

There are corresponding definitions for the even and odd parts of analog signals as well.

Proposition (Even/Odd Decomposition). If $x(n)$ is a discrete signal, then

(i) $x_e(n)$ is even;
(ii) $x_o(n)$ is odd;
(iii) $x(n) = x_e(n) + x_o(n)$.

Proof: Exercise. ■

Examples. $\sin(t)$ is odd; $\cos(t)$ is even; and the Gaussian, $g_{\mu,\sigma}(t)$ of mean μ and standard deviation σ (1.14), is symmetric about μ.

Of course, some signals are neither even nor odd. For complex-valued signals, we often look at the real and imaginary components for even and odd symmetries.

1.6.1.2 Finitely Supported Signals.

The set of time values over which a signal x is nonzero is called the *support* of x. *Finitely supported* signals are zero outside some finite interval. For analog signals, a related concept is also useful—compact support.

Definition (Finite Support). A discrete signal $x(n)$ is *finitely supported* if there are integers $M < N$ such that $x(n) = 0$ for $n < M$ and $n > N$.

If $x(n)$ is finitely supported, then it can be specified via square brackets notation: $x = [k_M, \ldots, \underline{kd_0}, \ldots, k_N]$, where $x(n) = k_n$ and $M \leq 0 \leq N$.

For analog signals, we define the concept of finite support as we do with discrete signals; that is, $x(t)$ is of *finite support* if it is zero outside some interval $[a, b]$ on the real line. It turns out that our analog theory will need more specialized concepts from the topology of the real number line [44, 70].

Definition (Open and Closed Sets, Open Covering, Compactness). A set $S \subseteq \mathbb{R}$ is *open* if for every $s \in S$, there is an open interval *(a, b)* such that $s \in (a, b) \subseteq S$. A set is *closed* if its complement is open. An *open covering* of a $S \subseteq \mathbb{R}$ is a family of open sets $\{O_n \mid n \in \mathbb{N}\}$ such that $\bigcup_{n=0}^{\infty} O_n \supseteq S$. Finally, a set $S \subseteq \mathbb{R}$ is *compact* if for every open covering of S, $\{O_n \mid n \in \mathbb{N}\}$, there is a finite subset that also contains S:

$$\bigcup_{n=0}^{N} O_n \supseteq S, \tag{1.64}$$

for some $\{n \in \mathbb{N}\}$.

Definition (Compact Support). An analog signal $x(t)$ has *compact support* if $\{t \in \mathbb{R} \mid x(t) \neq 0\}$ is compact.

It is easy to show that a (finite) sum of finitely supported discrete signals is still of finite support; that is, the class of finitely supported signals is closed under addition. We will explore this and other operations on signals, as well as the associated closure properties in Chapters 2 and 3. The following theorem connects the idea of compact support for analog signals to the analogous concept of finite support for discrete signals [44, 70].

Theorem (Heine–Borel). $S \subseteq \mathbb{R}$ is compact if and only if it is closed and contained within some finite interval $[a, b]$ (that is, it is *bounded*).

Proof: The exercises outline the proof. ■

1.6.2 Summable and Integrable Signals

Compact support is a very strong constraint on a signal. This section introduces the classes of absolutely summable (discrete) and absolutely integrable (analog) signals. Their decay is sufficiently fast so that they are often neglible for large time values. Their interesting values are concentrated near the origin, and we can consider them as having localized shape.

Definition (Absolutely Summable Signals). A discrete signal $x(n)$ is *absolutely summable* (or simply *summable*) if the sum of its absolute values is finite:

$$\sum_{n=-\infty}^{\infty} |x(n)| < \infty. \tag{1.65}$$

Another notation for this family of discrete signals is l^1. Finite support implies absolutely summability.

Definition (Absolutely Integrable Signals). A signal $x(t)$ is *absolutely integrable* (or simply *integrable*) if the integral of its absolute value over $\mathbb{R}$ is finite:

$$\int_{-\infty}^{\infty} |x(t)|\, dt < \infty\,. \tag{1.66}$$

Other notations for this analog signal family are L^1 or $L^1[\mathbb{R}]$. Signals that are integrable of an interval $[a, b]$ are in $L^1[a, b]$. They satisfy

$$\int_{a}^{b} |x(t)|\, dt < \infty\,. \tag{1.67}$$

1.6.3 Finite-Energy Signals

The most important signal classes are the discrete and analog finite energy signals.

Definition (Finite–Energy Discrete Signals). A discrete signal $x(n)$ has *finite energy* or is *square-summable* if

$$\sum_{n\,=\,-\infty}^{\infty} |x(n)|^2 < \infty\,. \tag{1.68}$$

Another notation for this family of discrete signals is l^2. Note that a discrete signal that is absolutely summable must also be finite energy. We require the square of the absolute value $|x(n)|^2$ in (1.68) to accomodate complex-valued signals.

Definition (Finite-Energy Analog Signals). An analog signal $x(t)$ is *finite-energy* (or *square-integrable*) if

$$\int_{-\infty}^{\infty} |x(t)|^2\, dt < \infty. \tag{1.69}$$

Alternative names for this family are L^2 or $L^2[\mathbb{R}]$. $L^1[a, b]$ signals satisfy

$$\int_{a}^{b} |x(t)|^2\, dt < \infty. \tag{1.70}$$

The term "finite-energy" has a physical meaning. The amount of energy required to generate a real-world signal is proportional to the total squares of its values. In classical electromagnetic theory, for example, a radio wave carries energy that is

proportional to the sum of the squares of its electric and magnetic fields integrated over the empty space through which the fields propagate.

Discrete and analog finite-energy signals are central to the later theoretical development. The next two chapters generalize the concept of a vector space to infinite dimensions. A discrete signal is like a vector that is infinitely long in both positive and negative directions. We need to justify mathematical operations on signals so that we can study the processes that operate upon them in either nature or in engineered systems. The goal is to find classes of signals that allow infinite support, yet possess all of the handy operations that vector space theory gives us: signal sums, scalar multiplication, inner (dot) product, norms, and so forth.

1.6.4 Scale Description

Only recently have we come to understand the advantages of analyzing signals by the size of their time-domain features. Before the mid-1980s, signal descriptions using frequency content predominated. Sometimes the frequency description was localized, but sometimes these methods break down. Other applications naturally invite an analysis in terms of the feature scales. At a coarse scale, only large features of the signal are evident. In a speech recognition application, for example, one does not perform a frequency decomposition or further try to identify phonemes if a coarse-scale inspection of the signal reveals only the presence of low-level background noise. At a finer scale, algorithms separate words. And at even higher resolution, the words may be segmented into phonemes that are finally subjected to recognition efforts. Although a frequency-domain analysis is necessary to identify phonemes, therefore, some kind of scale-domain analysis may be appropriate for the initial decomposition of signal.

Figure 1.32 shows an example from image analysis. One-dimensional analysis is possible by extracting lines from the image. In fact, many image analysis applications approach the early segmentation steps by using one-dimensional methods at a series of coarse scales. The time-consuming, two-dimensional analysis is thus postponed as long as possible.

1.6.5 Scale and Structure

Signal description at many scales is one of the most powerful methods for exposing a signal's structure. Of course, a simple parsing of a signal into time-domain subsets that do and do not contain useful signal represents a structural decomposition. However, when this type of signal breakdown is presented a different scales, then an artificial intelligence algorithm can home in on areas of interest, perform some goal-directed interpretation, and proceed—based upon the coarse scale results—to focus on minute details that were ignored previously. Thus, the structural description of a signal resembles a tree, and this tree, properly constructed, becomes a guide for the interpretation of the signal by high-level algorithms.

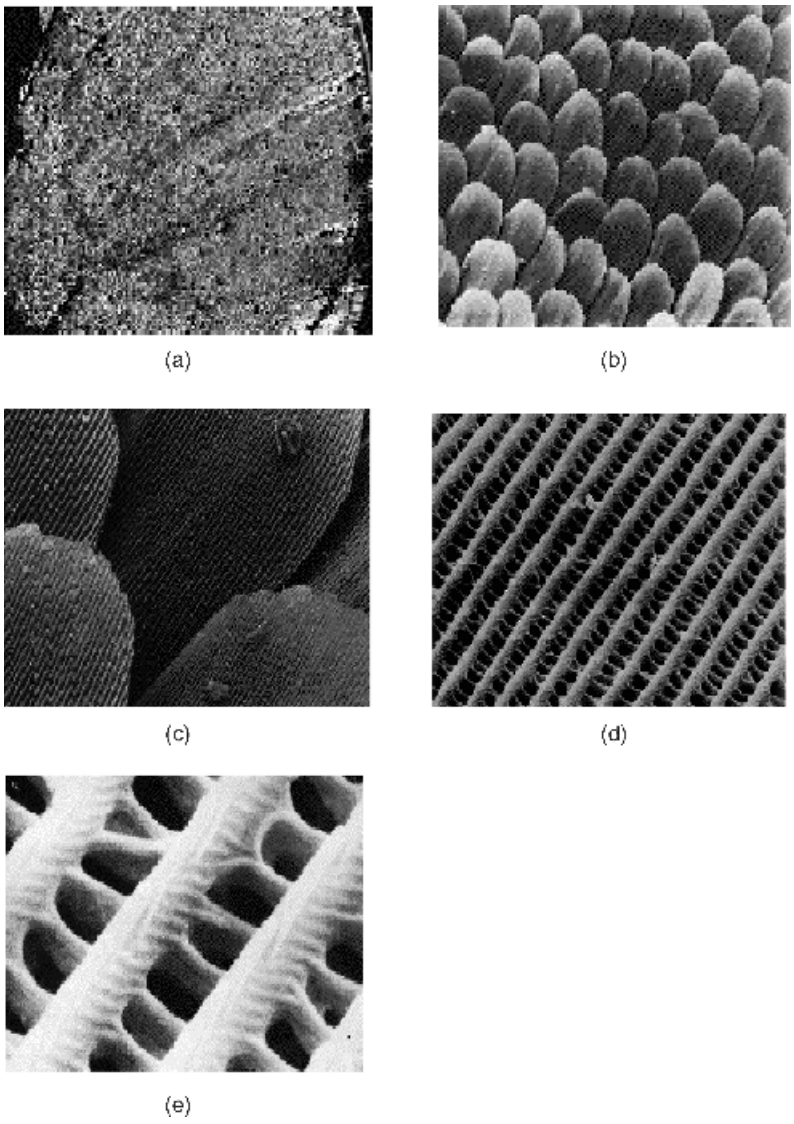


Fig. 1.32. A butterfly and a scanning electron microscope (SEM) teach us the concept of scale. The first image (a), which looks like a butterfly wing is taken at a magnification of 9×. (b) The SEM, at a power of 330×, reveals a scaly pattern. (c) At 1700× the scales appear to possess striations. (d) This confirms the existence of striations at 8500× and hints of small-scale integuments between the principal linear structures. (e) This exposes both the coarse-scale striations and the fine-scale integuments between them at 40,000× magnification.

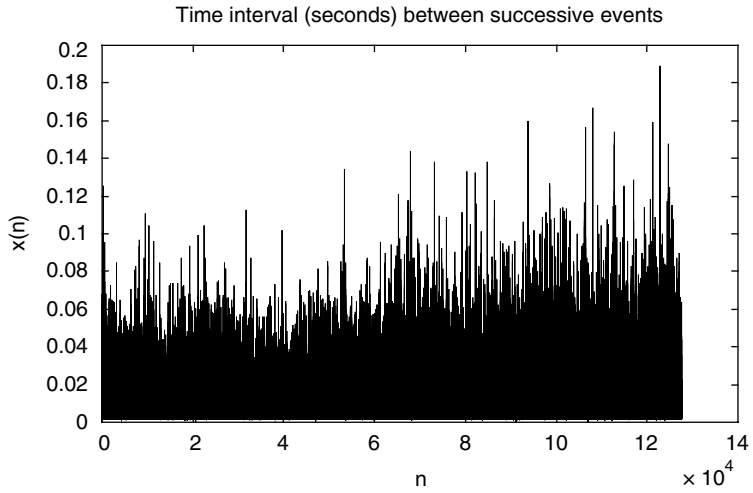


Fig. 1.33. Fractal behavior of an auditory neuron. From a single neuron in the hearing pathways of an anesthesized cat, researchers recorded the time intervals between successive neural discharge events in the presence of a tone stimulus. Note in particular that the independent variable represents not the flow of time or an ordering by distance but rather an enumeration of successive neural discharge events.

Two methods of structural description rely on signal shape models:

(i) Using self-similar shape models over a range of scales
(ii) Using a library of special shapes

Applications using self-similar shape models have begun to draw the attention of researchers in recent years. This approach makes possible a fractal analysis of a signal pattern. A fractal application attempts to understand to what degree the same shape occurs within a signal at different scales.

Sometimes, naturally occurring signals unexpectedly reveal fractal properties (Figure 1.33). This signal monitors neurons in the auditory pathway of an anesthetized cat. The signal's independent variable is not temporally dimensioned. Rather, each signal instant represents the next neural event, and the value of the signal is the time interval between such events. This twist in the conventional way of sketching a signal is key to arriving at the fractal behavior of the neuron.

Representing the signal in this special way provides insight into the nature of the neural process [71]. The discharge pattern reveals fractal properties; its behavior at large scales (in this case, over a span of many discharge events) resembles its behavior over small scales. Recently, signal analysts have brought their latest tools—among them the wavelet transform (Chapter 11), which breaks down a signal according to its frequency content within a telescoping family of scales—to bear on fractal analysis problems [53, 69].

1.7 SIGNALS AND COMPLEX NUMBERS

Complex numbers are useful for two important areas of signal theory:

(i) Computation of timing relationships (phase) between signals;
(ii) Studying the frequency content of signals.

1.7.1 Introduction

To understand why complex numbers are computationally useful, let us consider the superposition of two sinusoidal waves. Earlier, we considered this important case. It occurs in electrical circuits, where two voltages are summed together; in speech recognition, where vowel phonemes, for example, are represented by a sinusoidal sum; and in optics, where two optical wavefronts combine and interfere with one another to produce an interference pattern. Introducing complex numbers into the mathematical description of signal phenomena makes the analysis much more tractable [72].

Let $x(t) = x_1(t) + x_2(t) = A_1\cos(\Omega_1 t + \phi_1) + A_2\cos(\Omega_2 t + \phi_2)$. An awkward geometric argument showed earlier that if $\Omega_1 = \Omega_2 = \Omega$, then $x(t)$ remains sinusoidal. Why should a purely algebraic result demand a proof idea based on rotating parallelograms? Complex numbers make it easier. We let $x_1(t)$ and $x_2(t)$ be the real parts of the complex exponentials: $x_1(t) = \text{Real}[z_1(t)]$ and $x_2(t) = \text{Real}[z_2(t)]$, where

$$z_1(t) = A_1\exp(j[\Omega t + \phi_1]), \tag{1.71a}$$

$$z_2(t) = A_2\exp(j[\Omega t + \phi_2]). \tag{1.71b}$$

Then $x(t) = \text{Real}[z_1(t) + z_2(t)] = \text{Real}[z(t)]$. We calculate

$$\begin{aligned} z(t) &= A_1\exp(j[\Omega t + \phi_1]) + A_2\exp(j[\Omega t + \phi_2]) \\ &= \exp(j\Omega t)[A_1\exp(j\phi_1) + A_2\exp(j\phi_2)] \end{aligned} \tag{1.72}$$

Notice that the sum (1.72) has radial frequency Ω radians per second; the $\exp(j\Omega t)$ term is the only one with a time dependence. To calculate $\|z(t)\|$, note that $\left\|e^{j\Omega t}\right\| = 1$, and so $z(t) = \left\|A_1\exp(j\phi_1) + A_2\exp(j\phi_2)\right\|$. As before, we find

$$\|z(t)\|^2 = A_1^2 + A_2^2 + 2A_1A_2\cos(\phi_2 - \phi_1). \tag{1.73}$$

Thus, complex arithmetic takes care of the phase term for us, and this is one reason why complex arithmetic figures in signal theory. Of course, we understand that only the real part of the complex-valued signal model corresponds to any physical reality.

Now let us consider complex-valued functions, $f: \mathbb{C} \to \mathbb{C}$, and develop the ideas of calculus for them. Such functions may at first glance appear to bear a strong resemblance to signals defined on pairs of real numbers: $f: \mathbb{R} \times \mathbb{R} \to \mathbb{C}$, for example. To pursue the idea for a moment, we can let $z = x + jy$, where $j^2 = -1$. Then, $f(z) = f(x + jy)$, and f is a function of the real pair (x, y). The function f is complex-valued, of course, but we can take its real and imaginary parts: $f(z) = \text{Real}[f(x + jy)] + j\text{Imag}[f(x + jy)]$. Now $f(z)$ looks suspiciously like a sum of two multidimensional signals—a sum of two images—one of which is scaled by the imaginary square root of -1. If we were to define differentiation of $f(z)$ with respect to x and y, where $z = x + jy$, then this would indeed be the case; our theory of complex analysis would look a lot like ordinary real analysis. But when we define differentiation with respect to the complex variable z, what a huge difference it makes! A seemingly innocuous distinction about how to define differentiation makes the calculus of complex variables rich, novel, and powerful.

1.7.2 Analytic Functions

The existence of a derivative is a very special and far-reaching property for a complex function $f(z)$.

Definition (Differentiation, Derivative). Let $S \subseteq \mathbb{C}$ and $f: S \to \mathbb{C}$. Then f is *differentiable* at a point $z \in S$ if the limit,

$$f'(z) = \lim_{w \to z} \frac{f(w) - f(z)}{w - z} \tag{1.74}$$

exists. As in calculus, the limit $f'(z) = \frac{d}{dz} f(z)$ is called the *derivative* of f at z.

Definition (Analyticity). Let $w \in \mathbb{C}$. If there is an $R > 0$ such that $f(z)$ is differentiable for all at all z such that $|w - z| < R$, then f is *analytic* at w. If $f(z)$ is analytic at every $w \in S$, then $f(z)$ is analytic in S.

Proposition (Differentiation). Let f and g be differentiable at $z \in \mathbb{C}$. Then

(i) f is continuous at z.
(ii) If $c \in \mathbb{C}$, then cf is differentiable at z, and $(cf)'(z) = cf'(z)$.
(iii) $f+g$ is differentiable at z, and $(f + g)'(z) = f'(z) + g'(z)$.
(iv) fg is differentiable at z, and $(fg)'(z) = fg'(z) + f'g(z)$.
(v) If $g(z) \neq 0$, then f/g is differentiable at z, and

$$(f/g)'(z) = \frac{g(z)f'(z) - f(z)g'(z)}{g^2(z)}. \tag{1.75}$$

Proof: As in calculus of a real variable [44]; also see complex analysis texts [73–75]. ■

Proposition (Chain Rule). Let f be differentiable at $z \in \mathbb{C}$ and let g be differentiable at $f(z)$. Then the composition of the two functions, $(g \circ f)(z) = g(f(z))$, is also differentiable, and $(g \circ f)'(z) = g'(f(z))f'(z)$.

Proof: Refer to Refs. [44] and [73–75]. ■

Power series of a complex variable are useful for the study of discrete *systems*, signal processing entities that modify discrete signals. A system takes a signal as an input and produces another signal as an output. A special complex power series, called the *z-transform* of a signal, is studied in Chapters 8 and 9.

Definition (Power and Laurent Series). A *complex power series* is a sum of scaled powers of the complex variable z:

$$\sum_{n=0}^{\infty} a_n z^n, \tag{1.76a}$$

where the a_n are (possibly complex) coefficients. Sometimes we expand a complex power series about a point $w \in \mathbb{C}$:

$$\sum_{n=0}^{\infty} a_n (z-w)^n. \tag{1.76b}$$

A *Laurent series* is two-sided:

$$\sum_{n=-\infty}^{\infty} a_n z^n \tag{1.77a}$$

and can be expanded about $w \in \mathbb{C}$:

$$\sum_{n=-\infty}^{\infty} a_n (z-w)^n. \tag{1.77b}$$

[The z-transform of $x(n)$ is in fact a Laurent expansion on the values of the discrete signal: $x(n) = a_n$ in (1.77a).]

We define special complex functions in terms of power series:

$$\sin(z) = z - \frac{z^3}{3!} + \frac{z^5}{5!} - \frac{z^7}{7!} + \cdots; \tag{1.78a}$$

$$\cos(z) = 1 - \frac{z^2}{2!} + \frac{z^4}{4!} - \frac{z^6}{6!} + \cdots; \tag{1.78b}$$

and the most important function in mathematics,

$$\exp(z) = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \frac{z^4}{4!} + \cdots = e^z. \tag{1.79}$$

Their convergence criteria are similar to those of real power series. The following theorem says that a convergent power series is differentiable and its derivative may be computed by termwise differentiation.

Proposition (Power Series Differentiation). Suppose that

$$p(z) = \sum_{n=0}^{\infty} a_n (z-w)^n \tag{1.80}$$

converges in $S = \{z : |z-w| < R\}$. Then $p(z)$ is analytic (differentiable at every point) inside S, and

$$p'(z) = \sum_{n=1}^{\infty} n a_n (z-w)^{n-1}. \tag{1.81}$$

Proof: References [44] and [73–75]. ■

The next theorem suggests that complex function calculus is very different from conventional real variable theory [73–75].

Theorem (Cauchy–Riemann Equations). Suppose that $f(z) = u(x, y) + jv(x, y)$, where u and v are the real and imaginary parts, respectively, of $f(z)$. If f is differentiable at $z = w$, then the partial derivative $\partial u/\partial x, \partial u/\partial y, \partial v/\partial x$, and $\partial v/\partial y$ all exist; furthermore,

$$\frac{\partial u}{\partial x}(w) = \frac{\partial v}{\partial y}(w), \tag{1.82a}$$

$$\frac{\partial u}{\partial y}(w) = \frac{\partial v}{\partial x}(w). \tag{1.82b}$$

Proof: We can compute the derivative $f'(w)$ in two different ways. We can approach w along the real axis or along the imaginary axis. Thus, we see that the following limit gives $f'(w)$ when we approach w from values $w + h$, where h is real:

$$\lim_{h \to 0} \frac{f(w+h) - f(w)}{h} = \lim_{h \to 0} \frac{u(x+h, y) - u(x, y)}{h} + j \lim_{h \to 0} \frac{v(x+h, y) - v(x, y)}{h}. \tag{1.83a}$$

But $f'(w)$ can also be computed by taking the limit along the imaginary axis; we now approach w from values $w+jk$, where k is real. Consequently,

$$\lim_{k \to 0} \frac{f(w+jk)-f(w)}{jk} = \lim_{k \to 0} \frac{u(x, y+k)-u(x, y)}{jk} + j \lim_{k \to 0} \frac{v(x, y+k)-v(x, y)}{jk}. \tag{1.83b}$$

The limit (1.83a) is

$$f'(w) = \frac{\partial u}{\partial x}(x, y) + j\left[\frac{\partial v}{\partial x}(x, y)\right], \tag{1.84a}$$

whereas (1.83b) is

$$f'(w) = -j\left[\frac{\partial u}{\partial y}(x, y)\right] + \frac{\partial v}{\partial y}(x, y). \tag{1.84b}$$

Since $f'(w)$ has to be equal to both limits, the only way to reconcile (1.84a) and (1.84b) is to equate their real and imaginary parts, which gives (1.82a)–(1.82b). ■

Remark. The Cauchy–Riemann equations imply that some surprisingly simple complex functions, $f(z) = f(x + jy) = x - jy$, for example, are not differentiable.

The converse to the theorem requires an additional criterion on the partial derivatives, namely that they be continuous.

Theorem (Cauchy–Riemann Converse). Let $f(z) = u(x, y) + jv(x, y)$, where u and v are the real and imaginary parts, respectively, of $f(z)$. Furthermore, let the partial derivatives $\partial u/\partial x$, $\partial u/\partial y$, $\partial v/\partial x$, and $\partial v/\partial y$ all exist and be continuous and satisfy the Cauchy–Riemann equations (1.82a)–(1.82b) at $z = w$. Then $f'(w)$ exists.

Proof: Not too difficult [73–75]. ■

Corollary. Let $f(z) = u(x, y) + jv(x, y)$, where u and v are the real and imaginary parts, respectively, of $f(z)$. If $f'(w)$ and $f''(w)$ exist, then the partial derivatives of u and v obey the Laplace equation:

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} = 0. \tag{1.85}$$

Proof: By the Cauchy–Riemann theorem, u and v satisfy (1.82a)–(1.82b). Applying the theorem again to the deriviatives, and using the fact from calculus that

mixed partial derivatives are equal where they are continuous, we find

$$\frac{\partial}{\partial x}\frac{\partial u}{\partial x}(w) = \frac{\partial}{\partial x}\frac{\partial v}{\partial y}(w) = \frac{\partial}{\partial y}\frac{\partial v}{\partial x}(w) = -\frac{\partial}{\partial y}\frac{\partial u}{\partial y}(w) \tag{1.86a}$$

Similarly,

$$\frac{\partial}{\partial y}\frac{\partial v}{\partial y}(w) = -\frac{\partial}{\partial x}\frac{\partial v}{\partial x}(w). \tag{1.86b}$$

The Laplace equations for both u and v follow. ■

This is an intriguing result. Complex differentiability leads to a second-order partial differential equation. That is, if a function *f(z)* is twice differentiable, then it is harmonic in a set $S \subseteq \mathbb{C}$. Thus, complex differentiation is already seen to be a much more restricted condition on a function than real differentiation. Laplace's equation appears in many applications of physics and mechanics: heat conduction, gravitation, current flow, and fluid flow, to name a few. The import of the corollary is that complex functions are a key tool for understanding such physical systems. For applications to the theory of fluid flow, for example, see Ref. 75.

Even stronger results are provable. The next section outlines the development of complex integration theory. It seems quite backwards to prove theorems about differentiation by means of integration theory; but in the exotic realm of complex analysis, that is exactly the course we follow. Using contour integration in the complex plane, it is possible to prove that an analytic function (differentiable in a region) has continuous derivatives of all orders. That is, every analytic function expands in a Taylor series.

1.7.3 Complex Integration

This section continues our sweep through complex analysis, turning now to integration in the complex plane. Given the results of the previous section, one might imagine that complex integration should also have special properties unlike anything in real analysis. Such readers will not be disappointed; the theory of complex integration is even more amazing than differentiation.

Definition (Contour). A *curve* in the complex plane is a function $s : [a, b] \to \mathbb{C}$, where $[a, b] \subset \mathbb{R}$. We say that *s parameterizes* its range. If the real and imaginary parts of $s(t)$ are continuously differentiable, then s is called an *arc*. If $s(a) = s(b)$, then the curve s is *closed*. And if $s(t_1) = s(t_2)$ on (a, b) implies $t_1 = t_2$, then the curve s is *simple*. A sequence of arcs $\{s_n(t) : [a_n, b_n] \to C : 1 \le n \le N\}$ is a *contour* if $s_n(b_n) = s_{n+1}(a_{n+1})$, for $n = 1, 2, \ldots, N-1$.

Remarks. A curve is a complex-valued analog signal, defined on a closed interval of the real line. An arc is a continuously differentiable, complex-valued analog signal. A simple curve does not intersect itself, save at its endpoints. We often denote

an arc in the complex plane by its range, $\mathrm{C} = \{z : z = s(t), \text{ for some } a \le t \le b\}$, and the defining curve function is implicit. Our purpose is to define integration along a contour [76].

Definition (Contour Integral). If the complex function $f(z)$ is continuous in a region containing an arc C, then the contour integral of f over C is defined by

$$\oint_C f(z)\, dz = \int_a^b f[s(t)]s'(t)\, dt\,, \tag{1.87}$$

where $s(t)$ is the function that parameterizes C.

Since $f(z)$, $s(t)$, and $s'(t)$ are all continuous, the integrand in (1.87) is Riemann integrable. The function $f[s(t)]s'(t)$ is complex-valued; we therefore perform the real integration (that is, with respect to t) twice, once for the real part and once for the imaginary part of the integrand. Observe that the change of integration variable, $z = s(t)$ and $dz = s'(t)\, dt$, converts the integral's definition with respect to z in (1.87) to one with respect to t.

The main result of this section is Cauchy's integral theorem. There is an interpretation of contour integration that provides an intuitive link to the familiar theory of integration from calculus and an informal argument for the theorem [77]. Readers seeking rigor and details will find them in the standard texts [73–76]. We first consider the case where C is a circle around the origin, which is a simple, closed arc. Then we shall argue the extension to general arcs, by supposing the arcs to be the limit of a local tiling of the region by adjacent triangles. From this, the extension to contours, which are a sequence of arcs, follows directly.

Theorem (Cauchy Integral for a Circle). Suppose $f(z)$ is analytic in a region containing the closed circle C, with radius R and center $z = (0, 0)$. Then,

$$\frac{1}{2\pi j}\oint_C z^m dz = \begin{cases} 0 & \text{if } m \neq 1, \\ 1 & \text{if } m = -1. \end{cases} \tag{1.88}$$

Proof: In calculus courses [44], Riemann integrals are the limits of Riemann sums. For the case of a contour integral, such a sum is a limit:

$$\oint_C f(z)dz = \lim_{N \to \infty} \sum_{n=1}^{N} f(w_n)[z_{n+1} - z_n]\,. \tag{1.89}$$

where $s : [a, b] \to \mathbb{C}$ parameterizes the arc C; $a = t_1 < t_2 < \cdots < t_N < t_{N+1} = b$ partitions $[a, b]$; $z_n = s(t_n)$; and $w_n = s(t)$ for some $t \in [t_n, t_{n+1}]$. Suppose further that we select the t_n so that $|z_{n+1} - z_n| = \varepsilon_N = \text{Length}(C)/N = L_C/N$. Then we have

$$\oint_C f(z)dz = \lim_{N\to\infty}\sum_{n=1}^{N} f(w_n)\frac{[z_{n+1}-z_n]}{|z_{n+1}-z_n|}|z_{n+1}-z_n|$$
$$= \lim_{N\to\infty}\varepsilon_N\sum_{n=1}^{N} f(w_n)\frac{[z_{n+1}-z_n]}{|z_{n+1}-z_n|} = L_C\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} f(w_n)\frac{[z_{n+1}-z_n]}{|z_{n+1}-z_n|}. \tag{1.90}$$

Note that as $N\to\infty$, we have $w_n\to z_n$, and $(z_{n+1}-z_n)/|z_{n+1}-z_n|$ approaches a complex value whose real and imaginary parts are the components of the unit tangent vector to C at z_n, $\boldsymbol{T}(z_n)$. Since C has radius R, $\boldsymbol{T}(z_n) = jz/R$ and $L_C = 2\pi R$. Therefore, the final sum in (1.90) approaches $L_C\times\{\text{average over } C \text{ of } f(z)\boldsymbol{T}(z)\}$. We conclude

$$\frac{1}{L_C}\oint_C f(z)\,dz = \text{Avg}_{z\in C}[f(z)\boldsymbol{T}(z)]. \tag{1.91}$$

Now suppose $m = -1$, so that $f(z) = z^{-1}$. Then

$$\frac{1}{2\pi R}\oint_C f(z)\,dz = \text{Avg}_{z\in C}\left[\frac{\boldsymbol{T}(z)}{z}\right] = \text{Avg}_{z\in C}\left[\frac{jz}{Rz}\right] = \text{Avg}_{z\in C}\left[\frac{j}{R}\right] = \frac{j}{R}. \tag{1.92}$$

To show the other possibility in (1.88), we let $m\neq -1$ and find

$$\frac{1}{2\pi R}\oint_C f(z)\,dz = \text{Avg}_{z\in C}\left[\frac{jz^{m+1}}{R}\right] = \frac{j}{R}\text{Avg}_{z\in C}[z^{m+1}]. \tag{1.93}$$

But, the average of all values z^{m+1} over the circle $|z| = R$ is zero, which demonstrates the second possibility of (1.88) and concludes the proof. ■

Note that the informal limit (1.89) is very like the standard calculus formulation of the Riemann integral. The definition of the contour integral is thus a plausible generalization to complex-valued functions.

We will apply this result in Chapter 8 to derive one form of the inverse z-transform. This produces discrete signal values $x(n)$ from the complex function $X(z)$ according to the rule:

$$x(n) = \frac{1}{2\pi j}\oint_C X(z)z^{n-1}dz. \tag{1.94}$$

The Cauchy residue theorem leads to the following concepts.

Definition (Poles and Zeros). A complex function $f(z)$ has a *pole* of order k at $z = p$ if there is a $g(z)$ such that $f(z) = g(z)/(z-p)^k$, $g(z)$ is analytic in an open set containing $z = p$, and $g(p)\neq 0$. We say that $f(z)$ has a *zero* of order k at $z = p$ if there is a $g(z)$ such that $f(z) = g(z)(z-p)^k$, $g(z)$ is analytic in a region about $z = p$, and $g(p)\neq 0$.

Definition (Residue). The *residue* of $f(z)$ at the pole $z = p$ is given by

$$\text{Res}(f(z), p) = \begin{cases} \left[\dfrac{1}{(k-1)!} f^{(k-1)}(p)\right] & \text{if } p \in \text{Interior}(C), \\ 0 & \text{if otherwise,} \end{cases} \tag{1.95}$$

where k is the order of the pole.

Theorem (Cauchy Residue). Assume that $f(z)$ is a complex function, which is analytic on and within a curve C; $a \notin C$; and $f(z)$ is finite (has no pole) at $z = a$. Then

$$\frac{1}{2\pi j}\oint_C \frac{f(z)}{(z-a)^{m-1}}\, dz = \begin{cases} \left[\dfrac{1}{(m-1)!} f^{(m-1)}(a)\right] & \text{if } a \in \text{Interior}(C), \\ 0 & \text{if otherwise.} \end{cases} \tag{1.96}$$

More generally, we state the following theorem.

Theorem (Cauchy Residue, General Case). Assume that C is a simple, closed curve; $a_m \notin C$ for $1 \le m \le M$; and $f(z)$ is analytic on and within C, except for poles at each of the a_m. Then

$$\frac{1}{2\pi j}\oint_C f(z)\, dz = \sum_{m=1}^{M} \text{Res}(f(z), a_m), \tag{1.97}$$

Proof: References 73–76. ■

1.8 RANDOM SIGNALS AND NOISE

Up until now, we have assumed a close link between mathematical formulas or explicit rules and our signal values. Naturally occurring signals are inevitably corrupted by some random noise, and we have yet to capture this aspect of signal processing in our mathematical models. To incorporate randomness and make the models more realistic, we need more theory.

We therefore distinguish between *random* signals and *deterministic* signals. Deterministic signals are those whose values are completely specified in terms of their independent variable; their exact time domain description is possible. The signal may be discrete or continuous in nature, but as long as there is a rule or formula that relates an independent variable value to a corresponding signal value, then the signal is deterministic. In contrast, a random signal is one whose values are not known in terms of the value of its independent variable. It is best to think of time-dependent signals to understand this. For a random signal, we cannot know the value of the signal in advance; however, once we measure the signal at a particular

time instant, only then do we know its value. Deterministic signals are good for carrying information, because we can reliably insert and extract the information we need to move in a reliable fashion. Nature is not kind, however, to our designs. A random component—for example, a measurement error, digitization error, or thermal noise—corrupts the deterministic signal and makes recovery of the signal information more difficult.

This situation often confronts electrical communication engineers. There are many sources of noise on telephone circuits, for example. If the circuits are physically close, electromagnetic coupling between them occurs. Faint, but altogether annoying, voices will interfere with a conversation. One might argue that this is really a deterministic interference: someone else is deliberately talking, and indeed, if the coupling is strong enough, the other coversation can be understood. However, it is in general impossible to predict when this will occur, if at all, and telephony engineers allow for its possibility by considering models of random signal interference within their designs. Thermal noise from the random motion of electrons in conductors is truly random. It is generally negligible. But it becomes significant when the information-bearing signals are quite weak, such as at the receiving end of a long line or wireless link.

An important signal analysis problem arises in communication system design. In a conversation, a person speaks about 35% of the time. Even allowing that there are two persons talking and that both may speak at once, there is still time available on their communication channel when nobody speaks. If circuitry or algorithms can detect such episodes, the channel can be reused by quickly switching in another conversation. The key idea is to distinguish voice signals from the channel's background noise. There is one quirk: When the conversation is broken, the telephone line sounds dead; one listener or the other invariably asks, "Are you still there?" In order to not distress subscribers when the equipment seizes their channel in this manner, telephone companies actually synthesize noise for both ends of the conversation; it sounds like the connection still exists when, in fact, it has been momentarily broken for reuse by a third party. This is called *comfort noise* generation. A further problem in digital telephony is to estimate the background noise level on a voice circuit so that the equipment can synthesize equivalent noise at just the right time.

Now let us provide some foundation for using random signals in our development. Our treatment is quite compact; we assume the reader is at least partly familiar with the material. Readers can gain a deeper appreciation for discrete and continuous probability space theory from standard introductory texts [78–81]. Random signal theory is covered by general signal processing texts [13, 14] and by books that specialize in the treatment of random signals [82, 83].

1.8.1 Probability Theory

This section introduces the basic principles and underlying definitions of probability theory, material that should already be familiar to most readers.

Consider the noise in the 12-lead electrocardiogram signal. Close inspection of its trace shows small magnitude jaggedness, roughness of texture, and spiky artifacts. Variations in the patient's physical condition and skin chemistry, imperfections in the sensors, and flaws in the electronic signal conditioning equipment impose an element of randomness and unknowability on the ECG's value at any time. We cannot know the exact voltage across one of the ECG leads in advance of the measurement. Hence, at any time t, the voltage across a chosen ECG lead $v(t)$ is a *random variable*. All of the possible activities of ECG signal acquisition constitute the *sample space*. An *event* is a subset of the the sample space. For instance, recording the ECG signal at a moment in time is an event. We assign numerical likelihoods or probabilities to the ECG signal acquisition events.

1.8.1.1 Basic Concepts and Definitions. In order that probability and random signal theory work correctly, the events must obey certain rules for separating and combining them.

Definition (Algebra and σ-Algebra). An *algebra* over a set Ω is a collection of subsets of Ω, $\Sigma \subseteq \wp(\Sigma) = \{A : A \subseteq \Omega\}$, with the following properties:

(i) The empty set is in Σ: $\varnothing \in \Sigma$.
(ii) If $A \in \Sigma$, then the complement of A is in Σ: $A' \in \Sigma$.
(iii) If $A, B \in \Sigma$, then $A \cup B \in \Sigma$.

A σ-*algebra* over a set Ω is an algebra Σ with a further property:

(iv) If $A_n \in \Sigma$ for all $n \in \mathbb{N}$, then their union is in Σ:

$$\bigcup_{n=0}^{\infty} A_n \in \Sigma. \tag{1.98}$$

It is easy to verify that in an algebra Σ, $\Omega \in \Sigma$, the union of any finite set of its elements is still in Σ, and Σ is closed under finite intersections. A σ-algebra is also closed under the intersection of infinite families of elements as in (1.98).

The probability measure must have certain mathematical properties.

Definition (Probability Measure). A *probability measure* on a σ-algebra Σ over Ω is a function $P: \Sigma \to [0, 1]$ such that

(i) $P(\Omega) = 1$;
(ii) P sums on disjoints unions; that is, if $\{A_n : n \in I\} \subseteq \Sigma$, where $I \subseteq \mathbb{N}$, and $A_n \cap A_m = \varnothing$, when $n \neq m$, then

$$P(\bigcup_{n \in I} A_n) = \sum_{n \in I} P(A_n). \tag{1.99}$$

Definition (Probability Space). A *probability space* is an ordered triple (Ω, Σ, P), where Ω is a set of experimental outcomes, called the *sample space*; Σ is a σ-algebra over Ω, the elements of which are called *events*; and P is a probability measure on Σ. The event $\varnothing$ is called the *impossible* event, and the event Ω is called the *certain* event.

Alternative approaches to probability exist. The earliest theories are drawn from the experiments of early gambler-mathematicians, such as Cardano and Pascal.[17] Their dice and card games, run through many cycles—sometimes to the point of financial ruin of the investigator—inspired an alternative definition of probability. It is the value given by the limiting ratio of the number of times the event occurs divided by the number of times the experiment has been tried:

$$P(X) = \lim_{n \to \infty} \frac{O_{X,n}}{n}, \tag{1.100}$$

where $O_{X,n}$ is the number of observations through n trials where X occurred. This intuition serves as a foundation for probability. The exercises supply some flavor of the theoretical development. More widely accepted, however, is the axiomatic approach we follow here. Soviet mathematicians—notably Kolmogorov[18]—pioneered this approach in the 1930s. Through William Feller's classic treatise [79] the axiomatic development became popular outside the Soviet Union. Most readers are probably familiar with this material; those who require a complete treatment will find [78–81] helpful.

1.8.1.2 Conditional Probability.

Conditional probability describes experiments where the probability of one event is linked to the occurrence of another.

Definition (Conditional Probability, Independence). Suppose A and B are two events. The probability that A will occur, given that B has occurred, is defined as

$$P(A \mid B) = P\frac{(A \cap B)}{P(B)}. \tag{1.101}$$

The quotient $P(A|B)$ is called the *conditional probability* of event A given B.

[17]Girolamo Cardano (1501–1576) led a scandalous life as a gambler, but learned enough to found the theory of probability decades before Fermat and Pascal. Blaise Pascal (1623–1662) was a French mathematician and philosopher. See O. Ore, *Cardano, The Gambling Scholar*, New York: Dover, 1953; also, O. Ore, Pascal and the invention of probability theory, *American Mathematical Monthly*, vol. 67, pp. 409–419, 1960.

[18]Andrei Nikolaevich Kolmogorov (1903–1987) became professor of mathematics at Moscow University in 1931. His foundational treatise on probability theory appeared in 1933.

B must occur with nonzero probability for the conditional probability $P(A|B)$ to be defined.

Definition (Independent Events). Suppose A and B are two events. If $P(A|B) = P(A)P(B)$, then A and B are said to be *independent* events.

Proposition. If A and B are independent, then

(i) A and $\sim B$ are independent;
(ii) $\sim A$ and $\sim B$ are independent.

Proof: Exercise. ■

Proposition (Total Probability). Suppose $\{B_n: 1 \le n \le N\}$ is a partition of Ω and $P(B_n) > 0$ for all n. Then for any A,

$$P(A) = \sum_{n=1}^{N} P(B_n)P(A \mid B_n). \tag{1.102}$$

Proof: $A = (A \cap B_1) \cup (A \cap B_2) \cup \ldots \cup (A \cap B_N)$, which is a disjoint union. The definition of conditional probability entails

$$P(A) = \sum_{n=1}^{N} P(A \cap B_n) = \sum_{n=1}^{N} P(B_n)P(A \mid B_n). \tag{1.103}$$

■

1.8.1.3 Bayes's Theorem. An important consequence of the total probability property, known as Bayes's[19] theorem, is central to a popular pattern classification scheme (Chapter 4).

Theorem (Bayes's). Suppose $\{C_n: 1 \le n \le N\}$ is a partition of Ω and $P(C_n) > 0$ for all n. If $P(A) > 0$ and $1 \le k \le N$

$$P(C_k \mid \mathrm{A}) = \frac{P(C_k)P(A \mid C_k)}{\sum_{n=1}^{N} P(C_n)P(A \mid C_n)}. \tag{1.104}$$

[19]A friend of Thomas Bayes (1703–1761) published the Nonconformist minister's theorem in a 1764 paper before the Royal Society of London.

Proof: The definition of conditional probability implies

$$P(C_k \mid \mathrm{A}) = \frac{P(A \cap C_k)}{P(A)} = \frac{P(C_k)P(A \mid C_k)}{P(A)} = \frac{P(C_k)P(A \mid C_k)}{\sum_{n=1}^{N} P(C_n)P(A \mid C_n)}. \quad (1.105)$$

■

Example (Phoneme Classification). Consider the application of Bayes's theorem to a phoneme classification system. Phonemes fall into a fixed number of classes, C_1, C_2, …, C_N, given by the application domain. There are also a set of signal features that the application computes for each candidate phoneme. Let us suppose that there are M features, A_1, A_2, …, A_M, and the application design is so well done that, for any phoneme-bearing signal, it is possible to both reliably distinguish the phonemes from one another and to assign one of the classes A_m as the principal feature of the signal. A typical feature might be a set of sinusoidal frequencies (formants) that dominate the energy contained in the signal. In any case, we are interested in the phoneme class C_n to which a given input signal belongs. Suppose that the dominant feature is $A = A_m$. We calculate each of the probabilities: $P(C_1|A)$, $P(C_2|A)$, …, $P(C_N|A)$. The highest of these probabilities is the answer—the *Bayes classification*.

How can we calculate these N probabilities? Evidently, we must know $P(C_n)$ for each n. But any of the features might be the dominant one within a signal. Therefore, we must know $P(A_m|C_n)$ for each m and n. And, finally, we must know $P(A_m)$ for each m. A working Bayes classifier requires many probabilities to be known in advance. It is possible to develop these statistics, however, a step called the classifier *training phase*. We gather a large, representative body of speech for the application. If we classify the phonemes manually, in an offline effort, then the relative frequencies of each phoneme can be used in the real-time application. This gives us $P(C_n)$, $1 \le n \le N$. Once we identify a phoneme's class, then we find its predominant feature. For each phoneme C_n, we calculate the number of times that feature A_m turns out to be its predominant feature, which approximates $P(A_m|C_n)$. Lastly, we compute the number of times that each feature is dominant and thus estimate $P(A_m)$. Now all of the numbers are available from the training phase to support the execution of the phoneme classifier on actual data. The more sample phonemes we process and the more genuinely the training data reflects the actual application sources, the better should be our probability estimates.

It is unfortunately often the case that one cannot discover any predominant feature from a set of signal data. What we usually encounter is a feature vector $\boldsymbol{a} = (a_1, a_2, ..., a_M)$, where the a_m represent numerical values or scores indicating the presence of each feature A_m. We can compute the probability of a vector of features, but that can only be done after a little more development.

1.8.2 Random Variables

A *random variable* is a function that maps events to numerical values.

Definition (Random Variable). Suppose that (Ω, Σ, P) is a probability space. A *random variable* x on Ω is a function $\boldsymbol{x} : \Omega \to \mathbb{R}$, such that for all $r \in \mathbb{R}$, $\{\omega \in \mathbb{R}: \boldsymbol{x}(\omega) \leq r\} \in \Sigma$.

Notation. $\boldsymbol{x} \leq r$ or $\{\boldsymbol{x} \leq r\}$ is standard for the event $\{\omega \in \mathbb{R}: \boldsymbol{x}(\omega) \leq r\} \in \Sigma$. Similarly, we write $\boldsymbol{x} > r$, $\boldsymbol{x} = r$, $r < \boldsymbol{x} \leq s$, and so on. Using the properties of a σ-algebra, we can show these too are events in S. It is also possible to consider complex-valued random variables, $\boldsymbol{z} : \Omega \to \mathbb{C}$.

Definition (Distribution Function). Suppose that (Ω, Σ, P) is a probability space and $\boldsymbol{x}$ is a random variable $\boldsymbol{x} : \Omega \to \mathbb{R}$. Then the *probability distribution function*, or simply the *distribution function*, for $\mathbf{x}$ is defined by $F_{\boldsymbol{x}}(r) = P(\boldsymbol{x} \leq r)$.

Since there is no ordering relation on the complex numbers, there is no distribution function for a complex-valued random variable. However, we can consider distribution functions of the real and imaginary part combined; this topic is explored later via the concept of multivariate distributions.

Proposition (Distribution Function Properties). Let $\boldsymbol{x} : \Omega \to \mathbb{R}$ be a random variable in the probability space (Ω, Σ, P), and let $F_{\boldsymbol{x}}(r)$ be its distribution function. Then the following properties hold:

(i) If $r < s$, then $F_{\boldsymbol{x}}(r) \leq F_{\boldsymbol{x}}(s)$.
(ii) $\lim_{r \to \infty} F_{\boldsymbol{x}}(r) = 1$ and $\lim_{x \to \infty} F_{\boldsymbol{x}}(r) = 0$.
(iii) $P(\boldsymbol{x} > r) = 1 - F_x(r)$.
(iv) $P(r < \boldsymbol{x} \leq s) = F_x(s) - F_x(r)$.
(v) $P(\boldsymbol{x} = r) = F_x(r) - \lim_{s > 0,\, s \to 0} F_x(r - s)$.
(vi) $P(r \leq \boldsymbol{x} \leq s) = F_x(s) - \lim_{t > 0,\, t \to 0} F_x(r - t)$.
(vii) If $F_x(r)$ is a continuous function of r, then $P(\boldsymbol{x} = r) = 0$ for all r.

Proof: Exercise [81]. ■

The proposition's first statement (i) is a *monotonicity* property.

The distribution function of a random variable may be computed by experiment or may be assumed to obey a given mathematical rule. Special mathematical properties are often assumed for the distribution function; this facilitates mathematical investigations into the behavior of the random variable. One common assumption is that the distribution function is differentiable. This motivates the next definition.

Definition (Density Function). Suppose that (Ω, Σ, P) is a probability space and $\boldsymbol{x}$ is a random variable on Ω. If $F_x(r)$ is differentiable, then the derivative with respect to r of $F_x(r)$, denoted with a lowercase letter f,

$$f_{\boldsymbol{x}}(r) = \frac{d}{dr} F_{\boldsymbol{x}}(r), \tag{1.106}$$

is called the *probability density* function or simply the *density* function of x.

Only functions with specific properties can be density functions. The exercises explore some specific cases.

Proposition (Density Function Properties). Let $\boldsymbol{x} : \Omega \to \mathbb{R}$ be a random variable in the probability space (Ω, Σ, P) with distribution function $F_{\boldsymbol{x}}(r)$. Then

(i) $0 \leq f_{\boldsymbol{x}}(r)$ for all $r \in \mathbb{R}$.

(ii)

$$\int_{-\infty}^{\infty} f_x(t)\, dt = 1. \tag{1.107}$$

(iii)

$$F_{\boldsymbol{x}}(r) = \int_{-\infty}^{r} f_x(t)\, dt. \tag{1.108}$$

(iv)

$$P(r < x \leq s) = F_x(s) - F_x(r) = \int_{s}^{r} f_x(t)\, dt. \tag{1.109}$$

Proof: Property (i) follows from the monotonicity property of the distribution function. Property (iv) follows from the fundamental theorem of calculus [44], where we let the lower limit of the integral pass to infinity in the limit. Properties (ii) and (ii) derive from (iv) via the distribution function limit properties. ■

In the proposition, (i) and (ii) are the conditions that a general function $f : \mathbb{R} \to \mathbb{R}$ must satisfy in order to be a density function. One may also prove an existence theorem that constructs a random variable from such a density function [81]. Random variables divide into two classes: discrete and continuous, based on the continuity of the distribution function. (There is also a *mixed distribution* that has aspects of both, but it is outside our scope.)

1.8.2.1 Discrete Random Variables. Discrete random variables prevail within discrete signal theory.

Definition (Discrete Random Variable). The random variable $\boldsymbol{x}$ is *discrete* if its distribution function is a step function.

In this case, there is a set $M = \{r_n : n \in \mathbb{Z}\}$, such that $m < n$ implies $r_m < r_n$, the set of half-open intervals $[r_m, r_n)$ partition $\mathbb{R}$, and $F_{\boldsymbol{x}}(r)$ is constant on each $[r_m, r_n)$.

Proposition (Discrete Random Variable Characterization). Let $\boldsymbol{x}$ be a random variable in the probability space (Ω, Σ, P) with distribution function $F_{\boldsymbol{x}}(r)$. Set $M = \{r \in \mathbb{R} : \mathrm{P}(\mathrm{x} = \mathrm{r}) > 0\}$. Then, $\boldsymbol{x}$ is discrete if and only if

$$\sum_{r \in M} P_x(\mathrm{x} = \mathrm{r}) = 1 . \tag{1.110}$$

Proof: By the definition, we see that $P(\boldsymbol{x} \le r_n) = P(\boldsymbol{x} < r_{n+1})$. This occurs if and only if $P(r_n \le \boldsymbol{x} \le r_{n+1}) = P(\boldsymbol{x} = r_n)$. Therefore the sum (1.110) is

$$\sum_{r \in M} P_x(\mathrm{x} = \mathrm{r}) = \lim_{r \to \infty} F_{\boldsymbol{x}}(r) + \lim_{r \to -\infty} F_{\boldsymbol{x}}(r) = 1 \tag{1.111}$$

by the distribution function properties. ■

If the random variable $\boldsymbol{x}$ is discrete, then the $F_{\boldsymbol{x}}(r)$ step heights approach zero as $r \to -\infty$ and approach unity as $r \to \infty$. Because a step function is not differentiable, we cannot define a density function for a discrete random variable as in the previous section. However, we can separately define the density function for a discrete random variable as discrete impulses corresponding to the transition points between steps.

Definition (Discrete Density Function). Suppose the random variable $\boldsymbol{x}$ is discrete, and its distribution function $F_x(r)$ is constant on half-open intervals $[r_n, r_m)$ that partition R. Its density function $f_x(r)$ is defined:

$$f_x(r) = \begin{cases} F_x(r_{n+1}) - F_x(r_n) & \text{if } r = r_n, \\ 0 & \text{if otherwise.} \end{cases} \tag{1.112}$$

Example (Dice). Consider an experiment where two fair dice are thrown, such as at a Las Vegas craps table. Each die shows one to six dots. The probability of any roll on one die is, given honest dice, 1/6. The throw's total is the sum, a random variable $\boldsymbol{x}$. The values of $\boldsymbol{x}$ can be 2, 12, or any natural number in between. There are 36 possible rolls, and the probability of the event that either 2 or 12 is rolled is 1/36. Lucky seven is the most common event—with probability 6/36—as it occurs through the following tosses: (1, 6), (2, 5), (3, 4), (4, 3), (5, 2), or (6, 1). Figure 1.34 shows the distribution function and the density functions for the dice toss.

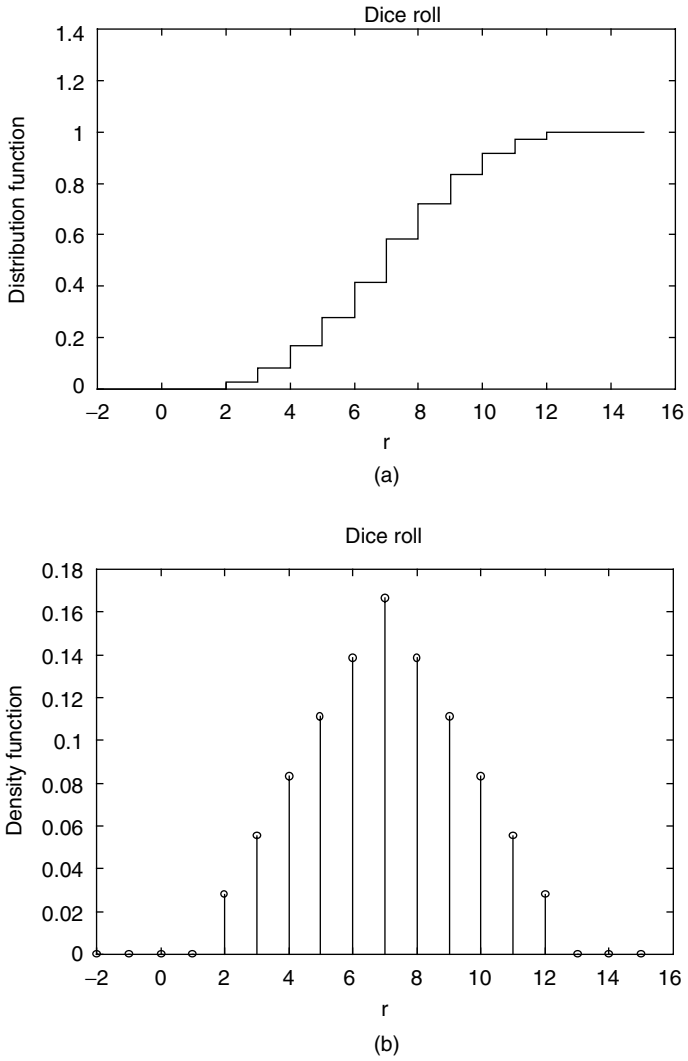


Fig. 1.34. Distribution (a) and density functions (b) for tossing a pair of dice.

Discrete signal theory commonly assumes a density or distribution function for a random variable. If the density or distribution is unknown, it can be measured, of course, but that is sometimes impractical. Instead, one typically approximates it by a distribution that has tractable mathematical properties.

Definition (Binomial Distribution). Suppose that the discrete random variable x has $\text{Range}(\boldsymbol{x}) \subseteq \{0, 1, 2, \ldots, n\}$. Then $\boldsymbol{x}$ has a *binomial distribution* of order n if

there are non-negative values p and q, such that $p + q = 1$, and

$$P(\boldsymbol{x} = k) = \binom{n}{k} p^k q^{n-k}. \tag{1.113}$$

Definition (Poisson[20] Distribution). Suppose the discrete random variable $\boldsymbol{x}$ has $\text{Range}(\boldsymbol{x}) \subseteq \{0, 1, 2, \ldots, n\}$ and $a > 0$. Then $\boldsymbol{x}$ has a *Poisson distribution* with parameter a if

$$P(\boldsymbol{x} = k) = \frac{a^k}{k!} e^{-a}. \tag{1.114}$$

We cannot know the value that a random variable will assume on an event before the event occurs. However, we may know enough about the trend of the random variable to be able to specify its average value over time and how well grouped about its average the random values tend to be. There are a variety of parameters associated with a random variable; these we calculate from its distribution or density functions. The most important of these are the mean and standard deviation.

Definition (Discrete Mean). If the random variable $\boldsymbol{x}$ is discrete and $M = \{r \in \mathbb{R}: P(\boldsymbol{x} = r) > 0\}$, then the *mean* or *expectation* of $\boldsymbol{x}$, written $E[x]$, is

$$E(x) = \sum_{r \in M} rP(\boldsymbol{x} = r). \tag{1.115}$$

Definition (Discrete Variance, Standard Deviation). Let the random variable $\boldsymbol{x}$ be discrete, $M = \{r \in \mathbb{R}: P(\boldsymbol{x} = r) > 0\}$, and $\mu = E[\boldsymbol{x}]$. Then the *variance* of $\boldsymbol{x}$, σ_x^2, is

$$\sigma_x^2 = \sum_{r \in M} (r - \mu)^2 P(\boldsymbol{x} = r). \tag{1.116}$$

The *standard deviation* of $\boldsymbol{x}$ is the square root of the variance: σ_x.

1.8.2.2 Continuous Random Variables. The distribution function may have no steps.

Definition (Continuous Random Variable). The random variable $\boldsymbol{x}$ is *continuous* if its distribution function $F_{\boldsymbol{x}}(r)$ is continuous.

Proposition (Continuous Random Variable Characterization). Let x be a continuous random variable in the probability space (Ω, Σ, P) with distribution function $F_{\boldsymbol{x}}(r)$. Then, $P(\boldsymbol{x} = r) = 0$ for all $r \in \mathbb{R}$.

[20]This distribution was first described in 1837 by French mathematician Siméon-Denis Poisson (1781–1840).

Proof: By continuity $F_x(r) = \lim_{s>0,\, s\to 0} F_x(r-s)$. But the distribution function properties entail that $P(\boldsymbol{x} = r) = F_x(r) - \lim_{s>0,\, s\to 0} F_x(r-s)$. So, $P(\boldsymbol{x} = r) = 0$. ■

Assuming a particular form of the density function and then integrating it to get a distribution is common in analytical work. The only restrictions are that the density function must be non-negative and that its integral over the entire real line be unity. This implies that density functions for continuous random variables are in fact absolutely integrable. There are many distribution functions useful for analog signal theory, but the normal or Gaussian distribution is of paramount importance.

Definition (Normal Distribution). The random variable x is *normally* or *Gaussian* distributed if its probability density function is of the form

$$f_x(r) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(r-\mu)^2}{2\sigma^2}\right). \tag{1.117}$$

where μ and σ are the mean and standard deviation of the Gaussian (1.117), respectively.

Definition (Exponential Distribution). The random variable $\boldsymbol{x}$ has an *exponential distribution* with parameter $a > 0$ if its density function is of the form

$$f_x(r) = \begin{cases} a\exp(-ar) & \text{if } r > 0, \\ 0 & \text{if } r \le 0. \end{cases} \tag{1.118}$$

Definition (Gamma Distribution). The random variable x has a *gamma distribution* with *scale* parameter $a > 0$ and *shape* parameter $b > 0$ if its density function is of the form

$$f_x(r) = \begin{cases} \dfrac{a^b r^{b-1}\exp(-ar)}{\Gamma(b)} & \text{if } r > 0, \\ 0 & \text{if } r \le 0, \end{cases} \tag{1.119}$$

where $\Gamma(t)$ is the gamma function [80]:

$$\Gamma(t) = \int_0^\infty s^{t-1}\exp(-s)\, ds, \tag{1.120}$$

defined for $t > 0$.

Definition (Continuous Mean). If the random variable $\boldsymbol{x}$ is continuous and has density function $f_x(r)$ and if $xf_x(r)$ is in $L^1(\mathbb{R})$, then the *mean* or *expectation* of $\boldsymbol{x}$,

written $E[\boldsymbol{x}]$, is

$$E(x) = \int_{-\infty}^{\infty} r f_x(r)\, dr. \tag{1.121}$$

Definition (Continuous Variance, Standard Deviation). Suppose that the random variable $\boldsymbol{x}$ is continuous, $xf_x(r)$ is in $L^1(\mathbb{R})$, $\mu = E[\boldsymbol{x}]$, and $x^2 f_x(r)$ is in $L^1(\mathbb{R})$. Then the *variance* of $\boldsymbol{x}$, σ_x^2, is

$$\sigma_x^2 = \int_{-\infty}^{\infty} (r-\mu)^2 f_x(r)\, dr. \tag{1.122}$$

1.8.2.3 Multivariate Distributions. This section considers the description of random vectors, entities that consist of two or more random components. Much of the development follows from a direct, albeit somewhat messy, extension of the ideas from the single random variables.

Definition (Multivariate Distributions). Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be random variables in the probability space (Ω, Σ, P). Their joint distribution function is defined by $F_{\boldsymbol{x},y}(r, s) = P[(\boldsymbol{x} \le r) \cap (\boldsymbol{y} \le s)]$. This generalizes to an arbitrary finite number of random variables, $\boldsymbol{r} = (\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_M)$. For continuous random variables, the *joint density* of $\boldsymbol{x}$ and $\boldsymbol{y}$ is

$$f_{x,y}(r, s) = \frac{\partial^2}{\partial x \partial y} F_{x,y}(r, s). \tag{1.123}$$

We can define joint probability density functions for families of random variables too. This requires vector and matrix formulations in order to preserve the properties of density and distribution functions. For example, for the multivariate normal density, we have the following definition.

Definition (Joint Normal Density). Suppose that $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M)$ is a vector of M random variables on the probability space (Ω, Θ, P). We define the *joint normal density* function $f_{\boldsymbol{X}}(r)$ by

$$f(r) = \frac{\exp\left[-\frac{1}{2}(\boldsymbol{r}-\mu)^T \Sigma^{-1} (\boldsymbol{r}-\mu)\right]}{\sqrt{\det(\Sigma)(2\pi)^M}}, \tag{1.124}$$

where $\boldsymbol{r} = (\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_M)$ is a vector of length M; $\mu = (E[\boldsymbol{x}_1], E[\boldsymbol{x}_2], \ldots, E[\boldsymbol{x}_M])$ is the vector of means; $(\boldsymbol{r}-\mu)^T$ is the transpose of $\boldsymbol{r}-\mu$; Σ is the $M \times M$ covariance matrix for $\boldsymbol{X}$, $\Sigma = [\sigma_{m,n}] = E[(x_m - \mu_m)(x_n - \mu_n)]$; $\det(\Sigma)$ is its determinant; and its inverse is Σ^{-1}.

Earlier we considered how to apply Bayes's theorem to the problem of signal classification. However, we noted that it is not easy to distinguish signals by one feature alone, and our one-dimensional statistical classification breaks down. Now let's consider how to use statistical information about feature vectors and classes of signals to develop statistical discriminant functions. Suppose that we know the *a priori* probability of occurrence of each of the classes C_k, $P(C_k)$. Suppose further that for each class C_k we know the probability density function for the feature vector $\boldsymbol{v}$, $p(\boldsymbol{v}|C_k)$. The conditional probability $P(\boldsymbol{v}|C_k)$ provides the likelihood that class k is present, given that the input signal has feature vector $\boldsymbol{v}$. If we could compute $P(C_k|\boldsymbol{v})$ for each C_k and $\boldsymbol{v}$, then this would constitute a statistical basis for selecting one class over another for categorizing the input signal f.

We can restate Bayes's theorem for the multivariate case as follows.

Theorem (Multivariate Bayes). Suppose that for K signal classes C_k we know the *a priori* probability of occurrence of each of the classes $P(C_k)$ and the probability density function for the feature vector $\boldsymbol{v}$, $P(\boldsymbol{v}|C_k)$. Then

$$P(C_k \mid \boldsymbol{v}) = \frac{p(\boldsymbol{v} \mid C_k)P(C_k)}{p(\boldsymbol{v})} = \frac{p(\boldsymbol{v} \mid C_k)P(C_k)}{\sum_{i=1}^{K} p(\boldsymbol{v} \mid C_i)P(C_i)}, \tag{1.125}$$

where $p(\boldsymbol{v})$ is the probability density function for feature vector $\boldsymbol{v}$.

1.8.3 Random Signals

The ideas of random variables, their distribution and density functions, and the principal parameters that describe them are the basis for a definition of a random signal.

When we say that a signal is random, that is not to say that we know nothing of its values; in fact, we might know that the value of the signal at a time instant is almost certain to be in a given range. We might know that the signal remains, at other times, in some other range. It should be possible to provide a table that specifies the possible ranges of the signal and furnishes rough measures for how likely the signal value is to fall within that range. Every time the signal is measured or evaluated at a time, the signal is different, but we have an approximate idea of how these values behave. We can find one set of signal values, one instance of the random signal, but the next instance will differ. Thus, our concept of a random signal is embodied by a family of signals, and each member of the family represents a possible measurement of the signal over its domain. In probabilty theory, this is known as a *random* or *stochastic process*.

Definition (Random Signal). Suppose that (Ω, Σ, P) is a probability space. Let $\boldsymbol{X} = \{x(r): t \in T\}$ be a family of random variables on (Ω, Σ, P) indexed by the set T. Then $\boldsymbol{X}$ is a *stochastic process* or *random signa*l. If the index set T is the integers, then we say that $\boldsymbol{X}$ is a *discrete* random signal. If T is the real numbers, then we call $\boldsymbol{X}$ an *analog* random signal.

1.9 SUMMARY

There are two distinct signal categories: those with a continuous independent variable and those with a discrete independent variable. Natural signals are generally analog, and they become discrete—or more precisely, digital—by means of a conversion apparatus or by the way in which they are collected. The electrocardiogram source is analog, and it could be displayed in such a mode on an oscilloscope, for example. But nowadays it is often digitized for computer analysis. Such digital signals have a finite range. They are the true objects of digital signal processing on computers, but they are awkward for theoretical development. The temperature of the earth is a continuous function of depth, and it could be continuously recorded on a strip chart. But since the changes in the geothermal signal are so slow, it is more practical to collect isolated measurements. It is therefore discrete from the beginning. We rely on mathematical model for signal theory: continuous time functions, defined on the real numbers, for analog signals, and discrete time functions, defined on the integers, for discrete signals.

There is also a notion of the units of the interval between numerical signal values; This is called the independent variable. It is often a time variable, measured in seconds, minutes, hours, and so on, and this is natural, because time of occurrence provides a strict, irreversible ordering of events. So often are signals based on time that we get imprecise and routinely speak in temporal terms of the independent variable. On occasion, the independent variable that defines earlier and later signal values is a distance measure. The geothermal signal has an independent variable typically measured in meters, or even kilometers, of depth into the earth. Despite the fact that the independent variable is a distance measure, we often casually refer to the list of the signal's values as its "time-domain" specification.

The dependent variable of the signal is generally a real value for analog and discrete signals, and it is an integral value for digital signals. These are the signal values, and we stipulate that they assume numerical values, so that we can apply mathematical methods to study them. So the terminology here follows that from the mathematical notions of the independent and dependent variable for a mathematical function. We reserve the idea of a sequence of symbols, which is sometimes called a signal in ordinary language, for our concept of a signal interpretation.

We are concerned mainly with signals that have a one-dimensional independent and dependent variables. It is possible for a signal's dependent measurement to depend on multiple independent measurements. Image processing performs conditioning operations on two-dimensional signals. Computer vision analyzes multidimensional signals and produces a structural description or interpretation of them. We distinguish between single channel and multiple channel signals. If a signal produces a one-dimensional value, then it is single channel. An example is the temperature versus depth measurement. Signals that generate multidimensional values are called multichannel signals. An example is the 12-lead ECG. Multichannel signals have multidimensional range values. They arise in many applications, but we confine our discussions primarily to single-channel signals and refer interested readers to the more specialized literature on sensor fusion.

1.9.1 Historical Notes

A popular introduction to signal theory with historical background is Ref. 84. One of the latest discoveries in signal analysis is wavelet theory—a relatively recent and exciting approach for time-scale analysis [85]. An overview of wavelets and related time-frequency transforms is Ref. 69.

Several practical inventions spurred the early development of signal processing and analysis. The telegraph, invented by the American portrait painter Samuel F. B. Morse (1791–1872) in 1835, transmitted messages comprised of a sequence of isolated pulse patterns, or symbols, standing for letters of the English alphabet. The symbols themselves were (and still are, for the Morse code has been amended slightly and standardized internationally) a finite sequence of short and long electrical pulses, called dots and dashes, respectively. Shorter symbols represent the more prevalent English letters. For example, single dot and single dash represent the most common English letters, E and T, respectively. Morse's signaling scheme is an essentially discrete coding, since there is no continuous transition between either the full symbols or the component dots and dashes. Moreover, as a means of communication it could be considered to be digital, since the code elements are finite in number. But it would eventually be supplanted by analog communication technologies—the telephone and voice radio—which relied on a continuously varying representation of natural language.

Alexander Graham Bell (1847–1922), the famed U.S. inventor and educator of the deaf, discovered the telephone in the course of his experiments, undertaken in the mid-1870s, to improve the telegraph. The telegraph carried a single signal on a single pair of conductors. Bell sought to multiplex several noninterfering telegraphic messages onto a single circuit. The economic advantages of Bell's Harmonic Telegraph would have been tremendous, but the results were modest. Instead, Bell happened upon a technique for continuously converting human voices into electrical current variations and accurately reproducing the voice sounds at a remote location. Bell patented the telephone less than a year later, in March 1876; verified the concept six months later in sustained conversations between Boston and Cambridgeport, Massachusetts; and built the first commerical telephone exhange, at New Haven, Connecticut in January 1878. Bell's patent application points to the analog nature of telephony as clearly distinguishing it from discrete telegraphy.

Wireless telegraphy—and eventually wireless telephony—were the fruit of persistent efforts by yet another scientific layperson, Guglielmo Marconi (1874–1937). The young Italian inventor was aware of both J. C. Maxwell's theory of electromagnetic waves[21]and H. R. Hertz's demonstration[22] of precisely this radiation with a spark coil transmitter and wire loop receiver. But Hertz's apparatus was too weak for practical use. Marconi's improvements—a telegraph key to control the firing of the spark gap, a long wire antenna and earth ground for greater signal strength, and

[21]Scottish physicist James Clerk Maxwell (1831–1879) announced his electromagnetic field theory to a skeptical scientific community in 1864.

[22]With a small spark coil, German physicist Heinrich Rudolf Hertz (1857–1894) generated the first electromagnetic waves at the University of Karlsruhe and verified Maxwell's theory.

the crystal detector for improved reception—enabled him to demonstrate radio telegraphy in 1895. Unheralded in his native Italy, Marconi took his technology to England, received a patent in 1896, formed his own company a year later, and received the Nobel prize in 1909. These techniques could only serve discrete modes of telecommunication, however. Analog communication awaited further improvements in radio communication, in particular the radio-frequency alternator.

These practical advances in analog technologies were complemented by the discovery of the richness within Jean-Baptiste Fourier's discovery, long past, that even signals containing discontinuities could be represented by sums of smoothly undulating sinusoids. Fourier developed his theory for the purpose of studying heat propagation. In particular, it remains a principal tool for solving the differential equations governing such phenomena as heat conduction, Fourier's original problem [1]. Thus, at the turn of the last century, the most important signal technologies and the most important signal theories revolved around analog methods.

Theory would not link the analog and discrete realms of signal processing until the early twentieth century, when Nyquist [2], Shannon [3], and Vladimir Kotelnikov[23] developed the sampling theory. Nyquist's original research article focused on telegraphy, and it established a first theoretical link between discrete and analog communication methods. In particular, he showed that a continuous domain signal containing but a limited variety of frequencies could be captured and regenerated with a discrete signal. But analog practice and analog theory ruled supreme, and Nyquist's contribution was largely overlooked. Only when Shannon proved that error-free digital communication—even in the presence of noise—was possible did the attention of scientists and engineers turn once more to discrete modes of communication. The contributions of Nyquist and Shannon did firmly establish signal theory as a distinct scientific and engineering discipline. Both analog and discrete signal theory were soundly fixed upon mathematical foundations and shared a link through the Shannon–Nyquist results.

One seemingly insurmountable problem remained. The frequency analysis of analog signals was possible using conventional analog instruments such as a frequency analyzer. But discrete signal frequencies could not be calculated fast enough to keep pace with the arrival of discrete values to a processing apparatus. Therefore, although mathematicians developed a considerable complement of tools for understanding discrete signals, engineers remained preoccupied with analog tools which could handle their signals in real time.

The discovery of the fast Fourier transform (FFT) by J. W. Cooley and J. W. Tukey in 1965 shattered the analog tradition in signal processing. By eliminating duplicate computations in the DFT, it became possible to produce the frequency spectrum of a signal with N data points in $N\log_2 N$ operations; real-time digital signal spectrum analysis became feasible [4–6].

[23]Vladimir A. Kotelnikov (1908–), a Russian communications engineer, independently discovered the sampling theorem in 1933. His work was largely unknown outside the Soviet Union.

1.9.2 Resources

A vast array of resources—commercial products, inexpensive shareware applications, and public-domain software—are available nowadays for studying signal theory. Researchers, university laboratories, private firms, and interested individuals have also made available signal and image processing data sets. Some of these have become standards for experimentation, algorithm development, and performance comparisons.

1.9.2.1 Signal Processing Tools. Commercial packages used for this chapter's examples include:

- *Matlab*, available from The MathWorks, Inc., 24 Prime Park Way, Natick, MA, 01760, USA [86].
- *Mathematica*, available from Wolfram Research, Inc., 100 Trade Center Drive, Champaign, IL, 61820, USA [87].

Public-domain packages include the following:

- *Wavelab*, which uses Matlab and includes several popular research data sets, available free of charge from Stanford University: `http://playfair.stanford.edu/~wavelab`.
- *Khoros*, available free of charge via anonymous ftp from the University of New Mexico: `ftp.eece.unm.edu` [88].

1.9.2.2 Data. There are many data sets available over the internet, including several smaller data archives, maintained by individual researchers, tapped for examples in this chapter.

Among the larger repositories are the following:

- Rice University, Houston, TX, in conjunction with the Institute of Electrical and Electronic Engineers (IEEE), supports the Signal Processing Information Base (SPIB): `http://spib.rice.edu/spib.html`. SPIB contains a variety of signal and image data sets, several of which found their way into the examples of this text.
- The University of California at Irvine, Irvine, CA supports a machine intelligence database.

Every effort has been made to use example data sets that are available to the reader. Readers should be able to find this chapter's signal data examples within the public domain. Figure 1.1 is on the web site of the Princeton Earth Physics Project (`http://www.gns.cri.nz/quaketrackers/curr/seismic_waves.htm`). The EEG signals of Figure 1.3 are from Krishna Nayak's Florida State University web

site, `http://www.scri.fsu.edu`. The aerial scenes of Figure 1.4 are from the Danish Center for Remote Sensing, `http://www.dcrs.dk`. The ECG signal of Figure 1.6 is from SPIB. The geothermal data of Figure 1.8 comes from the Appalachian Deep Core Hole project and is available at the ADCOH web site [36]. The auditory neuron pulse train data are from SPIB.

1.9.3 Looking Forward

Now that we have introduced the basic raw material, signals, we proceed in Chapters 2 and 3 to introduce the machinery, systems. The term "system" is a very broad term, but in signal theory it is used in a quite specific sense. A *system* is the mathematical entity that accomplishes signal processing; it takes a signal as input and produces a signal as output. A system is a function that operates on signals.

An understanding of signals requires ideas from basic mathematics, algebra, calculus, a dose of complex analysis, and some random variable theory. In contrast, a firm understanding of the ideas of systems—the mechanisms that convert one signal into another, signal processing in other words—depends upon ideas from advanced mathematical analysis. In particular, we must draw upon the concepts of functional analysis—especially Hilbert space theory—topics normally taught at the university graduate mathematics level. For practical-minded scientists and engineers, this seems ominous. But the good news is that this development is straightforward for discrete signals. Thus, in Chapter 2 we concentrate exclusively on discrete signal spaces, of which discrete Hilbert spaces are a special case.

To most of us, the mastery of analog signal processing theory comes less readily than a thorough understanding of discrete theory. Readers need to understand both developments, even though the analog theory is more mathematically involved. However, scientists, applied mathematicians, and engineers who are looking further toward modern mixed-domain signal processing methods need a good foundation in signal spaces and an advanced presentation of analog signal analysis. Chapter 3 presents the prerequisite background in continuous-domain signal spaces.

1.9.4 Guide to Problems

All of the chapters provide problems. They range in difficulty from simple exercises that recall basic ideas from the text to more complicated problems that extend and develop the chapter's material. Some of them are outlines of research projects that may involve several weeks of work. The student may need to make simplifying assumptions, discover constraints, and—quite likely—will not arrive at a once-and-for-all answer to the problems posed.

REFERENCES

1. J.-B. J. Fourier, *The Analytical Theory of Heat*, translated by A. Freeman, New York: Dover, 1955.
2. H. Nyquist, Certain topics in telegraph transmission theory, *Transactions of the AIEE*, vol. 47, pp. 617–644, 1928.
3. C. E. Shannon, A mathematical theory of communication, *Bell Systems Technical Journal*, vol. 27, pp. 379–423 and pp. 623–656, 1948.
4. J. W. Cooley and J. W. Tukey, An algorithm for the machine calculation of complex Fourier series," *Mathematics of Computation*, vol. 19, pp. 297–301, April 1965.
5. J. W. Cooley, P. A. Lewis, and P. D. Welch, Historical notes on the fast Fourier transform, *IEEE Transactions on Audio and Electroacoustics*, vol AU-15, pp. 76–79, June 1967.
6. J. W. Cooley, How the FFT gained acceptance, *IEEE SP Magazine*, pp. 10–13, January 1992.
7. H. Baher, *Analog and Digital Signal Processing*, New York: Wiley, 1990.
8. J. A. Cadzow and H. F. van Landingham, *Signals, Systems, and Transforms*, Englewood Cliffs, N J: Prentice-Hall, 1989.
9. L. B. Jackson, *Signals, Systems, and Transforms*, Reading, MA: Addison-Wesley, 1991.
10. A. V. Oppenheim, A. S. Willsky, and S. H. Nawab, *Signals and Systems*, 2nd ed., Englewood Cliffs, NJ: Prentice-Hall, 1989.
11. R. E. Ziemer, W. H. Tranter, and D. R. Fannin, *Signals and Systems*: *Continuous and Discrete*, New York: Macmillan, 1989.
12. L. B. Jackson, *Digital Filters and Signal Processing*, Boston: Kluwer Academic Publishers, 1989.
13. A. V. Oppenheim and R. W. Shafer, *Discrete-Time Signal Processing*, Englewood Cliffs, NJ: Prentice-Hall, 1989.
14. J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications*, 2nd ed., New York: Macmillan, 1992.
15. L. R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing*, Englewood Cliffs, NJ: Prentice-Hall, 1975.
16. K.-S. Lin, ed., *Digital Signal Processing Applications with the TMS320 Family*, vol. 1, Dallas, TX: Texas Instruments, 1989.
17. R. J. Simpson, *Digital Signal Processing Using the Motorola DSP Family*, Englewood Cliffs, NJ: Prentice-Hall, 1994.
18. Motorola, Inc., *DSP56000/DSP56001 Digital Signal Processor User's Manual*, Phoenix, AZ: Motorola Literature Distribution, 1990.
19. R. J. Higgins, *Digital Signal Processing in VLSI*, Englewood Cliffs, NJ: Prentice-Hall, 1990.
20. A. Mar, ed., *Digital Signal Processing Applications Using the ADSP-2100 Family*, Englewood Cliffs, NJ: Prentice-Hall, 1990.
21. V. K. Madisetti, *VLSI Digital Signal Processors: An Introduction to Rapid Prototyping and Design Synthesis*, Piscataway, NJ: IEEE Press, 1995.
22. B. A. Bolt, *Earthquakes and Geologic Discovery*, New York: Scientific American Library, 1990.

23. O. Kulhanek, *Anatomy of Seismograms*, Amsterdam: Elsevier, 1990.
24. J. F. Claerbout, *Fundamentals of Geophysical Data Processing: With Applications to Petroleum Prospecting*, Boston: Blackwell Scientific, 1985.
25. M. Akay, *Biomedical Signal Processing*, San Diego, CA: Academic Press, 1994.
26. J. C. Russ, *The Image Processing Handbook*, Boca Raton, FL: CRC Press, 1995.
27. A. Rosenfeld and A. C. Kak, *Digital Picture Processing*, vols. 1 and 2, Orlando, FL: Academic Press, 1982.
28. A. K. Jain, *Fundamentals of Digital Image Processing*, Englewood Cliffs, NJ: Prentice-Hall, 1989.
29. R. J. Schalkoff, *Digital Image Processing and Computer Vision*, New York: Wiley, 1989.
30. D. H. Ballard and C. M. Brown, *Computer Vision*, Englewood Cliffs, NJ: Prentice-Hall, 1982.
31. R. M. Haralick and L. G. Shapiro, *Computer and Robot Vision*, vols. 1 and 2, New York: Addison-Wesley, 1992.
32. B. K. P. Horn, *Robot Vision*, Cambridge, MA: MIT Press, 1986.
33. M. J. Goldman, *Principles of Clinical Electrocardiography*, Los Altos, CA: Lange Medical Publications, 1986.
34. D. B. Geselowitz, On the theory of the electrocardiogram, *Proceedings of the IEEE*, vol. 77, no. 6, pp. 857–872, June 1989.
35. M. Unser and A. Aldroubi, A review of wavelets in biomedical applications, *Proceedings of the IEEE*, vol. 84, no. 4, pp. 626–638, April 1996.
36. J. K. Costain and E. R. Decker, Heat flow at the proposed Appalachian Ultradeep Core Hole (ADCOH) site: Tectonic implications, *Geophysical Research Letters*, vol. 14, no. 3, pp. 252–255, 1987.
37. C .F. Gerald and P. O. Wheatley, *Applied Numerical Analysis*, Reading, MA: Addison-Wesley, 1990.
38. F. Attneave, Some informational aspects of visual perception, *Psychological Review*, vol. 61, pp. 183–193, 1954.
39. D. Marr, *Vision*, New York: W. H. Freeman and Company, 1982.
40. H. Asada and M. Brady, The curvature primal sketch, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 1, pp. 2–14, January 1986.
41. I. Biedermann, Human image understanding: Recent research and a theory," *Computer Vision, Graphics, and Image Processing*, vol. 32, pp. 29–73, 1985.
42. A. P. Witkin, Scale-space filtering, *Proceedings of the 8th International Joint Conference on Artificial Intelligence*, Karlsruhe, W. Germany, 1983. See also A. P. Witkin, Scale-space filtering, in *From Pixels to Predicates*, A. P. Pentland, ed., Norwood, NJ: Ablex, 1986.
43. T. Lindeberg, Scale space for discrete signals, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 3, pp. 234–254, March 1990.
44. M. Rosenlicht, *Introduction to Analysis*, New York: Dover, 1978.
45. D. Gabor, Theory of communication, *Journal of the Institute of Electrical Engineers*, vol. 93, pp. 429–457, 1946.
46. D. A. Pollen and S. F. Ronner, Visual cortical neurons as localized spatial frequency filters, *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-13, no. 5, pp. 907–916, September–October 1983.

47. J. J. Kulikowski, S. Marcelja, and P. O. Bishop, Theory of spatial position and spatial frequency relations in the receptive fields of simple cells in the visual cortex, *Biological Cybernetics*, vol. 43, pp. 187–198, 1982.
48. A. H. Zemanian, *Distribution Theory and Transform Analysis*, New York: Dover, 1965.
49. M. J. Lighthill, *Introduction to Fourier Analysis and Generalized Functions*, New York: Cambridge University Press, 1958.
50. J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, 2nd ed., New York: Springer-Verlag, 1993.
51. S. Lang, *Linear Algebra*, Reading, MA: Addison-Wesley, 1968.
52. G. Strang, *Linear Algebra and Its Applications*, 2nd ed., New York: Academic Press, 1980.
53. S. G. Mallat, A theory for multiresolution signal decomposition: The wavelet representation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, July 1989.
54. J. Carson and T. Fry, Variable frequency electric circuit theory with applications to the theory of frequency modulation, *Bell Systems Technical Journal*, vol. 16, pp. 513–540, 1937.
55. B. Van der Pol, The fundamental principles of frequency modulation, *Proceedings of the IEE*, vol. 93, pp. 153–158, 1946.
56. J. Shekel, Instantaneous frequency, *Proceedings of the IRE*, vol. 41, p. 548, 1953.
57. F. Jelinek, Continuous speech recognition by statistical methods, *Proceedings of the IEEE*, vol. 64, no. 4, pp. 532–556, April 1976.
58. S. Young, A review of large-vocabulary continuous-speech recognition, *IEEE Signal Processing Magazine*, vol. 13, no. 5, pp. 45–57, September 1996.
59. L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Englewood Cliffs, NJ: Prentice-Hall, 1993.
60. N. Morgan and H. Bourlard, Continuous speech recognition, *IEEE Signal Processing Magazine*, vol. 12, no. 3, pp. 25–42, May 1995.
61. B. Malmberg, *Phonetics*, New York: Dover, 1963.
62. J. L. Flanagan, *Speech Analysis, Synthesis, and Perception*, 2nd ed., New York: Springer-Verlag, 1972.
63. N. Deshmukh, R. J. Duncan, A. Ganapathiraju, and J. Picone, Benchmarking human performance for continuous speech recognition, *Proceedings of the Fourth International Conference on Spoken Language Processing*, Philadelphia, pp. SUP1-SUP10, October 1996.
64. J. R. Cox, Jr., F. M. Nolle, and R. M. Arthur, Digital analysis of electroencephalogram, the blood pressure wave, and the electrocardiogram, *Proceedings of the IEEE*, vol. 60, pp. 1137–1164, 1972.
65. L. Khadra, M. Matalgah, B. El-Asir, and S. Mawagdeh, Representation of ECG-late potentials in the time frequency plane, *Journal of Medical Engineering and Technology*, vol. 17, no. 6, pp. 228–231, 1993.
66. F. B. Tuteur, Wavelet transformations in signal detection, in *Wavelets: Time-Frequency Methods and Phase Space*, J. M. Combes, A. Grossmann, and P. Tchamitchian, eds., 2nd ed., Berlin: Springer-Verlag, pp. 132–138, 1990.

67. G. Olmo and L. Lo Presti, Applications of the wavelet transform for seismic activity monitoring, in *Wavelets: Theory, Applications, and Applications*, C. K. Chui, L. Montefusco, and L. Puccio, eds., San Diego, CA: Academic Press, pp. 561–572, 1994.
68. J. L. Larsonneur and J. Morlet, Wavelets and seismic interpretation, in *Wavelets*: *Time-Frequency Methods and Phase Space*, J. M. Combes, A. Grossmann, and P. Tchamitchian, eds., 2nd ed., Berlin: Springer-Verlag, pp. 126–131, 1990.
69. Y. Meyer, *Wavelets: Algorithms and Applications*, Philadelphia: SIAM, 1993.
70. A. N. Kolmogorov and S. V. Fomin, *Introductory Real Analysis*, New York: Dover, 1975.
71. M. C. Teich, D. H. Johnson, A. R. Kumar, and R. Turcott, Fractional power law behavior of single units in the lower auditory system, *Hearing Research*, vol. 46, pp. 41–52, May 1990.
72. R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics*, vol. 1, Reading, MA: Addison-Wesley, 1977.
73. L. V. Ahlfors, *Complex Analysis*, 2nd ed., New York: McGraw-Hill, 1966.
74. E. Hille, *Analytic Function Theory*, vol. 1, Waltam, MA: Blaisdell, 1959.
75. N. Levinson and R. M. Redheffer, *Complex Variables*, San Francisco: Holden-Day, 1970.
76. R. Beals, *Advanced Mathematical Analysis*, New York: Springer-Verlag, 1987.
77. A. Gluchoff, A simple interpretation of the complex contour integral, *Teaching of Mathematic*s, pp. 641–644, August–September 1991.
78. A. O. Allen, *Probability, Statistics, and Queueing Theory with Computer Scinece Applications*, Boston: Academic, 1990.
79. W. Feller, *An Introduction to Probability Theory and Its Applications*, New York: Wiley, 1968.
80. E. Parzen, *Modern Probability Theory and Its Applications*, New York: Wiley, 1960.
81. A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, New York: McGraw-Hill, 1984.
82. R. E. Mortensen, *Random Signals and Systems*, New York: Wiley, 1984.
83. W. B. Davenport and W. L. Root, *An Introduction to the Theory of Random Signals and Noise*, New York: McGraw-Hill, 1958.
84. J. R. Pierce and A. M. Noll, *Signals: The Science of Telecommunications*, New York: Scientific American Library, 1990.
85. B. B. Hubbard, *The World According to Wavelets*, Wellesley, MA: A. K. Peters, 1996.
86. A. Biran and M. Breiner, *Matlab for Engineers*, Harlow, England: Addison-Wesley, 1995.
87. S. Wolfram, *The Mathematica Book*, 3rd ed., Cambridge, UK: Cambridge University Press, 1996.
88. K. Konstantinides and J. R. Rasure, The Khoros software development environment for image and signal processing, *IEEE Transactions on Image Processing*, vol. 3, no. 3, pp. 243–252, May 1994.

PROBLEMS

1. Which of the following signals are analog, discrete, or digital? Explain.

(a) The temperature reading on a mercury thermometer, as a function of height, attached to a rising weather balloon.

(b) The time interval, given by a mechanical clock, between arriving customers at a bank teller's window.

(c) The number of customers that have been serviced at a bank teller's window, as recorded at fifteen minutes intervals throughout the workday.

2. Which of the following constitute time domain, frequency domain, or scale domain descriptions of a signal? Explain.

(a) A listing of the percentages of 2-kHz, 4-kHz, 8-kHz, and 16-kHz tones in a ten second long tape recording of music.

(b) The atmospheric pressure readings reported from a weather balloon, as it rises above the earth.

(c) From a digital electrocardiogram, the number of QRS pulses that extend for 5, 10, 15, 20, 25, and 30 ms.

3. Sketch the following signals derived from the unit step $u(t)$:

(a) $u(t-1)$

(b) $u(t+2)$

(c) $u(-t)$

(d) $u(-t-1)$

(e) $u(-t+2)$

(f) $u(t-2) - u(t-8)$

4. Sketch the following signals derived from the discrete unit step $u(n)$:

(a) $u(n-4)$

(b) $u(n+3)$

(c) $u(-n)$

(d) $u(-n-3)$

(e) $u(-n+3)$

(f) $u(n-2) - u(n-8)$

(g) $u(n+6) - u(n-3)$

5. Describe the difference between the graphs of a signal $x(t)$; the shifted version of $x(t)$, $y(t) = x(t-c)$; and the reflected and shifted version, $z(t) = x(-t-c)$. Consider all cases for $c > 0$, $c < 0$, and $c = 0$.

6. Suppose that an N-bit register stores non-negative digital values ranging from 0 (all bits clear) to all bits set. The value of bit b_n is 2^n, $n = 0, 1, ..., N-1$. Show that the largest possible value is $2^N - 1$.

7. Consider the two's complement representation of a digital value in an N-bit register. If the bits are $b_{N-1}, b_{N-2}, ..., b_1, b_0$, then the digital value is $-b_{N-1}2^{N-1} + b_{N-2}2^{N-2} + \cdots + b_2 2^2 + b_1 2^1 + b_0 2^0$.

(a) Find the largest positive value and give its bit values.

(b) Find the most negative value and give its bit values.

(c) Show that the dynamic range is 2^N.

8. Suppose that an N-bit register uses the most significant bit b_{N-1} as a sign bit: If $b_{N-1} = 1$, then the value is -1 times the value in first $N-1$ bits; otherwise the value is positive, 1 times the value in first $N-1$ bits. The remaining $N-1$ bits store a value as in Problem 6.

(a) Again, find the largest possible positive value and the most negative value.

(b) What is the dynamic range for this type of digital storage register? Explain the result.

9. Suppose discrete signal $x(n)$ is known at distinct points $(n_k, x(n_k)) = (n_k, y_k)$, where $0 \le k \le N$. Suppose too that there are interpolating cubic polynomials over the $[n_k, n_{k+1}]$:

$$p(t) = a_k(t-n_k)^3 + b_k(t-n_k)^2 + c_k(t-n_k) + d_k. \tag{1.126}$$

(a) If the interpolants passes through the knots (n_k, y_k), then show $y_k = d_k$.

(b) Compute the derivatives, $p_k'(t)$ and $p_k''(t)$ for each k, and show that if $D_k = p_k''(n_k)$ and $E_k = p_k''(n_{k+1})$, then a_k and b_k can be written in terms of D_k and E_k.

(c) Suppose that for some k, we know both D_k and E_k. Show that we can then give the coefficients of the interpolating cubic, $p_k(t)$ on $[n_k, n_{k+1}]$.

10. Let $x(t) = 5\sin(2400t + 400)$, where t is a (real) time value in seconds. Give:

(a) The amplitude of x

(b) The phase of x

(c) The frequency of x in Hz (cycles/second)

(d) The frequency of x in radians/second

(e) The period of x

11. Consider a discrete signal $x(n) = A\cos(\Omega n + \phi)$ for which there is an $N>0$ with $x(n)=x(n+N)$ for all n.

(a) Explain why the smallest period for all discrete signals is $N=1$, but there is no such lowest possible period for the class of analog signals.

(b) Show that if $x(n)$ is a sinusoid, then the largest frequency it can have is $|\Omega| = \pi$ or, equivalently, $|F| = 1$, where $\Omega = 2\pi F$.

12. Let $s(n) = -8\cos\left(\frac{21}{2}n + 3\right)$ be a discrete signal. Find the following:

(a) The amplitude of s

(b) The phase of s

(c) The frequency of s in radians/sample

(d) The frequency of s in Hz (cycles/sample)

(e) Does s have a period? Why?

13. Find the frequency of the following discrete signals. Which ones are even, odd, finitely supported? Which ones are equal?

(a) $a(n) = 5\cos\left(n\frac{\pi}{4}\right)$.

(b) $b(n) = 5\cos\left(-n\frac{\pi}{4}\right)$.

(c) $c(n) = 5\sin\left(n\frac{\pi}{4}\right)$.

(d) $d(n) = 5\sin\left(-n\frac{\pi}{4}\right)$.

14. Prove that a signal decomposes into its even and odd parts. If $x(n)$ is a discrete signal, then show that:

(a) $x_e(n)$ is even.

(b) $x_o(n)$ is odd.

(c) $x(n) = x_e(n) + x_o(n)$.

15. Consider the signal $x(n) = [3, 2, 1, -1, -1, \underline{-1}, 0, 1, 2]$. Write $x(n)$ as a sum of even and odd discrete functions.

16. Show the following:

(a) $\sin(t)$ is odd.

(b) $\cos(t)$ is even.

(c) $g_{\mu,\sigma}(t)$ of mean μ and standard deviation σ (1.14) is symmetric about μ.

(d) Suppose a polynomial $x(t)$ is even; what can you say about $x(t)$? Explain.

(e) Suppose a polynomial $x(t)$ is odd; what can you say about $x(t)$? Explain.

(f) Show that the norm of the Gabor elementary function $\left\|G_{\mu,\sigma,\omega}(t)\right\|$ (1.20) is even.

(g) Characterize the real and imaginary parts of $G_{\mu,\sigma,\omega}(t)$ as even or odd.

17. Show that rational signal $x(t) = 1/t$ is neither integrable nor square-integrable in the positive real half-line $\{t : t > 0\}$. Show that $s(t) = t^{-2}$, however, is integrable for $\{t : t > 1\}$.

18. Show that $f(z) = f(x + jy) = x - jy$, the complex conjugate function, is not differentiable at a general point $z \in \mathbb{C}$.

19. Suppose that $f(z) = z$ and C is the straight line arc from a point u to point v in the complex plane.

(a) Find the contour integral

$$\oint_C f(z)\, dz. \tag{1.127}$$

(b) Suppose that $f(z) = z^{-1}$; again evaluate the contour integral in part (a); what assumptions must be made? Explain.

20. Suppose Σ is an algebra over a set Ω.

(a) Show that $\Omega \in \Sigma$.

(b) Show that Σ is closed under finite unions; that is, show that a finite union of elements of Σ is still in Σ.

(c) Show that Σ is closed under finite intersections.

(d) Supposing that Σ is a σ-algebra as well and that $S_n \in \Sigma$ for all natural numbers $n \in \mathbb{N}$, show that

$$\bigcap_{n \in \mathbb{N}} S_n \in \Sigma. \tag{1.128}$$

21. Suppose that (Ω, Σ, P) is a probability space. Let S and T be events in Σ. Show the following:

(a) $P(\varnothing) = 0$.

(b) $P(S) = 1 - P(S')$, where S' is the complement of S inside Ω.

(c) If $S \subseteq T$, then $P(S) \leq P(T)$.

(d) $P(S \cup T) = P(S) + P(T) - P(S \cap T)$.

22. Suppose that Ω is a set.

(a) What is the smallest algebra over Ω?

(b) What is the largest algebra over Ω?

(c) Find an example set Ω and an algebra Σ over Ω that is not a σ-algebra;

(d) Suppose that every algebra over Ω is also a σ-algebra. What can you say about Ω? Explain.

23. If A and B are independent, show that

(a) A and $\sim B$ are independent.

(b) $\sim A$ and $\sim B$ are independent.

24. Let $\boldsymbol{x}$ be a random variable and let r and s be real numbers. Then, by the definition of a random variable, the set $\{\omega \in \Omega : \boldsymbol{x}(\omega) \leq r\}$ is an event. Provide definitions for the following and show that they must be events:

(a) $\boldsymbol{x} > r$.

(b) $r < \boldsymbol{x} \leq s$.

(c) $\boldsymbol{x} = r$.

25. Find constants A, B, C, and D so that the following are probability density functions:

(a) $x(n) = A \times [\underline{4}, 3, 2, 1]$.

(b) $f(r) = B[u(r) - u(r-2)]$, where $u(t)$ is the analog unit step signal.

(c) The Rayleigh density function is

$$f(r) = \begin{cases} Cr\exp\left(-\dfrac{r^2}{2}\right) & \text{if } r \geq 0, \\ 0 & \text{if } r < 0. \end{cases} \tag{1.129}$$

(d) $f(r)$ is defined as follows:

$$f(r) = \begin{cases} \dfrac{D}{\sqrt{1-r^2}} & \text{if } |r| < 1, \\ 0 & \text{if } r \geq 1. \end{cases} \tag{1.130}$$

The following problems are more involved and, in some cases, expand upon ideas in the text.

26. Let $x(t) = A\cos(\Omega t)$ and $y(t) = A\cos(\Phi t)$ be continuous-domain (analog) signals. Find conditions for A, B, Ω, and Φ so that the following statement is true, and then prove it: If $x(t) = y(t)$ for all t, then $A = B$ and $\Omega = \Phi$.

27. Explain the following statement: There is a unique discrete sinusoid $x(n)$ with radial frequency $|\omega| \leq \pi$.

28. The following steps show that the support of a signal $x(t)$ is compact if and only if its support is both closed and bounded [44, 70].

(a) Show that a convergent set of points in a closed set S converges to a point in S.

(b) Prove that a compact $S \subset \mathbb{R}$ is bounded.

(c) Show that a compact $S \subset \mathbb{R}$ has at least one cluster point; that is, there is a t in S such that any open interval (a, b) containing t contains infinitely many points of S.

(d) Using (a) and (b), show that a compact set is closed.

(e) If $r > 0$ and $S \subset \mathbb{R}$ is bounded, show S is contained in the union of a finite number of closed intervals of length r.

(f) Show that if $S \subset \mathbb{R}$ is closed and bounded, then S is compact.

29. The *average power* of the discrete signal $x(n)$ is defined by

$$P_x = \lim_{N \to \infty} \left[\frac{1}{2N+1} \sum_{n=-n}^{n} |x(n)|^2 \right]. \tag{1.131}$$

If the limit defining P_x exists, then we say that $x(n)$ has *finite average power.* Show the following.

(a) An exponential signal $x(n) = Ae^{j\omega n}$, where A is real and nonzero, has finite average power, but not finite energy.

(b) If $x(n)$ is periodic and $x(n)$ is non-zero, then $x(n)$ is neither absolutely summable nor square summable.

(c) If $x(n)$ is periodic, then $x(n)$ has finite average power.

30. Show that under any of the following conditions, the differentiable function $f(z)$ must be constant on $\mathbb{C}$.

(a) Real$[f(z)]$ is constant.

(b) Imag$[f(z)]$ is constant.

(c) $|f(z)|$ is constant.

31. Show that a discrete polynomial $p(k)$ may have consecutive points, k_0, k_1, ..., and so on, where the discrete second derivative is zero.

(a) For a given degree n, what is the limit, if any, on the number of consecutive points where the discrete second derivative is zero? Explain.

(b) For a discrete polynomial $p(k)$ of degree n, find formulas for the first, second, and third derivatives of $p(k)$.

(c) Show that a polynomial $p(t)$ of degree $n > 1$ has only isolated points, t_0, t_1, ..., t_N, where the second derivative is zero. What is N?

32. Prove the proposition on distribution function properties of Section 1.8.2 [81].

33. Suppose the discrete random variable $\boldsymbol{x}$ has a binomial distribution (1.113).

(a) Find the density function $f_{\boldsymbol{x}}(r)$.

(b) Find the distribution function $F_{\boldsymbol{x}}(r)$.

(c) Find the mean $E[\boldsymbol{x}]$.

(d) Find the variance $(\sigma_{\boldsymbol{x}})^2$.

(e) Discuss the case where p or q is zero in (1.113).

34. Suppose the discrete random variable **x** has a Poisson distribution with parameter $a > 0$ (1.114).

(a) Find the density function $f_{\boldsymbol{x}}(r)$.

(b) Find the distribution function $F_{\boldsymbol{x}}(r)$.

(c) Find the mean $E[\boldsymbol{x}]$.

(d) Find the variance $(\sigma_{\boldsymbol{x}})^2$.

The next several problems consider electrocardiogram processing and analysis.

35. Develop algorithms for calculating the running heart rate from a single ECG lead.

(a) Obtain the ECG trace of Figure 1.6 from the Signal Processing Information Base (see Section 1.9.2.2). Plot the data set using a standard graphing package or spreadsheet application. For example, in Matlab, execute the command lines: `load ecg.txt; plot (ecg)`. As an alternative, develop C or C++ code to load the file, plot the signal, and print out the time-domain values. Identify the QRS complexes and give a threshold value M which allows you to separate QRS pulses from noise and other cardiac events.

(b) Give an algorithm that reads the data sequentially; identifies the beginning of a QRS complex using the threshold M from (a); identifies the end of the QRS pulse; and finds the maximum value over the QRS event just determined.

(c) Suppose two successive QRS pulse maxima are located at n_0 and n_1, where $n_1 > n_0$. Let the sampling interval be T seconds. Find the elapsed time (seconds) between the two maxima, $v(n_1)$ and $v(n_0)$. Give a formula for the heart rate from this single interval; let us call this value $H(n_1)$.

(d) Critique the algorithm for instantaneous heart rate above. Explain any assumptions you have made in the algorithm. Calculate the instantaneous heart rate $H(n_i)$, for all successive pairs of QRS pulses beginning at n_{i-1}. Plot this $H(n_i)$ value over the entire span of the signal. What do you observe? What if the threshold you choose in (a) is different? How does this affect your running heart rate value?

(e) Suppose that the running heart rate is computed as the average of the last several $H(n_i)$ values—for example, $H_3(n_i) = [H(n_i) + H(n_{i-1}) + H(n_{i-2})]/3$. Is the instantaneous heart rate readout better? Is there a practical limit to how many past values you should average?

36. Explore the usefulness of signal averaging when computing the instantaneous heart rate.

(a) Use a symmetric moving average filter on the raw ECG trace:

$$w(n) = \frac{1}{3}[v(n-1) + v(n) + v(n+1)]\,. \tag{1.132}$$

Calculate the running heart rates as in the previous problem using $w(n)$ instead of $v(n)$.

(b) How does an asymmetric smoothing filter,

$$w(n) = \frac{1}{3}[v(n) + v(n-1) + v(n-2)]\,, \tag{1.133}$$

affect the results? Explain.

(c) Sketch an application scenario which might require an asymmetric filter.

(d) Try symmetric moving average filters of widths five and seven for the task of part (a). Graph the resulting ECG traces. Are the results improved? Is the appearance of the signal markedly different?

(e) Why do signal analysts use symmetric filters with an odd number of terms?

(f) When smoothing a signal, such as the ECG trace $v(n)$, would it be useful to weight the signal values according to how close they are to the most recently acquired datum? Contemplate filters of the form

$$w(n) = \frac{1}{4}v(n-1) + \frac{1}{2}v(n) + \frac{1}{4}v(n+1)\,, \tag{1.134}$$

and discuss their practicality.

(g) Why do we choose the weighting coefficients in the equation of part (f) to have unit sum? Explain.

(h) Finally, consider weighted filter coefficients for asymmetric filters. How might these be chosen, and what is the motivation for so doing? Provide examples and explain them.

37. Develop algorithms for alerting medical personnel to the presence of cardiac dysrhythmia. A statistical measure of the variability of numerical data is the standard deviation,

(a) What is the average heart rate over the entire span of the ECG trace, once again using the distance between QRS pulse peaks as the basis for computing the instantaneous heart rate.

(b) Calculate the standard deviation of time intervals between QRS pulses. How many pulses are necessary for meaningful dysrhythmia computations?

38. Find algorithms for detecting the presence of the P wave and T wave in an ECG. One approach is to again identify QRS pulses and then locate the P wave pulse prior to the detected QRS complex and the T wave pulse subsequent to the detected QRS complex.

(a) Find the presence of a QRS pulse using a threshold method as before. That is, a QRS pulse is indicated by signal values above a threshold T_q.

(b) However, to locate P and T waves adjacent to the QRS complex, we must develop an algorithm for finding the time domain extent, that is the *scale*, of QRS pulses in the ECG trace. Develop an algorithm that segments the signal into the QRS pulse regions and non-QRS pulse regions. How do you handle the problem of noise that might split a QRS region? Is the method robust to extremely jagged QRS pulses—that is, *splintering* of the QRS complex?

(c) Show how a second, smaller threshold T_p can be used to find the P wave prior to the QRS complex. Similarly, a third threshold T_t can be used to find the T wave after the falling edge of the QRS complex.

(d) Should the thresholds T_p and T_t be global constants, or should they be chosen according to the signal levels of the analog and discrete signal acquisition procedures? Explain.